





RESEARCH PAPER

 OPEN ACCESS 

Differentially methylated regions within lung cancer risk loci are enriched in deregulated enhancers

Marina Laplana ^a, Matthias Bieg ^{b,c}, Christian Faltus^{a,d}, Svitlana Melnik^a, Olga Bogatyrova^a, Zuguang Gu^{c,e}, Thomas Muley^{f,g}, Michael Meister^{f,g}, Hendrik Dienemann^{g,h}, Esther Herpelⁱ, Christopher I. Amos^j, Matthias Schlesner ^{e,k}, Roland Eils^{b,c,l}, Christoph Plass^a, and Angela Risch ^{a,d,g,m}

^aDivision of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany; ^bCenter for Digital Health, Berlin Institute of Health and Charité – Universitätsmedizin Berlin, Berlin, Germany; ^cHeidelberg Center for Personalized Oncology (DKFZ-HIPO), Heidelberg, Germany; ^dDepartment of Biosciences, Allergy-Cancer-BioNano Research Centre, University of Salzburg, Salzburg, Austria; ^eDivision of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany; ^fTranslational Research Unit, Thoraxklinik-Heidelberg gGmbH, University of Heidelberg, Heidelberg, Germany; ^gTranslational Lung Research Center Heidelberg (TLRC), Member of the German Center for Lung Research (DZL), Heidelberg, Germany; ^hDepartment of Thoracic Surgery, Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany; ⁱTissue Bank of the National Center for Tumor Diseases (NCT) and Institute of Pathology, Heidelberg University Hospital, Germany; ^jDepartment of Medicine, Baylor College of Medicine, Houston, TX, United States; ^kBioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany; ^lHealth Data Science Unit, University Hospital Heidelberg, Heidelberg, Germany; ^mCancer Cluster Salzburg, Austria

ABSTRACT

Genome-wide association studies (GWAS) have identified SNPs linked with lung cancer risk. Our aim was to discover the genes, non-coding RNAs, and regulatory elements within GWAS-identified risk regions that are deregulated in non-small cell lung carcinoma (NSCLC) to identify novel, clinically targetable genes and mechanisms in carcinogenesis. A targeted bisulphite-sequencing approach was used to comprehensively investigate DNA methylation changes occurring within lung cancer risk regions in 17 NSCLC and adjacent normal tissue pairs. We report differences in differentially methylated regions between adenocarcinoma and squamous cell carcinoma. Among the minimal regions found to be differentially methylated in at least 50% of the patients, 7 candidates were replicated in 2 independent cohorts ($n = 27$ and $n = 87$) and the potential of 6 as methylation-dependent regulatory elements was confirmed by functional assays. This study contributes to understanding the pathways implicated in lung cancer initiation and progression, and provides new potential targets for cancer treatment.

ARTICLE HISTORY

Received 31 July 2020
Revised 19 December 2020
Accepted 7 January 2021

KEYWORDS


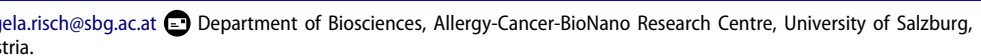
Risk SNPs; bisulphite sequencing; DNA methylation; enhancers; lung cancer; DMR

Introduction


Lung cancer is the major cause of cancer-related deaths in both men and women worldwide [1]. GWAS and meta-analyses have reported a number of SNPs associated with lung cancer risk for adenocarcinoma (AD) and squamous cell carcinoma (SCC), the two major non-small cell lung cancer subtypes [2–8]. Some risk SNPs for lung cancer have been reported in multiple independent studies and in populations with different origins (e.g., 5p15.33, 6p21.33, 8p11, 15q25.1 and 19q13.2) while others seem to be population specific (e.g., 11q22, 13q12, 18p11 and 21q22). Most of these regions contain dozens to hundreds of

SNPs that are highly associated with lung cancer risk, and few of the variants have known mechanisms by which they affect lung cancer development. This shows that we are still far from understanding the role of risk SNPs and their contribution to the disease and the complexity of the risk regions and the genes and regulatory elements that are located therein.

Aside from genomic alterations, epigenetic changes also have an important role in complex diseases such as cancer. DNA methylation is usually interrogated as a surrogate marker for epigenetic alterations. Recently, a few genome-wide studies have been addressed at evaluating DNA

CONTACT Angela Risch  angela.risch@sbg.ac.at 

Marina Laplana: Departament de Ciència Animal, Universitat de Lleida, 25,198 Lleida, Spain.

 Supplemental data for this article can be accessed [here](#)

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

methylation changes in lung cancer. The Cancer Genome Atlas Research Network (TCGA) performed a comprehensive characterization of AD and SCC subtypes, including a genome-wide CpG methylation analysis using Infinium HumanMethylation450 BeadChip array (Illumina450K) [9,10]. Their results supported a high degree of heterogeneity present in lung cancer comprising a large number of mutations, copy number alterations and differences in methylation and gene expression levels within and between cancer subtypes.

The availability of SNPs and Illumina450K data on the same normal lung tissue samples was used by Shi et al. to investigate the effect of genetic variation on DNA methylation by analysing methylation quantitative trait loci (meQTL) [11]. The authors identified cis-meQTLs for 34,304 CpG Illumina probes and 585 trans-meQTLs, including inter- and intrachromosomal associations. Heyn et al. combined Illumina450K and Genome-Wide Human SNP Array 6.0 (Affymetrix) pan-cancer data from TCGA to identify genome-wide associations between methylation and risk genetic variants for the most common solid tumour entities [12]. Investigating cis-meQTLs for 109 GWAS associated SNPs in 13 different cancer types, they found cis-meQTLs in the promoter region of *TP63* and *TERT* for rs10937405 and rs2736100, respectively, in lung adenocarcinoma samples. These studies demonstrate the importance of studying the epigenome to help us to understand lung carcinogenesis.

Most of the recent studies conducted to evaluate the effect of genetic variation on the methylome make use of Illumina450K arrays that are commonly used as a tool for genome-wide screening of DNA methylation on a large number of samples; however, the resolution of the arrays is limited and restricted to the pre-designed regions. Using quantitative methylation analysis targeted to specific genomic regions, Scherf *et al.* demonstrated that previously reported lung cancer risk SNPs on 15q25 were associated with DNA methylation levels in the promoter of *CHRNA4*. Furthermore, *CHRNA4* promoter hypomethylation in the tumours was linked with gene over-expression [13], and *in vitro* knock-down of the

gene reduced cell proliferation and the capacity for colony formation. In addition, Jones et al. reported a particular CpG site (cg17028067) for which methylation is associated with plasma levels of Lipoprotein(a) [14]. Further analysis revealed the causative effect as a low frequency G/A SNP, rs76735376 within the cg17028067, which affects the methylation status of the CpG, thus altering enhancer activity. These types of studies demonstrate the clear interconnection between the genome and the epigenome and emphasize the necessity of studying lung cancer risk regions in detail in order to decipher the link between risk SNPs and the causal changes affecting disease susceptibility or progression.

Our study aims to contribute to the understanding of lung cancer by exploring DNA methylation patterns at the disease-implicated loci defined by GWAS. We used a targeted bisulphite sequencing approach in order to perform a comprehensive methylome characterization of the lung cancer risk regions. This is the first time that DNA methylation changes are analysed in lung cancer risk regions in a comprehensive manner, with single CpG resolution and a high power to detect deregulated regions in lung cancer.

Material and methods

Sample collection

All subjects included in the discovery cohort and replication cohort 1 were male smokers with stage I lung cancer. Replication cohort 2 was composed of a higher number of subjects including men and women with stages ranging from IA to IIIA. Tumour histology was classified according to the 3rd edition of the World Health Organization classification system [15]. All cohorts comprised subjects with lung adenocarcinoma (AD) or lung squamous cell carcinoma (SCC). DNA samples from fresh frozen lung tumours and adjacent normal tissues from lung cancer patients were retrieved from the tissue bank of the Thoraxklinik in Heidelberg, Germany. Clinical and demographic characteristics of patients included in the study are described in Supplementary Table 1. All

participants provided written informed consent. The study was approved by the Ethics committee of the Medical Faculty of the University of Heidelberg (Nr. 270/2001).

