https://doi.org/10.1038/s41586-019-1198-z

Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia

Federico Gaiti[1,2,10], Ronan Chaligne[1,2,10], Hongcang Gu[3,10], Ryan M. Brand[1,2], Steven Kothen-Hill[1,2], Rafael C. Schulman[1,2], Kirill Grigorev[2], Davide Risso[2,4], Kyu-Tae Kim[1,2], Alessandro Pastore[5], Kevin Y. Huang[1,2], Alicia Alonso[2], Caroline Sheridan[2], Nathaniel D. Omans[1,2], Evan Biederstedt[1,2], Kendell Clement[3], Lili Wang[6], Joshua A. Felsenfeld[2], Erica B. Bhavsar[2], Martin J. Aryee[3,7], John N. Allan[2], Richard Furman[2], Andreas Gnirke[3], Catherine J. Wu[3,8], Alexander Meissner[3,9] & Dan A. Landau[1,2]*

Genetic and epigenetic intra-tumoral heterogeneity cooperate to shape the evolutionary course of cancer[1]. Chronic lymphocytic leukaemia (CLL) is a highly informative model for cancer evolution as it undergoes substantial genetic diversification and evolution after therapy[2,3]. The CLL epigenome is also an important disease-defining feature[4,5], and growing populations of cells in CLL diversify by stochastic changes in DNA methylation known as epimutations[6]. However, previous studies using bulk sequencing methods to analyse the patterns of DNA methylation were unable to determine whether epimutations affect CLL populations homogeneously. Here, to measure the epimutation rate at single-cell resolution, we applied multiplexed single-cell reduced-representation bisulfite sequencing to B cells from healthy donors and patients with CLL. We observed that the common clonal origin of CLL results in a consistently increased epimutation rate, with low variability in the cell-to-cell epimutation rate. By contrast, variable epimutation rates across healthy B cells reflect diverse evolutionary ages across the trajectory of B cell differentiation, consistent with epimutations serving as a molecular clock. Heritable epimutation information allowed us to reconstruct lineages at high-resolution with single-cell data, and to apply this directly to patient samples. The CLL lineage tree shape revealed earlier branching and longer branch lengths than in normal B cells, reflecting rapid drift after the initial malignant transformation and a greater proliferative history. Integration of single-cell bisulfite sequencing analysis with single-cell transcriptomes and genotyping confirmed that genetic subclones mapped to distinct clades, as inferred solely on the basis of epimutation information. Finally, to examine potential lineage biases during therapy, we profiled serial samples during ibrutinib-associated lymphocytosis, and identified clades of cells that were preferentially expelled from the lymph node after treatment, marked by distinct transcriptional profiles. The single-cell integration of genetic, epigenetic and transcriptional information thus charts the lineage history of CLL and its evolution with therapy.

To measure the intra-sample variability in the epimutation rate, we profiled single-cell DNA methylation (DNAme) of 831 normal B cells from six healthy donors, including B cells across the maturation spectrum, and 1,821 cells from 12 primary CLL samples with or without mutations in the gene encoding the immunoglobulin heavy-chain variable region (*IGHV*) (M-CLL or U-CLL, respectively; Fig. 1a, b; Extended Data Figs. 1, 2; Supplementary Tables 1–4). The average epimutation rate (measured by the proportion of discordant reads (PDR)[6]; Fig. 1c) was higher in B cells from patients with CLL than from healthy donors (Mann–Whitney *U*-test, $P = 0.0003$; Fig. 1d), in line with previous bulk DNAme sequencing results[6]. Notably, the single-cell measurement showed that the CLL epigenome exhibited consistently increased rates of epimutation (that is, low cell-to-cell variation in epimutation rates), irrespective of their *IGHV* mutational status, compared with $CD19^+$ B cells (Mann–Whitney *U*-test, $P = 0.0006$; Fig. 1e; Extended Data Fig. 3a). Lower variability in the epimutation rate in CLL than in normal B cells was observed across all genomic regions, including regions that were hypermethylated (such as CpG islands (CGIs)) or hypomethylated (intergenic regions) in CLL (Extended Data Fig. 3b–e). The common origin of CLL cells from a single, transformed cell is thus reflected in minimal cell-to-cell epimutation rate variability. By contrast, normal B cells represent an admixture of cells with different replicative histories, with newly formed naive cells intermixed with long-lived post-germinal centre memory B cells, and have highly variable epimutation rates. The epimutation rates of index-sorted B cell subsets progressively increased during B cell maturation (Fig. 1f; Extended Data Fig. 3f, g). Notably, the CLL epimutation rate showed lower cell-to-cell variation than even these well-defined B cell subsets, especially those from low- to high-maturity memory B cells, which more closely resemble CLL in their epigenetic profiles[4] (Extended Data Fig. 3h). These results are consistent with the epimutation rate correlating with the proliferative history of the cell, and serving as an epigenetic molecular clock[7–9].

To extend the assessment of epimutation beyond DNAme concordance within single sequencing reads[6,7], we measured the concordance odds ratio of DNAme between pairs of neighbouring CpGs as a function of their genomic distance (Extended Data Fig. 4a). We observed a faster concordance decay in CLL at genomic regions with known regulatory roles, such as promoter CGIs, suggesting an erosion of CGI spatial organization (Mann–Whitney *U*-test, $P = 0.0013$; Extended Data Fig. 4b). Faster concordance decay involved promoters of TP53 targets, genes differentially methylated across cancer, and genes associated with cell stemness (Extended Data Fig. 4c, e), previously reported to exhibit a high epimutation rate[6], but not promoters of housekeeping genes (Extended Data Fig. 4d). Therefore, CLL epimutation also alters DNAme at larger scales[10], in addition to local methylation disorder[6].

Although stochastic diversification by epimutation occurs in CLL, a minority of CpGs may maintain stable DNAme owing to an active role in the leukaemia regulatory code. To identify CpGs with a low epimutation rate, we adapted the four-gamete test[11] to measure the epimutation rate at single-CpG resolution (Fig. 1g; see Methods). As expected, the frequency of four gametes was positively correlated with the PDR measurement of epimutation (Spearman's rho = 0.32, $P = 3.263 \times 10^{-14}$). Across the 12 CLL patient samples, 166,720 CpGs exhibited a lower four-gamete frequency than expected based on their DNAme level, representing 1.22% ± 0.42 (mean ± s.e.m.) of assessable CpGs per sample (Fig. 1h; Extended Data Fig. 5a–c; Supplementary Table 5). Consistent with the key role of transcription factors in the patterning of DNAme in CLL[4], we identified enrichment in gene promoters for binding motifs of transcription factors with established roles in CLL progression at sites that surround low epimutation CpGs (±25 base pairs (bp)), including

[1]New York Genome Center, New York, NY, USA. [2]Weill Cornell Medicine, New York, NY, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Statistical Sciences, University of Padova, Padova, Italy. [5]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Beckman Research Institute, City of Hope, Monrovia, CA, USA. [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Dana-Farber Cancer Institute, Boston, MA, USA. [9]Max Planck Institute for Molecular Genetics, Berlin, Germany. [10]These authors contributed equally: Federico Gaiti, Ronan Chaligne, Hongcang Gu. *e-mail: dlandau@nygenome.org

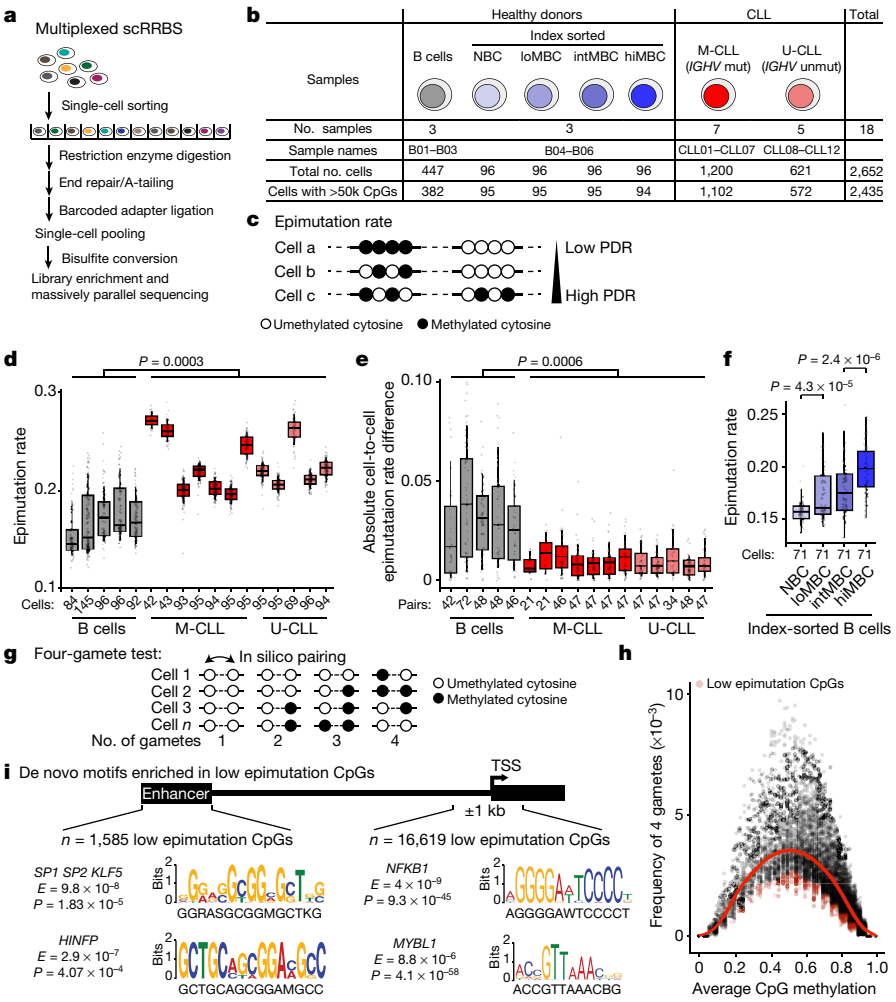


Fig. 1 | CLL epigenomes show increased epimutation rate with low cell-to-cell variation. a, Schematic of the protocol for multiplexed single-cell RRBS (scRRBS). See also Extended Data Fig. 1a. **b**, Summary of samples from healthy donors and patients with CLL. NBC, naive B cells; loMBC, intMBC and hiMBC, low-, intermediate- and high-maturity memory B cells; M-CLL, *IGHV*-mutated CLL; U-CLL, *IGHV*-unmutated CLL. **c**, Epimutations are measured as the proportion of discordant reads (PDR). **d**, Single-cell epimutation rate across B cells from healthy donors (samples B01–B02, B04–B06; $n = 5$) and patients with CLL (samples CLL01–CLL12; $n = 12$). **e**, The difference in cell-to-cell epimutation rate across normal B (samples B01–B02, B04–B06; $n = 5$) and CLL (samples CLL01–CLL12; $n = 12$) cells. **f**, Single-cell epimutation rate across index-sorted normal B (samples B04–B06; $n = 3$) cells. **g**, Schematic of the procedure for the four-gamete test (see Methods). **h**, Frequency of four gametes according to the level of average methylation of each CpG across CLL cells (sample CLL04 is shown as a representative example; $n = 29{,}114$ low epimutation CpGs out of a total of 1,835,994 CpGs assessed; see also Extended Data Fig. 5a). Red line denotes smooth local regression line. Low epimutation CpGs are indicated in red. **i**, Sequence logos of the DNA motifs significantly overrepresented in low epimutation CpGs (±25 bp) at promoters or enhancers, across CLL samples. For each motif, the *E* value and the TOMTOM *P* value are shown. See Methods for details on de novo motif enrichment analysis, and Extended Data Fig. 5d for additional motifs. TSS, transcription start site. In all figures, box plots represent the median, bottom and upper quartiles, whiskers correspond to 1.5× the interquartile range. *P* values were determined by Mann–Whitney *U*-test (**d**–**f**), comparing the median values across samples (**d**, **e**).

NFKB1[12] and *MYBL1*, which encodes a transcription factor that is involved in MYC activation in lymphoid neoplasms[13] (Fig. 1i, right; Extended Data Fig. 5d, e; Supplementary Table 6).

DNAme of enhancers can also affect transcriptional activity and cellular phenotypes in CLL[14]. Low epimutation enhancer CpGs ($n = 1{,}585$; Supplementary Table 7) were located in proximity to genes implicated in lymphoproliferation, including *NOTCH1*, *NFATC1* and *FOXC1*, and genes involved in key CLL pathways (for example, the WNT and MAPK signalling pathways[15]; Benjamini–Hochberg false discovery rate (FDR) adjusted $P < 0.2$). Low epimutation enhancer CpGs were also enriched for binding sites of *SP1*, a component of the CLL regulatory network[16], and the transcriptional repressor *HINFP* that is involved in DNAme-mediated gene silencing[17] (Fig. 1i, left; Extended Data Fig. 5d, e; Supplementary Table 8). This suggests that conserved CpG sites are protected from alterations in DNAme by transcription factor binding, by either direct exclusion of methylases or negative selection due to a disruption of the CLL regulatory code.

To examine the effect of epimutation on gene expression at the single-cell level, we integrated multiplexed single-cell reduced-representation bisulfite sequencing (scRRBS) with whole-transcriptome sequencing (Fig. 2a; Extended Data Fig. 6a). Although the expected relationship between promoter DNAme and gene silencing was preserved in both CLL and normal B cells (Extended Data Fig. 6b), a higher single-cell epimutation rate in CLL was associated with higher transcriptional entropy—a measure of heterogeneity of gene expression within cells[18]—than in normal B samples, consistent with transcriptional dysregulation in CLL (Fig. 2b; Extended Data Fig. 6c–e). A negative correlation between promoter DNAme and gene expression was observed at the single-cell level in both CLL and normal B cells (Fig. 2c–f; Extended Data Fig. 6f–n), but was more pronounced in CLL (Fig. 2e; Extended Data Fig. 6j, n), suggesting that, at least in part, the decreased epigenetic–transcriptional coordination observed in bulk CLL sequencing[6] results from intra-leukaemic epigenetic diversity. A subset of genes exhibited positive correlation between expression and

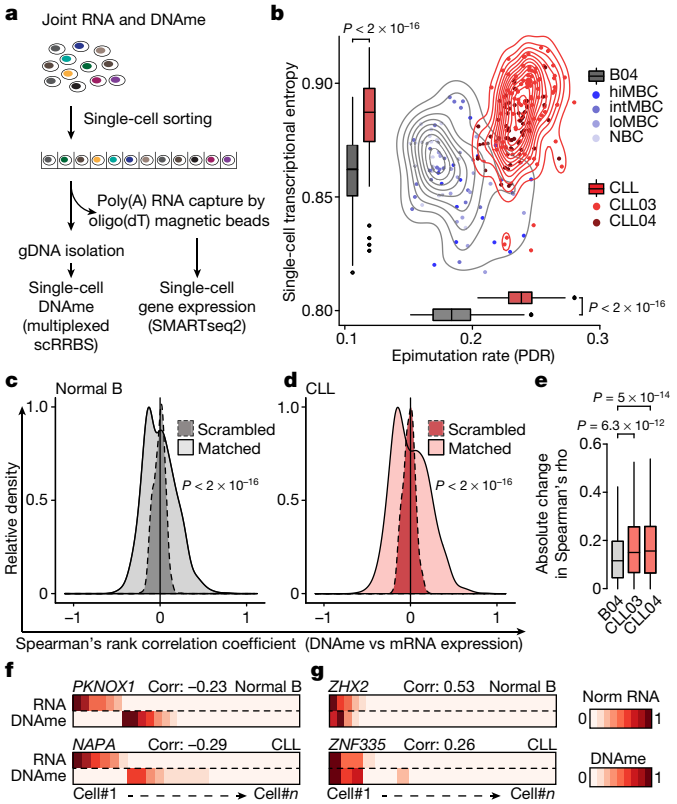


Fig. 2 | CLL higher epimutation rate is associated with higher transcriptional entropy consistent with transcriptional dysregulation. **a**, Schematic of the joint methylomics and transcriptomics method, applied to normal B cells (B04, $n = 84$) and CLL cells (CLL03, $n = 94$; CLL04, $n = 92$). **b**, Single-cell transcriptional entropy[18] and epimutation rate for normal B and CLL cells. See also Extended Data Fig. 6c. **c**, **d**, Distribution of Spearman's rho of expression and promoter DNAme correlation (genes expressed >5 cells, DNAme >5 CpGs per promoter) across normal B cells (**c**; B04, 3,239 genes), or CLL cells (**d**; CLL03, 2,699 genes). Spearman's rho value distribution was compared to the distribution of values obtained by randomly permuted cell labels. See also Extended Data Fig. 6g–n. **e**, Absolute change in Spearman's rho when comparing matched versus permuted DNAme and RNA single-cell data in CLL and normal B cells, across 990 genes overlapping between the three samples. **f**, **g**, Representative genes with negative (**f**) or positive (**g**) correlation in single-cell expression methylation, determined by Spearman correlation coefficient. Scale bar represents promoter methylation and RNA read counts scaled by maximal value. *P* values were determined by Mann–Whitney *U*-test (**b**), Kolmogorov–Smirnov test (**c**, **d**) or Wilcoxon signed-rank test (**e**).

