Cancer Cell
Article

Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia

Dan A. Landau,[1,2,3,13] Kendell Clement,[3,4,5,13] Michael J. Ziller,[3,4] Patrick Boyle,[3] Jean Fan,[6] Hongcang Gu,[3] Kristen Stevenson,[7] Carrie Sougnez,[3] Lili Wang,[1,2] Shuqiang Li,[8] Dylan Kotliar,[1] Wandi Zhang,[1] Mahmoud Ghandi,[3] Levi Garraway,[2,3] Stacey M. Fernandes,[2] Kenneth J. Livak,[8] Stacey Gabriel,[3] Andreas Gnirke,[3] Eric S. Lander,[3] Jennifer R. Brown,[2,9] Donna Neuberg,[7] Peter V. Kharchenko,[6,10] Nir Hacohen,[3,11] Gad Getz,[3,12,14] Alexander Meissner,[3,4,14,*] and Catherine J. Wu[1,2,10,14,*]
[1]Cancer Vaccine Center
[2]Department of Medical Oncology
Dana-Farber Cancer Institute, Boston, MA 02115, USA
[3]Broad Institute, Cambridge, MA 02139, USA
[4]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
[5]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA 02139, USA
[6]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[7]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, MA 02115, USA
[8]Fluidigm, South San Francisco, CA 94080, USA
[9]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[10]Division of Hematology/Oncology, Children's Hospital, Boston, MA 02115, USA
[11]Center for Immunology and Inflammatory Diseases
[12]Cancer Center and Department of Pathology
Massachusetts General Hospital, Boston, MA 02114, USA
[13]Co-first author
[14]Co-senior author
*Correspondence: alexander_meissner@harvard.edu (A.M.), cwu@partners.org (C.J.W.)
http://dx.doi.org/10.1016/j.ccell.2014.10.012

SUMMARY

Intratumoral heterogeneity plays a critical role in tumor evolution. To define the contribution of DNA methylation to heterogeneity within tumors, we performed genome-scale bisulfite sequencing of 104 primary chronic lymphocytic leukemias (CLLs). Compared with 26 normal B cell samples, CLLs consistently displayed higher intrasample variability of DNA methylation patterns across the genome, which appears to arise from stochastically disordered methylation in malignant cells. Transcriptome analysis of bulk and single CLL cells revealed that methylation disorder was linked to low-level expression. Disordered methylation was further associated with adverse clinical outcome. We therefore propose that disordered methylation plays a similar role to that of genetic instability, enhancing the ability of cancer cells to search for superior evolutionary trajectories.

INTRODUCTION

Cancer evolution is a central obstacle to achieving cure, as treatment-resistant disease often emerges even in the context of highly effective therapies. Recent studies by us and others have demonstrated the contribution of genetic heterogeneity within each individual cancer to clonal evolution and its impact on clinical outcome (reviewed in Landau et al., 2014). In addition

Significance

Although it is well established that genetic intratumoral diversity fuels tumor evolution, relatively little is known about epigenetic diversity in primary cancer samples and its impact on evolution and outcomes. Using a variety of molecular platforms, we demonstrated a higher degree of intratumoral heterogeneity of DNA methylation in CLL. We have further shown that this heterogeneity stems from seemingly stochastic variation, reminiscent of the model of genetic heterogeneity in cancer, wherein stochastic variation is subjected to selection in tumor evolution. These data transform the way we view methylation differences between normal and cancer cells and will facilitate the crucial distinction between epigenetic alterations that result from background stochastic variation versus positive selection, in this leukemia and other cancers.

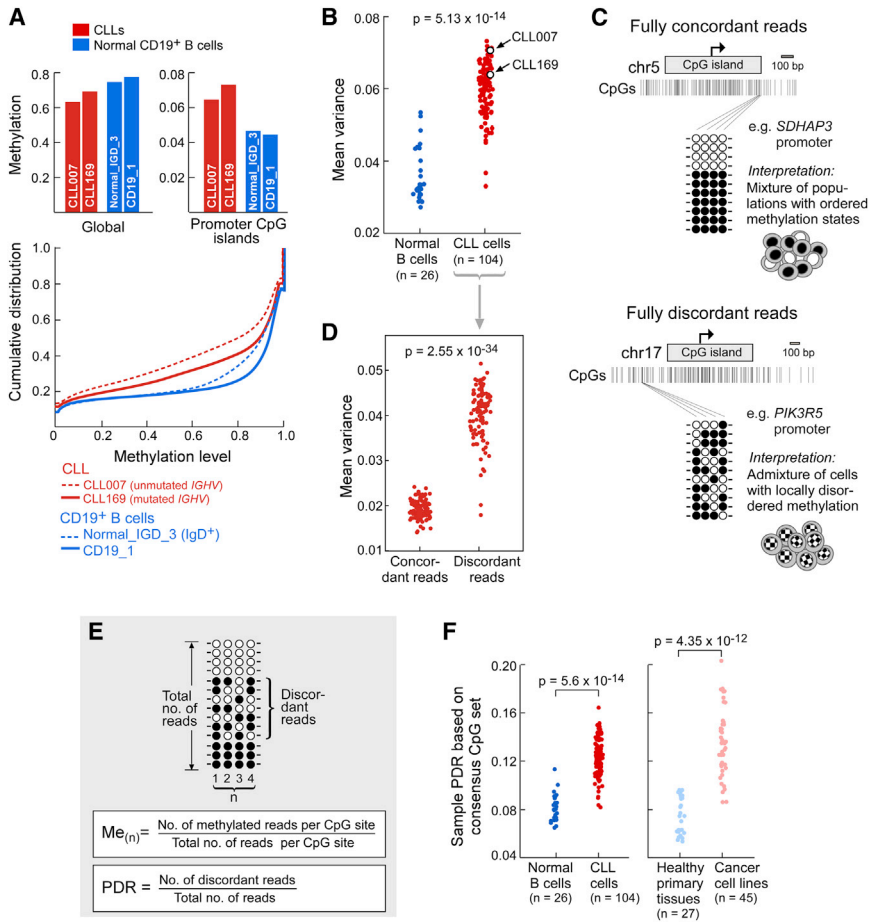


Figure 1. Higher DNA Methylation Intrasample Heterogeneity in CLL Arises from Locally Disordered Methylation

(A) CLL global and CGI methylation compared with normal B cells, measured with WGBS (top). Cumulative distribution analysis (bottom) enables the comparison of the proportion of intermediate methylation values in WGBS data of CLL and B cells from healthy adult volunteers (also see Figure S1).

(B) Mean intrasample CpG variance measured with RRBS.

(C) Methylation patterns from RRBS data of a CLL sample (CLL007) show two patterns of methylation (black circles, methylated CpGs; white circles, unmethylated): (1) a pattern compatible with a mixture of cell populations with clear but distinct methylation states for a particular nonimprinted locus (left-*SDHAP3* promoter [chr5:1594239-1594268]) and (2) a pattern compatible with an admixture of cells with locally disordered methylation (right-*PIK3R5* promoter [chr17:8869616-8869640]).

(D) A comparison between the intrasample CpG variance that arises from discordant compared with concordant reads across the 104 CLLs.

(E) CpG methylation and the PDR were calculated as shown.

(F) Sample average PDR for CLL, cancer cell lines, normal B cells, and a collection of primary healthy human tissues. To enable an accurate comparison between samples, sample average PDR is calculated on the basis of a consensus set of 63,443 CpGs that are covered with greater than ten reads in >75% of all 202 RRBS samples.

See also Figure S1 and Tables S1 and S2.

to genetic mutations, somatic epigenetic alterations are also drivers of neoplastic transformation and fitness (Baylin, 2005; Baylin and Jones, 2011). Moreover, genetically uniform cells exhibit phenotypic variation in essential properties such as survival capacity and proliferative potential (Kreso et al., 2013; Spencer et al., 2009), likely reflecting epigenetic variation. Hence, a priority in cancer biology is to measure intratumoral heterogeneity at the epigenetic level and determine how somatic genetic and epigenetic heterogeneity together affect tumor evolution.

To examine these critical questions, we focused on chronic lymphocytic leukemia (CLL), a malignancy of mature B cells with well-documented epigenetic dysregulation of CLL-associated genes (Raval et al., 2007; Yuille et al., 2001). Stable differences have been observed in DNA methylation across CLL samples compared with normal B cells as well as between subtypes of CLL (e.g., with mutated versus unmutated *IGHV*) (Cahill et al., 2013; Harris et al., 2010; Kulis et al., 2012; Pei et al., 2012). We were motivated to perform an integrative study of intraleukemic genetic and DNA methylation heterogeneity in CLL because (1) recent studies have suggested that both epigenetic marks and genetic alterations can improve prognostic models of CLL (Kulis et al., 2012; Rossi et al., 2013); (2) higher methylation variability has been detected across cancer subtypes compared with healthy tissue-matched samples, including in other B cell malignancies (Berman et al., 2012; De et al., 2013; Hansen et al., 2011); and (3) the availability of whole-genome bisulfite sequencing (WGBS) and reduced-representation bisulfite sequencing (RRBS) now enables genome-wide investigation of DNA methylation at single base pair resolution and with local sequence context. In particular, RRBS constitutes a cost-effective approach that allows the study of large patient cohorts (Boyle et al., 2012).

We thus performed WGBS and RRBS on a large cohort of primary patient samples that were previously characterized by whole-exome sequencing (WES) (Landau et al., 2013), to assess intraleukemic DNA methylation heterogeneity in CLL.

RESULTS

Increased Intrasample DNA Methylation Heterogeneity in CLL Arises from Locally Disordered Methylation

To measure intrasample CLL DNA methylation heterogeneity, we compared WGBS data generated from two CLL cases and two healthy donor B cell samples (Figure 1A). We observed globally decreased methylation in CLL compared with normal B cells, with focally increased methylation of CpG islands (CGIs) (Figure 1A, top; Figures S1A–S1C available online), as previously reported in CLL and other cancers (Baylin and Jones, 2011; Kulis et al., 2012), but also a markedly increased frequency of

intermediate methylation values in CLL (Figure 1A, bottom; Figures S1A–S1D), pointing to a large proportion of CpGs that are methylated in some cells in the sample and unmethylated in others. We reanalyzed published WGBS and Illumina 450 K methylation array data (Kulis et al., 2012) and confirmed the increased cell-to-cell variability in CpG methylation in CLL compared with normal B cells (Figures S1E–S1H).

We next applied RRBS to 104 primary CLL samples that had been previously characterized by WES (Landau et al., 2013) (Tables S1 and S2) and examined mean CpG variance. Consistent with the WGBS data, a greater than 50% increase in intrasample methylation heterogeneity was detected in CLL cells compared with 26 normal B cell samples (Figure 1B). We considered two possible sources for intrasample heterogeneity: variability between concordantly methylated fragments (i.e., whereby CpGs in an individual fragment are consistently methylated or unmethylated; Figure 1C, left) and variability within DNA fragments (i.e., discordant methylation by which CpGs in an individual fragment are variably methylated; Figure 1C, right).

On the basis of established observations that short-range methylation is highly correlated in normal physiological states (Eckhardt et al., 2006; Jones, 2012), we initially hypothesized that intrasample heterogeneity in CLL stems from variability between concordantly methylated fragments, reflecting a mixture of subpopulations with distinct but uniform methylation patterns. To test this, we focused on CpGs covered by reads containing four or more neighboring CpGs, as previously suggested (Landan et al., 2012), and with sufficient read depth (greater than 10 reads per CpG, with ~6.5 million CpGs/sample covered by 100-mer WGBS reads, and an average of 307,041 [range 278,105–335,977] CpGs/sample covered by 29-mer RRBS reads). Contrary to the expected hypothesis, we found that 67.6 ± 3.2% (average ± SD) of the intratumoral methylation variance resulted from discordantly methylated reads across the 104 CLL samples ($p = 3.24 \times 10^{-35}$; Figure 1D). Similarly, the CLL WGBS confirmed a higher proportion of heterogeneously methylated CpGs in the discordant reads compared with the concordant reads (Figure S1E, right). These results demonstrate that methylation heterogeneity in CLL arises primarily from variability within DNA fragments, which we have therefore termed "locally disordered methylation."

We performed several analyses to exclude potential alternative explanations to these findings, including the impact of contaminating nonmalignant cells (Figure S1I), allele-specific methylation (Figures S1J–S1L), the contribution of reads that cover an ordered transition point from one methylation state to another (Figure S3L), and technical biases (see Supplemental Experimental Procedures). The sex chromosomes were excluded from this analysis to avoid possible confounding sex chromosome-specific effects. In addition, CLL genomes are near diploid (Brown et al., 2012), and therefore the analysis was not significantly affected by somatic copy number variations (see Supplemental Experimental Procedures and Figure S1O).

To quantify the magnitude of this phenomenon across large collections of normal and malignant human tissues, we analyzed RRBS data not only from the 104 CLL and 26 B cell samples but also from 45 solid and blood cancer cell lines and from 27 primary human tissue samples (Table S2). We then calculated the proportion of discordant reads (PDR) as the number of discordant over the total number of reads for each CpG in the consensus set (Figure 1E). As expected, we found that the average PDR was higher in CLL compared with normal B cells ($p = 5.60 \times 10^{-14}$). Similarly, we found higher PDR in cancer cell lines compared with a diverse collection of healthy human tissue samples ($p = 4.35 \times 10^{-12}$; Figure 1F). These results support the idea that locally disordered methylation is a general property of the malignant process.

Locally Disordered Methylation Broadly Affects the CLL Genome

To determine whether specific elements in the genome harbor higher levels of locally disordered methylation in CLL compared with normal B cells, we calculated the average PDR across the 104 CLL samples and 26 healthy donor B cell samples (Table S3).

In normal B cells, PDR levels were lowest in regions with major roles in gene regulation (promoters, CGIs, exons, enhancers) and higher in regions with presumably less of a regulatory role (CGI shelves and shores, intergenic regions). In CLL, PDR was higher across all measured regions (Figure 2A), regardless of whether they were relatively hypermethylated (e.g., CGIs) or hypomethylated (e.g., intergenic regions) compared with normal B cells (Figure 2B). This phenomenon appeared to be neither specific to a subregion of CGIs or promoters (e.g., CGI borders; Figure 2C) nor restricted to a subtype of CGI (Figure S2A). Increased PDR in CLL was also observed in highly repetitive DNA sequences (e.g., long interspersed elements [LINE] and long terminal repeat retrotransposons; Figure 2A, RRBS data, and Figure S2B, WGBS data), which largely account for the global DNA hypomethylation observed in cancer (Ehrlich, 2009).

Alterations in the DNA methylation regulatory machinery could affect PDR. Unlike other hematological malignancies (Ley et al., 2010), somatic mutations affecting direct DNA methylation modulators in CLL are rare (Landau et al., 2013). Nonetheless, three CLL samples with such somatic mutations (DNMT3A-Q153*, TET1-N789I, and IDH1-S210N) showed increased PDR compared with the 101 CLL samples wild-type for these genes (Figure S2C).

Locally Disordered Methylation Appears to Be a Largely Stochastic Process

Two observations in the data suggest that PDR measures a process that stochastically increases variation in methylation, a notion that was recently conceptualized as a feature of the cancer epigenome (Pujadas and Feinberg, 2012). First, the pervasiveness of locally disordered methylation across every region evaluated in CLL compared with B cells supports a stochastic genome-wide process. Second, consistent with a stochastic process, wherein the expected rate of increase in PDR would be related to the starting level of disorder, we observed a larger relative PDR increase in CLL in regions with lower PDR in normal B cells. To formally measure the level of disorder, we undertook a parallel analysis to calculate Shannon's information entropy of intrasample methylation variation (Figure S3A). We determined this entropy to be higher in CLL than in normal B cells (as well as higher in cancer cell lines compared with normal tissues), consistent with an increase in stochastic "noise" (Figures S3B and S3C).

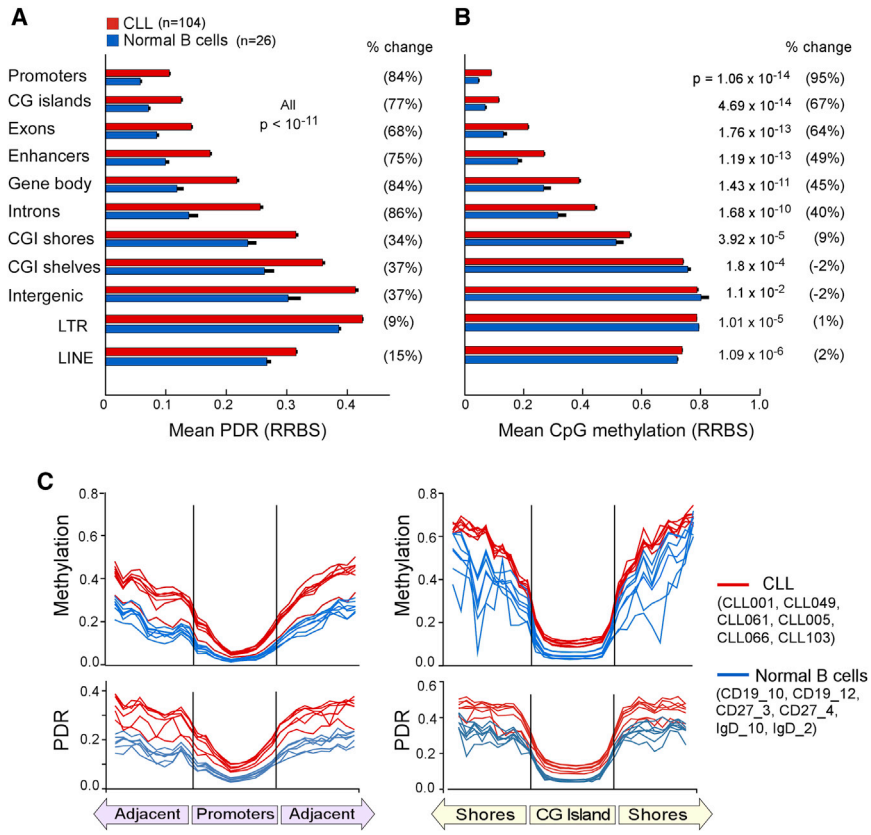


Figure 2. Locally Disordered Methylation Affects All Genomic Regions in CLL, Including CGIs and Repeat Regions

(A and B) Comparison of mean PDR (A) and mean CpG methylation (B) per genomic region between CLLs and normal B cells using RRBS data (Table S3 provides the average number of CpGs analyzed for each genomic region). Error bars represent upper 95% CI of the mean.

(C) Top: the distribution of PDR and methylation across all promoters covered by RRBS for randomly selected six CLL and six normal B cell samples. The distribution was derived by dividing each promoter into 100 bins and then averaging methylation and PDR for CpGs falling into each bin across all promoters in the sample. The PDR and methylation values in the adjacent 2KB upstream and downstream are also shown. Bottom: an analogous analysis of CGIs and adjacent shore regions.

See also Figure S2 and Table S3.

To model the relationship between methylation and PDR under completely stochastic conditions, we plotted the expected distribution of PDR for any level of methylation assuming a purely random assignment of methylation states at each individual CpG (Figure 3A; see Supplemental Experimental Procedures). Strikingly, the distribution of measured PDR and methylation values of ~14,000 individual promoter CGIs from CLL WGBS data closely followed the pattern of the modeled stochastic process (Figure 3B). In outlier genes (i.e., those with less promoter PDR than expected on the basis of the promoter methylation level; n = 195 [1.4%]; Table S4 and Figure S3D), imprinted genes were enriched (Morison et al., 2005) as expected, because these are hemimethylated under normal physiological conditions (n = 10, Fisher's exact test $p = 1.94 \times 10^{-6}$). In addition, the outlier genes contained at least three tumor suppressor genes (WIF1, DUSP22, and DCC) that have established roles in hematopoietic malignancies (Chim et al., 2008; Inokuchi et al., 1996; Jantus Lewintre et al., 2009) and also had >10% higher methylation in the CLL169 sample compared with the normal $CD19^+$ B cell sample.

Similar to promoters, methylation of ~1,900 LINE repeat elements also displayed a similar relationship between methylation and PDR (Figure 3C). A comparable distribution was observed for other genomic features (Figure S3E) and with RRBS data (Figure S3F). This pattern was also found in promoter CpGs of tumor suppressor genes implicated in lymphoproliferation, such as WT1 (Menke et al., 2002) and DAPK1 (Raval et al., 2007) (Figure S3G).

Altogether, these data support the hypothesis that the most commonly described cancer-related methylation alterations (Baylin and Jones, 2011)—increased methylation of CGIs and decreased methylation in repeat regions—are generated largely through a seemingly stochastic process. Indeed, across the 104 CLLs, sample average promoter CGI PDR was highly correlated with an increase in sample average promoter CGI methylation (Pearson's correlation coefficient r = 0.90, $p = 1.01 \times 10^{-38}$; Figure 3D). When this analysis was repeated with genes grouped on the basis of their average methylation level across the samples, this strong correlation was positive for genes with methylation < 0.5 and negative for genes with methylation > 0.5, as expected from the previously described distribution in Figure 3B (Figure S3H). Overall, a key implication of this analysis is that a change in CGI methylation in CLL does not arise from alteration in a relatively small proportion of cells with uniformly methylated alleles but rather from a larger proportion of cells with randomly scattered methylation. We likewise observed sample average LINE repeat elements PDR to be correlated with a decrease in methylation ($r = -0.32$, $p = 6.99 \times 10^{-4}$; Figure 3E).

