Inventory of Online Supplemental Data

Supplemental Methods

Additional details about methods are described here.

Supplemental Tables

Supplemental Figures

Supplemental References

Additional References are listed here.

Supplemental Methods

Library preparation for whole exome sequencing

Briefly, prior to library preparation, 200ng of DNA was fragmented using Covaris sonification to 250 bp and further purified using Agentcourt AMPure XP beads. Size-selected DNA was then ligated to specific adaptors during library preparation using the standard HTP Kapa library preparation kit and the libraries were quantified using a MiSeq Nano.

Mutational Signature Analysis

Computational Algorithms. The mutational signatures discovery is a process of de-convoluting cancer somatic mutations counts, stratified by mutation contexts or biologically meaningful subgroups, into a set of characteristic patterns (signatures) and inferring the activity of each of the discovered signatures across samples[1]. For this purpose, we exploited a Bayesian variant of non-negative matrix factorization (Bayesian NMF) recently implemented and applied to several cancer genome projects (see [2-4] for additional background and technical details regarding the Bayesian NMF methodology). Bayesian NMF exploits a *shrinkage* or *automatic relevance determination* (ARD) technique to allow a sparse representation for both signatures and activities as well as an optimal inference for the number of signatures (*K*) by iteratively pruning away irrelevant components in balancing between a data-fidelity and a complexity[4]. The same parameters set as previously described were used[2,3]. All SNVs were classified to 96 possible mutation types or categories based on six base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) within the tri-nucleotide sequence context including the base immediately 5' and 3' to the mutated base. In addition to 96 tri-nucleotide mutation types, we also considered the clustering information of mutations as an additional feature to capture a signal of the mutational process related to the activation-induced cytidine deaminase (AID signature). As was previously demonstrated[3], there was a substantial difference in mutation spectra between clustered and non-clustered mutations due to a differential activity of both canonical and non-canonical AID signatures. For this reason we first computed NMDs (Nearest Mutation Distance) for all SNVs, a minimum genomic distance to all other mutations on the same chromosome in the same patient, and partitioned them into 'clustered' (NMD <= 10kb) and 'non-clustered' groups (NMD > 10kb) (supplemental Figure 3A). The threshold (10kb) was manually chosen from a bimodal feature of the NMD distribution (supplemental Figure 3A). Then, we separately counted clustered and non-clustered mutations across 96 mutation channels and split mutations in each sample into two columns representing clustered and non-clustered mutational groups, giving rise to the mutation count matrix $\boldsymbol{X}$ (96 by 2*M*, *M* is the number of samples). This mutation count matrix was ingested as an input for the BayesNMF and factored into two matrices, $\boldsymbol{W'}$ (96 by *K*) and $\boldsymbol{H'}$ (*K* by 2*M*), approximating $\boldsymbol{X}$ by $\boldsymbol{W'H'}$. It should be noted that clustered and non-clustered mutations from the same patient were separately handled to capture a characteristic signal from clustered mutations. Through a scaling transformation, $\boldsymbol{X} \sim \boldsymbol{W'H'} = \boldsymbol{WH}$, $\boldsymbol{W} = \boldsymbol{W'U^{-1}}$ and $\boldsymbol{H} = \boldsymbol{UH'}$ where $\boldsymbol{U}$ is a *K*-by-*K* diagonal matrix with the element corresponding to the 1-norm of column vectors of $\boldsymbol{W'}$, resulted in the final signature loading matrix $\boldsymbol{W}$ and the activity loading matrix $\boldsymbol{H}$.

A. De-novo signature discovery in 37 PMBL WES samples. A de-novo signature extraction for 37 PMBL WES samples for SNVs stratified by 96 tri-nucleotide mutation contexts with the clustering identified four major mutational processes (Figure 2A). The similarity of these signatures to known 30 COSMIC signatures (http://cancer.sanger.ac.uk/cosmic/signatures) was computed by a cosine similarity. The first signature most resembling COSMIC1 (cosine similarity

0.93) was characterized by C>T mutations at CpG sites with a background broad spectrum of base substitutions. The activity of this signature was pervasive across samples, explaining about 38% of all mutations. The second signature was most similar to COSMIC2 (cosine similarity 0.92), but spiked C>T and C>G mutations at TCW (W = A/T) suggest that this signature is a mixture of COSMIC2 and COSMIC13 corresponding to the APOBEC mutagenesis. The third signature, which didn't match any known 30 COSMIC signatures with a cosine similarity > 0.58, had characteristic peaks of C>T/G mutations at GCT context corresponding to one of canonical-AID known hotspot motifs at RCY (R = A/G, Y= C/T), and T>A/C/G at TW (W=A/T) context corresponding to non-canonical AID hotspot motifs, and its activity was significantly higher in clustered mutations consistent to known AID biology. The fourth signature most resembled COSMIC26 (cosine similarity 0.93), which is known to be associated with defective DNA mismatch repairs and found in microsatellite unstable tumors, and its activity was exclusive to three hyper-mutant tumors (Figure 2B). Interestingly, we found that the PMBL tumors LS2287 and c_M_1403 had frame-shift mutations in *MLH1* (chr3:37050314_37050315 Frame_Shift_InsT) and the tumors c_M_1403 (chr7:6038873A>C [missense]; chr7:6036955A>C [Splice_Site]; chr7:6026786_6026787 Frame_Shift_InsC) and c_M_07 (chr7: 6029539G>A [nonsense]; chr7: 6042256A>G [missense]) bi-allelic, likely inactivating mutations of *PMS2*, suggesting a possible link of these three hyper-mutant tumors to MSI phenotype.