Risk region definition and sureselect methyl seq custom capture library design

Using the PubMed database, the literature was screened for studies describing loci associated with lung cancer risk from familial cases, case-control studies of candidate genes, GWAS and meta-analyses. A full description of the custom capture library design is shown in the Supplementary

Material. Shortly, 84 unique lung cancer risk SNPs were extracted from selected studies. A total of 54 candidate regions spanning 12.078Mb were defined as regions in linkage disequilibrium (LD) with the risk SNPs (Figures 1 & Figures 2a, Supplementary Table 3). Coordinates of selected regions were uploaded to SureDesign software (Agilent) with the following bait tailing parameters: +strand, density 2x, balanced boosting and most stringent masking of repetitive elements, centred, length 120bp and an allowed overlap of 20bp. After removal of 34,048 non-CpG containing baits, the custom library was composed of 57,676 probes and spanned 4.038Mb.

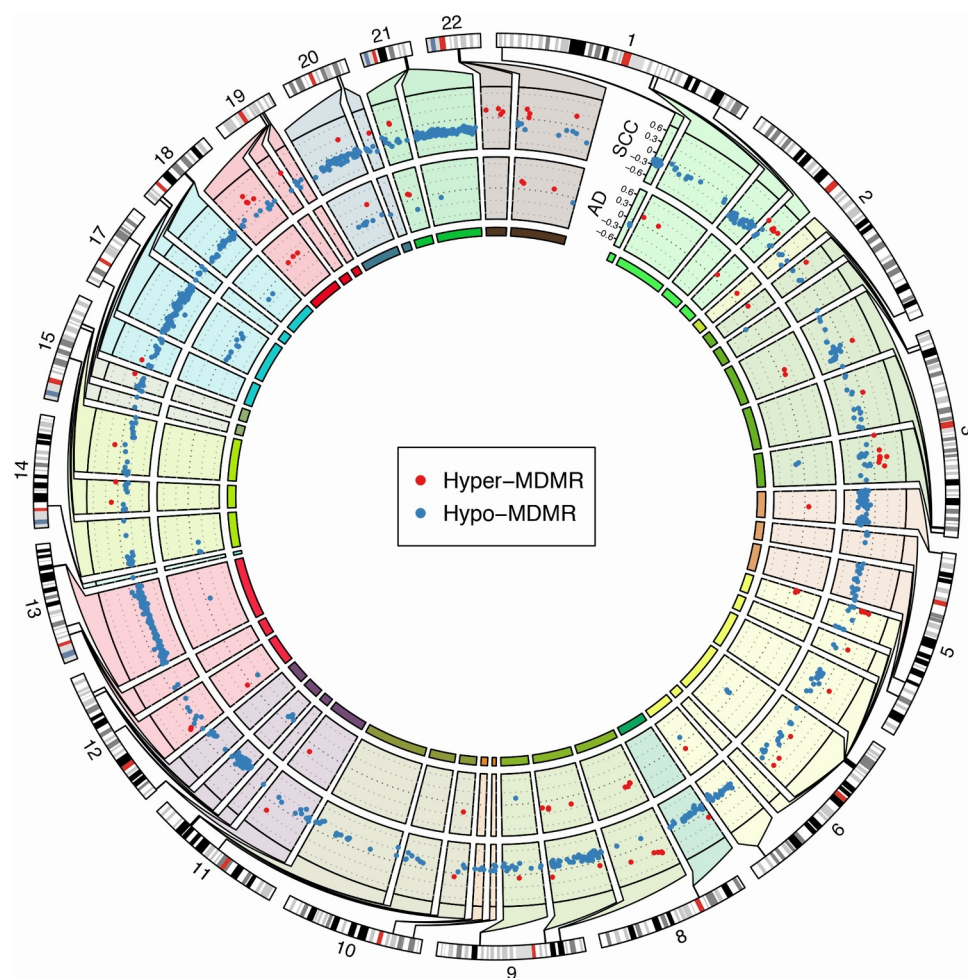


Figure 1. Distribution of minimal differentially methylated regions (MDMRs) in lung cancer risk regions. Nineteen human autosomes are depicted with MDMRs distribution in 54 lung cancer risk regions (coordinates are listed in Supplementary Table 3). The innermost track depicts the lung cancer risk regions per chromosome in different colours. The second and third tracks depict changes in MDMRs DNA methylation in squamous cell carcinoma (SCC) and adenocarcinoma (AD) patients, respectively. Blue denotes hypomethylated MDMRs and red denotes hypermethylated MDMRs.

DNA-targeted enrichment and methylation analysis

SureSelectXT MethylSeq (Agilent) was used to perform targeted enrichment of selected loci according to the manufacturer's instructions (version B January 2013) in the discovery cohort, including 11 AD and 13 SCC tumour and adjacent normal tissue samples. A complete description of the protocol is shown in the Supplementary Material. Briefly, the protocol was modified as follows: 6 µg of DNA were used as starting material and sheared in a 50 µl volume using a CovarisS1 instrument with the following settings: duty cycle, 10%; intensity, 5; cycles per burst, 200; time, 4 cycles of 60 seconds and 1 cycle of 30 seconds; set mode, frequency sweeping and temperature, 4 to 7°C. Amplification and indexing PCRs of the bisulphite-treated libraries (EZ DNA Methylation-Gold™ Kit, Zymo Research) were carried out in a LightCycler 480 (Roche) including 1x SybrGreen (10,000x, Invitrogen) adjusting PCR cycle number when necessary. Libraries were multiplexed in groups of 12 and sequenced in IlluminaHiSeq2000 with 100bp paired-end sequencing at DKFZ Genomics and Proteomics Core Facility, Heidelberg.

Bisulphite sequencing data analysis

Reads were mapped against two reverse complementary *in-silico* bisulphite converted strands of the human reference sequence (hg19) using methylCtools [16]. First adaptor sequences were trimmed using SeqPrep followed by the modification of all sequencing reads, such that cytosines were masked by thymines and guanines by adenines. Afterwards, the converted reads were mapped against the reference sequence using BWA-ALN (version 0.6.2) [17] and the non-default parameters -q20-s. PCR duplicates were removed using PICARD MarkDuplicates. After alignment and duplicate removal, sequencing reads were translated back into their original state and for each CpG position the number of methylated (and unmethylated) cytosines was calculated by counting the number of reads showing evidence for methylated (and unmethylated) state at the guanine position. Reads with a mapping

quality less than 1 and bases with a PHRED-scaled quality less than 20 were discarded. To avoid accuracy issues and SNP-induced variability in CpG methylation, we only considered on-target CpG positions with a sequencing depth greater or equal to 20, not overlapping a known SNP from dbSNP135 [18]. Of the 71,787 CpGs in the targeted regions, 15,854 CpGs overlapping SNPs were excluded from further analysis.

SNP calling within the target regions was performed using the bisulphite conversion aware SNP-caller Bis-SNP [19] using the non-standard parameters `-out_modes EMIT_VARIANTS_ONLY -toCoverage1000`. To enhance the reliability of the called polymorphisms, we only considered SNPs known from dbSNP135 with a PHRED-scaled confidence score greater or equal to 50.

Samples were then subjected to quality control and excluded from further analysis if: median on-target CpG coverage <14x or estimated bisulphite conversion efficiency <96.5%. Mean bisulphite conversion efficiency per sample was estimated as 1 - the mean chromosomal methylation in non-CpG sites (mCH). Only SNP positions which were consistently covered by more than 9 reads among all individuals were considered. Seven samples were excluded after quality control giving rise to the drop of 7 tumour/normal tissue pairs and leaving 17 pairs for further analysis.

Differentially methylated positions (DMPs) were calculated with the R-package methylKit [20] using the following criteria: minimum sequencing depth at CpG positions of 20x, methylation difference between tumour and normal tissue samples ≥10% and Fisher's exact test p-value ≤0.05. Differentially methylated regions (DMRs) were called with the R-package eDMR [21] using the following criteria: DMP methylation difference ≥10%, q-value ≤0.05, minimum of 3 DMPs in the same direction, distance between neighbouring CpGs ≤300bp, DMR methylation difference ≥10% and q-value <0.05. For both DMP and DMR analysis, we performed multiple testing correction using a Benjamini-Hochberg approach that controls for false discovery rate (FDR). DMRs were not identical in all patients but rather they tended to overlap with different

ends. Thus, in order to focus on the most important deregulated regions, we defined Minimal Differentially Methylated Regions (MDMRs), i.e., regions where a DMR occurred in more than 50% of the patients.

Further description of methods regarding enrichment analysis and TCGA data analysis as well as description of MassArray Epityper and dual-luciferase assays are provided in Supplementary Materials.

