Methylome-based cell-of-origin modeling (Methyl-COOM) identifies aberrant expression of immune regulatory molecules in CLL

Justyna A. Wierzbinska[1,2,3], Reka Toth[1], Naveed Ishaque[3], Karsten Rippe[3,4], Jan-Philipp Mallm[3,4], Lara Klett[3,4], Daniel Mertens[3,5], Thorsten Zenz[6], Thomas Hielscher[7], Marc Seifert[8], Ralf Küppers[8], Yassen Assenov[1], Pavlo Lutsik[1], Stephan Stilgenbauer[9], Philipp M. Roessner[10], Martina Seiffert[10], John Byrd[11], Christopher C. Oakes[11,12], Christoph Plass[1,3,#,§], Daniel B. Lipka[3,13,14,15,#,§]

Affiliations:

[1] Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2] Faculty of Biosciences, Heidelberg University, Heidelberg, Germany

[3] The German Cancer Consortium (DKTK)

[4] Division of Chromatin Networks, DKFZ, Heidelberg, Germany

[5] Mechanisms of Leukemogenesis, DKFZ, Heidelberg, Germany

[6] Experimental Hematology Lab, University Hospital Zurich, Switzerland

[7] Biostatistics, DKFZ, Heidelberg, Germany

[8] Group Molecular Genetics, Essen University Hospital, Essen, Germany

[9] Department of Internal Medicine, Ulm University, Ulm, Germany

[10] Division of Molecular Genetics, DKFZ, Heidelberg, Germany

[11] Department of Internal Medicine, Division of Hematology, The Ohio State University, Columbus, USA

24   [12] Department of Biomedical Informatics, The Ohio State University, Columbus, USA

25   [13] Translational Cancer Epigenomics, Division of Translational Medical Oncology, German

26   Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

27   [14] National Center for Tumor Diseases (NCT), 69120 Heidelberg, Germany

28   [15] Faculty of Medicine, Medical Center, Otto-von-Guericke-University, 39120 Magdeburg,

29   Germany

30

31   # Joint senior authors

32   § Corresponding authors

33

34   **Corresponding Authors:**

35   Christoph Plass

36   Division of Cancer Epigenomics

37   German Cancer Research Center (DKFZ)

38   Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

39   Tel. +49 6221 42 3330, Fax +49 6221 42 3359, E-mail: c.plass@dkfz.de

40

41   Daniel B. Lipka

42   Section Translational Cancer Epigenomics

43   Division of Translational Medical Oncology

44   German Cancer Research Center (DKFZ) & National Center for Tumor Diseases (NCT)

45   Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

46   Tel. +49 6221 42 1603 or +49 6221 56 38287, E-mail: d.lipka@dkfz.de

47

48

49   **Running Title**: Methylome-based cell-of-origin modeling identifies immune molecule

50   dysregulation in CLL

51 **ABSTRACT**

52 **Background:** In cancer, normal epigenetic patterns are disturbed and contribute to gene

53 expression changes, disease onset and progression. The cancer epigenome is composed of the

54 epigenetic patterns present in the tumor-initiating cell at the time of transformation, and the

55 tumor-specific epigenetic alterations that are acquired during tumor initiation and progression.

56 The precise dissection of these two components of the tumor epigenome will facilitate a better

57 understanding of the biological mechanisms underlying malignant transformation. Chronic

58 lymphocytic leukemia (CLL) originates from differentiating B cells, which undergo extensive

59 epigenetic programming. This poses the challenge to precisely determine the epigenomic

60 ground-state of the cell-of-origin in order to identify CLL-specific epigenetic aberrations.

61 **Methods:** We developed a linear regression model, methylome-based cell-of-origin modeling

62 (Methyl-COOM), to map the cell-of-origin for individual CLL patients based on the continuum of

63 epigenomic changes during normal B cell differentiation.

64 **Results:** Methyl-COOM accurately maps the cell-of-origin of CLL and identifies CLL-specific

65 aberrant DNA methylation events that are not confounded by physiologic epigenetic B cell

66 programming. Furthermore, Methyl-COOM unmasks abnormal action of transcription factors,

67 altered super-enhancer activities, and aberrant transcript expression in CLL. Among the

68 aberrantly regulated transcripts were many genes that have previously been implicated in T cell

69 biology. Flow cytometry analysis of these markers confirmed their aberrant expression on

70 malignant B cells at the protein level.

71 **Conclusions:** Methyl-COOM analysis of CLL identified disease-specific aberrant gene

72 regulation. The aberrantly expressed genes identified in this study might play a role in immune-

73 evasion in CLL and might serve as novel targets for immunotherapy approaches. In summary,

4

74   we propose a novel framework for *in silico* modeling of reference DNA methylomes and for the

75   identification of cancer-specific epigenetic changes, a concept that can be broadly applied to

76   other human malignancies.

77

78   **KEY WORDS**

79   DNA methylation; cell-of-origin; chronic lymphocytic leukemia; T cell antigens

BACKGROUND

In cancer, normal epigenetic patterns are disturbed and contribute to gene expression changes, disease onset and progression [1]. This seems to be a universal characteristic of all cancers, including chronic lymphocytic leukemia (CLL). CLL originates from rapidly differentiating B cells. Although several mutations creating a pre-leukemic clone, including variants in *SF3B1*, *NOTCH1* or *TP53*, have been identified in the hematopoietic stem cell (HSC) compartment of CLL patients, additional genetic or epigenetic driver events are required for full transformation[2]. Normal B cells undergo extensive epigenetic programming during differentiation [3,4]. The epigenetic fingerprint of the B cell that has acquired the transforming hit is 'frozen' and stably propagated in the leukemic cells [4]. This demonstrates that two factors contribute to the epigenomic landscape of CLL: first, epigenetic patterns that were present in the tumor-initiating B cell at the time of transformation, and second, CLL-specific epigenetic alterations that are acquired during leukemia initiation and progression. For the purpose of this study, we define the cell-of-origin of CLL as the normal B cell differentiation stage with the highest overlap to the CLL methylome. Consequently, the cell-of-origin of CLL represents the differentiation stage at which the clonal B cells deviate significantly from the normal differentiation trajectory and therefore the cell-of-origin defines the first cell that has acquired sufficient oncogenic hits to initiate leukemic transformation[5].

Numerous publications have reported extensive epigenetic alterations in CLL resulting in deregulation of protein coding genes [6–11] or miRNAs [12–19]. In this context, most studies used the epigenome of $CD19^+$ B cells as controls, but such an approach neglects the epigenetic programming occurring during B cell differentiation. As a result, the genes found to be deregulated mainly reflected the changes occurring during normal B cell differentiation rather than CLL-specific pathogenic events. Refined analyses should aim at discriminating between

104   epigenetic changes occurring during normal B cell differentiation and CLL-specific epigenetic

105   aberrations. Here we outline a novel framework for cancer methylome analysis, termed

106   methylome-based cell-of-origin modeling (Methyl-COOM). We show how Methyl-COOM can be

107   applied to epigenomic datasets from CLL patients to identify disease-specific epigenetic events

108   and demonstrate its power to detect epigenetically deregulated transcripts which encode for

109   proteins that are involved in immune regulatory processes.

110

METHODS

Flow cytometry analysis

Patients' samples were obtained from the Department of Internal Medicine III of Ulm University after approval of the study protocol by the local ethics committee according to the declaration of Helsinki, and after obtaining informed consent of patients. Patients met standard diagnosis criteria for CLL. Patients' characteristics such as age, gender, mutational state and Binet stage are depicted in Table 1.

Peripheral blood was drawn using Ethylenediaminetetraacetic acid (EDTA)-coated tubes (Sarstedt, Nümbrecht, Germany). PBMCs were isolated by Ficoll (Biochrom, Berlin, Germany) density gradient centrifugation. PBMCs were viably frozen and, when needed, thawed and further processed.

After blockade of Fc-receptors using Human TruStain FcX™ (BioLegend, London, United Kingdom), $5*10^6$ PBMCs were stained with fluorescently labelled antibodies in phosphate-buffered saline (PBS) with addition Fixable Viability Dye eFluor® (ThermoFisher Scientific, Dreieich, Germany) for 30 min at 4°C. Cells were fixed using eBioscience™ IC Fixation Buffer (ThermoFisher Scientific, Dreieich, Germany) for 30 min at room temperature. The antibodies used are listed in Table 2. If necessary, cells were permeabilized with eBioscience™ Permeabilization Buffer (ThermoFisher Scientific) and stained intracellularly for 30 min at room temperature. CTLA-4 was stained as surface as well as intracellular marker. Samples were stored at 4°C in the dark until acquisition. Data was acquired using a BD LSR Fortessa (BD Biosciences, Heidelberg, Germany) FACS analyzer. Flow cytometric data was analyzed using FlowJo X 10.0.7 software (FlowJo, Ashland, OR, USA). Paired Wilcoxon signed-rank test was used to determine statistical significance of changes between CLL B cells and normal B cells.

136 **Table 1:** Characteristics of the CLL patients used for flow cytometric analysis.

# of patients	7
Age [years]	57.1 (mean) 52 (median)
Sex	5/7 male 2/7 female
Prior therapies	7/7 no prior treatment
Binet stage	7/7 A 0/7 B 0/7 C
IGHV status	6/7 mutated 1/7 unmutated
Genetics (FISH)	1/7 Trisomy 12 5/7 del(13q) 1/7 no aberration
TP53 mutation status	4/7 WT 3/7 not tested

137

138

139 **Table 2:** List of FACS antibodies and reagents.

Reagent	Clone	Supplier	Cat #
APC anti-human CD5	UCHT2	BioLegend	300612
eBioscience™ CD152 (CTLA-4) PerCP-eFluor 710	14D3	Thermo Fisher Scientific	46-1529-42
eBioscience™ CD276 (B7-H3) PE-Cyanine7	7-517	Thermo Fisher Scientific	25-2769-41
eBioscience™ Fixable Viability Dye eFluor™ 506		Thermo Fisher Scientific	65-0866-14
eBioscience™ IC Fixation Buffer		Thermo Fisher Scientific	00-8222-49
eBioscience™ Permeabilization Buffer (10X)		Thermo Fisher Scientific	00-8333-56
eBioscience™ TIGIT PE-Cyanine7	MBSA43	Thermo Fisher Scientific	25-9500-42
Human TruStain FcX™ (Fc Receptor Blocking Solution)		BioLegend	422302
PE anti-human CD85k (ILT3, LILRB4) Antibody	ZM4.1	BioLegend	333007
PE/Dazzle™ 594 anti-human CD19 Antibody	HIB19	BioLegend	302252
PerCP/Cyanine5.5 anti-human CD2	RPA-2.10	BioLegend	300215

140

Analysis of RNA-seq / sncRNA-seq data

142 Expression data (RNA-Seq) from CLLs were obtained from our previous study [4]. RNA-Seq

143 data from normal B cells was obtained from International Cancer Genome Consortium (ICGC).

144 Reads per kilo base per million mapped reads (RPKM) normalized values were used for the

145 comparison of gene expression levels. sncRNA-seq data from CLLs was obtained from our

146 previous study [20]. Differential miRNA expression was assessed using normalized counts,

147 reads per million (RPM).

148

Analysis of 450k methylome array data

150 450K data from B cells was obtained from Oakes et al. [4]. CLL 450k data for the discovery and

151 validation cohorts were both obtained from previous studies [4,21]. The analysis of 450K data

152 was performed using RnBeads software [22]. Both datasets (normal B cells and CLLs) were

10

153 processed simultaneously. Briefly, raw 450K data for both CLL and healthy B cell sample sets

154 were normalized by the BMIQ method [23] without the background subtraction. The probes

155 overlapping SNPs and the X and Y chromosomes were removed and remaining probes

156 (n=464,743 CpGs) were considered for the downstream analysis, for the identification of CLL-

157 specific methylation events (Method Section: 'Identification of disease-specific methylation

158 events').

159

160 **Inference of the cell-of-origin and identification of disease-specific methylation events**

161 We studied the DNA methylation programming during normal B cell differentiation, using six

162 discrete B cell subpopulations including naïve to mature B cells: referred to as naïve B cells

163 (NBCs), germinal center founder cells (GCFs), low- and intermediate-memory B cells (loMBCs,

164 intMBC), splenic marginal zone B cells (sMGZs), and high maturity memory B cells (hiMBCs).

165 DNA methylomes from 2-4 donors per normal B cell subpopulation. In addition 34 CLL samples

166 were analyzed using Illumina 450k Bead Chip arrays.

167

168 **Cell-of-origin based methylome analysis, Methyl-COOM**

169 For analysis, we determined the DNA methylation dynamics during normal B cell differentiation

170 (differentiation axis). Here we assumed that changes in DNA methylation during the cellular

171 differentiation process are reminiscent of the DNA nucleotide changes over the evolutionary

172 time. CpG sites showing a statistically significant gain or loss of methylation of more than 20%

173 during B cell differentiation defined our set of so-called B cell-specific CpGs (n=74,333 CpGs;

174 student's t-test). A Manhattan distance matrix was calculated and used to build a methylation-

175 based phylogenetic tree of normal B cell differentiation by applying the minimum evolution

176 method (*fastme.bal* function, R package "ape"; Desper and Gascuel, [24]). Each node in the

177 phylogenetic tree corresponds to a certain differentiation stage reached by the B cell. Using this

178   approach, we observed a non-branched differentiation trajectory of normal B cell differentiation.

