


RESEARCH

Open Access



5-Hydroxymethylcytosine profiles of cfDNA are highly predictive of R-CHOP treatment response in diffuse large B cell lymphoma patients

Hang-Yu Chen^{1†}, Wei-Long Zhang^{2†}, Lei Zhang⁶, Ping Yang², Fang Li², Ze-Ruo Yang⁶, Jing Wang², Meng Pang², Yun Hong², Changjian Yan², Wei Li², Jia Liu², Nuo Xu¹, Long Chen¹, Xiu-Bing Xiao³, Yan Qin⁴, Xiao-Hui He⁴, Hui Liu⁵, Hai-Chuan Zhu⁸, Chuan He⁷, Jian Lin^{1*}  and Hong-Mei Jing^{2*}

Abstract

Background: Although R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) remains the standard chemotherapy regimen for diffuse large B cell lymphoma (DLBCL) patients, not all patients are responsive to the scheme, and there is no effective method to predict treatment response.

Methods: We utilized 5hmC-Seal to generate genome-wide 5hmC profiles in plasma cell-free DNA (cfDNA) from 86 DLBCL patients before they received R-CHOP chemotherapy. To investigate the correlation between 5hmC modifications and curative effectiveness, we separated patients into training ($n = 56$) and validation ($n = 30$) cohorts and developed a 5hmC-based logistic regression model from the training cohort to predict the treatment response in the validation cohort.

Results: In this study, we identified thirteen 5hmC markers associated with treatment response. The prediction performance of the logistic regression model, achieving 0.82 sensitivity and 0.75 specificity ($AUC = 0.78$), was superior to existing clinical indicators, such as LDH and stage.

Conclusions: Our findings suggest that the 5hmC modifications in cfDNA at the time before R-CHOP treatment are associated with treatment response and that 5hmC-Seal may potentially serve as a clinical-applicable, minimally invasive approach to predict R-CHOP treatment response for DLBCL patients.

Keywords: Epigenetics, 5-Hydroxymethylcytosine (5hmC), Diffuse large B cell lymphoma, R-CHOP, Logistic regression modeling

Introduction

Diffuse large B cell lymphoma (DLBCL) is the primary type of invasive lymphoid tissue tumor, accounting for about 30% of non-Hodgkin's lymphoma [1]. Although the majority of the DLBCL patients are elderly patients, this disease is found in all ages [2]. Since rituximab (R) joined cyclophosphamide, adriamycin, vincristine, and prednisone (CHOP) chemotherapy regimen ten years ago, the

*Correspondence: linjian@pku.edu.cn; hongmeijing@bjmu.edu.cn

[†]Hang-Yu Chen and Wei-Long Zhang have contributed equally to this work

¹ Synthetic and Functional Biomolecules Center, Beijing National Laboratory for Molecular Sciences, Key Laboratory of Bioorganic Chemistry and Molecular Engineering of Ministry of Education, College of Chemistry and Molecular Engineering, Innovation Center for Genomics, Peking University, Beijing 100871, People's Republic of China

² Department of Hematology, Lymphoma Research Center, Peking University Third Hospital, Beijing 100191, People's Republic of China
Full list of author information is available at the end of the article



overall survival rate of DLBCL patients has improved significantly [3].

However, 30–50% of patients are not sensitive to this standard treatment [4], and existing methods fail to predict the treatment response before R-CHOP treatment accurately or efficiently [5, 6]. Currently, positron emission tomography (PET)-CT is the gold standard to evaluate the efficacy of different treatment regimens for DLBCL. However, it is generally used after the treatment and thus cannot predict the treatment response [7]. The International Prognostic Index (IPI) is the primary prognostic risk assessment method for DLBCL, especially in high-risk patients, and is used for R-CHOP chemotherapies [8–10]. However, IPI cannot accurately predict the therapeutic effect of R-CHOP in DLBCL patients [11]. Furthermore, recent studies have demonstrated that the detection of the apoptosis inhibitor, survivin [12], activation-induced cytidine deaminase (AID) [13], plasma miRNA [14], exosome miRNA [15], and genes polymorphism [16, 17], as well as the presence of CD3 and FoxP3 in the immune microenvironment [18], were all potential indicators of treatment efficacy in DLBCL patients. However, these predictors showed contradictory results that have not been well solved. Therefore, an accurate and effective method to predict the response of R-CHOP regimen is highly necessary.

In recent years, cell-free DNA (cfDNA) in the circulating blood, which carries genetic and epigenetic information from cells of origin, has emerged as a promising noninvasive approach for the diagnosis and prognosis in cancer [19]. 5-Methylcytosines (5mCs) of DNA is an important epigenetic feature that plays an important role in gene expression and cancer development [20]. Kristensen et al. [21] found that the methylation of *DAPKI* in cfDNA from patients with DLBCL can be used to assess the effect of R-CHOP treatment. In the human genome, 5-methylcytosines (5mCs) in cfDNA are dynamic and reversible [22, 23] and can be oxidized into 5-hydroxymethylcytosines (5hmCs) through the ten-eleven translocation (TET) enzymes in an active DNA-demethylation process [24, 25]. Therefore, 5hmC, as an oxidation product of DNA demethylation (5mC), may also be used to assess the effect of R-CHOP treatment. Recently, a study has also shown that 5hmC is associated with the prognosis of DLBCL [26]. However, its role in the prediction of treatment response of R-CHOP scheme for DLBCL patients is not established.

In this study, we used 5hmC-Seal technique to obtain genome-wide 5hmC profiles in plasma cfDNA from 86 DLBCL patients, before they received R-CHOP chemotherapy. Our results demonstrated that responders and non-responders of R-CHOP treatment had distinct 5hmC profiles and that 5hmC markers selected by

bioinformatics tools and machine learning algorithms could be used to predict treatment response of R-CHOP treatment in DLBCL patients.

