Novel insights into the genetics and epigenetics of MALT lymphoma unveiled by next generation sequencing analyses

Luciano Cascione,[1,2] Andrea Rinaldi,[1] Alessio Bruscaggin,[1] Chiara Tarantelli,[1] Alberto J. Arribas,[1,2] Ivo Kwee,[1,2,3] Lorenza Pecciarini,[4] Afua A. Mensah,[1] Valeria Spina,[1] Elaine Y.L. Chung,[1] Lodovico Terzi di Bergamo,[1,2] Stephan Dirnhofer,[5] Alexandar Tzankov,[5] Roberto N. Miranda ,[6] Ken H. Young,[6] Alexandra Traverse-Glehen,[7] Gianluca Gaidano,[8] Steven H. Swerdlow,[9] Randy Gascoyne,[10] Raul Rabadan,[11] Maurilio Ponzoni,[4] Govind Bhagat,[12] Davide Rossi,[1,13] Emanuele Zucca[13] and Francesco Bertoni[1]

[1]Università della Svizzera italiana, Institute of Oncology Research, Bellinzona, Switzerland; [2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland; [3]Dalle Molle Institute for Artificial Intelligence (IDSIA), Manno, Switzerland; [4]San Raffaele Scientific Institute, Milan, Italy; [5]Institute of Pathology and Medical Genetics, University Hospital Basel, University of Basel, Basel, Switzerland; [6]Department of Hematopathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [7]Anatomie-pathologique, Hospices Civils de Lyon Centre, Lyon, France; [8]Division of Hematology, Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy; [9]Division of Hematopathology, UPMC Presbyterian, Pittsburgh, PA, USA; [10]British Columbia Cancer Agency, Vancouver, BC, Canada; [11]Department of Systems Biology, Department of Biomedical Informatics, Columbia University College of Physicians & Surgeons, New York, NY, USA; [12]Department of Pathology and Cell Biology, Columbia University Medical Center and New York Presbyterian Hospital, New York, NY, USA and [13]Oncology Institute of Southern Switzerland, Bellinzona, Switzerland

Correspondence: FRANCESCO BERTONI - frbertoni@mac.com

Novel insights into the genetics and epigenetics of MALT lymphoma unveiled by next generation sequencing analyses

Luciano Cascione [1,2], Andrea Rinaldi [1], Alessio Bruscaggin [1], Chiara Tarantelli [1], Alberto J. Arribas [1,2], Ivo Kwee [1,2,3], Lorenza Pecciarini [4], Afua A. Mensah [1], Valeria Spina [1], Elaine Y.L. Chung [1], Lodovico Terzi di Bergamo [1,2], Stephan Dirnhofer [5], Alexandar Tzankov [5], Roberto N. Miranda [6], Ken H. Young [6], Alexandra Traverse-Glehen [7], Gianluca Gaidano [8], Steven H. Swerdlow [9], Randy Gascoyne [10], Raul Rabadan [11], Maurilio Ponzoni [4], Govind Bhagat [12], Davide Rossi [1,13], Emanuele Zucca [13], Francesco Bertoni [1].

[1] Università della Svizzera italiana, Institute of Oncology Research, Bellinzona, Switzerland; [2] Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland; [3] Dalle Molle Institute for Artificial Intelligence (IDSIA), Manno, Switzerland; [4] San Raffaele Scientific Institute, Milan, Italy; [5] Institute of Pathology and Medical Genetics, University Hospital Basel, University of Basel, Basel, Switzerland; [6] Department of Hematopathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [7] Anatomie-pathologique, Hospices Civils de Lyon Centre, Lyon, France; [8] Division of Hematology, Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy; [9] Division of Hematopathology, UPMC Presbyterian, Pittsburgh, PA, USA; [10] British Columbia Cancer Agency, Vancouver, BC, Canada; [11] Department of Systems Biology, Department of Biomedical Informatics, Columbia University College of Physicians & Surgeons, New York, NY, USA; [12] Department of Pathology and Cell Biology, Columbia University Medical Center and New York Presbyterian Hospital, New York, NY, USA; [13] Oncology Institute of Southern Switzerland, Bellinzona, Switzerland.

