Hotspot DNMT3A mutations in Clonal Hematopoiesis and Acute Myeloid Leukemia sensitize cells to Azacytidine via viral mimicry response

Marina Scheller

Anne Kathrin Ludwig

Stefanie Göllner

Christian Rohde

Stephen Krämer

Sina Stäble

Maike Janssen

James-Arne Müller

Lixiazi He

Nicole Bäumer

Christian Arnold

Joachim Gerß

Maximilian Schönung

Christian Thiede

Christian Niederwieser

Dietger Niederwieser

Hubert Serve

Wolfgang E. Berdel

Ulrich Thiem

Inga Hemmerling

Florian Leuschner

Christoph Plass

Matthias Schlesner

Judith Zaugg

Micheal Milsom

Andreas Trumpp

Caroline Pabst

Daniel B. Lipka

Carsten Müller-Tidow (✉ Carsten.Mueller-Tidow@med.uni-heidelberg.de)

Method Article

Abstract

Somatic mutations in DNA Methyltransferase 3A (*DNMT3A*) are among the most frequent alterations in clonal hematopoiesis (CH) and Acute Myeloid Leukemia (AML), with a hotspot in exon 23 at arginine 882 (*DNMT3A*-R882). Here we demonstrate that *DNMT3A*-R882H-dependent CH- and AML cells are specifically susceptible to the hypomethylating agent azacytidine (AZA). Addition of AZA to chemotherapy prolonged AML survival solely in patients with *DNMT3A*-R882 mutation, suggesting its potential as a novel predictive marker for AZA response. AML and CH mouse models confirmed AZA susceptibility specifically in *DNMT3A*-R882H-expressing cells. Hematopoietic stem and progenitor cells expressing *DNMT3A*-R882H exhibited cell-autonomous viral mimicry response as a result of focal DNA hypomethylation at retrotransposon sequences. Administration of AZA boosted hypomethylation of retrotransposons specifically in *DNMT3A*-R882H expressing cells and maintained elevated levels of canonical interferon-stimulated genes (ISGs), thus leading to suppressed protein translation and increased apoptosis.

Introduction

Reagents

Equipment

Procedure

1. Whole-genome bisulfite sequencing (WGBS)

Samples for DNA methylome analysis were collected 10 days after the last AZA or control treatment. At this time, widespread re-methylation occurs after AZA exposure, a phenomenon that is clinically known and which has been described in multiple experimental settings. We focused on DNA methylation changes that persisted after AZA exposure of wildtype and mutant DNMT3A as these effects may determine the long-term consequences of AZA treatment in DNMT3A mutants.

For mouse LSK cells DNA input amount was limited, which is why we used the tagmentation-based whole genome bisulfite sequencing (TWGBS) that has been described previously[1,2]. Genomic DNA from FACS-sorted LSK cells (10 - 30 ng) was used as input to prepare two sequencing libraries, each. Each of the sequencing libraries was subjected to 125 bp paired-end sequencing on one lane of an Illumina HiSeq 2000 instrument. TWGBS raw data can be accessed in Gene Expression Omnibus (GSE146907).

For human AML blasts, DNA was isolated from $1.2 - 5.8 \times 10^5$ flow-sorted CD34+ cells and submitted to the Genomics and Proteomics Core Facility of the German Cancer Research Center. WGBS libraries were prepared using the Accel-NGS® Methyl-Seq DNA Library Kit (Swift Biosciences) according to the

manufacturer's recommendations. Human WGBS raw-data have been deposited in the European Genome Archive (EGAS00001004825).

