Computational and Structural Biotechnology Journal 20 (2022) 1726–1742

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Integrating whole genome sequencing, methylation, gene expression, topological associated domain information in regulatory mutation prediction: A study of follicular lymphoma

Amna Farooq [a], Gunhild Trøen [a], Jan Delabie [c], Junbai Wang [b,d,*]

[a] Department of Pathology, Oslo University Hospital - Norwegian Radium Hospital, Oslo, Norway
[b] Department of Clinical Molecular Biology, Institute of Clinical Medicine, University of Oslo, Norway
[c] Laboratory Medicine Program, University Health Network and University of Toronto, Toronto, Ontario, Canada
[d] Department of Clinical Molecular Biology (EpiGen), Akershus University Hospital, Lørenskog, Norway

ARTICLE INFO

Article history:
Received 7 January 2022
Received in revised form 22 March 2022
Accepted 22 March 2022
Available online 23 March 2022

Keywords:
Genome
Epigenome
3D chromatin domain
Regulatory mutation
Cancer
Machine learning
Integrative data analysis

ABSTRACT

A major challenge in human genetics is of the analysis of the interplay between genetic and epigenetic factors in a multifactorial disease like cancer. Here, a novel methodology is proposed to investigate genome-wide regulatory mechanisms in cancer, as studied with the example of follicular Lymphoma (FL). In a first phase, a new machine-learning method is designed to identify Differentially Methylated Regions (DMRs) by computing six attributes. In a second phase, an integrative data analysis method is developed to study regulatory mutations in FL, by considering differential methylation information together with DNA sequence variation, differential gene expression, 3D organization of genome (e.g., topologically associated domains), and enriched biological pathways. Resulting mutation block-gene pairs are further ranked to find out the significant ones. By this approach, BCL2 and BCL6 were identified as top-ranking FL-related genes with several mutation blocks and DMRs acting on their regulatory regions. Two additional genes, CDCA4 and CTSO, were also found in top rank with significant DNA sequence variation and differential methylation in neighboring areas, pointing towards their potential use as biomarkers for FL. This work combines both genomic and epigenomic information to investigate genome-wide gene regulatory mechanisms in cancer and contribute to devising novel treatment strategies.

1. Introduction

Carcinogenesis involves epigenomic, genomic, transcriptomic and proteomic changes. The integration of these changes is important for gaining a better insight into cancer molecular biology and may lead to improved diagnosis and finding novel strategies for cure. Epigenetic alterations such as DNA methylation and chromatin modifications are interlinked [1]. DNA hypomethylation or loss of DNA methylation on CpG dinucleotides was the first epigenetic anomaly to be recognized in cancer cells [2]. DNA methylation represses gene expression. Hypomethylation of DNA by contrast, can induce expression of genes, including oncogenes. Further, it can result in activation of transposable elements and loss of genomic imprinting. Hypermethylation of DNA is also a common feature of cancer. DNA hypermethylation is often seen at promoter regions of tumor suppressor genes, inducing their epigenetic silencing. Key gatekeeper genes like cyclin-dependent kinase inhibitor 2A(CDKN2A) and BRCA1, for example, are silenced in this way [3,4].

Gene expression profiles of tumor cells have been useful in classification, prognostication and prediction of multiple types of cancers such as breast, colorectal and lung cancer etc. [5–7]. It has been demonstrated in multiple studies that differential expression of certain gene sets is linked to cancer progression [8]. This has led to development of gene signatures to predict prognosis of cancer [9]. Gene expression profiling can not only predict clinical outcome

Abbreviations: differentially methylated region, DMR; topologically associated domain, TAD; follicular lymphoma, FL; single nucleotide variation, SNV; differentially expressed gene, DEG; principal component analysis, PCA; T-distributed stochastic neighbor embedding, t-SNE; group mean difference, GMD.

* Corresponding author at: Institute of Clinical Medicine, University of Oslo, Norway.
E-mail address: junbai.wang@medisin.uio.no (J. Wang).

https://doi.org/10.1016/j.csbj.2022.03.023

but also be used to select optimal personalized therapy. Nonetheless, expression profiles of cancer can be highly variable limiting the use of expression profiling in clinical practice. [10]. Incorporation of DNA sequence variations like Single Nucleotide Variations (SNVs) is therefore warranted, as they are typically present in cancer and are stably detected [11]. Multiple cancer genetic studies have reported SNVs that can contribute to cell transformation as gain-of-function or loss-of-function, for example by activating an oncogene or reducing the expression of a tumor suppressor [12]. However, it remains difficult to associate any SNV functionally to a gene, and consequently to cancer because of the regulatory complexity of higher order genomes. Due to presence of long-distance regulatory elements, physical proximity of any SNV to an oncogene or tumor suppressor gene is not sufficient evidence for it to be labeled as cancer driver SNV [13]. Chromatin architecture also comes into play here. The majority of the long-range chromatin interactions happen within and are limited by Topologically Associated Domains (TAD) boundaries. Coupling SNVs with the genes present in a similar TAD can give us an indication about the functional relevance of a SNV, and thereby its contribution as a cancer driver mutation [14,15].

Follicular lymphoma (FL) is a recurring lymphoma for which chromosomal translocation was identified [16]. However, this translocation alone is not sufficient to cause FL. Recurrent mutations have also been reported in FL in multiple developmental, signaling pathway and chromatin regulator genes [17]. In addition, epigenetic alterations have been observed in FL [18]. DNA methylation of tumor suppressor genes have been reported in FL [19]. DNA hypermethylation may also cause transcriptional repression of functionally important genes in FL [20].A gene-expression profiling study predicted the risk of progression of patients with follicular lymphoma using a 23-gene score [21]. Of note, the most previous research conducted in FL does not comprehensively address all genetic changes as explained earlier. For instance, an integrated data analysis pipeline [22] was developed previously to identify putative functional regulatory mutations in FL by considering both the gene expression profiles and the clustered distribution of SNVs. However, it was only able to predict regulatory mutations near the promoter region of genes. To have a comprehensive picture of oncogenesis of FL, we are motivated to design a new integrated data analysis method that takes into account epigenomic (e.g., DNA methylation), genomic (e.g., chromatin architecture- TAD), and transcriptomic (e.g., gene expression) information together with the distribution of genome wide SNVs in patients. In this way, our method makes it possible to predict functional regulatory mutations that affect gene regulation through a long-distance. This will be a great leap forward for the investigation of non-coding mutations in cancer or disease, by utilizing genome wide sequence technology in clinical studies.

Publicly available biological data sets allow comprehensive analysis of data. However, efficient statistical and bioinformatic methods for such integrative analysis are lacking. A few tools like sTRAP and is-rSNP predict functional non-coding variants by hypothesizing that non-coding variants can affect gene expression by altering protein-DNA binding [23,24]. Another method searched for SNP combinations for disease on the basis of the energy distribution difference considering an individual's genotype data as a point with a unit of energy [25]. CADD and FunSeq2 can integrate some data types like predicted transcription factor binding sites, measured ChIP-Seq peaks of TFs, chromatin state marks, conservation scores and protein–protein interactions [26,27]. Another study integrates DNA methylation, gene expression and somatic mutations to infer tissue-of-origin of a tumor [28]. In short, there are multiple integrative studies conducted on cancer, but none integrates all the genetic and epigentic changes (i.e. DNA methylation, gene expression, DNA sequence variation and Topologically associated domains) in cancer.[29–32]. Moreover, there is less focus on using differential methylation for functional annotation of SNVs. Previous studies have reported a correlation between the regional methylation level and the rate of mutation at CpG sites in genomic regions [33]. Especially in the presence of allele-specific methylation, mutations in the driver genes can be inherently connected with the aberrant DNA methylation landscape in cancer [34–36]. This points towards the potential use of differential methylation to gauge the functional relevance of a SNV.

Although many studies have highlighted the link of the DNA methylation and SNVs, there is no method to our knowledge which employs DNA methylation data to identify disease-related SNVs. While some methods for differential methylation detection are already available, it is important to have a score or rank explaining the magnitude of the differential methylation. In particular, if we want to profit from differentially methylated regions (DMRs) in an integrative study. To overcome this limitation, we have first devised a new method for significant differential methylation detection, which was used to analyze DMRs in FL through parallel computation of six attributes (Fig. 1). Some of the computed attributes are used to identify high confidence DMRS (hcDMRs). HcDMRs then serve as a standard to rank the remaining methylated regions. The resulting model reports a set of DMRs with respective scores depicting the significance of a particular DMR.

Acknowledging the role of differential methylation and chromatin architecture at the epigenetic level and DNA sequence variations like SNVs and gene expression profiles in cancer, we broaden the scope of our study by using a newly developed integrative data analysis method to investigate regulatory mutations in FL at a genome wide level. First, genomic blocks having a high SNV concentration were identified (we will address them as mutation blocks), and mapped to the DMRs and differentially expressed genes (DEG) that are present in the same TAD. Then, the frequency of occurrence of mutation blocks and its annotation to the genomic elements, differential expression level of the associated genes and the related DMR score are used as features to identify significant mutation block-gene pairs in FL. From the analysis, a final set of genes associated to mutation blocks is obtained which is further evaluated by a robustness analysis, based on an independent source (e.g., chromatin state segmentations) that was not used in the prediction. Mutation block-gene associations related to FL that passed robustness analysis with high statistical and biological support, are reported in this study. These set of mutation block-genes can be seen as cancer drivers. Similarly, the set of hcDMRs can have a strong potential of being used as biomarkers for diagnosis, prognosis, prediction and potential treatment of FL. Our study presents a robust method for DMR detection and integrative genomic analysis of regulatory mutations that can be applied to any malignancy.

2. Material and methods

2.1. Diverse high throughput sequencing data for follicular lymphoma patients

Genome-wide sequencing data of 14 tumor-normal paired FL patients was obtained from a previous study [37], by getting access to controlled data kept on ICGC. Samples were downloaded from European Genome-phenome Archive [38] (https://www.ebi.ac.uk/ega/) under accession numbers EGAD00001000645 and EGAD00001000355. RNA-Seq data of four control samples (Germinal center B-cell - GCB) from healthy people was downloaded from GEO database under accession number GSE4598265 [39]. DNA methylation data of whole-genome bisulfite sequencing (WGBS) for 8 FL patients and 4 GCB control samples was acquired from an earlier work [40]. For data sets not available specifically for

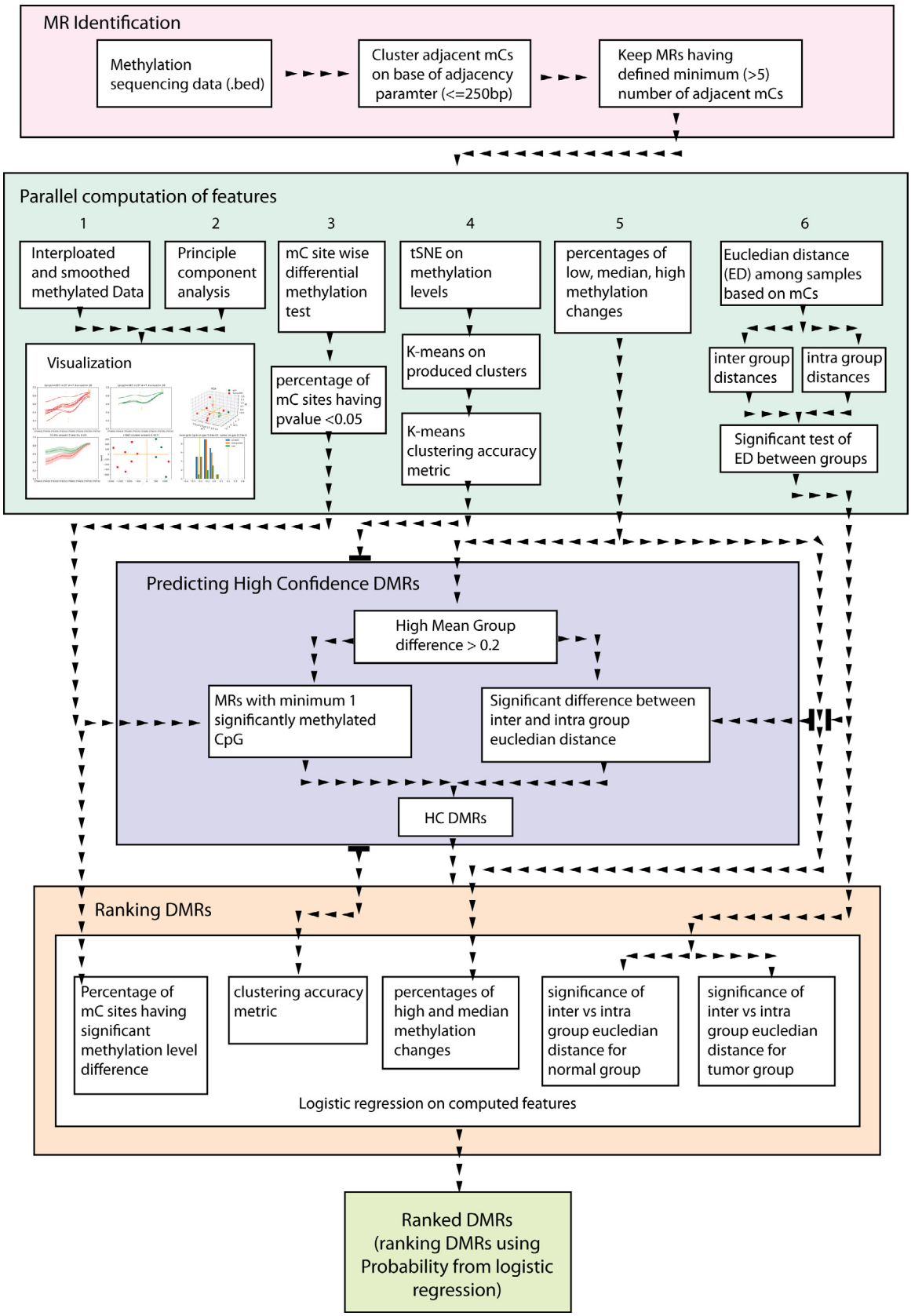


Fig. 1. Work flow for ranking differentially methylated regions between two groups by a new machine learning approach. This figure describes a new machine learning approach for predicting and ranking high quality differentially methylated regions (DMRs) with four steps: 1) search for methylated region (MR) in a genome-wide manner based on predefined criterias, 2) parallel computation of six attributes in each MR, 3) four of the attributes (e.g., the percentage of differentially methylated CpG methylation sites, the clustering accuracy of predicted sample group label based on 2-D t-SNE map, the percentage of the high and median methylation level changes, and the significance of Euclidean distance difference between the intra-group and the inter-group) are used to identify highly confidence DMRs (hcDMRs), 4) fits a logistic regression model for all available MRs by using hcDMRs as true targets. Probability value of each MR (from logistic regression model) is used to rank the DMRs per their significance.

