Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evolution of the human methylome in blood from embryo to ageing and cancer

Jyoti Nangalia

jn5@sanger.ac.uk

Wellcome Sanger Institute https://orcid.org/0000-0001-7122-4608

Xin Wang

Wellcome Sanger Institute https://orcid.org/0000-0002-0322-8970

Jing Guo

Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK https://orcid.org/0000-0003-1923-6614

Joe Lee

Wellcome Sanger Institute

Sean Laidlaw

Wellcome Sanger Institute https://orcid.org/0000-0001-7658-1428

Lori Kregar

Wellcome Sanger Institute https://orcid.org/0000-0001-7485-4509

Pierre Murat

Wellcome Sanger Institute

Xiaotian Liao

Wellcome Sanger Institute

Kevin Dawson

The Cancer, Ageing and Somatic Mutation Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA

Karim Mane

Wellcome Sanger Institute https://orcid.org/0000-0002-9892-2999

Michael Quail

The Wellcome Trust Sanger Institute https://orcid.org/0000-0002-3861-0483

Matthew Mayho

Wellcome Sanger Institute

Nicole Müller-Sienerth

Wellcome Sanger Institute https://orcid.org/0000-0002-7527-5127

Michael Spencer Chapman

Wellcome Sanger Institute https://orcid.org/0000-0002-5320-8193

Aditi Vedi

Cambridge University Hospitals https://orcid.org/0000-0001-6761-6510

Paul Milne
Newcastle University

Clarissa Pacyna
Wellcome Sanger Institute; Harvard Medical School https://orcid.org/0000-0002-4404-8183

Ken To
Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Kudzai Nyamondo
Wellcome Sanger Institute https://orcid.org/0009-0008-3312-732X

Aleksandra Kamizela
Wellcome Sanger Institute

Emily Mitchell
Wellcome Sanger Institute

Joanna Baxter
Cambridge Stem Cell Bank

Daniel Leongamornlert
WELLCOME SANGER INSTITUTE https://orcid.org/0000-0002-3486-3168

Steven Allen
Northwell and Northwell Health Cancer Institute https://orcid.org/0000-0002-3482-3182

Elisa Laurenti
University of Cambridge https://orcid.org/0000-0002-9917-9092

Peter Campbell
Wellcome Sanger Institute https://orcid.org/0000-0002-3921-0510

Matthew Collin
Newcastle University https://orcid.org/0000-0001-6585-9586

Nicholas Williams
Wellcome Trust Sanger Institute https://orcid.org/0000-0003-3989-9167

Biological Sciences - Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11370327/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. PJC is a cofounder and employee of Quotient Therapeutics. JL is a current employee of GSK

Evolution of the human methylome in blood from embryo to ageing and cancer

*Xin Wang[1], *Jing Guo[1], *Joe Lee[1], Sean Laidlaw[1], Lori Kregar[1], Pierre Murat[1], Xiaotian Liao[1], Kevin Dawson[1], Karim Mane[1], Michael A. Quail[1], Matthew Mayho[1], Nicole Muller-Sienerth[1], Michael Spencer Chapman[1], Aditi Vedi[2,3], Paul Milne[4], Clarissa N. Pacyna[1,5], Ken To[1,3], Kudzai Nyamondo[1], Aleksandra Kamizela[1], Emily Mitchell[3], E. Joanna Baxter[6], Daniel Leongamornlert[1], Steven L. Allen[7], Elisa Laurenti[2], Peter Campbell[1,8], Matthew Collin[4], Nicholas Williams[1], Jyoti Nangalia[1-3].

1. Wellcome Sanger Institute, Hinxton, UK
2. Cambridge Stem Cell Institute & Department of Haematology, University of Cambridge, Cambridge, UK
3. Cambridge University NHS Trust, Cambridge, UK
4. Newcastle University, Newcastle, UK
5. Harvard Medical School, Boston, MA, USA
6. Cambridge Blood and Stem Cell Biobank, Cambridge, UK
7. Northwell, New Hyde Park, NY, USA and Northwell Health Cancer Institute, Lake Success, NY, USA
8. Quotient Therapeutics, Saffron Walden, UK

* Joint first authors

Corresponding author: Jyoti Nangalia jn5@sanger.ac.uk

ABSTRACT

DNA methylation establishes and maintains cellular identity, is copied at every cell division, and changes during ageing and in cancer. Yet how it evolves within individual somatic lineages over the human lifespan has not been observable. We reconstruct the evolutionary history of the methylome in human blood from matched whole-methylome and whole-genome sequencing of 1,183 single-cell derived haematopoietic colonies from 13 individuals, spanning 17 weeks post conception to 77 years of age, healthy haematopoiesis to malignancy, and including monozygotic twins. Heritable methylation change exceeds the rate of single base substitution by six orders of magnitude before birth and four orders thereafter, often affecting whole tracts of CpGs at once. Each blood cell carries ~300,000 somatic methylation changes before birth, including >3000 at large segments that remain stable for decades. We term these large segments embryonic epialleles: CpG tracts at which lineages of the same cell type within an individual carry different heritable methylation states, of which we identify ~250,000 across the cohort, with many recurring between individuals. Against this developmental mosaic, we time X-chromosome inactivation to the first few divisions of life. We show that epigenetic age reports the division history of whichever clones dominate blood, rather than the age of the individual, reframing the interpretation of methylation biomarkers. In clonal expansions and blood cancer, many methylation differences between an expansion and its background prove to be embryonic states inherited by the founder cell. Expansions do, however, converge on methylation change at Polycomb-associated regulatory regions, gaining methylation in most clades but losing it where *DNMT3A* is mutated. A few changes at established cancer genes, including *CDKN2B*, *ABL1* (in *BCR::ABL1*) and *CACNA1D*, arise only within clonal expansions, and bear the hallmarks of epigenetic drivers. Methylation states determine which transcript a gene produces, and whether it is expressed at all. Together with their heritability and variability between blood cells, these phenotypic consequences establish methylation mosaicism as a pervasive additional substrate for Darwinian somatic evolution.

INTRODUCTION

DNA methylation of cytosines at CpG dinucleotides is the most abundant modification of the human genome. Of around 28 million CpGs in the haploid genome, 60-80% are methylated in somatic cells[1]. Methylation establishes and maintains cellular identity, suppresses one X chromosome in every female cell, and preserves parental imprints over a lifetime[2]. Given that methylation is copied by maintenance methyltransferases each time a cell divides[3], it should, in principle, also hold a record of ancestry.

In haematopoiesis, somatic mutations in *DNMT3A* and *TET2,* both regulators of methylation, are the commonest drivers of clonal haematopoiesis (CH)[4] and contribute to myeloid malignancies[5,6]. Methylation variation at selected CpGs has been exploited to estimate biological ageing[7] and clonal dynamics[8,9]. Yet fundamental properties of somatic methylation evolution *in vivo* remain poorly understood. It is unknown how often heritable methylation changes are acquired, where in the genome or when in life they occur, whether changes affect individual CpGs or tracts, or how these processes are altered during clonal expansion. Addressing these questions requires methylation to be followed across the genome at every CpG, across cells of known ancestral relationships, and across time.

Somatic mutations provide the most durable record of lineage history. They accumulate through life in all cells and are faithfully inherited by descendants, allowing ancestral relationships between cells to be traced back to the zygote. Our previous phylogenetic studies have provided insights into embryonic development[10], healthy ageing[11,12], and cancers[13–15]. We reasoned that if the methylome of each cell could also be captured and read alongside its genome, then every CpG methylation state could be placed on the phylogeny and every methylation change assigned to the branch on which it arose. Here, we present data from 1,183 single-cell derived haematopoietic colonies with matched genomes and methylomes from 13 donors, from a 17-week foetus to 77 years of age, spanning healthy haematopoiesis, CH and haematological cancer, as well as a pair of monozygotic twins. This allowed us to walk through every CpG in the genome, in every cell, from the earliest divisions of life to old age and cancer.

RESULTS

Genome-wide architecture of the human methylome at lineage resolution

Reconstructing methylation inheritance requires the complete methylome of individual cells of known ancestry. Bulk methods, such as arrays and reduced-representation bisulphite sequencing, sample a biased 2-3% of CpGs across mixtures of cells[16–18], while single-cell methylation recovers a sparse, stochastic subset of CpGs per cell[19–21], sufficient to group cells by shared methylation profiles[22]. We therefore expanded single haematopoietic stem and progenitor cells (HSPCs) into colonies and undertook low-DNA input whole-genome sequencing (WGS)[12–14] and enzymatic-methylation sequencing (EM-seq) on each. Enzymatic conversion, unlike bisulphite treatment, preserves DNA integrity, and modifications to the oxidation chemistry (**Methods**) yielded genome-wide methylome coverage from nanogram inputs of DNA[23]. The methylome of each colony could then be placed on a lineage tree reconstructed from its genome (**Fig.1a**).

The scale of these data presented a particular challenge. Whereas each colony carries tens to over a thousand single nucleotide variants (SNV), methylation is measured at nearly all CpGs in the genome. Across >1000 colonies, this yielded ~30 billion observations. We therefore built *cpgAccessData*, a computational framework for efficient, indexed, phylogeny-aware access to methylation counts, including allele-resolved counts near informative heterozygous SNPs (**Methods**). Together with laboratory automation, this made genome-wide reconstruction of methylation evolution tractable across a thousand colonies, and in principle, much larger cohorts.

Following quality control, 1,183 colonies were analysed from 13 donors **(Fig.1b)**, capturing healthy haematopoiesis[11], slow clonal expansions of CH[4,11], faster growth of myeloproliferative neoplasms (MPN)[14], and the explosive expansion of *BCR::ABL1*-positive chronic myeloid leukaemia (CML)[13]. Median depth was 15x for WGS (range 7–55x), and 14x per CpG dyad (range 4–39x) (**Fig.1b**), with a median of >95% of 26.8 million autosomal CpG dyads per colony covered by at least five reads, most residing in CpG-poor open seas (**Fig.1c**).

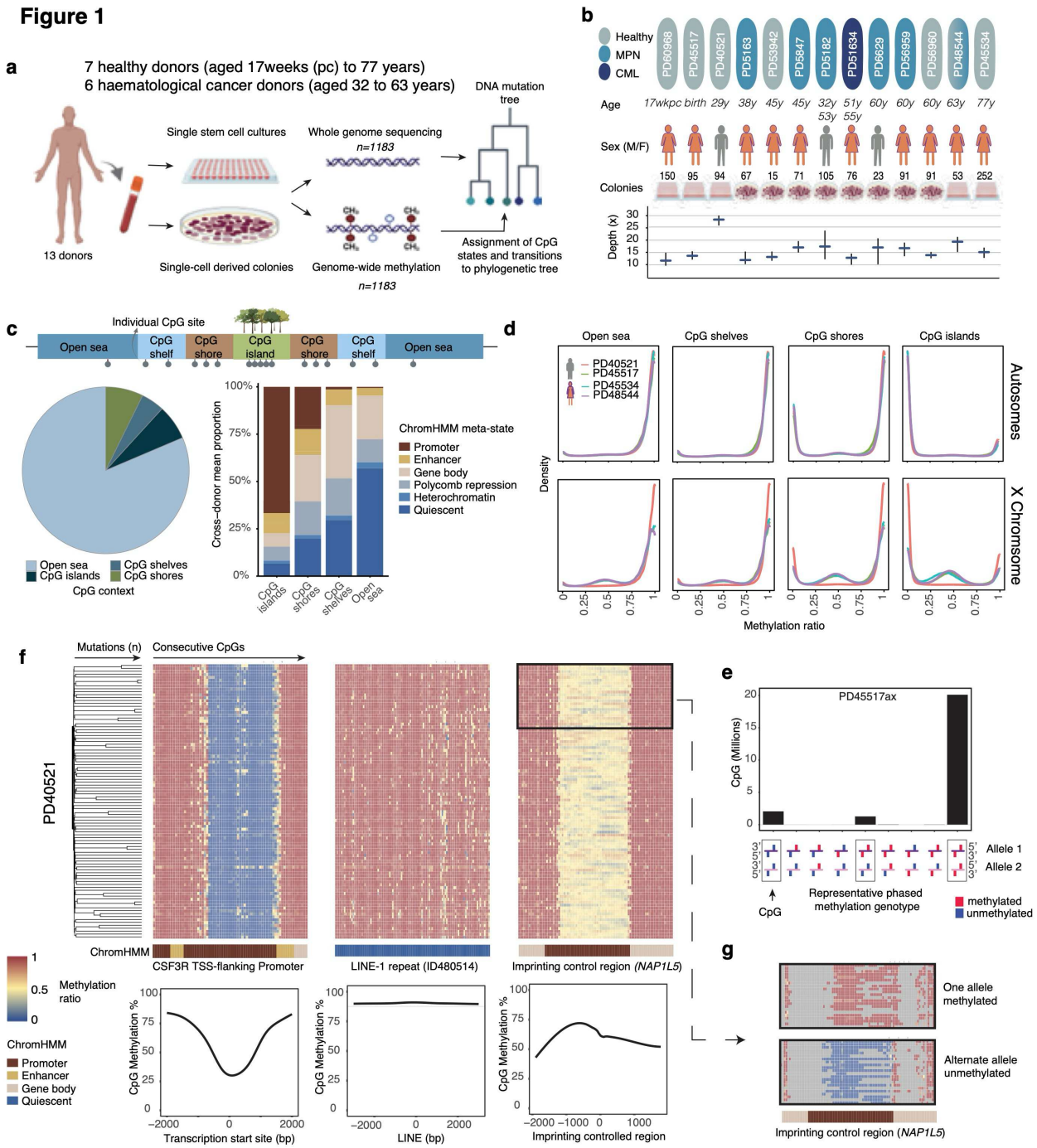


Figure 1. Genome-wide methylation at clonal resolution. a. *Study design. Peripheral blood or bone marrow from each donor was expanded into single-cell-derived haematopoietic colonies (n=1,183). Each colony underwent whole-genome sequencing (WGS) and whole-genome EM-seq methylome sequencing (WGMS, ~28 million CpGs). Somatic DNA mutations were used to build a donor-specific phylogenetic tree, onto which genome-wide CpG methylation states and their transitions were subsequently mapped.* ***b.*** *Cohort overview of 13 donors including 7 healthy and 6 with*

haematological malignancy (MPN/CML), spanning 17-weeks post-conception to 77 years, and one monozygotic twin pair. Age at sampling, sex, culture methods (liquid or MethoCult semisolid culture), number of colonies sequenced, WGMS median depth and interquartile range (IQR; 25th–75th percentile) are shown. ***c.*** *Genome-wide CpG distribution by CpG context and grouped chromatin state with CpG context modelled above.* ***d.*** *Density of per-CpG methylation ratio across CpG contexts for two male and female representative donors, shown separately for autosomes (top) and the X chromosome (bottom).* ***e.*** *Genome-wide distribution of phased methylation genotypes across the two parental alleles for a representative colony (PD45517ax), showing that CpGs are overwhelmingly confined to three states: unmethylated on both alleles, methylated on both alleles, or allele-specific methylation.* ***f.*** *Representative single-colony methylation heatmaps (donor PD40521) at three loci spanning three canonical regimes: an unmethylated CpG-island promoter (CSF3R TSS), a fully methylated LINE-1 element, and an allele-specific imprinting control region (NAP1L5). The boxed rows at NAP1L5 are expanded in g.* ***g.*** *Phased methylation at the boxed NAP1L5 imprinting control region, split by parental alleles, showing methylated and unmethylated alleles.*

Each colony grows from a single cell that divides and differentiates in culture, with opportunity for methylation change. Colonies grown in liquid culture nevertheless retained a bimodal methylome, with 91.1% of autosomal CpGs at β <0.2 or >0.8 (**Fig.1d, Table S1**). The sharpness of these peaks reflects methylation changes in culture: gain was negligible, but methylation loss (γ) was higher in colonies grown in Methocult (mean γ 0.21 versus 0.095, **Ext.DataFig.1a**), reflecting their predominant erythroid composition (**Ext.Data Fig.1b)**. This loss was concentrated in CpG-poor regions rather than CpG islands (**Ext.Data Fig.1c-d)**, consistent with replication-coupled loss of maintenance during erythroid maturation[24,25].

Within each colony, CpG methylation was overwhelmingly confined to three states - unmethylated on both alleles (0), fully methylated on both (2), or methylated on one allele only (allele-specific methylation, ASM; 1). The two cytosines of individual CpG dyads were highly concordant, with hemi-methylation (asymmetry within a single CpG dyad) being rare (**Fig. 1e**), consistent with efficient maintenance[26,27]. Downstream analysis was therefore performed at the dyad level (hereafter, the CpG) considering these 3 states (0, 1 or 2) and phasing ASM where possible. This trinary architecture recapitulated canonical genomic features: promoters and CpG islands were largely unmethylated, repetitive elements such as LINEs were highly methylated (**Fig.1d, f**), and imprinting control regions (ICR) showed the expected ASM confirmed by phasing (**Fig.1f,g**). Female X-chromosomes also showed ASM, concentrated at CpG islands and shores and absent from males (**Fig.1d**).

Heritable segmental organisation of the methylome

To retrace methylome evolution, we reconstructed somatic phylogenies for each donor from WGS[12–14] (**Ext.DataFig.2)**. Here, each tip is a single colony, shared branches carry mutations inherited by all descendants, terminal branches represent private mutations, and branch length reflects molecular time. The trees therefore provide both the ancestral relationships and the temporal scaffold on which methylation changes can be placed. Changes arising in culture affect single colonies and so fall on terminal branches, whereas changes shared by several colonies must have arisen in an ancestral cell *in vivo.*

To assign methylation changes to branches (see **Supplementary note**), we developed *DAPHNE*, a hidden Markov tree model**.** Neighbouring CpGs frequently shared methylation states and were assigned the same transition on the same branch, suggesting that the heritable unit of methylation is often not a single CpG but a CpG segment. We therefore developed *LAUREL* to first partition the genome into runs of CpGs with homogenous methylation within each colony, and conserved boundaries across. Segments ranged from single CpGs to over a thousand, the largest enriched at promoters and gene bodies (**Fig.2a-c**). Applying *DAPHNE* to these segments accounted for segmental methylation transitions on branches but single-CpG events nonetheless remained common (**Fig.2d**).

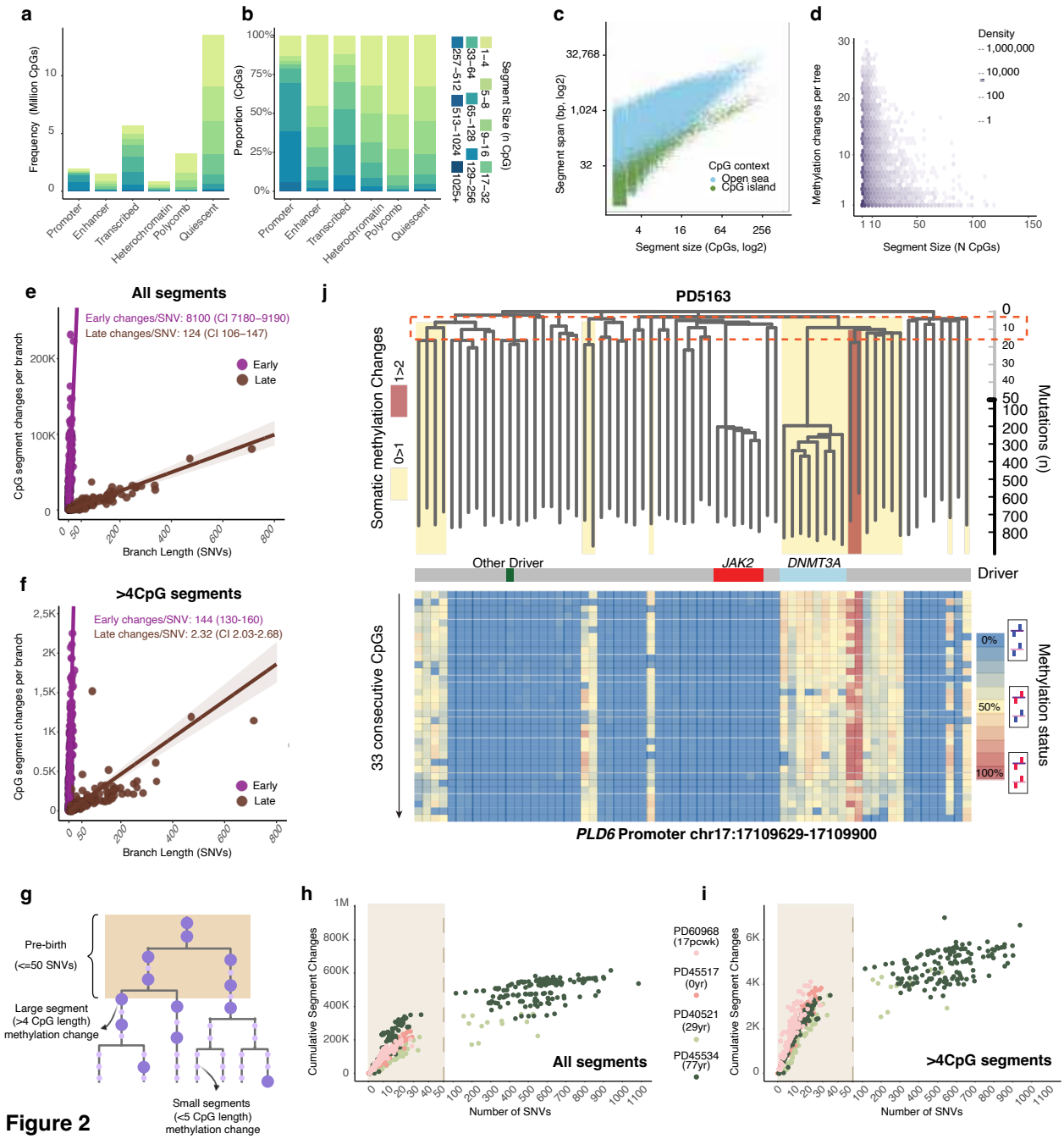


Figure 2

Figure 2 Legend. Methylation changes over human lifespan. ***a, b.*** *Distribution of genome-wide CpGs across grouped chromatin states, stratified by CpG segment size. Absolute frequency (millions of CpGs) (a) and proportion of CpGs (b) per chromatin state.* ***c.*** *Genomic span (bp) of segments scales with CpG number. Each point is one autosomal segment on chromosome 1: y is genomic span (bp, $\log_2$), and x is size (number of CpGs, $\log_2$), coloured by CpG context: open sea (blue, median 54bp/CpG) and CpG island (green, median7bp/CpG).* ***d.*** *Segment size versus methylation changes. Hexagonal binning of segments by size (CpG number, x) and the number of methylation changes inferred on internal branches of the phylogenies (y). Colour indicates the number of donor-segment pairs per hexagon (log scale).* ***e.*** *Methylation change rate for all segments. Each point is an internal branch from healthy donors: y, the number of CpG-segment methylation changes on the branch, x, branch length (SNVs), with branches split by epoch - early (branches finishing ≤50 SNVs, magenta); late (branches starting >50 SNVs, brown). Lines depict the rate of changes per branch, fitted jointly across donors; shading, 95% confidence intervals.* ***f.*** *Methylation change rate for large (>4 CpG) segments.* ***g.*** *Schematic phylogeny with methylation changes marked on branches. Large circles, changes in segments with >4 CpG; small circles, changes in segments with ≤ 4 CpGs. Shading marks before birth (≤ 50 SNVs). For each internal branch, the*

*cumulative burden is the number of methylation changes on all branches from the root to that branch, plotted against its cumulative SNV count as shown in **h-i.** Cumulative accrual of all segmental methylation changes **(h)**, and large (>4 CpG) segmental changes **(i)** with molecular time (x). Each point is a phylogenetic branch from one of four donors: PD60968 (17-week foetus, pink); PD45517 (cord blood, salmon); PD40521 (29 yr, light green); PD45534 (77 yr, dark green). The x-axis is expanded before birth (shaded region). **j.** Example **embryonic epiallele** at PLD6 (promoter) in PD5163. Top: the colony somatic mutation phylogeny; each tip is a single-cell-derived colony with the vertical axis in molecular time, expanded for mutations acquired before birth. Branches and clade shading are coloured where the model infers a somatic methylation state change at the promoter segment: cream shades lineages that acquired a 0>1 transition (unmethylated to allele-specific/half-methylated) and the clade that inherits it, and maroon a 1>2 transition (to fully methylated). The red dashed box marks the early molecular-time window around gastrulation (~10-20 mutations). The bar beneath the tree indicates driver status. Bottom: per-CpG methylation across the 33 consecutive CpGs of the PLD6 promoter (rows; chr17:17,109,629-17,109,900) for every colony (columns), coloured from 0% (blue, unmethylated) through 50% (cream) to 100% (red, methylated).*

Epimutation exceeds somatic mutation and is concentrated before birth

With methylation changes placed on 'time-stamped' branches, we could estimate how often they occur and when in life they arise. Using healthy donors of different ages and excluding terminal branches, we found that a single constant rate of methylation change (hereafter, epimutation) was rejected **(Supplementary Note).** Allowing for one rate before birth and another after, reconciled the cohort, with rates from young donors (cord blood and foetus) now matching those from the same periods (tops of the trees) in adult donors. The true transition here is likely gradual but our two epochs are a practical approximation.