179   Therefore, we initially used all B cell-specific CpGs to generate a linear regression model of

180   DNA methylation programming during normal B cell differentiation. Linear behavior between the

181   differentiation stage of every B cell subset and the methylation profiles at B cell-specific CpGs

182   were tested at the single CpG level using F-test. The majority of the B cell-specific CpGs

183   (79.8%, n=59,326 CpGs) showed linear methylation dynamics across the six B cell

184   differentiation states. To exclude a potential bias on differentiation stage assignment, we re-

185   created both the phylogeny and the regression model of normal B cell differentiation, this time

186   using the linearly behaving B cell specific CpGs, only. The final regression model was designed

187   to infer DNA methylation levels of all CpGs included in our analysis.

188   Next, we mapped all CLL samples onto the normal B cell differentiation trajectory in order to

189   infer the closest virtual normal B cell methylome (cell-of-origin) defined as the position of the

190   closest normal B cell node in the phylogenetic tree. Then, we applied the linear regression

191   model to infer the DNA methylation levels for each CpG site in the putative cell-of-origin for

192   every patient, according to the formula:

193

194    $M = \alpha + \beta * d.s.$

195   , where

196   M denotes the calculated beta methylation value for a CpG site of cell-of-origin,

197   d.s. denotes the differentiation stage (defined as the distance between the NBC and the cell-of-

198   origin nodes as determined by the phylogenetic analysis),

199   β denotes the slope of the regression line,

200   α denotes the vertical (y-axis) intercept.

201

12

202   To test our cell-of-origin assignment, we applied a cross-validation model on our phylogenetic

203   analysis. The patient cohort was repeatedly divided into two subgroups; 70% and 30% (5000

204   repetitions). To minimize the likelihood of selecting the same sample multiple times, a random

205   sampling was allowed in the 70%-group, while sample replacement was restricted only to the

206   30%-group. Using this approach, we observed that our original cell-of-origin is located between

207   interquartile ranges of the cross-validation assignments, confirming the robustness of the cell-of-

208   origin definition (**Supplementary Figure S2 f**).

209

210   *Identification of CLL-specific DNA methylation*

211   Subsequently, the inferred DNA methylome of the cell-of-origin was used as a reference to

212   determine aberrantly methylated CpG sites in each sample. Disease-specific CpGs were

213   defined as sites with significant deviation from the expected methylation levels as compared to

214   the corresponding cell-of-origin.

215

216   *Sites with epigenetic B cell programming*

217   Sites undergoing epigenetic B cell programming (i.e. B cell-specific CpGs) could still show

218   disease-specific methylation events if their actual methylation status massively deviates from

219   what would be expected based on the regression model (sites with "epigenetic B cell

220   programming"). We used a conservative cut-off of more than 20% methylation loss (class A) or

221   gain (class B) relative to the calculated cell-of-origin methylation value (M value) in at least 75%

222   of the CLL patients.

223

224

13

225   *Sites without epigenetic B cell programming*

226   Sites with no epigenetic B cell programming (i.e. non-B cell-specific CpGs) were defined to have

227   CLL-specific aberrant DNA methylation if they displayed either methylation loss (class C) or gain

228   (class D) of more than 20% relative to the cell-of-origin in at least 75% of the CLL patients.

229

230   **Identification of CLL-specific protein-coding genes**

231   To identify CLL-specific protein-coding genes, disease-specific methylation events were

232   overlapped with promoter regions (-2.5kb, +0.5kb to TSS) of protein-coding genes. Next,

233   correlation between aberrant DNA methylation and gene expression was determined (Pearson

234   correlation test, p-value <0.05; correlation coefficient < -0.7). A full list of identified CLL-specific

235   protein-coding genes is available in **Supplementary Table S1**.

236

237   **Identification of CLL-specific SE-associated genes**

238   To identify CLL-specific Super-enhancer (SE)-associated genes, SE data from DKFZ PRECiSe

239   consortium was used [28]. All statistically significant, differential super-enhancers being gained

240   in CLLs ("gained", p<0.05, FC>0) and consensus super-enhancers shared between normal B

241   cells and CLLs ("stable") were used for the analysis. Firstly, SEs were associated with the

242   closest gene in the vicinity. CLL-specific methylation events were then overlapped with SE

243   coordinates. Next, correlation between aberrant DNA methylation in SE region and gene

244   expression of the SE-closest gene (Pearson correlation test, p-value <0.05; correlation

245   coefficient < -0.7) was used to identify CLL-specific Super-enhancer (SE)-associated genes. A

246   full list of identified SE-associated genes is available in **Supplementary Table S2**.

247

Super-enhancer (SE) enrichment analysis

For the super-enhancer enrichment analysis two sets of super-enhancers were used, SE data from DKFZ PRECiSe consortium [25] and SE data from Ott et al. [26]. From DKFZ PRECiSe consortium all statistically significant, differential super-enhancers being gained in CLLs ("gained", p<0.05, FC>0) and consensus super-enhancers shared between normal B cells and CLLs ("stable") were used for the analysis. From Ott et al. paper a unified SE region was created using reduce function in GenomicRanges package, providing a SE data from individual CLL patients (n=18). All CpG probes present on the 450k array were used as a background in the enrichment analysis.

Identification of micro-RNA promoters

To identify miRNA promoters, the promoter segmentation data from CLLs (DKFZ PRECiSe consortium; promoter segmentation data is deposited under GSE113336; raw ChIP-seq data can be found in the European Genome-phenome Archive under the accession number EGAS00001002518) and normal cell lines (Encyclopedia of DNA Elements – ENCODE; ENCODE Mar 2012 Freeze, UCSC accession numbers: wgEncodeEH000784, wgEncodeEH000785, wgEncodeEH000790, wgEncodeEH000789, wgEncodeEH000788, wgEncodeEH000786, wgEncodeEH000787, wgEncodeEH000791, wgEncodeEH000792) was used. To define constant promoter segments, the *reduce* function from the "GenomicRanges" R package was used to create simplified promoter regions, present in all datasets (CLL and ENCODE segmentation data). Putative promoters of pri-miRNAs were assigned based on their distance to the pri-miRNA TSSs. The genomic coordinates of pri-miRNAs/miRNAs were downloaded from miRBase (version 20; v20). Any promoter located within 100kb upstream of a pri-miRNA TSS was considered as a putative pri-miRNA promoter. The distance of 100 kb was chosen based on similar approaches that have been used in the past by Corcoran *et al.*, Fujita

15

273   *et al.* and Fukao *et al.* [27–29]. The larger distance of putative promoters to pri-miRNA TSSs is

274   especially important in case of intergenic miRNAs, which are originating from intronic

275   sequences and which are considered to be transcribed together with their host gene.

276

277   **Identification of CLL-specific micro-RNAs**

278   To identify CLL-specific microRNAs, disease-specific methylation events were overlapped with

279   potential pri-miRNA promoters. To identify candidate CLL-specific miRNAs, correlation between

280   aberrant DNA methylation and pri-miRNA expression was determined (Spearman correlation

281   test, p-value <0.05; abs(correlation coefficient ρ) ≥ 0.35). Since many mature miRNAs are

282   derived from the same pri-miRNAs, correlations were calculated using pri-miRNA expression

283   levels determined by sncRNA-seq. A full list of identified CLL-specific microRNAs is available in

284   **Supplementary Table S3**.

285

286   **Target genes of CLL-specific microRNAs**

287   To link CLL-specific microRNAs with their pathogenetic effects, two databases of experimentally

288   validated microRNA-target gene interactions were used, TarBase v8.0 and miRTarBase. A full

289   list of experimentally validated CLL-specific microRNA targets is included in **Supplementary**

290   **Table S4**. To find whether CLL-specific microRNAs are targeting epigenetic regulators, the

291   comprehensive list of epigenetic regulators was used (**Supplementary Table S5**). The list of

292   epigenetic regulators was further used as a query for the list of CLL-specific microRNA targets

293   defined above. The epigenetic regulators targeted by CLL-specific microRNAs are included in

294   the **Supplementary Table S6**.

295

296

16

297 **Transcription factor enrichment analysis**

298 Transcription factor motif analysis in disease-specific methylation events was performed using

299 HOMER software v4.5 [30] using only the results for the 'known motifs' analysis. All CpGs

300 present on the 450k array were used as a background and adjustment for GC- and CpG-content

301 was used. Furthermore, enrichment of actual binding events of TFs and other DNA-binding

302 proteins was analyzed using available ChIP-seq data from the tier 1 ENCODE cell line

303 GM12878 (for a complete list of datasets used for this analysis, please refer to **Supplementary**

304 **Table S7**). The ChIP-seq enrichment analysis was performed using the LOLA tool [31] providing

305 all CpG probes present on the 450k array as the 'universe'. Unsupervised hierarchical clustering

306 and data visualization were performed using R.

307

17

RESULTS

Modeling of normal B cell differentiation

CLL epigenomes are shaped by two major components. The first component constitutes

signatures that stem from the leukemia-initiating B cell. The second component is formed by

epigenetic alterations acquired during leukemogenesis and progression of the disease. To

discriminate these components, we developed an *in silico* approach to infer DNA methylation

dynamics during normal B cell differentiation and to model the epigenome of the cell-of-origin,

utilizing previously published Illumina 450k array DNA methylome data from six distinct B cell

subpopulations [4] and from 34 CLL samples [21](**Figure 1a**). Our approach to this was based

on classical phylogeny analysis (minimum evolution method, Desper and Gascuel [24]), which is

typically used to reconstruct evolutionary processes based on inherent characters. Similarly to

copy number or mutational studies [32,33], phylogeny analysis on DNA methylation has been

used successfully to reconstruct the developmental processes occuring during cell proliferation

and differentiation [4,34]. Therefore, to model B cell differentiation, we inferred the hierarchical

relationship between normal B cell subsets ranging from naïve to memory B cells based on their

DNA methylation patterns. The normal B cell methylomes were used to identify CpG sites that

show dynamic DNA methylation during B cell differentiation (B cell-specific CpGs; see also

Methods). A total of 74,333 B cell-specific CpGs were identified (≥ 20% DNA methylation

change between naïve and differentiated memory B cells, Student's t-test, p-value<0.05 [4,35]).

Pairwise Manhattan distances based on DNA methylation profiles at B cell-specific CpGs for

normal B cell subsets were used to build a methylation-based phylogenetic tree revealing a non-

branched trajectory of B cell differentiation (**Supplementary Figure S1a)**. This suggested that

linear regression might be suitable to model DNA methylation dynamics. The initial linear

regression model of B cell differentiation considered all B cell-specific CpGs. Testing the

18

332   linearity between the differentiation stage of every normal B cell subset and the methylation

333   profiles at B cell-specific CpGs, revealed that the vast majority of the differentiation-specific

334   CpGs (79.8%, n=59,326 CpGs) showed linear behavior across all B cell differentiation states (F-

335   test, p-value <0.05; **Supplementary Figure S1b-g, Supplementary Table S8**). To exclude a

336   potential bias on the model from the non-linear CpG sites, we re-generated both the phylogeny

337   and the regression model of normal B cell differentiation using only the linearly behaving B cell-

338   specific CpGs.

Identification of disease-specific DNA methylation patterns in CLL

340   This B cell differentiation model was applied to a CLL patient cohort (n= 34) in order to

341   determine the closest virtual normal B cell methylome (i.e. cell-of-origin or B cell differentiation

342   stage) for each CLL case (**Figure 1b**). As expected, our model confirmed that good-prognosis

343   *IGHV* mutated CLL originates from more mature B cells, as opposed to *IGHV* unmutated CLL,

344   which develops from more immature B cells (**Supplementary Figure S2a-e**). Next, we tested

345   the stability of cell-of-origin assignment using a cross-validation model (5000 repetitions; for

346   details see **Methods** section). Using this approach, we observed that the predicted cell-of-origin

347   is located between interquartile ranges of the cross-validation assignments, confirming the

348   robustness of the cell-of-origin definition (**Supplementary Figure S2f**). The linear regression

349   model was then used to infer DNA methylation levels for all 464,743 CpG sites in the predicted

350   cell-of-origin of every patient. These inferred cell-of-origin methylomes were subsequently used

351   as controls to identify aberrant (i.e. CLL-specific) DNA methylation patterns for each sample

352   individually (see **Figure 1a** for a schematic overview of Methyl-COOM). CLL-specific aberrant

353   DNA methylation was defined as CpG sites with >20% deviation from the expected DNA

354   methylation level of the cell-of-origin, and which were aberrantly methylated in at least 75% of

355   patients. This analysis revealed two categories of CLL-specific DNA methylation events; 1)

19

356   aberrant DNA methylation occurring at sites undergoing epigenetic programming during B cell

357   differentiation ('Sites with epigenetic B cell programming') and 2) aberrant DNA methylation

358   occurring at CpG sites that normally do not change during B cell differentiation ('Sites with no

359   epigenetic B cell programming') (see **Figure 1c**). The first category was further subdivided into

360   class A, showing a loss, and class B, showing a gain of DNA methylation relative to the

361   differentiation stage achieved. The second group of CpG sites without DNA methylation

362   programming during normal B cell differentiation was subdivided into class C and class D

363   displaying hypo- and hypermethylation, respectively (**Figure 1c**). Overall, only 2.2% of all CpG-

364   sites (10,335 CpGs) represented on the 450k array were affected by disease-specific

365   methylation programming, the majority of which were 'sites with epigenetic B cell programming'

366   (class A & B, 5,940 CpG sites; **Figure 1c, Supplementary Table S9**). The majority of CLL-

367   specific DNA methylation events were characterized by hypomethylation (9,995 hypomethylated

368   CLL-specific CpGs; class A: 5,757 CpGs, class C: 4,238 CpGs), while only a small proportion of

369   CpGs were hypermethylated as compared to their inferred cell-of-origin (340 hypermethylated

370   CLL-specific CpGs; class B: 183 CpGs, class D: 157 CpGs) (**Figure 1c, Supplementary**

371   **Figure S2g, h**).