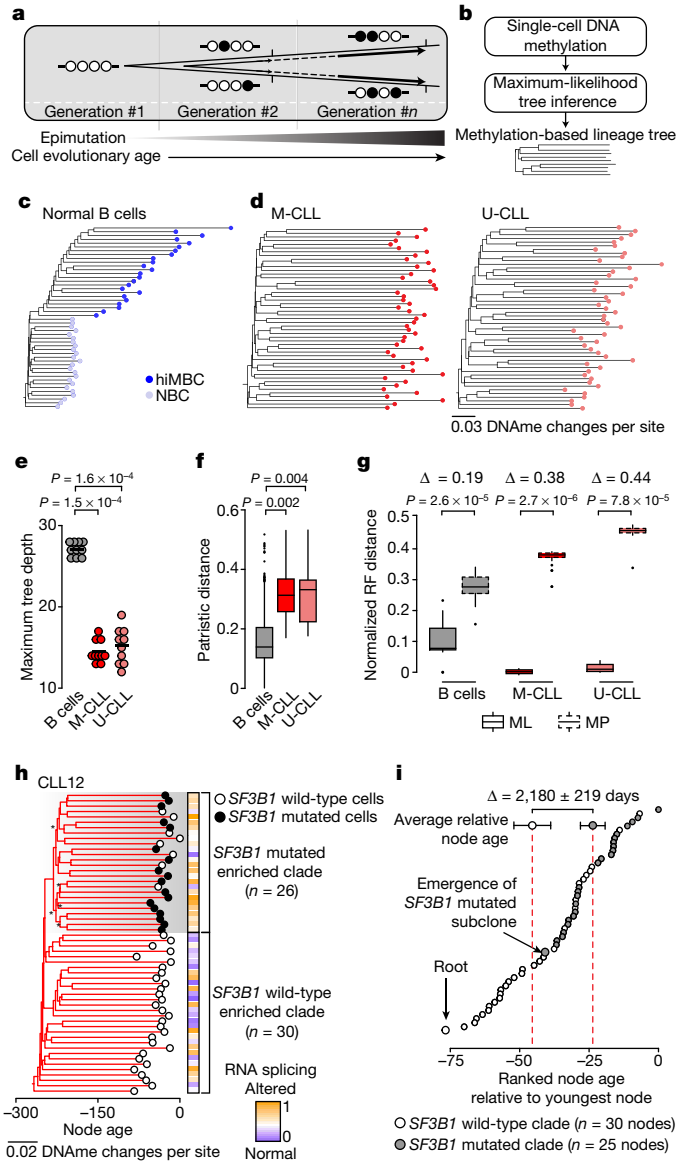


Fig. 3 | Lineage relationships inferred from single-cell DNA methylomes. **a**, Epimutation as a somatic cell molecular clock, in which each division has a given likelihood of generating epimutations. **b**, Schematic of methylation-based lineage tree inference (see Methods). **c**, **d**, Representative (random cell subsampling) methylation-based lineage trees of index-sorted normal B cells (B05, NBC and hiMBC cells) and CLL cells (M-CLL (CLL07); U-CLL (CLL10)). See also Extended Data Fig. 7. **e**, Maximum tree depth of lineage trees (one value per tree) between normal B (B05) and CLL (M-CLL (CLL07); U-CLL (CLL10)) samples ($n = 10$ tree replicates; see Methods). Black lines represent averages. **f**, Patristic distances between CLL samples (M-CLL, $n = 7$; U-CLL, $n = 5$) and normal B samples ($n = 6$). One representative tree of randomly sampled cells for each sample was used. **g**, Normalized Robinson–Foulds (RF) distances between any two trees ($n = 30$ tree replicates; see Methods) of CLL (M-CLL (CLL07); U-CLL (CLL10)) and normal B (B05) cells reconstructed by maximum-likelihood (ML) versus maximum-parsimony (MP). Differences (Δ) are indicated. **h**, Methylation-based lineage tree of CLL12 with projection of the *SF3B1* mutation status. Asterisk indicates bootstrap values <50%. Tree was rooted by including five normal B cells as an outgroup (not shown). **i**, Ranked node ages relative to the youngest node in the tree in **h**. Difference in average node ages between the wild-type and the mutated *SF3B1* clade is shown. Node ages (estimated number of divisions before present) are calculated by dividing node heights by a rate of 0.0005 changes per CpG site per division[29] (see Methods). A total of 21.8 divisions translate to 2,180 days at a 1% cell proliferation rate per day[30]. Error bars represent 95% confidence interval. *P* values were determined by Mann–Whitney *U*-test (**e**–**g**).

promoter DNAme (Fig. 2g; Extended Data Fig. 6f, right), enriched in genes marked by cytosine hydroxymethylation, which is known to be positively correlated with gene expression[19] (Extended Data Fig. 6o).

Because epimutations may serve as a molecular clock[9], we used the heritable epimutation information to reconstruct methylation-based lineage relationships in CLL and normal B cells (Fig. 3a–d; Extended Data Fig. 7a–e; see Methods). CLL lineage trees exhibited early branching with a lower maximum tree depth (Fig. 3e; Extended Data Fig. 7f) and homogeneous root-to-tip branch lengths (Extended Data Fig. 7g), consistent with rapid drift after the initial malignant transformation ('big bang' cancer evolutionary framework[20]). Moreover, the homogeneous branch length is inconsistent with a notable contribution from cancer stem cells in CLL, in contrast to data that revealed highly divergent replicative histories in acute myeloid leukaemia[21], in which cancer stem cells have been well described. The greater proliferative histories in CLL were reflected in increased accumulation of epimutation resulting in higher patristic distances (that is, the sum of the lengths of branches that link two tips in a tree) than in normal B cell trees

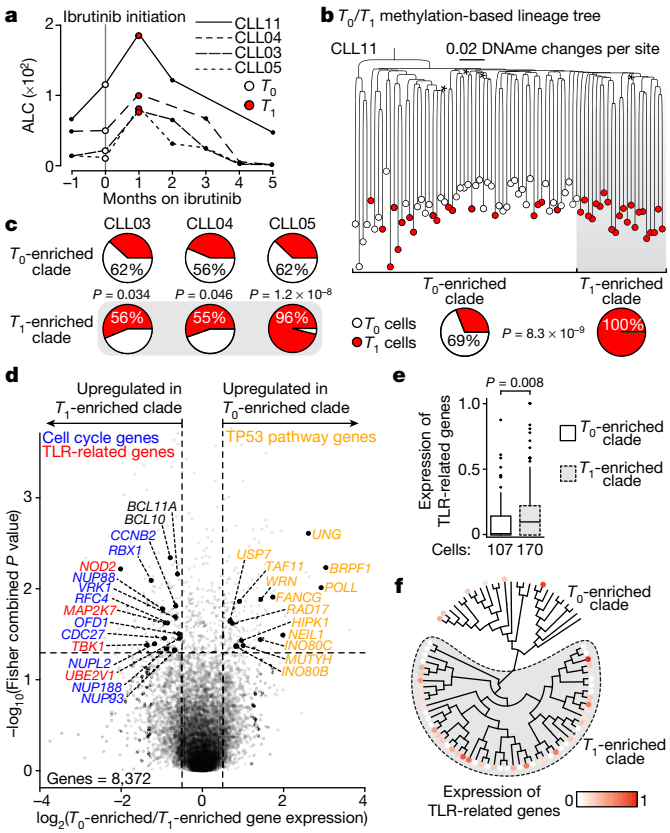


Fig. 4 | Joint single-cell methylomics and RNA sequencing link lineage and transcriptional information in CLL evolution. a, Absolute lymphocyte counts (ALC) for the first few months of ibrutinib treatment. Serial multiplexed scRRBS and joint multiplexed scRRBS and RNA sequencing (RNA-seq) were performed before (T_0) and 1 month after (T_1) the initiation of ibrutinib treatment. **b**, Top, representative lineage tree integrating CLL11 cells before treatment (T_0; white circles) and during treatment (T_1; red circles). Each sample is represented by 40 out of 96 randomly sampled cells. Asterisk indicates bootstrap values <50%. Bottom, percentage of T_1 cells in each of the two clades inferred from the lineage tree. **c**, As in **b** (bottom) for samples CLL03, CLL04 and CLL05. **d**, Volcano plot of gene expression comparing T_1 cells from T_1-enriched clades and T_1 cells from T_0-enriched clades (CLL03–CLL05, CLL11) ($n = 8{,}372$ genes expressed >5 cells in ≥3 samples). **e**, Scaled average gene expression across TLR pathway genes from **d** for each T_1 cell from T_1-enriched clades and each T_1 cell from T_0-enriched clades (CLL03–CLL05, CLL11). **f**, Gene expression projections on lineage tree for genes of the TLR pathway from **d** for sample CLL03. Scale bar represents RNA read counts scaled by maximal value. Expression value projection is performed only for T_1 cells, comparing T_1- versus T_0-enriched clades. Asterisk indicates cell without RNA information. *P* values were determined by Fisher's exact test (**b**, **c**), Fisher's combined probability test (**d**; combined across patient samples) or Mann–Whitney *U*-test (**e**).

(Fig. 3f; Extended Data Fig. 7h, i). By contrast, normal B cell clades followed a pattern that was consistent with normal B cell differentiation by exhibiting late branching and deeper tree topology, with younger naive CD27$^-$ B cells showing shorter branches than CD27$^+$ memory terminally differentiated B cells (Fig. 3c; Extended Data Fig. 7b). As expected, normal B cell lineage trees resulted in a smaller increase in fidelity compared with parsimony trees (based on DNAme mismatches between cells; see Methods) than CLL trees, consistent with their non-clonal growth (Fig. 3g).

To validate tree topology inferred via epimutation, we integrated single-cell DNAme and whole-transcriptome sequencing with targeted sequencing of known somatic mutations in the cDNA (Extended Data Fig. 8a). We sampled a CLL that contains a subclonal driver mutation in *SF3B1* (K666N; variant allele frequency of 0.23) and inferred its lineage tree from single-cell DNAme (Fig. 3h; Extended Data Fig. 8b). The *SF3B1*-mutated cells mapped accurately to a distinct clade inferred solely based on epimutation information (Fisher's exact test, $P = 7.4 \times 10^{-9}$; Extended Data Fig. 8c, d). This accurate mapping was probably not due to distinct DNAme profiles of *SF3B1*-mutated cells, given the small number of differentially methylated regions (Extended Data Fig. 8e), but instead due to the ability of stochastic epimutation to trace lineage histories. Cells belonging to the *SF3B1*-mutated clade showed higher alternative 3′ splicing than their wild-type counterparts (Mann–Whitney *U*-test, $P = 0.015$; Extended Data Fig. 8f), consistent with the known *SF3B1*-mediated splicing defect[22], and were marked by a distinct transcriptional profile (Extended Data Fig. 8g, h; Supplementary Table 9). We further observed decreased transcriptional similarity between cells as a function of their lineage distance, providing a direct measurement of the heritability of the transcriptional profile in a human sample (Mann–Whitney *U*-test, $P = 0.044$; Extended Data Fig. 8i). Notably, cells in the *SF3B1*-mutated clade showed lower node heights (that is, the sum of branch lengths of the longest downward path to a leaf from a given node; Extended Data Fig. 8j) and longer root-to-tip branch lengths than cells in the wild-type *SF3B1* clade (Extended Data Fig. 8k), consistent with *SF3B1* mutation as a late subclonal event in CLL[15]. The molecular clock feature of epimutations further enabled the timing of the subclonal divergence in the evolutionary history of CLL, estimated to have occurred 2,180 ± 219 days after the emergence of the parental clone (Fig. 3i; Extended Data Fig. 8l).

Next, we applied joint single-cell DNAme and whole-transcriptome sequencing analysis to study the dynamic changes that occur during therapy with ibrutinib—a targeted agent that abrogates B cell receptor (BCR) signalling. This treatment results in a transient rise in the peripheral blood leukaemic cell burden owing to forced migration of cells from the lymph node niche[23]. To examine potential lineage biases in ibrutinib-induced CLL migration, we profiled four CLLs, without subclonal genetic drivers, before (T_0) and during (T_1) ibrutinib-associated lymphocytosis (Fig. 4a). Lineage trees that integrated T_0 and T_1 cells identified major clades enriched for T_1 cells in each of the CLL samples (Fig. 4b, c; Extended Data Fig. 9a–c; see Methods), despite few differences in DNAme between the T_1-enriched clades and other T_1 cells (Extended Data Fig. 9d). These data suggest that different CLL lineages may be preferentially affected by ibrutinib and expelled from the lymph node after treatment. Projection of transcriptomic data onto the lineage trees revealed that the T_1-enriched clade cells were marked by increased expression of *BCL11A*—a proto-oncogene with expression restricted to the lymph node[24]—and increased expression of *BCL10*—an upstream regulator of the NF-κB pathway in the BCR signalling cascade. Genes related to cell cycle and proliferation pathways (Fig. 4d; Extended Data Fig. 10a; Supplementary Tables 10, 11) were also overexpressed in T_1-enriched clades compared with other T_1 cells. As the lymph node is the primary anatomical site of CLL proliferation[25], these findings are consistent with the recent expulsion of cells of T_1-enriched clades from the lymph node after the initiation of treatment. T_1-enriched clades across patients were also found to have upregulation of genes of the Toll-like receptor (TLR) pathway (Fig. 4d–f; Extended Data Fig. 10b). The TLR pathway is known to interact with the ibrutinib-inhibited BCR signalling pathway, as it has been identified in functional genomics screens for ibrutinib sensitivity[26], and the pathway is specifically activated in CLL cells in the lymph node niche, triggering activation of the pro-survival NF-κB pathway[27,28], which was also upregulated in T_1-enriched clades (Extended Data Fig. 10c). Because the abnormal activation of the TLR pathway may disrupt lymph node trafficking, these results are consistent with clades enriched in ex-migrating cells, and also suggest the potential for dual inhibition of the BCR and TLR pathways, as described ex vivo[27,28].

Collectively, by leveraging the heritable information captured by epimutation, we have retraced the evolutionary histories of CLL and charted its evolution after therapy, demonstrating how different lineages may be preferentially affected by a therapeutic intervention, even in the absence of genetic subclonal drivers. We foresee that future application of multi-modality single-cell sequencing will enable the

annotation of intra-tumoral disparities in transcription in response to therapy with precise lineage history information, as well as the integration of genetic, epigenetic and transcriptional information at the atomic unit of somatic evolution—the single cell.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1198-z.

Received: 21 December 2017; Accepted: 12 April 2019;
Published online 15 May 2019.

1. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017).
2. Burger, J. A. et al. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nat. Commun.* **7**, 11589 (2016).
3. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
4. Beekman, R. et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).
5. Oakes, C. C. et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–264 (2016).
6. Landau, D. A. et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).
7. Landan, G. et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
8. Shipony, Z. et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).
9. Shibata, D. Mutation and epigenetic molecular clocks in cancer. *Carcinogenesis* **32**, 123–128 (2011).
10. Hansen, K. D. et al. Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
11. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
12. Chen, T. L. et al. NF-κB p50 (*nfkb1*) contributes to pathogenesis in the Eμ-TCL1 mouse model of chronic lymphocytic leukemia. *Blood* **130**, 376–379 (2017).
13. Arsura, M., Hofmann, C. S., Golay, J., Introna, M. & Sonenshein, G. E. A. A-myb rescues murine B-cell lymphomas from IgM-receptor-mediated apoptosis through c-myc transcriptional regulation. *Blood* **96**, 1013–1020 (2000).
14. Qu, Y. et al. Cancer-specific changes in DNA methylation reveal aberrant silencing and activation of enhancers in leukemia. *Blood* **129**, e13–e25 (2017).
15. Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
16. Rendeiro, A. F. et al. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat. Commun.* **7**, 11938 (2016).
17. Sekimata, M. & Homma, Y. Sequence-specific transcriptional repression by an MBD2-interacting zinc finger protein MIZF. *Nucleic Acids Res.* **32**, 590–597 (2004).
18. Grün, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
19. Caron, G. et al. Cell-cycle-dependent reconfiguration of the DNA methylome during terminal differentiation of human B cells into plasma cells. *Cell Reports* **13**, 1059–1071 (2015).
20. Sottoriva, A. et al. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
21. Shlush, L. I. et al. Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. *Blood* **120**, 603–612 (2012).
22. Wang, L. et al. Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. *Cancer Cell* **30**, 750–763 (2016).
23. Herman, S. E. M. et al. Ibrutinib-induced lymphocytosis in patients with chronic lymphocytic leukemia: correlative analyses from a phase II study. *Leukemia* **28**, 2188–2196 (2014).
24. Satterwhite, E. et al. The *BCL11* gene family: involvement of *BCL11A* in lymphoid malignancies. *Blood* **98**, 3413–3420 (2001).
25. Herndon, T. M. et al. Direct in vivo evidence for increased proliferation of CLL cells in lymph nodes compared to bone marrow and peripheral blood. *Leukemia* **31**, 1340–1347 (2017).
26. Phelan, J. D. et al. A multiprotein supercomplex controlling oncogenic signalling in lymphoma. *Nature* **560**, 387–391 (2018).
27. Herishanu, Y. et al. The lymph node microenvironment promotes B-cell receptor signaling, NF-κB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood* **117**, 563–574 (2011).
28. Dadashian, E. L. et al. TLR signaling is activated in lymph-node resident CLL cells and is only partially inhibited by ibrutinib. *Cancer Res.* **79**, 360–371 (2019).
29. Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavaré, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl Acad. Sci. USA* **106**, 4828–4833 (2009).
30. Messmer, B. T. et al. In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. *J. Clin. Invest.* **115**, 755–764 (2005).