Results

Targeted enrichment of lung cancer candidate DNA regions

Our aim to study comprehensively the epigenetic changes occurring in lung cancer risk regions required a different approach to those commonly being applied for DNA methylation analysis, such as Illumina450K or MeDiPseq, as these would not properly cover our regions of interest. From the literature, we defined 54 lung cancer candidate deregulated regions spanning a total of 12,078 Mb (Figure 1, Supplementary Table 3). In order to establish DNA methylation profiles in these regions, we designed a custom capture library to enrich for genomic DNAs covering the selected sequences. Library design is summarized in Figure 2a. Briefly, the initial 91,950 probes of the SureSelect DNA library (Agilent) covering all 54 candidate regions were narrowed down by removing non-CpG containing baits and repetitive sequences. The final custom capture library consisted of 57,676 probes spanning 4.038Mb and covering a total of 71,787 CpG dinucleotides of which 22.3% are located in promoter regions (defined as 2Kb upstream and 1 Kb downstream of the TSS), 4.3% in 5'UTR, 23% in exons, 50.2% in introns, 2.4% in 3'UTR, and 28.8% in intergenic regions (Figure 2b). The number of captured CpGs per targeted region is shown in Supplementary Table 3.

For the DNA methylation analysis, we isolated DNA from lung tumour tissue and adjacent normal tissue from lung cancer patients. The clinical and demographic characteristics of patients included in the discovery cohort are listed in Supplementary Table 1. DNA samples were

enriched for our candidate regions, sodium bisulphite converted and then analysed by next-generation sequencing. After applying quality standards described in the methods section, we analysed methylation patterns of 17 patients including 7 AD and 10 SCC matched tumour-normal tissue pairs. Bisulphite sequencing data for samples that passed the quality control had a median of 10.88 million analysable reads and a median read depth at on-target CpGs of 127x. The median on-target reads-ratio of 91.6% indicated a high level of enrichment for the targeted regions that also showed high conversion efficiency (median bisulphite conversion efficiency of 98.9%).

DNA methylation analysis identifies differences in the DMRs from AD and SCC

DNA methylation analysis of the risk regions revealed lower methylation levels in tumour tissues in comparison to the adjacent normal tissue (median of the median methylation per sample 71.67% and 78.6% in tumour and normal tissue, respectively). Broad DNA methylation patterns in the targeted regions are shown in Supplementary Materials, Supplementary Figure 1. Interestingly, the median methylation for lung AD samples was 75.6%, closer to the methylation levels found in normal tissue (78.6%) than to the ones found in lung SCCs (63.78%) (Figure 2c). Unsupervised hierarchical clustering of the 5,000 most variable CpGs showed 3 clusters of samples corresponding to the normal tissues, AD and SCC samples (Figure 2d) revealing that methylation patterns in the analysed regions were globally disturbed in the tumours in comparison to normal tissue, but also differed significantly between tumour subtypes.

In order to validate our results, we compared our data with publicly available Illumina 450 K methylation data of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) from The Cancer Genome Atlas (TCGA) and we implemented a high-throughput technology for replication of our results (Supplementary Material, Supplementary Figure 3).

To exclude inter-individual differences from the analyses, we analysed intrapair methylation changes for the tumour and normal tissue

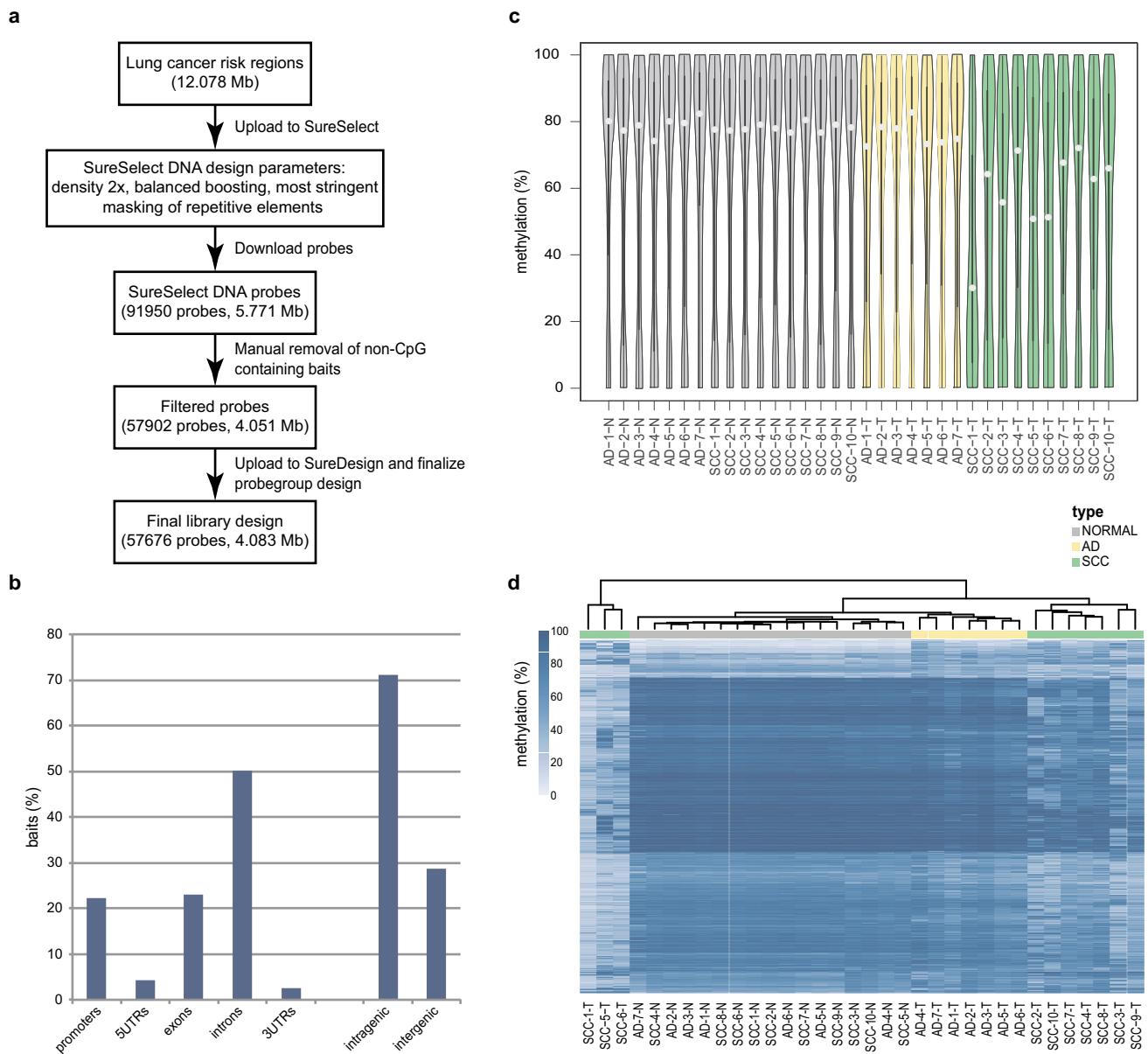


Figure 2. Scheme for custom capture library and results from overall methylation analysis. (A) Workflow for custom capture library design. (B) Distribution of capture baits in genomic regions. (C) Violin plot depicting distribution of methylation levels and median values per sample. (D) Heatmap depicting the unsupervised hierarchical clustering of 5,000 most variable CpG sites between 17 tumour and adjacent normal tissue samples. Sample type is colour coded as follows: normal samples in grey, lung adenocarcinoma (AD) samples in yellow and lung squamous cell carcinoma (SCC) samples in green.

matched samples. We detected a median of 16,706 differentially methylated positions (DMPs) in SCC tumour-normal pairs, consisting of a median of 12,517 hypomethylated DMPs and 3,419 hypermethylated DMPs, with a median methylation difference between tumour and normal tissue of -28.69% and 21.16% , respectively. For AD tissue pairs, we detected a median of 7,259 DMPs comprising 2,757 hypomethylated and 2,741 hypermethylated DMPs, with a median methylation

difference of -18.93% and 18.27% , respectively. The number of DMPs per patient is shown in Supplementary Table 4.