372

373   **CLL-specific aberrant DNA methylation patterns are independent of the differentiation**

374   **stage achieved**

375   CLL-specific DNA methylation changes were quantified for each CpG site in each sample as

376   compared to the cell-of-origin and inspected by unsupervised hierarchical clustering. For all

377   classes, consistent patterns of either loss or gain in methylation relative to the cell-of-origin were

378   observed, irrespective of the differentiation stage achieved (**Figure 2a**, **Supplementary Figure**

379   **S2i**). Hypomethylation at class A sites resulted from an exaggerated loss of DNA methylation at

380   sites which show loss of methylation during normal B cell differentiation (**Figure 2b, c**,

381 **Supplementary Figure S2i**; class A, hypomethylation). Aberrant hypermethylation observed at

382 class B sites results from exaggeration of hypermethylation normally occurring during B cell

383 differentiation, and from failed hypomethylation during normal B cell programming (**Figure 2b, c**,

384 **Supplementary Figure S2i**; class B, hypermethylation). Class C and class D sites do not

385 undergo any significant DNA methylation programming during normal B cell differentiation,

386 highlighting the potential importance of these sites for CLL pathogenesis (**Figure 2a-c**

387 **Supplementary Figure S2i;** class C, class D). Overall, the observed CLL-specific aberrant

388 methylation patterns are largely independent of the differentiation stage achieved by the CLL

389 cell-of-origin.

390

391 **CLL-specific DNA methylation affects super-enhancers**

392 To test for functional implications of CLL-specific DNA methylation events, we tested their

393 enrichment in ENCODE ChromHMM genome segments in the GM12878 lymphoblastoid cell

394 line. Aberrantly methylated CpG sites from classes A, B & C were enriched for enhancer

395 elements (**Figure 3a**). A recent systematic assessment of transcription factor dependencies in

396 CLL has implicated super-enhancer (SE) based transcription factor (TF) rewiring in CLL

397 pathogenesis [26,36]. In line with this, enrichment of CLL-specific CpGs was detected in SE

398 regions identified in a recently published CLL data set from Ott et al. (**Supplementary Figure**

399 **S3a**) [26]. Using another SE data set from Rippe and colleagues [25,37] enabled us to

400 distinguish between SEs that are either present in normal B cells ("stable") or that have been

401 acquired *de novo* in CLL ("gained"). Enrichment of *de novo* SEs was found in class A and class

402 C sites (**Figure 3b**). *De novo* SEs overlapping with CLL-specific CpG sites harbor many known

403 genes with relevance in CLL biology (e.g. *CD5*, *CLLU1*, *IRF2*; **Supplementary Figure S3b,**

404 **Supplementary Table S2**).

405

CLL-specific DNA methylation differences result from aberrant transcription factor

programming

Recent SE perturbation studies implicated rewiring of TF regulatory circuitries in CLL pathogenesis [26]. These findings motivated us to ask whether CLL-specific DNA methylation patterns would be indicative of aberrant TF programming. To address this hypothesis, we used ATAC-seq to test whether CLL-specific DNA methylation patterns were reflected at the level of chromatin accessibility. Indeed, we found that CLL-specific hypo- and hypermethylation events were associated with inverse changes in chromatin accessibility in CLL as compared to normal B cells (**Figure 3c**). These concomitant changes in DNA methylation and chromatin accessibility indicated that CLL-specific DNA methylation patterns reflect global epigenomic changes and further demonstrated that disease-specific DNA methylation changes identify functionally relevant *cis*-regulatory sequences in CLL. In line with this, transcription factor (TF) binding sites enriched in class A (e.g. IKZF1, BATF, NFAT, EGR1/2) and in class C sequences (e.g. NFAT, EGR1/2, E2A) were predominantly associated with B cell biology, e.g. BATF controling the expression of activation-induced cytidine deaminase (AID) and of I_H-C_H germline transcripts or E2A controlling B cell lineage commitment. This suggested involvement of altered TF binding patterns in CLL pathogenesis: class A CpG sites are characterized by stronger than normal TF binding and class C sites are likely de novo bound by B cell specific TFs (**Figure 3d**, **e**). Class B sites were enriched in motifs for EBF, NKX6-1 and PAX5, but overall the motif enrichment as well as the associated changes in chromatin accessibility were only moderate (**Figure 3c-e**). Binding of proteins related to genome architecture (CTCF, RAD21, SMC3) was overrepresented in class D sites (**Figure 3d, e**). Aberrant DNA methylation patterns at TF binding sites in CLL might be associated with disturbed TF expression levels. TF expression analysis revealed transcriptional deregulation of MAFB, JUN, KLF14, KLF4, IRF2 and EBF1, none of which showed major changes in their promoter DNA methylation status (**Supplementary Figure S4a,**

431 **b**). Among the deregulated TFs, EBF1 showed the strongest and most consistent transcriptional

432 deregulation with almost complete loss of expression in CLL samples (log2-FC: -7.98 [CLL -

433 hiMBC]; **Supplementary Figure S4a**). The EBF1 downregulation potentially explains the

434 observed CLL-specific hypermethylation at class B sites, as EBF1 has been shown to possess

435 pioneering activity [38]. Similarly, upregulation of KLF4, JUN and IRF2 (**Supplementary Figure**

436 **S4a**) could explain hypomethylation programming observed at class A and C CpG sites as all of

437 these TFs have been reported to possess pioneering activity [39–41].

438

439 **Class D hypermethylation is associated with reduced CTCF binding and potentially**

440 **deregulates expression of neighboring genes**

441 The enrichment of CTCF binding sites and motifs as well as the enrichment of ChromHMM

442 insulator regions (**Figure 3a, d, e**) led us to investigate the effects of aberrant CTCF binding in

443 CLL in more detail. We found that class D sites had lower CTCF occupancy and reduced

444 chromatin accessibility in CLL samples as compared to normal B cells (**Figure 3f**) while

445 globally, these patterns were identical (**Supplementary Figure S5a, b**). The differences in

446 CTCF binding were associated with changes in gene expression of neighboring genes (**Figure**

447 **3g**). This further highlights the importance of aberrant CTCF binding at class D CpGs and might

448 point towards a novel pathogenetic mechanism in CLL. Unfortunately, the low absolute number

449 of class D sites does not allow a comprehensive analysis of associated gene expression

450 changes and further studies involving whole-genome bisulfite sequencing will be required to

451 systematically address this observation.

452

453 **Identification of epigenetically deregulated transcripts in CLL**

454 The promoter DNA methylation status is widely used as a marker for gene regulation and

455 significant correlation of promoter DNA methylation with gene expression has been

23

456 demonstrated before [12,42–44]. Previous studies in CLL identified many epigenetic events

457 potentially deregulating the expression of protein-coding genes and miRNAs. However, all of the

458 work published so far used CD19$^+$ B cells as controls to call aberrant DNA methylation

459 [6,9,11,45–53]. To stress the importance of using appropriate controls to delineate disease-

460 specific DNA methylation events, we compared our cell-of-origin model to the classical

461 approach using bulk CD19$^+$ B cells as a reference. We correlated DNA methylation levels of all

462 aberrant promoter CpGs with gene-expression. The classical approach resulted in a ~1.5-fold

463 overcalling of epigenetically deregulated protein-coding genes (**Supplementary Figure S6a**).

464 For miRNAs this difference was even more pronounced (about 5- to 7-fold; **Supplementary**

465 **Figure S6b**). Interestingly, previously identified differentially methylated promoters of *TCL1*,

466 *HOXA4*, *TWIST2* or *DAPK1* did not pass the stringent filtering criteria of our correlation analysis.

467 This suggested that applying Methyl-COOM results in the identification of a more relevant set of

468 epigenetically deregulated candidate genes.

469 Using the cell-of-origin model, corrleation between promoter DNA methylation and miRNA

470 expression levels identified 8 CLL-specific miRNAs (**Figure 4a, b**). Seven out of these miRNAs

471 have been demonstrated to regulate epigenetic key players, and, even more importantly, they

472 regulate genes that have been shown to be recurrently mutated in CLL, namely *ARID1A*,

473 *ASXL1*, *CHD2*, *SETD1A*, *SETD2* and *KMT2D*. Reasoning that miRNA binding to their target

474 genes results in gene expression changes, we compared expression levels between miRNAs

475 and their target genes in CLL and normal B cells. Indeed, concordant with the pattern of miRNA

476 promoter hypomethylation and subsequent upregulation of miRNA transcript levels, we found

477 that known target genes of CLL-specific miRNAs were significantly downregulated in CLL as

478 compared to normal B cells while non-target genes were unaffected (**Figure 4c**).

479

24

480   A similar correlation analysis on protein-coding genes revealed statistically significant

481   correlations between DNA methylation and gene expression for 491 (class A), 20 (class B), 390

482   (class C), and 20 (class D) genes. The majority of correlations observed were negative (i.e. a

483   decrease in DNA methylation was associated with an increase in gene expression and vice

484   versa; **Supplementary Figure S6c**), and, as expected, the negative correlation with gene

485   expression was most unambiguous for hypermethylation events (59% class A, 95% class B,

486   70% class C, 85% class D; **Figure 5a**, **Supplementary Figure S6d**). A detailed analysis of the

487   top correlating genes (Pearson correlation test, p-value <0.05; correlation coefficient < -0.7)

488   encompassing 102 transcripts demonstrated a tight link between CLL-specific aberrant DNA

489   methylation and the expression levels of the corresponding genes (**Figure 5b**; **Supplementary**

490   **Figure S6a**). Normal B cell differentiation-related epigenetic and transcriptional changes were

491   exaggerated in class A and B whereas the changes detected in class C and D were observed

492   exclusively in CLL. Aberrantly methylated CpGs of class A and C converged in promoters of

493   12/102 transcripts (*TIGIT, SH3D21, LAX1, LILRB4, CD5, NOD2, POLR3GL, IGFBP4, ZAP70,*

494   *KSR2, XXYLT1–AS2,* and *LAG3*), highlighting the potential functional relevance of the

495   associated genes in CLL pathogenesis. In order to validate our findings, we applied Methyl-

496   COOM to 107 CLL samples that have been published previously by Oakes and colleagues

497   (**Supplementary Figure S7a**; [4]). This analysis identified 11,059 CLL-specific CpGs, of which

498   8,440 (76%) overlapped with the 10,339 CpGs identified in our discovery cohort

499   (**Supplementary Figure S7b**). Furthermore, CLL-specific CpGs identified in our validation

500   cohort recapitulated 92/102 (90%) of the top correlating candidate genes found in the discovery

501   cohort (**Supplementary Figure S7c**.

502

503

504 **Epigenetically deregulated transcripts are enriched for T cell-related and immune-**
505 **modulating genes**

506 Some of the top correlating genes have already been implicated to play a role in CLL biology,

507 e.g. *ZAP70*, *CD5*, *LCK*, *LAG3* or *CLLU1* (**Supplementary Figure S8a, b**), while for others their

508 role in CLL pathogenesis is currently unknown. To gain insights into the potential functional role

509 of these epigenetically deregulated genes, we performed enrichment analysis of known

510 biological functions, interactions, or pathways. MSigDB and GO analysis revealed strong

511 enrichment of gene sets related to immune response, immune system processes, hematopoietic

512 stem cells, CLL, and NOTCH signaling (**Supplementary Figure S8a, b**). Ingenunity Pathway

513 Analysis (IPA) and Metascape analysis resulted in enrichment of T-lymphocyte-related

514 processes (*Metascape:* 'Reguation of T cell activation' , 'Reguation of T cell receptor signaling

515 pathway', 'T cell costimulation', 'T cell differentiation'; *IPA:* 'Cell Proliferation of T Lymphocytes',

516 'T cell homeostasis', 'Proliferation of lymphocytes' (**Supplementary Figure S8a**). These

517 findings are in line with recent reports demonstrating that CD8[+] T cells from patients with

518 chronic lymphocytic leukemia exhibit features of T cell exhaustion, i.e. lower proliferative and

519 cytotoxic capacity and increased expression of inhibitory receptors (e.g. CTLA-4, TIGIT, Lag3,

520 PD-1), suggesting both CLL and T cell specific changes leading to decreased ability to eliminate

521 malignant cells [54–57].

