Discovery of candidate DNA methylation cancer driver genes

Heng Pan, Loic Renaud, Ronan Chaligne, Johannes Bloehdorn, Eugen Tausch, Daniel Mertens, Anna Maria Fink, Kirsten Fischer, Chao Zhang, Doron Betel, et al.

HHS Public Access

Author manuscript

Cancer Discov. Author manuscript; available in PMC 2022 March 01.

Published in final edited form as:

Cancer Discov. 2021 September ; 11(9): 2266–2281. doi:10.1158/2159-8290.CD-20-1334.

Discovery of candidate DNA methylation cancer driver genes

Heng Pan[1,2,3], **Loïc Renaud**[4,5,6,7], **Ronan Chaligne**[4,5,6], **Johannes Bloehdorn**[8], **Eugen Tausch**[8], **Daniel Mertens**[9], **Anna Maria Fink**[10], **Kirsten Fischer**[10], **Chao Zhang**[3,6], **Doron Betel**[3,6], **Andreas Gnirke**[11], **Marcin Imielinski**[1,3,4,5,12], **Jérôme Moreaux**[13,14,15,16], **Michael Hallek**[10], **Alexander Meissner**[11,17], **Stephan Stilgenbauer**[8], **Catherine J. Wu**[11,18], **Olivier Elemento**[1,2,3,5], **Dan A. Landau**[3,4,5,6,19,*]

[1]Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA.

[2]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA.

[3]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA.

[4]New York Genome Center, New York, NY, USA

[5]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA.

[6]Division of Hematology and Medical Oncology, Department of Medicine, Weill Cornell Medicine, New York, NY, USA

[7]Inserm, UMR-S 1172, Lille, France

[8]Department of Internal Medicine III, Ulm University, Ulm, Germany

[9]German Cancer Research Center (DKFZ), Heidelberg, Germany

[10]German CLL Study Group, and Department I of Internal Medicine, and Center of Integrated Oncology ABCD, University of Cologne, Cologne, Germany

[11]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[12]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA

[13]IGH, CNRS, Univ Montpellier, France

[14]CHU Montpellier, Department of Biological Hematology, Montpellier, France

[*]Correspondence: Dan A. Landau, Weill Cornell Medicine, 413 East 69th Street, BB1428, New York, NY, 10021. Phone: +1-646-962-6311; dlandau@nygenome.org.

Conflict of interest statement

E. Tausch reports grants and personal fees from Roche, grants, personal fees, and non-financial support from Abbvie, and personal fees and non-financial support from Janssen outside the submitted work. A. Fink reports grants from Celgene, personal fees from AbbVie, and personal fees from JAnssen outside the submitted work. K. Fischer reports other from Roche and personal fees from Roche during the conduct of the study; personal fees from Abbvie outside the submitted work. C.J. Wu reports other support from Pharmacyclics outside the submitted work. O. Elemento reports personal fees and other from Volastra Therapeutics, other from OneThree Biotech, other from Freenome, personal fees from Champions Oncology, and other from Owkin during the conduct of the study. D.A. Landau reports grants from NIH and grants from LLS during the conduct of the study; personal fees from C2i Genomics, personal fees from 10x Genomics, and personal fees from Illumina outside the submitted work.

Accession codes: RRBS data of CLL8 are available via GEO accession number GSE143673. RRBS data of CLL-DFCI are available via dbGap accession number phs000435.v3.p1. RRBS data of DCIS-MDACC are available via GEO accession number GSE69994. RRBS data of GBM-MUV are available via EGA accession number EGAS00001002538.

Code availability: The MethSig pipeline is available on GitHub at https://github.com/HengPan2007/MethSig.

[15]Univ Montpellier, UFR de Médecine, Montpellier, France

[16]Institut Universitaire de France (IUF), France

[17]Max Planck Institute for Molecular Genetics, Berlin, Germany

[18]Dana-Farber Cancer Institute, Boston, MA, USA

[19]Lead Contact

Abstract

Epigenetic alterations such as promoter hypermethylation may drive cancer through tumor suppressor genes inactivation. However, we have limited ability to differentiate driver DNA methylation (DNAme) changes from passenger events. We developed DNAme driver inference – MethSig – accounting for the varying stochastic hypermethylation rate across the genome and between samples. We applied MethSig to bisulfite sequencing data of chronic lymphocytic leukemia (CLL), multiple myeloma, ductal carcinoma in situ, glioblastoma, and to methylation array data across 18 tumor types in TCGA. MethSig resulted in well-calibrated Quantile-Quantile plots and reproducible inference of likely DNAme drivers with increased sensitivity/specificity compared to benchmarked methods. CRISPR/Cas9 knockout of selected candidate CLL DNAme drivers provided a fitness advantage with and without therapeutic intervention. Notably, DNAme driver risk score was closely associated with adverse outcome in independent CLL cohorts. Collectively, MethSig represents a novel inference framework for DNAme driver discovery to chart the role of aberrant DNAme in cancer.

Keywords

DNA methylation; Cancer drivers; Cancer epigenetics; Statistical inference framework; CLL

Introduction

DNA methylation (DNAme) is a central epigenetic modification of the human genome (1,2). DNAme is also thought to be an important disease-defining feature in many cancers (3-6), pointing to the cancer's cell-of-origin and predictive of the outcome. Indeed, several tumor types harbor frequent mutations in genes that encode components of the methylation machinery (2).

DNAme changes in cancer have been described along two principal axes: global hypomethylation impacting retroviral elements and genome stability, and focal hypermethylation at promoters of tumor suppressor genes (TSGs) (1,2). Promoter hypermethylation of TSGs has been surveyed across cancer in The Cancer Genome Atlas (TCGA) as well as other studies (1-3), and revealed that a plethora of cancer-related cellular pathways are disrupted by hypermethylation of TSG promoters, such as DNA repair (*MLH1*, *RBBP8*), cell cycle (*CDKN2A*, *CDKN2B*), P53 network (*CDKN2A*, *TP73*), apoptosis (*WIF1*, *SFRP1*), Ras signaling (*RASSF1*), Wnt signaling (*SOX17*) and tyrosine kinase cascades (*SOCS3*) (7-9).

While it is tempting to assume that all observed DNAme changes occur deterministically and drive the cancer phenotype, *in vitro* models and human cancers have shown that DNAme changes overwhelmingly follow a stochastic process (5,6,10,11). While these changes are stochastic, they occur at different rate across the genome, correlated with features such as low gene expression and late DNA replication (5). Thus, stochastic DNAme changes in the growing malignant population result in a cancer methylome that displays locally disordered methylation and high intra-tumoral heterogeneity (5,6). These data underscore the challenge of identifying candidate DNAme changes that are likely to be linked to the cancer phenotype among the highly abundant stochastic DNAme events across the genome, reminiscent of the challenge of distinguishing driver from passenger mutations in cancer.

However, unlike the field of cancer genomics where increasingly sophisticated tools have been developed to distinguish between driver and passenger mutations, accounting for confounding covariates (12,13), inference tools in cancer epigenomics largely rely on uniform background models. Thus, widely used statistical methods produce hundreds or thousands of candidate promoter hypermethylation sites, likely overshadowing a much smaller number of DNAme changes that impact oncogenesis (referred to here as DNAme drivers).

To address this challenge, we developed a statistical inference framework accounting for varying stochastic hypermethylation rate across the genome and between patients – MethSig, analogous to leading approaches for cancer driver gene inference (12). MethSig estimates expected tumor promoter hypermethylation with an inference model that includes biological features known to affect the stochastic rate of DNAme changes (5). We applied MethSig to reduced representation bisulfite sequencing (RRBS) (14) data across blood and solid tumor malignancies. Compared to benchmarked methods, MethSig delivers well-calibrated Quantile-Quantile (Q-Q) plots and more reproducible identification of DNAme drivers in independent cohorts. Importantly, MethSig achieved higher sensitivity and specificity in the inference of likely DNAme drivers compared to extant methods. Finally, the MethSig framework was extended to methylation array data (MethSig-array) and demonstrated the ability to identify candidate DNAme drivers, enriched in TSGs and associated with gene silencing, as well as disease outcomes. Thus, MethSig represents a novel statistical framework to infer DNAme drivers of cancer genesis and relapse, paving the way towards enhanced understanding of the role of epigenetic changes in cancer.

Results

MethSig infers putative DNAme drivers through the application of an optimized background model for stochastic hypermethylation

Promoter hypermethylation was measured using differentially hypermethylated cytosine ratio (DHcR), defined as the ratio of hypermethylated cytosines (HCs) to the total number of CpGs profiled in promoters (Fig. 1A; Methods). We reasoned that a large number of high DHcR promoters may result from passenger hypermethylation due to the non-uniform distribution of stochastic hypermethylation rate across the genome (5). Extant inference tools relying on uniform background models will thus lead to spuriously high number of

significantly affected promoters. To illustrate this point, we compared simplified scenarios of constant versus varying HC rate across the genome, both sharing the same average promoter HC rate (Fig. 1B). This analysis demonstrated that when the HC rate varies across the genome, the uniform background assumption leads to many of the highly hypermethylated genes being falsely determined as significantly altered.

To overcome this challenge, we devised a model to estimate promoter-specific background hypermethylation rate (Fig. 1C; Supplementary Fig. S1A-E). We included covariates known to impact hypermethylation rate such as gene expression and replication time (Fig. 1D; Supplementary Fig. S1D-E), as well as promoter proportion of discordant reads, PDR, (Fig. 1A; Supplementary Fig. S1B-C), a metric developed to characterize stochastic DNAme changes (5). Intuitively, loci and samples with high PDR (Fig. 1A, locus B) suggest lower reliability in DNAme driver identification compared to those with low PDR (Fig. 1A, locus A), akin to the role of background mutation rates in cancer driver gene inference.