Acknowledgements We thank the Epigenomics Core Facility at Weill Cornell Medicine for technical help. R.C. is supported by Leukemia Research Foundation (LRF) and Marie Skłodowska-Curie fellowships. A.G. is supported by Broad Institute SPARC funding. D.A.L. is supported by the Burroughs Wellcome Fund Career Award for Medical Scientists, American Society of Hematology (ASH) Scholar Award, Pershing Square Sohn Prize for Young Investigators in Cancer Research, and the National Institutes of Health (NIH) Director's New Innovator Award (DP2-CA239065). This work was also supported by the Starr Foundation, the Max Planck Society, Leukemia & Lymphoma Society (LLS) Translational Research Program, National Cancer Institute (R01-CA229902), and Stand Up To Cancer Innovative Research Grant (SU2C-AACR-IRG-0616).

Reviewer information *Nature* thanks Ken Duffy and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions F.G., R.C., C.J.W., A.M. and D.A.L. conceived and designed the project. R.C., H.G., A.A., C.S., N.D.O., L.W., J.A.F., E.B.B., J.N.A., R.F. and A.G. performed patient selection and prepared samples for sequencing. R.C., H.G., A.G., D.A.L. and A.M. designed and developed multiplexed scRRBS and joint multiplexed scRRBS and single-cell RNA sequencing. F.G., R.M.B., S.K.-H., R.C.S., K.G., D.R., K.T.K., A.P., K.Y.H., E.B., K.C., M.A. and D.A.L. performed the computational genomics analyses. F.G., R.C., C.J.W., A.M. and D.A.L. wrote the manuscript with comments and contributions from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1198-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1198-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.A.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Human subjects, sample collection and genotyping. The study was approved by the local ethics committee and by the Institutional Review Board (IRB) and conducted in accordance to the Declaration of Helsinki protocol. Blood samples were collected in EDTA blood collection tubes (BD Biosciences) from patients and healthy adult volunteers enrolled on clinical research protocols at the Dana-Farber/Harvard Cancer Center (DF/HCC) and New York-Presbyterian/Weill Cornell Medical Center (NYP/WCMC), approved by the DF/HCC and NYP/WCMC IRBs. We note that the IRB does not permit collection of demographic information of healthy donors. The diagnosis of CLL according to WHO (World Health Organization) criteria was confirmed in all cases by flow cytometry, or by lymph node or bone marrow biopsy. Informed consent on DF/HCC and NYP/WCMC IRB-approved protocols for genomic sequencing of patient samples was obtained before the initiation of sequencing studies. B cells from healthy donors and CLL patient samples were isolated from blood samples using Ficoll-Paque Plus (GE Healthcare) density gradient centrifugation and red blood cell lysis, followed by EasySep Human B Cell Enrichment Kit (STEMCELL Technologies, Vancouver, Canada) as per the manufacturer's recommendations. *IGHV* homology was determined[31] (unmutated was defined as greater than or equal to 98% homology to the closest germline match). Cytogenetics were primarily evaluated by FISH analysis for the most common CLL abnormalities (del(13q), trisomy 12, del(11q), del(17p), del(6q), amp(2p)); if FISH was unavailable, genomic data were used (Supplementary Table 12). The presence and location of recurrent somatic mutations were detected in the genes tested by Genoptix clinical grade CLL gene panel testing (Genoptix; Supplementary Table 13).

Multiplexed scRRBS library construction. Single-cell methylome profiling was performed with multiplexed scRRBS—an adaption of a previous scRRBS protocol[32,33] that enables throughput to be increased by the addition of cell barcodes early in the scRRBS protocol. Specifically, single-cell experiments were performed by sorting DAPI-negative cells into 96-well plates in 3 μl of 0.1× CutSmart buffer (New England Biolabs) per well using a BD Influx sorter (Becton Dickinson). Normal B cells for sample B04, B05 and B06 were further index-sorted using the following sorting strategy: NBC ($CD27^-$, IgM^+, IgD^{+++}), loMBC ($CD27^-$, IgM^+, IgD^+), intMBC ($CD27^+$, IgM^+, IgD^{++}) and hiMBC ($CD27^+$, IgG^+). The antibodies used were: FITC mouse anti-human IgD (clone IA6-2, BD Pharmingen), APC mouse anti-human IgG (clone G18-145, BD Biosciences), APC/Cy7 anti-human IgM (clone MHM-88, BioLegend) and PE/Cy7 anti-human CD27 antibody (clone O323, Bio Legend). Plates were then stored at −80 °C until further processing. The day of the experiment, cells were lysed for 2 h at 50 °C in 1× CutSmart buffer supplemented with Proteinase K (0.2 U, NEB) and Triton X-100 (0.3%, Sigma Aldrich) for a final volume of 5 μl. Proteinase K was heat-inactivated for 30 min at 75 °C. DNA was incubated with 10 units of the restriction enzyme Msp1 (Fermentas) in 6.5 μl final volume reaction during 90 min at 37 °C. Heat-inactivation was performed for 10 min at 70 °C. Digested DNA was filled-in and A-tailed at the 3′ sticky ends in 8.5 μl final volume of 1× CutSmart with 2.5 units of Klenow fragment (Exo-, Fermentas). Reaction was supplemented with 1 mM dATP and 0.1 mM dCTP and 0.1 mM dGTP (NEB) and performed as follows in a thermocycler: 30 °C for 25 min, 37 °C for 25 min and heat-inactivation at 70 °C for 10 min. Custom barcoded methylated adaptors (0.1 μM) were then ligated overnight at 16 °C with the dA-tailed DNA fragments in the presence of 800 units of T4 DNA ligase (NEB) and 1 mM ATP (Roche) in a final volume of 11.5 μl of 1× CutSmart buffer. T4 DNA ligase heat-inactivation was performed at 70 °C for 15 min the next day. Genomic DNA from 24 individual cells was pooled together according to their barcodes, giving, for a 96-well plate, 4 pools of 24 cells. Pooled genomic DNA was cleaned-up and concentrated using 1.8× SPRI beads (Agencourt AMPure XP, Beckman Coulter). Each pool was then sodium bisulfite-converted (Fast Epitect Bisulphite, Qiagen) following the manufacturer's recommendations. To ensure full bisulfite conversion, two cycles of conversion were performed. The double-stranded DNA was first denatured for 10 min at 98 °C and then incubated for 20 min at 60 °C. Dephosphorylated and sheared bacterial DNA (100 ng) was added as carrier to every pool before conversion. Converted DNA was then amplified using primers containing Illumina i7 and i5 index. Following Illumina pooling guidelines, a different i7 index was used for every 24-cell pool, allowing multiplexing of 96 cells for sequencing on one Illumina HiSeq lane. Library enrichment was done using KAPA HiFi Uracil+ master mix (Kapa Biosystems) and the following PCR condition was used: 98 °C for 45 s; 6 cycles of: 98 °C for 20 s, 58 °C for 30 s, 72 °C for 1 min; followed by 12 cycles of: 98 °C for 20 s, 65 °C for 30 s, 72 °C for 1 min. PCR was terminated by an incubation at 72 °C for 5 min. Enriched libraries were cleaned-up and concentrated using 1.3× SPRI beads. DNA fragments between 200 bp and 1 kb were size-selected and recovered after resolving on a 3% NuSieve 3:1 agarose gel. Library molarity concentration calculation was obtained by measuring concentration of double-stranded DNA (Qubit) and quantifying the average library size (base pairs) using an Agilent Bioanalyzer. Every 24-cell pool was mixed with the others pool in an equimolar ratio. All cells from a 96-well plate were sequenced as paired-end on HiSeq 2500 with 10% PhiX spike-in. Negative controls (empty wells with no cells) were used to control for non-specific amplification of the libraries.

Multiplexed scRRBS read alignment. Each pool of 96 cells was first demultiplexed by Illumina i7 barcodes (Supplementary Table 1), resulting in four pools of 24 cells. Each pool of 24 cells was further demultiplexed by unique cell barcodes (Supplementary Table 2). Reads were assigned to a given cell if they matched 80% of the template adapters. Adapters and adaptor reverse complements (6 bp) were trimmed from the raw sequence reads. After adaptor removal, reads were trimmed from their 3′ end for read quality by applying a 4-bp sliding window and removing bases until the mean base quality of the window had a Phred quality score greater than 15. Read pairs with a read shorter than 36 bp after trimming were discarded. We aligned trimmed reads to the hg19 human genome assembly using Bismark[34] (v.0.14.5; parameters: -multicore 4 -X 1000 -un -ambiguous) running on bowtie2-2.2.8 aligner[35]. Bismark methylation extractor (-bedgraph -comprehensive) was used to determine the methylation state of each individual CpG. For downstream analyses, a site was considered methylated or unmethylated only if there was 90% agreement of the methylation state for all reads mapped to the site. Cells with coverage of at least 50,000 unique CpGs were retained for downstream analyses ($n = 2{,}435$ cells; 92% of the total; Fig. 1b; Extended Data Fig. 1b; Supplementary Table 4), with bisulfite conversion rates of 99.8% ± 0.09 (median ± median absolute deviation) and an average of 276,165 ± 3,765 (mean ± s.e.m.) unique CpGs per cell (Supplementary Table 4). We note that the analysis for Extended Data Fig. 2c was performed before the implementation of this filtering procedure to confirm that single-cell methylation values predominately equal 0 or 1, consistent with the random sampling of a single allele.

Joint multiplexed scRRBS and single-cell RNA-seq library construction. Single cells were sorted by flow cytometry, as above-described, into 5 μl of RLT Plus buffer (Qiagen) supplemented with 1 U μl^{-1} of RNase inhibitor (Lucigen). Sorted cells were immediately store at −80 °C. Genomic DNA (gDNA) and mRNA have been separated manually as previously described[36]. In brief, a modified oligo-dT primer (5′-biotin-triethyleneglycol-AAGCAGTGGTATCAACGCAGAGTACT30VN-3′, in which V is either A, C or G, and N is any base; IDT) was conjugated to streptavidin-coupled magnetic beads (Dynabeads, Life Technologies) according to the manufacturer's instructions. To capture polyadenylated mRNA, we added the conjugated beads (10 μl) directly to the cell lysate and incubated them for 20 min at room temperature with mixing to prevent the beads from settling. The mRNA was then collected to the side of the well using a magnet, and the supernatant, containing the gDNA, was transferred to a fresh plate. Single-cell complementary DNA was amplified from the tubes containing the captured mRNA according to the Smart-seq2 protocol[37]. After amplification and purification using 0.8× SPRI beads, 0.5 ng cDNA was used for Nextera Tagmentation and library construction. Library quality and quantity were assessed using Agilent Bioanalyzer 2100 and Qubit, respectively. gDNA present in the pooled supernatant and wash buffer from the mRNA isolation step was concentrated on 0.8× SPRI beads and eluted directly into the reaction mixtures for Msp1 (±HaeIII) (Fermentas) enzymatic reaction (10 μl final reaction). The multiplexed scRRBS protocol was then performed on the digested gDNA after the restriction enzyme digestion step. To obtain higher coverage single-cell DNA methylomes, we performed double digestion with HaeIII in addition to MspI on cells from patient sample CLL11, increasing coverage to an average of 2,298,281 ± 86,699 (mean ± s.e.m.) reads per cell, and yielding 790,951 ± 24,098 unique CpGs per cell.

Single-cell RNA-seq read-alignment and differential gene expression quantification. The sequenced read fragments were mapped against the hg19 human genome assembly using the 2pass default mode of STAR[38] (v.2.5.2a) with the annotation of GENCODE[39] (v.19). The number of read counts overlapping with annotated genes were quantified applying the 'GeneCounts' option in the STAR alignment. The single-cell transcriptomes recovered an average of 552,201 ± 19,808 reads per cell and 4,211 ± 69 genes per cell, comparable to previous stand-alone single-cell whole-transcriptome data in CLL[6].

Comparison of transcriptional distances as a function of lineage distance between cell pairs was performed by first normalizing the read counts by scaling for the total number of counts per cell. We then performed principal component analysis on the log of the normalized counts and used the first three components to compute the Euclidean distance between each pair of cells (Extended Data Fig. 8i).

Differential expression analyses (Fig. 4d; Extended Data Fig. 8g) were performed using a negative binomial model with observational weights to account for zero inflation[40]. Specifically, we used ZINB-WaVE[41] (v.1.0.0) to estimate a set of observational weights and edgeR (v.3.20.1) to test for differential expression using a weighted F statistic approach, as previously described[42].

In Extended Data Fig. 8g, we defined differentially expressed genes by adjusting nominal P values using a Benjamini–Hochberg FDR procedure (cut-off of adjusted $P < 0.2$), with an additional criterion of an absolute $\log_2$(*SF3B1* mutated-enriched/wild-type-enriched clade gene expression) > 0.5. In Fig. 4d, although the differentially expressed genes were examined individually for each patient (CLL03,

CLL04, CLL05 and CLL11; Supplementary Table 10), they were also examined in combination across the four patients by combining the nominal P values for the differentially expressed genes via Fisher's combined probability test and averaging the fold change in gene expression (Supplementary Table 11). We used Fisher's combined $P < 0.05$ and absolute $\log_2(T_0\text{-enriched}/T_1\text{-enriched gene expression}) > 0.5$ to nominate candidate genes for subsequent gene set enrichment analysis (see 'Gene set enrichment analysis' section). The gene set analysis was then followed by a Benjamini–Hochberg FDR adjustment, correcting the nominal P values for multiple hypothesis testing (cut-off of adjusted $P < 0.2$). Gene expression projections of transcriptomic data onto the lineage trees for differentially expressed genes belonging to TLR pathways in Fig. 4f and Extended Data Fig. 10b was performed by averaging gene expression across genes for each cell. Average gene expression was subsequently scaled by the maximum expression value to bring values into a 0–1 range.

Genome annotations definitions. Promoters were defined as 1 kb upstream and 1 kb downstream of hg19 RefGene gene transcription start sites, unless stated otherwise. The set of CGIs was defined using biologically verified CGIs[43]. Enhancer regions were defined using FANTOM5 human robust enhancer set[44]. To verify the suitability of FANTOM5 human robust enhancer set in the context of CLL, we produced genome-wide maps of H3K27ac by bulk chromatin immunoprecipitation followed by sequencing (ChIP–seq) of two *IGHV*-mutated and two *IGHV*-unmutated CLL patient samples. We observed a large overlap (72%) between FANTOM5 human robust enhancers and the CLL H3K27ac ChIP–seq peaks. In addition, 85% of the low epimutation CpGs at enhancers overlapped with CLL H3K27ac ChIP–seq peaks (1,360 out of 1,585). In Extended Data Fig. 1d, CTCF-binding sites were annotated based on published CTCF binding ChIP–seq experiments generated by the ENCODE Consortium from the GM12878 lymphoblastoid cell line[45]. We curated a list of CTCF-binding sites based on sites that were detected in at least 75% of these samples. The location of long terminal repeats was identified on the basis of the RepBase database[46] for hg19.

ChIP–seq analysis. Antibody used for ChIP included anti-H3K27ac (2 mg for 25 mg of chromatin; ab4729, Abcam). A minimum of 2 million purified human CLL cells were used. In brief, cells were fixed in a 1% methanol-free formaldehyde solution and then resuspended in sodium dodecyl sulfate (SDS) lysis buffer. Lysates were sonicated in an E220 focused-ultrasonicator (Covaris) to a desired fragment size distribution of 100–500 bp. ChIP assays were processed on a SX-8G IP-STAR Compact Automated System (Diagenode) using a direct ChIP protocol[47]. In brief, immunoprecipitation reactions were performed with the above-indicated antibody, each on approximately 500,000 cells, and incubated overnight at 4 °C. The immune complex was collected with protein A/G agarose or magnetic beads and washed sequentially in the low-salt wash buffer (20 mM Tris pH 8, 150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA), the high-salt wash buffer (20 mM Tris pH 8, 500 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA), the LiCl wash buffer (10 mM Tris pH 8, 250 mM LiCl, 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA) and Tris-EDTA buffer. Chromatin was eluted with elution buffer (1% SDS, 0.1 M $NaHCO_3$), and then reverse cross-linked with 0.2 M NaCl at 65 °C for 4 h. DNA fragments were purified using Agencourt AMPure XP beads (Beckman Coulter). Barcoded immunoprecipitated DNA and input DNA were prepared using the NEBNext ChIP-seq Library Prep Master Mix Set for Illumina (E6240, New England Biolabs) and TruSeq Adaptors (Illumina) according to the manufacturer's protocol on a SX-8G IP-STAR Compact Automated System (Diagenode). Phusion High-Fidelity DNA Polymerase (New England Biolabs) and TruSeq PCR Primers (Illumina) were used to amplify the libraries, which were then purified to remove adaptor dimers using AMPure XP beads and multiplexed on the HiSeq 2000 (Illumina). ChIP–seq data were processed according to the ENCODE Histone ChIP-seq Data Standards and Processing Pipeline (https://www.encodeproject.org/chip-seq/histone/). Raw reads were mapped to the human genome hg19 assembly using Burrows–Wheeler Aligner[48] (BWA v.0.7.17). Duplicate reads were removed using Picard (https://broadinstitute.github.io/picard/). Peaks were identified with MACS2[49] (v.2.0.10) with a q value threshold of 0.01. Peaks overlapping with satellite repeat regions and Encode blacklist were discarded.