FL, we chose data sets closest possible to FL. Human common Topologically Associating Domains (TAD) and boundaries information was downloaded from supplementary Tables 1 and 4 of [41]. The common boundaries are from five human cell lines that represent three distinct embryonic germ layers (GM12878 and HMEC, mesoderm; IMR90, endoderm; HUVEC and NHEK, ectoderm). Those TAD boundaries were reported in the original study as significantly (83–85% with pval $\langle 10^{-7}$) conserved between normal and malignant cells and were thus used in this study. Human enhancer annotations of 197 tissue/cell types were download from EnhancerAtlas 2.0 [42]. Annotations from 5 out of 197 cell lines (DOHH2, GC B cell, Namalwa, OCI-ly1 and OCI-Ly7) were later grouped as FL-related cell lines [43]. Information of KEGG, BIOCARTA, and GO Biology Process pathways was retrieved from DAVID functional annotation tool [44]. Here, all sequencing data were aligned to hs37D5, a variant of GRCh37 human genome assembly used by the 1000 Genomes project [45]. Genome-wide mutations were called by using Strelka [46] and MuTect [47] with the default parameters. An intersection of mutation calls from both programs was used in the further data analysis for each patient [48]. From mutations called by these programs, we only considered SNVs. For identifying the transcripts of all protein-coding genes, we used gene annotation from the UCSC hg19 [49]. Annotation to the reference genome were performed on four defined genomic regions: the TSS/TES regions between −5kb and + 1kp to the TSS (transcription start site)/TES (transcription end site) of protein-coding genes, and the gene body region between TSS and TES, and the 5′distance regions were calculated from 1 Mb to 5 kb upstream of the TSS. These four defined regions are similar to the previous publications [50,51] in differential methylation analysis and identification of promoter-distal loops. Gene expression levels are measured as reads per kilobase of transcript per million mapped reads (RPKM) of RNA-Seq experiments and were computed by applying the featureCounts [52] and our in-house Python code on aligned BAM files. In this study, the genetic (SNVs), epigenetic (DNA methylation), and transcriptomic (gene expression) data are from the same FL patient cohort [40]. A brief description of these FL samples and the procedures for obtaining them are provided in supplementary Stable 1. The tumor cell content in the cryopreserved sample material was at least 60% in all cases [40]. More information of these FL samples and the basic characterization including histopathological panel review and immunohistochemical and FISH analyses can be seen in the previous publication [37].

2.2. Segmentation of human genome in functional regions based on chromatin features

Chromatin modifications are important epigenic makers in genome, which can be used to characterize functional regions (e.g., enhancer and TSS et al). We obtained predicted segmentation of human genome based on chromatin features from an earlier publication [53]. This segmentation of human genome is based on the predictions from two machine-learning methods (ChromHMM [54] and Segway [55]), by using multiple chromatin marks (e.g., H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, and H3K20me) known to be involved in enhancer, repression, or promoter regions across six human cell-lines (e.g., GM12878, H1 hESC, HeLa-S3, HepG2, HUVEC, and K562). In addition to chromatin marks, other genomic marks such as Pol2, CTCF, and nucleosome density were also considered. The final combined segmentation from the two predictions uses only seven chromatin states to segment the human genome in functional regions (e.g., TSS – predicted promoter region including TSS; PF – predicted promoter flanking region; E – predicted enhancer; WE – predicted weak enhancer or open chromatin cis regulatory element; CTCF – CTCF enriched element; T – predicted transcribed region; R – predicted repressed or low activity region).

2.3. Identifying methylation regions

There are a few popular methods available for differential methylation analysis, but with their own limitations. For example, MethylKit predicts differential methylation at a single base pair resolution, neglecting the confounding effects of any neighboring methylated site [56]. This limitation is addressed by HMST-Seq-Analyzer by predicting DMRs instead of differentially methylated sites [51]. However, HMST-Seq-Analyzer can only make one-to-one or one-to-many comparisons and cannot make group comparisons (many-to-many). Therefore, a new method for rigorous differential methylation detection is developed. It takes into account the confounding effect of neighboring methylated sites and can perform group comparisons with multiple samples in each group. As a foremost step, it tries to search for Methylated Regions (MR) in the genome. The selection criteria for MRs are that there should be a minimum number of CpG methylation sites in a MR (default of minimum 5), and any two neighboring CpGs must not be any further than a specified distance (by default 250 bp). Gathering MRs from genome-wide data can produce hundreds and thousands of MRs depending upon the type of sequencing method used for methylation detection. A brute force attempt for differential methylation analysis can be computationally exhaustive at the genome-wide scale. Therefore, once all MRs passing aforementioned filtering conditions are acquired from the sequencing data, a parallel computational algorithm is used to assess the significance of differential methylation between the two groups in each MR. The total number of observed MRs is equally divided on the available computer processers (e.g., 10 or 20 processes) and is ran in parallel, which significantly speeds up the calculation. Lastly, a rigorous set of DMRs is reported with their scores which can be used as input to any further integrative study.

2.4. Parallel computation of six attributes in each methylated region

Six major attributes are computed and evaluated at the MR level in order to identify Differentially Methylated Region (DMR): 1) Interpolated and smoothed data curves of original methylation levels are computed for both tumor and control/normal groups, and a 95% confidence interval from the group mean is graphically illustrated. 2) Based on the smoothed methylation curves of all samples, the first three principal component elements of PCA (Principal Component Analysis - a linear dimensionality reduction method) are calculated and visualized in a 3D plot (e.g., samples are colored by their group label), from which the difference among samples due to methylation variation in a MR is revealed. 3) As a third attribute, a two sampled T-test is performed to evaluate the significance of differential methylation at each CpG site, respectively. The percentage of CpG methylation sites in a MR that reaches a predefined significant level (e.g., T-test P value < 0.05) is recorded. The higher the percentage the better the differentially methylated region. 4) As a fourth attribute, T-distributed Stochastic Neighbor Embedding (t-SNE, a non-linear dimensionality reduction method) is applied on the smoothed methylation profiles of each MR, and the corresponding 2-D plot of samples is generated (e.g., samples are colored by their group label). Subsequently, k-means clustering is applied on this 2-D t-SNE map, to estimate the clustering accuracy by comparing the predicted sample group label against the true sample group label. The higher the clustering accuracy the more significant the DMR. 5) The fifth attribute is categorizing the levels of differential methylation changes (or absolute group mean difference - GMD) at each CpG site of a MR. As a default, three levels are defined i.e., 0.07 < GMD<=0.1, 0.1 < GM

D<=0.2, and GMD > 0.2 for low, median and high methylation changes, respectively, between tumor and control group. The percentages of CpG sites in a MR with low, median, and high methylation changes are recorded. The higher the percentage of high methylation changes the more significant the DMR will be. 6) Finally, Euclidean distances for samples within a group (e.g., intra-group in either tumor or normal group) and between groups (e.g., inter-group such as between tumor and normal group) are calculated based on the methylation profiles, respectively. A two sampled T-test is used to evaluate the significance of the difference between the intra-group distance and the inter-group distance (e.g., Euclidean distances of samples within a tumor group vs. Euclidean distance of samples between tumor and normal). The corresponding P-values are recorded for each MR. The underlying hypothesis for computing this attribute is that the intra-group distance is usually significantly different from inter-group distance for a DMR. Thus, the more significant the difference of distance (between the intra and the inter groups), the more significant the DMR. Here, the first two attributes (attribute 1 and 2) are computed for the purpose of visualization and giving an impression of the distribution of data at a broader level. The rest of the four attributes and the recorded summary statistics are directly used as features for identifying high confidence DMRs (hcDMRs).

2.5. Predicting high confidence differentially methylated regions by a two-levels approach

After completing the parallel computation of six major attributes in all the identified MRs, a two-level filtering approach is adopted to determine a set of hcDMRs. The first filter is applied on the attribute 5 i.e., the percentage of high methylation changes (e.g., mean group difference > 0.2) between two groups should be greater than zero. Then a second level filter is applied, where a putative hcDMR should meet at least one of the two conditions: either the percentage of CpG sites in a MR that are significantly different between two groups (e.g., T-test P value < 0.05 in attribute 3) is greater than a threshold value (e.g., default > 0), or there should be a significant difference of Euclidean distances between the intra-group and the inter-group (T-test P value < 0.05 in attribute 6). After the two levels filtering, a set of putative hcDMRs are obtained. The strength of these DMRs is controlled by three key parameters such as 1) the percentage of high methylation changes (e.g., mean group changes > 0.2; default) between two group and 2) the percentage of methylation sites show differential methylation in a MR are greater than zero, and 3) a significant difference (P value < 0.05 in default) of Euclidean distances between the intra-group and the inter-group.

2.6. Ranking differentially methylated regions through logistic regression

Although hcDMRs are a solid set of DMRs, to improve the sensitivity of the method and to sort DMRs based on their significance, a ranking approach through logistic regression method is further introduced. Here, hcDMRs are used as true target sites in logistic regression to fit all the available MRs with pre-computed attributes. The four attributes used as regressors are: 1) the percentage of methylation sites in a MR that have passed significant level of differential methylation (from attribute 3), 2) the clustering accuracy for predicting group labels in a MR based on a 2-D projected t-SNE methylation profiles (from attribute 4), 3) the percentage of group mean methylation changes in the high and median level changes (from attribute 5) and 4) the log10 transformed P values of the significance of difference in Euclidean distance between the intra-group and the inter-group (from attribute 6). A probability value is assigned to each MR after fitting the logistic regression model to all MRs by using the high confidence DMRs as true target. These probability values can be used to sort and select the most significant DMRs. For example, if a probability of logistic regression model is P>=0.7, then around 90% of the hcDMRs from the initial two-step filtering (e.g., the mean group changes > 0.2 and, either the percentage of differentially methylated CpG sites is > 0 or there is a significant difference in Euclidean distance between the intra-group and the inter-group) will be included, endorsing that the initial two-steps filtration is powerful for identifying sturdy DMRs. Unlike other popular methods for differential methylation analysis, the length of DMRs predicted by the current method can range from dozens of bp to hundred thousand bp which is similar to actual behavior of DMRs. Moreover, only the distribution of methylation sites is considered in defining a MR and the same trend of methylation level changes in a MR is not forced. Therefore, three types of DMRs (hyper, hypo, and mix) are reported in the prediction. HyperDMRs indicate increase in methylation levels as compared to control/normal samples. HypoDMRs have decreased methylation levels as compared to control/normal samples. It can be misleading to assume that methylated regions, showing differential methylation, will either show increase or decrease of the methylation level. Some DMRs can show both increasing and decreasing levels of methylation at different sites within the region. The current method captures such a possibility as well and reports them as mixed DMRs. All DMRs can also be manually explored through the plots exported by our method. More information about both the identification of hcDMRs and the ranking of DMRs are shown in Fig. 1.

2.7. Integrating differential Methylation, differential gene expression and topologically associated domain information in regulatory mutation prediction

Based on the aforementioned new method for analyzing and ranking DMRs between two groups of samples, it is possible to integrate the differential methylation with differentially expressed genes (DEG) data in predicting functional non-coding mutation in disease. First, SNVs from whole-genome-sequencing (WGS) data for the disease are identified by using Strelka and Mutect. An intersection of the SNVs predicted by the both programs was used for further analysis to strengthen the evidence, as performed in a previous publication [57]. A region harboring multiple SNVs can have more regulatory potential as compared to a single SNV. Hence, genome-wide identification of mutation blocks in patient samples is done by using BayesPI-BAR2 [58]. We first identify mutation clusters and then group them into mutation blocks. Mutation cluster will be a genomic region having a certain number of consecutive SNVs present in any of the patient sample. Here, any regions having at least one SNV in any of the patient samples were selected and grouped into mutation clusters to keep low stringency. In case of more than one SNVs the distance between adjacent SNVs should be<30 bp to be included in the same cluster. The mutation clusters were further grouped into mutation blocks. Mutation block will be a group of mutation clusters, where any consecutively located clusters are not>500 bp apart from each other. Then, these mutation blocks were annotated to multiple genomic regions based on annotated HG19 reference genome (i.e., Gene, TSS, TES, 5′distance and enhancers). In this study, only mutation blocks that were either overlapping with a DMR or associated with a DEG (e.g., a mutation block located in a the TSS, TES, 5′distance region or gene body of the DEG) in FL patients are considered. However, mutations can impact the expression profiles of target genes from long range interaction as well. To ensure the confidence in relevance of relationship between a mutation block and its long-distance target gene (e.g., DEG), the search for mutation block-DEG pair in 5′distance region is confined within the same topologically associated

domains (TAD). The information of TAD that are common in five cell lines, is obtained from a recent paper [41].

