SCIENTIFIC REPORTS
natureresearch

OPEN

Novel genes exhibiting DNA methylation alterations in Korean patients with chronic lymphocytic leukaemia: a methyl-CpG-binding domain sequencing study

Miyoung Kim[1], Eunyup Lee[1], Dae Young Zang[2], Hyo Jung Kim[2], Ho Young Kim[2], Boram Han[2], Han-Sung Kim[1], Hee Jung Kang[1], Seungwoo Hwang[3*] & Young Kyung Lee[1*]

Chronic lymphocytic leukaemia (CLL) exhibits differences between Asians and Caucasians in terms of incidence rate, age at onset, immunophenotype, and genetic profile. We performed genome-wide methylation profiling of CLL in an Asian cohort for the first time. Eight Korean patients without somatic immunoglobulin heavy chain gene hypermutations underwent methyl-CpG-binding domain sequencing (MBD-seq), as did five control subjects. Gene Ontology, pathway analysis, and network-based prioritization of differentially methylated genes were also performed. More regions were hypomethylated (2,062 windows) than were hypermethylated (777 windows). Promoters contained the highest proportion of differentially methylated regions (0.08%), while distal intergenic and intron regions contained the largest number of differentially methylated regions. Protein-coding genes were the most abundant, followed by long noncoding and short noncoding genes. The most significantly over-represented signalling pathways in the differentially methylated gene list included immune/cancer-related pathways and B-cell receptor signalling. Among the top 10 hub genes identified via network-based prioritization, four (*UBC*, *GRB2*, *CREBBP*, and *GAB2*) had no known relevance to CLL, while the other six (*STAT3*, *PTPN6*, *SYK*, *STAT5B*, *XPO1*, and *ABL1*) have previously been linked to CLL in Caucasians. As such, our analysis identified four novel candidate genes of potential significance to Asian patients with CLL.

Chronic lymphocytic leukaemia (CLL) is characterized by the co-expression of CD5 and CD23 in monomorphic small B-cells; it is the most common leukaemia among adults in Western countries, especially the elderly[1]. CLL among Asians differs from that among Caucasians in terms of the incidence rate, median age at onset, phenotype, and the genetic profile. The incidence rates per 100,000 person-years in Korea and Japan are 0.04 and 0.48, respectively, but the rate is 3.83 in Western countries[2]. The median age at initial CLL diagnosis is 61 and 70 years among Asians and Caucasians, respectively[2]. A single-institution study in South Korea found that 56% of patients had an atypical immunophenotype with high frequencies of FMC7 positivity and strong CD22 positivity[3]. A Chinese study showed that *TP53* mutations are more common in Chinese patients with CLL than in Caucasian patients, whereas *SF3B1* mutations are less common[4]. Furthermore, a Korean study found that the frequencies of mutations in *ATM, TP53, KLHL6, BCOR*, and *CDKN2A* tend to be higher in Koreans than in Caucasians, while those in *SF3B1, NOTCH1, CHD2*, and *POT1* tend to be lower[2].

DNA methylation directly impacts human genome function, and multiple studies have demonstrated the existence of aberrant epigenetic changes that play important roles in tumour initiation and progression in Western patients with CLL[5–8]. Recent advances in high-throughput techniques have enabled genome-wide methylation profiling in Caucasians with CLL. For example, an array study identified methylation in seven known or candidate

[1]Department of Laboratory Medicine, Hallym University Sacred Heart Hospital, Anyang, Republic of Korea. [2]Department of Internal Medicine, Hallym University Sacred Heart Hospital, Anyang, Republic of Korea. [3]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Republic of Korea. *email: swhwang@kribb.re.kr; lyoungk@hallym.or.kr

tumour suppressor genes (including *VHL, ABI3*, and *IGSF4*) as well as eight unmethylated genes involved in cell proliferation and tumour progression (including *ADORA3* and *PRF1*) in Swedes with immunoglobulin heavy chain gene variable region (*IGHV*)-unmutated CLL[9]. Another study of the same cohort found 2,239 CpG sites that were differentially methylated in *IGHV*-mutated and unmutated patients; DNA methylation over time was relatively stable, implying that aberrant methylation is an early leukaemogenic event[10]. Another Spanish study of 139 patients that included in-depth interrogation using whole-genome bisulfite sequencing showed that *IGHV*-mutated and -unmutated CLL had differing DNA methylomes that represented epigenetic imprints from distinct normal B-cell subpopulations[11]. Additionally, an American cohort study using reduced-representation bisulfite sequencing showed that CLL cells consistently displayed higher intra-sample variability in DNA methylation patterns across the genome, implying that disordered methylation is akin to genetic instability, thereby enhancing the ability of cancer cells to follow superior evolutionary trajectories[12].

Methyl-CpG binding domain (MBD) sequencing (MBD-seq), a next-generation sequencing technique for genome-wide methylation profiling, sequences the DNA captured by the MBD[13]. MBD-seq is an affinity enrichment-based method in which methyl-CpG binding proteins link to methylated CpGs via the MBD. This technique has several advantages; for example, since DNA methylation occurs primarily within CpG dinucleotides (which represent approximately 1% of the genome), MBD-seq is efficient as it comprehensively interrogates only regions of relevance[14]. Furthermore, MBD-seq does not show restriction enzyme-dependent bias toward CpG-rich regions, resulting in greater genome-wide coverage (61%) than that using reduced-representation bisulfite sequencing (12%)[15]. Compared to array platforms, the genome-wide coverage of MBD-seq is also not restricted to the fixed array content. MBD-seq use has been increasing owing to the aforementioned reasons; however, only one MBD-seq-based study of CLL has been published to date[16]. That study showed that 40% and 60% of hypermethylated and hypomethylated genes, respectively, were mapped to noncoding RNAs. It was also observed that the major repetitive elements (such as the short and long interspersed elements) have a high percentage of differentially methylated regions in *IGHV*-mutated subgroups compared to normal controls.

In contrast to Caucasians, from whom abundant data are available, the methylation profiles of CLL in Asians has never been reported. Hence, we aimed to elucidate the role of aberrant methylation in the pathogenesis of Asian CLL for the first time, and to investigate the differences in methylation patterns between Asian and Caucasian patients with CLL. We also sought to identify novel candidate genes with potentially high functional relevance in CLL using protein-protein and protein-DNA network-based approaches that link the differentially methylated genes (DMGs) to known CLL-related genes.

Results

Genome-wide distribution of differentially methylated regions.

Our approach using MBD-seq enabled a comprehensive genome-wide interrogation of differentially methylated regions in CLL. Stringent criteria (false discovery rate <0.01) resulted in a total of 2,839 windows of 250 nucleotides that were differentially methylated between the CLL and normal groups. There were more hypomethylated regions (2,062 windows) than hypermethylated ones (777 windows). The genomic annotations of all windows created by binning the human genome into adjacent 250-nucleotide windows are shown in Fig. 1A; introns and distal intergenic regions constituted most (86%) of the human genome, followed by promoter regions (7%). The genomic annotations of the differentially methylated regions are shown in Fig. 1B; they most frequently overlapped with distal intergenic regions, followed by introns and promoters. Hypomethylation was more prevalent than hypermethylation regardless of the type of annotation. The proportions of the differentially methylated regions were obtained by dividing the numbers in Fig. 1B by those in Fig. 1A; this revealed that promoters contained the highest proportion of the differentially methylated regions (0.08%) and were thus the preferential targets of differential methylation in CLL (Fig. 1C). Nevertheless, distal intergenic and intron regions contained the largest numbers of differentially methylated regions (Fig. 1B), highlighting the importance of differential methylation in these regions.