Online Supplementary Materials and Methods

Targeted next generation sequencing

Cases were selected based on the availability of frozen material with a fraction of neoplastic cells in the specimen representing more than 70% of overall cellularity as determined by morphologic and/or immunophenotypic studies, as previously reported [1]. DNA (50 ng) was extracted from fresh/frozen material using the QIAamp mini kit (Qiagen, Germantown, USA) and underwent target capture and enrichment performed with the Nextera Rapid Capture Enrichment (Illumina, USA), with probes designed using Illumina DesignStudio (list of genes in Table S1). Libraries were then quantified using qPCR, diluted to 2 nM and pooled, prior to cluster generation and analysis on Illumina MiSeq sequencer, using MiSeq Reagent kit v3 (600cycles) (San Diego, USA) and 300 bp paired-end reads (up to 9 indexed samples per run). Fastq files of reads where over 70% of reads were above Q30, were demultiplexed with the Consensus Assessment of Sequence and Variation (CASAVA) tool (Illumina), and samples with at least 6Gb of data were used for mapping and variant calling using Burrows-Wheeler Transform [2] and VarScan 2 [3], using the following criteria: 0 allowable ambiguous alignments, at least 90% of a read having to match the reference genome, software set to detect large indels, hiding the unmatched ends of reads, with at least 10% variant allelic fraction, and at least three variant reads to call a variant. We used Picard tools version 1.10 (http://broadinstitute.github.io/picard/) to remove PCR duplicates and GATK version v3.5 for local indel realignment and base quality recalibration, as recommended in GATK best practices. [4-6]. After annotation, the variants were cross referenced with those in the 1000 Genomes Project (accessed in October 2017) [7], dbSNP (version 137) [8], and the Exome Variant Server(ESP 6500) of the NHLBI GO Exome Sequencing Project (URL: http://evs.gs.washington.edu/EVS/): variants with an allele prevalence >1% in the 1000 genomes project were excluded, as well as common variants identified in prior constitutional exome analyses, non-pathogenic variants reported in dbSNP, and low quality calls were filtered out. The remaining variants underwent manual curation and variant prioritization with visual review of alignments. Synonymous variants and intronic variants that were more than 2 bp from the exon/intron junction were excluded. Variants were manually cross-referenced with the Catalog of Somatic Variants in Cancer

(COSMIC v82) [9] and cBioPortal (accessed in February 2018) [10]. Any variant present in the normal samples (N = 9 paired, 2 unpaired) was excluded from analysis.

<u>Sanger Sequencing</u>

Recurrent mutations deemed to be pathogenic were confirmed by Sanger sequencing as previously described [11]. Primers flanking the mutations of interest were: 5'-CCCCAGAAGGACACTCAAAA-3' and 5'-ACAGCTTGCAGGTGGATTCT-3' for 106157970 frameshift deletion; 5'-AGACTTGCCGACAAAGGAAA-3' and 5'-GGGGGCAAAACCAAAATAAT-3' for the 106194057 stop gain; 5'-ACAGACTGCAGGGACAATGA-3' and 5'-GCCTTCAATTCAATCCATCC-3' for the 106156675 stop gain. Data were evaluated using Mutation Surveyor (Softgenetics, State College, USA).

<u>Copy number variation analysis</u>

Genome-wide DNA profiling was performed using the Affymetrix SNP6 (*N* = 10), the HumanOmni2.5 Beadchip (*N* = 14), or the HumanOmni5-Quad BeadChip (*N* = 14), following manufacturers' protocols. Already published profiles were used for 34 cases [1]. The raw copy number was extracted from CEL files as previously reported [12]. Genomic profiles were segmented with the 'Fast first-derivative segmentation' (FFSEG) algorithm [13, 14]. Profiles were considered of poor quality and discarded from further analyses if they showed severe over-segmentation or no aberrations at all upon evaluation by two independent investigators [14]. Focal aberrations were additionally identified using the Genomic Identification of Significant Targets in Cancer (GISTIC) 2.0 algorithm following the default criteria on the filtered and merged segments [15]. Changes affecting known tumor suppressor genes, which were also part of the targeted sequencing panel (KMT2C, KMT2D, CREBBP, TET2, LRP1B, SPEN, TNFAIP3, PRDM1, PTPRB, TBL1XR1, TNFRSF14, EP300, ITPKB, TP53, CD58, KLF2, PTPRD, RB1, B2M), were manually reviewed by visual inspection.

