

Embryonic origins of epigenetic mosaicism in the human colorectal epithelium

Young Seok Ju

ysju@kaist.ac.kr

Korea Advanced Institute of Science and Technology (KAIST) <https://orcid.org/0000-0002-5514-4189>

Hyein Won

Korea Advanced Institute of Science and Technology

Yunah Lee

Korea Advanced Institute of Science and Technology <https://orcid.org/0000-0003-1691-0474>

Chang Hyun Nam

Korea Advanced Institute of Science and Technology (KAIST) <https://orcid.org/0009-0006-5429-7585>

Beomki Lee

Korea Advanced Institute of Science and Technology <https://orcid.org/0000-0003-3463-2121>

Joonoh Lim

Inocras

Ji-Hyung Park

Korea Advanced Institute of Science and Technology (KAIST)

Jisong An

Korea Advanced Institute of Science and Technology (KAIST)

Soo A Oh

Inocras Inc.

Myungsuk Choi

Korea Advanced Institute of Science and Technology (KAIST)

Jeong Yeon Kim

Division of Rheumatology, Department of Internal Medicine, Seoul National University College of Medicine

Jung Woo Park

Korea Institute of Science and Technology Information

Junehawk Lee

<https://orcid.org/0000-0001-8908-0884>

Jeonghwan Youk

Seoul National University Hospital

Ji Won Park

Seoul National University Hospital <https://orcid.org/0000-0003-0046-8175>

Seung-Yong Jeong

Seoul National University Hospital

Hyun Woo Kwon

Korea University <https://orcid.org/0000-0002-0597-837X>

Min Jung Kim

Seoul National University Hospital

Biological Sciences - Article

Keywords:

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11351803/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. Young Seok Ju is a co-founder of Inocras Inc.. The other authors declare no competing interests.

Embryonic origins of epigenetic mosaicism in the human colorectal epithelium

Authors:

Hyein Won^{1*}, Yunah Lee^{1*}, Chang Hyun Nam¹, Beomki Lee¹, Joonoh Lim², Ji-Hyung Park¹, Jisong An¹, Soo A Oh², Myungsuk Choi¹, Jeong Yeon Kim², Jung Woo Park³, Junehawk Lee³, Jeonghwan Youk¹, Ji Won Park⁴, Seung-Yong Jeong⁴, Hyun Woo Kwon⁵, Min Jung Kim^{4,#} and Young Seok Ju^{1,2,#}

Affiliations:

1. Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
 2. Inocras Inc., San Diego, CA 92121, USA
 3. Korea Institute of Science and Technology Information (KISTI), Daejeon 34141, Republic of Korea
 4. Department of Surgery, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
 5. Department of Nuclear Medicine, Korea University Ansan Hospital, Republic of Korea
- * Co-first authors with equal contribution

Present addresses:

- Ji-Hyung Park: Dept. of Neurology, Yonsei University Severance Hospital, Seoul, Republic of Korea
- Jisong An: Dept. Genome Sciences, University of Washington, Seattle, WA, USA
- Jeonghwan Youk: Dept. of Internal Medicine, Seoul National University Hospital, Seoul 03080, Republic of Korea
- Seung-Yong Jeong: Sheikh Khalifa Specialty Hospital. CEO Office. P.O. Box 6355 Ras Al-Khaimah. United Arab Emirates

Correspondence to (#):

Min Jung Kim MD PhD

minjungkim@snuh.org

Associate Professor, Department of Surgery, Seoul National University College of Medicine, Seoul 03080, Republic of Korea

Young Seok Ju MD PhD

ysju@kaist.ac.kr

Professor, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea, and
Scientific co-founder, Inocras Inc., San Diego, CA 92121, USA

Abstract

Somatic cells can acquire epigenetic changes throughout life, generating epigenetic mosaicism among cells of the same type. However, the extent, timing and persistence of this somatic epimosaicism remain poorly understood. Here, by integrating whole-methylome and whole-genome sequencing of 406 single-cell-derived colorectal epithelial clones from 16 donors, we show that, by ~60 years of age, a somatic colorectal lineage accumulates ~1.2 million CpG methylation differences from the consensus epiallelic state. Remarkably, we find that the majority of these differences (~60%) are established during early embryogenesis and are subsequently stably propagated across somatic lineages throughout life, rather than representing age-associated CpG epimutations. These early CpG methylation differences frequently clustered and recurred at specific genomic regions, consistent with embryonic lineage-specific commitment to one of several possible epiallelic states rather than stochastic epimutation. We identified 14,402 such embryonic epiallelic commitment (**EEC**) regions, which were enriched at CpG islands and gene promoters, particularly those of homeobox genes and zinc-finger/KRAB-domain-containing genes, consistent with a role in early developmental regulation. Notably, EECs were often associated with differences in gene expression across the adult colorectal epithelium, suggesting that their functional consequences may persist into adulthood. EECs also accounted for >70% of LINE-1 promoter loss-of-methylation frequently observed in adult colorectal epithelium and enabled more accurate reconstruction of early developmental phylogenies than somatic mutations alone. This principle extended to the DNA methylation landscape underlying X-chromosome inactivation and escape in female cells. Together, our study reveals early embryogenesis as a major source of lifelong somatic epigenetic mosaicism and establishes clonal multi-omics as a high-resolution framework for tracing its origins, persistence and functional consequences.

Introduction

The adult human body contains ~36 trillion cells spanning more than 400 cell types, all descended from a single fertilized egg^{1,2}. Through successive rounds of cell proliferation combined with cell fate decision processes, this single cell gives rise to hundreds of specialized cell types and states³. With the exception of somatic mutations acquired during development and ageing⁴, and programmed genomic alterations such as antigen-receptor rearrangement in lymphocytes, these cells retain nearly identical DNA sequences. The cellular diversity in morphology, gene expression and function is established through epigenetic remodeling, including DNA methylation, histone modifications, chromatin accessibility, and lineage-defining transcription-factor occupancy and network⁵. Large-scale efforts, including ENCODE, the Roadmap Epigenomics Mapping Consortium, and Human Cell Atlas, have catalogued these epigenetic landscapes across tissues and cell types, providing a molecular framework for understanding differences in cellular identity⁶⁻⁸.

By contrast, much less is known about epigenomic heterogeneity among cells that share the same differentiated identity, a phenomenon that we refer to throughout this study as 'somatic epimosaicism'. Cells of the same type, such as colonic epithelial cells, are often implicitly assumed to carry broadly constant epigenomes. In principle, however, cells within a given cell type accumulate distinct histories of cell division, DNA damage and repair, environmental exposure and stochastic regulatory change in ageing. Age-associated alterations in DNA methylation underpin epigenetic clocks^{9,10}, although the biological processes that generate these patterns remain incompletely understood. In addition, somatic epimosaicism may also originate much earlier in life, as suggested from LINE-1 promoter methylation heterogeneity¹¹. Numerous cells within a given cell type, such as the millions of colorectal crypts in the adult colon, descend from distinct embryonic progenitors that undergo cell-fate programming independently (**Fig. 1a**). If this programming is imperfectly constrained, lineage-specific epigenetic differences may arise early in development, persist into adulthood and confer phenotypic differences among otherwise similar epithelial cells.

Despite extensive efforts to characterize genomic and epigenomic heterogeneity^{6,7}, the extent of epigenetic heterogeneity within a cell type and their acquisition timing remain unclear. Two technical limitations have hindered progress. First, single-cell epigenomic methods, including DNA methylation and ATAC sequencing^{12,13}, generate sparse and uneven coverage^{14,15}. Signals here were therefore mostly aggregated across clusters of cells, which focused on cell-type classification but masked subtle variation within a cell type. Second, most, if not all, current single-cell epigenome techniques do not reveal early

developmental lineages, making it difficult to extract variation established in specific progenitors during embryogenesis from that acquired through ageing, environmental exposure or technical noise.

To explore somatic epimosaicism among adult single cells and trace their acquisition in human lifetime, here we combined clonal expansion with multi-omics profiling in the human colorectal epithelium as introduced previously¹¹ (**Fig. 1b**). We collected multiple colorectal epithelial stem cells from each donor, expanded individual cells to ~100,000 cells using clonal organoid techniques under conditions that suppress *in vitro* differentiation, and subjected each clone to whole-genome, whole-methylome and whole-transcriptome sequencing. This approach enabled near-complete profiling of ~28 million CpG sites genome-wide in each clone, providing substantially greater sensitivity for detecting methylation differences than sparse single-cell assays. By analysing DNA methylation profiles from 406 colorectal epithelial clones across 16 donors and using somatic mutations to reconstruct developmental phylogenies, we inferred the timing of DNA methylation differences among adult colorectal epithelial cells. Integration with transcriptome sequencing further enabled us to assess the functional consequences of lineage-associated epigenetic variation. Together, these analyses reveal that early embryogenesis is a major source of epigenomic heterogeneity in the adult human colorectal epithelium.

Results

Multi-omics of clones reveal epigenomic mosaicism among adult single cells

In total, we established and analysed paired whole-genome and whole-methylome (Enzymatic Methyl-Seq, EM-Seq)¹⁶ sequences from 406 colorectal organoid clones derived from 16 donors (**Fig. 1b**; **Supplementary Table 1**), extending our previous work¹¹. All donors were patients with colorectal cancer in their 40s to 80s, and apparently non-tumour crypts were isolated from surgical specimens collected >5 cm from the tumour margin. Paired whole-transcriptome sequencing was also available for 374 clones. From the whole-genome data, we identified 1,296,700 somatically acquired single-nucleotide variants (**SNVs**) and indels. Their variant allele fraction (**VAF**) distributions confirmed that nearly all clones originated from single founder cells and that most mutations were acquired *in vivo* before clonal expansion, with minimal contribution from culture-associated artefacts (**Extended Data Fig. 1a**). The absence of known cancer driver mutations and copy number changes supported the conclusion that these clones were not tumour-associated. On

average, the colorectal clones accumulated ~50 mutations per year during aging¹¹ (**Extended Data Fig. 1b**).

Using a total of 33,029 shared somatic mutations by multiple clones as endogenous lineage barcodes, we reconstructed the early developmental phylogenies using methods described previously¹⁷ (**Extended Data Fig. 2,3**). The combined VAFs of mutations assigned to the earliest phylogenetic branches of endodermal cells (colorectal epithelium) approached 50% in matched bulk blood tissues (mesoderm), strongly supporting that these phylogenies captured lineages arising during the earliest, pre-gastrulation stages of human development, as previously discussed^{17,18} (**Extended Data Fig. 2,3**).

The human reference genome (GRCh37) contains 26,752,702 CpG dinucleotides in autosomes (**Fig. 1c**). Of these, 9.6% are included within CpG islands¹⁹, whereas 8.2%, 4.2%, and 78.0% are located within CpG shores (regions up to 2 kb from a CpG island), CpG shelves (located 2–4 kb from a CpG island), and open sea (more than 4kb from the nearest CpG island), respectively²⁰. CpG islands, which are often depleted of DNA methylation in somatic cells, frequently overlap gene promoters (**Fig. 1c**). Of the 26.8 million autosomal CpGs, ~24.9 million (93%) per clone were robustly profiled by EM-seq; the remaining ~1.9 million were excluded owing to donor-specific germline variants disrupting reference CpGs, poor mappability, or consistently insufficient coverage. As exemplified by a representative clone (HC01-09), methylation β values (the fraction of methylated reads at each CpG site) showed prominent peaks near 0 and 1, corresponding predominantly to biallelic-unmethylated (referred to as U/U) and biallelic-methylated (referred to as M/M) CpGs, respectively, with distinct distributions across CpG islands, shores, shelves and open-sea regions (i.e., mostly U/U for CpG islands, mostly M/M for open sea; **Fig. 1d**). A smaller subset of CpGs exhibited intermediate β values, which could be interpreted to be monoallelic-methylated (referred to as U/M).

To facilitate exploration of CpG methylation β values and expression levels of nearby genes across all clones analysed in this study, we developed a web-based methylation browser (<https://methylome.kaist.ac.kr>; **Extended Data Fig. 4**).

For subsequent analyses, we converted continuous β values of a CpG site into three discrete epigenotypes using Gaussian-Mixture Model (**Methods**; **Extended Data Fig. 5a-c**). Across the 406 clones, the overall distribution of CpG epigenotypes was highly consistent, with ~75% classified as M/M, 15% as U/U and 10% as U/M epigenotypes (**Fig. 1e**). For example, the promoter of *POU5F1* (known as OCT4), which is active during early

development but silenced in adult colonic stem cells, was M/M in all clones (Fig. 1f). By contrast, the promoter of the actively expressed housekeeping gene *ACTB* was consistently U/U epigenotypes (Fig. 1f). At the imprinted *H19* locus, which is expressed from the maternal allele in somatic cells, the regulatory region showed the expected U/M pattern (Fig. 1f).

Although our clone-based DNA methylation profiling was generally accurate, several factors could introduce epigenotyping errors. First, technical factors, including stochastic allelic sampling, incomplete enzymatic conversion during EM-seq and read misalignment, could lead to erroneous epigenotype calls. Second, epigenetic drift during clonal expansion may occur and confound epigenotype assignment despite the short culture period and conditions designed to suppress *in vitro* cell differentiation. To validate the technical reproducibility of our clone-based approach, we independently performed EM-seq twice on four clones as technical replicates. In addition, although most clone pairs diverged during early embryogenesis, 11 pairs descended from common ancestors that had accumulated more than 2,000 shared somatic mutations, indicating divergence after >40 years of age. Because these pairs shared most of their lifespan, we expected their DNA methylation profiles to be more similar than those of randomly selected clone pairs (biological replicates; Extended Data Fig. 2,3). As expected, technical and biological replicate pairs showed highly concordant β -value profiles, with 97.8% and 95.9% of randomly selected CpGs, respectively, differing by less than 0.25 in β value, compared with 86.1% in distinct within-donor clone pairs. (Fig. 1g). Concordance increased further when CpGs were analysed as local clusters rather than as individual sites. For example, the mean β values of CpG islands differed by less than 0.25 in 99.8% of technical replicates and 99.9% of biological replicates, compared with 96.5% of random within-donor clone pairs (Fig. 1g). Using discrete epigenotypes, concordance at individual CpGs and CpG islands was ~94% and ~98%, respectively, in both technical and biological replicates, substantially higher than in random within-donor clone pairs (84.3% for individual CpGs and 93.7% for CpG islands; Fig. 1h). Of note, most discrepancies within CpG islands in biological/technical replicates did not reflect technical errors in our approach, but rather arose from the conventional definition of CpG islands, which can encompass CpGs with heterogeneous epigenotypes (Extended Data Fig. 5d-f). Together, the substantially higher concordance observed in technical and biological replicates than in random clone pairs indicates that true epigenetic heterogeneity across clones exceeds the variability introduced by technical limitations, particularly when neighboring CpGs are analysed collectively.

After accounting for discordance between technical replicates as technical error, two random clones from the same 64-year-old donor differed at an average of ~2.4 million CpG epigenotypes, suggesting that each colorectal epithelial lineage had accumulated >1.2 million CpG epimutations over the course of life, corresponding to 4.9% of accessible CpGs.

Early developmental epiallelic commitment stably maintained in the adult colorectal epithelium

Intriguingly, a substantial fraction of DNA methylation epimutations was shared among descendants of the same developmental lineage, indicating that they were acquired in embryonic ancestral cells and stably maintained in descendant somatic cells into adulthood. For example, across 52 consecutive CpGs in an open-sea region (chr22:45,831,580–45,834,074), 18 clones descending from 5 embryonic ancestor nodes in HC50 shared a biallelic-to-monoallelic methylation shift (**Fig. 2a, left**). A second interval containing 58 CpGs within a CpG island (chr10:82,294,747–82,297,315) showed a similar pattern in multiple individuals including HC21 and HC50, with three possible methylation states strongly concordant with the early developmental phylogeny (**Fig. 2a, right**). Notably, these consecutive CpGs shared across multiple clones are unlikely to reflect technical noise, as independent epigenotyping errors would rarely recur across consecutive CpGs in multiple clones while conforming to the same phylogenetic pattern.

As an alternative scenario, these changes could have arisen independently in descendant cells of a common ancestor during later, postnatal life, that is, through identity by state rather than identity by descent, if local niches at different anatomical sites imposed distinct epigenetic states. However, clones sharing the same epigenotypes were distributed across multiple sampling sites, arguing against this possibility (**Fig. 2b**).

These representative examples provide three key insights into DNA methylation heterogeneity in the colorectal epithelium. First, somatic epimosaicism observed in adult cells can arise during early embryogenesis. Second, these early epigenetic states can persist throughout life, despite lineage specification of the colorectal epithelium and decades of postnatal stem-cell self-renewal. Third, their recurrence across multiple developmental lineages and donors suggests that they are not purely stochastic epimutations, but instead reflect a systematic process of embryonic epiallelic commitment (**EEC**).

To quantify EEC events and estimate their acquisition timing, we systematically integrated CpG methylation states with somatic mutation-based developmental phylogenies from 15 of

16 donors (except for HC14, which contained too few clones to provide sufficient statistical power; **Methods; Extended Data Fig. 6**). Briefly, at each CpG site, the most frequent epigenotype across clones was designated as the consensus methylation state. We then identified CpGs whose epiallelic commitment events were significantly associated with the phylogenetic structure than randomly expected, and assigned each inferred epiallelic commitment event to the corresponding developmental node (**Fig. 2c**). Of note, this analysis included only epigenotypes shared across multiple clones and excluded CpG changes confined to terminal branches, making it less sensitive to technical errors.

Further, the number of somatic mutations accumulated at each phylogenetic node provided an independent molecular clock for developmental and ageing time^{17,18}, accumulating at approximately one mutation per cell generation during early embryogenesis^{17,21}, reaching ~50 mutations per colorectal epithelial lineage by birth^{22,23}. The timing of an EEC event assigned to an early node can be inferred to fall between the molecular times of its parent node and the node to which it was assigned (**Fig. 2c**).

Our approach estimated that, by birth (corresponding to ~50 somatic mutations), a colorectal epithelial lineage had accumulated a median of 0.65 million EECs across the 15 donors (IQR, 0.62–0.76 million). In HC50, for which the largest number of colorectal clones was available and thus afforded the greatest temporal resolution, ~0.75 million EECs were inferred to have been established by birth (**Fig. 2d; Extended Data Fig. 7a**). The rate was markedly higher during the earliest developmental stages, reaching commitments of ~40,000 CpGs per somatic mutation before accumulation of the first ten mutations and declining towards a plateau by birth. By contrast, nine late-stage nodes that branched after accumulation of >2,000 somatic mutations (**Extended Data Fig. 2**) enabled us to estimate the rate of postnatal age-associated methylation change, 144 CpGs per endogenous mutation from HC50 (from 8 lineages) and 112 CpGs per endogenous mutation from HC21 (from 1 lineage; **Fig. 2d; Extended Data Fig. 7b**). These rates correspond to ~6,000 CpG methylation changes annually, and assuming a constant accumulation rate throughout life, a single colorectal epithelial lineage would acquire ~0.5 million postnatal epimutations by 80 years of age. Collectively, these findings indicate that EEC events are comparable in magnitude to, or more abundant than, those acquired over a lifetime of postnatal ageing (**Fig. 2d**).

We further examined EEC events according to CpG context and the inferred direction of methylation change, using the consensus epigenotype as the reference state (**Fig. 2e**). At face value, loss-of-methylation events from biallelically methylated CpGs were approximately

2.5-fold more frequent than gain-of-methylation events from biallelically unmethylated CpGs, particularly in open-sea regions. The greater absolute number of loss-of-methylation events in open-sea regions largely reflects the greater abundance of methylated CpGs in this genomic context. In contrast, CpG islands, which are predominantly unmethylated, showed a relatively greater contribution of gain-of-methylation events, whereas shores and shelves showed more comparable contributions from the two directions. Notably, these inferred directions are defined relative to the consensus methylation state of each CpG in adult colorectal epithelium and therefore may not necessarily represent the actual direction of methylation change during development *in vivo*.

Embryonic epiallelic commitment events frequently cluster

As briefly mentioned above, the observed EEC CpGs were markedly clustered together (Fig. 2a), with stretches of five or more consecutive CpGs occurring substantially more frequently than expected (Extended Data Fig. 7c). CpGs in an EEC cluster were typically separated by only approximately 10 bp, substantially shorter than the background median CpG distance (40 bp; $P = 0.0099$), suggesting the preferential formation of EEC in CpG-dense regions (Extended Data Fig. 7d).

Defining an EEC segment as a stretch of five or more consecutive CpGs showing epimutations in the same direction, we identified 1,900–6,500 EEC segments per donor (Fig. 2f). Of note, donor HC50 yielded the largest number of segments, consistent with the greater sensitivity afforded by the larger number of clones in the phylogeny ($n=110$). An EEC segment contained a median of 14 CpGs, with the largest comprising 298 CpGs.

Merging overlapping segments across lineages and donors yielded 14,402 non-redundant autosomal EEC regions. Of these, 6,569 (45.6%) and 9,648 (67.0%) were recurrent across multiple donors and multiple lineages, respectively, suggesting these changes are not purely stochastic events (Figs. 2g,h). A single EEC region was often composed of multiple segments with distinct homogeneous epigenotype states (Fig. 2i). We defined each contiguous stretch of CpGs sharing the same epigenotype state as an EEC element, identifying 22,550 EEC elements in total. Of the 14,402 EEC regions, 5,123 (35.6%) contained multiple EEC elements (Supplementary Tables 2, 3).

Recurrent EEC segments within an EEC element predominantly mapped to nearly identical genomic positions across donors, showing ~82% reciprocal CpG overlap (Figs. 2i,j). Their boundaries frequently aligned with CpG island boundaries (Fig. 2k), consistent with preferential localization to CpG-rich regions, particularly CpG islands and shores (Fig. 2l).

EEC regions were also enriched ~2-fold near imprinting control regions (ICRs) and 1.30-fold in H3K27me3-marked Polycomb-repressed chromatin, but depleted from H3K27ac-marked active regulatory regions (Fig. 2m; Supplementary Table 4).

Functional impacts of embryonic epiallelic commitment regions

DNA methylation at gene promoters can influence gene transcription and thereby have functional consequences²⁴. Of 15,873 promoters regulating 20,953 genes in autosomes²⁵, 3,328 (21%) overlapped EEC regions, encompassing 4,472 genes (Fig. 3a; Supplementary Table 5; Methods). Functional enrichment analysis of these genes revealed significant enrichment of several gene families, including homeobox genes, zinc finger/KRAB domain-containing genes, non-coding RNAs (ncRNAs), and H2B histone genes (Fig. 3b; Supplementary Table 6). Homeobox genes are key developmental regulators that establish embryonic body patterning and specify cell and tissue identities during early development²⁶. KRAB zinc-finger genes are key epigenetic regulators that repress transposable elements and shape gene-regulatory and chromatin states during early development²⁷. These findings suggest that EEC events preferentially affect regulatory gene networks active during early development.

We then assessed whether EEC events also have functional consequences in adult somatic cells. Of the 4,472 genes associated with EEC elements, 2,341 showed substantial gene expression and promoter methylation variance among adult clones (Methods). Correlating promoter methylation with gene expression revealed a significant negative association for 725 genes (31%), including *MAP3K7CL*, consistent with the predominantly repressive effect of promoter methylation (linear regression, FDR < 0.05; Figs. 3c,d; Supplementary Table 7). By contrast, 58 genes (2%), including *HOXD8*, showed a significant positive association, consistent with previous observations that methylation within promoter regions can positively correlate with transcriptional activity in specific regulatory contexts²⁸ (Fig. 3c,e). Together, these findings demonstrate that EEC regions acquired during early development can exert persistent transcriptional effects that remain detectable in adult somatic cells.

Notably, several EEC regions overlapped cancer-associated genes and were associated with epigenotype-dependent differences in gene expression. At the *MGMT* locus, for example, promoter methylation within a developmental lineage was accompanied by reduced *MGMT* expression in late-stage somatic cells (Fig. 3f). *MGMT* encodes a DNA repair protein that removes alkylated DNA lesions, and its epigenetic silencing has been implicated in tumorigenesis^{29,30}. These findings suggest that persistent developmental

epimutations can alter the expression of genes involved in cancer-related pathways and may thereby contribute to lineage-specific differences in cancer susceptibility.

Across the 406 clones, using a cutoff of at least 10 consecutive altered CpGs, each clone harbored an average of 493 DNA methylation segment changes relative to the consensus epiallele state (range, 265–864). Of these, 477 (96.8%) overlapped EEC regions, indicating that the vast majority of segmental methylation changes in the adult colorectal epithelium are of embryonic origin (**Fig. 3g**).