The proportions in Fig. 1C were tested with the two-tailed Fisher's exact test (Table 1). Promoters and introns were the most significant; differential methylation was observed more often in promoters than would be expected by chance and less often in introns. Distal intergenic regions were the next significant genomic annotation and showed enrichment with differential methylation.

When visualized as a heatmap (Fig. 2), the differential methylation profiles indicated more hypomethylated regions than hypermethylated ones. Moreover, the CLL and normal groups were correctly separated by supervised hierarchical clustering with the methylation profile of the differentially methylated regions, as expected.

DMGs.

The list of DMGs (i.e., genes that overlapped with differentially methylated regions) is provided as Supplementary Data S1; there were 1,507 DMGs (1,241 hypomethylated and 315 hypermethylated). Rarely, some of the DMGs showed both hypomethylation and hypermethylation in their genic regions (48 genes; 3%). For brevity, the top 40 hypermethylated and hypomethylated genes are shown in Supplementary Tables 1 and 2, respectively.

The DMGs were classified with respect to gene type (Fig. 3). Protein-coding genes were the most abundant, followed by long noncoding genes such as long intronic noncoding RNAs, short noncoding genes such as microRNAs and small nucleolar RNAs, and others such as pseudogenes and unannotated nucleotides. Their proportions in the hypomethylated and hypermethylated DMG sets were similar.

Gene ontology (GO) biological processes and pathways over-represented in the DMG list.

Functional enrichment analyses of the DMG lists identified a number of significantly over-represented GO terms and pathways; up to 30 categories were selected for each condition to examine functional categories specific to either hyper- or hypomethylation.

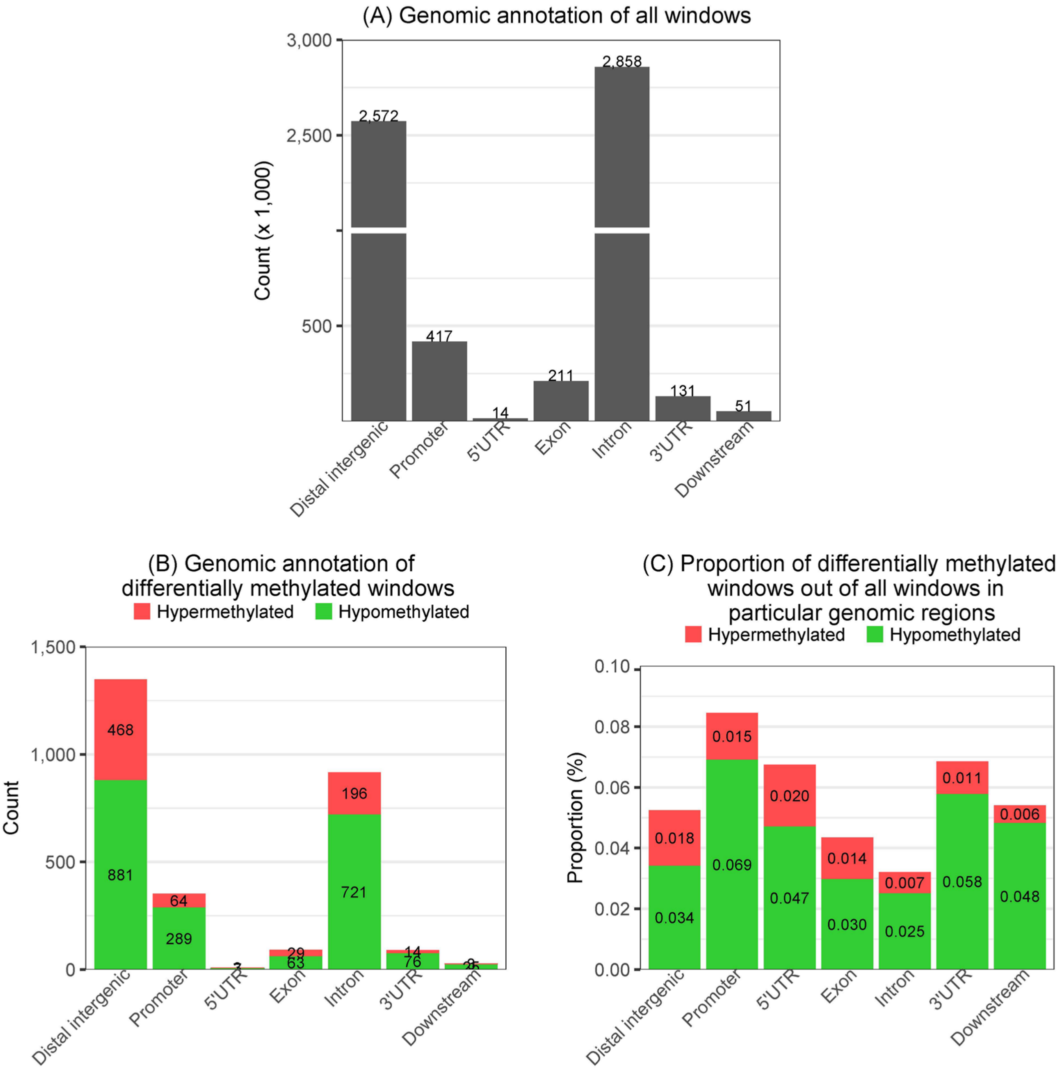

Figure 1. Genomic annotations of differentially methylated regions and their relative proportions with respect to background occurrences in the human genome. **(A)** Genomic annotation of all 250 nucleotide-long windows along the human genome. **(B)** Genomic annotations of the differentially methylated windows. **(C)** Proportion of differentially methylated windows out of all windows in particular genomic regions. UTR, untranslated region.

Genomic annotation	P-value	Odds ratio	95% confidence interval
Promoter	<2.2e-16	1.987	1.772–2.222
Intron	<2.2e-16	0.567	0.524–0.614
Distal intergenic	5.39e-12	1.297	1.204–1.397
3′UTR	0.000178	1.528	1.225–1.886
5′UTR	0.2395	1.488	0.713–2.742
Downstream	0.3491	1.195	0.793–1.731
Exon	0.7550	0.958	0.770–1.180

Table 1. The extent of association between genomic annotation and differential methylation as tested by two-tailed Fisher's exact test.

Numerous immune processes and cancer-related GO terms were over-represented in both the hypo- and hypermethylated gene lists (Fig. 4, highlighted in bold), with the former being more prevalent.