Usually, mutations or SNVs are more influential if present in the regulatory region such as enhancers, as they may disrupt or create a binding site for a transcription factor. The identified mutation blocks are further mapped to enhancer regions provided by EnhancerAtlas2.0 [42]. Once all mutation blocks are linked to either DMR or DEG, the strength of genes associated (e.g., through gene, TSS, TES, or 5′distance) to relevant mutation blocks is inferred by a method similar to weighted voting systems [59], where four features are used in ranking: 1) the number of patients having affected mutation blocks in a gene, 2) the probability of logistic regression fitting for DMRs associated to a gene, 3) the absolute log10 transformed P-values of DEG, and 4) the annotated genomic region that is linked to a mutation block (e.g., a mutation block locates in TSS, Gene, Enhancer, TES, and 5′distance region will be assigned a weight 4, 4, 3, 2, and 1, respectively). An average of normalized four features' scores (min–max normalization) is being used to rank the strength of associations between a mutation block and a gene (e.g., normalized score spans from 0 to 1). Finally, mutation block-gene pairs with an average of normalized feature scores>0.5 are extracted, and subjected to GO and pathway enrichment analysis (P-value < 0.05) by using DAVID functional annotation tool [60]. In Fig. 2, a workflow or pipeline for such integrated analysis (DMR, DEG, TAD information, and DNA sequence variation) of mutation blocks in FL is presented.

2.8. A robustness analysis of mutation block-gene associations by evaluating seven chromatin states

Here, a new robustness analysis is developed to evaluate the top ranked mutation block-gene associations obtained from an average of normalized feature scores (e.g., 327 genes in Fig. 2). The new analysis is based on independent information i.e. segmentation of human genome to seven functional regions based on chromatin features [53] (e.g., active or repressive histone modifications and nucleosome density), which was not used in the prediction. The robustness analysis utilized a permutation test to assess the significance of associated mutation blocks in seven chromatin states (e.g., R, T, TSS, enhancer, CTCF, WE, PF), respectively. The segmentation of human genome in chromatin states was predicted by two machine learning methods ChromHMM and Segway. First, for mutation blocks associated to each predicted target gene, if there are N number of mutation blocks in a gene, then the percentage of blocks (the actual percentage) located in the seven chromatin states are calculated, respectively. Then, we randomly draw 10,000 times of N mutation blocks from all $\sim$ 66467 blocks of 14 FL patients that overlap with the seven chromatin states, where the N mutation blocks predicted in the first step are excluded. Subsequently, for each sampled N random mutation blocks, their percentage (the expected percentage) in the seven chromatin states is computed, respectively. In each of the chromatin states, if the expected percentage obtained from the randomly sampled mutation blocks is greater than the actual percentage in the first step, then the chromatin state is incremented by one. Finally, for each tested gene, an expected P-value of its associated mutation blocks located in one of the seven chromatin states is calculated (e.g., P-value = the total number of the expected percentages greater than the actual ones divided by the number of samplings such as 10000), respectively. A filtered list of top ranked mutation block-gene associations will be obtained by assuming that the functional regulatory mutation blocks are significantly enriched (e.g., expected P-value < 0.05) in either TSS or enhancer regions. Thus, a final top ranked (e.g., within the top 20) mutation block-gene associations will be more reliable, if they passed such robust analysis in multiple predictions based on the same data.

3. Results

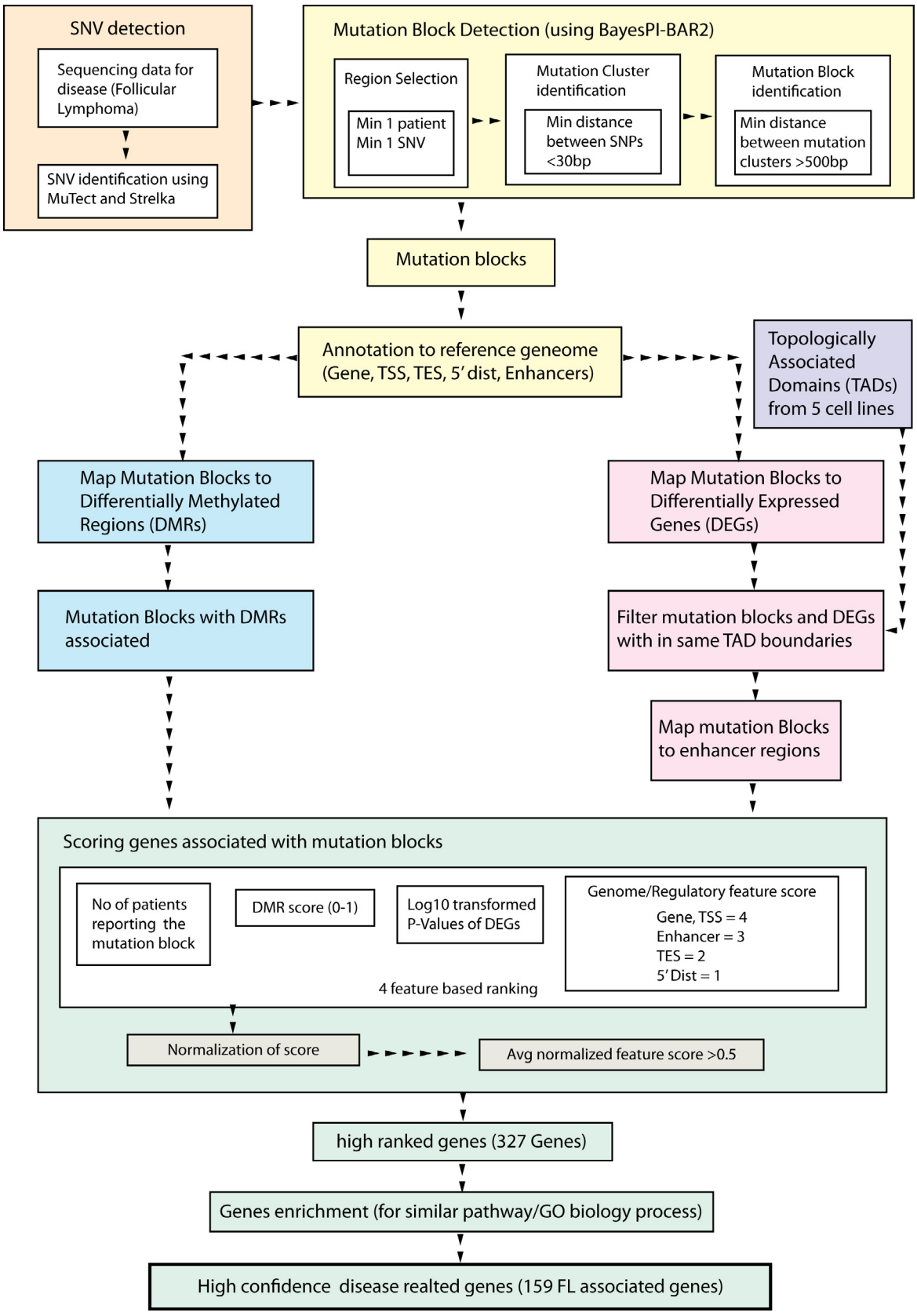
3.1. Predicting DMRs with a new machine learning method

To identify DMRs between FL patients and normal samples, a newly developed machine learning approach (default parameters) was applied on the WGBS data of 8 tumors and 4 control samples, where multiple samples from each group were analyzed simultaneously. A probability P > 0.7 of logistic regression model was used as a cutoff value for detecting DMRs between case (FL) and control (normal) samples. Total 275,949 DMRs were predicted. These DMRs were annotated to four defined genomic regions (TSS, TES, gene, and 5′distance), as well as to human enhancers obtained from EnhancerAtlas [42] by using intersection function of BEDTools [61]. Example of predicted DMRs is: mr37 in chromosome 1 (SFig. 1), having probability = 1 in the logistic regression model. A counter example of a region not considered DMR (mr60385 in chromosome three; probability = 5.582354e-16 in logistic regression model) is illustrated in SFig. 2. In SFig. 1, the predicted DMR is roughly 250 bp long with six attributes illustrated. A difference in the trend of smoothed methylation profiles between tumor and normal samples can be seen in the upper panel of SFig. 1. Upon calculating the PCA based on the smoothed methylation profiles, tumor samples clearly separate from the normal ones in a 3-D plot of the first three principal components. Nevertheless, it is crucial to see the shape of methylation level distribution in the lower panel of SFig. 1, where $\sim$ 53.8% of the methylated sites in the MR are differentially methylated (P-value < 0.05) between tumor and normal group. Clustering accuracy of K-means clustering for tumor and normal samples is high (equals $\sim$ 0.92), according to the two-dimensional t-SNE projections for all available methylation sites in the MR. Especially, the difference between intra-group and inter-group Euclidian distance is marginally significant for normal group (P-value < 0.06) but significant for tumor group (P-value < 0.0083), and the peak for mean group methylation changes is centered around -0.1. These six attributes in the SFig. 1 reiterate the significant differential methylation in mr37 on chromosome one.

3.2. Comparison of predicted DMRs between the new method and the HMST-seq-Analyzer

Though HMST-Seq-Analyzer is a tool to predict DMRs from data obtained from multiple methylation detection methods including

Fig. 2. Work flow for integrative analysis of regulatory mutations in follicular lymphoma by using whole genome sequencing, differential methylation, differential expression and TAD information. First, SNVs from the whole genome sequencing (WGS) data of 14 follicular lymphoma (FL) patients were used to identify genome-wide mutation blocks in FL by using BayesPI-BAR2. Then, the mutation blocks were annotated to four genomic regions (gene, TSS, TES, and 5′distance regions) as well as to enhancers collected from EnhancerAtlas 2.0. Subsequently, mutation blocks overlapping with either differentially methylated region (DMR) or differentailly expressed gene (DEG) are recorded. For a mutation block-DEG association that is linked by a 5′distance region of gene, the method requests that both the mutation block and the gene are located in the same topologically associated domain (TAD). The importance of mutation block-gene associations is ranked by four features: 1) the number of patients having affected mutation blocks, 2) the significance of a DMR, 3) the significance of the differentially expressed gene, and 4) a weighted score for annotated genomic region (e.g., gene, TSS, TES, 5′distance or enhancers) that a mutation block is mapped to. Finally, a set of top ranked mutation block-gene associations are extracted, where the selected genes are enriched in a common pathway or GO biology process.


SNV detection
Sequencing data for disease (Follicular Lymphoma)
SNV identification using MuTect and Strelka
Mutation Block Detection (using BayesPI-BAR2)
Region Selection
Min 1 patient
Min 1 SNV
Mutation Cluster identification
Min distance between SNPs <30bp
Mutation Block identification
Min distance between mutation clusters >500bp
Mutation blocks
Annotation to reference geneome (Gene, TSS, TES, 5′ dist, Enhancers)
Topologically Associated Domains (TADs) from 5 cell lines
Map Mutation Blocks to Differentially Methylated Regions (DMRs)
Map Mutation Blocks to Differentially Expressed Genes (DEGs)
Mutation Blocks with DMRs associated
Filter mutation blocks and DEGs with in same TAD boundaries
Map mutation Blocks to enhancer regions
Scoring genes associated with mutation blocks
No of patients reporting the mutation block
DMR score (0-1)
Log10 transformed P-Values of DEGs
Genome/Regulatory feature score
Gene, TSS = 4
Enhancer = 3
TES = 2
5′ Dist = 1
4 feature based ranking
Normalization of score
Avg normalized feature score >0.5
high ranked genes (327 Genes)
Genes enrichment (for similar pathway/GO biology process)
High confidence disease realted genes (159 FL associated genes)

Whole Genome Bisulfite Sequencing (WGBS) [51], in the case of group comparisons, it performs one-to-many comparison only. In this study, mean methylation levels of 4 normal GCB WGBS samples were used to compare to 8 FL patient data [40], respectively, by using the default parameters of HMST-seq-Analyzer. Only those DMRs (∼60322) predicted in all 8 samples (∼78–90% of each prediction) were used for further comparison to the new results (∼259963 with probability > 0.8 in logistic regression model of new machine learning approach). Around 92% (239325) of the DMRs from the new prediction method were completely overlapping with that from the HMST-Seq-Analyzer. If the probability cutoff of logistic regression model is varied between 0.5 and 0.7, the overlap between the two results remains ∼ 91 to 92%. It is worthy to note that HMST-Seq-Analyzer generates much longer DMRs than the new method, which explains why more DMRs are predicted by the new method than the HMST-Seq-Analyzer while maintaining a high percentage of overlap between the both. Thus, DMRs detected by the new method are robust and their shorter length makes their annotation and integrated data analysis further easier.