We extended our methylation analysis to identify differentially methylated regions (DMRs) between tumour and normal tissue samples. We found a median of 1,086 DMRs among SCC patients (median of 937.5 hypo- and 124 hypermethylated) and 350 DMRs among AD patients (median of 111 hypo- and 59 hypermethylated). In

both tumour subtypes hypomethylated DMRs were predominant, with a median frequency of 90.6% in SCC and 52.1% in AD; however, AD patients had a bigger proportion of

hypermethylated DMRs compared to SCC patients (Figure 3a). The number of hyper- and hypomethylated DMRs per subject is shown in Supplementary Table 4. For the complete list of

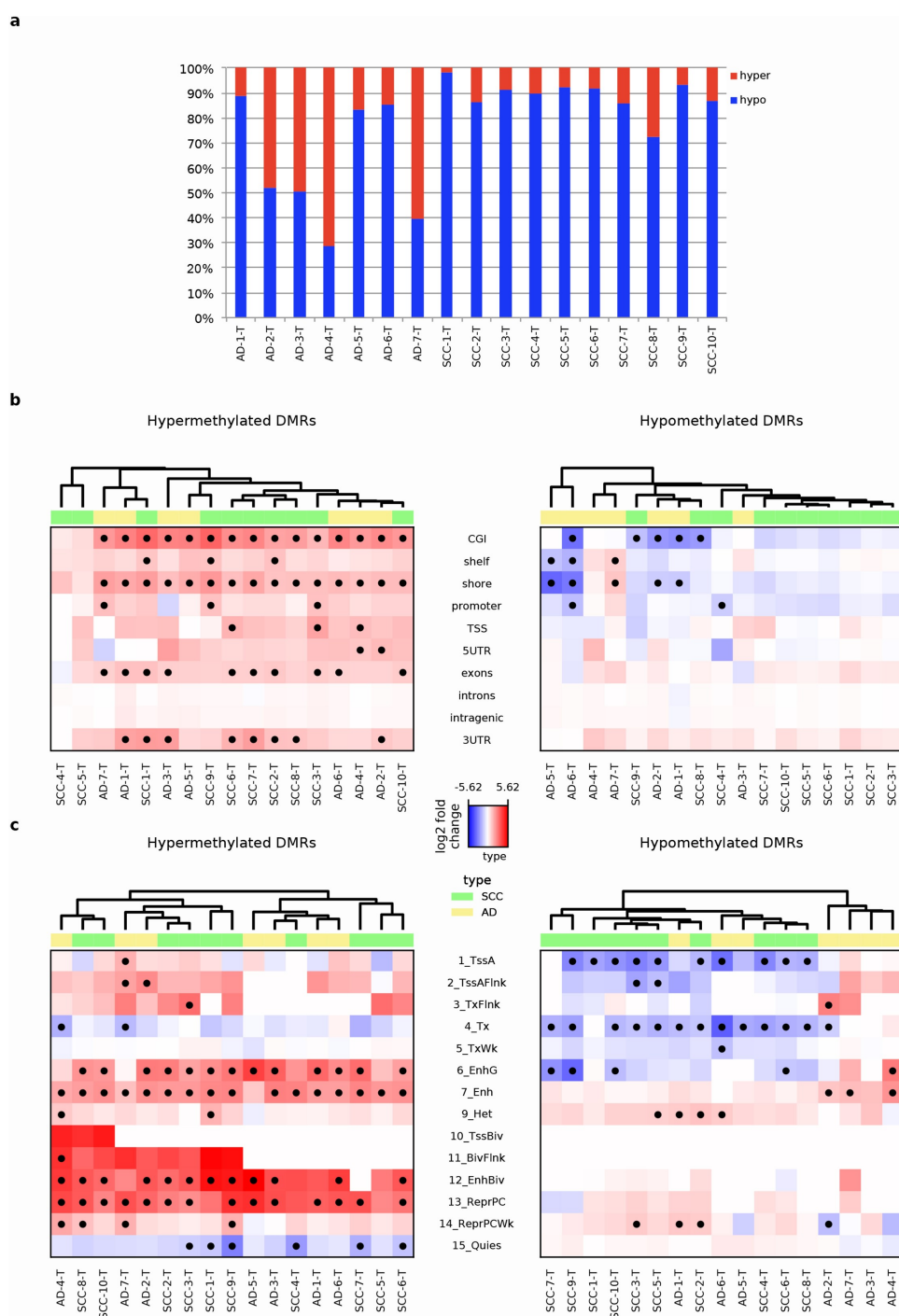


Figure 3. Differentially methylated regions (DMRs) and DMR enrichment analyses. (A) Proportion of hyper- (in red) and hypomethylated DMRs (in blue) from pairwise tumour-normal comparison per subject. (B) DMRs enrichment analysis by genomic location using RefSeq annotation. Left hypermethylated, right hypomethylated DMRs. (C) DMRs enrichment analysis in chromatin states from lung tissue using 15-states ChromHMM definition from roadmap epigenomics consortium data. Dots depict significant enrichments (p -value < 0.01 and fold change > 2 or < 0.5). Left hypermethylated, right hypomethylated DMRs. Sample type is colour coded as follows: adenocarcinoma (AD) samples in yellow and squamous cell carcinoma (SCC) samples in green.

DMRs per patient, including p-values, see Supplementary Table 5.

The median CpG content of DMRs was 6 in both tumour subtypes, and the median length was slightly longer in AD compared to SCC patients (431bp vs. 401.75bp, respectively). In accordance with the DMPs data, methylation difference of DMRs was smaller in AD than in SCC, with a mean DMR methylation difference per patient of 18.3% and 24.8% in AD and SCC, respectively.

To understand the genomic localization of the DMRs, we performed an enrichment analysis of DMRs in genomic locations using RefSeq gene annotation. Additionally, in order to examine their possible influence on gene expression or in the activity of regulatory elements, we analysed their enrichment in lung tissue chromatin states (15-states of ChromHMM) defined by their deposition of different histone marks (Roadmap Epigenomics Consortium) [22] (Figures 3b and c). 8-ZNF/Rpts chromatin state (ZNF genes & repeats), including repetitive elements, was not analysed as those regions had been specifically excluded from our library, thus keeping 14-states for the analysis. Enrichment was considered when at least 50% of the patients (4 of 7 AD subjects and 5 of 10 SCC subjects) showed a p-value ≤ 0.01 and fold change >2 or <0.5 for a certain region of interest. Interestingly, the genomic localization of hyper- and hypomethylated DMRs was different. SCC and AD hypermethylated DMRs were enriched in CpG islands (CGI), CGI shores and exons. In addition, 5 SCCs showed an enrichment of DMRs in 3'UTRs. On the other hand, for hypomethylated DMRs, no major enrichment was observed, but they are depleted in CGI shores for AD (4/7 subjects) (Figure 3b). The analysis of the distribution of hypermethylated DMRs in the 14 lung chromatin states showed a depletion of DMRs in the Quiescent State (Quies) for 6 out of 10 SCC patients and enrichment for both AD and SCC subjects in the repressive states Bivalent Enhancer and Repressed by Polycomb (EnhBiv and ReprPC) and also in active regulatory enhancer states (Enh and EnhG) (Figure 3c). Similarly to the results observed in the enrichment in genomic locations, hypomethylated DMRs did not

show major enrichments, but depletions were found in SCC for DMRs in the active states Strong transcription and Flanking active TSS (Tx and TssA) for 9 out of 10 individuals.

Validation and replication of top deregulated regions

In order to focus on the most interesting deregulated regions, we extracted the minimal differentially methylated regions (MDMRs) defined as regions where a DMR occurred in the same direction in more than 50% of the patients (Figure 4a). A full list of the MDMRs and descriptive statistics is provided in Supplementary Table 6. We found 1,102 MDMRs for SCC (94.5% hypomethylated and 5.5% hypermethylated) and 80 MDMRs for AD (57.5% hypomethylated and 42.5% hypermethylated). The number of MDMRs per targeted region is shown in Figure 1 and Supplementary Table 3. To gain an overview of the methylation differences between both tumour subtypes we evaluated the number of overlapping MDMRs (Figure 4a). We found 59 MDMRs in both AD and SCC datasets, of which 17 were hypermethylated. Two MDMRs were hypermethylated in AD and hypomethylated in SCC indicating their putative different roles in tumour subtypes, and 40 hypomethylated MDMRs were defined with the exact same coordinates in both tumour subtypes giving rise to a total number of identified unique MDMRs of 1,142. The median CpG content of MDMRs was 13.92 in SCC and 12.56 in AD, and the median length was slightly longer in AD compared to SCC, similar to the previous finding for DMRs (408bp vs. 402bp). In addition, in both tumour subtypes hypermethylated MDMRs were longer than hypomethylated ones (AD: 489bp vs. 359.5bp and SCC: 600bp vs. 398bp). The mean methylation difference per MDMR was -24.46% and 19.3% for SCC hypo- and hypermethylated MDMRs, and a little smaller for AD MDMRs, with -16.05% and 16.4% , respectively.