To determine if a single gene promoter is significantly hypermethylated in each sample, MethSig first generates the expected promoter hypermethylation (expected DHcR) based on a beta regression model and relevant covariates (Fig. 1C, step 1). Second, MethSig tests observed promoter hypermethylation (tumor DHcR) against the expected DHcR (Fig. 1C, step 2). Of note, the beta regression model was found to deliver good fits to tumor DHcR in most genes and patients (Supplementary Fig. S1F-G). Moreover, candidate DNAme drivers were not preferentially nominated by a small subgroup of patients (Supplementary Fig. S1H). Third, hypermethylation signal is aggregated across the cohort as a stronger candidate DNAme driver is likely to affect a larger number of patients. This cross-patient aggregation procedure enables the estimation of hypermethylation enrichment at the cohort level (Fig. 1C, step 3).

To compare MethSig's performance to the currently used methods, we applied three widely used methods to identify hypermethylation in cancer: t-test, methylKit (15) and globalTest (16) (see Methods and Supplementary Data for details). These methods were applied to prospective RRBS profiling of the CLL8 cohort (17,18) (Fig. 1E; Supplementary Table S1). Notably, a comparison of top candidates across methods showed relatively limited overlap, reinforcing the need to develop better statistical models to nominate candidate DNAme drivers (Supplementary Fig. S1I; Supplementary Table S2-S3).

In a well-calibrated statistical model, p-values are uniformly distributed when the null hypothesis is true, and all other assumptions are met. We thus evaluated the performance of MethSig and benchmarked methods through Q-Q plots (19), an established method to assess the uniformity of p-value distribution in statistical genetics. As we anticipate only a small number of DNAme drivers (as compared to the much larger number of genome-wide stochastic changes), a well-calibrated Q-Q plot will mostly adhere to the diagonal, with few outliers with extreme p-values. Benchmarked methods showed inflated Q-Q plots when applied to the CLL8 dataset, which deviated from the expected line (dashed grey line) across the range of p-values (Fig. 2A, 1st row). Nearly half of gene promoters were identified as candidate DNAme drivers, likely reflecting an underlying global phenomenon such as elevated passenger DNAme alteration in CLL compared to normal B cells, rendering the

task of pinpointing candidate DNAme drivers, with biological and clinical significance, highly challenging. In contrast, MethSig exhibited a well-calibrated Q-Q plot, with a deviation factor that more closely approximated 1, and only few candidate DNAme driver p-values deviated from expected (Fig. 2A, 1st row).

Next, to test whether candidate DNAme driver nomination with MethSig is robust across datasets, we applied MethSig to an independent, previously published CLL RRBS dataset (CLL-DFCI; Fig. 1E). Similarly, MethSig resulted in a well-calibrated Q-Q plot and a deviation factor closer to 1, compared to benchmarked methods (Fig. 2A, 2nd row). To further test the generalizability of MethSig, we applied MethSig to available RRBS datasets of three additional tumor types (Fig. 1E; Supplementary Table S1). The performance of MethSig was maintained in a multiple myeloma cohort (MM-CNRS) and two solid tumor datasets (ductal carcinoma in situ, DCIS-MDACC (20); glioblastoma, GBM-MUV(21)), resulting in well-calibrated Q-Q plots (Fig. 2A, 3rd to 5th row). Here too, benchmarked methods showed inflated Q-Q plots, suggesting that these methods are challenged to distinguish oncogenic DNA hypermethylation from global DNAme changes.

We performed extensive model optimization to ensure the robustness of DNAme driver inference by MethSig, confirming that MethSig included informative covariates, parameters and methodology (Supplementary Fig. S2A-H and S3A-G; see Supplementary Data for details). Notably, model optimization was also performed for benchmarked methods (e.g., using over-dispersion correction option in methylKit), however improvements of Q-Q plots were subtle (Supplementary Fig. S3H).

MethSig provides reproducible and transcription-relevant candidate DNAme drivers, enriched in genes dysregulated across cancer types

Unlike passenger changes, DNAme drivers are anticipated to affect a large proportion of tumors and associated with silenced gene expression. Thus, we hypothesized that accurate inference of DNAme drivers can be assessed through reproducibility across independent patient cohorts, and association with gene silencing.

Considering varied numbers of candidate DNAme drivers identified by different methods using identical p-value cutoff, we compared an equal number of top ranking DNAme drivers to test the reproducibility of DNAme drivers nominated by different methods across the two CLL cohorts. MethSig resulted in a significantly higher overlap across the two cohorts, compared to benchmarked methods (Fig. 2B; Supplementary Fig. S4A).

Next, we tested whether DNAme drivers nominated by MethSig are more frequently linked to gene silencing compared to other methods (see Supplementary Data for details). Area under the receiver operating characteristic (AUROC) showed that MethSig achieved higher performance compared to benchmarked methods in identifying DNAme drivers associated with gene silencing (Fig. 2C; Supplementary Fig. S4B). Indeed, candidate DNAme drivers identified by MethSig were significantly more enriched in silenced genes compared to benchmarked methods or to randomly selected genes (Fig. 2C). Similar findings were observed in the DCIS and GBM cohorts, where matched DNAme and RNA-seq data are available (Supplementary Fig. S4C-D).

Integrating data across the two CLL cohorts, MethSig nominated 189 candidate DNAme drivers out of 9,661 promoters captured by RRBS and with available input covariates (Supplementary Table S2; see Supplementary Fig. S5A-B, Supplementary Table S4 and Supplementary Data for additional analyses to rule out confounders including CpG density, B cell subtype specific epigenetic profiles, copy number changes and driver mutations). Samples where candidate DNAme drivers were found to be hypermethylated have a higher fraction of highly methylated promoters (DHcR > 0.75) compared to samples without hypermethylation (Supplementary Fig. S5C), suggesting high clonality level consistent with positive selection (see Supplementary Fig. S5D-F and Supplementary Data for further characterization of DNAme drivers). While known transcription factor (TF) binding motifs did not show enrichment in DNAme drivers, we observed significantly higher H3K27me3 signal at putative driver loci compared with non-driver loci, suggesting that MethSig candidates may in part conform to the model of promoting cancer development due to locking-in of repression by H3K27me3 (Supplementary Fig. S5G; Supplementary Data).

To interrogate their biological significance, we performed a pathway enrichment analysis of candidate DNAme drivers. Candidate CLL, MM, DCIS and GBM DNAme drivers were enriched in genes hypermethylated or silenced across tumor types, and associated with poor clinical outcome (22) (Supplementary Table S5; Benjamini-Hochberg false discovery rate, BH-FDR Q < 0.25). CLL and DCIS DNAme drivers were also enriched in genes downregulated by Myc and genes upregulated by p53 (22). Specifically, DCIS DNAme drivers were enriched in genes silenced in breast ductal carcinoma versus normal ductal breast cells (22), consistent with DNAme drivers-mediated repression of corresponding genes.

Candidate CLL DNAme drivers include established TSGs, and were functionally validated to enhance cancer cell fitness

Candidate CLL DNAme drivers included a well-established TSG, *DUSP22*, whose function as a TSG is silenced through promoter hypermethylation, as demonstrated previously in CLL (23). In addition to *DUSP22*, MethSig also identified other TSGs as putative CLL DNAme drivers such as *RPRM* and *SASH1.* RPRM is known to cooperate with p53 leading to cell cycle arrest at G2 phase and has been reported to be hypermethylated or inactivated in carcinomas (24,25). *SASH1* encodes a scaffold protein involved in the TLR4 signaling pathway (26), which has been demonstrated to be a key signaling pathway in CLL (27).

To functionally validate candidate DNAme drivers identified with MethSig, given the limitations of demethylation agents and current dCas9 guided DNAme modification (Supplementary Data), we generated CRISPR/Cas9 mediated knockout (KO) to mimic gene silencing via promoter hypermethylation. We generated KO of three candidate DNAme drivers – *DUSP22, RPRM* and *SASH1* (see Supplementary Data for selection criteria). Of note, these candidates were suitable for functional validation given baseline gene expression, and minimal promoter methylation in the HG3 cell line (Supplementary Fig. S6A-F).

After transduction with Cas9 and locus-specific targeting sgRNAs, HG3 cells were cultured with three leading CLL therapeutic agents: ibrutinib (a targeted BTK inhibitor), fludarabine (a key chemotherapy backbone in CLL chemoimmunotherapy regimens), and venetoclax (a

BH3 mimetic) (17,28,29). HG3 cells transduced with a non-targeting sgRNA (HG3-mock) were used as control. After 11 days (~7 doubling times) of ibrutinib treatment, we observed higher fitness in cells with sgRNAs targeting all three candidate DNAme drivers (Fig. 3A). In contrast, only the *DUSP22* KO cells showed higher proliferation after fludarabine treatment (Fig. 3B), and none of the KO led to greater proliferation with venetoclax, suggesting that DNAme drivers may have context-specific effects (Fig. 3C).

HG3 cells are known to show clonal diversity (30), which may impact bulk CRISPR/Cas9 KO. Furthermore, the *DUSP22* locus is present in only one copy in HG3 cells due to a partial loss of the chromosome 6p (30), which may contribute to the greater effect in *DUSP22* KO with fludarabine compared to *RPRM* (Fig. 3B). We therefore further generated stable KO HG3 clones of *RPRM* and *DUSP22* through single cell cloning (Fig. 3D; Methods; Supplementary Data). For *RPRM*, we identified a clone with bi-allelic frameshift inducing indels and a second clone with mono-allelic frameshift deletion (Fig. 3E). As *DUSP22* locus is present in only one copy in the HG3 cell line, we generated two separate clones with complete gene KO by introducing frameshift indels in the remaining allele (Fig. 3F). A single cell derived clone with a non-targeting sgRNA was used as a control (mock cell line). After culturing all clones without treatment for 7 days, we observed faster growth for the *RPRM* KO clones with a gene dose effect, compared to controls (Fig. 3G). Similarly, a KO clone for *DUSP22* showed a significantly higher proliferation (Fig. 3G). These data are consistent with a fitness advantage in the absence of treatment, and the enrichment of these DNAme drivers in the previously untreated CLL8 cohort.

