Corrupted coordination of epigenetic modifications leads to diverging chromatin states and transcriptional heterogeneity in CLL

Alessandro Pastore1,10, Federico Gaiti2,3,4,10, Sydney X. Lu1, Ryan M. Brand2,3,4, Scott Kulm3, Ronan Chaligne2,3,4, Hongcang Gu5, Kevin Y. Huang2,3,4, Elena K. Stamenova5, Wendy Béguelin3, Yanwen Jiang3, Rafael C. Schulman2,3,4, Kyu-Tae Kim2,3,4, Alicia Alonso3, John N. Allan3, Richard R. Furman3, Andreas Gnirke5, Catherine J. Wu5,6, Ari M. Melnick3, Alexander Meissner5,7, Bradley E. Bernstein5,8, Omar Abdel-Wahab1,9,10 & Dan A. Landau2,3,4

Cancer evolution is fueled by epigenetic as well as genetic diversity. In chronic lymphocytic leukemia (CLL), intra-tumoral DNA methylation (DNAme) heterogeneity empowers evolution. Here, to comprehensively study the epigenetic dimension of cancer evolution, we integrate DNAme analysis with histone modification mapping and single cell analyses of RNA expression and DNAme in 22 primary CLL and 13 healthy donor B lymphocyte samples. Our data reveal corrupted coherence across different layers of the CLL epigenome. This manifests in decreased mutual information across epigenetic modifications and gene expression attributed to cell-to-cell heterogeneity. Disrupted epigenetic-transcriptional coordination in CLL is also reflected in the dysregulation of the transcriptional output as a function of the combinatorial chromatin states, including incomplete Polycomb-mediated gene silencing. Notably, we observe unexpected co-mapping of typically mutually exclusive activating and repressing histone modifications, suggestive of intra-tumoral epigenetic diversity. Thus, CLL epigenetic diversification leads to decreased coordination across layers of epigenetic information, likely reflecting an admixture of cells with diverging cellular identities.
Cancer growth, progression, and relapse are the result of an evolutionary process fueled by intra-tumoral diversity. Chronic lymphocytic leukemia (CLL)—a common B cell malignancy—serves as a highly informative model for cancer evolution as it undergoes substantial genetic diversification and evolution with therapy.

In addition to genetic changes, the CLL epigenome is an important disease-defining feature linked to its cell-of-origin and is predictive of outcome. In fact, the stable propagation of the ancestral epigenome allowed the use of DNA methylation patterns to precisely retrace the initially transformed cell-of-origin from which different CLLs emerge. In addition to the largely stably inherited epigenome, we have previously shown that growing CLL populations also undergo ongoing somatic DNAme changes akin to the process of genetic diversification through ongoing mutations, leading to high intra-leukemic epigenetic heterogeneity, greater clonal evolution, and adverse outcome, as has been shown for other malignancies.

However, DNAme constitutes only a single layer of the epigenetic information encoding cell identity. Given the importance of histone modifications to lineage plasticity in cancer, we reasoned that intra-leukemic epigenetic heterogeneity may extend to histone modifications, likely promoting lineage plasticity by enabling permissive chromatin states. To address this question, we complemented bulk reduced representation bisulphite sequencing (RRBS) analysis with a chromatin immunoprecipitation sequencing (ChIP-seq) compendium of histone post-translational modifications and gene expression, together with joint DNAme and transcriptome single cell analysis in a cohort of 22 primary CLL and 13 healthy B lymphocytes samples. Our integrative analysis revealed a markedly decreased coordination between different layers of the CLL epigenome, whereby ongoing epigenetic diversification leads to an admixture of cells with diverging epigenetic identities, thus providing a novel perspective into the epigenetic dimension of cancer evolution.

Results
Super-enhancer and associated DNAme alteration in CLL. To comprehensively study the epigenetic landscape of evolving CLL and its relationship to intra-leukemic diversity, we generated genome-wide maps of histone marks with non-overlapping regulatory functions (H3K4me3, H3K27ac, and H3K27me3) and transcriptome sequencing (bulk RNA-seq) in a cohort of 20 primary IGHV mutated and unmutated CLL (corresponding to the major known disease subtypes; n = 14 and n = 6, respectively), as well as 12 healthy B lymphocytes samples (CD19+/CD23-/CD27−-negative tonsillar naïve B cells [NBCs]; CD19+CD23+CD27−IgD+), n = 2; peripheral blood NBCs [CD19+CD23+CD27−IgD+]), n = 4; CD19/CD23+CD27-positive IgD-negative tonsillar germinal center B cells [GCs]; CD19+CD23+CD27+IgD−], n = 2; peripheral blood memory B cells [GCBs; CD19+CD23+CD27+IgD−], n = 3; CD20+ tonsillar B cells [CD20+], n = 1; Supplementary Fig. 1a, b).