We compared our data with LUAD and LUSC datasets from TCGA and found that 85.8% of the identified MDMRs would be missed in 450 K data while the biggest proportion of the captured

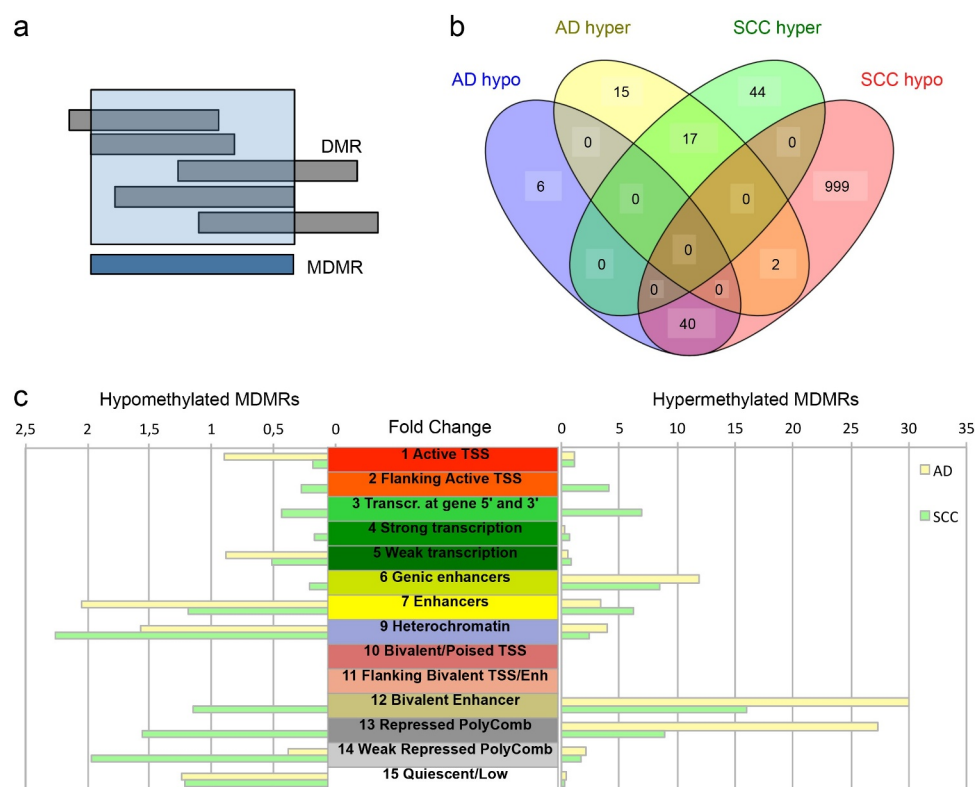


Figure 4. Identification of minimal differentially methylated regions (MDMRs) and MDMR enrichment analyses. (A) Scheme of MDMR annotation. MDMRs are defined as the regions where, within at least one tumour subtype, DMRs overlap in more than 50% of the subjects. (B) Venn diagram showing the number of overlapping MDMRs between hyper- and hypomethylated MDMRs in squamous cell carcinoma (SCC) and adenocarcinoma (AD). (C) Enrichment of hypomethylated (left) and hypermethylated (right) MDMRs in chromatin states from lung tissue using 15-states ChromHMM definitions.

MDMRs would be covered only by 1 probe. After data normalization and quality control, only probes covering 140 MDMRs were kept. The absolute mean MDMR methylation difference with our approach was 21.4% while the TCGA mean methylation difference was only 4.53%. These data illustrate the poor coverage of 450 K in the most interesting regions revealed by our analysis within lung cancer risk regions, and indicate that most studies will have probably missed those methylation changes.

Furthermore, we analysed the distribution of MDMRs in 14 different ChromHMM states from lung tissue (Figure 4c). Interestingly, we found an enrichment of hypermethylated MDMRs in repressive chromatin states from lung tissue, such as Bivalent Enhancer and Repressed by Polycomb, and in the active enhancer states, with a fold change higher

than 3 and p -value <0.05 . For hypomethylated MDMRs we did not find major enrichments (Figure 4c). We also evaluated the location of transcription factor motifs from HOCOMOCO database among the MDMRs. A full list of all transcription factor motifs for each tumour type is reported in Supplementary Table 8. Briefly, we found an enrichment of Fos family members and Jun in AD hypomethylated MDMRs that have been implicated as regulators of cell proliferation, differentiation and transformation (Supplementary Materials, Supplementary Figure 5B). Importantly, Sp subfamily members are enriched in all four MDMR sets (AD and SCC, hyper- and hypomethylated) (Supplementary Materials, Supplementary Figure 5). This subfamily has been described as potential activators or repressors of expression in different promoters. Other relevant transcription factor motifs

involved in apoptosis, DNA replication, and carcinogenesis have also been found enriched in the MDMRs (Supplementary Materials, Supplementary Figure 5).

Validation and replication of selected MDMRs

For candidate selection, we ranked the MDMRs according to their absolute mean methylation difference among AD or SCC patients. MDMR enrichment in ChromHMM states indicated a putative role of the MDMRs in the deregulation of regulatory elements. Thus, we selected a total of 7 MDMRs from the top 15 in each tumour subtype to test their potential role as regulatory elements (Supplementary Table 2). Three MDMRs were selected from the SCC list, including 1 hypermethylated MDMR (MDMR_1) and 2 hypomethylated ones (MDMR_5 and MDMR_7). For AD 4 MDMRs were selected comprising 3 hypermethylated MDMRs (MDMR_2, MDMR_3 and MDMR_6) and MDMR_4 which is hypomethylated in the tumours. MDMR_1 was located on chromosome 3, in the intron of *Xyloside Xylosyltransferase 1 Antisense RNA 2 (XXYLT1-AS2)*, a non-coding RNA that overlaps with the 3rd intron of *XXYLT1*. MDMR_1 showed a methylation difference of 42.50% between tumour and normal tissue and contains 11 CpGs. MDMR_2 and MDMR_3 are in the *CDKN2B-CDKN2A* gene cluster on chromosome 9p21 located in the 1st intron of the *Cyclin Dependent Kinase Inhibitor 2A (CDKN2A)* and the *CDKN2B Antisense RNA 1 (CDKN2B-AS1)* with a difference in methylation of 23.6% and 21.5%, respectively. MDMR_4 showed a difference in methylation of -25.2% and it is located on chromosome 13 in the 18th intron of *Mitochondrial Intermediate Peptidase (MIPEP)* gene. MDMR_5 located on chromosome 18, in intron 2 of *Piezo Type Mechanosensitive Ion Channel Component 2 (PIEZO2)* exhibited a hypomethylation of -48.42%. MDMR_6 is located in intron 3 of the *Nuclear Factor Of Activated T-Cells 2 (NFATC2)* on chromosome 20 and showed a difference in methylation of 27.3%. Finally, MDMR_7 is located in the 3' UTR of the *HORMA Domain Containing 2 (HORMAD2)* gene on chromosome 22 and showed a difference in methylation of -38.4%.

We validated all MDMR results from the discovery sample set by MassArray EpiTyper assay (Agena Bioscience) using primers listed in Supplementary Table 7. A candidate MDMR was considered as technically validated when in the same tumour-subtype, the mean methylation difference was >20%, in the same direction as reported in the discovery cohort and the p-value <0.05. Thereby, all MDMRs from the discovery sample set were validated (Figure 5a). Next, MDMRs were replicated in 2 larger independent lung cancer cohorts. Clinical and demographic characteristics of the patients included in both replication cohorts are presented in Supplementary Table 1. First, an independent cohort comprised by 14 AD and 13 SCC patients with the same characteristics as the discovery cohort (Replication cohort 1) was used to replicate our findings for the 7 MDMRs by MassArray (Figure 5b). Then, a larger cohort of 44 AD and 43 SCC patients (Replication cohort 2), comprising both male and female subjects as well as different lung cancer stages ranging from IA to IIIA was used to replicate our results in a more comprehensive lung cancer cohort (Supplementary Table 1, Figure 5c). All candidate MDMR methylation values were replicated in both cohorts indicating the general deregulation of the candidate regions in lung cancer.