Single-cell DNA methylation–gene expression correlation analysis. For each sample, we filtered out poor quality cells when the number of detected CpGs was below 50,000, the number of detected genes in the transcriptomes was below 2,000 or the fraction of mitochondrial or ribosomal gene counts was higher than 20% of the library size (total number of read counts). We randomly downsampled the vector of RNA read counts per cell such that the total number of read counts equated to the bottom quartile of the library size distribution for all cells in the sample (cells below this threshold were dropped). Mitochondrial genes, genes encoding ribosomal proteins, and genes with RNA-seq expression in less than 5 cells were then removed from the analysis. At single-cell resolution, the methylation rate of a gene promoter was represented by the proportion of methylated CpGs in the region 1 kb upstream/downstream of the transcription start site. Genes with fewer than 5 CpG observations in the promoter region were excluded. Spearman's rank correlation coefficient between expression and promoter methylation rate was then calculated across available cells for each gene. The observed Spearman's rho was validated by a non-parametric permutation test, in which we compared the correlation of promoter DNAme with gene expression against a null distribution obtained by randomly permuting cell labels for the methylation values (such that RNA and DNAme are no longer linked at the single-cell level) and then computing the Spearman's rank correlation coefficient ($n = 26$ permutations for normal B sample (B04) and $n = 16$ permutations for CLL samples (CLL03 and CLL04) were used to obtain comparable numbers of genes between samples; see Fig. 2c, d; Extended Data Fig. 6f). We note that the same result was obtained when equalizing number of permutations ($n = 16$) and/or number of genes ($n = 2{,}500$) between samples in the analysis (see Extended Data Fig. 6g–n).

Single-cell transcriptional entropy analysis. Transcriptional entropy in Fig. 2b and Extended Data Fig. 6c, d was computed as previously described[18]. In brief, for a given cell we divided each element of the downsampled vector of gene expression counts by the cell's library size to obtain the corresponding proportion of overall expression attributable to each gene. These gene proportions were used to compute Shannon's information entropy for each cell using the standard formula:

$$S = -\Sigma_i P_i \ln(P_i)$$

Where S is Shannon's information entropy, and P_i is the proportion of overall expression attributable to gene i within that single cell. This value was subsequently scaled by the maximum obtainable entropy to bring each value into a 0–1 range. We note that the analyses in Fig. 2b and Extended Data Fig. 6d were performed with downsampling to create a balanced dataset by matching the total number of RNA read counts for all cells in each sample ($n = 50{,}000$ reads per cell).

Gene set enrichment analysis. Gene set enrichment analysis was limited to the Molecular Signature Database[50] (MSigDB; http://www.broad.mit.edu/gsea/) CGP (expression signatures of genetic and chemical perturbations) and CP (canonical pathways derived from KEGG, Reactome, and BIOCARTA) curated gene set collections. In Fig. 4d, genes with a Fisher's combined $P < 0.05$ and absolute $\log_2(T_0\text{-enriched}/T_1\text{-enriched gene expression}) > 0.5$ were used for the subsequent gene set enrichment analysis ($n = 336$). A hypergeometric test was used to measure the enrichment of these genes in each gene-set, followed by a Benjamini–Hochberg FDR procedure (cut-off of adjusted $P < 0.2$).

PDR analysis. Epimutation rates are quantified by assessing the concordance of adjacent CpGs within the same sequencing read (both methylated and unmethylated CpGs on a single sequencing read) and are measured with multiplexed scRRBS as the proportion of discordant reads per cell (single-cell PDR) as previously described[6], with minor modifications. In brief, at each CpG, PDR is equal to the number of discordant reads (reads containing both methylated and unmethylated sites) divided by the total number of reads. To calculate PDR for each individual cell, all reads with greater than four CpGs were evaluated for discordance, and the sum of discordant reads was divided by total number of reads with greater than four CpGs within that cell. To determine region-specific PDR, each cell's reads were intersected with the genomic coordinates of the region of interest before PDR calculation. To compute cell-to-cell PDR differences, pairs of cells were randomly sampled without replacement and the absolute difference between the two cells was measured. This procedure was repeated until all pairs of cells within a sample were exhausted. We note that for the analyses in Fig. 1d, e and Extended Data Fig. 3b, we excluded 175 cells (6.5%) with a bisulfite conversion rate <0.99, to remove incomplete conversion as a technical source of epimutation, from the total of 1,721 cells profiled with stand-alone multiplexed scRRBS (see Extended Data Fig. 1b). In addition, we also excluded cells from sample B03, as these are $CD19^+CD27^-$ index-sorted B cells.

To exclude technical artefacts as a potential cause of lower PDR dispersion in CLL compared with normal B cells, a multivariable generalized linear model regression analysis was performed, confirming that the observed low cell-to-cell epimutation rate variability was strongly associated with CLL versus normal B cell status. Cell-to-cell PDR difference was used as dependent variable. Number of unique CpGs, bisulfite conversation rate, number of reads, and cell type status (CLL versus normal B cells) were used as explanatory variables. P values for the generalized linear model coefficients (Student's t-test) of less than 0.05 were considered significant (Extended Data Fig. 3e).

Concordance odds ratio analysis. We present a CpG auto-correlation metric known as the concordance odds ratio (COR). CpG observation (CpG_a) is considered concordant with another CpG observation (CpG_b) at genomic base pair distance, d, away if both CpG_a and CpG_b are methylated, or both are unmethylated, otherwise they are labelled as discordant. The COR at each base pair distance d is the quotient between the concordance empirical likelihood at d and the background concordance empirical likelihood. For a given distance d, all pairs of CpGs covered in a single cell i that are separated by d base pairs are obtained. The COR for distance d in a given single cell i is then computed by measuring the

ratio of concordant pairs separated by distance d out of all pairs of CpGs that are at a distance d and dividing it by the expected background ratio of concordance determined by average methylation in the given genomic region of interest in cell i (for example, CGI, see formula in Extended Data Fig. 4a). This provides a vector for cell i of COR values as a function of d, in the range of 100 bp (that is, beyond the length of a single sequencing read) to 1,000 bp for the region of interest. Owing to differences in length of the assessed genomic regions of interest, we corrected for the length of these genomic regions by dividing each genomic location into equal-sized bins and averaging the COR values within each bin. For visualization clarity, COR values were subsequently scaled to bring all values into the range of 0–1. We then fitted a linear curve to this vector of COR by d and computed the slope as a measure of concordance decay for each independent cell. All cells belonging to CLL01–CLL12 and B01–B06 samples profiled with multiplexed scRRBS were used in the analysis. Finally, P values were computed for two-tailed Mann–Whitney U-test by comparing the average rate of decay in COR of healthy donor samples ($n = 6$) with the average rate of decay in COR of CLL samples ($n = 12$), to test whether CLL samples lose DNA methylation concordance at a different rate compared with healthy donor samples.

Four-gamete analysis. We present a CpG epimutation metric based on the four-gamete test[11]. We will refer to this metric as four gametes. This test relies on the fact that detecting four gametes defies the assumptions of the infinite site mutation model[51] and therefore is likely to reflect a high epimutation rate. Moreover, this test allows us to estimate epimutation rate at single CpG site resolution in CpG-sparse regions, such as enhancers, in contrast to methods that rely on capturing multiple CpGs on the same read[6,7]. For each sample (samples CLL01–CLL12 were used in the analysis), the number of gametes between two CpGs, CpG_a, and CpG_b, was determined by counting how many of the four possible combinations of methylation and unmethylation were observed across all cells in a given sample where both CpG_a and CpG_b were obtained. This process was repeated by pairing each individual CpG_a with all CpGs further than 100 bp away (to exclude CpGs contained within a single sequencing read) and enumerated the number of gametes observed in each pair of sites in all cells. A binary mask was applied to the resulting counts to exclude the pairing of a site with itself. After all pairings, as a measure of CpG epimutation, we computed the frequency of observing four gametes at CpG_a by dividing the number of observed pairs with four gametes by the total number of pairings. As the direct implementation of such an algorithm has time complexity of $O(m \times n^2)$, in which m is the number of cells and n is the number of sites, the number of pairings analysed for each CpG was randomly downsampled by a 100× factor to speed up the calculation. To validate this approach, five runs with random 100-fold downsampling were performed for the same dataset and the frequencies of observing four gametes were compared. The results were highly concordant (Pearson correlation coefficient $r = 0.93$), supporting the validity of this approach. Notably, by pairing individual CpGs to all other CpGs across the genome, the four-gamete test enabled the determination of epimutation rate even for CpGs that are not in close genomic proximity to other CpGs, which is required for methods such as PDR and epigenetic polymorphism for calculation of epimutation[6,7]. We note that the assumption of independence between CpGs in the four-gamete test is probably valid here, as multiplexed scRRBS captures approximately 10% of the targeted methylome per single cell owing to the sparsity of the single-cell data. Therefore, the four-gamete test is based on a nearly unique combination of CpGs per cells for each CpG pairing. Only CpG sites covered by at least five cells in each sample were used in the analysis (range 156,662–2,371,498 CpGs per sample). Within each sample (CLL01–CLL12), CpG sites with lower four-gamete rate than expected based on their methylation level (that is, low epimutation CpGs) were defined as being 1.5× median absolute deviation away from the median frequency of four gametes in each DNAme window size of 0.05 (from 0.1 to 0.9). A total of 166,720 unique CpGs across all the 12 CLL patient samples (average of 1.22% ± 0.42 (mean ± s.e.m.)); range 0.04–2.9%) exhibited a lower frequency of four gametes than expected based on their DNAme level and were used for downstream analyses.

BEDTools[52] v.2.25.0 was used to calculate overlaps between low epimutation CpGs and gene promoters or FANTOM5 human robust enhancers[44]. De novo motif enrichment analyses were performed using MEME-ChIP[53] against JASPAR CORE vertebrates and UniPROBE Mouse databases (-order 2, -meme-minw 6, -meme-maxw 15, -meme-nmotifs 5, -dreme-e 0.05, -meme-mod zoops). Specifically, we performed a discriminative motif discovery to find motifs within gene promoters or enhancers that were overrepresented at sites surrounding low epimutation CpGs (±25 bp around CpG) relative to a control set consisting of randomly selected CpGs (±25 bp around CpG), matched for methylation values and cell coverage to the low epimutation CpGs. To control for possible CpG content biases further (for example, as MspI cut site is C^CGG), a two-order background model was used to normalize for biased distribution of trimer nucleotides in our sequences. Only motifs with an $E \leq 0.05$ were reported, and each motif was then matched to its most similar motif in the TOMTOM database[54] or literature if available. The E value is an estimate of the expected number of motifs with the given log-likelihood ratio (or higher), and with the same width and site count, that one would find in a similarly sized set of random sequences[53]. We also report the TOMTOM P value, defined as probability that a random motif of the same width as the target would have an optimal alignment with a match score as good as or better than the target[53].

Lineage tree inference and support values. Because epimutations mark cell divisions[9], the heritable DNAme information captured by multiplexed scRRBS can inform the reconstruction of cellular lineages. Given that the maintenance methylation machinery has an error rate estimated to be four orders of magnitude higher than that observed for DNA replication[55,56], the phylogenetic information content of single-cell DNAme data are higher than that of single-cell nucleotide variants. Moreover, although single-cell copy number variations[57,58], IGH transcript sequences[59], somatic microsatellite[21] and mitochondrial DNA[60,61] mutations allow for the reconstruction of cancer lineages, they may have limited resolution given the smaller number of events that can be detected with current single-cell sequencing approaches, limited applicability across cancer types, or have not been adapted for large scale multi-modality single-cell sequencing. Specifically, reconstruction of cancer lineages from copy number aberrations is not applicable to near-euploid cancers, such as CLL. We therefore generated methylation-based lineage trees by applying a tree-searching maximum-likelihood algorithm based on binary methylation values. We used the MPI version of IQ-TREE[62] v.1.5.3, which exhibits improved performance compared to other maximum-likelihood fast phylogenetic programs in identifying trees of higher likelihood scores[63]. We selected a substitution model based on the binary alignment, inferred a maximum-likelihood tree, and computed bootstrap support values (1,000 bootstrap replicates). We opted for the new model selection procedure[64] (-m TESTNEW), which additionally implements the FreeRate heterogeneity model inferring the site rates directly from the data (mixture of four gametes and technical errors permitted in phylogeny reconstruction) instead of being drawn from a gamma distribution[65]. General time reversible model 'GTR2' consistently outperformed the other model tested (Jukes–Cantor type model) for our methylation binary data. IQ-TREE also incorporates an approach for calculating ultrafast bootstraps (UFBoot)[66]. We complemented UFBoot analysis with the Shimodaira–Hasegawa-like (SH-like) approximate likelihood ratio test (SH-aLRT) and the approximate Bayes test to further assess support for single branches. In brief, we initialized different tree search runs per batch of cells, each with a different random starting seed. In each run, a maximum-parsimony tree is first constructed directly from the alignment (methylation state mismatches between cells). Then, parameters of the given binary substitution models are estimated based on the maximum-parsimony tree. The log-likelihoods of this initial maximum-parsimony tree are computed for the many different given models along with the Akaike information criterion, corrected Akaike information criterion, and the Bayesian information criterion. The model that minimizes the Bayesian information criterion score (the best-fit model) is then selected. The estimated model parameters are now used for initializing candidate tree set and further maximum-likelihood optimizations using an iterative, 'hill-climbing' optimization technique. Maximum-likelihood tree search starts by generating 100 trees. From these 100 trees, all unique topologies are collected, and their approximate likelihoods computed. From the ranked list of maximum-likelihood values, the top 20 trees are selected and NNI are performed on each tree to obtain the locally optimal maximum-likelihood trees. The top five topologies with highest likelihood (the candidate tree set) are then retained for further maximum-likelihood optimizations. An important weakness of pure hill-climbing methods is that they can be easily trapped in local optima. The locally optimal trees in the candidate tree set are, thus, randomly perturbed to allow escape from local optima. IQ-TREE keeps the best maximum-likelihood tree while it searches the tree parameter space and stops searching after going through a user-defined number of trees. We extended this number to 1,000 trees to better explore tree parameter space. The final optimized best maximum-likelihood tree is then printed in NEWICK format. Trees were visualized with FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Lineage tree structures were validated by cross-validation by restricting phylogeny reconstruction to only autosomes or chromosome X, holding-out chromosomes (three at a time), or downsampling the number of CpGs per cell to equal numbers, confirming the robustness of the lineage tree inference (Extended Data Fig. 7). The inferred lineage trees were also found to be >3-fold more robust than maximum-parsimony-based reconstruction trees (Extended Data Fig. 7e), confirming that the lineage tree structure adds new information to the simple comparison of the DNAme profiles. The haploid X chromosome in male patient samples showed an even greater robustness when compared with maximum-parsimony trees, probably owing to the removal of the confounding random sampling of the two alleles in autosomes.

Methylation-based lineage trees integrating cells before treatment (T_0) and during treatment (T_1) for patient samples CLL03, CLL04, CLL05 and CLL11 from

joint multiplexed scRRBS and single-cell RNA-seq were reconstructed by maximum-likelihood, followed by ultra-fast bootstrapping branch support analysis with 1,000 replicates (Fig. 4b; Extended Data Fig. 9a). T_1-enriched clades were defined based on clades occurring after the first major split in the lineage tree. Differential expression was compared between T_1 cells that map to the T_1-enriched clades and T_1 cells that map to the T_0-enriched clades. We further matched the cells belonging to the T_1-enriched clade identified from these T_0–T_1 lineage trees, by integrating the two groups of T_1 cells into a maximum-likelihood tree search and computing bootstrapping branch support analysis with 1,000 replicates, as described above. In Extended Data Figs. 8e, 9d, we defined genes with an absolute weighted average DNAme difference > 0.3 and a two-sided non-parametric permutation test $P < 0.05$ as differentially methylated.