2. Sequence alignment and methylation calling

Sequence alignment was performed using an updated version of the workflow published in [1] that was implemented as a Roddy Workflow (https://github.com/DKFZ-ODCF/AlignmentAndQCWorkflows) and run using OTP[2]. For TWGBS data raw reads were aligned against the mm10 reference genome extended with the PhiX and lambda phage sequences and with the human *DNMT3A* coding sequence (CCDS33157, CCDS release 22). Briefly, before alignment, trimmomatic[3] was used for adapter trimming and reads were *in silico* converted (C>T for the first read in the pair, G>A for the second). Alignments of the individual libraries were performed against the concatenated C>T and G>A *in silico* converted reference genomes using bwa mem (https://arxiv.org/abs/1303.3997) with default parameters. After alignment, reads were converted back to their original state. PCR duplicate removal was performed per library using Picard MarkDuplicates (http://broadinstitute.github.io/picard) before merging reads from all libraries per replicate. Mapping rates (computed with samtools flagstat[4]), insert size distributions and genome coverage statistics were checked to ensure alignment quality. Methylation calling and M-bias trimming was performed using the bistro methylation caller (version 0.2.0) (https://github.com/stephenkraemer/bistro). Automatic M-bias detection was done using the 'binomp' algorithm to remove both the gap repair nucleotides introduced by the tagmentation reaction (first 9 bp following sequencing primer 2) and additional read positions with M-bias if necessary. Only reads with a mapping quality ≥25 and bases with Phred-scaled quality score of ≥25 were considered. Bisulfite conversion rates were estimated using the non-CpG methylation levels.

For human WGBS data, alignment was performed against the hs37d5 reference genome extended with the PhiX and lambda phage sequences. The workflow was equivalent to the alignment of the mouse TWGBS data, with the following exceptions: i) due to the different library preparation protocol, only a single library per sample was created and no merging step was necessary ii) methylation calling and read trimming (first 10bp of both reads) was performed using methylctools, version 1.0.0 (https://github.com/hovestadt/methylCtools).

3. Detection of differentially methylated regions

Mouse data: Differentially methylated regions (DMR) were called between all conditions using DSS[5]. Differential methylation at each CpG site was tested without prior smoothing to avoid artificial erosion of sharp DMR boundaries. DMR calling was performed with i) a minimum methylation delta of 10% ii) a minimal DMR length of 50bp iii) a minimal number of 3 CpGs iv) a minimum fraction of differentially methylated CpGs of 50% (p-value < 0.05). DMRs separated by less than 50 bp were joined. For integrated

analyses, DMRs from all pairwise comparisons were combined (union of the genomic intervals with 50bp slack). This approach led to the identification of 15,566 DMRs, covering 72,369 CpGs (2,929,913 bp). To obtain DMRs with a high signal-to-noise ratio, we filtered for DMRs with a total coverage >= 30 and an absolute average methylation difference >= 10% between the lowest and highest methylated condition, leading to a final DMR set comprising 15,468 DMRs.

Human data: Pairwise DMR calling was performed using DSS as described above with the exception that DNA methylation levels were smoothed using a smoothing span of 500bp prior to differential methylation testing, to account for the lower available coverage. DSS was used to perform pairwise DMR calls for the following comparisons: i) DNMT3A-WT (control, untreated) vs. DNMT3A-WT (AZA-treated), ii) DNMT3A-R882H (control, untreated) vs. DNMT3A-R882H (AZA-treated), iii) DNMT3A-WT (control, untreated) vs. DNMT3A-R882H (control, untreated), iv) DNMT3A-WT (AZA-treated) vs. DNMT3A-R882H (AZA-treated). To reduce the number of false positives, the required methylation delta was increased to 20% for the DNMT3A-WT (control, untreated) vs DNMT3A-WT (AZA-treated) and DNMT3A-R882H (control, untreated) vs DNMT3A-R882H (AZA-treated) comparisons.

4. PCA and Clustering

Principal component analysis (PCA) was performed based on the average DNA methylation levels present in the mm10 ensembl regulatory regions[6] for the CH12-LX and MEL cell lines. DMRs were clustered using hierarchical clustering (Ward's method) of the z-score normalized average DMR methylation levels. Partitioning was performed with the cutreeHybrid algorithm of the Dynamic Tree Cut package[7] using the python implementation of the algorithm (https://github.com/kylessmith/dynamicTreeCut). Partitioning was performed with parameters maxCoreScatter=0.95, minClusterSize=0.005 * number of DMRs, pamStage=False and minGap=0.3.

5. Gene annotation

DMR gene annotations were used for enrichment analysis and for integration of human and mouse DMR sets (see below). Annotation was performed using the gtfanno package (https://github.com/stephenkraemer/gtfanno). Promoters were defined as regions 5000 bp upstream and 1000 bp downstream of the TSS. Mouse and human DMRs were annotated using Gencode (release 19).