Evidence from in vitro assays for regulatory activity of selected MDMRs

To confirm the role of the candidate MDMRs as regulatory elements that are aberrantly methylated in lung cancer, we used a dual-luciferase reporter assay. Both the plus and minus strands of the candidate MDMRs were cloned into the reporter firefly constructs pGL4.10 and pGL4.23, to test their role as promoter and enhancer regulatory elements. Six out of the 7 tested regions showed regulatory activity (Figure 6a). MDMR_3 within *CDKN2B-AS1* did not show any potential as regulatory element under the experimental conditions tested. MDMR_1 in *XXYLT1-AS2* seemed to have both promoter and enhancer activity in both DNA strands. Three MDMRs at *CDKN2A*, *MIPEP*, and *NFATC2* (MDMR_2, MDMR_4, and MDMR_6,

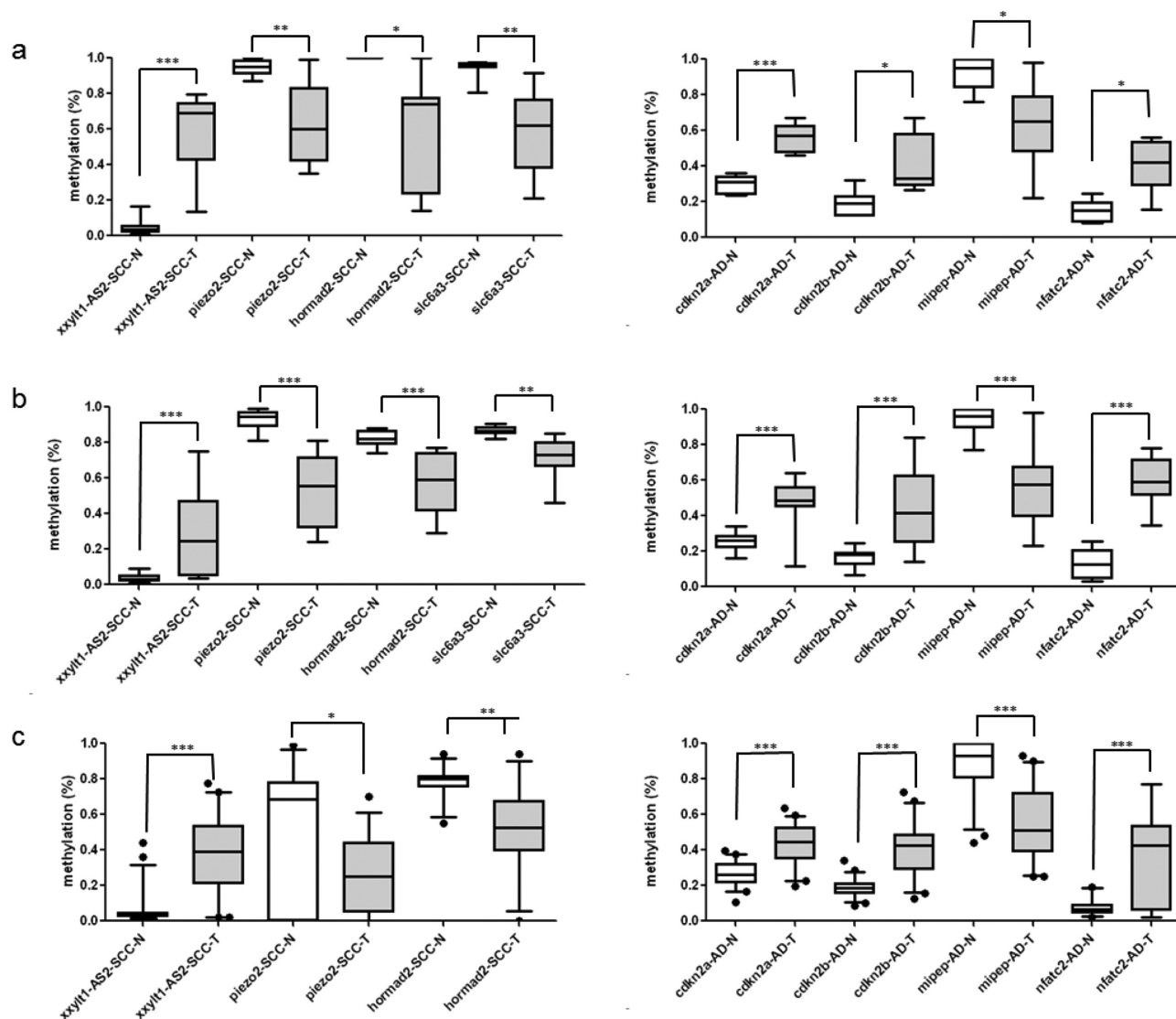


Figure 5. (A) Candidate MDMRs DNA methylation in the discovery cohort validated by MassArray epityper assay. (B) Replication of candidate MDMRs in the replication cohort 1 by MassArray. (C) Replication of candidate MDMRs in the replication cohort 2 by MassArray. * p-value < 0.05. ** p-value < 0.005. *** p-value < 0.0005. Whiskers indicate 5–95% percentiles.

respectively) showed both promoter and enhancer activity in the plus strand, and MDMRs located within *PIEZO2* and *HORMAD2* (MDMR_5 and MDMR_7) showed both promoter and enhancer activity in the minus strand. Collectively, these *in vitro* data confirm that among 7 identified MDMRs, only MDMR_3 displayed no potential as a regulatory region.

Next, we investigated the possible role of DNA methylation in the regulatory activity of candidate promoters and enhancers within the identified

MDMRs using the pCpG-free-promoter Lucia vector as a reporter. Based on the regulatory activity determined in the previous experiment, candidate regions were cloned in forward or reversed directions. In line with previous results, we found that all candidate regions except for the MDMR_1 at *CDKN2B-AS1* showed overexpression of the reporter when they were un-methylated (Figure 6b). This indicates that DNA methylation at candidate MDMRs may contribute to a change in the regulatory activity of these regions which could be affecting the expression of