These data reveal that DNA methylation changes in this cancer predominately arise from a disordered change in methylation, resulting in a strong correlation between difference in PDR (ΔPDR) and difference in methylation (ΔMeth). Because previous reports have indicated that a large degree of methylation disorder occurs during normal differentiation (Landan et al., 2012), we sought to compare the correlation between ΔPDR and ΔMeth among pairs of cancer and normal samples with the correlation between pairs of healthy human tissues. Indeed, the correlation coefficient between ΔPDR and ΔMeth was significantly higher when CLL samples were paired to either normal B cells or to other healthy primary tissue samples, compared with the pairing of healthy primary tissues against either normal B cells or other

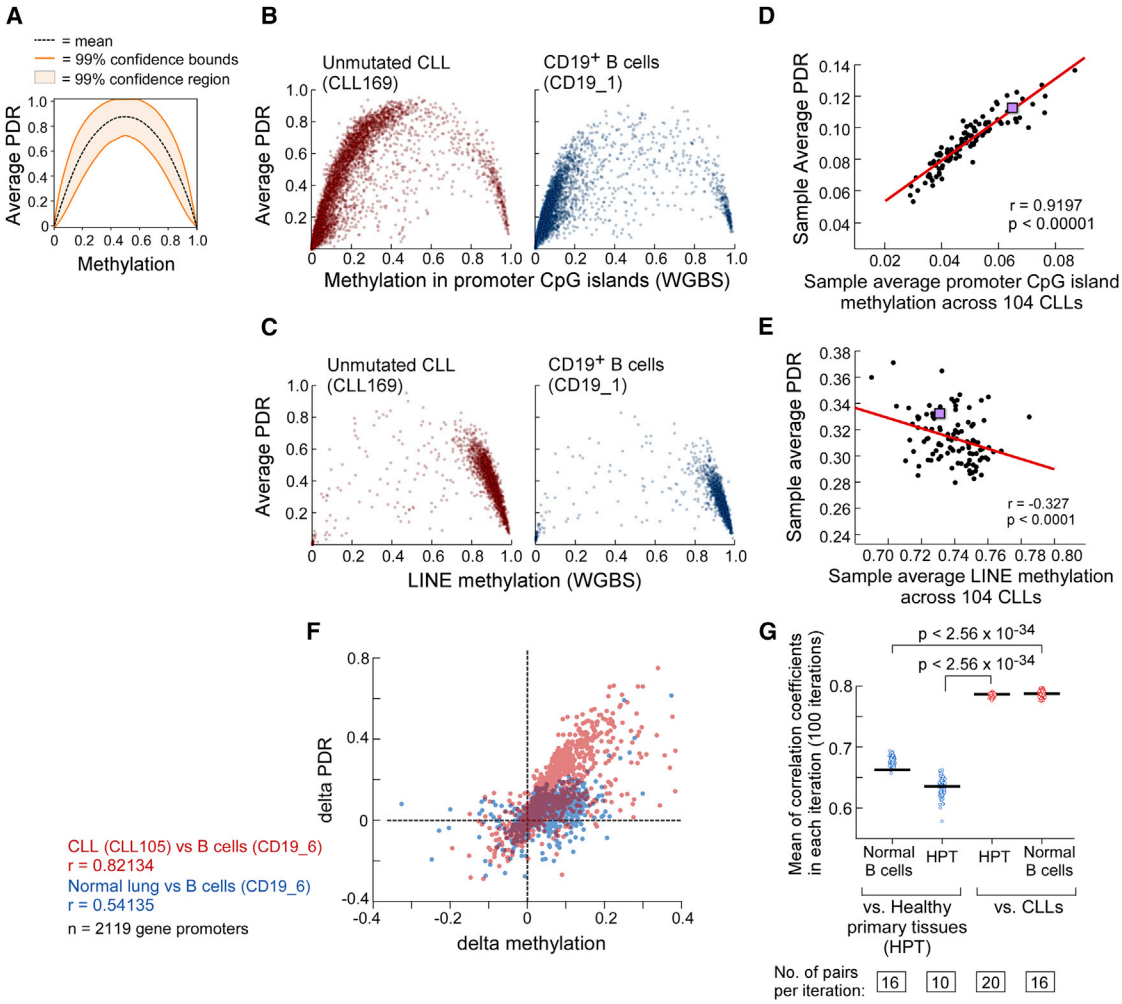


Figure 3. Locally Disordered Methylation in CLL Is Consistent with a Stochastic Process

(A) We developed a model to determine the probability of observing any PDR value in a random CpG methylation state model, given (1) the total number of reads that cover the locus, (2) the number of neighboring CpGs contained in individual reads, and (3) the locus methylation level. The plot demonstrates the case in which a locus is covered at a read depth of 30 and each read contains four neighboring CpGs. The expected PDR value is shown by the dashed line, and the shaded region represents methylation-PDR tuples with a probability greater than 0.01 under the random model.

(B) The CLL methylation data are consistent with the stochastic pattern shown in (A). Average promoter CGI methylation and PDR were calculated for 13,943 CGIs covered by WGBS (more than ten CpGs per island) in both the CLL and the normal B cell samples. Outliers represent 1.4% of events (see Figure S3D and Table S4).

(C) Average LINE element methylation and PDR were calculated for 1,894 elements covered by WGBS (>20 CpGs per element) in the same samples as in (B).

(D) The correlation in CLL between sample average of CGI methylation and PDR is shown (8,740.2 ± 3,102.8 promoter CGIs per sample were evaluated; see also Figure S3E).

(E) Similarly, the correlation in CLL between sample average LINE element methylation and PDR are also shown. The RRBS-based results of CLL169 are highlighted with a purple square.

(F) To study the correlation between ΔPDR and ΔMeth, we paired representative CLL and normal B cell samples. For each promoter (>20 CpGs per promoter, n = 2,119), ΔMeth and ΔPDR were plotted (red). An identical procedure was performed with a pairing of the same normal B cell sample to an adult lung sample (Lung_normal_BioSam_235, blue). These data enable the comparison between the Pearson's coefficient for the correlation between ΔPDR and ΔMeth in cancer-related changes versus normal physiological state changes.

(G) To confirm this finding across the entire data set, random pairings were performed in each category listed on the x axis, avoiding repeated use of any individual sample within a category. This procedure was repeated 100 times, and the means of the correlation coefficients for each iteration are plotted and compared. See also Figure S3 and Table S4.

healthy tissue samples (Figures 3F and 3G). Thus, methylation changes associated with the malignant process differ substantially from those that occur during changes in physiological cellular states and show a significantly higher degree of methylation disorder.

Increased Susceptibility to Locally Disordered Methylation in Gene-Poor Regions and Silent Genes

Some regions of the genome may be more prone to stochastic variation in methylation (Pujadas and Feinberg, 2012). We found 3-fold higher promoter PDR in regions with the lowest gene

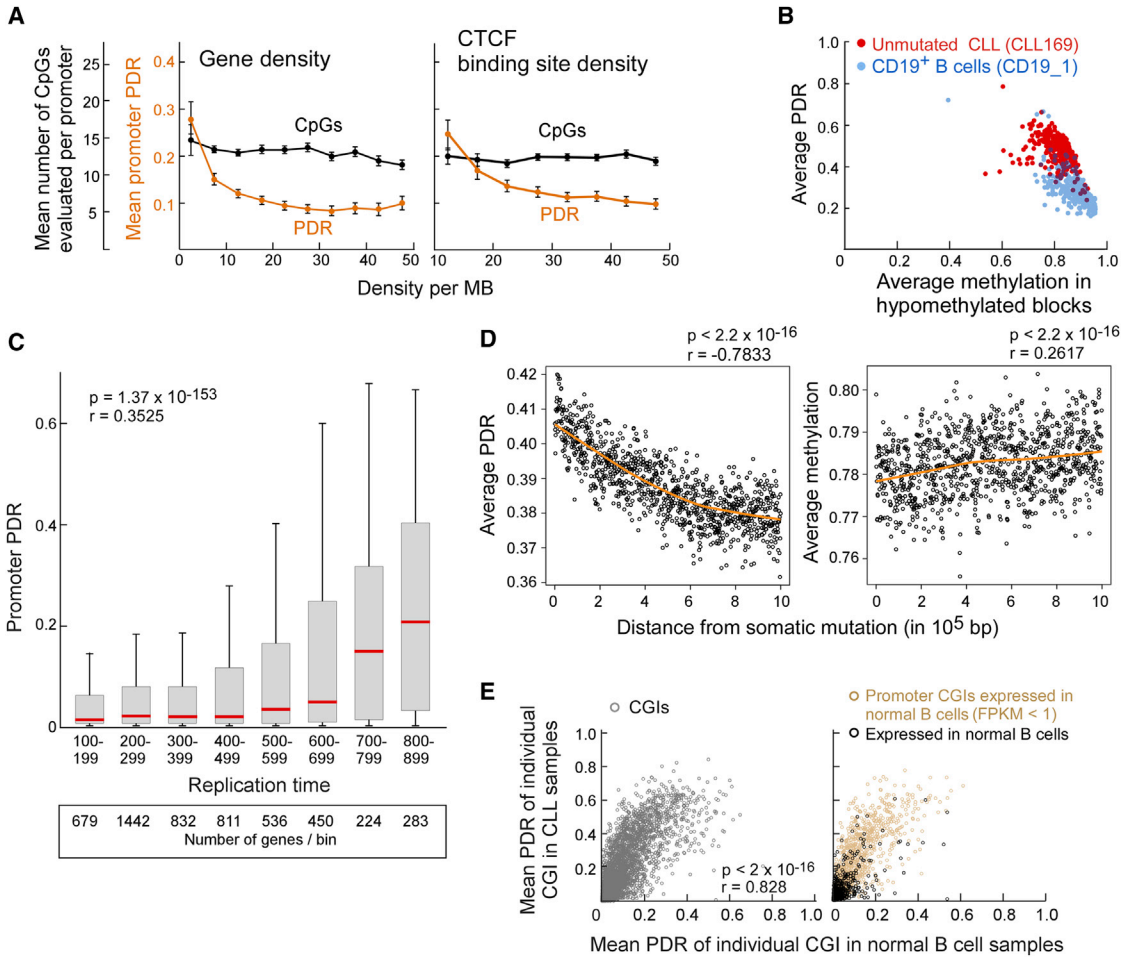


Figure 4. Locally Disordered Methylation Affects Preferentially Gene-Poor Regions and Can Be Traced Back to Nonexpressed Genes in Normal B Cells

(A) Promoter PDR (orange, error bars represent 95% CIs of means) in relation to gene density (genes/MB, left) and CTCF binding site density (right) regions. As reference, the CpG content is also provided (black).

(B) PDR and methylation in hypomethylated blocks (Hansen et al., 2011) is plotted for CLL and normal B cells (shown are blocks with >1,000 CpGs in WGBS; see also Figure S4A for comparison with a matched set of control genomic blocks).

(C) Replication time and PDR are correlated; PDR was averaged for each promoter covered in >70% of 104 CLLs, and these values were grouped in replication time bins.

(D) To assess the relationship between somatic mutations and PDR, sSNVs were identified with whole-genome sequencing of matched tumor and germline DNA (CLL169). Average PDR (left) and methylation (right) were measured in 1,000 bp increments from each somatic mutation. Values of CpGs in each 1,000 bp bin were averaged over 4,973 sSNVs and plotted as a function of the distance from the somatic mutation. Orange lines denote the locally weighted scatterplot smoothing. See Figures S4B and S4C for an analysis performed separately for clonal and subclonal mutations.

(E) Left: promoter CGI PDR is correlated between CLL and normal B cell samples (Pearson, evaluated with 5,811 consistently covered CGIs). Right: promoter CGI PDR in B cells and CLLs is shown for genes expressed and not expressed in normal B cells (FPKM < 1, n = 1,002 from RNA-seq data of seven healthy donor B cell samples).

See also Figure S4.

density compared with those with highest gene density (with similar correlations to CTCF density; Figure 4A). In addition, previously described hypomethylated blocks are regions notable for their association with the nuclear lamina and furthermore are enriched with genes that have high expression variability in cancer and impact critical cellular processes such as mitosis and cell cycle control (Hansen et al., 2011; Timp and Feinberg, 2013). In these regions as well, we observed a significant PDR increase in CLL (Figures 4B and S4A). Finally, in concert with these findings, we observed higher promoter PDR in genes with later replication time across the 104 CLL samples ($r = 0.35$, $p = 1.3 \times 10^{-153}$; Figure 4C), in agreement with other recent reports (Berman et al., 2012; Shipony et al., 2014). Notably, late replication time is closely associated with increased somatic mutation rate (Lawrence et al., 2013). Thus, similar genomic regions may share lower genetic and epigenetic fidelity, as we observed in a joint analysis of somatic single-nucleotide variants (sSNVs) and locally disordered methylation (Figures 4D, S4B, and S4C).

As many features of chromatin and spatial organization may be shared between the CLL and normal B cell genomes, we

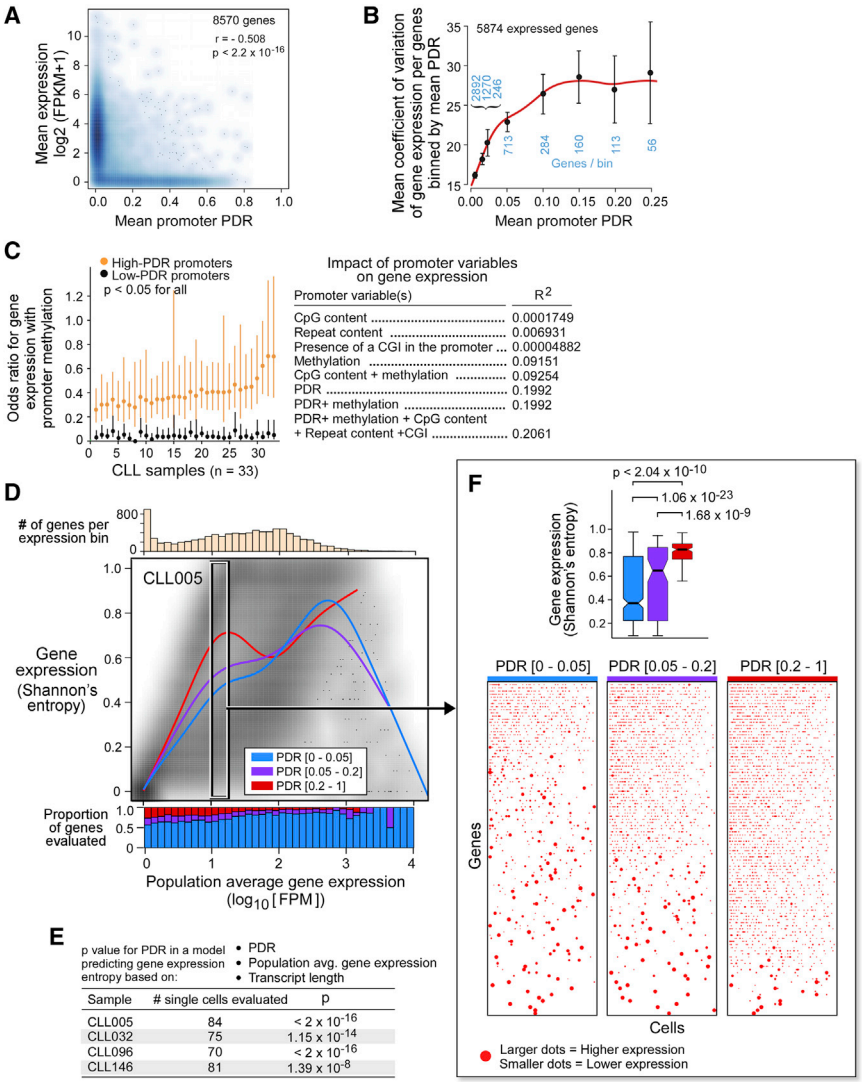


Figure 5. Locally Disordered Methylation Is Associated with Transcriptional Variation

(A) Mean promoter PDR and gene expression are correlated (evaluated with 8,570 genes that had promoter RRBS coverage in >70% of 33 samples with matched RRBS and RNA-seq, the number of genes evaluated within each expression range provided in Figure S5A, and mean expression and methylation correlation is provided in Figure S5B.

(B) PDR and expression variability as measured with CV of 5,874 transcribed genes (FPKM > 1). Black circles (brackets) denote mean CV (95% CI) for genes within PDR bins (number of genes per bin in blue). Red line is the cubic smoothing spline of CV and PDR values (unbinned). Note that the analysis was limited to transcribed genes to avoid an artificial enhancement of the CV that occurs with very low mean expression values. Because >97.5% of transcribed genes had PDR < 0.3, we limited the x axis to PDR < 0.3.

(C) Left: OR (bars denote 95% CI) for gene expression (FPKM > 1) with a methylated promoter (average methylation > 0.8) versus an unmethylated promoter (average methylation < 0.2) is calculated for genes with high (orange, 27.5 ± 2.6% of genes) or low promoter PDR (black). Right: linear models that combine information from all 33 CLLs as continuous variables to predict expression.

(D) PDR and intrasample gene expression heterogeneity (assessed by Shannon's information entropy) across the range of population average expression (fragments per million [FPM]), by single-cell RNA-seq of 84 cells from CLL005 (see Figure S5D for analysis of three additional CLL samples). Local regression lines for genes with low PDR (0–0.05, blue), intermediate PDR (0.05–0.2, purple), and high PDR (0.2–1.0, red) are shown.

(E) Results of generalized additive regression tests that model single-cell gene expression Shannon's information entropy on the basis of PDR, population average expression, and transcript length across the four CLL samples.

(F) Single-cell gene expression patterns for genes within a narrow population average expression range of 1.0 to 1.2 (black rectangle in D). Consistent with the higher gene expression Shannon's information entropy observed in genes with higher PDR (top), genes with low PDR (bottom left) tend to be expressed at high magnitude (larger dot size) in fewer cells, whereas genes with high PDR (bottom right) are frequently expressed at low expression magnitudes across many cells.

See also Figure S5 and Tables S5 and S6.

hypothesized that some degree of locally disordered methylation might exist in normal B cells in regions with high PDR in CLL. In fact, average PDR of individual CGI in CLL and B cell samples was highly correlated ($r = 0.83$, $p < 2 \times 10^{-16}$; Figure 4E, left). Thus, the promoters with highest PDR in CLL already have increased PDR in normal B cells. Consistent with the notion that nonexpressed genes are the most vulnerable to aberrant methylation (Meissner et al., 2008), promoter CGIs with a high PDR in both CLL and normal B cells were often found in genes not expressed in normal B cells (Figure 4E, right).

Locally Disordered Methylation and Gene Expression

To examine the relationship between locally disordered DNA methylation and gene expression in more detail, we analyzed matched RRBS and RNA sequencing (RNA-seq) profiles of 33 CLL samples (Table S5; PDR and methylation calculated on the basis of an average ± SD of 12.1 ± 4.8 CpGs per promoter). As in normal B cells, in the 33 CLL samples, PDR was inversely correlated with gene expression ($r = -0.51$, $p < 2 \times 10^{-16}$; Figures 5A, S5A, and S5B). Notably, whereas promoter PDR was negatively correlated with mean transcript levels, it was positively correlated with intersample variation in transcript levels (Figure 5B). Although it may be difficult to definitively deconvolute the positive correlation between PDR and expression variation from the strong negative correlation of mean expression and expression variation, both low gene expression and high promoter PDR levels were predictive of higher coefficient of variation (CV) of gene expression in a linear model ($p < 2 \times 10^{-16}$ for both).

To further examine the impact of locally disordered methylation in CLL on expression levels, we calculated the odds ratio

(OR) of gene expression (defined as fragments per kilobase of exon per million fragments mapped (FPKM) > 1) with a methylated promoter (defined as methylation > 0.8, unmethylated defined as < 0.2). Promoters with low PDR (i.e., lower than the mean PDR [mean ± SD promoter PDR was 0.10 ± 0.01]) tended to preserve the expected relationship between promoter methylation and expression and rarely generated transcripts in the presence of a methylated promoter. Across 33 CLL samples, the average OR was 0.043 (range 0.036–0.050). In contrast, genes with high PDR promoters (greater than the mean PDR) had a greater likelihood of undergoing transcription (OR 0.396, range 0.259–0.698, Wilcoxon $p = 6.5 \times 10^{-11}$; Figure 5C), despite comparable promoter methylation levels. As a representative example, we show ZNF718 in two samples with comparable levels of promoter methylation (0.82 in CLL062, 0.87 in CLL074) but low promoter PDR (0.04) in the former and high promoter PDR (0.24) in the latter. Consistent with the OR analysis above, we observed undetectable expression in CLL062 (FPKM of 0.03) and measurable RNA expression in CLL074 (FPKM of 5.6) (Figure S5C).

These observations demonstrate how locally disordered methylation and epigenetic heterogeneity may contribute to increased transcriptional variation. To assess the relationship between PDR and gene expression as continuous variables, we used linear models to predict expression on the basis of methylation information. Across the 33 samples, a univariate model that predicts expression on the basis of average promoter methylation yielded an adjusted R^2 value of 0.092, whereas one using promoter PDR yielded an average adjusted R^2 value of 0.202. Inclusion of additional features such as CpG and repeat content only modestly improved the predictive power of the model (average adjusted $R^2 = 0.214$; Table S6). Indeed, the addition of PDR information to a model that uses promoter methylation to predict gene expression as a continuous variable (evaluated for 320,574 matched values of expression and methylation from 33 CLLs) resulted in a significant improvement, with more than doubling of the model's explanatory power (increase in adjusted R^2 value from 0.0915 to 0.1992, likelihood ratio test $p < 1 \times 10^{-16}$). This held true when the model included only genes with lowly methylated or only genes with highly methylated promoters ($p < 1 \times 10^{-16}$). Even after adding additional variables such as repeat element content, the presence of a CGI in the promoter, and CpG content, PDR remained the strongest predictor of expression (Figure 5C, right).

Single-Cell Gene Expression Patterns of Genes with Disordered Promoter Methylation

We next isolated 96 individual cells from four $CD19^+CD5^+$ purified CLL samples and generated single-cell full-length transcriptomes using SMART-seq (Clontech; 75–84 cells analyzed per sample after excluding cells with $< 1 \times 10^4$ aligned reads; Table S2). Promoter PDR was associated with significantly higher intratumoral expression information entropy in all four samples ($p < 1.4 \times 10^{-8}$; Figures 5D, 5E, and S5D), in a model that included transcript length as well as population average gene expression (see Supplemental Experimental Procedures), which is the variable associated most closely with technical noise in single-cell transcriptome analyses (Shalek et al., 2014). These results remained significant even after the addition of promoter methylation to the model (Figure S5E). Because expression information entropy may be affected by variation in sampling of lowly expressed transcripts, we compared the single-cell expression patterns of genes with low or high promoter methylation disorder but with similar population average expression levels (Figure 5F). We observed that high promoter PDR genes tend to be expressed in larger numbers of cells at lower expression magnitude, whereas low promoter PDR genes tend to be expressed in smaller numbers of cells at higher expression magnitude. Thus, promoter methylation disorder correlates with an intermediate transcriptional state that interferes with both complete silencing and high-level expression.

Locally Disordered Methylation Affects Stem Cell Genes and May Facilitate Leukemic Evolution

Increased epigenetic "noise" would be expected to generate a more plastic evolutionary landscape that facilitates the emergence of fitness-enhancing genetic and epigenetic alterations. To explore the potential relationship between locally disordered methylation and selection, we identified differentially methylated regions (DMRs) in promoters and CGIs, because the presence of recurrent epigenetic alterations might signal the presence of evolutionary convergence. In fact, these DMRs were associated with significantly higher PDR, suggestive of positive selection operating against a backdrop of stochastic epigenetic heterogeneity (Figure S6A).

Furthermore, a gene set enrichment analysis of genes with consistently high promoter PDR across CLL samples compared with genes with consistently low promoter PDR revealed enrichment in TP53 targets (Perez et al., 2007), in genes differentially methylated across various malignancies (Acevedo et al., 2008; Sato et al., 2003), and in gene sets associated with stem cell biology (Lim et al., 2010; Wong et al., 2008) (BH-FDR Q < 0.1; Figures 6A and S6B; Table S7). Finally, regions that are specifically hypomethylated in human embryonic stem cells compared with a diverse collection of differentiated cells (Ziller et al., 2013) also showed decreased methylation and increased PDR in CLL compared with normal B cells, suggestive of a drift toward a more stem-cell-like state (Figure 6B). Collectively, these findings suggest that locally disordered methylation creates a rich substrate for CLL evolution by stochastic variation amenable to positive selection and by increasing the number of cells that carry the potential to propagate new genotypes to progeny populations. Indeed, CLLs with a higher number of subclonal mutations also exhibit higher PDR ($p = 0.002$; Figure 6C).