Analysis of H3K27ac, a histone modification known to be a marker of active gene regulatory regions, revealed core enhancer and super-enhancer (as defined in ref.; see Methods; Supplementary Fig. 1c) reprogramming in CLL. A total of 297 super-enhancers were differentially regulated in CLL compared with normal B cells (absolute log2[H3K27ac fold-change] >2 and Wald test BH-FDR <0.01; see Methods), with increased H3K27ac in proximity to genes critical for lymphocyte proliferation and differentiation, including BCL2, LEF1, and CTLA4 (Fig. 1a–c; Supplementary Fig. 1d) and involved in pathways previously reported to play key roles in CLL (e.g., B cell receptor, NF-kB and MAPK inflammatory signaling pathways; Fig. 1d). As ChIP-seq experiments are prone to technical variation, we further demonstrated the reproducibility of H3K27ac derangements in CLL by analyzing additional CLL and normal B cell samples from the Blueprint Initiative (Supplementary Data 1), showing high pairwise correlations across our cohort and the Blueprint initiative samples at super-enhancers (Supplementary Fig. 1e). Fewer differences in the super-enhancer landscape were observed between the two major known CLL subtypes (IGHV mutated and unmaturated; n = 27 super-enhancers differentially regulated; Supplementary Fig. 1f; see Methods), and with chromosome 13q deletion (del(13q); n = 25 super-enhancers differentially regulated; Supplementary Fig. 1g; see Methods), consistent with previous studies showing more subtle chromatin differences between CLL subtypes. In line with prior studies that profiled epigenomic features of a large CLL cohort and discrete B cell subtypes along the differentiation program, this extensive chromatin rewiring at super-enhancers is mediated by specific transcription factors, as evidenced by enrichment of their motifs in activated super-enhancers, including NFAT, a deregressed gene with functional and therapeutic potential in CLL, and TCF7L2, a downstream target of the WNT pathway overexpressed in CLL (Fig. 1e; Supplementary Data 2, 3).

DNAme changes at enhancers and super-enhancers impact their transcriptional activity. Therefore, to assess the relationship between DNAme and enhancers, we profiled bulk DNAme of normal B cell populations (peripheral blood naïve B cells [CD19+CD23+CD27−IgD+], n = 3; peripheral blood memory B cells [CD19+CD23+CD27−IgD−], n = 2) and CLL patient samples (IGHV unmaturated, n = 2; IGHV mutated, n = 3) using a targeted bisulphite sequencing capture assay, which preferentially evaluates dynamic CpGs at gene-regulatory elements (Supplementary Fig. 1a, b; Supplementary Fig. 2a, b; Supplementary Data 4, 5). Consistent with prior reports, we observed a global decrease in DNAme in CLL compared with normal B samples (Supplementary Fig. 2c, left; Supplementary Fig. 2d), with a focal increase in methylation of CpG islands (CGI; Supplementary Fig. 2c, right).

In addition, we identified 41,057 differentially methylated regions (DMRs; absolute change in DNAme >0.3 and Fisher’s exact test P <0.05; see Methods) between CLL and normal B samples, most of which were hypomethylated in CLL (Supplementary Fig. 2e; Supplementary Data 6–8). Interestingly, hypomethylation preferentially affected H3K27ac-enriched regions, including super-enhancers (Fisher’s exact test P <0.0001; Fig. 2a; Supplementary Fig. 2e, f). This extensive focal hypomethylation at super-enhancers was observed in proximity to genes involved in pathways previously reported to play key roles in CLL (e.g., B cell receptor activation, Notch signaling, and cell proliferation; Supplementary Fig. 2g; Supplementary Data 9). Additionally, CLL-specific super-enhancers showed a strong decrease in DNAme compared to normal B samples (Mann–Whitney U-test, P <0.0001; Fig. 2b, c), presented for the BCL2 gene locus (Fig. 2d). In contrast, super-enhancers that become inactive in CLL did not gain DNAme compared to normal B samples (Mann–Whitney U-test, P >0.05; Fig. 2b–d), supporting the concept that DNAme is slow to accumulate with CLL progression.

Notably, we observed that hypomethylation at super-enhancers resulted preferentially in intermediate DNAme levels in CLL (Fisher’s exact test P <0.0001; Fig. 2e; Supplementary Fig. 2h). These data demonstrated that cancer-associated hypomethylation is not limited to previously described intermediately methylated blocks in heterochromatin and lamina associated domains, but may also involve regions of active chromatin.

Decreased epigenetic-transcriptional coordination in CLL. The observed intermediate bulk DNAme patterns at H3K27ac
regulatory regions are reminiscent of our previous observation of intermediate DNAme in promoters stemming from stochastic DNAme intra-leukemic diversification during CLL evolution. Therefore, to examine whether enhancer rewiring is also associated with disordered methylation leading to reduced coordination between DNAme and H3K27ac, we drew on a well-established metric in information theory—mutual information (MI)—which measures how much can be learned from one variable about another (see Methods). Consistent with disrupted coordination across these two layers of the CLL epigenome, we observed lower pairwise MI between bulk DNAme and H3K27ac in CLL samples (irrespective of their IGHV mutational status) compared with normal B cell samples at super-enhancer regions (Welch’s t-test, *P* < 0.0001; Fig. 3a; Supplementary Fig. 3a). This decrease in MI also corresponded to a weaker correlation between DNAme and H3K27ac in CLL samples compared with normal B cell samples at super-enhancers (linear regression *R*^2^ of 0.558 normal B vs. 0.471 CLL samples, t-test *P* < 0.0001).

The decrease in MI was observed more broadly, including a 13% decrease in MI between DNAme at transcription start sites (TSSs) and gene expression, in CLL relative to normal B cells (Supplementary Fig. 3b). This decrease in MI may result from greater intra-leukemic cell-to-cell heterogeneity that is not captured in bulk population sequencing assays. To directly test this hypothesis, we performed joint single-cell DNAme sequencing and whole transcriptome sequencing on additional normal B and CLL samples (*n* = 96 cells [1 sample], *n* = 288 cells [2 samples], respectively; Fig. 3b; Supplementary Fig. 3c). While MI was higher across samples in matched vs. scrambled single-cell DNAme and RNA-seq data (paired t-test, *P* < 0.0001), the matched single-cell MI increase was higher in CLL compared with normal B cells (43 ± 5.2% vs. 29 ± 4.8%, respectively; Mann–Whitney U-test, *P* = 0.036; Fig. 3c). These data suggest that, at least in part, the decreased epigenetic-transcriptional coordination observed in CLL is the result of cell-to-cell epigenetic diversification.