With respect to genomic annotation, most GO terms were over-represented in the genes that showed differential methylation in introns (Supplementary Fig. 1) owing to the presence of many differentially methylated

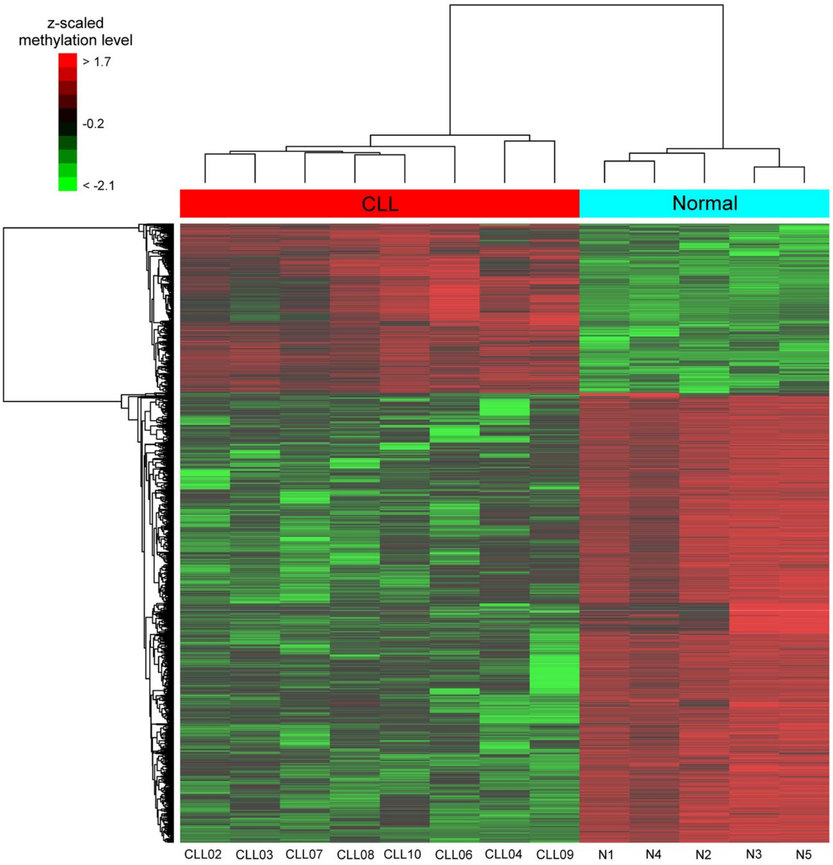

Figure 2. Hierarchical clustering of samples reflecting the profiles of normal and chronic lymphocytic leukaemia (CLL) samples regarding differentially methylated regions. Supervised hierarchical clustering correctly separates the CLL and normal (N) groups. The colour intensity is scaled within each row so that the highest methylation value corresponds to bright red and the lowest to bright green.

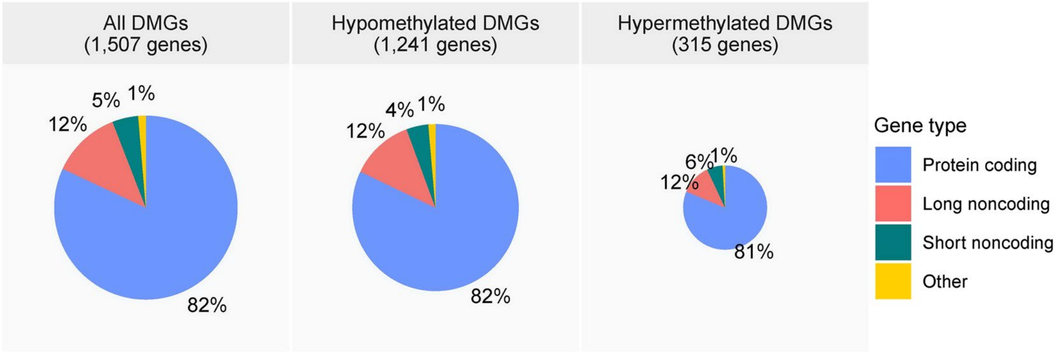

Figure 3. Gene type classification of differentially methylated genes (DMGs).

regions. Some immune- and cancer-related GO terms were also over-represented in the genes that showed differential methylation in promoters and exons.

Over-represented pathways also included those that are directly related to immune processes and cancer (Fig. 5; the Kyoto Encyclopaedia of Genes and Genomes [KEGG] pathway database does not include CLL-related content). Overall, most pathways were over-represented in the gene lists that showed hypomethylation (Fig. 5) or differential methylation of introns (Supplementary Fig. 2). Some immune- and cancer-related pathways were also over-represented in the gene lists that showed differential methylation of promoters and exons.

The B-cell receptor signalling pathway, which plays a crucial role in the pathogenesis of CLL and is a therapeutic target, was the most significantly over-represented (Fig. 5). The DMGs in this pathway are shown in Fig. 6.

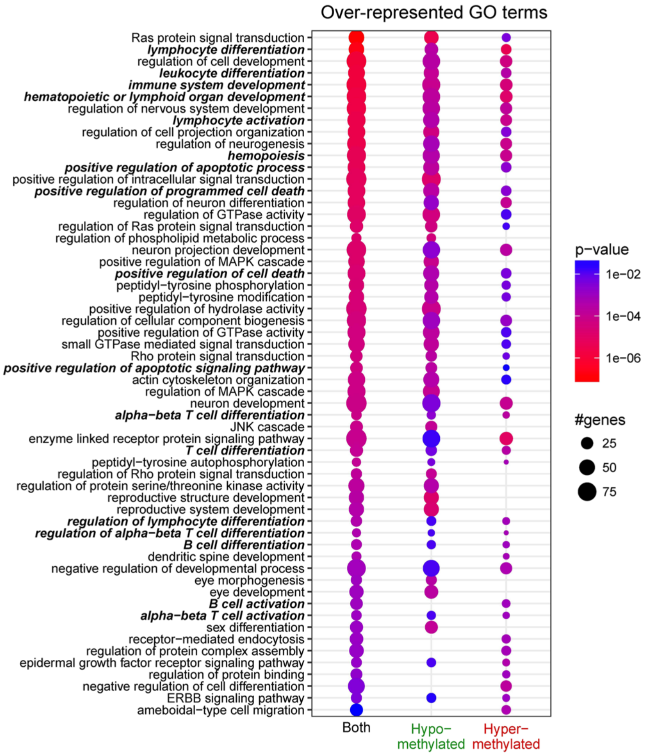

Figure 4. Most significantly over-represented Gene Ontology (GO) terms analysed with respect to differential methylation. A maximum of 30 GO terms were selected for each track; the dot plot shows their p-values according to the displayed colour code. The statistical significance of a GO term increases with a redder tint. The number of differentially methylated genes (DMGs) belonging to a GO term increases with its size. The absence of a dot denotes insignificant over-representation of the corresponding GO terms under that particular condition. GO terms directly related to immune processes and cancer are shown in bold-type.

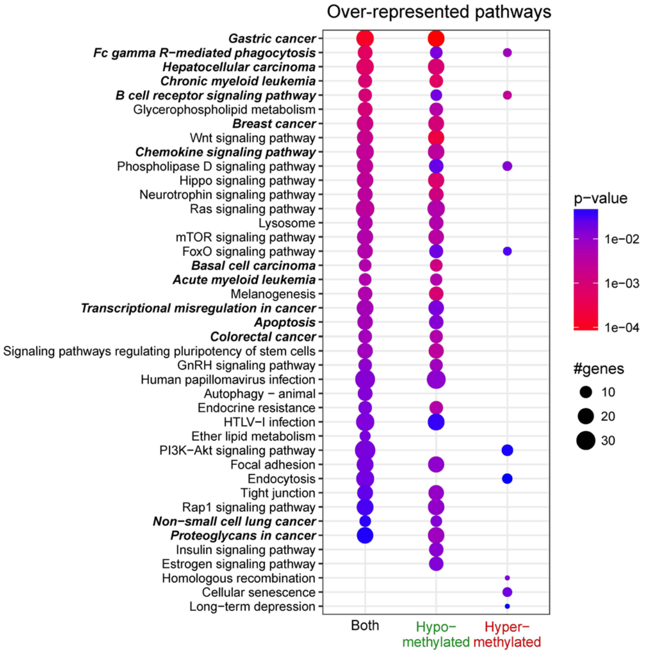

Figure 5. Most significantly over-represented pathways analysed with respect differential methylation. The plot was prepared in the same manner as in Fig. 4.