522

523 **Epigenetically deregulated transcripts show aberrant protein expression in CLL**

524 Cancer cells express immune regulatory molecules that might represent potential targets for

525 novel immunotherapies. These proteins modulate the activity of tumor-infiltrating immune cells

526 and mediate immune-escape of tumor cells. Among the epigenetically deregulated genes we

527 identified several with immune regulatory function. Therefore, we aimed to determine whether

528 these are also aberrantly expressed at the protein level in CLL cells. We selected 5 candidates

529 from the list of top correlated genes which are known to be involved in lymphocyte/T-lymphocyte

530 related processes (*TIGIT*, *CTLA-4*, *CD276*, *LILRB4*, and *CD2*; **Figure 6a**). Flow cytometry was

531 utilized for the differential analysis of protein expression in malignant (CD19$^+$CD5$^+$) and normal

532 (CD19$^+$CD5$^-$) B cells of 7 CLL patients' blood samples (gating strategy in **Supplementary**

533 **Figure S9a**). We found that CTLA-4, TIGIT, LILRB4 and CD276 showed statistically significant

534 increased expression in malignant B cells as compared to normal B cells (CTLA-4, p-val=0.047;

535 TIGIT, p-val=0.016; CD276, p-val=0.016; LILRB4, p-val=0.016 [Wilcoxon paired signed-rank

536 test]), while CD2 surface expression was not detectable neither in normal nor CLL B cells

537 (**Figure 6b**; **Supplementary Figure S9b**). Despite the fact that the functional relevance of some

538 of these aberrantly expressed proteins (TIGIT, CD276 or LILRB4) still remains to be elucidated

539 in the context of CLL, our observation is of particular interest for the development of new

540 therapeutic strategies in CLL. Options to interfere with the signaling of these receptors are

541 currently investigated as potential novel therapeutic strategies in several cancer entities.

542

543 **DISCUSSION**

544 Applying Methyl-COOM analysis to CLL cells, we identified a number of microRNAs and

545 protein-coding genes that are epigenetically deregulated and validated the CLL-specific

546 epigentic deregulation for the vast majority of target genes in an independent patient cohort.

547 These epigenetically deregulated transcripts are likely involved in the pathogenesis or

548 maintenance of CLL and are functionally enriched for immune system- and lymphocyte-related

549 processes. The expression levels of these transcripts are very low in normal B cells, which is in

550 stark contrast to the strong overexpression observed in CLL cells. These epigenetically

551 deregulated transcripts are further expressed and detectable on the surface of malignant B

552 cells. CLL patients are known to progressively develop an immunosuppressive state including

553 dysfunctional T cells [57] and our data suggest that CLL cells contribute to the

27

554 immunosuppressive microenvironment as well as T cell exhaustion by expressing immune

555 regulatory molecules. Immune dysregulation is known to worsen over the course of the disease,

556 e.g. effector T cells are increased in early-stage disease and show progressive accumulation

557 and exhaustion in the late-stage [57,58]. This, together with the fact that CLL frequently affects

558 older patients with co-morbidities, makes CLL an ideal candidate for the development of

559 effective immunotherapies. CD276, TIGIT and LILRB4 would be of particular interest, since to

560 our knowledge they were not yet considered as immunotherapeutic targets in CLL. TIGIT is a

561 recently identified inhibitory receptor expressed on T cells and natural killer (NK) cells. In T cells,

562 TIGIT expression inhibits cell proliferation, cytokine production, and T cell receptor signaling

563 [59]. In tumors, TIGIT is involved in mediating a T cell exhaustion phenotype, which is

564 manifested by poor effector function of T cells and, consequently, decreased ability to eliminate

565 tumor cells. In non-Hodgkin B cell lymphomas, PD1- and TIGIT-expressing intratumoral T cells

566 were shown to mark dysfunctional or exhausted effector T cells [60]. CLL patients with an

567 advanced disease stage display elevated numbers of TIGIT[+] CD4[+] T cells compared to low risk

568 patients [61]. In preclinical models of colorectal and breast carcinoma, TIGIT blockade was

569 shown to reverse the exhaustion phenotype of cytotoxic T cells and to inhibit tumor growth [62].

570 Another immune inhibitory receptor, LILRB4, was reported as tumor-associated antigen that is

571 highly expressed on monocytic AML cells [63,64]. It was also reported as a selective marker of

572 neoplastic B cells and HSCs from CLL patients [65]. LILRB4 targeting, either by antibodies or by

573 CAR-T cells, impeded AML development [55,56]. CD276 overexpression, on the other hand,

574 was linked to anti-apoptosis in colorectal cancer through activation of Jak2-STAT3 signaling

575 pathway, and as a result, increased expression of anti-apoptotic protein Bcl-2 [66]. High CD276

576 expression levels were already linked to poor prognosis in CLL, prostate and pancreatic cancer

577 [67–70]. Altogether, TIGIT, LILRB4 and CD276 represent attractive therapeutic targets for

578 treatment of CLL.

28

579   The present study demonstrates that Methyl-COOM delineates cancer-specific DNA methylation

580   patterns and identifies deregulated pathways involved in the pathogenesis or maintenance of

581   CLL. Our work serves as a proof-of-concept that tracing the cell-of-origin by comparison to

582   normal differentiation trajectories is of great conceptual importance in cancer epigenetics.

583   Identifying the cell-of-origin is not only crucial for the precise analysis of epigenetic data, but it is

584   also important for clinical translation. The cell-of-origin impacts on tumor biology, affects chemo-

585   and radiosensitivity and influences disease outcome. For instance, studies in a murine model of

586   MLL-rearranged AML have shown that the cell-of-origin can influence the phenotype and the

587   aggressiveness of the resulting leukemia [71]. Likewise, glioma subtypes vary in their response

588   to therapy and share molecular signatures with different normal neural lineages, suggesting a

589   difference in their cellular origin [72–76]. So far, the identification of a cancer's cellular origin is

590   based on genetic lineage-tracing experiments in mice, like the ones from Blanpain and

591   colleagues demonstrating the presence of distinct cells-of-origin for two types of skin cancer

592   [77]. In colorectal cancer the cell-of-origin has been studied intensively, pointing towards three

593   potential cell types as founder cells: intestinal stem cells [78–82], transit amplifying cells [78,83],

594   and differentiated villus cells [83]. In most instances, however, the precise cell-of-origin, in which

595   transformation occurs, remains undefined.

596   Methyl-COOM can, in principle, be applicable to any type of DNA methylation data as a source

597   of epigenetic information. In contrast to previous reports in CLL and other malignancies,

598   epigenetic pathomechanisms were investigated using an approach that systematically avoids

599   confounding factors introduced by epigenome dynamics occurring in the context of physiological

600   differentiation processes. It has been demonstrated that similar concepts apply to other

601   lymphatic neoplasms, e.g. T-ALL, DLBCL or MCL [84–87]. However, for other tumors, including

602   myeloid malignancies, the knowledge on the cell-of-origin is still scarce. Therefore, beyond the

29

603   field of CLL research this study could serve as a template for the analysis of epigenomic data in

604   other cancer entities.

605

606   **CONCLUSIONS**

607   Our work describes a new analytical framework, Methyl-COOM, to delineate cancer-specific

608   DNA methylation patterns, a concept that should, in principle, be applicable to all tumor entities.

609   Using Methyl-COOM, we interrogated DNA methylomes of CLL samples in the context of

610   normal B cell differentiation. This enabled us to unmask abnormal transcription factor and super

611   enhancer activities, as well as to identify aberrant transcript expression in CLL. Furthermore, we

612   were able to demonstrate that epigenetically deregulated transcripts are enriched in immune

613   regulatory molecules which are also expressed at the protein level in CLL cells, suggesting that

614   CLL cells contribute to immunosuppression and T cell exhaustion by upregulation of immune

615   regulatory molecules. This finding might serve as a starting point for the development of novel

616   therapeutic strategies to overcome immune evasion of CLL cells.

617

618   **DECLARATIONS**

619   **Ethics approval and consent to participate**

620   The study was conducted in accordance with the declaration of Helsinki and was approved by

621   the Ethics Committee Heidelberg (University of Heidelberg, Germany; S-206/2011; S-356/2013)

622   and by the Ethics Committee Ulm (Ulm University; 130/2002). Samples were taken after

623   patients gave their written informed consent.

624

625   **Availability of data and materials**

626   The datasets used and analysed in the current study were published previously as indicated in

627   Table 3. The Methyl-COOM framework is accessible via GitHub

628   (https://github.com/justannwska/Methyl-COOM)[88].

629

630   **Table 3:** List of datasets used in the manuscript.

Dataset	Source
Illumina 450 data, normal B cells	Oakes et al. [4]
Illumina 450 data, CLLs discovery cohort	Dietrich et al. [21]
Illumina 450 data, CLLs validation cohort	Oakes et al. [4]
RNAseq, CLLs	Dietrich et al. [21]
RNAseq, normal B cells	International Cancer Genome Consortium (ICGC); EGAD00001000258
sncRNAseq, CLL	Blume et al. [20]
ENCODE TF ChIP-seq GM12878	ENCODE project [89]
ATAC-seq, normal B cells and CLLs	DKFZ PRECiSE consortium [25]

ChIP-seq, normal B cells and CLLs	DKFZ PRECiSE consortium [25]; EGAS00001002518
Promoter segmentation data CLL	DKFZ PRECiSE consortium [25]; GSE113336
ChromHMM GM12878 data	ENCODE ENCSR212BHV [89]

631

632 **URLs**

633 Bioconductor http://bioconductor.org/ [90]

634 Human genome (hg19, GRCh37) http://genome.ucsc.edu/downloads.html

635 LOLA https://bioconductor.org/packages/release/bioc/html/LOLA.html [31]

636 ENCODE https://www.encodeproject.org/ [89]

637 HOMER http://homer.ucsd.edu/homer/ [30]

638 miRTarBase: http://mirtarbase.mbc.nctu.edu.tw/php/index.php [91]

639 TarBase v8.0 http://carolina.imis.athena-

640 innovation.gr/diana_tools/web/index.php?r=tarbasev8%2Findex [92]

641 microRNA.org http://www.microrna.org [93]

642 miRBase v.18.0 http://www.mirbase.org [94]

643

644 **Competing interests**

645 The authors declare that they have no competing interests.

646

647

Author contributions

J.A.W, C.P. and D.B.L. developed the research concept, designed the analysis workflow and

experiments, and collected and interpreted the data. J.A.W., R.T., N.I., T.H., Y.A. and P.L.

analyzed data. J.A.W. performed experiments. K.R., J.M., L.K., D.M., T.Z., Marc.S., R.K., S.S.,

J.B. and C.C.O. provided clinical samples or data. P.M.R. and Mart.S. performed flow cytometry

experiments and analyzed data. J.A.W, C.P. and D.B.L. prepared the figures and wrote the

manuscript. C.P. and D.B.L. jointly supervised the project. All authors contributed to the writing

of the manuscript and approved the final version.

Corresponding authors

Correspondence to Christoph Plass (c.plass@dkfz.de) or to Daniel B. Lipka (d.lipka@dkfz.de)

33

REFERENCES

1. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer. 2011;11:726–34.

2. Damm F, Mylonas E, Cosson A, Yoshida K, Della Valle V, Mouly E, et al. Acquired initiating mutations in early hematopoietic cells of CLL patients. Cancer Discov. 2014;4:1088–101.

3. Kulis M, Merkel A, Heath S, Queiros AC, Schuyler RP, Castellano G, et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nat Genet. 2015;47:746–56.

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 Genet. 2016;48:253–64.

5. Visvader JE. Cells of origin in cancer. Nature. 2011;469:314–22.

6. Raval A, Tanner SM, Byrd JC, Angerman EB, Perko JD, Chen SS, et al. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. Cell. 2007;129:879–90.

7. Raval A, Byrd JC, Plass C. Epigenetics in chronic lymphocytic leukemia. Semin Oncol. 2006;33:157–66.

8. Claus R, Lucas DM, Ruppert AS, Williams KE, Weng D, Patterson K, et al. Validation of ZAP-70 methylation and its relative significance in predicting outcome in chronic lymphocytic leukemia. Blood. 2014;124:42–8.

9. Claus R, Lucas DM, Stilgenbauer S, Ruppert AS, Yu L, Zucknick M, et al. Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia. J Clin Oncol. 2012;30:2483–91.

10. Rush LJ, Raval A, Funchain P, Johnson AJ, Smith L, Lucas DM, et al. Epigenetic profiling in chronic lymphocytic leukemia reveals novel methylation targets. Cancer Res. 2004;64:2424–33.

11. Corcoran M, Parker A, Orchard J, Davis Z, Wirtz M, Schmitz OJ, et al. ZAP-70 methylation status is associated with ZAP-70 expression status in chronic lymphocytic leukemia. Haematologica. 2005;90:1078–88.

12. Baer C, Claus R, Frenzel LP, Zucknick M, Park YJ, Gu L, et al. Extensive promoter DNA hypermethylation and hypomethylation is associated with aberrant microRNA expression in chronic lymphocytic leukemia. Cancer Res. 2012;72:3775–85.

13. Pallasch CP, Patz M, Park YJ, Hagist S, Eggle D, Claus R, et al. miRNA deregulation by epigenetic silencing disrupts suppression of the oncogene PLAG1 in chronic lymphocytic leukemia. Blood. 2009;114:3255–64.

14. Wang LQ, Kwong YL, Kho CS, Wong KF, Wong KY, Ferracin M, et al. Epigenetic inactivation of miR-9 family microRNAs in chronic lymphocytic leukemia--implications on constitutive activation of NFkappaB pathway. Mol Cancer. 2013;12:173.