By contrast, only a minor fraction, an average of 16 segmental changes per clone, was confined to terminal branches and likely accumulated predominantly during postnatal ageing. Across all clones from 15 donors (except for HC14), we identified 2,691 such later-acquired epigenetic changes (**LECs; Supplementary Table 8**; 20 representative EEC and LEC regions are shown in **Related Manuscript File 1**). In contrast to EEC elements, genes whose promoters overlapped LECs showed no significant enrichment for specific gene families, consistent with more stochastic accumulation in the absence of strong selective pressure (**Fig. 3h; Supplementary Table 6**). Intriguingly, promoter 1A of *APC*, a well-established tumour suppressor in colorectal tumorigenesis³¹, was independently affected by LECs in five clones from four donors (**Fig. 3i**). Clones with promoter 1A hypermethylation showed reduced relative expression of the *APC* 1A isoform transcript (**Fig. 3i**). Promoter hypermethylation-mediated silencing of *APC*, particularly reduced transcription from promoter 1A, has been implicated in Wnt pathway activation and colorectal tumorigenesis³¹, suggesting potential functional relevance of the methylation-associated transcriptional changes observed here.

Molecular timing of X-chromosome inactivation in females

We next examined DNA methylation across the X chromosome. Excluding the pseudoautosomal regions (**PARs**), which are shared by the X and Y chromosomes, females carry two X-chromosome copies, whereas males are hemizygous. In non-PAR X chromosome regions, male CpGs in single cells are expected to be either hemizygous-X-allele methylated ($\beta = 1.0$; referred to as M/-) or unmethylated ($\beta = 0$; referred to as U/-; **Figs. 4a,b**). In females, transcriptional dosage compensation is primarily achieved through epigenetic inactivation of one X chromosome (known as X-chromosome inactivation, **XCI**), with DNA methylation contributing to the maintenance of transcriptional silencing^{32,33}. Consistent with this, CpGs in the non-PAR X chromosome predominantly exhibited one of two methylation states: (1) full methylation in open-sea regions, with $\beta \approx 1.0$ in both females (M/M) and males (M/-); or (2) single-allele unmethylation at CpG islands, with $\beta \approx 0.5$ in

females (U/M) and $\beta \approx 0$ in males (U/-; **Figs. 4b,c**). CpG-island methylation, with most islands overlapping promoters, predominantly followed the latter pattern (**Fig. 4d**). In females, the unmethylated allele at U/M CpG sites consistently phased to the same homologous X chromosome and showed transcription by RNA sequencing, consistent with promoter demethylation and active gene expression from the active X chromosome (**Xa**) under XCI (**Figs. 4d,e**).

By integrating active and inactive X-chromosome assignments with developmental phylogenies in females (**Xa** and **Xi**; **Fig. 4e**), we inferred the timing of XCI during early embryogenesis. In all 10 female donors, the choice of **Xi** appeared to be fixed before the lineage had accumulated 18 embryonic mutations (**Fig. 4f**). Probability-weighted inference across phylogenetically compatible XCI scenarios further localized the timing of X-inactivation, with the overall timing distribution starting not earlier than ~5 mutations of molecular time, peaking at ~15 somatic mutations (50% highest-density interval, 10.2–18.8 mutations; **Fig. 4g**). Early mutations in the molecular time were frequently shared in matched peripheral blood, albeit at low VAFs, indicating that ancestral cells with established **Xi** choice were present before the completion of gastrulation (**Fig. 4f**). This timing, inferred retrospectively from adult somatic cells, is consistent with previous experimental observations from *in vitro* cultures of developing human embryos³⁴.

Embryonic epiallelic commitment contributes to escape from X-chromosome inactivation

Despite the predominantly dosage-concordant DNA methylation patterns observed between males and females, a substantial proportion of CpGs exhibited alternative, sex-discordant configurations (20.2%; **Fig. 4c**). Among these, regions exhibiting U/U epigenotype in females (an additional unmethylated allele relative to males) were strongly concentrated within CpG islands, but sharply transitioned to M/M epigenotype outside the islands ($\beta \approx 1$; **Fig. 4h**; **Extended Data Fig 8a**). Of the 750 CpG islands on the non-PAR X chromosome, 476 (62.7%) exhibited a U/U epigenotype in at least one of the 278 female clones, overlapping 315 promoters associated with 391 X-chromosome genes (**Supplementary Table 9**).

Using the frequency of demethylation across clones and the uniformity of CpG methylation within each demethylated island, we classified the 476 U/U CpG islands into four distinct categories (**Figs. 4d,i**; **Supplementary Table 9**; **Extended Data Fig. 8b**). **Category A** islands (n=33; 6.9%) comprised CpGs that were completely demethylated across all female clones, indicating universal epigenetic desilencing. These islands were enriched in proximity

to the PAR and short arm of X chromosome (**Extended Data Fig. 8c**). Among informative genes harbouring at least one heterozygous germline variant and sufficient expression, both alleles were transcribed at comparable levels, indicating complete biallelic expression (**BAE**; **Figs. 4j,k; Supplementary Table 10**). **Category B** islands (n=25; 5.3%; **Fig. 4d**) showed complete CpG demethylation, but only in a subset of female clones. Affected clones clustered according to the developmental phylogeny, indicating that these desilencing events were not universal but instead reflected embryonic lineage-specific epiallelic commitment. Unlike Category A, promoter demethylation in Category B produced weaker transcriptional desilencing and often failed to yield complete BAE (**Figs. 4j,k; Extended Data Fig. 9a,b**), suggesting that loss of DNA methylation alone was insufficient to overcome other repressive mechanisms, such as Polycomb-mediated silencing³⁵. Presumably, these Category B islands may underlie the variable XCI escape genes previously reported³⁶, such as *TBL1X*, *SHROOM2* and *NAA10*. **Category C** islands (n = 45; 9.5%) were observed in most clones but, in contrast to Category A, showed interspersed and incomplete CpG demethylation within the affected CpG islands, consistent with partial epigenetic desilencing in affected clones (**Fig. 4d**). **Category D** islands (n = 373; 78.4%) were found only in sporadic clones without association with the developmental phylogeny, consistent with stochastic late-stage epigenetic desilencing during ageing (**Fig. 4d**). Category C and D did not consistently result in BAE of the target genes (**Figs. 4j,k; Supplementary Table 10**). Of note, widespread culture-induced XCI erosion³⁷ is unlikely to have affected our results, as all female clones retained evidence of XCI (**Extended Data Fig. 9c,d**).

Starting from transcriptomic profiles primarily and subsequently annotating the corresponding DNA methylation states, we identified 64 genes with substantial BAE in some clones despite lacking promoter demethylation, suggesting that XCI escape can arise through mechanisms independent of promoter demethylation (**Extended Data Fig. 9a,b**).

Intriguingly, three neighboring Category B promoter CpG islands within chrX:153,188,452–153,239,901 (hereafter R1, R2 and R3) showed non-random co-phasing categories (**Fig. 4l**). Across all 278 female clones from 10 donors, phasing of R1–R2–R3 methylation states revealed only 4 of the 8 possible configurations, representing a significant departure from the independent assortment of methylation states (permutation test based on the Pearson χ^2 statistic, $P < 1 \times 10^{-5}$). For example, R2 demethylation occurred exclusively in clones with concurrent R3 demethylation (**Fig. 4l**). This restricted distribution indicates that demethylation establishment across the three regions is not an independent process. Notably, this locus contains one of the highest densities of CTCF-binding sites on the X

chromosome, suggesting that CTCF-mediated chromatin architecture may hierarchically insulate neighboring CpG islands from XCI-associated DNA methylation (**Fig. 4l**)^{38,39}.

Embryonic epiallelic commitment predominantly determines L1 activity in colorectal epithelium

Long interspersed nuclear element-1 (**L1**) retrotransposons constitute approximately 17% of the human genome⁴⁰. Although L1 elements have been evolutionarily successful by copying and inserting themselves at new genomic sites, most of the 500,000 L1 copies in the human reference genome are truncated or otherwise incapable of further mobilization⁴¹, with only 276 identified or suggested as retrotransposition-competent (referred to as rc-L1)¹¹. Somatic L1 retrotransposition is frequent in a few types of cancers as well as in normal colorectal epithelium^{11,41}.

Because L1 retrotransposition requires transcription of an rc-L1 source, promoter demethylation is a necessary, although not sufficient, condition for successful retrotransposition (**Fig. 5a**)^{11,42}. However, the timing and causes of the loss-of-methylation remain poorly understood, with ageing and environmental exposure proposed as possible contributors^{43,44}.

Of note, many CpGs within the promoters of rc-L1s were not included in the conventional CpG methylation analysis described above because they are located within repetitive L1 sequences (i.e., approximately +1 to +250 bp). We therefore performed a separate L1-specific methylation epigenotyping analysis (**Methods**). In 276 potential rc-L1 sources (99 referenced and 177 non-referenced rc-L1s in our cohort), 72 showed promoter loss of methylation in at least one of the 406 clones (**Supplementary Tables 11, 12**). The frequency of loss-of-methylation varied markedly among L1 sources. Some sources, including 4p12, showed promoter loss of methylation only rarely, whereas others, including 22q12.1-2, were unmethylated in nearly all informative clones (**Fig. 5b**).

Interestingly, these locus-specific methylation patterns were strongly associated with early developmental phylogeny, with closely related clones displaying highly concordant states (**Fig. 5c**). Approximately 70% of rc-L1 promoter loss-of-methylation compared from consensus epigenotypes were shared among clonally related cells (**Fig. 5b, 5c**). Across all individuals, as exemplified by donor HC50 (**Fig. 5d**), concordance of rc-L1 promoter loss-of-methylation states between paired clones began to increase at an embryonic molecular time of ~5 mutations and reached ~70% by ~50 mutations (approximately corresponding to birth) after correcting for chance concordance expected from the locus-specific distribution of

methylation states (**Fig. 5e**). Collectively, our findings indicate that the majority of rc-L1 promoter loss-of-methylation events observed in adult colorectal epithelium are established during early embryogenesis, with relatively few acquired during postnatal ageing¹¹. Furthermore, for all somatic L1 transduction events in which the source rc-L1 could be identified, the source locus showed promoter loss of methylation in each clone, suggesting that once established, rc-L1 promoter hypomethylation remains stable throughout life (**Fig. 5d**). Together, these findings suggest that the developmental history of a somatic lineage largely determines which rc-L1 source elements remain epigenetically permissive for transcriptional activation throughout life.

Intriguingly, we observed that acquisition of loss of methylation in the promoter was not fully independent across rc-L1 loci. For example, 11 descendant clones from a common ancestor harbouring 12 early mutations showed coordinated promoter hypomethylation at multiple rc-L1 loci, including 17p13.3, 15q25.2-1, 3q25.1, 11q21-3, 4p12 and 5q14.1-1 (**Fig. 5d**; the lineage is highlighted in red). This pattern suggests that rc-L1 promoter re-methylation was globally compromised in the ancestral cell during early embryogenesis. Notably, this lineage also showed a significantly higher frequency of late-stage cell divisions or clonal expansions than other lineages ($P = 0.00014$, Fisher's exact test), raising the possibility that impaired embryonic control of L1 remethylation may be associated with subsequent clonal expansion in the adult colorectal tissue.

Embryonic epiallelic commitments as lineage barcodes for developmental phylogenies

Currently, early developmental phylogenies were predominantly reconstructed using spontaneous somatic mutations^{17,18,21}. Although this approach enables direct lineage tracing, early embryonic mutations accumulate at a rate of only ~1 mutation per cell division, and not every division generates a detectable somatic mutation¹⁷. Consequently, mutation-based lineage tracing often lacks sufficient resolution to resolve individual cell divisions, resulting in multifurcations, or polytomies¹⁷. More recently, DNA methylation profiles obtained directly from developing embryonic cells have been used to reconstruct early developmental lineages⁴⁵. We therefore asked whether EEC events could similarly provide sufficient information to infer early developmental relationships and help resolve polytomies in phylogenies reconstructed from adult cells. Although EECs represent only the subset of early embryonic epigenetic states that remain stably maintained into adulthood, they may provide complementary lineage information.

We first evaluated this approach using lineages with known bifurcations supported by early somatic point mutations (**Extended Data Fig. 10a–c; Supplementary Table 13; Methods**). With EEC elements identified from multiple donors, the accuracy of EEC-based topology prediction was low for the earliest branches (≤ 5 shared somatic mutations; 13.3%; 2 out of 15), but increased markedly for branches arising after >5 shared mutations (75.7%; 28 out of 37) (**Extended Data Fig. 10d**). Prediction accuracy after >5 shared mutations increased significantly with the confidence score, defined by the number of EECs compatible with each topology, reaching 93.8% for confidence scores ≥ 0.7 (15 out of 16; **Fig. 5f**). The lower accuracy of the earliest branches (i.e., ≤ 5 shared somatic mutations) likely reflects the high recurrence of DNA methylation changes during early embryogenesis, making them less reliable as lineage markers.

To assess the utility of this approach in real-world cases, we applied the framework to 57 unresolved polytomies arising after >5 mutations of molecular time across the 15 donor phylogenies. Using confidence scores ≥ 0.7 , we resolved 25 polytomies into bifurcating relationships (**Extended Data Fig. 11; Supplementary Table 13**). One predicted topology was independently validated by identifying an early embryonic tandem-repeat mutation that had not been detected by routine genomic analysis (**Fig. 5g, 5h**). Our approach demonstrates the potential of EEC events to refine early developmental phylogenies beyond the resolution achievable with somatic mutations alone. Notably, our method used only recurrent EEC elements identified across multiple individuals. Incorporating individual-specific EEC elements could further improve the resolution of polytomies, although this would require case-specific analyses and greater computational complexity.

Discussion

Here, we provide one of the most comprehensive genome-wide assessments of CpG epimutation heterogeneity among individual colorectal epithelial cells by combining organoid-based clonal expansion with whole-genome methylation profiling. Clonal expansion generated sufficient material for high-coverage methylome, genome and transcriptome sequencing, enabling near-complete epigenotyping to be integrated with developmental phylogenies and gene expression.

Our findings reveal extensive epigenetic heterogeneity among colorectal epithelial cells. More importantly, a predominant fraction of this methylation heterogeneity in the adult colorectal epithelium originates during early embryogenesis rather than from age-associated stochastic changes acquired in terminal lineages. We estimate that early embryonic events

account for ~60% of altered individual CpGs, ~97% of regions containing ≥ 10 consecutive altered CpGs, and ~70% of rc-L1 promoter methylation changes in individuals in their 60s. Practically, EEC regions include numerous gene promoters, loci associated with escape from X-chromosome inactivation in females and rc-L1 sources that have undergone loss of methylation. These early embryonic epimutations were stably inherited through lineage specification and decades of epithelial stem-cell self-renewal, and some were associated with persistent transcriptional differences. These findings extend the concept of early developmental mosaicism beyond genomic mutations and suggest that lineage history contributes substantially to functional variation among otherwise similar adult cells.

Conceptually, EECs extend the framework of early-established epiallelic variation exemplified by metastable epialleles and correlated regions of systemic interindividual variation (CoRSIVs)^{46,47}. Metastable epialleles are thought to reflect alternative epigenetic states established early in embryogenesis and subsequently maintained across somatic tissues, whereas CoRSIVs similarly capture systemic interindividual methylation variation that is correlated across tissues derived from different germ layers. In both cases, the predominant variation focuses on between individuals rather than between lineages within an individual. By contrast, EECs reveal epiallelic states partitioned among embryonic lineages within the same individual, thereby generating intra-individual epigenetic heterogeneity within the same cell type. These phenomena need not be mutually exclusive: an EEC established sufficiently early, or in a lineage that contributes disproportionately to multiple tissues, could manifest at the organismal level as systemic interindividual variation resembling a metastable epiallele or CoRSIV. In this sense, EEC may represent a broader framework that could encompass at least a subset of metastable epialleles. Indeed, *VTRNA2-1*, a well-characterized metastable epiallele⁴⁸, was identified as an EEC region in our dataset. At this locus, we observed both lineage-specific epiallelic divergence within individuals and interindividual differences associated with skewed epiallelic commitment (**Extended Data Fig. 12**).

Why certain genomic regions are particularly susceptible to early epigenetic alteration, and what drives their commitment to specific epigenetic states, remain unclear. Their recurrence across lineages and donors indicates that they are not randomly distributed across the genome. EEC regions are likely depleted from genomic regions undergoing frequent epigenetic changes throughout life, but may be enriched at hotspots where early epigenetic programming is less tightly controlled or intrinsically heterogeneous. The frequent alignment of EEC segment boundaries with CpG islands and other sharply defined regulatory domains is consistent with domain-level remodelling and suggests that local sequence composition,

transcription-factor occupancy and three-dimensional chromatin architecture constrain the establishment and spread of EECs during development. Potentially, commitment to a particular epigenetic state at EECs may be probabilistic across embryonic lineages and influenced by the periconceptional environment, as suggested by effects of maternal diet and/or by season-of-conception effects. Future studies will be necessary to reveal such associations.

Unlike DNA sequence mutations, methylation states can be lost, regained or repeatedly altered over the lifetime of a lineage. It therefore remains unclear how often individual CpGs change state during a human lifespan, how frequently such changes revert, and whether recurrent changes at the same site represent independent events or restoration of an ancestral state. We believe that EECs identified in this study represent only a subset of a much larger number of epigenetic changes established during early embryogenesis, specifically, those that remain stably maintained in colorectal epithelial lineages. Extending this analysis to other cell types will therefore be particularly informative, both for determining which EECs are shared across tissues and for uncovering additional embryonic epigenetic states that may have been erased from colorectal lineages but retained elsewhere.

DNA methylation explored in this study represents only one layer of epigenetic regulation. Similar lineage-resolved approaches could be extended to histone modifications, chromatin accessibility, transcription-factor occupancy and three-dimensional genome organization. Integrating these features with phased long-read sequencing and somatic phylogenies should reveal whether early developmental variation is coordinated across epigenetic layers and how it shapes cell identity, transcriptional potential and disease susceptibility throughout life.

Finally, cancer cells commonly exhibit profound epigenome-wide remodelling, characterized by global DNA hypomethylation together with focal CpG-island hypermethylation; a subset of colorectal cancers additionally exhibits the CpG island methylator phenotype (**CIMP**)⁴⁹. Neither widespread hypomethylation nor a CIMP-like hypermethylation pattern was observed across the 406 normal colorectal epithelial clones analysed here, suggesting that these cancer-associated methylation states are not simply extensions of the epigenetic heterogeneity present in normal colorectal epithelium. Instead, neoplastic transformation may engage distinct mechanisms accelerating remodelling the methylome on a genome-wide scale. Defining these mechanisms, and how they emerge from the normal epigenetic landscape described here, will be an important direction for future studies.

References (main text)

1. Hatton, I. A. *et al.* The human cell count and size distribution. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2303077120 (2023).
2. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
3. Pijuan-Sala, B., Guibentif, C. & Göttgens, B. Single-cell transcriptional profiling: a window into embryonic cell-type specification. *Nat. Rev. Mol. Cell Biol.* **19**, 399–412 (2018).
4. Samuels, M. E. & Friedman, J. M. Genetic mosaics and the germ line lineage. *Genes (Basel)* **6**, 216–237 (2015).
5. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17**, 487–500 (2016).
6. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
7. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
8. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).
9. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
10. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
11. Nam, C. H. *et al.* Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature* **617**, 540–547 (2023).
12. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
13. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).

14. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358**, 69–75 (2017).
15. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
16. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).
17. Park, S. *et al.* Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).
18. Coorens, T. H. H. *et al.* Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).
19. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3740–3745 (2002).
20. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
21. Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
22. Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).
23. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
24. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38 (2012).
25. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol.* **16**, 56 (2015).
26. Garcia-Fernández, J. The genesis and evolution of homeobox gene clusters. *Nat. Rev. Genet.* **6**, 881–892 (2005).

27. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
28. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).
29. Pegg, A. E. Repair of O(6)-alkylguanine by alkyltransferases. *Mutat. Res.* **462**, 83–100 (2000).
30. Esteller, M. *et al.* Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* **343**, 1350–1354 (2000).
31. Zhu, L., Li, X., Yuan, Y., Dong, C. & Yang, M. APC promoter methylation in gastrointestinal cancer. *Front. Oncol.* **11**, 653222 (2021).
32. Wutz, A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat. Rev. Genet.* **12**, 542–553 (2011).
33. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
34. Zhou, F. *et al.* Reconstituting the transcriptome and DNA methylome landscapes of human implantation. *Nature* **572**, 660–664 (2019).
35. Żylicz, J. J. *et al.* The implication of early chromatin changes in X chromosome inactivation. *Cell* **176**, 182–197.e23 (2019).
36. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).
37. Mekhoubad, S. *et al.* Erosion of dosage compensation impacts human iPSC disease modeling. *Cell Stem Cell* **10**, 595–609 (2012).
38. Filippova, G. N. *et al.* Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev. Cell* **8**, 31–42 (2005).
39. Wiehle, L. *et al.* DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* **29**, 750–761 (2019).

40. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
41. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
42. Tubio, J. M. C. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
43. Bollati, V. *et al.* Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech. Ageing Dev.* **130**, 234–239 (2009).
44. Barchitta, M. *et al.* Mediterranean Diet and particulate matter exposure are associated with LINE-1 methylation: Results from a cross-sectional study in women. *Front. Genet.* **9**, 514 (2018).
45. Chen, M., Fu, R., Chen, Y., Li, L. & Wang, S.-W. High-resolution, noninvasive single-cell lineage tracing in mice and humans based on DNA methylation epimutations. *Nat. Methods* **22**, 488–498 (2025).
46. Rakyán, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in mammals. *Trends Genet.* **18**, 348–351 (2002).
47. Gunasekara, C. J. *et al.* A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol.* **20**, 105 (2019).
48. Silver, M. J. *et al.* Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol.* **16**, 118 (2015).
49. Jung, G., Hernández-Illán, E., Moreira, L., Balaguer, F. & Goel, A. Epigenetics of colorectal cancer: biomarker and therapeutic potential. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 111–130 (2020).

Main figure legends

Figure 1 | Profiling somatic DNA methylation mosaicism in normal human colorectal epithelium

a, Schematic illustrating the developmental origins of somatic epimosaicism in the colorectal epithelium, including imperfect epigenetic programming across embryonic lineages and age-associated stochastic changes. **b**, Cohort and study design for multi-omic profiling of single-crypt-derived colorectal organoid clones with matched peripheral blood. Donor age, sex and number of analysed clones are indicated. **c**, Genomic distribution of CpGs and definitions of CpG islands, shores, shelves and open seas in relation to gene promoters, gene bodies and intergenic regions. **d**, Distribution of β values for all accessible autosomal CpGs in a representative clone (HC01-09), shown overall and stratified by CpG context. **e**, Genome-wide proportions of biallelic-methylated (M/M), monoallelic-methylated (U/M) and biallelic-unmethylated (U/U) epigenotypes across 406 clones and 16 donors. **f**, CpG methylation patterns at the promoters of *POU5F1* (silenced in adult cells), *H19* (imprinted gene) and *ACTB* (constitutively active housekeeping gene) in clones from HC01 and HC08. Rows represent individual clones ordered according to somatic mutation-derived phylogenies, and columns represent consecutive CpGs within each promoter. Genomic-region and CpG-context annotations are shown above each heatmap. **g**, Cumulative distributions of absolute β -value differences at individual-CpG and CpG-island levels for technical replicate pairs, biological replicate pairs and within-donor clone pairs. **h**, Pairwise epigenotype concordance at individual-CpG and CpG-island levels in technical replicate pairs, biological replicate pairs and all within-donor clone pairs, shown as the mean frequencies of each pairwise combination of U/U, U/M and M/M epigenotypes.

