Supplemental Methods and Figures

Identification of somatic mutations

Sequence reads from WGS/WES were aligned to the human reference genome with the Burrows-Wheeler Aligner software[1] and single-nucleotide variants (SNVs) were detected using Varscan[2], whereas somatic insertions and deletions (InDels) were identified by GATK (WES)[3] or Platypus (WGS)[4]. On average, 94% of the exonic regions were covered by WES and 98% of whole genomes were covered by WGS with at least 10 sequencing reads (Table S11). Somatic mutations were considered when: 1). at least 10 reads in coverage; 2). at least 10% of mutation allele frequency in tumors; 3). less than 1% of minor allele frequency in the paired controls. All reported SNV and InDels passed visual inspection using Integrative Genomics Viewer (IGV)[5].

Identification of significantly mutated genes

Three prediction methods were performed to define cancer-driver genes in our DLBCL discovery cohort[6-8]. The algorithm from Kan et al. considers the mutation prevalence in the context of the background mutation rate and gene sequence length, as well as evaluation of functional impact[6], the *MutSig* algorithm considers the sample-specific mutation rate, the ratio of nonsynonymous to synonymous mutations in a given gene and the median expression level of each gene in the tumors[7], and the *OncodriveFML* approach provides top-ranking genes based on the functional impact of the mutations[8]. The *Mutsig* algorithm has a limitation in defining the cancer-driver genes in the region of kataegis or clustered mutations, and these features have been observed in the genome of DLBCL[9-11].

Mutation signature analysis

Mutational signatures were extracted as follows: 1). Somatic base substitutions of each data set were classified into 96 possible mutated trinucleotides, 6 types of substitution (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G and T:A>G:C) × 4 types of 5' base (A, C, G, T) × 4 types of 3' base (A, C, G, T), to generate a mutational catalogue. The prevalence of each type of substitution was subsequently calculated for each sample. 2). Signatures of 10 mutational processes from the mutational catalogue were deciphered using the mutational signature framework. The number of signature extracted (N) is determined by estimating the signature reproducibility and reconstruction error rate as described previously[12]. N here is the number where the lowest reconstruction error is achieved without decreasing the reproducibility. 3). The minimal set of mutational signatures was then determined to optimally explain the proportion of each mutation type found in the catalogue, based on reproducibility of their signatures and low error for reconstruction.

Identification of mutations by the targeted-sequencing panel lymphochip

The final list of variants was generated using the following filtering criteria. After sequencing, sequencing reads were discarded if they contained: (1) adaptor reads; (2) low-quality reads, with >10% Ns); (3) low-quality base (>50% bases with quality <10). High quality paired-end reads were then mapped to the UCSC human reference genome (hg19) using BWA-MEM (v0.7.12) with default parameters[1]. Picard (v1.87; http://broadinstitute.github.io/picard/) was used to sort and mark duplicate reads caused by PCR. More than 99.87% of exonic regions of 212 targeted genes were covered with at least 100× in the lymphochip analysis (Table S11). VarScan (v2.3.9) was used to detect substitutions and Indels using the defined parameters[2]. All substitutions and Indels were then annotated using ANNOVAR. To identify somatically occurring, nonsilent mutations in the tumors and to remove potential contamination from the germline polymorphisms and sequencing errors, the identified SNPs and Indels were filtered using the following steps: 1). SNPs annotated as synonymous were removed; 2). SNPs or Indels with a mutation allele frequency (MAF) ≥1% in databases of 1000 genome all, 1000 genome East Asian, or Esp6500, or with a MAF≥0.01% in databases of ExAC all or ExAC East Asian were removed. 3). SNPs or Indels defined as benign in ClinVar database were filtered out; 4). SNPs or Indels detected in tumor samples, which were also detected in YH cell line tested in parallel, or detected in our in-house Chinese DLBCL health control database with MAF ≥1% were filtered out; 5). SNPs or Indels detected in more than 50% samples in our cohort with MAF ≥10% was discarded; 6). All remaining SNPs and Indels with MAF between 10%-90% were kept for further analysis. The performance of the lymphochip was evaluated by including 19 DLBCL samples that had already been characterized by WES or WGS. When considering the nonsilent mutations in the targeted coding region, 93.10% (162/174) of nonsilent mutations identified by WES/WGS can be readily called by the lymphochip using the above filtering strategy and the remaining 12 mutations could also be validated based on manual inspection by IGV, but were filtered out as the MAF was just below the 10% cutoff.