Maximum tree depths—defined as number of nodes along the longest path from the root node down to the farthest leaf—of lineage trees of CLL and normal B cells were computed by initializing ten independent tree search replicates per batch of randomly sampled 50 cells, each with a different random starting seed. Patristic distances—defined as the sum of the lengths of the branches that link two tips in a given tree—between CLL and normal B cells were computed by analysing one representative methylation-based lineage tree of randomly sampled cells for each sample. To compare between inferred lineage trees, we computed the pairwise Robinson–Foulds distance—a measure of tree structure similarity between two given trees[67]— between them. Specifically, 30 independent tree search replicates per batch of randomly sampled 50 cells were initialized, each with a different random starting seed. To compute the Robinson–Foulds distances, pairs of trees were then randomly sampled without replacement and the Robinson–Foulds distance between the two trees computed. The Robinson–Foulds distances were normalized by the total number of internal edges in respective pairs of trees (normalized Robinson–Foulds distance). Node ages—the estimated number of divisions before present—were calculated by dividing node height (defined as the length of the longest downward path to a leaf from a given node) values by a rate of 0.0005 changes per CpG site per division[29].

Statistical methods. Statistical analysis was performed with Python 2.7.13 and R version 3.4.2. Categorical variables were compared using the Fisher's exact test. Continuous variables were compared using the Mann–Whitney *U*-test, Welch's *t*-test, Wilcoxon signed-rank test, non-parametric permutation test or Kolmogorov–Smirnov test as appropriate. *P* values were adjusted for multiple comparisons by Bonferroni family-wise error rate or Benjamini–Hochberg FDR adjustment procedure, as appropriate. All *P* values are two-sided and considered significant at the 0.05 level unless otherwise noted.

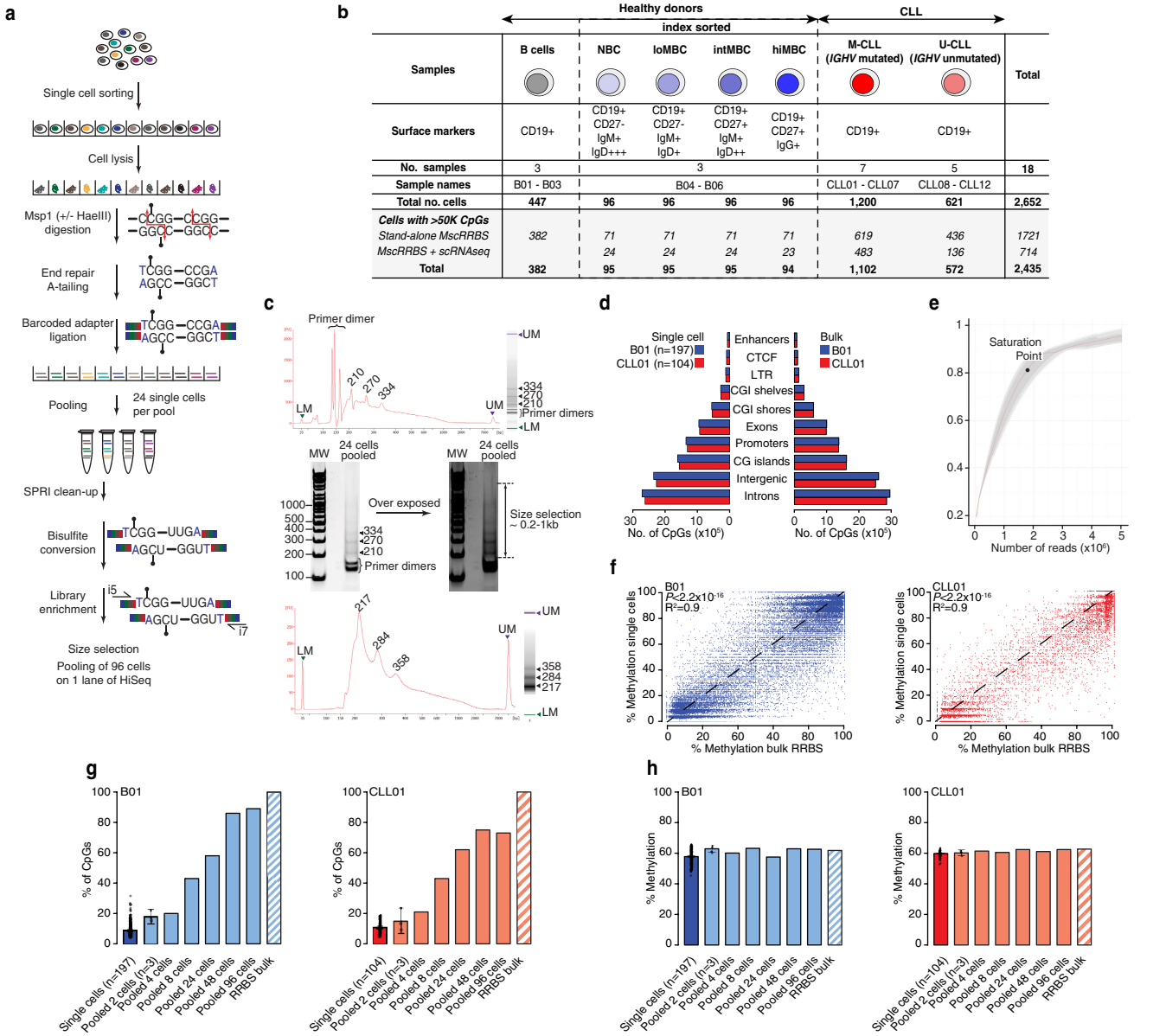
No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

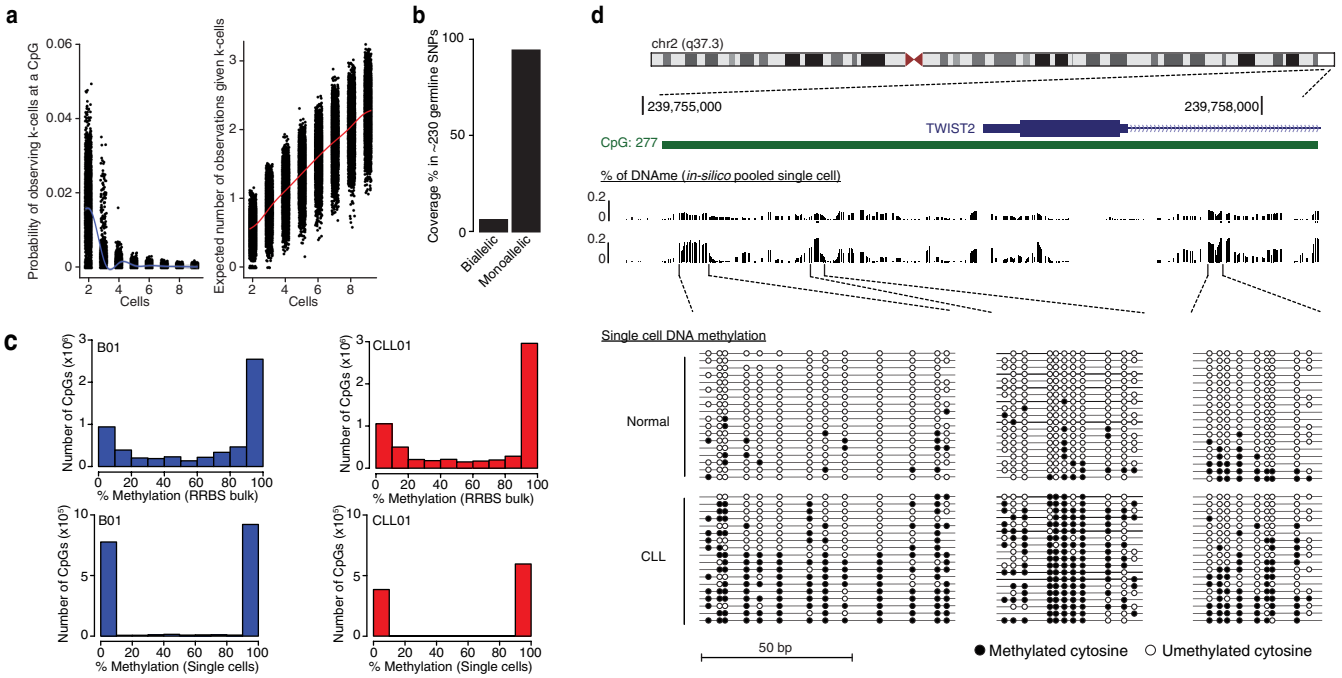
Data availability

Multiplexed scRRBS and single-cell Smart-seq2 datasets have been deposited to the NCBI Gene Expression Omnibus (GEO) under accession number GSE109085. ChIP–seq datasets have been deposited to the NCBI GEO under accession number GSE119103. Other data are available from the corresponding author upon reasonable request.

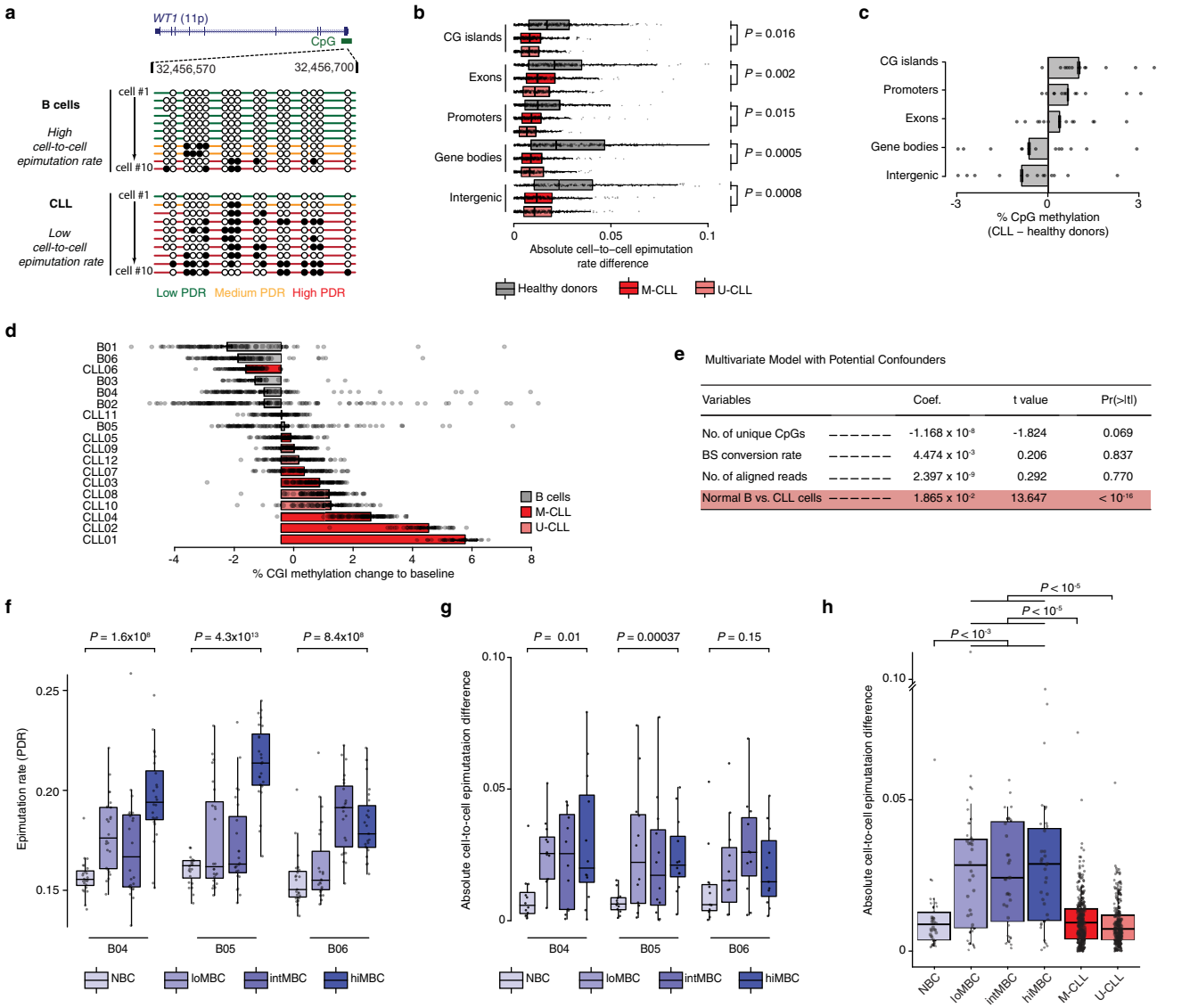
31. Rassenti, L. Z. et al. Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia. *Blood* **112**, 1923–1930 (2008).
32. Guo, H. et al. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat. Protocols* **10**, 645–659 (2015).
33. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
34. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
36. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
37. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
38. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
39. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
40. Van den Berge, K. et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **19**, 24 (2018).
41. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
42. Van den Berge, K., Soneson, C., Robinson, M. D. & Clement, L. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biol.* **18**, 151 (2017).
43. Illingworth, R. S. et al. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134 (2010).
44. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
45. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
46. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
47. O'Geen, H., Echipare, L. & Farnham, P. J. in *Epigenetics Protocols* (ed. Tollefsbol, T. O.) 265–286 (Humana Press, 2011).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
50. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
51. Tajima, F. Infinite-allele model and infinite-site model in population genetics. *J. Genet.* **75**, 27 (1996).
52. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
53. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
54. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
55. Ushijima, T. et al. Fidelity of the methylation pattern and its variation in the genome. *Genome Res.* **13**, 868–874 (2003).
56. Biezuner, T. et al. A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res.* **26**, 1588–1599 (2016).
57. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
58. Bian, S. et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060–1063 (2018).
59. de Bourcy, C. F. A. et al. Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc. Natl Acad. Sci. USA* **114**, 1105–1110 (2017).
60. Xu, J. et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife* **8**, e45105 (2019).
61. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339 (2019).
62. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
63. Zhou, X., Shen, X.-X., Hittinger, C. T. & Rokas, A. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol. Biol. Evol.* **35**, 486–503 (2018).
64. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
65. Soubrier, J. et al. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol.* **29**, 3345–3358 (2012).
66. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2017).
67. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
68. Raval, A. et al. TWIST2 demonstrates differential methylation in immunoglobulin variable heavy chain mutated and unmutated chronic lymphocytic leukemia. *J. Clin. Oncol.* **23**, 3877–3885 (2005).
69. Perez, C. A., Ott, J., Mays, D. J. & Pietenpol, J. A. p63 consensus DNA-binding site: identification, analysis and application into a p63MH algorithm. *Oncogene* **26**, 7363–7370 (2007).
70. Hsiao, L.-L. et al. A compendium of gene expression in normal human tissues. *Physiol. Genomics* **7**, 97–104 (2001).
71. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).