To more broadly examine the relationship between epigenetic states (i.e., combinatorial interactions of epigenetic marks) and transcriptional output, we modeled the combinatorial patterns of histone modifications (H3K4me3, H3K27ac, H3K27me3) and DNAme (based on bulk bisulfite sequencing), with or without gene expression (based on bulk RNA-seq), using a Dirichlet
Process Mixture (DPM) approach, which allows learning de novo the number of combinatorial states. We observed a significantly higher number of states across CLL samples (in both IGHV mutated and unmutated samples), compared with normal B cells when adding RNA information into the DPM analysis, indicating that the transcriptional output of epigenetic states is less uniform in CLL (Mann–Whitney U-test, \(P < 0.0001\); Fig. 3d; Supplementary Fig. 3d). Specifically, while H3K27me3\(^{hi}\)/H3K4me3\(^{low}\)/H3K27ac\(^{low}\)–marked genes were indeed associated with uniform gene silencing in B cells, they were associated with variable expression in CLL (Mann–Whitney U-test, \(P < 0.0001\); Fig. 3e), suggesting that disrupted Polycomb repression in CLL results in leaky silencing allowing partial reactivation of these genes. Notably, variable expression of H3K27me3\(^{hi}\)/H3K4me3\(^{low}\)/H3K27ac\(^{low}\)–marked genes preferentially affected genes related to the critical B-cell receptor (BCR) signaling pathway (Supplementary Fig. 3e). In addition, we observed enrichment of specific transcription factor binding motifs in H3K27me3\(^{hi}\)/H3K4me3\(^{low}\)/H3K27ac\(^{low}\)–marked regions, including NAF7\(^{8}\) and MYB, a proto-oncogene overexpressed in CLL (Hypergeometric test \(P < 0.0001\); Supplementary Fig. 3f). These data suggest that CLL epigenomes are associated with less uniform transcriptional outputs compared with normal B cell epigenomes.

Transcriptional variation in genes with similar epigenetic patterns may stem from cell-to-cell transcriptional heterogeneity. To test this, we computed gene expression information entropy, a measure of cell-to-cell gene expression heterogeneity\(^9\) in our single-cell whole transcriptome data, and found that in CLL single cells (\(n = 94\)) H3K27me3\(^{hi}/\)H3K4me3\(^{low}/\)H3K27ac\(^{low}\)–marked genes were indeed associated with significantly higher intra-leukemic expression information entropy compared to normal B cells (\(n = 84\)), or compared to a set of genes with matched mean expression but not marked by H3K27me3\(^{hi}/\)H3K4me3\(^{low}/\)H3K27ac\(^{low}\) (Mann–Whitney U-test, \(P = 0.0003\) and \(P = 0.005\), respectively; Fig. 3f, g; Supplementary Fig. 3g). Our data therefore suggest a model in which H3K27me3–marked genes in CLL are incompletely silenced, resulting in greater cell-to-cell transcriptional heterogeneity.

Corrupted coherence across layers of the CLL epigenome. An alternative approach to assess the coordination between layers of the epigenome involves capturing their overlapping and mutually exclusive combinatorial patterns\(^{28}\). We pursued this orthogonal approach by training a multivariate Hidden Markov Model (HMM) on CLL and normal B cells data based on three of the different histone modifications (H3K4me3, H3K27ac, H3K27me3), DNAme (based on bulk bisulfite sequencing), and gene expression information (based on bulk RNA-seq). We identified 12 distinct epigenetic states that fell into two broad categories. First, a category that correlated with active transcription including active promoters (“Active flanking TSS”, “TSS”), enhancers (“Enhancer”, “H3K4me3/H3K27ac”), and 5' and 3' boundaries of transcribed genes (“1-IV transcription”). Second, a category of genes with no or little detectable transcription, including bivalent or poised (“Bivalent/
Poised TSS”, repressed Polycomb ("PRC"), and mCpG-rich ("mCpG") states (Fig. 4a). CLL overall showed high resemblance to normal B cells, and no significant differences in genomic coverage were observed between IGHV mutated and unmutated CLL (Supplementary Fig. 4a). Importantly, HMM analysis revealed a chromatin state simultaneously marked by H3K27ac and H3K27me3, modifications which are typically mutually exclusive, with a >2-fold enrichment in CLL compared with normal B cells (Hypergeometric test \(P < 0.0001\); Fig. 4a, b), and affecting ~1.6 M 200 bp genomic segments including non-first introns and distal regulatory elements (Supplementary Fig. 4b). We further validated this chromatin state by analyzing an additional 6 CLL and 7 normal B cell samples from the Blueprint Initiative\(^18\) and obtained high pairwise correlation (Spearman’s rho correlation coefficient = 0.47) between H3K27ac and H3K27me3 marks at H3K27ac-H3K27me3 segments identified in our data (Supplementary Fig. 4c). Evaluation of these H3K27ac-H3K27me3 segments from CLL revealed that a notable fraction (46.7%) of these regions possessed repressive chromatin modifications in normal B cells (Fig. 4c), suggesting these are genomic regions that are subject to CLL-specific activation by gaining activating acetylation marks (H3K27ac). Gaining H3K27ac in the transition from a healthy to disease state may be associated with upregulation of neighboring genes. Consistent with this scenario, RNA gene expression was increased in proximity to regions that gain H3K27ac in CLL (Fig. 4d; Supplementary Fig. 4d). A gene set enrichment
**Fig. 4** Corrupted coherence across layers of CLL epigenome leads to cell-to-cell transcriptional heterogeneity.  