B. Semi-supervised signature discovery in 34 PMBL WES samples. To minimize a possible interference of the MSI signature with other signatures and enable a separation of the APOBEC signal into COSMIC2 and COSMIC13 we excluded three putative MSI samples and repeated a de-novo signature extraction for 34 PMBL WES samples, while enforcing two APOBEC signatures (COSMIC2 and COSMIC13) in the signature extraction. To enforce APOBEC, we created two simulated samples with a predominant activity of COSMIC2 and COSMIC13 with 10,000 mutations each and added them to the mutation count matrix of 34 PMBL samples. The mutation counts along 96 contexts in two simulated samples were proportionally distributed according to the normalized profiles of COSMIC2 and COSMIC13 signatures. Bayesian NMF was run on the combined real 34 PMBL samples plus two simulated APOBEC samples. Following post-process removal of the simulated samples, Bayesian NMF identified (supplemental Figure 3B), COSMIC1 (cosine similarity 0.92, 76% overall mutations) and the AID signature (9.1% overall mutations), in addition to two APOBEC signatures, COSMIC2 (9.8% mutations) and COSMIC13 (3.5% overall mutations). As expected the AID signature explained a majority of clustered mutations (overall 63%) as seen in supplemental Figure 3C. We used this semi-supervised signature analysis in all downstream analyses.

C. Semi-supervised signature discovery for the combined cohort of 37 PMBL WES and three PMBL cell lines. As described in ***B*** we performed a semi-supervised signature discovery for the combined cohort of 37 PMBL tumors and three PMBL cell lines (Karpas1106, Farage, U2940). In addition to the enforced two APOBEC signatures (COSMIC2 and COSMIC13) we identified five additional signatures (supplemental. Figure 3D); COSMIC1 (cosine similarity 0.92), AID, and two MSI signatures of COSMIC15 (cosine similarity 0.9) and COSMIC26 (cosine similarity 0.94), and COSMIC11 (cosine similarity 0.96). The attribution of COSMIC11, known to be associated with the treatment with alkylating agents, was exclusive to a single cell line (U2940), while the attribution of two MSI signatures (COSMIC15 and COSMIC26) was mostly present in three PMBL tumors and three PMBL cell lines (supplemental Figure 3E).

D. Gene-level Signature enrichment analysis. We annotated each mutation with the probability (likelihood of association) that it was generated by each of the discovered mutational signatures, $P_{m,s}$, where '*m*' denoted a mutation and '*s*' refers to the signature. More specifically, the likelihood of association to the *k*-th signature for a set of mutations corresponding to *i*-th mutation context and *j*-th clustered or non-clustered mutation group was defined as $[\boldsymbol{w}_k\boldsymbol{h}_k / \sum(\boldsymbol{w}_k\boldsymbol{h}_k)]_{ij}$, where $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$ correspond to the *k*-th column vector and *k*-th row vector of $\boldsymbol{W}$ and $\boldsymbol{H}$, respectively. The relative activity enrichment for candidate driver genes in Figure 2C was determined by taking an average of P_{ms} for all mutations in each driver gene.

Immunohistochemistry (IHC) and scoring

Immunohistochemical staining for PAX5 (BD Biosciences, clone 24, 1:350), B2M (Dako, A0072, 1:6,000), MHC class I (Abcam, EMR8-5, 1:6,000), and MHC class II (Dako, CR3/43 M0775, 1:750) was performed using an automated staining system (Bond III, Leica Biosystems) according to the manufacturer's protocol following antigen retrieval (Bond, ER2 solution). Hematoxylin counterstain was subsequently applied. Staining on two or more adjacent lymphoma cells was used to determine membrane expression of the above-mentioned antigen presentation proteins. Stained slides were scored separately for each of the markers by an expert hematopathologist (S.J.R.) blinded to the genetic data. β2M, MHC class I and MHC class II expression were specifically assessed in $PAX5^+$ malignant cells and average intensity of staining (0 = no staining, 1 = weak staining, 2 = moderate staining, 3 = strong staining) was reported. The percentage of positively stained malignant cells in each case was also estimated (0% to 100%). Thereafter, a H-score was generated by multiplying percentage of malignant cells with positive staining (0% to 100%) and average intensity of positive staining in cells (1 to 3+).

Multiplex Immunofluorescence Staining

Multiplex Immunofluorescent staining to characterization the tumor microenvironment was performed on a BOND RX autostainer (Leica Biosystems) as previously described[5]. Briefly, 4-µm thick FFPE tissue sections were baked for 3 hours at 60°C before loading into the BOND RX (Leica Biosystems). Slides were deparaffinized using BOND DeWax Solution (Leica Biosystems) then rehydrated. Antigen retrieval was performed in BOND Epitope Retrieval Solution 1 (ER1, Leica Biosystems) at pH 6 for 10 minutes at 98°C. Next, slides were serially stained with primary antibodies. Incubation time per primary antibody was 40 minutes. Subsequently, anti-rabbit Polymeric Horseradish Peroxidase (Poly-HRP, BOND Polymer Refine Detection Kit, Leica Biosystems) was applied as a secondary label with an incubation time of 10 minutes. Signal for antibody complexes was labeled and visualized by their corresponding Opal Fluorophore Reagents (Akoya) by incubating the slides for 5 minutes. The same process was repeated for the following antibodies / fluorescent dyes. Slides were air dried, mounted with Prolong Diamond Anti-fade mounting medium (#P36965, Life Technologies) and stored in a light-proof box at 4℃ prior to imaging. The target antigens, antibody clones, and dilutions for each marker are listed below.