3.3. Including differential methylation and differential gene expression in regulatory mutation analysis

There are 118867 and 81812 single nucleotide variants (SNVs) in 14 FL patients, called by MuTect and Strelka, respectively. About 71235 SNVs (∼87% overlap) were detected by both methods. They were selected to identify genome-wide mutation blocks in FL by using BayesPI-BAR2 [58,62]. The result spans to 66868 mutation blocks, which requests minimum one patient and one SNV in a cluster. The SNP cluster and block distance was kept as 30 bp and 500 bp, respectively. Here, a built-in mutation background model from BayesPI-BAR2 was not applied to select highly mutated blocks (i.e., SNVs from multiple patients are located in the same mutation block). Instead, a putative functional mutation block was selected based on a different criterion: either it overlaps with a DMR or triggers a nearby gene activity (e.g., DEG). This assumes that gene regulation or TF binding and DNA methylation often affect each other [63,64]. For example, impact of SNV on TF binding may cause a variation of DNA methylation levels in neighboring regions or a dysregulation of gene expressions in the nearby location. Following this assumption, ∼13143 mutation blocks were found overlapping with the predicted DMRs. After considering genes that were associated to these DMRs through a gene body, TSS, TES, or 5′distance regions, and mutation blocks that are overlapping with either DMR (∼4603 mutation blocks) or their associated genes are differentially expressed (∼1831 DEG; $P < 0.05$), previously published mutation block-gene associations in FL are recovered in this initial analysis: for example, three known regulatory mutation blocks near the promoters of dysregulated BCL6 and BCL2 genes [57] (up and down regulated in FL compared to normal, respectively). This is a result that supports the hypothesis that functional regulatory mutation may affect DNA methylation level and/or gene expression activity in the nearby region.

3.4. Ranking mutation block-gene association by considering diverse information

Though it is possible to narrow down the number of mutation block-gene association by considering both DMR and DEG information, it is a challenge to evaluate their significance. The problem is further complicated by the fact that a mutation block may be assigned to multiple 5′distance regions of different genes (e.g., from 1 Mb to 5 kb upstream of the TSS). Thus, an additional evaluation of mutation block-DEG associations through 5′distance regions is needed: for example, if a mutation block and a gene are not located in the same TAD, then their association through 5′distance region will be ignored. In this study, information of common TADs/boundaries in five cell lines (GM12878, HUVEC, IMR90, HMEC and NHEK) were obtained from a previous publication [41]. After thus filtering mutation block-gene pairs, ∼1453 differentially expressed genes (DEG in TSS, TES, gene, or 5′distance) are associated with ∼ 5756 mutation blocks (e.g., either overlapping with DMRs or not). Among these mutation blocks, ∼3684 of them are located in enhancer regions [42]. To rank the significance of these inferred mutation block-gene associations, a weighted vote approach was used to rank them by integrating normalized four feature scores (e.g., the number of patients affected by mutation blocks, DMR significance, P-value to DEG, and the weighted genomic feature for a mutation block; Fig. 2).

In this work, an average of normalized feature scores was used to export the top ranked mutation block-genes associations: for example, with a mean feature score>=0.5, ∼327 genes are selected. Among these the BCL2 gene is ranked at the top. The number of mutation blocks located in enhancers/TES/TSS/Gene remains stable (∼76%) when a mean feature score cutoff value is decreased (e.g., <0.4). These 327 top ranked genes were used in further pathway analysis because functional mutations often influence multiple genes in the same pathway or biological process [65,66]. Thus, GO enrichment analysis is applied on these top ranked genes by using DAVID functional annotation tool [60]. The enriched GO biological process and KEGG/BIOCARTA pathways (e.g., $P < 0.05$) are extracted for defining a final list of mutation block-gene associations, which includes 159 genes involved in several important signaling pathways and biological processes related to FL. For example, intrinsic apoptotic signaling pathway in response to DNA damage, T cell receptor signaling pathway, B cell receptor signaling pathway, NF-kappa B signaling pathway, transcriptional misregulation in cancer, and immune response etc. The aforementioned enriched pathways are affected by mutation blocks from at least 13 FL patients (data in supplementary website). Notably, our previously predicted putative functional regulatory mutation blocks near BCL2 and BCL6 genes [57] are ranked in the top 10 (e.g., ranked 1 and 9 for BCL2 and BCL6, respectively; supplementary Stable 2) by this new analysis. Additionally, several novel mutation block-gene associations in FL are also identified (e.g., mutation block associated to CTSO and CDCA4 are ranked at top 2 and 5, respectively; supplementary Stable 2). The presence of numerous mutation blocks and DMRs in the vicinity of these genes urges also to investigate the role of long-range interaction in their regulation. Coming sections will discuss the genes and the respective distal elements in detail.

3.5. Hypomethylation of mutation blocks and enhancers in the BCL2 promoter region can contribute to BCL2 overexpression in follicular lymphoma.

The BCL2 gene is located at chromosome 18q21 and codes for BCL-2 protein which inhibits apoptosis and is important for normal B-cell development and differentiation. Follicular lymphoma shows the t(14;18) chromosomal translocation. This translocation involves BCL2 and causes its overexpression, thus providing survival advantage to the malignant B-cells [67]. A previous study aimed at predicted two mutation blocks in TSS region of BCL2 [57]. The same two mutation blocks of 3113 bp and 726 bp (block_66303, block_66304 respectively) in the TSS region (Fig. 3; stable 3) are detected by this new genome-wide analysis. However, the new method expands its scope beyond the promoter region, hence there are additional eleven mutation blocks predicted in gene body of BCL2 and one in 5′ distance of BCL2 (Fig. 3). Especially, 9 mutation blocks out of total 14 (Table 1) were found overlapping with enhancers from 197 tissue/cell types. Tar-

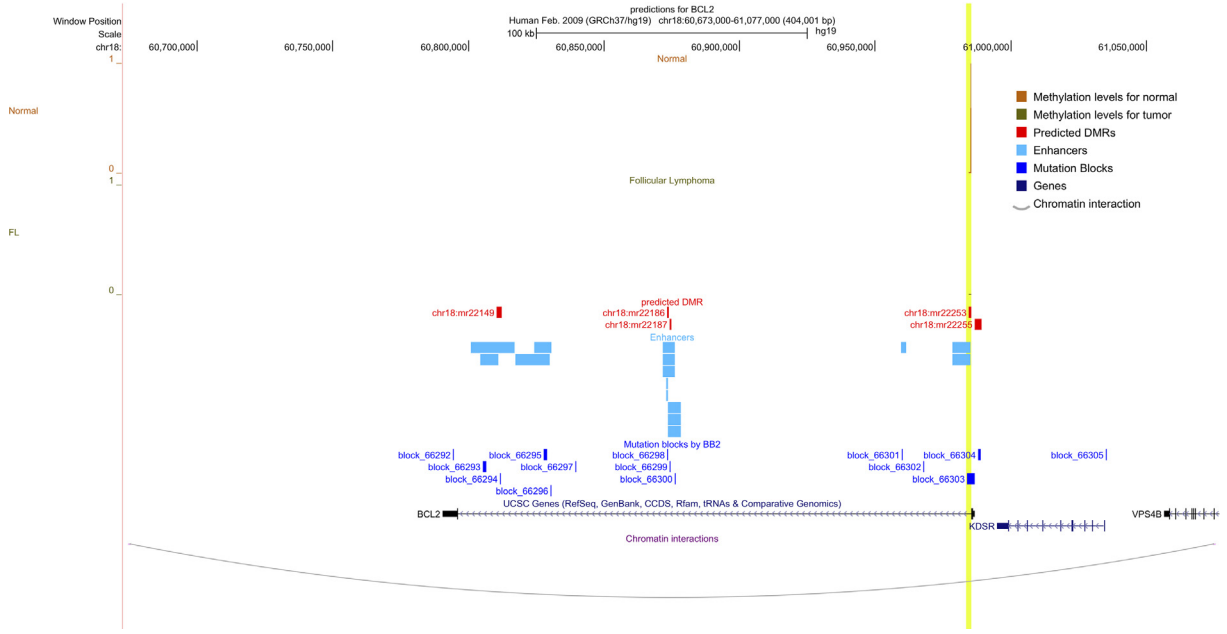


Fig. 3. Mutation blocks, differentially methylated regions, and enhancers in a single TAD around BCL2 identified by the new integrative data analysis. This figure displays mutation blocks, DMRs and enhancers that are identified in a single TAD containing the BCL2 gene. First panel presenting brown color horizontal bars presents mean methylation levels of normal samples in the DMR that are predicted around BCL2 and overlaps with the mutation block and enhancers. Second panel containing green coloured horizontal bars presents mean methylation levels of the same DMR in FL samples. Third panel presents all predicted DMRs with respective DMR IDs in the region in form of red tiles. Fourth panel presents overlapping enhancers (with mutation blocks or DMRs) in the region from the 5 FL related cell lines (DOHH2, GC B cell, Namalwa, OCI-ly1 and OCI-Ly7) in light blue color. Fifth panel presents mutation blocks (with respective block IDs) predicted by BayesPI-BAR2 in dark blue tiles. Sixth panel presents the RefSeq genes (BCL2, KDSR) present in the region. Seventh panel presents the TAD boundaries around the region, linking the start of TAD with its end with a grey curve. A yellow vertical bar across the figure highlights the important overlapping mutation blocks, DMRs, enhancers discussed in result section. Coordinates for all these genomic features are mentioned in stable4. (For interpretation of the references of color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Information of the selected four top ranking FL related mutation block-gene pairs identified by new integrative data analysis. Table contains details for the selected top ranking four genes (BCL2, BCL6, CDCA4 and CTSO). Total mutation blocks and DMRs (differential methylated regions) identified around the four genes and the enhancers overlapping with any of these are mentioned. TAD (topological associated domains) boundaries around the genes, identifies genomic regions are mentioned, and DEG (differential gene expression) levels are also described. Gene name marked by * means it did not pass the robustness analysis by using chromatin state segmentations (Stable 7) that was not used in the prediction.

Gene name	Mutation blocks	DMRs	Overlapping enhancer	TAD	Gene expression
BCL2	14 total (9 overlapping to enhancers)	5 total (2 in TSS, 1 in 5′dist, 2 in gene) 3 hypo, 1 hyper, 1 mixDMR.	With mutation blocks = 2 (16 cell lines), 14 (FL related) With DMRs = 8 FL related	chr18:60675000–61075000 Low-active	Upregulation (pval < 1.7e-6)
BCL6	23 total (22 overlapping with enhancers)	1 total (in 5′dist) hyperDMR	With mutation blocks = 89 (7 cell lines), 17 (FL related) With DMRs = 1 FL related	chr3:187400000–189600000 Low-active	Down regulation (pval < 0.0008)
CDCA4	30 total (14 overlapping with enhancers)	13 total (in 5′dist) 4 hypo,9 hyperDMRs	With mutation blocks = 1 (5 cell lines), 10 (FL related) With DMRs = 17 FL related	chr14:105225000–107374540 Heterochromatin	down regulation (pval < 0.001)
CTSO*	47 total (7 overlapping with enhancers)	3 in total (5′dist) 1 hypo, 1 hyper, 1 mixDMR	With mutation blocks = 18 (6 cell lines), 0 (FL related) With DMRs = 2 non FL related	chr4:156850000–158025000 Low	Upregulation (pval < 5.5e-10)

get genes of the majority of these mutation blocks related enhancers are VPS4B and KDSR. Both genes are nearby BCL2. This indicates that there is a potential impact of these mutation blocks on long ranged regulation of these two genes. VPS4B is differentially expressed in diffuse large B-cell lymphoma [68]. KDSR is significantly differentially expressed between FL patient and normal samples (e.g., P-value < 0.0003; upregulated in FL compared to normal). Detailed enhancer target gene information of these mutation blocks is provided in the supplementary data.