<u>Whole-transcriptome sequencing (RNA-seq).</u>

RNA was isolated from frozen lymphoma samples (lung, n=10; salivary glands, n.=4; thyroid, n.=3; stomach, n=2; ocular adnexal and others, n.=1 each) by Trizol (Invitrogen - Thermo Fisher Waltham MA, USA) and then DNAse treated using RNase-free DNase Kit (Qiagen, Germantown, MD, USA). Initial RNA Quality control for was performed on the Agilent BioAnalyzer (Agilent Technologies, California, USA) using the RNA 6000 Nano kit (Agilent Technologies) and concentration was determined by the Invitrogen Qubit (Thermo Fisher Scientific) using the RNA BR reagents (Thermo Fisher Scientific). The TruSeq RNA Sample Prep Kit v2 for Illumina (Illumina, San Diego, CA, USA) was used for cDNA synthesis and addition of barcode sequences. The sequencing of the libraries was performed via a paired end run on an HiSeq Illumina sequencer (Illumina, San Diego, CA, USA). The raw reads were quality assessed using fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)[16]. For each sample, the distribution of unique, multi- and unmapped reads was checked for high proportion of unmapped or multi mapped reads. Reads obtained from RNA sequencing were mapped against the *human* hg38 genome build using the Genecode version 22 annotation [17]. Alignment was done with STAR (v2.4.0h) [18], counting of reads overlapping gene features with HTSeq-Count [19]. Differential gene expression analysis was performed using the voom/limma [20] R package. Transcripts that were expressed at > = 1 count per million mapped reads in > = 10 samples were considered for furthers analyses. Immunoglobulin and T cell receptor transcripts were discarded. Chimeric transcripts were detected by deFuse [21] and Chimerascan [22] and Pegasus [23] was applied for fusion annotation and functional selection. Functional annotation was done using Gene Set Enrichment Analysis (GSEA) on fold-change pre-ranked lists with the MSigDB 5.2 Hallmark genesets collection (hallmark) [24] and with genesets obtained from different publications as reported, applying as thresholds an absolute normalized enrichment score (NES) > 1.5 and *P* and false discovery rate (FDR) values <0.05. Profiling data are available at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) database.

<u>Methylation profiling</u>

DNA samples (lung, n=9; ocular adnexal, n.=7; salivary glands, stomach, thyroid and others, n.=1 each) underwent bisulphite treatment with the EZ DNA methylation kit (Zymo Research, Irvine, CA, USA) and then were hybridized on the Infinium 450K BeadChip arrays (Illumina, San Diego, CA, USA), following the manufacturers' protocols. Signal intensities and beta values were exported from *Beadstudio* 2.0 software (Illumina Inc.), applying default settings. All bioinformatic processing was performed with R version 3.2.0. Raw intensity signals were imported and processed using the minfi package [25]. All samples had an average detection p-value < 0.001, indicating good quality data. For quality control (QC), histograms and boxplots were plotted for signal intensities and beta values. Therefore, no sample was removed from the analysis. All samples were Illumina and quantile normalized to reduce technical bias between Type 1 and Type 2 probes. β-values and M-values were calculated in minfi. No further batch correction method was performed. Because the presence of SNPs inside the probe body or at the nucleotide extension can have important consequences on the downstream analysis, we removed such probes. For clustering and statistical tests, probes mapped on sex chromosomes were excluded. Unsupervised analysis was performed on all the retained probes using hierarchical clustering with Euclidean distance and Ward linkage on the beta-values. Supervised analysis of differential methylation between groups was performed using both Limma and Fisher's exact test on the whole CpG-probe set β values, treating the latter as continuous or categorical data, respectively. For the Fisher's exact test, the probes were classified as "methylated" (β value ≥0.3) or "unmethylated" (β value <0.3). The Benjamini-Hochberg multiple test correction was applied to control for false positives. Probes showing an unadjusted FDR < 0.05 were considered differentially methylated.

Morphology

Assessment of plasmacytic differentiation was assessed in hematoxylin and eosin (H&E) stained formalin-fixed paraffin-embedded tissue sections from 19 cases, as previously described [26].

References

1.   Rinaldi A, Mian M, Chigrinova E, et al. Genome-wide DNA profiling of marginal zone lymphomas identifies subtype-specific lesions with an impact on the clinical outcome. Blood. 2011;117(5):1595-1604.
2.   Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-595.
3.   Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-576.
4.   DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-498.
5.   McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303.
6.   Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(11 10 11-33.
7.   The Genomes Project C. A global reference for human genetic variation. Nature. 2015;526(68.
8.   Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-311.
9.   Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research. 2017;45(D1):D777-D783.
10.   Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1.
11.   Mensah AA, Rinaldi A, Ponzoni M, et al. Absence of NOTCH1 gene mutations in MALT lymphomas. Br J Haematol. 2012;157(3):382-384.
12.   Chigrinova E, Rinaldi A, Kwee I, et al. Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. Blood. 2013;122(15):2673-2682.