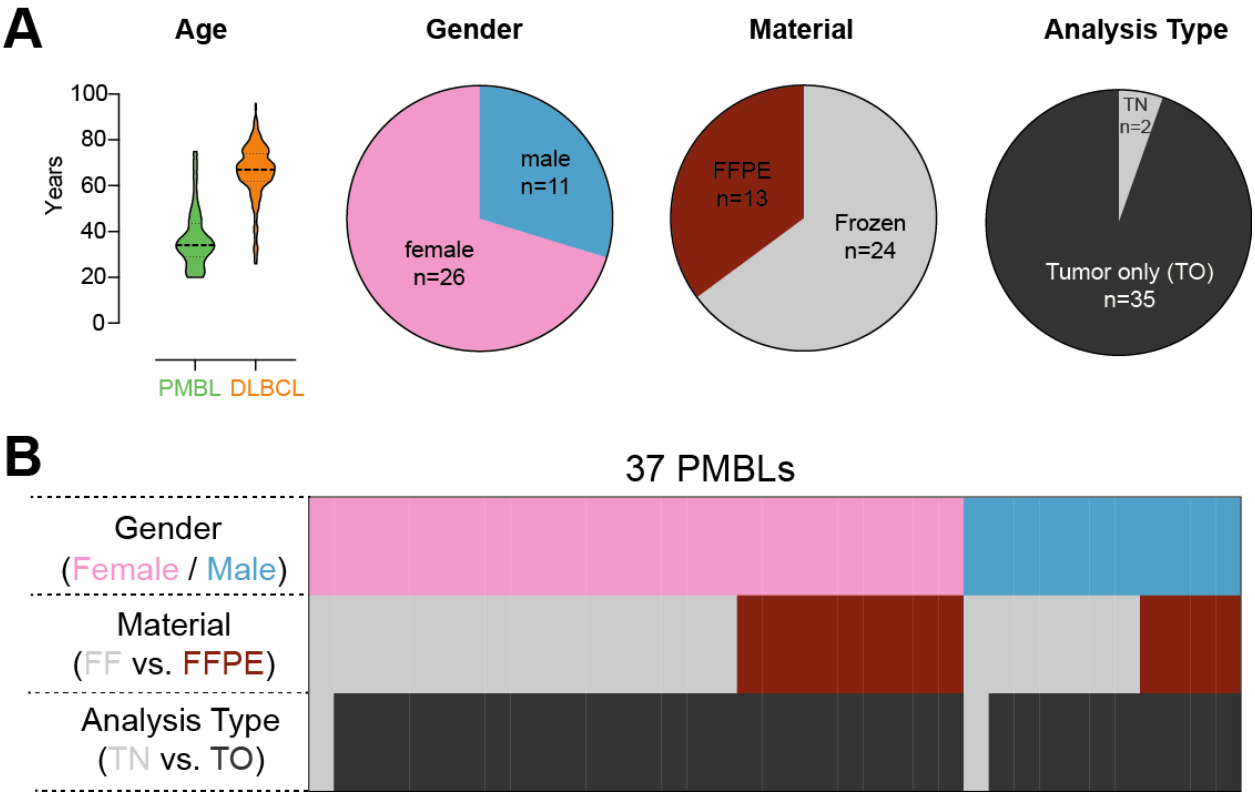
Primary Antibody	Clone ID/Company	Dilution	Opal Kit Fluor	Opal Fluor Dilution
CD4	4B12, DAKO	1:250	Opal 520	1:100
CD3	Polyclonal, DAKO	1:1000	Opal 540	1:100

CD68	PGM1, DAKO	1:2000	Opal 570	1:200
CD8	C8/144B, DAKO	1:5000	Opal 650	1:100
PAX5	24/PAX5, BD Biosciences	1:100	Opal 690	1:50

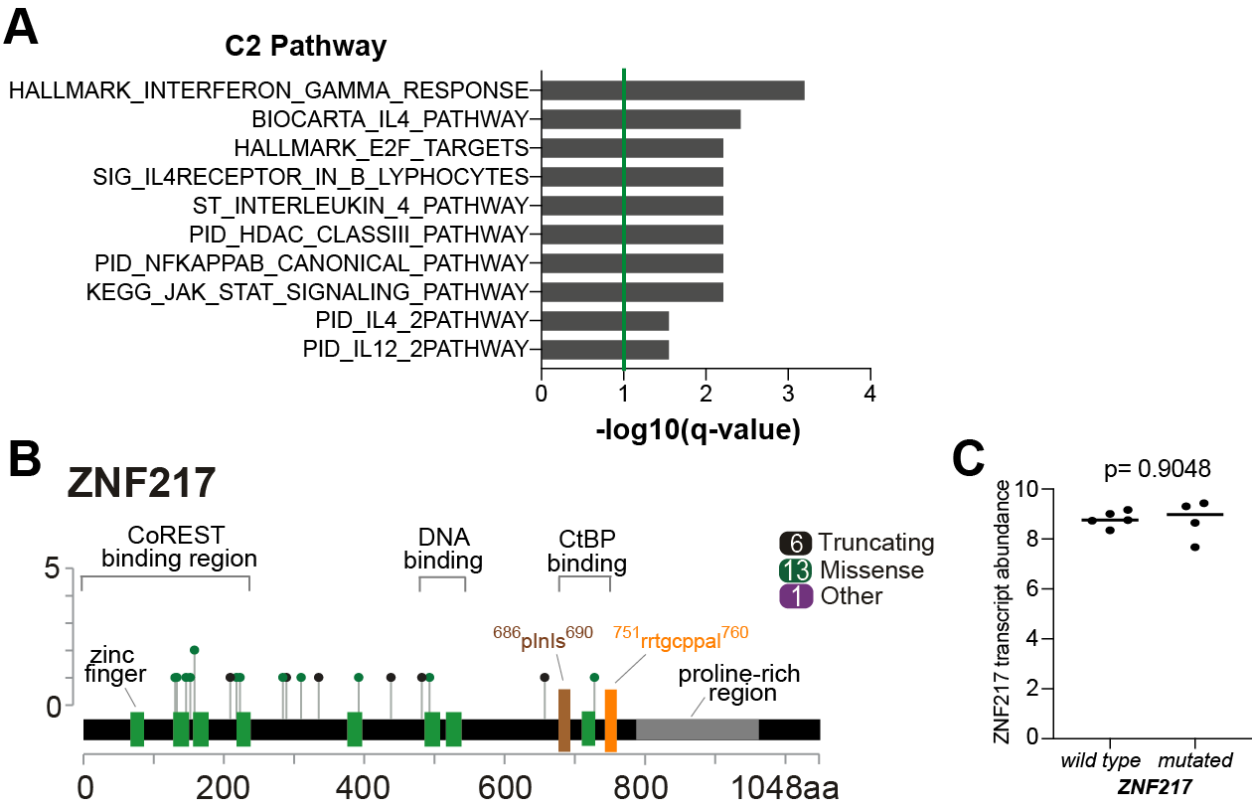
Image Acquisition and Analysis

Image acquisition was performed using the Mantra multispectral imaging platform (Akoya) as previously described[5]. Representative regions of interest were chosen by the pathologist, and 3-5 fields of view (FOVs) per case were acquired at 20x resolution as multispectral images. Images were spectrally unmixed and analyzed using supervised and trained machine learning algorithms within Inform 2.4.2 Image Analysis Software (Akoya). Each cell is assigned an identity based on the combination of immunostained markers that are expressed.