In agreement with our above results showing that candidate DNAme driver disruption confers resistance to treatment with leading CLL agents, we observed greater survival for the *RPRM* KO clones under ibrutinib and fludarabine with a gene dose effect (Fig. 3H; Supplementary Fig. S6G). Supporting the role of *DUSP22* as a candidate DNAme driver, both KO clones for *DUSP22* showed improved survival under ibrutinib and fludarabine treatment (Fig. 3I; Supplementary Fig. S6H). However, similar to the bulk transduction experiments, no fitness differences were observed with venetoclax treatment across KO clones (Supplementary Fig. S6G-H).

MethSig-nominated CLL DNAme drivers provide independent prognostic information, and are associated with adverse outcome

We next sought to test the clinical significance of candidate DNAme drivers in the well-annotated CLL cohorts. Promoters whose hypermethylation is associated with failure-free survival (FFS) were defined as true positives (see Supplementary Data for details) while other promoters were defined as true negatives. In CLL8, MethSig resulted in highest AUROC compared to benchmarked methods, and DNAme drivers identified by MethSig were enriched in genes associated with outcome compared to the benchmarked methods or randomly selected genes (Fig. 4A). We further validated this association with clinical outcome in the independent CLL-DFCI cohort (Supplementary Fig. S7A). Notably, MethSig also achieved higher AUROC compared to other methods when we combined two key features that are likely to be associated with DNAme drivers (i.e., either silenced by promoter hypermethylation or associated with FFS; Supplementary Fig. S7B). These results

confirm that MethSig provides a non-incremental advance in the ability to identify likely DNAme drivers with high sensitivity and specificity.

Taking advantage of the large sample size in CLL8 cohort, we sought to further triage the list of DNAme drivers by evaluating the clustering of methylated CpG positions. Intuitively, promoters with a non-random distribution of methylated CpGs are more likely to exert repression on corresponding genes and result in a substantial phenotypic impact (Fig. 4B). We used the maximum number of consecutive methylated CpGs to quantify the clustering degree of methylated CpGs (see Supplementary Fig. S7C and Supplementary Data for details). Indeed, we observed higher clustering of methylated positions in samples where the gene was predicted to be hypermethylated compared to other samples (Fig. 4B), and therefore applied an additional criterion of higher level of clustering, decreasing the number of nominated CLL DNAme drivers to 122 (Supplementary Table S2).

To examine the prognostic value of DNAme drivers, we developed a clinical prediction score based on candidate DNAme drivers ($n = 122$). Elastic net regression (31) with a Cox proportional hazards model was used (see Supplementary Fig. S7D and Supplementary Data for model selection) to assign weights (coefficients) in terms of their contribution to the prediction of FFS to each candidate DNAme driver. To safeguard against overfitting and poor generalizability, CLL8 was used as the training set to select candidate DNAme driver coefficients, while CLL-DFCI was designated as an independent, test cohort not used in the training process.

Candidate DNAme drivers selected by the regression model included all three functionally validated TSGs, whose hypermethylation defined the subset with the least favorable prognosis (Fig. 4C). Higher risk score (greater than median) was significantly associated with shorter FFS in the training set (Fig. 4D, median FFS was 41.2 months in patients with high risk, and not reached in patients with low risk, hazard ratio 2.9, 95% confidence interval [CI] 2.1 to 4.0). Notably, the model was also highly significant in distinguishing patients with high versus low risk of FFS in the test set (Fig. 4E; Supplementary Fig. S7E), and a regression model including established CLL risk indicators demonstrated that DNAme drivers contribute to adverse clinical outcome independently of previously established risk factors (Fig. 4F; Supplementary Fig. S7F-H; Supplementary Data).

MethSig identifies relapse-specific DNAme drivers in CLL

Our data demonstrated that MethSig is an effective tool to nominate cancer DNAme drivers through the comparison of primary malignant (T1) versus controls. We sought to extend the application of MethSig to identify DNAme drivers of relapse disease after fludarabine based chemotherapy through the comparison of relapse (T2) versus control samples (Fig. 1E; CLL8). Notably, the application of MethSig within the context of relapsed CLLs resulted in an equally well-calibrated Q-Q plot (Fig. 4G), consistent with its ability to identify the infrequent DNAme changes that likely contribute to the relapse phenotype.

We identified T2 specific ($n = 32$), T1 and T2 shared ($n = 88$), and T1 specific DNAme drivers ($n = 101$) (Fig. 4H-I; Supplementary Table S2). In addition to previously observed DNAme drivers (e.g., T1 and T2 shared, *DUSP22*, *SASH1*), T2 specific DNAme drivers

involve additional genes with potential tumor suppressor function, such as *G0S2*. G0S2 can promote apoptosis through BCL2, the therapeutic target of the BH3 mimetic venetoclax in CLL (29,32). A pathway enrichment analysis of T2 specific DNAme drivers revealed enrichment in TP53 targets and DNA damage pathway (22) while T1 and T2 shared or T1 specific DNAme drivers were not enriched in these pathways (Fig. 4I; Supplementary Table S5). The enrichment in TP53 targets and DNA damage pathway of T2 specific DNAme drivers indicates that CLL relapse after chemotherapy may follow an alternative path compared to CLL progression in the absence of therapy, offering novel insights for therapeutic strategies to address drug-resistant or relapsed cancer.

MethSig-array infers candidate DNAme drivers with methylation arrays

Considering wide availability of methylation array data, we designed MethSig-array under the same statistical framework proposed by MethSig. Of note, promoter PDR cannot be estimated by array data, which does not provide read-level methylation information, and as shown in the covariate analysis, promoter PDR provides an important contribution to the model (Supplementary Fig. S3B).

Nonetheless, we applied MethSig-array to Infinium HumanMethylation450 arrays of 18 tumor types in TCGA Pan-Cancer analysis project (33) (Supplementary Table S6). As anticipated, the deviation factors of Q-Q plots derived from MethSig-array were closer to 1, compared to higher deviation factors of benchmarked methods (Fig. 5A). To further evaluate the performance of MethSig-array, AUROC was used to assess the sensitivity and specificity in the inference of likely DNAme drivers, which were defined following three key readouts: association with gene silencing, association with disease outcome, and enrichment with TSGs using different published catalogues (Supplementary Data). MethSig-array achieved higher AUROC compared to benchmarked methods in the inference of likely DNAme drivers associated with gene silencing (Fig. 5B) and clinical outcome (Fig. 5C). MethSig-array also resulted in highest AUROC compared to benchmarked methods in the inference of TSGs (Supplementary Fig. S7I, OncoKB (34) or the TCGA cancer driver study (35)). For example, *SOX17* was identified as a DNAme driver in 13 different tumor types (Supplementary Data), which encodes a TF involved in embryonic development and cell fate (9). Hypermethylation and downregulation of *SOX17* have been described in multiple cancer types, which implies the broad tumor suppression function of *SOX17* gene (9). Another important TSG is *RASSF1*, which was identified as a DNAme driver in 6 tumor types (Supplementary Data). RASSF1 is a microtubule-associated and multitasking scaffold protein communicating with the RAS pathway, estrogen receptor signaling and Hippo pathway (36). *RASSF1* methylation is proposed as a candidate maker in many cancer types (36). Other identified important TSGs included genes in a plethora of cancer-related cellular pathways such as DNA repair (*MLH1*, *RBBP8*), apoptosis (*WIF1*, *SFRP1*) and tyrosine kinase cascades (*SOCS3*) (7,8). Collectively, these data confirm that MethSig can accurately infer likely DNAme drivers across cancer with both array and next-generation sequencing (NGS) based methylation assays.

Discussion

Aberrant gene function due to acquired epigenetic abnormalities have been highlighted as key features of cancer over the last decade, implicated in cancer initiation, progression and treatment resistance (1,2). Although the causal role of DNAme in cancer remains to be conclusively determined (37), genome-wide DNAme analyses have provided comprehensive surveys of the cancer epigenome and tumor-associated DNAme changes, and have proposed that these changes fuel the malignant process through TSG silencing and other mechanisms (1,2).

However, in the context of steadily growing DNAme sequencing datasets, a major challenge remains: to identify the DNAme changes involved in tumor progression among the abundant stochastic DNAme changes that occur in cancer cells. This challenge is reminiscent of the challenge of distinguishing driver from passenger mutation in cancer exome or genome data. While for the latter challenge progress has been achieved through increasingly sophisticated inference tools that model the varying background mutation rate across the genome (12), inference tools in cancer epigenomics largely rely on uniform background models (38). Given the recent observation that stochastic DNAme varies widely in different genomic regions (5), statistical models relying on a uniform background assumption are anticipated to lead to spuriously high numbers of significantly affected regions.

Drawing on lessons learned in cancer genomics, we posited that robust nomination of oncogenic DNAme changes requires a rethinking of the statistical inference process to enable the differentiation of driver promoter hypermethylation changes (DNAme drivers) from the far larger number of stochastic DNAme changes without biological consequences (passenger DNAme changes). To address this challenge, we developed a statistical inference framework accounting for varying stochastic hypermethylation rate across the genome and between samples – MethSig. The model provides the expected promoter DNAme changes between tumor and control samples, allowing the identification of loci where the observed hypermethylation significantly exceeds expectation, potentially reflecting positive selection of fitness-enhancing candidate DNAme drivers.