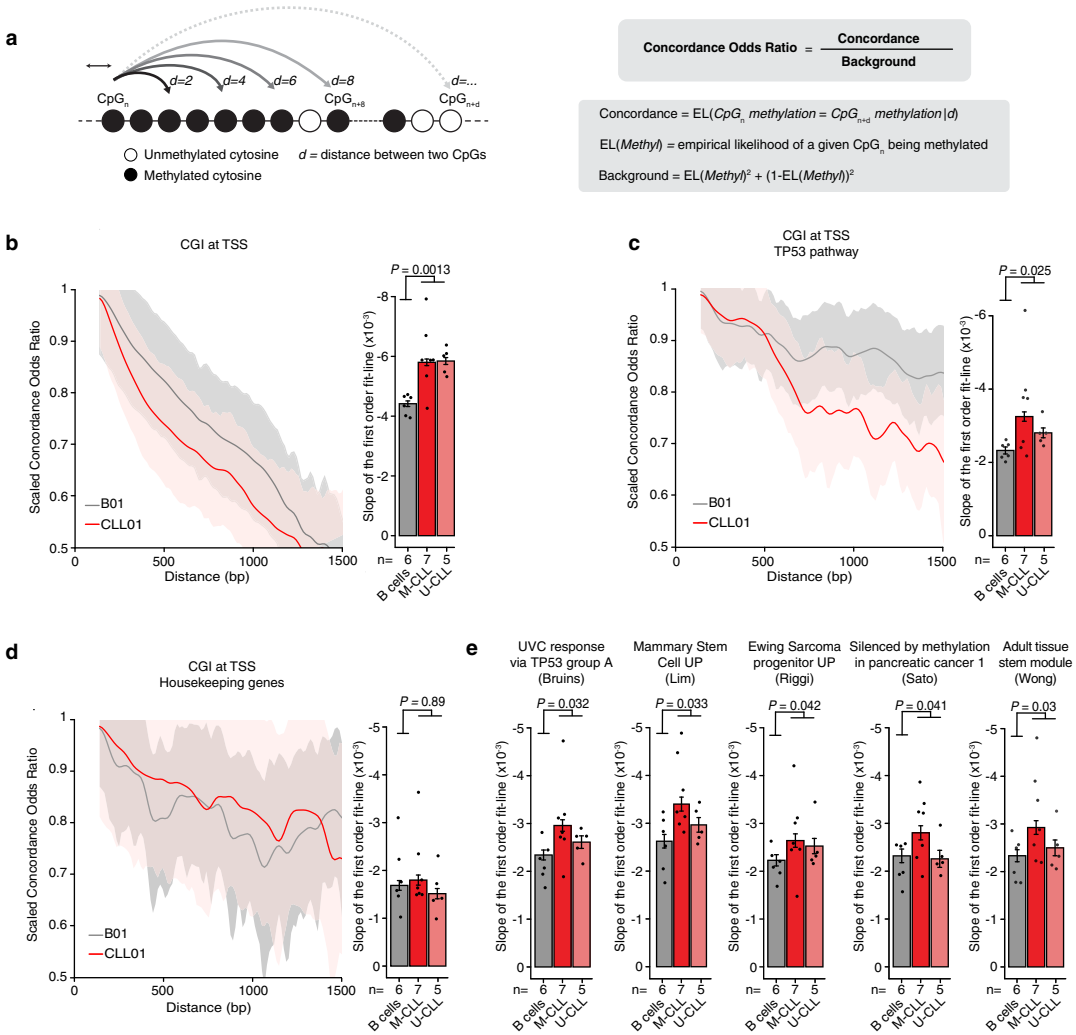
Extended Data Fig. 1 | Multiplexed scRRBS is an accurate and reproducible method for single-cell DNAme analysis. a, Detailed schematic of the protocol for multiplexed scRRBS. **b**, Summary table of the healthy donor and CLL patient samples used in this study. **c**, Representative size distribution of the multiplexed scRRBS libraries assessed by Agilent Bioanalyzer before and after primer dimers removal. The DNA fragment size in multiplexed scRRBS libraries is typically 200–1,000 bp, with some visible peaks corresponding to the MspI fragments for repeat elements, and primer dimer contaminants (approximately 170 bp). LM, lower marker; UM, upper marker; MW, molecular-weight size marker. **d**, Number of CpGs observed in multiplexed scRRBS libraries across relevant genomic regions comparing multiplexed scRRBS (left) and bulk RRBS (right) assays for normal B (B01) and CLL (CLL01) cells. The enrichment in exons, promoters and CpG islands (CGIs) observed in multiplexed scRRBS libraries corresponded to approximately 40% of the total sequenced CpGs, akin to bulk RRBS assays. **e**, Downsampling analysis showing that around 1.7 million paired-end reads per cell provided roughly 85% of unique CpGs with further sequencing, resulting in a marginal increase in coverage. **f**, Correlation of average CpG methylation across in silico merged single cells and bulk RRBS obtained from matched samples for normal B (B01, $n = 40{,}257$ CpGs) (left) and CLL (CLL01, $n = 9{,}578$ CpGs) (right) cells. *P* values are indicated for two-sided Pearson's correlation test. **g**, Pooling individual single cells together rapidly increases the number of CpGs recovered, approaching bulk RRBS coverage with more than 48 cells. The percentage of CpG sites detected in single-cell data (blue and red for normal B and CLL cells, respectively), the in vitro pooled single-cell datasets (light blue and light red, respectively) and matched bulk RRBS libraries (striped bars) are shown. Error bars represent 95% confidence interval. **h**, As in **g** for the percentage of average CpG DNA methylation. Single, pooled cells and bulk RRBS showed a similar percentage of CpG methylation, suggesting measured genome-wide DNAme profiles of individual cells accurately recapitulate bulk methylation profiles in the same cell type.



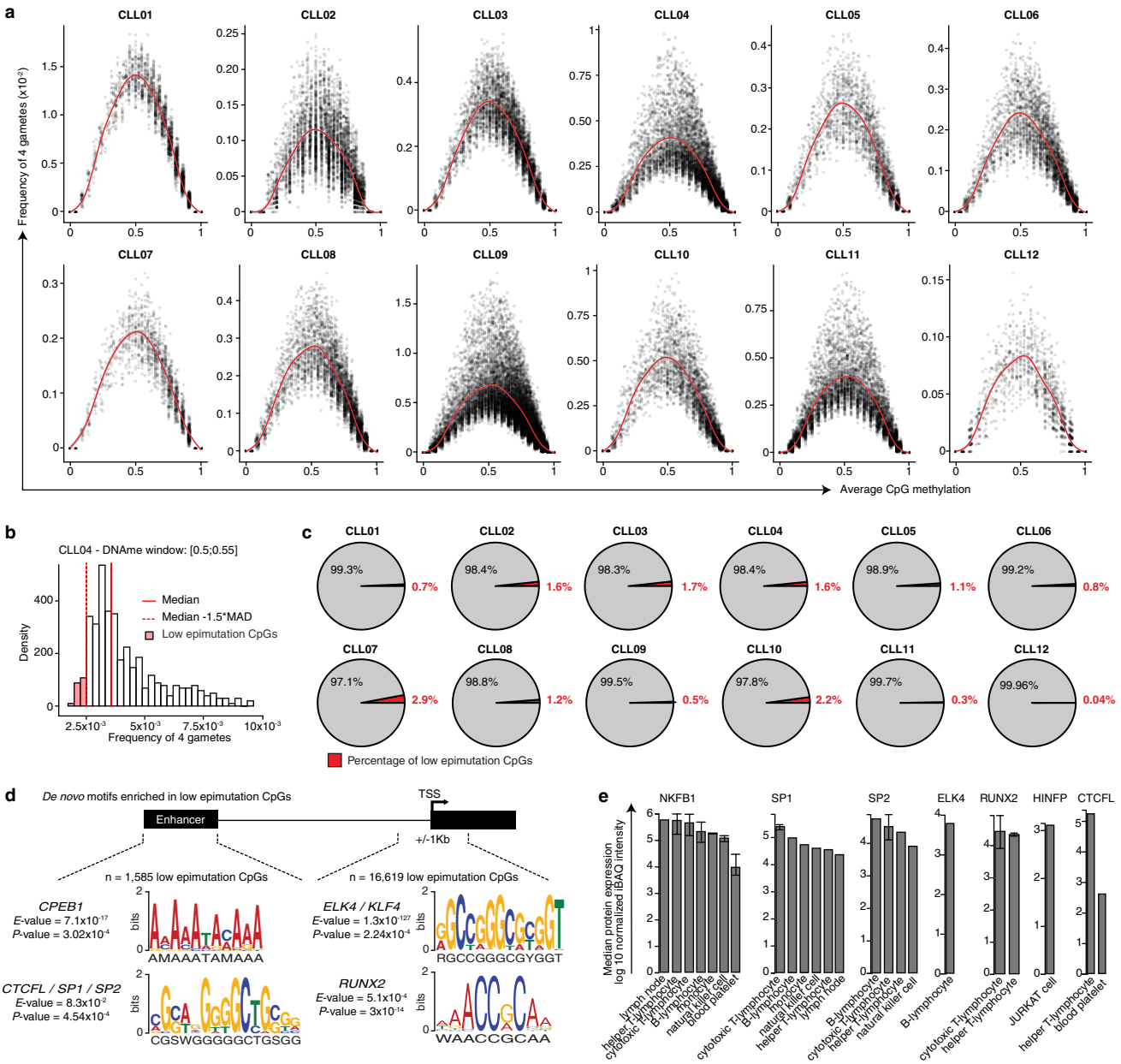
Extended Data Fig. 2 | Single-cell DNA methylation coverage analysis. **a**, The approximately 10% sampling of the multiplexed scRRBS DNA methylome leads to intersection decrease of individual CpGs across cells. Left, expected number of times of observing a given CpG across all *k*-cells (matching *k* (number of cells) indicated in the *x*-axis value). Right, expected number of measured CpGs given *k*-cells. **b**, Biallelic coverage within a given single cell was detected in only 4.6 ± 2% of approximately 230 germline single nucleotide polymorphisms (SNPs) available for analysis, suggesting that the observed single-cell CpG data largely represents only one of the two alleles of the near-diploid CLL genome. **c**, Histograms of the distribution of CpG methylation values for single normal B (blue) and CLL (red) cells and matched bulk RRBS libraries showing highly digitized patterns of DNAme in single cells (that is, CpG sites either methylated or unmethylated) in contrast to bulk RRBS, which shows intermediate DNAme values. **d**, Representative analysis for three non-contiguous genomic windows around the promoter region of *TWIST2*, previously shown to be implicated in CLL pathogenesis[68]. Shown from top to bottom are the annotation of the *TWIST2* promoter locus with CGI sites indicated (green); the estimated methylation rate of in silico pooled single cells for healthy donors and CLL; and the CpG methylation patterns (black circles: methylated; white circles: unmethylated) of single cells. Note the higher level of DNAme percentage in CLL compared with healthy donor cells at these selected regions.



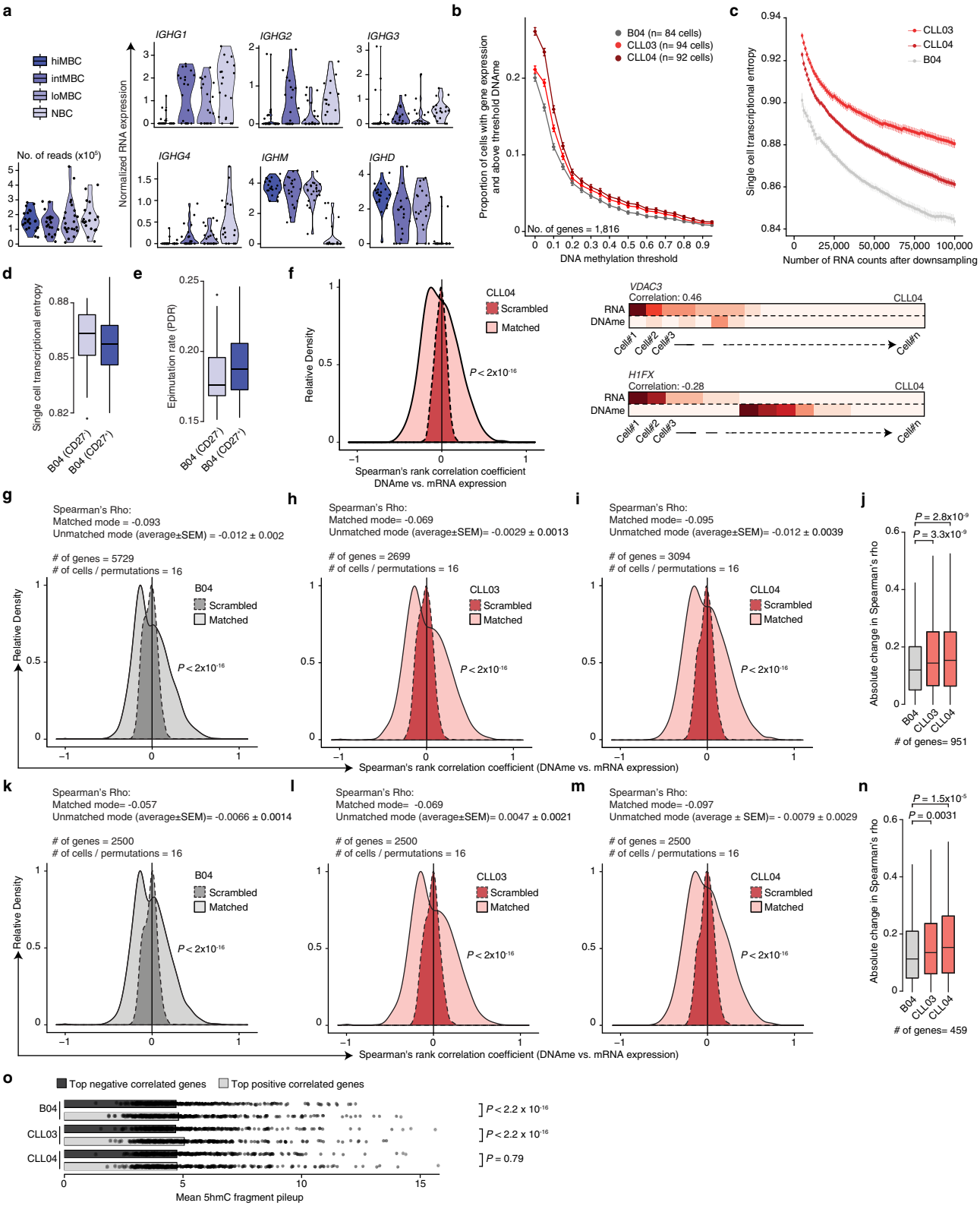
Extended Data Fig. 3 | CLL epigenomes show an increased epimutation rate across all genomic regions, with low cell-to-cell variability in epimutation rates. a, Representative analysis of the *WT1* locus. CpG island is indicated in green, along with the CpG methylation patterns (black circles denote methylated; white circles denote unmethylated) in single cells. We note that CLL cells exhibit lower cell-to-cell variation in the epimutation rate than normal B cells. **b**, Comparison of cell-to-cell epimutation rate difference per genomic region between CLL cells ($n = 12$; M-CLL (CLL01–CLL07), $n = 309$ pairs; U-CLL (CLL08–CLL12), $n = 218$ pairs) and healthy B cells ($n = 5$; B01–B02, B04–B06, $n = 256$ pairs). **c**, Difference in average CpG methylation per genomic region between CLL samples ($n = 12$; CLL01–CLL12 (M-CLL, $n = 619$ cells; U-CLL, $n = 436$ cells)) and normal B samples ($n = 6$; B01–B06 ($n = 666$ cells)). **d**, Percentage of change in CpG methylation at CGIs when comparing the DNAme level of individual cells in each sample to the baseline (defined as the average DNAme level across all samples) for CLL cells ($n = 12$; CLL01–CLL12 (M-CLL, $n = 619$ cells; U-CLL, $n = 436$ cells)) and normal B cells ($n = 6$; B01–B06 ($n = 666$ cells)). **e**, Multivariable linear regression model that accounts for potential technical confounders (bisulfite conversion rate, number of aligned reads, number of covered CpGs) in CLL samples ($n = 12$; CLL01–CLL12 (M-CLL, $n = 619$ cells; U-CLL, $n = 436$ cells)) and normal B samples ($n = 6$; B01–B06 ($n = 666$ cells)). **f**, Single-cell epimutation rate across index-sorted normal B cells (B04, $n = 96$ cells; B05, $n = 96$ cells; B06 = 92 cells). **g**, As in **f** for the difference in cell-to-cell epimutation rate (B04, $n = 48$ pairs; B05, $n = 48$ pairs; B06 = 46 pairs). **h**, Direct comparison of difference in cell-to-cell epimutation rate between CLL cells ($n = 12$; M-CLL (CLL01–CLL07)), $n = 309$ pairs; U-CLL (CLL08–CLL12), $n = 218$ pairs) and index-sorted B cells ($n = 3$; B04–B06; NBC, $n = 35$ pairs; loMBC, $n = 35$ pairs; intMBC, $n = 35$ pairs; hiMBC, $n = 35$ pairs). Box plots are as defined in Fig. 1. Error bars represent 95% confidence interval. *P* values were determined by two-sided Mann–Whitney *U*-test (**b**, **f**–**h**), followed by a Bonferroni adjustment procedure (**b**).



Extended Data Fig. 4 | Long-range DNA methylation concordance decay. a, Concordance odds ratio (COR) of the DNA methylation state between any two neighbouring CpGs as a function of their genomic distance (see Methods for details). **b**, Left, scaled COR (0–1) for CGIs at transcription start sites (TSS) (the B01 and CLL01 samples are shown as representative examples). Right, average rate of decay (slope of the first order fit line) in the COR for normal B samples ($n = 6$) and CLL samples ($n = 12$) for CGIs at TSS (B01–B06 ($n = 666$ cells; $n = 48{,}065{,}000$ CpGs) and CLL01–CLL12 (M-CLL, $n = 619$ cells, $n = 38{,}968{,}846$ CpGs; U-CLL, $n = 436$ cells, $n = 37{,}464{,}310$ CpGs)). **c**, As in **b** for CGIs at TSS of genes belonging to the *TP53* gene set[69]. Healthy donor B cell samples ($n = 6$): $n = 666$ cells, $n = 6{,}308{,}174$ CpGs; CLL samples ($n = 12$): M-CLL, $n = 619$ cells, $n = 5{,}113{,}493$ CpGs; U-CLL, $n = 436$ cells, $n = 4{,}982{,}039$ CpGs. **d**, As in **b** for CGIs at the TSS of housekeeping genes[70]. Healthy donor B cell samples ($n = 6$): $n = 666$ cells, $n = 2{,}087{,}432$ CpGs; CLL samples ($n = 12$): M-CLL, $n = 619$ cells, $n = 1{,}686{,}295$ CpGs; U-CLL, $n = 436$ cells, $n = 1{,}620{,}802$ CpGs. **e**, Average rate of decay in the COR for normal B ($n = 6$) and CLL ($n = 12$) samples for CGIs at the TSS of genes belonging to gene sets previously reported to be affected by a high epimutation rate[6]. Healthy donor B cell samples ($n = 6$): $n = 666$ cells, $n = 48{,}065{,}000$ CpGs; CLL samples ($n = 12$): M-CLL, $n = 619$ cells, $n = 38{,}968{,}846$ CpGs; U-CLL, $n = 436$ cells, $n = 37{,}464{,}310$ CpGs. Error bars represent 95% confidence interval. *P* values were determined by two-sided Mann–Whitney *U*-test.