Materials and methods

Study participants

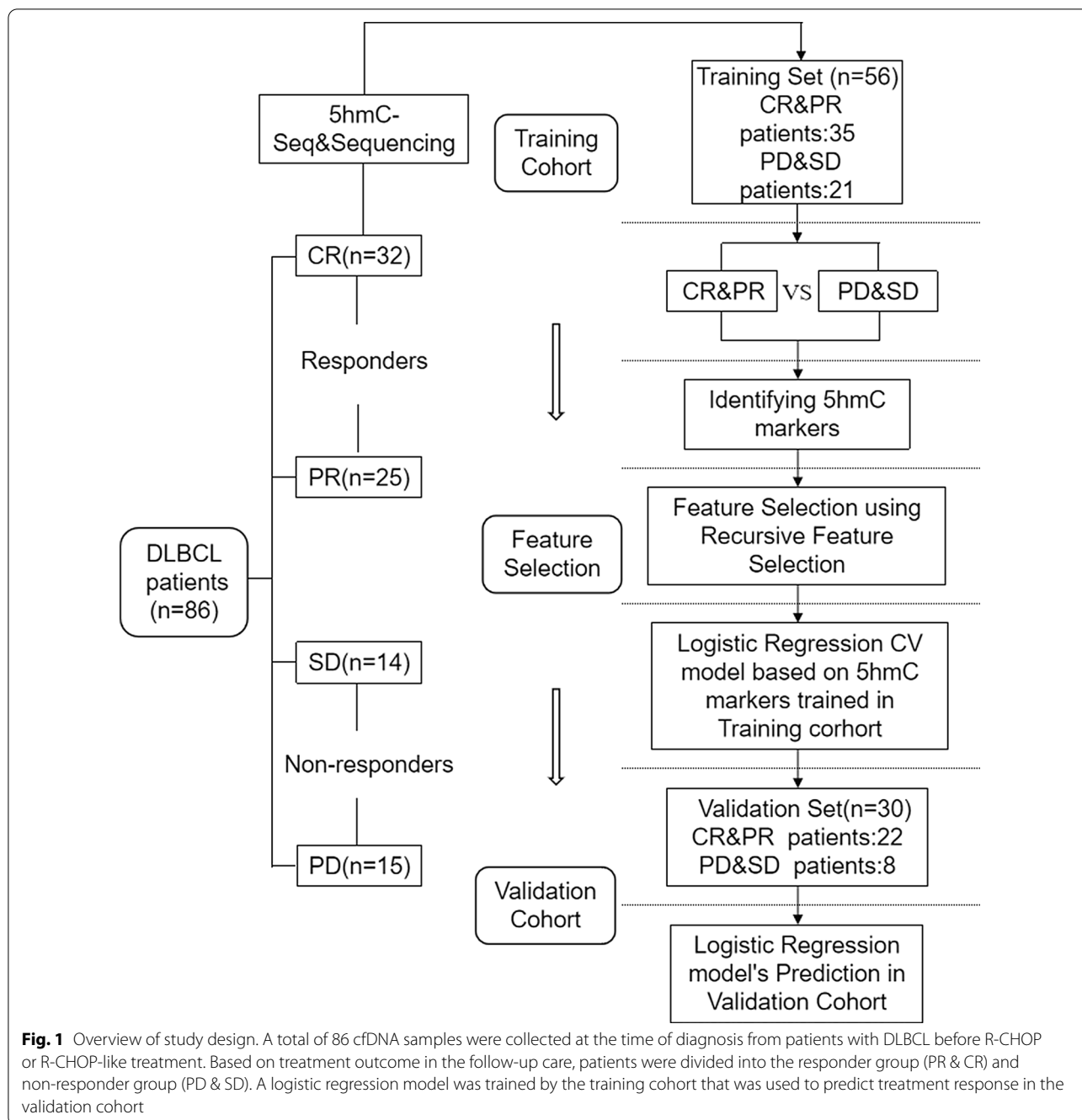
From 2017 to 2019, 86 diffuse large B cell lymphoma (DLBCL) patients from multicenter studies including Peking University Third Hospital, Fifth Medical Center of PLA General Hospital, and Cancer Hospital Chinese Academy of Medical Sciences were included in this study. All patients had signed the patient consent form. In all cases, the diagnosis of DLBCL was made using appropriate diagnostic criteria from the 2016 WHO classification of lymphoid tumors with combinations of histologic, immunohistochemical, and cell of origin (coo) defined according to the Hans algorithm [27]. Medical records were reviewed for demographic and clinical data. Laboratory tests, white blood cell count (WBC), renal and hepatic function examinations, lactate dehydrogenase (LDH), and β 2 microglobulin (β 2MG) and cfDNA from peripheral blood samples were collected before any treatment. Then, all patients received standard R-CHOP chemotherapy. Other baseline assessments including bone marrow biopsy and PET/CT were conducted in all patients in the follow-up care. The disease stage was defined by the Ann Arbor staging system. Treatment efficacy was evaluated after four cycles of treatment according to Lugano 2014 criteria [28], and patients were divided into PD (progressive disease), SD (stable disease), PR (partial response), and CR (complete response) based on the treatment outcome. This study was conducted in accordance with the Declaration of Helsinki.

Study design

This study aimed to discover 5hmC markers to predict the curative effectiveness of R-CHOP scheme through high-efficiency hmC-Seal technology. Among the 86 patients recruited, PR and CR patients were grouped as responders ($n=57$), and PD and SD patients were grouped as non-responders ($n=29$) to R-CHOP treatment. We split 86 patients into a training and validation cohort. The objective of the first part of the study was to screen candidate genes with differential 5hmC modifications in these two groups from the training cohort. The objective of the second part of the study was to predict treatment outcome, using the model developed in the first part, in the validation cohort (Fig. 1).

Clinical samples collection and cfDNA preparation

Eight milliliters of peripheral blood from DLBCL patients was collected into Cell-Free DNA Collection



Tubes (Roche). Within 24 h, plasma was prepared by centrifuging twice at 1350×g for 12 min at 4 °C and 13,500×g for 12 min at 4 °C. Then, the plasma samples were immediately stored at −80 °C. The plasma cfDNA was extracted using the Quick-cfDNA Serum & Plasma Kit (ZYMO) and then stored at −80 °C. The fragment size of all the cfDNA samples was verified by nucleic acid electrophoresis before library preparation.

5hmC library construction and high-throughput sequencing

5hmC libraries for all samples were constructed with high-efficiency hmC-Seal technology [29]. Due to the highly sensitive nature of the chemical labeling method, the input cfDNA can be as low as 1–10 ng. According to the requirements of next-generation sequencing, the cfDNA extracted from plasma was end-repaired, 3'-adenylated using the KAPA Hyper Prep Kit (KAPA

Biosystems), and then ligated with the Illumina compatible adapters. The ligated cfDNA was added in a glycosylation reaction in 25 μ L solution containing 50 mM HEPES buffer (pH 8.0), 25 mM $MgCl_2$, 100 μ M UDP-6-N3-Glc, and 1 μ M β -glucosyltransferase (NEB) for 2 h at 37 °C. Next, the cfDNA was purified using DNA Clean & Concentrator Kit (ZYMO). The purified DNA was incubated with 1 μ L of DBCO-PEG4-biotin (Click Chemistry Tools, 4.5 mM stock in DMSO) for 2 h at 37 °C. Similarly, the DNA was purified using the DNA Clean & Concentrator Kit (ZYMO). Meantime, 2.5 μ L streptavidin beads (Life Technologies) in 1 \times buffer (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl, and 0.2% Tween 20) was added directly to the reaction for 30 min at room temperature. Finally, the beads were subsequently washed eight times for five minutes with buffer 1–4. All binding and washing steps were performed at room temperature with gentle rotation. Then, the beads were resuspended in RNase-free water and amplified with 14–16 cycles of PCR amplification. The PCR products were purified using AMPure XP beads (Beckman), according to the manufacturer's instructions. The concentration of libraries was measured with a Qubit 3.0 fluorometer (Life Technologies). Paired-end 39-bp high-throughput sequencing was performed on the NextSeq 500 platform.

Mapping and identifying 5hmC-enriched regions

FastQC (version 0.11.5) was used to assess the sequence quality. Raw reads were aligned to the human genome (version hg19) with bowtie2 (version 2.2.9) [30] and further filtered with SAMtools (version 1.3.1) [31], (parameters used: SAMtools view -f 2 -F 1548 -q 30 and SAMtools rmdup) to retain unique non-duplicate matches to the genome. Pair-end reads were extended and converted into BedGraph format normalized to the total number of aligned reads using bedtools (version 2.19.1) [32], and then converted to bigwig format, using bedGraphToBigWig from the UCSC Genome Browser for visualization in the Integrated Genomics Viewer. Potential 5hmC-enriched regions (hMRs) were identified using MACS (version 1.4.2), and the parameters used were macs 14 -p 1e-3 -f BAM -g hs [33]. Peak calls were merged using bedtools merge, and only those peak regions that appeared in more than 10 samples and that were less than 1000 bp were retained. Blacklisted genomic regions that tend to show artifact signals, according to ENCODE, were also filtered. The hMRs for each patient were generated by intersecting the individual peak call file with the merged peak file. The hMRs within chromosome X and Y were excluded and used as an input for the downstream analyses.

Feature selection, model training, and validation

A two-step procedure was used to select optimal hMRs for distinguishing the non-responder group from the responder group prior to R-CHOP treatment. In step 1, DLBCL patients were randomly divided into training and validation cohorts in a stratified manner, using train_test_split in Scikit-Learn (version 0.22.1) [34] package in Python (version 3.6.10). In the training cohort, we identified differentially modified 5hmC regions (DhMRs) using EdgeR package (version 3.24.3) [35] in R (version 3.5.0), with the filtering threshold (p value < 0.01 & log₂FoldChange > 0.5). In step 2, the dhMRs were further filtered using the recursive feature elimination algorithm (RFECV) in Scikit-Learn (parameters used: estimator = LogisticRegressionCV (class_weight = 'balanced', cv = 2, max_iter = 1000), scoring = 'accuracy').

Then, we trained the logistic regression CV model (LR) with the features selected from step 2 (parameter used: max_iter = 100, method = "lbfgs"). The trained LR model was used to predict the treatment outcome for patients in the validation cohort. Receiver operating characteristics (ROC) analysis was used to evaluate model performance. Area under the curve (AUC), best cutoff point, sensitivity, and specificity were computed with sklearn.metrics module.