We applied MethSig to methylation sequencing data of two CLL cohorts, including 304 CLLs from a prospective clinical trial, as well as to other malignancies with available DNAme data (MM, DCIS and GBM), and benchmarked against state-of-the-art methods. Compared with benchmarked methods, MethSig resulted in well-calibrated Q-Q plots, higher reproducibility in DNAme driver inference across independent cohorts, and increased sensitivity/specificity in the inference of likely DNAme drivers. These observations confirm that MethSig allows to separate specific cancer related DNAme drivers, which cause gene downregulation and phenotypic changes associated with tumoral progression, from stochastic passenger DNAme changes. Notably, the performance of MethSig was maintained across both hematological malignancies (CLL and MM) and solid tumors (DCIS and GBM), suggesting broad applicability for creating catalogues of candidate DNAme drivers of cancer genesis and relapse. Moreover, while MethSig was extended to array data and provided a non-incremental improvement in DNAme driver inference, we anticipate that future shift

towards NGS data will leverage the even higher performance of MethSig with read-level data.

Our data showed that MethSig can also account for broad phenomena that alter DNAme profiles in identifying gene-specific DNAme drivers. For example, DNAme of the cell-of-origin represents one of the strongest sources of variation in the cancer epigenome (4). Indeed, in CLL, DNAme has been shown to strongly encode normal B cell epigenetic reprogramming during differentiation, allowing high resolution inference of the differentiation state of the initially transformed B cell (4). It is therefore notable that MethSig candidate DNAme drivers were not significantly enriched in the most variable methylated regions identified between naïve and class-switched memory B cells (Supplementary Data).

CLL candidate DNAme drivers included TSGs inactivated through hypermethylation, such as *DUSP22*, *RPRM*, and *SASH1*, which may play important roles in the initiation and relapse of CLL. To functionally validate these candidate DNAme drivers in CLL cells, we generated single or double allele frameshift KO. While transformed cells showed superior fitness in the absence of drug selection and with therapy, DNAme drivers showed context-specific effects, which may underlay some of the heterogeneity in CLL clinical course (5). These data also show that venetoclax therapy may uniquely overcome these mechanisms, providing rationale for future DNAme driver guided trials. Of note, our data demonstrated improved risk stratifications based on candidate DNAme drivers, independent of known prognostic factors in CLL, suggesting that DNAme drivers contribute to adverse clinical outcome.

While we believe that this work presents a transformative advance in identifying DNAme drivers with higher sensitivity and specificity, further work will be needed to improve performance and reduce false positive candidates. This may be achieved through expanded datasets and future discovery of additional informative covariates, following the example of genetic driver inference where successive versions resulted in a continuous improvement in performance and reduction in false positives (12,13). Given the limitations of RRBS, including poor coverage of distal regulatory elements or other regions that are CpG poor, the future exploration of DNAme changes in other genomic areas will be greatly empowered by larger whole genome DNAme sequencing datasets. Finally, hypomethylation may also play important roles in cancer genesis by impacting genome stability. Further efforts will be needed to enable statistical inference of those epigenetic events.

Collectively, our data support a novel framework for the analysis of DNAme changes in cancer to specifically identify DNAme drivers of disease progression and relapse, empowering the discovery of candidate epigenetic mechanisms that may enhance cancer cell fitness. This work addresses a central gap between cancer epigenetics and genetics, where such tools have had a transformative impact in precision oncology and cancer gene discovery. We envision that inference tools such as MethSig, coupled with novel DNAme sequencing modalities (39) and emerging tools for epigenetic editing (40), may herald a new era in cancer epigenomics in which large cohort studies will provide precision identification of oncogenic DNAme drivers for improved patient stratification and therapeutic targeting.

Methods

Sample acquisition:

For CLL8 cohort, blood was obtained from previously untreated patients enrolled in a prospective, randomized, open-label CLL8 trial (17,18) before the first cycle of treatment. Written informed consent for genomic sequencing of patient samples was obtained prior to the initiation of sequencing studies. For MM-CNRS cohort, bone marrow of patients presenting with previously untreated MM ($n = 24$) or at relapse ($n = 20$) was obtained after patients' written informed consent in accordance with the IRB and the Montpellier University Hospital Centre for Biological Resources (DC-2008-417). Genomic DNA was extracted from CLL and MM cells.

RRBS:

RRBS libraries were generated by digesting genomic DNA with MspI to enrich for CpG-rich fragments, and then ligated to barcoded TruSeq adapters (Illumina) to allow immediate subsequent pooling. It was followed by bisulfite conversion and PCR, as previously described (14). Libraries were sequenced and aligned to the bisulfite-converted hg19 reference genome using Bismark v0.15.0 (RRID: SCR_005604) (41).

Promoter hypermethylation:

Promoter (defined as ± 2 kb windows centered on RefSeq transcription start site) hypermethylation was measured using DHcR, defined as the ratio of HCs to the total number of CpGs profiled in promoter. HCs of each sample were defined as CpGs at which DNAme is statistically higher than the control (FDR = 20%, Chi-squared test) (6). Only CpGs with read depth greater than 10 were included in the analysis. DHcR of each normal sample was calculated in the same way as for the tumor samples, testing against all normal sample controls. This was followed by averaging DHcR of all the normal samples as the normal DHcR.

Promoter PDR:

If all the CpGs on a specific read are methylated or 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 divided by the total number of reads that cover that location. Promoter PDR is given by averaging the values of individual CpGs, as calculated for all CpGs within the promoter of interest that are covered by a minimum of 10 reads that contain at least 4 CpGs. The normal PDR was calculated by averaging PDR of all the normal samples.

Algorithmic procedure:

The superscripts n, t, e are shorthand for normal, tumor and expected. The subscripts i and j represent gene i and sample j. The model was processed as following steps:

1. Estimate expected hypermethylation of tumor samples ($DHcR^e$): The independent variable matrix (X) has number of genes times number of patients rows (Equation 1). The beta regression model was implemented by R package

betareg (42) (Equation 2). Next, predicted distribution of $DHcR^t$ (used as distribution of $DHcR^e$ in the following analysis) was estimated using $\hat{\alpha}$ and $\hat{\beta}$ derived from the above beta regression model (Equation 3).

$$X_{i,j} = \left(DHcR_i^n, PDR_i^n, gexp_i^n, reptime_i, PDR_{i,j}^t, depth_{i,j}^t, ncpg_{i,j}^t\right) \quad (1)$$

$$DHcR_{i,j}^t = beta(\alpha + \beta X_{i,j}) \quad (2)$$

$$DHcR_{i,j}^e = beta(\hat{\alpha} + \hat{\beta} X_{i,j}) \quad (3)$$

2. Evaluate if $DHcR^t$ (observed) is significantly higher than $DHcR^e$ (expected): $DHcR^t$ was tested against the distribution of $DHcR^e$. The patient-specific p-value indicates the probability that observed promoter hypermethylation is significantly higher than expected (Equation 4). Only genes whose promoter hypermethylation significantly exceeded expectation will be assigned as patient-specific DNAme drivers ($P < 0.05$).

$$p_{i,j} = P(DHcR_{i,j}^e > DHcR_{i,j}^t) \quad (4)$$

3. Determine if promoter hypermethylation is overrepresented in patients (DNAme driver): Wilkinson p-value combination method was used to combine p-values from different patients to identify those frequently recurring DNAme drivers (43). To eliminate the effect of cohort size on p-value combination results, MethSig randomly sampled equal number of patients ($K = 10$) iteratively ($S = 100$) and used lower quartile of combined p-values to identify DNAme drivers. Wilkinson p-value combination was performed by R package *metap* (https://cran.r-project.org/web/packages/metap/).

Benchmarked methods:

In the evaluation of benchmarked methods, *t.test* of R 3.3.2, methylKit 1.0.0 (RRID: SCR_005177) and globalTest 5.28.0 (RRID: SCR_001256) were used.

Pathway enrichment analysis:

Pathway enrichment analysis was limited to the chemical and genetic perturbations of the C2 gene set collection (22), which includes gene sets are more specific to cancer processes in different cancer types with or without perturbation. DNAme drivers were tested against all MethSig inferred non-drivers and only pathways with at least 10 inferred genes were included. A hypergeometric test was used to measure the enrichment of DNAme drivers in each gene set, followed by a BH-FDR procedure.

Cell lines:

HG3 (DSMZ #ACC-765, RRID: CVCL_Y547), PGA1 (DSMZ #ACC-766, RRID: CVCL_Y545) and MEC1 (DSMZ #ACC-497, RRID: CVCL_1870) cells were provided by Leibniz Institute DSMZ in August 2018. HEK293T cells (ATCC #CRL-3216, RRID: CVCL_0063) were provided by the American Tissue Collection Center (ATCC, Manassas, VA) in January 2017. Cells were routinely tested for mycoplasma using the MycoAlertTM Mycoplasma Detection Kit (Lonza #LT07-318). All cell lines used for described experiments came from early frozen batches between 3 and 8 passages after cell reception. HG3, PGA1 and MEC1 cells were used for RT-qPCR (Supplementary Data). HEK293T cells were used for production of lentivirus and transduction (Supplementary Data). HG3 cells were used for all the other described experiments.

CRISPR/Cas9 design and cloning:

sgRNA was designed using CHOPCHOP (RRID: SCR_015723) (44), in order to minimize *in silico* predicted off target activity (≤ 1), target the first exon of the genes of interest and have a good predicted efficiency (> 56) based on the PAM sequence and the 3' nucleotide (45). Two sgRNAs have been designed for each targeted locus and the empirically defined most efficient sgRNA was then used for the CRISPR/Cas9 experiments. The sgRNAs were cloned into lentiCRISPRv2 puro (Addgene #98290) as described previously (46).

Generation of Cas9-expressing KO HG3 clones:

To create single and double allele KO HG3 clones, we have transduced cells with CRISPR/Cas9 and a gene-specific sgRNA, performed a 10-day puromycin [0.5 μg ml^{-1}] selection, confirmed Cas9 efficiency using T7E1 – EnGen® Mutation Detection Kit (NEB), created single cell colony by cell sorting and selected single cell clones showing single or double allele KO (indels) by targeting the predicted CRISPR cutting site through NGS.