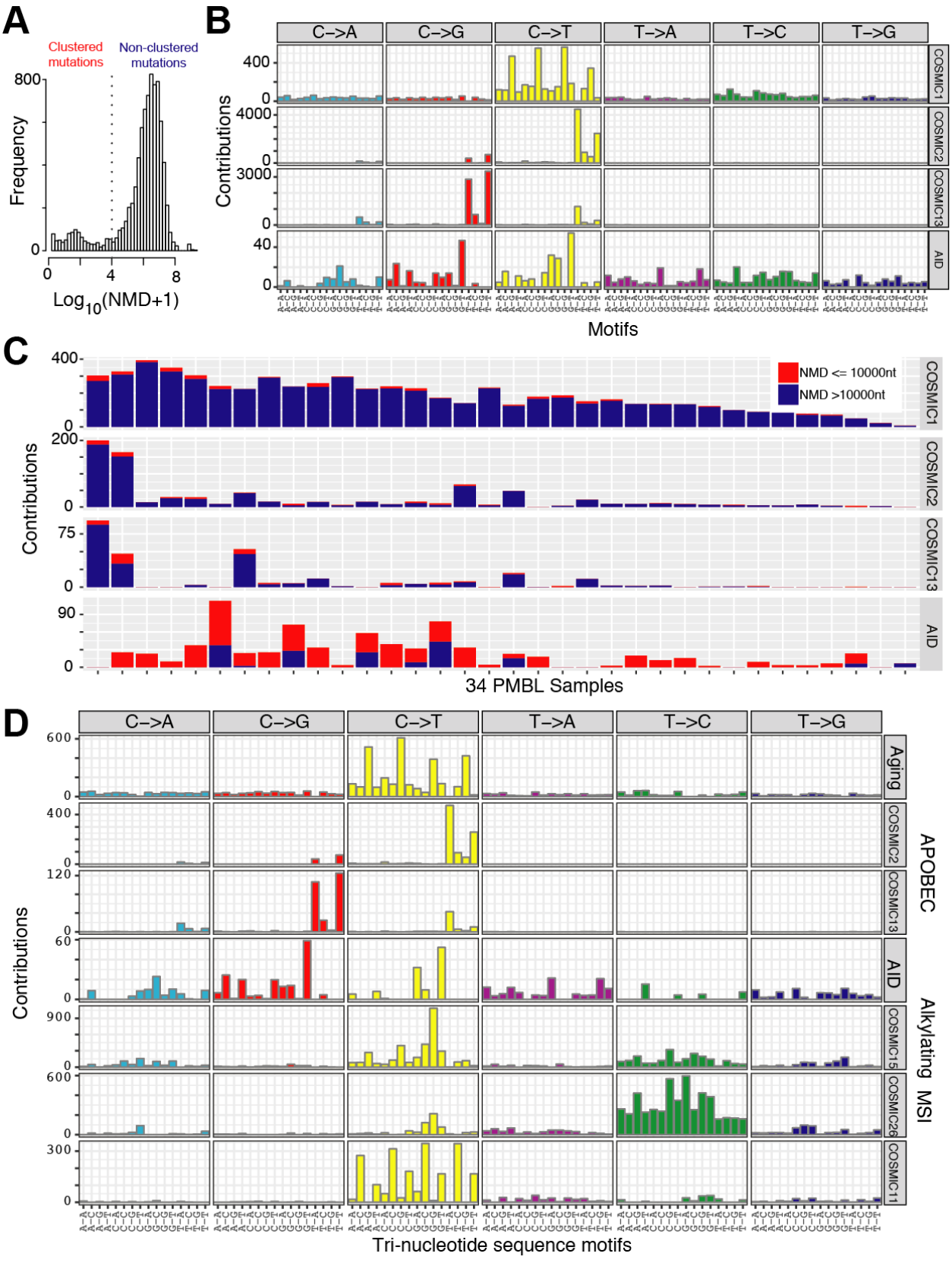
Supplemental Figures

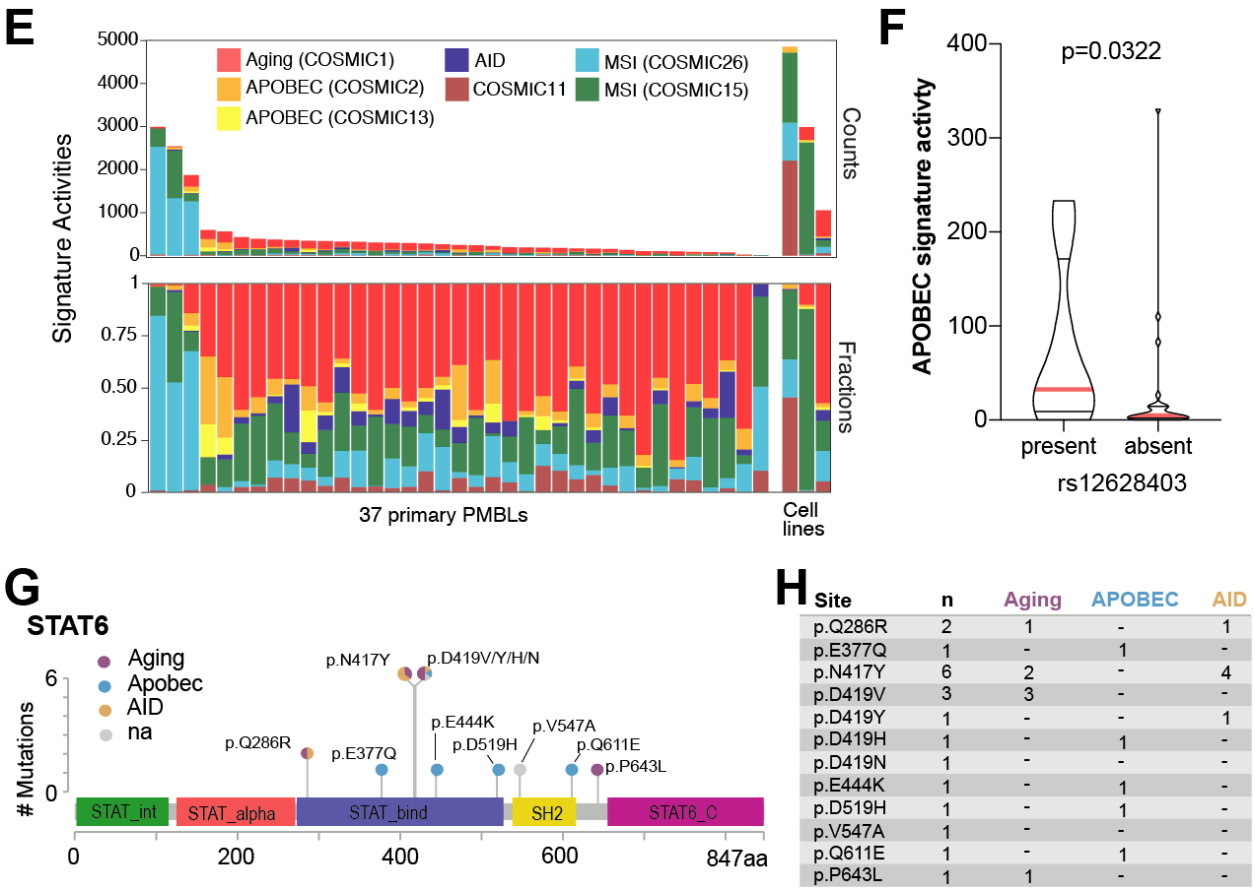


Supplemental Figure 1: Cohort composition. (A) The PMBL cohort has a median age of 34 years, which is significant younger than our recently published DLBCL cohort with a median age of 65 and has a female predominance (left two panels)[6]. Sixty five percent of patients in this series (24/37) had fresh frozen (FF) samples available, the remainder had formalin fixed paraffin embedded (FFPE) tissue as source for their DNA; the majority of tumors (94%) had no patient-matched normal (tumor only, TO; tumor normal pairs, TN) (right two panels). **(B)** Distribution of gender, material type and analysis type by individual patients.



Supplemental Figure 2: Pathway enrichment of significantly mutated genes and ZNF217 expression. (A) Bar graph reflects the FDR-corrected p-value obtained from the hypergeomentric distribution of CCGs in the C2 MSigDB pathways over all genes. **(B)** Mutations in *ZNF217* are visualized in the ZNF217 protein highlighting the functional domains. Putative functions of the protein regions are indicated above. The two putative CtBP-binding domains are indicated in brown and orange (see Ref [7] for details). **(C)** ZNF217 transcript abundance in PMBLs with and without *ZNF217* mutations (Gene expression profiling from [8]). The p-value was obtained using a two-sided Mann-Whitney U-test.