Sanger sequencing

Validation of the *KLF2, TP73* and *ZFP36L1* mutations identified by WGS/WES was performed by Sanger sequencing. In addition, the non-coding exon 1 of *BCL6*, which was not included in the design of the lymphochip, was screened in 179 samples. Primers were designed with Primer3 and primer sequences and PCR conditions for each gene are available upon request. PCR products were purified and sequenced at Macrogen (Amsterdam, Netherlands) or Eurofins MWG Operon (Ebersberg, Germany).

Identification of potential AID off-targets

Potential AID off-targets are proposed if one of the following criteria is fulfilled: 1). have been reported as an AID targeted gene in mouse studies[11,13-16] 2). have been reported as target of aberrant SHM in human studies[9,11] ; 3). showed a significant SHM indicator ($p<0.1$) or a significantly higher mutation density within a 2kb region downstream of transcription start site ($p<0.00001$)[9] ; 4). identified as AID targets in mouse B-cells by target sequencing[17].

PCR for HBV detection in tumor DNA

A Taqman-based real-time-PCR assay was used to detect the sequences encoding HBV polymerase and the primer and probe sequences have been described previously[18].

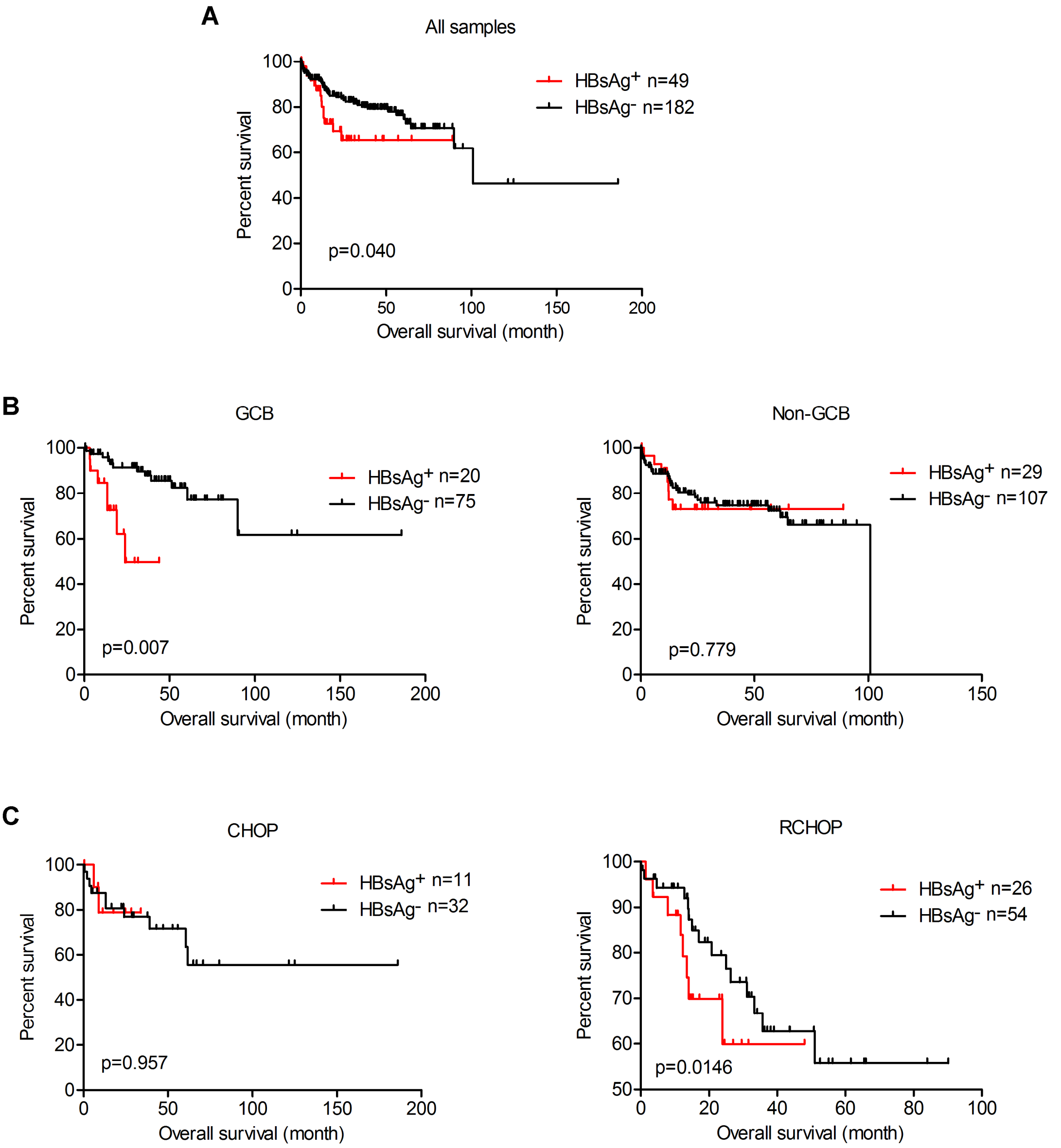
Characterization of V(D)J rearrangements in DLBCL tumors