**a** Chromatin state definitions and enrichments for a 12-state Hidden Markov Model based on three histone marks (H3K4me3, H3K27ac, H3K27me3), DNAme, and RNA information. *P*-values of a given HMM state between CLL and normal B cells are shown for two-sided hypergeometric test.  

**b** Epigenomic profiling of the FYN gene locus, demonstrating “H3K27ac-H3K27me3” state increase in CLL compared with normal B cells across our cohort and Blueprint initiative samples.  

**c** Sankey diagram showing that ~47% of the regions in a “H3K27ac-H3K27me3” state in CLL carried repressive chromatin modifications in B cells.  

**d** Fold-change gene expression between CLL and normal B cells in relation to genomic distance from regions that gain H3K27ac (orange; *n* = 11,740 genes) or H3K27me3 (blue; *n* = 8867 genes) in CLL. Mann–Whitney U-test.  

**e** Top 3 motifs in H3K27ac-H3K27me3 genomic segments in CLL. Position weight matrices of the top three motifs over-represented in CLL “H3K27ac-H3K27me3” regions. Motif enrichment hypergeometric test *P*-value and the best reference motif match (JASPAR core database) are shown.  

**f** Expression levels (log2(TPM)) of MYC target genes (containing promoter MYC binding motif, as in analysis in e) compared with non-MYC target genes in “H3K27ac-H3K27me3” regions.  

**g** Population average single-cell gene expression (log2(TPM)) in CLL. Presence of a single-cell gene expression value in the range of [-1.85, -1.35] is shown.  

**h** Gene expression magnitudes of genes in “Repressed Polycomb (PRC)” state and genes in “H3K27ac-H3K27me3” state. Cumulative distribution (right) showing the proportion of intermediate single-cell gene expression Shannon’s information entropy values at these genes is also shown.  

**i** Well-coordinated chromatin programs stabilize gene expression and cellular identities in normal B cells (left). On the contrary, intra-leukemic epigenetic diversity results in a permissive chromatin state in CLL cells (right), enhancing cell-to-cell transcriptional variation. Boxplots represent median and bottom and upper quartile; whiskers correspond to 1.5×IQR; error bars represent 95% confidence interval.
analysis of closest genes to these regions revealed enrichment in gene sets associated with stem cell identity\(^29,30\) (Hypergeometric test BH-FDR <0.05), linking regulatory chromatin variability to stem-like cell programs, according to the notion that epigenetic variability in cancer may lead to a drift toward a hybrid stem-somatic cell state\(^9,31\) (Supplementary Fig. 4e; Supplementary Data 10).

Epigenetic factors, such as aberrant regulation of H3K27 methylation\(^12\) and sporadic TF activation\(^25\), have been recently implicated in promoting lineage plasticity in cancer. Thus, to identify which TFs may carry the potential to rewire CLL cells and promote lineage plasticity in CLL, we further mined the regions marked by H3K27ac-H3K27me3 for transcription factor motif enrichment and identified a significant enrichment of the proto-oncogene MYC motif, a TF associated with lineage plasticity and CLL transformation to aggressive large B cell lymphoma\(^33\) (Hypergeometric test \(P < 0.0001\); Fig. 4e). RNA gene expression of genes with a MYC binding motif at their promoters was increased compared with non-MYC target genes, in the region marked by H3K27ac-H3K27me3 (median [IQR] of 9.44 [4.34] vs. 8.23 [5.17] log[TPM], respectively; Mann–Whitney U-test, \(P < 0.0001\); Fig. 4f).

Notably, the observed co-mapping of H3K27ac and H3K27me3 to the same genomic locus in CLL may arise from cell-to-cell divergence in histone modification, rather than co-occurrence of these mutually exclusive marks in the same cells. Consistent with this hypothesis, we analyzed our single-cell whole-transcriptome data and observed that genes neighboring H3K27ac-H3K27me3 regions in CLL were associated with higher intra-leukemic expression information entropy in single cells compared with genes neighboring Polycomb repressed regions (Mann–Whitney U-test, \(P < 0.0001\); Fig. 4g; h; Supplementary Fig. 4f, g). Collectively, these data suggest that CLL cell populations lose effective Polycomb repression of MYC targets, likely enabling an exploration of transcriptional stem-like cell programs in CLL evolution.

**Discussion**

While cancer evolution investigations have focused on genetic alterations, emerging data across cancer also highlighted the contribution of heritable epigenetic changes to cancer evolution\(^11,12,32\). In this study, we provided an integrative analysis of the epigenetic landscape of CLL and its relationship to intra-leukemic epigenetic and transcriptional diversity.

We observed extensive chromatin rewiring at H3K27ac regulatory regions mediated by specific transcription factor families, in particular NFAT and TCF/LEF transcription factor families\(^4,19,20\).

Through targeted bisulfite sequencing capture assay, we further showed these regulatory regions to display the highest degree of change in DNAme. Notably, enhancer hypomethylation is preferentially associated with intermediate DNAme levels, likely reflecting intra-leukemic cell-to-cell heterogeneity\(^9,10\). Thus, intermediately methylated regions in cancer may not be limited to heterochromatin as previously described\(^25,26\), affecting also regions of regulatory chromatin.