Exploring functional relevance of the 5hmC markers

We annotated the dhMRs from step 1 using the ChIP-seeker package (version 1.20.0) [36], and genes that were closest to the marker regions were used for the following functional analyses. The GO enrichment analysis (Biological Process) was done by the ClueGO (version 2.5.5) and CluePedia (version 1.5.5) plug-in from Cytoscape software (version 3.7.2) (parameters used: medium network specificity, Bonferroni step-down p V correction and two-sided hypergeometric test). We used the Search Tool for the Retrieval of Interacting Genes (STRING) database (version 10.0, <https://string-db.org>) to find protein–protein interactions for 5hmC markers. Then, the Cytoscape software was used to construct the PPI network.

Survival analysis and gene expression correlation analysis in TCGA-DLBC

For survival analysis, we downloaded the mRNA HTseq-FPKM data of 48 DLBCL patients from the TCGA-DLBC dataset [37] in the GDC Data Portal using gdc-client (version 1.5.0) and downloaded manually curated clinical data, including overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI) from UCSC Xena [38]. Survminer package (version 0.4.6) and Surviva packages (version 2.44-1.1) in R were used for survival analysis. Forty-eight

patients were divided into the high-expression group and low-expression group according to the cutoff points determined by the maximally selected rank statistics algorithm (maxstat) [39]. Survival analysis of each gene was assessed by Kaplan–Meier curves [40] and the log-rank test [41]. For the survival analysis, p value < 0.05 was considered statistically significant. For gene expression correlation analysis, we used a web tool called TIMER2.0 [42], which incorporated all TCGA expression data, to explore the mRNA expression relationship between 5hmC markers and other genes of interests in the TCGA-DLBC dataset. The correlation analysis was done using Spearman rank correlation.

Statistical analysis

For clinical data, continuous variables are presented as mean (SD) and categorical variables are presented as count (percentages). To understand the relationship between categorical/continuous variables and treatment outcome, Kruskal–Wallis test by ranks [43] and χ^2 test [44] were used, respectively. A two-sided p value of < 0.05 was considered to indicate statistical significance. The predictive power of clinical data was estimated by glm function in R-base and pROC package (version 1.15.3) in R.

Results

Clinical characteristics of Diffuse large B Cell lymphoma (DLBCL) patients

The clinical summary, including baseline characteristics and laboratory data, of all 86 patients is shown in Table 1. Of the 86 patients with DLBCL patients, 46 were male and 40 were female. The median age of all the patients was 54.6 years, and 63.9% of patients had advanced disease (including stage III and stage IV). Importantly, all patients were newly diagnosed with DLBCL and received standard R-CHOP chemotherapy. Treatment efficacy was evaluated in all patients after 4 cycles of treatment. According to the efficacy standard of Lugano 2014 criteria, the treatment response of patients was as follows: CR in 32 patients (37.2%), PR in 25 patients (29.1%), SD in 14 patients (16.3%), and PD in 15 patients (17.4%). Besides, according to the Hans model, 23 patients (26.7%) were germinal center B cell (GCB), 61 patients (70.9%) had non-GCB and 2 patients (2.3%) had an unknown cell of origin. The results of the international prognostic index (IPI) score showed that 52.3% of patients (IPI score > 2) belonged to the high–intermediate-risk/high-risk group. Finally, the mean of WBC, LDH, and $\beta 2$ MG for all patients was $6.94 \times 10^4/L$, 364.33 U/L, and 2.84 mg/L, respectively.

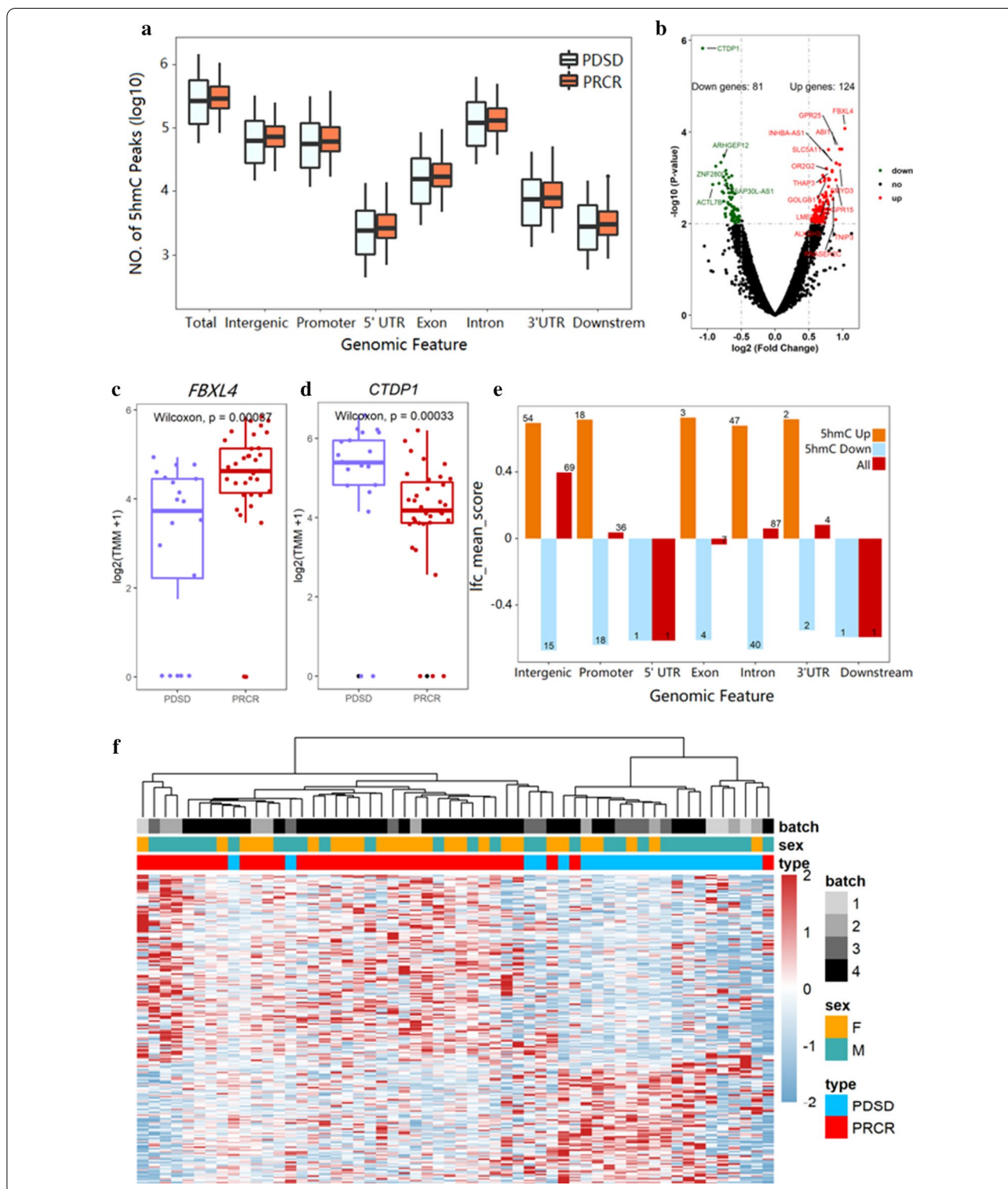
Table 1 Diffuse large B cell lymphoma (DLBCL) patient characteristics

Characteristics	Level/type	Value
<i>n</i>		86
Sex (%)	F	40 (46.5)
	M	46 (53.5)
Age (mean (SD))		54.59 (15.56)
Diagnosis (%)	DLBCL	86 (100.0)
Therapy (%)	R-CHOP	86 (100.0)
Response (%)	CR	32 (37.2)
	PD	15 (17.4)
	PR	25 (29.1)
	SD	14 (16.3)
Ann Arbor stage (%)	I	6 (7.0)
	II	18 (20.9)
	III	7 (8.1)
	IV	48 (55.8)
	Unknown	7 (8.1)
Cell of origin (%)	GCB	23 (26.7)
	Non-GCB	61 (70.9)
	Unknown	2 (2.3)
IPI (%)	0	8 (9.3)
	1	12 (14.0)
	2	18 (20.9)
	3	28 (32.6)
	4	15 (17.4)
	5	2 (2.3)
	Unknown	3 (3.5)
Mean LDH (SD)		364.33 (326.72)
Mean $\beta 2$ MG (SD)		2.84 (2.77)