CLL patients risk model:

The regression model was implemented by R package *glmnet* (RRID: SCR_015505) (47,48). When evaluating the performance of the model in the training and test set, CLL cases were divided into two subgroups based on their predicted risk scores (patients with high versus low risk, median risk score of the cohort was used as the cutoff) and the FFS difference between groups was evaluated using log-rank test.

MethSig-array:

Promoter DHcR and PDR cannot be estimated by array data, which does not provide read-level methylation information. MethSig-array was designed by using average promoter methylation instead of DHcR in the beta regression model and leaving out promoter PDR, under the same statistical framework proposed by MethSig.

Statistical methods:

Statistical analysis was performed with R version 3.3.2 (https://www.R-project.org/). All p-values were two-sided and considered significant at the 0.05 level unless otherwise noted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

R. Chaligne is supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 750345. D.A. Landau is supported by the Burroughs Wellcome Fund Career Award for Medical Scientists, Vallee Scholar Award, Pershing Square Sohn Prize for Young Investigators in Cancer Research, the Sontag Foundation Distinguished Scientist award, and the National Institutes of Health (NIH) Director's New Innovator Award (DP2-CA239065). J. Moreaux is supported by Institut Universitaire de France (IUF), by grants from INCa (Institut National du Cancer; PLBIO2018-160 PIT-MM), ANR (PLASMADIFF-3D), ANR (TIE-Skip; 2017-CE15-0024-01) and SIRIC Montpellier Cancer INCa_Inserm_DGOS_12553. S. Stilgenbauer is supported by the DFG, SFB1074 subprojects B1 and B2. This work was supported by the NCI (1R01CA229902), the Leukemia, Lymphoma Society, Quest for Cures program and LLS-SCOR 7012-16. The authors thank all patients and their physicians for trial participation and donation of samples, the DCLLSG; Sabrina Schrell and Christina Galler for their excellent technical assistance on CLL8 sample work-up.

References

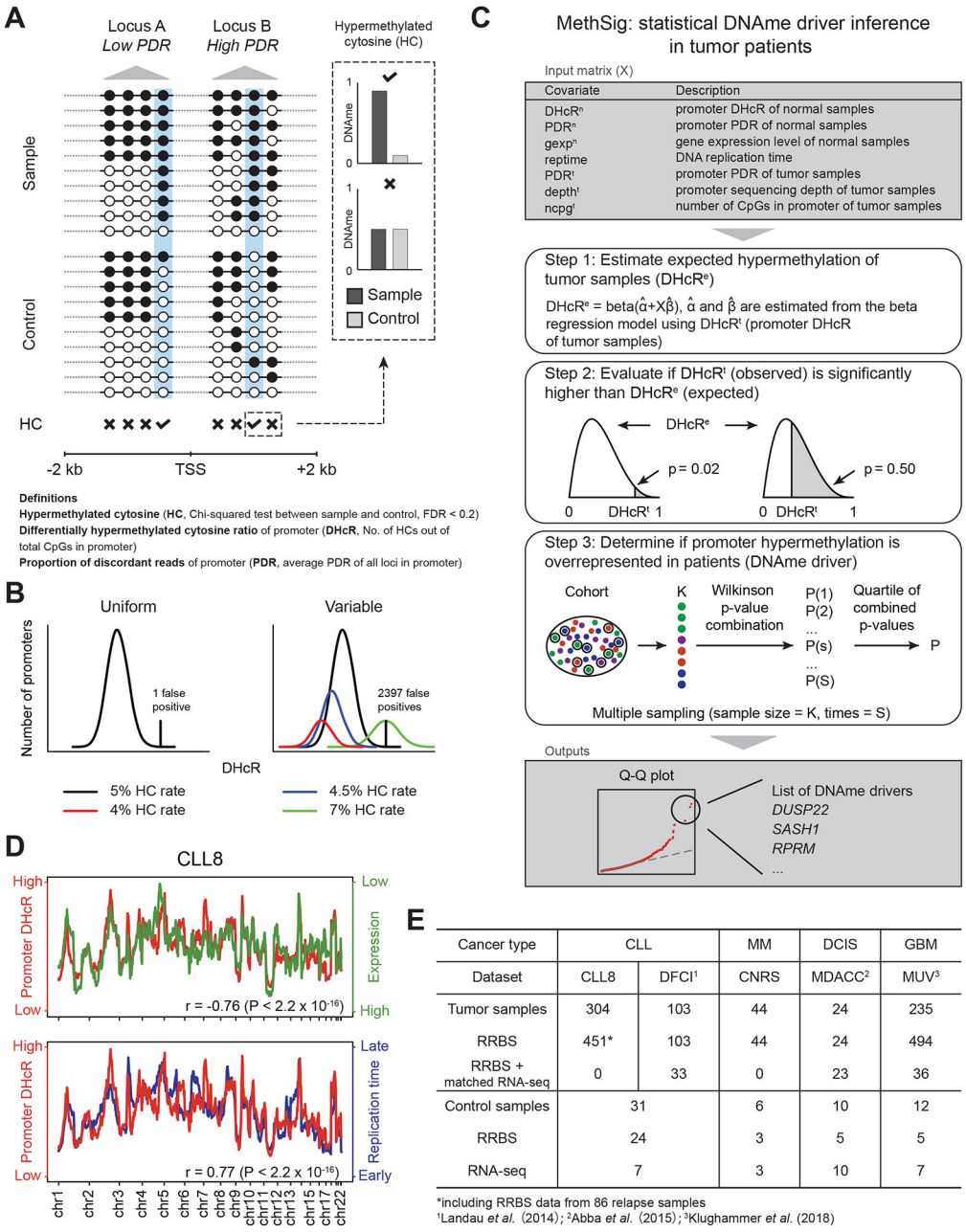
1. Robertson KD. DNA methylation and human disease. Nat Rev Genet2005;6:597–610. [PubMed: 16136652]
2. Baylin SB, Jones PA. Epigenetic determinants of cancer. Cold Spring Harb Perspect Biol2016;8:a019505. [PubMed: 27194046]
3. Koch A, Joosten SC, Feng Z, De Ruijter TC, Draht MX, Melotte V, et al.Analysis of DNA methylation in cancer: Location revisited. Nat Rev Clin Oncol2018;15:459–66. [PubMed: 29666440]
4. Oakes CC, Seifert M, Assenov Y, Gu L, Przekopowitz M, Ruppert AS, et al.DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat Genet2016;48:253–64. [PubMed: 26780610]
5. Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, et al.Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. Cancer Cell2014;26:813–25. [PubMed: 25490447]
6. Pan H, Jiang Y, Boi M, Tabbò F, Redmond D, Nie K, et al.Epigenomic evolution in diffuse large B-cell lymphomas. Nat Commun2015;6:6921. [PubMed: 25891015]
7. Esteller MEpigenetic gene silencing in cancer: The DNA hypermethylome. Hum Mol Genet2007;16:R50–9. [PubMed: 17613547]
8. Yu Y, Chen L, Zhao G, Li H, Guo Q, Zhu S, et al.RBBP8/CtIP suppresses P21 expression by interacting with CtBP and BRCA1 in gastric cancer. Oncogene2020;39:1273–89. [PubMed: 31636387]
9. Li L, Yang WT, Zheng PS, Liu XF. SOX17 restrains proliferation and tumor formation by down-regulating activity of the Wnt/β-catenin signaling pathway via trans-suppressing β-catenin in cervical cancer. Cell Death Dis2018;9:741. [PubMed: 29970906]
10. Landan G, Cohen NM, Mukamel Z, Bar A, Molchadsky A, Brosh R, et al.Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet2012;44:1207–14. [PubMed: 23064413]
11. Shipony Z, Mukamel Z, Cohen NM, Landan G, Chomsky E, Zeliger SR, et al.Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature2014;513:115–9. [PubMed: 25043040]
12. Lawrence MS, Stojanov P, Polak P, Kryukov GV., Cibulskis K, Sivachenko A, et al.Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature2013;499:214–8. [PubMed: 23770567]
13. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al.Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science2019;364:eaaw2872. [PubMed: 31249028]

14. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, et al.Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol2012;13:R92. [PubMed: 23034176]
15. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al.methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol2012;13:R87. [PubMed: 23034086]
16. Goeman JJ, Van De Geer SA, Van Houwelingen HC. Testing against a high dimensional alternative. J R Stat Soc Ser B Stat Methodol2006;68:477–93.
17. Hallek M, Fischer K, Fingerle-Rowson G, Fink AM, Busch R, Mayer J, et al.Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: A randomised, open-label, phase 3 trial. Lancet2010;376:1164–74. [PubMed: 20888994]
18. Fischer K, Bahlo J, Fink AM, Goede V, Herling CD, Cramer P, et al.Long-term remissions after FCR chemoimmunotherapy in previously untreated patients with CLL: Updated results of the CLL8 trial. Blood2016;127:208–15. [PubMed: 26486789]
19. Wilk MB, Gnanadesikan R. Probability Plotting Methods for the Analysis of Data. Biometrika1968;55:1–17. [PubMed: 5661047]
20. Abba MC, Gong T, Lu Y, Lee J, Zhong Y, Lacunza E, et al.A molecular portrait of high-grade ductal carcinoma in situ. Cancer Res2015;75:3980–90. [PubMed: 26249178]
21. Klughammer J, Kiesel B, Roetzer T, Fortelny N, Nemc A, Nenning KH, et al.The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. Nat Med2018;24:1611–24. [PubMed: 30150718]
22. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst2015;1:417–25. [PubMed: 26771021]
23. Arruga F, Gizdic B, Bologna C, Cignetto S, Buonincontri R, Serra S, et al.Mutations in NOTCH1 PEST domain orchestrate CCL19-driven homing of chronic lymphocytic leukemia cells by modulating the tumor suppressor gene DUSP22. Leukemia2017;31:1882–93. [PubMed: 28017968]
24. Morris MR, Ricketts C, Gentle D, Abdulrahman M, Clarke N, Brown M, et al.Identification of candidate tumour suppressor genes frequently methylated in renal cell carcinoma. Oncogene2010;29:2104–17. [PubMed: 20154727]
25. Xu M, Knox AJ, Michaelis KA, Kiseljak-Vassiliades K, Kleinschmidt-DeMasters BK, Lillehei KO, et al.Reprimo (RPRM) is a novel tumor suppressor in pituitary tumors and regulates survival, proliferation, and tumorigenicity. Endocrinology2012;153:2963–73. [PubMed: 22562171]
26. Blonder J, Park Y-J, Hussainkhel A, Yang C, Veenstra TD, Fuller ME, et al.SASH1 Is a Scaffold Molecule in Endothelial TLR4 Signaling. J Immunol2013;191:892–901. [PubMed: 23776175]
27. Dadashian EL, McAuley EM, Liu D, Shaffer AL, Young RM, Iyer JR, et al.TLR signaling is activated in lymph node–resident CLL cells and is only partially inhibited by ibrutinib. Cancer Res2019;79:360–71. [PubMed: 30498085]
28. Byrd JC, Brown JR, O'Brien S, Barrientos JC, Kay NE, Reddy NM, et al.Ibrutinib versus ofatumumab in previously treated chronic lymphoid leukemia. N Engl J Med2014;371:213–23. [PubMed: 24881631]
29. Fischer K, Al-Sawaf O, Bahlo J, Fink AM, Tandon M, Dixon M, et al.Venetoclax and obinutuzumab in patients with CLL and coexisting conditions. N Engl J Med2019;380:2225–36. [PubMed: 31166681]
30. Quentmeier H, Pommerenke C, Ammerpohl O, Geffers R, Hauer V, MacLeod RAF, et al.Subclones in B-lymphoma cell lines: Isogenic models for the study of gene regulation. Oncotarget2016;7:63456–65. [PubMed: 27566572]
31. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol2005;67:301–20.
32. Welch C, Santra MK, El-Assaad W, Zhu X, Huber WE, Keys RA, et al.Identification of a protein, G0S2, that lacks Bcl-2 homology domains and interacts with and antagonizes Bcl-2. Cancer Res2009;69:6782–9. [PubMed: 19706769]

33. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al.The cancer genome atlas pan-cancer analysis project. Nat Genet2013;45:1113–20. [PubMed: 24071849]
34. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al.OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol2017;1:00011.
35. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al.Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell2018;173:371–385.e18. [PubMed: 29625053]
36. Malpeli G, Innamorati G, Decimo I, Bencivenga M, Nwabo Kamdje AH, Perris R, et al.Methylation Dynamics of RASSF1A and Its Impact on Cancer. Cancers (Basel)2019;11:959.
37. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet2017;18:441–51. [PubMed: 28555657]
38. Klein HU, Hebestreit K. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. Brief Bioinform2016;17:796–807. [PubMed: 26515532]
39. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods2017;14:407–10. [PubMed: 28218898]
40. Lei Y, Huang Y-H, Goodell MA. DNA methylation and de-methylation using hybrid site-targeting proteins. Genome Biol2018;19:187. [PubMed: 30400938]
41. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics2011;27:1571–2. [PubMed: 21493656]
42. Cribari-Neto F, Zeileis A. Beta regression in R. J Stat Softw2010;34:1–24.
43. Wilkinson BA statistical consideration in psychological research. Psychol Bull1951;48:156–8. [PubMed: 14834286]
44. Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res2019;47:W171–4. [PubMed: 31106371]
45. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al.Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol2014;32:1262–7. [PubMed: 25184501]
46. Stringer BW, Day BW, D'Souza RCJ, Jamieson PR, Ensbey KS, Bruce ZC, et al.A reference collection of patient-derived cell line and xenograft models of proneural, classical and mesenchymal glioblastoma. Sci Rep2019;9:4902. [PubMed: 30894629]
47. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw2010;33:1–22. [PubMed: 20808728]
48. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw2011;39:1–13.
49. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al.The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov2012;2:401–4. [PubMed: 22588877]
50. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al.Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal2013;6:pl1. [PubMed: 23550210]

Statement of significance

MethSig provides a novel statistical framework for the analysis of DNA methylation changes in cancer, to specifically identify candidate DNA methylation driver genes of cancer progression and relapse, empowering the discovery of epigenetic mechanisms that enhance cancer cell fitness.



E

Cancer type	CLL		MM	DCIS	GBM
Dataset	CLL8	DFCI[1]	CNRS	MDACC[2]	MUV[3]
Tumor samples	304	103	44	24	235
RRBS	451*	103	44	24	494
RRBS + matched RNA-seq	0	33	0	23	36
Control samples	31		6	10	12
RRBS	24		3	5	5
RNA-seq	7		3	10	7

*including RRBS data from 86 relapse samples
[1]Landau *et al.* (2014); [2]Abba *et al.* (2015); [3]Klughammer *et al.* (2018)

Figure 1. Overview of MethSig and datasets.

A, HC, promoter DHcR and PDR were calculated as shown. Methylation patterns of sample and control are shown (black circles, methylated CpGs; white circles, unmethylated). HCs are highlighted in light blue while DNA methylation (DNAme) of sample and control are detailed inside dashed lines. **B,** Simplified illustration of the challenge on DNAme driver detection due to non-uniform hypermethylation rate across the genome. Here, we simulated a simplified methylome consisting of 20,000 promoters. The average promoter HC rate is 5%, and all hypermethylation is assumed to be due to stochastic processes (i.e., there are no functional DNAme drivers). Two variants of this scenario are compared. Left panel, uniform model whereby all promoters have a constant 5% HC rate. The plot shows a histogram of

promoters by their observed DHcR. The vertical black line indicates a significance threshold that allows a single false positive promoter, which corresponds to 7% HC rate. In contrast, right panel shows a variable model consistent with prior observation of varying HC rate across the genome. In this scenario one quarter of promoters (red) have a HC rate equal to 4%; another half of promoters (blue) have a HC rate of 4.5%; and the final quarter of promoters (green) have a HC rate of 7%. Applying the same threshold for significance (vertical black line corresponding to 7%), 2,397 promoters will be determined as having significant DNAme changes. **C**, An overview of the MethSig statistical inference model. n ~ normal; t ~ tumor; e ~ expected. **D**, Top panel, average promoter DHcR and gene expression level plotted across the entire genome in CLL8. Bottom panel, average promoter DHcR and replication time plotted across the entire genome in CLL8. The average values were calculated based on a sliding window across the human genome with 50 Mb window size and 2.5 Mb step size. Note that gene expression is reversed in the figure in order to emphasize the correlation with methylation (low gene expression at the top and high gene expression at the bottom). Statistical analysis was performed by two-sided Pearson correlation. **E**, Description of datasets.

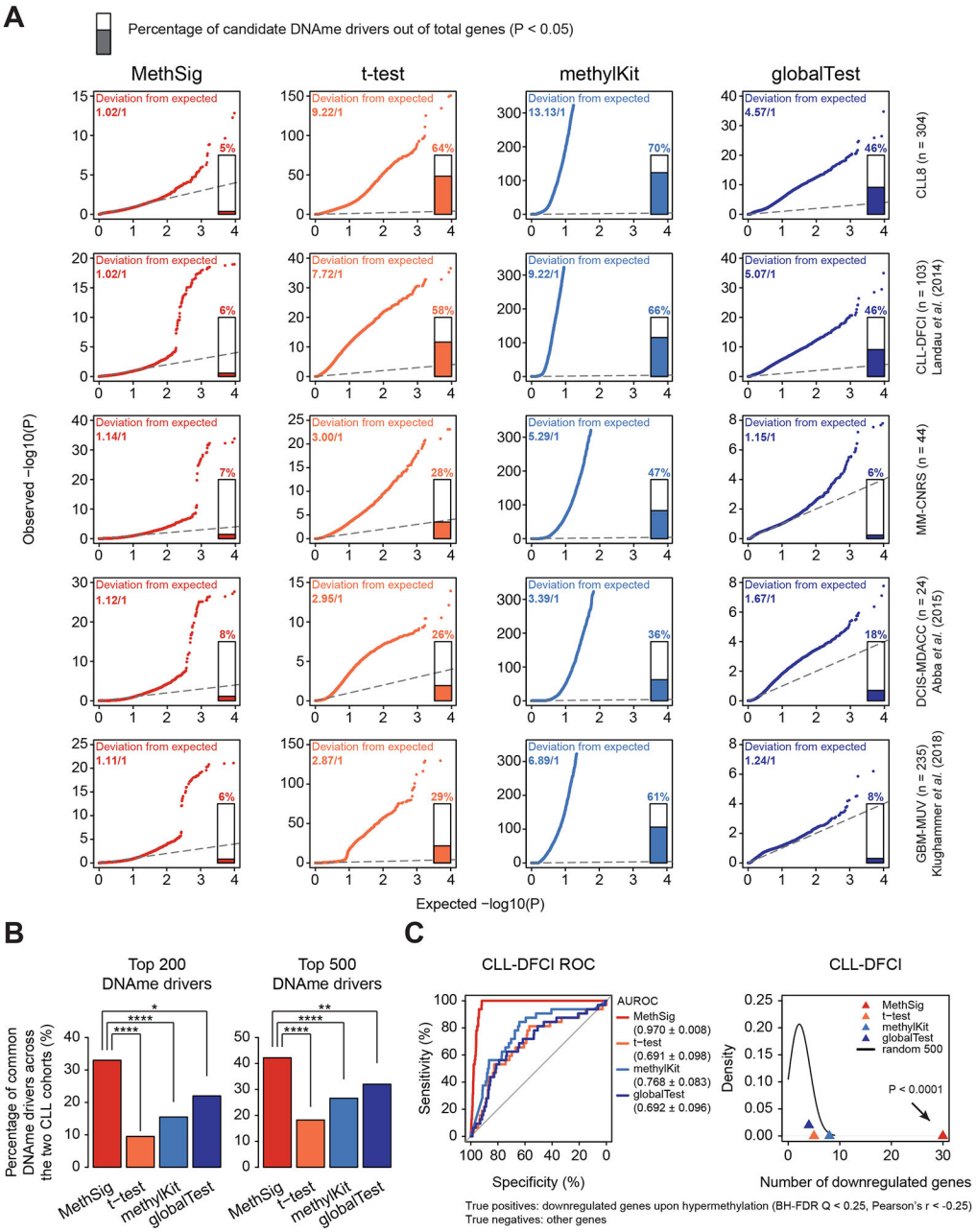


Figure 2. MethSig provides statistically robust, reproducible and functionally relevant DNAme driver inference.