Rearranged *IGHV-IGHD-IGHJ* genes were amplified, cloned and sequenced from the tumor DNA[19] or cDNA[20] samples as described previously. For the cloning by using DNA as template, primers for framework region 1 or 2 (FR1/FR2) were used and V(D)J rearrangements were amplified from 11 samples, including 6 HBsAg^{+} and 5 HBsAg^{-} DLBCL cases. For the cloning by using cDNA as template, V(D)J rearrangements were amplified from 9 samples, including 8 HBsAg^{+} and 1 HBsAg^{-} DLBCL cases. *IGH* gene usage, somatic mutations in the *IGHV* genes and composition of the CDR3 regions were analyzed by the IMGT/V-QUEST tools[21]. Major clone was identified as: 1). more than 9 clones were sequenced; 1). at least 5 clones were identical. Major clones were identified from 9 samples, including 4 HBsAg^{+} and 5 HBsAg^{-} samples. Two of samples were also tested by the high throughput methods (Table S9).

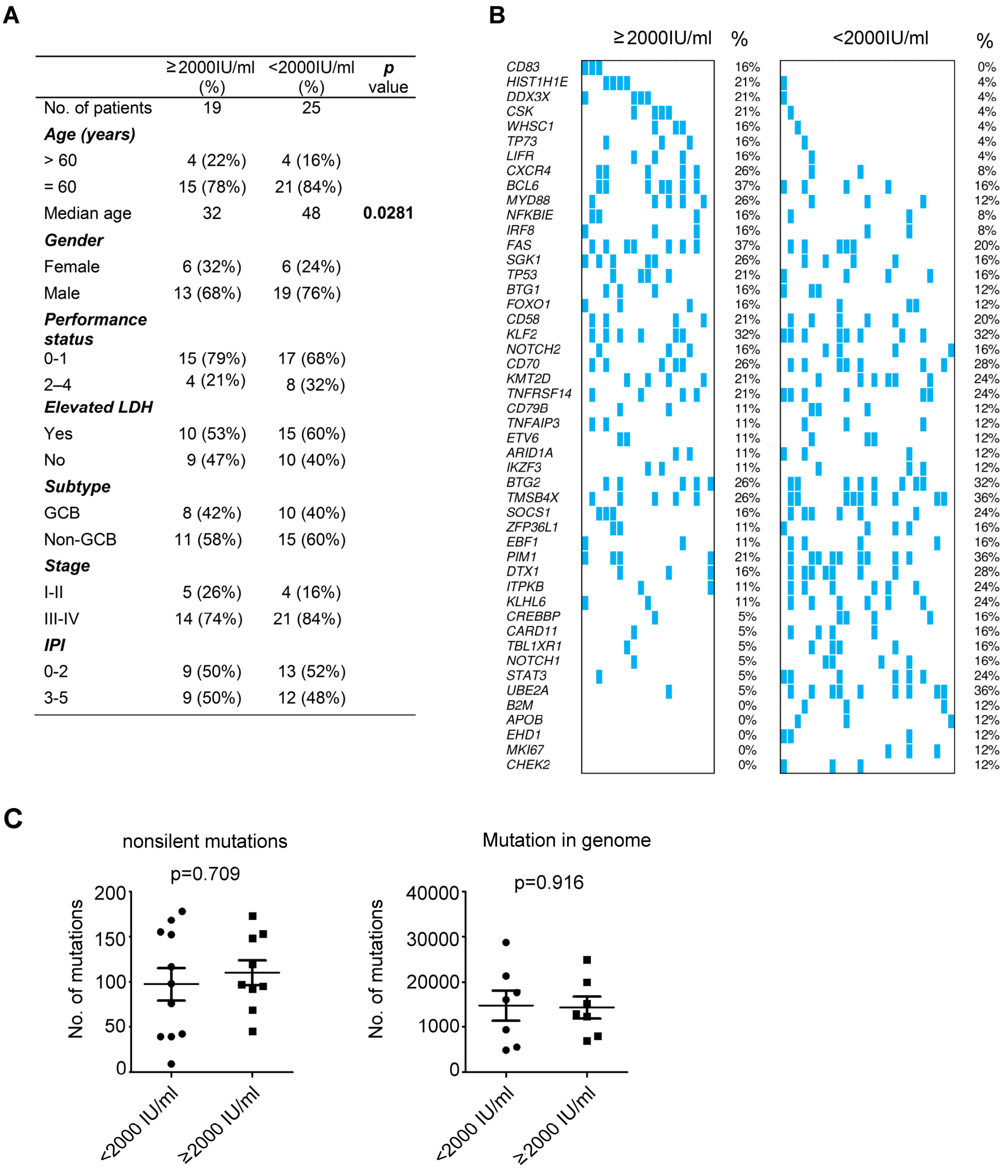
For the high throughput sequencing of V(D)J rearrangements, primers for FR1 were used[19] and VDJ rearrangements were successfully amplified from 36 samples, including 19 HBsAg^{+}