Moreover, while normal B cells exhibit coordinated epigenetic-transcriptional regulation resulting in higher pairwise mutual information, CLL samples have a substantial decrease in DNAme-RNA mutual information. This finding is consistent with intra-leukemic heterogeneity decreasing the mutual information of these two variables when measured at the population level. To directly examine this scenario, we applied matched DNAme and mRNA single-cell information and found a greater increase in single-cell mutual information in CLL compared with normal B cells. This observation confirms that the relatively small contribution of promoter DNAme to explaining transcriptional variation in bulk cancer studies\(^8\) results, at least in part, from intra-leukemic epigenetic diversity.

To further extend the evaluation of epigenetic co-ordination beyond two epigenetic layers, we modeled the combinatorial patterns of histone modifications, DNAme, and gene expression. Interestingly, we observed a dysregulation of the transcriptional output as a function of the combinatorial chromatin states. Specifically, while in normal B cells H3K27me3\(^30\)/H3K4me3\(^30\)/H3K27ac\(^30\)–marked genes were generally associated with a uniform transcriptional output, in CLL these genes were associated with variable expression level. As H3K27me3 is typically deposited at gene promoters by Polycomb Repressive Complex 2 (PRC2) via its catalytic Ezh2/Ezh1 subunit\(^33\), these results are consistent with CLL epigenetic landscape being marked by incomplete Polycomb complex-mediated gene silencing resulting in permissive chromatin states in a fraction of cells. Furthermore, as DNAme is important for appropriate retargeting of PRC2 and H3K27me3 histone modification across cell divisions\(^35\), stochastic DNAme alterations during CLL evolution\(^9\) may lead to redistribution of the repressive activity of the PRC2 complex and the H3K27me3 mark, and cell-to-cell variation in the efficiency of PRC2 transcriptional silencing\(^36\).

Lastly, we observed an unexpected co-occurrence of typically mutually exclusive activating (H3K27ac) and repressing (H3K27me3) histone modifications, closely associated with activation of stem-like programs and greater cell-to-cell transcriptional heterogeneity. Notably, the co-mapping of these typically mutually exclusive histone modifications was previously observed in the context of embryonic stem cell neural differentiation, reflecting cellular heterogeneity due to admixture of differentiated and undifferentiated cells\(^37\). Thus, epigenetic diversification leads to corrupted coherence across the different layers of the epigenome in CLL, consistent with ongoing epigenetic diversification leading to an admixture of cells with diverging epigenetic identities (Fig. 4i).

While genetic heterogeneity plays a key role in cancer growth, progression, and evolution with therapy\(^5,38,39\), epigenetic evolutionary routes are a major emerging theme across cancer, including prostate cancer, lung cancer and melanoma\(^11,40\). Cancer cells can display profound non-genetically mediated transcriptional variability, which may enable adaptive changes such as therapeutic resistance, persistence or lineage plasticity. Notably, these states are efficiently propagated to progeny cells suggesting stable epigenetic encoding. Indeed, in CLL, non-genetic persistence as well as lineage transformation have been reported as potential routes of escape from therapeutic inhibition\(^38,41\). Our data demonstrates that these adaptive capacities may be fueled by significant intra-tumoral epigenetic diversity resulting in permissive chromatin states across cells, leading to greater cell-to-cell transcriptional variation (Fig. 4i). Thus, intra-tumoral epigenetic diversity may permit leukemic cells to stochastically activate alternate gene regulatory programs, facilitating the emergence of novel cell states, ultimately fostering CLL’s ability to efficiently explore the fitness landscape for superior evolutionary trajectories during tumorigenesis and in response to therapy.

**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), Memorial Sloan Kettering Cancer Center (MSKCC), and New-York Presbyterian/Weill Cornell Medical Center (NYP/WCMC). We note that the IRB does not permit collection of demographic information of healthy donors. Informed consent on DF/HCC, MSKCC and WCMC IRB-approved protocols for genomic sequencing of patient samples was obtained prior to the initiation of
sequencing studies. The diagnosis of CLL according to World Health Organization (WHO) criteria was confirmed in all cases by flow cytometry, or by lymph node or bone marrow biopsy. Enhancers on the right of the initiation point were defined as super-enhancers (see Supplementary Fig. 1c). We note that the number of super-enhancers identified in our CLL cohort (range [279–964]; median of 474 across samples) is in line with a recent study that investigated enhancer architecture in a distinct CLL cohort48. To identify variable super-enhancer domains enriched in hypermethylated regions compared to normal B cells, we performed bisulphite sequencing on DNA isolated from our CLL patient samples which were discovered across the CLL and normal B cell cohorts. Differentially regulated super-enhancers in CLL compared with normal B cells (n = 297; see Fig. 1a–c and Supplementary Fig. 1d) were identified with DESeq249 as those with absolute log2 fold-change > 2 and Benjamini-Hochberg adjusted P-value < 0.01. To determine the methylation state of all CpGs captured and assess the bisulphite conversion rate, we imported the CpG level methylation call files into R using the methylKit function (http://bioconductor.org/packages/release/bioc/html/methylKit.html) version 2.8.0. Bisulphite sequencing data were processed according to the ENCODE Histone ChIP-seq Data Standards and Processing Pipeline (https://www.encodeproject.org/chip-seq/histone-seq-data-standards-processing-pipeline/) version 2.8.2, following the instructions provided by Illumina (HiSeq 2500 protocol). Bisulphite sequencing data were processed using Bismark v2.9.753 with the following parameters: bsmap -n 10 --1 0.1 -s 1 -n 1 -q 20 -r 0. Subsequently, we used Picard tools (http://picard.sourceforge.net) to filter bisulphite sequencing data (Fig. 2a). To determine the methylation state of all CpGs captured and assess the bisulphite conversion rate, we used the mcCL module in the MOABS46 software suite with further parameters. Then, we converted the resulting Cpg level files to bigwig files for visualization in IGV55, filtering out all Cpgs that were covered with less than five reads. Analysis of targeted bisulphite sequencing capture assay data was conducted using the methylKit package56 and a 500-bp tiling of the target capture set. Briefly, we imported the Cpg level methylation call files from mcCL into R using the methylKit function “methRead” and then computed the weighted methylation mean for each CpG using the “getMethData”, weighting the CpG mean by the number of observed reads. We then identified all the CpGs with the tiling level methylation information across all samples and retained only those tiling covered with more than 10 reads in 70% or more of all samples. To compute differently