A, Q-Q plots comparing observed −log10 p-values of MethSig and benchmarked methods to expected −log10 p-values. Results of CLL8, CLL-DFCI, MM-CNRS, DCIS-MDACC and GBM-MUV cohorts are listed from top to bottom. For each row, results of MethSig, t-test, methylKit, and globalTest are listed from left to right. Bar plot in each figure represents the percentage of genes with p-values less than 0.05. Deviation from expected factor was the slope value derived from linear regression through zero, modeling the relationship between observed and expected −log10 p-values for each method. Expected p-values used for all the methods were sampled from uniform distribution (*runif* function in R starting from

identical random seed number). **B**, Percentage of shared candidate DNAme drivers between CLL8 and CLL-DFCI cohorts. Top 200 or 500 DNAme drivers ranked by p-values of each method were used. Statistical analysis was performed by Chi-squared test: ****P < 0.0001; **P<0.01; *P < 0.05. **C**, Left panel, MethSig showed higher AUROC compared to benchmarked methods in the inference of likely DNAme drivers in CLL-DFCI (one-sided DeLong's test, compared to method t-test, $P = 1 \times 10^{-8}$; method methylKit, $P = 1 \times 10^{-6}$; method globalTest, $P = 8 \times 10^{-9}$). ROC ± 95% CI are shown. The same list of genes was used among all four methods. Right panel, number of true positives in top 500 candidate DNAme drivers identified by MethSig and benchmarked methods, as well as 500 randomly selected genes in CLL-DFCI cohort. The black curve indicates the density of true positive number estimated by 10,000 times of random selection. Empirical p-value was calculated according to the probability that the number found in top MethSig candidates is greater than the number found in an equal number of randomly selected genes from all the inferred promoters.

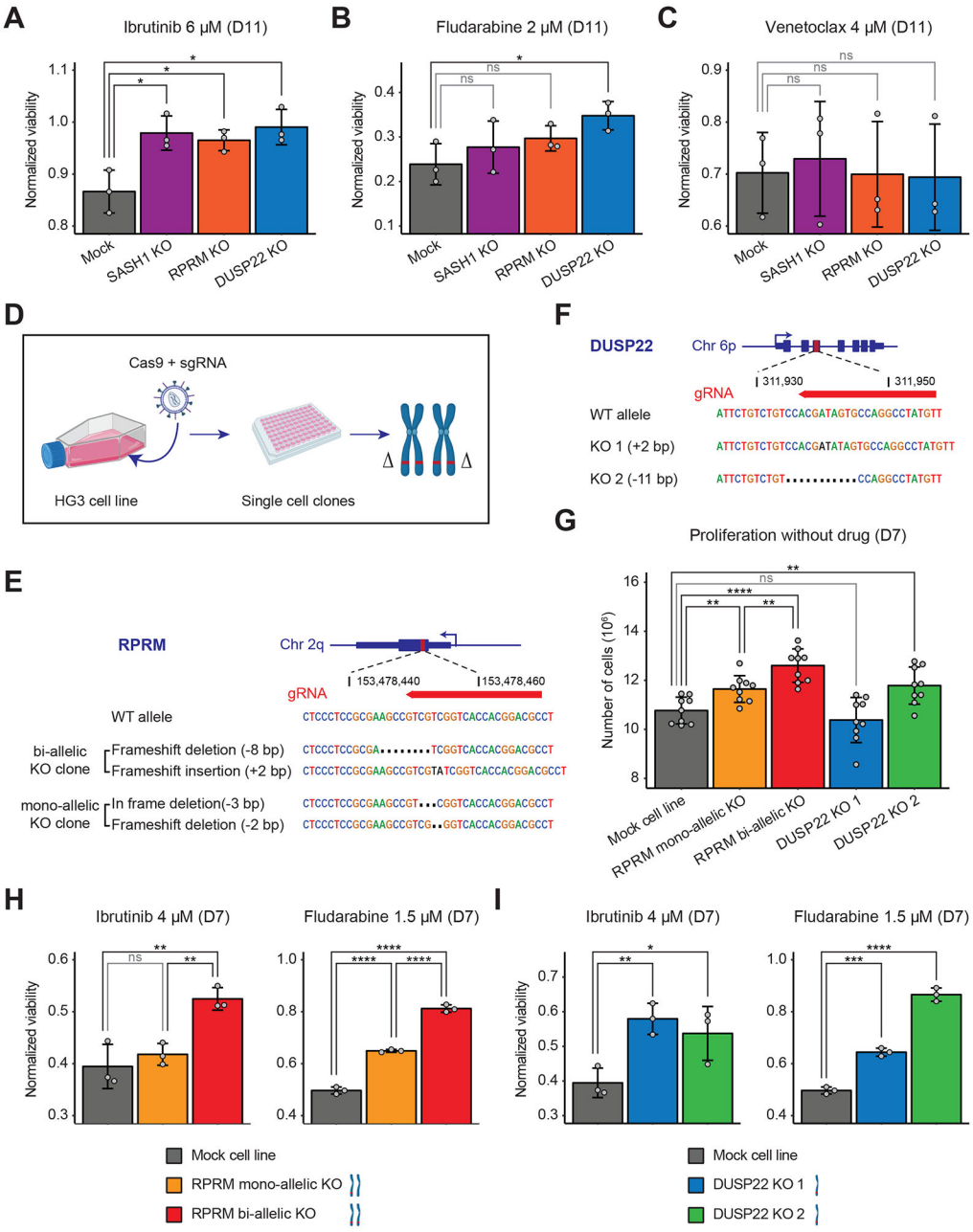


Figure 3. Candidate DNAme driver KO CLL cell lines show superior fitness in drug treatment compared with controls.

A-C, CellTiter-Glo Viability assay of four different cell lines: *SASH1* KO, *RPRM* KO, *DUSP22* KO and HG3-mock as a control after 11 days of exposure to ibrutinib, fludarabine and venetoclax. Triplicates were performed for each condition. **D**, Workflow of the CRISPR KO single cell clone experiment: transduction of the HG3 cell line with a lentivirus containing CRISPR/Cas9 and a targeting sgRNA followed by puromycin selection; isolation and expansion of single cell clones; NGS assessment of KO. **E-F**, Representation of the sgRNA and the indels found in different clones compared to the wild type (WT) allele. For *RPRM* (**E**), 2 single cell clones are shown: a clone with bi-allelic frameshift inducing

indels and a second clone with a mono-allelic frameshift deletion (2 bp deletion inducing a frameshift on the first allele and 3 bp non-frameshift deletion on the second allele). For *DUSP22* (**F**), 2 clones with indels inducing complete KO as the *DUSP22* locus in HG3 is present in only one copy as the result of a partial loss of the chromosome 6p. **G**, CellTiter-Glo Viability assay of 2 single cell *RPRM* KO clones, 2 single cell *DUSP22* KO clones and control (cf method) after 7 days of growth without any drug. Nine replicates were performed for each cell line. **H-I**, CellTiter-Glo Viability assay after 7 days of exposure to ibrutinib and fludarabine. Triplicates were performed for each condition. In **A-C** and **G-I**, data are presented as means ± s.d. Statistical analysis was performed by one-way ANOVA: ****P < 0.0001; ***P < 0.001; **P < 0.01; *P < 0.05; ns, not significant.