To directly observe the relationship between genetic and epigenetic evolution, we studied RRBS data from 14 longitudinally sampled CLL patients with characterized patterns of genetic evolution (median time between samples 3.45 years; 9 CLLs with and 5 without evidence of genetic evolution; Table S8). CLLs that underwent genetic clonal evolution also had increased average promoter PDR over time (paired t test, $p = 0.037$; Figure 6D), which may indicate a higher PDR in the subclone that expanded over time. In addition, genes with promoters that were demethylated over time, were significantly enriched for the same aforementioned stem cell-related gene sets (Boquest et al., 2005; Jaatinen et al., 2006; Lim et al., 2010; Wong et al., 2008) (Figure 6E; Table S9). Importantly, the correlation

A

Gene sets enriched in genes with high promoter PDR	Q_{high}	$Q_{high\ vs\ low}$
TP63 targets (Perez)	4.83×10^{-25}	1.80×10^{-2}
TP53/TP63 targets (Perez)	4.05×10^{-18}	5.70×10^{-2}
Silenced by methylation in pancreatic cancer 1 (Sato)	1.98×10^{-28}	3.83×10^{-7}
Ewing sarcoma progenitor UP (Riggi)	3.68×10^{-23}	7.63×10^{-3}
Methylated in liver cancer DN (Acevedo)	4.91×10^{-22}	2.99×10^{-7}
Adult tissue stem module (Wong)	1.52×10^{-36}	1.12×10^{-2}
UVC response via TP53 group A (Bruins)	6.89×10^{-26}	4.22×10^{-2}
Mammary stem cell UP (Lim)	5.42×10^{-38}	2.93×10^{-7}
Methylated in cancer (Hoque)	3.78×10^{-10}	4.22×10^{-2}
Neuronal crest stem cell UP (Lim)	2.02×10^{-9}	4.18×10^{-2}

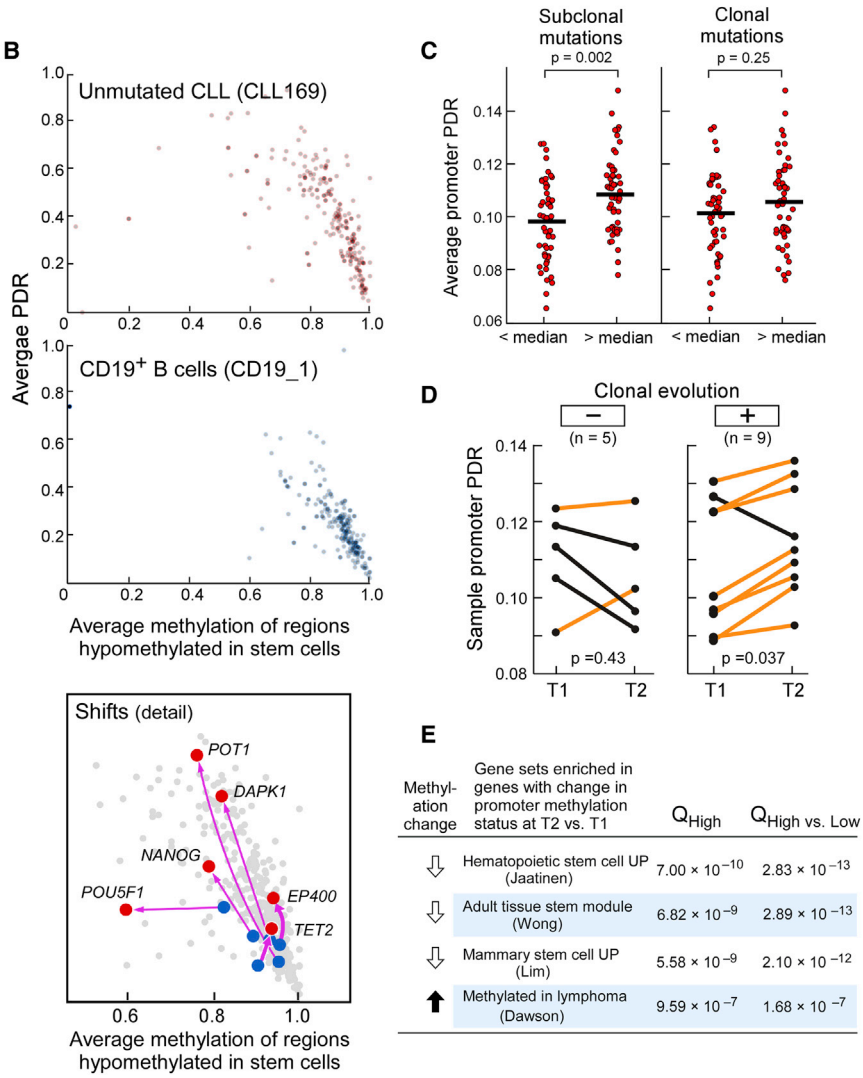


Figure 6. Locally Disordered Methylation May Interact with Evolution through Drift toward a Stem-like State

(A) Gene set enrichment analysis comparing 1,668 genes with consistently high promoter PDR (>0.1 in >75% of samples) with 5,392 genes with consistently low promoter PDR (<0.1 in >75% of samples, selected ten gene sets displayed; see Table S7 for the top 30 enrichments). Enrichment in genes with consistently high PDR was calculated for hypergeometric distribution followed by BH-FDR ("Q(high)"). In addition, enrichment in high-PDR genes versus low-PDR genes was calculated using Fisher's exact test followed by BH-FDR ("Q(high versus low)").

(B) PDR and methylation in regions hypomethylated in embryonic stem cells (Ziller et al., 2013), in CLL compared with normal B cells (WGBS data). Regions include 91 enhancers (e.g., *POU5F1*, *NANOG*), 41 enhancer CGIs (e.g., *TET2*, *EP400*), six CGIs (e.g., *DAPK1*), six promoters, and 84 other putative regulatory elements (e.g., *DEC1* and *POT1*) (Ziller et al., 2013). The inset shows individual changes of selected regions.

(C) PDR in CLLs with high versus low numbers of subclonal (median 7.5 sSNVs) and clonal mutations (median 10 sSNVs).

(D) Fourteen CLLs were sampled longitudinally at two time points (T1 and T2; median interval time 3.5 years), and change in PDR over time was compared between CLLs that underwent genetic clonal evolution (n = 9) and those without genetic evolution (n = 5) (paired t test).

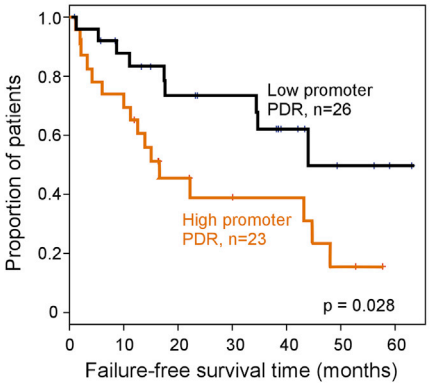
(E) Gene set enrichment of the 899 genes from the 14 cases with significant promoter methylation change between time points T1 and T2 (absolute change > 10%, FDR BH Q < 0.1) in genes with promoter demethylation over time (456 genes), and in genes with promoter methylation over time (443 genes; see Table S9 for top 30 enrichments). See also Figure S6 and Tables S7–S9.

coefficient between ΔPDR and ΔMeth was markedly lower for gene promoters that were significantly demethylated or hypermethylated over time (r = 0.0937 and r = 0.0987, respectively), compared with the correlation coefficient for gene promoters without significant changes in methylation (r = 0.4163; 144,161 promoters across 14 CLLs). These results suggest that gene promoters with significant changes in methylation over time were enriched for genes that underwent ordered methylation change, as expected from positive selection.

Locally Disordered Methylation Affects Clinical Outcome

The presented data support a model in which locally disordered DNA methylation facilitates tumor evolution through increased genetic and epigenetic plasticity. Thus, we hypothesized that increased PDR would be associated with a shorter remission time after treatment, which we previously linked with clonal evolution (Landau et al., 2013).

We therefore examined failure-free survival after treatment (FFS; failure defined as retreatment or death) in 49 patients included in the cohort who were treated after tumor sampling for RRBS. A higher mean sample promoter PDR (greater than the mean for the cohort) was significantly associated with shorter FFS (median FFS of 16.5 versus 44 months, hazard ratio 2.5, 95% confidence interval [CI] 1.1 to 5.7, p = 0.028, Figure 7A; 52% and 65% of patients, respectively, were treated with fludarabine-based immunochemotherapy, p = 0.39). A regression



B

Variable	HR	95%CI	p
High promoter PDR	2.81	1.05 - 7.53	0.039
IGHV unmutated status	4.55	1.45 - 14.20	0.009
del(17p)	2.50	0.83 - 7.50	0.099
del(11q)	1.19	0.41- 3.45	0.746
⬇			
High promoter PDR	1.44	0.49 - 4.19	0.503
IGHV unmutated status	3.69	1.15 - 11.74	0.027
del(17p)	3.20	0.99 - 10.53	0.050
del(11q)	1.09	0.33 - 3.04	0.98
subclonal driver present	4.086	1.31 - 12.73	0.015

Figure 7. Locally Disordered Methylation Is Associated with Adverse Clinical Outcome

(A) Kaplan-Meier plot showing FFS time (failure defined as retreatment or death from the time of first therapy after RRBS analysis) in CLLs with higher versus lower than average promoter PDR. Note that the analysis could be performed only for the 49 patients who received therapy after RRBS sampling.

(B) Multivariate analysis for this association with the addition of well-established poor outcome predictors in CLL (*IGHV* unmutated status, *del*[17p] and *del*[11q]), as well as with the addition of the presence of a subclonal driver (including somatic copy number changes, sSNVs, and indels), as previously described (Landau et al., 2013), to the model.

See also Table S10.

model including established CLL risk indicators (IGHV unmutated status, del[17p] and del[11q]) showed an adjusted hazard ratio of 2.81 (95% CI 1.05–7.53, p = 0.039, Figure 7B) for high promoter PDR. Similar results were obtained after the inclusion of additional variables in the model, including mutation burden and average promoter methylation (Table S10). Samples with higher promoter PDR were also more likely to have a subclonal driver mutation as previously defined (Landau et al., 2013) (p = 0.01). When the presence of a subclonal driver was added to the regression model, the increased risk associated with the elevated PDR was no longer preserved (Figure 7B). These results support the notion that epigenetic "noise" may function primarily as a facilitating feature, allowing the emergence of subclonal drivers, which then contribute to the adverse clinical outcome.

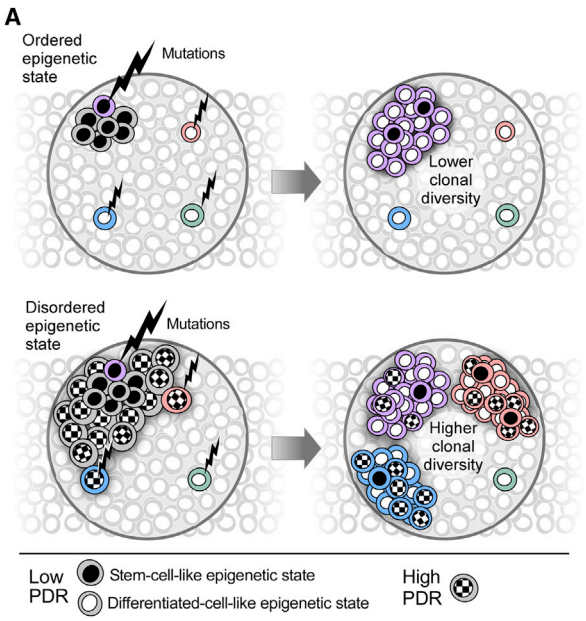
DISCUSSION

Cancer epigenomes have been long appreciated to differ from their normal tissue counterparts (Baylin and Jones, 2011). Global hypomethylation of cancer DNA was described as early as the 1980s, with frequent focal hypermethylation of key regulatory regions (Jones and Baylin, 2007). Recent genome-wide mapping have further highlighted alterations likely to contribute to the malignant process such as the epigenetic silencing of tumor suppressor genes and the activation of genes in stem-like cellular programs (Akiyama et al., 2003; Jones and Baylin, 2007; Widschwendter et al., 2007).

We now report the analysis of DNA methylation in primary leukemia cells that reveals another fundamental difference between cancer and normal methylomes: locally disordered methylation arising from a stochastic process, which leads to a high degree of intrasample methylation heterogeneity. These findings further advance key concepts described in several prior reports (Berman et al., 2012; Hansen et al., 2011; Landan et al., 2012; Maegawa et al., 2014; Pujadas and Feinberg, 2012; Siegmund et al., 2009). Thus, as previously suggested (Timp and Feinberg, 2013), cancer epigenomes may accommodate a higher amplitude of epigenetic "noise" and thereby allow cancer cells a greater degree of population diversity. Analogous to the role of genetic instability, which fuels cancer plasticity by facilitating the acquisition of somatic alterations at random locations across the genome (Hanahan and Weinberg, 2011), we propose that stochastic methylation changes enhance epigenetic plasticity and likewise enable tumor cells to better explore the evolutionary space in search of superior fitness trajectories.

These data alter the way we understand differential methylation in cancer. First, the insight that stochastic variation underlies the bulk of CLL methylome heterogeneity signifies that changes in methylation measured between cancer and normal cells do not likely reflect a uniform change in methylation state of a given region but rather a disordered methylation change involving differing, isolated CpGs, affecting many cells in the cancer population. Second, these data suggest improved methods from which we can identify fitness-enhancing DMRs. We can draw from the lessons of the computational analyses of large cancer genome sequencing data sets, in which a better understanding of the variation in the distribution of gene mutations has led to an improved ability to distinguish "driver" from "passenger" mutations (Lawrence et al., 2013). In an analogous fashion, we anticipate that appreciation of the extent of locally disordered methylation provides an appropriate background model against which a departure from the stochastic regime would indicate positively selected DMRs. We note that only a small proportion of methylation events fall outside the predictions of the stochastic model, suggesting very few of the changes in methylation undergo positive selection.

These data moreover demonstrate that locally disordered methylation is associated with a more "noisy" transcriptional landscape, with a decoupling of the relationship between promoter methylation and gene expression. Our analysis suggests that some of the epigenetic variability is likely associated with stemlike cell programs, which have been implicated in cancer (Kim et al., 2010; Ohnishi et al., 2014). Indeed, we detected a concurrent decrease in methylation and an increase in PDR, affecting regions that were identified to be hypomethylated in human embryonic stem cells, consistent with the notion that stochastic noise may lead to a drift toward a hybrid stem-somatic cell state (Timp and Feinberg, 2013). Furthermore, in CLLs that were directly observed to undergo genetic diversification and evolution over time, stem cell-related genes with higher promoter PDR also underwent demethylation over time. Thus, increased stochastic variation may blur the lines between populations with different proliferative potentials and thus increase



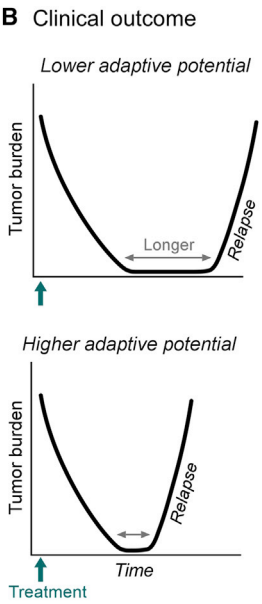


Figure 8. Proposed Interaction between Methylation Disorder and Clonal Evolution

A novel somatic mutation (depicted with lightning bolts) would have to coincide with an epigenetic state that will be permissive to the propagation of the new genotype to a progeny population. In a cellular population with limited stochastic methylation changes (top), the proportion of cells that are therefore able to actively participate in the evolutionary process is small. However, in a more malleable epigenetic landscape, such as expected to result from a high level of locally disordered methylation, a greater proportion of cells can give birth to new subclones, increasing the diversity and the adaptive capacity of the cancer population, resulting in adverse clinical outcome with therapy.

the diversity of adaptive mechanisms available to the cell, a hedging strategy for enhanced survival (Balázsi et al., 2011).

A further extension of this model proposes that locally disorder methylation enhances the evolutionary capacity of CLL by optimizing the process of genetic diversification. This framework would necessitate coincidence of a novel somatic mutation with an epigenetic state permissive to the propagation of the new genotype to a progeny population. In cellular populations with a preserved epigenetic landscape (Figure 8, top), the proportion of cells capable of actively participating in the evolutionary process is predicted to be small. On the other hand, in a more malleable epigenetic landscape (Figure 8, bottom) as is expected with a high level of locally disordered methylation, a greater proportion of cells can give birth to new subclones. This process would accelerate genetic evolution, provide a greater adaptive capacity for the cancer population, and result in adverse clinical outcome with therapy, as we saw in our CLL cohort.

What is the basis of increased locally disordered methylation in CLL? Although the exact mechanism remains to be fully elucidated, we speculate that the considerably higher replication rate in CLL compared with their normal differentiated counterparts could contribute to accumulation of stochastic lapses in methylation inheritance in cancer cells, given the estimated error rate of 0.08% to 4% for a given CpG per cell division (Bird, 2002; Ushijima et al., 2003). This maybe further compounded by the occurrence of genetic lesions in essential components of the methylation machinery. In addition, the finding that locally disorder methylation in CLL tended to be highest in gene-poor and late-replicating regions suggests that some genomic regions exhibit even higher error rates, consistent with the previously observed high cancer intersample methylation variability in these regions (Hansen et al., 2011).

Our data suggest that evolution and diversity of DNA methylation in CLL result from stochastic events. This insight should improve our model for background methylation changes in cancer and allow more rigorous identification of positively selected methylated regions. Locally disordered DNA methylation is likely to have a similar role to genetic instability, providing a mechanism for cancer cells to find superior evolutionary trajectories during tumorigenesis and in response to therapy.

EXPERIMENTAL PROCEDURES

Sample Acquisition

Peripheral blood samples were obtained from patients with CLL and healthy adult volunteers. Informed consent on Dana-Farber Cancer Institute institutional review board-approved protocols for genomic sequencing of patients' samples was obtained prior to the initiation of sequencing studies. Genomic DNA was extracted from CLL cells or normal B cell populations.

WGBS

WGBS was performed as described in Supplemental Experimental Procedures. Subsequently, CpG methylation calls were made using custom software, excluding duplicate and low-quality reads. Previously published WGBS data for two CLL samples and three normal B cell samples (Kulis et al., 2012) were downloaded with permission and processed in identical fashion to the in-house-produced WGBS libraries.

RRBS

RRBS was performed as described in Supplemental Experimental Procedures. RRBS of primary diverse human tissue samples were previously reported (http://www.roadmapepigenomics.org). Reads were aligned, and methylation was determined using identical protocols to the rest of the samples.

RNA-Seq

RNA-seq of CLL and normal B cell samples was performed as previously described (Landau et al., 2013). For single-cell RNA-seq, the C1 Single-Cell Auto Prep System (Fluidigm) was used to perform SMARTer (Clontech) whole-transcriptome amplification (WTA), on up to 96 individual cells per sample from four primary CLL patient samples. WTA products were then converted to Illumina sequencing libraries using Nextera XT (Illumina) (Ramsköld et al., 2012).

Statistical Analysis

Statistical analysis was performed with MATLAB (The MathWorks), R version 2.15.2 (R Foundation for Statistical Computing), and SAS version 9.2 (SAS Institute). A complete description of the materials and methods is provided in Supplemental Experimental Procedures. The CLL and normal B cell

sequencing data were deposited in the database of Genotypes and Phenotypes (dbGaP) (phs000435.v2.p1) and the processed data deposited in Gene Expression Omnibus (GEO) (GSE58889).

ACCESSION NUMBERS

The GEO accession number for the data reported in this paper is GSE58889. The dbGaP accession number for the sequencing data reported in this paper is phs000435.v2.p1.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and ten tables and can be found with this article online at http://dx.doi.org/10.1016/j.ccell.2014.10.012.

AUTHOR CONTRIBUTIONS

D.A.L, K.C., N.H., G.G., A.M., and C.J.W. conceived and designed the experiments. D.A.L, P.B., H.G., L.W., D.K., W.Z., K.J.L., S.L., and A.G. performed the experiments. D.A.L, K.C., M.Z., J.F., K.S., M.G., D.N., and P.V.K. analyzed the data. Additional contribution to funding and sample collection and processing was provided by C.S., L.G., E.S.L., S.M.F., and J.R.B. The paper was written by D.A.L, K.C., A.M., and C.J.W.

ACKNOWLEDGMENTS

We thank all members of the Broad Institute's Biological Samples, Genetic Analysis, and Genome Sequencing Platforms, who made this work possible (NHGRI-U54HG003067). We thank Rahul Satija, Angela Brooks, Scott Carter and Brad Bernstein for their valuable input and insights. We thank John Daley, Suzan Lazo-Kallanian, Jonna Grimsby, and Niall Lennon for their assistance in the single-cell RNA-seq. We thank Adam Kiezun for his guidance regarding germline SNP detection and Michael Lawrence for the replication time data. D.A.L. is supported by a Postdoctoral Fellowship from the American Cancer Society (ACS) and by the NIH Big Data to Knowledge initiative (BD2K, 1K01ES025431-01). K.C. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 112237. L.W. is supported by a Lymphoma Research Foundation Postdoctoral Fellowship. J.F. is supported by the National Science Foundation Graduate Research Fellowship (DGE1144152). J.R.B. is a Scholar in Clinical Research of the Leukemia & Lymphoma Society (LLS) and is supported by a LLS Translational Research Program award as well as an ACS Research Scholar Grant and the Melton and Rosenbach Funds. A.M. is a New York Stem Cell Foundation Robertson Investigator and received support from NIH grants U01ES017155 and 1R01DA036898. C.J.W. is a Scholar of LLS and the recipient of a Quest for Cures Award from LLS. She acknowledges support from the Blavatnik Family Foundation, the American Association for Cancer Research (Stand Up to Cancer Innovative Research Grant), the National Heart, Lung, and Blood Institute (1RO1HL103532-01), and the National Cancer Institute (1R01CA155010-01A1). K.J.L. and S.L. are employees of Fluidigm Corporation.

Received: March 4, 2014
Revised: September 16, 2014
Accepted: October 24, 2014
Published: December 8, 2014

REFERENCES

Acevedo, L.G., Bieda, M., Green, R., and Farnham, P.J. (2008). Analysis of the mechanisms mediating tumor-specific changes in gene expression in human liver tumors. Cancer Res. *68*, 2641–2651.

Akiyama, Y., Watkins, N., Suzuki, H., Jair, K.W., van Engeland, M., Esteller, M., Sakai, H., Ren, C.Y., Yuasa, Y., Herman, J.G., and Baylin, S.B. (2003). GATA-4 and GATA-5 transcription factor genes and potential downstream antitumor target genes are epigenetically silenced in colorectal and gastric cancer. Mol. Cell. Biol. *23*, 8429–8439.

Balázsi, G., van Oudenaarden, A., and Collins, J.J. (2011). Cellular decision making and biological noise: from microbes to mammals. Cell *144*, 910–925.

Baylin, S.B. (2005). DNA methylation and gene silencing in cancer. Nat. Clin. Pract. Oncol. *2* (*Suppl 1*), S4–S11.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nat. Rev. Cancer *11*, 726–734.

Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C.P., van Dijk, C.M., Tollenaar, R.A., et al. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat. Genet. *44*, 40–46.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Boquest, A.C., Shahdadfar, A., Frønsdal, K., Sigurjonsson, O., Tunheim, S.H., Collas, P., and Brinchmann, J.E. (2005). Isolation and transcription profiling of purified uncultured human stromal stem cells: alteration of gene expression after in vitro cell culture. Mol. Biol. Cell *16*, 1131–1141.

Boyle, P., Clement, K., Gu, H., Smith, Z.D., Ziller, M., Fostel, J.L., Holmes, L., Meldrim, J., Kelley, F., Gnirke, A., and Meissner, A. (2012). Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol. *13*, R92.

Brown, J.R., Hanna, M., Tesar, B., Werner, L., Pochet, N., Asara, J.M., Wang, Y.E., Dal Cin, P., Fernandes, S.M., Thompson, C., et al. (2012). Integrative genomic analysis implicates gain of PIK3CA at 3q26 and MYC at 8q24 in chronic lymphocytic leukemia. Clin. Cancer Res. *18*, 3791–3802.

Cahill, N., Bergh, A.C., Kanduri, M., Göransson-Kultima, H., Mansouri, L., Isaksson, A., Ryan, F., Smedby, K.E., Juliusson, G., Sundström, C., et al. (2013). 450K-array analysis of chronic lymphocytic leukemia cells reveals global DNA methylation to be relatively stable over time and similar in resting and proliferative compartments. Leukemia *27*, 150–158.

Chim, C.S., Pang, R., and Liang, R. (2008). Epigenetic dysregulation of the Wnt signalling pathway in chronic lymphocytic leukaemia. J. Clin. Pathol. *61*, 1214–1219.

De, S., Shaknovich, R., Riester, M., Elemento, O., Geng, H., Kormaksson, M., Jiang, Y., Woolcock, B., Johnson, N., Polo, J.M., et al. (2013). Aberration in DNA methylation in B-cell lymphomas has a complex origin and increases with disease severity. PLoS Genet. *9*, e1003137.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. *38*, 1378–1385.

Ehrlich, M. (2009). DNA hypomethylation in cancer cells. Epigenomics *1*, 239–259.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nat. Genet. *43*, 768–775.

Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L., Johnson, B.E., Fouse, S.D., Delaney, A., Zhao, Y., et al. (2010). Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat. Biotechnol. *28*, 1097–1105.

Inokuchi, K., Miyake, K., Takahashi, H., Dan, K., and Nomura, T. (1996). DCC protein expression in hematopoietic cell populations and its relation to leukemogenesis. J. Clin. Invest. *97*, 852–857.

Jaatinen, T., Hemmoranta, H., Hautaniemi, S., Niemi, J., Nicorici, D., Laine, J., Yli-Harja, O., and Partanen, J. (2006). Global gene expression profile of human cord blood-derived CD133+ cells. Stem Cells *24*, 631–641.

Jantus Lewintre, E., Reinoso Martín, C., Montaner, D., Marín, M., José Terol, M., Farrás, R., Benet, I., Calvete, J.J., Dopazo, J., and García-Conde, J. (2009). Analysis of chronic lymphotic leukemia transcriptomic profile: differences between molecular subgroups. Leuk. Lymphoma *50*, 68–79.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Jones, P.A., and Baylin, S.B. (2007). The epigenomics of cancer. Cell *128*, 683–692.

Kim, J., Woo, A.J., Chu, J., Snow, J.W., Fujiwara, Y., Kim, C.G., Cantor, A.B., and Orkin, S.H. (2010). A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. Cell *143*, 313–324.

Kreso, A., O'Brien, C.A., van Galen, P., Gan, O.I., Notta, F., Brown, A.M., Ng, K., Ma, J., Wienholds, E., Dunant, C., et al. (2013). Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. Science *339*, 543–548.

Kulis, M., Heath, S., Bibikova, M., Queirós, A.C., Navarro, A., Clot, G., Martínez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., et al. (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat. Genet. *44*, 1236–1242.

Landan, G., Cohen, N.M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D.A., Goldfinger, N., Zundelevich, A., et al. (2012). Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat. Genet. *44*, 1207–1214.

Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714–726.

Landau, D.A., Carter, S.L., Getz, G., and Wu, C.J. (2014). Clonal evolution in hematological malignancies and therapeutic implications. Leukemia *28*, 34–43.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., Kandoth, C., Payton, J.E., Baty, J., Welch, J., et al. (2010). DNMT3A mutations in acute myeloid leukemia. N. Engl. J. Med. *363*, 2424–2433.

Lim, E., Wu, D., Pal, B., Bouras, T., Asselin-Labat, M.L., Vaillant, F., Yagita, H., Lindeman, G.J., Smyth, G.K., and Visvader, J.E. (2010). Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. Breast Cancer Res. *12*, R21.

Maegawa, S., Gough, S.M., Watanabe-Okochi, N., Lu, Y., Zhang, N., Castoro, R.J., Estecio, M.R., Jelinek, J., Liang, S., Kitamura, T., et al. (2014). Age-related epigenetic drift in the pathogenesis of MDS and AML. Genome Res. *24*, 580–591.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770.

Menke, A.L., Clarke, A.R., Leitch, A., Ijpenberg, A., Williamson, K.A., Spraggon, L., Harrison, D.J., and Hastie, N.D. (2002). Genetic interactions between the Wilms' tumor 1 gene and the p53 gene. Cancer Res. *62*, 6615–6620.

Morison, I.M., Ramsay, J.P., and Spencer, H.G. (2005). A census of mammalian imprinting. Trends Genet. *21*, 457–465.

Ohnishi, K., Semi, K., Yamamoto, T., Shimizu, M., Tanaka, A., Mitsunaga, K., Okita, K., Osafune, K., Arioka, Y., Maeda, T., et al. (2014). Premature termination of reprogramming in vivo leads to cancer development through altered epigenetic regulation. Cell *156*, 663–677.

Pei, L., Choi, J.H., Liu, J., Lee, E.J., McCarthy, B., Wilson, J.M., Speir, E., Awan, F., Tae, H., Arthur, G., et al. (2012). Genome-wide DNA methylation analysis reveals novel epigenetic changes in chronic lymphocytic leukemia. Epigenetics *7*, 567–578.

Perez, C.A., Ott, J., Mays, D.J., and Pietenpol, J.A. (2007). p63 consensus DNA-binding site: identification, analysis and application into a p63MH algorithm. Oncogene *26*, 7363–7370.

Pujadas, E., and Feinberg, A.P. (2012). Regulated noise in the epigenetic landscape of development and disease. Cell *148*, 1123–1131.

Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. *30*, 777–782.

Raval, A., Tanner, S.M., Byrd, J.C., Angerman, E.B., Perko, J.D., Chen, S.S., Hackanson, B., Grever, M.R., Lucas, D.M., Matkovic, J.J., et al. (2007). Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. Cell *129*, 879–890.

Rossi, D., Rasi, S., Spina, V., Bruscaggin, A., Monti, S., Ciardullo, C., Deambrogi, C., Khiabanian, H., Serra, R., Bertoni, F., et al. (2013). Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. Blood *121*, 1403–1412.

Sato, N., Fukushima, N., Maitra, A., Matsubayashi, H., Yeo, C.J., Cameron, J.L., Hruban, R.H., and Goggins, M. (2003). Discovery of novel targets for aberrant methylation in pancreatic carcinoma using high-throughput microarrays. Cancer Res. *63*, 3735–3742.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363–369.

Shipony, Z., Mukamel, Z., Cohen, N.M., Landan, G., Chomsky, E., Zeliger, S.R., Fried, Y.C., Ainbinder, E., Friedman, N., and Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature *513*, 115–119.

Siegmund, K.D., Marjoram, P., Woo, Y.J., Tavaré, S., and Shibata, D. (2009). Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proc. Natl. Acad. Sci. U S A *106*, 4828–4833.

Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M., and Sorger, P.K. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature *459*, 428–432.

Timp, W., and Feinberg, A.P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nat. Rev. Cancer *13*, 497–510.

Ushijima, T., Watanabe, N., Okochi, E., Kaneda, A., Sugimura, T., and Miyamoto, K. (2003). Fidelity of the methylation pattern and its variation in the genome. Genome Res. *13*, 868–874.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D.J., Campan, M., Young, J., Jacobs, I., and Laird, P.W. (2007). Epigenetic stem cell signature in cancer. Nat. Genet. *39*, 157–158.

Wong, D.J., Liu, H., Ridky, T.W., Cassarino, D., Segal, E., and Chang, H.Y. (2008). Module map of stem cell genes guides creation of epithelial cancer stem cells. Cell Stem Cell *2*, 333–344.

Yuille, M.R., Condie, A., Stone, E.M., Wilsher, J., Bradshaw, P.S., Brooks, L., and Catovsky, D. (2001). TCL1 is activated by chromosomal rearrangement or by hypomethylation. Genes Chromosomes Cancer *30*, 336–341.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.

Cancer Cell, Volume *26*

Supplemental Information

Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia

Dan A. Landau, Kendell Clement, Michael J. Ziller, Patrick Boyle, Jean Fan, Hongcang Gu, Kristen Stevenson, Carrie Sougnez, Lili Wang, Shuqiang Li, Dylan Kotliar, Wandi Zhang, Mahmoud Ghandi, Levi Garraway, Stacey M. Fernandes, Kenneth J. Livak, Stacey Gabriel, Andreas Gnirke, Eric S. Lander, Jennifer R. Brown, Donna Neuberg, Peter V. Kharchenko, Nir Hacohen, Gad Getz, Alexander Meissner, and Catherine J. Wu

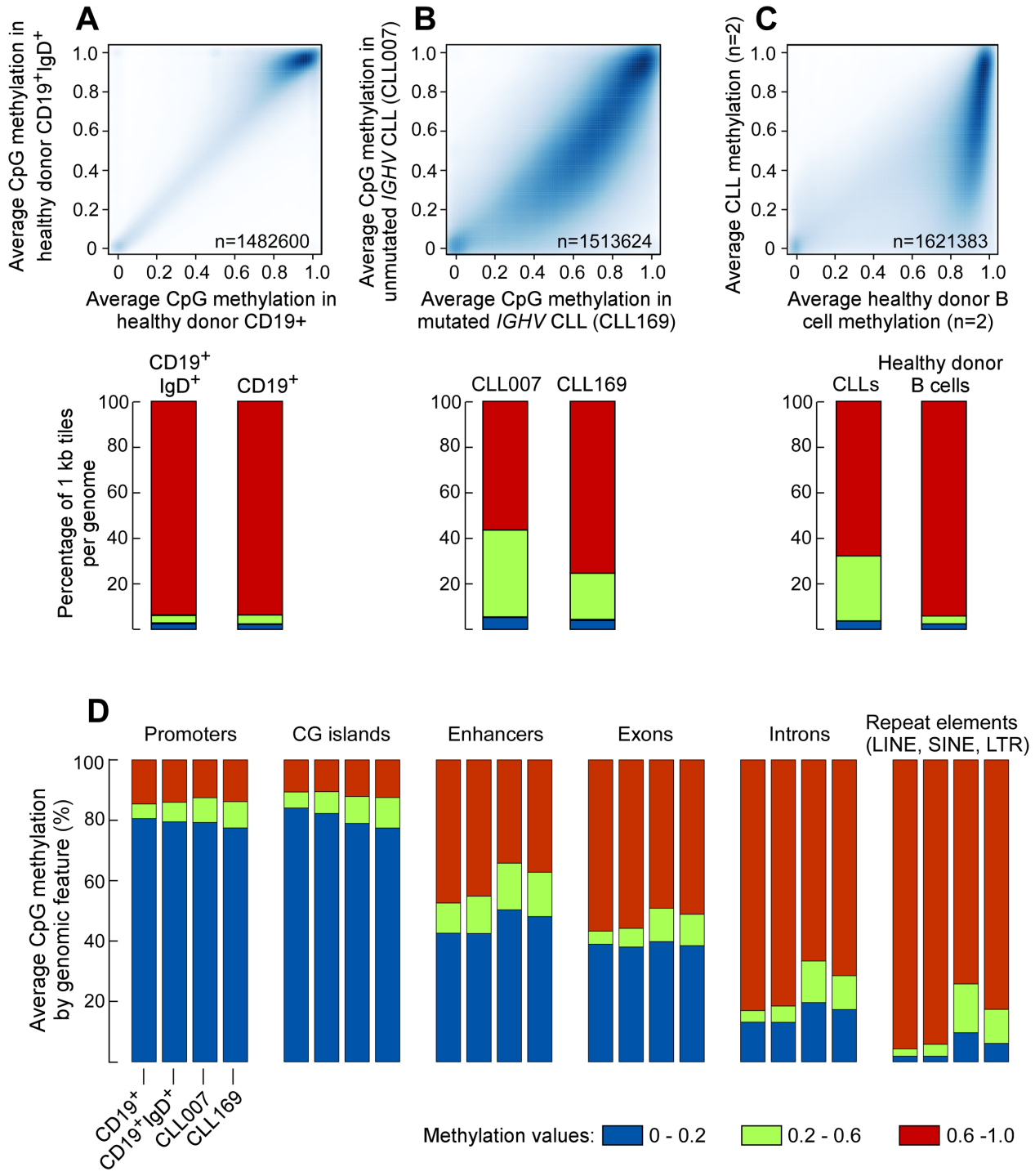
SUPPLEMENTAL DATA

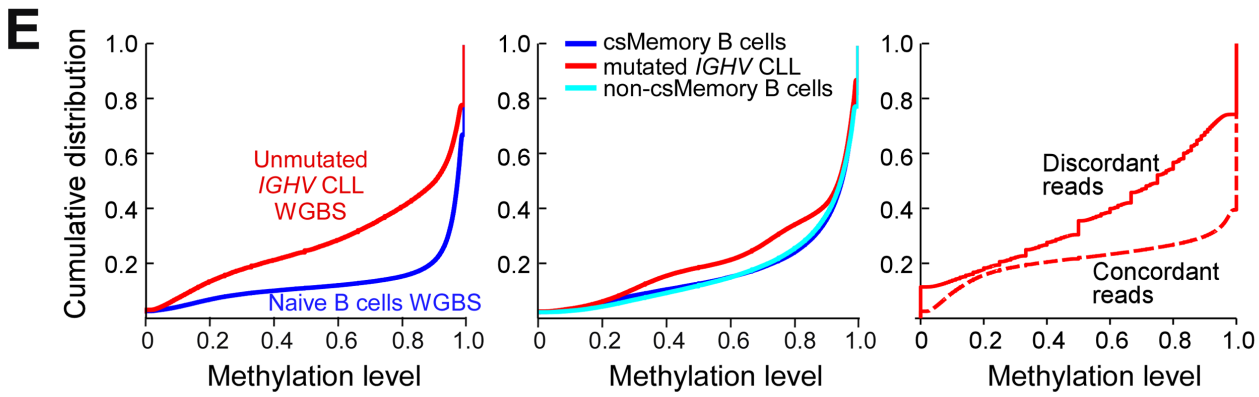
Table S1, related to Figure 1. Characteristics and mean promoter PDR of the 104 CLL patients whose DNA were analyzed by WES and RRBS.

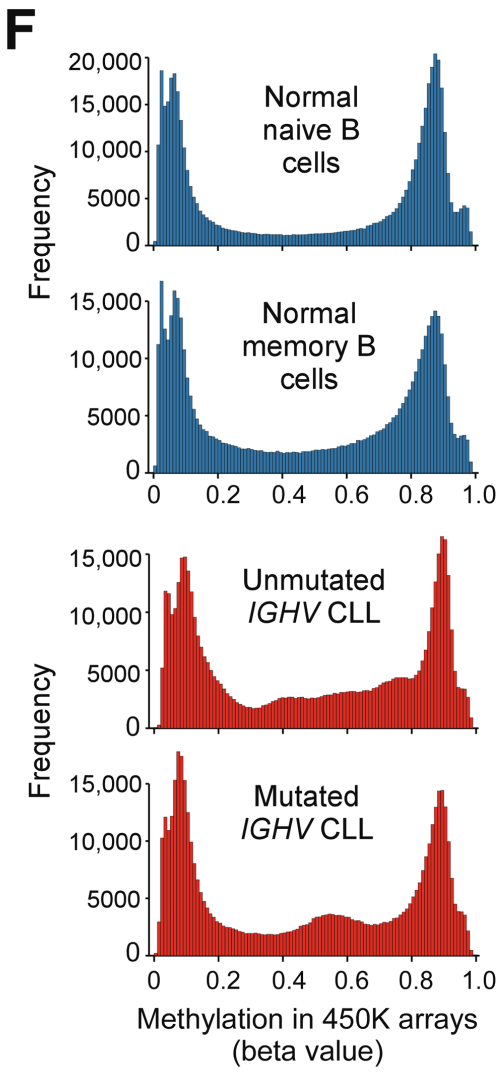
	N (%)	Mean Promoter PDR (SD)	p value†
N	104		
Age (median = 54yrs)			
<54 yrs.	46 (44)	0.101 (0.016)	0.15
≥54 yrs.	58 (56)	0.105 (0.016)	
Sex			
Female	38 (37)	0.106 (0.016)	0.30
Male	66 (63)	0.102 (0.016)	
Rai Stage at Sample			
0-1	78 (75)	0.102 (0.016)	0.049
2-4	26 (25)	0.109 (0.015)	
Treatment Status at time of			
Chemotherapy naïve	82 (79)	0.103 (0.017)	0.59
Prior Treatment	22 (21)	0.105 (0.014)	
IGHV status			
Mutated	57 (55)	0.107 (0.017)	0.035
Not Mutated	34 (33)	0.0996	
Unknown	13 (13)	0.0977	
FISH Cytogenetics††			
del(13q) present	67 (67)	0.105 (0.016)	0.059
absent	33 (33)	0.099 (0.016)	
Trisomy 12 present	18 (18)	0.099 (0.015)	0.21
absent	82 (82)	0.104 (0.016)	
del(11q) present	18 (18)	0.095 (0.013)	0.019
absent	82 (82)	0.105 (0.016)	
del(17p) present	14 (14)	0.105 (0.016)	0.62
absent	86 (86)	0.103 (0.016)	
Mutational Status			
Subclonal Mutation Present	49 (47)	0.105 (0.016)	0.25
Absent	55 (53)	0.102 (0.016)	
TP53 Present	15 (14)	0.110 (0.016)	0.091
Absent	89 (86)	0.102 (0.016)	
*NOTCH1*Present	11 (11)	0.096 (0.016)	0.097
Absent	93 (89)	0.104 (0.016)	
*SF3B1*Present	9 (9)	0.108 (0.015)	0.36
Absent	95 (90)	0.103 (0.016)	
MYD88 Present	8 (8)	0.111 (0.009)	0.19
Absent	96 (92)	0.103 (0.016)	
ATM Present	6 (6)	0.112 (0.017)	0.18
Absent	98 (94)	0.103 (0.016)	

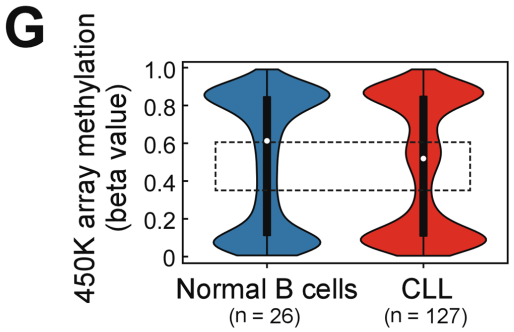
†Testing excludes unknown categories; Welch t-test (variances were not significantly different) ††N=100

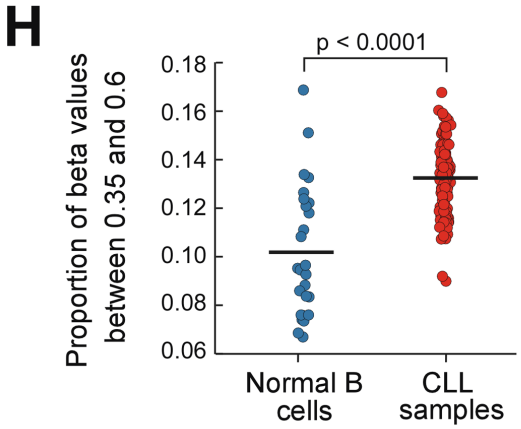
Table S2, related to Figure 1. Sample annotation and sequencing metrics for RRBS, WGBS and RNAseq data. Provided as an Excel file.