IPI, International Prognostic Index; GCB, germinal center B cell; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease; LDH, lactate dehydrogenase; $\beta 2$ MG, beta2 microglobulin; WBC, white blood cell

5hmC profiles differ between responders and non-responders to R-CHOP treatment in the training cohort

Eighty-six DLBCL patients were randomly divided into the training cohort ($n = 56$) and validation cohort ($n = 30$) (Fig. 1). We used hmC-Seal to generate genome-wide 5hmC profiles for patients in the training set, including 35 responders and 21 non-responders to R-CHOP treatment. The overall 5hmC enrichment (all hMRs) was most common in intronic, intergenic, and promoter regions for both responders and non-responders, even though no statistically significant difference was found between these two groups for any genomic feature types (Fig. 2a). Meanwhile, we conducted differential analysis (EdgeR; $p < 0.01$, fold change > 0.5) and observed 205 DhMRs, including upregulate ($n = 124$) and downregulate ($n = 81$) regions in responders compared to non-responders (Fig. 2b). For instance, *FBXL4* (Fig. 2c) was



highly enriched in hydroxymethylation for responders ($p=0.00087$), and *CTDPI* (Fig. 2d) was highly enriched in hydroxymethylation for non-responders ($p=0.00033$). In addition, for the top 205 DhMRs, the most significant enrichment was found in intronic, intergenic, and

promoter regions, consistent with previous studies [45, 46] (Fig. 2e). Finally, heatmap results, using default clustering methods, demonstrated that these 205 DhMRs could effectively separate responders from non-responders (Fig. 2f).

(See figure on previous page.)

Fig. 2 Characteristics of 5hmC distribution in plasma cfDNA of DLBCL patients in the training cohort ($n = 56$). **a** Genome-wide 5hmC distribution in different genomic features grouped by R-CHOP treatment response (PDS vs PRCR). **b** Volcano plot. Significantly altered genes (\log_2 Foldchange ≥ 0.5 ; p value < 0.01) are highlighted in red (up) or green (down) using the responder group (PRCR) as the reference ($n = 205$). Black dots represent the genes that are not differentially expressed. **c, d** Boxplots of *FBXL4* and *CTDP1* grouped by treatment response (PDS vs PRCR). Log₂ transformed of TMM normalized 5hmC enrichment values were plotted, and the Wilcoxon t test was used. **e** Mean log₂ Foldchange value of 205 DhMRs across different genomic features (Orange for 124 5hmC-up DhMRs, blue for 81 5hmC-down DhMRs, red for all 205 DhMRs). **f** Heatmap of 205 DhMRs markers with treatment response, batch, and sex information labeled. Unsupervised hierarchical clustering was performed across genes and samples

Pathway analysis and function exploration

Pathway analysis of 205 5hmC markers (Additional file 1: Table 1) in DLBCL patients suggested functional enrichment in certain canonical pathways. The top enriched

GO biological pathways included signaling like alpha-beta T cell differentiation, protein-lysine N-methyltransferase activity, and histone H3-K9 modification (Fig. 3a). Among these pathways, signaling by alpha-beta T cell

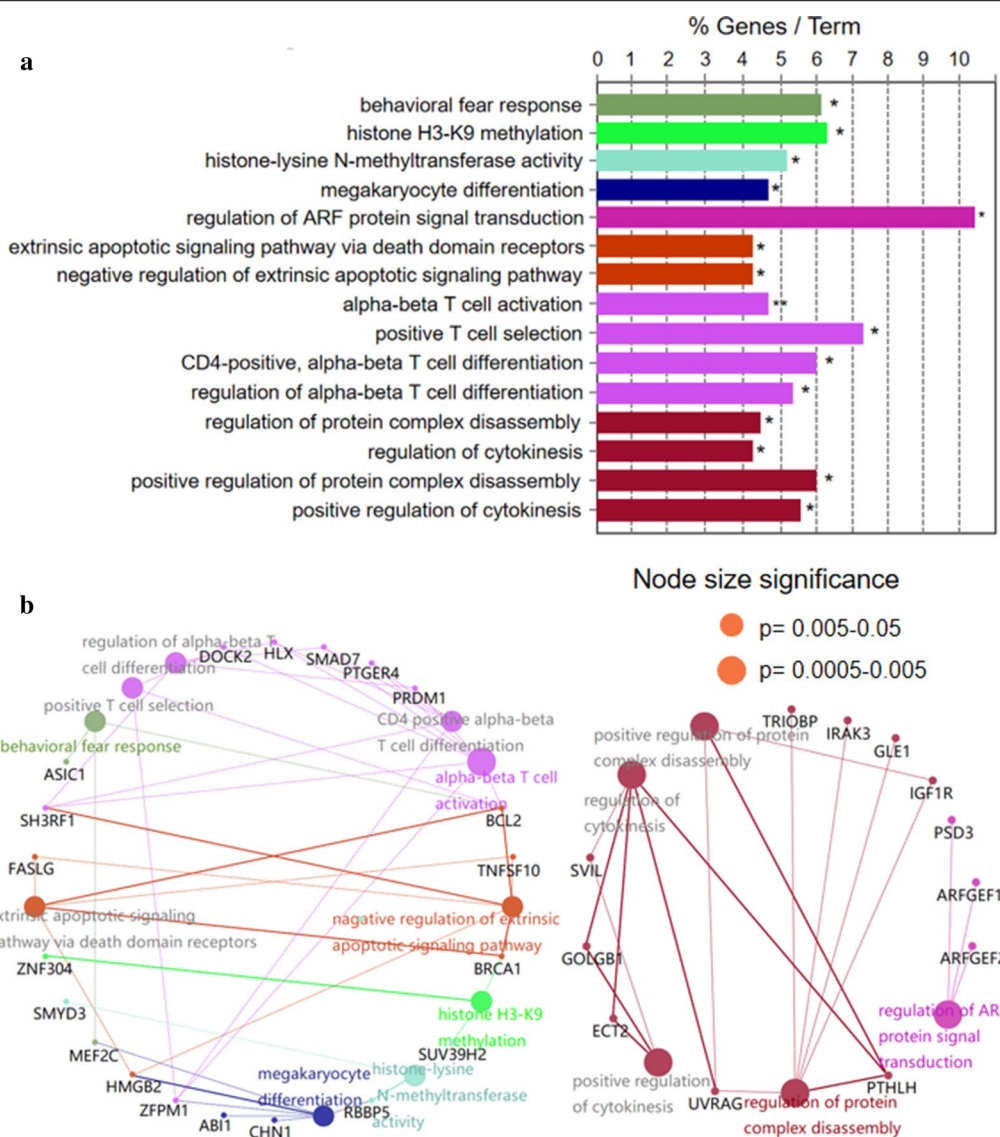


Fig. 3 GO enrichment analysis and function exploration of 205 5hmC markers using Cytoscape software. **a** GO enrichment bar plot ($*p = 0.005-0.05$, $**p = 0.0005-0.005$). **b** GO enrichment and Gene-Concept Network. The node size is proportional to the p value calculated from the network

differentiation was known to be relevant to tumor growth and apoptosis, which suggested that the DhMRs might be involved in the immunity system [47–49]. Meanwhile, the hubs of the GO functional interaction networks (Fig. 3b) showed that these genes, including BCL2 apoptosis regulator (*BCL2*), PR/SET domain 1 (*PRDM1*), prostaglandin E receptor 4 (*PTGER4*), SMAD family member 7 (*SMAD7*), H2.0 like homeobox (*HLX*), dedicator of cytokinesis 2 (*DOCK2*) and SH3 domain containing ring finger 1 (*SH3RF1*), participated in the regulating T cell activation and differentiation pathway.