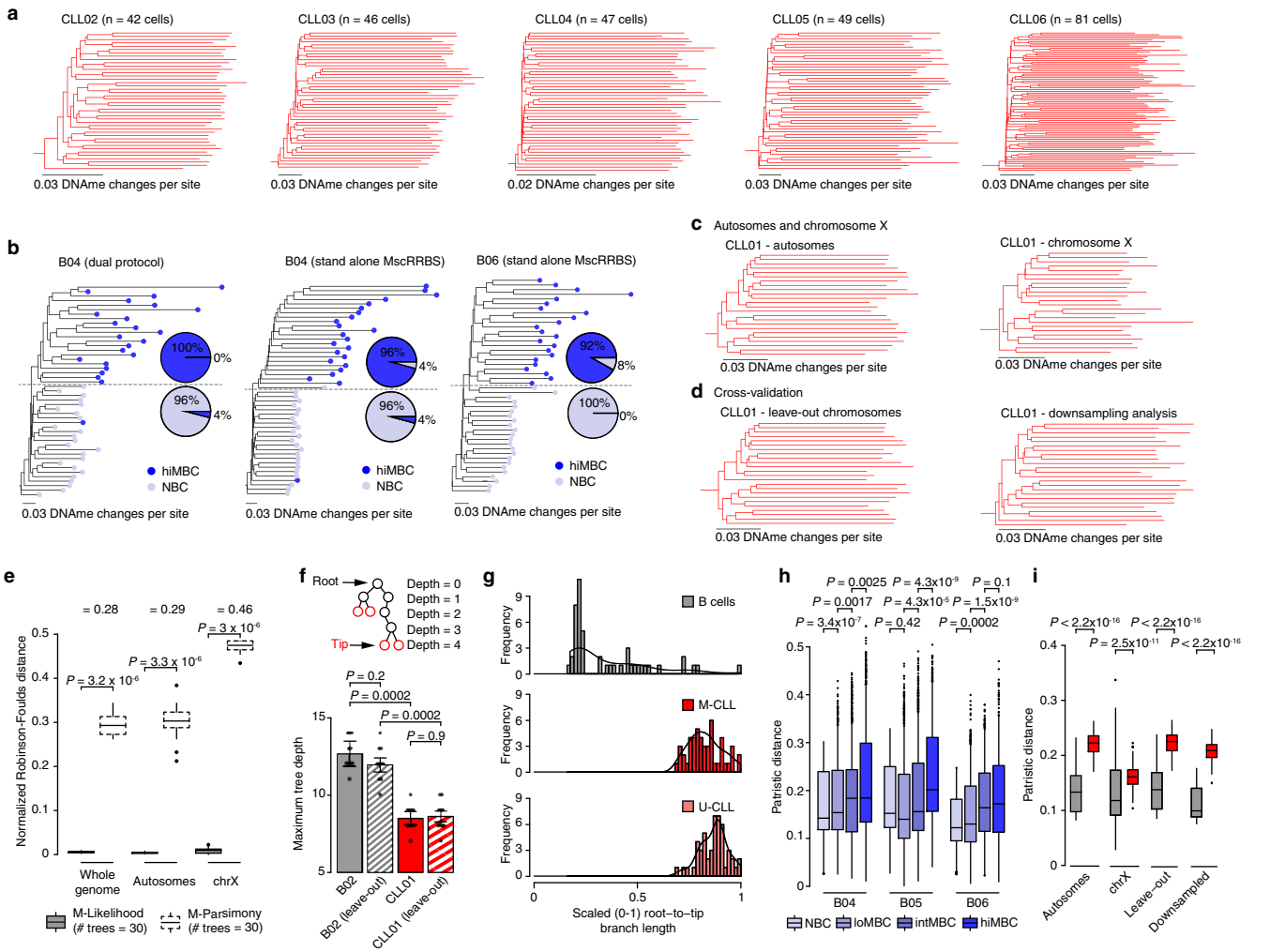


Extended Data Fig. 5 | Epimutations at single CpG resolution. **a**, Frequency of four gametes according to the level of average methylation of each individual CpG site in each CLL sample (CLL01–CLL12; randomly sampled CpGs shown out of the total CpGs assessed in each CLL sample; range 156,662–2,371,498 CpGs per sample covered in >5 cells in each sample). Smooth local regression line (LOESS) is shown in red. **b**, Low epimutation (loEpi) CpGs are defined as being 1.5× median absolute deviation (MAD) away from the median frequency of four gametes in each DNAme window of 0.05 (range 0.1–0.9) for a given sample. Shown is a representative example of this procedure for DNAme window of 0.5–0.55 in the CLL04 patient sample. **c**, Percentage of low epimutation CpGs (average of 1.22% ± 0.42 (mean ± s.e.m.); range 0.04–2.9%) out of the total CpGs assessed in each CLL sample. CLL01, $n = 14{,}711$ loEpi CpGs; CLL02, $n = 2{,}573$ loEpi CpGs; CLL013, $n = 25{,}270$ loEpi CpGs; CLL04, $n = 29{,}114$ loEpi CpGs; CLL05, $n = 16{,}603$ loEpi CpGs; CLL06, $n = 11{,}413$ loEpi CpGs; CLL07, $n = 19{,}330$ loEpi CpGs; CLL08, $n = 19{,}916$ loEpi CpGs; CLL09, $n = 11{,}440$ loEpi CpGs; CLL10, $n = 18{,}614$ loEpi CpGs; CLL11, $n = 7{,}067$ loEpi CpGs; CLL12, $n = 308$ loEpi CpGs. **d**, Additional sequence logos of the DNA motifs determined to be significantly overrepresented in low epimutation CpGs (±25 bp around CpGs at promoters (TSS ± 1 kb) or at enhancers) across all CLL samples. For each motif, the *E* value and the TOMTOM *P* value are shown. See Methods for details on the de novo motif enrichment analysis and the statistical tests used. **e**, Median protein expression ($\log_{10}$(normalized intensity-based absolute quantification (iBAQ))) of transcription factors for which motifs were enriched in regions with low epimutation CpGs, confirming that the identified transcription factors are expressed at the protein level in B cells and/or haematopoietic compartments. Error bars represent 95% confidence interval. All available human proteome data from lymphoid/haematopoietic lineages are displayed[71].

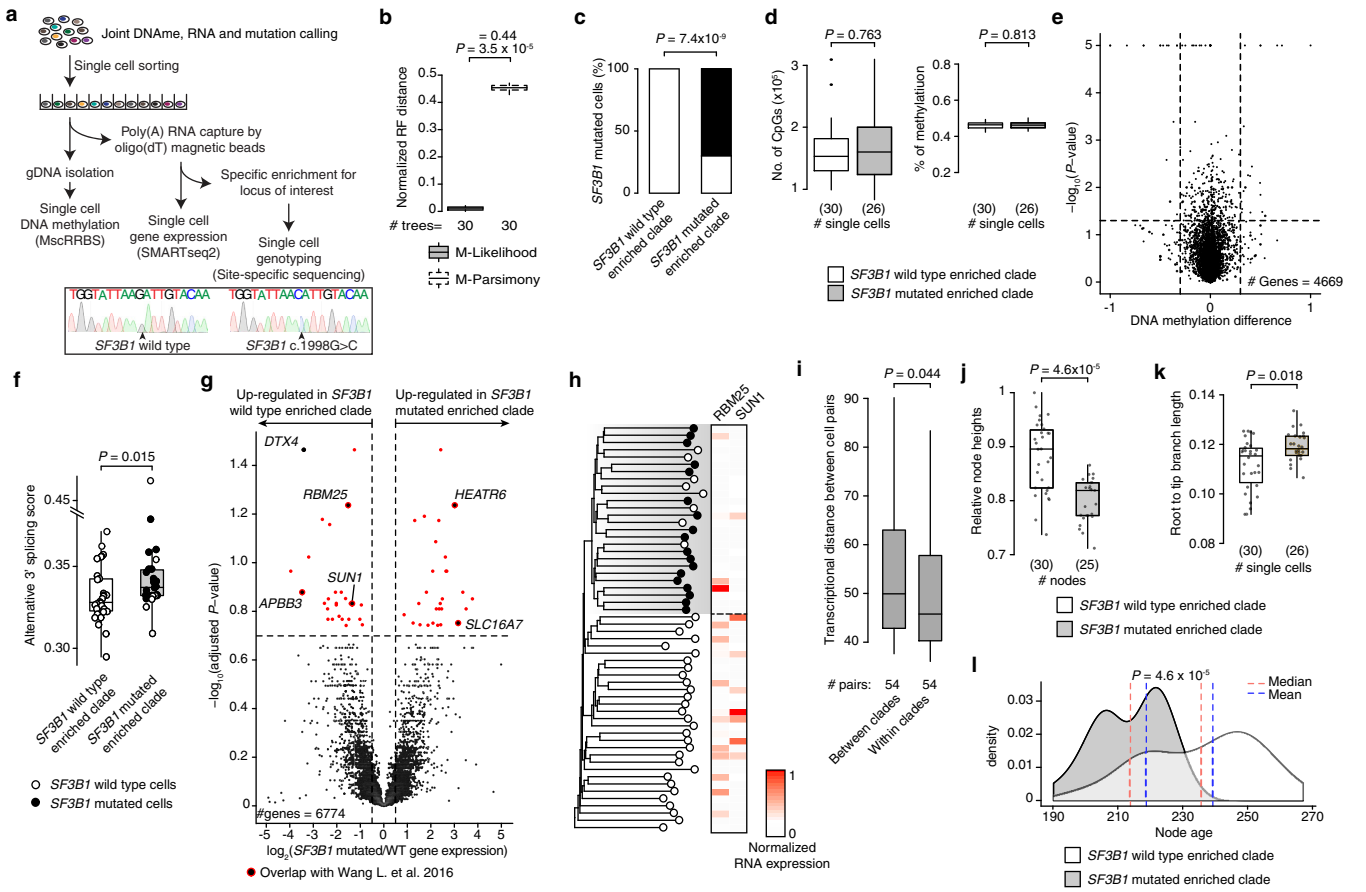


Extended Data Fig. 6 | See next page for caption.

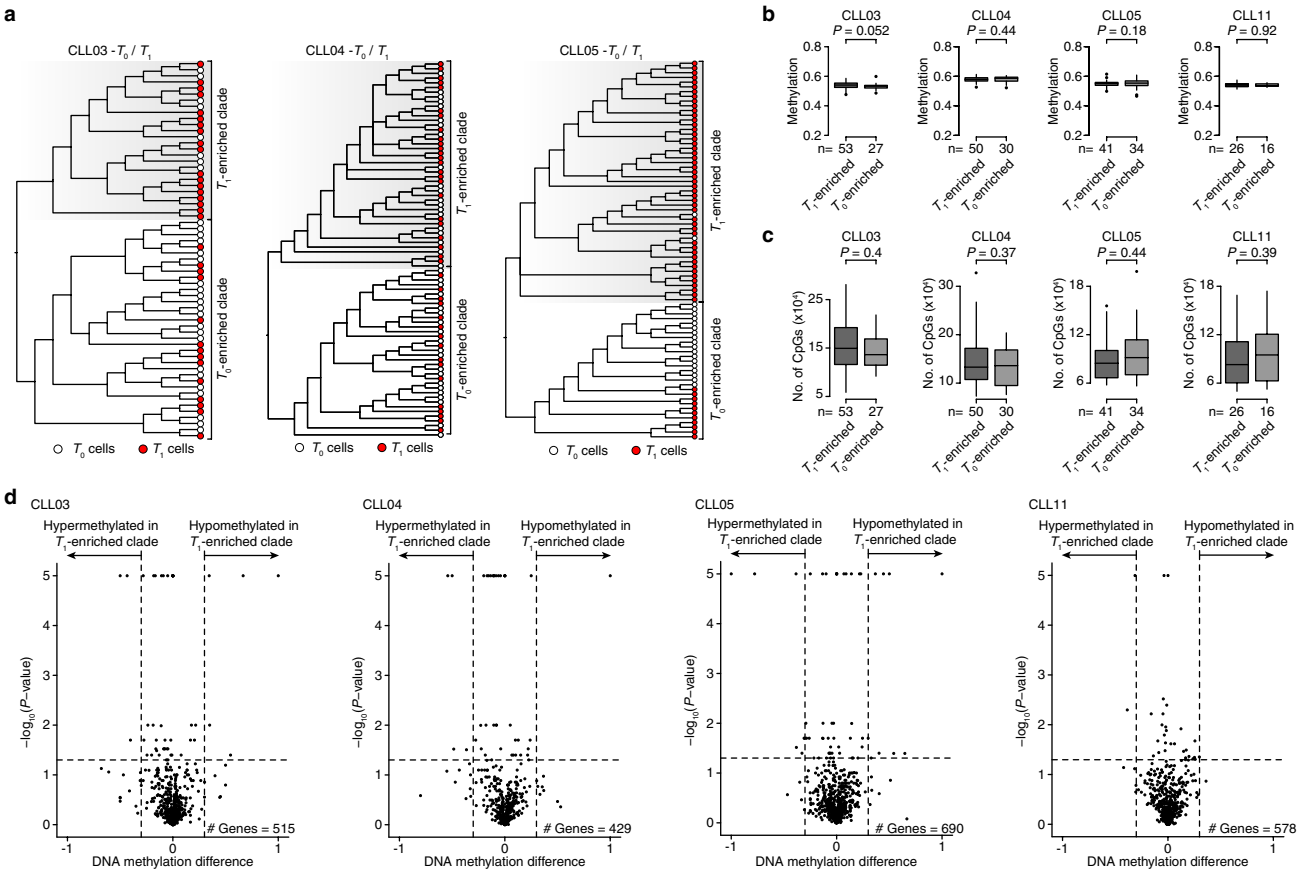
Extended Data Fig. 6 | Methylation-transcription relationships at the single-cell level. a, Number of reads (left) and expression of *IGH* genes (right) in index-sorted B cells, validating our index-sorting strategy ($CD27^{-}IgM^{+}IgD^{+++}IgG^{-}$ (NBC, $n = 24$ cells), $CD27^{-}IgM^{+}IgD^{+}IgG^{-}$ (loMBC, $n = 24$ cells), $CD27^{+}IgM^{+}IgD^{++}IgG^{-}$ (intMBC, $n = 24$ cells), and $CD27^{+}IgG^{+}$ (hiMBC, $n = 23$ cells)). Violin plots represent kernel density estimation showing the distribution shape of the data. **b**, Proportion of cells with gene expression (read count > 0) and exhibiting above-threshold DNAme. Data are mean ± s.e.m. across all genes with sufficient RNA (expression seen in >5 cells) and DNAme (>5 CpGs per promoter) information across the three samples ($n = 1{,}816$ genes). **c**, Transcriptional entropy across cells (see Methods) showing higher transcriptome entropy in CLL cells (CLL03, $n = 94$; CLL04, $n = 92$) than in healthy donor B cells (B04, $n = 84$) across various downsampling regimes (range 5,000–100,000; step-size of 1,000). Data are mean ± s.e.m. **d**, **e**, Single-cell transcriptional entropy (**d**) and epimutation rate (**e**) between normal $CD27^{-}$ B (NBC and loMBC) and $CD27^{+}$ B (intMBC and hiMBC) cells. **f**, Left, distribution of the Spearman's rho between expression and promoter DNAme rate ($n = 3{,}094$ genes with sufficient RNA (expression seen in >5 cells) and DNAme (>5 CpGs per promoter) information) in CLL04. The observed Spearman's rho values were compared to values obtained by randomly permuting cell labels for the methylation values (see Methods). Right, heat maps of Spearman's rank-order correlation for representative genes with positive or negative single-cell expression-methylation correlation. Scale bar represents promoter methylation and RNA read counts scaled by maximal value. **g**, As in **f** for individual normal B cells ($n = 5{,}729$ genes; $n = 16$ permutations; see Methods for details). **h**, As in **g** for CLL03 ($n = 2{,}699$ genes; $n = 16$ permutations). **i**, As in **g** for CLL04 ($n = 3{,}094$ genes; $n = 16$ permutations). **j**, Absolute change in Spearman's rho when comparing matched versus scrambled DNAme and RNA single-cell data in CLL (CLL03 and CLL04) and normal B (B04) cells. From the pool of genes used in **g**–**i**, only overlapping genes ($n = 951$) across the three samples were used in the comparison. **k**, As in **f** for individual normal B cells ($n = 2{,}500$ most variable genes with sufficient RNA (expression seen in >5 cells) and DNAme (>5 CpGs per promoter) information; $n = 16$ permutations; see Methods for details). **l**, As in **k** for CLL03. **m**, As in **k** for CLL04. **n**, Absolute change in Spearman's rho when comparing matched versus scrambled DNAme and RNA single-cell data in CLL (CLL03 and CLL04) and normal B (B04) cells. From the pool of genes used in **k**–**m**, only overlapping genes ($n = 459$) across the three samples were used in the comparison. **o**, Hydroxymethylation (5hmC) level at genes with positive correlation between expression and promoter DNA methylation (top correlated 10% of genes) compared with negatively correlated genes (top anti-correlated 10% of genes) in both normal B (B04; $n = 336$ and 330 genes, respectively) and CLL (CLL03 ($n = 290$ and 278 genes, respectively); CLL04 ($n = 320$ and 314 genes, respectively)) cells. Error bars represent 95% confidence interval. Published 5hmC data were used for the analysis[19]. Box plots are as defined in Fig. 1. *P* values were determined by two-sided Kolmogorov–Smirnov test (**f**–**i**, **k**–**m**), two-sided Wilcoxon signed-rank test (**j**, **n**) or two-sided Welch's *t*-test (**o**).