For every somatic mutation acquired, a cell gained ~8000 heritable methylation changes before birth, and ~120 thereafter (**Fig.2e**). Larger segments changed less often, but showed the same ~60-fold reduction after birth (**Fig.2f**). Per site, methylation changed a million times more often than DNA mutates before birth (1.3×10^{6}, CI 1.1×10^{6}–1.4×10^{6}), and 20,000 times more often subsequently ($2.0\text{x}10^{4}$, CI $1.7\text{x}10^{4}$-$2.3\text{x}10^{4}$), a disparity also reported in plants[28]. Most heritable methylation diversity in blood has therefore already accumulated by birth, with each blood cell carrying ~300,000 somatic methylation changes, of which ~1% affects larger segments (**Fig.2h,i**).

Yet the methylome was also largely stable. Two in every five CpGs never changed on any branch, and half of all changes arose at just 5% of CpGs. When changes occurred, they were overwhelmingly monoallelic, with changes from homozygous states to ASM, rather than switching directly between fully methylated and unmethylated. Methylation is thus gained or lost one parental allele at a time, at rates that vary with genomic context and segment size **(Ext.DataFig.3, Table S2**). The *PLD6* promoter illustrates this clearly **(Fig.2j).** Over 30 consecutive CpGs transitioned together from unmethylated to ASM independently in several early branches of the phylogeny. Across donors, two-thirds of such events arise 10-20 mutations from conception, around gastrulation[10], before being faithfully maintained for the remainder of life.

We term these *embryonic epialleles*: large CpG segments at which independent lineages from an individual carry different heritable methylation states, established in the embryo and inherited for life. They form a large repertoire. Each cell has >3000 such changes per cell by birth, and across the cohort we identify ~250,000, of which ~40% recur between individuals. Embryonic epialleles are independently reported in colorectal epithelium by Won, Lee et al[29]; and we explore them further in three healthy donors from this study in Kregar, Lee et al.

Retracing X chromosome inactivation *in vivo*

Some epigenetic decisions are made only once in development and persist for life. Given the stable record of early, segmental methylation changes, we asked whether known developmental programmes could be retraced from the phylogenies. X-chromosome inactivation (XCI) is one such programme, achieving gene dosage compensation in female mammals through chromosome-wide chromatin compaction, visible as the Barr body under microscopy, as well as methylation. On the inactive X (Xi), this methylation is allele-specific and concentrated at CpG-rich islands (**Fig.1d**), which we exploited to reconstruct when XCI occurs *in vivo.*

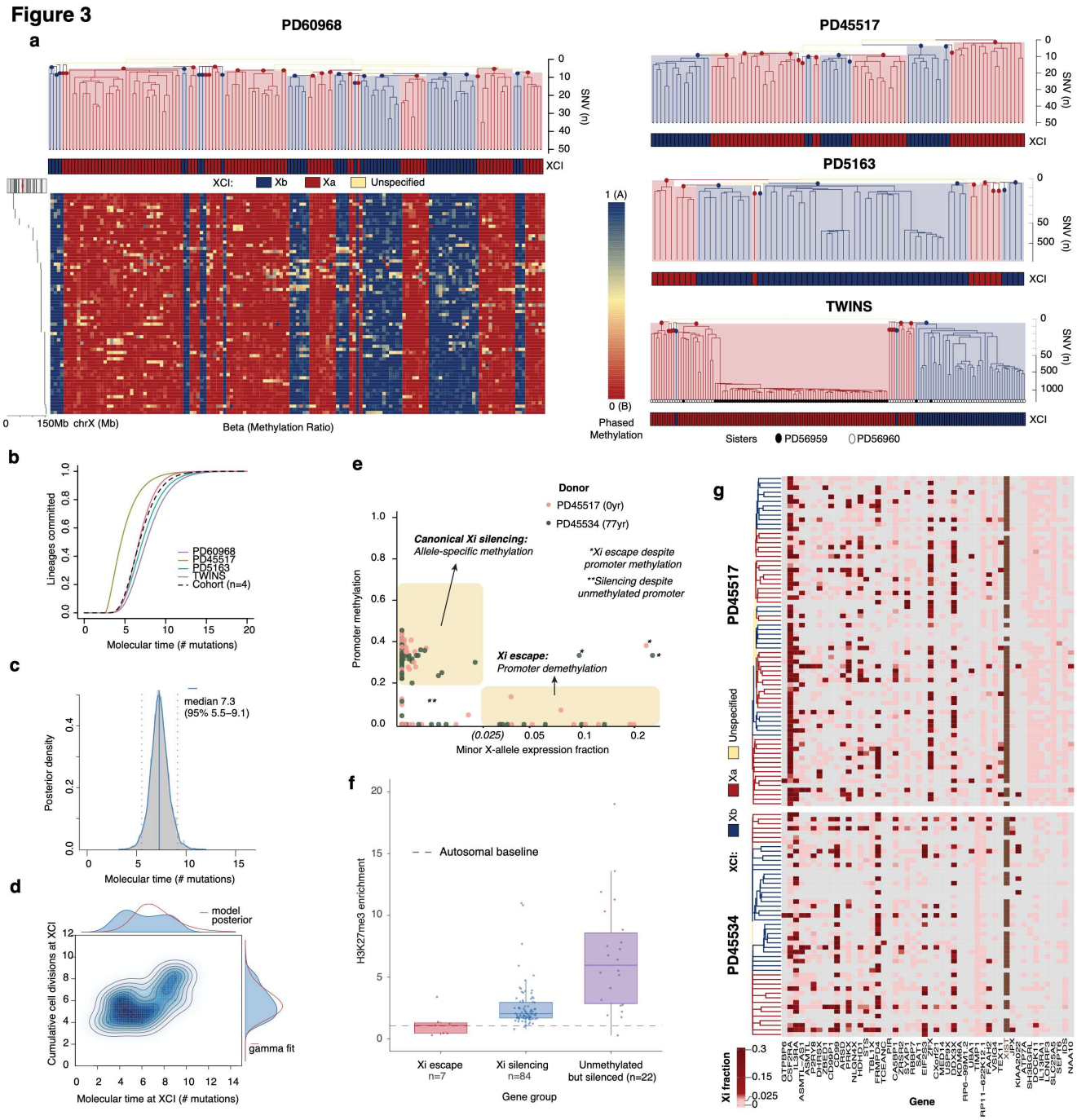


Figure 3 Legend. X chromosome inactivation (XCI) and escape. a. *Somatic phylogenies in female donors with branches coloured by XCI choice inferred from phased methylation (Xa, red; Xb, blue). Shaded clades carry the same XCI choice. Filled dots mark inferred ancestral commitment events (uncommitted, yellow; Xa, red; Xb, blue). For TWINS, filled versus open tip symbols denote the two sisters. For PD60968 (left), the tree sits above an allele-aligned methylation heatmap*

where rows are X-linked CpGs near transcription start sites ordered along the X chromosome (left ruler, chrX Mb position, non-PAR), and colour reflects allelic methylation skew (1, Xa, red; 0, Xb, blue; 0.5, no skewing). ***b.*** *Timing of XCI across the cohort. Fitted cumulative probability curves showing the fraction of lineages committed to XCI as a function of molecular time for individual female donors (coloured curves) and the pooled cohort (dashed black). Curves depict posterior median parameters.* ***c.*** *Posterior distribution of mean XCI commitment time for PD60968 (grey histogram, posterior draws; blue curve, density). Solid vertical line, posterior median (7.3 mutations); dotted lines, 95% credible interval (5.5–9.1).* ***d.*** *Joint distribution of commitment timing and cell divisions in foetus PD60968.* *Each point is an inferred X-inactivation commitment event placed on the phylogeny. Horizontal axis, molecular time of commitment event; vertical axis, estimated cumulative number of cell divisions at commitment event. The central panel shows a two-dimensional kernel-density estimate (darker represents more events). The top strip shows the density distribution (blue) and posterior-predictive commitment time (red). The right strip shows the distribution of division counts (blue) with an empirically based fitted gamma curve (red).* ***e.*** *Promoter methylation versus allelic expression for X-linked genes. Each point is an X-linked gene in PD45517 (salmon) and PD45534 (green) showing promoter methylation against the fraction of its expression coming from the inactive X. Cream boxes highlight canonical silencing by methylation (top-left) and XCI escape (bottom-right). Genes departing from these are marked by * (escape despite a methylated promoter) and ** (silencing despite an unmethylated promoter).* ***f.*** *H3K27me3 enrichment at core promoters of X-linked genes with Xi escape, methylated promoters (Xi silencing), and genes without promoter methylation but nonetheless silenced. Boxes show median and interquartile range.* ***g.*** *Colony-level allele-specific expression of X-linked genes in PD45534 and PD45517. Tree branches are coloured by XCI choice (Xa, red; Xb, blue). Columns are a set of X-linked genes ordered by genomic position. Cell colour shows the expression fraction from the inactive X for that gene (pink, <0.025 as the escape threshold; darkening red for increasing escape; grey, insufficient allele coverage; XIST, expressed only from the inactive X, shown in brown).*

Embryonic development leaves two traces in the phylogenies: branch lengths provide a molecular clock, and nodes capture ancestral cell divisions. Nodes whose descendants carry different Xi must predate XCI, whereas lineages with a shared Xi imply either an ancestral cell that arose after XCI commitment or multiple independent but identical commitment choices. We thus assigned Xi choice to 715 colonies from six female donors, focusing on four with sufficient resolution at the top of the trees (**Fig.3a**). Using a three-state Markov chain model, we inferred XCI commitment at ~7 mutations of molecular time (posterior median, 95% CI 5.5-8.4, **Fig.3b**), with estimates from individual donors being remarkably similar to that from the 17-week foetus with the highest resolution at the top of the tree (**Fig.3c)**. Counting the preceding cell divisions (nodes) on the tree of the foetus, placed XCI between 3 and 9 cell divisions (**Fig.3d, Table S3)**, consistent with classical estimates from twin concordance[30] and chimeric mosaicism[31]. The monozygotic twins shared ancestral cells well beyond XCI commitment, to ~30 mutations of molecular time (**Fig.3a**), corresponding to the first trimester, and indicating exchange of HSPCs through monochorionic placentation.

We were intrigued by how reliably methylation maintained silencing on X. Using single-cell RNA sequencing from the cord blood and 77-year old females, genes with Xi promoter methylation showed monoallelic expression as expected, while genes with unmethylated Xi promoters accounted for the majority of biallelically expressed genes (**Fig.3e**). However, the relationship was not absolute. Two genes, *NEXMIF* and *STK26*, showed detectable biallelic expression despite Xi promoter methylation, and a larger set of genes remained silent despite unmethylated Xi promoters (**Fig.3e**). The latter were strongly enriched for H3K27me3, the mark deposited by PRC2, in keeping with methylation and Polycomb providing two complementary routes to gene silencing (**Fig.3f)**. Lack of methylation is thus permissive, but not sufficient, for gene expression. Overall, Xi escape was a property of genes[32] rather than lineages, recurring across the trees with no lineage losing Xi silencing across many loci (**Fig.3g)**.

Genomic imprinting is maintained lifelong but somatically diversified in early development

In contrast to the selective establishment of XCI, imprinting is a challenge of selective maintenance, requiring parent-of-origin methylation marks to survive the demethylation that follows fertilisation. Intermediate methylation characteristic of ICRs in bulk samples can reflect several things at the single cell level - a mixture of methylation states across cells, a mixture of methylation states across different CpGs within individual cells, or genuine ASM. Our clonal, phased methylomes distinguish between these. Across 28 canonical ICRs[33,34], methylation within colonies was consistently allele-specific, which we confirmed by phasing where available (**Fig.1g, Ext.DataFig4a**). However, candidate ICRs from broader catalogues[35] often behaved in blood as mixtures of methylation states across cells, or mixtures of unmethylated and methylated CpGs within a cell (**Ext.DataFig.4b)**.

Canonical ICRs were maintained as ASM in every healthy donor from foetus to old age, but upon this stable backdrop, epimutations nonetheless accumulated, following the same temporal structure as the wider methylome. Changes were concentrated before birth, but also elevated within clonal expansions (**Ext.DataFig.4c**). Both sets of branches had a higher proportion of C>T mutations at CpGs, a mutational footprint of spontaneous deamination of methylated cytosines[13] that is elevated when cell division is more rapid relative to elapsed time (**Ext.DataFig.4c**)[13]. Due to the heritability of these methylation changes, individual lineages thus carried distinct CpG configurations at ICRs that recapitulated the phylogeny, becoming readily visible following lineage expansion (**Ext.DataFig.4d**).

ICR methylation in donors with haematological cancers resembled that of healthy donors, with one exception. PD51634, with embryonic NF1 loss followed by *BCR::ABL1*-driven CML, showed extensive loss of methylation at *DIRAS3, FAM50B* and *GNAS* ICRs (**Ext.DataFig.4e)**. Colonies retained methylation at one locus but not for another, and vice versa; such reciprocal discordance was incompatible with a single imprinting error in one ancestral cell (**Ext.DataFig.4f)**. Instead, the embryo appears to have passed through a window of vulnerability where imprint protection was deficient across several early cell generations. The minimum number of ancestral events required to explain the losses placed this window as early as the first three cell divisions of life (**Ext.DataFig.4f)**. All three affected ICRs are maternally methylated, consistent with multi-locus imprinting disturbance (MLID), and usually attributable to constitutional defects in proteins that shield methylated ICRs after fertilisation[36]. Here, we show that such errors can arise post-zygotically.

Epigenetic clocks record cell divisions rather than chronological age

Two classes of CpGs have attracted attention as biomarkers of ageing and clonal dynamics. The Horvath clock uses a weighted set of CpGs whose combined methylation in bulk predicts chronological age[7]; while fluctuating CpGs (fCpGs), defined by intermediate mean methylation but high variance, are interpreted as switching at a fixed rate per cell division and used to infer stem cell dynamics[8,37]. We asked what these biomarkers record in our phylogenies, analysing both with a constant epimutation rate model, so that any concentration in development could not arise from our two-epoch model by construction.

Published fCpG panels were concentrated in single CpGs and small segments (**Ext.DataFig.5a**). As expected for sites selected for high variance, they changed faster than the genome-wide background average (**Ext.DataFig.5b**), but followed the same temporal pattern as the methylome at large. Two-thirds of fCpG divergence accumulated before birth (**Ext.DataFig.5c**), closely matching the whole-methylome background (**Fig.2h-i**) as well as background CpGs matched for fCpG context (**Ext.DataFig.5d**). As at ICRs, fCpG change was highest on branches with elevated

C>T proportions at CpGs (**Ext.DataFig.5e-f**). Change at these sites is therefore division-coupled rather than time-coupled.

Being heritable, methylation states established early in life at reported fCpGs persisted. An example is three fCpGs within one segment that changed repeatedly in the foetal sample, on different parental alleles (**Ext.DataFig.5g**). Such inherited states became prominent later in life as they hitchhiked with CH (**Ext.DataFig.5h**). In longitudinal samples taken 4 and 12 years apart, these fCpGs remained unchanged from early life, despite intervening clonal sweeps (**Ext.DataFig.5i-j)**, and the twins also shared their developmentally established fCpG methylation states **(Ext.DataFig.5k**) despite one developing cancer. fCpGs can therefore track clones over decades[9], as can the methylome at large. They are best understood as a mitotic readout, informative within a clone, but read against a stable backdrop of developmental mosaicism, and therefore, sensitive to the clonal composition of a tissue.

The same applies to the Horvath clock. Aggregating colony methylation into pseudo-bulks, epigenetic age of donors closely tracked chronological age (slope 1.01+/- 0.19, **Ext.DataFig.6a**). Two donors, however, lay 23-33 years above their chronological age, and both harboured rapidly expanding clones: *BCR::ABL1*-driven CML (PD51634) and an advanced MPN clone (PD5847). To ask whether this was a property of the individuals or their clones, we compared Horvath residuals of clonal expansions against their matched wild-type colonies (**Ext.DataFig.6b**). Acceleration was confined to clonal expansions in MPN and CML donors, reaching +40 years in PD5182, whose disease had rapidly progressed, in keeping with epigenetic age acceleration previously reported in MPN[38]. Most CH clones showed little acceleration, consistent with both their slower expansion[11,39] and with a clock calibrated in ageing populations in which CH is common. The Horvath clock has been argued not to report cell division[40,41]. Yet within donors, where every cell shares the same chronological age and tissue, Horvath residuals increased with increasing proportion of C>T at CpGs (Pearson r=0.73; **Ext.DataFig.6c**). In other words, among cells of the same chronological age, differences in cell division largely account for differences in epigenetic age.

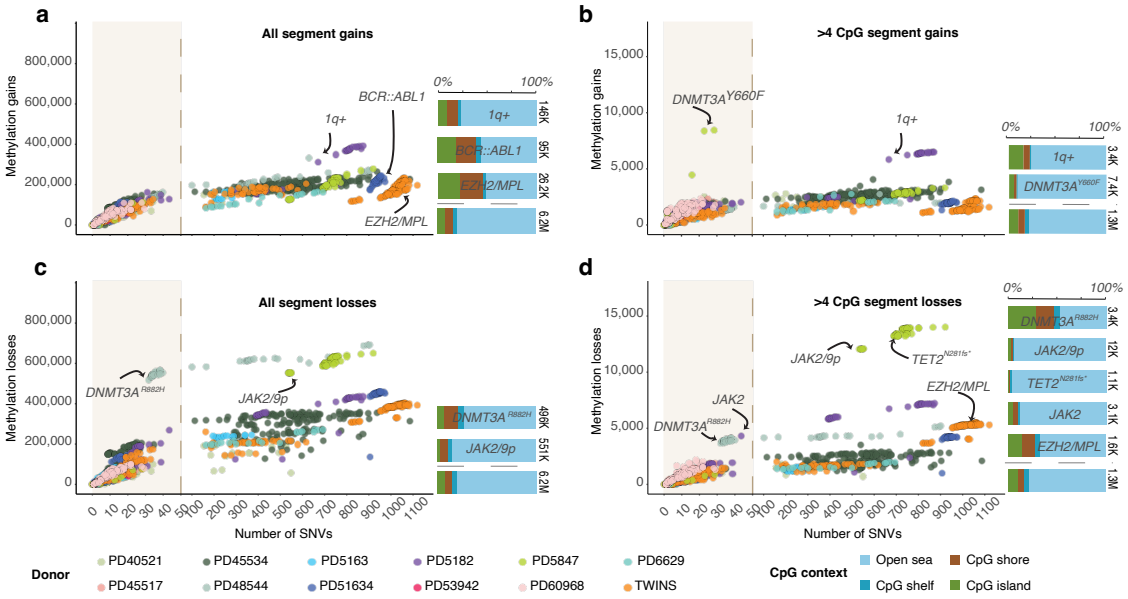
Two further observations establish this clearly. Monozygotic twins, identical in germline genome and age, differed in epigenetic age by 15 years, the twin with MPN appearing older (**Ext.DataFig.6a,b**). In PD51634, epigenetic age fell from 74 to 39 years over 4 years of successful CML treatment; the patient becoming epigenetically younger over four years of chronological ageing. Epigenetic age thus reports the division history of the cellular composition of the haematopoietic pool, rather than the age of the individual, and can move in either direction as that composition changes. Taken together, Horvath sites and fCpGs are two readouts of one process, and that process is neither elapsed time nor a mechanism peculiar to either set of sites. Both sample division-coupled methylation change effectively, but are a measure of years only where division rate is itself predictable from time, as in adult homeostasis, and not during development or clonal expansion.

Methylation changes in clonal expansions: acquired, inherited and epigenetic drivers

Methylation differences between clonal expansions and their contemporaries could have been inherited by the founding cell prior to expansion, or acquired in the lineage as it expanded. The phylogenies distinguish between these, and allow us to ask whether any such changes could drive clonal expansion itself. Several clonal expansions showed step changes in methylation on their trunks (**Fig.4a-d**). 1q+ in PD5182 was accompanied by methylation gain, whereas *DNMT3A*^{R882H} and JAK2/9pUPD trunks (PD48544 and PD5847) showed methylation loss**.** Across 34 trunks associated with diverse drivers (*DNMT3A, TET2, EZH2, MPL, JAK2, chr 1q+, NF1, BCR::ABL1*), and some lacking a driver (**Ext.DataFig.2a)**, we identified 301,747 truncal methylation changes that were stably inherited by descendants. Over 5,000 changes were at larger segments **(Fig.4e)** - these were enriched for CpG islands, bivalent and Polycomb-associated promoters and enhancers, and EZH2 and SUZ12 binding sites (**Fig.4f-h**). Every expansion

with sufficient truncal changes showed this pattern (**Fig.4i)** irrespective of the driver, whether a driver was present at all, or when the expansion arose; indeed, the 1q+ clade, which arose later in life, showed the strongest enrichment. The direction of change, however, depended on genotype. Most expansions gained methylation at these regions, as long recognised in cancer[42] (**Fig.4i**), with *DNMT3A*-mutant clades being the exception. *DNMT3A*^{R882H} (PD48544) lost methylation at these regions, as reported in CH and AML[43,44], despite being as strongly enriched for changes at these sites as other expansions. Two further *DNMT3A*-mutant clades changed in mixed directions (*DNMT3A*$^{c.1015\text{-}1G>C}$, *DNMT3A*^{F752S}), while a third resembled unmutated expansions (*DNMT3A*^{P904L})(**Fig.4i**). That the direction of methylation change depends on *DNMT3A* genotype suggests DNMT3A is itself depositing this methylation.