Figure 2 | Embryonic epiallelic commitments and their developmental and genomic characteristics.

a, Two representative loci showing EEC events in HC21 and HC50. Heatmaps show CpG methylation β values across individual clones. Developmental phylogenies reconstructed from somatic mutations are shown on the left, with branches carrying epigenetic states that differ from the consensus epiallele highlighted in red. CpG context, genomic location and the number of consecutive CpGs within each EEC segment are indicated. **b**, Spatial distribution of HC50 clones carrying the epiallelic states observed at the two loci shown in **a**, demonstrating that embryonic lineages carrying a specific epiallele are not restricted to particular anatomical sites within the colon. Numbers in parentheses indicate the number of

HC50 clones in each epiallelic state. D1–D3, descending colon; PS, MS and DS, proximal, middle and distal sigmoid colon, respectively. **c**, Schematic of the timing analysis for EEC events, with representative phylogenetically assigned EECs in HC21. Bars below indicate the somatic-mutation-based molecular-time intervals during which each epiallelic commitment was inferred to have arisen. **d**, Cumulative epiallelic CpG changes as a function of somatic-mutation-based molecular time in HC50. For each molecular-time interval, lineage-specific CpG-change estimates were averaged across contemporaneous lineages and cumulatively summed to generate the trajectory. The schematic above shows representative shared branches used to estimate embryonic and postnatal changes. The early developmental period is enlarged on the left; the dashed line indicates birth (~50 mutations), and the lower axis indicates approximate chronological age. **e**, Cumulative gain- and loss-of-methylation events during early development in HC50, stratified by CpG context. **f**, Number of donor-specific EEC segments identified in each donor, stratified by segment size, defined as the number of consecutive CpGs within each segment. **g**, Distribution of donor-specific EEC segments according to recurrence across donors. **h**, Recurrence of EEC regions across embryonic lineages. Overall, 67% of EEC regions were observed in two or more embryonic lineages. **i**, Schematic illustrating the organization of an EEC region and its constituent EEC elements. Consensus methylation state and EEC-associated epigenotypes are indicated by colour. **j**, Pairwise reciprocal CpG overlap between donor-specific EEC segments assigned to the same EEC element. The histogram shows the distribution of the minimum reciprocal overlap. **k**, Alignment of EEC segment boundaries relative to CpG-island (CGI) boundaries. Segment starts and ends were oriented in a common direction, with the CGI boundary set to 0 and negative and positive coordinates indicating positions outside and inside the island, respectively. Grey lines represent individual segments, and the black curve shows the density of aligned EEC segment boundaries. **l**, Enrichment of EEC regions across CpG contexts relative to CpG-number-matched random regions, expressed as $\log_2(\text{observed/expected})$. Error bars indicate 95% permutation intervals. Significance was assessed by permutation testing with Benjamini–Hochberg correction; *adjusted* $P < 0.05$, $** < 0.01$, $*** < 0.001$. **m**, Enrichment of EEC regions across genomic annotations, epigenomic features and replication-timing domains, expressed as $\log_2(\text{observed/expected})$. ICR, imprinting control region; DHS, DNase I hypersensitive site.

Fig. 3 | Functional consequences of embryonic epiallelic commitment regions

a, Fraction of autosomal gene promoters and associated genes overlapping EEC regions. **b**, Enrichment of EEC-associated promoter genes across HUGO Gene Nomenclature Committee (HGNC) gene groups. Bars indicate fold enrichment relative to the expected

frequency in the background gene set; the dashed line indicates no enrichment (fold enrichment = 1). Representative EEC-associated genes are shown for each gene group. **c**, Volcano plot showing associations between promoter EEC methylation and gene expression in adult colorectal clones. Regression coefficients (β) and FDR values are shown, with representative genes showing negative (*MAP3K7CL*) or positive (*HOXD8*) associations indicated. **d,e**, Representative associations between promoter EEC epigenotype and gene expression for *MAP3K7CL* (**d**) and *HOXD8* (**e**) in donors with matched methylation and expression data. Points represent individual clones and are coloured by donor. Regression coefficients and *P* values are indicated. **f**, Representative promoter EEC at *MGMT* in HC50. Left, somatic mutation-based phylogeny and promoter CpG methylation states, with EEC-associated clones highlighted in yellow. Right, *MGMT* expression according to promoter epigenotype (Wilcoxon rank-sum test, $P = 4.1 \times 10^{-5}$; $**P < 0.001$). **g**, Number of segmental DNA methylation changes comprising ≥ 10 consecutive CpGs per adult colorectal clone, and the proportion overlapping EEC regions. **h**, HGNC gene-group enrichment among genes associated with EEC elements or LECs, with LEC and EEC enrichment shown on the left and right, respectively. Significant gene groups are highlighted (two-sided Fisher's exact test, FDR < 0.05). **i**, Recurrent LECs at the *APC* promoter 1A. Left, somatic mutation-based phylogenies and promoter CpG methylation states across four donors harbouring independent LECs. Right, relative usage of *APC* transcript 1A according to promoter 1A LEC status (Wilcoxon rank-sum test, $P = 0.03$).

Figure 4 | X chromosome DNA methylation and inactivation in female colorectal epithelium

a, Schematic of X-chromosome methylation β values in males and females across non-pseudoautosomal (non-PAR) regions. **b**, Distribution of consensus CpG epigenotypes across PAR and non-PAR regions in male and female clones ($n = 128$ and 278 , respectively). **c**, Possible combinations of consensus CpG epigenotypes between males and females, stratified by CpG context. **d**, Landscape of transcription and sex-specific DNA methylation across X-chromosomal CpG islands ($n = 750$) in colorectal epithelium. Transcriptional features include allele-specific expression (ASE), frequency of biallelic expression (BAE), frequency of U/U clones showing BAE (BAE freq_{U/U}), extent of BAE among affected U/U clones (BAE extent_{U/U}), and female-to-male expression differences. Four categories of CpG islands showing complete demethylation in females are indicated, with representative examples from HC50. **e**, Allele-resolved promoter methylation in HC15, showing the somatic mutation-based developmental phylogeny and methylation states of phased X-chromosome haplotypes (HP1 and HP2) determined by long-read sequencing of

clones. Blood-shared mutations and mutations informative for XCI timing are highlighted in yellow. **f**, Concordance of XCI allelic choice among descendant clones as a function of somatic-mutation-based molecular time (top), and mean VAF of the corresponding early mutations in bulk blood (bottom). Mutations shared between colorectal epithelium (endoderm) and blood (mesoderm) indicate a pre-gastrulation origin. Error bars indicate s.e.; the red line marks 18 shared mutations. **g**, Probability-weighted inference of XCI timing across 10 female donors. Grey curves represent individual donors, and the blue curve shows the equally weighted combined distribution. **h**, DNA methylation epigenotypes flanking U/U CpG islands in HC15-01, aligned by relative CpG position from CpG-island boundaries. Top, proportions of U/U, U/M and M/M epigenotypes; bottom, CpG-context composition. **i**, Numbers of X-chromosomal CpG islands in Categories A–D among the 476 islands showing U/U epigenotypes in female clones. **j**, Relationship between promoter demethylation to U/U and allelic expression across Categories A–D. For genes with sufficient expression and at least one exonic heterozygous germline variant, BAE and monoallelic expression (MAE) were distinguished. Genes were considered non-informative when expression was insufficient (LowDepth) or no informative heterozygous germline variant (w/oSNP_{Het}) was present. **k**, Clone-level comparison of promoter DNA methylation and allele-specific expression for genes associated with U/U CpG islands. In total, 163 genes had informative ASE data in U/U clones. BAE frequency indicates the proportion of informative U/U clones showing BAE, whereas BAE extent indicates the mean minor-allele VAF among BAE clones. Genes are coloured according to U/U CpG-island category (A–D), with the overall two-dimensional density shown in light grey. **l**, DNA methylation landscape across chrX:153,188,452–153,239,901 in female clones, showing three regions of coordinated CpG-island methylation (R1–R3).

Figure 5 | Embryonic epiallele commitments at LINE-1 promoters and the utility of EEC for reconstructing developmental phylogenies.

a, Schematic of LINE-1 activation, comprising (1) promoter loss of methylation (LoM), (2) transcription and (3) retrotransposition, with a 3' transduction event shown as an example. Filled and open circles denote methylated and unmethylated CpGs, respectively. **b**, Germline composition, promoter DNA methylation and sharing of LINE-1 promoter LoM across clones. The analysis includes 72 retrotransposition-competent LINE-1 (rc-L1) sources showing promoter LoM in at least one clone. The upper panel shows the germline source copy-number distribution across 15 donors, classified as homozygous (2 copies), heterozygous (1 copy) or absent (0 copies). The middle panel shows the frequencies of promoter methylation states (1, 0.5 and 0) among informative clones for each source. The lower panel shows the proportion of promoter LoM events classified as shared, when

present in multiple descendant clones of an early phylogenetic node, or non-shared. Sources are ordered by the frequency of promoter LoM. **c**, CpG-level promoter methylation at two representative rc-L1 sources at 22q12.1–2 and 4p12 in HC50. Heatmaps show CpG methylation β values across the rc-L1 promoter and flanking regions in individual clones ordered according to the somatic mutation-based phylogeny. Green boxes denote the analysed promoter interval (+1 to +250 bp relative to the L1 transcription start site), with coordinates oriented according to the direction of L1 transcription. **d**, Clone-level mean promoter methylation β values across rc-L1 sources in HC50, shown alongside the developmental phylogeny. Columns represent individual clones ordered according to the somatic mutation-based phylogeny above, including late-stage branches, and rows represent rc-L1 sources showing promoter LoM in more than five HC50 clones. Red branches highlight a lineage of clones descended from a common ancestor and showing a higher number of promoter LoM across multiple rc-L1 loci. Source annotations indicate the previously reported allele frequency in human population (population allele-frequency; PAF)¹¹, germline allele composition in HC50 and consensus promoter methylation state across HC50 clones. Yellow cells indicate detected transduction events originating from the corresponding source in each clone. The bar plot below shows, for each clone, the number of rc-L1 promoters showing additional LoM relative to the consensus epigenetic state. **e**, Prevalence-adjusted concordance of rc-L1 promoter methylation states between clone pairs with promoter LoM in at least one clone, plotted as a function of their shared somatic mutation burden. Points and the line indicate mean adjusted concordance within bins of shared mutations; shading indicates the 95% bootstrap confidence interval. **f**, Utility of EEC information for improving developmental phylogeny reconstruction in addition to early embryonic somatic mutations. The plot shows the association between prediction confidence score from our method and the accuracy of methylation-based phylogenetic inference for resolving branches with >5 shared somatic mutations. The line indicates logistic regression with 95% confidence interval ($n = 37$; $P = 0.022$). **g**, Representative methylation-based refinement of a multifurcating node. EEC elements supported the inferred bifurcation with a confidence score of 0.8, which was independently validated by an additional somatic mutation in a repetitive region (chr18:20,071,460 TGATA>T). The heatmap shows supporting EEC epigenotypes across clones. **h**, IGV validation of the somatic mutation supporting the bifurcation in **g**. Read alignments confirm the chr18:20,071,460 TGATA>T variant in clones assigned to the inferred lineage.

Methods

Human tissue sampling

Colorectal tissues were obtained from patients undergoing surgical resection. Fourteen donors included in this study were derived from a previously reported cohort¹¹. As described previously, macroscopically normal colorectal mucosa was collected from surgical specimens of patients undergoing elective colorectal tumour resection, with normal tissue sampled from regions located more than 5 cm from the primary tumour. Matched peripheral blood was also collected for bulk-tissue whole-genome sequencing.

An additional patient (donor ID HC50) undergoing surgical resection for rectal cancer was prospectively sampled for this study. The resected specimen extended from the descending colon to the rectum, and macroscopically normal mucosa was collected from six spatially defined sites along the specimen. Three adjacent samples were obtained from the descending colon, and one sample each was obtained from the proximal, middle, and distal sigmoid colon. At each site, thin pieces of mucosa containing colorectal crypts were dissected while excluding the underlying muscular layer. These tissues were transported to the laboratory for clonal organoid establishment, followed by genomic and methylation analyses.

All procedures were approved by the Institutional Review Board of Seoul National University Hospital (approval no. [H-2304-116-1425 and 1103-125-357]) and KAIST (approval no. [KH2023-111]), and written informed consent was obtained from all participants.

Organoid culture of colorectal crypts

Colorectal crypts were isolated and cultured using a protocol adapted from previously described human intestinal organoid culture methods, with modifications⁵⁰. Fresh mucosal tissues were transported to the laboratory in ice-cold PBS and processed with minimal delay. After briefly washing with PBS, the tissues were finely minced into small pieces. A small portion of each tissue specimen was retained for bulk DNA sequencing, and the remaining tissue fragments were repeatedly washed with PBS until the supernatant became clear. To isolate colorectal crypts, tissue fragments were incubated for 30 min at 4°C with gentle tilting in PBS containing 10 mM EDTA (Invitrogen), using 50 ml conical tubes. Following incubation, the tissue suspension was vigorously shaken to release crypts, and the presence of isolated crypts in the supernatant was confirmed microscopically. The crypt-containing

supernatant was collected, centrifuged at $150 \times g$ for 5 min at 4°C , and washed twice with PBS. The resulting crypt pellet was resuspended in growth-factor-reduced Matrigel (Corning) with gentle mechanical dissociation. A small aliquot was initially plated to estimate crypt density, after which the suspension was further diluted in Matrigel as needed and plated as approximately 10- μl droplets in prewarmed 6-well plates. Matrigel droplets were allowed to solidify for 5–10 min before addition of organoid culture medium, prepared as described in **Supplementary Table 14**. Cultures were maintained with medium replacement two to three times per week and monitored regularly. Individual organoids were manually isolated and passaged before contact or fusion with neighboring organoids, whereas bulk cultures were passaged upon reaching confluence.

For 15 donors derived from the previously reported cohort¹¹, clonal colorectal organoids had already been established and cryopreserved; therefore, crypt isolation and primary organoid establishment were not repeated for these donors.

Clonal expansion of organoid

For newly established cultures, primary organoids were monitored until individual organoids derived from spatially isolated crypts had grown sufficiently for clonal expansion. Individual organoids were manually picked before contact or fusion with neighboring organoids and transferred to separate wells. For the 15 donors from the previously reported cohort, individual cryopreserved clonal organoid stocks were thawed and recovered for expansion in the present study.

For subsequent passage, culture medium was removed and Matrigel was recovered using Cell Recovery Solution (Corning), followed by incubation on ice for 40–50 min. Organoids were collected by centrifugation at $300 \times g$ for 5 min, and the supernatant containing dissolved Matrigel was removed. Organoid pellets were resuspended in 50 μl of TrypLE Express (Gibco) by gentle pipetting and incubated for 4 min at 37°C with agitation at 1,100 rpm. TrypLE activity was terminated by addition of 300 μl ADF+++ medium, consisting of Advanced DMEM/F12 supplemented with 10 mM HEPES, $1\times$ GlutaMAX, and 1% penicillin–streptomycin. The suspension was centrifuged at $300 \times g$ for 5 min at 4°C , and the resulting pellet was mechanically dissociated by pipetting and resuspended in growth-factor-reduced Matrigel (Corning).

Organoids were seeded into prewarmed 24- or 12-well plates, and the Matrigel was allowed to solidify for 5–10 min at 37°C before addition of prewarmed organoid culture medium. Cultures were maintained with medium replacement every 2–3 days and serially passaged

until each clone was expanded to two 12-well cultures. One well was cryopreserved as an organoid stock, whereas the other was harvested for simultaneous DNA and RNA extraction.

DNA and RNA extraction from clonal organoids

Genomic DNA and total RNA were simultaneously extracted from expanded clonal organoids using the AllPrep DNA/RNA Kit (Qiagen) according to the manufacturer's instructions, with minor modifications. Organoids were lysed in RLT buffer supplemented with dithiothreitol (DTT), and DNA and RNA were purified using the corresponding AllPrep DNA and RNeasy MinElute spin columns. Purified DNA and RNA were used for downstream sequencing analyses.

Library preparation and sequencing

For the 15 donors derived from the previously reported cohort, previously generated WGS data from clonal colorectal organoids and matched peripheral blood were used in the present study¹¹. For newly established clonal organoids, genomic DNA was used to generate PCR-free WGS libraries using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina), followed by paired-end sequencing on the NovaSeq 6000 platform. Matched peripheral blood was processed and sequenced in parallel for WGS.

For whole-genome DNA methylation sequencing, genomic DNA from expanded clonal organoids was subjected to enzymatic methylation sequencing (EM-seq). Libraries were prepared from 200 ng of input DNA using the NEBNext Enzymatic Methyl-seq Kit (New England Biolabs), with CpG-methylated pUC19 and unmethylated lambda DNA included as methylation and conversion controls. Paired-end sequencing was performed on the Illumina NovaSeq 6000 platform. EM-seq libraries were newly generated for the clonal organoids analyzed in the present study, including clones recovered from the 15 previously reported donors.

For whole-transcriptome sequencing, total RNA from expanded clonal organoids was used to construct strand-specific RNA-sequencing libraries using the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs), according to the manufacturer's instructions. Libraries were sequenced using paired-end sequencing on an Illumina platform. Previously generated whole-transcriptome sequencing data from organoid clones of HC15 and HC16, prepared using the TruSeq Stranded Total RNA Gold Kit (Illumina), were also included in the present study.

973

974 **Variant calling and filtering of WGS data**

975 Sequencing reads were mapped to the human reference genome (GRCh37) using the
976 Burrows–Wheeler Aligner (BWA)-MEM algorithm⁵¹, coordinate-sorted using SAMtools⁵², and
977 duplicate reads were removed using GATK4 MarkDuplicates (available at
978 <http://broadinstitute.github.io/picard>). Local realignment around known indels was
979 subsequently performed using GATK3 RealignerTargetCreator and IndelRealigner⁵³. SNVs
980 and short indels were independently called using VarScan2⁵⁴ and GATK HaplotypeCaller⁵⁵.
981 Variant calls from the two callers were combined, decomposed and normalized using VT⁵⁶,
982 and variants overlapping satellite regions were excluded. Candidate variants were further
983 filtered against the panel of normal. Variants present at a frequency of $\geq 5\%$ in any PoN were
984 excluded, whereas variants absent from the PoN or present at $< 5\%$ were retained. The
985 remaining candidates were annotated using read-level information from the corresponding
986 BAM files. High-confidence variants were required to have at least two alternate-supporting
987 reads ($AD \geq 2$), a clipped-read fraction ≤ 0.50 , mean mapping quality of alternate-supporting
988 reads ≥ 40 , a strand-bias metric > 0.001 and a mean relative read position between 0.1 and
989 0.9. Variants with a deletion-associated read fraction ≥ 0.25 were excluded. Additional
990 filtering was applied to variants within long homopolymer contexts to remove sites showing
991 excessive sequencing-error signatures. Variants with a VAF ≥ 0.10 in the matched bulk blood
992 sample were excluded as putative germline variants. Variants showing recurrent alternate-
993 read evidence in other clones were additionally removed to minimize sequencing and
994 mapping artefacts. Clonality of each organoid culture was then assessed from the VAF
995 distribution of the retained somatic variants, with a dominant peak around 0.5 considered
996 consistent with clonal expansion of diploid cells.

997 **Mutational signature analysis**

998 Mutational signature analysis was performed separately for single-base substitutions (SBSs)
999 and small insertions and deletions (indels) using the SigProfiler framework. For each clone,
1000 mutation matrices were generated from the corresponding VCF using
1001 SigProfilerMatrixGenerator⁵⁷ with the GRCh37 reference genome. SBSs were represented in
1002 the 96-channel SBS context, whereas indels were represented in the 83-channel ID context.
1003 Signature exposures were estimated by non-negative least-squares fitting with the signature
1004 decomposition procedure implemented in SigProfilerExtractor. Candidate signatures were
1005 restricted to SBS1, SBS5, SBS18, SBS40, SBS88 and SBS89 for SBSs, and ID1, ID2, ID5,

1006 ID9, ID18 and IDB for indels. Signature contributions were reported as the proportion of the
1007 total fitted exposure for each clone.

1008 **Reconstruction of phylogenetic trees based on somatic mutations**

1009 Phylogenetic trees were reconstructed for each donor using somatic SNVs and short indels
1010 identified in clonally expanded colorectal organoids, using an approach adapted from
1011 previously described methods for reconstructing developmental phylogenies from somatic
1012 mutations^{11,17}. High-confidence variants identified in individual clones were merged within
1013 each donor, and read-level information at all candidate sites was re-extracted from the BAM
1014 files of all corresponding clones. For each variant, the genotype of each clone was binarized
1015 according to its variant allele frequency (VAF), with $VAF > 0.20$ assigned as present (1) and
1016 $VAF \leq 0.20$ as absent (0). Variants detected in all clones were excluded from phylogenetic
1017 reconstruction.

1018 Mutations were grouped according to their presence–absence patterns across clones, and
1019 mutation-sharing patterns were visualized as heatmaps to identify shared lineage
1020 relationships and remove inconsistent variant patterns. Hierarchical relationships among
1021 mutation groups were established such that a mutation group present in a subset of clones
1022 carrying another mutation group was assigned to a descendant lineage. Phylogenetic trees
1023 were reconstructed to best represent these hierarchical relationships, with each organoid
1024 clone assigned to a terminal node. The number of mutations belonging to each mutation
1025 group was used as the branch length, representing somatic-mutation-based molecular time.

1026 **Processing and quality control of whole-genome methylation sequencing data**

1027 Paired-end EM-seq reads were aligned to the human reference genome (GRCh37) using
1028 Bismark⁵⁸ with Bowtie2⁵⁹ after adaptor sequences were removed by using Cutadapt⁶⁰. PCR
1029 duplicates were removed using the Bismark deduplication module, and deduplicated reads
1030 with a mapping quality < 20 were excluded using SAMtools⁵². CpG methylation calls were
1031 extracted using the Bismark methylation extractor, excluding the first five bases of read 2.
1032 Genome-wide CpG coordinates were defined from the GRCh37 reference genome using the
1033 cytosine position of each CpG dinucleotide. For each CpG, methylated and unmethylated
1034 read counts from the two complementary strands were summed, and the methylation
1035 fraction was calculated as the fraction of methylated reads among all informative reads.
1036 CpG-level measurements from all organoid clones within each donor were then merged
1037 according to the reference CpG coordinates, with uncovered sites retained as missing
1038 values.

Methylation fractions supported by fewer than five total reads were set to missing. For locus-level quality control, the average depth of each CpG was calculated as the total number of methylated and unmethylated reads across all clones divided by the number of clones. CpGs with an average depth <5 or >50 , or with missing methylation measurements in ≥ 150 clones, were excluded. CpG sites overlapping donor-specific germline variants were excluded from downstream analyses. Germline variants were identified from matched blood samples using a VAF threshold of >0.10 .

Sample-level methylation quality was assessed using sequencing depth and the genome-wide distribution of CpG methylation fractions. Mean sequencing depth for each organoid clone was calculated using `mosdepth`⁶¹. In differentiated human somatic cells, approximately 70–80% of CpGs are methylated, and genome-wide methylation profiles typically show a bimodal distribution with a prominent highly methylated peak near a methylation fraction of $1^{62,63}$. We therefore used the position of the highly methylated component as an additional indicator of methylation-profile quality. For each clone, CpG methylation fractions were ranked, and the mean methylation level of the upper 60% of CpGs was calculated. Clones with a mean sequencing depth of less than 10-fold or an upper-60% mean methylation level <0.97 were designated as lower quality and excluded from the epigenotype assignment procedure described in the following section.

Epigenotype assignment

To convert continuous CpG methylation fractions (β values) into discrete methylation states (epigenotypes), we applied a CpG-specific Gaussian mixture model (GMM) across organoid clones using the `mclust` R package⁶⁴. Only methylation measurements supported by at least five reads were considered, and CpGs with fewer than five non-missing measurements among high-quality samples were not subjected to state assignment. Samples designated as lower quality on the basis of sequencing depth and genome-wide methylation-distribution metrics were excluded from GMM parameter estimation. For each CpG, β values from high-quality samples were modelled using the Gaussian mixture framework with one to three components and equal variance across components. Model initialization was guided by assigning each methylation value to the nearest of three expected epigenotypes (U/U, U/M, and M/M), and the optimal number of mixture components was selected by the model-based clustering procedure. CpGs showing only a single unique methylation value were treated as a single-component distribution. The fitted Gaussian components were mapped to methylation states according to their estimated means, using boundaries of 0.25 and 0.75 to define epigenotypes (U/U, U/M and M/M), respectively. Individual samples were initially

assigned according to their fitted mixture component. To reduce misclassification of observations lying far from the centre of a fitted component, values differing by more than 0.25 from the median of their assigned cluster were reassigned to the nearest theoretical epigenotype (U/U, U/M or M/M). Empirical peak positions were subsequently recalculated as the mean methylation fraction of samples assigned to each state. Samples designated as lower quality were not used to estimate the mixture distributions. Their epigenotypes were subsequently assigned using the empirical state peaks inferred from high-quality clones; when the corresponding empirical state peak was not observed at a given CpG, assignment reverted to the nearest theoretical state. Missing methylation measurements remained unassigned.

For chromosome X, epigenotype assignment was performed separately for female and male donors. In females, CpGs were classified using the same three-state GMM framework (U/U, U/M and M/M) as for autosomes. In males, CpGs outside the pseudoautosomal regions (PARs) were modelled using one or two components corresponding to the two expected methylation states on the single X-chromosomal allele, unmethylated (U) and methylated (M), with a boundary of $\beta = 0.5$. CpGs within PAR1 (chrX:60,001–2,699,520) and PAR2 (chrX:154,931,044–155,260,560; GRCh37) were treated as diploid and assigned using the same three-state framework as autosomes.

For each donor, the consensus epigenotype at each CpG was defined as the most frequent state among all evaluable colorectal clones from that donor. Deviations from this donor-specific WT state in individual clones were considered methylation-state changes and were used for subsequent phylogenetic reconstruction. To assess the overall stability of CpG methylation states across the cohort, a cohort-wide consensus methylation state was additionally defined at each CpG as the most frequent state among high-quality clones. The proportion of evaluable clones matching this consensus methylation state was calculated for each CpG, and the cumulative proportion of CpGs exceeding each level of clone agreement was determined.