5hmC markers showed prediction performance superior to clinical indicators for R-CHOP treatment response

Similarly, we generated genome-wide 5hmC profiles for patients in the validation set, including 22 responders and 8 non-responders to R-CHOP treatment. By using the recursive feature elimination algorithm based on the logistic regression CV estimator, we further reduced the number of 5hmC markers from 205 to 13, which achieved the best cross-validation score (Additional file 2: Figure S1). Further, we found that the 13 5hmC markers (Table 2), selected by the LR model, could distinguish responders from non-responders in both the training and validation cohorts (Fig. 4a, b). Meantime, these 13 5hmC markers could effectively predict responders and non-responders to R-CHOP treatment in the training (AUC = 1.00) and the validation cohorts (AUC = 0.78) (Fig. 4c), achieving 0.82 sensitivity and 0.75 specificity in the validation cohort (Fig. 4d). Finally, we also calculated the individual AUC for each of the 13 5hmC markers in the training and validation cohorts (Additional file 2: Figure S2A, B). Among these, *ARHGEF12* and *ZNF280D*

showed the best predictive performance, yielding an AUC of 0.76 in the validation cohort.

We also investigated the association between available clinical indicators, including stage, pathology, IPI, LDH, β 2MG and WBC, and R-CHOP treatment response. Among all those clinical indicators, only LDH (continuous variable, $p=0.03474$) and stage (categorical variable, $p=0.004453$) showed a significant association with treatment response (Additional file 2: Table 3). Thus, we used these two indicators to build logistic regression models to predict treatment response. As expected, LDH level, stage, and LDH combined with stage (LDH + stage) could also predict treatment response to a certain level. However, the AUC of LDH level (AUC = 0.646), stage (AUC = 0.658), and LDH combined with stage (AUC = 0.669) were lower than that of 5hmC markers (AUC = 0.78) (Fig. 4e).

Potential associations between 5hmC markers and R-CHOP treatment response in DLBCL patients

To further understand the potential associations between those 13 5hmC-modified marker genes and R-CHOP treatment response, we investigated their mRNA expression profiles and compared them to that of B-lymphocyte antigen CD20 (*MS4A1*), a rituximab target gene, in 48 DLBCL patients from the TCGA-DLBC dataset. Among those 13 marker genes, we found that the mRNA expression of *MS4A1* was positively correlated with the mRNA expression of *ARHGEF12* ($\rho=0.385$), *FBXL4* ($\rho=0.376$), *GOLGB1* ($\rho=0.434$), *LMBR1* ($\rho=0.45$) (Additional file 2 Figure S3A–D). We decided to further investigate the potential mechanism of Rho Guanine Nucleotide Exchange Factor 12 (*ARHGEF12*),

Table 2 Coefficients for 13 5hmC markers in the logistic regression model trained by the training cohort

Markers	GeneID	Coefficients	SE	z value	p value
Intercept		-5.5704	0.867	2.652	<0.01
chr1_6721489_6721898	THAP3	0.7712	0.145	1.865	<0.05
chr1_246290825_246291238	SMYD3	0.39	0.149	1.955	<0.05
chr1_247755954_247756505	OR2G2	3.1779	0.108	1.344	<0.05
chr11_43905400_43905804	ALKBH3	3.3423	0.128	1.306	<0.05
chr11_65511519_65512429	RNASEH2C	1.5211	0.072	2.061	<0.05
chr11_120211662_120212234	ARHGEF12	-3.8797	0.115	-3.225	<0.001
chr15_56982146_56982638	ZNF280D	-1.2266	0.177	-3.250	<0.001
chr16_24916341_24916920	SLC5A11	0.6683	0.076	0.149	<0.05
chr18_77500908_77501376	CTDP1	-2.573	0.103	-3.182	<0.001
chr3_98270705_98271079	GPR15	0.1052	0.167	2.348	<0.05
chr3_121430838_121431239	GOLGB1	0.8526	0.178	2.982	<0.01
chr6_99461404_99461922	FBXL4	1.7188	0.101	2.165	<0.05
chr7_156700537_156701031	LMBR1	1.0942	0.078	0.579	<0.05

SE, standard errors of coefficients; z value, Wald z-statistic value

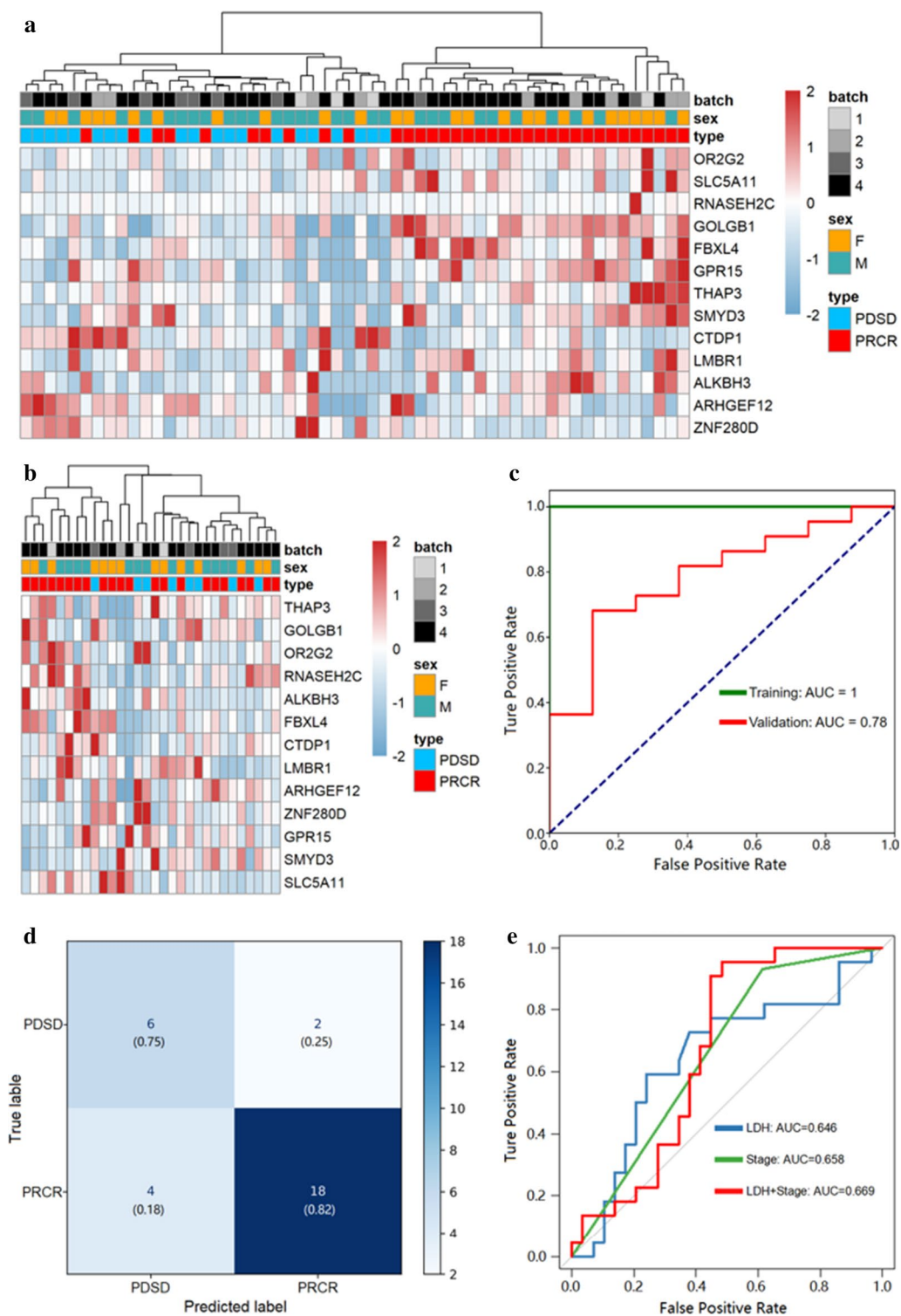


Fig. 4. 5hmC markers' prediction for treatment response in the training and validation cohort. **a, b** Heatmaps of 13 5hmC markers with treatment response, batch and sex information labeled in the training and validation cohorts. Unsupervised hierarchical clustering was performed across genes and samples. **c** Receiver operating characteristic (ROC) curve of the classification model with 13 5hmC markers in the training and validation cohorts. The true-positive rate (sensitivity) is plotted in function of the false-positive rate (1-specificity). **d** Confusion matrix that shows the model performance in the validation cohort (responders: 22, non-responders:8). **e** ROC curve of the classification model with LDH, stage and LDH combined with stage for DLBCL patients

for its mRNA expression was positively associated with *MS4A1*, and it achieved the highest AUC in the validation cohort among 13 5hmC-modified marker genes. According to recent studies, 5hmC enrichment in promoter regions was positively associated with gene expression levels [25, 50]. In our study, *ARHGEF12* was highly enriched in hydroxymethylation in the non-responders ($p=0.022$) (Fig. 5a), and the hydroxymethylation site was in the promoter region (Additional file 3: Table 2). Therefore, we speculated that the change in 5hmC enrichment in the promoter region of *ARHGEF12* might lead to the change in the mRNA expression of this gene.