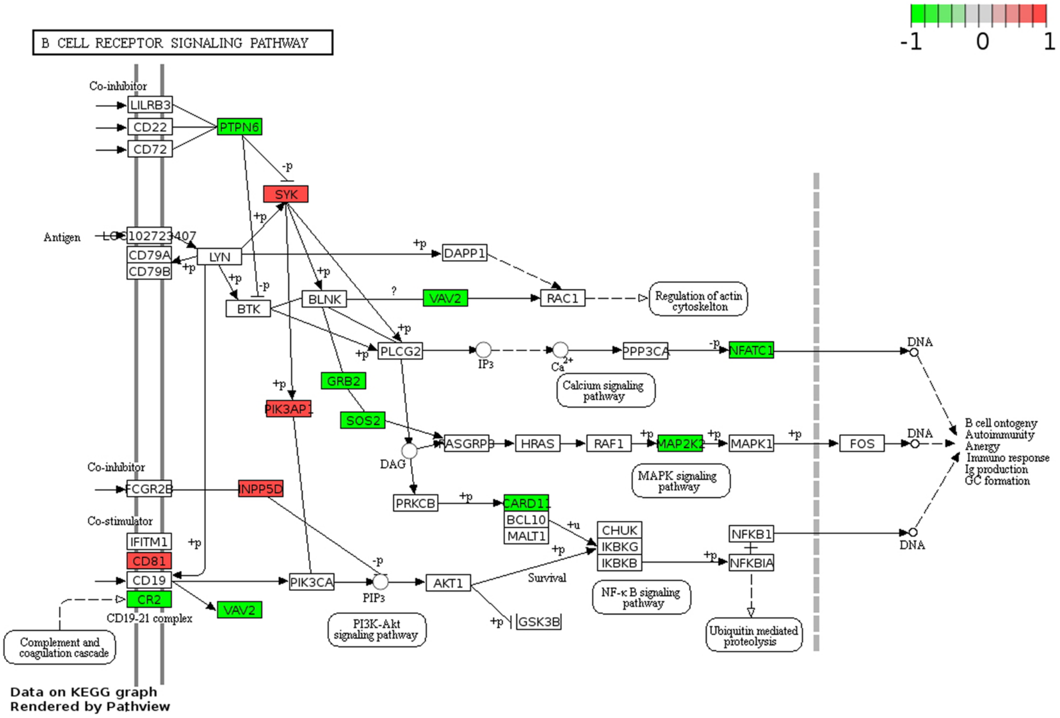

Figure 6. Differential methylation of genes in the B-cell receptor signalling pathway. Hypo- and hypermethylated genes are indicated in green and red, respectively.

Network-based prioritization of DMGs. We performed a network-based prioritization of the DMGs based on the assumption that a DMG that interacts with many known CLL-related genes is also likely to be functionally relevant to CLL itself. A network of 1,094 nodes and 3,304 edges was obtained by linking DMGs to known CLL-related genes by protein-protein and protein-DNA interactions. For easier navigation, we first selected the 10 hub DMGs with the strongest links to known CLL-related genes, then extracted a subnetwork that consisted of the top 10 hub DMGs and their known CLL-related gene interactors. The subnetwork comprised 203 nodes and 512 edges (Fig. 7).

The hub DMGs have the potential to influence or be influenced by the activities of their interacting partners; hence, differential methylation of the hub DMGs may have functional implications in CLL. Among the 10 hub DMGs revealed, six (*STAT3*, *PTPN6*, *SYK*, *STAT5B*, *XPO1*, and *ABL1*) had known CLL relevance (blue tint), whereas the remaining four (*UBC*, *GRB2*, *CREBBP*, and *GAB2*) did not. The former corroborated known data, while the latter demonstrates a novel finding.

Discussion

Our MBD-seq and subsequent bioinformatics analysis showed that Korean CLL shares similar features with Caucasian CLL in terms of frequent hypomethylation in intragenic regions[11,17] and aberrant methylation in pathways related to B-cell development, immune processes, and cancer[17,18]. The network analysis identified 10 hub DMGs, 4 of which (*UBC*, *GRB2*, *CREBBP*, and *GAB2*) were reported as being CLL-related for the first time in our study, suggesting potential methylome differences between Korean and Western CLLs.

We identified a total of 2,839 differentially methylated regions, which comprise approximately 0.06% of the human genome. This proportion is comparable to that of previous studies, although most such studies did not investigate proportions and used different methodologies[6,9,18]. Differential methylation was observed not only in promoter regions but also in intragenic and distal intergenic regions, with a higher proportion of hypomethylation than hypermethylation; this was consistent with the findings of a comparable MBD-seq-based study by Subhash *et al.*[16] as well as a whole-genome bisulfite sequencing study by Kulis *et al.*[19]. While promoter hypermethylation is a well-known tumour suppressor mechanism, hypomethylation has not been as thoroughly investigated. Nevertheless, global hypomethylation (particularly in the intragenic regions) has been noted as a hallmark of CLL and presumed to contribute to genomic instability and gene activation during the pathogenesis of the disease[18]. Examples of hypomethylation in oncogenesis include (i) cancer-linked hypomethylation in gene regulatory regions[20,21] and (ii) hypomethylation of interspersed and tandem repeats that promote tumour formation or progression by fostering DNA rearrangements[22,23]. It has also been shown that CpG islands that are not associated with the 5′ region but are located in intergenic or intragenic CpG islands of any gene can perform important biological functions[24].

Noncoding genes that were frequently hyper- or hypomethylated accounted for up to 20% of DMGs, while the remaining 80% were protein-coding. DNA methylation is a relatively stable modification causing transcriptional inactivation of both protein coding genes and non-coding regulatory microRNA genes[25–27] and is therefore a