Furthermore, to take into account the tissue specificity of enhancers, tissues/cell lines closer to FL were selected from the LL-100 panel which covers 100 cell lines for human leukemias and lymphomas [43]. There are 5 FL-related cell lines (DOHH2, GC B-cell, Namalwa, OCI-Ly 1 and OCI-Ly 7) with enhancer information in this study. Interestingly, 8 mutation blocks out of the total 14 are overlapping with 14 different enhancers from these 5 lymphoma cell lines. Additionally, 5 DMRs were found in and around BCL2, and one of them (mr22253) overlaps with the muta-

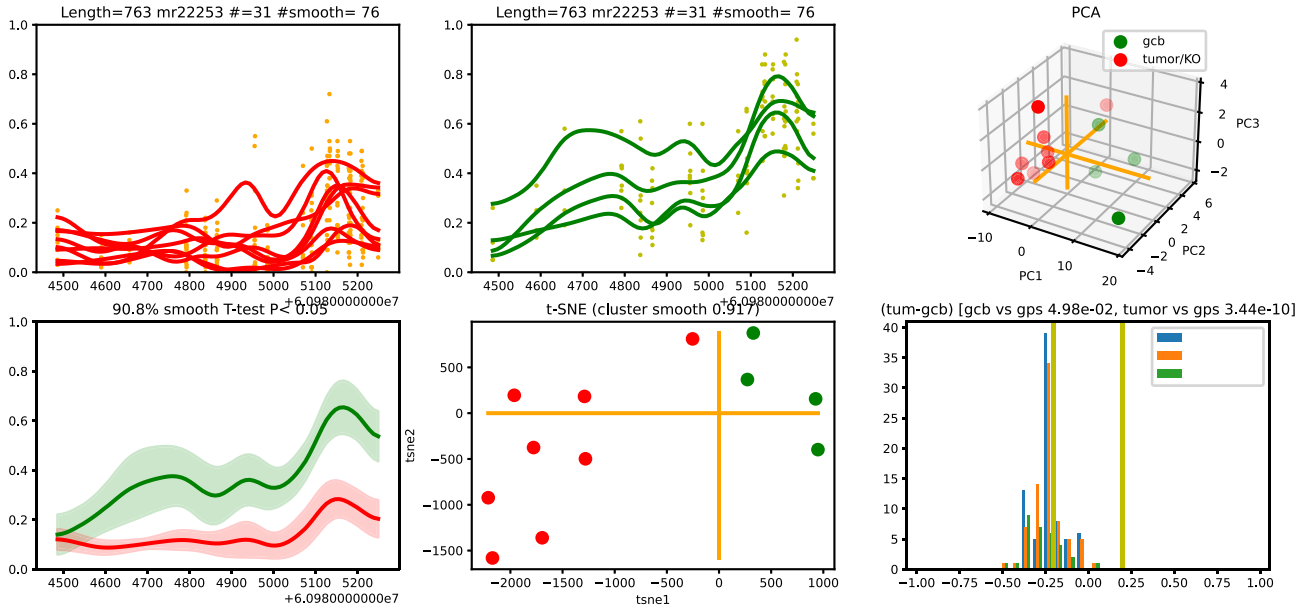


Fig. 4. A hypoDMR (mr22253) overlapping with mutation block (block_66303) identified in TSS region of BCL2. Figure displays six attributes of a predicted hypoDMR (mr22253 in chromosome 18; probability equals one in logistic regression model) in follicular lymphoma. Upper panel of the figure, from the left to the right side is the methylation profiles (both original 31 and smoothed 76 data points) of mr22253 in tumor (n = 8) and normal samples (n = 4), respectively, as well as a 3-D plot of the first three principal components for tumor and normal samples. Lower panel of the figure, from the left to the right side is the smoothed methylation profiles of tumor and normal groups where shadow arear represent a 95% confidence interval (~90.8% of data points are differentially methylated; P-value < 0.05), a 2-D plot of t-SNE map for 14 FL samples (the clustering accuracy of predicted sample group labels is ~ 0.917), and a histogram plot of mean methylation changes (centered around −0.25) between the two groups, respectively. Here, gcb/tumor vs gps means intra-group Euclidean distance of normal (P-value < 0.05)/tumor (P-value < 4e-10) samples versus inter-group Euclidean distance between tumor and normal samples, and tumor and normal samples are illustrated by red and green points (or lines), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tion block (block_66303) in TSS and an enhancer from OCI-LY1 cell line (Fig. 3, Stable 3). Since BCL2 is significantly over expressed (P-value < 1.66 e-06) in FL patients, one can expect a rise of hypomethylation in the regulatory region of BCL2. The predicted DMR (mr22253 in chromosome 18) successfully meets the expectation, by reporting it as a hypoDMR with a high score (probability equals 1 from logistic regression model), and the significant difference in methylation levels between tumor and normal samples is

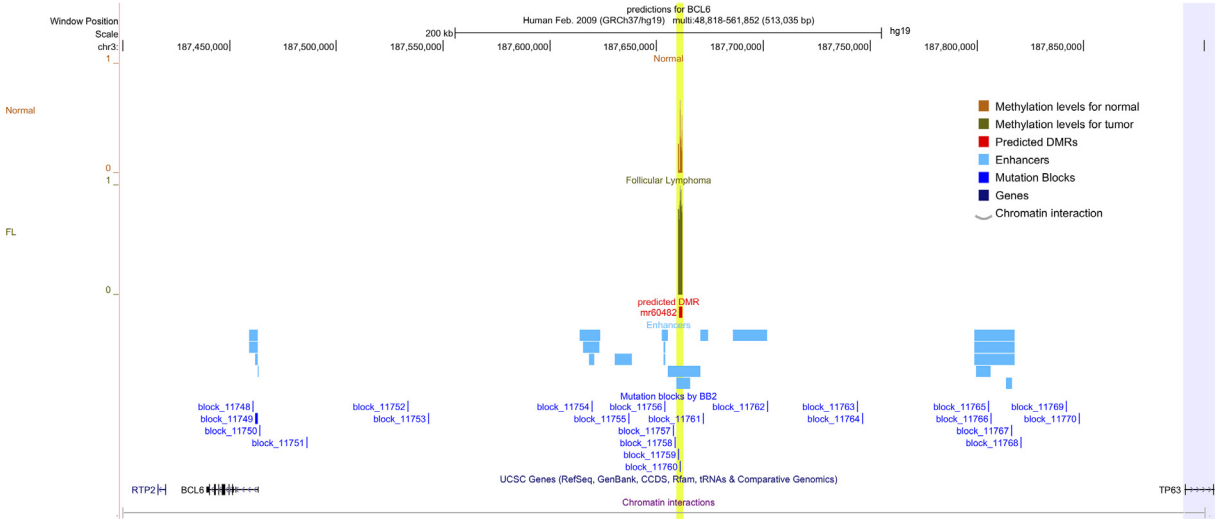


Fig. 5. Mutation blocks, differentially methylated regions, and enhancers in a single TAD around BCL6 identified by the new integrative data analysis. This figure displays mutation blocks, DMRs and enhancers that are identified in a single TAD containing BCL6 as well. First panel presenting brown color horizontal bars presents mean methylation levels of normal samples in the DMR that are predicted around BCL6 and overlaps with the mutation block and enhancers. Second panel containing green colored horizontal bars presents mean methylation levels of the same DMR in FL samples. Third panel presents all predicted DMRs with respective DMR IDs in the region in form of red tiles. Fourth panel presents overlapping enhancers (with mutation blocks or DMRs) in the region from the 5 FL related cell lines (DOHH2, GC B cell, Namalwa, OCI-ly1 and OCI-Ly7) in light blue color. Fifth panel presents mutation blocks (with respective block IDs) predicted by BayesPI-BAR2 in dark blue tiles. Sixth panel presents the RefSeq genes (BCL6 etc) present in the region. Seventh panel presents the TAD boundaries around the region, linking the start of TAD with its end with a grey curve. Since the TAD was large, the region of the interest is zoomed in and the end boundary of the TAD is shown in the right light blue vertical section of the image. A yellow vertical bar across the figure highlights the important overlapping mutation blocks, DMRs, enhancers discussed in result section Coordinates for all these genomic features are mentioned in stable5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

clearly illustrated in Fig. 4. Thus, an over expression of BCL2 in FL can be credited to the hypomethylation of the enhancer and mutation blocks present in promoter region of BCL2 (Figs. 3 and 4).

3.6. Differentially methylated region and mutation block in 5′ distance of BCL6.

Another key regulator in FL is the BCL6 gene. Located on locus 3q27, it codes for the BC-6 protein which is a transcriptional repressor expressed by germinal center B cells, and is crucial for GC formation [69]. In this work, 23 mutation blocks are identified in and around the BCL6 gene. Except for 2 mutation blocks that reside near TSS and gene body, all other mutation blocks are located in the 5′distance region of BCL6 (e.g., the long non-coding region in the 5′region of BCL6; Fig. 5, Stable4). There is a large gene desert stretch in 5′ distance region of the BCL6 (Fig. 5). Our previous study focusing on promoter regions only, reported one mutation block in promoter region of BCL6 [57]. This time two mutation blocks (sizes 1501 bp and 506 bp for block_11749 and block_11750, respectively; Fig. 5) are identified in the promoter region of BCL6, in which block_11749 is overlapping with the previously reported block [57] and is present in 11 out of 14 patients. While examining these 23 mutation blocks in TSS and 5′distance region of BCL6, 22 of those are overlapping with 17 enhancers from 5 lymphoma related cell lines. Additionally, a 1638 bp long DMR (mr60482; Fig. 6) is identified in the 5′distance (or gene desert) region of BCL6. This DMR is especially interesting because it covers a mutation block (block_11760; Fig. 5) and overlaps with enhancers from lymphoma related cell lines (OCI-LY1 and Namalwa). Target genes for these enhancers are long non coding RNAs like RP11-132 N15.3 and RP11-430L16.1, which involved in B-cell malignancies [70]. As shown in Fig. 6, hypermethylation of this DMR can result into inaccessibility of the affected enhancer region, having a negative impact on regulation of BCL6. As expected, a significant lower expression of BCL6 is observed in FL patients [57] as compared to that of the normal samples (p-value < 0.0008; down-regulated in FL compared to normal). This suggests that the DMR, present in the large gene desert stretch in the 5′ distance region of BCL6, may contribute in the oncogenesis of FL. Thus, hypermethylation of the region and the presence of a mutation block can inactivate the underlying enhancer by restricting TF binding, which results into lower expression of BCL6 in FL patients.

3.7. Cell division cycle associated protein 4 (CDCA4) as a putative new biomarker for FL

Both BCL2 and BCL6 genes are well known in FL, and they are ranked in the top 10 of the mutation block-gene associations in this new genome-wide mutation analysis. Among the top 10 predictions (Stable 2), a mutation block (block_61069) in chromosome 14 near CDCA4, has the highest mutation frequency across all 14 FL patients. Apart from this particular block, CDCA4 is also associated with 30 mutation blocks through its 5′distance region (Stable 5). Additionally, the 5′ distance region not only had large number of mutation blocks, but also showed extensive differential methylation (Fig. 7). There are 13 DMRs identified in 5′distance region of CDCA4, and all of them overlapped with mutation blocks (Stable 5). Additionally, 16 of 30 mutation blocks were also found overlapping with multiple enhancers from different cell lines. Interestingly, 14 out of these 16 mutation blocks overlap with 17 enhancers from the selected follicular lymphoma related cell lines. The target genes of all these enhancers are mainly from IGH family (i.e., IGHA1/2, IGHG2, IGHGP, IGHEP1, IGHG3, IGHD, IGHM, IGHJ4/5 etc). Moreover, all of these mutation blocks, DMRs, genes, and enhancer are located in a single TAD, which strengthens the idea of long-ranged interaction between the mutation blocks/DMRs and the genes (Fig. 7). For example, a mutation block (block_61069) overlaps with both enhancers (from all 5 FL related

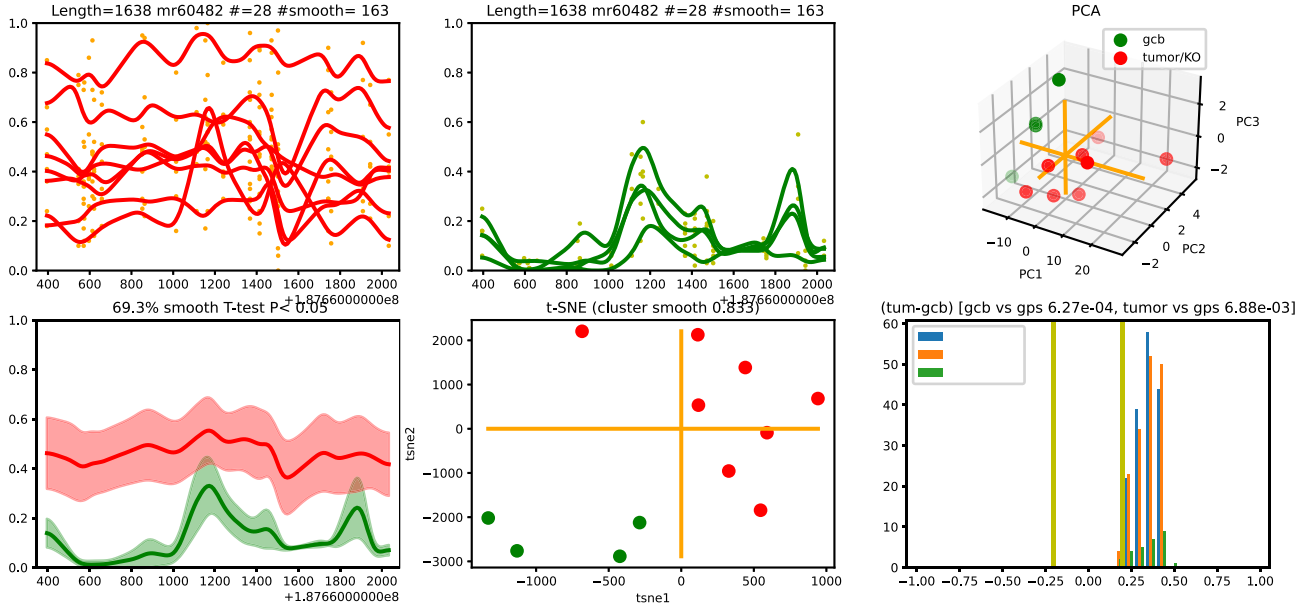


Fig. 6. A hyperDMR (mr60482) overlapping with a mutation block (block_11760) identified in 5′ distance region of BCL6. Figure displays six attributes of a predicted hyperDMR (mr60482 in chromosome 3; probability equals one in logistic regression model) in follicular lymphoma. Upper panel of the figure, from the left to the right side is the methylation profiles (both original 28 and smoothed 163 data points) of mr60482 in tumor and normal samples, respectively, as well as a 3-D plot of the first three principal components for tumor and normal samples. Lower panel of the figure, from the left to the right side is the smoothed methylation profiles of tumor and normal groups where shadow arear represent a 95% confidence interval (~69.3% of data points are differentially methylated; P-value < 0.05), a 2-D plot of t-SNE map for 14 FL samples (the clustering accuracy of predicted sample group labels is ~ 0.833), and a histogram plot of mean methylation changes (centered around 0.35) between the two groups, respectively. Here, gcb/tumor vs gps means intra-group Euclidean distance of normal (P-value < 0.00063)/tumor (P-value < 0.007) samples versus inter-group Euclidean distance between tumor and normal samples, and tumor and normal samples are illustrated by red and green points (or lines), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