A
Average CpG methylation in healthy donor CD19+IgD+
Average CpG methylation in healthy donor CD19+
n=1482600
B
Average CpG methylation in unmutated IGHV CLL (CLL007)
Average CpG methylation in mutated IGHV CLL (CLL169)
n=1513624
C
Average CLL methylation (n=2)
Average healthy donor B cell methylation (n=2)
n=1621383
CD19+ IgD+
CD19+
CLL007
CLL169
CLLs
Healthy donor B cells
Percentage of 1 kb tiles per genome
D
Promoters
CG islands
Enhancers
Exons
Introns
Repeat elements (LINE, SINE, LTR)
Average CpG methylation by genomic feature (%)
CD19+
CD19+IgD+
CLL007
CLL169
Methylation values:
0 - 0.2
0.2 - 0.6
0.6 -1.0


E
Cumulative distribution
Unmutated
IGHV CLL
WGBS
Naive B cells WGBS
Methylation level
csMemory B cells
mutated IGHV CLL
non-csMemory B cells
Methylation level
Discordant
reads
Concordant
reads
Methylation level


F
Frequency
Normal
naive B
cells
Normal
memory B
cells
Unmutated
IGHV CLL
Mutated
IGHV CLL
Methylation in 450K arrays
(beta value)


G
450K array methylation
(beta value)
Normal B cells
(n = 26)
CLL
(n = 127)


H
p < 0.0001
Proportion of beta values
between 0.35 and 0.6
Normal B
cells
CLL
samples

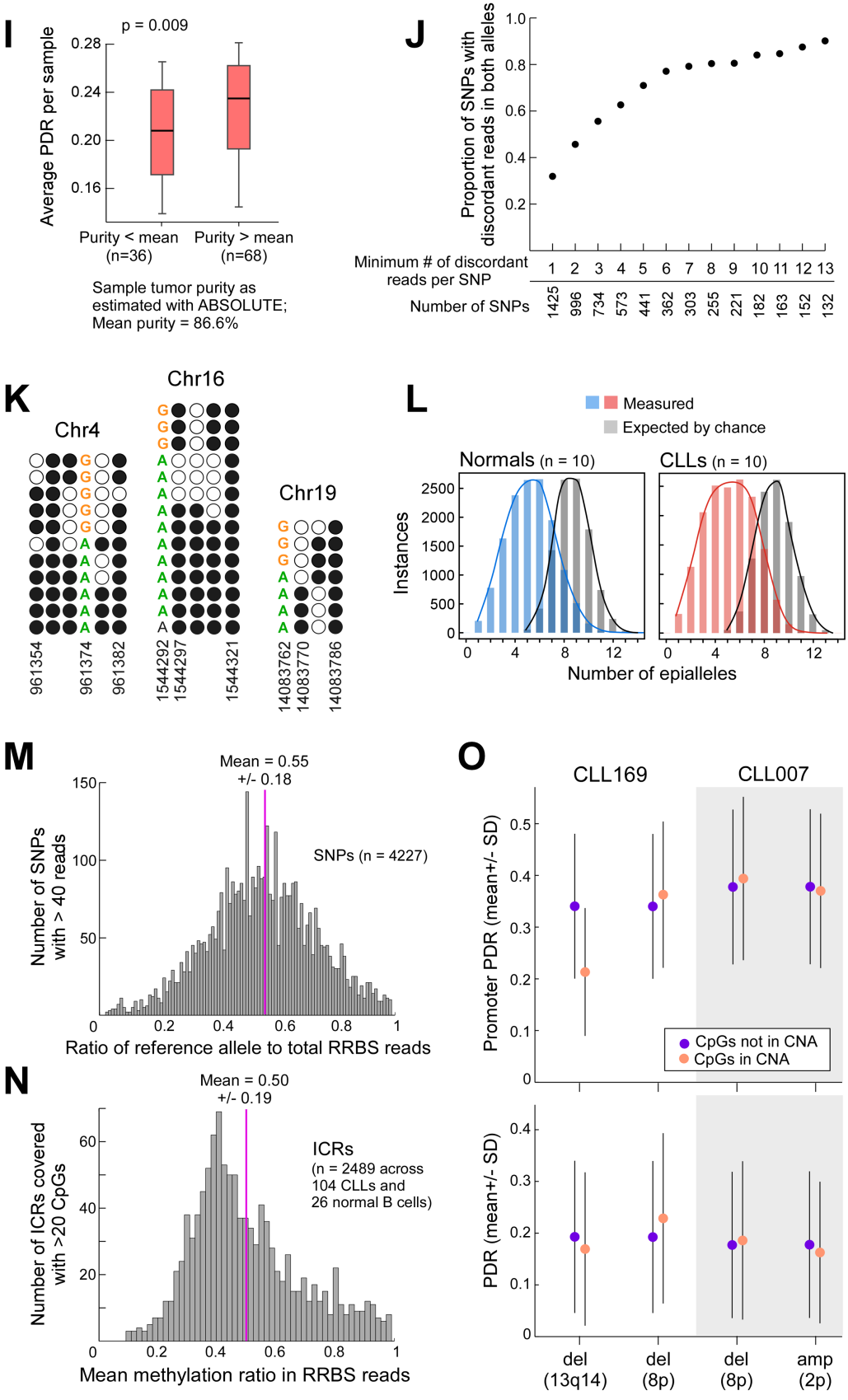
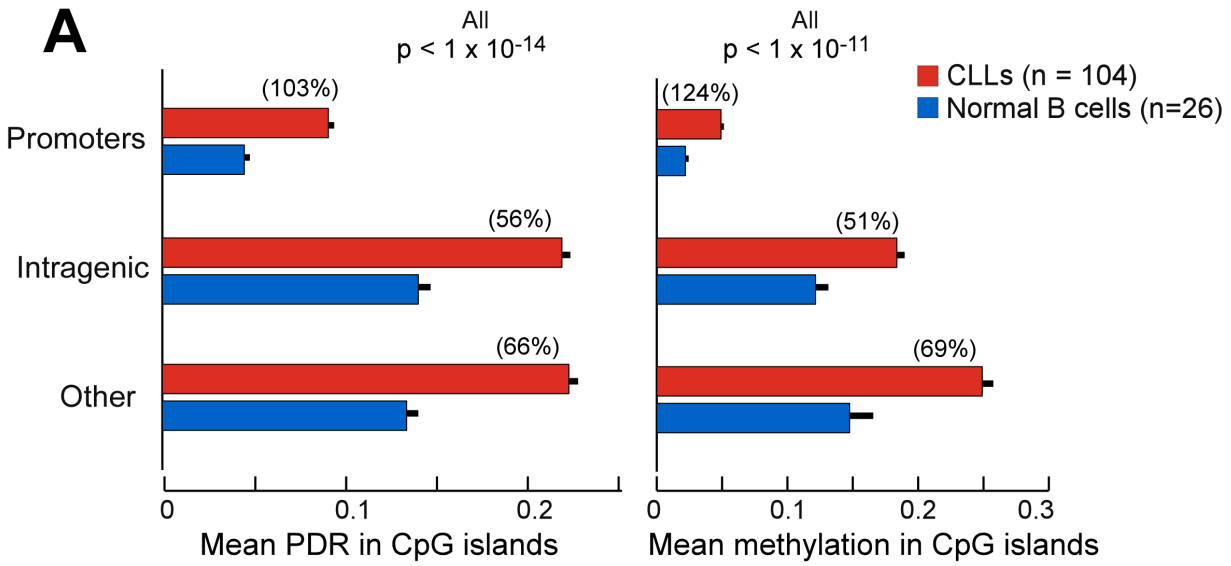

I
p = 0.009
Average PDR per sample
0.28
0.24
0.20
0.16
Purity < mean (n=36)
Purity > mean (n=68)
Sample tumor purity as estimated with ABSOLUTE; Mean purity = 86.6%
J
Proportion of SNPs with discordant reads in both alleles
1.0
0.8
0.6
0.4
0.2
Minimum # of discordant reads per SNP
1 2 3 4 5 6 7 8 9 10 11 12 13
Number of SNPs
1425 996 734 573 441 362 303 255 221 182 163 152 132
K
Chr16
Chr4
Chr19
961354
961374
961382
1544292
1544297
1544321
14083762
14083770
14083786
L
Measured
Expected by chance
Normals (n = 10)
CLLs (n = 10)
Instances
2500
2000
1500
1000
500
0
0 4 8 12
Number of epialleles
M
Mean = 0.55
+/- 0.18
SNPs (n = 4227)
Number of SNPs with > 40 reads
150
100
50
0 0.2 0.4 0.6 0.8 1
Ratio of reference allele to total RRBS reads
N
Mean = 0.50
+/- 0.19
ICRs
(n = 2489 across 104 CLLs and 26 normal B cells)
Number of ICRs covered with >20 CpGs
60
40
20
0 0.2 0.4 0.6 0.8 1
Mean methylation ratio in RRBS reads
O
CLL169
CLL007
Promoter PDR (mean+/- SD)
0.5
0.4
0.3
0.2
0.1
0
CpGs not in CNA
CpGs in CNA
PDR (mean+/- SD)
0.4
0.3
0.2
0.1
0
del (13q14)
del (8p)
del (8p)
amp (2p)

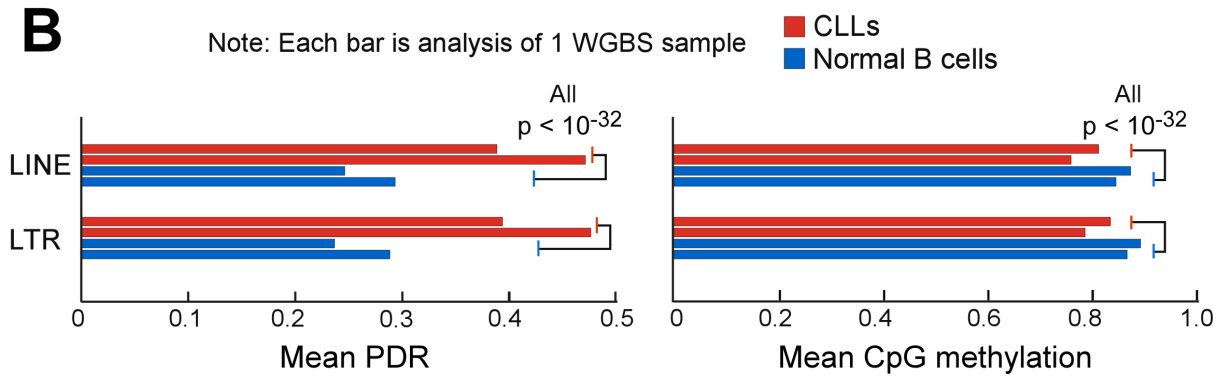
Figure S1*, related to Figure 1*: WGBS and RRBS data from CLLs and normal B cells shows higher Intratumoral DNA methylation heterogeneity that arises from locally disordered methylation. A-C (top). The genome was divided into 1KB tiles. The analysis was limited to tiles that contained at least 5 CpGs covered with greater than 5 reads. The scatter plots enable the examination of the methylation patterns consistency of the two B cell samples (**A**), and the two CLL samples (**B**). Note that the somewhat decreased methylation in CLL007 perhaps results from the DNMT3A nonsense mutation affecting this tumor. A comparison between the average methylation values across the genome in CLL and normal B cells is also shown (**C**). **A-C (bottom)**. The proportion of genomic 1KB tiles with intermediate values is compared between CLL and normal B cells. **D**. The percentage of methylation values falling within each category (0-0.2, 0.2-0.6, 0.6-1.0) is shown for the 4 WGBS samples for different genomic features. Number of CpGs per sample per feature (mean [range]): Promoters – 1,737,131 [1,728,620-1,747,890], CG islands – 2,031,560 [2,025,376-2,044,203], enhancers – 865,820 [860,997-870,134], exons – 1,489,549 [1,483,138-1,493,987], introns – 6,691,529 [6,599,956-6,739,995] and repeat elements – 7,301,495 [7,163,887-7,368,831]. **E.** Reanalysis of WGBS data (**Kulis et al.**, 2012) for the frequency of CpGs with intermediate methylation in CLL samples compared with B cells from healthy adult volunteers. Shown are cumulative distributions of CpG methylation values in unmutated *IGHV* CLL compared with naive B cells (left), as well as for mutated *IGHV* CLL vs. memory B cells (cs – class switched) (middle). The intratumoral DNA methylation heterogeneity in CLL from discordant reads (solid line) versus concordant reads (dashed) (right, analysis of WGBS data from CLL169). **F.** Histograms of individual normal B cell samples (blue) show bimodal distribution in methylation values as measured by DNA 450K methylation arrays (**Kulis et al.**, 2012), while CLL samples (red) show more CpGs with intermediate methylation values, diverging from a pure bimodal distribution. **G-H**. Violin plots comparing the proportion of intermediate methylation values from 450K array data (Kulis et al.(**Kulis et al.**, 2012)) from 127 CLL samples and 26 normal B cell samples (beta methylation values between 0.35 and 0.65, average ± SEM, 13.7 ± 0.002% vs. 10.1 ± 0.01%, respectively, $p = 5 \times 10^{-8}$, Wilcoxon rank sum test)**. I.** While overall purity of the CLL samples was consistently high (median of 90.2%), contaminating non-malignant cells in samples may contribute to the PDR, we therefore compared the PDR in CLL samples with high vs. low tumor purity (above and below the overall average; 86.6%). **J.** Stochastic disorder in methylation patterns is expected to yield discordant reads that involve both parental alleles in a given locus (in contrast to an allele-specific methylation (ASM) phenomenon). We therefore measured the proportion of germline SNPs for which a discordant read is found to involve both parental alleles (Y axis). As expected, with an increasing number of discordant reads in the studied locus (X axis), the proportion of SNPs with a discordant read involving both parental alleles increases and converges towards 1. **K.** Even within a given genotype, different methylation patterns were seen. For example, in the left most panel, 3 distinct methylation patterns are seen to affect both the A genotype parental allele and the G genotype parental allele. **L**. We measured the number of distinct discordant methylation patterns found in each locus (similar to a previous analysis (**Landan et al.**, 2012)). Presence of 1 or 2 patterns of discordancy across all reads covered for a particular locus would be expected of ASM. The plot shows the distribution of the number of methylation patterns in loci with 10-20 discordant reads across 10 randomly selected CLL and normal B cell samples. The distribution shows that there are generally more than 2 discordant methylation patterns per locus for both normal (blue) and CLL (red) samples. In addition, the high number of distinct methylation profiles per locus excludes also the possibility that PDR arises from reads that cover an ordered transition point from one methylation state to another. The shaded distribution (grey) shows the number of distinct patterns if the state of CpG methylation was purely random (with equal frequencies of the number of reads as in the experimental data). The finding that the measured distribution demonstrates less distinct patterns than purely random is consistent with inheritance of discordant patterns to progeny cells. **M.** To assess for possible amplification biases, the allelic frequencies of germline SNP not involving

CpGs was measured and shows a tight distribution around 0.5 compatible with limited amplification biases. **N.** To assess for possible amplification biases, the methylation of imprinted control regions was measured and shows a tight distribution around 0.5 compatible with limited amplification biases. **O.** Similar PDR values are seen in regions of somatic copy number variations (sCNV) in the two CLLs that underwent WGBS (CLL169 and CLL007), both for promoter CpGs (top) and for all CpGs (bottom).

Table S3, related to Figure 2. Average number of CpGs covered by RRBS with 4 or more CpGs per read, and read depth greater than 10, given by genomic feature.

Genomic feature	CLL samples		Normal B cell samples		Total # of CpGs in the human genome
	Mean	***Standard Dev.***	***Mean***	***Standard Dev.***	
Promoters	129212.20	81086.22	163485.80	107845.00	1954610
CpG islands	171342.80	108393.30	215941.40	143384.80	2124041
Exons	71536.45	45162.34	83029.35	54823.08	1954610
Enhancers	92322.87	62354.34	91093.69	62145.70	1176256
Introns	155736.40	102363.70	164236.30	110325.50	14479789
Genes	29397.06	18483.37	34193.31	22639.84	26917396
LTR	14222.88	9679.65	9102.58	6533.52	2133049
LINE	5256.03	3754.65	3055.23	2309.95	3516060
Shores	2488.20	1755.74	2349.08	1801.29	3886809
Shelves	3094.95	2403.47	4682.19	3636.86	1259327
Intergenic	35881.12	26900.36	18867.27	15753.56	5087650


A
All
p < 1 x 10-14
All
p < 1 x 10-11
CLLs (n = 104)
Normal B cells (n=26)
Promoters
Intragenic
Other
(103%)
(56%)
(66%)
(124%)
(51%)
(69%)
0
0.1
0.2
0.3
Mean PDR in CpG islands
Mean methylation in CpG islands


B
Note: Each bar is analysis of 1 WGBS sample
CLLs
Normal B cells
All
p < 10-32
All
p < 10-32
LINE
LTR
0
0.1
0.2
0.3
0.4
0.5
0.2
0.4
0.6
0.8
1.0
Mean PDR
Mean CpG methylation

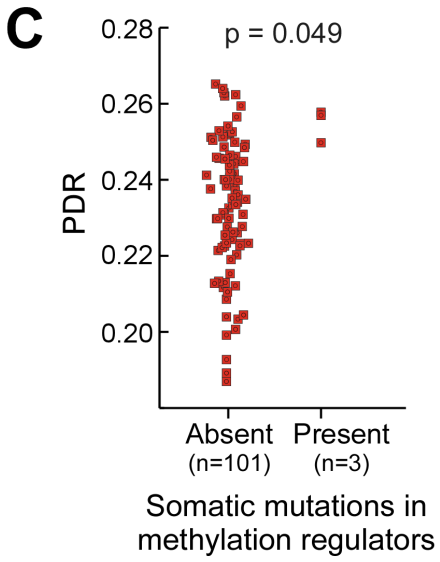
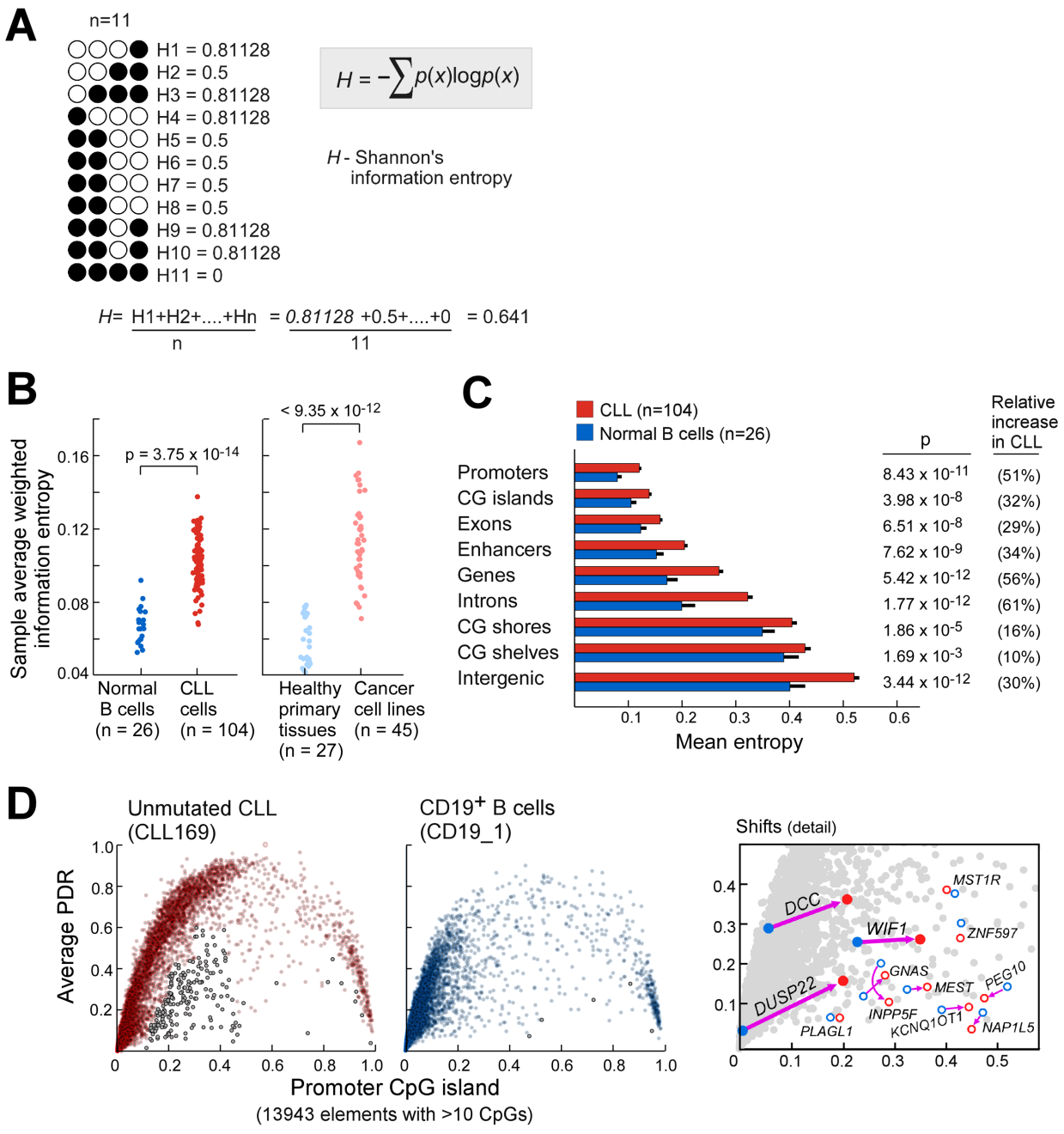

C
p = 0.049
PDR
0.28
0.26
0.24
0.22
0.20
Absent
(n=101)
Present
(n=3)
Somatic mutations in
methylation regulators

Figure S2, *related to Figure 2*. Genomic characterization of locally disordered methylation including analysis of CpG island subtypes and repeat elements based on WGBS. A. Increase in PDR with concomitant increase in methylation in 104 CLLs compared to 26 B cell samples affects all 3 major categories of CpG islands (CGIs: promoters, intragenic, other). **B**. WGBS based analysis of the 2 CLL samples (CLL007 and CLL169), compared with 2 normal B cell samples (Normal_CD19_1 and Normal_IGD_3) showing increased PDR in repeat elements concomitant with decreased methylation. Comparison between PDR and methylation values individually between each CLL sample and each normal B cell sample yielded a statistically significant difference ($p < 1 \times 10^{-32}$). **C**. The three of 104 CLL samples with nonsilent mutations in methylation modulators (*DNMT3A-Q153*, TET1-N789I, IDH1-S210N)* revealed high average PDR by RRBS compared to samples with wildtype alleles for these genes.