13.	Mian M, Rinaldi A, Mensah AA, et al. Del(13q14.3) length matters: an integrated analysis of genomic, fluorescence in situ hybridization and clinical data in 169 chronic lymphocytic leukaemia patients with 13q deletion alone or a normal karyotype. Hematol Oncol. 2012;30(1):46-49.

14.	Rinaldi A, Kwee I, Young KH, et al. Genome-wide high resolution DNA profiling of hairy cell leukaemia. Br J Haematol. 2013;162(4):566-569.

15.	Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.

16.	Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 2014.

17.	Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760-1774.

18.	Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

19.	Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.

20.	Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

21.	McPherson A, Hormozdiari F, Zayed A, et al. deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. PLoS Comput Biol. 2011;7(5):e1001138.

22.	Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27(20):2903-2904.

23.	Abate F, Zairis S, Ficarra E, et al. Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. BMC Syst Biol. 2014;8(97.

24.	Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-15550.

25.	Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363-1369.

26.	Ganapathi KA, Jobanputra V, Iwamoto F, et al. The genetic landscape of dural marginal zone lymphomas. Oncotarget. 2016;7(28):43052-43061.

27.	Rossi D, Trifonov V, Fangazio M, et al. The coding genome of splenic marginal zone lymphoma: activation of NOTCH2 and other pathways regulating marginal zone development. J Exp Med. 2012;209(9):1537-1551.

28.	Piva R, Deaglio S, Fama R, et al. The Kruppel-like factor 2 transcription factor gene is recurrently mutated in splenic marginal zone lymphoma. Leukemia. 2015;29(2):503-507.

29.	Spina V, Khiabanian H, Messina M, et al. The genetics of nodal marginal zone lymphoma. Blood. 2016;128(10):1362-1373.

30.	Chng WJ, Remstein ED, Fonseca R, et al. Gene expression profiling of pulmonary mucosa-associated lymphoid tissue (MALT) lymphoma identifies new biological insights with potential diagnostic and therapeutic applications. Blood. 2009;113(3):635-645.
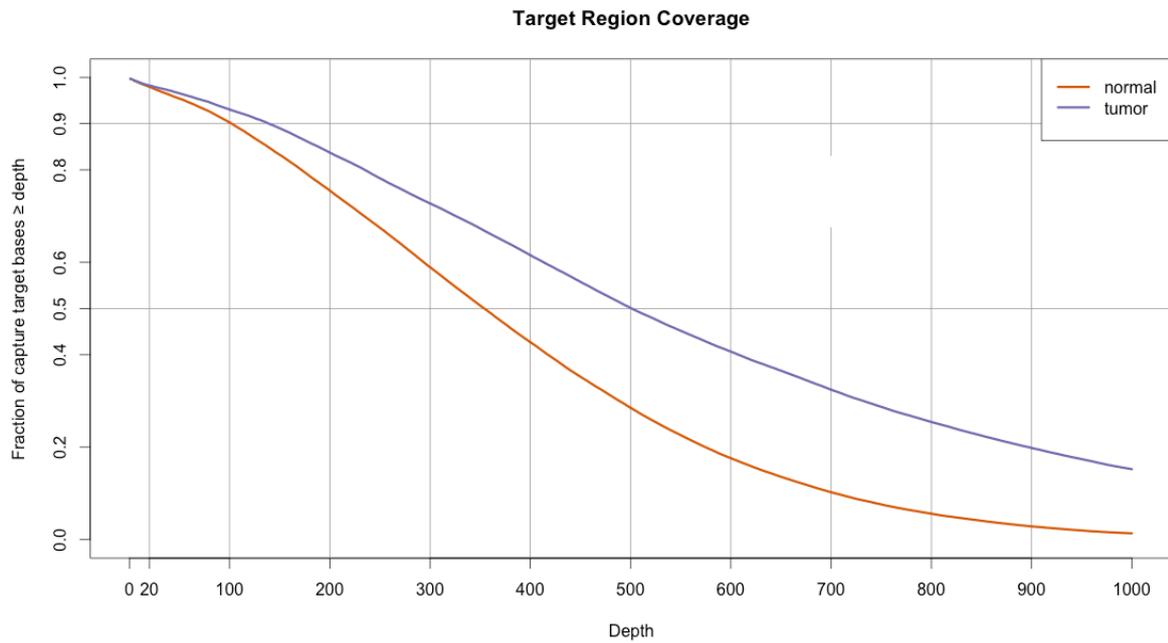
Figure S1. **Coverage of target regions.** The percent of total targeted bases covered with more than or equal to specified depths in normal and tumor samples. For target regions, ~90% had at least one hundred times sequence coverage.
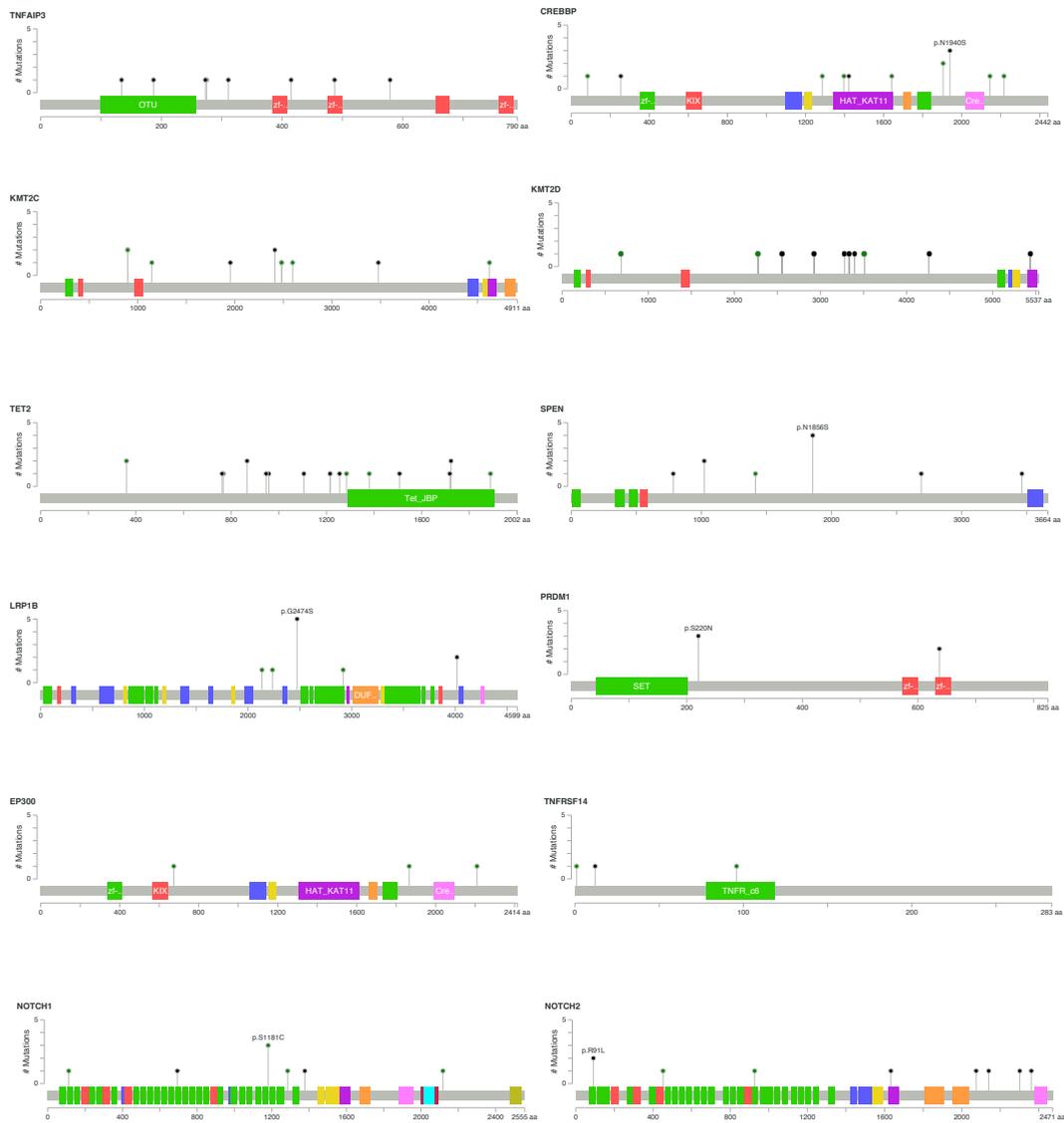
Figure S2. Lollipop plots with the distribution of mutations on the linear proteins and domains of the most affected