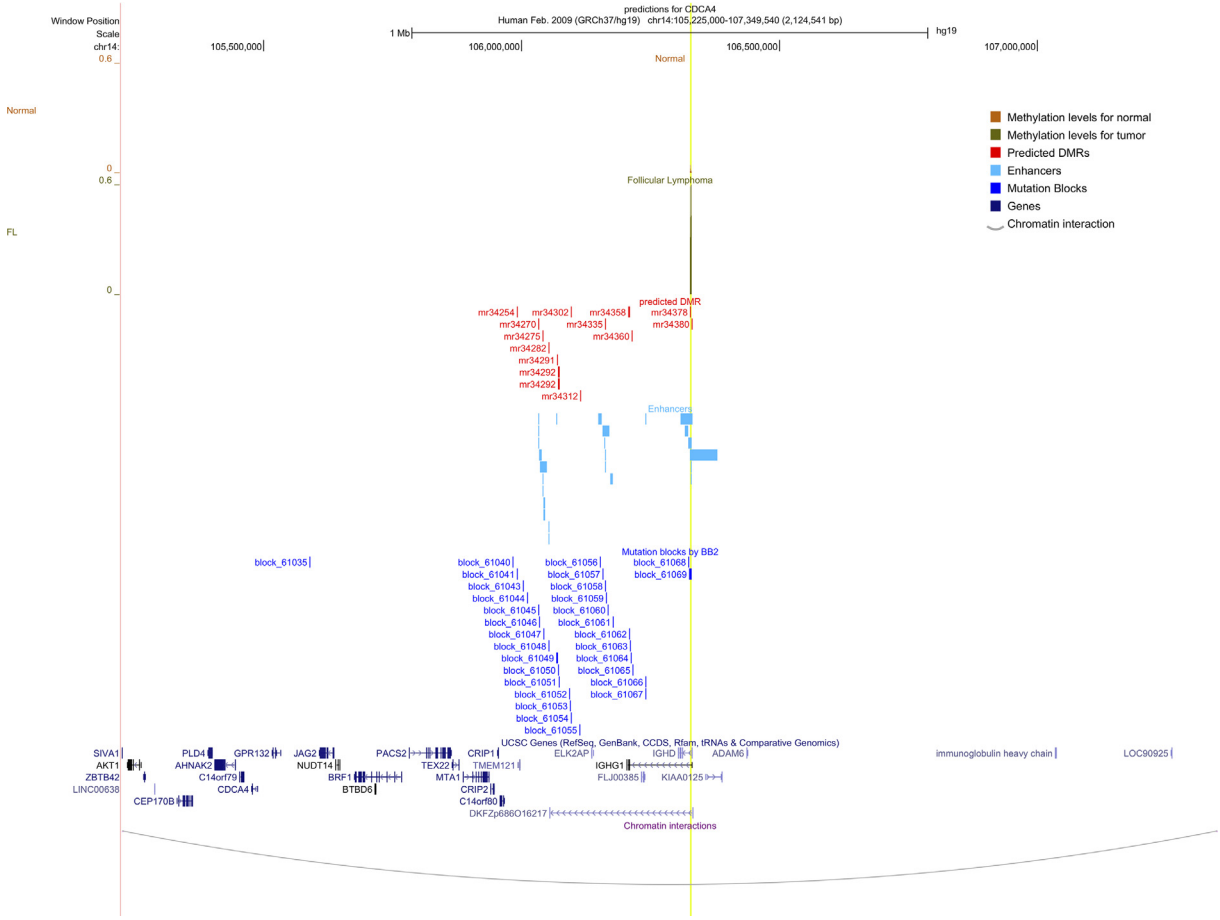


Fig. 7. Mutation blocks, differentially methylated regions, and enhancers in a single TAD around CDCA4 identified by the new integrative data analysis. This figure displays mutation blocks, DMRs and enhancers that are identified in a single TAD containing CDCA4 as well. First panel presenting brown color horizontal bars presents mean methylation levels of normal samples in the DMR that are predicted around CDCA4 and overlaps with the mutation block and enhancers. Second panel containing green coloured horizontal bars presents mean methylation levels of the same DMR in FL samples. Third panel presents all predicted DMRs with respective DMR IDs in the region in form of red tiles. Fourth panel presents overlapping enhancers (with mutation blocks or DMRs) in the region from the 5 FL related cell lines (DOHH2, GC B cell, Namalwa, OCI-ly1 and OCI-Ly7) in light blue color. Fifth panel presents mutation blocks (with respective block IDs) predicted by BayesPI-BAR2 in dark blue tiles. Sixth panel presents the RefSeq genes (CDCA4, IGHG, IGHD etc) present in the region. Seventh panel presents the TAD boundaries around the region, linking the start of TAD with its end with a grey curve. A yellow vertical bar across the figure highlights the important overlapping mutation blocks, DMRs, enhancers discussed in result section. Coordinates for all these genomic features are mentioned in stable 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cell lines) and two DMRs (mr34380 and mr34378; Fig. 8 and SFig. 3), and both DMRs have near to zero methylation in normal samples but high methylation levels in FL patients. Mutation blocks in the enhancer region with high methylation levels (e.g., hypermethylation) can alter the transcription factor binding in the region, which may result in altered target gene expressions levels (e.g., decreased CDCA4 gene expression due to hypermethylation in the regulatory region). Indeed, the expression levels of CDCA4 were significantly decreased in FL patients (e.g., P-value < 0.001). Thus, both the presence of mutation blocks and the variation of DNA methylation in the same region might disrupt the binding of transcription factors that are regulating CDCA4 expression. For that reason, BayesPI-BAR2 was used to predict TF binding affinity changes at DNA sequences of 2 DMRs (mr34380 and mr34378), which overlap with both a mutation block and enhancers (details in supplementary methods). The result suggests that there are several transcription factors showing significant binding affinity changes (SFigs. 4 and 5) between the tumor and the normal DNA sequences. For example, in region of mr34378, TFs (HOXD9, BAPX1, PBX1, MSX3, MAFB, HOXA2, TST-1, EN1) have significantly reduced binding affinity on sequences with the presence of mutation blocks (SFig. 4). For mr34380, the binding affinities of both TFAP2A and HIC1 are negatively impacted by the mutations in FL (SFig. 5). TFAP2A has been reported previously as hypermethylated in diffuse large B-cell lymphoma [71].

3.8. CTSO as potential biomarker for follicular lymphoma

CTSO (Cathepsin O) gene encodes for a proteolytic enzyme which plays an important role in cell death and apoptosis. Cathepsins are generally considered as house-keeping enzymes which are ubiquitously expressed in human tissues. CTSO and other members of the same enzyme family have already been recognized as a biomarker in other cancers such as breast cancer. However, in our study we found that a non-coding region upstream of CTSO can harbour regulatory potential for FL. We detected 47 mutation blocks in and around CTSO, and 4 mutation blocks out of these overlapped with enhancers from different cell lines and tissues (SFig. 6, Stable 6). Interestingly, we found 3 DMRs in the same 5′ region. All three of them overlapped with mutation blocks, and two (mr43894 with block_27288 in SFig. 7 and mr_43917 with block 27297 in SFig. 8) of them also overlapped with 18 enhancers from 6 different cell lines. All three predicted DMRs are hypoDMRs and lie within the same TAD as the gene and mutation blocks (SFig. 6). One of the DMRs (mr43894) predicted by our method. This is shown in detail in (SFig. 7) where there is a clear difference

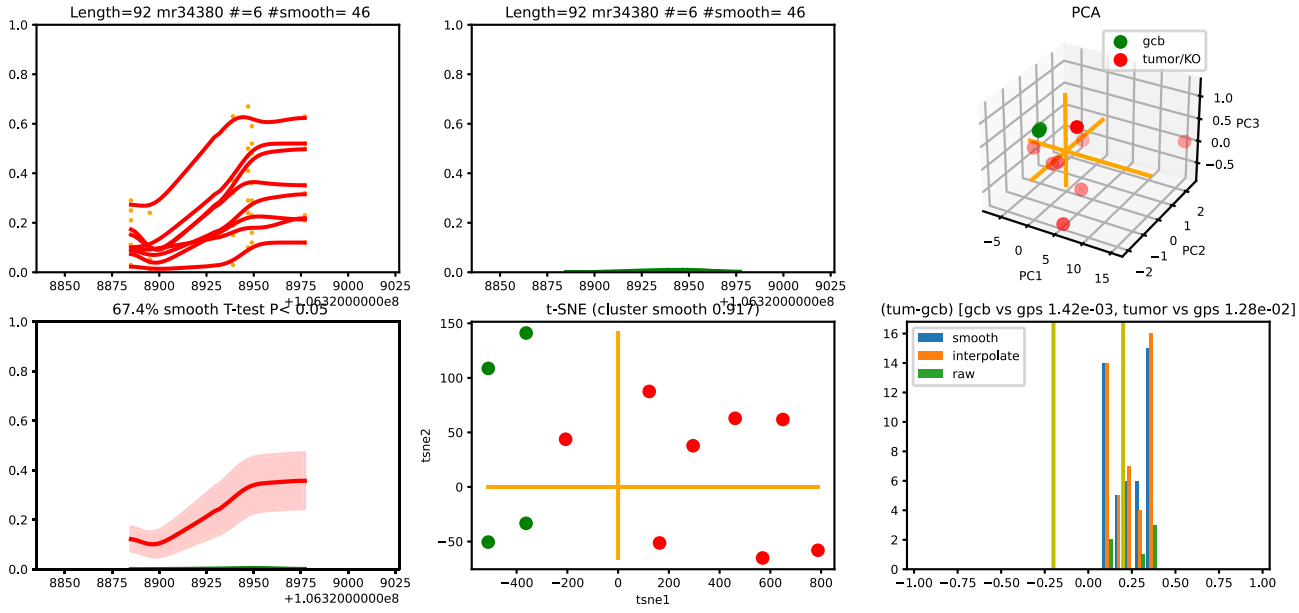


Fig. 8. A hyperDMR (mr34380) identified in 5′ distance region of CDCA4 overlapping with the highest ranked mutation block (block_61069). Figure displays six attributes of a predicted hyperDMR (mr34380 in chromosome 14; probability equals one in logistic regression model) in follicular lymphoma. Upper panel of the figure, from the left to the right side is the methylation profiles (both original 6 and smoothed 46 data points) of mr34380 in tumor and normal samples, respectively, as well as a 3-D plot of the first three principal components for tumor and normal samples. Lower panel of the figure, from the left to the right side is the smoothed methylation profiles of tumor and normal groups where shadow arear represent a 95% confidence interval (∼67.4% of data points are differentially methylated; P-value < 0.05), a 2-D plot of t-SNE map for 14 FL samples (the clustering accuracy of predicted sample group labels is ∼ 0.917), and a histogram plot of mean methylation changes (centered around 0.2) between the two groups, respectively. Here, gcb/tumor vs gps means intra-group Euclidean distance of normal (P-value < 0.0015)/tumor (P-value < 0.013) samples versus inter-group Euclidean distance between tumor and normal samples, and tumor and normal samples are illustrated by red and green points (or lines), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between methylation patterns of tumor and normal samples. Of interest, there is a dip of the methylation level at the center of this 90 bp long DMR. There is a possibility that the lower methylation level indicates an underlying transcription factor binding site in the enhancer that overlaps with this DMR. We discuss this possibility because methylation is a repressive mark and for a house keeping enzyme like CTSO, DNA methylation levels must be lower in the regulatory region of CTSO where relevant *trans*-acting elements are expected to bind. Existence of this hypomethylation on top of enhancers that contain the mutation predicts upregulation of CTSO in FL samples. Indeed, we found significant high expression level of CTSO in FL sample as compared to normal samples (e.g., p-value < 5.5e-10). Hence, we conclude that the presence of DMRs and mutation blocks in the upstream regulatory region of CTSO may result in over expression of CTSO gene in FL tumors.

3.9. Robustness analysis of the predicted mutation block-gene associations in follicular lymphoma

To evaluate the robustness of aforementioned four mutation block-gene associations (BCL2, BCL6, CDCA4, and CTSO) based on an integrative analysis of DMR, DEG and TAD information (e.g., 327 genes in Fig. 2), we repeated analysis three times with different setting or parameters and obtained three sets of top ranked mutation block-gene associations from the same FL data: 1) top ranked 327 block-gene associations predicted based on 2105 DEG, where RPKM values were quantile normalized and computed from gene region; 2) top ranked 959 block-gene associations by using 4603 DEG, where RPKM values were log transformed quantile normalized and calculated from exon; 3) top ranked 758 block-gene associations selected based on the same 4603 DEG but with different genome/regulatory feature scores (e.g., TSS = 4, Enhancer = 3, Gene = TES = 2, and 5′Dist = 1). Then, a newly developed robustness analysis was applied on these three sets of top ranked mutation block-gene associations, the fraction/percentage of mutation blocks of a gene located in seven types of chromatin segmentation in human genome (Fig. 9) was computed, respectively. Results from Fig. 9 and supplementary SFigs. 9 and 10 illustrate that the majority of mutation blocks of a gene are overlapping to R (repressed regions > 80%), T (transcribed regions > 25%), and E (enhancer > 10%). Often, the mutation blocks are significantly (e.g., the mean of expected P-values is ∼ 0.05 by 10,000 random samplings) associated with R, E, TSS, and PF (promoter flanking) regions. After filtering mutation blocks without significant associations with either TSS or E region, the top 20 ranked genes (supplementary Stable 7) were reported from the three results, respectively. The mutation blocks associated with BCL2, CDCA4, and BCL6 passed such robustness analysis in all of the three tests (e.g., the mutation blocks significantly enriched in either TSS or enhancer region based on a new permutation test; the expected P-value < 0.05), and are all ranked in top 15 from the three results. None of the three results include CTSO, meaning it does not pass our robustness analysis. Therefore, the predicted mutation block-gene association for CTSO may not be as robust as the other three ones, which requires further confirmation from newly evaluated clinical samples.