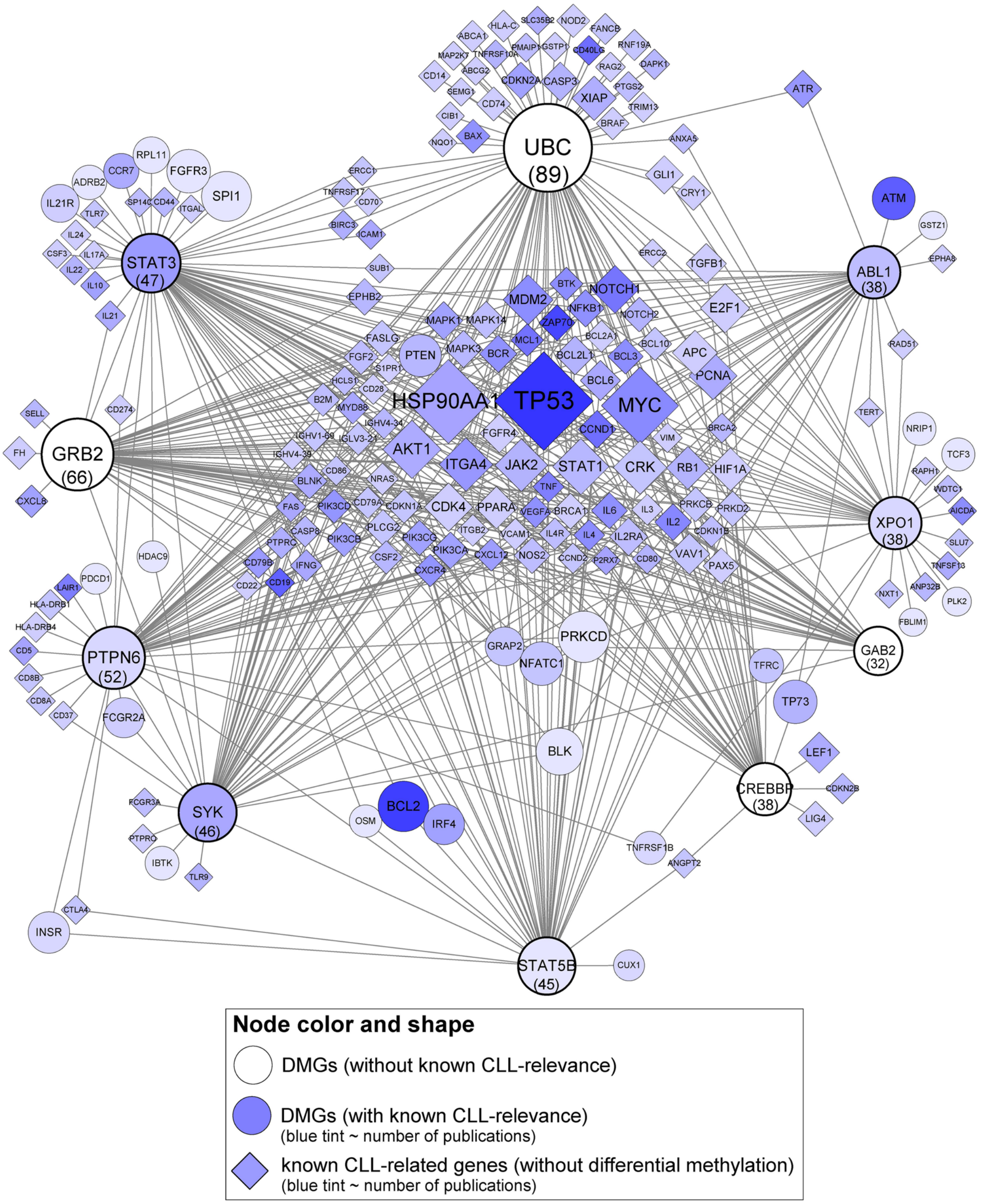

Figure 7. Interaction subnetwork between the top 10 hub differentially methylated genes (DMGs) and known chronic lymphocytic leukaemia (CLL)-related genes. The top 10 hub DMGs are shown along the periphery of the network. Numbers in parentheses denote the node degree of the hub DMGs; the node size is scaled by its node degree.

main mechanism of aberrant gene silencing in cancer[17]. The dysregulation of long noncoding RNAs such as long intronic noncoding RNAs promotes carcinogenesis, disease progression, and metastasis in various cancers[28,29] including CLL[30]. These long noncoding RNA genes encode non-protein-coding transcripts of $>$200 nucleotides generated by RNA polymerase II, and their expression is tightly regulated in a cell type-specific and/or cellular differential stage-specific manner[31]. They comprised 12% of the DMGs in our study, and some were listed in the top 40 hypermethylated (including *LINC00273* and *LINC00839*) as well as in the top 40 hypomethylated genes (such as *LINC00348*); none of these three genes have been reported in previous Western studies. Meanwhile, microRNA genes, which encode a class of single-stranded noncoding RNAs 19–25 nucleotides in length[32], can either be oncogenic or tumour suppressive[17], and their aberrant methylation has clearly been implicated in CLL pathogenesis in previous studies[17,32]. Previously reported microRNA genes that elicit epigenetic changes in Caucasian CLL include *miR15a, miR16-1, miR-21, miR-29a, miR-34a, miR-139, miR-155, miR-574, miR-582*, and *miR1204*[17,32]. Aggressive and indolent CLLs exhibit a different microRNA profile, and high levels of miR-21 and miR-155 are associated with a greater mortality rate[32]. In our study, microRNA genes represented 5% of the DMGs, and four (*MIR4436A*, *MIR4537*, *MIR4715*, and *MIR7850*) that were among the top 40 hypermethylated genes have never been reported in previous Western studies. Further investigation on aberrant methylation of these novel long noncoding RNA and microRNA genes will provide a better insight into their roles in the pathogenesis of CLL in Korean individuals.

GO and pathway analyses showed findings consistent with those of previous CLL studies in Caucasians; the most significantly over-represented GO terms included those for lymphocyte differentiation, immune system development, lymphocyte activation, B-cell differentiation, and/or B-cell activation[17,18]. The roles of the B-cell receptor signalling pathway genes *SYK, PIK3AP1, PTPN6, MAP2K2*, and *NFATC1* in CLL pathogenesis have also been previously described[33–37]. *SYK* is a tyrosine kinase and is involved in the CD38 signal transduction pathway in CLL, and a selective Syk inhibitor is currently undergoing a clinical trial[34]. An expression study revealed that *PIK3AP1* is involved in the B-cell receptor signalling pathway in CLL as shown via functional enrichment analysis[37]. *PTPN6* encodes SHP-1 and is an important negative modulator of antigen-receptor signalling in lymphocytes; it is activated by *NOTCH1*, which has an important pathogenic role in CLL[33]. *MAP2K2* is involved in the RAS-BRAF-MAPK-ERK pathway, and mutations in this gene have been observed in CLL[34]. *NFATC1* activation by DNA hypomethylation in CLL correlates with clinical staging and can be inhibited by ibrutinib[36]. Our data demonstrated that CLL in Koreans shares common features with CLL in Caucasians in this regard.

Our network-based prioritization analysis identified genes that were differentially methylated and that are linked to many known CLL-related genes via protein-protein and protein-DNA interactions. Among the 10 hub DMGs revealed in our analysis, six (*STAT3, PTPN6, SYK, STAT5B*,and *XPO1*, and *ABL1*) had known CLL relevance; hence, our results corroborated previous data. The remaining four (*UBC, GRB2, CREBBP*, and *GAB2*) have never been reported in previous CLL studies[38–48]. *UBC* represents a ubiquitin gene (ubiquitin C) and has been described in cancers infrequently. In a previous study, interaction analysis of biomarker genes revealed that *UBC* may have a major role in renal cancer[38]. *GRB2* encodes growth factor receptor-bound protein 2 and has been described in cancers relatively frequently; as such, anti-cancer therapeutics targeting *GRB2* are currently in development[48]. *CREBBP* encodes chromatin-modifying enzymes such as the histone acetyl-transferases and has been studied in diffuse large B cell lymphoma, acute lymphoblastic leukaemia, and lung cancer[41–45]. *GAB2* encodes the *GRB2* associated binding protein 2[46] and has been studied in breast cancer, ovarian cancer, hepatocellular carcinoma, lung cancer, and melanoma[47,48]. The top three genes most relevant to CLL in our network were *TP53*, *BCL2*, and *ZAP70*. Among them, *TP53* and *ZAP70* interacted with the four novel hub DMGs. The interactions of *UBC* and *CREBBP* with *TP53* represent post-translational regulation of the p53 protein via ubiquitination and acetylation[49]. *GRB2* and *GAB2* interacted with *ZAP70*[50]; the ZAP-70-mediated phosphorylation of the GRB2/GAB2 protein complex serves as a scaffold for the assembly of downstream signalling proteins[50]. Taken together, the interactions between the four hub DMGs and the well-known CLL-related genes underscore their biological significance.