and 17 HBsAg$^-$ DLBCL cases. Major clones were identified from 18 samples, including 11 HBsAg$^+$ and 7 HBsAg$^-$ samples (Table S9). Initial PCR was performed according to the BIOMED-2 Concerted Action protocols[19]. The PCR product was purified using Agencourt AMPure XP PCR clean up (Beckman Coulter) with size selection in the range of 200-500 bp. Indexing PCR was performed on 20μL purified PCR product using primers from illumina TruSeq DNA HT dual-index kit according to the manufacturer's protocol. After clean up, DNA concentrations were measured in each sample using Qubit dsDNA high sensitivity kit (Thermo Fisher Scientific) and then samples were pooled equimolar to the final concentration of 4nM. Sequencing was performed using MiSeq paired-end 600 cycles reagent kit v3 (Illumina). Sequencing data was analysed using MIXCR software[22].

Supplemental Figures



Supplemental Figure 1. Overall survival in HBsAg^{+} DLBCLs compared to HBsAg^{-} DLBCLs. (**A**) the overall survival analysis of all DLBCL patients in our cohort. (**B**) The overall survival analysis of GCB or non-GCB DLBCLs. (**C**) the overall survival analysis of CHOP- or RCHOP-treated DLBCL patients. Kaplan-Meier method was used and the differences between two groups were compared by log-rank test.