3.10. Chromosome translocation and regulatory mutation blocks in BCL2

BCL2 t(14; 18) translocation is found in 85–90% of FL cases, which is considered to be the main cause of high BCL2 expression in FL. In the current study, two regulatory mutation blocks (block_66303 and block_66304; Fig. 3) near BCL2 are found to affect>70% of 14 FL patients (Stable 1), and both of them are located near a DMR at the promoter of BCL2 (hypoDMR mr22253; Fig. 4). We suspect this regulatory arrangement has a significant impact on BCL2 expression. Thus, it is very interesting

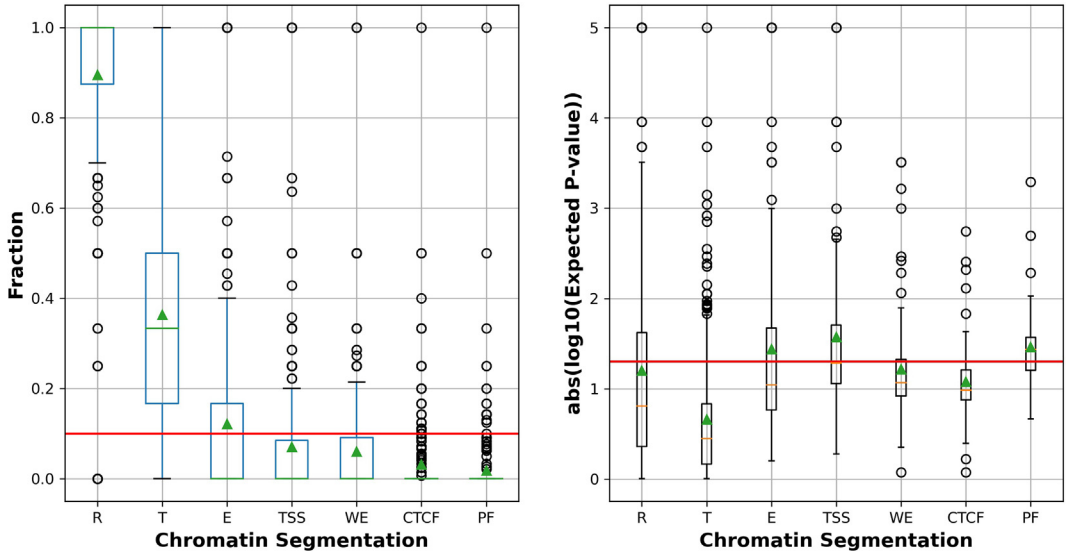


Fig. 9. Boxplots for the percentage and the expected P-values of mutation blocks associated with 327 genes that overlapping in the seven types of chromatin segmentation. Left panel of figure shows a box plot of the fraction/percentage of mutation blocks that are overlapping with the seven types of chromatin segmentation (or chromatin states) in human genome, the percentage for each gene is calculated based on the number of its associated mutation blocks overlapping to the seven types of segmentation of genome, respectively. A red horizontal line represents the 10% of mutations blocks of a gene, and the green triangles are the mean values in each box plot. Right panel of the figure displays the absolute log10 of expected p-values of mutation blocks associated to a gene are enriched in the seven types of chromatin segmentation, respectively. The p-values are calculated on base of a permutation test of 10,000 times randomly sampled mutation blocks. A red horizontal line indicates an absolute log10 of expected p-value = 1.3 (e.g., expected p-value = 0.05), and the green triangles are the mean values in each box plot. The seven types of chromatin segmentation of human genome (or chromatin states) are R, T, E, TSS, WE, CTCF, and PF that represent the predicted repressed/low activity region, transcribed region, enhancer, promoter region/transcription start site, weak enhancer/open chromatin region, CTCF enriched element, and promoter flanking region, respectively. The mutation blocks associated with 327 genes are selected by a weighted vote approach (e.g., mean feature score>=0.5; Fig. 2). (For interpretation of the references of color in this figure legend, the reader is referred to the web version of this article.)

Table 2
A two-way contingency table for both the two BCL2 regulatory mutation blocks and the BCL2 t(14; 18) translocation in 14 follicular lymphoma samples. This table shows the distribution of 14 follicular lymphoma samples with either two BCL2 regulatory mutation blocks (both block_66303 and block_66304; Fig. 3) or BCL2 t(14;18) translocation. A two-sided Fisher's exact test of this table is p-value = 0.01, which indicates that there is a significant correlation between the two BCL2 regulatory mutation blocks and the BCL2 t(14;18) translocation in FL, though the number of sample size is small.

	Two BCL2 regulatory mutation blocks (Yes)	Two BCL2 regulatory mutation blocks (No)	Total
BCL2 t(14;18) translocation (Yes)	10	1	11
BCL2 t(14;18) translocation (No)	0	3	3
Total	10	4	14

to study the relationship between the two regulatory mutation blocks and the BCL2 t(14;18) translocation in FL. We first tested all 14 FL samples for presence of BCL2 t(14; 18) by using DELLY2 [72] program. We found 10 out of the 14 samples (71%) have BCL2 t(14;18) translocation (Stable 1). As Supplementary SFig. 11 shows, all BCL2 t(14; 18) translocations appear near the 3′ end, which means the translocated chromosomal segment may contain the two regulatory mutation blocks when IGH-BCL2 translocation happens. Supplementary STable 1 provides a summary of the distribution of the two regulatory mutation blocks and the predicted BCL2 t(14;18) translocation in 14 FL patients. In Table 2, a two-way contingency table for the two regulatory mutation blocks and the BCL2 t(14; 18) translocation in 14 FL samples is illustrated, where a Fisher's exact test of the table is p-value = 0.01. It indicates that there is a significant correlation between the two regulatory mutation blocks (blocks 66303 and 66304; Fig. 3) and the BCL2 t(14;18) translocation in FL. Results of BCL2 differential expression analysis between the four groups of FL samples (e.g., with/without the two regulatory blocks or with/without BCL2 t(14;18) translocation) and the normal control samples are shown in supplementary SFig. 12, which suggest that both the two regulatory mutation blocks (P-value < 0.001) and the BCL2 t(14;18) translocation (P-value < 0.002) contribute to high BCL2 expression in FL. Though there is a significant difference (P-value < 0.029) of BCL2 expression between the FL samples without BCL2 t(14; 18) translocation and the normal ones, the differential expression of BCL2 between the FL samples without the two regulatory mutation blocks and the normal ones is marginal (P-value < 0.057). Thus, the two regulatory mutation blocks around BCL2 may have an effect additive to BCL2 t(14; 18) translocation that causes abnormally high BCL2 expression in FL. Nevertheless, the sample size is small in the current study, which needs more data in future for confirmation.

4. Discussion

The regulation of the genome is complex and cannot be explained by a single type of data set. Integration of the major genomic changes like SNVs, DNA methylation, gene expression and genomic structures like TADs, holds the promise of a more thorough understanding of the genome in human disease. Complex disease phenotypes like cancer are regulated by multiple genomic and epigenomic factors. Differential gene expression can be caused by changes in a nearby regulatory region like enhancer regions.

Enhancer regions can show epigenetic change such as hypomethylation, increasing the chance of transcription factor binding or can show genetic changes like SNVs resulting into altered binding of transcription factors. Both of these phenomena can cause over expression of a gene. However, long range interaction between potential regulatory regions like enhancers and mutation blocks are difficult to be mapped to their target genes. TADs restrict such long-range interactions within the TAD boundaries and can be used to map the regulatory regions to their respective target genes. TADs also are fundamental 3D genomic structures, which are important for gene regulation and known to restrict long-distance gene regulation in their boundaries (e.g., by topologically limiting enhancer's approach to its target genes) [73]. Usually, TADs are stable across different cell-types or conditions and TAD boundaries are more evolutionarily constrained as compared to TADs [74–76]. Thus, it is reasonable to apply inferred common TADs from cell line data, and to assume that both the long-distance gene regulatory region (e.g., 5′distance region) and the putative target genes are positioned in the same TAD. In this way, many non-significant associations between a gene and a mutation block in its neighboring non-coding region like 5′distance regions can be removed.

In this work, a novel method for integration of aforementioned diverse information is developed to understand the importance of regulatory mutation in FL, by analyzing genetic, epigenetic, and transcriptomic data collected from the same FL cohort. Gene expression profiles of immune cell markers (e.g., T-cells (CD3), macrophage (CD68), and B-cells (CD19; PAX5); SFig. 13) from the 14 FL samples shows that there is a similar expression level between T-cells/macrophage genes and B-cell genes. This indicates that these FL samples were representative of the tumour. Thus, a new machine learning method (Fig. 1) was designed first to identify and rank DMR between two groups of samples. Then, the extracted differential methylation information from genome-wide DNA methylation data is integrated with mutation blocks related to complex phenotype like FL. Mutation blocks are also mapped to differentially expressed genes (DEG) and enhancers that lie in the same TAD boundaries. Finally, a list of high confidence disease related genes associated with mutation blocks and DMRs is obtained by using a new weighted voting approaching that considers four feature scores (e.g., the number of patients affected by mutation blocks, DMR significance, P-value to DEG, and the weighted genomic feature of a mutation block; Fig. 2). Based on this new integrative data analysis pipeline, ∼66868 mutation blocks (initial) are initially identified in genome-wide manner, from 14 FL patients by BayesPI-BAR2, but are reduced to ∼ 45570 blocks by considering the mutation blocks overlapping with either DMR or DEG (DMR-or-DEG). This number is further decreased to ∼ 1272 mutation blocks with strong associations with 159 genes through TSS/TES/gene body/5′Distance, when including TAD information and GO biological process/pathway enrichment information (DMR-or-DEG-Pathway). The number of mutation blocks located in enhancers (e.g., ∼ from 197 tissue/cell lines, retrieved from EnhancerAtlas 2.0) is ∼ 58% for initial analysis. But it increases to 60%, and then 76% for DMR-or-DEG, and DMR-or-DEG-Pathway analysis, respectively. Which means that when more information is included in the analysis, a higher number of mutation blocks with regulatory potential (through their overlap with enhancers) are detected. Therefore, an integrative analysis of mutation blocks by including diverse information (e.g., DMR, DEG, TAD, and GO/pathway enrichment information) can significantly improve the prediction of functional regulatory mutations in disease.

Among the top 10 (Stable 2) of the final 159 genes having strong associations with mutation blocks in FL patients, BCL2 is the highest scoring gene which shows over expression in FL. Additionally, two mutation blocks are detected in the promoter region of BCL2. A novel hypoDMR was also found covering the two mutation blocks (Fig. 3) and an enhancer. In the DMR, the high methylation levels in the normal samples (Fig. 4; noticeably form two peaks in the normal samples) indicate hypomethylation at the locations of the two mutation blocks. However, in tumor samples the methylation levels are decreased, allowing possibility of transcription factor binding in the region which can be a reason for high expression of BCL2 in FL. Our previous study reported altered TF binding in the same region, justifying the higher expression levels of BCL2 during lymphomagenesis [57]. BCL6 is also in the top ten scoring genes from the current study, which is considered as a key regulator in FL. In this work, a novel mutation block in the gene desert stretch in the 5′distance region of BCL6 is identified by the new integrative data analysis (Fig. 5). Of interest, the mutation block exists in an enhancer region and also shows significant hypermethylation in the lymphoma samples (Fig. 6). The presence of the hypermethylated DMR encircling a mutation block in an enhancer region of BCL6 may explain the low expression of this gene in follicular lymphoma.