In addition, from the PPI network constructed from the STRING database, we identified several genes linked to *ARHGEF12*, including Ras Homolog Family Member A (*RHOA*), Ras Homolog Family Member B (*RHOB*), Ras Homolog Family Member C (*RHOC*), Cell Division Cycle 42 (*CDC42*), Rho Associated Coiled-Coil Containing Protein Kinase 1 (*ROCK1*), G Protein Subunit Alpha 12 (*GNA12*) and G Protein Subunit Alpha 13 (*GNA13*) (Fig. 5b). Interestingly, we found that all of these gene expressions (*RHOA* ($\rho=0.667$), *RHOB* ($\rho=0.604$), *CDC42* ($\rho=0.676$), *ROCK1* ($\rho=0.832$), *GNA12* ($\rho=0.721$), *GNA13* ($\rho=0.784$)) were highly positively associated with that of *ARHGEF12* (Fig. 5c–h). Moreover, from survival analysis results in the TCGA-DLBC dataset, we found that the overall survival time (OS, days) of patients with high expression of *ARHGEF12* and *CDC42* was significantly lower than that of patients with low expression in these 2 genes (Fig. 5i, j). Also, we found that the mRNA expression of *ARHGEF12* was positively associated with several immune-related genes, such as *CD44*, *CD47*, *CD53*, *CD59*, and *CD274* (Additional file 2: Figure S4A–E). Finally, we conducted a GO enrichment analysis (Fig. 5k) for all the genes associated with *ARHGEF12* (Fig. 5b) and found that the main GO enrichment was in the Rho signaling pathway which was consistent with the PPI network constructed from the STRING database.

Discussion

Even though previous studies have reported that 5hmC modifications could serve as potential diagnostic and prognostic markers in DLBCL patients [26], its role in

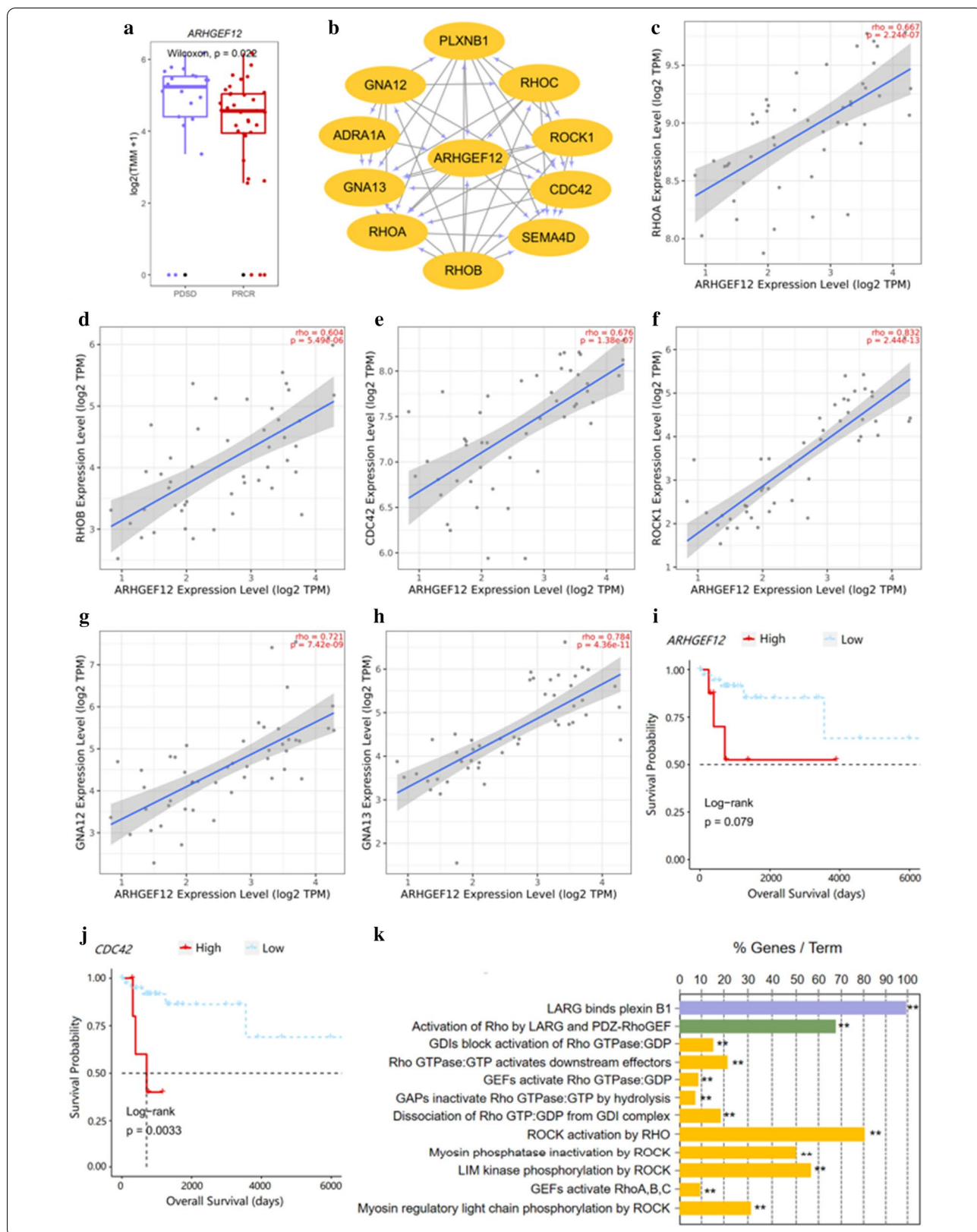
the prediction of treatment response of R-CHOP scheme was not fully studied. Therefore, an accurate, noninvasive prediction test for treatment response of R-CHOP is highly desirable, and to this end, the emergence of liquid biopsy technology has shown to be a promising approach. In this study, we aimed to develop a model to predict R-CHOP scheme treatment response for DLBCL patients based on the 5hmC profiles derived from plasma cfDNA before R-CHOP treatment using hmC-Seal sequencing method.

In our cohort, we found that responders and non-responders to R-CHOP scheme had distinctive differences in 5hmC enrichment, containing 205 DhMRs detected by differential analysis method. Additionally, pathway analysis of the 205 marker genes with differentially modified 5hmC between responders and non-responders suggested enrichment in alpha–beta T cell activation and differentiation signaling pathway. As we all known, tumor progression and drug resistance are highly associated with the physiological state of the tumor microenvironment (TME), and thus, the tumor microenvironment (TME) represents an attractive therapeutic target and closely related to the curative effect of tumor therapy [47]. The composition of tumor microenvironment is complex, which mainly include tumor cells, stromal elements, extracellular matrixes, inflammation, and immune cells [48], which are closely related to tumor development, metastasis, and tumor therapy [49]. Importantly, cfDNA is not only derived from tumor cells, but also from the tumor microenvironment [51]. Therefore, these 5hmC marker genes could be related to the effect of R-CHOP treatment.

Furthermore, we found that 13 5hmC markers filtered by machine learning algorithms could well distinguish non-responders from responders in both the training and validation cohorts. Meantime, the prediction performance of the logistic regression model, established by 13 5hmC markers, achieving 0.82 sensitivity and 0.75 specificity (AUC=0.78), was superior to existing clinical indicators, such as LDH (AUC=0.646) and stage (AUC=0.658). Furthermore, when combining the LDH and stage, the AUC was also lower than 13 5hmC markers. Taken together, these findings indicated that 5hmC markers derived from cfDNA may serve as effective biomarkers for

(See figure on next page.)