A
n=11
H1 = 0.81128
H2 = 0.5
H3 = 0.81128
H4 = 0.81128
H5 = 0.5
H6 = 0.5
H7 = 0.5
H8 = 0.5
H9 = 0.81128
H10 = 0.81128
H11 = 0
$H = -\sum p(x)\log p(x)$
H - Shannon's information entropy
$H = \frac{H1+H2+....+Hn}{n} = \frac{0.81128 +0.5+....+0}{11} = 0.641$
B
Sample average weighted information entropy
p = 3.75 x 10-14
< 9.35 x 10-12
Normal B cells (n = 26)
CLL cells (n = 104)
Healthy primary tissues (n = 27)
Cancer cell lines (n = 45)
C
CLL (n=104)
Normal B cells (n=26)
p
Relative increase in CLL
Promoters 8.43 x 10-11 (51%)
CG islands 3.98 x 10-8 (32%)
Exons 6.51 x 10-8 (29%)
Enhancers 7.62 x 10-9 (34%)
Genes 5.42 x 10-12 (56%)
Introns 1.77 x 10-12 (61%)
CG shores 1.86 x 10-5 (16%)
CG shelves 1.69 x 10-3 (10%)
Intergenic 3.44 x 10-12 (30%)
Mean entropy
D
Unmutated CLL (CLL169)
CD19+ B cells (CD19_1)
Average PDR
Promoter CpG island
(13943 elements with >10 CpGs)
Shifts (detail)
DCC
WIF1
DUSP22
PLAGL1
MST1R
ZNF597
GNAS
MEST
PEG10
INPP5F
KCNQ1OT1
NAP1L5

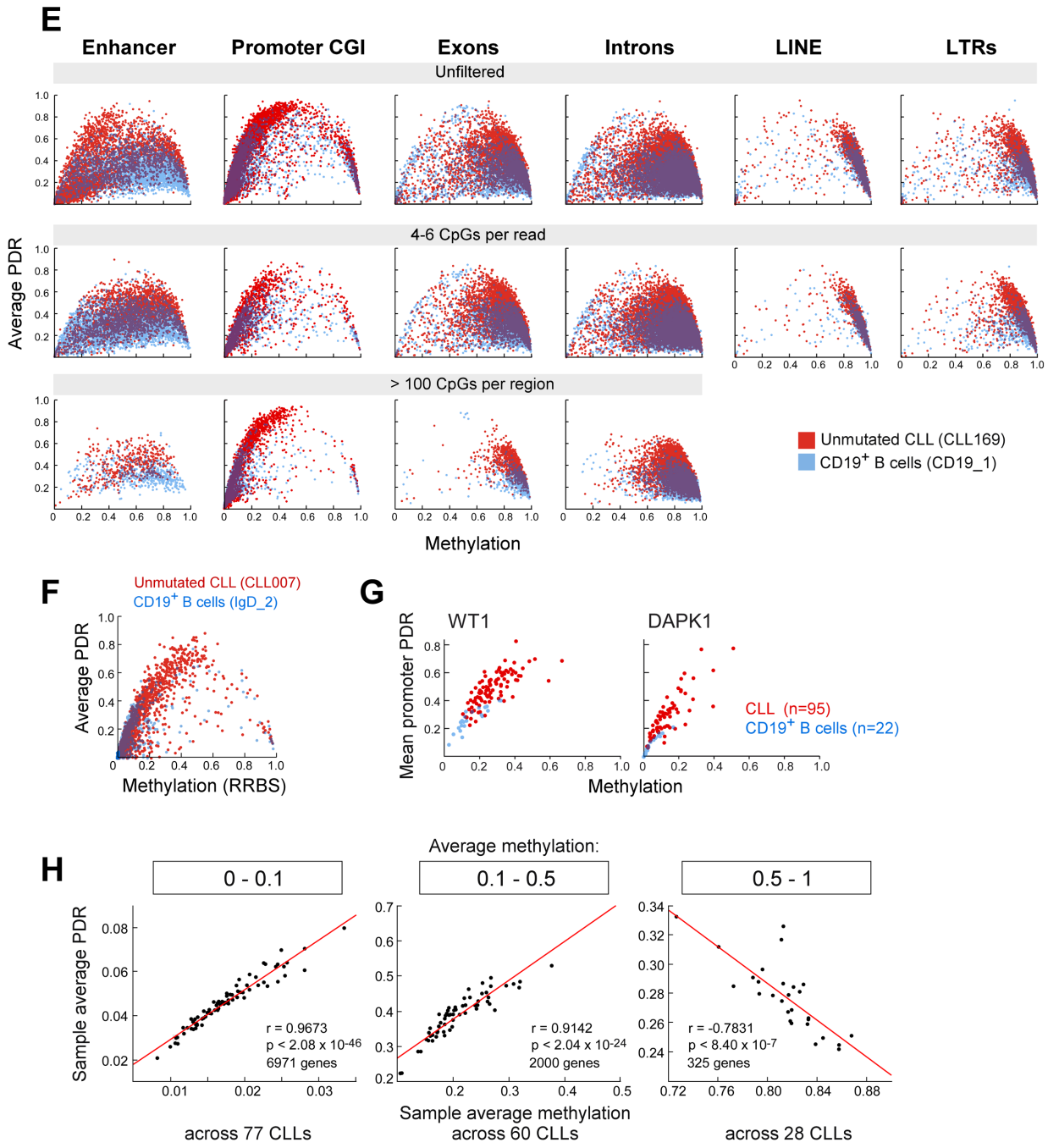


Figure S3, *related to Figure 3*. Locally disordered methylation in CLL is consistent with a stochastic process. A. As an additional measure of methylation disorder in individual reads, we calculated Shannon's information entropy (**Shannon**, 1948) for intra-sample methylation variation. Information entropy was calculated for each read and then averaged across all reads for each CpG as shown. **B**. Increased average Shannon's information entropy was observed in CLL and cancer cell line samples compared with normal B cells and primary healthy diverse human tissue samples, demonstrating an increase in stochastic methylation variation. **C**. An increase in information entropy is seen across all measured regions in RRBS data from CLL samples (red) compared with B cells from healthy adult volunteers (blue). Error bars indicate upper 95% CI. Relative increase in

average entropy from B cells to CLL samples and p value for Wilcoxon rank sum test are shown. **D.** Analysis of outlier genes falling outside of the expected distribution of PDR in relation to methylation level. Left panels – Outlier genes (black) were identified by the Tukey method in which promoter CGI PDR was lower then expected given the methylation level. Right panel - the comparative location of selected gene promoters in CLL (red) compared with normal B cells (blue). This plot highlights the considerable CLL hypermethylation without a significant concomitant change in PDR in tumor suppressor genes (*WIF1, DCC, DUSP22*; solid circles). In contrast, imprinted genes (empty circles, e.g., *GNAS*) show relative little difference between CLL and normal B cells. **E.** Scatter plots for methylation and PDR values were generated for a CLL sample (CLL169) and a normal B cell sample (Normal_CD19_1). Values were calculated for each element (enhancers**,** promoter CGIs, exons, introns, LINE family repeat elements and LTR family repeat elements) as long as at least 20 evaluable CpGs were contained in the specific element, with at least 4 CpG per read and read depth >10 ('unfiltered'). The same data were analyzed with filtering such that only CpGs covered by reads with 4-6 CpGs per read (similar to RRBS data) were examined (second row), or such that a more stringent criteria on the number of evaluable CpGs (>100) per evaluated element was used. Together the plots follow the same distribution of PDR to methylation values suggestive of a stochastic change in methylation (***Figure 3A***). **F**. A scatter plot for methylation and PDR values for promoter CGIs utilizing RRBS data (CLL007 compared with Normal_IGD_2). **G.** Similar distribution can be seen for the methylation and PDR values of promoter regions of the key tumor suppressor genes *DAPK1* and *WT1* across CLL samples. **H.** The strong correlation between average promoter CGI PDR and methylation across 104 CLL samples in shown separately for 3 groups of genes, arranged according to their average methylation values across 104 CLLs (0-0.1, left; 0.1-0.5, center; 0.5-1.0, right).

Table S4, related to Figure 3. 195 outlier genes identified by the Tukey method (i.e., PDR lower than 1.5 times the IQR below the lowest quartile for the methylation value bin), based on CLL169 WGBS data (please see ***Figure S3D***). Provided as an Excel file.

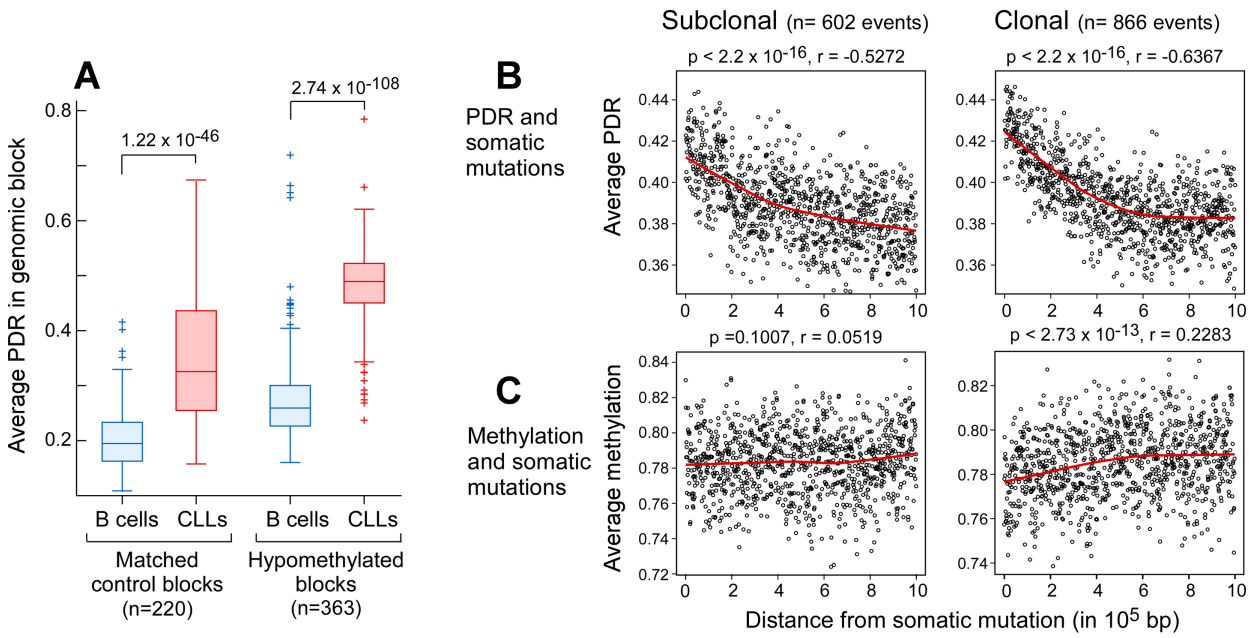
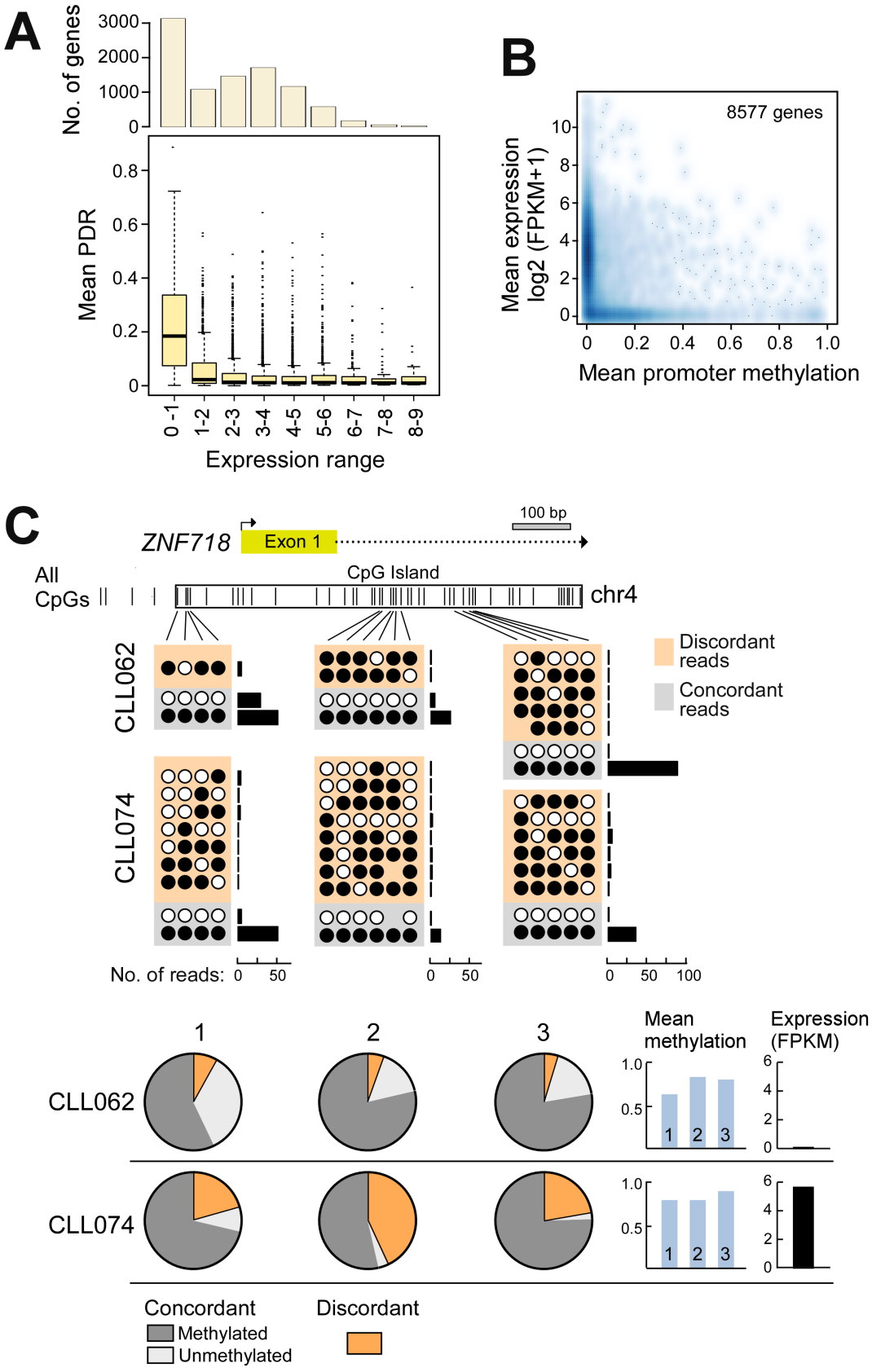
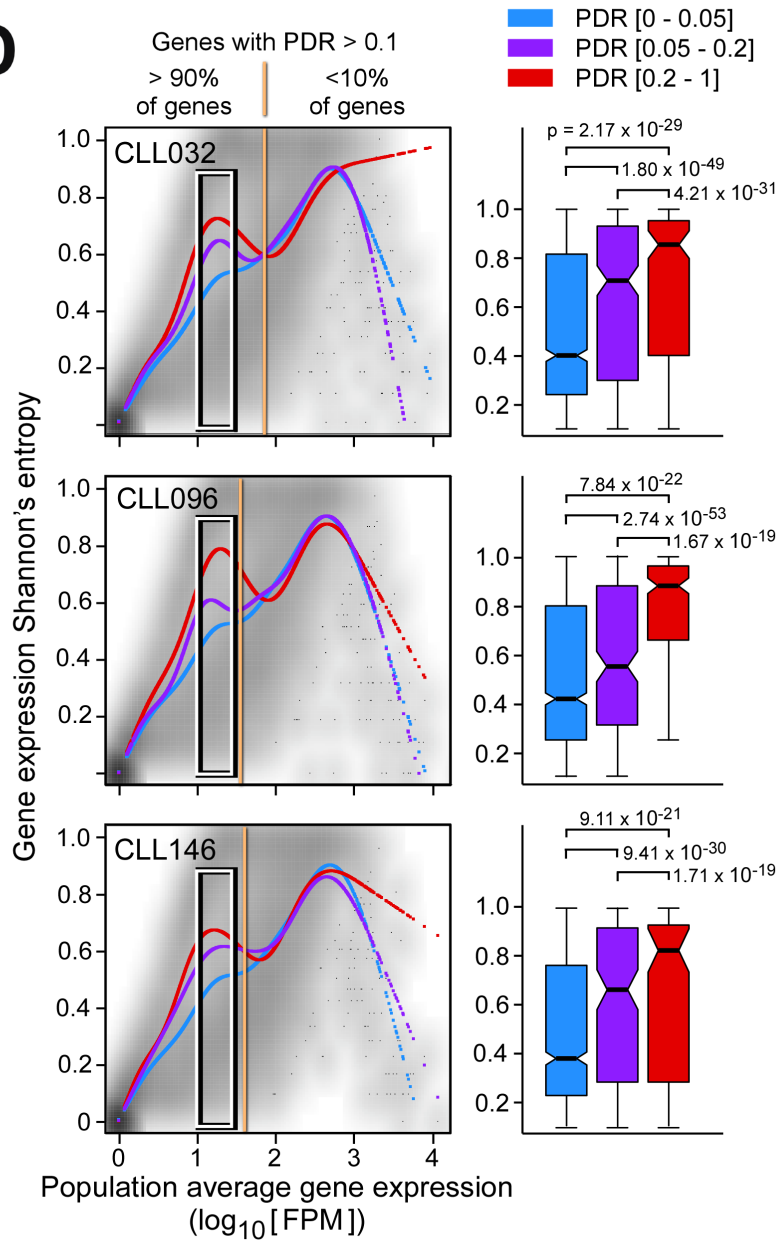


Figure S4, *related to Figure 4*. The association between PDR and distance from somatic mutation is similar for clonal and subclonal mutations. A. To study the specificity of the PDR increase in the previously defined hypomethylated blocks (**Hansen et al.**, 2011), we identified size, GC and repeat content matched regions at random from the genome. Of these regions, we retained only those that harbored more than 1000 CpGs each (covered with greater than 10 reads and 4 or more CpGs per read in the CLL169 and Normal_CD19_1 WGBS data). Compared to the control genomic regions, the hypomethylated blocks exhibit higher PDR in both CLL and normal B cells, as well as a greater increase in CLL compared to the normal B cells. To assess for a relationship between somatic mutations and PDR, somatic single nucleotide variants (sSNVs) were identified with WGS of CLL169 and matched germline DNA. Subsequently, sSNVs with sufficient read depth (>40) were classified as clonal (n=866) or subclonal (n= 602) based on the allelic frequency (above or below 0.2, respectively, analysis limited to sSNVs with greater than 40 reads and that do not involve sCNVs to enhance the confidence in the clonal vs. subclonal classification). Average PDR (**B**) and methylation (**C**) were measured in 1000bp increments from each somatic mutation. Values for each 1000bp bin were averaged over sSNVs, and plotted as a function of the distance from the somatic mutation. Red lines -- the LOWESS (locally weighted scatterplot smoothing).


A
No. of genes
3000
2000
1000
0
Mean PDR
0.8
0.6
0.4
0.2
0
0 -1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
Expression range
B
8577 genes
Mean expression log2 (FPKM+1)
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1.0
Mean promoter methylation
C
ZNF718
Exon 1
100 bp
All CpGs
CpG Island
chr4
CLL062
CLL074
Discordant reads
Concordant reads
No. of reads:
0
50
0
50
0
50
100
1
2
3
Mean methylation
Expression (FPKM)
CLL062
CLL074
1.0
0.5
6
4
2
0
Concordant
Methylated
Unmethylated
Discordant

D



E

Sample	# single cells evaluated	p value for PDR in a model predicting gene expression entropy
CLL005	84	2.27×10^{-10}
CLL032	75	1.51×10^{-7}
CLL096	70	5.16×10^{-11}
CLL146	81	1.64×10^{-4}

A model for predicting single cell gene expression information entropy based on PDR, methylation, average gene expression and transcript length

Figure S5, *related to Figure 5*. Locally disordered methylation is linked to transcriptional variation A. Genes were divided into 9 bins according to their mean expression over 33 samples (starting from 0, and then in increments of 1 until 9; log2[FPKM+1]). PDR is shown for each bin in boxplots, demonstrating that PDR is highest in genes with low expression values (bottom). The number of genes in each expression bin is shown (top). **B.** Density scatter plot of mean promoter methylation in relation to mean expression (log_2(FPKM+1)), showing that these features are negatively correlated. 8,570 genes were evaluated that had promoter RRBS coverage in at least 70% of 33 samples with matched RRBS and RNAseq. **C.** An example is shown of the promoter region of *ZNF718* from two samples (CLL062 and CLL074) with similar promoter methylation values but different PDR and different expression as measured by RNAseq (bottom right)**.** *ZNF718* promoter RRBS reads for CLL062 and CLL074 are shown (top). The number of concordantly methylated (grey background) or discordantly methylated (orange background) sequencing reads for each distinct methylation pattern is indicated to the right of each read pattern. **D**. Gene expression Shannon's information entropy (y-axis) in relation to the population average gene expression (x-axis, log_{10}[FPM]) for each gene covered in single cells of CLL032, CLL096 and CLL146, evaluated by single cell transcriptome sequencing. Colored lines - local regression curves for genes with low PDR (0-0.05, blue), intermediate PDR (0.05 – 0.2, purple), and high PDR (0.2-1.0, red). 90% of genes with higher promoter PDR (PDR >0.1) have lower population average expression (bounded by the yellow highlighted line). Right panels - Boxplots of the gene expression Shannon's information entropy for each of the three PDR bins for genes with population average gene expression of 1.0-1.5 (to control for differences in this variable). **E.** Generalized additive regression tests that model gene expression Shannon's information entropy based on: PDR, population average gene expression (locally smoothed), transcript length and promoter methylation across the 4 CLL samples that underwent single-cell transcriptome sequencing.

Table S5, related to Figure 5. Promoter PDR and methylation values for 104 CLLs as well as gene expression data from RNAseq of bulk RNA from 33 CLLs. These data were utilized in the linear models for prediction of expression based on methylation information. Provided as an Excel file.

Table S6, related to Figure 5. Results of models of prediction of gene expression for 33 CLL samples with matched RNAseq and RRBS: Values represent the adjusted R squared for the model.

Sample	PDR+Meth+CpG_content +Repeat_content	Meth	PDR	Meth+PDR	Genes measured per sample
CLL146	0.223	0.087	0.210	0.210	6110
CLL124	0.170	0.068	0.160	0.161	7016
CLL131	0.168	0.062	0.155	0.156	7264
CLL170	0.191	0.080	0.179	0.179	7636
CLL097	0.213	0.073	0.192	0.193	8511
CLL141	0.222	0.089	0.217	0.217	8818
CLL117	0.285	0.128	0.276	0.276	8926
CLL140	0.196	0.088	0.182	0.182	9005
CLL096	0.222	0.093	0.209	0.209	9016
CLL003	0.237	0.118	0.226	0.226	9044
CLL041	0.231	0.104	0.220	0.220	9045
CLL074	0.217	0.086	0.208	0.208	9566
CLL120	0.215	0.099	0.203	0.203	9871
CLL138	0.268	0.133	0.255	0.256	9914
CLL129	0.221	0.093	0.206	0.206	9976
CLL068	0.248	0.112	0.235	0.235	10029
CLL038	0.117	0.038	0.100	0.101	10058
CLL062	0.163	0.066	0.143	0.143	10141
CLL105	0.226	0.074	0.211	0.211	10311
CLL119	0.239	0.128	0.232	0.232	10351
CLL100	0.230	0.115	0.221	0.221	10387
CLL153	0.256	0.132	0.247	0.248	10426
CLL069	0.205	0.082	0.196	0.196	10600
CLL123	0.189	0.069	0.182	0.182	10655
CLL067	0.242	0.116	0.230	0.230	10684
CLL057	0.237	0.110	0.227	0.228	10745
CLL128	0.171	0.068	0.158	0.158	10750
CLL054	0.208	0.083	0.198	0.198	10755
CLL152	0.219	0.096	0.210	0.210	10828
CLL007	0.210	0.101	0.199	0.199	10883
CLL126	0.203	0.075	0.181	0.181	11051
CLL005	0.218	0.111	0.202	0.203	11064
CLL049	0.193	0.086	0.180	0.180	11138

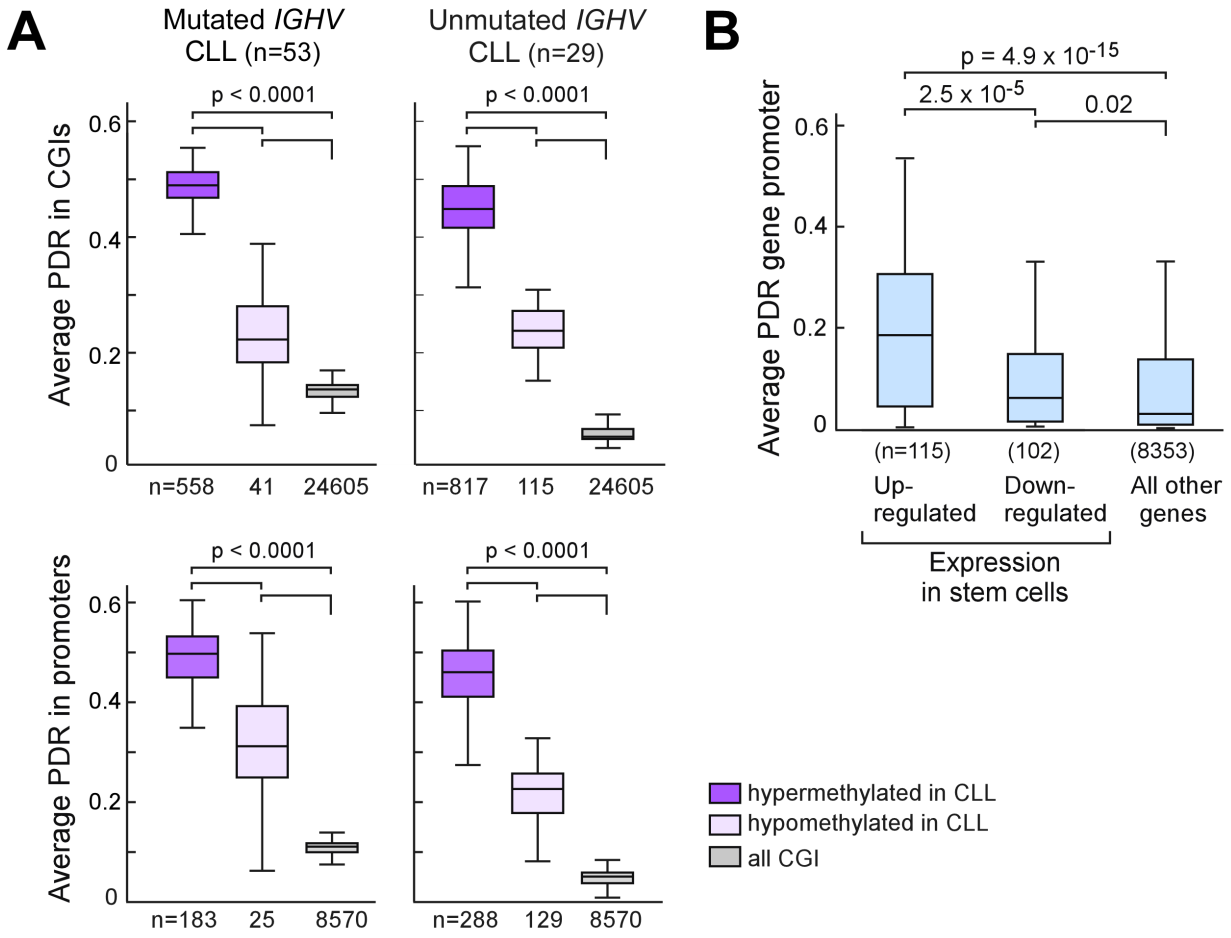


Figure S6, *related to figure 6*. Increased locally disordered methylation involves differentially methylated regions and affects stem cell related genes. A. Two sets of differentially methylated CpG islands and promoter regions were identified by comparing methylation across: i) unmutated *IGHV* CLL vs. normal naive B cell samples, and ii) across mutated *IGHV* CLL vs. normal memory B cell samples. Significantly differentially methylated regions were defined as having a >10% average methylation change with a t-test p value < 0.01. Average PDR was then calculated for each one of these regions. Higher PDR was measured in differentially methylated (both increased and decreased methylation) promoters and CpG islands compared with regions that are not differentially methylated between CLL and normal B cells (Wilcoxon rank sum test). **B**. Average promoter PDR is highest in promoters of 115 genes up-regulated in stromal stem cells compared with 102 genes down-regulated in stromal stem cells (**Boquest et al.**, 2005) as well as the average for 8,353 genes without a differential expression in stem cells (all comparisons by Wilcoxon rank sum test). Boxes represent median and interquartile range (IQR). Whiskers represent 1.5 times IQR.