Figure 4



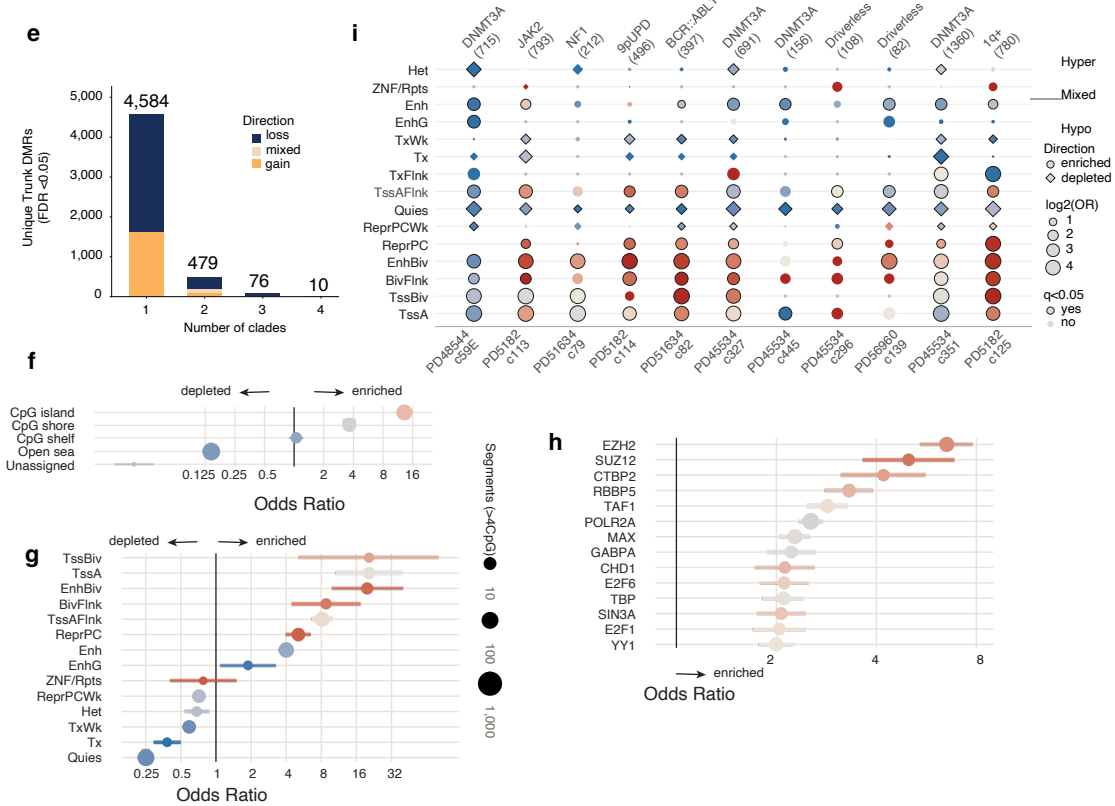


Figure 4 Legend. Recurrent methylation changes in clonal expansions. a-d. *Cumulative methylation changes with molecular time, stratified by gains **(a,b)** and losses **(c,d),** shown for all segments **(a,c)** and only larger (>4 CpG) segments **(b,d).** Each point is an internal phylogenetic branch from a donor (colour) showing the cumulative number of methylation-segment changes accrued from the zygote to the end of that branch against the corresponding cumulatively accrued SNVs. Arrows indicate branches showing an excessive number of methylation changes causing a sudden hike in methylation change that are often observed in association with acquisition of a driver mutation on the trunk of a clonal expansion. Bars beside each panel show the CpG context of these excessive changes on highlighted branches in **(a-d)** together with absolute number of methylation changes on these branches. CpG context of the genome-wide background is shown as the lowest bar. **e.** Recurrent differentially methylation regions (DMR) on shared branches that immediately precede clonal expansions (trunk-DMRs) were identified for 34 clonal expansions (see Extended Data Fig.2) across phylogenies, focussing on segments with >4 CpGs. Number of per-clade Bonferroni-significant trunk-DMRs, their recurrence across clades (from 1 to 4), coloured by dominant direction of methylation change (loss, gain, mixed). Of 5,149 segments significant in at least one clade, 565 are recurrent in >1 clades. **f–h**, Odds-ratio enrichment at trunk-DMRs for CpG context (**f**), Roadmap 15-state chromatin state (**g**) and ENCODE transcription factor binding sites (**h**). **f** and **g** are tested against 1.3M genome-wide segments as the background, **h** against segments carrying at least one binding site. **i.** Enrichment of Roadmap Epigenomics ChromHMM state among trunk DMRs shown at a per clade level. Columns, clades with trunk DMRs; rows, chromatin states. Shape gives the direction of enrichment. Strength of enrichment is shown by circle size, and strength of depletion by diamond size, specifically, area is $|log_2(OR)|$, capped at the 95th percentile. Fill gives the dominant direction of change). Cells outlined in black are significant at BH-adjusted P<0.05, corrected across states within each clade. Grey dots mark clade-state combinations with no DMR calls.*

Truncal methylation changes are, by construction, only changes that occured in the founder cell of the expansion. To also capture methylation changes during the expansion, we compared promoter methylation between all colonies of clonal expansions and their matched background. *DNMT3A*^{R882H} clades (PD48544, PD6629) showed widespread hypomethylation predominantly affecting bivalent promoters, including at developmental regulators *EGFL7* and *NSD1* (**Ext.DataFig.7**). *DNMT3A*-mutant clades also showed differential methylation at genes, such as *NFIX*, *TCF4, CTNNA1, RAF1* and *IRF8,* with known roles in myeloid neoplasia and CH[45–50]. However, many DMRs were actually segments that changed recurrently, and early, across lineages and clonal expansions. We found 129 such DMRs, including the *PLD6* embryonic epiallele identified earlier (**Fig.2j**, **Ext.DataFig.8)**. Where all three methylation states occurred within embryonic epialleles, expansions appeared as hypermethylated or hypomethylated depending on the state they inherited, as at *FOCAD* (**Fig.5a)** or *EN2* (**Ext.DataFig.9a**). Where embryonic alleles only occupied two states, changes between clonal expansions and background lineages appeared unidirectional (**Ext.DataFig.9b,c**). At some segments, almost every lineage carried its own unique pattern of methylation, akin to a somatic barcode (**Ext.DataFig.9d**). Lastly, reported metastable epialleles, such as, *VTRNA2-1*[51], which differed between some of our donors as expected, also changed early between lineages, thus, demonstrating shared properties with embryonic epialleles.

A methylation change that drives clonal expansion must lie in the founding lineage. At three DMRs, changes were confined to clonal expansions, with no background changes were observed in lineages of the cord blood or foetus. First, the *BCR::ABL1* clade of PD51634 harboured ASM at the alternate promoter (Pa) of *ABL1* (**Fig.5b**), a region fully unmethylated in every other colony in the cohort and previously reported as methylated in CML[52–54]. Pa lies 3' of this patient's breakpoint and is therefore on the fusion itself. Secondly, the *CDKN2B* promoter in the same clade also transitioned from unmethylated to ASM. It also occurred in a late branch of PD6629, around 400 mutations of molecular time, confirming that it can arise as an independent methylation event outside of development (**Fig.5c**). *CDKN2B* encodes p15(INK4b), and promoter hypermethylation at this locus has been reported in several myeloid malignancies and blast crisis CML that this donor's disease subsequently transformed to[53,55]. Thirdly, methylation loss at the *CACNA1D* gene body was observed in two independent clades carrying

JAK2^{V617F}/9pUPD (PD5847, PD5182, resulting in *JAK2*^{V617F} homozygosity, **Fig.5d**); methylation change at the same region has previously been shown to drive overexpression of the CaV1.3 calcium channel in prostate cancer[56,57]. This methylation change also occurred later in life in one lineage elsewhere in the cohort (eg. PD56959), indicating that it is not a consequence of homozygosity but may instead promote selection. Each of these methylation changes arose at established cancer genes, moved against the methylation state of the cohort, and two arose in more than one donor with cancer - *CDKN2B* in two unrelated donors, and *CACNA1D* in two clades carrying the same driver combination. Taken together, these methylation changes bear the hallmarks of epigenetic drivers.

Methylation and impact on gene expression

For methylation to be a substrate of clonal selection, it must be heritable and variable between cells, both of which we show, but it also needs to alter phenotype. In two donors with matched expression (cord blood, PD45517; 77 years, PD45534), methylated promoters were almost uniformly silent at the colony level (**Fig.5e,f**). In contrast, lack of promoter methylation did not imply gene expression, as 1 in 7 genes with unmethylated promoters were nonetheless silent. Such promoters were strongly enriched for bivalent and Polycomb-associated regions and H3K27me3 (**Fig.5g,h)**, confirming at clonal resolution that methylation and Polycomb repression are complementary routes to gene silencing, as on the inactive X[58–60].(**Fig.3f**).

Differentially methylated segments in clonal expansions and embryonic epialleles, often lay at alternative promoters, enhancers and gene bodies, where methylation has an established regulatory role[61]. Grouping colonies from the same donor by their methylation state at these sites revealed associations with transcript length, transcription initiation zone usage, and gene expression levels (**Ext.DataFig.10)**. At the *FOCAD* embryonic epiallele, which lay directly over a transcription initiation zone (IZ), absence of methylation promoted transcription initiation at IZ-1, whereas methylation redistributed initiation towards IZ-2 and IZ-3 (**Fig.5i)**. This relationship was independently replicated in both cord blood and the 77 year old donor (**Ext.DataFig.10a,b)**. At the gene level, allele-specific methylation at the promoter of *PLD6,* our first discovered embryonic epiallele, resulted in undetectable gene expression compared to colonies with unmethylated promoters (**Ext.DataFig.10c).** Clonal expansion related DMRs from the 77-year old donor, were initially identified within *DNMT3A*-mutated clades but different methylation states were subsequently also observed in the cord blood. Here too, changes in transcript and transcription initiation zone usage was observed for different methylation states (**Ext. DataFig.10d-f)**. At *TCF4* and *CTNNA1,* methylation at a DMR was accompanied by additional initiation of transcription leading to longer transcripts that were absent in unmethylated colonies. In contrast, at *RAF1,* a DMR within an H3K27me3 domain, this was reversed with longer transcripts seen only in unmethylated colonies. These genes have known roles in haematopoietic stem cell renewal and cancer[49,50,62].

Taken together, methylation changes, whether arising early in life or within clonal expansions, determine which transcript a gene produces, or whether it produces any at all. We lacked expression data from the donors carrying methylation changes in *CDKN2B, ABL1* and *CACNA1D*, but each of these methylation changes lay where methylation would be expected to be consequential. *CDKN2B* is unmethylated and expressed in cord blood and the older donor, and therefore methylation gain would be expected to lead to transcriptional repression, as at *PLD6*. *CACNA1D* is stably methylated and not expressed, so loss of gene-body methylation would be expected to de-repress it. By contrast, *ABL1* Pa is methylated within the *BCR::ABL1* fusion only, where methylation would be expected to suppress internal initiation in favour of transcription driven by the promoter of *BCR,* as at *FOCAD* and *RAF1*.

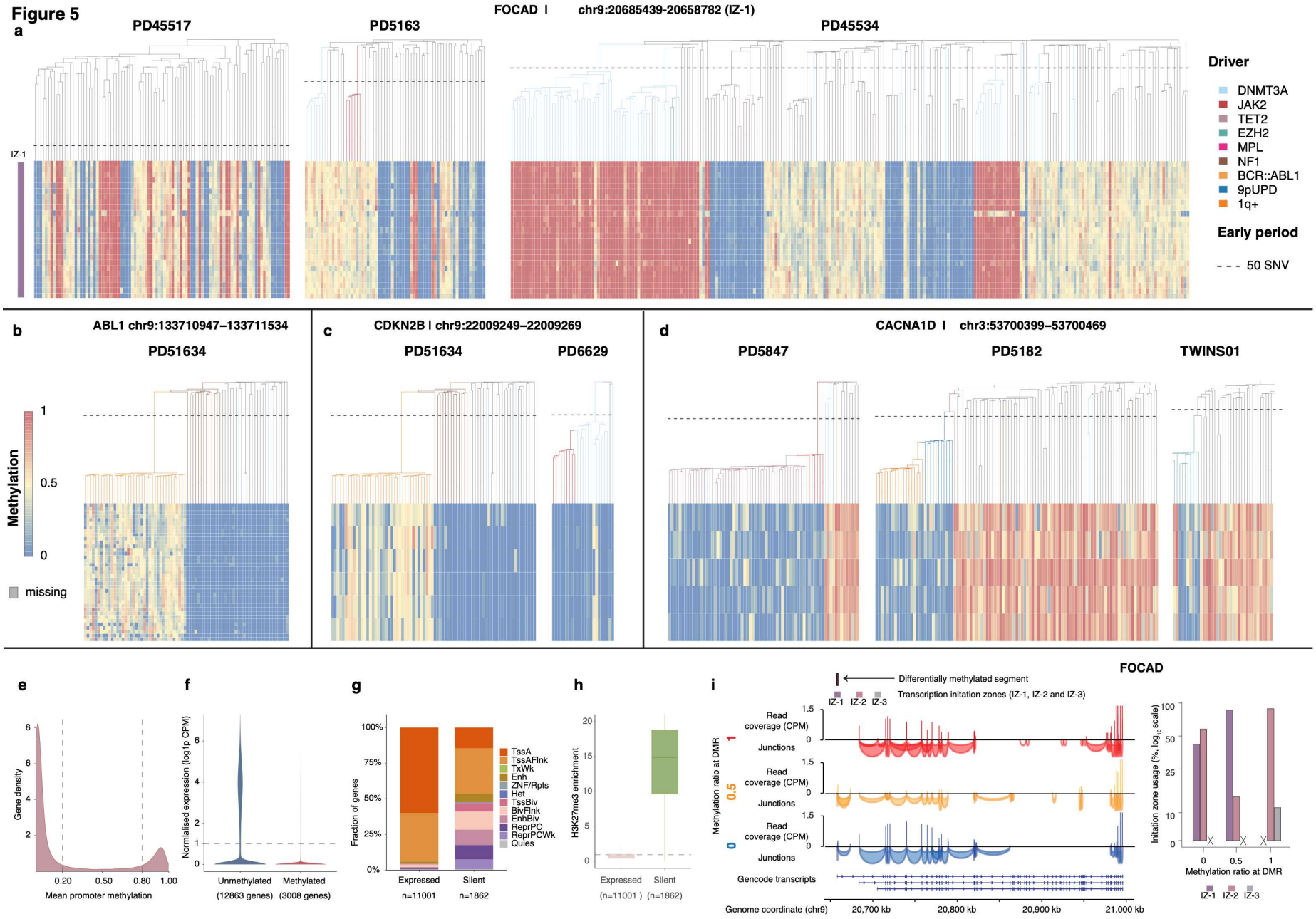

Figure 5
a
FOCAD | chr9:20685439-20658782 (IZ-1)
PD45517
PD5163
PD45534
IZ-1
Driver
DNMT3A
JAK2
TET2
EZH2
MPL
NF1
BCR::ABL1
9pUPD
1q+
Early period
50 SNV
b
ABL1 chr9:133710947–133711534
PD51634
Methylation
1
0.5
0
missing
c
CDKN2B | chr9:22009249–22009269
PD51634
PD6629
d
CACNA1D | chr3:53700399–53700469
PD5847
PD5182
TWINS01
e
Gene density
Mean promoter methylation
f
Normalised expression (log1p CPM)
Unmethylated (12863 genes)
Methylated (3008 genes)
g
Fraction of genes
Expressed n=11001
Silent n=1862
TssA
TssAFlnk
TxWk
Enh
ZNF/Rpts
Het
TssBiv
BivFlnk
EnhBiv
ReprPC
ReprPCWk
Quies
h
H3K27me3 enrichment
Expressed (n=11001)
Silent (n=1862)
i
FOCAD
Differentially methylated segment
Transcription initation zones (IZ-1, IZ-2 and IZ-3)
IZ-1 IZ-2 IZ-3
Read coverage (CPM)
Junctions
Methylation ratio at DMR
Gencode transcripts
Genome coordinate (chr9)
20,700 kb
20,800 kb
20,900 kb
21,000 kb
Initation zone usage (%, log10 scale)
Methylation ratio at DMR

Figure 5 Legend. Promoter methylation and transcription in matched colonies. a. *Per-CpG, per-colony DNA methylation across the FOCAD promoter locus in three donors (PD45517, PD5163, PD45534) shown below each donor's phylogeny. Colour shows methylation fraction, 0, blue; 0.5, yellow; 1, red; grey, no coverage.* ***b-d.*** *Selected promoter methylation heatmaps shown beneath somatic phylogenies.* ***b****. The ABL1 promoter gains methylation (0>1) specifically on the BCR::ABL1 lineage in PD51634 and is unmethylated elsewhere in the cohort.* ***c****. The CDKN2B promoter gains methylation on the BCR::ABL1 lineage in PD51634 and in PD6629 (0>1).* ***d****. Fully methylated segment within CACNA1D loses methylation in two independent JAK2/9pUPD lineages in PD5847 and PD5182.* ***e-h.*** *Relationship between promoter DNA methylation, gene expression and chromatin states in an example healthy donor (PD45517).* ***e****. Distribution of mean promoter methylation (aggregated methylation β values over the canonical transcription start site +/- 500 bp, averaged across colonies); promoters are predominantly unmethylated (<0.2, dashed line) or fully-methylated (>0.8, dashed line).* ***f****. Per-colony monocyte expression (log1p counts per million, CPM) for genes with unmethylated versus methylated promoters.* ***g.*** *15-state ChromHMM composition of unmethylated promoters, split by expression groups, silent vs expressed (monocyte log1p(CPM) >= 1 in any colony) genes or silent.* ***h****. BLUEPRINT monocyte H3K27me3 enrichment at the same promoters compared to the autosomal baseline for unmethylated and expressed genes versus unmethylated yet silent genes.* ***i****. Methylation-dependent regulation of transcription initiation at the FOCAD locus. Top left, position of the differentially methylated segment relative to the gene's three transcription initiation zones (IZ-1, IZ-2 and IZ-3). Below, RNA-seq coverage and splice junctions are shown for colonies grouped by their methylation state at this segment (1, fully methylated; 0.5, half-methylated; 0, unmethylated), with GENCODE transcripts and genomic coordinates aligned beneath. Right, percentage usage of each initiation zone in each methylation group (log_{10} scale; crosses denote no detectable usage).*

DISCUSSION

Our somatic phylogenies overlaid with lineage- and time-stamped methylation transitions reveal the dynamics of CpG methylation across the genome and human lifespan. They provide a lens into the earliest divisions of life, otherwise inaccessible in humans, dating X-chromosome inactivation and an instance of imprinting failure to within the first divisions from the zygote. Tracing methylation change along the trees shows how it outpaces somatic mutation by six orders of magnitude before birth and four orders thereafter. Much like SNVs that arise on one allele at a time, methylation transitions move into or out of allele-specific methylation rather than switching directly between homozygous states; unlike SNVs which alter single bases, they often affect whole tracts of CpGs at once. Every cell therefore carries two records of its own history: a sparse one written in mutations, and a much richer one written in methylation.

Most of this change, in highly similar HSC-derived clones, is written early. The burst of early, stable, and often segmental methylation change highlights how each cell's methylome retains a rich and enduring record of its embryonic journey. Indeed, over half the methylation changes a cell will ever carry are acquired within the first ~30 mutations of molecular time, corresponding to the end of the first trimester[10], these are then held for the remaining decades of life. This developmental excess is not explained by the higher division rates earlier in life. Adult HSCs acquire 15-17 SNVs per year of which roughly one is division-related[63]; a foetus accumulates ~40 by 18 weeks post conception[10], compatible with a few dozen cell divisions in addition to the elapsed time. This equates to around 10,000 epimutations per cell division before birth, against a few thousand thereafter. So while division is more frequent in early life, it seems insufficient to explain the 60-fold fall in epimutation rates after birth. Methylation remodelling during development and cell specification may be an additional source. Why the embryo should write so much methylation diversity, whether it is the cost of haste that is simply tolerated, or whether it serves important functions, is unclear. Regardless of the origin, extensive intercellular epigenetic mosaicism is established before birth, and where this affects larger CpG segments, lineages take their embryonic epialleles through life.

This stable mosaic is the backdrop against which all subsequent evolution should be read. Horvath age tracks the division history of whichever clones comprise the bulk sample rather than the age of the individual, allowing it to fluctuate by 15 years between identical twins where one has a blood cancer, or fall by 35 years following eradication of a rapidly dividing CML clone. Fluctuating CpGs are likewise sites of high local change, with two-thirds of their divergence established before birth. The same conflation runs through cancer epigenetics. Many of the methylation differences between clonal expansions and their background did not emerge during expansion, but were inherited by the founder cell and made visible by its subsequent expansion, something distinguishable only at lineage resolution.

What lineages do acquire are consistent methylation gains at Polycomb-regulated regions, in both early and late clonal expansions, and irrespective of the initiating driver, or indeed if one is even present. Cell division may displace PRC2, exposing these sites to *de novo* methyltransferases[42]. Mutated *DNMT3A,* however, resulted in either an inability for such sites to gain methylation in the first place, or its loss, suggesting that DNMT3A is responsible for these changes, and raising the possibility that mutant cells do not acquire the constraints that other expanding cells do. How much this convergence contributes to increased fitness rather than simply recording expansion is unknown.

A few changes, however, are the clone's own. At three loci otherwise stable across the cohort, methylation changed in a manner bearing the hallmarks of epigenetic drivers. *CDKN2B*, a tumour-suppressor known to be hypermethylated in haematological cancers[64,65], gained promoter methylation prior to leukaemic transformation in one donor, and arose independently within a clonal expansion in another. The Pa promoter of *ABL1* gained methylation only in a *BCR::ABL1* clone, potentially allowing it to suppress internal initiation in favour of *BCR*-driven expression of the fusion. Loss of methylation at *CACNA1D* accompanied *JAK2* homozygosity in two different donors, at the same site where demethylation increases transcription in solid tumours[56,66]. That these converge on genes already implicated in cancer, and recur independently in separate lineages, suggests they confer a selection advantage rather than drift.

Somatic methylation, then, is not merely a record of a cell's history. It is heritable, it differs between similar cells within one person, and it can have direct phenotypic consequences, influencing which transcripts a gene produces or whether a gene is expressed at all. This applies to changes established in embryogenesis, which are therefore not inert. At the risk of over-extending the metaphors in this field, embryonic epialleles are not simply tide marks on the methylome but changes to the sandbanks themselves. Together with the methylation changes clones acquire as they expand, they establish somatic methylation mosaicism as a pervasive additional substrate for Darwinian evolution.

REFERENCES

1. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
2. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
3. Sharif, J. *et al.* The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908–912 (2007).
4. Bernstein, N. *et al.* Analysis of somatic mutations in whole blood from 200,618 individuals identifies pervasive positive selection and novel drivers of clonal hematopoiesis. *Nat. Genet.* **56**, 1147–1155 (2024).
5. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–2433 (2010).
6. Delhommeau, F. *et al.* Mutation in TET2 in myeloid cancers. *N Engl J Med* **360**, 2289–301 (2009).
7. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
8. Gabbutt, C. *et al.* Fluctuating DNA methylation tracks cancer evolution at clinical scale. *Nature* **645**, 764–773 (2025).
9. Scherer, M. *et al.* Clonal tracing with somatic epimutations reveals dynamics of blood ageing. *Nature* **643**, 478–487 (2025).
10. Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
11. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
12. Kapadia, C. D. *et al.* Clonal dynamics and somatic evolution of haematopoiesis in mouse. *Nature* **641**, 681–689 (2025).
13. Kamizela, A. E. *et al.* Timing and trajectory of BCR::ABL1-driven chronic myeloid leukaemia. *Nature* **640**, 982–990 (2025).
14. Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).

15. Machado, H. E. *et al.* Convergent somatic evolution commences in utero in a germline ribosomopathy. *Nat. Commun.* **14**, 5092 (2023).
16. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
17. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
18. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
19. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
20. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
21. Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
22. Gaiti, F. *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).
23. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).
24. Shearstone, J. R. *et al.* Global DNA demethylation during mouse erythropoiesis in vivo. *Science* **334**, 799–802 (2011).
25. Yu, Y. *et al.* High resolution methylome analysis reveals widespread functional hypomethylation during adult human erythropoiesis. *J. Biol. Chem.* **288**, 8805–8814 (2013).
26. Xu, C. & Corces, V. G. Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. *Science* **359**, 1166–1170 (2018).
27. Laird, C. D. *et al.* Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 204–209 (2004).
28. Becker, C. *et al.* Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* **480**, 245–249 (2011).

29. Ju, Y. S. *et al.* Embryonic origins of epigenetic mosaicism in the human colorectal epithelium. Preprint at https://doi.org/10.21203/rs.3.rs-11351803/v1 (2026).
30. Monteiro, J. *et al.* Commitment to X inactivation precedes the twinning event in monochorionic MZ twins. *Am. J. Hum. Genet.* **63**, 339–346 (1998).
31. Payer, B., Lee, J. T. & Namekawa, S. H. X-inactivation and X-reactivation: epigenetic hallmarks of mammalian reproduction and pluripotent stem cells. *Hum. Genet.* **130**, 265–280 (2011).
32. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).
33. Rosenski, J. *et al.* Atlas of imprinted and allele-specific DNA methylation in the human body. *Nat. Commun.* **16**, 2141 (2025).
34. Skaar, D. A. *et al.* The Human Imprintome: Regulatory Mechanisms, Methods of Ascertainment, and Roles in Disease Susceptibility. *ILAR J.* **53**, 341–358 (2012).
35. Jima, D. D. *et al.* Genomic map of candidate human imprint control regions: the imprintome. *Epigenetics* **17**, 1920–1943 (2022).
36. Mackay, D. J. G. *et al.* Multi-locus imprinting disturbance (MLID): interim joint statement for clinical and molecular diagnosis. *Clin. Epigenetics* **16**, 99 (2024).
37. Gabbutt, C. *et al.* Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nat. Biotechnol.* **40**, 720–730 (2022).
38. Vieri, M. *et al.* Cellular aging is accelerated in the malignant clone of myeloproliferative neoplasms. *Blood Cancer J.* **13**, 164 (2023).
39. Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
40. Yang, Z. *et al.* Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol.* **17**, 205 (2016).
41. Bell, C. G. *et al.* DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* **20**, 249 (2019).
42. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.* **39**, 232–236 (2007).

43. Nam, A. S. *et al.* Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. *Nat. Genet.* **54**, 1514–1526 (2022).
44. Spencer, D. H. *et al.* CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. *Cell* **168**, 801-816.e13 (2017).
45. Kulkarni, R. *et al.* Regulation of hematopoietic stem cell (HSC) proliferation by Epithelial Growth Factor Like-7 (EGFL7). Preprint at https://doi.org/10.1101/2025.01.21.634107 (2025).
46. Qiu, Q., Liu, P., Liu, F. & Ren, R. Irf8 Regulates Cytokine Signaling and Proliferation of Hematopoietic Stem Cells. *Blood* **130**, 2432 (2017).
47. Holmfeldt, P. *et al.* Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood* **122**, 2987–2996 (2013).
48. Cytlak, U. *et al.* Differential IRF8 Transcription Factor Requirement Defines Two Pathways of Dendritic Cell Development in Humans. *Immunity* **53**, 353-370.e8 (2020).
49. Hout, F. E. M. in 't, Reijden, B. A. van der, Monteferrario, D., Jansen, J. H. & Huls, G. High expression of transcription factor 4 (TCF4) is an independent adverse prognostic factor in acute myeloid leukemia that could guide treatment decisions. *Haematologica* **99**, e257–e259 (2014).
50. Li, M. *et al.* CTNNA1 hypermethylation, a frequent event in acute myeloid leukemia, is independently associated with an adverse outcome. *Oncotarget* **7**, 31454–31465 (2016).
51. 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).
52. Sun, B. *et al.* ABL1 Promoter Methylation Can Exist Independently of BCR-ABL Transcription in Chronic Myeloid Leukemia Hematopoietic Progenitors. *Cancer Res.* **61**, 6931–6937 (2001).
53. Nguyen, T. T. *et al.* Quantitative measure of c-abl and p15 methylation in chronic myelogenous leukemia: biological implications. *Blood* **95**, 2990–2992 (2000).
54. Zion, M. *et al.* Progressive de novo DNA methylation at the bcr-abl locus in the course of chronic myelogenous leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10722–10726 (1994).
55. Janssen, J. J. W. M. *et al.* Methylation patterns in CD34 positive chronic myeloid leukemia blast crisis cells. *Haematologica* **95**, 1036–1037 (2010).