707   15. Wong KY, Yim RL, Kwong YL, Leung CY, Hui PK, Cheung F, et al. Epigenetic inactivation of the MIR129-
708   2 in hematological malignancies. J Hematol Oncol. 2013;6:16.

709   16. Wang LQ, Kwong YL, Wong KF, Kho CS, Jin DY, Tse E, et al. Epigenetic inactivation of mir-34b/c in
710   addition to mir-34a and DAPK1 in chronic lymphocytic leukemia. J Transl Med. 2014;12:52.

711   17. Deneberg S, Kanduri M, Ali D, Bengtzen S, Karimi M, Qu Y, et al. microRNA-34b/c on chromosome
712   11q23 is aberrantly methylated in chronic lymphocytic leukemia. Epigenetics. 2014;9:910–7.

713   18. Baer C, Oakes CC, Ruppert AS, Claus R, Kim-Wanner SZ, Mertens D, et al. Epigenetic silencing of miR-
714   708 enhances NF-kappaB signaling in chronic lymphocytic leukemia. Int J Cancer. 2015;137:1352–61.

715   19. Wang LQ, Wong KY, Rosen A, Chim CS. Epigenetic silencing of tumor suppressor miR-3151
716   contributes to Chinese chronic lymphocytic leukemia by constitutive activation of MADD/ERK and
717   PIK3R2/AKT signaling pathways. Oncotarget. 2015;6:44422–36.

718   20. Blume CJ, Hotz-Wagenblatt A, Hullein J, Sellner L, Jethwa A, Stolz T, et al. p53-dependent non-coding
719   RNA networks in chronic lymphocytic leukemia. Leukemia. 2015;29:2015–23.

720   21. Dietrich S, Oles M, Lu J, Sellner L, Anders S, Velten B, et al. Drug-perturbation-based stratification of
721   blood cancer. J Clin Invest. 2018;128:427–45.

722   22. Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA
723   methylation data with RnBeads. Nat Methods. 2014;11:1138–40.

724   23. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture
725   quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA
726   methylation data. Bioinformatics. 2013;29:189–96.

727   24. Desper R, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-
728   evolution principle. J Comput Biol. 2002;9:687–705.

729   25. DKFZ PRECiSE consortium. DKFZ PRECiSE consortium data resources. 2018.

730   26. Ott CJ, Federation AJ, Schwartz LS, Kasar S, Klitgaard JL, Lenci R, et al. Enhancer Architecture and
731   Essential Core Regulatory Circuitry of Chronic Lymphocytic Leukemia. Cancer Cell. 2018;34:982-995.e7.

732   27. Fujita S, Iba H. Putative promoter regions of miRNA genes involved in evolutionarily conserved
733   regulatory systems among vertebrates. Bioinformatics. 2008;24:303–8.

734   28. Corcoran DL, Pandit KV, Gordon B, Bhattacharjee A, Kaminski N, Benos PV. Features of mammalian
735   microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. PLoS One.
736   2009;4:e5279.

737   29. Fukao T, Fukuda Y, Kiga K, Sharif J, Hino K, Enomoto Y, et al. An evolutionarily conserved mechanism
738   for microRNA-223 expression revealed by microRNA gene profiling. Cell. 2007;129:617–31.

739 30. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-
740 Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell
741 Identities. Molecular Cell. 2010;38:576–89.

742 31. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R
743 and Bioconductor. Bioinformatics. 2016;32:587–9.

744 32. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by
745 single-cell sequencing. Nature. 2011;472:90.

746 33. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-
747 nucleotide mutation characteristics of a kidney tumor. Cell. 2012;148:886–95.

748 34. Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C, et al. Intratumor DNA methylation
749 heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell Rep. 2014;8:798–806.

750 35. Lipka DB, Witte T, Toth R, Yang J, Wiesenfarth M, Nollke P, et al. RAS-pathway mutation patterns
751 define epigenetic subclasses in juvenile myelomonocytic leukemia. Nat Commun. 2017;8:2126.

752 36. Lipka DB, Lutsik P, Plass C. From Basic Knowledge to Effective Therapies. Cancer Cell. 2018;34:871–3.

753 37. Mallm J-P, Iskar M, Ishaque N, Klett LC, Kugler SJ, Muino JM, et al. Linking aberrant chromatin
754 features in chronic lymphocytic leukemia to transcription factor networks. Molecular Systems Biology.
755 2019;15:e8339.

756 38. Boller S, Ramamoorthy S, Akbas D, Nechanitzky R, Burger L, Murr R, et al. Pioneering Activity of the
757 C-Terminal Domain of EBF1 Shapes the Chromatin Landscape for B Cell Programming. Immunity.
758 2016;44:527–41.

759 39. Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, et al. Transcription factor AP1
760 potentiates chromatin accessibility and glucocorticoid receptor binding. Molecular cell. 2011;43:145–55.

761 40. Ren G, Cui K, Zhang Z, Zhao K. Division of labor between IRF1 and IRF2 in regulating different stages
762 of transcriptional activation in cellular antiviral activities. Cell Biosci. 2015;5:17.

763 41. Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS. Pioneer transcription factors
764 target partial DNA motifs on nucleosomes to initiate reprogramming. Cell. 2015;161:555–68.

765 42. Bock C, Beerman I, Lien WH, Smith ZD, Gu H, Boyle P, et al. DNA methylation dynamics during in vivo
766 differentiation of blood and skin stem cells. Mol Cell. 2012;47:633–47.

767 43. Cabezas-Wallscheid N, Klimmeck D, Hansson J, Lipka DB, Reyes A, Wang Q, et al. Identification of
768 regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and
769 DNA methylome analysis. Cell Stem Cell. 2014;15:507–22.

770 44. Lipka DB, Wang Q, Cabezas-Wallscheid N, Klimmeck D, Weichenhan D, Herrmann C, et al.
771 Identification of DNA methylation changes at *cis* -regulatory elements during early steps of HSC
772 differentiation using tagmentation-based whole genome bisulfite sequencing. Cell Cycle. 2014;13:3476–
773 87.

774   45. Bichi R, Shinton SA, Martin ES, Koval A, Calin GA, Cesari R, et al. Human chronic lymphocytic
775   leukemia modeled in mouse by targeted TCL1 expression. Proc Natl Acad Sci U S A. 2002;99:6955–60.

776   46. Raval A, Lucas DM, Matkovic JJ, Bennett KL, Liyanarachchi S, Young DC, et al. TWIST2 demonstrates
777   differential methylation in immunoglobulin variable heavy chain mutated and unmutated chronic
778   lymphocytic leukemia. J Clin Oncol. 2005;23:3877–85.

779   47. Yuille MR, Condie A, Stone EM, Wilsher J, Bradshaw PS, Brooks L, et al. TCL1 is activated by
780   chromosomal rearrangement or by hypomethylation. Genes Chromosomes Cancer. 2001;30:336–41.

781   48. Cahill N, Rosenquist R. Uncovering the DNA methylome in chronic lymphocytic leukemia.
782   Epigenetics. 2013;8:138–48.

783   49. Melki JR, Vincent PC, Brown RD, Clark SJ. Hypermethylation of E-cadherin in leukemia. Blood.
784   2000;95:3208–13.

785   50. Bechter OE, Eisterer W, Dlaska M, Kuhr T, Thaler J. CpG island methylation of the hTERT promoter is
786   associated with lower telomerase activity in B-cell lymphocytic leukemia. Exp Hematol. 2002;30:26–33.

787   51. Chantepie SP, Vaur D, Grunau C, Salaun V, Briand M, Parienti JJ, et al. ZAP-70 intron1 DNA
788   methylation status: determination by pyrosequencing in B chronic lymphocytic leukemia. Leuk Res.
789   2010;34:800–8.

790   52. Wiestner A, Rosenwald A, Barry TS, Wright G, Davis RE, Henrickson SE, et al. ZAP-70 expression
791   identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior
792   clinical outcome, and distinct gene expression profile. Blood. 2003;101:4944–51.

793   53. Strathdee G, Sim A, Parker A, Oscier D, Brown R. Promoter hypermethylation silences expression of
794   the HoxA4 gene and correlates with IgVh mutational status in CLL. Leukemia. 2006;20:1326–9.

795   54. Zenz T. Exhausting T cells in CLL. Blood. 2013;121:1485–6.

796   55. Hanna BS, Roessner PM, Scheffold A, Jebaraj BMC, Demerdash Y, Ozturk S, et al. PI3Kdelta inhibition
797   modulates regulatory and effector T-cell differentiation and function in chronic lymphocytic leukemia.
798   Leukemia. 2019;33:1427–38.

799   56. Lewinsky H, Barak AF, Huber V, Kramer MP, Radomir L, Sever L, et al. CD84 regulates PD-1/PD-L1
800   expression and function in chronic lymphocytic leukemia. J Clin Invest. 2018;128:5465–78.

801   57. Hanna BS, Roessner PM, Yazdanparast H, Colomer D, Campo E, Kugler S, et al. Control of chronic
802   lymphocytic leukemia development by clonally-expanded CD8(+) T-cells that undergo functional
803   exhaustion in secondary lymphoid tissues. Leukemia. 2019;33:625–37.

804   58. Forconi F, Moss P. Perturbation of the normal immune system in patients with CLL. Blood.
805   2015;126:573–81.

806   59. Joller N, Kuchroo VK. Tim-3, Lag-3, and TIGIT. Curr Top Microbiol Immunol. 2017;410:127–56.

807   60. Josefsson SE, Beiske K, Blaker YN, Forsund MS, Holte H, Ostenstad B, et al. TIGIT and PD-1 Mark
808   Intratumoral T Cells with Reduced Effector Function in B-cell Non-Hodgkin Lymphoma. Cancer Immunol
809   Res. 2019;7:355–62.

810   61. Catakovic K, Gassner FJ, Ratswohl C, Zaborsky N, Rebhandl S, Schubert M, et al. TIGIT expressing
811   CD4+T cells represent a tumor-supportive T cell subset in chronic lymphocytic leukemia.
812   Oncoimmunology. 2017;7:e1371399.

813   62. Johnston RJ, Comps-Agrar L, Hackney J, Yu X, Huseni M, Yang Y, et al. The immunoreceptor TIGIT
814   regulates antitumor and antiviral CD8(+) T cell effector function. Cancer Cell. 2014;26:923–37.

815   63. Deng M, Gui X, Kim J, Xie L, Chen W, Li Z, et al. LILRB4 signalling in leukaemia cells mediates T cell
816   suppression and tumour infiltration. Nature. 2018;562:605–9.

817   64. John S, Chen H, Deng M, Gui X, Wu G, Chen W, et al. A Novel Anti-LILRB4 CAR-T Cell for the
818   Treatment of Monocytic AML. Mol Ther. 2018;26:2487–95.

819   65. Zurli V, Wimmer G, Cattaneo F, Candi V, Cencini E, Gozzetti A, et al. Ectopic ILT3 controls BCR-
820   dependent activation of Akt in B-cell chronic lymphocytic leukemia. Blood. 2017;130:2006–17.

821   66. Zhang T, Jiang B, Zou ST, Liu F, Hua D. Overexpression of B7-H3 augments anti-apoptosis of
822   colorectal cancer cells by Jak2-STAT3. World J Gastroenterol. 2015;21:1804–13.

823   67. Ramsay AG, Clear AJ, Fatah R, Gribben JG. Multiple inhibitory ligands induce impaired T-cell
824   immunologic synapse function in chronic lymphocytic leukemia that can be blocked with lenalidomide:
825   establishing a reversible immune evasion mechanism in human cancer. Blood. 2012;120:1412–21.

826   68. Inamura K, Takazawa Y, Inoue Y, Yokouchi Y, Kobayashi M, Saiura A, et al. Tumor B7-H3 (CD276)
827   Expression and Survival in Pancreatic Cancer. J Clin Med. 2018;7.

828   69. Roth TJ, Sheinin Y, Lohse CM, Kuntz SM, Frigola X, Inman BA, et al. B7-H3 ligand expression by
829   prostate cancer: a novel marker of prognosis and potential target for therapy. Cancer Res.
830   2007;67:7893–900.

831   70. Wang L, Kang FB, Shan BE. B7-H3-mediated tumor immunology: Friend or foe? Int J Cancer.
832   2014;134:2764–71.

833   71. Krivtsov AV, Figueroa ME, Sinha AU, Stubbs MC, Feng Z, Valk PJ, et al. Cell of origin determines
834   clinically relevant subtypes of MLL-rearranged AML. Leukemia. 2013;27:852–60.

835   72. Alcantara Llaguno S, Chen J, Kwon C-H, Jackson EL, Li Y, Burns DK, et al. Malignant Astrocytomas
836   Originate from Neural Stem/Progenitor Cells in a Somatic Tumor Suppressor Mouse Model. Cancer Cell.
837   2009;15:45–56.

838   73. Lai A, Kharbanda S, Pope WB, Tran A, Solis OE, Peale F, et al. Evidence for sequenced molecular
839   evolution of IDH1 mutant glioblastoma from a distinct cell of origin. J Clin Oncol. 2011;29:4482–90.

840 74. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis
841 identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1,
842 EGFR, and NF1. Cancer Cell. 2010;17:98–110.