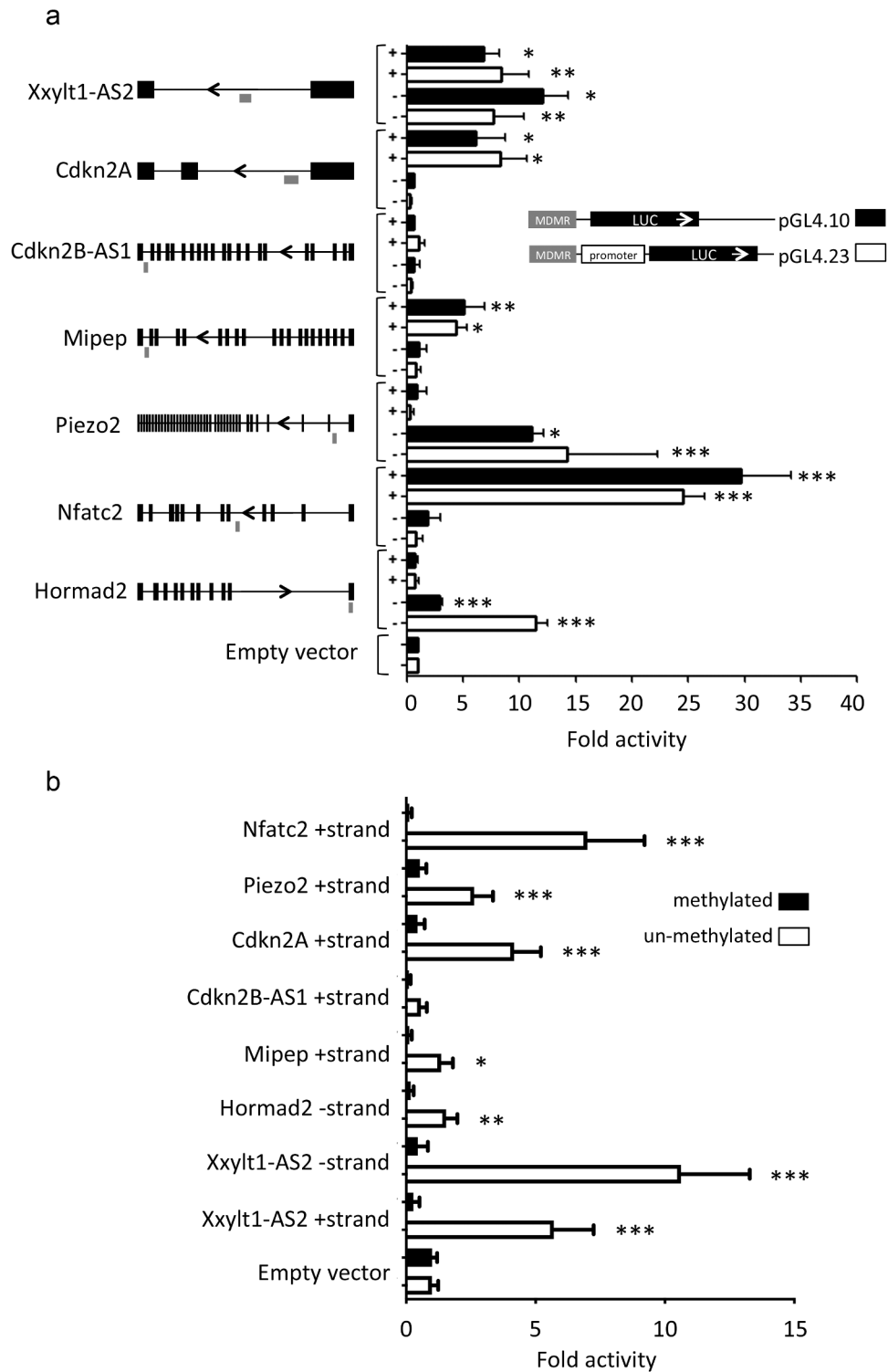


Figure 6. *In vitro* analysis of promoter and enhancer function, and effect of methylation on regulatory activity. (A) Bar plot of the dual-luciferase assay in HEK293T cell line to test promoter and enhancer regulatory effect of candidate MDMRs on both the plus (+) and minus (-) strand. (B) Bar plot of the dual-luciferase assay to test the effect of DNA methylation on the regulatory activity of candidate MDMRs in HEK293T cells. * p-value < 0.05. ** p-value < 0.005. *** p-value < 0.0005. Error bars represent standard deviation.

genes and non-coding RNAs, possibly, but not necessarily, in the vicinity, with a potential role in the initiation or progression of lung cancer.

Discussion

Throughout the last decades, GWAS have led to the discovery of lung cancer risk SNPs that are associated with disease susceptibility or its progression. However, the causative variants for these associations and the genes or regulatory elements playing a central role in lung carcinogenesis are still far from being completely understood. Today, we are able to study the methylome as a marker of epigenetic changes that may contribute to lung cancer. With the aim of expanding the understanding of lung cancer and of identifying new causative regulatory regions for the known associations, we investigated DNA methylation changes in previously described lung cancer risk regions. We conducted methylation analysis by targeted DNA bisulphite sequencing and demonstrated the power of this approach to detect aberrantly methylated loci previously overlooked by other approaches. The comparison of methylation patterns of lung tumour and matched normal tissues as basis for the identification of differentially methylated positions allowed us to focus on lung cancer-specific methylation alterations, i.e., discarding interindividual methylome heterogeneity.

We detected a large number of DNA methylation changes occurring in SCC and AD lung cancer subtypes within the defined lung cancer risk regions. Both tumour entities showed global hypomethylation when compared with the normal samples, in line with previously described data for cancer tissues [23]. However, we found that for the defined lung cancer risk regions AD accumulated fewer methylation aberrations and thus resembles normal tissue more closely than SCC. Although our targeted approach covered only a small portion of the genome, our DNA methylation data allowed for discrimination of 3 clusters of samples (Normal, AD tumour, and SCC tumour tissues). The two lung cancer subtypes investigated showed different methylation profiles in the lung cancer risk regions, which could be due to the different origin and evolution of the tumour entities, but may also be related to the definition of the risk regions.

The present study provides a comprehensive list of differentially methylated regions detected in the different AD and SCC patients that can be used as a resource to identify links between risk SNPs and the underlying molecular cause of disease. In order to focus on the most interesting deregulated regions and to provide functional characterization for some candidates, we combined the DMRs of the different subjects into MDMRs. We found a different distribution of hyper- and hypomethylated MDMRs in genomic regions and ChromHMM states indicating differences in their behaviour and a role in lung cancer. Interestingly, we found an enrichment of MDMRs in regulatory elements such as enhancers. Our results are in line with those from Shi *et al.*, who found enrichment of meQTLs in regulatory regions analysed in TCGA datasets [11]. Thus, we tested the role of 7 MDMRs as putative regulatory elements. Of these, 6 MDMRs showed regulatory activity, with their effects occurring mostly when DNA was not methylated. Taken together, these results confirm the potential of the MDMRs as methylation-dependent regulatory regions that can contribute to lung cancer initiation and/or progression.

One of the selected candidates, *XXYLT1-AS2*, is an antisense-RNA of the *XXYLT1* gene. *XXYLT1* has been implicated in the cancer-related Notch pathway as a regulatory gene product of the Notch receptor activation [24]. A role of anti-sense RNAs in the regulation of gene expression has previously been described [25]. MDMR_1 might affect both the expression of *XXYLT1-AS2* and *XXYLT1*. MDMR_4 is located in the intron of *MIPEP* gene that has been reported as a modulator of Notch activity. Both MDMRs could alter the Notch pathway, with effects in cancer cell signalling.

Different regions from the gene cluster *CDKN2A-CDKN2B* have been implicated in cancer and in particular in lung cancer. In our analysis, several DMRs were defined in this gene cluster, summarized in several MDMRs, with two of them highly ranked among our candidates (MDMR_2 and MDMR_3). MDMR_3 located in the *CDKN2B-AS1* did not show regulatory potential under the tested conditions which does not, however, exclude a possible regulatory activity in lung cells.

The *PIEZO2* gene could be under the regulation of MDMR_5. This gene is a pseudogene of *PIEZO1* (*FAM38A*) which has been associated with a reduction of cell adhesion and an increase in cell migration in small cell lung cancer [26]. In addition, *PIEZO1* has been postulated as a good candidate for diagnosing gastric cancer [27]. We hypothesize that *PIEZO2* might play a central role in lung cancer. MDMR_6 might be involved in regulation of the *NFATC2* gene. *NFATC2* is a DNA-binding protein that is present in the cytosol and translocates to the nucleus upon T-cell receptor activation. *NFATC2* has a key role in immune response by inducing gene transcription. However, *NFAT* isoforms present in the tumour microenvironment also contribute to the regulation of the interactions between compartments and have a function in cell growth, survival, invasion, differentiation, proliferation and angiogenesis [28–31]. *NFATC2* has been shown to be pro-invasive and pro-migratory in breast carcinoma [32,33].

MDMR_7 at the *HORMAD2* locus also showed a potential regulatory effect. *HORMAD2* is predominantly expressed in human testis; however, it is ectopically expressed in nearly 10% of lung cancer samples from Chinese Han individuals [34].

We acknowledge the limited power of our study due to the small sample size in the discovery cohort that could potentially limit the number of identified deregulated regions. However, samples were selected as homogenous as possible to limit the effect of confounding factors. Furthermore, some candidates were validated in 2 independent cohorts, including a heterogeneous larger one, indicating the reliability of our results. There are a variety of mechanisms by which SNPs can affect methylation (reviewed in Wang et al. [35]). Unfortunately, the limited number of samples and the design of the custom library (which resulted in the capture of a unique strand thus preventing discrimination between methylation status of the CpG or the C/T alleles of an overlapping SNP) did not allow us to analyse the connection between DNA methylation alteration and genetic variants. However, we were able to narrow down lung cancer risk loci to the most interesting deregulated regions that can be

further investigated to decipher their role in lung carcinogenesis.

The MDMRs we studied *in vitro* pointed us towards candidate genes that have a putative role in the molecular mechanisms of lung tumour initiation or progression as denoted by the pathways affected, their interactions with cancer-related genes or their previously described role in other cancer types. Thus, these candidates and the MDMR associated loci need to be investigated in depth in further studies, which consider their functional effects, including a potential impact on chromatin structure in distant regulatory regions.

Conclusions

This is the first study that comprehensively explored DNA methylation patterns at single CpG resolution at the GWAS-identified lung cancer risk loci. Aberrantly methylated regions in tumour tissue were identified, and an enrichment of differentially methylated regions in regulatory elements was shown. Furthermore, *in vitro* functional assays demonstrated a methylation-dependent enhancer activity for 6 out of 7 tested MDMRs. Taken together, these results confirm the potential of the MDMR regions as methylation-dependent regulatory elements that can contribute to lung cancer initiation and/or progression. The data presented here will be very valuable to understand the pathways implicated in lung cancer and provide new targets for further functional characterization as potential druggable targets for cancer treatment.