**RNA-seq** RNA was extracted using Qiagen (Hilden, Germany) RNeasy columns according to the manufacturer’s instructions. Subsequently, 500 ng of total RNA was used for polyA selection and TruSeq library preparation according to the instructions provided by Illumina (TruSeq RNA Sample Prep Kit v2), with 8 cycles of PCR. Samples were barcoded and run on a HiSeq 4000 in a 125 bp paired-end mode, using the TruSeq SBS Kit v3 (Illumina, San Diego, CA). An average of 75 million paired reads was generated per sample. Raw reads were mapped to the human genome GRCh37 using STAR (v2.5.2a)aligner50. We used several QC metrics for the RNA-seq assay, including intron–exon ratio, intragenic reads fraction, and GC bias. We quantified exon and gene expression using Salmon51 against the Homo sapiens transcriptome GRCh37.

**Reduced representation bisulphite sequencing (RRBS)** Genomic DNA from CLL and normal cell samples was used to produce RRBS libraries. These were generated by digesting genomic DNA with MspI to enrich for CpG-rich fragments, and then were ligated to barcoded TruSeq adapters (Illumina). Millions of paired-end reads were sequenced on an Illumina HiSeq 2500 instrument in fast mode together with a pool of 2 nanomoles of the targets listed in Supplementary Data 4. Hybrid-selected sequencing libraries were sequenced on an Illumina HiSeq 2500 instrument in fast mode in a 18 to 20 cycle sequencing run using 3 flow cells (Kapa Biosystems). SeqCap Epi hybridization reactions contained a total of 1 μg of a pool of 2 to 4 PCR-amplified pre-capture libraries, a total of 1 nmol of 2 indexespecific blocking oligonucleotides, and the custom SeqCap probe pool designed for the targets listed in Supplementary Data 4. Hybrid-selected sequencing libraries were sequenced on an Illumina HiSeq 2500 instrument in fast mode in a 10 cycle sequencing run using 2 flow cells (Kapa Biosystems).

**Targeted bisulphite sequencing capture assay** Hybrid-selected sequencing libraries were prepared combining Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences) with the NimbleGen SeqCap Epi Enrichment System (Roche NimbleGen), enabling lower input DNA quantities while maintaining library complexity. Pre-capture libraries were constructed following the NimbleGen SeqCap Epi Enrichment System and sequenced on an Illumina HiSeq 2500 instrument using 3 flow cells (Kapa Biosystems). SeqCap Epi hybridization reactions contained a total of 1 μg of a pool of 2 to 4 PCR-amplified pre-capture libraries, a total of 1 nmol of 2 indexespecific blocking oligonucleotides, and the custom SeqCap probe pool designed for the targets listed in Supplementary Data 4. Hybrid-selected sequencing libraries were sequenced on an Illumina HiSeq 2500 instrument in fast mode in a 10 cycle sequencing run using 2 flow cells (Kapa Biosystems).
methylated tiles, we performed Fisher’s exact test on pooled CLL vs. normal B samples for each tile. Subsequently, we corrected the resulting p-values using Benjamini-Hochberg and defined differential methylated regions (DMRs) if they were less than 400 bp apart22. To begin, cells with fewer than 0.4% or 0.27% of the population were removed from the analysis. At single-cell resolution, it was required to have at least 10 cells with sufficient expression (n = 240; 99.6%), with a negligible number of cells being in either G2/M (n = 0; 0%; AUC > 0.27) or G1 phase (n = 0; 0%; AUC > 0.27) or G1 phase (n = 0; 0%; AUC > 0.27). We note that the analysis for Fig.3c was applied applying the “GeneCounts” option in the STAR alignment. We filtered out poor quality cells when the detected number of genes was below 500 or the fraction of mitochondrial gene counts was higher than 20%. To compare the cells in terms of their transcriptional differences, we normalized the read counts by making each sample equal to 1000 bp, with each sample being normalized to the number of counts per cell. To detect differentially methylated regions, we first demultiplexed by Illumina pooling guidelines, a different i7 index was used for each sample. Following Illumina pooling guidelines, we assigned a cell with any detected methylation or expression for a given gene set. Cells expressing many genes from the gene set will have higher AUC score than cells expressing fewer. The highest AUC threshold is used to consider a gene set “active” in a given cell. The Molecular Signature Database29 (MSigDB; http://software.broadinstitute.org/gsea/index.jsp) and gSEA were used as input in the enrichment analysis. We observed that the vast majority of cells are classified as being non-cycling cells (n = 240; 99.6%), with a negligible number of cells being in either G2/M (n = 0; 0%; AUC > 0.27) or G1 phase (n = 0; 0%; AUC > 0.27), consistent with the majority of CLL cells being in a resting non-cycling stage28.