genes. Each lollipop denotes a unique mutation location and its height represents the number of observed mutations (Table S1). Green circles indicate a frameshift, nonsense, or a splice site mutation (missense). Black circles denote truncating mutations. Grey circles indicate either silent or nonsynonymous mutations. Colored bars indicate the individual protein domains.
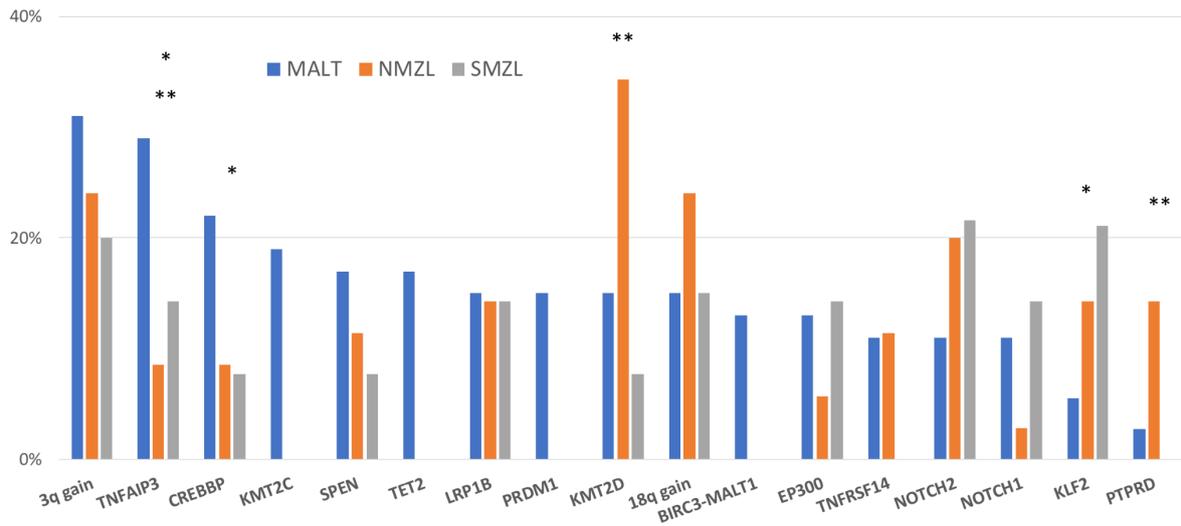
Fig S3. Distribution of genetic lesions across the three MZL types. *, P < 0.05 when MALT lymphomas are compared to SMZL, **, P < 0.05 when MALT lymphomas are compared to NMZL. Data for MALT lymphoma from the current paper, data for SMZL and NMZL from literature [1, 27-29].