Table S7, related to Figure 6. Gene set enrichments of genes with promoters with consistently high PDR across 104 CLL samples (top 30 enrichments shown).

Gene Set Name	Q_{high}	$Q_{highvslow}$
ZWANG_TRANSIENTLY_UP_BY_2ND_EGF_PULSE_ONLY	5.55E-52	2.89E-28
YOSHIMURA_MAPK8_TARGETS_UP	1.73E-43	5.92E-11
ACEVEDO_METHYLATED_IN_LIVER_CANCER_DN	4.91E-22	3.32E-10
LIM_MAMMARY_STEM_CELL_UP	5.42E-38	3.32E-10
SATO_SILENCED_BY_METHYLATION_IN_PANCREATIC_CANCER_1	1.98E-28	2.00E-09
MCBRYAN_PUBERTAL_BREAST_4_5WK_UP	4.44E-25	2.23E-08
DURAND_STROMA_MAX_UP	1.11E-25	2.23E-08
LIU_PROSTATE_CANCER_DN	3.43E-27	2.25E-07
ONDER_CDH1_TARGETS_2_UP	8.01E-21	1.78E-06
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	5.67E-16	5.10E-06
WU_CELL_MIGRATION	1.44E-13	6.49E-05
SMID_BREAST_CANCER_RELAPSE_IN_BONE_DN	2.06E-14	9.04E-05
MIKKELSEN_ES_ICP_WITH_H3K4ME3	1.47E-06	9.04E-05
SERVITJA_ISLET_HNF1A_TARGETS_UP	2.34E-13	1.19E-04
WONG_ADULT_TISSUE_STEM_MODULE	1.52E-36	1.28E-04
ACEVEDO_LIVER_CANCER_WITH_H3K27ME3_DN	5.36E-12	1.65E-04
MIYAGAWA_TARGETS_OF_EWSR1_ETS_FUSIONS_DN	1.17E-16	1.67E-04
SMID_BREAST_CANCER_LUMINAL_B_DN	1.50E-12	1.71E-04
RIGGI_EWING_SARCOMA_PROGENITOR_UP	3.68E-23	2.14E-04
SCHAEFFER_PROSTATE_DEVELOPMENT_48HR_DN	3.02E-23	2.55E-04
MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_DN	9.08E-15	2.73E-04
VART_KSHV_INFECTION_ANGIOGENIC_MARKERS_UP	4.12E-11	2.73E-04
SCHAEFFER_PROSTATE_DEVELOPMENT_48HR_UP	1.90E-22	2.93E-04
KATSANOU_ELAVL1_TARGETS_UP	1.37E-09	3.65E-04
MOHANKUMAR_TLX1_TARGETS_DN	4.84E-11	3.79E-04
PEREZ_TP63_TARGETS	4.83E-25	6.73E-04
BRUINS_UVC_RESPONSE_VIA_TP53_GROUP_A	6.89E-26	6.73E-04
GOZGIT_ESR1_TARGETS_DN	1.42E-25	7.30E-04
YAUCH_HEDGEHOG_SIGNALING_PARACRINE_UP	6.46E-10	1.16E-03
YAUCH_HEDGEHOG_SIGNALING_PARACRINE_DN	2.31E-09	1.33E-03

Table S8, related to Figure 6. Clinical characteristics of 14 patients for whom longitudinal samples were studied.

CLL IDs	Age	Therapy		*IGHV* mutation status	*Genetic evolution*	*ZAP70* status	FISH Cytogenetics	Years between samples
		Prior to timepoint 1	Between timepoints 1 & 2					
CLL018	71	None	None	Y	N	-	del(13q)	2.4
CLL020	54	None	None	Y	N	+	del(13q)	2.5
CLL019	52	None	None	Y	Y	-	del(13q)	3.2
CLL030	54	None	None	Y	N	+	del(13q)	3.5
CLL011	41	None	FCR	N	Y	+	del(13q)	5
CLL088	60	None	FCR, Alem+R	N	Y	-	tri12	4.5
CLL169	69	None	FR	Y	Y	+	del(13q)	4.7
CLL167	56	None	FR	Y	Y	-	del(13q),tri12	2.7
CLL016	59	None	FR	N	Y	+	del(13q)	3.4
CLL001	58	None	FR	N	Y	+	del(11q, 13q)	3.5
CLL006	67	FC, Chloram	Alem+R, FR, exp.	N	Y	-	del(13q),del(11q)	4.6
CLL014	65	R	FR	Y	N	-	del(13q)	2.9
CLL066	70	FR, Chloram	R-CVP	Y	N	-	del(13q)	3.5
CLL040	60	FCR	FCR, Alem+R	N	Y	+	del(13q),del(11q)	3

Abbreviations: Y- Yes, N- No, Mut.- Mutated, FISH-Fluorescence In Situ Hybridization, F-Fludarabine, C- Cyclophosphamide, R-Rituximab, V-Vincristine, Chloram- Chlorambucil, Alem – Alemtuzumab; Rev – Revlimid; exp - experimental

Table S9, related to Figure 6. Gene set enrichments of genes with significant promoter methylation changes over time (top 30 enrichments for demethylation and methylation are shown).

Gene Set Name	Q_{high}	Q_{high} vs. genes with no change	Methylation change
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	1.82E-17	3.42E-22	decrease
BENPORATH_SUZ12_TARGETS	3.76E-15	9.10E-21	decrease
BENPORATH_ES_WITH_H3K27ME3	4.89E-14	4.70E-19	decrease
PEREZ_TP53_TARGETS	2.45E-11	3.96E-17	decrease
ACEVEDO_METHYLATED_IN_LIVER_CANCER_DN	3.16E-08	4.60E-17	decrease
BENPORATH_EED_TARGETS	4.47E-12	7.05E-17	decrease
DODD_NASOPHARYNGEAL_CARCINOMA_UP	5.63E-08	2.73E-16	decrease
MIKKELSEN_MCV6_HCP_WITH_H3K27ME3	6.78E-12	1.06E-14	decrease
SMID_BREAST_CANCER_BASAL_DN	3.88E-09	3.35E-14	decrease
MEISSNER_NPC_HCP_WITH_H3K4ME2_AND_H3K27ME3	6.78E-12	4.87E-14	decrease
MIKKELSEN_MEF_HCP_WITH_H3K27ME3	7.35E-11	5.53E-14	decrease
ZWANG_TRANSIENTLY_UP_BY_2ND_EGF_PULSE_ONLY	9.93E-05	2.06E-13	decrease
JAATINEN_HEMATOPOIETIC_STEM_CELL_UP	6.96E-10	2.83E-13	decrease
WONG_ADULT_TISSUE_STEM_MODULE	6.82E-09	2.89E-13	decrease
BENPORATH_PRC2_TARGETS	6.96E-10	6.13E-13	decrease
MEISSNER_NPC_HCP_WITH_H3K4ME2	3.91E-10	1.28E-12	decrease
LIM_MAMMARY_STEM_CELL_UP	5.58E-09	2.10E-12	decrease
MIKKELSEN_NPC_HCP_WITH_H3K27ME3	6.18E-10	4.62E-12	decrease
GOZGIT_ESR1_TARGETS_DN	1.56E-06	1.56E-11	decrease
ONDER_CDH1_TARGETS_2_UP	1.99E-08	2.55E-11	decrease
CUI_TCF21_TARGETS_2_DN	1.62E-07	6.88E-11	decrease
MEISSNER_BRAIN_HCP_WITH_H3K27ME3	3.98E-08	1.52E-09	decrease
ZWANG_TRANSIENTLY_UP_BY_1ST_EGF_PULSE_ONLY	1.70E-03	1.91E-09	decrease
DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	3.23E-06	8.12E-09	decrease
CHYLA_CBFA2T3_TARGETS_UP	3.83E-05	2.48E-08	decrease
CHEN_METABOLIC_SYNDROM_NETWORK	5.55E-04	3.21E-08	decrease
LEE_BMP2_TARGETS_UP	5.39E-05	5.16E-08	decrease
MEISSNER_NPC_HCP_WITH_H3_UNMETHYLATED	7.13E-06	6.43E-08	decrease
LIU_PROSTATE_CANCER_DN	5.55E-05	7.34E-08	decrease
GOBERT_OLIGODENDROCYTE_DIFFERENTIATION_DN	9.93E-05	1.08E-07	decrease
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	1.82E-17	3.42E-22	decrease

Gene Set Name	Q_{high}	Q_{high} vs. genes with no change	Methyla-tion change
BENPORATH_SUZ12_TARGETS	1.49E-19	2.99E-25	increase
BENPORATH_ES_WITH_H3K27ME3	2.27E-18	2.05E-23	increase
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	4.03E-18	6.63E-22	increase
BENPORATH_EED_TARGETS	3.55E-15	1.21E-19	increase
MIKKELSEN_MCV6_HCP_WITH_H3K27ME3	5.84E-14	2.80E-16	increase
DODD_NASOPHARYNGEAL_CARCINOMA_UP	6.53E-08	1.40E-15	increase
BENPORATH_PRC2_TARGETS	1.72E-11	3.77E-14	increase
MEISSNER_NPC_HCP_WITH_H3K4ME2	7.06E-12	5.84E-14	increase
MIKKELSEN_MEF_HCP_WITH_H3K27ME3	2.32E-11	6.76E-14	increase
MEISSNER_NPC_HCP_WITH_H3K4ME2_AND_H3K27ME3	4.49E-10	2.21E-11	increase
ZWANG_TRANSIENTLY_UP_BY_2ND_EGF_PULSE_ONLY	7.57E-04	7.91E-11	increase
MARTENS_TRETINOIN_RESPONSE_UP	1.48E-05	1.04E-10	increase
GOZGIT_ESR1_TARGETS_DN	8.29E-06	6.73E-10	increase
SCHAEFFER_PROSTATE_DEVELOPMENT_48HR_UP	9.59E-07	7.99E-10	increase
MEISSNER_NPC_HCP_WITH_H3_UNMETHYLATED	9.09E-08	8.06E-10	increase
LEE_BMP2_TARGETS_UP	1.01E-06	8.06E-10	increase
CHEMNITZ_RESPONSE_TO_PROSTAGLANDIN_E2_DN	1.71E-06	6.84E-09	increase
ZWANG_TRANSIENTLY_UP_BY_1ST_EGF_PULSE_ONLY	2.06E-03	8.66E-09	increase
MEISSNER_BRAIN_HCP_WITH_H3K27ME3	1.84E-07	1.31E-08	increase
BLALOCK_ALZHEIMERS_DISEASE_UP	1.40E-04	2.97E-08	increase
MIKKELSEN_NPC_HCP_WITH_H3K27ME3	1.01E-06	6.16E-08	increase
SMID_BREAST_CANCER_BASAL_UP	8.09E-05	7.55E-08	increase
GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_UP	7.07E-05	1.05E-07	increase
DAWSON_METHYLATED_IN_LYMPHOMA_TCL1	9.59E-07	1.68E-07	increase
WONG_ENDMETRIUM_CANCER_DN	1.16E-05	2.55E-07	increase
BRUINS_UVC_RESPONSE_VIA_TP53_GROUP_A	5.02E-04	2.68E-07	increase
GINESTIER_BREAST_CANCER_ZNF217_AMPLIFIED_DN	1.48E-05	3.12E-07	increase
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_5	3.55E-05	3.62E-07	increase
MARTINEZ_TP53_TARGETS_DN	1.91E-04	9.25E-07	increase
BENPORATH_SUZ12_TARGETS	1.49E-19	2.99E-25	increase
BENPORATH_ES_WITH_H3K27ME3	2.27E-18	2.05E-23	increase

Table S10, related to Figure 7. Stepwise regression model for prediction of clinical outcome.

	Unadjusted HR [95% CI]	**Stepwise selection Final model (without subclonal driver) HR [95% CI]**	**Stepwise selection Final model (including subclonal driver as candidate) HR [95% CI]**
Promoter PDR: cutpoint at the Mean > 0.1033 vs. ≤0.1033	2.51 [1.10-5.17] p=0.029	3.48 [1.37-8.86] p =0.009	
IGVH Mutated vs. Unmutated	0.29 [0.11-0.77] p =0.013	0.16 [0.05-0.47] p =0.0009	0.20 [0.07-0.58] p =0.003
Presence of del11q	1.26 [0.55-2.86] p =0.58		
Presence of del17p	3.46 [1.39-8.62] p =0.008	2.51 [0.84-7.51] p =0.10	3.24 [0.99-10.54] p =0.051
Presence of a subclonal driver	4.80 [1.79-12.92] p =0.002	NA	6.54 [2.16-19.86] p =0.0009
Promoter methylation: cutpoint at mean >0.0735 vs. ≤ 0.0735	1.81 [0.83-3.99] p =0.14		
Mutation number: cutpoint at mean > 18.8 vs. ≤ 18.8	1.89 [0.85-4.23] p =0.012	2.57 [1.04-6.35] p =0.040	3.42 [1.39-8.39] p =0.007

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Human samples: Heparinized blood samples were obtained from patients and healthy adult volunteers enrolled on clinical research protocols at the Dana-Farber/Harvard Cancer Center (DF/HCC), approved by the DF/HCC Human Subjects Protection Committee. The diagnosis of CLL according to WHO criteria was confirmed in all cases by flow cytometry, or by lymph node or bone marrow biopsy. Peripheral blood mononuclear cells (PBMC) from normal donors and patients were isolated by Ficoll/Hypaque density gradient centrifugation. Mononuclear cells were cryopreserved with FBS/10% DMSO and stored in vapor-phase liquid nitrogen until the time of analysis. The patients included in the cohort represent the broad clinical spectrum of CLL (***Table S1***). Informed consent on DFCI IRB-approved protocols for genomic sequencing of patients' samples was obtained prior to the initiation of sequencing studies.

Established CLL prognostic factor analysis: *Immunoglobulin heavy-chain variable* (*IGHV*) homology (unmutated was defined as greater than or equal to 98% homology to the closest germline match) and *ZAP-70* expression (high risk defined as >20% positive) were determined (**Rassenti et al.**, 2008). Cytogenetics were evaluated by FISH for the most common CLL abnormalities (del(13q), trisomy 12, del(11q), del(17p), all probes from Vysis, Des Plaines, IL, performed at the Brigham and Women's Hospital Cytogenetics Laboratory, Boston MA). Samples were scored positive for a chromosomal aberration based on consensus cytogenetic scoring (**Smoley et al.**, 2010).

DNA isolation from CLL and normal B-cell subpopulations: Genomic DNA was extracted from CLL cells or normal B cell populations utilizing the ROCHE DNA Isolation Kit (Roche Applied Science, Indianopolis, IN). Control $CD19^+$ B cell samples were isolated from buffy coats of healthy adult volunteers using a two-step enrichment procedure. B cells were first enriched using the RosetteSep Human B cell Enrichment System (StemCell Technologies Inc., Vancouver, British Columbia, Canada) and then further purified by immunomagnetic bead selection ($CD19^+$ beads, Miltenyi Biotec, Cambridge, MA). From these purified $CD19^+$ cells, naive B cells ($CD19^+CD27^-IgD^+$) and memory B cells ($CD19^+CD27^+IgD^-$) were isolated by flow cytometric sorting (FACSAria II, BD Biosciences) using CD27-PC5 (Beckman Coulter, Brea, CA) and IgD-CY7 (Biolegend, San Diego, CA) antibodies. Standard protocols for DNA quality control for genomic studies were applied, as recently described (**Berger et al.**, 2011; **Chapman et al.**, 2011; **Landau et al.**, 2013).

Reanalysis of whole-exome DNA sequencing (WES) data from CLL samples: We re-analyzed WES from 104 of 160 previously reported CLLs and their matched germline samples (**Landau et al.**, 2013), deposited in dbGaP (phs000435.v2.p1). Details of whole-exome library construction and analysis have been detailed elsewhere (**Fisher et al.**, 2011; **Landau et al.**, 2013)**.** Briefly, output from Illumina software (Illumina, San Diego, CA) was processed by the "Picard" data processing pipeline to yield BAM files containing aligned reads with well-calibrated quality scores (**Chapman et al.**, 2011; **DePristo et al.**, 2011). Somatic alterations were identified using a set of tools within the "Firehose" pipeline, developed at the Broad Institute (www.broadinstitute.org/cancer/cga) (**Berger et al.**, 2011; **Chapman et al.**, 2011). Somatic single nucleotide variations (sSNVs) were detected using MuTect (**Cibulskis et al.**, 2013). We used the ABSOLUTE algorithm to calculate the purity, ploidy, and absolute DNA copy-numbers of each sample (**Carter et al.**, 2012) and clonal/subclonal status of each alteration inferred using a probabilistic approach (**Escobar and West**, 1995; **Landau et al.**, 2013). We note that

the spectrum of mutations in these samples was consistent with prior publications (**Quesada et al.**, 2012), with C>T transitions constituting the most frequent sSNVs (average of 41.8 ± 15% of all sSNV across all 104 CLL WES analyzed in this study). There was no significant correlation between the proportion per sample of any specific subtype of sSNV and PDR ($-0.1 < r < 0.1$, $p > 0.3$).

Whole Genome Sequencing of CLL sample CLL169 and CLL007: Library construction was performed using 1–3 micrograms of native DNA from primary tumor (peripheral blood) and germline (saliva) samples. The DNA was sheared to a range of 101–700 bp using the Covaris E210 Instrument and was then phosphorylated and adenylated according to the Illumina protocol. Adaptor ligated purification was done by preparatory gel electrophoresis, and size was selected by excision of two bands (500–520 bp and 520–540 bp, respectively), yielding two libraries per sample with average of 380 bp and 400 bp, respectively. The libraries were then sequenced with the Illumina GA-II or Illumina HiSeq sequencer with 76 or 101 bp reads, achieving an average of ~30X coverage depth. The resulting data were analyzed with the current Illumina pipeline, which generates data files (BAM files) that contain the reads and quality parameters. Sequencing data are available in the dbGaP database (http://www.ncbi.nlm.nih.gov/gap) under accession number phs000435.v2.p1. Somatic single nucleotide variations (sSNVs) were detected using MuTect (**Cibulskis et al.**, 2013). Replication times were adopted from Chen et al. (**Chen et al.**, 2010). S50 values (for a defined genome region, S50 corresponds to the fraction of the S phase at which 50% of the sequence reads that map in this region were obtained) were rescaled to vary from 100 (early) to 1000 (late) as previously described (**Lawrence et al.**, 2013). Although replication times reported by Chen et al., were not measured directly in CLL cells or B cells, previous studies have shown that replication time is fairly consistent across different cell types (**Karnani et al.**, 2007). Furthermore, Chen and colleagues confirmed a high correlation with previously measured replication time in other cell types including human lymphocytes.

RNA-sequencing of CLL samples and analysis: 5mg of total RNA was poly-A selected using oligo-dT beads to extract the desired mRNA, and used to construct dUTP libraries as previously described (**Landau et al.**, 2013). Samples were pooled and sequenced using either 76 or 101bp paired end reads. RNAseq BAMs were aligned to the hg19 genome using the TopHat suite. FPKM values were generated with the Cufflinks suite (http://cufflinks.cbcb.umd.edu/). These data are deposited in dbGaP (phs000435.v2.p1).

RRBS: Genomic DNA from CLL samples, normal B cell samples and cancer cell line samples were 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) to allow immediate subsequent pooling. This was followed by bisulfite conversion and PCR, as previously described (**Boyle et al.**, 2012). Libraries were sequenced and 29mers were aligned to the hg19 genome using MAQ version 0.6.6 (**Li et al.**, 2008). Reads were further filtered if: i) The read did not align to an autosome, ii) The read failed platform/vendor quality checks (samtools flag 0x200), and/or iii) the read did not align to an MspI cut site.