Some recent epigenetic studies of CLL provided new insights into the chromatin landscape of this disease[19,51,52]; our data can be regarded as building on such knowledge. Previous studies that aimed to identify CLL-specific methylation events compared CLL cells to normal CD19+ B cells in order to pinpoint the specific features that represent the epigenetic characteristics of CLL[9,10,53,54]. Recent epigenetic studies also compared the chromatin landscape of CLL cells and B-cells from different maturation stages[19,51,52] and observed that a large proportion of the differentially methylated sites overlap with those undergoing dynamic methylation during normal differentiation, mainly those of memory B-cells and bone marrow plasma cells[19,51,52]. This suggests that virtually all reported 'CLL-specific' differences reflected normal B cell maturation and are likely not causative of the disease[51]. They also reported that (i) early differentiation stages mainly displayed enhancer demethylation, which was associated with the upregulation of key B-cell transcription factors, and affected multiple genes involved in B-cell biology[19]; and (ii) CpGs losing methylation at any B-cell maturation stage were preferentially located in introns, intergenic regions, and repetitive elements[19]. As we did not compare the methylation patterns of our CLL cells to those of memory B-cells isolated from control individuals (as performed in previous Western studies), we infer that most of our differentially methylated regions (which were also identified in Western studies) might overlap with those of normal memory B-cells, and are maturation stage-specific rather than disease-specific. Nevertheless, our observation of over-represented B-cell receptor signalling pathway components and prevalent hypomethylation in the distal intergenic and intron regions were consistent with data from studies in which B-cells of different maturation stages were analysed separately, thereby affirming the credibility of our data. We infer that the common features shared with previous Western studies might have been derived from the chromatin landscape of normal memory B-cells, although the unique findings in our CLL samples (i.e., those which have not been reported in previous Western studies) differentiate Korean CLL from its Western counterpart. As Kulis *et al*. concluded, the changes shared during neoplastic transformation and normal differentiation may be epigenetic 'passengers', whereas those exclusively occurring in CLL cells, as we observed in our study, were likely epigenetic 'drivers' with a potential role in CLL development[19].

A limitation of our approach is that we were unable to assess the expression statuses of genes speculated to be affected by differential hypo- or hypermethylation owing to the lack of RNA samples. Even though the main consequence of aberrant promoter methylation is dysregulated gene expression, the consequences of aberrant hypo- or hypermethylation are not limited to the alteration of transcription. Kulis *et al*. also reported that they rarely observed a direct correlation between gene expression and DNA methylation, even in regulatory elements, which was similar to previous observations[11,19,55,56]. In their study comparing *IGHV*-mutated and *IGHV*-unmutated CLLs, Beekman *et al*. suggested that the different cellular origins of these two types of CLL do not necessarily imply differential chromatin activation, likely owing to the fact that the differential DNA methylation in the *IGHV*-mutated and *IGHV*-unmutated CLL-originating cells is independent of the differential expression of the target genes[51]. In fact, dysregulated methylation outside the promoter can affect carcinogenesis and other conditions such as embryonic development, atherosclerosis, aging, and neural development via chromosomal instability[57] or alternative splicing[58]. Thus, our methylation profiling data ought to be valuable even without considering gene expression, as they provide a comprehensive picture of aberrant hypo- and hypermethylation in Koreans with CLL.

In summary, we used MBD-seq to perform global methylation profiling of CLL in an Asian population for the first time. Our results showed that promoters were the preferential targets of differential methylation, as were distal intergenic and intron regions. Along with protein-coding genes, long intronic noncoding RNAs and microRNAs were frequently affected as well. Pathways related to immune processes and cancer were the main targets of aberrant methylation in Koreans with CLL, which is consistent with data from Caucasians. We also revealed novel candidate CLL-associated genes (*UBC*, *GRB2*, *CREBBP*, and *GAB2*) that closely interact with *TP53* and *ZAP70*, implying the existence of differences between CLLs afflicting Asians versus Caucasians.

Methods

Study populations. Eight ethnically Korean patients diagnosed with CLL without *IGHV* mutations between May 2008 and July 2014 at Hallym University Sacred Hospital, Republic of Korea, were enrolled. CLL was diagnosed based on the World Health Organisation[59,60] and 2008 International Workshop on Chronic Lymphocytic Leukemia-National Cancer Institute criteria[61]. Collected laboratory data included complete blood counts, bone marrow pathology, immunophenotyping, conventional karyotyping, and *IGHV* somatic hypermutation status. Five age-matched, voluntary donors were examined as healthy controls. The study was performed according to the guidelines of the Declaration of Helsinki and was approved by the Ethics Committee of Hallym University (No. HALLYM 2019-01-004-002). All subjects provided written informed consent to participate in this study.

MBD-seq library preparation and sequencing. Bone marrow buffy coats were collected from the patients; the median lymphoid cell percentage was 85.75% (range, 41.60–99.00%). CD19-positive B cells were collected from five healthy donors using magnetic bead sorting (EasySep^{TM}; STEMCELL Technologies, Inc., Vancouver, Canada). Purity was confirmed using flow cytometry analysis (>95.0%). Genomic DNA was isolated using the Promega Maxwell® 16 MDx Instrument. Genomic DNA (1 µg) was sheared to 200–400 bp using a Covaris LE220 sonicator; the fragments were then subject to methyl-CpG enrichment using the Invitrogen MethylMiner Methylated DNA Enrichment Kit, which uses a recombinant form of human MBD protein-2. The enriched methylated DNA fragments were eluted as a single enriched population with a 2,000 mM NaCl elution buffer. The eluted DNA was then used to generate libraries according to the standard Illumina protocol. Briefly, the DNA fragments were subject to end repair, A-tailing of the 3′ end, Illumina adapter ligation, size selection (aiming for 300–500 bp), PCR amplification, and validation using an Agilent Bioanalyzer. The libraries were then sequenced on the Illumina HiSeq 2000 platforms.

Pre-processing of sequencing data. FastQC was used to check the sequence quality of the 100 bp paired-end sequencing reads. Trimmomatic[62] was used to clean the reads by removing adapter sequences, bases from the ends of the reads with quality <3, sliding windows of four bases with a mean quality <15, and reads shorter than 36 bp. The cleaned reads were aligned to the hg38 human genome with Bowtie2[63] using default parameters. The resulting mapped data in BAM format served as an input to subsequent differential methylation analysis.

Differential methylation analysis. The BAM files were inputted into the MEDIPS program[64]. The data for chromosomes 1–22 and M were selected for analysis. Sex chromosomes were excluded to avoid potential biases arising from X chromosome inactivation in samples from women. The genome was binned into adjacent windows 250 nucleotides in length. Differential methylation analysis between the CLL and control groups was performed at the window level. The parameter for data normalization and differential methylation analysis was set to edgeR and that for multiple testing correction was set according to the Benjamini-Hochberg procedure; default values were used for all other parameters. Windows with false discovery rates <0.01 were deemed differentially methylated. The methylation data are available at the Gene Expression Omnibus under the accession number GSE136986. The methylation levels of the differentially methylated windows were used for hierarchical clustering analysis to examine the separability of the CLL and normal sample groups. The methylation level of each window, as measured in log-counts per million values obtained from the edgeR R package[65], was standardized using a z-transformation such that the row mean and variance were set to 0 and 1, respectively. Next, hierarchical clustering was performed in the EMA R package[66] using the average linkage method with Pearson correlation analysis as the similarity metric.