Fig. 5 *ARHGEF12* and its potential relevance in DLBCL patients and treatment response. **a** Boxplot of *ARHGEF12* grouped by treatment response (PDS vs PRCR). Log₂ transformed of TMM normalized 5hmC enrichment values were plotted, and Wilcoxon *t* test was used. **b** Functional protein–protein interaction networks (PPI) from the STRING database. **c–h** Correlation plots of the mRNA expression of *ARHGEF12* with the mRNA expressions of genes in the RHO pathway, including *RHOA*, *RHOB*, *CDC42*, *ROCK1*, *GNA12* and *GNA13* in DLBCL in the TCGA-DLBC dataset. **i, j** Overall survival curves of DLBCL patients with low or high gene expressions in *ARHGEF12* or *CDC42* in the TCGA-DLBC dataset. The x-axis represents the OS time (days), and the y-axis represents the survival probability. **(k)** GO enrichment bar plot for genes associated with *ARHGEF12* as shown in the PPI network (* $p=0.005-0.05$, ** $p=0.0005-0.005$)



minimally noninvasive prediction for treatment response of DLBCL patients with R-CHOP scheme.

According to recent studies, 5hmC enrichment in promoter regions can promote gene transcription [25]. In our study, the hydroxymethylation of *ARHGEF12* is enriched in the promoter region in non-responders. Notably, among 13 5hmC marker genes, *ARHGEF12* showed the best predictive performance, and its mRNA expression was positively associated with that of *MS4A1* in the TCGA-DLBC dataset. Meantime, *ARHGEF12* expression was highly positively correlated with Rho-related genes, such as *RHOA*, *RHOB*, *CDC42*, *ROCK1*, *GNAI2*, and *GNAI3*. Previous research suggested that Rho signaling pathway was linked to cancer microenvironment, cancer initiation, proliferation, and metastasis, and might incorporate novel biological implications and therapeutic opportunities [52]. These genes related to *ARHGEF12* all play crucial roles in the Rho signaling pathway, and their functions in cancer initiation, proliferation, metastasis, and drug resistance are well supported by previous research [52–57]. In addition, the potential functions of *ARHGEF12* were also reported in various researches. For instance, *ARHGEF12* is a well-studied activator of Rho signaling downstream of G-protein-coupled receptors (GPCRs) and has essential roles in chemokine-driven tumor cell invasion [58, 59]. More importantly, *ARHGEF12* expression was also positively associated with immune-related genes, such as *CD44*, *CD47*, *CD53*, *CD59*, and *CD274*. Therefore, we suspected that, like its related genes, *ARHGEF12* was crucial in the Rho signaling pathway and might be related to diffuse large B cell lymphoma initiation, proliferation, metastasis and treatment, but the deep biological basis of these relationships needs further study. Taken together, the above evidence provided by previous research suggested that *ARHGEF12* might serve as a potential drug target that was related to the treatment response of R-CHOP treatment.

Nevertheless, this study still has some limitations. First, the sample size is relatively small and may not fully represent all DLBCL patients. The performance of our model still needs to be tested in larger study cohorts. Second, this study only focuses on Chinese patients and may not represent DLBCL patients in other races. Thirdly, the regulatory mechanism of 5hmC in *ARHGEF12* and its relevance in R-CHOP treatment effectiveness are still not clear. Thus, further studies are required. In the future, we aim to increase the sample size of DLBCL patients and find more stable and reliable 5hmC marker genes to predict the treatment response of R-CHOP scheme.

In conclusion, our results suggested that 5hmC markers derived from plasma cfDNA can be used to predict treatment response of DLBCL patients treated with

R-CHOP scheme. Meanwhile, hmC-Seal might serve as a minimally noninvasive technique to unveil potential drug targets related to the treatment response of R-CHOP in DLBCL patients.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13148-020-00973-8>.

Additional file 1. 205 5hmC markers.anoation.

Additional file 2. Supplementary material.

Additional file 3. 13 5hmC feature markers.anoation.

Abbreviations

R-CHOP: Rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone; DLBCL: Diffuse large B cell lymphoma; IPI: International Prognostic Index; AID: Activation-induced cytidine deaminase; 5hmC: 5-Hydroxymethylcytosine; 5mC: 5-Methylcytosine; TET: Ten-eleven translocation; cfDNA: Cell-free DNA; WBC: White blood cell count; LDH: Lactate dehydrogenase; β 2MG: β 2 Microglobulin; PD: Progressive disease; SD: Stable disease; PR: Partial response; CR: Complete response; hMRs: 5hmC-enriched regions; RFECV: Recursive feature elimination algorithm; ROC: Receiver operating characteristic; AUC: The area under ROC curves; STRING: Search Tool for the Retrieval of Interacting Genes; OS: Overall survival; DSS: Disease-specific survival; DFI: Disease-free interval; PFI: Progression-free interval.

Acknowledgments

We would like to acknowledge the essential contributions of all staffs and students who participated in this work.

Authors' contributions

H-YC and W-LZ conceived the study and designed the experiments. H-YC performed the experiments with the help from LZ. H-YC analyzed data with help from Z-RY, PY, FL, JW, MP, YH, CY, WL, JL, X-BX, YQ, X-HH, and HL recruited patients, collected blood, and organized clinical information. H-YC, W-LZ, and Z-RY wrote the manuscript with input and comments from PY, NX, and LC. All authors read and approved the final manuscript.

Funding

This research was funded by Beijing Natural Science Foundation (7132183 and 7182178), China Health Promotion Foundation (CHPF-zlkysx-001), Key Clinical Projects of Peking University Third Hospital (BYSYZD2019026), Natural Science Foundation of China (81800195, 31471299 and 81522046), and National Science and Technology Major Projects for "Major New Drugs Innovation and Development" (No. 2018ZX09711003).

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. All other datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

The study was conducted according to the guidelines of the Helsinki Declaration and was approved by the Ethics Committee of Peking University Third Hospital. Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interests.

Author details

¹ Synthetic and Functional Biomolecules Center, Beijing National Laboratory for Molecular Sciences, Key Laboratory of Bioorganic Chemistry and Molecular

Engineering of Ministry of Education, College of Chemistry and Molecular Engineering, Innovation Center for Genomics, Peking University, Beijing 100871, People's Republic of China. ² Department of Hematology, Lymphoma Research Center, Peking University Third Hospital, Beijing 100191, People's Republic of China. ³ Lymphoma Head and Neck Oncology, Fifth Medical Center of PLA General Hospital, Beijing 100039, People's Republic of China. ⁴ Department of Medical Oncology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, People's Republic of China. ⁵ Department of Hematology, Beijing Hospital, National Center of Gerontology, Beijing 1000730, People's Republic of China. ⁶ Yang Sheng Tang Natural Medicine Research Institute, Hangzhou 310024, People's Republic of China. ⁷ Department of Chemistry, University of Chicago, Chicago, IL 60637, USA. ⁸ Institute of Biology and Medicine, College of Life and Health 20 Sciences, Wuhan University of Science and Technology, Hubei 430081, People's Republic of China.