Single-cell DNAme-gene expression MI analysis. To begin, cells with fewer than 500 detected genes or a proportion of mitochondrial or ribosomal reads above 20% were removed from the analysis for quality control. Constitutively highly expressed mitochondrial genes and genes encoding ribosomal proteins across all cells were then removed. Then, cells in the bottom 10th percentile of total read counts for a given sample were discarded, and each of the remaining cells was probabilistically down-sampled to the number of reads at this cutoff. Subsequently, genes with reads discarded across more than five cells were removed from the analysis. At single-cell resolution, a gene’s promoter methylation rate was represented by the proportion of methylated CpG sites in the region 2500 base pairs upstream and downstream of the transcription start site. Genes with less than 10 CpG observations in the promoter region for a given cell were removed. We then computed the mutual information between the promoter methylation rate and gene expression for each cell using a threshold of zero, implying that any detected methylation or expression for a given gene was treated as having a value of 1 for that cell, and 0 otherwise. For a gene to be included in the final analysis, it was required to have at least 10 cells with sufficient CpG data for a methylation call (10 CpG observations) as well as greater than 10% non-zero expression across all cells to mitigate the impact of dropout. The approach was validated by a non-parametric premutation test, in which we randomly permuted the cell methylation values for each gene while holding the corresponding expression vector constant (such that RNA and DNAme are no longer linked at the single-cell level) and computed an unmatched version of the mutual information. This was followed by a permutation test, which was computed by taking the median of the matched and unmatched mutual information value provided corresponds to the median of the result for each of these permutations. We note that the analysis for Fig. 3c was
performed with downsampling to create a balanced dataset by matching the number of genes between CLL and normal B cells (n = 759 genes).

**Gene set enrichment analysis.** Gene set enrichment analysis was performed using GSEA software, and Molecular Signature Database (MSigDB, [http://www.broad.mit.edu/gsea/](http://www.broad.mit.edu/gsea/)). Specifically, we used the C2 curated gene sets and Benjamini–Hochberg FDR adjusted P-value cut-off of 0.05.

**Chromatin hidden Markov model (HMM).** Chromatin states across the genome were defined using EpigSeq2, which is based on a multivariate HMM, using H3K4me3, H3K27ac, H3K27me3, whole cell extract, RNA-seq and DNAm e (based on bulk RRBS) datasets as input. ChIP-seq reads were shifted in the 5′-3′ direction by 100 bp. Reads counts were computed in 200 bp non-overlapping bins. Normalized raw counts were then modeled with an HMM assuming that the hidden state vector followed a negative binomial distribution. We trained several HMM models in parallel with the number of states ranging from 5 states to 25 states and chose a 12-state model as the best model that captures all the key interactions between the epigenetic marks and cover all possible genomic elements (promoter, enhancer, gene body) that we expected to resolve given the selected datasets we used (H3K4me3, H3K27ac, H3K27me3, RNA-seq, and DNAm e). Genomic regions were then annotated with the state with the maximum posterior probability in the 200 bp bin. State enrichment in different genomic features was calculated dividing the percentage of nucleotides occupied by a state in a particular genomic feature by the percentage of nucleotides that this genomic feature represents in the entire genome.

**Chromatin Dirichlet process Gaussian mixture model.** Infinite mixture model with the Dirichlet process was used to model the normalized signal count matrix and to derive a segmentation of the chromatin tracks. The scikit-learn Python library ([sklearn.mixture.BayesianGaussianMixture](https://scikit-learn.org/stable/modules/generated/sklearn.mixture.BayesianGaussianMixture.html)) was used to generate an independent model for each sample. Cross-validation for each sample was performed training on a random 1/10 of the genome, applying the cross-validation model to the sample and repeating this procedure 100 times. Subsequently, a leave-one-out procedure was implemented to assess the contribution of each chromatin and transcriptome track independently. Unsupervised hierarchical clustering of state emission was performed to identify unique states.

**Single-cell entropy analysis.** To test for significance of association of chromatin state status with expression heterogeneity in Figs. 3f and 4g, single cell RNA-seq read counts observed in each cell were normalized by the effective library size and transcript length, and the fraction of positive cells (fpc) was calculated per gene (a cell is defined as positive if > 0 reads aligned to the gene). Subsequently, Shannon’s information entropy (ent) was calculated for each gene as followed:

\[
\text{ent} = - \sum \left( \frac{f}{n} \log_{2} \frac{f}{n} \right) = - \left( \frac{1}{f} \sum f \log_{2} \frac{f}{n} \right)
\]

The association with chromatin state status was tested using a generalized additive model (implemented by gam R package). The following type of model was tested:

\[
\text{ent} \sim \beta(n) \text{population average expression} + \text{chromatin state status} \quad \text{where } n \text{ indicates local regression.}
\]

The population average expression values were entered into the model on a log2 scale.

**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, paired t-test, non-parametric permutation test or Kolmogorov–Smirnov test as appropriate. P-values were adjusted for multiple comparisons by Benjamini–Hochberg FDR procedure, as appropriate. All P-values are two-sided and considered significant at the 0.05 level unless otherwise noted.

**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

ChIP-seq, RNA-seq, and DNAm e datasets have been deposited to the NCBI Gene Expression Omnibus ([GEO](https://www.ncbi.nlm.nih.gov/geo/)) under accession number GSE119103. MscRRBS and single-cell Smart-seq2 datasets have been deposited to the NCBI GEO under accession number GSE109085. The dGAP accession number for the whole-exome sequencing data were entered into the model on a log10 scale.