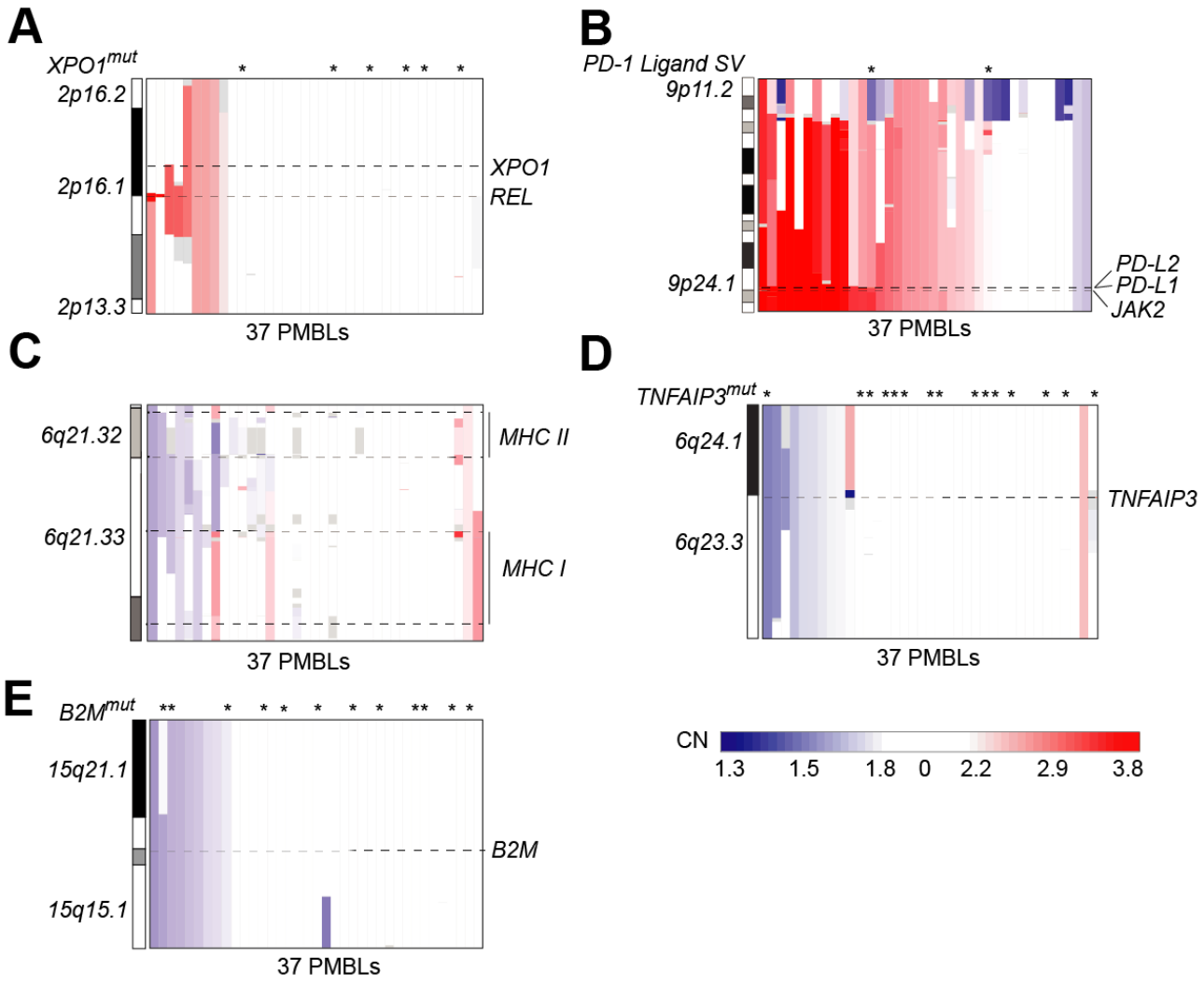

A
Clustered mutations
Non-clustered mutations
Frequency
Log10(NMD+1)
B
C–>A
C–>G
C–>T
T–>A
T–>C
T–>G
COSMIC1
COSMIC2
COSMIC13
AID
Contributions
Motifs
C
NMD <= 10000nt
NMD >10000nt
Contributions
34 PMBL Samples
D
Aging
APOBEC
Alkylating
MSI
COSMIC15
COSMIC26
COSMIC11
Contributions
Tri-nucleotide sequence motifs



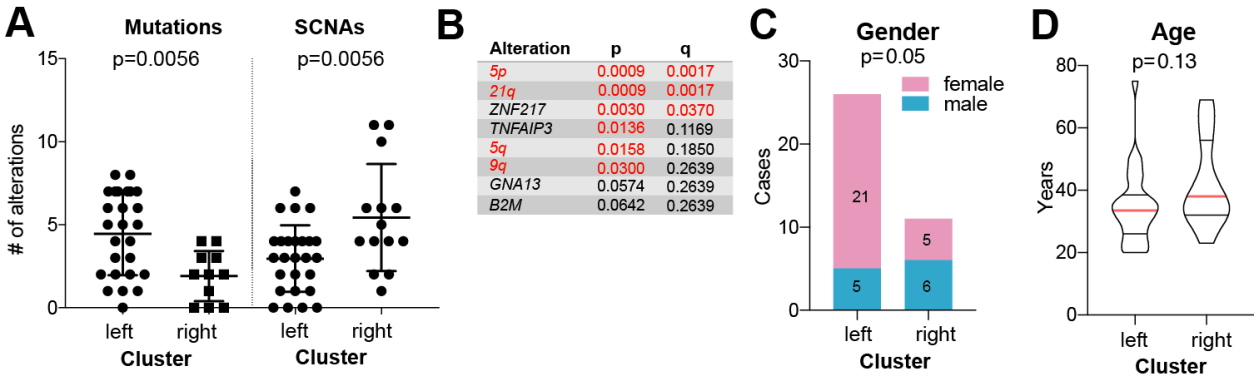
H

Site	n	Aging	APOBEC	AID
p.Q286R	2	1	-	1
p.E377Q	1	-	1	-
p.N417Y	6	2	-	4
p.D419V	3	3	-	-
p.D419Y	1	-	-	1
p.D419H	1	-	1	-
p.D419N	1	-	-	-
p.E444K	1	-	1	-
p.D519H	1	-	1	-
p.V547A	1	-	-	-
p.Q611E	1	-	1	-
p.P643L	1	1	-	-