Technical and biological reproducibility of epigenotype assignment

Technical reproducibility of CpG epigenotype assignment was assessed using independently prepared and sequenced EM-seq replicates from the same organoid DNA samples. Four colorectal organoid clones were independently subjected to EM-seq twice. Biological reproducibility was additionally assessed using 11 pairs of closely related colorectal clones that diverged after accumulating more than 2,000 shared somatic mutations along late developmental branches. All other pairs of distinct colorectal clones from the same donor,

1107 excluding these designated biological replicate pairs, were used as within-donor
1108 comparisons to provide a reference for interindividual methylation difference in DNA
1109 methylation.

1110 For CpG-level comparisons, only CpGs with assigned epigenotypes in both members of a
1111 pair were considered evaluable. Donor-specific germline variant-overlapping CpGs were
1112 additionally excluded from within-donor comparisons. For each pair, evaluable CpGs were
1113 classified into the nine possible pairwise combinations of the three epigenotypes (U/U, U/M
1114 and M/M). Concordance was defined as the proportion of evaluable CpGs with identical
1115 epigenotypes in the two samples, whereas all other state combinations were considered
1116 discordant.

1117 Reproducibility was also evaluated at the CpG-island level. For each sample, mean β values
1118 were calculated across evaluable CpGs within individual CpG islands, and island
1119 epigenotypes were assigned as U/U, U/M or M/M using mean- β thresholds of 0.25 and 0.75.
1120 Only islands evaluable in both samples of a pair were included. Island-level concordance
1121 was calculated analogously to the CpG-level analysis. For visualization of epigenotype
1122 transitions, the numbers of observations in each of the nine possible state combinations
1123 were calculated separately for each pair and then averaged across technical replicate,
1124 biological replicate and within-donor clone pairs.

1125 To assess quantitative methylation differences independently of discrete epigenotype
1126 assignment, absolute β value differences ($|\Delta\beta|$) were calculated between paired samples at
1127 both individual CpGs and CpG islands. At the CpG level, differences were calculated only at
1128 loci evaluable in both samples after depth and locus-level quality filtering; at the CpG-island
1129 level, differences were calculated from island mean β values. For each pair, $|\Delta\beta|$ values
1130 were binned at intervals of 0.01 and converted to within-pair proportions. These proportions
1131 were then averaged across pairs so that each pair contributed equally, and cumulative
1132 distributions were generated by cumulatively summing the pair-averaged proportions.

1133 For CpG-island comparisons, mean β values from paired samples were also compared
1134 directly using Pearson correlation coefficients. Correlations for technical and biological
1135 replicate pairs were calculated from all evaluable island-pair observations, whereas 2,000
1136 observations per pair type were randomly sampled for visualization. For within-donor clone
1137 pairs, island observations were sampled across pairs to generate a comparison dataset
1138 comparable in size to the technical-replicate dataset, after which up to 2,000 observations
1139 were randomly selected for plotting. Islands were classified as concordant or discordant
1140 according to whether the two samples received the same island epigenotype.

Phylogenetic reconstruction and identification of embryonic epiallelic commitment (EEC) events

Somatic-mutation-based phylogenetic trees reconstructed from WGS data were used as fixed tree topologies for phylogenetic reconstruction of CpG methylation states. For each donor, the discrete epigenotypes (U/U, U/M, and M/M) of individual colorectal clones at autosomal CpGs were arranged according to the corresponding clone tips in the somatic phylogeny. The donor-specific WT epigenotype defined above was appended as a reference pseudo-tip adjacent to the root. Methylation-state matrices were converted to phyDat objects with 0, 0.5 and 1 specified as the possible character epigenotypes, and ancestral methylation states at internal nodes were reconstructed by maximum parsimony using the MPR procedure implemented in the phangorn package⁶⁵. Because maximum-parsimony reconstruction can yield multiple equally parsimonious states at an internal node, ambiguous ancestral states were retained as sets of possible methylation states and further resolved using local phylogenetic information. For each ambiguous node, support for the three epigenotypes was first aggregated across the node and its immediate descendant branches, and the node was assigned when a single state showed uniquely maximal support. Remaining ambiguous nodes were subsequently reevaluated using the parent node together with its descendant branches, and were resolved only when a unique epigenotype was supported; otherwise, the ancestral state remained ambiguous.

Methylation changes were then annotated along individual edges of the fixed phylogeny by comparing the reconstructed states of each parent–child pair. A transition between two different unambiguous states was counted as one change, whereas identical states were assigned a change score of zero. To account for uncertainty in partially resolved ancestral states, fractional change scores were assigned to transitions involving ambiguous states. An unambiguous state compatible with an adjacent ambiguous state was assigned a score of 0.5, whereas an incompatible transition was assigned a score of 1. Transitions between distinct partially ambiguous states were assigned a score of 0.25, and transitions involving a fully ambiguous ancestral state were assigned a score of 0.5. Changes assigned to internal branches were summed as the lineage-associated change score (lin), whereas changes on terminal clone branches were summed separately as the terminal change score (rand). The artificial rooting edge introduced by the WT reference was excluded from the lineage-associated score.

To determine whether an apparent lineage-associated methylation change represented a phylogenetically structured event rather than a pattern arising by chance from the marginal

frequencies of methylation states, a donor-specific permutation test was performed for CpGs with at least one inferred internal-branch change ($\text{lin} \geq 1$). At each tested CpG, the observed methylation states across clone tips were randomly permuted 1,000 times while preserving the numbers of clones assigned to states 0, 0.5 and 1, as well as the number of missing measurements; the donor-specific WT state was kept fixed. Each permuted dataset was subjected to the same maximum-parsimony ancestral reconstruction, ambiguity resolution and branch-wise change scoring as the observed data. For each observed and permuted CpG pattern, the total phylogenetic change score was calculated as the sum of internal- and terminal-branch change scores ($T = \text{lin} + \text{rand}$). Because a methylation pattern that follows the somatic lineage structure is expected to require fewer independent state changes than a randomly distributed pattern, smaller total change scores were considered stronger evidence of phylogenetic association.

To avoid redundant computation, one permutation distribution was generated for each unique combination of the numbers of clones in states 0, 0.5 and 1 and was reused for CpGs with the same state composition. For each CpG j , the permutation P value was calculated as

$$P_j = \frac{r_j}{B + 1}$$

where r_j is the average rank of the observed total change score (T) among the observed and $B = 1,000$ permuted scores, with lower change scores receiving lower ranks. CpGs with an internal-branch methylation change ($\text{lin} \geq 1$) and a permutation $P < 0.05$ were considered independently supported lineage-associated CpGs and were subjected to cross-donor integration and rescue as described below.

Donor-specific permutation results were subsequently integrated across all donors at the CpG level. For each donor, CpGs with a permutation P value < 0.05 were initially considered independently supported lineage-associated CpGs, and the union of all tested CpG positions across donors was generated. CpGs failing the locus-level quality-control criteria were excluded before downstream analysis. Because the phylogenies of HC21 and HC50 contained predefined internal branches representing later developmental or postnatal periods, methylation changes confined exclusively to these branches were distinguished from early lineage-associated changes. For these two donors, lineage-change scores attributable to the predefined late branches were subtracted from the total internal-branch change score to obtain an early-lineage change score. Calls arising exclusively on the

designated late branches were not allowed to contribute evidence of cross-donor recurrence or rescue.

To recover CpGs with reproducible lineage-associated methylation patterns across individuals that did not independently reach the permutation threshold in every donor, a cross-donor rescue procedure was applied. For each CpG, two measures of independent support were calculated. First, the recurrence count was defined as the number of donors in which the CpG independently satisfied the permutation criterion for an early lineage-associated change. Second, a cumulative lineage-support score was calculated by summing the internal-branch change scores only from donors in which the CpG had independently satisfied this criterion. For HC21 and HC50, only the early-lineage component of the change score contributed to these quantities. Cross-donor rescue thresholds were empirically determined by permutation, supporting a cutoff of 2 for both recurrence and cumulative lineage-support criteria. For each donor, a CpG meeting the donor-specific permutation criterion was classified as an original candidate. A CpG that did not independently meet this criterion in that donor was rescued when either the cumulative lineage-support score across independently supporting donors was ≥ 2 or the CpG was independently supported in at least two donors. CpGs overlapping a donor-specific germline variant were excluded from both original and rescued calls for that donor. Hereafter, CpG-level methylation changes assigned to early internal branches and supported by either the donor-specific permutation test or the cross-donor rescue procedure were termed embryonic epiallelic commitment (EEC). Contiguous clusters of these EEC-associated CpGs were subsequently used to define EEC segments and regions.

Molecular timing of EEC events

Using the branch-wise methylation-state changes reconstructed above, we estimated the molecular timing of EEC events separately for each donor. CpG-level EEC events were first identified through the donor-specific permutation and cross-donor rescue procedures described above. For the final timing analysis, however, phylogenetically reconstructed methylation changes at all CpGs located within the genomic intervals defined as EEC regions by the segmentation procedure described below were included. For each colorectal clone, the ancestral path from the root to the corresponding terminal node was identified. We focused on methylation changes that arose on an ancestral branch and were retained in the descendant clone without an additional change on the terminal branch. Each retained change was assigned to the earliest ancestral node at which the transition was inferred, with

fractional scores from ambiguous ancestral-state reconstruction retained as corresponding weights.

Molecular time was defined by the number of somatic mutations accumulated along each lineage, using branch lengths of the somatic mutation-derived phylogenetic trees. For each internal node, EEC-associated changes were averaged across its descendant clones, restricting the denominator to clones belonging to the corresponding lineage. Branches were then projected onto intervals of accumulated somatic mutations. When a branch spanned multiple mutation intervals, its inferred methylation changes were distributed across the intervals in proportion to their lengths, and estimates from contemporaneous branches were averaged. The cumulative number of EEC-associated methylation changes was obtained by summing these interval-specific estimates along increasing somatic-mutation time.

To examine the genomic context and directionality of early methylation changes, the same analysis was performed separately for CpG islands, shores, shelves and open-sea regions. Methylation gains were defined as transitions from U/U to U/M, U/M to M/M or U/U to M/M, whereas methylation losses were defined as transitions from U/M to U/U, M/M to U/M or M/M to U/U.

Construction of donor-specific EEC segments

To determine whether CpG-level EEC events were spatially clustered into coherent genomic domains, we defined **EEC segments** as contiguous genomic intervals containing neighboring CpGs with the same type of lineage-associated methylation changes occurring in the same or hierarchically related descendant lineages. EEC segments were identified separately for each donor by examining CpG-level EEC events together with the reconstructed methylation states and branch-specific change scores of neighboring CpGs. Consecutive CpGs supporting a common internal-lineage change were first grouped into candidate blocks, requiring evidence from at least three CpGs. Candidate blocks were then further refined according to the similarity of methylation-state profiles across adjacent CpGs. Blocks were split at positions showing low concordance between neighboring CpGs, whereas locally concordant blocks containing a core run of highly similar CpG profiles were retained. Short two-CpG blocks were additionally recovered when two neighboring CpGs showed highly concordant clone and reconstructed lineage-state profiles. Adjacent blocks were subsequently connected over relatively long genomic distances when they showed a shared internal-branch change, concordant clone-state profiles, and only a limited intervening CpGs. The resulting segments containing at least one CpG identified by the permutation or cross-donor rescue procedure were retained.

1274 To exclude spurious internal-branch changes driven by missing methylation measurements,
1275 CpGs with insufficient support among descendant clones were filtered out, and candidate
1276 segments lacking sufficient support from consecutive CpGs were removed. Neighboring
1277 donor-specific segments and isolated EEC CpGs were then re-evaluated to recover locally
1278 fragmented regions. Units separated by ≤ 150 bp and by no more than one intervening
1279 reference CpG were connected when they shared the same dominant donor-specific WT
1280 state. Connected units were merged only when the resulting component contained an
1281 anchor segment comprising at least five CpGs; components without such an anchor
1282 remained separate. This reuniting step was introduced to reduce artificial fragmentation of
1283 spatially coherent EEC segments caused by isolated discordant or missing CpGs that
1284 interrupted the stringent consecutive-CpG and concordance criteria used during initial
1285 segmentation.

1286 **Assessment of spatial clustering of EEC events**

1287 To determine whether EEC events were spatially clustered beyond that expected by chance,
1288 we compared the observed distribution of consecutive EEC-associated CpGs with
1289 chromosome-matched randomized distributions. For each clone, CpGs within donor-specific
1290 EEC segments supported by that clone, together with clone-specific singleton EEC events,
1291 were considered EEC-associated CpGs. The number of EEC-associated CpGs on each
1292 chromosome was preserved, and their positions were randomly reassigned among
1293 reference CpGs on the same chromosome 100 times. Consecutive runs of EEC-associated
1294 CpGs were counted in the observed and permuted datasets, and the observed run-length
1295 distribution was compared with the permutation-derived expectation.

1296 To further assess the local spatial organization of EEC events, we compared the genomic
1297 distances between adjacent CpGs within EEC segments with those expected from the
1298 genome-wide CpG distribution. Consecutive reference CpG pairs located within the same
1299 donor-specific EEC segment were identified, and, for each donor and chromosome, an equal
1300 number of adjacent reference CpG pairs was randomly sampled from the corresponding
1301 chromosome. This procedure was repeated 100 times to generate the expected inter-CpG
1302 distance distribution. The observed median inter-CpG distance was compared with the
1303 permutation distribution using a one-sided empirical test.

1304 Based on the permutation-derived background distribution of consecutive EEC-associated
1305 CpGs, donor-specific EEC segments containing at least five CpGs were carried forward for
1306 subsequent cross-donor analyses.

1307 **Construction and annotation of cross-donor EEC regions**

1308 Donor-specific EEC segments containing at least five CpGs were then combined across
1309 donors, and overlapping intervals were merged to define cross-donor **EEC regions**.
1310 Because a single EEC region could contain distinct patterns of methylation change, each
1311 region was further partitioned into **EEC elements** according to clone-level methylation
1312 patterns. For each clone, CpGs differing from the corresponding donor-specific WT state
1313 were identified, and local change blocks were defined from clusters of at least three changed
1314 CpGs, allowing up to three intervening CpGs. Within each EEC region, these clone-specific
1315 change blocks were represented as a binary CpG-by-clone matrix, and preliminary element
1316 boundaries were defined where the pattern of changed clones shifted across neighboring
1317 CpGs. To reduce over-segmentation, adjacent preliminary elements with similar affected-
1318 clone and WT-state profiles were iteratively remerged, while very short elements were
1319 absorbed into neighboring elements.

1320 Final EEC elements were annotated for CpG-island context, genomic regulatory context, and
1321 gene-promoter overlap based on CpG-level genomic annotations. For EEC regions
1322 overlapping gene promoters, continuous raw methylation fractions were averaged across
1323 CpGs within each element separately for each clone. These element-level mean methylation
1324 fractions were subsequently used for promoter methylation–gene expression correlation
1325 analyses described in the corresponding Methods section.

1326 **Spatial recurrence and genomic organization of EEC segments**

1327 To assess the recurrence and positional concordance of EEC segments across individuals,
1328 each donor-specific EEC segment was assigned to its corresponding cross-donor EEC
1329 element. For each EEC element, donor support was defined by the presence of a donor-
1330 specific segment overlapping the element, and the number of supporting donors was
1331 recorded.

1332 Positional concordance of recurrent EEC segments was quantified for all pairs of segments
1333 from different donors belonging to the same EEC element. For each pair, the number of
1334 reference CpGs contained within each segment and within their genomic intersection was
1335 determined. Reciprocal CpG overlap was calculated as the fraction of CpGs in each
1336 segment covered by the other segment, and the smaller of the two fractions was used as the
1337 pairwise reciprocal overlap.

1338
$$R(A, B) = \min\left(\frac{|A \cap B|_{\text{CpG}}}{|A|_{\text{CpG}}}, \frac{|A \cap B|_{\text{CpG}}}{|B|_{\text{CpG}}}\right),$$

1339 where $|A|_{CpG}$ and $|B|_{CpG}$ denote the numbers of reference CpGs within segments A and B ,
1340 respectively.

1341 To examine the relationship between EEC-segment boundaries and CpG islands, donor-
1342 specific EEC segments overlapping CpG islands were identified and normalized relative to
1343 the corresponding CpG-island span, with its start and end boundaries defined as 0 and 1,
1344 respectively. When multiple CpG islands overlapped a segment, their outermost boundaries
1345 were used to define the span. For each segment, the nearer CpG-island boundary was
1346 selected, and segments aligned to the end boundary were mirrored ($x' = 1 - x$) so that both
1347 island boundaries were superimposed at 0. Thus, negative and positive coordinates
1348 represented positions outside and inside the CpG island, respectively. Aligned segment
1349 intervals were stacked, and the distribution of segment-boundary positions was visualized by
1350 kernel density estimation. Up to 4,000 segments were randomly sampled for display.

1351 **Genomic annotation and CpG context enrichment analysis of EEC regions**

1352 EEC regions were annotated using CpG-level genomic annotations. All reference CpGs
1353 located within each EEC region were classified according to CpG context as CpG island,
1354 shore, shelf, or open sea and assigned their corresponding genomic annotations. For each
1355 EEC region, the proportion of CpGs belonging to each CpG context was calculated, and the
1356 dominant CpG context and genomic annotation were defined as the most frequent
1357 annotation among CpGs within the region.

1358 Enrichment of EEC regions across CpG contexts was assessed using chromosome- and
1359 CpG-number-matched random regions. A genome-wide set of callable CpGs was defined
1360 from CpGs passing methylation-quality filtering. For each EEC region, 1,000 background
1361 intervals were generated on the same chromosome, each containing the same number of
1362 callable CpGs as the corresponding EEC region; candidate intervals overlapping observed
1363 EEC regions were excluded. CpG-context composition of both observed and randomized
1364 regions was then determined using the same reference CpG annotations. Regions for which
1365 complete matched background sets could not be generated were excluded from the
1366 comparison. For each CpG context, the observed proportion was calculated as the mean
1367 fraction of CpGs assigned to that context across EEC regions. For each permutation, the
1368 corresponding mean fraction was calculated across the matched background regions, and
1369 fold enrichment was defined as the observed proportion divided by the mean expected
1370 proportion and reported on a log2 scale. Two-sided empirical P values were calculated from
1371 the permutation distribution and adjusted across CpG contexts using the Benjamini–

1372 Hochberg method. The 95% confidence interval of the fold enrichment was defined by the
1373 2.5th and 97.5th percentiles of the permutation-derived fold-enrichment distribution.

1374 **Association with genome features**

1375 To assess the association between EEC regions and genomic features, we generated
1376 matched background regions for each early methylation region. Background regions were
1377 sampled exclusively from CpG sites interrogated by EM-seq to ensure that only assay-
1378 accessible genomic regions were considered. For each early methylation region, candidate
1379 background regions were generated as contiguous CpG windows on the same chromosome
1380 containing the same number of CpGs as the corresponding early methylation region. From
1381 the resulting candidate pool, 100 background regions were randomly sampled without
1382 replacement whenever possible; sampling with replacement was performed only when fewer
1383 than 100 eligible candidate regions were available. These matched background regions
1384 were used as permutation controls for all downstream enrichment analyses. This matching
1385 strategy preserved chromosome-specific genomic context, local CpG density, and genomic
1386 span, thereby minimizing potential biases in the enrichment analysis.

1387 To evaluate the local genomic environment surrounding early methylation regions, genomic
1388 annotations and chromatin features were quantified within a ± 10 -kb window centered on
1389 each early methylation region. Replication timing was evaluated within a ± 100 -kb window to
1390 reflect its broad domain-scale organization. Promoter and enhancer annotations were
1391 obtained from the Ensembl Regulatory Build (GRCh37). CTCF-binding sites were obtained
1392 from ENCODE colonic mucosa CTCF ChIP-seq data (ENCFF663OEV) and lifted over from
1393 GRCh38 to GRCh37. Imprinting control regions (ICRs) were obtained from the human
1394 imprintome⁶⁶. Histone modification (H3K4me3, H3K27ac and H3K27me3) and DNase I
1395 hypersensitivity signal tracks were obtained from the NIH Roadmap Epigenomics Project
1396 using the colonic mucosa reference epigenome (E075). Replication timing was quantified
1397 using the ENCODE/UW IMR90 Repli-seq track.

1398 For genomic annotations, including promoters, enhancers, CTCF-binding sites and ICRs, the
1399 number of overlapping features within each window was counted. For continuous signals,
1400 including histone modifications, DNase I hypersensitivity and replication timing, the overlap-
1401 length-weighted mean signal was calculated from the corresponding bigWig tracks. The
1402 observed value was defined as the mean feature value across all early methylation regions.
1403 For each of the 100 matched background sets, the same feature was quantified using an
1404 identical procedure to generate an empirical null distribution. Fold enrichment was calculated
1405 as the ratio of the observed mean to the mean of the permutation-derived background

means. Statistical significance was assessed using a two-sided empirical permutation test. The 95% permutation interval was defined as the 2.5th and 97.5th percentiles of the fold-enrichment values calculated by dividing the observed mean by each permutation-derived background mean.

Whole-transcriptome sequencing data processing

Sequencing reads were trimmed using Cutadapt⁶⁰ to remove adapter sequences and aligned to the GRCh37 human reference genome using STAR (v2.7.5a)⁶⁷. Gene expression was quantified using RSEM (v1.3.1)⁶⁸, which generated gene-level expected counts and transcripts per million (TPM) values.

To account for differences in library preparation methods between HC15 and HC16 (TruSeq Stranded Total RNA Gold Kit, Illumina) and the remaining donors (NEBNext Ultra II Directional RNA Library Prep Kit, New England Biolabs), batch effect correction was performed on RSEM-derived expected counts using ComBat-seq from the sva R package⁶⁹. The batch-corrected counts were subsequently converted to TPM values using gene-specific effective lengths estimated by RSEM. Both batch-corrected TPM values and original RSEM-derived TPM values are provided in **Supplementary Table 15** (Sheets 1 and 2, respectively).

All downstream quantitative gene expression analyses were performed using batch-corrected TPM values, except for the comparison of XIST expression between male and female clones (**Extended Data Fig. 9c**), for which original RSEM-derived TPM values were used, as batch correction tended to attenuate the pronounced sex-specific expression differences of XIST.

Promoter–gene annotation

Gene coordinates and transcription start sites (TSSs) were obtained from the same Ensembl gene annotation (Homo sapiens, GRCh37 release 87) used for RNA-seq alignment. Promoter elements were extracted from the Ensembl Regulatory Build annotation for GRCh37. Promoter–gene associations were determined using a custom script by identifying promoter elements located within 1 kb upstream of each TSS in a strand-aware manner. When multiple promoters were identified for a gene, the promoter closest to the TSS was retained. Using this approach, promoters were assigned to 21,830 genes.

Regression analysis of promoter associated EEC methylation and gene expression

To investigate the relationship between early methylation changes and gene expression while minimizing inter-donor differences in baseline expression, we integrated element-level EEC epigenotypes with matched RNA-seq data across clonally derived samples. For each EEC element associated with a gene promoter, its epigenotype (U/U, U/M, or M/M) was paired with the expression level (TPM) of the corresponding gene. For regression analysis, the three ordered epigenotypes were numerically encoded as 0, 0.5, and 1 for U/U, U/M, and M/M, respectively. To minimize inter-donor variation in baseline gene expression, expression values were converted to percentile ranks within each donor, ranging from 0 to 1, thereby representing the relative expression level of each gene among clones from the same donor.

To ensure that the relationship between EEC epigenotype and expression could be evaluated within individual donors, donors were retained for each EEC–gene pair only when at least two distinct epigenotypes were each represented by at least two clones. Donors were additionally required to have at least two clones with expression of the corresponding gene at TPM ≥ 1 . EEC–gene pairs represented by fewer than four clones after donor-level filtering were excluded.

The relationship between EEC epigenotype and donor-normalized expression rank was assessed using linear regression, with expression rank as the dependent variable and the numerically encoded EEC epigenotype as the independent variable. The regression coefficient (β) therefore represents the direction and magnitude of the association across the ordered epigenotypes, with a negative β indicating progressively lower relative expression from U/U to U/M to M/M and a positive β indicating the opposite relationship. When multiple EEC elements were associated with the same gene, the element with the smallest P value among those showing a negative association was retained for gene-level analysis; when no element showed a negative association, the element with the smallest P value was retained. P values for the resulting gene-level associations were adjusted for multiple testing using the Benjamini–Hochberg procedure, and associations with FDR < 0.05 were considered statistically significant.

Identification of later-acquired epigenetic changes (LECs)

Later-acquired epigenetic changes (LECs) were identified as spatially coherent methylation-state changes observed in individual colorectal clones relative to the corresponding donor-specific WT state and occurring outside previously defined EEC regions. For each donor and

clone, CpGs with non-missing methylation states that differed from the donor-specific WT state were identified, and runs of at least 10 consecutive changed CpGs were selected as initial LEC candidates. Candidate runs overlapping any previously defined EEC region were excluded, and the remaining overlapping candidate intervals across clones and donors were reduced into non-redundant candidate regions.