References

1. Puente, X. S. et al. Whole genome sequencing identifies recurrent mutations in chronic lymphocytic leukemia. *Nature* 475, 101–105 (2011).
2. Schuh, A. et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* 120, 4191–4196 (2012).
3. Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152, 714–726 (2013).
4. Fabbrini, G. et al. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *J. Exp. Med.* 210, 2273–2288 (2013).
5. Burger, J. A. et al. Clonal evolution in patients with chronic lymphocytic leukemia developing resistance to btk inhibition. *Nat. Commun.* 7, 11589 (2016).
6. Oakes, C. C. et al. Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia. *Cancer Discov.* 4, 348–361 (2014).
7. Kulis, M. et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* 44, 1236–1242 (2012).
8. Oakes, C. C. et al. DNA methylation dynamics during b cell maturation underlie a continuum of diseases: phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* 48, 253–264 (2016).
9. 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).
10. 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).
11. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* 357, eaal2380 (2017).
12. Läus, B. B. et al. Adaptive chromatin remodeling drives glioblastoma stem cell plasticity and drug tolerance. *Science Adv.* 6, eaau9029 (2020).
13. Damle, R. N. et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 94, 1840–1847 (1999).
14. Pott, S. & Lieb, J. D. What are super-enhancers? *Nat. Genet.* 47, 8–12 (2014).
15. Wang, L. et al. Somatic mutation as a mechanism of wnt/b-catenin pathway activation in cll. *Blood* 124, 1089–1099 (2014).
16. Robertson, L. E. et al. Bcl-2 expression in chronic lymphocytic leukemia and its correlation with the induction of apoptosis and clinical outcome. *Leukemia* 10, 456–459 (1996).
17. Kosmaczewska, A. et al. CIta-4 overexpression in cld19/+/cl5+ cells correlates with the level of cell cycle regulators and disease progression in b-ll patients. *Leukemia* 19, 301–304 (2004).
18. Adams, D. et al. Blueprint to decode the epigenetic signature written in blood. *Nat. Biotechnol.* 30, 224–226 (2012).
19. Beekman, R. et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* 24, 868–880 (2018).
20. Gutierrez, A. J. et al. Lef-1 is a prosurvival factor in chronic lymphocytic leukemia and is expressed in the preleukemic state of monoclonal b-cell lymphocytosis. *Blood* 116, 2975–2983 (2010).
21. Qu, Y. et al. Cancer specific changes in DNA methylation reveal aberrant silencing and activation of enhancers in leukemia. *Blood* 129, 13–25 (2016).
22. Ziller, M. J. et al. Targeted bisulfite sequencing of the dynamic DNA methylation landscape. *Epigenet. Chromatin* 9, 55 (2016).
23. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome—biological and translational implications. *Nat. Rev. Cancer* 11, 726–734 (2011).
24. Kulis, M. et al. Whole-genome fingerprint of the DNA methylome during human b cell differentiation. *Nat. Genet.* 47, 746–756 (2015).
25. Berman, B. P. et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* 44, 40–46 (2012).
26. Timp, W. et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med.* 6, 61 (2014).
27. Vargova, K. et al. MB2 transcriptionally regulates the mib-155 host gene in chronic lymphocytic leukemia. *Blood* 117, 3816–3825 (2011).
28. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49 (2011).
29. Lim, E. et al. Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* 12, R21 (2010).
30. Wong, D. J. et al. Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell* 2, 333–344 (2008).
31. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* 13, 497–510 (2013).
32. Mu, P. et al. Sox2 promotes lineage plasticity and antiangiogenesis in p53 and r1-deficient prostate cancer. *Science* 355, 84–88 (2017).
33. Delgado, M. D. & León, J. Myc roles in hematopoiesis and leukemia. *Genes Cancer* 1, 605–610 (2010).

34. Hoffmann, A. et al. Switch-like roles for Polycym proteins from neurodevelopment to neurodegeneration. *Epigenomes* 1, 21 (2017).

35. Reddington, J. P. et al. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycym target genes. *Genome Biol.* 14, R25 (2013).

36. Gal-Yam, E. N. et al. Frequent switching of Polycym repressive marks and DNA hypomethylation in the PC3 prostate cancer cell line. *PNAS* USA 105, 12979–12984 (2008).

37. Rehimi, R. et al. Epigenomics-based identification of major cell identity regulators within heterogeneous cell populations. *Cell Rep.* 17, 3062–3076 (2016).

38. Landau, D. A. et al. The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. *Nat. Commun.* 8, 2185 (2017).

39. Landau, D. A. et al. Mutations driving cll and their evolution in progression and relapse. *Nature* 526, 525–530 (2015).

40. Shaffer, S. M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* 546, 431–435 (2017).

41. Byrd, J. C. et al. Ibrutinib in relapsed chronic lymphocytic leukemia. *Cancer Cell* 23, 677–692 (2013).

42. Love, M. I. et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Statistical Applications in Genetics and Molecular Biology* 15, Article 39 (2016).

43. Heinz, S. et al. Simple combinations of lineage-determining transcription regulators. *Cell* 148, 1158–1171 (2012).

44. McLean, C. Y. et al. Great improves functional interpretation of cis-regulatory regions. *Nature Biotechnol.* 36, 215–221 (2018).

45. Béguelin, W. et al. EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell* 23, 677–692 (2013).

46. Ott, C. et al. Enhancer architecture and essential core regulatory circuitry of cancer. *Nature* 501, 517–522 (2013).

47. Xi, Y. & Li, W. BSMAP: whole genome bisulfate sequencing data mapping. *BMC Bioinformatics* 10, 332 (2009).

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

49. Rehimi, R. et al. Mutations driving cll and their evolution in progression and relapse. *Nature* 526, 525–530 (2015).

50. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Genet.* 47, 494 (2015).