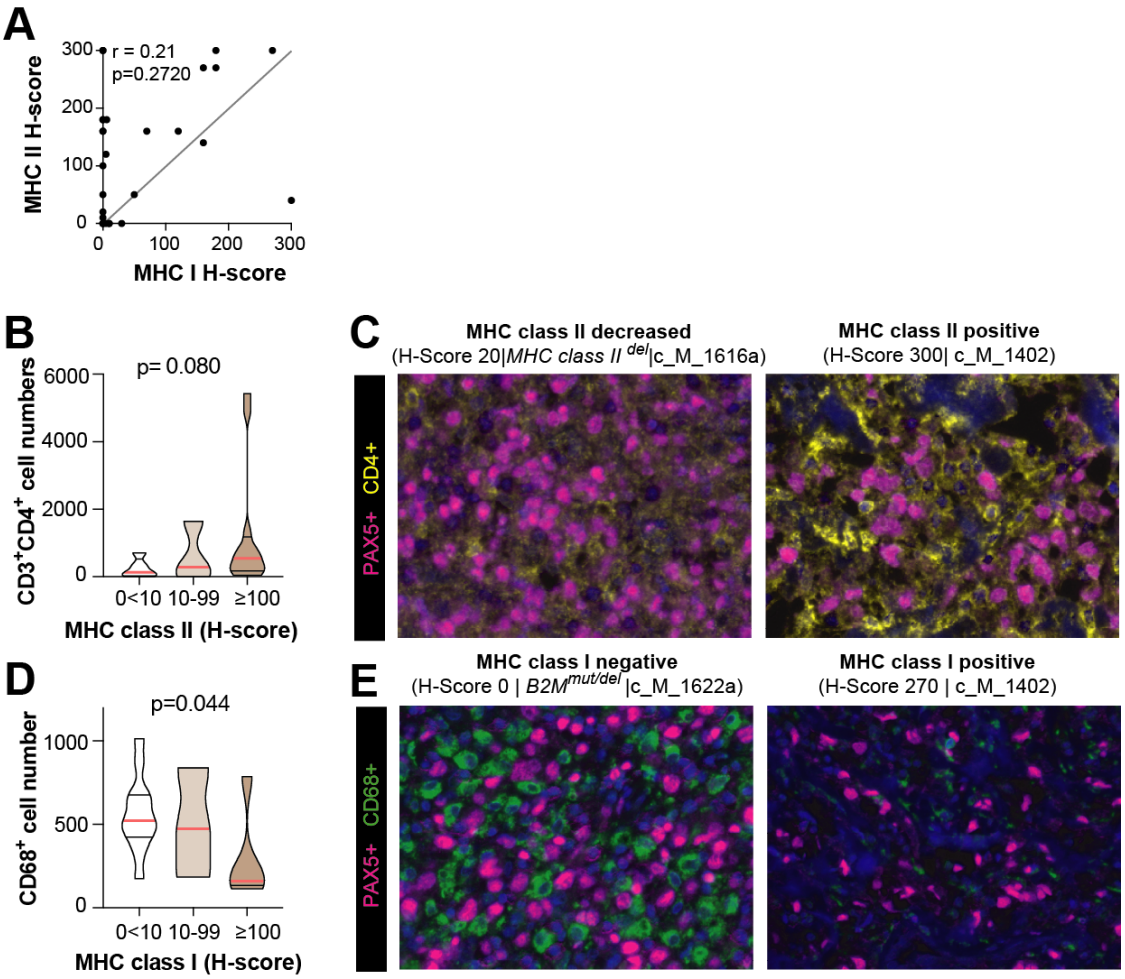
Supplemental Figure 3: Supplemental data for mutational signatures. (A) Nearest mutational distance (NMD) of all mutations reveal a bimodal distribution. Mutations with a NMD ≤ 10 kb are referred to as "clustered" and the one with NMD ≥ 10 kb are referred to as non-clustered, respectively. **(B)** Semi-supervised mutational signature discovery after removing the 3 MSI cases. **(C)** Signature activity of the mutational signatures in the semi-supervised mutational signature discovery in (B). **(D)** De novo mutational signature discovery in 37 PMBLs and 3 PMBL cell lines. **(E)** Distribution of discovered mutational signature activity in the 37 PMBLs and 3 PMBL cell lines. **(F)** APOBEC signature activity by presence or absence of the APOBEC-signature associated SNP rs12628403. The difference in the median APOBEC signature level was