56. Geybels, M. S. *et al.* Epigenomic profiling of prostate cancer identifies differentially methylated genes in TMPRSS2:ERG fusion-positive versus fusion-negative tumors. *Clin. Epigenetics* **7**, 128 (2015).
57. Chen, R. *et al.* Cav1.3 channel α1D protein is overexpressed and modulates androgen receptor transactivation in prostate cancers1. *Urol. Oncol. Semin. Orig. Investig.* **32**, 524–536 (2014).
58. Kondo, Y. *et al.* Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat. Genet.* **40**, 741–750 (2008).
59. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
60. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
61. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).
62. Rahimpour, S. K. *et al.* Diagnostic and prognostic relevance of RAF1 gene in acute myeloid leukemia. *Leuk. Res. Rep.* **24**, 100528 (2025).
63. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
64. Cameron, E. E., Baylin, S. B. & Herman, J. G. p15(INK4B) CpG island methylation in primary acute leukemia is heterogeneous and suggests density as a critical factor for transcriptional silencing. *Blood* **94**, 2445–2451 (1999).
65. Paul, T. A., Bies, J., Small, D. & Wolff, L. Signatures of polycomb repression and reduced H3K4 trimethylation are associated with p15INK4b DNA methylation in AML. *Blood* **115**, 3098–3108 (2010).
66. Chen, R. *et al.* Cav1.3 channel α1D protein is overexpressed and modulates androgen receptor transactivation in prostate cancers. *Urol. Oncol.* **32**, 524–536 (2014).
67. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* 1–16 (2017) doi:10.1200/PO.17.00011.

68. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* 1–31 (2020) doi:10.1038/s41596-020-00437-6.

69. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).

70. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

71. Babraham Bioinformatics - Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

72. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

73. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

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

75. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

76. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).

77. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

78. Martin, M., Ebert, P. & Marschall, T. Read-Based Phasing and Analysis of Phased Variants with WhatsHap. *Methods Mol. Biol.* **2590**, 127–138 (2023).

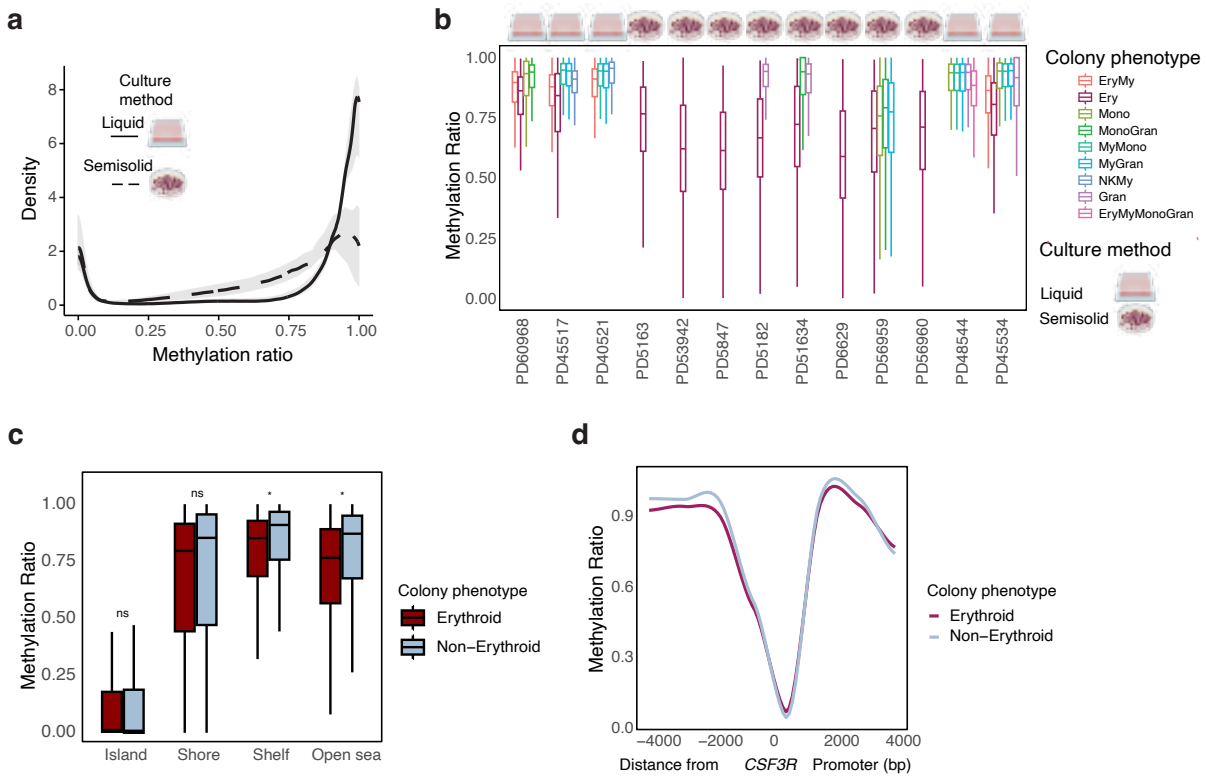
79. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).

80. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

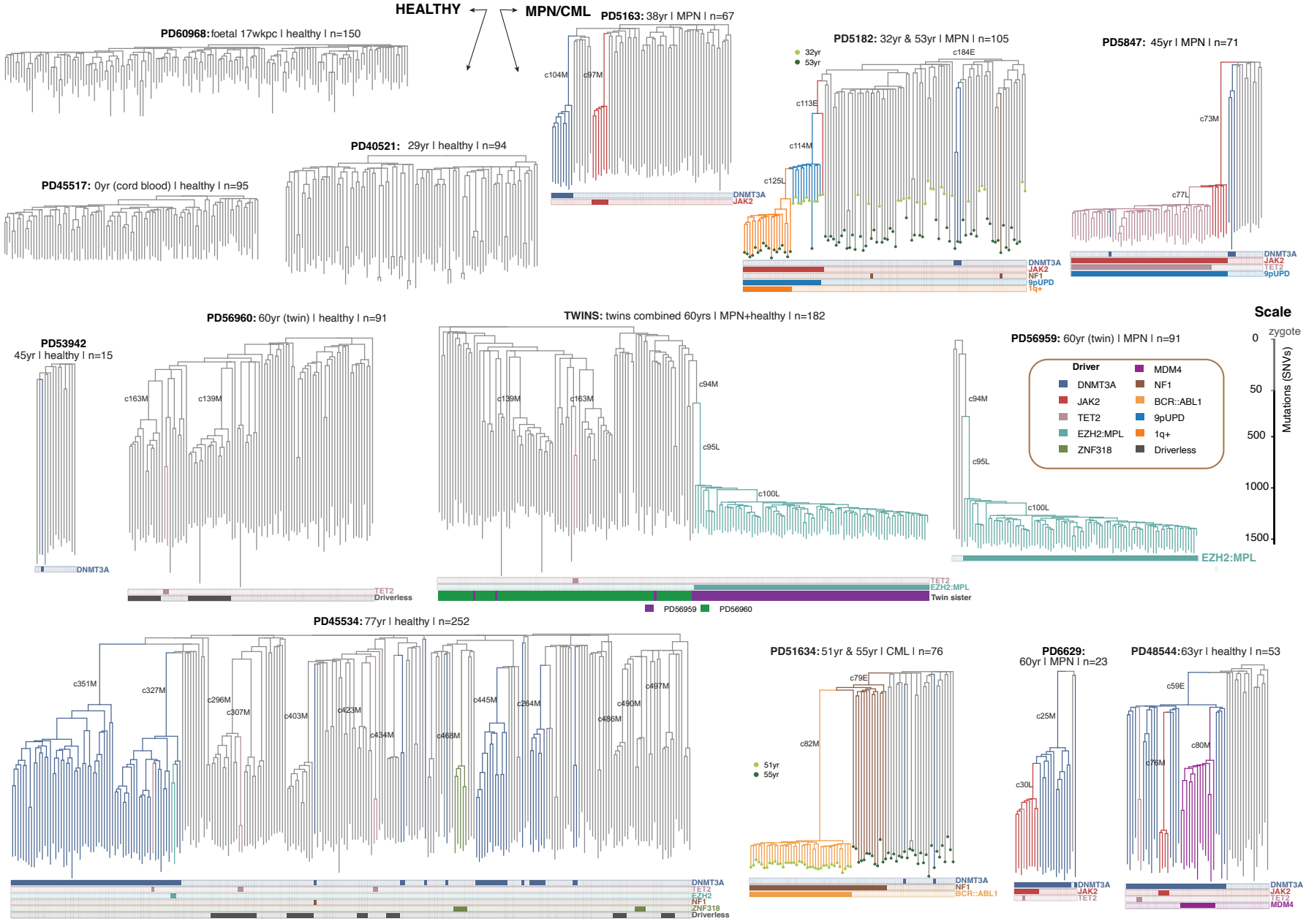
81. Twelve years of SAMtools and BCFtools | GigaScience | Oxford Academic. https://academic.oup.com/gigascience/article/10/2/giab008/6137722.

82. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).
83. Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.* **53**, D966–D975 (2025).
84. Fernández, J. M. *et al.* The BLUEPRINT Data Analysis Portal. *Cell Syst.* **3**, 491-495.e5 (2016).
85. Tung Ho, L. si & Ané, C. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Syst. Biol.* **63**, 397–408 (2014).
86. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
87. Cotto, K. C. *et al.* Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat. Commun.* **14**, 1589 (2023).
88. deepTools2: a next generation web server for deep-sequencing data analysis | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/44/W1/W160/2499308.

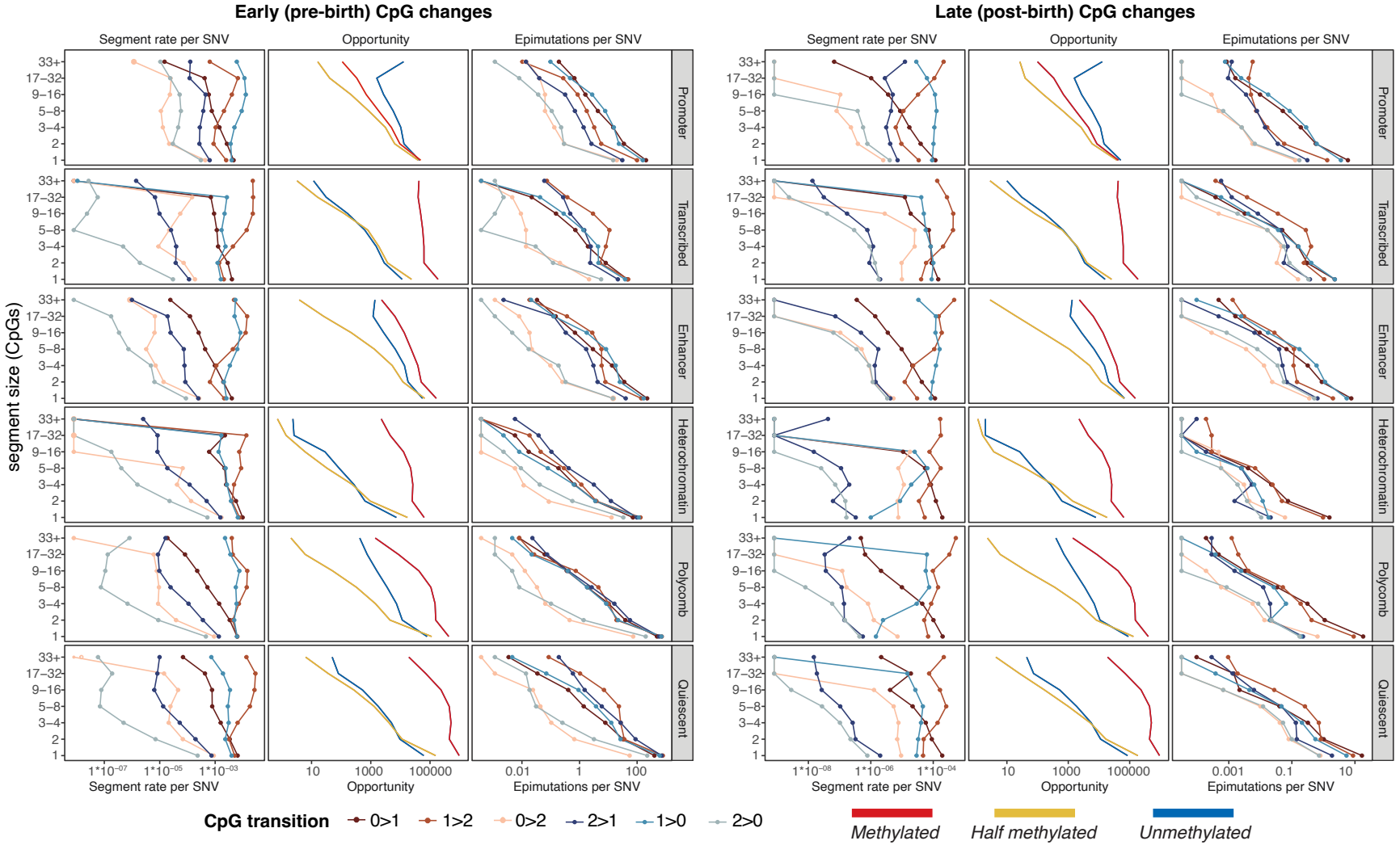
EXTENDED DATA FIGURES



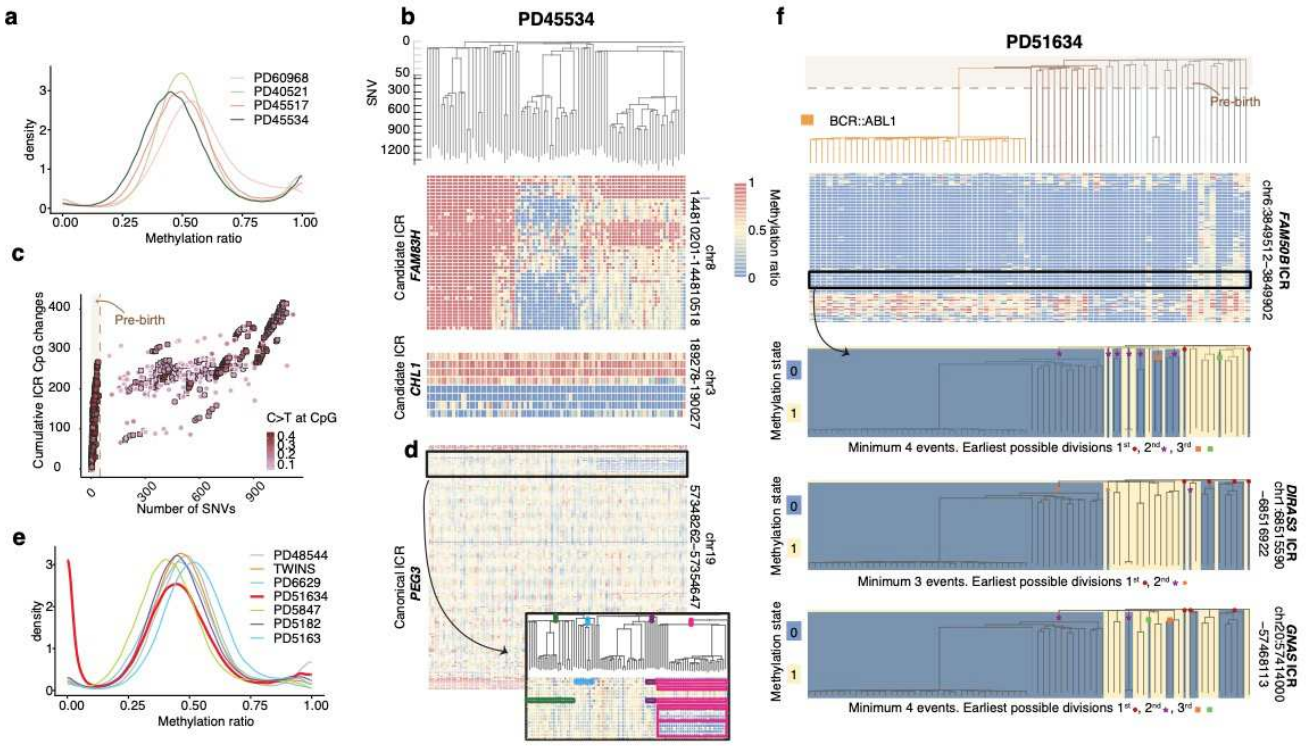
Extended Data Figure 1. Erythroid differentiation and genome-wide methylation changes. a. Density of per-CpG methylation ratio, liquid culture (solid line) vs MethoCult semisolid culture (dashed line), showing the shifted distribution underlying MethoCult colonies' higher mean culture-associated methylation loss. **b.** Per-donor, per-colony-phenotype (erythroid, myeloid subtypes, mixed lineage, etc.) pseudobulk methylation ratio distributions. **c.** Pseudobulk CpG methylation ratio, pooled across donors, by CpG context, split by erythroid (Ery/EryMy) vs. non-erythroid colonies. In both (**b**) and **(c)**, boxes show median, IQR, and whiskers extending to 1.5×IQR; outlying points beyond the whiskers are not shown individually. * $p < 0.05$; ns, not significant (Bonferroni-corrected, two-sided Wilcoxon signed-rank tests). **d.** CpG methylation profile across the *CSF3R* promoter and flanking region in PD40521. Points show per-CpG pseudobulk methylation ratio in Erythroid vs. non-erythroid colonies. In **(b)**, **(c)** and **(d)**, a thinned dataset was used, selecting CpGs at least 1000 base pairs apart.


HEALTHY
MPN/CML
PD60968: foetal 17wkpc I healthy I n=150
PD45517: 0yr (cord blood) I healthy I n=95
PD40521: 29yr I healthy I n=94
PD5163: 38yr I MPN I n=67
c104M
c97M
DNMT3A
JAK2
PD5182: 32yr & 53yr I MPN I n=105
32yr
53yr
c184E
c113E
c114M
c125L
DNMT3A
JAK2
NF1
9pUPD
1q+
PD5847: 45yr I MPN I n=71
c73M
c77L
DNMT3A
JAK2
TET2
9pUPD
PD53942
45yr I healthy I n=15
DNMT3A
PD56960: 60yr (twin) I healthy I n=91
c163M
c139M
TET2
Driverless
TWINS: twins combined 60yrs I MPN+healthy I n=182
c139M
c163M
c94M
c95L
c100L
TET2
EZH2:MPL
Twin sister
PD56959
PD56960
PD56959: 60yr (twin) I MPN I n=91
c94M
c95L
c100L
EZH2:MPL
Driver
DNMT3A
JAK2
TET2
EZH2:MPL
ZNF318
MDM4
NF1
BCR::ABL1
9pUPD
1q+
Driverless
Scale
zygote
0
50
500
1000
1500
Mutations (SNVs)
PD45534: 77yr I healthy I n=252
c351M
c327M
c296M
c307M
c403M
c423M
c434M
c468M
c445M
c264M
c486M
c490M
c497M
DNMT3A
TET2
EZH2
NF1
ZNF318
Driverless
PD51634: 51yr & 55yr I CML I n=76
c79E
c82M
51yr
55yr
DNMT3A
NF1
BCR::ABL1
PD6629:
60yr I MPN I n=23
c25M
c30L
DNMT3A
JAK2
TET2
PD48544: 63yr I healthy I n=53
c59E
c76M
c80M
DNMT3A
JAK2
TET2
MDM4

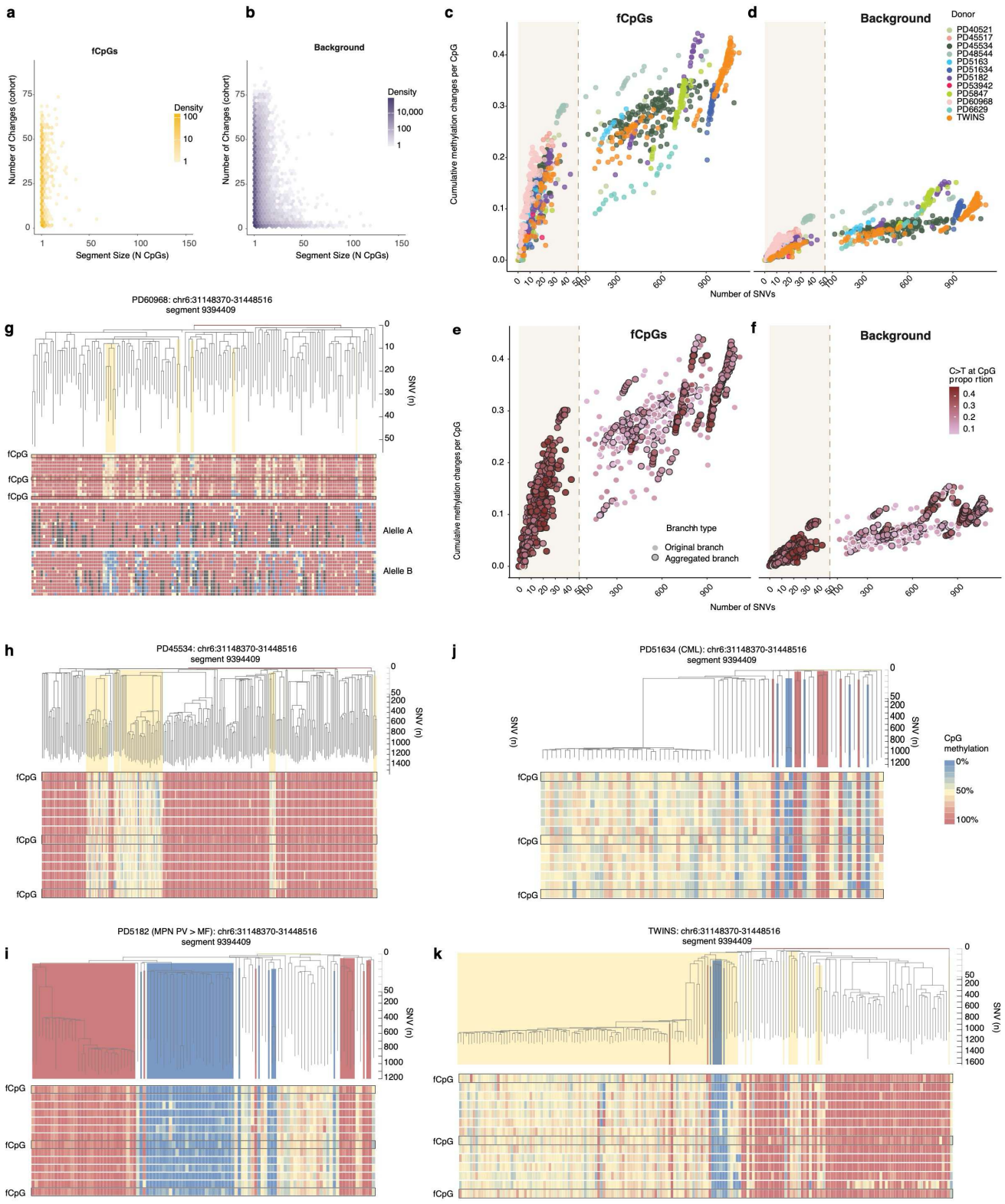
Extended Figure 2. Somatic mutation phylogenies of 13 donors. Phylogenies reconstructed from somatic DNA mutations, split by health (left) and disease (right) status. The twins PD56959 and PD56960 are shown individually and combined (TWINS), in which a strip marks each colony's twin of origin. Branch lengths are numbers of SNVs, with the first 50 SNVs magnified; tree width reflects the number of colonies. Branches are coloured by driver mutation, also shown below trees (grey, none). Colony tips are coloured by age at sampling for donors with multiple timepoints (PD5182, PD51634). Each tree is labelled with the individual, age at sampling, health status and number of colonies. Clades analysed for differential methylation are marked by "c_node_number_E/M/L". c_, clade; E, early – trunk ends before 50 SNV; L, late - the trunk commences after 50 SNV; M, mid – trunk timing spans early and late.