843 75. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based
844 classification of central nervous system tumours. Nature. 2018;555:469–74.

845 76. Sturm D, Witt H, Hovestadt V, Khuong-Quang DA, Jones DT, Konermann C, et al. Hotspot mutations
846 in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer Cell.
847 2012;22:425–37.

848 77. Blanpain C. Tracing the cellular origin of cancer. Nat Cell Biol. 2013;15:126–34.

849 78. Barker N, Ridgway RA, van Es JH, van de Wetering M, Begthel H, van den Born M, et al. Crypt stem
850 cells as the cells-of-origin of intestinal cancer. Nature. 2009;457:608–11.

851 79. Zhu L, Gibson P, Currle DS, Tong Y, Richardson RJ, Bayazitov IT, et al. Prominin 1 marks intestinal
852 stem cells that are susceptible to neoplastic transformation. Nature. 2009;457:603–7.

853 80. Sangiorgi E, Capecchi MR. Bmi1 is expressed in vivo in intestinal stem cells. Nat Genet. 2008;40:915–
854 20.

855 81. Westphalen CB, Asfaha S, Hayakawa Y, Takemoto Y, Lukin DJ, Nuber AH, et al. Long-lived intestinal
856 tuft cells serve as colon cancer-initiating cells. J Clin Invest. 2014;124:1283–95.

857 82. Powell AE, Wang Y, Li Y, Poulin EJ, Means AL, Washington MK, et al. The pan-ErbB negative regulator
858 Lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. Cell. 2012;149:146–58.

859 83. Schwitalla S, Fingerle AA, Cammareri P, Nebelsiek T, Goktuna SI, Ziegler PK, et al. Intestinal
860 tumorigenesis initiated by dedifferentiation and acquisition of stem-cell-like properties. Cell.
861 2013;152:25–38.

862 84. Queiros AC, Beekman R, Vilarrasa-Blasi R, Duran-Ferrer M, Clot G, Merkel A, et al. Decoding the DNA
863 Methylome of Mantle Cell Lymphoma in the Light of the Entire B Cell Lineage. Cancer Cell. 2016;30:806–
864 21.

865 85. Shaknovich R, Geng H, Johnson NA, Tsikitas L, Cerchietti L, Greally JM, et al. DNA methylation
866 signatures define molecular subtypes of diffuse large B-cell lymphoma. Blood. 2010;116:e81–9.

867 86. Liu Y, Easton J, Shao Y, Maciaszek J, Wang Z, Wilkinson MR, et al. The genomic landscape of pediatric
868 and young adult T-lineage acute lymphoblastic leukemia. Nat Genet. 2017;49:1211–8.

869 87. Soulier J, Clappier E, Cayuela JM, Regnault A, Garcia-Peydro M, Dombret H, et al. HOXA genes are
870 included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). Blood.
871 2005;106:274–86.

872 88. Wierzbisnka JA. Methyl-COOM Framework [Internet]. Available from:
873 https://github.com/justannwska/Methyl-COOM

874   89. ENCODE project ChIP-seq data matrix. ENCODE project ChIP-seq data.

875   90. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-
876   throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21.

877   91. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, et al. miRTarBase: a database curates
878   experimentally validated microRNA-target interactions. Nucleic Acids Res. 2011;39:D163-169.

879   92. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al.
880   DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions.
881   Nucleic Acids Research. 2018;46:D239–45.

882   93. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and
883   expression. Nucleic Acids Res. 2008;36:D149-153.

884   94. Griffiths-Jones S. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids
885   Research. 2006;34:D140–4.

886

887

Figure 1
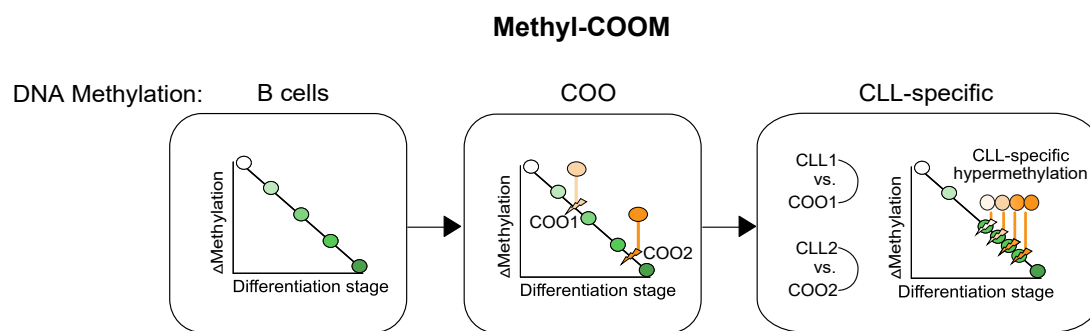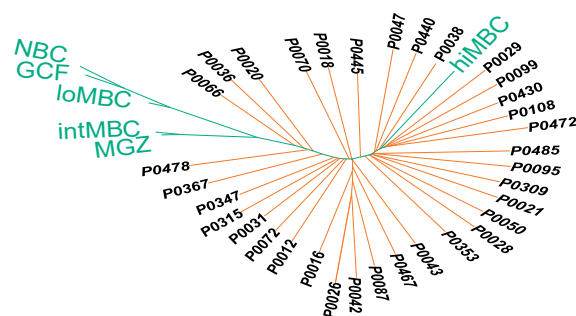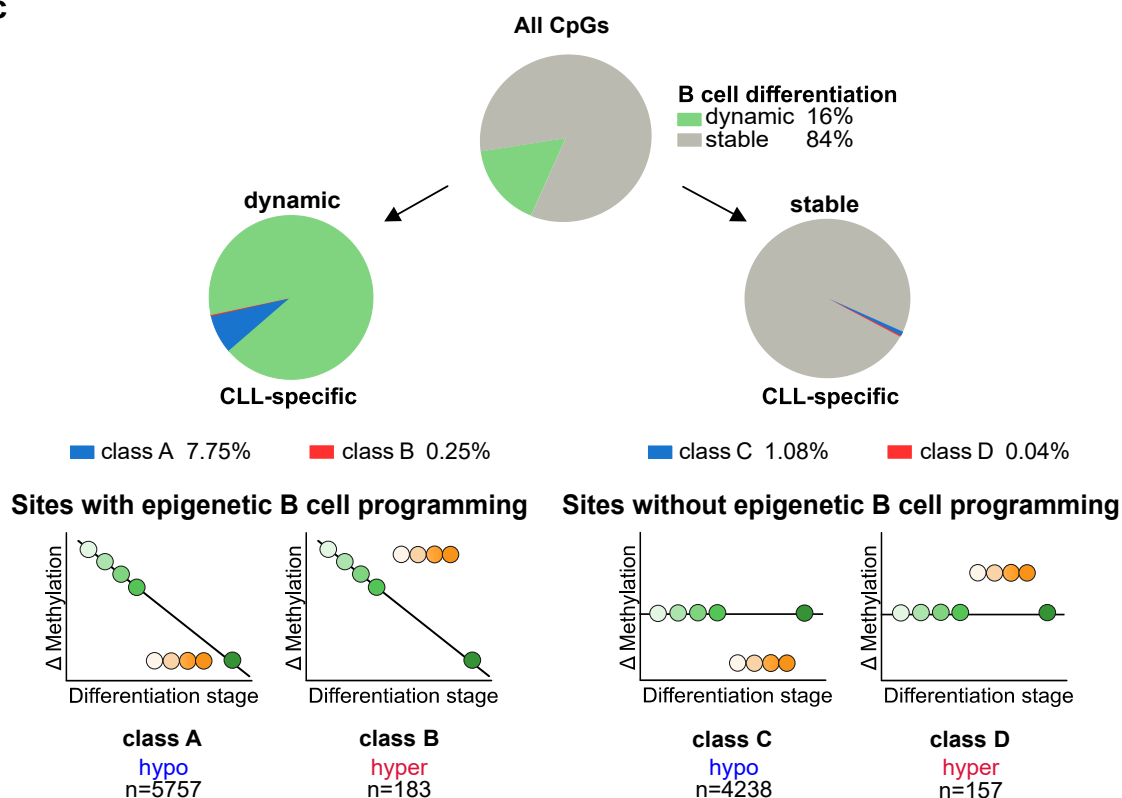
Figure Legends

Figure 1. Identification of CLL-specific DNA methylation events using Methyl-COOM.

a) Schematic outline of the Methyl-COOM pipeline used for the identification of CLL-specific DNA methylation events. Methylome data of six distinct B cell subpopulations, representing different stages of B cell differentiation were used to infer normal B cell differentiation. A linear regression model was applied to model DNA methylation dynamics during normal B cell differentiation ('DNA methylation: B cells'). DNA methylomes of 34 primary CLL samples were used to identify the closest virtual normal B cell (cell-of-origin; COO) based on phylogeny analysis. The linear regression model was then used to infer the DNA methylome of the COO ('DNA methylation: COO'). Next, the DNA methylome of each CLL was compared to the DNA methylome of its COO. CLL-specific aberrant DNA methylation was defined as a significant deviation from the inferred COO methylome ('DNA methylation: CLL-specific').

b) Identification of the cell-of-origin in CLL samples using phylogenetic analysis. A phylogenetic tree was generated using a set of linear CpG sites that show dynamic DNA methylation changes during normal B cell differentiation (linear B cell-specific CpGs, 59,326 CpGs). Pairwise Manhattan distances were calculated between DNA methylation profiles of normal B cells and CLL samples at B cell-specific CpGs and were subsequently used to assign the closest normal (virtual) B cell methylome (location of the node on the phylogenetic tree = differentiation stage of the cell-of-origin) to each CLL case. NBCs - naïve B cells; GCFs – germinal center founder B cells; loMBCs – early non class-switched memory B cells; intMBCs – non class-switched memory B cells; sMGZs – splenic marginal zone B cells; hiMBCs – class-switched memory B cells (mature B cells). CLL samples are depicted in orange color. Normal B cells are represented in green.

41

912   **c)** Summary of CLL-specific DNA methylation events. Top: pie chart displays the frequency of

913   CpGs that are either dynamic (green) or stable (grey) during normal B cell differentiation.

914   Middle: pie charts depict the frequency of CLL-specific DNA methylation events as fractions of

915   the dynamic (class A and B; left), and stable (class C and D; right) sites. Bottom: schematic

916   depicting the classification of CLL-specific DNA methylation events. We identified two groups:

917   'sites with epigenetic B cell programming' and 'sites without epigenetic B cell programming'.

918   'Sites with epigenetic B cell programming' undergo DNA methylation programming during

919   normal B cell differentiation, encompassing hypomethylation (class A) and hypermethylation

920   events (class B) relative to the DNA methylome of the COO. 'Sites without epigenetic B cell

921   programming' are defined as CpG sites without significant DNA methylation changes during

922   normal B cell differentiation and are classified as either hypo- or hypermethylation (class C and

923   D, respectively). Numbers of CLL-specific DNA methylation events (CLL-specific CpGs)

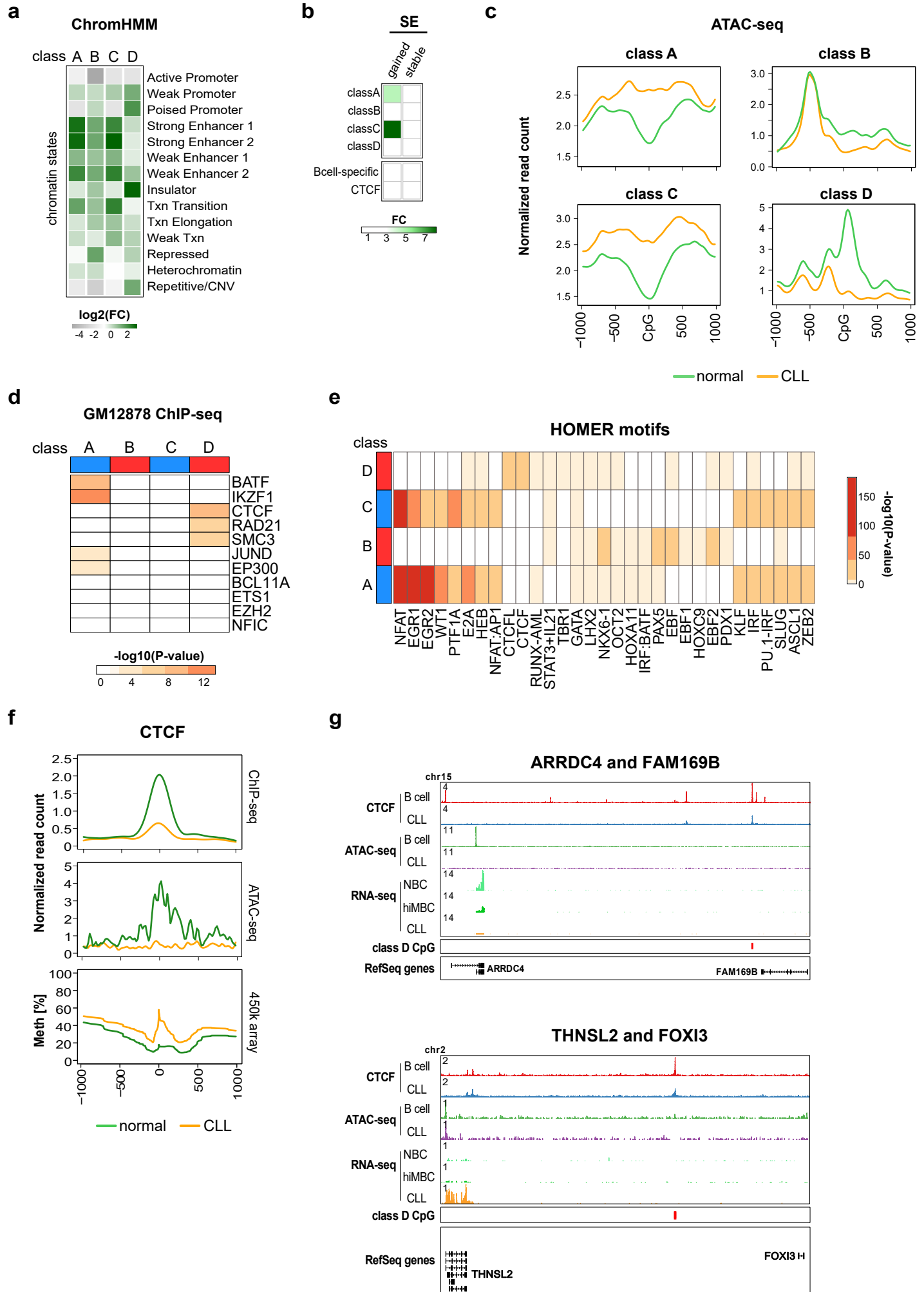924   resolved by class are indicated at the bottom.