6. Enrichment analysis

Gene-set enrichment and genomic region set enrichment analysis was performed using *regionset_profiler* (https://github.com/stephenkraemer/regionset_profiler). Enrichment of genomic region sets in individual

DMR clusters against the background of all other DMRs was tested using Fisher's exact test (two-sided). Repeat region sets were retrieved from the mm10 repeat masker track provided by the UCSC genome browser. Genome-wide transcription factor binding site predictions based on the scanMotifGenomeWide.pl algorithm from the HOMER toolbox[8] were obtained from the HOMER download page: http://homer.ucsd.edu/homer/data/motifs/homer.KnownMotifs.mm10.170917.bed.gz. Curated ChIP-seq transcription factor binding sites as determined by the Encode and Codex projects were obtained from the LOLA database (http://big.databio.org/regiondb/LOLACore_180412.tgz)[9]. For gene-set enrichment, only promoter DMRs were considered and gene-set membership for each DMR was determined based on its gene annotation. Association between cluster membership and gene-set membership was tested using Fisher's exact test (two-sided). Gene-sets were obtained from the Molecular Signatures Database (MSigDB, v6.2)[10] (http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp)

7. RNA sequencing

RNA was isolated from sorted LSK cells using the RNeasy Micro kit (Qiagen). RNA quality was checked using the Bioanalyzer Pico kit. RNA with a RIN value of <8 was excluded from library preparation. Double-stranded cDNA was synthesized and amplified using the SMARTer Ultra Low RNA kit (Clontech). Sequencing libraries were generated with the Illumina Nextera XT kit and sequenced the 150bp paired-end, mode of an Illumina NovaSeq 6000 instrument. Read demultiplexing was performed using Illumina Casava Software v1.8.

Reads were filtered for adapter contamination and quality and aligned to mus musculus genome GRCm38 using Hisat2 version 2.0.5. Fragment counts of mapped read pairs for genes and repetitive elements were extracted with featureCounts (v1.6.2 with the following parameters: -s 0 -p -B -Q 20) using a custom GTF file consisting of exonic gene regions from Ensembl Mus_musculus.GRCm38.98.chr.gtf.gz and repetitive elements (repeatmasker track for mm10 downloaded from UCSC genome Browser at 2018-07-06) excluding simple repeats. Strand unspecific fragment counting was used, since libraries were generated in a non-strand specific manner. Furthermore, we excluded repeat annotations overlapping with exonic regions. All genes with coverage were included into the analysis. Read inspection was performed using IGV (https://igv.org/, v2.5.2). Differential expression and normalized FPKM values were calculated with edgeR (edgeR_3.26.6).

8. Analysis of repetitive element RNA expression

We downloaded the annotation of repetitive regions for mouse genome mm10 from UCSC genome browser. We excluded simple repeats and added the remaining regions to the ENSEMBL gene model GTF file Mus_musculus.GRCm38.92.chr.gtf. We used featureCounts to extracted read counts. The read count

matrix of all genes and repeat families was subject to RPKM calculation and differential gene expression analysis with edgeR.

9. GSEA of TCGA RNA-seq data

We downloaded level 3 processed (normalized RPKM) RNA-seq data from TCGA. We used the available mutation status information from the TCGA MAF file and stratified the patients into two groups: DNMT3A-WT and DNMT3A-R882mut. This data set was analyzed with Gene set enrichment Analysis (GSEA) software.

References

References

1.   Wang, Q*., et al.* Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc* **8**, 2022-2032 (2013).

2.   Reisinger, E*., et al.* OTP: An automatized system for managing and processing NGS data. *J Biotechnol* **261**, 53-62 (2017).

3.   Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

4.   Li, H*., et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

5.   Feng, H., Conneely, K.N. & Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* **42**, e69 (2014).

6.   Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T. & Flicek, P.R. The ensembl regulatory build. *Genome Biol* **16**, 56 (2015).

7.   Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719-720 (2008).

8.   Heinz, S*., et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).

9.   Sheffield, N.C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587-589 (2016).

10.   Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005).