Extended Figure 3. Genome-wide epimutation rates. Attribution of genome-wide chromatin state specific segmental methylation transitions per SNV in terms of rate of transitions (first column), opportunity, ie., number of epoch specific segments with the parental starting state (second column; methylated, red; half-methylated, yellow; unmethylated, blue), and outcome, ie., number of transitions per SNV (third column, product of the first two columns). Within each panel, the y-axis is segment size (CpGs); rows are the six chromatin states; coloured lines are the 6 transition types in a segment's CpG methylation state. Estimates pool the 3 healthy donors PD45517, PD40521 and PD45534. **Left,** Early changes (at or before 50 mutations molecular time). **Right,** Later changes (after 50 mutations molecular time).

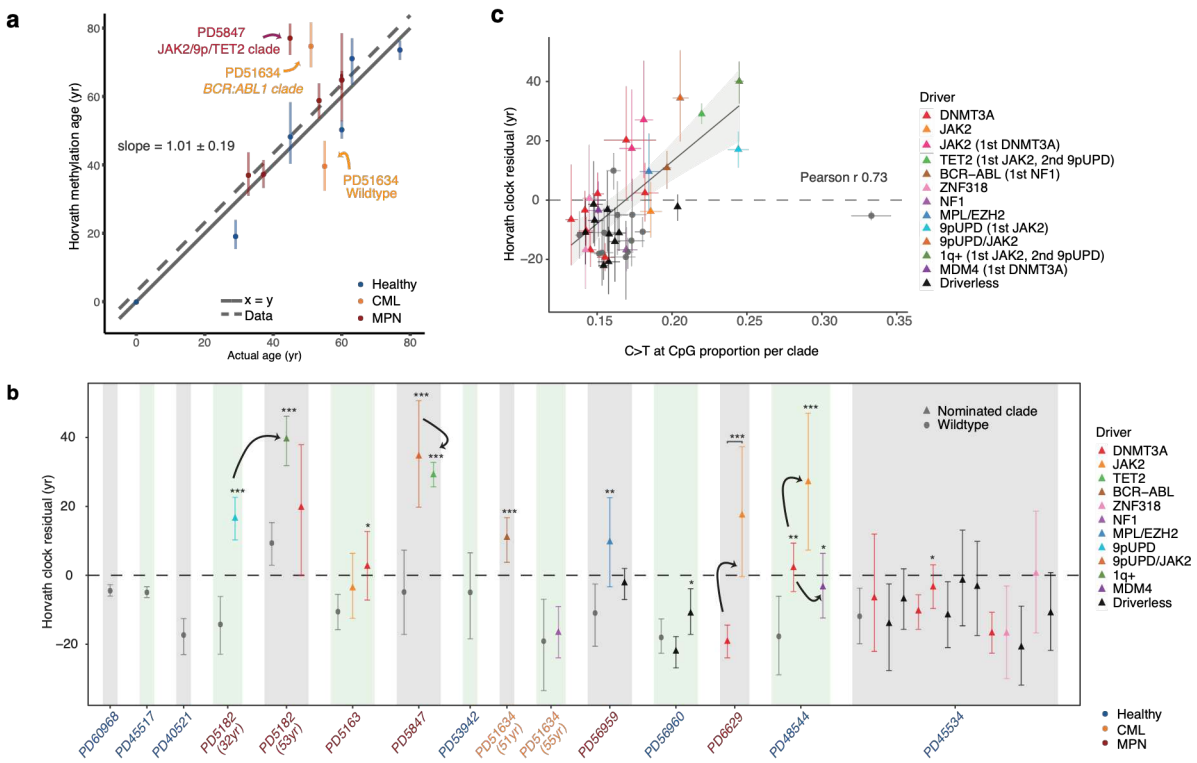


Extended Figure 4. Imprinting control regions (ICRs). a. Methylation ratio density plot at CpGs within 28 canonical ICRs in four healthy donors, from 17 weeks post-conception to 77 years, pooled across colonies. **b**. Methylation ratio heatmaps for two candidate ICRs[35] in PD45534 to demonstrate that they did not show allele-specific methylation in blood, instead showing a mixture of methylation states across individual colonies (*FAM83H,* top heatmap), or a mixture of CpG methylation states within the region within individual colonies (*CHL1*, bottom heatmap). Somatic phylogeny (partial) is shown above heatmaps. **c**. Cumulative methylation changes at canonical ICR CpGs against molecular time (SNV). Each point is one branch of a phylogeny, cumulative number of methylation state changes is calculated from the root to branch. Points are coloured by the branch's C>T at CpG fraction. Where branches have fewer than ~40 SNVs, neighbouring branches are pooled to estimate proportions (black circle outlines). Shaded band and dashed line mark before birth (<50 SNV). **d.** *PEG3* canonical ICR in PD45534 demonstrates subtle methylation gains and losses across the ICR region. The inset shows how these changes represent distinct events on embryonic branches of the phylogeny that are made visible by subsequent clonal expansions. **e**. As in **a,** but for donors with haematological malignancy. The red line shows one donor with ICR methylation loss. **f.** Loss of ICR methylation in PD51634. Top, phylogeny showing the *BCR::ABL1* clade in orange. Methylation ratio heatmap below for *FAM50B* with the boxed CpGs shown as methylation states inferred by *DAPHNE* on a tree below (0, unmethylated; 1, allele-specific methylation). Symbols mark the minimum number of independent losses needed to explain colony states, at their earliest possible cell divisions from the zygote (circle, 1st; star, 2nd; square, 3rd). Colours distinguish distinct methylation loss events at the locus, and repeated symbols mark lineages that could be explained by a single event at an unresolved node (polytomy). Methylation events are similarly shown for *DIRAS3* and *GNAS* ICRs, below *FAM50B.*

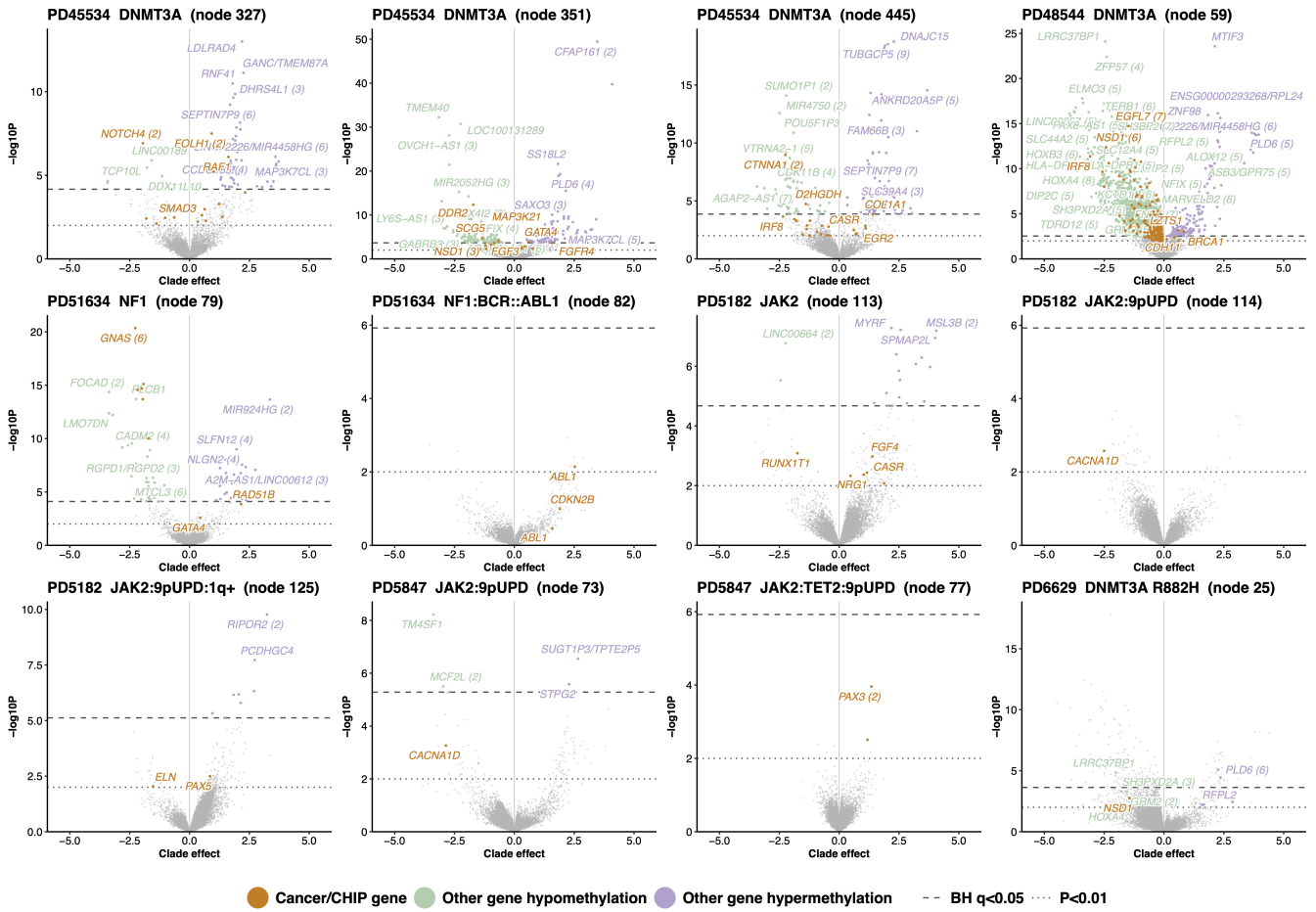


Extended Figure 5. Methylation changes at fluctuating CpGs. a,b. Segment size against the number of methylation changes on internal branches summed across the cohort, for segments containing at least one published fCpG (a) and all other segments (b). Only donor-segment pairs with a mean coverage of >=1 per CpG contribute. Colour indicates the number of segments per hexagon (log scale). **c,d.** Cumulative methylation changes per CpG from the root to each internal branch, against the branch's cumulative SNVs, for published fCpGs (c) and a background of CpGs matched for chromatin state and CpG context (d). Changes were inferred by a single-CpG model with time-constant rates. Points are coloured by donor; the first 50 SNVs are expanded. **e,f.** As in (c) and

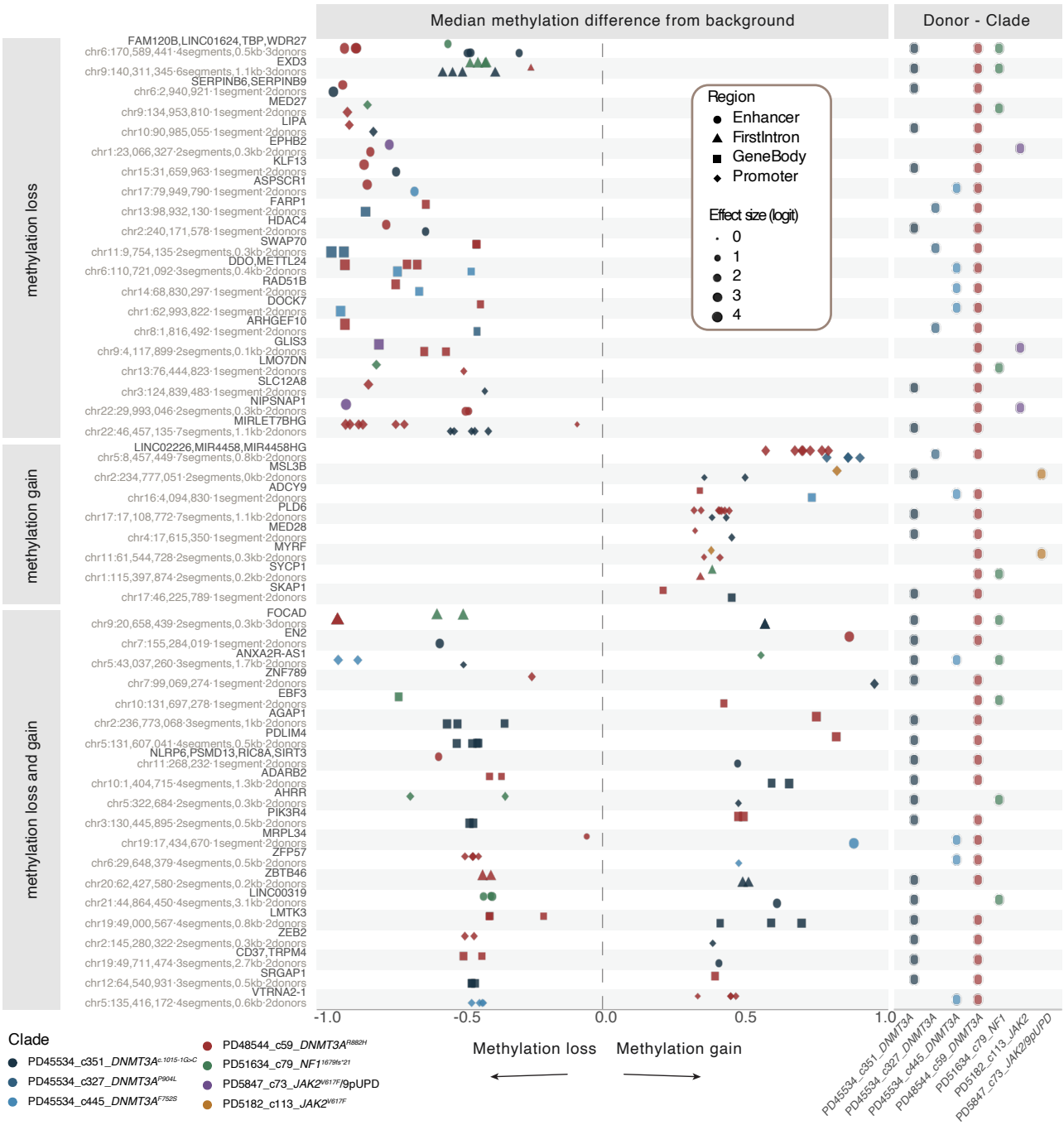
(d), with points coloured by the proportion of SNVs on the branch that are C>T at CpG, a proxy for cell divisions. Short branches were pooled into bins of >= 40 SNVs; outlined points mark branches in such bins. **g-k.** An fCpG-rich segment (chr6:31,148,370-31,148,516 harbouring 14 CpGs, three of which are published fCpGs in blood, boxed) in PD60968 (g), PD45534 (h), PD5182 (i), PD51634 (j) and TWINS (k). Each panel shows the phylogeny with the segment's inferred methylation state (shaded clades: blue, unmethylated; yellow, half-methylated; red, methylated) and per-CpG methylation of each colony (rows, CpGs; columns, colonies; black, no reads). Laurel defines this region as a single segment. In (g), phased per-CpG methylation is shown separately for the two parental alleles.



Extended Data Figure 6. Horvath CpGs. a. Horvath methylation age plotted against chronological age for donor-pseudobulk samples for each of the 13 individuals. Two patients, PD51634 and PD5182, were sampled at two time points. Points are coloured by health status (Healthy, blue; MPN, red; CML, orange). The dashed line shows the linear fit to the data (slope = 1.01 ± 0.19). The solid line shows identity (x = y). Error bars: 95% bootstrap CI (B = 200, tip-cell resampling). **b.** Horvath clock residuals for each nominated clonal expansion (clonal haematopoiesis (CH), MPN or CML clades) and the wildtype (WT) tips of each donor, relative to the donor-level chronological age fit shown in (a). Triangles, nominated clades; circles, WT. Colour indicates the clade's driver mutation as shown in Extended Data Figure 2 that also shows any nesting between clades. Error bars show the 95% bootstrap CI (B = 5000). Asterisks show the one-sided bootstrap p-value comparing each clade's residual to its donor-matched WT with *** $p < 0.0005$, ** $p < 0.005$, * $p < 0.05$. Black arrows show nesting between clades (ancestral>recent clone) **c.** Proportion of C>T-at-CpG (computed per nominated clade and per WT sample) plotted against each clade/WT's Horvath clock residual relative to its own donor (as in **b**). Triangles, nominated clades; circles, WT; colour indicates the driver mutation as shown in Extended Data Figure 2. The shaded band represents the linear fit to clonal expansions only, with 95% CI (Pearson r = 0.73). Vertical error bars indicate the 95% parametric bootstrap CI on the residual (B = 200). Horizontal error bars show 95% bootstrap CI on mean C>T proportions in cross-colony resampling. Driver mutations with preceding drivers in brackets.

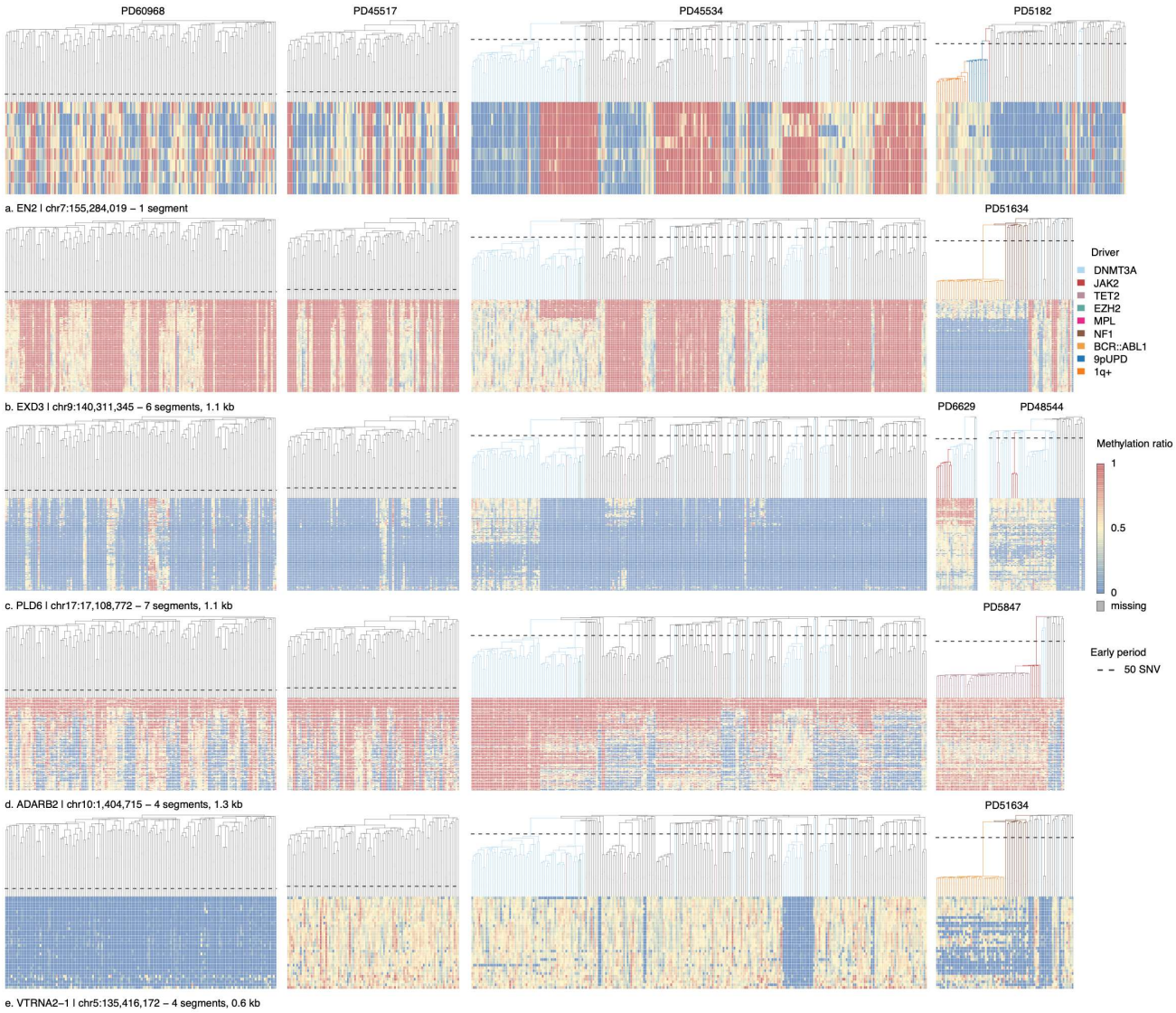


Extended Data Figure 7. Promoter methylation differences in driver clades. Each panel shows one nominated driver clade, titled with the donor, the drivers carried by the clade and the acquisition branch ID. Points are promoter segments with > 4 CpGs (segments within +/- 500 bp of the transcription start site of a UCSC knownGene transcript or of the GENCODE v48 MANE Select transcript): x, clade effect (difference in methylation between carrier and other colonies on the bounded logit scale, from phylogenetic regression by *phylolm*; see Methods); y, -log10(P-value). Shown are clades with >= 10 carrier and >= 10 non-carrier colonies except for PD6629, which has 2 non-carrier colonies but tested and shown for comparison with another *DNMT3A*^{R882H} clade in PD48544. Promoters of cancer[67] or clonal-haematopoiesis genes[4] are highlighted at P < 0.01 (dotted line; orange), and other segments at Benjamini-Hochberg q < 0.05 within the clade (dashed line; green, hypomethylated; purple, hypermethylated). Labels give gene names, with the number of significant segments in parentheses; the most significant segments of each class and genes with several significant segments are labelled. In PD6629, segments are instead highlighted if P < 0.01 and the gene is labelled in the PD48544 panel. In the BCR::ABL1 clade of PD51634, the *CDKN2B* and *ABL1* core-promoter segments discussed in the main text are also highlighted.

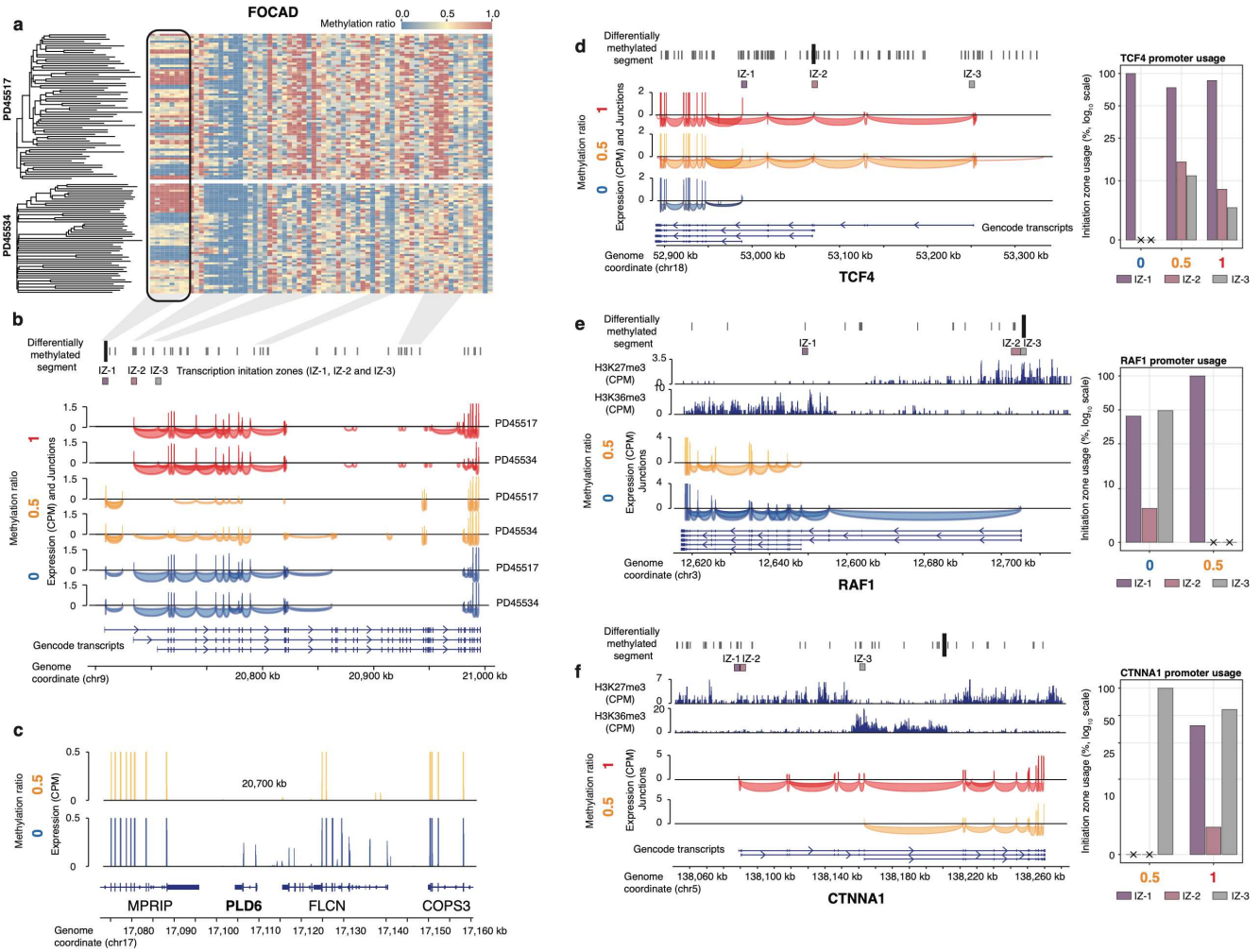


E

Extended Data Figure 8. Candidate epidrivers occurring early in life. Each row is a locus: early CH DMR segments within 3 kb are grouped and labelled with every gene their segments overlap. A DMR is shown where a segment is both a CH associated DMR within a clade (*phylolm*, Benjamini-Hochberg $p < 0.05$, $|\Delta \text{ methylation}| \geq 0.05$) and flagged as an early change in that clade's donor. Left, one point per (segment, clade) DMR, placed on the x-axis according to its Δ methylation versus normal. dashed line indicates no change; colour corresponds to clade; shape, to gene region; area to |effect size| on the logit scale. Right, a filled mark indicates for each locus, which clades carry the early DMR segments comprising the locus.