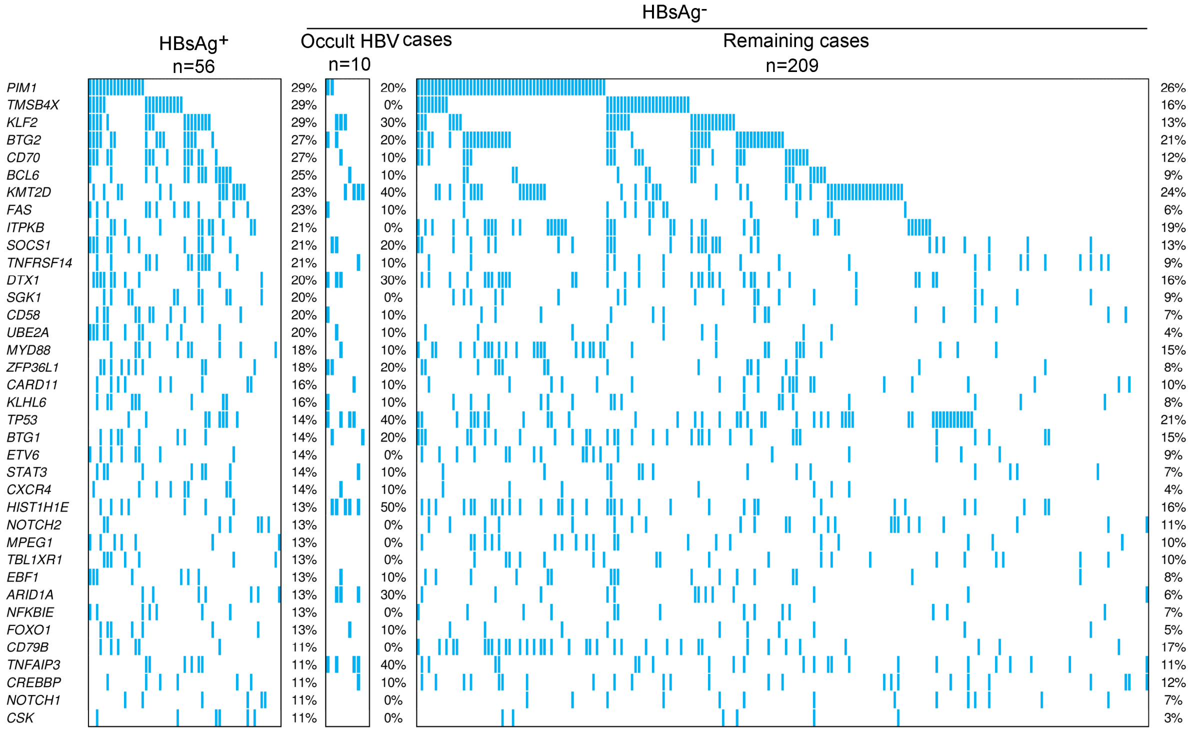
A

	≥2000IU/ml (%)	<2000IU/ml (%)	*p* value
No. of patients	19	25	
Age (years)			
> 60	4 (22%)	4 (16%)	
= 60	15 (78%)	21 (84%)	
Median age	32	48	**0.0281**
Gender			
Female	6 (32%)	6 (24%)	
Male	13 (68%)	19 (76%)	
Performance status			
0-1	15 (79%)	17 (68%)	
2–4	4 (21%)	8 (32%)	
Elevated LDH			
Yes	10 (53%)	15 (60%)	
No	9 (47%)	10 (40%)	
Subtype			
GCB	8 (42%)	10 (40%)	
Non-GCB	11 (58%)	15 (60%)	
Stage			
I-II	5 (26%)	4 (16%)	
III-IV	14 (74%)	21 (84%)	
IPI			
0-2	9 (50%)	13 (52%)	
3-5	9 (50%)	12 (48%)	