925

926
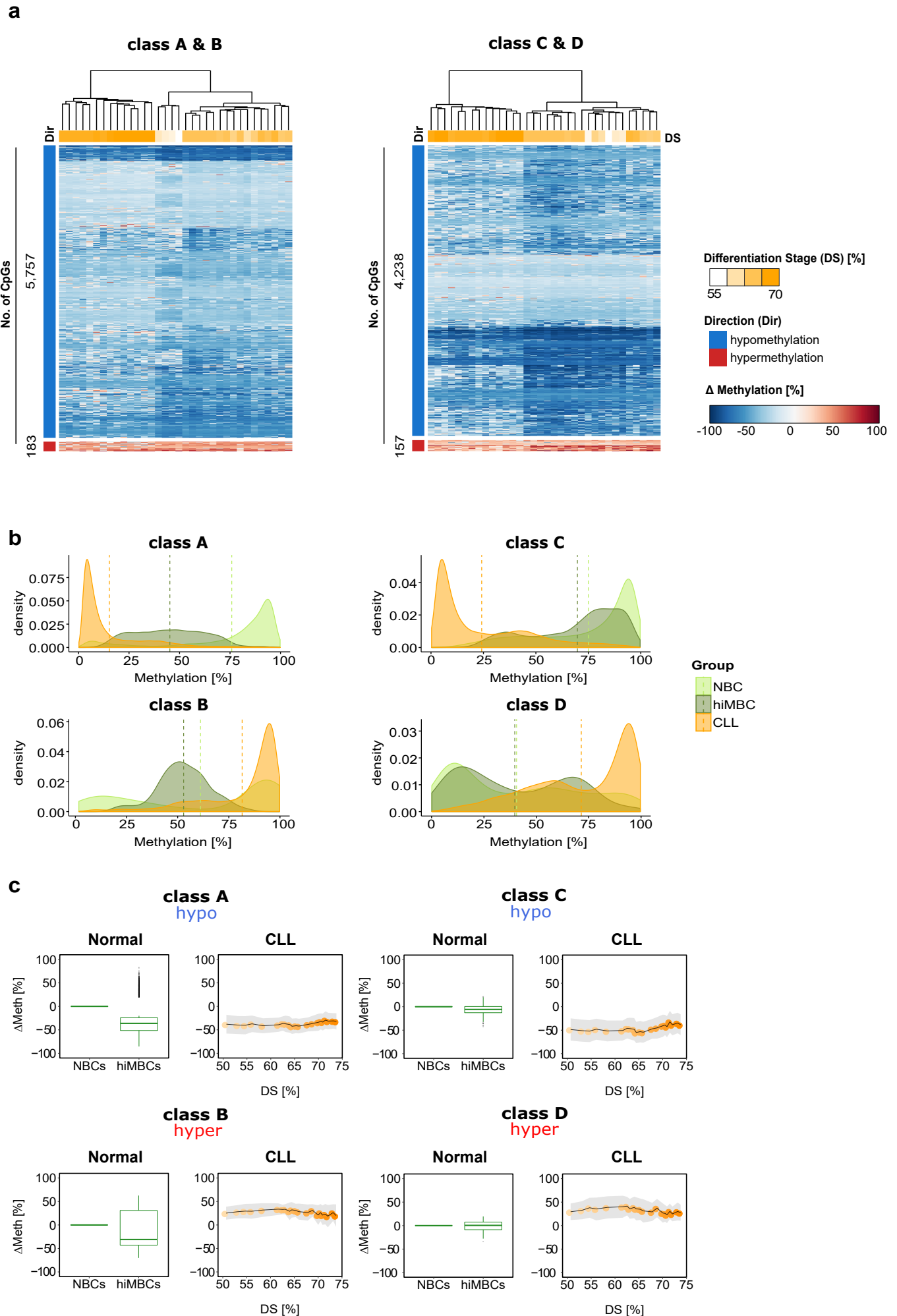
Figure 2

927 **Figure 2. Programming of disease-specific DNA methylation patterns in CLL**.

928 **a)** Heatmap depicting DNA methylation changes (ΔMethylation [%]) at CLL-specific CpG sites

929 relative to the samples' COO. Unsupervised hierarchical clustering of CLL-specific CpGs, class

930 A and B sites (left), class C and D sites (right). The direction of DNA methylation change (Dir

931 [%]) is indicated as blue and red bars for hypo- and hypermethylation, respectively, and the

932 numbers of CpG sites plotted are indicated next to the bars. Differentiation stages (DS) are

933 denoted as a color gradient (white-orange), where CLL samples with immature COO are

934 represented in white and samples with a more mature COO in orange. DS refers to % normal

935 differentiation programming achieved (relative to hiMBCs).

936 **b)** Density plots summarizing the distribution of absolute DNA methylation levels for all CLL-

937 specific CpG sites stratified by class (classes A - D). CLL patients (CLL): orange, naïve B cells

938 (NBC): light green, class-switched memory B cells (hiMBC): dark green.

939 **c)** Box plots and ribbon plots displaying the average DNA methylation change for each class of

940 CLL-specific alterations across normal B cells and CLLs. Left (normal): average DNA

941 methylation change (ΔMeth) of CLL-specific CpGs during normal B cell differentiation from

942 naïve B cells (NBCs) to class-switched memory B cells (hiMBCs) plotted for all classes (classes

943 A [n=5757 CpG sites], B [n=183 CpG sites], C [n=4238 CpG sites], and D [n=157 CpG sites]).

944 Right (CLL): ΔMeth for CLL-specific CpGs in CLL. ΔMeth [%] is represented as the mean DNA

945 methylation change relative to the expected DNA methylation level of the COO. Standard

946 deviation is depicted as grey shaded ribbons. DS refers to % normal differentiation

947 programming achieved (relative to hiMBCs).

948

949

Figure 3

Figure 3. CLL-specific DNA methylation differences result from aberrant transcription factor programming.

a) Enrichment of chromatin states in sequences representing CLL-specific DNA methylation. Chromatin states were defined using the 15-state ChromHMM model from immortalized B cells [88] for CLL-specific methylation sites of the classes A - D. The enrichment in category 'Repetitive/CNV' represents the averaged enrichment value of ChromHMM states called 'Repetitive/CNV'. Log2 fold change (log2 FC) was calculated using all 450k probes as a background.

b) Enrichment of super-enhancers (SE) in sequences representing CLL-specific DNA methylation. SE were defined as either being gained in CLLs (gained) or consensus between CLLs and B cells (stable). Fold change (FC) was calculated using all 450k probes as a background.

c) ATAC-seq read density (normalized read counts $*10^{-3}$) at CLL-specific CpG sites (±1kb) for categories of classes A, B, C and D. CLL samples (n=18) are represented in orange, normal CD19+ B cells (n=3) in green.

d) Transcription factor enrichment analysis using ENCODE ChIP-seq peaks from the B-cell lymphoblastoid cell line, GM12878. Displayed are $-\log_{10}$(p-values) resulting from Fisher's exact test with false discovery rate correction.

e) Transcription factor motif enrichment analysis using HOMER. The top 10 most enriched TF motifs for each class are displayed. The colors represent $-\log_{10}$(p-values) derived from a cumulative binomial distribution function as implemented in HOMER.

f) ATAC-seq & ChIP-seq read density (normalized read counts $*10^{-3}$) and DNA methylation profiles at class D CpGs co-locating with CTCF motifs (23 CpGs) (±1kb). CLL samples (n=7

44

973   CTCF ChIP-seq, n=18 ATAC-seq) are represented in orange, normal CD19+ B cells (n=4 CTCF

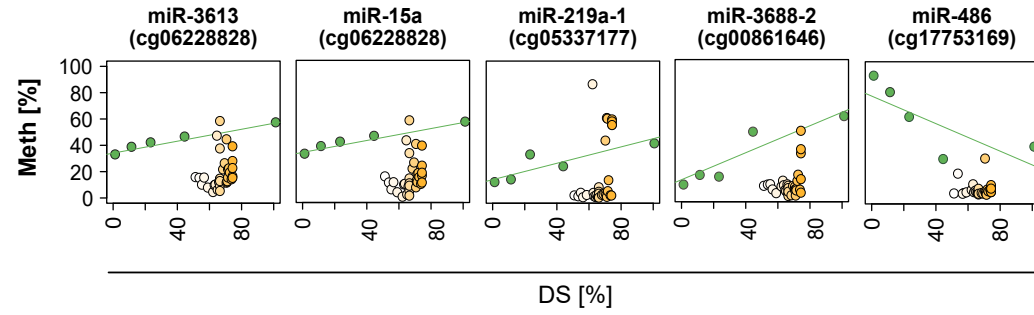974   ChIP-seq, n=3 ATAC-seq) in green.

975   **g)** Locus plots of exemplary genes associated with CTCF/classD events. Locus plots include

976   data from CTCF ChIP-seq on normal B cells (red) and CLL (blue); ATAC-seq on normal B cells

977   (green) and CLL (purple); RNA-seq on NBC (light green), hiMBC (dark green) and CLL
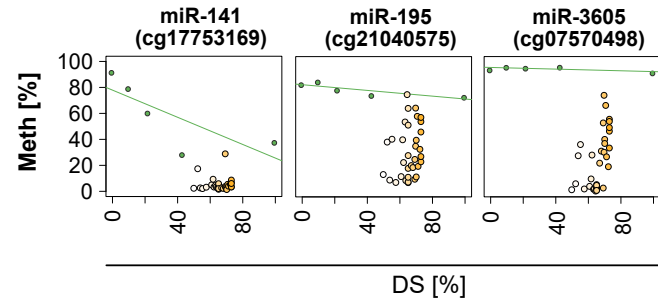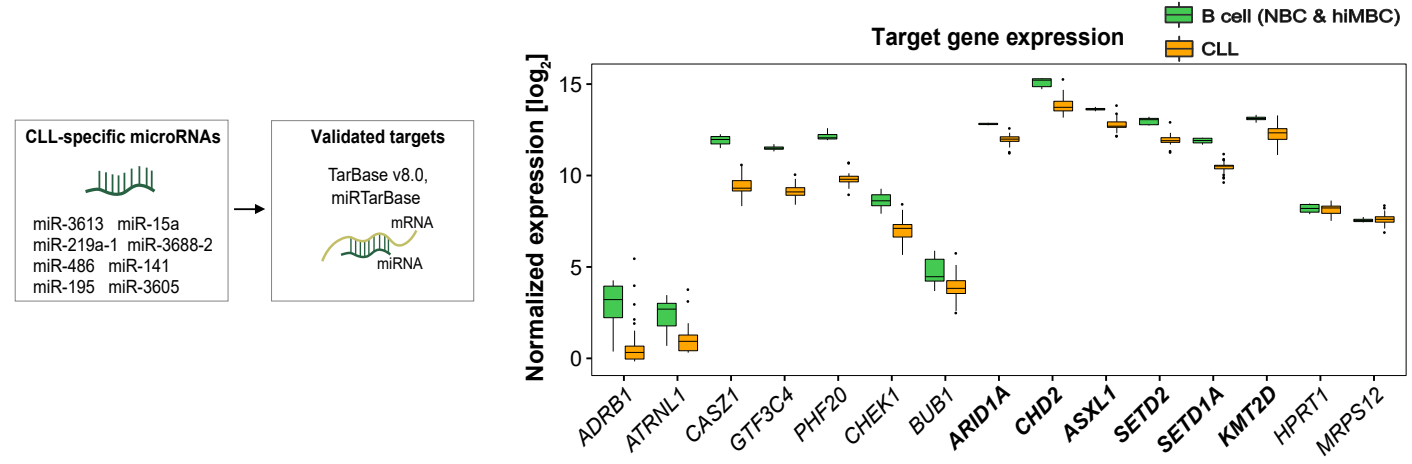
978   (orange). The class D CpGs are annotated in red.