Extended Data Figure 9. Embryonic epialleles. a-e, Five loci: *EN2* (a), *EXD3* (b), *PLD6* (c), *ADARB2* (d) and *VTRNA2-1* (e), each labelled with the position of its first CpG (hg19), the number of segments in the locus and their total span. Each column is one donor: the first three columns are a foetal donor (PD60968), a cord-blood donor (PD45517) and a 77-year-old donor (PD45534), followed by other donors in which the locus changed within a driver clade. Top, the donor's phylogenetic tree, branches are coloured by the most recent driver carried (grey, none); the first 50 SNVs are magnified, and the dashed line marks the end of this early period. Bottom, methylation ratio (methylated reads/total reads) of each CpG of the locus (rows, in genomic order) in each colony (columns, aligned with the tree tips); grey, no reads.



Extended Data Figure 10. DNA methylation state at DMRs impacts promoter usage. a. Heatmap of DNA methylation (beta values) across differentially methylated segments (columns, ordered by genomic position) spanning the *FOCAD* locus, for individual colonies (rows) from donors PD45517 and PD45534, according to donor-matched phylogenetic trees (dendrograms, left). Outlined segments correspond to the DMRs overlapping the main transcription initiation zone (IZ-1). **b**. Differentially methylated segments (top track, highlighting the DMR used to group colonies according to their DNA methylation state) and transcription initiation zones (IZ-1 to IZ-3) across the *FOCAD* locus (chr9, hg19, top strand, arrows indicate the direction of transcription). Sashimi plots show RNA-seq expression (CPM) and splice junctions for colonies grouped according to methylation state at the representative DMR, expressed as beta-value ratios (1, fully methylated; 0.5, half-methylated; 0, unmethylated), shown separately for each donor. GENCODE transcript models are shown below. This representation highlights the impact of differential DNA methylation on promoter usage in both donors. In this example, absence of methylation promotes transcription initiation at IZ-1, whereas methylation redistributes initiation events towards IZ-2 and IZ-3. **c**. Expression (CPM) across the *PLD6* locus (chr17, hg19), including the neighbouring genes *MPRIP*, *PLD6*, *FLCN* and *COPS3*, for colonies grouped according to methylation ratios at the *PLD6* promoter. *PLD6* expression is detected in unmethylated but not methylated colonies, whereas expression of the flanking genes *MPRIP*, *FLCN* and *COPS3* is unaffected by the methylation state at this locus. **d**. As in (**b**), for *TCF4* (chr18, hg19, bottom strand), shown across all three methylation ratios (0, 0.5 and 1), combining colonies from both donors. **e**. As in (**b**), for *RAF1* (chr3, hg19, bottom strand). Only two methylation groups were represented at this locus; no colonies with a fully methylated call were observed. H3K27me3 (ENCODE ENCFF195WDJ) and H3K36me3

(ENCODE ENCFF116UUS) ChIP-seq (ENCODE) (CPM) are shown above the expression tracks for chromatin-state context: H3K27me3, a repressive mark deposited by Polycomb repressive complex 2 (PRC2), is enriched broadly across the alternative initiation zones (IZ-2 and IZ-3), consistent with silencing or PRC2-mediated poising of these sites, while H3K36me3, a mark associated with active transcriptional elongation, is enriched specifically over the locus's main transcription initiation zone. **f**. As in (**e**), for *CTNNA1* (chr5, hg19, top strand). Only two methylation groups were represented at this locus; no colonies with a fully unmethylated call were observed. As for *RAF1*, H3K27me3 enrichment across the locus indicates PRC2-mediated silencing or poising of its transcription initiation zones, while H3K36me3 marks the locus's main transcription initiation zone, consistent with active transcription there. For *TCF4* (**d**), *RAF1* (**e**) and *CTNNA1* (**f**), promoter usage is quantified as the percentage of transcripts initiated at each initiation zone for each methylation state (log_{10} scale; crosses denote no detectable usage).

METHODS

Study cohort and samples. All participants provided written informed consent. Ethical approval was obtained from Cambridge research ethics committees 07/MRE05/44, 18/EE/0199, and 15/EE/0152. Samples were from six sources: (i) cord blood mononuclear cells (MNC) purchased from STEMCELL Technologies (#70007); (ii) peripheral blood or bone marrow MNC, Cambridge Blood and Stem Cell Biobank (NHS Cambridgeshire 4 Research Ethics Committee 07/MRE05/44 for samples pre-November 2019 and Cambridge East Ethics Committee 18/EE/0199 for samples post November 2019); (iii) peripheral blood or bone marrow MNCs, Cambridge Biorepository for Translational Medicine from deceased organ donors at the time of organ harvest (NHS Cambridgeshire 4 Research Ethics Committee reference 15/EE/0152), (iv) peripheral blood (PD48544) was received from the Princess Margaret Cancer Centre, University Health Network (UHN), Canada (UHN IRB protocol 01-0573) via the Weizmann Institute for Medical Science, and (v) Hematopoietic stem cells (HSCs) colonies from bone marrow (4), placenta (28), spleen (34), meninges (42), and liver (42) were received from Newcastle University (PD60968). Human fetal tissues were obtained from the MRC/Wellcome Trust-funded Human Developmental Biology Resource (HDBR; http://www.hdbr.org) with written informed consent and ethical approval from the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (23/NE/0135). The HDBR is regulated by the UK Human Tissue Authority and operates in accordance with relevant HTA Codes of Practice. (vi) The Feinstein Institutes for Medical Research provided the twin samples following local ethics approval (IRB study number: 08-202A). Patients with MPN and CML were recruited at Addenbrooke's Hospital, Cambridge (*Clonal Disorders of Haematopoiesis* study (07/MRE05/44)). Clinical information was from electronic health records at Cambridge University Hospitals NHS Foundation Trust. The cohort comprised 13 donors: seven haematologically normal individuals spanning 17 weeks post conception to 77 years, and six with a haematological malignancy (MPN or CML), including one monozygotic twin pair (PD56959 with MPN and the healthy twin PD56960). PD48544 had normal blood counts at sampling but carried a *JAK2* V617F clone detectable by whole-genome sequencing and was diagnosed with essential thrombocythaemia several years later.

Sample processing and *in vitro* colony culture. Colonies from eight donors (PD5163, PD5182, PD40521, PD45534, PD45517, PD53942, PD5847 and PD6629) were obtained from previously published studies that used either MethoCult[13,14] or single-cell fluorescence-activated cell sorting (FACS) followed by liquid culture[11]. New colonies were generated here using the same published protocols: MethoCult for PD56959 and PD56960, and liquid culture for PD48544. Briefly, mononuclear cells (MNCs) were isolated using Lymphoprep™ (STEMCELL Technologies), and $CD34^+$ HSPCs were enriched from MNCs using the EasySep™ Human Whole Blood CD34 Positive Selection Kit (STEMCELL Technologies). Germline DNA was obtained from buccal swabs or from $CD3^+$ T cells isolated from MNCs (Miltenyi Biotec), with or without in vitro expansion for up to 14 days using CD2/CD3/CD28-coated beads in RPMI medium supplemented with 10% human albumin and IL-2; genomic DNA was extracted with the QIAamp DNA Mini Kit (Qiagen). For MethoCult cultures, cells were grown in MethoCult™ H4034 Optimum for 14 days at 37 °C and 5% CO_2, and colonies were picked manually and lysed in RLT buffer (Qiagen). For liquid cultures, $Lin^-CD34^+CD38^-CD45RA^-$ cells were index-sorted by FACS from MNC or CD34-enriched samples into 96-well plates containing supplemented StemPro™ medium and cultured for 21 days. Larger colonies (≥ 3,000 cells) were split between flow-cytometric immunophenotyping and DNA extraction (DNeasy 96 Kit, Qiagen), whereas smaller colonies were processed with the Arcturus PicoPure™ DNA Kit (Thermo Fisher Scientific). For PD60968, clonal cultures were established from single haematopoietic cells, sorted on a BD FACSFusion (BD Biosciences) with a 100 µm nozzle into 96-well U-bottom plates containing 100 µl StemPro™-34 SFM supplemented with 25 ng/ml TPO, 10 ng/ml IL-3, 20 ng/ml GM-CSF, 50 ng/ml FLT3L, 50 ng/ml SCF, 10 ng/ml IL-6, 5 ng/ml IL-11 and 3 U/ml EPO (PeproTech); DNA was extracted in 20 µl Arcturus PicoPure™ extraction buffer according to the manufacturer's instructions. Single-cell RNA sequencing (scRNA-seq) data were available for a subset of colonies from two donors (PD45517, cord blood; PD45534, bone marrow), in which individual cells picked from each colony were profiled with a plate-based full-length protocol (NEBNext® Single Cell/Low Input RNA Library Prep Kit, NEB).

Parallel whole-genome and whole-methylome sequencing. Genomic DNA and whole-methylome libraries were generated from the same starting material using an optimized low-input dual WGS and enzymatic methyl-sequencing (EM-seq) workflow established at the Wellcome Sanger Institute in collaboration with New England Biolabs (NEB). Colonies were submitted either as extracted DNA or directly as cell lysate, without upfront DNA extraction. Genomic DNA was sheared by Covaris sonication, end-repaired, A-tailed, and ligated to methylated sequencing adapters. Adapter-ligated material was split 3:2, with the larger fraction proceeding to enzymatic conversion and 12 cycles of PCR to generate EM-seq libraries. The smaller fraction was amplified by standard PCR to generate matched WGS libraries. 50 pg of fully methylated plasmid pUC19 was spiked into each library to monitor conversion efficiency. Low DNA input initially caused incomplete oxidation and artefactual hypomethylation, and a 1:5 dilution of the oxidation enhancer restored pUC19 methylation to >95%. Thus the diluted enhancer was used for all samples below 20 ng input. Libraries were pooled with 10% PhiX spike-in to provide base diversity for cluster generation on the low-complexity converted genome, and sequenced on Illumina NovaSeq 6000 or HiSeqX 2500 to generate 100-150-bp paired-end reads. Sequencing quality control included assessment of mapping efficiency, duplicate rates, CpG methylation levels, and genome-wide coverage metrics. Libraries with >30% duplicate reads were excluded. EM-seq instead uses iterative TET2-mediated oxidation and APOBEC deamination to distinguish methylated from unmethylated cytosines, thus preserving DNA integrity during conversion. Direct comparison between EM-seq and low-input WGBS at 20 ng DNA input demonstrated comparable alignment rates and duplicate fractions, with EM-seq libraries covering up to 53.1 million CpGs, approaching near-complete strand-specific genome-wide representation. EM-seq libraries generated from as little as 1 ng DNA also retained broad genome-wide coverage, with up to 31.3 million CpGs covered, although duplicate rates increased with input below 10ng. Mapping efficiencies remained consistently >72–75% across low-input libraries. We adapted the workflow to operate directly from colony lysates without DNA extraction and enabled simultaneous whole-genome sequencing and methylome profiling from the same colony. WGS libraries followed standard low-input DNA protocols as previously published[68].

Read processing and alignment. FASTQ files were processed using an in-house built automated pipeline developed at the Wellcome Sanger Institute, based on a disassembled version of nf-core/methylseq 1.4[69]. Raw reads were assessed with FastQC[70], and Trim Galore[71,72] was used to remove adapter sequences and low-quality bases from read terminals. Reads were then aligned to GRCh37 using Bismark[73] with Bowtie2[74] as the underlying aligner, in paired-end, directional mode. Duplicate reads were removed using the deduplicate_bismark module. Methylation calls at individual cytosines were extracted using bismark_methylation_extractor with the flags --ignore_r2 2 --ignore_3prime_r2 2 --bedGraph --counts --gzip -p --no_overlap, which trims biased positions at read ends and avoids double-counting overlapping mate pairs. Aggregate QC was summarised with MultiQC[75]. Library-level QC metrics included mapping efficiency, duplication rate, mean CpG dyad depth, global CpG methylation percentage, and non-CpG methylation rate.

Methylation data representation. To support efficient genome-wide analysis of whole-genome methylation profiles (~28 million CpG dinucleotides per sample), we built an R-based data framework. The end-to-end workflow — from Bismark output through phasing, matrix construction and analysis — is orchestrated by *wgms_phasing* (https://gitlab.internal.sanger.ac.uk/nw14/wgms_phasing). Bismark methylation calls were converted into indexed binary matrices using the cpg_build_all function in the *cpgAccessData* R package (https://gitlab.internal.sanger.ac.uk/nw14/cpgaccessdata). These matrices store per-CpG, per-sample counts at single-strand resolution: methylated-forward (MF), methylated-reverse (MR), unmethylated-forward (UF), and unmethylated-reverse (UR), where forward and reverse refer to the original top (OT, Watson) and original bottom (OB, Crick) strands, respectively. At CpG dyad level, forward and reverse counts can be combined to give total methylated (M = MF + MR) and unmethylated (U = UF + UR) counts incorporating reads from both strands.

Genomic annotations (eg. CpG island context, relation to gene features, HMMChrom status) were added using the *annotatr* R package[76]. Overall, *cpgAccessData* represents each donor's methylation data as a CpgData object, which provides fast, memory-efficient access to the indexed binary matrices, supporting retrieval of CpG- or region-level methylation calls, generation of dyad-level matrices, and integration of annotations and sample metadata. All downstream analyses were performed in R using these processed methylation matrices.

Germline phasing and allele-specific methylation. Where a CpG dinucleotide lies near a heterozygous SNP, bisulfite sequencing reads overlapping that SNP can be assigned to the allele they carry. If the SNPs are phased into chromosome-spanning haplotypes, methylation status can be measured at haplotype resolution across extended contiguous regions. The phasing workflow, also orchestrated by *wgms_phasing*, proceeds as follows. We first called germline variants from a deeply sequenced (>15×) bulk normal WGS sample from each donor using DeepVariant 1.3.0[77] (--model_type=WGS). Additional mapped WGS samples from the same individual (e.g. colony WGS) were then used to perform read-backed phasing with the WhatsHap 2.3 phase command[78], grouping SNPs into local phase sets saved in VCF format. This VCF was then statistically phased using SHAPEIT4 (v4.2)[79] with the 1000 Genomes haplotype reference panel and the --use-PS 0.0001 flag to incorporate the read-backed phase set constraints. The output is a VCF of phased SNPs in which each heterozygous genotype is represented as 1|0 or 0|1. This method requires directional libraries in which reads are designated as original top strand (OT) or original bottom strand (OB). The Bismark BAM is split into OT-only and OB-only BAMs. The phased VCF is filtered to retain only biallelic heterozygous SNPs, excluding variants that overlap a reference CpG dinucleotide or that create a new CpG via the alternative allele, since methylation-dependent bisulfite like conversion renders such sites ambiguous. The filtered VCF is then bisulfite-recoded into strand-specific versions: for OT, both reference and alternative alleles are converted C->T; for OB, both are converted G->A. Variants where recoding collapses the reference and alternative alleles to the same base are removed, as they are uninformative in bisulfite converted space. OT reads are then haplotagged against the OT-recoded VCF, and OB reads against the OB-recoded VCF, using the WhatsHap haplotag command. This assigns each read to one of two haplotypes, conventionally labelled maternal (MT) and paternal (PT), or unphased. In the absence of duo or trio data, these labels denote arbitrary haplotype partitions (analogous to allele A/B) rather than true parental origin, though the framework could accommodate parental assignment where such data are available. Corresponding OT and OB BAMs are merged per haplotype (e.g. OT-MT + OB-MT = MT), and bismark_methylation_extractor is run on each merged BAM to produce separate methylation count files for the MT, PT, and unphased fractions. These allele-resolved counts are integrated into the CpgData representation as additional strand-level fields (MF_MT, MR_MT, UF_MT, UR_MT and equivalently for PT and unphased), with corresponding dyad-level summaries (M_MT, U_MT, M_PT, U_PT), enabling allele-specific methylation analysis within the same cpgAccessData framework used for aggregate methylation.

Rather than separately assessing the maternal (allele A) and paternal (allele B) methylation it is sometimes convenient to calculate the Allelic Methylation Skew (AMS). This measures the extent to which the phased methylation counts indicate that methylation is skewed towards A or B it is calculated as the read-weighted mean of allele A's methylation β and one-minus allele B's β (≈ mean(β_A, 1 - β_B)):

$$AMS = (M_A + U_B)/(M_A + U_A \quad + M_B + U_B)$$

Where M and U indicate methylation and unmethylation counts respectively. Values close to 1, 0.5 and 0 are respectively expected for a site exhibiting A-allele specific methylation, no allele specific methylation and B-allele specific methylation.

Sample and site quality control. Following review of the pipeline QC outputs we excluded samples with non-CpG methylation in excess of 1% and those with a positive control (puc19 spike-in) methylation of less than 90%. Having built the CpgData objects we then excluded the samples with an autosomal median coverage combined across both strands (CpG depth) < 8. We then flagged sites as failing site QC as those where more than 30% of their samples have CpG depth < 8, where aggregate depth differs from the median by a factor of 3, and where aggregate forward and reverse depth differ by a factor of 2. We also flagged CpGs that overlap SNPs. For generating colony-level distributions, we further restricted to the sites with at least 1.5 times of the cohort average depth and at most 20.

Genomic functional annotations. Chromatin state annotations were obtained from the Roadmap Epigenomics Project 15-state core-marks ChromHMM model[80], using the segmentation for E035 (primary haematopoietic stem cells) downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/Chromhmm Segmentations/ChmmModels/coreMarks/jointModel/final/E035_15_coreMarks_mnemonics.bed.gz. Each of the 28.16 million CpGs in the CpGData objects was assigned the state of the ChromHMM interval containing it. The 15 states were collapsed into six functional groups: Promoter (TssA, TssAFlnk, TssBiv, BivFlnk), Enhancer (Enh, EnhG, EnhBiv), Transcribed (Tx, TxWk, TxFlnk), Polycomb (ReprPC, ReprPCWk), Heterochromatin (Het, ZNF/Rpts) and Quiescent (Quies).

Methylation segmentation using *LAUREL*. CpGs were partitioned into contiguous co-methylation segments by multi-sample penalised segmentation, adapting the tissue-atlas approach of Loyfer et al. (2023) to our single-cell-like colony data (full description in Supplementary Note, Multi-sample segmentation of DNA methylation data). Per-CpG methylated and total read counts across all colonies (hg19, phased depth-8 QC blood data) were modelled as $M_{ik} \sim Binomial(D_{ik}, \theta_{sk})$, and the sites partitioned into S contiguous segments minimising the penalised negative log-likelihood using a dynamic-programming algorithm. We used the discrete model, in which each colony independently takes one of a small set of predefined methylation states (unmethylated/half/full, θ = {0.05, 0.50, 0.95}, adjusted per protocol group reflecting typical protocol γ and δ, with a two-state {0.05, 0.95} set for the male X non-PAR region). Segmentation was hierarchical: the whole cohort was segmented first (per-sample per-segment penalty β = 2.5) and then refined within each donor (β = 4.0, raised to limit boundary churn; monozygotic twins pooled as one donor), capturing both shared and donor-specific boundaries. Autosomes and the three X regions (PAR1, non-PAR, PAR2) were segmented separately. This resulting genome-wide segment set defines the co-methylation "segments" referred to throughout the paper.

Segment annotation. Because a segment may span more than one chromatin state (2.0% of the 6.75 million segments did so), each segment was assigned a single consensus group by tallying its CpGs by group and taking the maximum of the weighted tallies, with relative weights of 3 for Promoter and Enhancer, 1 for Transcribed, Polycomb and Heterochromatin, and 1/3 for Quiescent; weights were offset marginally from exact multiples (3.001, 0.333) so that ties resolve in favour of the functional annotation. This up-weighting prevents small promoters and enhancers embedded in large quiescent segments from being outvoted on CpG count alone.

Phylogenetic tree construction. Phylogenetic trees construction followed methods from our previously published studies: *Rtreefit* (*https://github.com/nangalialab/rtreemut*) and *Rtreemut* (*https://github.com/nangalialab/treemut*).

Assignment of methylation changes to branches using *DAPHNE*. Methylation state was inferred on each donor's colony phylogeny with a Hidden Markov tree model (**Supplementary Note 2**). The observed data are the methylation read counts at the tips of the tree while the underlying methylation state is treated as a hidden variable that descends from the zygote through the descendant lineages. Along each branch this state evolves as a continuous-time Markov chain, and its instantaneous transition rates are held piecewise-constant within molecular-time epochs, so that transitions acquired during early development (<50 mutations molecular time) occur at a different rate from those accrued later in life. The number of methylated reads $\boldsymbol{M_{ik}}$ at site k in sample i was modelled as $\boldsymbol{M_{ik} \sim Binomial(D_{ik}, \theta_{sk})}$ or, when read-level over-dispersion was estimated, as the corresponding beta-binomial with a per-sample overdispersion - conditional on the hidden methylation state, s, of the sample i, where $\boldsymbol{D_{ik}}$ is the read depth and $\boldsymbol{\theta_{sk}}$ the state-specific methylation probability. Observed reads depart from $\boldsymbol{\theta_{sk}}$ through our two error rates, γ and $\boldsymbol{\delta}$, which were estimated in a site-specific manner (**Supplementary Note 2**). These rates were trained on the healthy liquid-culture donors (PD40521, PD45517, PD45534): the per-epoch methylation transition rates were estimated jointly with the site-specific error rates and beta-binomial overdispersion, separately for each chromatin-context chunk (e.g. Promoter, Transcribed, under the Roadmap E035 grouped annotation) on segment-aggregated methylation counts.

Cohort-wide segment assignment. Holding these rates fixed, we then assigned methylation-state histories to every segment across the full donor cohort (twins pooled). For each segment the HMT was solved on the relevant donor tree by Viterbi EM, re-estimating only the per-donor error rates. This two-stage design which first estimates the site-specific error and epochal transition rates once on the healthy liquid-culture reference donors, then apply them to assign states genome-wide — decouples rate estimation from assignment, so that every donor, including shallow or low-coverage trees, is assigned under a common, stably estimated rate model. Assignments were generated per chromatin-context chunk on the segment-aggregated data and combined across chunks for downstream burden analysis.

Epimutation rate analysis. The methylation changes assigned by DAPHNE were then subject to post-hoc analysis to assess whether the rate of assigned changes themselves differed by epoch. These inferred changes (each one an event) form the basis for a negative binomial regression based estimates of epimutations per SNV:

```
MASS::glm.nb(nevents ~ 0 + timing + offset(log(branch_length_in_snv)))
```

Where nevents is the number of segment methylation changes on each qualifying phylogeny branch across PD45517, PD40521 and PD45534; timing is either “Early” (branch ends <=50 mutations) or “Late” (branch starts after 50 mutations) all other branches excluded from the analysis.

X chromosome inactivation (XCI) Timing Analysis. For each donor we filtered the X-chromosome to CpGs within 500bp of gene transcription start sites and then filtered to sites that exhibited aggregate half-methylation and then further filtered to sites where across all the donor’s colonies an average of at least 50% of reads were phased and exhibited good phase quality. The colonies were then k-means clustered into two groups based on Allelic Methylation Skew (AMS). The two groups were then designated as having XCI state A or B.

To infer the timing of XCI commitment we used an approach similar to that used in our previous study[12]. We defined three unobservable (‘hidden’) ancestral states, (**U**) Uncommitted, (**A**) Committed to Allele A and (**B**) Committed to Allele B, and used the observed outcomes XCI state (Allele A methylated/inactive or Allele B methylated/inactive) to infer parameters of a model underpinning transition probabilities U->A and U->B and then from that the most likely timing of the transition out of the Uncommitted state. The transitions between states are modelled by a discrete time Markov chain with one step in time representing one mutation in molecular time.