Clone-level support for each candidate region was evaluated using discrete methylation-state calls. Clones with measurements for <50% of CpGs were classified as low coverage; among the remaining clones, those showing a state different from or matching the donor-specific WT at $\geq 80\%$ of evaluable CpGs were classified as changed or WT-like, respectively, and the remainder as ambiguous. Candidate regions were further validated using continuous methylation fractions. For regions with at least five WT-like clones and seven evaluable CpGs, a CpG-specific WT baseline was defined as the median methylation fraction across WT-like clones, and both background stability and the magnitude and consistency of methylation shifts in changed clones were assessed relative to this baseline. A changed clone was considered validated when the background methylation profile was stable (regional MAD ≤ 0.05 and $\geq 90\%$ stable non-changed clones), its median methylation shift was ≥ 0.20 , $\geq 80\%$ of CpGs changed in the same direction, and $\geq 70\%$ differed from the WT baseline by at least 0.15. Final LEC regions were required to contain at least one validated changed clone, with $\leq 10\%$ ambiguous and $< 30\%$ low-coverage clones.

Finally, adjacent LEC regions separated by ≤ 150 bp were merged when they shared at least one changed clone in order to reduce over-fragmentation, and the corresponding changed donors and clones were combined across the merged interval. For downstream promoter analyses, LEC regions were annotated using CpG-level promoter annotations, assigning the promoter contributing the largest number of overlapping promoter CpGs when multiple promoters overlapped a region.

Gene family enrichment analysis

Gene-family enrichment analysis was performed separately for genes whose promoter regions overlapped EEC regions and LECs. Genes were first standardized to HGNC-approved gene symbols to ensure consistent matching between the promoter-associated gene sets and HGNC gene-group annotations. When a gene was represented by an outdated or alternative symbol in the input data, the symbol was converted to its corresponding HGNC-approved symbol based on HGNC records of previously used gene symbols. Genes that could not be unambiguously matched to an HGNC-approved symbol

were excluded. Pseudogenes were excluded, and the remaining successfully mapped genes constituted the gene universe for the enrichment analysis.

Each gene was subsequently annotated according to the gene groups defined by the HGNC Gene Groups resource. Because HGNC gene groups are not mutually exclusive, a gene annotated to more than one group was retained in each applicable group. Enrichment was assessed independently for the EEC- and LEC-associated gene sets. For each HGNC gene group, the observed number of genes belonging to the EEC- or LEC-associated set was compared with that expected based on the proportion of EEC- or LEC-associated genes in the corresponding gene universe. Specifically, the expected number was calculated as

$$E_g = \frac{N_{target} \times N_{universe}}{N_g}$$

where N_{target} is the total number of EEC- or LEC-associated genes, N_g is the number of genes belonging to a given HGNC gene group, and $N_{universe}$ is the total number of genes included in the corresponding analysis. Fold enrichment was calculated as the ratio of the observed to expected number of genes in each group.

Statistical significance was evaluated for each gene group using a two-sided Fisher's exact test based on a 2×2 contingency table comparing membership in the gene group with membership in the EEC- or LEC-associated gene set. Multiple-testing correction was performed separately for the EEC and LEC analyses using the Benjamini–Hochberg procedure, and gene groups with a false discovery rate (FDR) < 0.05 were considered significantly enriched.

Quantification of alternative APC transcript usage

Relative usage of the alternative APC transcripts initiated from promoters 1A and 1B was quantified directly from RNA-seq alignments. Because the two transcripts differ in their promoter-specific first exons but share the downstream exon structure, reads mapping to the shared portion of the gene cannot distinguish between transcripts 1A and 1B. We therefore quantified transcript usage using splice-supporting reads originating from their transcript-specific first-exon regions.

For each clone, aligned RNA-seq BAM files were queried at the two alternative first-exon regions on chromosome 5 (GRCh37): chr5:112,043,196–112,043,579 and chr5:112,073,582–112,073,623 (1-based coordinates). A read was assigned as supporting

the corresponding transcript when its leftmost aligned reference position fell within the respective first-exon interval and its CIGAR string contained an N operation, indicating a spliced alignment from the first-exon region toward a downstream exon. Unmapped reads, PCR/optical duplicates and reads with mapping quality <1 were excluded. Read names were collapsed within each first-exon region to prevent the same read from being counted more than once.

Because the remainder of the exon structure is shared between the two APC transcripts, only reads supporting the alternative first-exon regions were used to estimate their relative usage. The fraction of transcript 1A was calculated as

$$\text{Transcript 1A fraction} = \frac{N_{1A}}{N_{1A} + N_{1B}}$$

where N_{1A} and N_{1B} represent the numbers of splice-supporting reads originating from the transcript-specific first exons of transcripts 1A and 1B, respectively. Clones with ≤ 10 combined transcript-specific supporting reads ($N_{1A} + N_{1B}$) were excluded to reduce uncertainty associated with low read depth. Transcript 1A fractions were then compared between clones with and without LECs at the APC 1A promoter using a two-sided Wilcoxon rank-sum test.

Haplotype-resolved X-chromosome DNA methylation analysis

To investigate allele-specific DNA methylation across the X chromosome, we constructed an extended haplotype map for donor HC15 by integrating long-read germline phasing with allele-specific expression (ASE). PacBio long-read sequencing was performed for clone HC15-16, and heterozygous germline variants on chromosome X were phased using WhatsHap v2.3⁷⁰. Because long-read phasing generated multiple local phase blocks that did not span the entire chromosome, X-chromosome ASE information was used to extend the allelic orientation across independently phased blocks.

For RNA-guided extension of haplotype phasing, we identified X-chromosome SNPs showing reciprocal allelic expression between HC15 clones carrying opposite inactive-X (X_i) alleles. These SNPs provided a common allelic reference across expressed regions, allowing independently phased blocks to be oriented relative to the same X-chromosome haplotypes. The RNA-derived allelic orientation was compared with the WhatsHap-derived phase within each phase block, and blocks showing the opposite orientation were inverted. In this way, phase blocks not directly connected by individual long reads could be

consistently oriented using ASE information, and an RNA-guided phased germline VCF was generated.

Because all clones from donor HC15 share the same germline haplotypes, the extended haplotype information derived from HC15-16 was subsequently applied to EM-seq data from all HC15 clones. EM-seq reads from each clone were haplotype-tagged using the RNA-guided phased germline VCF and separated into haplotype 1 (HP1) and haplotype 2 (HP2) BAM files. Haplotype-specific DNA methylation was then quantified independently from the HP1 and HP2 BAM files using Bismark v0.22.3 (bismark_methylation_extractor). For visualization, CpGs within CpG islands were extracted using BEDTools v2.30.0, and their methylation levels were displayed separately for the two haplotypes across HC15 clones.

Estimation of X-chromosome inactivation timing

The timing of X-chromosome inactivation (XCI) was inferred by integrating somatic mutation-based phylogenies with inactive-X allelic states determined from allele-specific expression (ASE). First, for each donor, all clone pairs were classified according to whether they carried the same or different inactive-X alleles, and their phylogenetic relatedness was measured by the number of shared somatic mutations. Because clones diverging after XCI are expected to inherit the same inactive-X state, we determined the phylogenetic depth beyond which clone pairs consistently shared the same inactive-X allele, providing an upper bound on the molecular timing of XCI. Across donors, clone pairs sharing approximately 18 or more embryonic mutations consistently carried the same inactive-X allele, indicating that XCI had been established by this stage.

We next estimated the most likely timing of XCI by considering all possible combinations of XCI events on each phylogeny that were compatible with the observed inactive-X alleles. Because the same allelic pattern could be explained by fewer events occurring early or by multiple independent events occurring later, each compatible scenario s , requiring k_s independent XCI events, was assigned a probability

$$P_s = \left(\frac{1}{2}\right)^{k_s} ,$$

assuming an equal probability of inactivating either parental X chromosome at each independent event. The probability assigned to each scenario was divided equally among its constituent XCI events, such that each event contributed a weight of P_s/k_s . These event weights were distributed across the somatic mutation positions encompassed by the corresponding phylogenetic branches and summed across all compatible scenarios to obtain

the probability-weighted support, $W_d(m)$, at each mutation point m for donor d . The resulting distribution was normalized within each donor as

$$P_d(m) = \frac{W_d(m)}{\sum_{m'=0}^{20} W_d(m')},$$

such that each donor contributed a total probability of one. Based on the pairwise analysis indicating consistent XCI-state concordance beyond approximately 18 shared embryonic mutations, probability-based timing inference was restricted to the first 20 somatic mutations.

For HC50, the substantially larger number of clones resulted in a combinatorial increase in the number of XCI-event combinations compatible with the observed inactive-X states, making exhaustive enumeration of the complete phylogeny computationally prohibitive. We therefore generated 100 random subtrees of 20 clones while preserving both major lineages arising from the first bifurcation, with at least three clones retained from each lineage. The same probability-based analysis was performed independently for each subtree, and each resulting timing distribution was normalized to unit probability before averaging across the 100 subsamplings to obtain a single HC50 distribution. Finally, the HC50 distribution and those of the remaining donors were combined with equal donor weights to obtain the overall XCI timing distribution. The resulting distribution was smoothed using kernel density estimation and summarized by its peak and the 50% highest-density intervals.

Identification of U/U CpG islands and methylation categories on chrX

To characterize constitutively unmethylated CpG islands on the X chromosome, CpGs located within the pseudoautosomal regions (PARs) and CpGs with a male WT epigenotype of 0.5, which were considered low-confidence sites, were excluded from all analyses. After filtering, a total of 1,132,874 CpGs located in the non-PAR region of chromosome X were retained.

We next analyzed all 750 annotated CpG islands located in the non-PAR region of chromosome X. For each female clone, the mean epigenotype of all CpGs within each CpG island was calculated. CpG islands with a mean epigenotype below 0.2 in at least one female clone were retained for further analysis, resulting in 476 candidate CpG islands. The threshold of 0.2 was chosen to identify CpG islands in which the vast majority of constituent CpGs were unmethylated, thereby focusing on regions exhibiting robust loss of DNA methylation.

The methylation patterns of the 476 candidate CpG islands were subsequently classified into four categories based on predefined criteria applied to clone-level methylation profiles. **Category A** comprises CpG islands that were uniformly unmethylated across nearly all female clones. **Category B** comprises CpG islands exhibiting coordinated methylation patterns that followed the inferred clonal phylogeny. **Category C** comprises CpG islands displaying heterogeneous methylation patterns, with multiple epigenotypes interspersed across clones without an apparent phylogenetic structure. **Category D** comprises CpG islands exhibiting sporadic unmethylation in only a small subset of female clones, whereas the remaining clones consistently retained the canonical XCI-associated monoallelic methylation pattern (epigenotype = 0.5).

CpG-context enrichment analysis for unmethylated CpGs in female X chromosomes

To characterize the genomic context of CpGs that remained unmethylated on both male and female X chromosomes, we examined CpGs in the non-pseudoautosomal region (non-PAR) of the X chromosome. At each CpG, the consensus epigenotype was defined separately for males and females as the most frequently observed epigenotype across clones of the corresponding sex. CpGs with a WT epigenotype of 0 in both males and females (male WT = 0 and female WT = 0) were selected as the target set for enrichment analysis.

Each CpG was annotated according to its CpG-island context as CpG island, shore, shelf or open sea. All quality-filtered CpGs in the non-PAR X chromosome were used as the background, thereby accounting for the baseline distribution of CpG contexts among CpGs measurable in our dataset. For each CpG context c , the observed fraction was calculated as the number of target CpGs in that context divided by the total number of target CpGs. The expected fraction was defined as the number of all non-PAR CpGs in the same context divided by the total number of non-PAR CpGs. Fold enrichment was then calculated as

$$E_c = \frac{N_{target,c} / N_{target}}{N_{background,c} / N_{background}}$$

where $N_{target,c}$ and $N_{background,c}$ denote the numbers of target and background CpGs, respectively, assigned to context c , and N_{target} and $N_{background}$ denote their corresponding total numbers. Enrichment values were log2-transformed for visualization, such that a value of zero represents the expected frequency based on the non-PAR background, whereas positive and negative values indicate enrichment and depletion, respectively.

Boundary analysis of unmethylated CpG islands on chrX

To characterize the genomic environment surrounding unmethylated CpG islands, boundary-centered meta-profiles were generated using a representative female colorectal epithelial clone (HC15-01). CpG islands with a mean methylation level below 0.2 in HC15-01 were defined as unmethylated CpG islands. For each island, the first and last CpGs were defined as the left and right boundaries, respectively. Starting from each boundary, neighboring CpGs outside the CpG island were sequentially collected in both upstream and downstream directions until either 1,000 CpGs had been collected or another CpG island was encountered. Sampling was terminated upon reaching an adjacent CpG island to prevent neighboring CpG islands, particularly adjacent unmethylated CpG islands, from contributing to the boundary profiles.

For each flanking CpG, its position relative to the nearest CpG island boundary was recorded as the relative CpG index, allowing flanking regions from all unmethylated CpG islands to be aligned and aggregated. The observed epigenotype of each CpG in HC15-01 (0, 0.5, or 1) was summarized at every relative CpG index. In parallel, the fraction of CpGs belonging to each CpG context category (island, shore, shelf, and open sea) was calculated at every relative CpG index to generate boundary-centered CpG context profiles.

RNA-seq variant calling for allele-specific expression analysis

RNA-seq variant calling was performed using the Genome Analysis Toolkit (GATK, v4.0.0.0)⁷¹ with reads aligned to the GRCh37 human reference genome. Duplicate reads were marked using MarkDuplicates, followed by SplitNCigarReads to process spliced alignments. Read group information was standardized using AddOrReplaceReadGroups, and base quality score recalibration was performed using BaseRecalibrator and ApplyBQSR. Single-nucleotide variants (SNVs) and small insertions and deletions (indels) were called using HaplotypeCaller. The resulting variants were filtered using VariantFiltration, excluding variants with Fisher Strand (FS) > 30.0 or Quality by Depth (QD) < 2.0. Germline heterozygous SNVs identified from matched whole-genome sequencing (WGS) data were used as informative sites for subsequent allele-specific expression (ASE) analysis of X-linked genes, as described below.

Allele-specific expression analysis of the X chromosome

Allele-specific expression (ASE) was assessed in female clones using RNA-sequencing reads overlapping informative germline heterozygous single-nucleotide variants (SNVs) on

the X chromosome. Candidate SNVs were filtered to remove somatic variants and were required to show a variant allele fraction (VAF) between 0.1 and 0.9 in matched blood DNA. Variants in intergenic, upstream, downstream or intronic regions and those associated with pseudogenes were excluded, and the analysis was further restricted to a manually curated set of informative X-chromosome SNPs. For each SNP, reference (RD) and alternative (AD) allele-supporting reads were quantified; allele counts supported by fewer than three reads were set to zero, after which total depth (TD) and VAF were recalculated.

Because biallelic expression from the X chromosome can span a broad range of allelic ratios, from near-balanced expression to highly asymmetric expression with only low-level transcription from the inactive X chromosome, a fixed allelic-ratio threshold could fail to detect weak expression from the inactive allele. To account for this variability, we estimated the predominant VAF of each gene across female clones using kernel density estimation. SNP-level ASE was then evaluated using thresholds adapted to the resulting gene-level VAF distribution: genes with a central VAF peak (>0.2 and <0.8) were tested against allele fractions of 0.2 and 0.8, whereas genes with more strongly allele-biased distributions were tested against 0.03 and 0.97. SNPs with a one-sided binomial $P < 0.05$ were classified as supporting monoallelic expression (MAE), whereas the remaining informative SNPs were classified as supporting biallelic expression (BAE). Depth-dependent filtering was applied to avoid assigning MAE from insufficiently covered monoallelic observations.

For each gene–clone pair, SNP-level classifications were integrated as the proportion of informative SNPs supporting BAE. Gene–clone pairs were classified as BAE when this proportion exceeded 0.66 for genes with a central VAF peak or 0.33 for genes with an allele-biased peak; remaining informative pairs were classified as MAE, and those lacking sufficient information were designated LowDepth.

Integration of promoter methylation and allele-specific expression

To assess whether promoter demethylation was accompanied by transcriptional reactivation, DNA methylation and ASE were integrated at the gene–clone level using female clones with matched methylation and RNA-sequencing data. A gene was considered to have a U/U promoter state in a given clone when at least one associated promoter CpG island had a mean methylation level <0.2 , and ASE was evaluated specifically in the corresponding U/U clones.

For each gene, BAE frequency (BAE freq) and BAE extent were quantified both across all RNA-profiled clones and specifically among clones with a U/U promoter state. **BAE freq_{RNA}**

was calculated as $N_{BAE} / (N_{BAE} + N_{MAE})$ across all informative RNA-profiled clones, whereas **BAE freq_{U/U}** was calculated using only clones with a U/U promoter state. LowDepth and uninformative observations were excluded from both calculations.

BAE extent quantified the degree of allelic balance among BAE-classified clones and was defined as the minor-allele fraction, $\min(VAF, 1 - VAF)$. Values were first averaged across informative SNPs within each gene–clone pair and subsequently across clones for each gene. **BAE extent_{RNA}** was calculated across all BAE-classified RNA-profiled clones, whereas **BAE extent_{U/U}** was calculated specifically among BAE-classified clones with a U/U promoter state.

L1 promoter methylation analysis

Promoter methylation was analyzed for previously identified reference and non-reference L1 source loci¹¹. L1 transcription is driven by an internal promoter located within the 5' region of the element, and promoter methylation was quantified using CpGs located from +1 to +250 bp relative to the L1 transcription start site (TSS), as previously described¹¹. For reference L1 loci, CpG methylation fractions within this interval were extracted directly from genome-wide EM-seq data according to the genomic position and orientation of each source element. For non-reference loci, donor-specific germline presence was determined from matched blood WGS, and customized reference genomes incorporating the corresponding L1 insertions were generated for donors carrying each locus. EM-seq reads were realigned to the customized references using Bismark⁵⁸, and methylation fractions were extracted across the same promoter interval. Only colorectal clones represented in the corresponding WGS-derived phylogenetic trees were retained for downstream analyses.

Loss of methylation (LoM) was evaluated for clone–source combinations with at least five evaluable promoter CpGs. A clone was classified as showing LoM when at least 80% of evaluable CpGs had methylation fractions ≤ 0.65 and the mean promoter methylation fraction was ≤ 0.65 . WT-like clones were defined as those for which at least 80% of evaluable CpGs had methylation fractions ≥ 0.80 and the mean promoter methylation fraction was ≥ 0.75 . For donor–source combinations with at least three WT-like clones, a candidate LoM was retained only when the difference between the median methylation of WT-like clones and the minimum mean methylation among LoM clones was ≥ 0.25 . Donor–source combinations with fewer than three WT-like clones and at least one LoM clone were retained as global-LoM candidates. Candidate loci were subsequently reviewed using clone-by-CpG promoter methylation profiles, and loci with ambiguous or inconsistent patterns were excluded before downstream analysis.

1757 L1 genotypes were independently determined from matched blood WGS. For both reference
1758 and non-reference loci, an L1 variant allele fraction (VAF) was calculated from supporting
1759 reads, with loci having VAF <0.15 classified as absent, VAF >0.85 as homozygous and
1760 intermediate values as heterozygous. For non-reference insertions, loci with an inferred
1761 insertion length <200 bp were additionally classified as absent. Genotypes were converted
1762 to estimated L1 copy number (0, 1 or 2 copies) and linked to the corresponding methylation
1763 measurements. Where multiple non-reference labels mapped to the same curated biological
1764 source, duplicate representations were collapsed before source-level analyses to avoid
1765 double counting.

1766 For each informative clone–source combination, the mean promoter methylation fraction was
1767 discretized into three methylation states: 0 for values <0.25 , 0.5 for values from 0.25 to
1768 <0.75 and 1 for values ≥ 0.75 . To maintain an interpretable relationship between promoter
1769 state and source copy number, states 0 and 1 were considered interpretable for both one-
1770 and two-copy loci, whereas state 0.5 was retained only for two-copy loci because its allelic
1771 interpretation was ambiguous at single-copy loci. States 0 and 0.5 were considered LoM
1772 states. Source-level methylation-state frequencies were calculated across copy-number-
1773 interpretable clone–source observations, whereas germline source-copy-number
1774 frequencies were calculated independently across genotyped donors. The donor-specific
1775 consensus promoter state for each source was defined as the modal epigenotype among
1776 informative colorectal clones.

1777 The developmental timing and clonal sharing of promoter methylation changes were inferred
1778 using the WGS-derived colorectal phylogenies. For each donor and L1 source, the
1779 consensus promoter state was added as a pseudo-tip adjacent to the root, and ancestral
1780 promoter states were reconstructed by maximum parsimony using the MPR algorithm
1781 implemented in phangorn⁶⁵. Ambiguous ancestral states were resolved, where possible,
1782 using local phylogenetic support according to the same procedure used for CpG-level EEC
1783 reconstruction. Branch-specific promoter-state changes were identified by comparing
1784 reconstructed states between connected nodes, and internal-branch changes for which at
1785 least half of the descendant clones had missing or unresolved promoter states were
1786 masked. Among copy-number-interpretable clone–source observations showing LoM, a
1787 change was considered ancestrally shared when the reconstructed root was already in a
1788 LoM state (0 or 0.5), or when a LoM-producing change occurred on an internal ancestral
1789 branch and was retained in the descendant clone without a subsequent terminal-branch
1790 change. LoM associated with a promoter-state change on the terminal branch was classified
1791 as non-shared. Observations whose terminal state did not differ from the reconstructed root

state, or for which timing could not be resolved, were treated separately and excluded from the shared-versus-non-shared timing comparison. Source-level proportions of shared and non-shared LoM were calculated across classified LoM clone–source observations.

L1 transduction events were identified following the previously described approach¹¹. Briefly, somatic L1 retrotranspositions were called using MELT⁷², TraFiC-mem⁴², DELLY⁷³ and xTea⁷⁴ and were manually reviewed for supporting features including poly(A) tails and target-site duplications. A retrotransposition event was classified as a 3' transduction when the inserted segment mapped to a unique, non-repetitive genomic sequence with a full-length L1 located within 15 kb upstream, allowing the corresponding source L1 to be assigned. For each source-assigned transduction event, the promoter state of the corresponding source L1 was examined in the same clone using the LoM criteria defined above.

To assess coordinated promoter demethylation across L1 sources, a consensus promoter state was defined for each source as the modal state across informative clones; in the case of a tie, the higher methylation state was selected. For each clone, a source was considered demethylated relative to the consensus when its promoter state was lower than the corresponding consensus state, and the number of such sources was counted across loci. Late-stage clonal expansion was defined as clonal divergence occurring after the ancestral lineage had accumulated at least 2,000 shared somatic mutations. In HC50, enrichment of late-stage clonal expansion within the lineage showing coordinated L1 promoter demethylation was assessed using a two-sided Fisher's exact test, comparing the numbers of late and non-late clones in the target lineage with those among the remaining HC50 clones.

To assess whether promoter methylation similarity reflected clonal ancestry, all informative colorectal clone pairs were evaluated separately for each donor and L1 source. The number of shared somatic mutations for a clone pair was defined as the WGS-tree branch length from the root to their most recent common ancestor (MRCA). To focus the analysis on sources exhibiting promoter methylation variation, only clone pairs in which at least one member showed LoM were retained. Observed concordance (C_{obs}) was defined as 1 when the two clones had identical promoter states (0, 0.5 or 1) and 0 otherwise. For each donor–source combination, the expected concordance (C_{exp}) was calculated as the mean observed concordance across the same eligible LoM-containing clone-pair pool, thereby accounting for the baseline state distribution of that source. Prevalence-adjusted concordance was calculated as

$$C_{adj} = \frac{C_{obs} - C_{exp}}{1 - C_{exp}},$$

with donor–source combinations for which $C_{exp} = 1$ excluded as non-informative. Clone-pair comparisons were grouped according to the number of shared somatic mutations. Adjusted concordance values were pooled across evaluable clone-pair–source comparisons within each interval, and 95% confidence intervals were estimated from 2,000 bootstrap resamplings with replacement within each shared-mutation interval.

Methylation-based refinement of multifurcating phylogenies

To evaluate whether early methylation changes could resolve unresolved phylogenetic relationships, we artificially collapsed selected bifurcating nodes of the somatic mutation-based phylogeny into multifurcations. Each multifurcation consisted of three or more direct descendant lineages, and all possible binary refinements were generated by merging every pair of descendant lineages while leaving the remaining lineages unchanged. To avoid circularity, we excluded early methylation elements that were identified as EEC elements exclusively in the donor being tested, retaining only elements supported independently in other donors.

For each candidate refinement, early methylation elements supporting the proposed bifurcation were identified. An element was considered supporting when (i) at least 80% of clones within the merged lineage had available methylation measurements, (ii) all observed clones within the merged lineage shared an identical methylation state (0, 0.5, or 1), (iii) at least 80% of clones outside the merged lineage were observed, (iv) at least 80% of these outside clones shared a common methylation state, and (v) the methylation state of the merged lineage differed from the majority state of the remaining clones.

To further evaluate each supporting element, we calculated the minimum number of methylation-state transitions required to explain its distribution across the candidate phylogeny using the Fitch parsimony algorithm. Candidate refinements were ranked according to the number of supporting elements exhibiting the minimum parsimony score among all candidate topologies. The candidate supported by the largest number of minimum-parsimony elements was selected as the predicted refinement.

Prediction accuracy was evaluated by comparing the selected refinement with the original somatic mutation-based phylogeny. A prediction was considered correct when the merged lineage formed a monophyletic clade in the original tree. For each predicted refinement, we also calculated the number of shared somatic mutations separating the merged lineage from

its ancestral branch by summing the branch lengths along the path from the root to the most recent common ancestor of the merged clones in the original phylogeny. To quantify prediction confidence, we defined a confidence score as

$$Confidence = \frac{N_{best}}{N_{best} + N_{second}}$$

where N_{best} and N_{second} denote the numbers of minimum-parsimony supporting elements for the best- and second-ranked candidate refinements, respectively. The association between prediction confidence and reconstruction accuracy was assessed using logistic regression.

References (Methods)

50. Fujii, M., Matano, M., Nanki, K. & Sato, T. Efficient genetic engineering of human intestinal organoids using electroporation. *Nat. Protoc.* **10**, 1474–1485 (2015).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
54. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
55. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
56. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
57. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
58. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for

1885 Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

1886 59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*
1887 *Methods* **9**, 357–359 (2012).

1888 60. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing
1889 reads. *EMBnet J.* **17**, 10 (2011).

1890 61. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes
1891 and exomes. *Bioinformatics* **34**, 867–868 (2018).

1892 62. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human
1893 genome. *Nature* **500**, 477–481 (2013).

1894 63. Vandiver, A. R., Idrizi, A., Rizzardi, L., Feinberg, A. P. & Hansen, K. D. DNA methylation
1895 is stable during replication and cell cycle arrest. *Sci. Rep.* **5**, 17911 (2015).

1896 64. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. Mclust 5: Clustering, classification
1897 and density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016).

1898 65. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

1899 66. Jima, D. D. *et al.* Genomic map of candidate human imprint control regions: the
1900 imprintome. *Epigenetics* **17**, 1920–1943 (2022).

1901 67. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
1902 (2013).

1903 68. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with
1904 or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

1905 69. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package
1906 for removing batch effects and other unwanted variation in high-throughput experiments.
1907 *Bioinformatics* **28**, 882–883 (2012).

1908 70. Patterson, M. *et al.* WhatsHap: Weighted haplotype assembly for future-generation
1909 sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).

1910 71. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
1911 analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

72. Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
73. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
74. Chu, C. *et al.* Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat. Commun.* **12**, 3836 (2021).

Data availability

Whole-genome, DNA methylation, and transcriptome sequencing data are deposited in the European Genome-phenome Archive (EGA) with accession EGAS00001008522 and available for general research use after publication of this manuscript. Human reference genome (GRCh37) is available at NIH websites (https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000001405.13).

Code availability

In-house scripts for analyses are deposited on GitHub (https://github.com/ju-lab/colon_epimosaicism). Our in-house scripts are available upon request of reviewers during the peer-review process.

Acknowledgements

We thank Jong-Eun Park and Jihun Kim at KAIST for their fruitful comments and discussions. This work was supported by grants, from the National Research Foundation (NRF) funded by the Ministry of Science and ICT and Ministry of Health & Welfare, Republic of Korea (grant #: RS-2026-25550010 to Y.S.J.); from the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare of Republic of Korea (grant #: RS-2024-00440005 to Y.S.J.); from the Korea-US Collaborative Research Fund (KUCRF), funded by the Ministry of Science and ICT and Ministry of Health & Welfare, Republic of Korea (grant #: RS-2025-16064302); and Suh Kyungbae Foundation (SUHF-18010082 to Y.S.J.).

Author contributions

H.W., Y.L., and Y.S.J. conceived the study. H.W., Y.L., C.H.N., M.C., J.Y.K., J.Y., H.W.K., conducted clonal expansion of the colorectal epithelial cells. Ji.W.P., S.-Y.J., and M.J.K. collected colorectal specimens and clinical histories of the patients. S.A.O. conducted

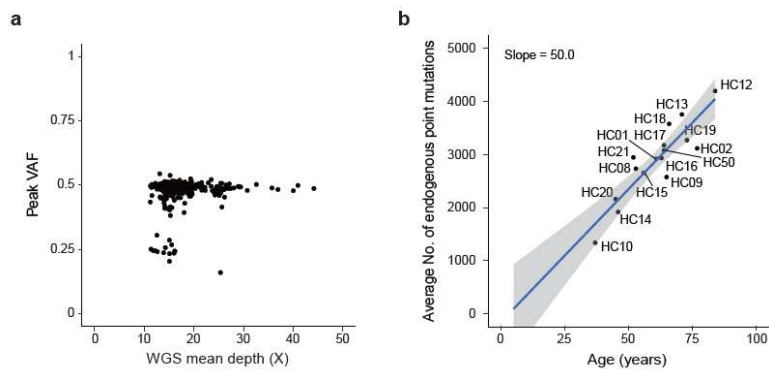
1942 genome sequencing. H.W. and Y.L. conducted most genome and statistical analyses with a
1943 contribution of C.H.N., B.L., Jo.L., J.-H.P., and J.A., J.W.P. and Ju.L. contributed to large-
1944 scale genome data management. D.-S.L., J.W.O., and J.H. participated in the data
1945 interpretation. B.L. developed epigenome browser. H.W., Y.L. and Y.S.J. wrote the
1946 manuscript with contributions from all the authors. Y.S.J. supervised the overall study.

1947 **Competing interests**

1948 Y.S.J. is a co-founder of Inocras Inc.. The other authors declare no competing interests.

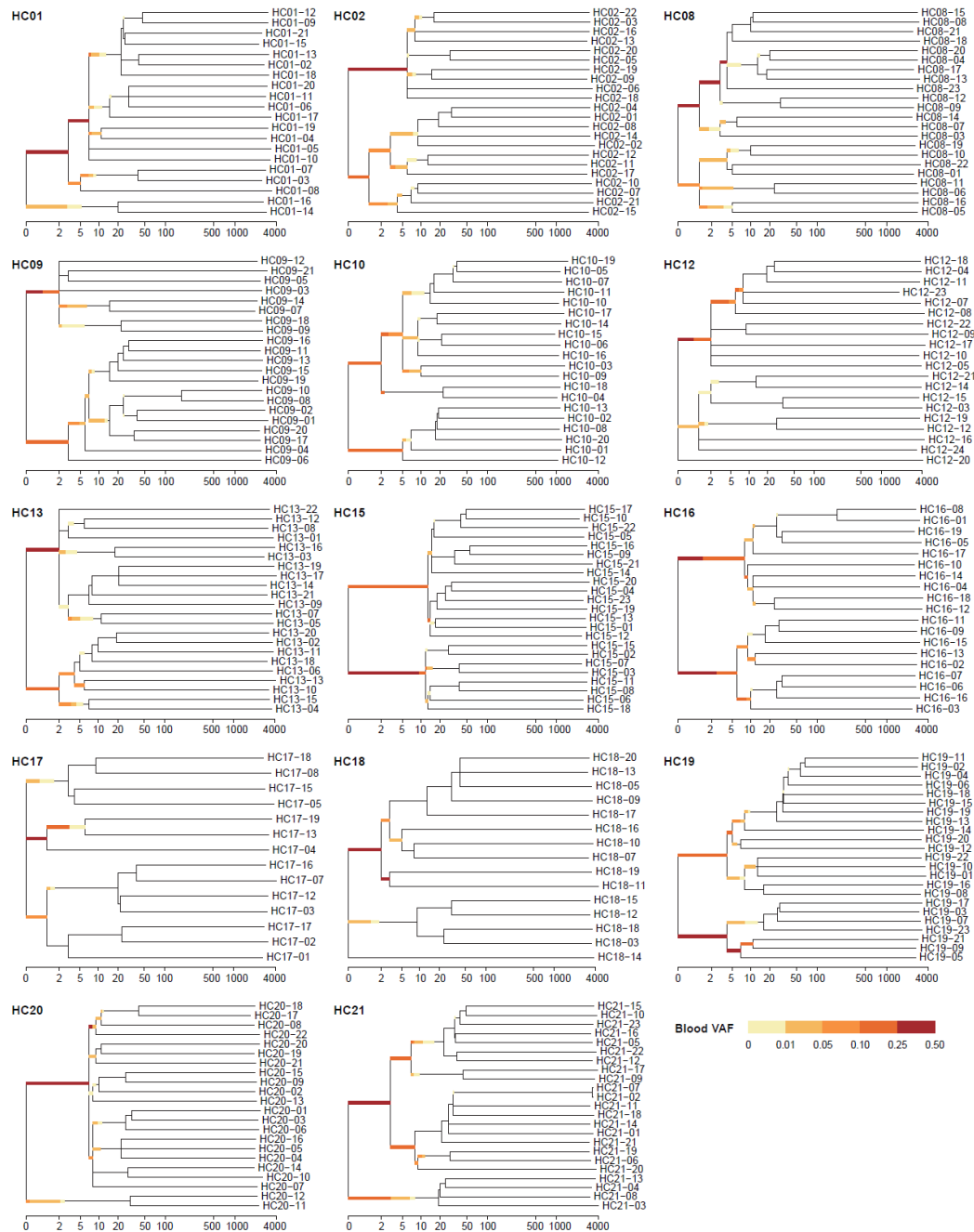
1949

Extended data legends



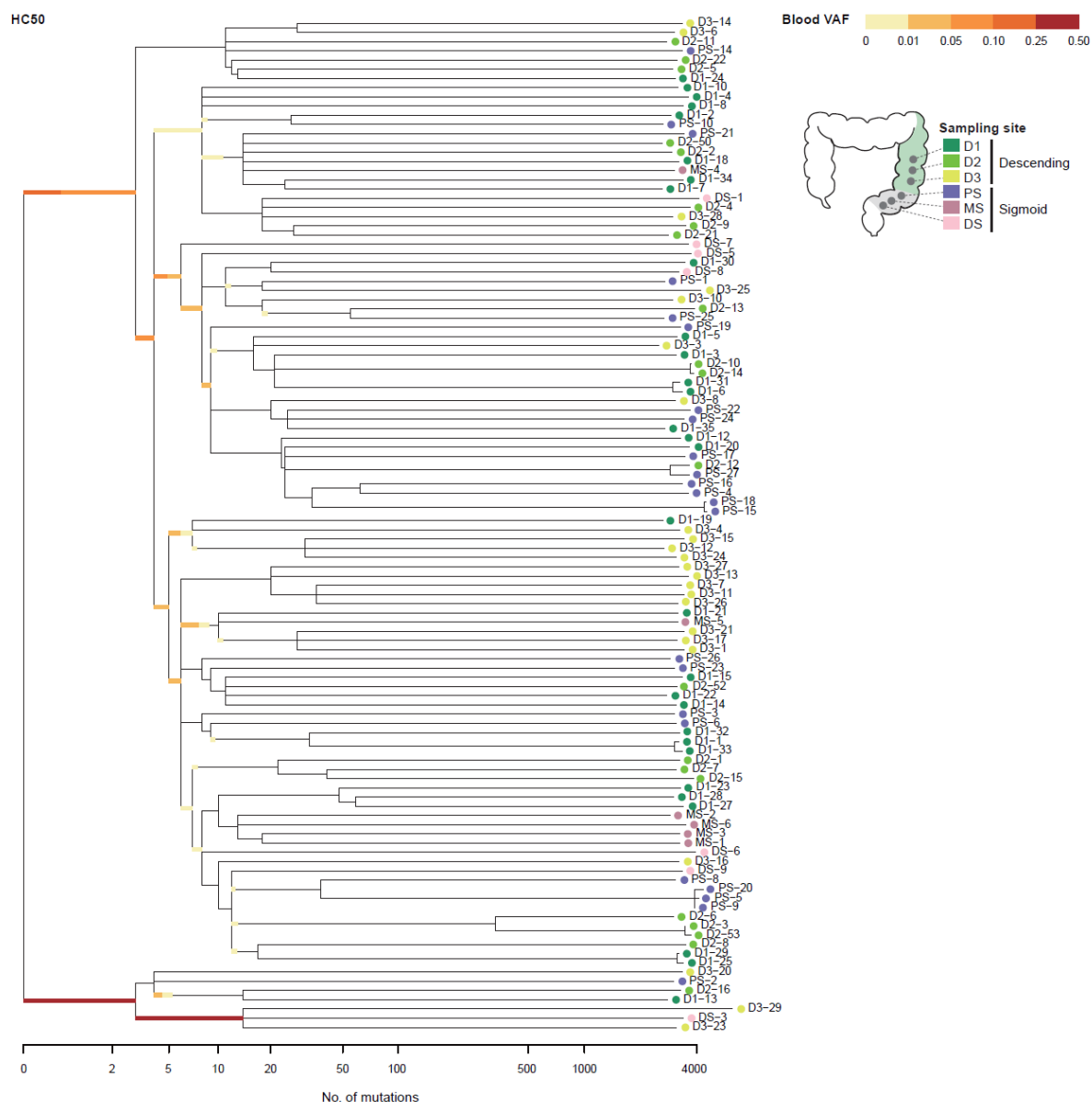
Extended Data Fig. 1 | Clonality and age-associated accumulation of somatic mutations in colorectal organoids.

a, Mean WGS sequencing depth and peak variant allele frequency (VAF) of somatic SNVs across colorectal organoid clones. Peak VAFs were predominantly centred around 0.5, consistent with clonal origin. **b**, Linear regression of donor age against the mean number of endogenous point mutations in colorectal organoid clones from 16 individuals. The blue line indicates the regression fit (50.0 point mutations per year), and the shaded area indicates the 95% confidence interval.



Extended Data Fig. 2 | Somatic-mutation-based developmental phylogenies of colorectal organoid clones.

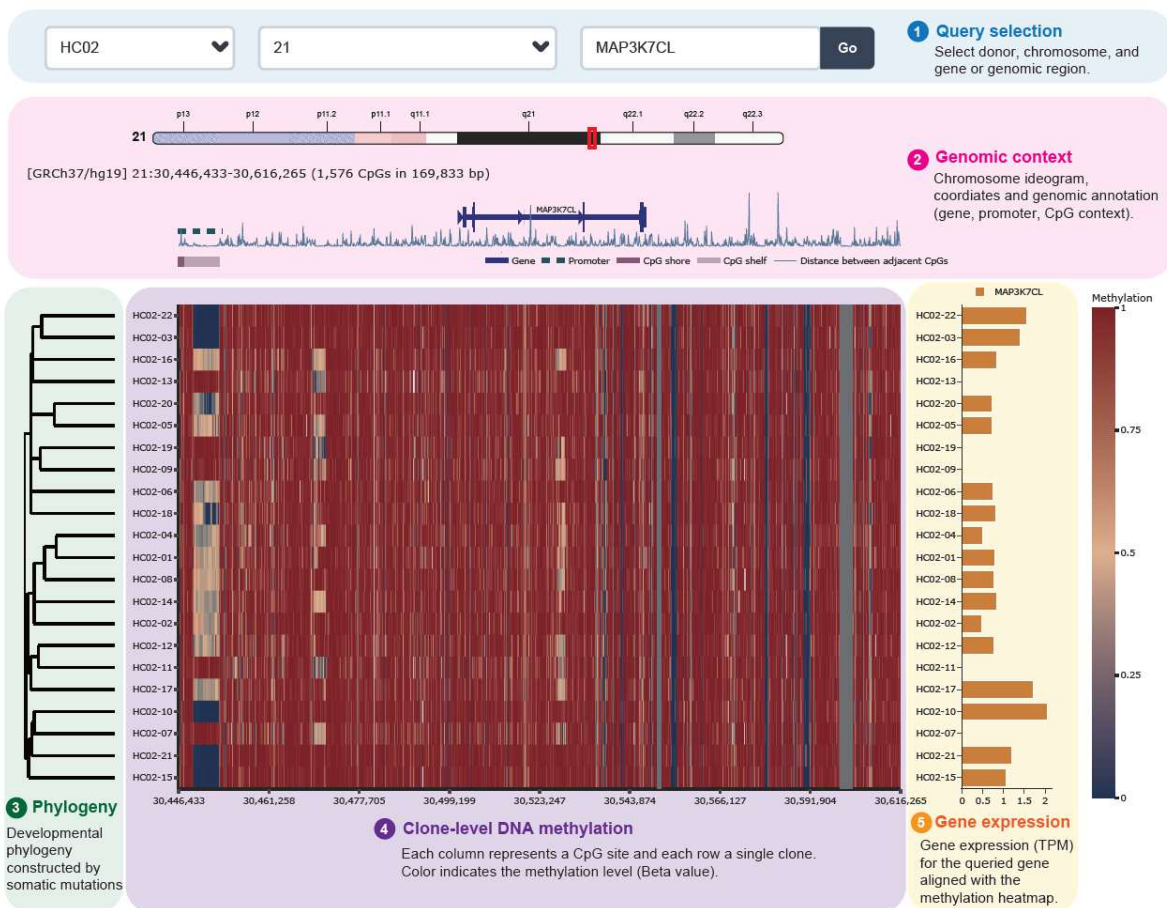
Phylogenetic trees reconstructed from somatic mutations in colorectal organoid clones from the remaining 14 donors included in downstream analyses. Each tip represents a clonally expanded colorectal organoid, and branch lengths indicate the number of somatic mutations assigned to each branch. Internal branches are coloured according to the variant allele frequencies (VAFs) of branch-defining mutations in matched peripheral blood, as indicated by the colour scale.



Extended Data Fig. 3 | Developmental phylogeny and spatial distribution of colorectal organoid clones from HC50.

The phylogenetic tree was reconstructed from somatic mutations shared among colorectal organoid clones from donor HC50. Each tip represents a clonally expanded colorectal organoid, and branch lengths indicate the number of somatic mutations assigned to each branch. Coloured circles adjacent to the tips indicate the anatomical sampling sites of individual clones, with sampling locations in the descending and sigmoid colon shown in the inset. Internal branches are coloured according to the VAFs of branch-defining mutations in matched peripheral blood, as indicated by the colour scale.

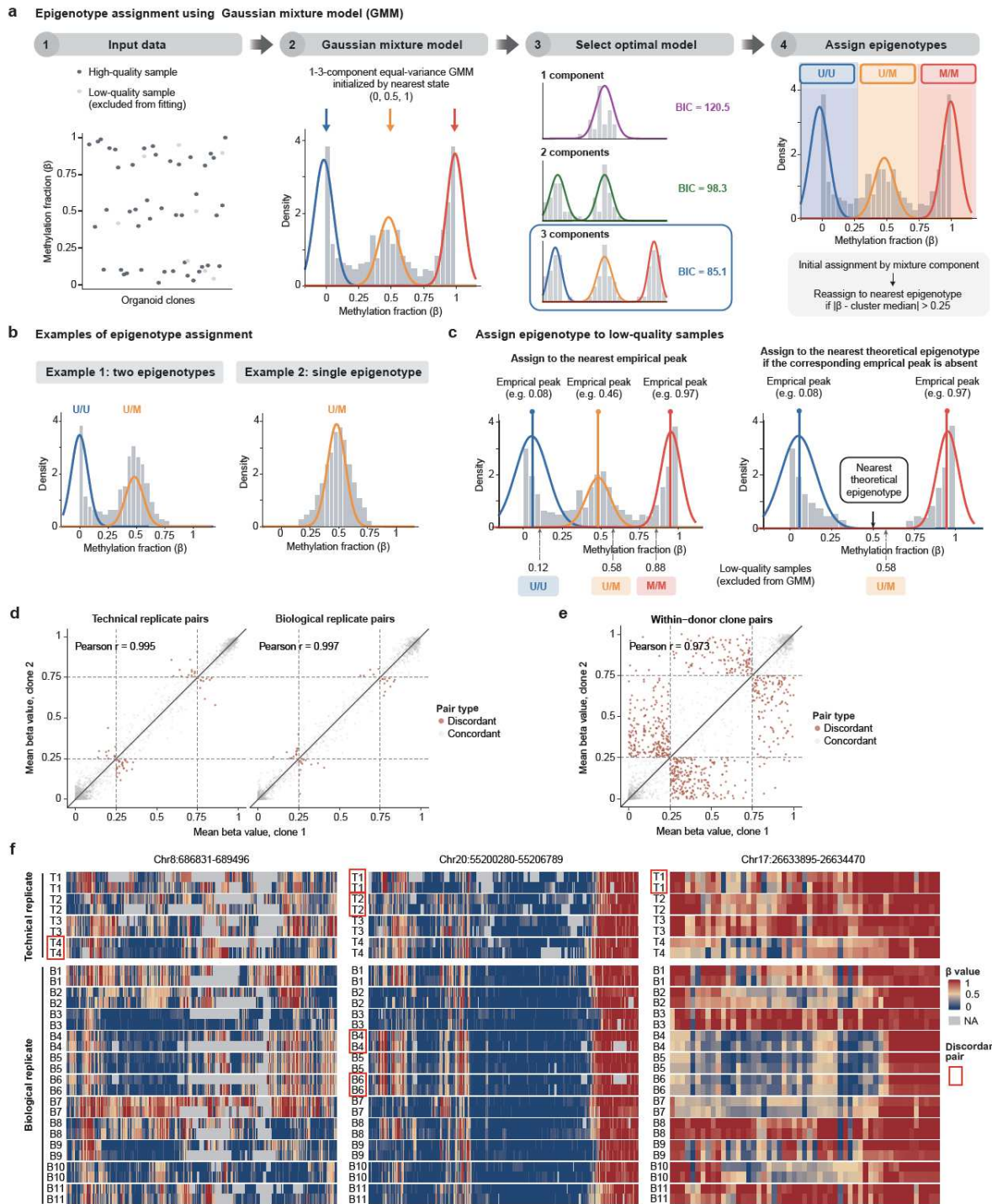
Methylation Browser



1980

1981 **Extended Data Fig. 4 | Interactive browser for clone-level DNA methylation**
 1982 **landscapes.**

1983 The Methylation Browser enables visualization of DNA methylation patterns at user-selected
 1984 genomic loci across individual clones. Users can select a donor, chromosome and gene or
 1985 genomic region and simultaneously explore genomic annotations, somatic mutation-based
 1986 phylogeny, clone-level CpG methylation and matched gene expression. The MAP3K7CL
 1987 locus in donor HC02 is shown as an example.



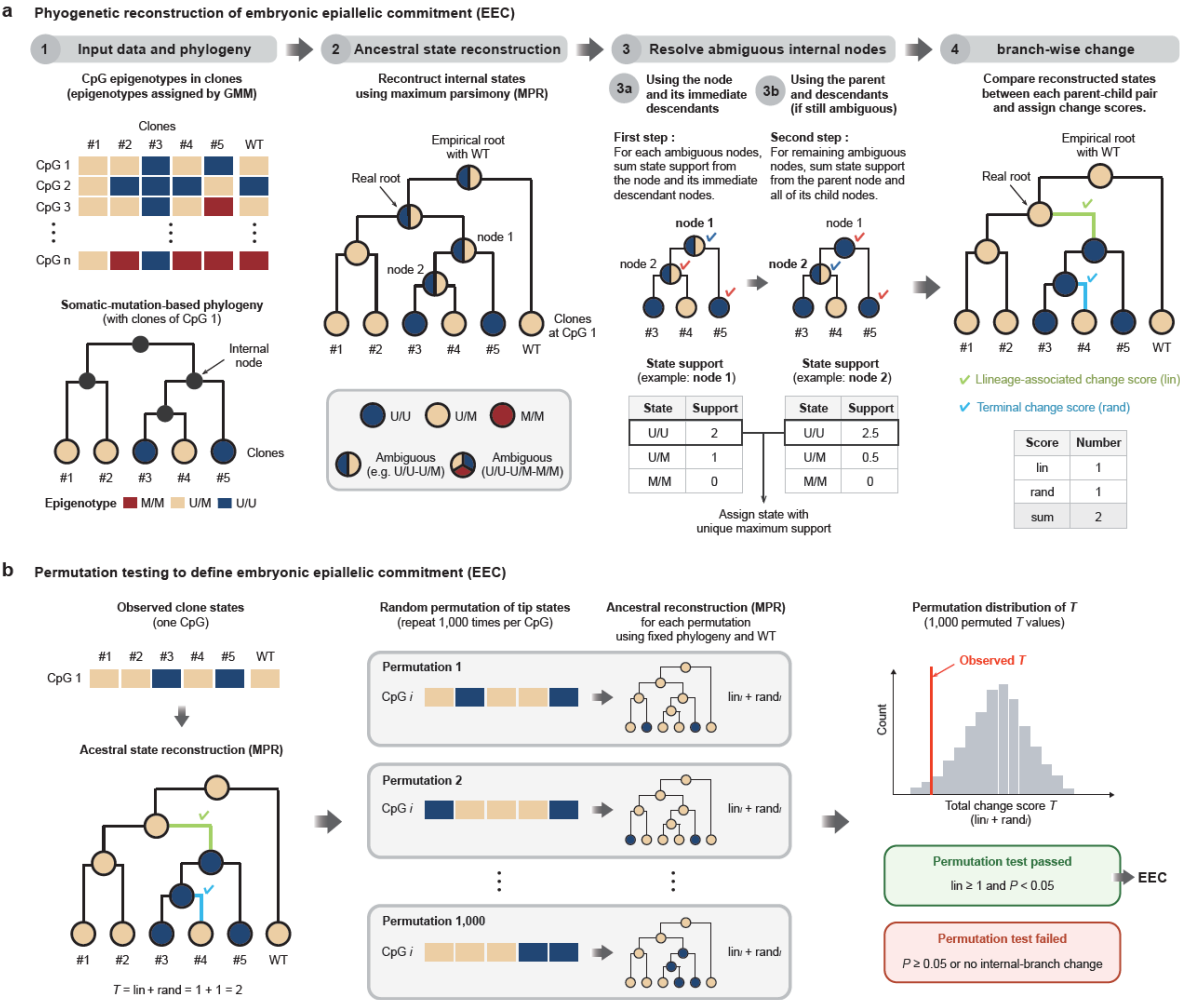
Extended Data Fig. 5 | Epigenotype assignment and concordance across replicate and within-donor clone pairs.

a, Workflow for CpG-specific epigenotype assignment. **b**, Representative CpGs resolved into two epigenotype states or a single epigenotype state using the Gaussian mixture model (GMM)-based approach. **c**, Post hoc assignment of low-quality samples using empirical methylation peaks estimated from high-quality clones. **d**, Comparison of mean β values for

1995 individual CpG islands between technical replicate pairs (left) and biological replicate pairs
1996 (right). Grey points indicate concordant island-level epigenotype assignments and red points
1997 indicate discordant assignments. Dashed lines mark the epigenotype classification
1998 boundaries at $\beta = 0.25$ and 0.75 , and the solid diagonal indicates identical β values between
1999 paired samples. For visualization, 2,000 island-pair observations were randomly sampled
2000 from all evaluable observations for each replicate type. Pearson correlation coefficients were
2001 calculated using the full datasets. **e**, Comparison of mean β values for individual CpG
2002 islands between within-donor clone pairs, displayed as in **d**. For visualization, 2,000 island-
2003 pair observations were randomly sampled; the Pearson correlation coefficient was calculated
2004 using the full dataset before subsampling for visualization. **f**, Representative CpG islands
2005 showing discordant epigenotype assignments in technical and biological replicate pairs.
2006 Heatmaps show β values of consecutive CpGs across each island for technical (T1–T4) and
2007 biological (B1–B11) replicate pairs, with paired samples displayed in adjacent rows. Red
2008 boxes highlight pairs with discordant island-level assignments.

2009

2010



2011

2012 **Extended Data Fig. 6 | Phylogeny-based identification of embryonic epiallelic**

2013 **commitment (EEC) events.**

2014 **a**, Reconstruction of CpG epigenotype changes along somatic phylogenies to identify

2015 lineage-associated changes assigned to internal branches and potentially established during

2016 early development. CpG epigenotypes were mapped onto fixed phylogenies, ancestral

2017 states were reconstructed by maximum parsimony, ambiguous internal nodes were

2018 resolved, and branch-wise lineage-associated (lin) and terminal (rand) changes were

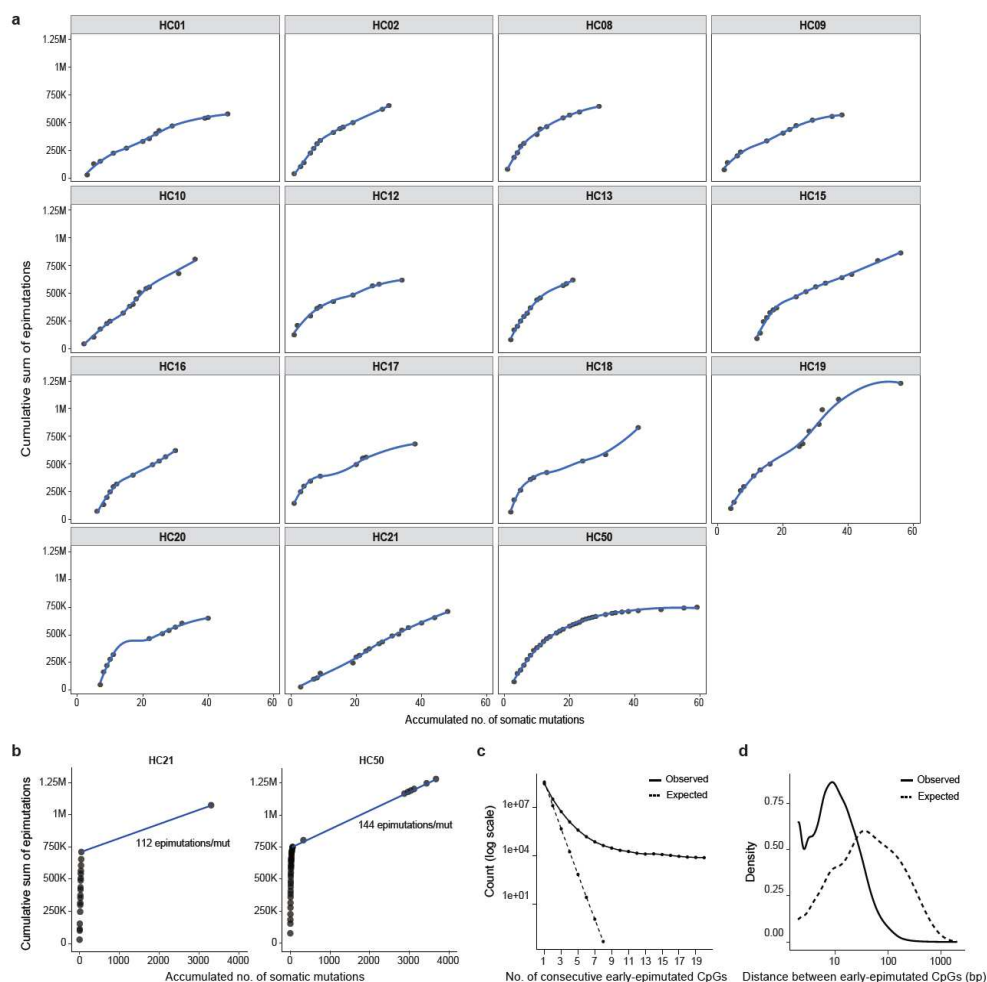
2019 assigned. **b**, Permutation testing to determine whether the observed lineage-associated

2020 pattern exceeded that expected by chance. For each CpG, clone epigenotypes were

2021 permuted 1,000 times on the same fixed phylogeny to generate a null distribution of the total

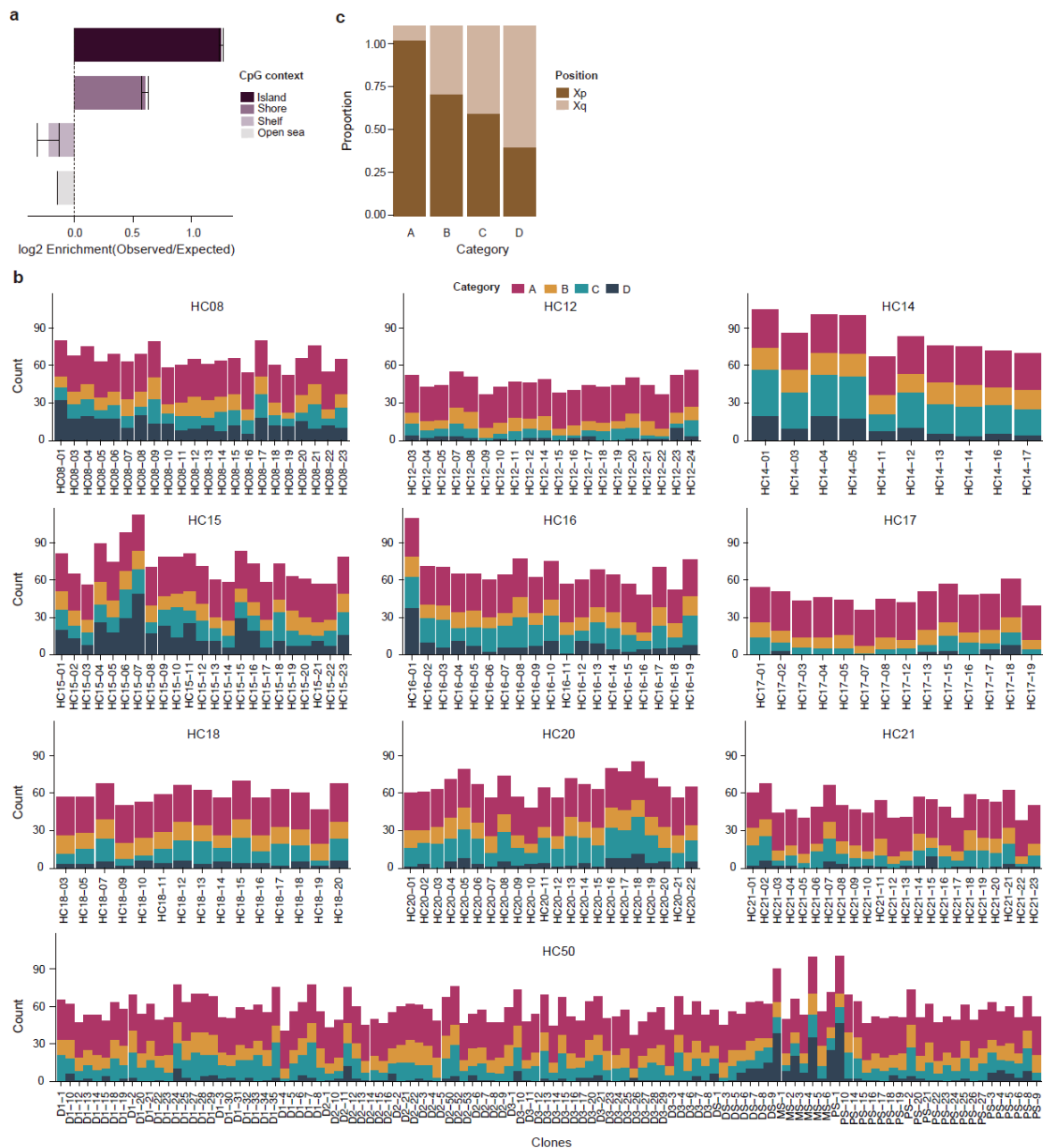
2022 change score, $T = \text{lin} + \text{rand}$. Smaller observed T values indicate stronger phylogenetic

2023 clustering.



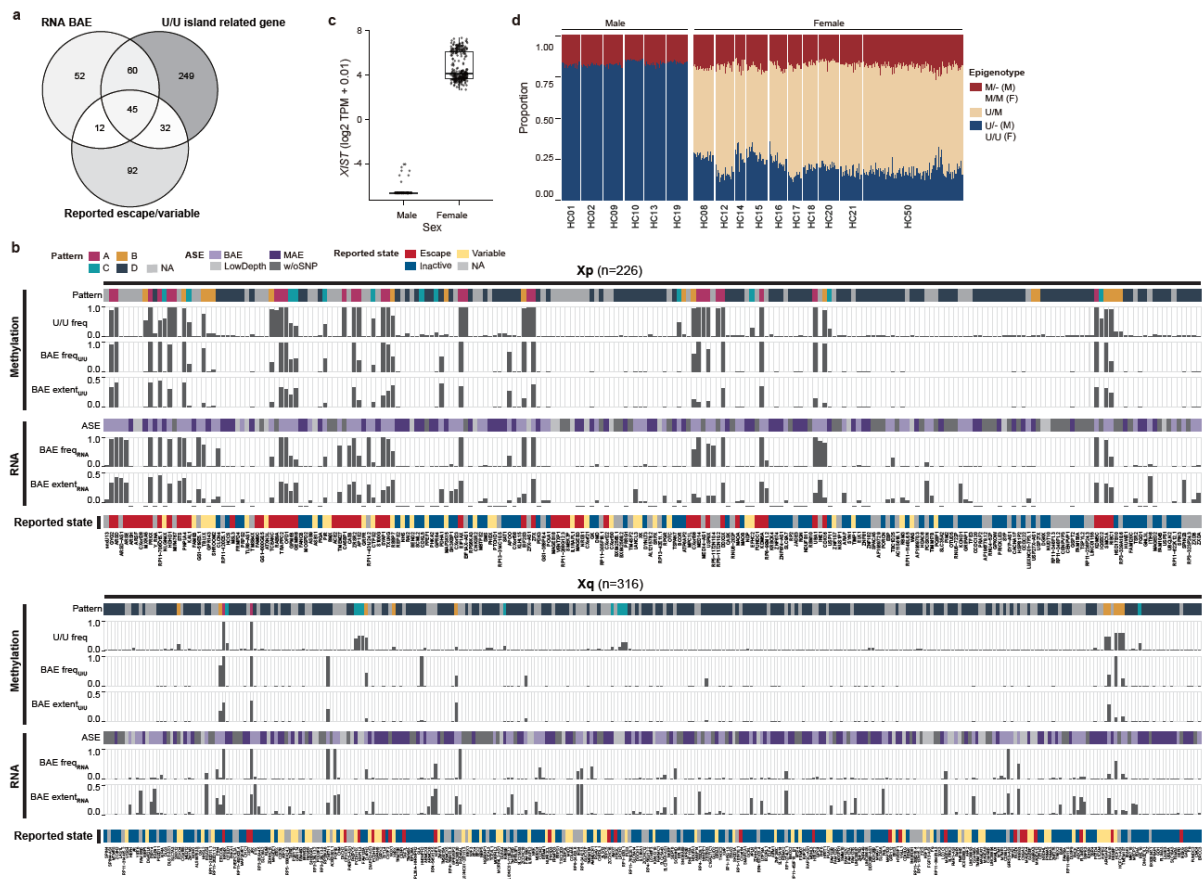
Extended Data Fig. 7 | Temporal dynamics and genomic clustering of epiallelic CpG changes

a, Cumulative epiallelic CpG changes plotted against accumulated somatic mutations for each of the 15 donors. Points indicate interval-specific estimates, and blue curves show the overall trajectories of cumulative methylation changes across developmental time. **b**, Cumulative epiallelic CpG changes plotted against accumulated somatic mutations for HC21 (left) and HC50 (right). Early developmental changes are concentrated within the first ~50 somatic mutations, whereas late-stage phylogenetic branches extend the trajectories into postnatal life. Blue lines indicate estimated postnatal accumulation rates of 112 and 144 CpG methylation changes per endogenous mutation in HC21 and HC50, respectively. **c**, Observed and randomly expected numbers of runs of consecutive EEC CpGs. Expected counts were obtained by permutation analysis. **d**, Distribution of genomic distances between adjacent EEC CpGs compared with random expectation ($P = 0.0099$).



Extended Data Fig. 8 | Clonal variation and genomic distribution of U/U CpG islands on the X chromosome.

a, Enrichment of CpG contexts among U/U CpG sites on the non-PAR X chromosome. Bars show log2 fold enrichment relative to all non-PAR X-chromosomal CpG sites; error bars indicate s.e.. **b**, Numbers of U/U CpG islands (mean $\beta < 0.2$) in individual clones across female donors ($n = 278$). Colours indicate the assigned categories (A–D), and each stacked bar represents an individual clone. **c**, Distribution of U/U CpG islands between the Xp and Xq arms, stratified by methylation category (A–D).

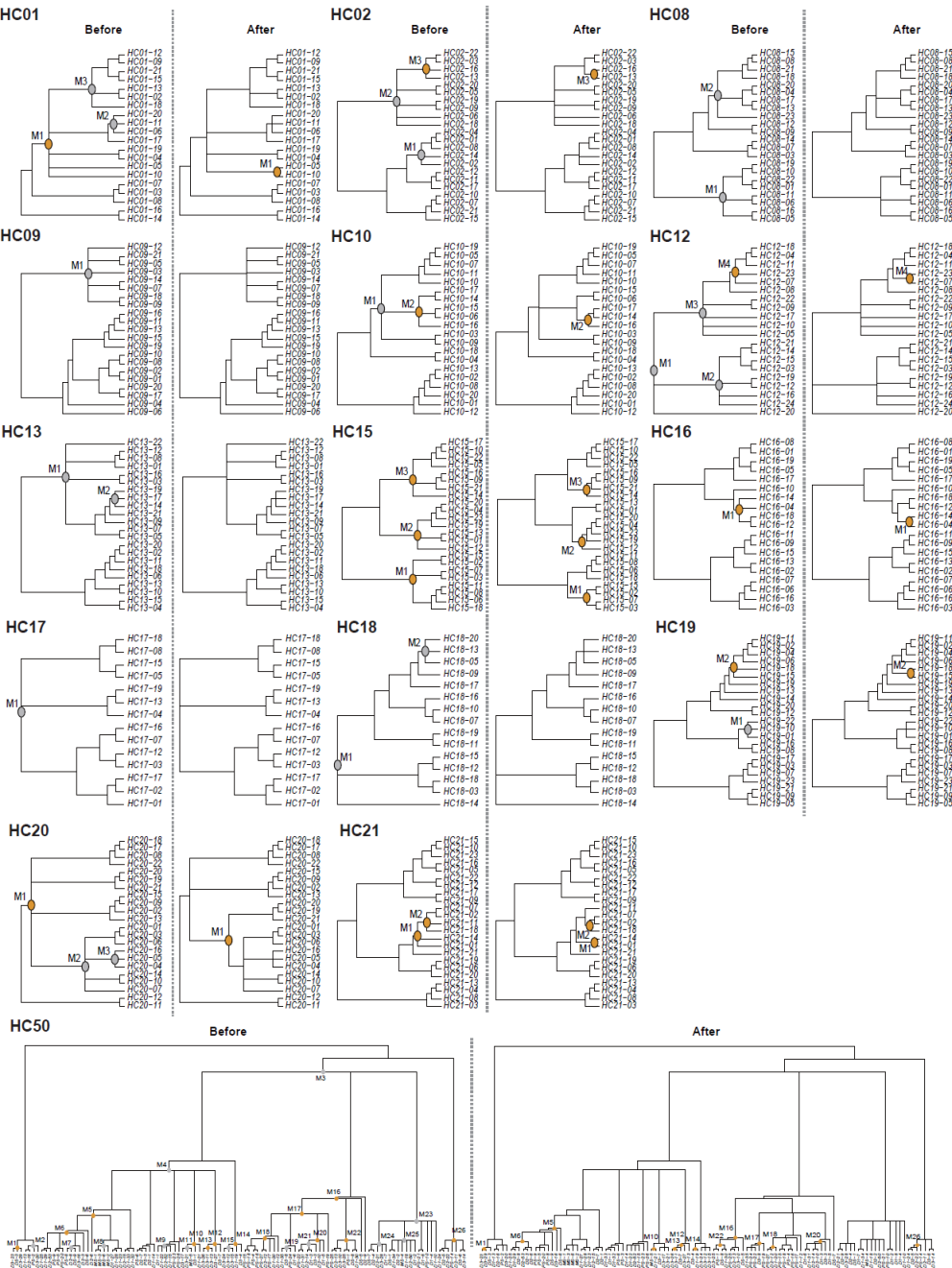


Extended Data Fig. 9 | Integration of promoter U/U CpG island, allele-specific expression and reported X-chromosome inactivation status.

a, Venn diagram showing the overlap among genes exhibiting biallelic expression (BAE) in at least one female clone (RNA BAE), genes associated with a promoter CpG island showing a U/U state in at least one female clone (U/U island-related genes), and genes previously reported as escape or variable X-chromosome inactivation (XCI) genes³⁶. **b**, Chromosome-wide integration of promoter methylation, allele-specific expression (ASE) and reported XCI status for non-PAR X-chromosome genes included in at least one of the three groups in **a**, ordered by genomic position and shown separately for Xp and Xq. **c**, *XIST* expression in male (n=105) and female (n=269) clones. *XIST* expression was quantified using RSEM-derived TPM values and is shown as log₂(TPM + 0.01). **d**, Epigenotype composition of CpGs within non-PAR X-chromosome CpG islands across male (n=128) and female (n=278) clones. Each bar represents an individual clone and shows the proportion of CpGs assigned to each epigenotype. M and U denote methylated and unmethylated alleles, respectively.

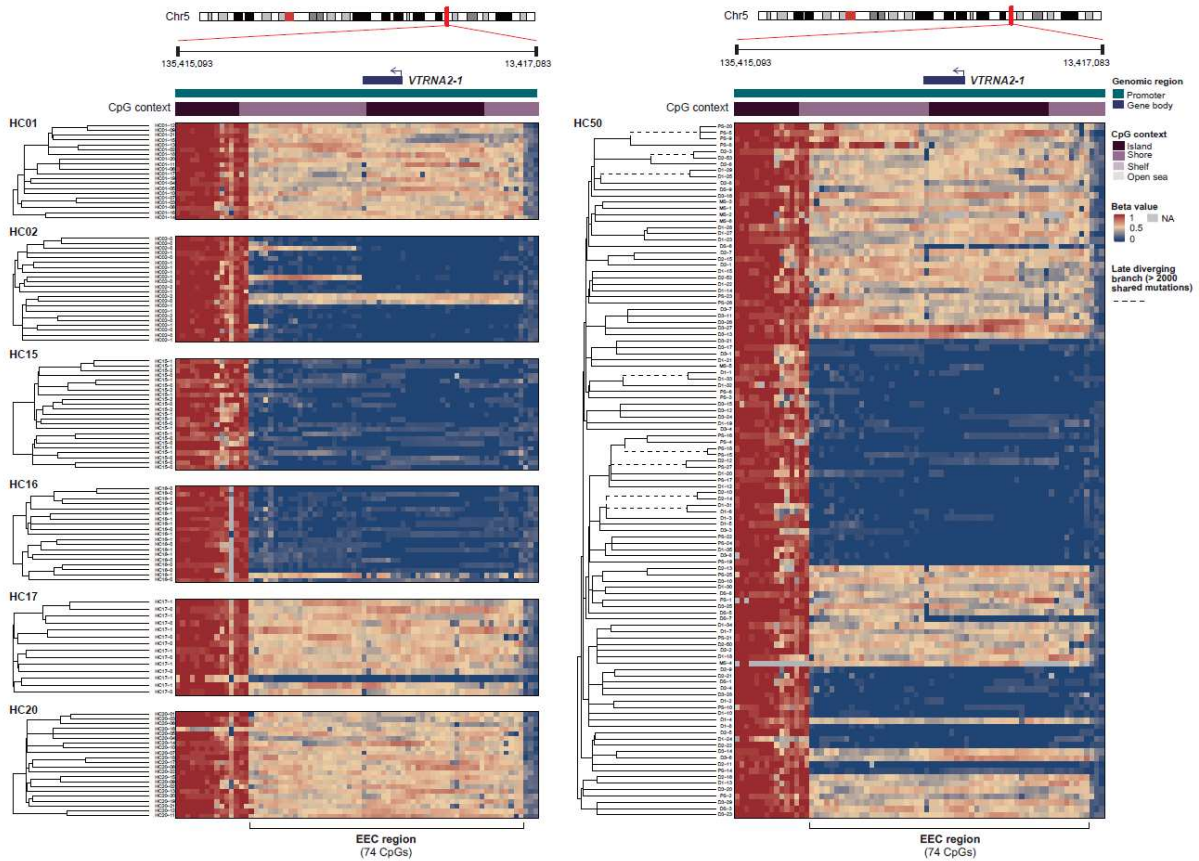
Extended data figure 11

Multifurcating node ● Unresolved ● Resolved



Extended Data Fig. 11 | Methylation-based refinement of multifurcating phylogenies.

Phylogenies before and after methylation-based refinement across 15 donors, with refinement resolving multifurcations in 10 donors. Multifurcating nodes were resolved using early methylation patterns when they occurred after more than five shared somatic mutations and had a prediction confidence of ≥ 0.7 . Nodes not meeting these criteria remained unresolved. Grey and orange circles indicate unresolved and resolved multifurcating nodes, respectively.



Extended Data Fig. 12 | VTRNA2-1 exemplifies lineage-specific and interindividual epiallelic variation.

VTRNA2-1, a well-characterized metastable epiallele, exhibits both lineage-specific EEC variation and interindividual epiallelic variation. Heatmaps show CpG methylation β values across individual clones, ordered according to developmental phylogenies reconstructed from somatic mutations. Most early lineage divergences occurred before the accumulation of ~50 somatic mutations, whereas dashed branches indicate late-diverging shared branches carrying >2,000 mutations. Genomic and CpG contexts are shown above, with the EEC span and number of consecutive CpGs indicated below.

2097 **Supplementary table legends**

2098 **Supplementary Table 1.**

2099 The table presents metadata and multi-omic sequencing characteristics for the 406
2100 colorectal organoid clones analysed in this study, including donor information, anatomical
2101 sampling location, availability and sequencing depth of WGS, EM-seq and RNA-seq data,
2102 and somatic SNV and indel burdens and mutational signatures.

2103 **Supplementary Table 2.**

2104 The table presents the genomic coordinates and constituent elements of non-redundant
2105 autosomal embryonic epiallelic commitment (EEC) regions identified across the cohort. For
2106 each element, donor-specific EEC coverage and the predominant genomic and CpG-island
2107 context are provided.

2108 **Supplementary Table 3.**

2109 The table presents the genomic coordinates and constituent elements of non-redundant
2110 EEC regions identified on chromosome X. For each element, donor-specific EEC coverage
2111 and the predominant genomic and CpG-island context are provided.

2112 **Supplementary table 4.**

2113 Enrichment of genomic and epigenomic features in the vicinity of embryonic epiallelic
2114 commitment (EEC) regions. The table presents detailed information on each feature
2115 examined, including reference data sources, biosample information, genomic window sizes,
2116 observed and expected feature counts or signals, fold enrichment, and statistical
2117 significance assessed by permutation analysis.

2118 **Supplementary table 5.**

2119 Annotation of autosomal promoters and their association with embryonic epiallelic
2120 commitment (EEC) regions. The table presents genomic annotations of autosomal
2121 promoters and their associated genes, including genomic coordinates, Ensembl Regulatory
2122 Build identifiers and gene assignments based on genomic proximity. Each promoter is
2123 further classified according to whether it overlaps an EEC region.

2124 **Supplementary table 6.**

2125 Enrichment of HGNC-defined gene groups among genes associated with embryonic
2126 epiallelic commitment (EEC) or later-acquired epigenetic changes (LEC). The table presents
2127 the observed and expected numbers of genes in each group, fold enrichment, associated
2128 gene names and statistical significance assessed using Fisher's exact test with Benjamini–
2129 Hochberg correction for multiple testing.

2130 **Supplementary table 7.**

2131 The table presents linear regression results for associations between promoter embryonic
2132 epiallelic commitment (EEC) epigenotypes and within-donor gene expression ranks. The first
2133 sheet includes all EEC element–gene pairs, whereas the second provides one
2134 representative element per gene, prioritizing negative associations with the smallest P value.
2135 Genomic annotations, regression coefficients, P values and sample sizes are provided.

2136 **Supplementary Table 8.**

2137 The table presents 2,691 non-redundant genomic regions showing later-acquired epigenetic
2138 changes (LECs) across the cohort. Genomic coordinates and the donors and colorectal
2139 clones carrying each LEC are provided.

2140 **Supplementary table 9.**

2141 CpG island methylation states and U/U categories across the X chromosome in females.
2142 The table presents genomic annotations and clone-specific DNA methylation levels of X-
2143 chromosome CpG islands, including associated promoters and genes, U/U epigenotype
2144 frequencies across female clones, and classification of U/U CpG islands into four categories
2145 (A–D).

2146 **Supplementary table 10.**

2147 Integrated analysis of promoter U/U states, allele-specific expression and reported X-
2148 chromosome inactivation status. The table presents genomic annotations, promoter-
2149 overlapping CpG island U/U categories and frequencies, biallelic expression (BAE)
2150 frequencies and extents across ASE-informative clones and promoter U/U clones, and
2151 overlap with previously reported XCI escape or variable genes.

2152 **Supplementary Table 11.**

2153 The table presents genomic and population characteristics of the 276 potential
2154 retrotransposition-competent L1 (rc-L1) source elements analysed in this study, including
2155 reference status, genomic location and orientation, population allele frequencies across
2156 ancestry groups and, where available, L1 subfamily, truncating mutations and insertion
2157 breakpoints as previously reported¹¹.

2158 **Supplementary Table 12.**

2159 The table presents genomic, epigenetic and retrotransposition characteristics of the 72 rc-L1
2160 sources showing promoter loss of methylation (LoM) in at least one colorectal clone. For
2161 each source, the table summarizes the phylogenetic timing of promoter LoM, clone-level
2162 promoter epigenotype and LoM frequencies, and clones harbouring source-assigned L1
2163 transduction events.

2164 **Supplementary table 13.**

2165 Validation and application of developmental phylogeny refinement using embryonic epiallelic
2166 commitment (EEC) methylation states. The first sheet presents validation results using
2167 artificially disrupted tree topologies, including predicted topologies, EEC element support,
2168 prediction confidence and outcomes. The second sheet provides refinement results for
2169 unresolved polytomies in reconstructed developmental lineage trees, including selected
2170 topologies, EEC element support, prediction confidence and refinement status.

2171 **Supplementary Table 14.**

2172 The table lists the components, suppliers, catalogue numbers and final concentrations used
2173 for culture of colorectal epithelial organoids.

2174 **Supplementary table 15.**

2175 Gene expression profiles of all 374 clones. The first sheet provides batch-corrected gene
2176 expression values expressed as transcripts per million (TPM), whereas the second sheet
2177 contains uncorrected TPM values directly obtained from RSEM quantification.

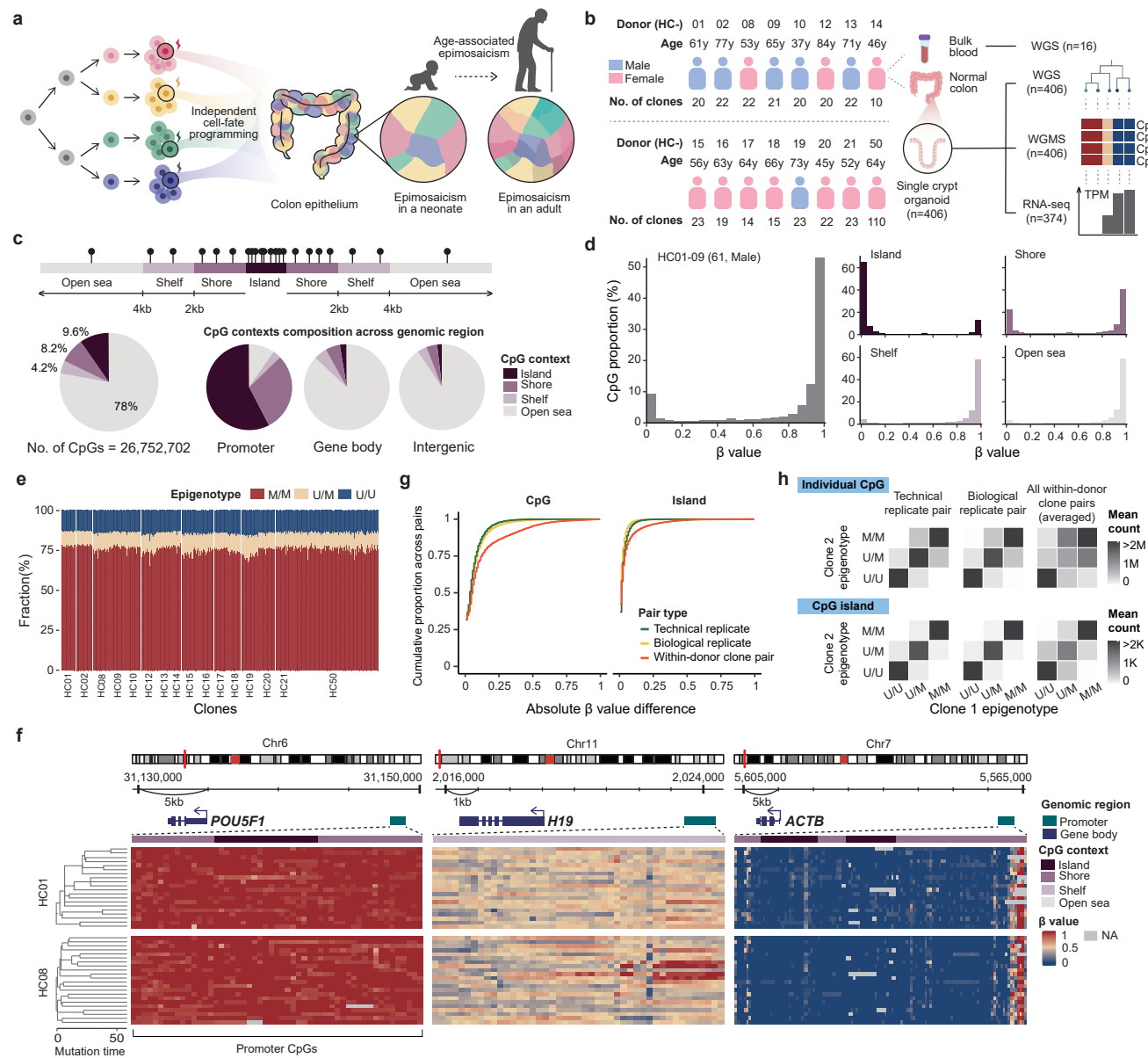
Fig 1

Fig 2

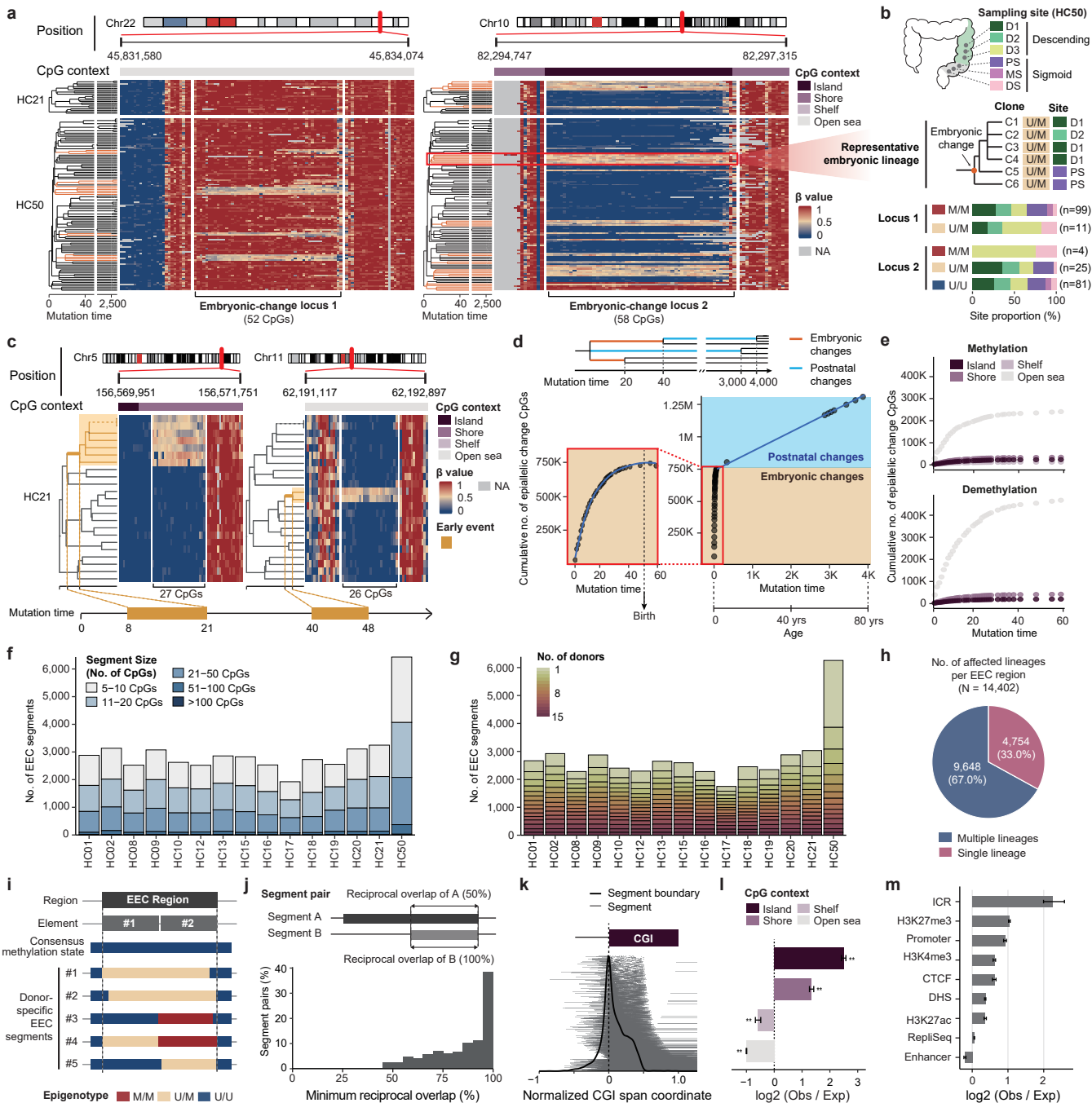


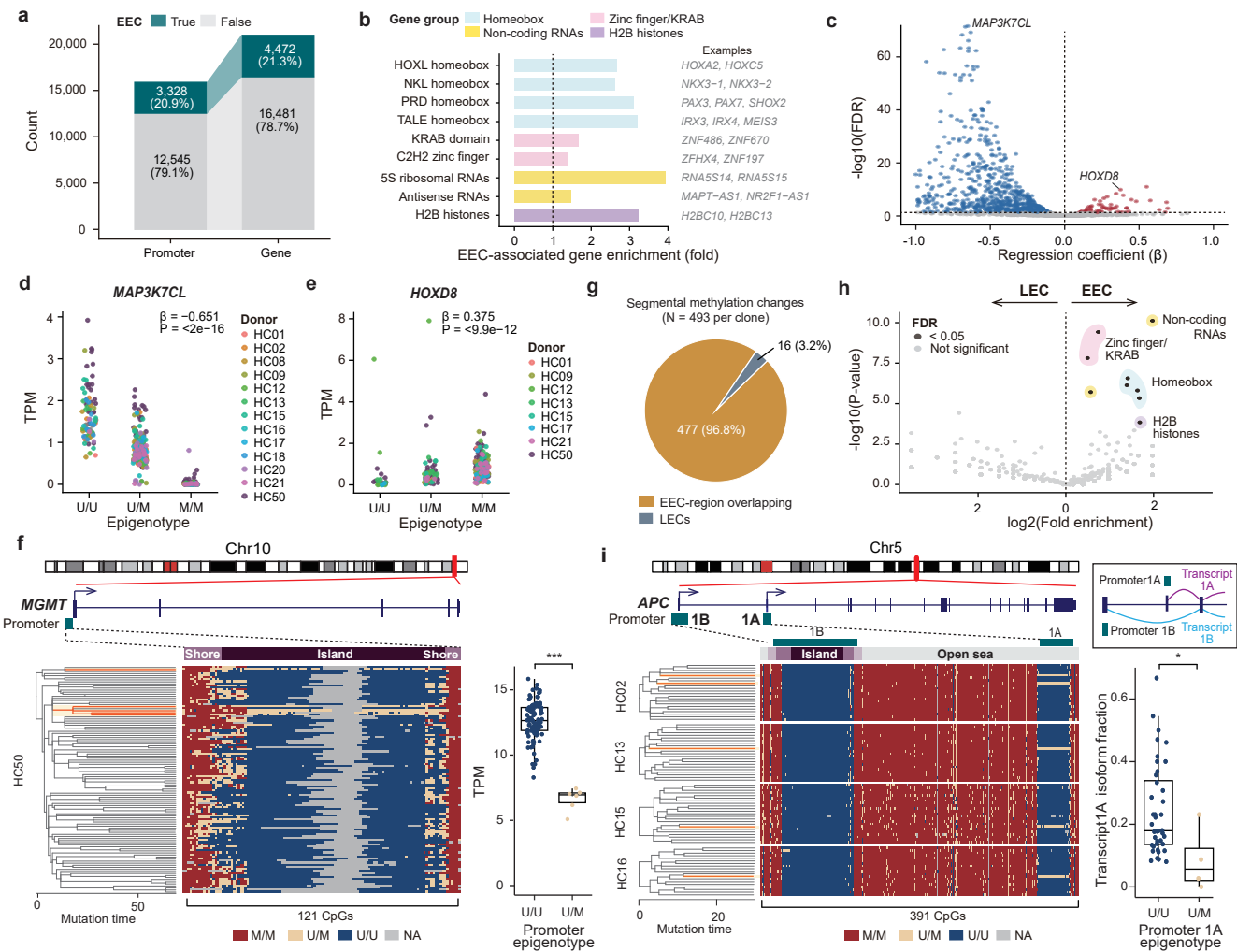
Fig 3

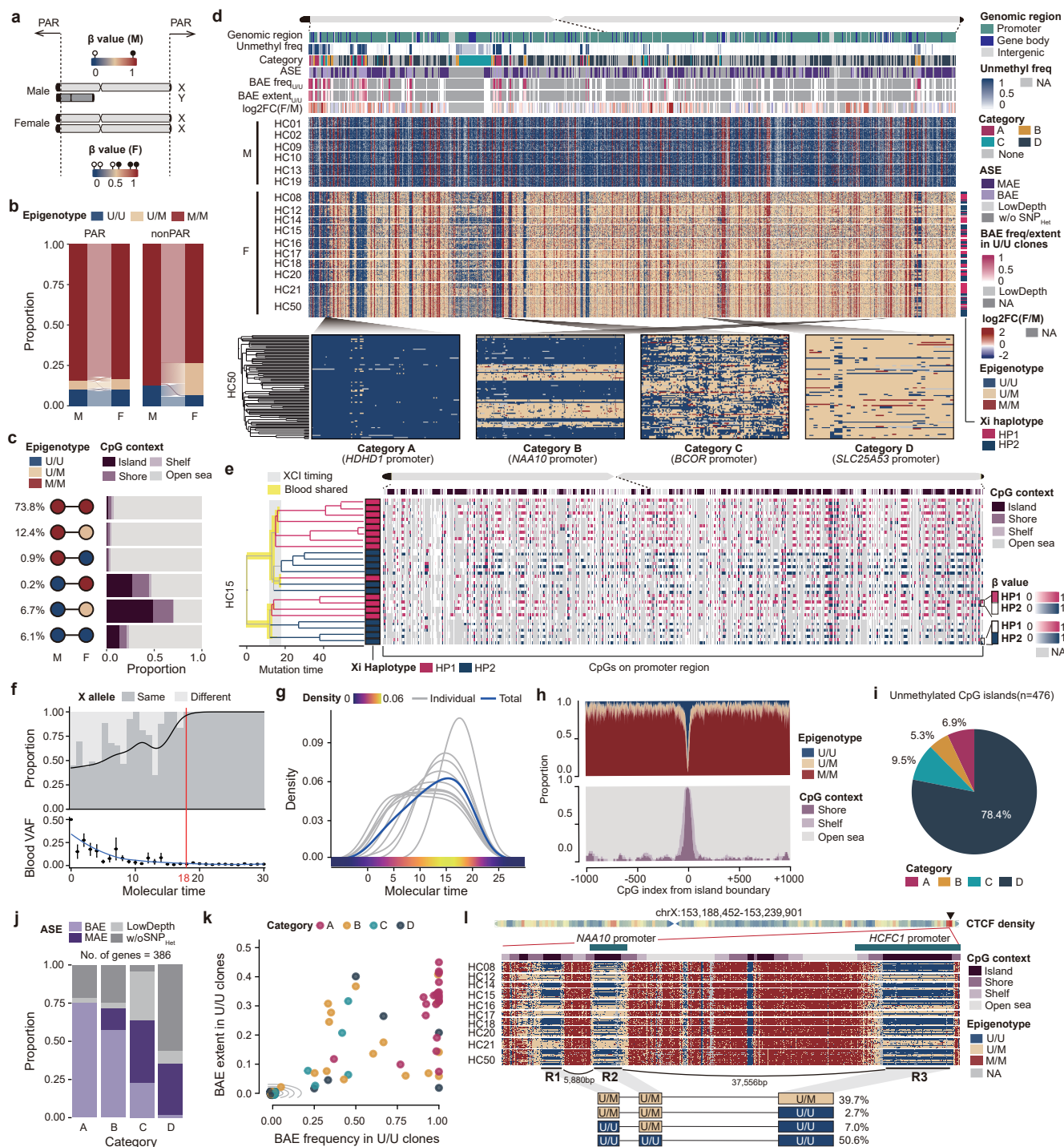
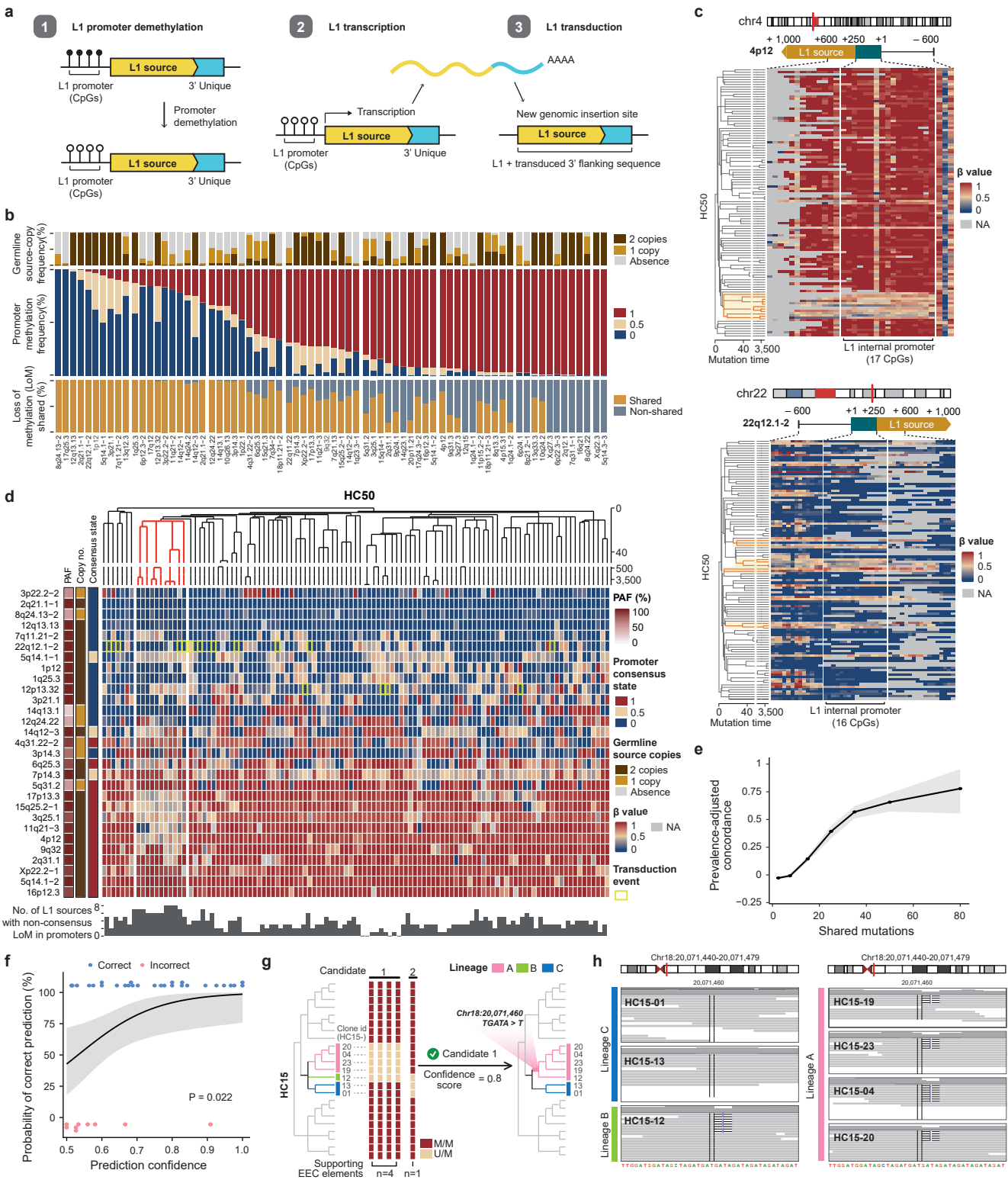
Fig 4

Fig 5

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarytable9.xlsx](#)
- [supplementarytable5.xlsx](#)
- [supplementarytable2.xlsx](#)
- [supplementarytable6.xlsx](#)
- [supplementarytable3.xlsx](#)
- [supplementarytable13.xlsx](#)
- [supplementarytable7.xlsx](#)
- [supplementarytable11.xlsx](#)
- [supplementarytable8.xlsx](#)
- [supplementarytable4.xlsx](#)
- [supplementarytable10u.xlsx](#)
- [supplementarytable1.xlsx](#)
- [supplementarytable12.xlsx](#)
- [supplementarytable14.xlsx](#)
- [SupplementaryTable15TPMre.xlsx](#)
- [Supplementarytablelegends.docx](#)