B

	≥2000IU/ml %	<2000IU/ml %
CD83	16%	0%
HIST1H1E	21%	4%
DDX3X	21%	4%
CSK	21%	4%
WHSC1	16%	4%
TP73	16%	4%
LIFR	16%	4%
CXCR4	26%	8%
BCL6	37%	16%
MYD88	26%	12%
NFKBIE	16%	8%
IRF8	16%	8%
FAS	37%	20%
SGK1	26%	16%
TP53	21%	16%
BTG1	16%	12%
FOXO1	16%	12%
CD58	21%	20%
KLF2	32%	32%
NOTCH2	16%	16%
CD70	26%	28%
KMT2D	21%	24%
TNFRSF14	21%	24%
CD79B	11%	12%
TNFAIP3	11%	12%
ETV6	11%	12%
ARID1A	11%	12%
IKZF3	11%	12%
BTG2	26%	32%
TMSB4X	26%	36%
SOCS1	16%	24%
ZFP36L1	11%	16%
EBF1	11%	16%
PIM1	21%	36%
DTX1	16%	28%
ITPKB	11%	24%
KLHL6	11%	24%
CREBBP	5%	16%
CARD11	5%	16%
TBL1XR1	5%	16%
NOTCH1	5%	16%
STAT3	5%	24%
UBE2A	5%	36%
B2M	0%	12%
APOB	0%	12%
EHD1	0%	12%
MKI67	0%	12%
CHEK2	0%	12%

Supplemental Figure 2. Comparison of clinical and molecular features in DLBCLs with high or low HBV viral load. Patients were divided into two groups based on the HBV DNA quantitation in the serum: high, ⩾ 2000IU/ml (n=19); low, <2000IU/ml (n=25). (**A**) Basic clinical data was compared. (**B**) Mutation frequency of genes mutated in at least 3 samples of either group was compared. Blue, nonsilent mutation. (**C**) Comparison of the mutation load in the coding region and the whole genome.



Supplemental Figure 3. Mutation pattern of DLBCLs with potentially occult HBV infection. The definition of occult HBV cases was based on: 1). HBsAg$^-$ but HBV DNA positive in tumors. Totally five samples were included. 2). HBsAg$^-$, HBcAb$^+$ and HBsAb$^-$. Totally five samples were included. Blue, nonsilent mutation.

Reference

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
2. Koboldt DC, Chen K, Wylie T, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009;25(17):2283-2285.
3. Li R, Li Y, Fang X, et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res*. 2009;19(6):1124-1132.
4. Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*. 2014;46(8):912-918.
5. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.
6. Kan Z, Jaiswal BS, Stinson J, et al. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*. 2010;466(7308):869-873.
7. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214-218.
8. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol*. 2016;17(1):128.
9. Khodabakhshi AH, Morin RD, Fejes AP, et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget*. 2012;3(11):1308-1319.
10. de Miranda NF, Georgiou K, Chen L, et al. Exome sequencing reveals novel mutation targets in diffuse large B-cell lymphomas derived from Chinese patients. *Blood*. 2014;124(16):2544-2553.

11. Qian J, Wang Q, Dose M, et al. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell*. 2014;159(7):1524-1537.

12. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-421.

13. Liu M, Duke JL, Richter DJ, et al. Two levels of protection for the B cell genome during somatic hypermutation. *Nature*. 2008;451(7180):841-845.

14. Chiarle R, Zhang Y, Frock RL, et al. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell*. 2011;147(1):107-119.

15. Hakim O, Resch W, Yamane A, et al. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature*. 2012;484(7392):69-74.

16. Meng FL, Du Z, Federation A, et al. Convergent Transcription at Intragenic Super-Enhancers Targets AID-Initiated Genomic Instability. *Cell*. 2014;159(7):1538-1548.

17. Alvarez-Prado AF, Perez-Duran P, Perez-Garcia A, et al. A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *J Exp Med*. 2018, *In press*.

18. Welzel TM, Miley WJ, Parks TL, Goedert JJ, Whitby D, Ortiz-Conde BA. Real-time PCR assay for detection and quantification of hepatitis B virus genotypes A to G. *J Clin Microbiol*. 2006;44(9):3325-3333.

19. van Dongen JJ, Langerak AW, Bruggemann M, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*. 2003;17(12):2257-2317.

20. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood*. 2010;116(7):1070-1078.

21. Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*. 2008;36:W503-508.

22. Bolotin DA, Poslavsky S, Mitrophanov I, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods*. 2015;12(5):380-381.