Received: 22 July 2020 Accepted: 9 November 2020

Published online: 11 February 2021

References

- Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. *J Clin Oncol*. 2014;32:3059–67.
- Oschlies I, et al. Diffuse large B-cell lymphoma in pediatric patients belongs predominantly to the germinal-center type B-cell lymphomas: a clinicopathologic analysis of cases included in the German BFM (Berlin–Frankfurt–Munster) Multicenter Trial. *Blood*. 2006a;107(10):4047–52.
- Fu K, et al. Addition of rituximab to standard chemotherapy improves the survival of both the germinal center B-cell-like and non-germinal center B-cell-like subtypes of diffuse large B-cell lymphoma. *J Clin Oncol*. 2008;26(28):4587–94.
- Coiffier B, Sarkozy C. Diffuse large B-cell lymphoma: R-CHOP failure—what to do? *American Society of Hematology*. 2016;1:366–78.
- Oschlies I, et al. Diffuse large B-cell lymphoma in pediatric patients belongs predominantly to the germinal-center type B-cell lymphomas: a clinicopathologic analysis of cases included in the German BFM (Berlin–Frankfurt–Munster) Multicenter Trial. *Blood*. 2006b;107(10):4047–52.
- Kim SH, et al. Prognostic impact of pretreatment albumin to globulin ratio in patients with diffuse large B-cell lymphoma treated with R-CHOP. *Leuk Res*. 2018;71:100–5.
- Barrington SF, Mikhael NG. PET scans for staging and restaging in diffuse large B-cell and follicular lymphomas. *Curr Hematologic Malig Rep*. 2016;11(3):185–95.
- International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med*. 1993;329(14):987–94.
- Flowers CR, Sinha R, Vose JM. Improving outcomes for patients with diffuse large B-cell lymphoma. *CA Cancer J Clin*. 2010;60(6):393–408.
- Zhou Z, et al. An enhanced International Prognostic Index (NCCN-IPI) for patients with diffuse large B-cell lymphoma treated in the rituximab era. *Blood*. 2014;123(6):837–42.
- Deng Y, et al. EZH2/Bcl-2 coexpression predicts worse survival in diffuse large B-cell lymphomas and demonstrates poor efficacy to rituximab in localized lesions. *J Cancer*. 2019;10(9):2006–17.
- Markovic O, et al. Survivin expression in patients with newly diagnosed nodal diffuse large B cell lymphoma (DLBCL). *Med Oncol*. 2012;29(5):3515–21.
- Arima H, et al. Prognostic impact of activation-induced cytidine deaminase expression for patients with diffuse large B-cell lymphoma. *Leuk Lymphoma*. 2018;59(9):2085–95.
- Song G, et al. Serum microRNA expression profiling predict response to R-CHOP treatment in diffuse large B cell lymphoma patients. *Ann Hematol*. 2014;93(10):1735–43.
- Feng Y, et al. Exosome-derived miRNAs as predictive biomarkers for diffuse large B-cell lymphoma chemotherapy resistance. *Epigenomics*. 2019;11(1):35–51.
- Jin X, et al. Homozygous A polymorphism of the complement C1qA276 correlates with prolonged overall survival in patients with diffuse large B cell lymphoma treated with R-CHOP. *J Hematol Oncol*. 2012;5(1):51.
- Kim DH, et al. FCGR3A gene polymorphisms may correlate with response to frontline R-CHOP therapy for diffuse large B-cell lymphoma. *Blood*. 2006;108(8):2720–5.
- Coutinho R, et al. Revisiting the immune microenvironment of diffuse large B-cell lymphoma using a tissue microarray and immunohistochemistry: robust semi-automated analysis reveals CD3 and FoxP3 as potential predictors of response to R-CHOP. *Haematologica*. 2015;100(3):363–9.
- Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer*. 2011;11:426–37.
- Dahl C, Gronbaek K, Guldborg P. Advances in DNA methylation: 5-hydroxymethylcytosine revisited. *Clin Chim Acta*. 2011;412(11–12):831–6.
- Kristensen LS, et al. Aberrant methylation of cell-free circulating DNA in plasma predicts poor outcome in diffuse large B cell lymphoma. *Clin Epigenetics*. 2016;8(1):95–95.
- Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*. 2009;324:930–5.
- Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*. 2009;324:929–30.
- Fu Y, He C. Nucleic acid modifications with epigenetic significance. *Curr Opin Chem Biol*. 2012;16:516–24.
- Branco MR, Ficiz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet*. 2011;13:7–13.
- Chiu BCH, et al. Prognostic implications of 5-hydroxymethylcytosines from circulating cell-free DNA in diffuse large B-cell lymphoma. *Blood Adv*. 2019;3(19):2790–9.
- Fang C, Xu W, et al. A systematic review and meta-analysis of rituximab-based immunochemotherapy for subtypes of diffuse large B cell lymphoma. *Ann Hematol*. 2010;89:1107–13.
- Van Heertum RL, Scarimbolo R, et al. Lugano 2014 criteria for assessing FDG-PET/CT in lymphoma: an operational approach for clinical trials. *Drug Des Dev Therapy*. 2014;2017(11):1719–28.
- Song CX, Szulwach KE, Fu Y, et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol*. 2011;29:68–72.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
- Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):137.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
- Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9:R137.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
- McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multi-factor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012;40(10):4288–97.
- Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*. 2015;31(14):2382–3.
- Schmitz R, Wright GW, Huang DW, Johnson CA, Phelan JD, Wang JQ, Roulland S, Kasbekar M, Young RM, Shaffer AL, Hodson DJ. Genetics and pathogenesis of diffuse large B-cell lymphoma. *N Engl J Med*. 2018;378(15):1396–407.
- Goldman M, Craft B, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. *BioRxiv* 2018; 326470.
- Hothorn T, Hothorn MT, Suggests TH. Package 'maxstat'. 2017.
- Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan–Meier estimate. *Int J Ayurveda Res*. 2010;1(4):274.
- Marubini E, Valsecchi MG. Estimation of survival probabilities. Analysing survival data from clinical trials and observational studies. Chichester: Wiley; 1995. p. 41–8.

42. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 2017;77(21):e108–10.
43. McKight PE, Najab J. Kruskal–Wallis test. *The Corsini Encyclopedia of Psychology* 2010;1-1.
44. Cochran WG. The χ^2 test of goodness of fit. *Ann Math Stat.* 1952;23:315–45.
45. Applebaum MA, Barr EK, et al. 5-Hydroxymethylcytosine profiles are prognostic of outcome in neuroblastoma and reveal transcriptional networks that correlate with tumor phenotype. *JCO Precis Oncol.* 2019;3:1–12.
46. Zhang Ji, Han X, et al. 5-hydroxymethylome in circulating cell-free DNA as A potential biomarker for non-small-cell lung cancer. *Genomics Proteomics Bioinformatics.* 2018;16:187–99.
47. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. *Nat Med.* 2013;19(11):1423–37.
48. Alexandra IC, Patricia IS, Liana S, et al. Tumor microenvironment in diffuse large B-cell lymphoma: role and prognosis. *Anal Cell Pathol.* 2019;2019:8586354.
49. Catarina RR, Rita M. Targeting tumor microenvironment for cancer therapy. *Int J Mol Sci.* 2019;20:840.
50. Song C-X, et al. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res.* 2017;27(10):1231–42.
51. Swaminathan R, Butt AN. Circulating nucleic acids in plasma and serum: recent developments. *Ann NY Acad Sci.* 2006;1075:1–9.
52. Venessa T, Adnan M, et al. Rho-associated kinase signalling and the cancer microenvironment: novel biological implications and therapeutic opportunities. *Expert Rev Mol Med.* 2015;17:1–14.
53. Müller PM, Rademacher J, Bagshaw RD, et al. Systems analysis of RhoGEF and RhoGAP regulatory proteins reveals spatially organized RAC1 signaling from integrin adhesions. *Nat Cell Biol.* 2020;22(4):498–511.
54. Patrick K, Stemmler LN, Madden JF, et al. A role for the G12 family of heterotrimeric G proteins in prostate cancer invasion. *J Biol Chem.* 2006;281(36):26483–90.
55. Kelly P, et al. The G12 family of heterotrimeric G proteins promotes breast cancer invasion and metastasis. *PNAS.* 2006;103(21):8173–8.
56. Yuan Bo, et al. G α 12/13 signaling promotes cervical cancer invasion through the RhoA/ROCK-JNK signaling axis. *Biochem Biophys Res Commun.* 2016;473(4):1240–6.
57. Suhail A, Hui Sun L, et al. GNA13 expression promotes drug resistance and tumor-initiating phenotypes in squamous cell cancers. *Oncogene.* 2018;37:1340–53.
58. Yagi H, et al. A synthetic biology approach reveals a CXCR4–G13–Rho signaling axis driving transendothelial migration of metastatic breast cancer cells. *Sci Signal.* 2011;4:ra60.
59. Struckhof AP, et al. PDZ-RhoGEF is essential for CXCR4-driven breast tumor cell motility through spatial regulation of RhoA. *J Cell Sci.* 2013;126:4514–26.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