Another interesting mutation block-gene association from the top ten list is CDCA4, which appears in all 14 patients. One of the mutation blocks is located in both the DMR and the enhancer region of the CDCA4 gene. There are two hypermethylated DMRs in this mutation block (Fig. 8 and SFig. 3), which may explain the observed down regulation of the CDCA4 gene. Upon investigation of the two DMRs in this mutation block, DNA sequence binding affinities of several transcription factors are predicted to be significantly altered due to mutations in FL patients (e.g., SFigs. 4 and 5; PBX1 and HOX family). Moreover, enhancers overlapping with the mutation blocks in the upstream 5′distance region of CDCA4 were targeting IgH genes which are a molecular hallmark of FL [77,78]. There are few IgH genes like IGHG and IGHD also present in the same TAD as seen in Fig. 7. It has been reported previously that HOX genes contribute to oncogenesis in FL because of BCL2/IGH rearrangements [79]. A similar pattern in neighborhood of CDCA4 is observed, where the DNA sequence binding affinities of HOXD9 and HOXA2 are negatively disrupted (SFig. 4; mr34378) in the enhancer, which can influence the expression of IGH. Notably, HOXD9 is also significantly differentially expressed (upregulated in FL compared to normal; p-value < 1.91e-5) between FL and normal samples. It is also worthy of note that hindrance in expression of CDCA4 in melanoma cells by siRNA causes inhibition of expression of BCL2 [80]. We suspect that CDCA4 has a functional collaboration with oncogenes and this could reflect its status as a suspected tumor progressor gene that can affect pathways common to multiple cancers. Similarly, CTSO as a key player in cell death and apoptosis, also harbours multiple mutation blocks (total 47) which overlap with enhancers and DMRs in its vicinity (SFig. 6). All three predicted DMRs around CTSO are hypoDMRs (two presented in SFigs. 7 and 8). In connection to hypomethylation in 5′ distance region of CTSO, CTSO was found over expressed in FL. BCL2 overexpression helps malignant cells to escape apoptosis [81]. Like BCL2, CTSO is also related to apoptosis. Both of these genes have been previously reported as downregulated in other cancers like melanoma and leukemia [82,83]. However, the mutation blocks associated to CTSO did not pass the robustness analysis (e.g., the significant enrichment in either enhancer or TSS regions; Stable 7), based on seven chromatin states predicted from chromatin modifications (Fig. 9). Therefore, without a confirmation from new clinical samples, CTSO may be a weaker candidate driver gene than CDCA4 for FL oncogenesis, though not reported previously as such.

In conclusion, the new integrative analysis of genome-wide SNVs in FL has uncovered candidate genes, mutation blocks and DMRs that can guide further research, regarding their functional

role in FL oncogenesis. This study also provides a rich data source for the further exploration and understanding of FL, by providing extensive lists of high confidence DMRs, of significant mutation blocks in enhancer regions (Stable 8) and of mutation block-gene pairs (supplementary website; https://amnfar.github.io/FL_project/FL_webpage.html). Demonstrating the robustness of the results, this work not only discovered mutation blocks reported in previous studies but also identified new mutation blocks in the non-coding regulatory region that are associated to FL. To strengthen the evidence of identified association of sequence variation like SNVs with the disease, differential methylation events were also assessed and DMRs were reported. From the regulatory perspective, identification of differential methylation events surrounding those mutation blocks, at 5′distance region of genes known for their relevance to FL, indicates that the mutation blocks and DMRs are involved in dysregulation of those genes. From a clinical perspective, DNA methylation changes identified in connection to FL represent an attractive therapeutic target because epigenetic changes are reversible in nature as compared to genetic changes. The method developed in the present study can be used to comprehensively identify driver elements of any other malignancy, provided that data is available. Nevertheless, there is a limitation in *in silico* predictions, the top ranked results may not always be reliable (e.g., only three of four proposed mutation block-gene associations passed a robustness analysis, by using an independent information that was not included in the prediction) and shall be confirmed by using newly evaluated clinical samples, which is extensive and will become a future project.

Funding

This work was supported by the South-Eastern Norway Regional Health Authority (HSØ 2017061 and HSØ 2018107), Radiumhospitalets Legater (project number 35279), and the Norwegian Research Council NOTUR project (nn4605k).

CRediT authorship contribution statement

Amna Farooq: Validation, Formal analysis, Visualization, Writing – original draft. **Gunhild Trøen:** Validation, Formal analysis. **Jan Delabie:** Validation, Writing – review & editing. **Junbai Wang:** Conceptualization, Methodology, Software, Data curation, Writing – review & editing, Visualization, Investigation, Formal analysis, Supervision, Validation, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the anonymous reviewers for their comments and suggestions that helped us improve the work, as well as ICGC for getting access to cancer genomics data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.03.023.

References

[1] Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer 2004;4(2):143–53.
[2] Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 1983;301(5895):89–92.
[3] Liu F et al. Beyond transcription factors: how oncogenic signalling reshapes the epigenetic landscape. Nat Rev Cancer 2016;16(6):359.
[4] Jones PA, Baylin SB. The epigenomics of cancer. Cell 2007;128(4):683–92.
[5] Nannini M et al. Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives. Cancer Treat Rev 2009;35(3):201–9.
[6] Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. The Lancet 2011;378(9805):1812–23.
[7] Prat A et al. Immune-related gene expression profiling after PD-1 blockade in non–small cell lung carcinoma, head and neck squamous cell carcinoma, and melanoma. Cancer Res 2017;77(13):3540–50.
[8] Kumar SU et al. Integrative Bioinformatics Approaches to Map Potential Novel Genes and Pathways Involved in Ovarian Cancer. Front Bioeng Biotechnol 2019;7(391).
[9] Alsaleem MA et al. A novel prognostic two-gene signature for triple negative breast cancer. Mod Pathol 2020;33(11):2208–20.
[10] Li Y et al. Exome analysis reveals differentially mutated gene signatures of stage, grade and subtype in breast cancers. PLoS ONE 2015;10(3):e0119383.
[11] Lawrence MS et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499(7457):214–8.
[12] Huang FW et al. Highly recurrent TERT promoter mutations in human melanoma. Science 2013;339(6122):957–9.
[13] Zhang X, Meyerson M. Illuminating the noncoding genome in cancer. Nature Cancer 2020;1(9):864–72.
[14] Zhang Y et al. A pediatric brain tumor atlas of genes deregulated by somatic genomic rearrangement. Nat Commun 2021;12(1):1–17.
[15] Dixon JR et al. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet 2018;50(10):1388–98.
[16] Tsujimoto Y et al. Cloning of the chromosome breakpoint of neoplastic B cells with the t (14; 18) chromosome translocation. Science 1984;226(4678):1097–9.
[17] Okosun J et al. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. Nat Genet 2014;46(2):176–81.
[18] Choi J-H et al. Genome-wide DNA methylation maps in follicular lymphoma cells determined by methylation-enriched bisulfite sequencing. PLoS ONE 2010;5(9):e13020.
[19] Hayslip J, Montero A. Tumor suppressor gene methylation in follicular lymphoma: a comprehensive review. Molecular cancer 2006;5(1):1–7.
[20] Bennett LB et al. DNA hypermethylation accompanied by transcriptional repression in follicular lymphoma. Genes Chromosom Cancer 2009;48(9):828–41.
[21] Huet S et al. A gene-expression profiling score for prediction of outcome in patients with follicular lymphoma: a retrospective training and validation analysis in three international cohorts. Lancet Oncol 2018;19(4):549–61.
[22] Batmanov K et al. Integrative whole-genome sequence analysis reveals roles of regulatory mutations in BCL6 and BCL2 in follicular lymphoma. Sci Rep 2017;7(1):1–15.
[23] Macintyre G et al. is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics 2010;26(18):i524–30.
[24] Manke T, Heinig M, Vingron M. Quantifying the effect of sequence variation on regulatory interactions. Hum Mutat 2010;31(4):477–83.
[25] Ding X et al. Searching high-order SNP combinations for complex diseases based on energy distribution difference. IEEE/ACM Trans Comput Biol Bioinf 2014;12(3):695–704.
[26] Kircher M et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46(3):310–5.
[27] Fu Y et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 2014;15(10):1–15.
[28] Liu H et al. Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-of-Origin. Front Cell Dev Biol 2021;9:886.
[29] Assié G et al. Integrated genomic characterization of adrenocortical carcinoma. Nat Genet 2014;46(6):607–12.
[30] Achinger-Kawecka J et al. Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. Nat Commun 2020;11(1):1–17.
[31] Cardoso MdFS et al. *Putative biomarkers for cervical cancer: SNVs, methylation and expression profiles.* Mutation Research/Reviews. Mutat Res 2017;773:161–73.
[32] Adeel MM et al. Structural Variations of the 3D Genome Architecture in Cervical Cancer Development. Front Cell Dev Biol 2021;9:1885.
[33] Zhou Y et al. *The impact of DNA methylation dynamics on the mutation rate during human germline development.* G3: Genes, Genomes. Genetics 2020;10(9):3337–46.
[34] Chen Y-C et al. Significant associations between driver gene mutations and DNA methylation alterations across many cancer types. PLoS Comput Biol 2017;13(11):e1005840.

[35] Shoemaker R et al. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res 2010;20(7):883–9.
[36] Stern JL et al. Allele-specific DNA methylation and its interplay with repressive histone marks at promoter-mutant TERT genes. Cell reports 2017;21 (13):3700–7.
[37] Richter J et al. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat Genet 2012;44(12):1316–20.
[38] Lappalainen I et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet 2015;47(7):692–5.
[39] Beguelin W et al. EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. Cancer Cell 2013;23 (5):677–92.
[40] Kretzmer H et al. DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. Nat Genet 2015;47(11):1316–25.
[41] Akdemir KC et al. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat Genet 2020;52(3):294–305.
[42] Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res 2020;48(D1):D58–64.
[43] Quentmeier H et al. The LL-100 panel: 100 cell lines for blood cancer studies. Sci Rep 2019;9(1):1–14.
[44] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37(1):1–13.
[45] 1000 Genomes Project Consortium, et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
[46] Saunders CT et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 2012;28(14):1811–7.
[47] Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31(3):213–9.
[48] Weinhold N et al. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet 2014;46(11):1160–5.
[49] Haeussler M et al. The UCSC genome browser database: 2019 update. Nucleic Acids Res 2019;47(D1):D853–8.
[50] Zhou Y et al. Modeling and analysis of Hi-C data by HiSIF identifies characteristic promoter-distal loops. Genome Med 2020;12(1):69.
[51] Farooq A et al. HMST-Seq-Analyzer: A new python tool for differential methylation and hydroxymethylation analysis in various DNA methylation sequencing data. Comput Struct Biotechnol J 2020;18:2877–89.
[52] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30 (7):923–30.
[53] Hoffman MM et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res 2013;41(2):827–41.
[54] Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 2017;12(12):2478–92.
[55] Hoffman MM et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 2012;9(5):473–6.
[56] Akalin A et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 2012;13(10):1–9.
[57] Batmanov K et al. Integrative whole-genome sequence analysis reveals roles of regulatory mutations in BCL6 and BCL2 in follicular lymphoma. Sci Rep 2017;7.
[58] Batmanov K, Delabie J, Wang J. BayesPI-BAR2: A New Python Package for Predicting Functional Non-coding Mutations in Cancer Patient Cohorts. Front Genet 2019;10:282.
[59] Nordmann L, Pham H. Weighted voting systems. IEEE Trans Reliab 1999;48 (1):42–9.
[60] Huang, D.W., et al., *DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W169-75.
[61] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.
[62] Wang JB, Batmanov K. BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations. Nucleic Acids Res 2015;43(21).
[63] Zhu H, Wang GH, Qian J. Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet 2016;17(9):551–65.
[64] Luo C, Hajkova P, Ecker JR. Dynamic DNA methylation: In the right place at the right time. Science 2018;361(6409):1336–40.
[65] Paczkowska M et al. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun 2020;11(1):735.
[66] O'Brien TD et al. Weak sharing of genetic association signals in three lung cancer subtypes: evidence at the SNP, gene, regulation, and pathway levels. Genome Med 2018;10.
[67] Tsujimoto Y et al. Involvement of the bcl-2 gene in human follicular lymphoma. Science 1985;228(4706):1440–3.
[68] Dierlamm J et al. Gain of chromosome region 18q21 including the MALT1 gene is associated with the activated B-cell-like gene expression subtype and increased BCL2 gene dosage and protein expression in diffuse large B-cell lymphoma. Haematologica 2008;93(5):688–96.
[69] Wagner SD, Ahearne M, Ferrigno PK. The role of BCL6 in lymphomas and routes to therapy. Br J Haematol 2011;152(1):3–12.
[70] Petri A et al. Long noncoding RNA expression during human B-cell development. PLoS ONE 2015;10(9):e0138236.
[71] Pike BL et al. DNA methylation profiles in diffuse large B-cell lymphoma and their relationship to gene expression status. Leukemia 2008;22(5):1035–43.
[72] Rausch T et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012;28(18):i333–9.
[73] Zhan Y et al. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Res 2017;27(3):479–90.
[74] Rao SSP et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 2014;159(7):1665–80.
[75] Li Y, Hu M, Shen Y. Gene regulation in the 3D genome. Hum Mol Genet 2018;27(R2):R228–33.
[76] McArthur E, Capra JA. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. Am J Human Genet 2021;108(2):269–83.
[77] Finn LS et al. Primary follicular lymphoma of the testis in childhood. Cancer: Interdiscip Int J Am Cancer Soc 1999;85(7):1626–35.
[78] Vinca B. Molecular characteristics and prognostic significance of Bcl-2/IgH gene rearrangement in Serbian follicular lymphoma patients. Neoplasma 2008;55:5.
[79] Nagel, S., et al., *Co-Activation of HOXB7 and BCL2/MYC Via Biallelic IGH Rearrangements in a B-Cell Lymphoma Cell Line*. 2004, American Society of Hematology.
[80] Liu J et al. Upregulation of miR-29c-3p Hinders Melanoma Progression by Inhibiting CDCA4 Expression. Biomed Res Int 2021;2021.
[81] Klanova M, Klener P. BCL-2 proteins in pathogenesis and therapy of B-cell non-Hodgkin lymphomas. Cancers 2020;12(4):938.
[82] Gobeil S et al. A genome-wide shRNA screen identifies GAS1 as a novel melanoma metastasis suppressor gene. Genes Dev 2008;22(21):2932–40.
[83] Taylor JM, Ghorbel S, Nicot C. Genome wide analysis of human genes transcriptionally and post-transcriptionally regulated by the HTLV-I protein p30. BMC Genomics 2009;10(1):1–14.