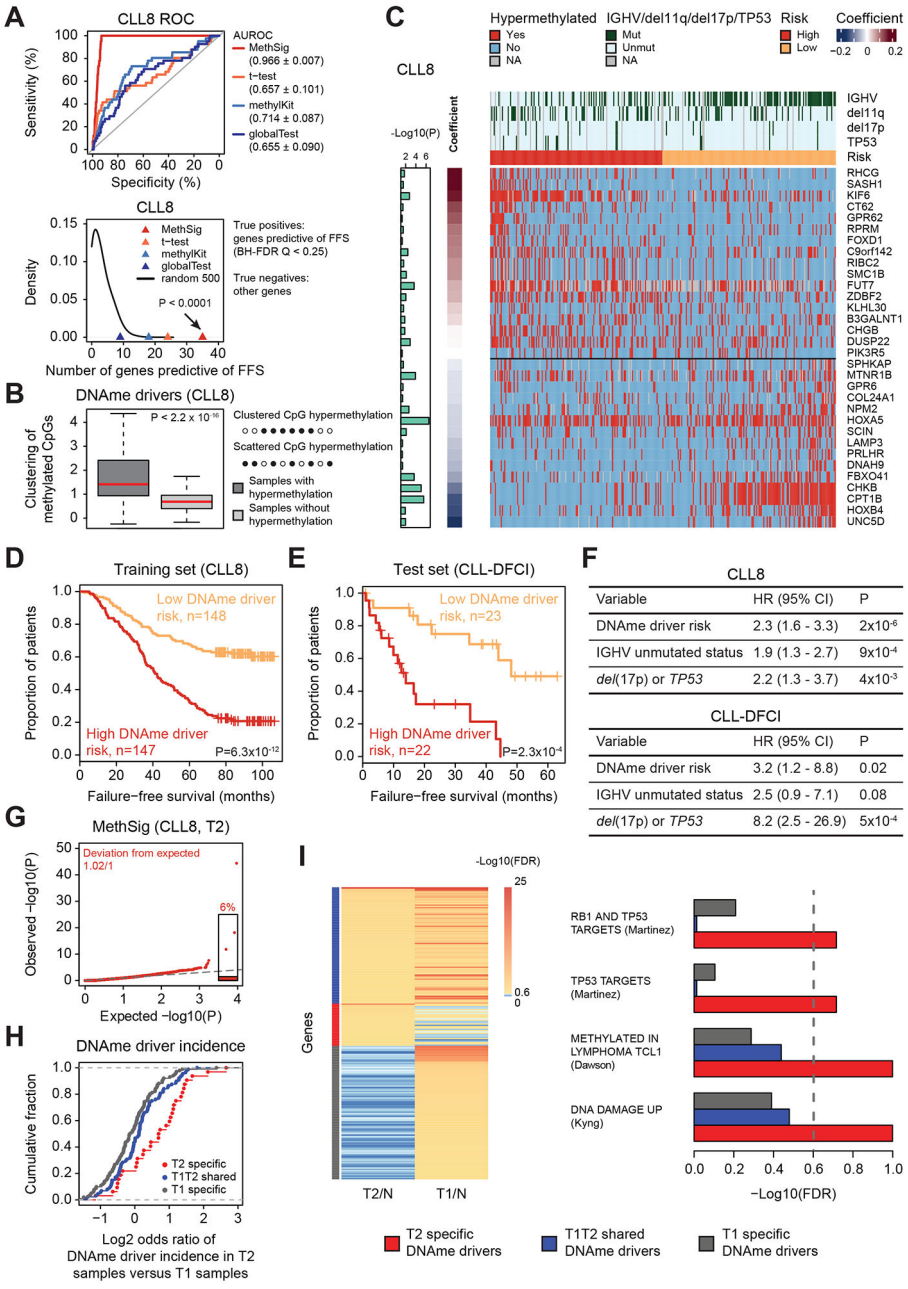


Figure 4. Candidate DNAme drivers provide independent prognostic information, are associated with adverse outcome in CLL, and define distinct alterations in relapsed CLL after fludarabine based therapy.

A, Top panel, MethSig resulted in highest AUROC compared to benchmarked methods in the inference of likely DNAme drivers in CLL8 (one-sided DeLong's test, compared to method t-test, $P = 1 \times 10^{-9}$; method methylKit, $P = 8 \times 10^{-9}$; method globalTest, $P = 9 \times 10^{-12}$). ROC ± 95% CI are shown. The same list of genes was used among all four methods. Bottom panel, number of true positives in top 500 candidate DNAme drivers identified by MethSig and benchmarked methods, as well as 500 randomly selected genes in CLL8 cohort. The black curve indicates the density of true positive number estimated by 10,000

times of random selection. Empirical p-value was calculated according to the probability that the number of genes found in top MethSig candidate genes is greater than the number of genes found in an equal number of randomly selected genes from all the inferred promoters. **B**, Different clustering of methylated CpGs between samples with or without hypermethylation in DNAme drivers. The median, upper and lower quartiles are shown. Whiskers represent upper quartile + 1.5 interquartile range (IQR) and lower quartile − 1.5 IQR. Clustered or scattered CpG hypermethylation are shown (black circles, methylated CpGs; white circles, unmethylated). Statistical analysis was performed by two-sided paired Mann-Whitney *U* test. **C**, Candidate DNAme drivers selected by the model are depicted in descending order of their association (coefficients) with poor FFS. Heatmap showing which selected DNAme drivers are hypermethylated in each patient in CLL8. Bar plot showing −log10 p-values in CLL8 cohort. **D**, Kaplan-Meier plot showing FFS in CLLs with high versus low risk in the training set (CLL8). **E**, Kaplan-Meier plot showing FFS in CLLs with high versus low risk in the independent test set (CLL-DFCI). In **C**-**D**, *alpha* equal to 0.1 was used in the elastic net regression. In **D**-**E**, statistical analysis was performed with log-rank test. **F**, Multivariable analyses for DNAme driver risk with the addition of well-established poor outcome predictors in CLL (IGHV unmutated status and *del*[17p] or *TP53* mutation status) in CLL8 and CLL-DFCI cohorts. **G**, Q-Q plot comparing observed −log10 p-values of MethSig to expected −log10 p-values. Result of relapsed (T2) patients in CLL8 cohort is presented. Bar plot represents the percentage of genes with p-values less than 0.05. Deviation from expected factor was the slope value derived from linear regression through zero, modeling the relationship between observed and expected −log10 p-values. Expected p-values were sampled from uniform distribution (*runif* function in R starting from identical random seed number). **H**, Cumulative distribution function plot of three subgroups of DNAme drivers in terms of log2 odds ratio of DNAme driver incidence in T2 over T1 samples. **I**, Left panel, heatmap of −log10 (BH-FDR Q) derived from DNAme driver identification of T2 or T1 samples over control samples. Right panel, the enrichment of three subgroup DNAme drivers in selected pathways. The dashed grey line indicates BH-FDR Q = 0.25, cutoff for significant enrichment of DNAme drivers in each pathway.

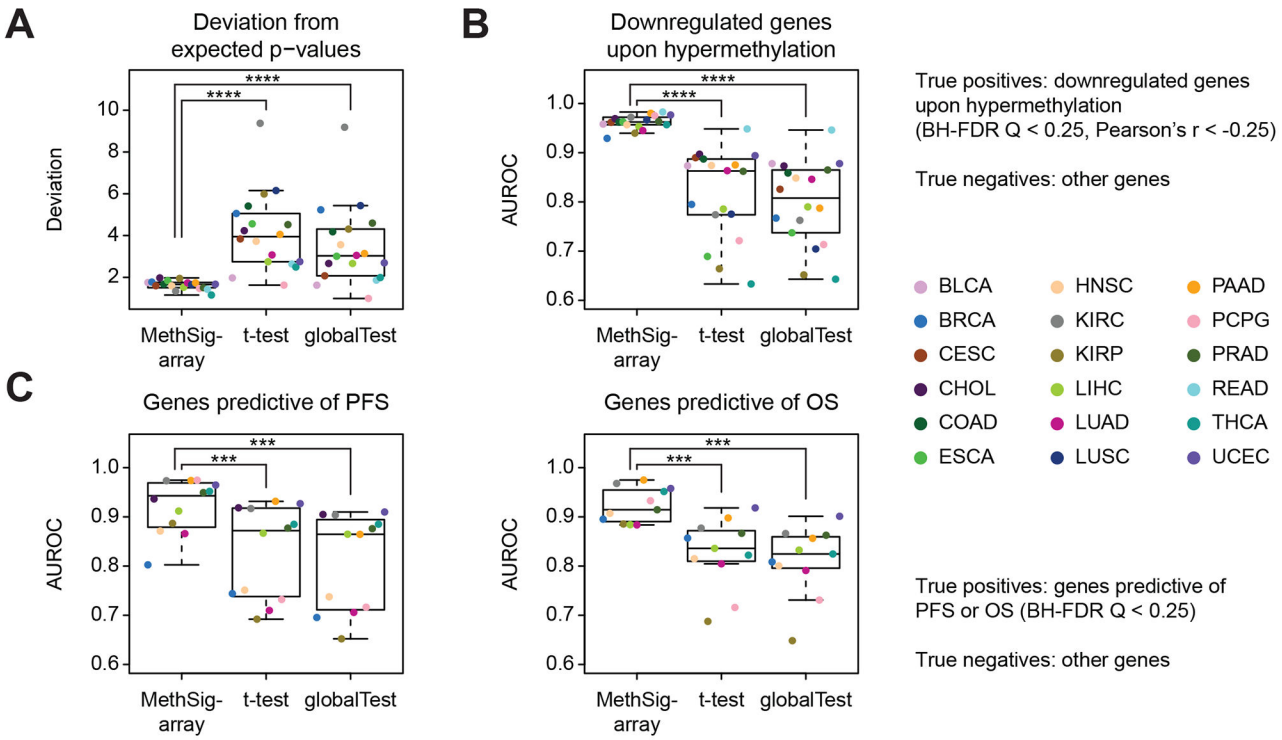


Figure 5. MethSig-array outperforms benchmarked methods in identifying candidate DNAme drivers.

A, Deviation from expected factors of MethSig-array, t-test and globalTest. Deviation from expected factor was the slope value derived from linear regression through zero, modeling the relationship between observed and expected −log10 p-values. **B**, AUROC of MethSig-array, t-test and globalTest in the inference of likely DNAme drivers associated with gene silencing. Matched tumoral DNAme array and RNA-seq data in TCGA Pan-Cancer analysis project (33) were used. **C**, AUROC of MethSig-array, t-test and globalTest in the inference of likely DNAme drivers associated with progression-free survival (PFS) or overall survival (OS). Clinical information obtained from cBioPortal (49,50). Statistical analysis was performed by two-sided paired Mann-Whitney *U* test: ****P < 0.0001; ***P < 0.001. The median, upper and lower quartiles are shown. Whiskers represent upper quartile + 1.5 IQR and lower quartile − 1.5 IQR. In **B-C,** only tumor types with a minimum of 10 likely DNAme drivers were included into the analysis. In each AUROC analysis, the same list of genes was used among all methods. One of the benchmarked methods – methylKit – could not be applied to TCGA dataset due to lacking count-based data.