The methylation state of each CpG was determined by comparing bisulfite-treated reads aligning to that CpG with the genomic reference sequence. The methylation level was computed by dividing the number of observed methylated cytosines (which did not undergo bisulfite conversion) by the total number of reads aligned to that CpG (***Figure 1E***). In addition, the number of CpG measurements on each read was noted. In order to identify locations in the genome where concordant methylation (in either methylated or unmethylated states) occurs, we

devised a measure called the ***Proportion of Discordant Reads*** (PDR). This measure can be computed for a specific genomic location or for the entire genome. After reads are aligned to the reference genome, the methylation state of each CpG on a read is determined. If all the CpGs on a specific read are methylated, or all of the CpGs on a read are unmethylated, the read is classified as *concordant*; otherwise it is classified as *discordant*. At each CpG, the PDR is equal to the number of discordant reads that cover that location divided by the total number of reads that cover that location (***Figure 1E***). The PDR across the entire genome or for a specified genomic region is given by averaging the values of individual CpGs, as calculated for all CpGs within the region of interest with read depth greater than 10 reads and that are covered by reads that contain at least 4 CpGs. It is important to note that PDR and variances were also calculated with means weighted by depth of coverage of a particular CpG with consistently similar results. For example, overall variance weighted by the number of read depth per CpG shows similar difference in variance of 0.0696 [0.0679- 0.0714] for CLL samples, vs. 0.0437 [0.0399- 0.0475] for normal B cell samples ($p = 2.61 \times 10^{-13}$). Weighted average of PDR for CLL samples was 0.2476 [0.2431- 0.2520] vs. 0.1402 [0.1275-0.1528] for normal B cell samples ($p = 1.06 \times 10^{-14}$). The CLL and normal B cell RRBS raw data are deposited in dbGaP (phs000435.v2.p1), and processed data format files containing PDR and methylation values for each CpG evaluated in the CLL and normal B cell samples are deposited in GEO (GSE58889). RRBS of primary diverse human tissue samples were previously reported (http://www.roadmapepigenomics.org). Reads were realigned and methylation was determined using identical protocols to the rest of the samples.

WGBS: Genomic DNA was fragmented to 100–500 bp fragments using a Covaris S2 sonicator (Woburn, MA). DNA fragments were cleaned-up, end-repaired, A-tailed and ligated with methylated paired-end adapters (from ATDBio, Southampton, UK). Libraries were sequenced and WGBS reads were aligned using BSMAP version 2.7 (**Xi and Li**, 2009) to the hg19/GRCh37 reference assembly. Subsequently, CpG methylation calls were made using custom software, excluding duplicate, low-quality reads, as well as reads with more than 10% mismatches. We note that as previously reported (**Kulis et al.**, 2012), non-CpG methylation levels were minimal (0.08% in both CLL samples). Only CpGs covered by > 10 reads were considered for further analysis. A methylation-calling pipeline was implemented in Perl and determines CpG methylation state by observing bisulfite conversion at read locations aligned to a CpG in the reference genome. Previously published WGBS data for 2 CLL samples and 3 normal B cell samples (**Kulis et al.**, 2012) were downloaded with permission from the European Genome-Phenome Archive. The raw sequencing reads were processed in identical fashion to the in-house produced WGBS libraries. Additional processing steps for WGBS reads included trimming by 4bp to ensure high data quality, and filtering out reads that: i) did not align to an autosome, ii) failed platform/vendor quality checks (samtools flag 0x200), iii) had poor alignment score (samtools flag 0x2), iv) had poor alignment of the read mate (samtools flag 0x8), v) aligned to the same location as another read (read duplicate), or vi) contained nucleotides at a CpG location that could not have been produced by bisulfite conversion. The determination of the concordant vs. discordant classification was performed in identical fashion as with RRBS reads. The CLL and normal B cell WGBS data are deposited in dbGaP (phs000435.v2.p1), and processed data format files containing PDR and methylation values for each CpG evaluated in the sample are deposited in GEO (GSE58889). **Table S2** contains a list of all the samples evaluated in this study along with the source of data and measurement type annotation.

Methylation array analysis: Data for previously published 450K methylation arrays (**Kulis et al.**, 2012) were downloaded with permission from the European Genome-Phenome Archive. Data from the 450k Human Methylation Array were analyzed by GenomeStudio (Illumina) and R using the lumi package available through Bioconductor.

Single cell RNA-Sequencing of CLL samples: Four primary cryopreserved peripheral blood CLL samples were thawed and stained with anti-CD19 FITC and anti-CD5 PE antibodies (Beckman Coulter, Indianapolis, IN). 7-AAD (Invitrogen, Grand Island, NY) was added before FACS sorting as a viability control. Live $CD19^{+}CD5^{+}$ tumor cells were preliminarily sorted into a collection tube. Subsequently, the bulk cell concentration was adjusted to 250 cell/ml and applied to the C1 Single-Cell Auto Prep System for single cell capture with a 5-10 micron chip (Fluidigm, San Francisco, CA). The capture rate was measured at > 80%. Following capture, whole transcriptome amplification (WTA) was immediately performed using the C1 Single-Cell Auto Prep System with the SMARTer Kit (Clontech, Mountain View, CA) on up to 96 individual cells. The C1 WTA products were then converted to Illumina sequencing libraries using Nextera XT (Illumina). RNA-Seq was performed on a MiSeq instrument (Illumina).

Analysis of single-cell RNA-seq data: Paired-ended reads were aligned against UCSC hg19 human annotation (March 6, 2013 version) using Tophat 2.0.10 (**Kim et al.**, 2013), and read counts for each gene were determined using HTSeq 0.5.4 (**Anders et al.**, 2014). A subset of cells with more than 10,000 total reads across all genes was selected for further analysis (73-87% of cells). To determine population average gene expression (performed separately for each of the 4 primary CLL samples), the read counts observed in each cell were normalized by the effective library size, determined by edgeR (**Robinson et al.**, 2010) 'calcNormFactors' method.

To test for significance of association of PDR with expression heterogeneity, first 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 ent = [-1 x (fpc x log2(fpc) + (1-fpc) x log2(1-fpc)]. The association with PDR was tested using generalized additive models (implemented by gam R package). The following types of models were tested:

- ent ~ s(population average expression) + PDR + transcript length
- ent ~ s(population average expression) + PDR + transcript length + methylation

where s() indicates local regression. The population average expression values were entered into the models on $\log_{10}$ scale (adding 1).

Genome annotations definitions: Promoters were defined as 1 Kb upstream and 1 Kb downstream of hg19 Refgene gene transcription start sites (TSSs). The set of CpG Islands (CGIs) were defined using biologically-verified CGIs (**Illingworth et al.**, 2010). Enhancer regions were defined as the union of the 'Distal Regulatory Modules' class from all cell types as previously identified (**Ramskold et al.**, 2012). CTCF binding sites were annotated based on published CTCF binding ChIP-seq experiments using 27 healthy donor transformed B cells ChIP-seq experiments (**Wang et al.**, 2012). We curated a list of CTCF binding sites based on sites that were detected in at least 75% of these B cell samples, and then calculated the CTCF binding site per megabase across the human genome. The location of repeat elements was identified based on the RepBase database version 18.09 for hg19 (http://www.girinst.org/server/archive/RepBase18.09/). Hypomethylated regions in embryonic stem cells were defined as previously described (**Ziller et al.**, 2013), and the analysis was limited to regions with at least 20 CpGs. Differentially-methylated regions (DMRs) were called using a two-sample t-test with significance of $p < 0.01$ and in which the difference between the weighted average region methylation levels was greater than 10%. Well-covered regions with at least 5 CpGs in at least 80% of the samples were used for the analysis, as previously described (**Bock et al.**, 2011).

Summary statistics of methylation across DMRs between CLL and normal B cells.

DMR category	Mean methylation		Standard deviation		Number of elements
	CLL	***Normal B cells***	***CLL***	***Normal B cells***	
Promoters hypermethylated in *IGHV* mutated CLL	3.06E-01	1.71E-01	1.51E-01	1.40E-01	213
Promoters hypomethylated in *IGHV* mutated CLL	4.95E-01	6.40E-01	2.11E-01	2.02E-01	28
CGIs hypomethylated in *IGHV* mutated CLL t	4.01E-01	5.54E-01	3.08E-01	2.97E-01	41
CGIs hypermethylated in IGHV mutated CLL	3.27E-01	1.87E-01	1.61E-01	1.51E-01	558
CGIs hypomethylated in *IGHV* unmutated CLL	4.70E-01	6.69E-01	3.15E-01	2.81E-01	115
CGIs hypermethylated in *IGHV* unmutated CLL	2.84E-01	1.24E-01	1.63E-01	1.55E-01	817
Promoters hypomethylated in *IGHV* unmutated CLL	5.57E-01	7.11E-01	2.55E-01	2.56E-01	145
Promoters hypermethylated in *IGHV* unmutated CLL	2.59E-01	1.02E-01	1.53E-01	1.42E-01	332

Modeling locally disordered methylation: In order to describe the expected PDR for a given set of reads covering the same set of CpGs, we developed a model to describe the likelihood of finding a certain number of discordant reads, given a methylation value for the set of reads. The input parameters for our model were the number of CpGs covered by the reads, the average methylation value of the covered CpGs, and the number of reads covering the CpGs. We modeled the methylation state of each CpG on each read as an independent Bernoulli trial, with the probability of getting a methylated CpG being set to the overall empirical methylation average. The probability of seeing a specified number of discordant reads was then unity minus the probability of observing a specified number of concordant reads (a probability derived directly from the independent Bernoulli trials for each CpG).

Using this model, we were able to predict the maximum likelihood for PDR for a set of reads covering a certain number of CpGs, with a certain methylation value. In addition to finding the maximum likelihood PDR, we were able to assign a *P*-value for the probability of finding a specified number of discordant reads, given the number of CpGs covered by the reads, the average methylation value, and the total number of reads. We plotted the 99% confidence interval using this model in ***Figure 3A***.

Germline variants detection for allele-specific analyses: Germline variants were detected using the UnifiedGenotyper in the Genome Analysis Toolkit (http://www.broadinstitute.org/gatk/), using default options, followed by the filtering of SNPs using Variant Quality Score Recalibration, and hard-filtering of indels (**DePristo et al.**, 2011; **McKenna et al.**, 2010). Germline variants were annotated using SeattleSeq137 (http://snp.gs.washington.edu/SeattleSeqAnnotation137/).

Excluding alternative explanations for high PDR other than locally disordered methylation: We considered several possible alternative explanations to these findings. First, the **contaminating non-malignant cell fraction** of samples may contribute to the PDR, even though the overall purity of the CLL samples was consistently high (90.2% median purity). However, when we compared samples with purity above and below the overall average (86.6%), PDR was higher in the former (mean ± SEM, 0.2259 ± 0.0047 vs. 0.2062 ± 0.0066, t-test p =0.009), indicating that indeed the malignant cells in the samples contribute to the high PDR (***Figure S1I***). Second, we considered the possibility that elevated PDR may affect only one allele in the sample as part of **allele-specific methylation** (ASM). To test this, we identified germline SNPs that did not involved CpGs across 53 randomly selected CLL samples in the cohort. Of these germline SNPs, 4,486 had equivalent coverage of both genotypes in the RRBS reads (ratio of 0.4-0.6 in variant reads/total reads). At these sites, discordant reads were found to contain both alternative genotypes in an increasing proportion of SNPs in association with an increased total number of discordant reads per locus (***Figure S1J***), converging towards 1. This result demonstrates that locally disordered methylation likely affects both parental alleles. Furthermore, even within a given genotype different discordancy patterns were seen (***Figure S1K***), revealing that high PDR results indeed from locally disordered methylation and not simply from allele-specific methylation patterns. In this context, it is important to note that X/Y chromosomes were excluded from the entire analysis.

In addition to the germline variants, we carried out a similar analysis with regards to somatic single nucleotide mutations, by integrating WGS and WGBS data for CLL007 and CLL169. After excluding C>T mutations, and limiting the analysis to regions with >4 CpGs per read on average (to ensure accurate estimation of PDR) and to mutations with >20X coverage in the WGS (to ensure accurately distinguishing clonal vs. subclonal events), we identified 52 and 66 high confidence mutations for analysis, respectively (91% and 79% of these mutations were either intronic or intergenic mutations in CLL007 and CLL169, respectively). The correlation between the average methylation values of the clonally mutated alleles and the matching germline alleles was high (CLL169 - number of clonal mutations evaluated = 30, $r = 0.96$, $p = 1.9 \times 10^{-17}$, CLL007 - number of clonal mutations evaluated = 10, $r = 0.94$, $p = 3.6 \times 10^{-5}$). Similarly, the correlation between the PDR of the clonally mutated alleles and the matched germline alleles was also high (CLL169: $r = 0.72$, $p = 5.6 \times 10^{-7}$; CLL007: $r = 0.65$, $p = 0.04$). While the correlation of average methylation values remained high between the mutated alleles and the matched germline alleles for subclonal mutations (CLL169 - number of subclonal mutations evaluated = 36, $r = 0.47$, $p = 0.008$, CLL007 - number of subclonal mutations evaluated = 42, $r = 0.81$, $p = 5.3 \times 10^{-11}$), the correlation between the PDR values of the two alleles was lower ($r = 0.09$ and 0.45, $p = 0.5$ and $p = 0.002$, respectively), with a trend towards higher PDR in the mutated subclonal allele (20.5% and 34.6% increase in PDR in mutated alleles, for CLL169 and CLL007, respectively, with $p = 0.2$ and 0.048). Collectively, these data show that disordered methylation involved both the mutant and germline alleles, with a trend towards higher PDR in subclonally mutated alleles.

Moreover, if high PDR results from ASM, then we would expect to find predominately 1 or 2 consistent patterns of discordancy, across all reads covered for a particular locus. However, a histogram of the number of distinct discordancy pattern in loci that have a significant number of discordant reads (10-20) across ten randomly selected CLL samples, shows a normal distribution centered at 5 discordant patterns, consistent with a model of stochastic disorder rather than ASM (***Figure S1L)***. This latter finding also confirms that most of the PDR does not result from reads that cover an ordered transition point from one methylation state to another, which is also expected to yield 1 recurrent discordancy pattern.

Another potential explanation for increased PDR could be related to **methQTL (Gibbs et al., 2010)**. This is unlikely to account for the genome-wide pervasive process we describe for the following reasons: i) this effect is expected to be of importance in a tumor with a high mutation load. However, CLL is a malignancy with one of the lowest mutational loads, 1000-2000 mutations per genome (**Wang et al.**, 2011). Extrapolating from the study by Gibbs et al., which evaluated ~1.5M germline SNPs and only found association with 4-5% of CpGs, the mutational load in CLL at best will only affect 0.005% of CpGs. This is expected to have a small effect in comparison to the pervasive disorder in methylation patterns (e.g., in CLL169 WGBS, 73.39% of CpGs have PDR >0.1). ii) Cancer cell lines, which harbor 1-3 orders of magnitude more somatic mutations than primary CLLs, harbor marginally higher rates of PDR. iii) Finally, the PDR pattern would more likely result from methQTLs of subclonal mutations, as clonal mutations would behave largely like germline SNPs and therefore are unlikely to result in increase in PDR in cancer vs. normal tissue, given their number in the CLL genome. To assess for the confounding effect of methQTL on PDR, which may be related to subclonal mutations, we compared the correlation to PDR between clonal mutations and subclonal mutations and found that the distance from clonal mutations shows a stronger negative correlation to PDR, compared to the distance from subclonal mutations (***Figure S4B-C***). Although methQTL may have long-range effects, at least a third supposedly act in *cis* (defined in Gibbs et al., as <1MB). These results, therefore, are not consistent with a significant impact of methQTL.

Finally, **technical artifacts** were also considered as a potential cause of locally disordered methylation. Incomplete bisulfite conversion is an unlikely explanation for these findings as bisulfite conversion rates were high in both CLL and normal B cell samples (average of 99.66% and 99.72%, respectively) as measured by the rate of unmethylated cytosines in a non-CpG context (**Bock et al.**, 2005). Furthermore, incomplete conversion is expected to decrease PDR preferentially in highly methylated region, however, we observed an increase in PDR in CLLs in regions with both low and high methylation.
PCR amplification biases in the RRBS procedure are not likely to contribute significantly to this result. First, we have no reason to expect differential impact on CLL samples and normal B cells. Second, the consistency of the finding in WGBS where duplicate reads were discarded makes this technical bias an unlikely source for locally disordered methylation. Indeed the Pearson's correlation of PDR in promoter CpGs covered by both RRBS and WGBS at >30X was high (CLL169; r= 0.856, CLL007; r = 0.855, and Normal_IGD_3; r=0.737). Finally, given that we have no reason to expect duplicate reads to affect concordant reads less than discordant reads, duplicate reads are expected to decrease PDR, as the overall number of concordant reads is higher than discordant reads (87.1 ± 2% of RRBS reads evaluated are concordant, evaluated in randomly selected 5 samples (CLL003, CLL005, CLL006_TP1, CLL001_TP1 and CLL001_TP2). To quantify PCR amplification biases, we measured the ratio of reads for each of the heterozygous SNP and found a similar representation of both parental alleles (***Figure S1M***). In addition, measured methylation values for germline Imprinted Control Regions (ICRs) (**Woodfine et al.**, 2011) and found that these loci approximated 50% methylation, as expected (***Figure S1N***).

Finally, although CLL genomes are mostly diploid (**Brown et al.**, 2012), and therefore the analysis is not expected to be significantly impacted by **somatic copy number variations** (sCNV), we examined the PDR in regions of sCNV in WGBS of CLL007 and CLL169. Altogether in these tumors, 4 sCNVs were detected (using SNP array analysis as described previously (**Landau et al.**, 2013)). As shown in ***Figure S1O,*** both the overall PDR and the promoter PDR do not differ substantially in the sCNVs compared to the remainder of the genome.

Gene set enrichment analysis: Gene set enrichment analysis was limited to the C2 gene set collection (**Subramanian et al.**, 2005). To assess gene set enrichments in genes that exhibit consistently elevated PDR (greater than mean promoter PDR of 0.1 in >75% of 104 CLL samples) a Fisher's exact test was used to measure the enrichment of these genes in each gene-set, followed by a Benjamini-Hochberg FDR procedure. Similarly, to compare enrichments between the set of genes with high promoter PDR and low promoter PDR (less than mean promoter PDR of 0.1 in >75% of 104 CLL samples), a Fisher's exact test was used, followed by a Benjamini-Hochberg FDR procedure. This latter procedure was done to avoid potential biases related to the CpG content of different promoters as previously described . By comparing enrichments of two gene sets both covered by RRBS, these biases are likely to have minimal impact. A similar procedure was undertaken for gene set enrichment analysis of genes with significant change in methylation in the longitudinal samples ($Q < 0.1$). By comparing these gene-sets with genes that did not have a significant change in methylation ($Q > 0.2$), we were able to assess the gene set enrichment while limiting the impact of biases related to CpG content of different gene promoters.

Statistical methods: Statistical analysis was performed with MATLAB (MathWorks, Natick, MA), R version 2.15.2 and SAS version 9.2 (SAS Institute, Cary, NC). Categorical variables were compared using the Fisher Exact test, and continuous variables were compared using the Student's t-test, Wilcoxon rank sum test, or Kruskal-Wallis test as appropriate. Linear modeling for expression as a predicted variable, based on methylation and PDR was performed using built in R linear model function. FFS (failure-free survival from first treatment after sampling) was defined as the time to the 2nd treatment or death from the 1st treatment following sampling, was calculated only for those patients who had a 1st treatment after the sample and was censored at the date of last contact for those who had only one treatment after the sample, and estimated using the method of Kaplan and Meier. The difference between groups was assessed using the log-rank test. Unadjusted and adjusted Cox modeling was performed to assess the impact of established CLL high-risk predictors and the presence of a subclonal driver. Models were adjusted for known prognostic factors including the presence of a 17p deletion, the presence of a 11q deletion and *IGHV* mutational status. Cytogenetic abnormalities were primarily assessed by FISH; if FISH was unavailable, genomic data were used. For unknown *IGHV* mutational status an indicator was included in adjusted modeling and was not found to be significant. Similarly, unadjusted and adjusted Cox modeling was performed to assess the impact of mutational burden and average promoter methylation in addition to established CLL prognostic factors. Given the large number of potential variables, a stepwise selection procedure was used to determine a final multivariable model considering all factors listed above. All p-values are two-sided and considered significant at the 0.05 level unless otherwise noted.

SUPPLEMENTAL REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2014). **HTSeq – A Python framework to work with high-throughput sequencing data**. bioRxiv.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C.*, et al.* (2011). **The genomic complexity of primary human prostate cancer**. Nature. *470*, 214-220.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H.*, et al.* (2011). **Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines**. Cell. *144*, 439-452.

Bock, C., Reither, S., Mikeska, T., Paulsen, M., Walter, J., and Lengauer, T. (2005). **BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing**. Bioinformatics. *21*, 4067-4068.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A.*, et al.* (2012). **Absolute quantification of somatic DNA alterations in human cancer**. Nat Biotechnol. *30*, 413-421.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M.*, et al.* (2011). **Initial genome sequencing and analysis of multiple myeloma**. Nature. *471*, 467-472.

Chen, C.L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O.*, et al.* (2010). **Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes**. Genome Res. *20*, 447-457.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples**. Nat Biotechnol. *31*, 213-219.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. Nat Genet. *43*, 491-498.

Escobar, M., and West, M. (1995). **Bayesian density estimation and inference using mixtures**. Journal of the American Statistical Association. *90*, 577-588.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L.*, et al.* (2011). **A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries**. Genome Biol. *12*, R1.

Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J.*, et al.* (2010). **Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain**. PLoS genetics. *6*, e1000952.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). **Orphan CpG islands identify numerous conserved promoters in the mammalian genome**. PLoS Genet. *6*, e1001134.

Karnani, N., Taylor, C., Malhotra, A., and Dutta, A. (2007). **Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas**. Genome research. *17*, 865-876.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. Genome biology. *14*, R36.

Li, H., Ruan, J., and Durbin, R. (2008). **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. Genome Res. *18*, 1851-1858.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. Genome Res. *20*, 1297-1303.

Quesada, V., Conde, L., Villamor, N., Ordonez, G.R., Jares, P., Bassaganyas, L., Ramsay, A.J., Bea, S., Pinyol, M., Martinez-Trillos, A.*, et al.* (2012). **Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia**. Nat Genet. *44*, 47-52.

Rassenti, L., Jain, S., Keating, M., Wierda, W., Grever, M., Byrd, J., Kay, N., Brown, J., Gribben, J., Neuberg, D.*, et al.* (2008). **Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia.** Blood. *112*, 1923-1930.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. Bioinformatics. *26*, 139-140.

Shannon, C.E. (1948). **A Mathematical Theory of Communication**. Bell System Technical Journal. *27*, 379–423.

Smoley, S.A., Van Dyke, D.L., Kay, N.E., Heerema, N.A., Dell' Aquila, M.L., Dal Cin, P., Koduru, P., Aviram, A., Rassenti, L., Byrd, J.C.*, et al.* (2010). **Standardization of fluorescence in situ hybridization studies on chronic lymphocytic leukemia (CLL) blood and marrow cells by the CLL Research Consortium**. Cancer Genet Cytogenet. *203*, 141-148.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E.*, et al.* (2005). **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** Proc Natl Acad Sci U S A. *102*, 15545-15550.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R.*, et al.* (2012). **Widespread plasticity in CTCF occupancy linked to DNA methylation**. Genome Res. *22*, 1680-1688.

Wang, L., Lawrence, M.S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D.S., Zhang, L.*, et al.* (2011). **SF3B1 and other novel cancer genes in chronic lymphocytic leukemia**. N Engl J Med. *365*, 2497-2506.

Woodfine, K., Huddleston, J.E., and Murrell, A. (2011). **Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue**. Epigenetics & chromatin. *4*, 1.

Xi, Y., and Li, W. (2009). **BSMAP: whole genome bisulfite sequence MAPping program**. BMC Bioinformatics. *10*, 232.