979

980

45

Figure 4

a

b

c

981   **Figure 4. microRNAs associated with CLL-specific DNA methylation.**

982   **a)** Candidate CLL-specific microRNAs deregulated by class A events in their promoter regions.

983   Epigenetic programming during normal B cell differentiation is represented as a green line.

984   Average DNA methylation values are represented as dots; normal B cell subpopulations (green

985   dots); CLL samples (white-orange dots). The y-axis represents DNA methylation levels (%),

986   while the x-axis depicts the differentiation stage of normal B cell subpopulations and of CLL

987   samples relative to hiMBCs (DS).

988   **b)** Candidate CLL-specific microRNAs deregulated by class C events in their promoter regions.

989   Epigenetic programming during normal B cell differentiation is represented as a green line.

990   Average DNA methylation values are represented as dots; normal B cell subpopulations (green

991   dots); CLL samples (white-orange dots). The y-axis represents DNA methylation levels (%),

992   while the x-axis depicts the differentiation stage of normal B cell subpopulations and of CLL

993   samples relative to hiMBCs (DS).

994   **c)** CLL-specific microRNAs target epigenetic regulators. <u>Left panel</u>: schematic outline of

995   microRNA-target gene prediction. Two databases of experimentally validated targets of

996   microRNAs, TarBase v8.0 and miRTarBase, were used to define a set of CLL-specific

997   microRNA targets. <u>Right panel</u>: normalized gene expression levels (rlog normalized) of

998   epigenetic regulators being targeted by CLL-specific microRNAs as well as gene expression

999   levels of non-target genes (negative controls; *HPRT1* and *MRPS12*) are shown. Recurrently

1000 mutated epigenetic regulators in CLL are presented in bold. Statistical significance of

1001 expression change between normal B cells (NBCs, hiMBCs) and CLLs was tested using

1002 Wilcoxon rank sum test (p-values: *ARDB1*=0.002; *ATRNL1*=0.0013; *CASZ1*= 0.000014;

1003 *GTF3C4*=0.000014; *PHF20*=0.000014; *CHEK1*= 0.000025; *BUB1*= 0.007; *ARID1A*=0.000014;

46
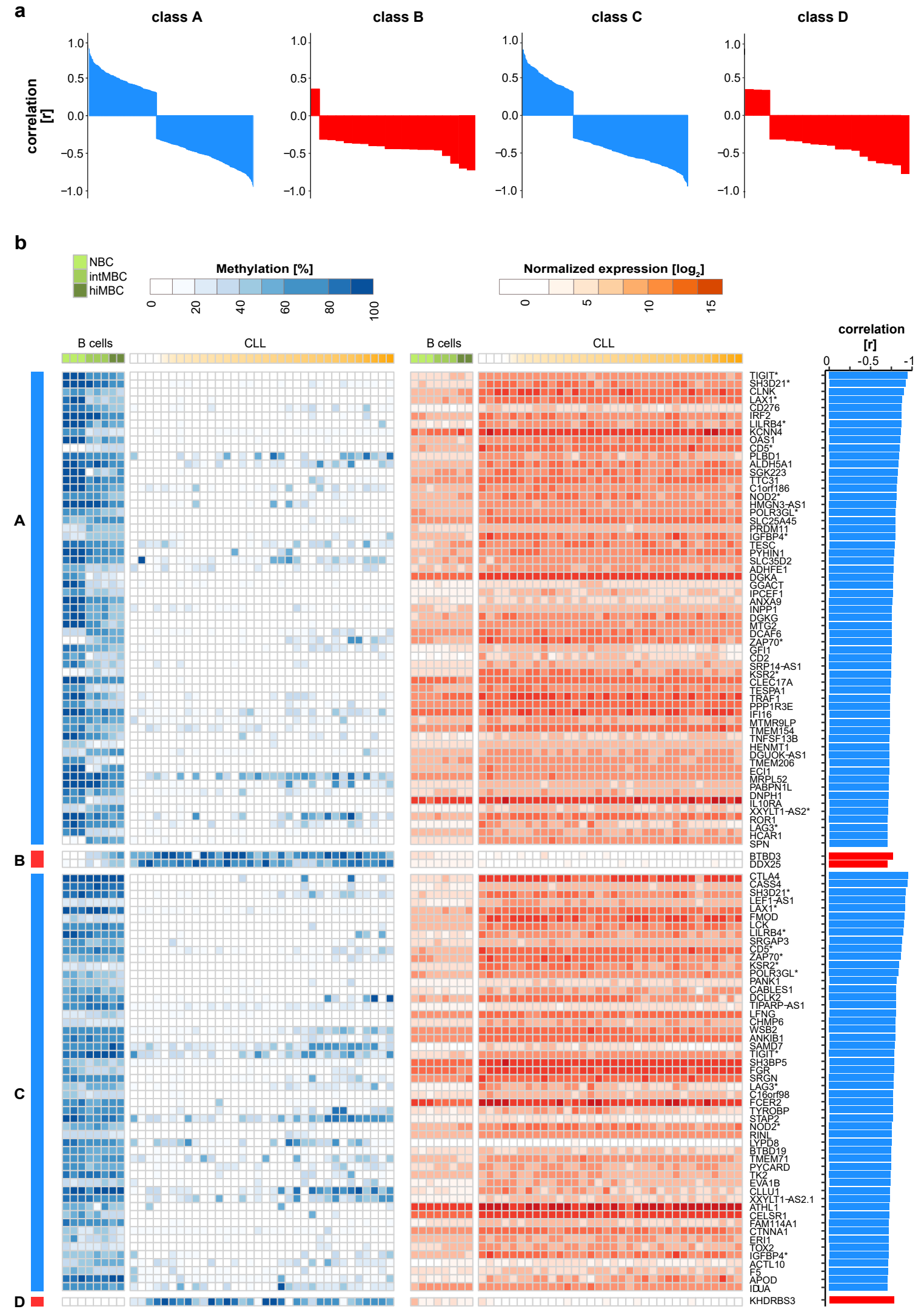
1004 *CHD2*=0.00003; *ASXL1*=0.00005; *SETD2*=0.00002; *SETD1A*=0.000014; *KMT2D*= 0.00007;

1005 *HPRT1*=0.43, *MRPS12*=0.45).

1006

1007

Figure 5

1008   **Figure 5. Protein-coding genes associated with CLL-specific aberrant DNA methylation.**

1009   **a)** Waterfall plots summarizing the correlation coefficients [r] between DNA methylation in the

1010   promoters and gene expression profiles of protein-coding genes for each class of CLL-specific

1011   alterations (classes A - D). The direction of DNA methylation change is indicated in blue and red

1012   for hypo- and hypermethylation, respectively.

1013   **b)** CLL-specific epigenetically deregulated transcripts. <u>Left panel</u>: heatmap depicting absolute

1014   DNA methylation levels [%] at CLL-specific CpG sites (classes A - D) in the promoter regions of

1015   protein-coding genes. Samples were sorted according to the differentiation stage. Differentiation

1016   stages are denoted as color gradients, CLLs (white - orange), normal B cells (light - dark green).

1017   <u>Middle panel</u>: heatmap depicting normalized gene expression levels (rlog normalization) of

1018   protein-coding genes in B cells (light - dark green) and CLLs (white - orange). Transcripts

1019   enriched for more than one class of CLL-specific events in their promoter regions are marked

1020   with asterisks. <u>Right panel</u>: barplots summarizing correlation coefficients [r] from Pearson

1021   correlation analysis between DNA methylation at CLL-specific CpGs in the promoter region and

1022   protein-coding gene expression levels. The direction of DNA methylation change is indicated in

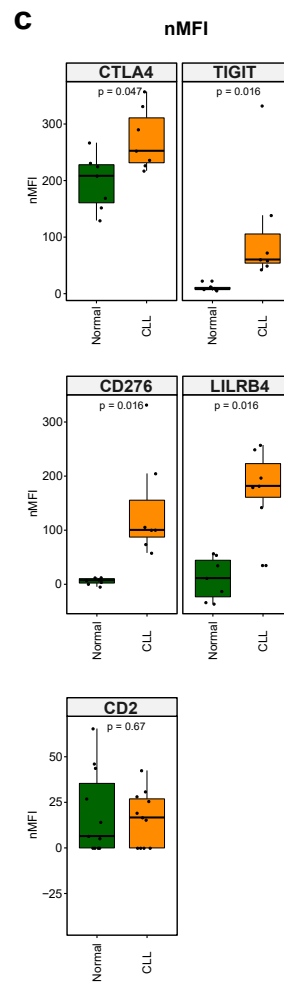1023   blue and red for hypo- and hypermethylation, respectively.

1024

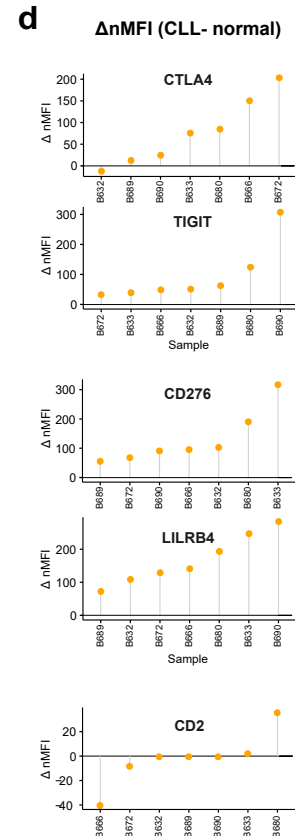1025

48
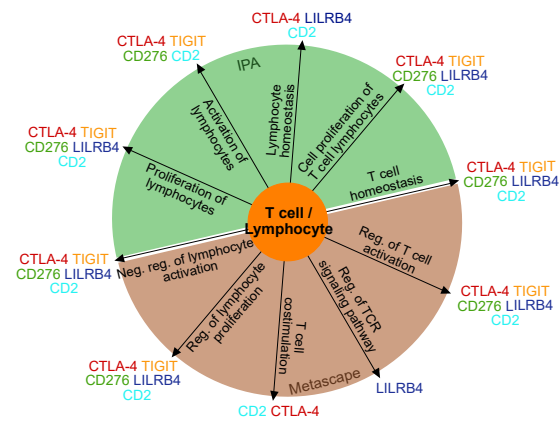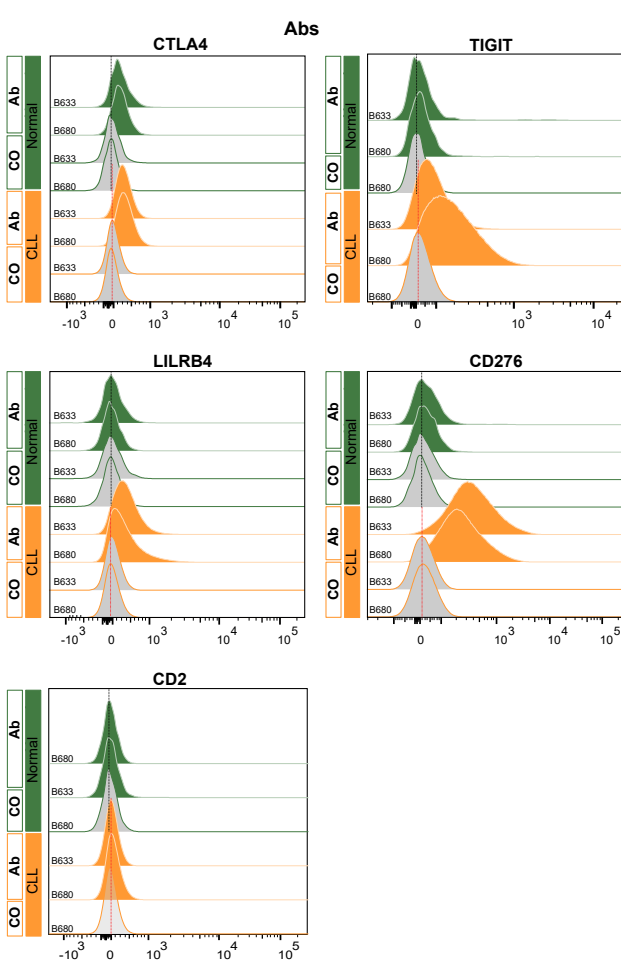
Figure 6

1026   **Figure 6. Flow cytometry analysis of T cell-/lymphocyte-specific markers on normal and**

1027   **malignant B cells from CLL patients.**

1028   **a)** Summary scheme representing functional implications of CLL-specific candidate genes

1029   selected for flow cytometric analysis.

1030   **b)** Flow cytometric analysis of expression of CTLA-4, TIGIT, CD276, LILRB4, and CD2 on

1031   peripheral blood B cells of CLL patients. The expression was determined for non-malignant B

1032   cells ('Normal'; CD19$^+$ CD5$^-$ B cells, represented in green) and neoplastic B cells ('CLL', CD19$^+$

1033   CD5$^+$ B cells, represented in orange) detected in the same samples. 'Co', no antibody staining

1034   control; 'Ab', staining with the antibody of interest as indicated.

1035   **c)** Normalized median fluorescence intensities (target MFI - MFI of negative control [Co]; nMFI).

1036   **d)** Δ normalized median fluorescence intensities between CLL cells and normal B cells (ΔnMFI

1037   (CLL-normal)) for each patient tested.

49