tested by a Man-Whitney U rank-sum test. **(G)** Mutation diagram (lollipop figure) for *STAT6* mutations. All non-synonymous mutations are visualized within the functional domains of the respective protein using *MutationMapper* v2.1.0[9,10]. The color key denotes the mutational mechanisms (causative probability of the indicated mechanism >0.75) for each site. **(H)** Table of STAT6 coding changes and putative underlying mutational mechanisms.



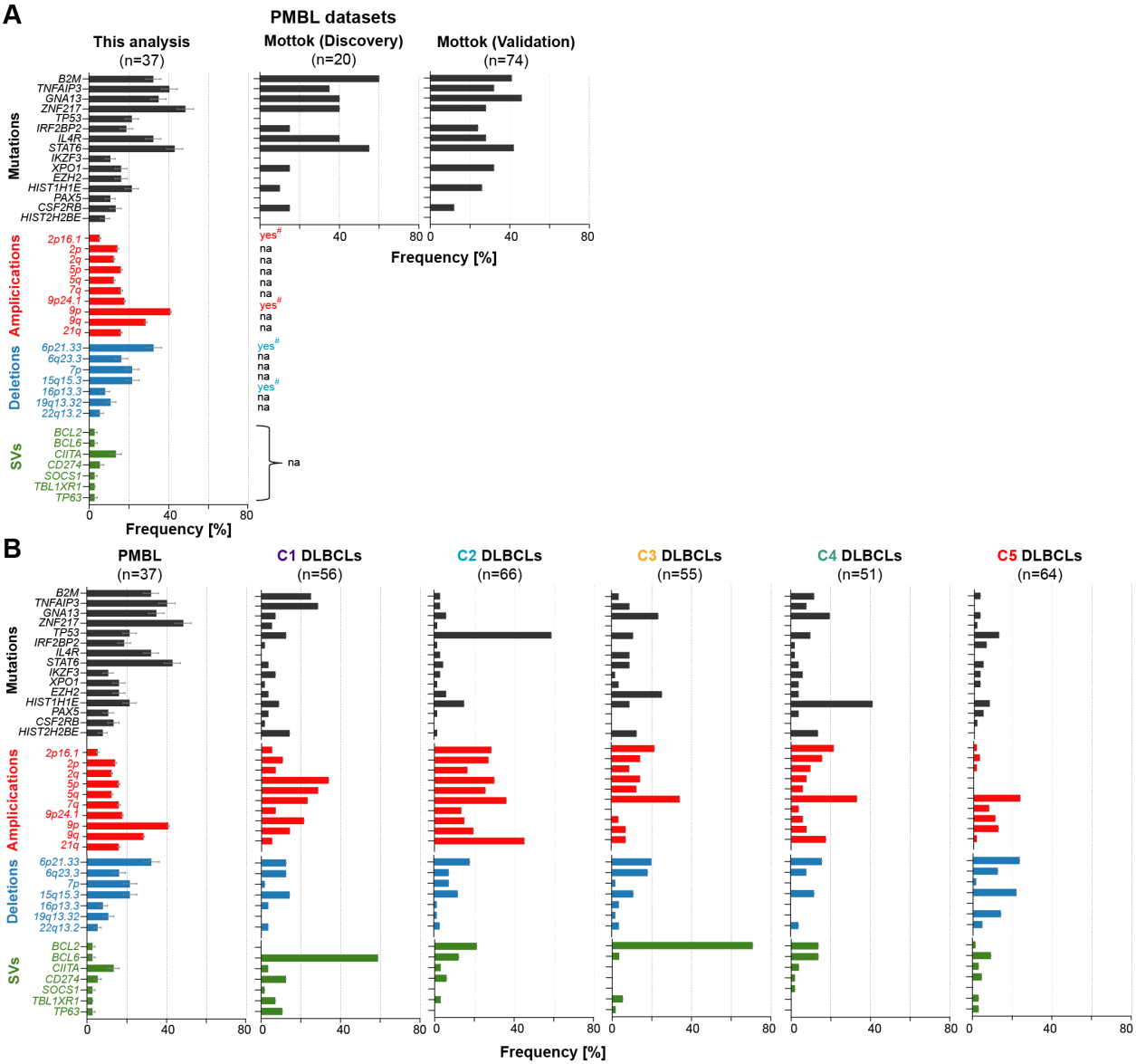
Supplemental Figure 4. IGV screenshots of focal SNCAs. IGV screenshots visualize the copy number (copy number ratio in log2(CNR) units) of the 2p16.2/*REL* **(A)**, 6q21.33/*MHC I/II* **(B),** 6q23.3/*TNFAIP3* **(C)**, 9p24.1/*PD-L1/PD-L2/JAK2* **(D)** and 15q15.1/*B2M* loci **(E)**. Cases with co-occurring mutations or SVs are highlighted at the top of the left panel with an asterisk.