Enrichment plot:
UP_IN_MALT_TRANSLOCATED_VS_NONTRANSLOCATED_MALT_PMID18974375

NES 2.4
P < 0.001
FDR < 0.001
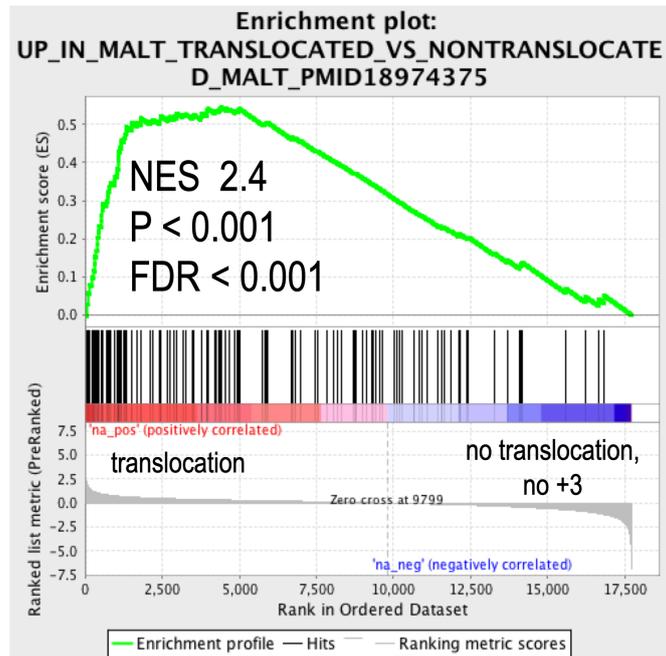
Figure S4. Enrichment of the *BIRC3-MALT1* gene expression signature in MALT lymphomas bearing the *BIRC3-MALT1* fusion compared versus cases lacking the lesions. Green line, enrichment score; bars in the middle portion of the plots show where the members of the gene set appear in the ranked list of genes; Positive or negative ranking metric indicates, respectively, correlation or inverse correlation with the profile; NES, normalized enrichment score.
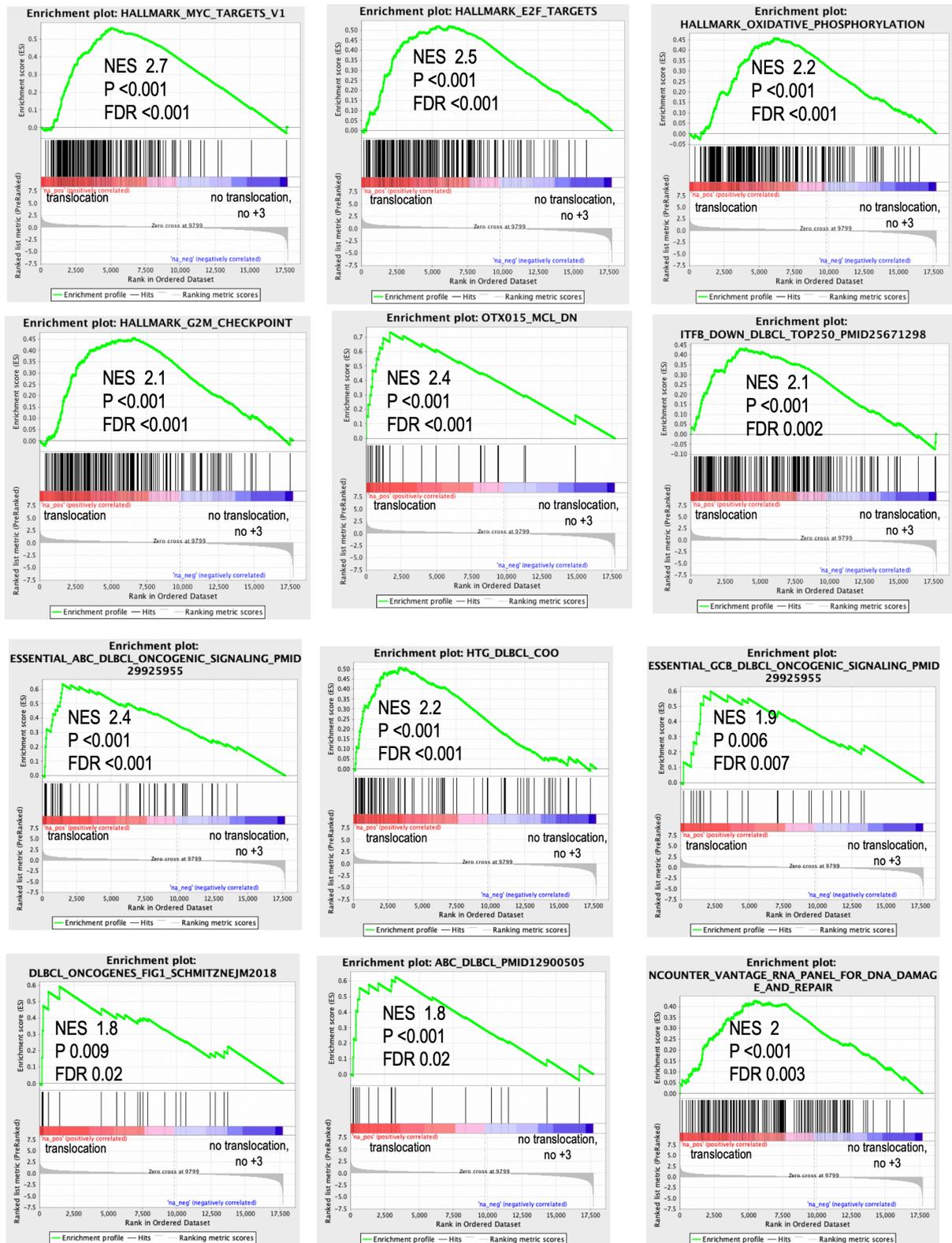
Figure S5. **GSEA plots of MALT lymphomas bearing the *BIRC3-MALT1* fusion compared versus cases lacking the lesions.** Green line, enrichment score; bars in the middle portion of the plots show where the members of the gene set appear in the ranked list of genes; Positive or negative ranking metric indicates, respectively, correlation or inverse correlation with the profile; NES, normalized enrichment score.
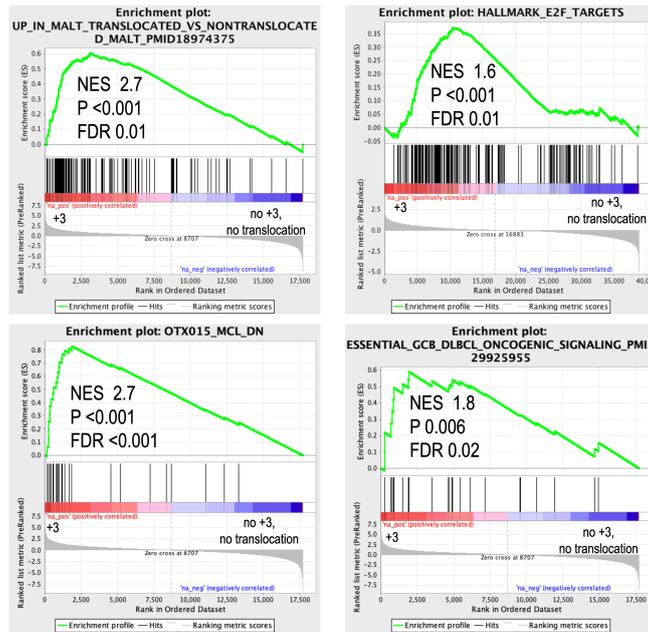
Figure S6. **GSEA plots of MALT lymphomas bearing the trisomy 3 and compared versus cases lacking the lesions.** Green line, enrichment score; bars in the middle portion of the plots show where the members of the gene set appear in the ranked list of genes; Positive or negative ranking metric indicates, respectively, correlation or inverse correlation with the profile; NES, normalized enrichment score.
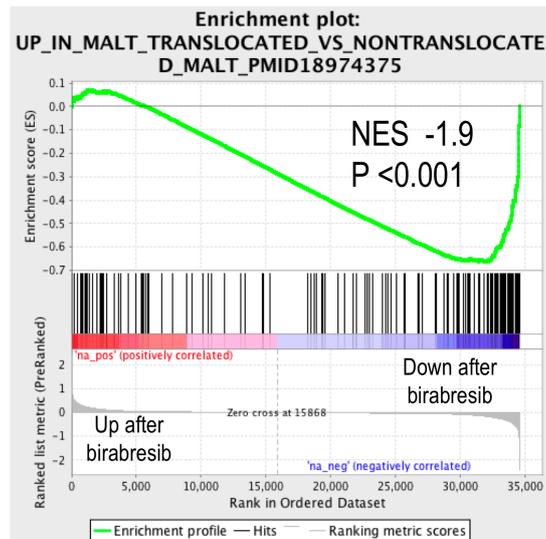
