Supplementary Material

UNRAVELING THE GENETICS OF TRANSFORMED SPLENIC MARGINAL ZONE LYMPHOMA

Marta Grau, Cristina López, Alba Navarro & Gerard Frigola et al.

SUPPLEMENTARY METHODS

Conventional and molecular cytogenetics

Cytogenetic analyses were done at diagnosis (n=12) and at transformation (n=11). Karyotypes were described according to the International System for Human Cytogenetic Nomenclature (ISCN).[1] Complex karyotype (CK) was considered when at least three clonal chromosomal abnormalities were detected. FISH was done using 7q32 (*IRF5*) locus-specific (Empire Genomics, New York, United States) at diagnosis and transformation, *BCL2* breakapart (BA), *BCL6* BA, and *MYC* BA probes at transformation with available material. *BCL3* BA probe was used in one patient at diagnosis and transformation, t(11;14)(q13;q32) was used in one patient at transformation. All FISH probes were provided by MetaSystems (Altlussheim, Germany). Hybridizations were performed according to the manufacturer's protocols. At least 100 nuclei were examined for each probe. Digital image acquisition, processing, and evaluation were performed using ISIS digital image analysis version 5.0 (MetaSystems).

Analyses of target next-generation sequencing (NGS) panel

Variant calling was performed using an updated version of our in-house pipeline.[2] Briefly, raw reads were trimmed using the SurecallTrimmer (v4.0.1, AGeNT, Agilent). Alignment of the trimmed reads was performed using minimap2 algorithm[3], PCR or optical duplicates were marked using MarkDuplicates from Picard (RRID: SCR_006525), and the base quality score recalibration was performed using GATK’s BaseRecalibrator and ApplyBQSR functions (RRID: SCR_001876 v4.0). Variant calling was performed in parallel using VarScan2 (v2.4.3)[4], Mutect2, VarDictJava (v1.4)[4], LoFreq (v2.1.3.1)[5], outLyzer (v1.0)[6], and freebayes (v1.1.0)[7]. Variants identified were annotated using snpEff/snpSift (v4.3t). Only variants that were identified as “PASS” by at least 4 of the algorithms. All mutations called by at least 4 algorithms were manually reviewed on Integrative Genomic Viewer (IGV). Variants that were likely false positive calls (noisy region, low quality reads) were filtered out. Mutations enriched in chimeric/hard-clipped reads supporting the mutation were flagged by considered in downstream analyses. Paired information (diagnosis and transformation) was used to curate and validate the mutations identified.

Variants reported in the 1000 Genome Project, ExAC and/or gnomAD with a population frequency >1% were considered polymorphisms and therefore removed from the analysis. To further filter out non-recurrent polymorphisms, variants were only considered somatic if 1) they were not reported as germline in our custom International Cancer Genome Consortium (ICGC) database of 506 chronic lymphocytic leukemia analyzed by whole-genome/exome sequencing[8];

and were 2) reported as somatic in lymphoid neoplasm in COSMIC database, 3) truncating, or 4) predicted as potentially damaging by at least one of the following algorithms: CADD (phred score > 10), PolyPhen2 (score > 0.9), SIFT (score < 0.1), MutationAssessor (score > 2) and Provean (deleterious).

Statistical modeling for recurrent alterations

After the preprocessing, filtering and curation of the NGS and copy number alteration (CNA) data, we had information about the presence or absence of 58 recurrent alterations in 59 splenic marginal zone lymphomas (SMZL) samples, 27 obtained at the time of diagnosis (SMZL) and 32 at the time of transformation (SMZL-T). These 59 samples belonged to 41 different patients, for 9 of them we only had the diagnostic sample, for 14 only the transformation sample, and for 18 patients at both time points. Additionally, only the 38 alterations with at least 5 altered samples were considered for the analyses explained in the current and the next section.

Statistical model. We assumed that the presence or absence of an alteration (*y*) could be explained by the following mixed effects logistic model:

$$
\begin{gathered}
y_{ij} \sim Bernoulli\left(p_{ij}\right) \\
logit\left(p_{ij}\right) = \alpha_{CASE[i]j} + \beta_j T_i \\
\alpha_{CASE[i]j} \sim Normal\left(\mu_j, \sigma_j^2\right) \\
log\sigma_j \sim Normal(\lambda, \tau^2)
\end{gathered}
$$

where the observed response variable y_{ij} is binary and takes value 1 when sample *i* presents the alteration *j* and 0 otherwise. According to this model, the probability that a sample presents the alteration *j* (p_{ij}) depends on two factors: (*i*) the random intercept $\alpha_{\mathrm{CASE[i]j}}$, which varies case to case and is shared by all samples of the same case, and (*ii*) whether the sample is transformed or not (T_i), scaled according to parameter β_j. The random intercepts (α) for the different cases were assumed to follow a normal distribution with mean μ_j and standard deviation σ_j. Finally, a common log-normal distribution was assumed for the 40 σ_j, which are difficult to estimate individually, in order to borrow information between them and aid in the estimation of each one. Parameters λ and τ control the shape of this log-normal distribution.

Interpretation. Focusing on the interpretation of some of the parameters, a large and positive β_j means that alteration *j* is much more likely to be present in a SMZL-T sample than in a diagnostic sample, whereas a large and negative value means that it is more likely to be present only in the diagnostic samples. The vector of parameters $\boldsymbol{\alpha}_j$ (one α for each case) controls how much the affinity of presenting alteration *j* varies case to case, cases with larger α are more likely to present the alteration in both samples and cases with lower α are more likely to not have it in any sample. Therefore, the variation in the values of vector $\boldsymbol{\alpha}_j$, controlled by the parameter σ_j, is an indication of how correlated are the diagnostic and SMZL-T samples in alteration *j*. A σ_j close to 0 would translate to all values of the vector $\boldsymbol{\alpha}_j$ being very similar and close to μ_j, thus presenting the alteration would not depend on the case and therefore no correlation would be observed between diagnosis and transformation. On the other hand, a high value of σ_j would translate in very different values in vector $\boldsymbol{\alpha}_j$, who would then define which samples present the alteration in both diagnosis and transformation and which ones do not, thus, observing a high correlation. It is important to note that α and β have an additive effect to the probability of presenting the alteration in the logit scale, so the magnitude of how much β changes the probability in the original scale depends upon the distribution of α.

Priors. The above model was estimated using a fully Bayesian approach, which requires that a prior distribution is specified for each parameter of the model. These prior distributions can summarize available previous information and help obtaining better estimations of the parameters, or just be very vague and let the data tell the full story. Using highly informative priors was not possible given the limited/absent literature in transformed SMZL cases, but using semi-informative priors to help bound the posterior distribution to plausible values was possible based on previous information in regular SMZL cases[9,10]. For example, the prior distribution assigned to each β_j was Normal with mean 0 and standard deviation 5, a distribution that contains most of its probability mass between -10 and 10. If β_j = 10, then a case with a 5% probability of presenting the alteration in the diagnosis, would have a probability of almost 99.9% in the transformation. For a β_j = -10, the probability in the transformation would be 0.0002%. This example highlights that values of β_j around -10 or 10 already have an extreme effect to the probabilities and are very unlikely to be true, at least for the majority of alterations, so low prior plausibility was assigned to these extreme values. Similar logic was used to specify priors for the other parameters, which are depicted below:

$$\beta_j \sim Normal(0, 5^2)\forall j$$
$$\mu_j \sim Normal(\text{-}3.5, 5^2)I_{[\text{-}12,2]}\forall j$$
$$\lambda \sim Uniform(0.5,3)$$
$$\tau \sim Uniform(0,1.2)$$

where $I_{[-12,2]}$ denotes that the distribution is truncated to the interval[-12, 2].

Technical details. Simulations of the posterior distribution were obtained with JAGS v4.3.0.[11] Specifically, we used three chains with an adaptation of 40000 simulations and a thinning interval of 6 for the next 600000, ending with a final 300000 ready to be used as an approximation of the posterior distribution once convergence was verified. The large number of posterior samples were needed for the statistical test explained in the next section (*co-ocurrence and mutual exclusivity*).

Results. The posterior distributions with the 95% credible intervals (CI) of λ, τ and each μ_j, σ_j and β_j are represented in Supplementary Figure 6. The σ_j distributions are relatively similar, where a large proportion of their mass is contained in the 3 to 7 range. The reason for this similarity is that each alteration has little information about its own σ_j, so the common distribution dominates all of them and applies a strong shrinkage. The σ_j and λ distributions are far from zero, an expected result that suggests a global moderate correlation between the diagnostic and transformation samples. Supplementary Table 8 contains the point estimate and 95% CI for each β_j, together with the classical *P*-value corresponding to test if $\beta_j=0$. Finally, the posterior distributions of the μ_j, σ_j and β_j parameters can be used to obtain, for each alteration, an estimation of the proportion of altered cases in the diagnosis and in the transformation.