Annotation of the differentially methylated windows. The window-level differential methylation analysis result from MEDIPS was inputted into the ChIPseeker program[67]. For each window, ChIPseeker assigns a corresponding Entrez Gene and one or more of the following annotations: distal intergenic, promoter (2 kb upstream to 0.5 kb downstream of transcription start site), 5′ untranslated region, exon, intron, 3′ untranslated region, and downstream. Default values were used for all other parameters. After annotating the differentially methylated windows, the list of DMGs was obtained. A gene was deemed a DMG if at least one of the windows that corresponded to it was differentially methylated, although there was usually more than one; as such, the DMG's p-value and log2 fold change were set to the smallest window-level p-value and the average of window-level log2 fold change values, respectively. The DMGs were classified with respect to gene type by the scheme adopted by 'Ensembl' into four categories: protein-coding, long noncoding, short noncoding, and others (pseudogenes and unannotated).

GO and pathway analysis.

Functional enrichment analysis was performed using the clusterProfiler program[68] to identify prevalent biological themes in the DMG list using GO and KEGG pathway analyses, with significance set at a p-value < 0.05. For GO analysis, *minGSSize* = 20 and *maxGSSize* = 1,000 were applied. To compare the functional enrichment results between the directions of differential methylation and between genomic regions, a maximum of 30 GO terms or pathways were selected for each track, and their p-values were displayed side by side on a dot plot using the 'clusterProfiler' function. The KEGG[69] pathway map for the B-cell receptor signalling pathway was rendered by the Pathview Web[70].

Network-based prioritization of DMGs.

We examined how well the DMGs interact with known CLL-related genes (per the DisGeNET database)[71] in human protein-protein and protein-DNA interaction networks obtained using Cytoscape[72] and the associated applications BisoGenet[73] and ReactomeFIViz[74].

We first downloaded the list of CLL-related genes by searching the DisGeNET website using the keyword 'chronic lymphocytic leukemia'; each of the CLL-related genes is annotated with supporting publications. From the obtained list, we removed genes that were annotated with only one publication to rule out potential false positives. Next, we constructed a network that links DMGs and known CLL-related genes; from BisoGenet, we retrieved the protein-protein interactions from all available sources as well as the protein-DNA interactions from the Biomolecular Interaction Network Database. We retrieved all available interactions from ReactomeFIViz, and excluded those that were only predicted. Gene symbols were used to query the interactions between genes in both applications; the two resultant networks were combined using Cytoscape's 'Merge' function, and duplicate edges were then removed to yield a consolidated human interaction network. Node degree was obtained using Cytoscape's 'NetworkAnalyzer' function[75].

Data availability

The methylation data are available at the Gene Expression Omnibus under the accession number GSE136986.

Received: 26 August 2019; Accepted: 6 January 2020;
Published online: 23 January 2020

References

1. Swerdlow, S. H. *et al*. Chronic lymphocytic leukaemia/small lymphocytic lymphoma. In *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* (eds. Campo, E. *et al*.) 216–219 (IARC, 2017).
2. Kim, J. A. *et al*. Genomic Profile of Chronic Lymphocytic Leukemia in Korea Identified by Targeted Sequencing. *PLoS One*. **11**, e0167641 (2016).
3. Jang, M. A. *et al*. Chronic lymphocytic leukemia in Korean patients: frequent atypical immunophenotype and relatively aggressive clinical behavior. *Int. J. Hematol*. **97**, 403–8 (2016).
4. Xia, Y. *et al*. Frequencies of SF3B1, NOTCH1, MYD88, BIRC3 and IGHV mutations and TP53 disruptions in Chinese with chronic lymphocytic leukemia: disparities with Europeans. *Oncotarget*. **6**, 5426–34 (2015).
5. Chen, S. S. *et al*. Silencing of the inhibitor of DNA binding protein 4 (ID4) contributes to the pathogenesis of mouse and human CLL. *Blood*. **117**, 862–71 (2011).
6. Corcoran, M. *et al*. ZAP-70 methylation status is associated with ZAP-70 expression status in chronic lymphocytic leukemia. *Haematologica*. **90**, 1078–88 (2005).
7. Raval, A. *et al*. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell*. **129**, 879–90 (2007).
8. Seeliger, B., Wilop, S., Osieka, R., Galm, O. & Jost, E. CpG island methylation patterns in chronic lymphocytic leukemia. *Leuk. Lymphoma*. **50**, 419–26 (2009).
9. Kanduri, M. *et al*. Differential genome-wide array-based methylation profiles in prognostic subsets of chronic lymphocytic leukemia. *Blood*. **115**, 296–305 (2010).
10. Cahill, N. *et al*. 450K-array analysis of chronic lymphocytic leukemia cells reveals global DNA methylation to be relatively stable over time and similar in resting and proliferative compartments. *Leukemia*. **27**, 150–8 (2013).
11. Kulis, M. *et al*. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet*. **44**, 1236–42 (2012).
12. Landau, D. A. *et al*. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*. **26**, 813–825 (2014).
13. Serre, D., Lee, B. H. & Ting, A. H. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res*. **38**, 391–9 (2010).
14. Aberg, K. A. *et al*. MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case–control samples. *Epigenomics*. **4**, 605–21 (2012).
15. Harris, R. A. *et al*. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol*. **28**, 1097–105 (2010).
16. Subhash, S., Andersson, P. O., Kosalai, S. T., Kanduri, C. & Kanduri, M. Global DNA methylation profiling reveals new insights into epigenetically deregulated protein coding and long noncoding RNAs in CLL. *Clin. Epigenetics*. **8**, 106 (2016).
17. Cahill, N. & Rosenquist, R. Uncovering the DNA methylome in chronic lymphocytic leukemia. *Epigenetics*. **8**, 138–48 (2013).
18. Pei, L. *et al*. Genome-wide DNA methylation analysis reveals novel epigenetic changes in chronic lymphocytic leukemia. *Epigenetics*. **7**, 567–78 (2012).
19. Kulis, M. *et al*. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet*. **47**, 746–56 (2015).
20. Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene*. **21**, 5400–5413 (2002).
21. Ogishima, T. *et al*. Increased heparanase expression is caused by promoter hypomethylation and upregulation of transcriptional factor early growth response-1 in human prostate cancer. *Clin. Cancer Res*. **11**, 1028–1036 (2005).
22. Gaudet, F. *et al*. Induction of tumors in mice by genomic hypomethylation. *Science*. **300**, 489–492 (2003).
23. Qu, G., Grundy, P. E., Narayan, A. & Ehrlich, M. Frequent hypomethylation in Wilms tumors of pericentromeric DNA in chromosomes 1 and 16. *Cancer Genet. Cytogenet*. **109**, 34–39 (1999).
24. Medvedeva, Y. A. *et al*. Intergenic, gene terminal, and intragenic CpG islands in the human genome. *BMC Genomics*. **11**, 48 (2010).
25. Baylin, S. B. & Herman, J. G. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends. Genet*. **16**, 168–74 (2000).
26. Esteller, M. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum. Mol. Genet*. **16**, R50–9 (2007).