Authors' contributions

AR, ML, SM, CA and CP conceived and designed the study; ML and SM designed the custom capture library; MB analyzed the bisulfite sequencing data, MS and RE supervised the sequencing data analyses; ZG performed the circos plot; ML performed the biological experiments and data analysis. ML and AR drafted the manuscript. ML and CF carried out the functional assays; OB contributed with TCGA data analysis and gave intellectual input for project discussion. TM and MM collected, organized and provided access to the samples in the Biobank and HD was involved in patient treatment. EH was in charge of the pathological classification of the samples. All authors contributed to the scientific discussion

of the data, read and confirmed the final version of the manuscript.

Acknowledgments

The scientific development and funding of this project were (in part) supported by the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network, and the German Center for Lung Research (DZL). This work was in part supported by the National Institute of Health (USA) [grant number CA148127] and earlier sample collection by the Deutsche Krebshilfe [grant numbers 70-2387; 70-2919, 106910]. We also thank the DKFZ Genomics and Proteomics Core Facility for their support of the scientific development of this project, and Michael Boutros for providing the pRL-Actin reporter construct.

Disclosure of interest

Dr. Amos is a CPRIT Research Scholar (this should be added to the grant as an accolade required by the Cancer Prevention Research Institute of Texas). Dr. Meister reports grants from BMBF, German Center for Lung Research, during the conduct of the study. Dr. Muley reports grants and personal fees from BMBF, German Center for Lung Research (DZL), during the conduct of the study; grants and personal fees from Roche, outside the submitted work. Dr. Risch reports grants from NIH, grants from German Center for Lung Research (DZL), during the conduct of the study. All other authors have nothing to disclose.

Funding

This work was supported by the Deutsche Krebshilfe [70-2387; 70-2919, 106910]; National Institutes of Health [CA148127]; Deutsches Zentrum für Lungenforschung (DZL).

Data and code availability

Datasets and code used in this study have been uploaded to ZENODO for reproducibility of the analysis DOI: 10.5281/zenodo.4327115. Datasets include Raw on target methylation profiles per sample (34 samples in total), DMR per patient, DMR enrichment results per patient, MDMRs, TF Motif enrichment results in MDMRs, TF Motif Scan results within target regions, Sequencing target regions, Reference Sequence against which the alignment was done and TF Motifs from HOCOMOCO database. Codes include: code used for Plotting: [Figure 1](#), [Figures 2 c and d](#), [Figure 3](#), [FigureS1](#), [FigureS2](#), [FigureS4](#), Code used to calculate MDMRs, Parameters for DMR calling (using MethylKit) and Code used for TF Motif Enrichment in MDMRs.

Ethics, consent and permissions

Tissue samples were provided by Lungbiobank Heidelberg (Thoraxklinik, University Hospital Heidelberg, Germany) a member of the Biomaterial Bank Heidelberg (BMBH) and the Biobank Platform of the German Centre for Lung Research (DZL) in accordance with the regulations of the tissue bank and the approval of the ethics committee of Heidelberg University. All participants that contributed samples for research purpose provided written informed consent. The study was approved by the Ethics committee of the Medical Faculty of the University of Heidelberg (No. 270/2001).

ORCID

Marina Laplana  <http://orcid.org/0000-0002-2548-0704>
 Matthias Bieg  <http://orcid.org/0000-0002-3606-4917>
 Matthias Schlesner  <http://orcid.org/0000-0002-5896-4086>
 Angela Risch  <http://orcid.org/0000-0002-8026-5505>

References

- [1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*. 2016;66(1):7–30.
- [2] Wang Y, McKay JD, Rafnar T, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet*. 2014;46(7):736–741.
- [3] Landi MT, Chatterjee N, Yu K, et al. A Genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet*. 2009;85(5):679–691.
- [4] Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008;452(7187):633–637.
- [5] McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet*. 2017;49(7):1126–1132.
- [6] Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*. 2008;40(5):616–622.
- [7] Hu Z, Wu C, Shi Y, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet*. 2011;43(8):792–796.
- [8] Lan Q, Hsiung CA, Matsuo K, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet*. 2012;44(12):1330–1335.
- [9] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519–525.

- [10] The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543–550.
- [11] Shi J, Marconett CN, Duan J, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat Commun*. 2014;5:3365.
- [12] Heyn H, Sayols S, Moutinho C, et al. Linkage of DNA methylation quantitative trait loci to human cancer risk. *CellReports*. 2014;7(2):331–338.
- [13] Scherf DB, Sarkisyan N, Jacobsson H, et al. Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of CHRN4. *Oncogene*. 2012;32(28):3329–3338.
- [14] Jones GT, Marsman J, Bhat B, et al. DNA methylation profiling identifies a high effect genetic variant for lipoprotein(a) levels. *Epigenetics*. 2020;15(9):949–958.
- [15] World Health Organization Classification of Tumours. Pathology and genetics of tumours of the lung, pleura, thymus and heart. International Agency for Research on Cancer IARC. 2004.
- [16] Hovestadt V, Jones DTW, Picelli S, et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*. 2014;510(7506):537–541.
- [17] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
- [18] Database of Single Nucleotide Polymorphisms (dbSNP). 2015. Bethesda (MD): National center for biotechnology information, national library of medicine. (dbSNP Build ID: 135). Available from: <http://www.ncbi.nlm.nih.gov/SNP/>[Internet]
- [19] Liu Y, Siegmund KD, Laird PW, et al. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol*. 2012;13(7):R61.
- [20] Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13(10):R87.
- [21] Li S, Garrett-Bakelman FE, Akalin A, et al. An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics*. 2013;14 (Suppl 5):S10.
- [22] Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, et al. Extensive variation in chromatin states across humans. *Science*. 2013;342(6159):750–752.
- [23] Ehrlich M. DNA hypomethylation in cancer cells. *Epigenomics*. 2009;1(2):239–259.
- [24] Yu H, Takeuchi M, LeBarron J, et al. Notch-modifying xylosyltransferase structures support an SNi-like retaining mechanism. *Nat Chem Biol*. 2015;11 (11):847–854.
- [25] Arab K, Park YJ, Lindroth AM, et al. Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A. *Mol Cell*. 2014;55(4):604–614.
- [26] McHugh BJ, Murdoch A, Haslett C, et al. Loss of the integrin-activating transmembrane protein Fam38A (Piezo1) promotes a switch to a reduced integrin-dependent mode of cell migration. *PLoS One*. 2012;7(7):e40346.
- [27] Cheng Y, Yan Z, Liu Y, et al. Analysis of DNA methylation patterns associated with the gastric cancer genome. *Oncol Lett*. 2014;7(4):1021–1026.
- [28] Hogan PG, Chen L, Nardone J, et al. Transcriptional regulation by calcium, calcineurin, and NFAT. *Genes Dev*. 2003;17(18):2205–2232.
- [29] Mancini M, Toker A. NFAT proteins: emerging roles in cancer progression. *Nat Rev Cancer*. 2009;9 (11):810–820.
- [30] Pan M-G, Xiong Y, Chen F. NFAT gene family in inflammation and cancer. *Curr Mol Med*. 2013;13 (4):543–554.
- [31] Shou J, Jing J, Xie J, et al. Nuclear factor of activated T cells in cancer development and treatment. *Cancer Lett*. 2015;361(2):174–184.
- [32] Jauliac S, López-Rodríguez C, Shaw LM, et al. The role of NFAT transcription factors in integrin-mediated carcinoma invasion. *Nat Cell Biol*. 2002;4(7):540–544.
- [33] Yoeli-Lerner M, Yiu GK, Rabinovitz I, et al. Akt blocks breast cancer cell motility and invasion through the transcription factor NFAT. *Mol Cell*. 2005;20 (4):539–550.
- [34] Liu M, Chen J, Hu L, et al. HORMAD2/CT46.2, a novel cancer/testis gene, is ectopically expressed in lung cancer tissues. *Mol Hum Reprod*. 2012;18 (12):599–604.
- [35] Wang H, Lou D, Wang Z. Crosstalk of genetic variants, allele-specific DNA methylation, and environmental factors for complex disease risk. *Front Genet*. 2019;9: 695.