The SMZL-SMZL-T bivariate 95% CI of each alteration are represented as contour lines in Supplementary Figure 7. Contour lines far from the red line (equal proportion in diagnosis and transformation) indicate larger differences in the proportions of both times.

Co-occurrence and mutual exclusivity analysis (COME)

The co-occurrence and mutual exclusivity between any pair of alterations is usually tested using the Fisher Exact Test. Due to the complex structure of our data, where there is a mix of independent samples and paired ones, this approach was not possible. Thus, we used a posterior predictive check of the above statistical model for dependence between alterations, given that the model was blind and ignorant about which samples had simultaneously the same alterations. Specifically, for each pair of alterations, we compared the observed co-occurrences in the data against the number of co-occurrences expected by the model, if two alterations are highly associated the observed number of co-occurrences would be much higher than the expected number (or lower in the case of mutual exclusivity).

The first step was to use the posterior distribution of the parameters to simulate replicated datasets with the same structure as the original dataset (same number of diagnosis and transformation samples, same number of paired and unpaired samples, etc.). Then, for each replicate and for each pair of alterations we calculated three variables: the number of altered cases in alteration 1 (X_1), the number of altered cases in alteration 2 (X_2), and the number of co-occurrences (C). So, for each pair of alterations we ended with 300000 replicates of a vector containing these three values ([X_1, X_2, C]). The final step was to compute a two sided *p*-value for each pair according to:

$P_{up}= P(C^{rep}>C^{obs} \mid X_1^{rep} = X_1^{obs}, X_2^{rep} = X_2^{obs})$

$P_{down} = P(C^{rep}<C^{obs} \mid X_1^{rep} = X_1^{obs}, X_2^{rep} = X_2^{obs})$

p-value = $2 \cdot \min(P_{up}, P_{down})$

where the super index *rep* references the replicated values and the super index *obs* references the unique observed value. Of note, the *p*-values were computed conditioning on the number of altered cases being equal to the observed values ($X_1^{rep} = X_1^{obs}$ and $X_2^{rep} = X_2^{obs}$). This conditioning was the reason to obtain a large number of posterior simulations, as we needed a reasonable number of replicates after selecting those that met the condition. For a specific pair of alterations, a low *p*-value would mean that the observed co-ocurrences are far from the co-ocurrences expected by a model that assumes independence, therefore, suggesting that those alterations are associated. This analysis was performed with the 38 alterations present in at least 5 samples. The obtained *P*-values are represented in Supplementary Figure 2.

Analysis of whole genome sequencing (WGS)

Raw reads were mapped to the human reference genome (GRCh37) using the BWA-MEM algorithm (v0.7.15)[12]. BAM files were generated, sorted, indexed and optical/PCR duplicates flagged using biobambam2 (https://gitlab.com/german.tischler/biobambam2,v2.0.65). FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc, v0.11.5) and Picard tools (https://broadinstitute.github.io/picard, v2.10.2) were used to extract quality control metrics.

Single nucleotide variants (SNV) were analyzed using CaVEMan (cgpCaVEManWrapper, v1.12.0)[13], Mutect2 (GATK v4.0.2.0)[14], and MuSE (v1.0 rc)[15] run in paired tumor-normal mode. SNV were normalized using bcftools (v1.8)[16] and intersected using custom scripts. We applied caller-specific filters to remove low quality variants identified by CaVEMan and Mutect2. Variants detected by CaVEMan with CLPM>0 and ASMD values <90, <120 or <140 for sequencing read lengths of 100, 125, or 150 bp, respectively, were excluded. Variants called by Mutect2 with MMQ<60 were eliminated. Mutations detected by at least two algorithms were considered. Short insertions/deletions (indels) were called by Pindel (cgpPindel, v2.2.3)[17], SvABA (v7.0.2)[18], Mutect2[14], and Platypus (v0.8.1)[19]. Pindel, SvABA, and Mutect2 were run in paired tumor-normal mode. The somaticMutationDetector.py script (https://github.com/andyrimmer/Platypus/blob/master/extensions/Cancer/somaticMutation

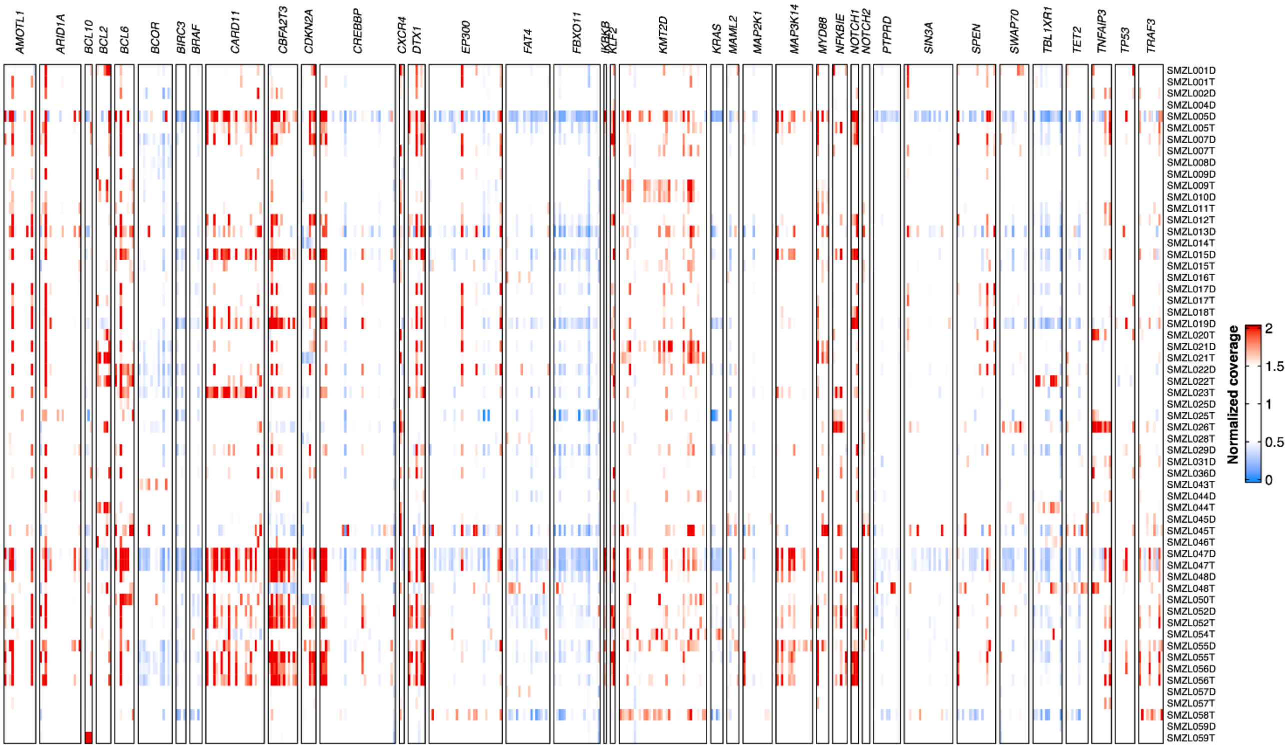
Detector.py) was used to identify somatic indels called by Platypus. Indels were left-aligned and normalized using bcftools[16] and intersected using custom scripts. Caller-specific filters were applied: indels with MMQ<60, MQ<60, and MAPQ<60 for Mutect2, Platypus, and SvABA, respectively, were removed. Only indels identified by at least two algorithms were retained for downstream analyses. Due to the longitudinal nature of the dataset analyzed, SNV called in one timepoint (either SMZL at diagnosis or at transformation) were considered in the second sample if at least one read with the mutation was found in the BAM file usingalleleCounter (min_map_qual=35, min_base_qual=20, https://github.com/cancerit/alleleCount, v4.0.0). Similarly, indels detected in one timepoint were added in the second sample if any of the algorithms detected the alteration independently of its filters. Finally, SNV and indels were annotated using snpEff/snpSift (v4.3t)[20,21] and RefSeq as a reference (GRCh37.p13.RefSeq).

CNA were called using Battenberg (cgpBattenberg, v3.2.2)[22] and ASCAT (ascatNgs, v4.1.0)[17]. CNA within any of the immunoglobulin loci (IGH, IGK, IGL) were filtered out. A consensus of CNA was performed by manual inspection and comparison with CNA data from copy number array data. Structural variants (SV) were extracted using BRASS (v6.0.5)[23], SvABA[18], and DELLY2 (v0.8.1)[24]. The SV identified by the different algorithms were intersected using a custom script considering a window of 300 bp around the breakpoints. For downstream analyses we kept the SV identified by at least two programs if at least one of the algorithms called the alteration with a high quality (MAPQ≥90 for BRASS, MAPQ=60 for SvABA and DELLY2). In addition, IgCaller (v1.2)[25] was used to call SV within any of the immunoglobulin loci. All SV were visually inspected using IGV[26].

SUPPLEMENTARY FIGURES

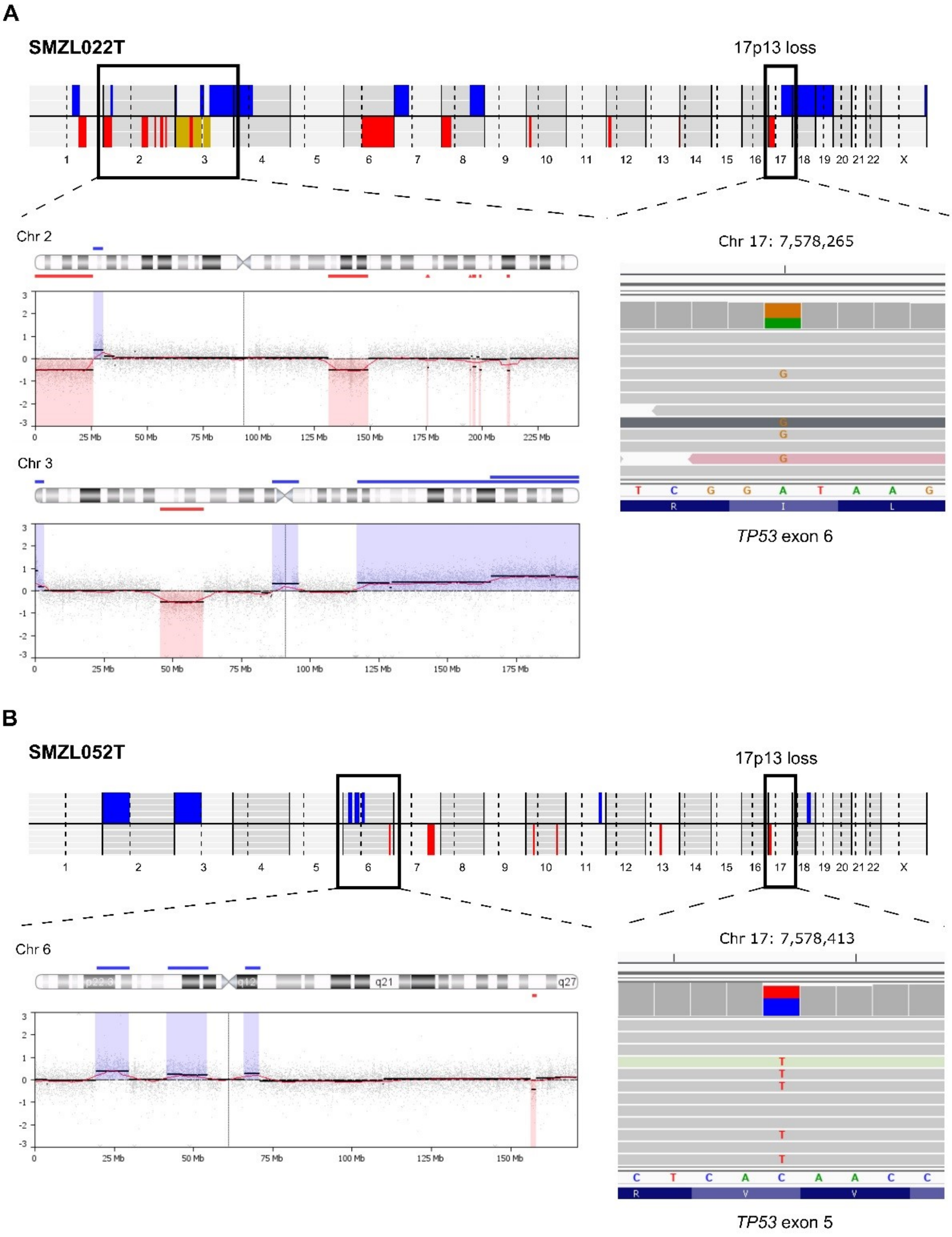
Supplementary Figure S1. Heatmap showing the normalized coverage across the regions captured for variant calling in the 59 samples analyzed by NGS.

Normalized coverage was calculated dividing the mean coverage of each target exon by the mean coverage of the sample. Therefore, a normalized coverage of approximately 1 means a uniform coverage across the studied regions. Samples are represented in rows, while the different exons of each gene are shown in columns, which are grouped by gene.



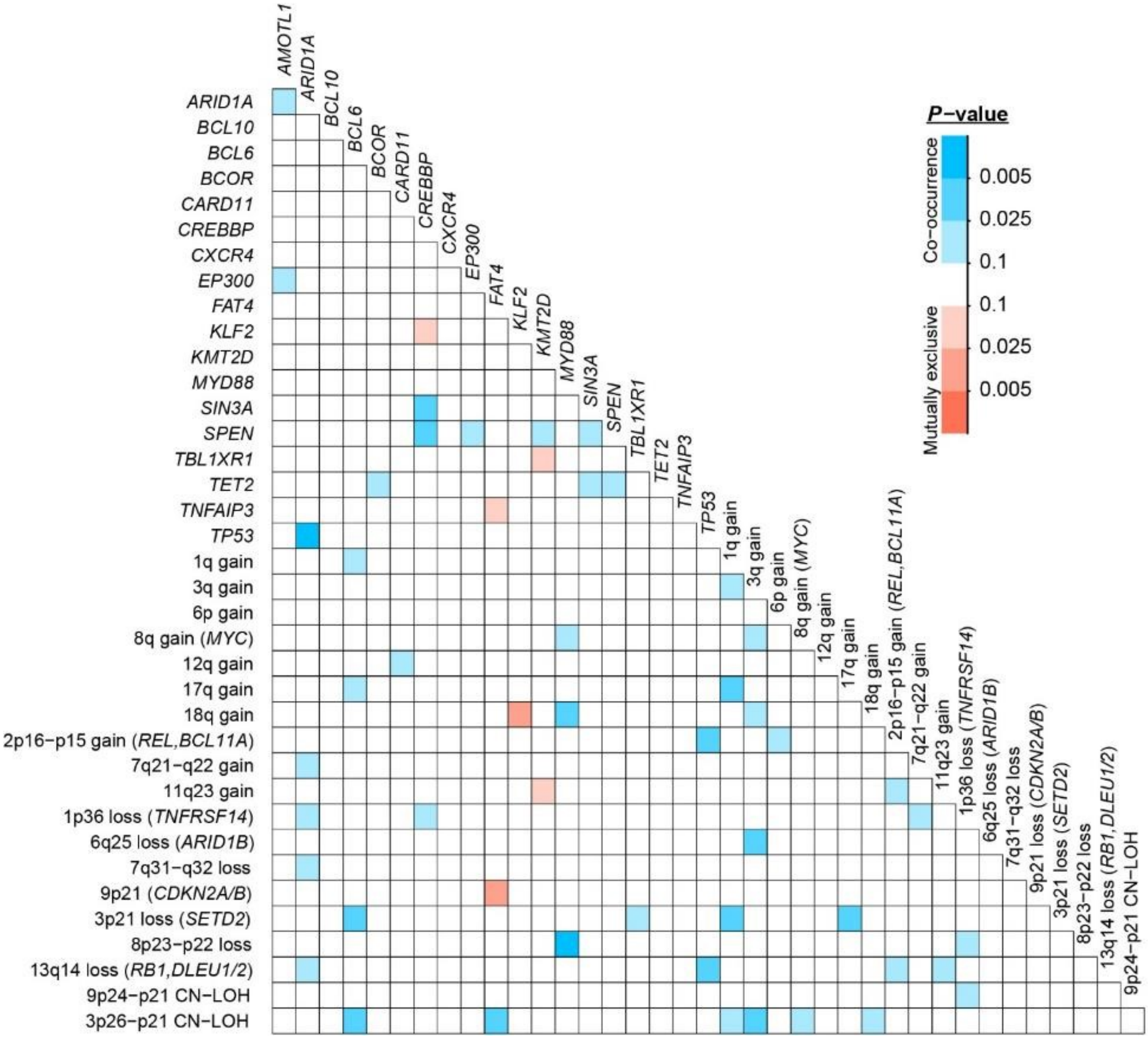
Supplementary Figure S2. SMZL-T cases with a chromothripsis pattern.

The genome-wide copy number profile of the two SMZL-T cases is represented from chromosome 1 to X, and from p-arm to q-arm (chromosome Y is excluded). Copy number gains (blue), losses (red) and copy neutral loss of heterozigosity (yellow) are displayed. An IGV window shows the altered reads (colored nucleotides) of *TP53* gene. (A) Copy number profile of SMZL022T. The chromosomes with chromothripsis (chromosomes 2 and 3) and the biallelic alteration (loss and SNV) on *TP53* are highlighted. (B) Copy number profile of SMZL52T. The chromosome 6 with chromothripsis and the biallelic alteration (loss and SNV) on *TP53* are highlighted.



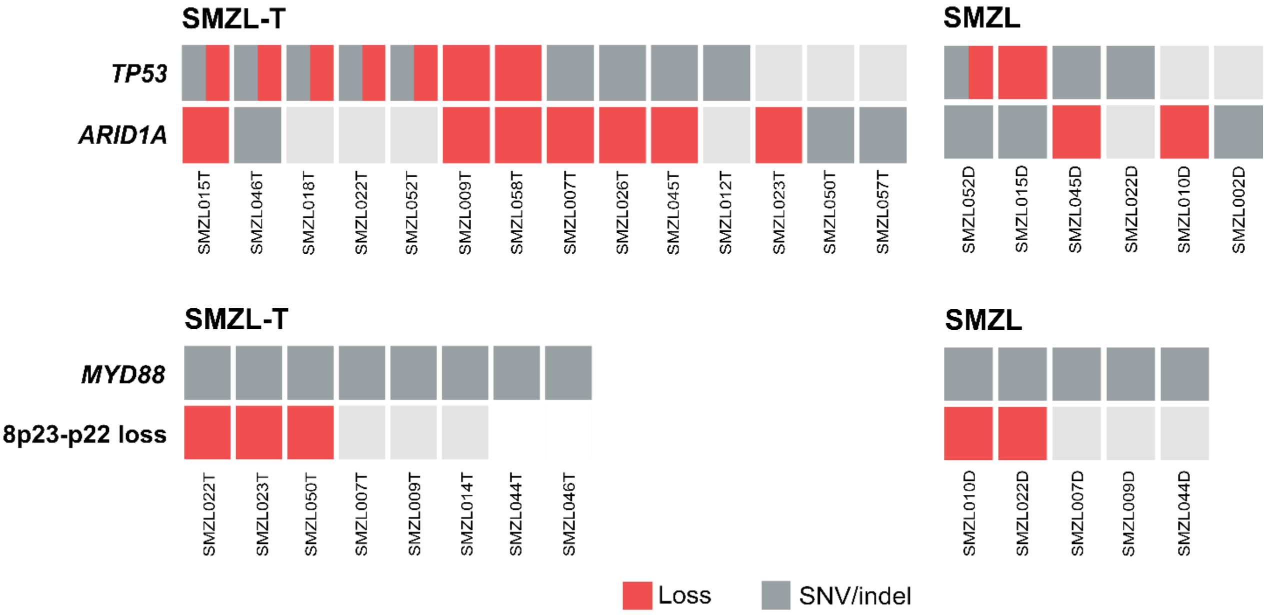
Supplementary Figure S3. Co-occurrence and mutual exclusivity plot.

Heatmap representing the co-occurrence (blue) and mutual exclusivity (red) between alterations (SNV/indels, and CNA). The higher intensity of the colors (blue and red) colors correspond to more significant *P*-values. Only alterations present in at least 5 samples are shown.



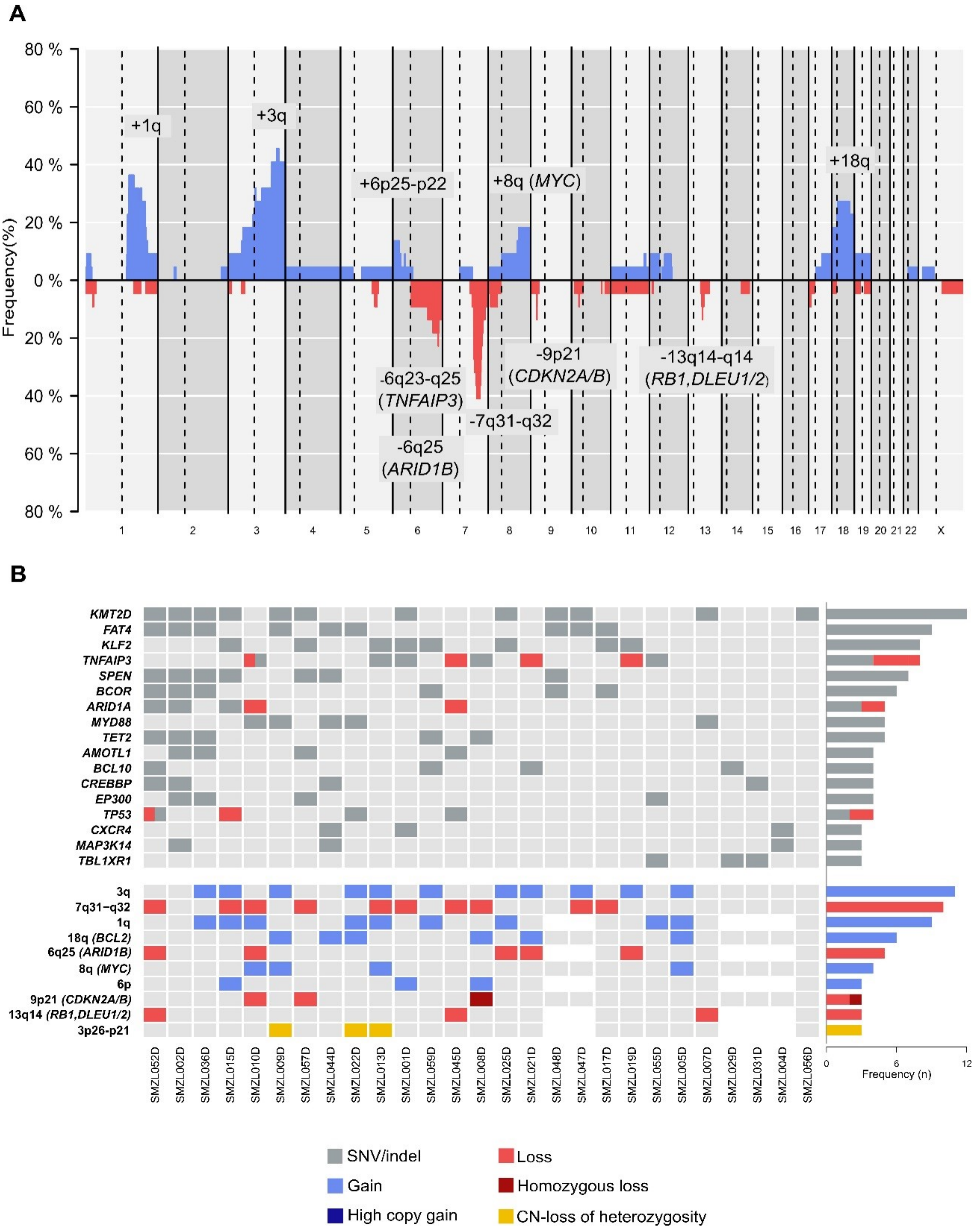
Supplementary Figure S4. Co-occurrence of genomic aberrations.

Oncoprint representations of individual cases with genomic aberrations at diagnosis (right) and SMZL-T (left). There are represented co-occurrences with *P*-value<0.005: *ARID1A* and *TP53; MYD88* and 8p23-p22 loss.



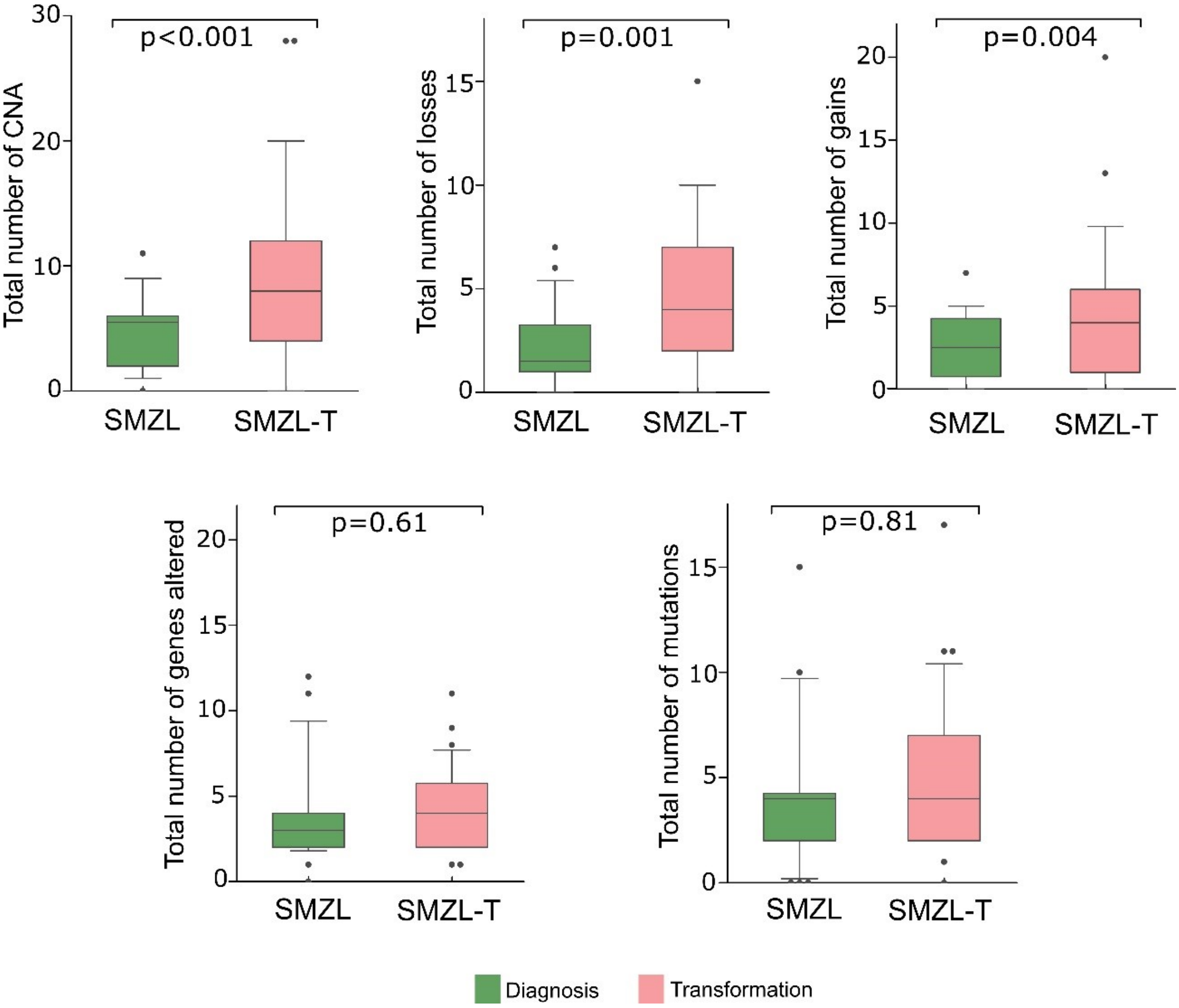
Supplementary Figure S5. Copy number alterations (CNA) and somatic variants (SNV/indel) identified in SMZL patients at diagnosis.

(A) Copy number profile identified by microarray in 22 SMZL patients at diagnosis. Gains (blue), and losses (red). Probes are aligned from chromosome 1 to X, from p-arm to q-arm (chromosome Y is excluded). The recurrently (n≥3) altered genomic regions and candidate target genes are indicated. (B) Oncoprint with the recurrent genetic alterations found in 27 SMZL cases at diagnosis. Upper panel: altered genes by decreasing frequency; bottom panel: altered genomic regions.



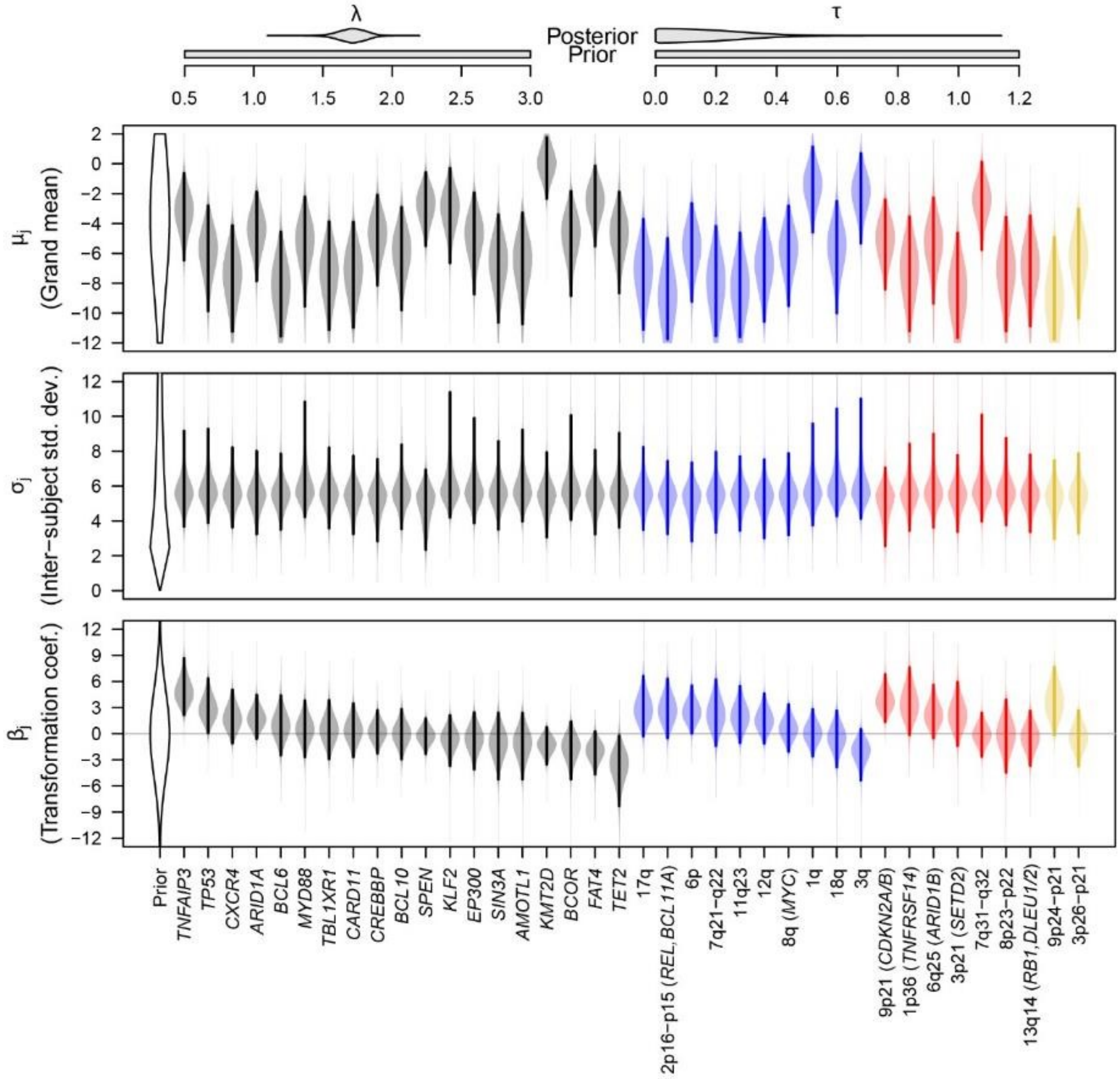
Supplementary Figure S6. Box-plots of copy number alterations (CNA) and SNVs/indels detected in SMZL.

Total number of CNA, losses, gains, genes altered and mutations detected in SMLZ cases at diagnosis (green), and SMZL-T (pink). *P*-values were obtained using mixed-effects negative binomial models, which account for the partially paired structure of the data. CNA, gains and losses, but not mutations, were more present at transformation.



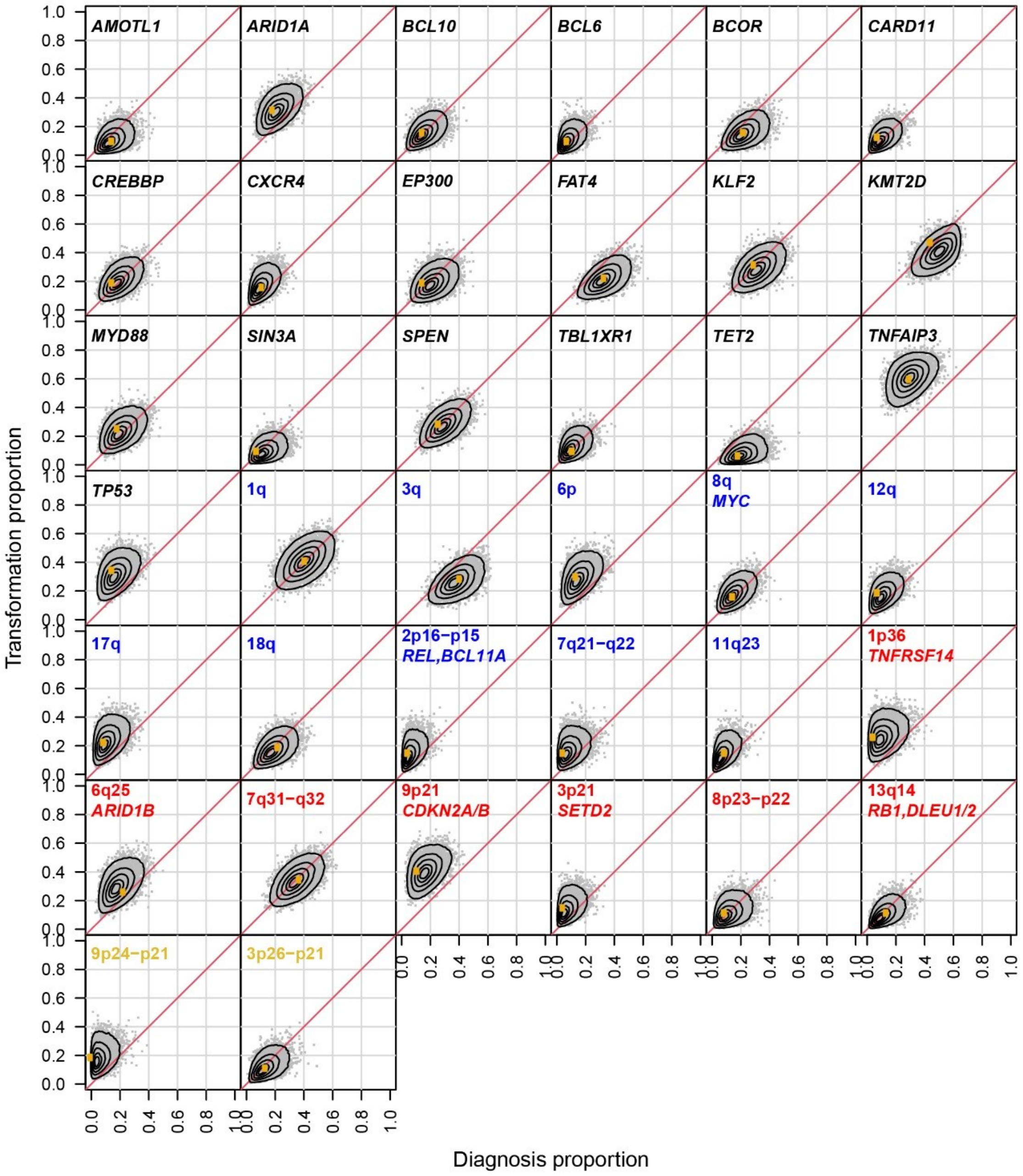
Supplementary Figure S7. Estimation Parameters.

Prior and posterior distributions of all parameters in the statistical model for the presence/absence of recurrent alterations in SMZL-T vs SMZL. The distributions are represented as violin plots and 95% credible intervals are highlighted with thick solid lines. The SNV/indels are labeled in grey, gains in blue, losses in red, and copy neutral loss of heterozigosity in yellow. Alterations in *TNFAIP3*, *TP53*, 6p gains and loss of 9p21 (*CDKN2A/B*) were enriched at SMZL-T.



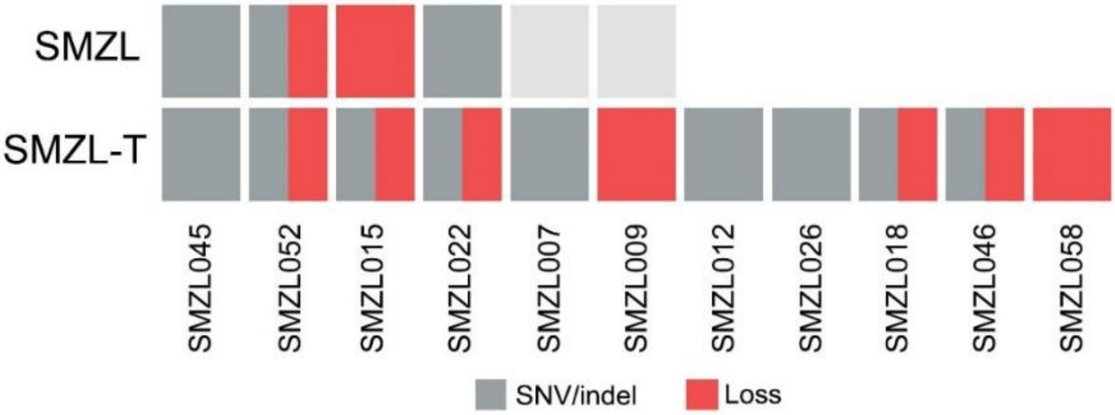
Supplementary Figure S8. Bivariate proportions.

Bivariate posterior distributions of the SMZL proportion of altered cases (x-axis) versus the SMZL-T proportion (y-axis). A different alteration is represented in each panel. Contour lines are drawn at different levels of the bivariate credible interval (95%, 75%, 50%, 25% and 10%). 5000 simulations of the posterior distribution are shown as gray points in the background. The observed proportions are shown as yellow points. SNV/indels are labeled in grey, and CNA are labeled in blue, red, and yellow indicating gain, loss and copy neutral loss of heterozigosity, respectively. Alterations in *TNFAIP3*, *TP53*, 6p gains and loss of 9p21 (*CDKN2A/B*) were enriched at SMZL-T.



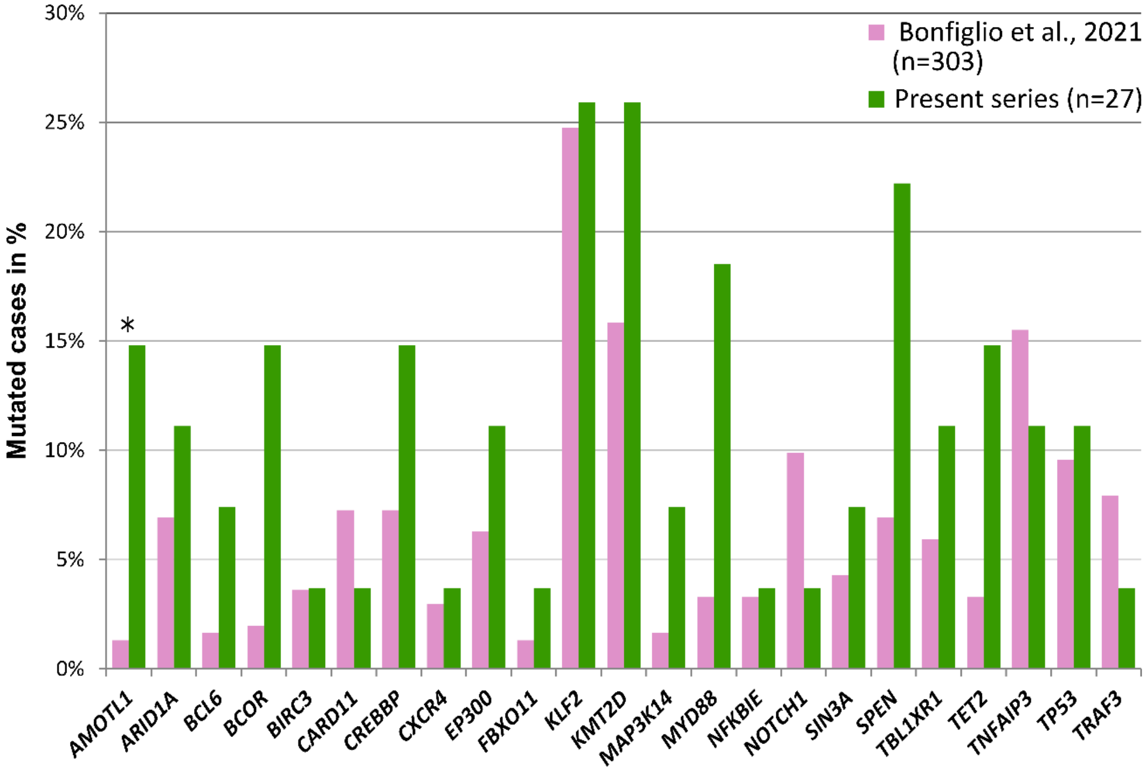
Supplementary Figure S9. Genomic alterations involving *TP53* in SMZL.

Genomic aberrations involving *TP53* gene, integrating SNVs, indels and CNA. Cases with at least one alteration in *TP53* are represented. The SNV/indels are labeled in grey and losses in red, and light gray indicates “not altered”.



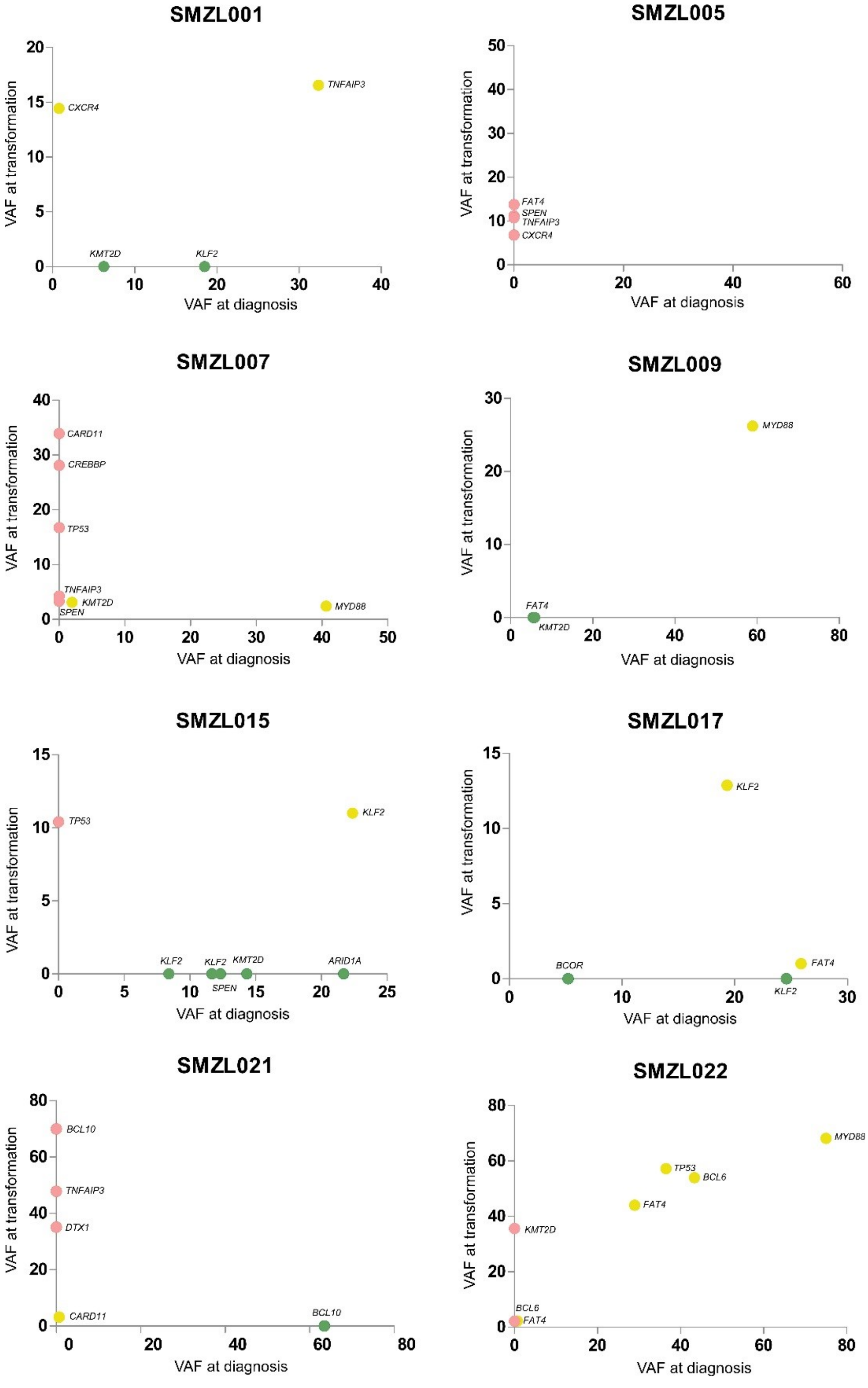
Supplementary Figure S10. Comparison of mutation frequencies between the present series at diagnosis and a published series of 303 SMZL samples.

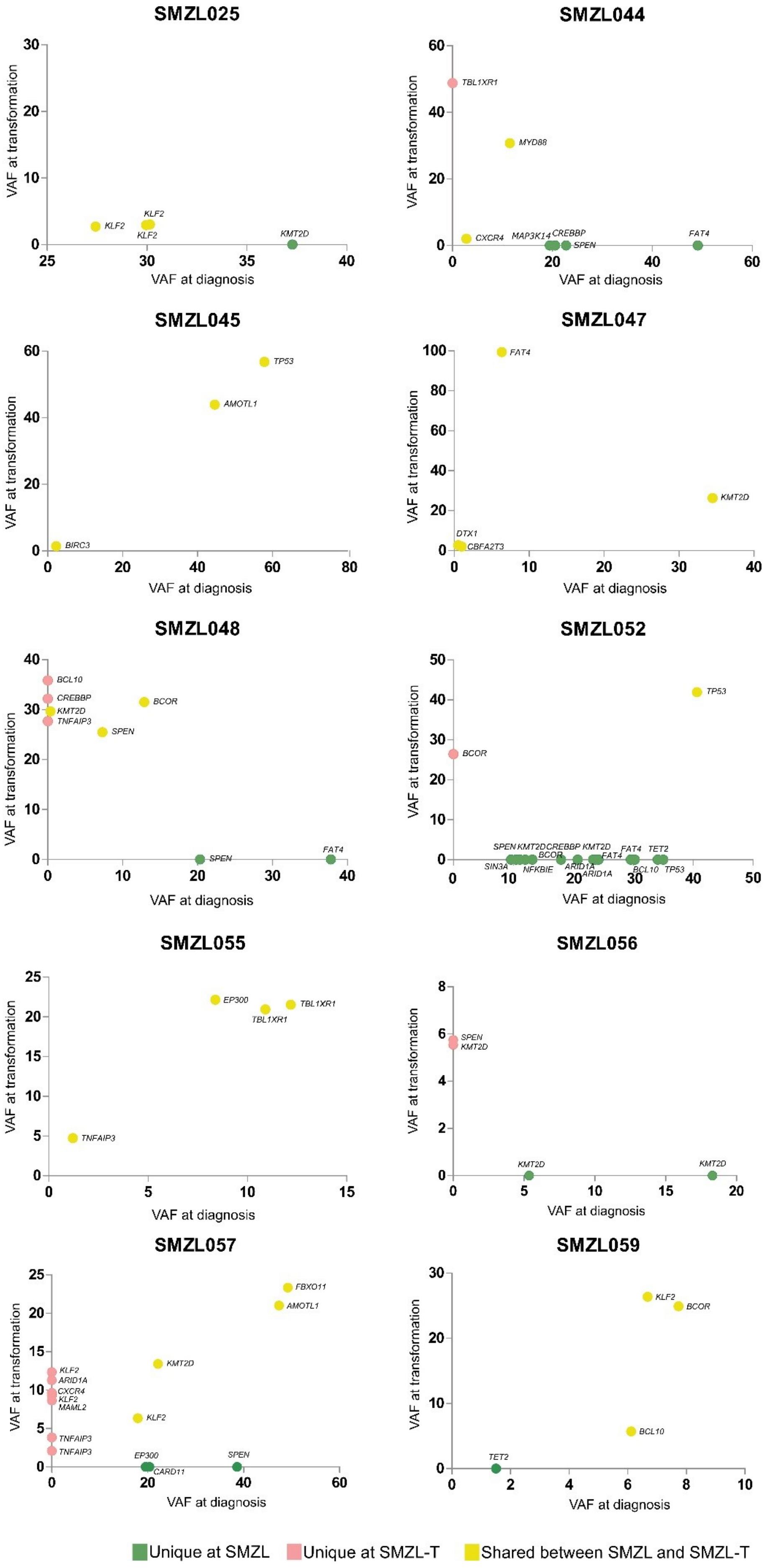
Comparison of the frequently mutated genes in the present series at diagnosis (n=27) and the series published by Bonfiglio et al., 2021[10]. The genes represented had at least 1 SNV/indel and a variant allele frequency higher than 10%.



Supplementary Figure S11. Dynamics of genomic aberrations during SMZL transformation.

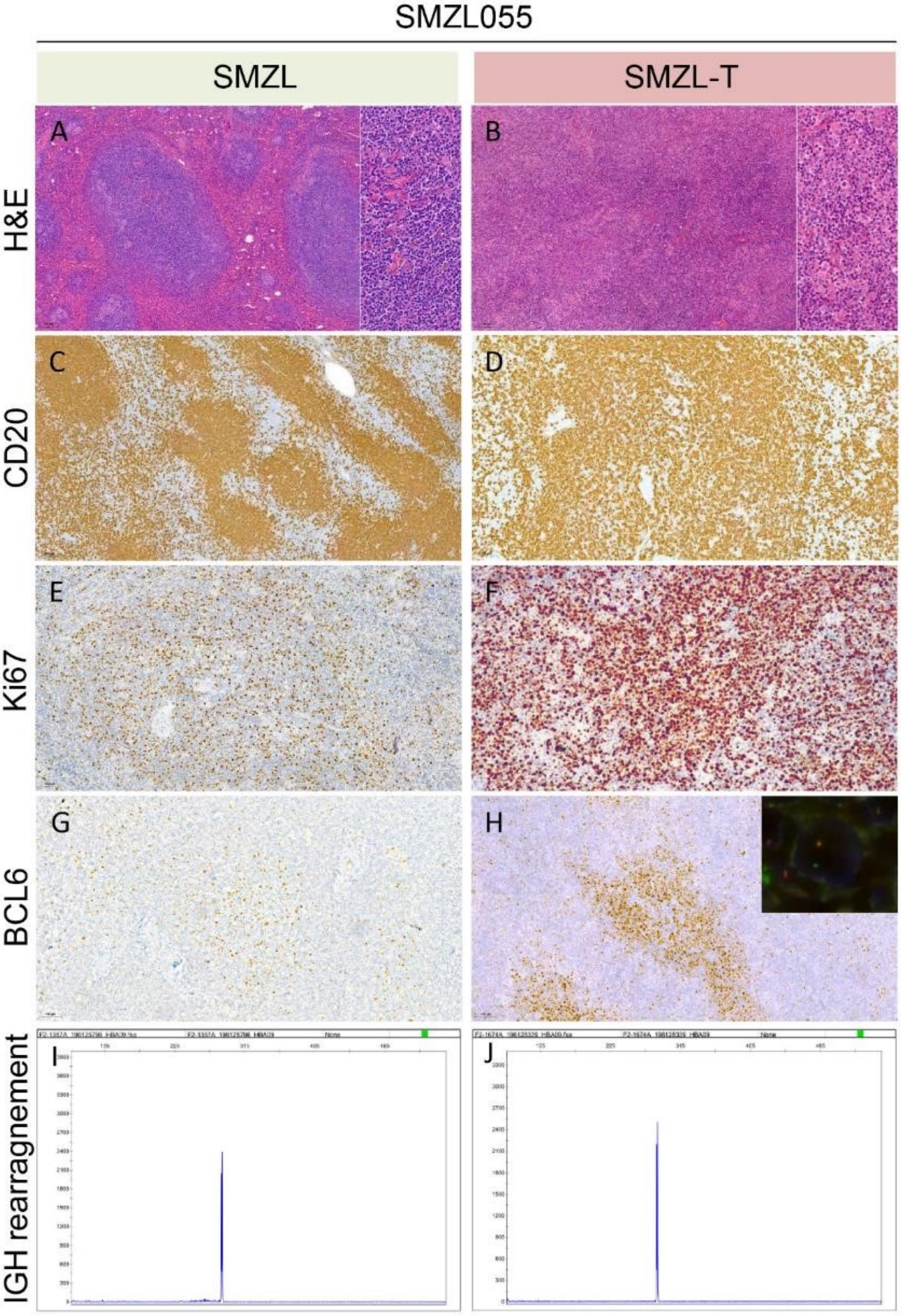
Representation of variant allele frequency (VAF) of each SNV/indel at diagnosis (x-axis) and at transformation (y-axis). Unique alterations at diagnosis (green) or at transformation (pink) or shared alterations (yellow) are represented.



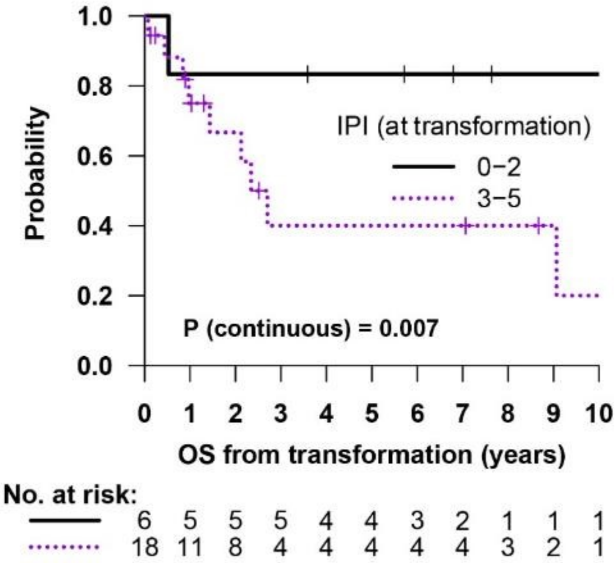

SMZL025
VAF at transformation
VAF at diagnosis
30
20
10
0
25
30
35
40
KLF2
KLF2
KLF2
KMT2D
SMZL044
VAF at transformation
VAF at diagnosis
60
40
20
0
20
40
60
TBL1XR1
MYD88
CXCR4
MAP3K14
CREBBP
SPEN
FAT4
SMZL045
VAF at transformation
VAF at diagnosis
60
40
20
0
20
40
60
80
TP53
AMOTL1
BIRC3
SMZL047
VAF at transformation
VAF at diagnosis
100
80
60
40
20
0
10
20
30
40
FAT4
KMT2D
DTX1
CBFA2T3
SMZL048
VAF at transformation
VAF at diagnosis
40
30
20
10
0
10
20
30
40
BCL10
CREBBP
KMT2D
TNFAIP3
BCOR
SPEN
SPEN
FAT4
SMZL052
VAF at transformation
VAF at diagnosis
50
40
30
20
10
0
10
20
30
40
50
TP53
BCOR
SPEN
KMT2D
CREBBP
KMT2D
BCOR
FAT4
FAT4
TET2
SIN3A
NFKBIE
ARID1A
ARID1A
BCL10
TP53
SMZL055
VAF at transformation
VAF at diagnosis
25
20
15
10
5
0
5
10
15
EP300
TBL1XR1
TBL1XR1
TNFAIP3
SMZL056
VAF at transformation
VAF at diagnosis
8
6
4
2
0
5
10
15
20
SPEN
KMT2D
KMT2D
KMT2D
SMZL057
VAF at transformation
VAF at diagnosis
25
20
15
10
5
0
20
40
60
FBXO11
AMOTL1
KMT2D
KLF2
ARID1A
CXCR4
KLF2
MAML2
KLF2
TNFAIP3
TNFAIP3
EP300
CARD11
SPEN
SMZL059
VAF at transformation
VAF at diagnosis
30
20
10
0
2
4
6
8
10
KLF2
BCOR
BCL10
TET2
🟩 Unique at SMZL
🟥 Unique at SMZL-T
🟨 Shared between SMZL and SMZL-T

Supplementary Figure S12. Histopathological features of case SMZL055 at diagnosis and transformation.

Images A-C-E-G-I correspond to the biopsy obtained at diagnosis, while images B-D-F-H-J correspond to the biopsy at the transformation. (A) Histologically, the diagnostic biopsy of the spleen showed a classic SMZL pattern, with small B cell nodules replacing the germinal centers from the white pulp, effacing the follicular mantles, and infiltrating the red pulp (H&E, 20X, high magnification field at 60X). (B) In the lymph node biopsy obtained at transformation, the normal architecture of the lymph node was effaced by a proliferation of centroblastic-looking cells arranged in a vaguely nodular pattern (H&E, 20X, high magnification field at 60X). (C-D) CD20 highlights the nodular pattern in the diagnosis (C; CD20, 5X), and an area of diffuse pattern in the transformation (D; CD20, 10X). (E-F; Ki67, 10X) Proliferation index assessed with Ki67 staining was higher in the transformation (F) compared to the diagnosis (E)". (G-H, BCL6, 10X) While Bcl6 in the diagnostic biopsy enhanced residual germinal center B-cells (G), Bcl6 expression in the transformation biopsy was more diffuse and intense (H). This increased expression could be attributed to a *BCL6* rearrangement detected by FISH with a breakapart *BCL6* probe (inset, 100X). (I-J) The study of the immunoglobulin heavy chain (*IGH*) gene showed the same clonal peak in both biopsies, confirming a clonal relationship.


SMZL055
SMZL
SMZL-T
H&E
A
B
CD20
C
D
Ki67
E
F
BCL6
G
H
IGH rearragnement
I
F2-1357A 196125786 HBA09.fsa
F2-1357A 196125786 HBA09
None
135
225
315
405
495
3900
3600
3300
3000
2700
2400
2100
1800
1500
1200
900
600
300
0
J
F2-1674A 196128339 HBA09.fsa
F2-1674A 196128339 HBA09
None
135
225
315
405
495
3300
3000
2700
2400
2100
1800
1500
1200
900
600
300
0

Supplementary Figure S13. Kaplan-Meier curve of survival from time of transformation (SFT) according to international prognostic index (IPI) score.



SUPPLEMENTARY REFERENCES

1. McGowan-Jordan J, Hastings RJ, Moore S, GmbH SK. ISCN 2020 an International System for Human Cytogenomic Nomenclature (2020). Basel, Switzerland: Freiburg Karger; 2020.

2. Nadeu F, Delgado J, Royo C, et al. Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood*. 2016;127(17):2122–2130.

3. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–3100.

4. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*. 2012;22(3):568–576.

5. Wilm A, Aw PPK, Bertrand D, et al. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*. 2012;40(22):11189–11201.

6. Muller E, Goardon N, Brault B, et al. OutLyzer: Software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. *Oncotarget*. 2016;7(48):79485–79493.

7. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012; arXiv:1207.3907 [q-bio.GN]

8. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519–524.

9. Parry M, Rose-Zerilli MJJ, Ljungström V, et al. Genetics and Prognostication in Splenic Marginal Zone Lymphoma: Revelations from Deep Sequencing. *Clinical Cancer Research*. 2015;21(18):4174–4183.

10. Bonfiglio F, Bruscaggin A, Guidetti F, et al. Genetic and phenotypic attributes of splenic marginal zone lymphoma. *Blood*. 2022;139(5):732–747.

11. Plummer M. JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. 2003; Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, 20-22 March 2003, 1-10.

12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.

13. Jones D, Raine KM, Davies H, et al. cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Current Protocols in Bioinformatics*. 2016;56(1):15.10.1-15.10.18.

14. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 2010;20(9):1297–1303.

15. Fan Y, Xi L, Hughes DST, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*. 2016;17(1):178.

16. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2).

17. Raine KM, Van Loo P, Wedge DC, et al. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current protocols in bioinformatics*. 2016;56(1):15.9.1-15.9.17.

18. Wala JA, Bandopadhayay P, Greenwald NF, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome research*. 2018;28(4):581–591.

19. Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*. 2014;46(8):912–918.

20. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in Genetics*. 2012;3:35.

21. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80–92.

22. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell*. 2012;149(5):994–1007.

23. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. 2016;534(7605):47–54.

24. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333–i339.

25. Nadeu F, Mas-de-les-Valls R, Navarro A, et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nature Communications*. 2020;11(1):3390.

26. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;29(1):24–26.