27. Weber, B., Stresemann, C., Brueckner, B. & Lyko, F. Methylation of human microRNA genes in normal and neoplastic cells. *Cell Cycle.* **6**, 1001–5 (2007).
28. Ling, H. *et al.* Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene.* **34**, 5003–5011 (2015).
29. Yang, G., Lu, X. & Yuan, L. LncRNA: a link between RNA and cancer. *Biochim. Biophys. Acta.* **1839**, 1097–1109 (2014).
30. Nobili, L., Ronchetti, D., Taiana, E. & Neri, A. Long non-coding RNAs in B-cell malignancies: a comprehensive overview. *Oncotarget.* **8**, 60605–60623 (2017).
31. Sattari, A. *et al.* Upregulation of long noncoding RNA MIAT in aggressive form of chronic lymphocytic leukemias. *Oncotarget.* **7**, 54174–54182 (2016).
32. Balatti, V., Acunzo, M., Pekarky, Y. & Croce, C. M. Novel mechanisms of regulation of miRNAs in CLL. Trends. *Cancer.* **2**, 134–143 (2016).
33. Fabbri, G. *et al.* Common nonmutational NOTCH1 activation in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA* **114**, E2911–E2919 (2017).
34. Giménez, N. *et al.* Mutations in the RAS-BRAF-MAPK-ERK pathway define a specific subgroup of patients with adverse clinical features and provide new therapeutic options in chronic lymphocytic leukemia. *Haematologica.* **104**, 576–586 (2019).
35. Sharman, J. *et al.* An open-label phase 2 trial of entospletinib (GS-9973), a selective spleen tyrosine kinase inhibitor, in chronic lymphocytic leukemia. *Blood.* **125**, 2336–43 (2015).
36. Wolf, C. *et al.* NFATC1 activation by DNA hypomethylation in chronic lymphocytic leukemia correlates with clinical staging and can be inhibited by ibrutinib. *Int. J. Cancer.* **142**, 322–333 (2018).
37. Yepes, S., Torres, M. M. & Andrade, R. E. Clustering of Expression Data in Chronic Lymphocytic Leukemia Reveals New Molecular Subdivisions. *PLoS One.* **10**, e0137132 (2015).
38. Bhalla, S. *et al.* Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci. Rep.* **7**, 44997 (2017).
39. Tang, Y. *et al.* Downregulation of ubiquitin inhibits the proliferation and radioresistance of non-small cell lung cancer cells *in vitro* and *in vivo*. *Sci. Rep.* **5**, 9476 (2015).
40. Ijaz, M. *et al.* The Role of Grb2 in Cancer and Peptides as Grb2 Antagonists. *Protein. Pept. Lett.* **24**, 1084–1095 (2018).
41. Dixon, Z. A. *et al.* CREBBP knockdown enhances RAS/RAF/MEK/ERK signaling in Ras pathway mutated acute lymphoblastic leukemia but does not modulate chemotherapeutic response. *Haematologica.* **102**, 736–745 (2017).
42. Gao, C. *et al.* Low CREBBP expression is associated with adverse long-term outcomes in paediatric acute lymphoblastic leukaemia. *Eur. J. Haematol.* **99**, 150–159 (2017).
43. Hashwah, H. *et al.* Inactivation of CREBBP expands the germinal center B cell compartment, down-regulates MHCII expression and promotes DLBCL growth. *Proc. Natl. Acad. Sci. USA* **114**, 9701–9706 (2017).
44. Jia, D. *et al.* Crebbp Loss Drives Small Cell Lung Cancer and Increases Sensitivity to HDAC Inhibition. *Cancer Discov.* **8**, 1422–1437 (2018).
45. Mullighan, C. G. *et al.* CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature.* **471**, 235–9 (2011).
46. Gu, H. & Neel, B. G. The "Gab" in signal transduction. *Trends Cell Biol.* **13**, 122–30 (2003).
47. Huang, E. *et al.* Gene expression predictors of breast cancer outcomes. *Lancet.* **361**, 1590–1596 (2003).
48. Chen, Y. *et al.* GAB2 promotes cell proliferation by activating the ERK signaling pathway in hepatocellular carcinoma. *Tumour Biol.* **37**, 11763–11773 (2016).
49. Coutts, A. S. & La Thangue, N. B. The p53 response: emerging levels of co-factor complexity. *Biochem. Biophys. Res. Commun.* **331**, 778–85 (2005).
50. Lev Maor, G., Yearim, A. & Ast, G. The alternative role of DNA methylation in splicing regulation. *Trends Genet.* **31**, 274–80 (2015).
51. Oakes, C. C. *et al.* DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–64 (2016).
52. Beekman, R. *et al.* The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).
53. Kanduri, M. *et al.* Distinct transcriptional control in major immunogenetic subsets of chronic lymphocytic leukemia exhibiting subset-biased global DNA methylation profiles. *Epigenetics.* **7**, 1435–42 (2012).
54. Ronchetti, D. *et al.* Distinct patterns of global promoter methylation in early stage chronic lymphocytic leukemia. *Genes Chromosomes Cancer.* **53**, 264–73 (2014).
55. Hovestadt, V. *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature.* **510**, 537–41 (2014).
56. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14**, R21 (2013).
57. Yamasaki, S. *et al.* Docking protein Gab2 is phosphorylated by ZAP-70 and negatively regulates T cell receptor signaling by recruitment of inhibitory molecules. *J. Biol. Chem.* **276**, 45175–83 (2001).
58. Wilson, A. S., Power, B. E. & Molloy, P. L. DNA hypomethylation and human diseases. *Biochim. Biophys. Acta.* **1775**, 138–62 (2007).
59. Jaffe, E. S., Harris, N. L., Stein, H. & Vardiman, J. W. Chronic lymphocytic leukaemia/small lymphocytic lymphoma. In *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* (eds. Muller-Hermelink, H. K., Montserrat, E., Catovsky, D. & Harris, N. L.) 127–130 (IARC, 2001).
60. Swerdlow, S. H. *et al.* Chronic lymphocytic leukaemia/small lymphocytic lymphoma. In *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* (eds. Muller-Hermelink, H.K. *et al.*) 180–182 (IARC, 2008).
61. Hallek, M. *et al.* Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood.* **111**, 5446–56 (2008).
62. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**, 2114–20 (2014).
63. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
64. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics.* **30**, 284–6 (2014).
65. Robinson, M. D., McCarthy, D. J & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* **26**, 139–40 (2010).
66. Servant, N. *et al.* EMA - A R package for Easy Microarray data analysis. *BMC Res. Notes.* **3**, 277 (2010).
67. Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics.* **31**, 2382–3 (2015).
68. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* **16**, 284–7 (2012).
69. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
70. Luo, W., Pant, G., Bhavnasi, Y. K., Blanchard, S. G. Jr. & Brouwer, C. Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Res.* **45**, W501–W508 (2017).

71. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
72. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
73. Martin, A. *et al.* BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics.* **11**, 91 (2010).
74. Wu, G., Dawson, E., Duong, A., Haw, R. & Stein, L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *Version* **2**, F1000Res (2014).
75. Doncheva, N. T., Assenov, Y., Domingues, F. S. & Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **7**, 670–85 (2012).

Acknowledgements

We appreciate the assistance from the KOBIC Research Support Program. S. Hwang was supported by a grant from the KRIBB Research Initiative Program.

Author contributions

M.K. prepared the manuscript. E.L. collected the patients' laboratory data. D.-Y.Z., H.-J.K., H.-Y.K. and B.H. collected the patients' clinical data. H.-S.K. and H.J.K. provided expert opinions. S.H. designed and performed the bioinformatics analysis and supervised the interpretation. Y.-K.L. diagnosed the patients and supervised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-57919-6.

Correspondence and requests for materials should be addressed to S.H. or Y.K.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.