Supplemental Figure 5. Additional features of PMBL genetic substructure. Bi-directional hierarchical clustering of the PMBL gene sample matrix identifies 2 major branches (left and right) (Figure 4A). **(A)** Numbers of CCGs (mutations) and recurrent SCNAs in the left and right branches. The p values are obtained using a Mann Whitney U test. **(B)** Relative enrichment of genetic alterations in the left and right branches. *ZNF217* and *TNFAIP3* mutations were significantly more frequent in the left branch (p=0.003 and p=0.0136, respectively, Fisher's Exact test). **(C)** Genders of patients in the left branch (81% [21/26] female and 19% [5/26] male) and right branch (45% [5/11] female and 55% [6/11] male; p=0.05, Fisher's Exact test). **(D)** Ages of patients in the left branch (median, 33.5 yo; 95% CI 28-37) and right branch (median, 38 yo; 95% CI 31-63; p=0.138, Mann Whitney U test).



Supplemental Figure 6. Features associated with tumor cell expression of MHC class I and MHC class II in PMBL. (A) Lack of correlation between tumor cell expression of MHC class I and MHC class II expression (H-scores) in the evaluated PMBLs (see Figure 6). The p-value was obtained with a Spearman correlation test. **(B)-(E)** Additional analyses of the cellular microenvironment of PMBLs evaluated for β2M, MHC class I and MHC class II protein expression by IHC (see Figure 6). These 28 PMBLs were analyzed for PAX5, CD3, CD4, CD8, CD68 and DAPI expression by multiparametric spectral imaging using established methods[5]. **(B)** $CD3^+CD4^+$ T-cells (y-axis) in PMBL cases with negative (0 - <10), decreased (10 - 99) or positive (≥ 100) MHC class II expression by H scores (x-axis). There was a trend towards higher levels of infiltrating $CD3^+CD4^+$ T-cells in PMBLs with higher levels of MHC class II expression (p=0.08, Cuzick's trend test). **(C)** Representative images of $CD3^+CD4^+$ T-cells (yellow) and $PAX5^+$ tumor cells (purple) in PMBLs with decreased MHC class II expression (left panel) or positive MHC class II expression (right panel). MHC class II H scores for these cases are from Figure 6 and indicated at top. **(D)** $CD68^+$ cells (macrophages) (y-axis) in PMBL cases with negative (0 - <10), decreased (10-99) or positive (≥ 100) MHC class I expression by H scores (x-axis). There were significantly more infiltrating $CD68^+$ cells (macrophages) in PMBLs with negative or decreased MHC class I expression (p=0.044, Cuzick's trend test). **(E)** Representative images of $CD68^+$ macrophages (green) and $PAX5^+$ (purple) primary PMBLs with either negative MHC class I expression (left panel) or positive MHC class I expression (right panel). MHC class I H scores for these cases are from Figure 6 and indicated at top.



Supplemental Figure 7. Comparison of the frequency of recurrent PMBL genetic alterations in this PMBL cohort (left panel and Figure 7) to a recently published PMBL dataset[11] (A) and our recently defined DLBCL genetic clusters, C1-C5[6] (B). Alterations are color coded: mutations, black; copy gains, red; copy losses, blue; SVs, green. Error bars are standard errors. See also legend of Figure 7A. The frequencies for the genetic alterations in the additional PMBL series (Discovery and Validation) are obtained from supplemental Table 2 of the recently published PMBL paper[11]; #, no quantitative information available; NA, not available.

References

1. Alexandrov LB, Jones PH, Wedge DC, et al. Clock-like mutational processes in human somatic cells. *Nat Genet*. 2015;47(12):1402-1407.
2. Kim J, Mouw KW, Polak P, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet*. 2016;48(6):600-606.
3. Kasar S, Kim J, Improgo R, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun*. 2015;6:8866.
4. Tan VY, Fevotte C. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(7):1592-1605.
5. Carey CD, Gusenleitner D, Lipschitz M, et al. Topological analysis reveals a PD-L1-associated microenvironmental niche for Reed-Sternberg cells in Hodgkin lymphoma. *Blood*. 2017;130(22):2420-2430.
6. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med*. 2018;24(5):679-690.
7. Quinlan KG, Verger A, Yaswen P, Crossley M. Amplification of zinc finger gene 217 (ZNF217) and cancer: when good fingers go bad. *Biochim Biophys Acta*. 2007;1775(2):333-340.
8. Chapuy B, Roemer MG, Stewart C, et al. Targetable genetic features of primary testicular and primary central nervous system lymphomas. *Blood*. 2016;127(7):869-881.
9. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401-404.
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. Mottok A, Hung SS, Chavez EA, et al. Integrative genomic analysis identifies key pathogenic mechanisms in primary mediastinal large B-cell lymphoma. *Blood Epub 2019/07/12*. 2019.