Figure S7. GSEA plot of the BIRC3-MALT1 signature [30] in MZL cell lines exposed to DMSO or to the BET inhibitor birabresib. Gene expression profiling analysis was performed in VL51, Karpas1718 and SSK41 cell lines exposed to DMSO or birabresib (500 nM) for 2, 4, 8 or 12 hours. Green line, enrichment score; bars in the middle portion of the plots show where the members of the gene set appear in the ranked list of genes; Positive or negative ranking metric indicates, respectively, correlation or inverse correlation with the profile; NES, normalized enrichment score.

Table legends

Table S1. Targeted region and achieved coverage.

Table S2. Detected genetic lesions in 72 MALT lymphoma cases. A, genes affected by at least one event (single nucleotide or copy number variation). B, individual detected SNVs.

Table S3. Distribution of genetic lesions across the three MZL types. Data for MALT lymphoma from the current paper, data for SMZL and NMZL from literature [1, 27-29].

Table S4. Anatomical distribution of the most common lesions.

Table S5. Methylation profiling on MALT lymphomas. A, Results of limma test based on the presence or absence of *TET2* mutations; B, Results of Fisher test based on the presence or absence of *TET2* mutations; C, Results of Fisher test based on the presence or absence of trisomy 3 or *BIRC3-MALT1* fusion; D, GSEA of genes differentially methylated based on the presence or absence of *TET2* mutations.

Table S6. Transcriptome analysis of MALT lymphomas. Differentially protein coding transcripts and their functional annotation using GSEA in MALT lymphoma subgroups based on the presence or absence of the *BIRC3-MALT1* fusion (A-B) or of the trisomy 3 (B-C).