We require the root of the tree, presumably the zygote, to start in the **U** state and to stay in that state until t_0 mutations in molecular time. After t_0 mutations, the cell then has a non-zero probability of transitioning out of the **U** state given by the Markov transition M(k):

M(k) =

	A	B	U
A	1	0	0
B	0	1	0
U	P(k)	P(k)	1-2 P(k)

We considered a constant rate P(k) where the waiting time distribution is exponential and a more general time dependent model where the waiting time follows a gamma distribution. The probability of an event transitioning between the start of a branch at time k and ending at k' is given by matrix multiplication of the time step dependent Markov transition matrix:

$$P(S_{k'} = i \mid S_k = j) = \left(\prod_{m=k}^{k'-1} M(m) \right)_{i,j}$$

The formula for calculating the model likelihood is as given in[12] and follows the same principles as our epimutation assignment algorithm (Supplementary Note 2). The $t_0$and the parameters of the waiting time distribution are estimated at either individual donor level or cohort level (the log-likelihood being straightforwardly additive) using *rstan*. The gamma shape had a normally distributed prior centred on 1 with a standard deviation of 5; gamma mean was half-normally distributed with standard deviation 50 mutations molecular time; $t_0$had a half normally distributed prior centred on 1 mutation molecular time (standard deviation 5) and strong quadratic log-density penalty on values less than 1.

XCI allele-specific expression analysis. For the two female donors with scRNA-seq data (PD45517, PD45534), heterozygous chrX variants were taken from phased germline genotypes, so that RNA and methylation haplotypes are directly comparable. For each colony, reads of all QC-passing cells were pooled and allelic counts obtained with bcftools[81] mpileup (MAPQ ≥ 30, base quality ≥ 20). A gene's inactivated allele (Xi) expression fraction in a colony is the sum of per-variant minor-allele counts divided by its total count (≥ 20 reads), which is insensitive to phasing switch within the gene and is more conservative at escape status calling. Colonies were retained if ≥ 90% of reads over genes that are inactive in Tukiainen et al[32]. came from one allele (≥ 50 reads over ≥ 5 such genes; ≥ 100 reads over all X-linked variants), leaving 181 colonies (121 matched to a WGMS colony). Of 1,005 X-linked genes (631 with an XCI status in Tukiainen et al. and 374 further annotated genes), 710 carried a heterozygous variant in at least one donor, 344 had ≥ 20 allelic reads in at least one retained colony, and 206 were covered in ≥ 5 colonies and tested. A tested gene was called escaping if its pooled Xi fraction exceeded a background of 0.025 (one-sided binomial test used in Tukiainen et al.; Benjamini–Hochberg q < 0.01) and ≥ 3 colonies escaped individually (P < 0.01); *XIST* gene was excluded. For comparison with promoter methylation (see XCI timing analysis), a gene was considered escaping in a donor if its Xi fraction pooled over WGMS-matched colonies was ≥ 0.025. Genes were then classed as concordant escape (unmethylated, escaping), concordant inactive (methylated, silenced) or unmethylated but inactive. H3K27me3 enrichment at each gene's canonical promoter (MANE[82] Select TSS +/- 500 bp, GENCODE v48[83]) was the mean BLUEPRINT[84] ChIP-seq signal over nine tracks from healthy female donors (five classical monocyte, three neutrophil, one erythroblast), each normalised to the median of 2,000 random autosomal promoters.

Epigenetic clock analysis. To measure epigenetic age, methylation data across colonies were aggregated into a pseudobulk, either at donor level, mutant clade level or for per-donor wildtypes. For each CH clade, per-donor wildtypes and donor level aggregates, a pseudobulk CpG methylation profile was constructed by the depth-weighted aggregation of per-site methylation ratios across all tip cells assigned to that group. Wildtype colonies were defined as haematopoietic terminal branches that were not in an expanded clade without known driver mutation. For donors sampled at multiple timepoints (PD5182 at ages 32 and 53 years; PD51634 at ages 51 and 55 years), pseudobulk profiles were constructed separately per timepoint. For Epigenetic clock scoring and bootstrap confidence intervals, pseudobulk profiles were scored using the Horvath multi-tissue clock by evaluating the published linear predictor across the 353 constituent CpG sites. The resulting score estimates methylation age for each profiled set. Tip-colony resampling bootstrap (200 times) provided 95% confidence intervals on each methylation age estimate. An ordinary least-squares linear model was fitted to donor pseudobulk methylation age values as a function of chronological age. The clock residual for each clade or wildtype colonies was the difference between its observed methylation age and the model-predicted value at the relevant chronological age. Residual confidence intervals were computed by parametric bootstrap (5000 times). For each replicate, the methylation age value of each clade or wildtype group was perturbed by Gaussian noise with standard deviation σ = (methage_hi − methage_lo) / (2 × 1.96), where methage_hi and methage_lo are the upper and lower bounds of the 95% bootstrap confidence interval for that group's methylation age estimate. The donor linear model was refit and residuals recomputed for each replicate. The one-sided bootstrap p-value for each clade was the proportion of replicates in which the clade residual did not exceed the matched wildtype residual (H1: clade > wildtype). All analyses used R version 4.4.0. Per-clade quantification of proportion of C>T somatic substitutions at CpG dinucleotides was obtained from the per-colony somatic mutation catalogue and aggregated per clade as the mean of the per-colony values per timepoint of sampling. Wildtype data were computed per donor per timepoint. To estimate the uncertainty in the mean arising from cross-colony variability, we generated a nonparametric bootstrap 95% confidence interval: colonies within a clade were resampled with replacement (n resamples, matching the original colony count) and the mean propCtAtCpG recalculated for each resample, repeated for 200 bootstrap iterations. The 2.5th and 97.5th percentiles of the resulting distribution of bootstrap means defined the lower and upper bounds of the 95% confidence interval. Clades with fewer than two colonies were excluded from this calculation.

Identification of expansion-associated methylation changes (Trunk-DMRs). For each of the 34 CH clades with at least 4 tip-level samples involved across 13 donors, all autosomal CpG segments were tested for clade-specific differential methylation relative to the matched wildtype (WT) pool. For each (clade, segment) pair, the segment was first required an HMT-assigned state transition on the trunk branch leading to the clade; and this new state needs to be maintained in at least 70% of descendant branches, meaning that the fraction of all edges within the clade subtree (trunk branch, internal branches, and tip branches) that carry the same HMT-inferred methylation state as the trunk branch should be equal to or greater less 0.7. Segments passing this filter were tested using a two-sided Wilcoxon rank-sum test comparing per-CpG beta values across clade versus WT samples. Per-clade Bonferroni correction was applied with a threshold of 0.05/n, where n is the number of segments tested within that clade. The Bonferroni-significant segments were defined as Trunk-DMRs. Among all 34 clades, 301,747 unique trunk-DMRs (5,149 were big segments; n CpGs ≥5) in 330,540 segment × clade pairs were identified, varying between 7 and 120,861 across 28 clades).

Enrichment of genomic annotations among DMRs. For each genomic annotation, enrichment was quantified for the 5,149 trunk DMRs (segments of >4 CpGs significant in at least one clade, pooled as unique segments across clades) against the genome-wide background of large segments, as a 2x2 odds ratio against the background. As DMR status of a segment is spatially autocorrelated along the genome (the autocorrelation of DMR status of segments decays below 0.05 at a lag of 33 segments), significance of enrichments was assessed against an

empirical null distribution of blocks of DMRs: segments were ordered genomically, grouped into contiguous 33-segment blocks (excluding 393 chromosome-end segments that did not fill a block), and the DMR status of all blocks shuffled 10,000 times within-chromosome to generate a null distribution of enrichments if DMRs were randomly assigned. Two-sided empirical p-values were taken as the proportion of permutations giving an odds ratio at least as extreme as the observed one and Benjamini–Hochberg corrected across the levels of each annotation; 95% confidence intervals were derived from the permutation s.d. of log(odds ratio).

Driver clade-associated methylation differences between carrier and wild-type samples. We tested each segment with >4 CpGs for methylation differences between driver clades and the other colonies of the same donor, using a phylogenetic linear regression (*phylolm*[85]). This accounts for the shared ancestry of colonies, since residuals are correlated according to shared branch length in the donor's phylogeny. For each colony, methylated (M) and total (D) read counts were summed over the segment's CpGs. The response was the empirical logit, logit((M + 0.5)/(D + 1)), passed through a smooth bounding function that maps it into (logit 0.05, logit 0.95), so that differences between near-fully methylated or unmethylated colonies are not inflated. Covariates include colony cell type (omitted when all colonies included for the segment shared the same cell type), SNV burden, one indicator per driver-acquisition node with ≥ 3 descendant colonies, so that nested clades were adjusted for each other. Colonies with a cell type annotation were included for a segment if they had ≥ 10 reads over it, and a segment was fitted only if ≥ 50% of colonies were included; the phylogeny was pruned to the included colonies. Effects were reported for clades defined by the acquisition of a curated driver (*DNMT3A*, *JAK2*, *TET2*, *EZH2*, *MPL*, *NF1*, *BCR::ABL1*, 9pUPD or 1q+) with ≥ 3 carrier and ≥ 3 non-carrier colonies. Each such effect is the clade coefficient on the bounded logit scale, with its two-sided P value. Donors without such a clade were not analysed.

Early CH DMRs. All methylation segments genome-wide carrying at least five CpGs (n = 1,352,700) were tested. A segment was defined as *early* in a donor where the phylogeny timed its methylation change before 50 somatic mutations of molecular time; 28,168 segments were early in at least one donor. Per-clade CH-versus-normal differential methylation was called with *phylolm*, and as DMR calling was performed per-clade, p-values were Benjamini–Hochberg adjusted within clade, and a segment was defined as a DMR for a clade where $q < 0.05$ and where the absolute difference in median methylation between CH and background branches was at least 0.05. A segment was an early CH DMR where both held in the same donor, that is, it was early in that donor and a DMR in one of that donor's clades. Restricted to donors with both, this resulted in 22,170 early segments carrying 4,297 early CH DMRs across 7 clades from 5 donors. As DMRs clustered within the same spatial neighbourhood, adjacent segments were chained together into larger multi-segment loci where the distance between the end of one and the start of another was less than 3kb. Loci were labelled with the genes their segments overlap or, for segments in distal enhancers, the genes those enhancers target. Loci containing early CH DMRs in more than one donor were retained (129 loci) and ranked by number of donors, then by the mean absolute methylation difference of their component Early DMRs. These were subsequently split by direction separating loci made up entirely of hypermethylated Early DMRs, hypomethylated, and into those discordant between donors.

Promoter methylation and transcription in matched colonies. Methylation and gene expression were compared in the colonies of donor PD45517 (healthy cord blood) that have both WGMS and scRNA-seq data. For each gene, the canonical promoter was defined as +/- 500 bp around the TSS of its MANE[82] Select transcript (GENCODE v48[83]), lifted over to hg19. Promoter methylation (beta value) in a colony was the fraction of methylated reads pooled over all CpGs in this window (both strands, ≥ 10 reads). Promoters with a mean beta across colonies < 0.2 were classified as unmethylated and > 0.8 as methylated. Gene expression in each colony was quantified as a monocyte pseudobulk: counts were summed over the colony's QC-passing monocytes and transformed to log1p(counts per million [CPM]). A gene was considered expressed if log1p(CPM) > 1 in at least one colony, and silent otherwise. Each promoter was assigned the majority Roadmap 15-state ChromHMM state (E035) of its CpGs. H3K27me3

enrichment at promoters was computed as described above for X-linked genes, using the five classical-monocyte BLUEPRINT tracks.

Promoter usage. Gene and transcript annotations were obtained from GENCODE release 50 and mapped to GRCh37/hg19 using the GRCh37 back-mapping pipeline (gencode-backmap; GTF and GFF3, GRCh37_mapping track). Transcript TSSs were identified in a strand-aware manner and unique TSSs consolidated per gene. TSSs on the same strand separated by <100 bp were merged into consensus transcription initiation zones (IZs). NEB scRNA-seq BAM files from individual clonal colonies were grouped according to methylation state based on previously identified DMRs and merged using `SAMtools`[86]. To identify the initiation zones used under each condition, we examined the 5′ ends of mature transcripts represented by sequencing reads supporting splice junctions. Splice junctions were identified using `regtools`[87] `junctions extract` (minimum anchor length, 8 bp; intron length, 50-500,000 bp; strand inferred from XS tags), and CPM-normalised coverage tracks were generated using `deepTools`[88] `bamCoverage`. 5′-proximal splice junctions were assigned to the nearest IZ using a strand-aware window extending 100 bp upstream and 1,000 bp downstream of the IZ boundary. When multiple IZs fell within the search window, the nearest IZ was selected. Promoter usage for each IZ was calculated as the proportion of junction-supported reads assigned to that IZ relative to the total number of junction-supported reads assigned to all IZs within the corresponding gene.

Fluctuating CpG sites (fCpGs). Two fCpG panels were used: a colon/intestine panel[37] and a blood panel[8], taken from supplementary tables of the respective publications. The colon panel lists Illumina probe IDs, which were mapped to GRCh37/hg19 using the Illumina EPIC annotation (Bioconductor IlluminaHumanMethylationEPICanno.ilm10b4.hg19). The blood panel provides hg19 coordinates. The panels comprise 1,794 and 978 autosomal CpGs (7 shared; 2,765 in total), all of which are covered in our data. At segment level, segments containing at least one published fCpG were compared with all other segments. For each segment we counted the methylation changes inferred by the segment *DAPHNE* on internal phylogenetic branches, summed across donors. To exclude the possibility that the early enrichment depends on the time-structured segment model (different rates before and after 50 SNVs), individual CpGs were refitted with a *DAPHNE* setting in which one time-constant transition-rate matrix was shared by all branches, with site-specific error rates. State changes were the most likely (Viterbi) assignments under this rate matrix. For each donor, the published fCpGs were modelled jointly with a background of CpGs drawn at random (fixed seed) from all other CpGs in our data, with the same composition of Roadmap 15-state ChromHMM state (E035) and CpG context. For each set, the cumulative number of methylation changes from the phylogeny root to each internal branch was divided by the number of CpGs in the set, and related to the branch's cumulative SNV count. As a proxy for cell division, internal branches were annotated with the proportion of their SNVs that are C>T at CpG. To stabilise this proportion, branches of each donor were pooled in order of their depth (separately within and beyond the first 50 SNVs) into bins of at least 40 SNVs.

Acknowledgements and Funding. We thank the Cambridge Blood and Stem Cell Biobank for their invaluable support with sample collection and colony growth. We thank the Somatic Genomics Programme, IT team and Sanger core facility for their support with next generation sequencing, and New England Biolabs for research and development together with Wellcome Sanger. The Nangalia lab is supported by CRUK, Wellcome core funding at Wellcome Sanger Institute, core funding at Wellcome-MRC Stem cell Institute, Alborada Trust, Blood Cancer United, Harrison foundation, Blood Cancer UK and Rosetrees Trust. JN is supported by CRUK. The Laurenti lab is supported by the British Council, core support grants by Wellcome and Medical Research Council (MRC) to the Wellcome-MRC Cambridge Stem Cell Institute, Alborada, Harrison foundation, Blood Cancer United and Blood Cancer UK. AV was a Gates Cambirdge Trust Scholar. We thank the patients for participation in this study. The

authors used Claude (Opus and Fable models, Anthropic) to assist with coding. All analysis scripts were carefully reviewed, verified, and edited by the authors. This research was funded in part by the Wellcome Trust (107630/Z/15/Z). For the purpose of Open Access, the author will apply a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We thank Dr Liran Shlush for identification and sourcing of sample PD48544.

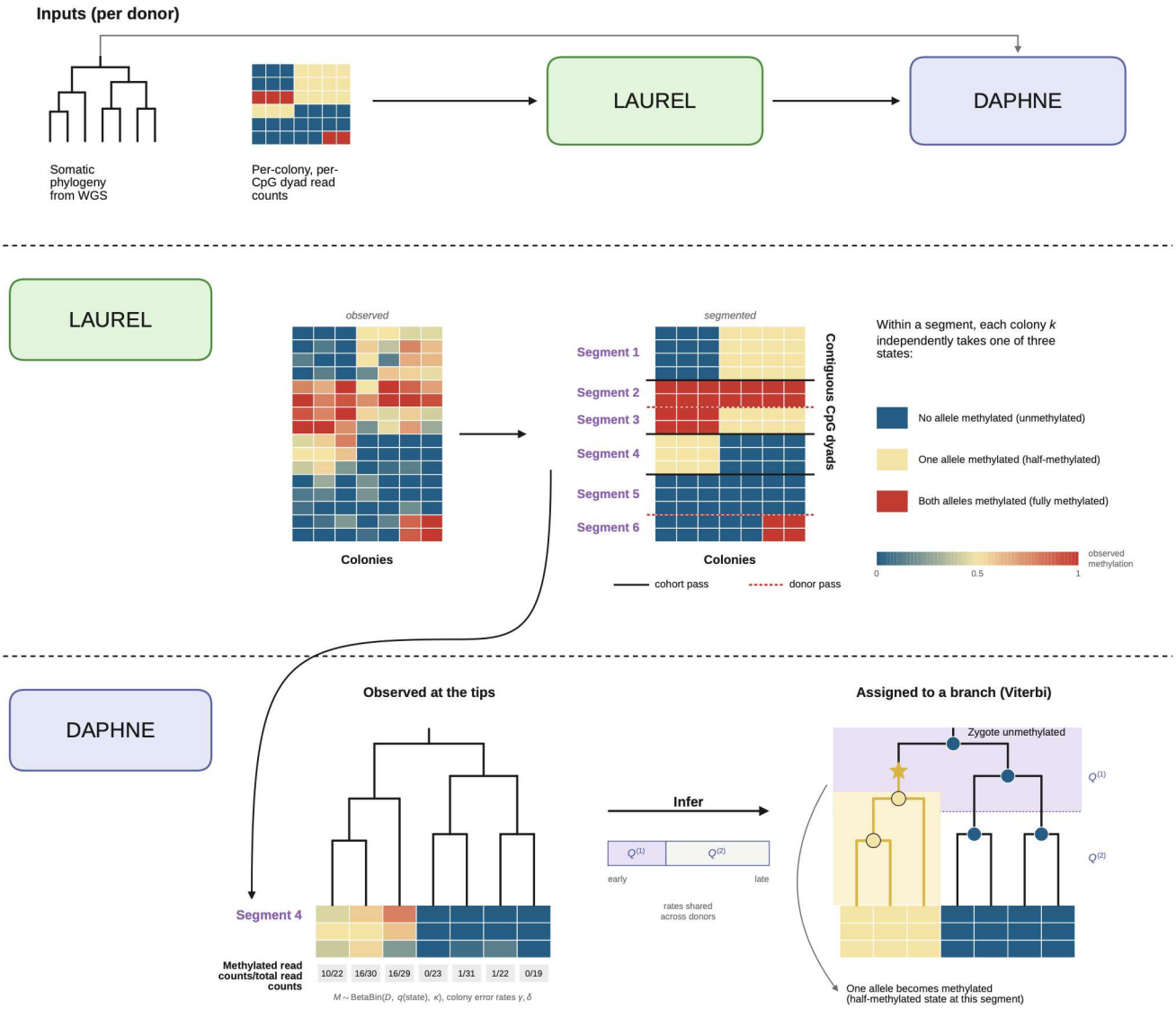
Authorship contributions. XW and JG performed analyses together with NW, following on from previous analyses undertaken by JL. JL developed the dual WGS and EMSeq pipeline together with MQ, MM and New England Biolabs, and discovered embryonic epialleles. JL and NW developed *cpgaccessdata* including phasing together with KM. JL and NW developed *Daphne* and performed analyses. NW developed *Laurel.* SL performed enrichment analysis and developed web visualisation. LK supported bench marking of *Daphne* together with NW and JG. PM analysed gene expression together with XW and KT. XL performed ICR analysis together with NW. KD analysed epimutation rates. EJB directed colony growth. NMS and MM supported NGS library preparation. AV provided samples for PD48544. PM together with CP and MSC provided PD60968. AK provided PD51634. EM provided healthy donor samples under EL. SA provided twin samples. JN wrote the manuscript with support from XW, JG, SL, XL, NW. PJC and MC provided scientific discussion. NW supported all analyses. All authors reviewed the manuscript. JN conceived the study and directed the research.

Conflicts of Interest. The authors have no competing interests to declare. PJC is a cofounder and employee of Quotient Therapeutics. JL is an employee of GSK.

Data availability and code. Code for analyses will be made available and deposited at https://github.com/nangalialab/Methylation. Sequencing files will be uploaded to EGA in line with Wellcome Sanger Institute data sharing.

SUPPLEMENTARY NOTE:

1. Schema for assigning methylation changes to somatic mutation phylogenies.



Overview of *LAUREL* and *DAPHNE*. The inputs for each donor are a WGS-based phylogeny and per-colony methylation read counts at every CpG dyad. *LAUREL* partitions the CpGs into contiguous groups ("segments") within which each colony shares one of three latent methylation states (unmethylated, half-methylated or fully methylated). It first segments at the cohort level (solid lines), then refines these into potentially smaller segments with a restricted donor-level segmentation designed to identify donor-specific changes relative to the background (dashed lines). Next, DAPHNE analyses each segment from its methylation counts aggregated per colony, modelled as beta-binomial given the colony's state and error rates. This state is treated as hidden on the phylogeny and evolves from zygote to tip as a two-epoch continuous-time Markov process (intensity matrices: early $Q^{(1)}$, late $Q^{(2)}$) with rates shared across donors. For each donor and segment, the maximum a posteriori (Viterbi) path is determined, as well as the posterior probability of each transition on each branch.

2: DAPHNE (Discrete **A**ncestral **P**hylogenetic **H**idden markov model of the **N**uclear **E**pigenome)

The state space. We consider a number of state spaces for the CpG dyad in a single cell. The most general set of states separately considers the methylation state of each allele, and the state of both forward and reverse strand cytosines in the dyad. For each allele we have 4 possible states of the two cytosines (00,01,10,11). Combining the two alleles we have 16 possible states e.g. for state "00|01" the first allele has both cytosines unmethylated (00) whereas the second allele has the forward strand cytosine unmethylated and reverse strand cytosine methylated (01). This state space can be simplified, firstly by assuming no hemi-methylation forcing the forward and reverse strand to have the same methylation state — this gives 4 possible states: 00|00, 00|11, 11|00 and 11|11. In the absence of phasing information we can further collapse the two allele specific states 00|11 and 11|00 into a single half methylated state. This gives a 3 state model $g = 0,1$ or 2 where g is the number of methylated alleles. In this study we use the 3 state model, but also do have read phase information that is incorporated as well. Note that by convention the states are always ordered by number of methylated alleles, so states $s = 1$ and $s = S$ are unmethylated and fully methylated respectively.

Modelling the ancestral unobserved methylation states with a hidden Markov tree. We defined S unobserved ("hidden") ancestral methylation states belonging to the state space of interest (as discussed above). For each individual we use the known phylogenetic tree in conjunction with the observed data (methylation counts and unmethylation counts) to jointly infer the global transition rates between these states and the most likely sequence of methylation states at ancestral nodes. The transitions between states are modeled as following a continuous time Markov chain model (see e.g. Moran 1984) with an intensity matrix Q that may vary across epochs of life (see the multi-epoch rate model below). Within each epoch the intensity matrix is constant. The intensity matrix is a square matrix where the off diagonal elements, $Q_{i,j}$ correspond to the rate at which a CpG in state i transitions to j i.e. $p(S(t+\delta t) = j \mid S(t) = i) \approx Q_{i,j}\delta t$. The diagonal elements $Q_{i,i}$ give the rate of exit from the specified state i.In the single-epoch case, the matrix of transition probabilities $P_{i,k}(u)$ for branch u, having duration t_u starting in state i and ending in state k is given by the solution to the Kolmogorov equations: $P(u) = exp\{t_u Q\}$. The multi-epoch generalisation is described below. Now conditional on the state of a node, the data in its subtree are independent of the rest of the tree. This conditional independence property underpins all three recursive algorithms presented below: the upward pass for the full data likelihood, the Viterbi algorithm for the most likely sequence of ancestral states, and the upward-downward algorithm for posterior marginal state probabilities at each node and posterior transition probabilities on each branch. The algorithm used is essentially an inhomogeneous special case of the approach detailed in Durand, Goncalves, and Guédon (2004). Also see Kapadia et al. (2025) for an implementation in a related context where the underlying transition probabilities were modeled using a molecular time based discrete Markov chain model and the emitted data was an observed cell type.