Extended Data Fig. 7 | Methylation-based lineage trees provide a native lineage tracing system. a, Additional representative (random cell subsampling) methylation-based lineage trees of CLL cells. **b**, As in **a** for index-sorted normal B cells, showing that naive CD27$^-$ B cells (NBC; CD27$^-$IgM$^+$IgD^{+++}IgG$^-$) precede CD27$^+$ memory terminally differentiated B cells (hiMBC; CD27$^+$IgG$^+$) in the lineage tree. **c**, Representative (cell subsampling) methylation-based lineage trees of CLL cells reconstructed using only autosomes or chromosome X. Tree topologies are similar to when using whole-genome information (see **a** and Fig. 3d), showing rapid drift after the initial malignant transformation. **d**, As in **c** for lineage trees of CLL cells obtained by holding-out chromosomes (hold-out three chromosomes at a time before phylogeny reconstruction; for example, excluding chromosomes 1–3, left), or downsampling the number of CpGs per cell to equal numbers (120,000 CpGs per cell; right). **e**, Normalized Robinson–Foulds distances between any two trees ($n = 30$ tree replicates; see Methods) of CLL01 reconstructed by maximum (M)-likelihood versus maximum-parsimony analyses. Differences (Δ) are indicated. **f**, Average maximum tree depth of lineage trees ($n = 10$ tree replicates; see Methods) of CLL (CLL01) and normal B (B02) cells when using whole-genome information compared to lineage trees obtained by holding-out chromosomes (hold-out three chromosomes at a time before phylogeny reconstruction). Error bars represent 95% confidence interval. **g**, Distribution of root-to-tip branch lengths (that is, the length from the root to each tip in the lineage tree) between CLL and normal B cells (M-CLL (CLL07), U-CLL (CLL10) and B05 are shown as representative examples). **h**, Patristic distances between index-sorted B cells from B04, B05 and B06 healthy donor samples (NBC, $n = 24$ cells for each sample; loMBC, $n = 24$ cells for each sample; intMBC, $n = 24$ cells for each sample; hiMBC, $n = 23$ cells for each sample). **i**, Patristic distances between CLL (CLL01) and normal B (B02) cells obtained from lineage trees reconstructed by using only autosomes, chromosome X, holding-out chromosomes (hold-out three chromosomes at a time before phylogeny reconstruction), or downsampling the number of CpGs per cell to equal numbers (120,000 CpGs per cell), respectively. Box plots are as defined in Fig. 1. *P* values were determined by two-sided Mann–Whitney *U*-test (**e**, **h**, **i**) or Welch's *t*-test (**f**).



Extended Data Fig. 8 | Multiplexed scRRBS integration with single-cell transcriptomes and genotyping. a, Schematic of the joint multiplexed scRRBS, transcriptome and genotyping capture protocol. **b**, Normalized Robinson–Foulds distances between any two trees ($n = 30$ tree replicates; see Methods) of CLL12 ($n = 56$ cells; see Fig. 3h) reconstructed by maximum-likelihood versus maximum-parsimony analyses. Differences (Δ) are indicated. **c**, Proportion of wild-type (white) and mutated (black) *SF3B1* cells in each clade identified from the lineage tree shown in Fig. 3h. **d**, Comparison of the number of unique CpGs (left) and the CpG methylation level (right) between the wild-type-enriched and the mutated-enriched *SF3B1* clade of cells identified from the lineage tree in Fig. 3h. **e**, Volcano plot of differentially methylated gene promoters (absolute weighted average DNAme difference > 0.3 and two-sided non-parametric permutation test $P < 0.05$) between the wild-type and mutated *SF3B1* cells from the lineage tree shown in Fig. 3h. **f**, Single-cell alternative 3′ splicing score (fraction of reads that map downstream to the 3′ end (up to 100 bp) of the exons versus within the exons) for cells belonging to wild-type ($n = 30$) and mutated ($n = 26$) *SF3B1* clades identified from the lineage tree shown in Fig. 3h. **g**, Volcano plot of differentially expressed genes between the wild-type-enriched and mutated-enriched *SF3B1* clade. Genes ($n = 57$) with absolute $\log_2$(*SF3B1* mutated-enriched/*SF3B1* wild-type-enriched gene expression) > 0.5 and Benjamini–Hochberg FDR-adjusted weighted *F*-test $P < 0.2$ are shown in red. Genes that were previously reported to be affected by *SF3B1* mutation[22] are also labelled. **h**, Gene expression projections on lineage trees for two representative genes identified in **g**. **i**, Comparison of transcriptional distances (measured as Euclidean distances of the first three principal components after principal component analysis) as a function of lineage distance between cell pairs from the lineage tree shown in Fig. 3h. **j**, Cells belonging to *SF3B1*-mutated enriched clade show significantly lower relative node heights (that is, height of internal tree nodes relative to the root node; see Methods) compared with wild-type *SF3B1*-enriched clade, consistent with *SF3B1* mutation being a later subclonal event in CLL[15]. **k**, As in **j** for root-to-tip branch lengths (that is, the length from the root to each tip in the lineage tree). **l**, Distribution of node ages (estimated number of divisions before present; see Methods) between the wild-type (white, $n = 30$ nodes) and mutated (grey, $n = 25$ nodes) *SF3B1* enriched clade. Box plots are as defined in Fig. 1. *P* values were determined by two-sided Mann–Whitney *U*-test (**b**, **d**, **f**, **i**, **j**, **l**) or two-sided Fisher's exact test (**c**).

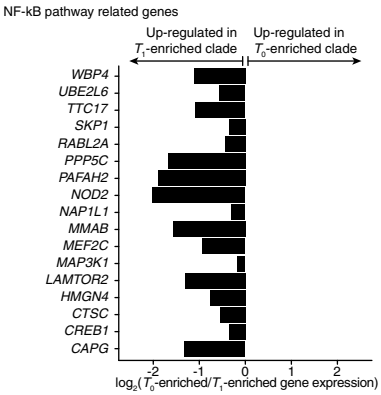


Extended Data Fig. 9 | Joint single-cell methylomics and RNA-seq link epigenetic and transcriptional information in CLL evolution with therapy. a, Representative methylation-based lineage trees integrating cells before treatment (T_0; white circle; $n = 40$ out of 96 randomly sampled cells) and during treatment (T_1; red circle; $n = 40$ out of 96 randomly sampled cells) for samples CLL03, CLL04 and CLL05. See Fig. 4c for the percentage of T_1 cells in each of the two clades (defined as the ones occurring after the first major split in the lineage tree) inferred from these lineage trees. **b**, Comparison of the CpG methylation level between the T_1-enriched clade of cells and the remaining T_1 cells identified from the lineage trees in **a** and in Fig. 4b and for samples CLL03, CLL04, CLL05 and CLL11, respectively. **c**, As in **b** for number of unique CpGs. **d**, Volcano plot of differentially methylated genes (absolute weighted average DNAme difference > 0.3 and two-sided non-parametric permutation test $P < 0.05$) between the T_1-enriched clade of cells and the remaining T_1 cells identified from the lineage trees in **a** and Fig. 4b for CLL03 ($n = 515$ genes), CLL04 ($n = 429$ genes), CLL05 ($n = 690$ genes) and CLL11 ($n = 578$ genes), respectively. Box plots are as defined in Fig. 1. *P* values were determined by two-sided Mann–Whitney *U*-test (**b**, **c**).

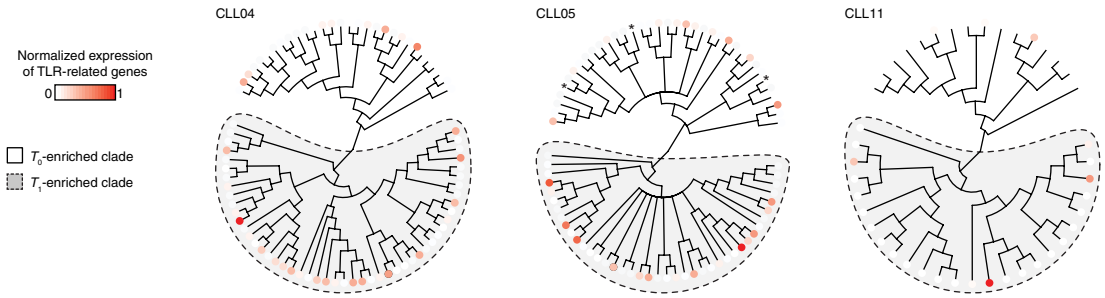
a

Description	*P*-value	*P*-adjust	BH-FDR
Up-regulated in T_1-enriched clade			
TLR related genes			
MyD88-independent TLR3/TLR4 cascade	0.0083	0.083	0.071
Toll Like Receptor 3 (TLR3) Cascade	0.0083	0.083	0.071
TRIF-mediated TLR3/TLR4 signaling	0.0083	0.083	0.071
JNK (c-Jun kinases) phosphorylation and activation mediated by activated human TAK1	0.0008	0.012	0.011
Cell cycle related genes			
Cell Cycle, Mitotic	0.0095	0.092	0.078
Nuclear Pore Complex (NPC) Disassembly	7.03×10^{-6}	0.0017	0.0014
Nuclear Envelope Breakdown	3.11×10^{-6}	0.0015	0.0013
Mitotic Prophase	0.0009	0.013	0.011
Up-regulated in T_0-enriched clade			
TP53 pathway genes			
Base-Excision Repair, AP Site Formation	6.7×10^{-5}	0.029	0.029
DNA Repair	0.00035	0.049	0.049
Regulation of TP53 Activity	0.0025	0.17	0.17
Base Excision Repair	0.0037	0.2	0.2

c



b



Extended Data Fig. 10 | Cells preferentially expelled from the lymph nodes are marked by a distinct transcriptional profile. a, Gene sets (canonical pathways; CP) enriched in differentially expressed genes ($n = 336$) between the T_1-enriched clade of cells and the remaining T_1 cells identified from the lineage trees in Fig. 4b and Extended Data Fig. 9a. A two-sided hypergeometric test was used to measure the enrichment of these genes in each gene set, followed by a Benjamini–Hochberg (BH) FDR procedure (cut-off of adjusted $P < 0.2$). **b**, Gene expression projections on lineage tree for TLR pathway genes from Fig. 4d for samples CLL04, CLL05 and CLL11, respectively. Scale bar represents RNA read counts scaled by maximal value. Expression value projection is performed only for T_1 cells, comparing T_1 versus T_0-enriched clades. Asterisks indicate cells without RNA information. **c**, Fold change in gene expression of NF-κB-related genes between the T_1-enriched clade of cells and the remaining T_1 cells identified from the lineage trees in Fig. 4b and Extended Data Fig. 9a.

nature research

Corresponding author(s): Dan Landau

Last updated by author(s): Mar 14, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Bismark (v.0.14.5); bowtie2-2.2.8; BWA v0.7.17; Picard; Macs2 (v2.0.10); STAR(version 2.5.2a); ZINBWaVE(v. 1.0.0); edgeR (v. 3.20.1); GSEA software and Molecular Signature Database (MSigDB) (http://www.broad.mit.edu/gsea/); BEDTools v2.25.0; MEME-ChIP; IQ-TREE v1.5.3; FigTreev1.4.3; Python 2.7.13; R version 3.4.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

MscRRBS and single-cell Smart-seq2 datasets have been deposited to the NCBI Gene Expression Omnibus (GEO) under accession number GSE109085. ChIP-seq datasets have been deposited to the NCBI GEO under accession number GSE119103. Additional supplementary data is available upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We sequenced 2,652 single cells in total from 6 and 12 B cell healthy donors and CLL patients, respectively, enabling us to do statistics both at the single cell and sample level, giving us enough statistical power to detect differences in all the analyses reported in this study (e.g., epimutation rates difference, concordance odds ratio analysis, four-gamete analysis). In addition, PAC learning analysis showed that ~400K CpGs enable trees with up to 350 leaves, allowing for information loss due to random sampling of two alleles. The dataset has >80% power to detect significant ($p < 0.05$), epigenetically derived subpopulations.
Data exclusions	No data were excluded from the study.
Replication	We performed 18 independent biological replicates, by applying multiplexed single-cell reduced representation bisulfite sequencing (MscRRBS) to 6 different B cells healthy donors and 12 CLL patients. This translates into a total of 2,652 cells profiled by single-cell methylome sequencing. All attempts at replication were successful.
Randomization	Randomization is not applicable as no experimental groups were used in our study.
Blinding	Blinding is not applicable as no experimental groups were used in our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	The antibodies used for index sorting of normal B cells were: FITC Mouse Anti-Human IgD (Clone IA6-2, BD Pharmingen), APC Mouse Anti-Human IgG (#562025, BD Biosciences), APC/Cy7 anti-human IgM Antibody (Clone MHM-88, BioLegend) and PE/Cy7 anti-human CD27 Antibody (clone O323, Bio Legend). Antibody used for ChIP is anti-H3K27ac (2 mg for 25 mg of chromatin; ab4729 Abcam, Cambridge, United Kingdom).
Validation	Expression of Immunoglobulin Heavy Chain (IGH) genes was assessed by scRNAseq in index-sorted B cell subpopulations validating our index-sorting strategy (CD27-IgM+IgD+++IgG- [NBC], CD27-IgM+IgD+IgG- [loMBC], CD27+IgM+IgD++IgG- [intMBC], and CD27+IgG+ [hiMBC]). In addition, all antibodies used were validated for their use in FACS or ChIP-seq experiments with human samples, as shown on the website provided by the respective companies.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Relevant information on human research participants is provided in Figure 1b, Extended Data Figure 1b and Supplementary Table 12.
Recruitment	The diagnosis of CLL according to World Health Organization (WHO) criteria was confirmed in all cases by flow cytometry, or by

Recruitment	lymph node or bone marrow biopsy. IRB-approved protocols for genomic sequencing of patients' samples was obtained prior to the initiation of sequencing studies. Blood samples were collected in EDTA blood collection tubes from patients and healthy adult volunteers enrolled on clinical research protocols at the Dana-Faber/Harvard Cancer Center (DF/HCC) and NewYork-Presbyterian/Weill Cornell Medical Center (NYP/WCMC), approved by the DF/HCC and NYP/WCMC Institutional Review Boards.
Ethics oversight	The study was approved by the local ethics committee and by the Institutional Review Board (IRB) and conducted in accordance to the Declaration of Helsinki protocol. We note that the IRB does not permit collection of demographic information of healthy donors.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119103
Files in database submission	GSM3358078 cll_175_h3k27ac_bam GSM3358082 cll_189_h3k27ac_bam GSM3358099 cll_248_h3k27ac_bam GSM3358103 cll_253_h3k27ac_bam
Genome browser session (e.g. UCSC)	No longer applicable

Methodology

Replicates	Two IGHV mutated and two IGHV unmutated CLL patient samples.
Sequencing depth	125 bp paired-end mode. An average of 75 million paired reads was generated per sample
Antibodies	Antibody used for ChIP is anti-H3K27ac (2 mg for 25 mg of chromatin; ab4729 Abcam, Cambridge, United Kingdom).
Peak calling parameters	Peaks were identified with Macs2 (v2.0.10) with a q-value threshold of 0.01, according to the ENCODE Histone ChIP-seq Data Standards and Processing Pipeline (https://www.encodeproject.org/chip-seq/histone/).
Data quality	Deeptools plotFingerprint v2 was used to assess ChIP-seq signal enrichment over background signal. In addition, we observed a large overlap (72%) between FANTOM5 human robust enhancers (defined by H3K27ac signal) and our CLL H3K27ac ChIP-seq peaks, confirming the reproducibility of our ChIP-seq data.
Software	ChIP-seq data were processed according to the ENCODE Histone ChIP-seq Data Standards and Processing Pipeline (https://www.encodeproject.org/chip-seq/histone/). Raw reads were mapped to the human genome hg19 assembly using Burrows-Wheeler Aligner (BWA v0.7.17). Duplicate reads were removed using Picard (https://broadinstitute.github.io/picard/). Peaks were identified with Macs2 (v2.0.10) with a q-value threshold of 0.01. Peaks overlapping with Satellite repeat regions and Encode Blacklist were discarded.