Upward algorithm for determining likelihood of the observed states given M and a prior probability of root state We denote by $L_u(i) \equiv P_u(D_u \mid s_u = i)$ the partial likelihood of all data in the subtree rooted at u, conditional on the end-of-branch state of u being i. Here S is the number of hidden states ($S = 3$ in our case) and D_u denotes the observed data descendant of u i.e. in our case the observed methylation counts of samples that descend from u. For tip nodes, corresponding to sequenced samples: $L_u(i) = P(M_u, U_u \mid s_u = i)$

For internal nodes: $L_u(i) = \prod_{v \in \text{children}(u)} \quad C_v(i)$ where we define the child contribution of branch v given parent state i as: $C_v(i) = \sum_{k=1}^{S} \quad P_{i,k}(v)\, L_v(k)$ Note that $C_v(i)$ is computed as part of the upward pass and will be reused in the downward pass for posterior computation. For the root, define the joint quantity $L^*_{root}(i) = \pi_i\, L_{root}(i)$. The full likelihood is then given by $L = \sum_{i=1}^{S} \quad L^*_{root}(i)$

A Viterbi-like algorithm to determine the most likely sequence of hidden end-of-branch states. This algorithm mirrors the upward pass but replaces summation over child states with maximisation, yielding the likelihood of the most likely (rather than the total) sequence of descendant states. Following Durand, Goncalves, and Guédon (2004) we denote this quantity $V_u(i)$ (δ in their notation). At tip nodes $V_u(i) = L_u(i)$, the emission likelihood

defined above. For internal nodes: $V_u(i) = \prod_{v \in \text{children}(u)} \left(\max_k \{V_v(k)\, P_{i,k}(v)\} \right)$ Additionally, for each node we store the most probable child states given that u is in state i: $\Psi_{u,v}(i) = \arg\max_k \{V_v(k)\, P_{i,k}(v)\}$ The above provides a recipe for recursively finding $V_{root}(i)$ and is combined with prior root probability π to give the most likely root state $\hat{s}_{root} = \arg max_k \{\pi_k\, V_{root}(k)\}$. The likelihood of the most likely sequence of states is $max_k \{\pi_k\, V_{root}(k)\}$. The most likely sequence of hidden end-of-branch states are then directly populated from Ψ.

Posterior marginals: per-node state probabilities and per-branch transition probabilities. To obtain posterior marginal probabilities at each node, and posterior transition probabilities on each branch, we supplement the upward pass with a downward pass. This is the tree analogue of the forward-backward algorithm for HMMs on chains. The upward pass computes the partial likelihoods $L_u(i)$ and child contributions $C_v(i)$ as defined in the Upward algorithm section above.

Downward (top-down) partial likelihoods. The downward pass computes, for each node u, the partial likelihood of all data *outside* the subtree rooted at u, conditional on the end-of-branch state of u. We denote this $R_u(i)$. **Root initialisation:** $R_{root}(i) = \pi_i$. **Recursion.** For an internal node u with children $v_1, v_2, \ldots, v_C$, we can express the partial likelihood of its child v_j as: $R_{v_j}(k) = \sum_{i=1}^{S} \; P_{i,k}(v_j)\, R_u(i) \prod_{l \neq j} \; \left(\sum_{k'=1}^{S} \; P_{i,k'}(v_l)\, L_{v_l}(k') \right)$

This expression has a natural interpretation: to compute the outside information for child v_j in state k, we sum over all possible parent states i, weighted by (a) the transition probability from i to k on branch v_j, (b) the outside information for u being in state i, and (c) the contributions from all sibling subtrees of v_j given parent state i. Using the child contribution $C_v(i)$ defined in the upward pass, this can be written more compactly as: $R_{v_j}(k) = \sum_{i=1}^{S} \; P_{i,k}(v_j)\, R_u(i) \prod_{l \neq j} \; C_{v_l}(i)$

Posterior node marginals. The posterior probability that node u is in state i, given all observed data D, is: $P(s_u = i \mid D) = \frac{R_u(i)\, L_u(i)}{\sum_{j=1}^{S} \; R_u(j)\, L_u(j)}$

Posterior branch transition probabilities. The posterior probability that branch v (connecting parent u to child v) has a transition from state i to state k, given all observed data, is: $P(s_u = i, s_v = k \mid D) = \frac{R_u(i)\, P_{i,k}(v)\, L_v(k) \prod_{l \neq v} \; C_{v_l}(i)}{L}$ where the sibling product is as defined in the downward recursion and L is the total data likelihood. The conditional transition probability on branch v is then:

$P(s_v = k \mid s_u = i, D) = \frac{P(s_u = i, s_v = k \mid D)}{P(s_u = i \mid D)}$ And the marginal probability of a *change* on branch v (i.e. the branch carries at least one transition) is: $P(change\ on\ v \mid D) = 1 - \sum_{i=1}^{S} \; P(s_u = i, s_v = i \mid D)$ This quantity gives the per-branch, per-site posterior probability that a methylation state change occurred, which can be aggregated across sites to estimate per-branch epimutation burden.

The emission probabilities, or the probability of observed read counts conditional on methylation state

Let us define two error rates:

- $\delta = p(\text{called methylated} \mid \text{original molecule unmethylated})$
- $\gamma = p(\text{called unmethylated} \mid \text{original molecule methylated})$

So γ reflects in-vitro demethylation and sequencing error, δ is a joint measurement of 1-conversion rate, sequencing error and in-vitro differentiation driven methylation.

Let $q(g) = p(\text{methylated} \mid g) \cdot (1-\gamma) + p(\text{unmethylated} \mid g) \cdot \delta$ giving:

$$q(0) = \delta, \qquad q(1) = \frac{1+\delta-\gamma}{2}, \qquad q(2) = 1-\gamma$$
$$P_u(M_u, U_u \mid s_u = i) = \binom{M_u + U_u}{M_u} q(g_i)^{M_u} \big(1 - q(g_i)\big)^{U_u}$$

The above can be generalised to incorporate reads with known phase as well as unphased reads. Let $M_u^{(k)}$ denote the methylation count attributable to reads of class k where $k \in \{Allele\ A, Allele\ B, Unphased\}$. Then the 4-state model (first number is Allele A state; second is Allele B state) 00, 01, 10, 11 has the following emission probabilities:

$$q^{(4)}(Allele\ A) = (\delta,\ \delta,\ 1-\gamma,\ 1-\gamma)$$
$$q^{(4)}(Allele\ B) = (\delta,\ 1-\gamma,\ \delta,\ 1-\gamma)$$
$$q^{(4)}(Unphased) = \left(\delta,\ \frac{1+\delta-\gamma}{2},\ \frac{1+\delta-\gamma}{2},\ 1-\gamma\right)$$
$$P_u\left(M_u, U_u \mid s_u^{(4)} = i\right) = \prod_{k \in \{A,B,U\}} \binom{M_u^{(k)} + U_u^{(k)}}{M_u^{(k)}} q^{(4)}(k)_i^{M_u^{(k)}} \left(1 - q^{(4)}(k)_i\right)^{U_u^{(k)}}$$

This can then be marginalised to a 3-state model by combining the two allele-specific terms: where we have introduced superscripts in brackets to indicate the relevant state space.

Estimation of Q, π, γ and δ. The genome is partitioned into 6 sets of CpGs, each representing a functional grouping of tissue-specific core 15-state Roadmap annotations; we denote such a set by B. Let B have n CpGs and let the phylogenetic tree have m branches and p tips each corresponding to sequenced single cell-like samples. We denote the temporal duration of the ith branch as t_i and its unique parent node as $parent_i$. The methylation state inferred by the HMT algorithm can be summarised as an $(m+1) \times n$ state matrix S. We index the rows of S from $0..m$ where row 0 corresponds to the root (zygote) and row i corresponds to the ith branch. $S = HMT(Q, \pi, \gamma, \delta; data)$ where γ and δ are $p \times n$ matrices.

Multi-epoch rate model. We allow the transition rate matrix Q to vary across life. Time is partitioned into E non-overlapping epochs $[T_0, T_1), [T_1, T_2), \ldots, [T_{E-1}, T_E)$ where $T_0 = 0$ (conception) and $T_E = \infty$. In the two-epoch model used in this study the boundary is set at $T_1 = 50$ somatic mutations from the root, roughly corresponding to the first year of life; this captures the well-established difference in early-life methylation remodelling rates relative to later life. Each epoch e has its own intensity matrix $Q^{(e)}$. The transition probability matrix for a branch that begins at time t_0 and ends at t_1 (measured in somatic mutations from the root) is the ordered product of the per-epoch matrix exponentials:

$$P(t_0, t_1) = \prod_{e:[T_{e-1}, T_e) \cap [t_0, t_1) \neq \emptyset} exp\left\{Q^{(e)} \cdot \big(min(T_e, t_1) - max(T_{e-1}, t_0)\big)\right\}$$

so that a branch spanning multiple epochs accumulates the appropriate contribution from each.

Iterative estimation procedure. The parameters $Q^{(1)}, \ldots, Q^{(E)}$, π, γ, δ and (optionally) κ are estimated by an iterative scheme combining Viterbi assignment, error rate estimation and rate fitting. We describe the full procedure with the empirical Bayes site-specific error model and posterior-based rate estimation; the simpler variants are special cases.

Initialisation:

$$Q_{i,j}^{(e,0)}(B) = 10^{-5}\ per\ mutation, \quad \forall\, i \neq j, \quad e = 1, \ldots, E$$

$$\pi_i^{(0)} = \frac{1}{S}, \quad i = 1, \ldots, S$$

$$\gamma_{site,sample}^{(0)} = 0.2, \qquad \delta_{site,sample}^{(0)} = 0.05$$

Outer loop ($K = 5$ rate iterations, indexed by k):

Phase 1 — Converge error rates per donor (inner loop). For each donor d we run G iterations (typically $G = 10$) of the following sub-loop:

(i) Viterbi assignment with current rates and error parameters to obtain genotypes $S^{(k)}$.

(ii) Two-step empirical Bayes estimation of γ and δ (described below).

(iii) Per-sample estimation of the Beta-Binomial concentration parameter κ (described below).

Phase 2 — Posterior marginals. With converged error parameters, compute the full upward-downward posterior for each donor. This yields the per-branch joint posterior $P(s_u = i, s_v = k \mid D)$ for all branches v and all sites in the bin.

Phase 3 — Multi-epoch rate estimation from posterior expected counts. For each transition type $i \to j$ and each branch b, we extract the posterior expected number of $i \to j$ transitions across all sites:

$$n_{b,i\to j} = \sum_{l=1}^{n} P\left(s_{parent_b} = i, s_b = j \mid D_l\right)$$

and the posterior expected number of sites with parent state i:

$$n_{b,i} = \sum_{j'} n_{b,i\to j'}$$

The per-branch, per-transition-type rate is $r_{b,i\to j} = n_{b,i\to j}/n_{b,i}$. To decompose this into epoch-specific rates we note that for each branch b the time spent in epoch e is

$$\tau_{b,e} = max\left(0, min\left(T_e, t_{1,b}\right) - max\left(T_{e-1}, t_{0,b}\right)\right)$$

where $[t_{0,b}, t_{1,b})$ is the temporal span of branch b. For each transition type $i \to j$ separately, we solve for epoch-specific rates $q_{i\to j}^{(e)}$ by fitting a non-negative linear model (via with , no intercept, and pseudocount regularisation):

$$r_{b,i\to j} \approx \sum_{e=1}^{E} q_{i\to j}^{(e)}\, \tau_{b,e}$$

pooling branches across all donors in the cohort. The fitted $q_{i\to j}^{(e)}$ values populate $Q^{(e)}$ for the next outer iteration.

Phase 4 — Update π. The root prior is set to the empirical frequency of root states:

$$\pi_i^{(k)} = \frac{1 + \sum_{l=1}^{n} 1\left[S_{0,l} = i\right]}{S + n}$$

Two step Empirical Bayes estimation of site-specific error rates. Error rates (γ and δ) were first estimated for each colony from its reads at sites assigned to the fully methylated and unmethylated states, respectively. We

then allowed each site to depart from the colony level rates, to capture local effects such as sequence context. Site-specific deviations on the logit scale were estimated using an empirical Bayes shrinkage procedure. It approximates ridge regression in the additive logistic model logit(γ) ~ colony + site, with the penalty applied to the site effects only and set from the data rather than by cross-validation (Efron and Morris 1973; Morris 1983). The between-site variance was estimated from well-covered sites (pooled depth >1,000 reads). Each site's deviation was then shrunk towards zero according to how much information that site carried i.e. its pooled read depth across colonies assigned to the relevant state. Thus, deeply covered sites with many contributing colonies retain their own rate, whilst sparse sites fall back towards the colony level rates. Where no between-site variation was detectable, the model reduced to colony-specific rates alone.

Beta-Binomial emission model and κ estimation. The standard binomial emission model can understate uncertainty at sites with overdispersed methylation counts. To handle this we optionally replace it with a Beta-Binomial model where the methylation fraction θ at a CpG in state g is drawn from a Beta distribution centred on $q(g)$ with concentration parameter κ, and the read counts are then binomial given θ:

$\theta \mid g \sim Beta\left(\kappa\, q(g),\ \kappa\big(1-q(g)\big)\right), \qquad M \mid D, \theta \sim Binomial(D, \theta)$ As $\kappa \to \infty$ this converges to the standard binomial emission. Separate concentration parameters κ_0 and κ_2 are estimated per sample by maximum likelihood, using consensus sites (CpGs where $\geq 80\%$ of samples agree on state 0 or state S respectively). The heterozygous state uses $\kappa_1 = \sqrt{\kappa_0 \kappa_2}$ as a geometric mean interpolation. These per-sample κ values are re-estimated at each inner (γ/δ convergence) iteration alongside γ and δ.

References

Durand, J-B, Paulo Goncalves, and Yann Guédon. 2004. "Computational Methods for Hidden Markov Tree Models-an Application to Wavelet Trees." IEEE Transactions on Signal Processing 52 (9): 2551–60.

Efron, Bradley, and Carl Morris. 1973. "Stein's Estimation Rule and Its Competitors - an Empirical Bayes Approach." Journal of the American Statistical Association 68 (341): 117–30.

Kapadia, Chiraag D, Nicholas Williams, Kevin J Dawson, Caroline Watson, Matthew J Yousefzadeh, Duy Le, Kudzai Nyamondo, et al. 2025. "Clonal Dynamics and Somatic Evolution of Haematopoiesis in Mouse." Nature, 1–9.

Moran, P. A. P. 1984. An Introduction to Probability Theory. Oxford Science Publications. Clarendon Press. https://books.google.co.uk/books?id=UPP90bGx_b4C.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." Journal of the American Statistical Association 78 (381): 47–55.

3: LAUREL (Local Assignment of Uniform Regions in Epigenetic Landscapes)

Introduction. In (Loyfer et al. 2023) they present an Atlas of CpG segments that represent contiguous groups of CpGs that with-in samples (or tissues) have a constant latent underlying methylation, but which can differ between samples (or tissues). Inspired by this we developed a similar approach that is appropriate for our single cell like data. The method jointly segments multiple samples efficiently using penalised dynamic programming. The method identifies an optimal partition of the CpGs into contiguous segments, assigning each sample a latent methylation state (or level) within each segment. The approach supports both single-cell like data via a discrete model, in which each sample is restricted to one of a small number of predefined methylation states, and a continuous model, in which per-sample methylation levels are estimated freely by maximum likelihood. Segmentation is performed using the Pruned Exact Linear Time (PELT) algorithm, which provides exact solutions with computational cost that scales near-linearly with the number of CpG sites. We employ this approach in a hierarchical workflow that first segments the full cohort, then refines within individual donors, enabling detection of both shared and donor-specific methylation changes.

Statistical model. Let M_{ik} and D_{ik} denote the methylated read count and total coverage at CpG site i in sample k, for $i = 1, \dots, m$ sites and $k = 1, \dots, n$ samples. We model $M_{ik} \sim Binomial(D_{ik},\ \theta_{sk})$ where s is the segment containing site i and θ_{sk} is the methylation probability for sample k within segment s. We seek a partition of the m sites into S contiguous segments that minimises the penalised negative log-likelihood$L = \sum_{s=1}^{S} \quad C(s) \ + \ \beta \cdot S$ where $C(s)$ is the total negative log-likelihood cost for segment s and β is a penalty per segment that controls the trade-off between fit and model complexity.

Discrete model. Each sample independently selects the best-fitting state from a predefined set Θ_k, typically $\{0.05,0.50,0.95\}$ representing unmethylated, half-methylated, and fully methylated. The segment cost is $C(s) = -\sum_{k=1}^{n} \quad \max_{\theta \in \Theta_k} l_{sk}(\theta)$ where $l_{sk}(\theta) = \widetilde{M}_{sk} log\theta + \left(\widetilde{D}_{sk} - \widetilde{M}_{sk}\right) log(1-\theta)$ is the binomial log-likelihood, with $\widetilde{M}_{sk}$ and $\widetilde{D}_{sk}$ the cumulative methylated and total counts over the segment. No parameter estimation is required — the maximisation is a lookup over $|\Theta_k|$ values. The state set can vary across samples, which is useful when males and females require different state spaces on the X chromosome.

Continuous model. The methylation level for each sample is estimated by maximum likelihood within each segment: $\hat{\theta}^{*}_{sk} = \frac{\widetilde{M}_{sk}}{\widetilde{D}_{sk}}$ The segment cost is the log-likelihood evaluated at $\hat{\theta}_{sk}$. The primary regularisation in the continuous model is provided by the penalty β, which prevents over-segmentation by requiring that each additional segment boundary yields a sufficient improvement in log-likelihood. A small pseudocount correction is applied for numerical stability at extreme methylation levels,[1] but plays no role in controlling segmentation complexity.

Optimisation by PELT. Both models are optimised exactly using the PELT algorithm (Killick, Fearnhead, and Eckley 2012). PELT exploits the fact that the segmentation problem satisfies an optimal substructure: if $F(t)$ is the minimum penalised cost for sites $1, \dots, t$, then $F(t) = \min_{\tau \in R_t}[F(\tau) + C(\tau + 1, t) + \beta]$ where R_t is the set of candidate changepoints. Without pruning, $R_t = \{0,1, \dots, t-1\}$ and the algorithm is $O(m^2 n)$. PELT reduces this by discarding candidates τ for which $F(\tau) + C(\tau + 1, t)$ already exceeds $F(t)$, since such candidates cannot improve on the current optimum at any future position. Under the assumption that the number of true changepoints grows linearly with m, the expected cost is $O(m \cdot n \cdot |\Theta|)$ for the discrete model and $O(m \cdot n)$ for the continuous model. Cumulative sums of methylation and depth counts are computed once per sample, allowing the cost of any segment $(a, b]$ to be evaluated in $O(n \cdot |\Theta|)$ or $O(n)$ time from the difference of cumulative sums at the endpoints. In the discrete model, precomputed log-odds coefficients $A_{ks} = log\theta_s - log(1-\theta_s)$ and $B_{ks} = log(1-\theta_s)$ reduce the inner loop to a multiply-accumulate operation per state. It is worth noting that the

[1] Specifically, M_{sk} and D_{sk} are replaced by $M_{sk+\alpha}$ and $D_{sk+2\alpha}$ respectively, with =0.01, preventing evaluation of log(0) when a segment is fully methylated or unmethylated.

pseudo-count based regularisation of the segments used in (Loyfer et al. 2023) does not satisfy the above PELT conditions, although in practical terms capping the length of segments makes their otherwise exact dynamic programming approach tractable.

Penalty calibration and detection sensitivity. The total penalty per segment is $\beta = \beta_0 \cdot n$, where β_0 is the user-specified per-sample penalty. This linear scaling ensures that the penalty grows with the amount of evidence available, so that the number of segments is controlled by signal strength rather than cohort size.

To understand when a methylation change is detectable, consider a segment of L consecutive CpGs at uniform coverage D, where a fraction x of samples carry an alternate methylation state at contrast $\delta(\theta)$ from the background (measured as the difference in log-likelihood between the two states per read). The aggregate log-likelihood gain from splitting this segment is approximately $x \cdot n \cdot L \cdot D \cdot \delta(\theta)$, and this must exceed the penalty $\beta = \beta_0 \cdot n$ to justify a new segment boundary. Rearranging gives the minimum detectable fraction of altered samples: $x_{min} \approx \frac{\beta_0}{L \cdot D \cdot \delta(\theta)}$ This expression is independent of n: increasing cohort size does not lower the detection threshold for a fixed per-sample penalty, because both signal and penalty scale linearly with n. Instead, sensitivity is improved by longer segments (more CpGs), higher coverage, or greater methylation contrast. For the discrete states $\{0.05, 0.95\}$, the contrast $\delta \approx 2.9$ per read, so at $\beta_0 = 2.5$, $L = 10$ CpGs, and $D = 15\times$ coverage, the minimum detectable fraction is roughly $x_{min} \approx 0.6\%$ of samples — equivalent to a single colony in a donor with 170 samples.

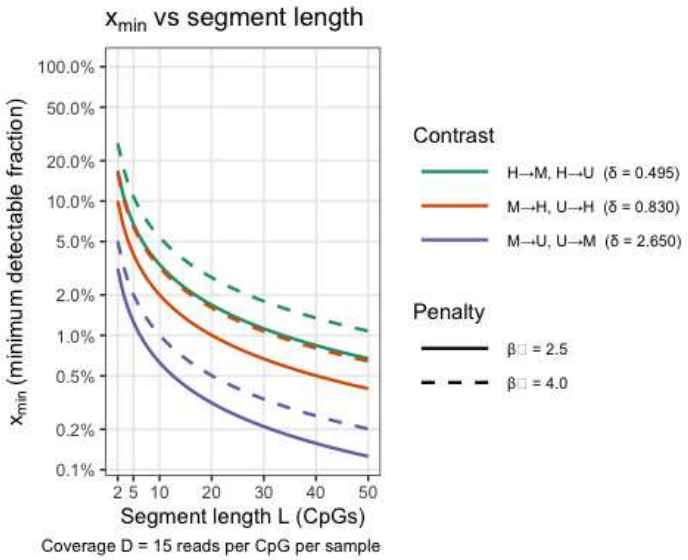


Figure: **Sensitivity of discrete-state methylation segmentation detection against homogeneous background.** The minimum fraction of samples (x_{min}) required to detect a segment that differs from an homogeneous background, plotted against segment length in CpGs. The curves show three classes of methylation state transition: high-contrast (Methylated->Unmethylated and vice-versa; δ = 2.650, purple), intermediate-contrast (Methylated or Unmethylated to half-methylated; δ = 0.830, orange), and low-contrast transitions (Half-methylated to either Methylated or Unmethylated; δ = 0.495, green). Solid and dashed lines correspond to per-sample penalties of β_0 = 2.5 and β_0 = 4.0 respectively. Coverage D = 15 reads per CpG per sample. H = half-methylated (θ = 0.5), M = methylated (θ = 0.95), U = unmethylated (θ = 0.05).

Hierarchical segmentation. Our cohort comprises multiple donors, each contributing tens to hundreds of samples (cell colonies). A single joint segmentation of the full cohort detects methylation domain boundaries shared across donors — for example, transitions between active and repressed chromatin — and is able to detect tracts that recurrently differ between samples across the cohort, but may lack power to detect changes that are private to a single donor, since these affect only a small fraction of the total sample count. We address this with a two-level hierarchical workflow. In the first step, all samples across all donors are segmented jointly, producing a cohort-level partition. In the second step, each donor is segmented independently, but constrained to respect the cohort-

level boundaries: the algorithm can introduce new breakpoints within cohort segments but cannot move or remove existing ones. This is implemented by running PELT independently within each cohort-defined region, using only the columns belonging to that donor. Because the penalty now scales with the donor's sample count rather than the full cohort, changes present in even a small number of that donor's colonies become detectable. The donor-level segmentations are then combined by taking the union of all breakpoints across donors: the refined segmentation has a boundary wherever any donor has one. Finally, each sample's methylation state (or MLE level) is classified within the refined segments, producing a single consistent annotation across the cohort. The result is a segmentation that captures global changes in methylation profile, recurrent changes at modest frequency that occur across donors and also changes (perhaps phylogenetically related) that occur within donors.

Implementation. The method is implemented in an R package LAUREL. The core PELT algorithms are implemented in C++ via Rcpp. Cumulative sums are computed internally in C++ for only the requested sample columns, avoiding unnecessary memory allocation when segmenting donor subsets. The dominant memory cost is $O(m \times n_{donor})$ for the two cumulative sum matrices. R-level wrappers provide a unified interface for both models, with arguments for boundary constraints and column subsetting, and the hierarchical workflow is orchestrated with parallel::mclapply for donor-level parallelism.

Summary. We have provided a principled framework for joint segmentation of multi-sample methylation data. The discrete and continuous models offer complementary strengths: the discrete model is naturally regularised and produces interpretable state assignments without parameter estimation, while the continuous model captures graded methylation differences and is more appropriate for bulk methylation samples. PELT ensures exact optimality with practical scalability to genome-wide data across large cohorts. The hierarchical workflow — cohort-level segmentation followed by boundary-constrained donor-level refinement — attempts to balance shared structure with individual sensitivity, enabling detection of both common methylation domains and rare somatic changes. The per-sample penalty β_0 provides a single, interpretable parameter that directly controls the minimum detectable signal in terms of the number of altered samples, segment length, and sequencing depth.

References.

K*illick, Rebecca, Paul Fearnhead, and Idris A Eckley. 2012. "Optimal Detection of Changepoints with a Linear Computational Cost." Journal of the American Statistical Association 107 (500): 1590–98.*

Loyfer, Netanel, Judith Magenheim, Ayelet Peretz, Gordon Cann, Joerg Bredno, Agnes Klochendler, Ilana Fox-Fisher, et al. 2023. "A DNA Methylation Atlas of Normal Human Cell Types." Nature 613: 355–64.

Supplementary Files

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

- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx