Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Schmitz R, Wright GW, Huang DW, et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. N Engl J Med 2018;378:1396-407. DOI: 10.1056/NEJMoa1801445

Appendix

Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma

List of investigators:

Roland Schmitz[1], Ph.D., George W. Wright[2], Ph.D., Da Wei Huang[1], M.D., Calvin A. Johnson[3], Ph.D., James D. Phelan[1], Ph.D., James Q. Wang[1], Ph.D., Sandrine Roulland[1,*], Ph.D., Monica Kasbekar[1], Ph.D., Ryan M. Young[1], Ph.D., Arthur L. Shaffer[1], Ph.D., Daniel J. Hodson[1,†], M.D., Ph.D., Wenming Xiao[1,§], Ph.D., Xin Yu[1], M.Sc., Yandan Yang[1], Ph.D., Hong Zhao[1], M.Sc., Weihong Xu[1], M.Sc., Xuelu Liu[3,‡], M.Sc., Bin Zhou[3], M.Sc., Wei Du[3], Ph.D., Wing C. Chan[4], M.D., Elaine S. Jaffe[5], M.D., Randy D. Gascoyne[6], M.D., Joseph M. Connors[6], M.D., Elias Campo[7], M.D., Armando Lopez-Guillermo[7], M.D., Andreas Rosenwald[8], M.D., German Ott[9], M.D., Jan Delabie[10], M.D., Ph.D., Lisa M. Rimsza[11], M.D., Kevin Tay Kuang Wei[12], M.D., Andrew D. Zelenetz[13,16], M.D., Ph.D., John P. Leonard[14,16], M.D., Nancy L. Bartlett[15,16], M.D., Bao Tran[17], M.Sc., Jyoti Shetty[17], M.Sc., Yongmei Zhao[17], M.Sc. Dan R. Soppet[17], Ph.D. Stefania Pittaluga[5], M.D., Wyndham H. Wilson[1], M.D., Ph.D., and Louis M. Staudt[1] , M.D., Ph.D.

[1]Lymphoid Malignancies Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, 20892, USA

[2]Biometric Research Program, Division of Cancer Diagnosis and Treatment, National Cancer Institute, National Institutes of Health, Bethesda, MD, 20892, USA

[3]Office of Intramural Research, Center for Information Technology, National Institutes of Health, Bethesda, MD, 20892, USA

[4]Departments of Pathology, City of Hope National Medical Center, Duarte, CA, 91010, USA

[5]Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, 20892, USA

[6]British Columbia Cancer Agency, Vancouver, British Columbia, Canada V5Z 4E6

[7]Hospital Clinic of Barcelona, University of Barcelona, Institute for Biomedical Research August Pi I Sunyer, 08036 Barcelona, Spain

[8]Institute of Pathology, University of Würzburg, and Comprehensive Cancer Center Mainfranken, 97080 Würzburg, Germany

[9]Department of Clinical Pathology, Robert-Bosch-Krankenhaus, and Dr. Margarete Fischer-Bosch Institute for Clinical Pharmacology, 70376 Stuttgart, Germany

[10]University Health Network, Laboratory Medicine Program, Toronto General Hospital and University of Toronto, Toronto, ON, M5G 2C4, Canada

[11]Department of Laboratory Medicine and Pathology, Mayo Clinic, Scottsdale, AZ, 85259, USA

[12]National Cancer Centre of Singapore, Singapore 169610

[13]Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA

[14]Weill Cornell Medicine, New York, NY, 10021, USA

[15]Department of Medicine, Washington University School of Medicine, St. Louis, MO, 63110, USA

[16]Alliance for Clinical Trials in Oncology, Chicago, IL, 60606, USA

[17]Cancer Research Technology Program, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, MD, 21702, USA

[*]Present address: Aix-Marseille Université, CNRS, INSERM, CIML, Centre d'Immunologie de Marseille-Luminy, Marseille, France

[†]Present address: Wellcome Trust-MRC Stem Cell Institute, Cambridge Biomedical Campus, Cambridge, UK

[§]Present address: Division of Bioinformatics and Biostatistics, National Center for Toxicological Research U.S Food and Drug Administration, Jefferson, AR, 72079, USA

[‡]Present address: OmniSeq, Inc., Buffalo, NY 14203, USA

Table of Contents

Supplemental Methods 4
Study design 4
Primary data processing 4
Exon-Seq and HaloPlex sequencing 4
Selection of candidate somatic mutations 5
RNA-Seq 6
Affymetrix SNP6.0 DNA copy number arrays 6
Analysis of RNA-seq data 6
Digital gene expression 6
Identification of cell of origin subtypes 7
Gene expression signature analysis 7
Gene-gene fusion analysis 7
Analysis of DNA copy number alterations 8
Creation and testing a predictive model for somatic mutations in DLBCL 10
Random forest model for detecting somatic mutations in tumor without matched normal 10
Preprocessing 11
Machine learning methodology 11
Cross-validation and holdout testing on DLBCL data 12
Somatic prediction on variants in the present study 12
Prediction of AID-dependent somatic hypermutation target genes 13
Multi-platform analysis 15
Genomic feature definition 15
The GenClass iterative genetic subtype algorithm 16
Application of the GenClass method to classify a single sample 18
Permutation testing 18
Random forest model for genetic subtype prediction 19
Additional statistical analysis 19

Supplemental Figures 20

Figure S1: Genetic distinctions between gene expression subgroups 20
Figure S2: The GenClass iterative genetic classifier used to create the DLBCL genetic subtypes 21
Figure S3: Gene expression signatures that distinguish the DLBCL genetic subtypes 22
Figure S4: Mutations shared by primary central nervous system lymphoma (PCNSL)[1-4] and DLBCL genetic subtypes 23
Figure S5: Oncogenic pathways in DLBCL subtypes 24
Figure S6: Comparison of Affymetrix U133+ based predictor score to RNAseq based DGE predictor score 25
Figure S7: MvsN plot of CGH segmentation for a typical sample 26
Figure S8: Relationship between signal values of single copy gains and single copy losses 27
Figure S9: ROC and PPV-sensitivity curves of the RF model in cross-validation tests 28

Figure S10: ROC and PPV-sensitivity curves of the RF model on holdout samples 29
Figure S11: Dimensionality-reduction plots of published and predicted somatic hypermutation genes 30
Figure S12: Overlap between predicted membership of the four genomic DLBCL subtypes 30

Supplemental Tables 31
Table S1: Mutation frequency of genes mutated in DLBCL subtypes (Excel spreadsheet) 31
Table S2: Mutation frequency of genes including subclonal mutations in DLBCL subtypes (Excel spreadsheet) 31
Table S3: Frequency of chromosomal amplifications of genes in DLBCL subtypes (Excel spreadsheet) 31
Table S4: Frequency of chromosomal gains of genes in DLBCL subtypes (Excel spreadsheet) 31
Table S5: Frequency of heterozygous losses of genes in DLBCL subtypes (Excel spreadsheet) 31
Table S6: Frequency of homozygous losses of genes in DLBCL subtypes (Excel spreadsheet) 31
Table S7: Frequency of individual mutations in DLBCL subtypes (Excel spreadsheet) 31
Table S8: Frequency of individual subclonal mutations in DLBCL subtypes (Excel spreadsheet) 31
Table S9: Characteristics of DLBCL patients (Excel spreadsheet) 32
Table S10: Statistical analysis addition DLBCL subtype distinction to International Prognostic Index (IPI) model (Excel spreadsheet) 32
Table S11: HaloPlex Design (Excel spreadsheet) 32
Table S12: Prediction values aberrant somatic hypermutation (Excel spreadsheet) 32
Table S13: Genetic features included in subtype predictors (Excel spreadsheet) 32
Table S14: Summary characteristics of the samples used for predictive model for somatic mutations in DLBCL 33
Table S15: List of annotation features used to create a Random Forest model of somatic mutations in DLBCL 33
Table S16: Performance of the leave-one out cross validation samples used in training the RF model 33
Table S17: Holdout testing performance of the RF model on 23 TCGA DLBCL samples not used in training 34
Table S18: Application of the RF model to predict somatic variants on the full set of filtered DLBCL variants used in the current study 34
Table S19: Leave-one-out cross validation results of models for predicting somatic hypermutations 35
Table S20: Known and predicted hypermutation genes in DLBCL 35

References 36

Supplemental Methods

Study design

The study was conducted using 574 DLBCL biopsy samples (fresh frozen tissue sections). 556 cases were analyzed using DNA sequencing data from whole exome capture (WES; n=556, average depth: 93X) or deep amplicon resequencing (HaloPlex; n=530, average depth: 896X) of 372 genes that are recurrently mutated in DLBCL based on both published data and our preliminary sequencing analysis (Fig. S1A). The majority of cases were analyzed by transcriptome sequencing (RNA-seq; n=562, average mapped reads per case: 45,532,882). For 11 additional DLBCL we used available U133plus2.0 gene data[5] to determine gene expression levels (see below). Genome-wide DNA copy number analysis was performed on 560 DLBCL (CGH; n=560, Affymetrix SNP6.0 arrays). Biopsies were obtained from patients at institutions in the Lymphoma/Leukemia Molecular Profiling Project (LLMPP) consortium, at the National Cancer Centre of Singapore, or from patients enrolled on the CALGB 50303 clinical trial[6] under IRB approved protocols. We utilized genomic profiling data 40 DLBCL biopsies generated by The Cancer Genome Atlas (TCGA)[7] initiative downloaded from the NCI Genomic Data Commons (GDC). For the majority of cases, we studied pre-treatment biopsies from cases of *de novo* DLBCL (n=554, 96.5%), with the remainder (n=20, 3.5%) from relapsed or refractory DLBCL tumors. Gene expression-based cell-of-origin classification of the tumors into ABC, GCB or Unclassified (Unclass) subgroups was achieved by generating concordant Bayesian predictors based on gene expression values from RNA-seq (n=562), Affymetrix U133plus2.0 arrays (n=11) or the Nanostring platform[8] (n=1). We deliberately enriched our sample set for ABC and Unclassified tumors to test the hypothesis that genetic heterogeneity in these cases is responsible for variable clinical responses to conventional and targeted therapy. The final data set consisted of 295 ABC cases (51.4%), 164 GCB cases (28.6%), and 115 Unclassified cases (20.0%). The clinical characteristics of these cases are presented in Table S9. Survival data was available on 240 cases treated with chemoimmunotherapy (Rituximab plus CHOP or CHOP-like chemotherapy). Mutation and copy number aberrations are shown in Tables S1-8. Primary sequencing data and copy number analysis from these cases will be made available through the NIH dbGAP system (accession numbers phs001444, phs001184 and phs000178) and the NCI Genomic Data Commons. Computer programs used will be made available from the investigators upon request.

The study was designed by L.M.S. and R.S., data were gathered by L.M.S. and R.S., data were analyzed by G.W., D.W.H., C.J., R.S. and L.M.S., L.M.S. vouches for the data and the analysis, L.M.S wrote the first draft of the paper with input from R.S., G.W., D.H.W., and C.J. and all authors reviewed the manuscript, and L.M.S. decided to publish the paper with review by all authors. There are no relevant legal agreements between the authors or their institutions.

Primary data processing

Exon-Seq and HaloPlex sequencing

DNA was extracted using the AllPrep kit following the manufactures instructions (QIAGEN). Sequencing libraries for Exome-sequencing were prepared using the Agilent SureSelectXT Human All Exon V5 target enrichment kit (Agilent). Paired-end 100 bp or 150 bp read sequencing was performed on a HiSeq 2500 or HiSeq3000 system using Illumina TruSeq V3 chemistry. Sequencing libraries for targeted amplicon

resequencing were prepared using the HaloPlex Target Enrichment kit – custom design (Agilent). Targeted genomic regions included the 5' untranslated region (UTR), coding region, and 3'UTR of 372 genes. Additionally, 230 chromosomal regions were included. Details are listed in Table S11. Paired-end 125 bp sequencing was performed on a HiSeq 2500 system using Illumina TruSeq V4 chemistry.

Paired-end reads were mapped to the human genome (NCBI build 37) using BWA-MEM 0.7.12[9] with default parameters. The alignment was further refined by the functions of local realignment, base quality recalibration and indel realignment provided by GATK software 3.3 -0[10]. The variants were called on the final alignment file (BAM) by VarSCan2 software[11] using the following criteria: variant read count >= 3 and variant read frequency >= 0.1 for Exon-Seq data; variant read count >= 15 and variant read frequency >=0.02 for HaloPlex data. We define HaloPlex variants with read frequency between 0.02 to 0.1 as subclonal variants.

Overall, we detected 54,168 protein-altering mutations (missense, inframe insertion/deletion, nonsense, frameshift, or splice donor/acceptor mutations), of which 45,497 (79.2%) were observed on more than one sequencing platform. In addition, for some analyses, we additionally considered subclonal mutations (mutant allele frequency < 10%) that were detected by the HaloPlex platform (n=1,952). The gene expression subgroups did not differ substantially in the prevalence of mutations, with an average of 99.6 protein-altering mutations per case.

Selection of candidate somatic mutations

To identify candidate somatic mutations from a large variant pool, we deployed a cascade of heuristic filtering steps, divided into inclusion, exclusion and rescue steps. Inclusion criteria were: 1) any variant in 372 genes studied using HaloPlex deep amplicon sequencing; 2) missense variants supported by Exome-Seq and RNA-Seq data; 3) truncating variants identified by Exome-Seq (nonsense, frameshift, splice donor or acceptor mutations). Exclusion criteria were: 1) variants identified by Exome-seq or RNA-seq analysis of in-house control DNA samples from normal B-cell subpopulations; 2) variants collected in dbSNP (version 138); 3) variants with population frequency >0.0001 in the ExAC database (release 2015); 4) variants present in an in-house curated blacklist that was generated by inspection of variants meeting the above criteria. The blacklisted variants were presumed to be artifacts generated either by the high throughput sequencing platform itself or due to errors in alignment or annotation of the sequencing reads by the analytical pipeline. These variants were identified by manual inspection and typically were those that were unusually prevalent, identified exclusively by one sequencing platform, and not recurrent among variants curated in-house from 57 next-generation sequencing studies of lymphoma and leukemia. Rescue criteria included: 1) variants detected recurrently among published somatic mutations from previous sequencing studies of lymphoma and leukemia; 2) variants at known mutational hotspots in cancer.[12,13] Evaluation of the sensitivity and specificity of these heuristic procedures is described below in the section "Creation and testing a predictive model for somatic mutations in DLBCL".

For certain lymphoma oncogenes, our analysis only included variants that target known functional regions of the proteins. For CD79A and CD79B, we only included variants that target the intracellular ITAM signaling regions, as such variants have been shown to promote BCR signaling.[14] For NOTCH1 and NOTCH2, we only included variants (mostly truncating) that target amino acids that are C-terminal to the intracellular domain (ICN), since such mutations stabilize NOTCH proteins by inactivating PEST

domains.[15] For NOTCH1, we additionally included 2 variants in the 3'UTR regions that have been shown to cause alternative splicing and disruption of the PEST domain.[16]

RNA-Seq

RNA was extracted using the AllPrep kit (QIAGEN). Sequencing libraries for RNA-sequencing were prepared using the TruSeq RNA Library Prep Kit V2 (Illumina). Paired-end 100 bp read sequencing was performed on a HiSeq 2500 system using Illumina TruSeq V3 chemistry.

Paired-end reads were mapped to the human genome (NCBI build 37) using the gapped aligner STAR 2.4.1[17], using the two-pass method and parameters recommended by NCI Genomic Data Commons (GDC)[18]. The alignment file was used for calculating the raw digital gene expression values by HTseq-count software 0.7.2[19], using the intersection-nonempty model, which were further analyzed to provide digital gene expression values (see below). The alignment file was also used for variant calling by VarScan2 with selection based on variant read count >= 3 and variant read frequency >= 0.1.

Affymetrix SNP6.0 DNA copy number arrays

DNA was extracted using the AllPrep kit (QIAGEN). Copy number was analyzed using Affymetrix SNP 6.0 arrays following manufacturer's instructions (Affymetrix).

The imaging signals in the 569 CEL files derived from Affymetrix SNP 6.0 arrays were analyzed using the Affymetrix Genotyping Console[20] with default parameters to obtain probe-level signal values. These were then analyzed using the DNAcopy Bioconductor tool[21] to generate genomic segments for which the probes had relatively uniform signal, and to calculate the mean signal values for the probes making up each of those segments.

Analysis of RNA-seq data

Digital gene expression

Counts for digital gene expression (DGE) were normalized and transformed according to the following equation:

$$y_{ij} = \max\left(0, \log_2\left(\frac{500 * c_{ij}}{Trim\ mean_{.90}(c_{1j} \ldots c_{Nj})}\right)\right)$$

where y_{ij} is the DGE value of gene i on sample j used for analysis, c_{ij} is the corresponding number of raw counts, and $Trim\ mean_{.90}(c_{1j} \ldots c_{Nj})$ is the average of the middle 90% of counts for genes in sample j.

For 12/574 samples, no RNAseq data was available. However, eleven of these samples did have expression data available from analysis of U133+ oligonucleotide arrays. Additionally, there were 381 samples which had matched RNAseq and U133+ array data available. This allowed us to impute pseudo-DGE values for samples with missing RNAseq data, based on their U133+ measured expression. Given the larger proportion of the data which had digital gene expression values exactly equal to 0, we found that simple linear regression performed poorly. Instead, a system which preserved the relative rank of expression for a given new sample between U133+ and DGE was used. For a given gene, let $u_1 \ldots u_N$ be

the U133+ expression values for the matched samples ordered from least to greatest, and similarly let $d_1 \ldots d_N$ be the DGE values for the matched sample for that gene. If x is the U133+ expression for that gene in a sample without DGE available, we substituted the imputed value y specified as follows:

$$if(x \le u_1) \qquad y = \max\left(0, d_1 - (u_1 - x)\left(\frac{d_{N/4} - d_1}{u_{N/4} - u_1}\right)\right)$$

$$if(u_k < x \le u_{k+1}) \quad y = d_k + (x - u_k)\left(\frac{d_{k+1} - d_k}{u_{k+1} - u_k}\right)$$

$$if(u_N < x) \qquad y = d_k + (x - u_N)\left(\frac{d_N - d_{3N/4}}{u_N - u_{3N/4}}\right)$$

Identification of cell of origin subtypes

We developed a DGE-based predictor of cell-of-origin subtype (COS) by mimicking the oligonucleotide cell of origin subtype previously developed [22] for the 381 samples for which we had matched U133+ and DGE available. The DGE model was a weighted average of the DGE values for 195 genes which were predictive of COS on the U133+ data and were highly correlated between the two data sets. The weights for these genes were given by the equation:

$$\text{weight} = \frac{Z}{\sqrt{1 + Z^2(\rho^{-2} - 1)}}$$

where Z is the average expression difference between ABC and GCB on the U133+ data, divided by the pooled within group standard deviation, and ρ is the Pearson correlation between the U133+ expression and the DGE expression. The weighted DGE averages were then linearly normalized, so that their mean and standard deviation matched that of the U133+ predictor scores. On the set of matched samples, the resulting scores were in very strong agreement with what was reported according to the U133+ gold standard (Fig. S6), and so we felt confident using the predictor score along with the U133+ cut-points previously defined[23] to define cell of origin subtypes for all samples.

Gene expression signature analysis

For each signature in a database of gene expression signatures[24] (https://lymphochip.nih.gov/signaturedb/), the DGE signal values of the signature genes were averaged to provide a signature average value for each sample. For the purpose of comparison between signatures, the values representing each signature were linearly normalized so that their median and interquartile range matched that of a standard normal distribution. In order to include samples for which the DGE was imputed based on U133+2.0 array data (see above), we restricted ourselves to those genes for which both RNAseq-based and array-based imputed digital gene expression were available. Overall, this represented 98% of the signature genes. Significance P-values for the differences in signature averages between DLBCL genetic subtypes were derived from Student t-tests.

Gene fusion analysis

Candidate gene-gene fusions were detected by using the alignment file (BAM) derived from RNA-Seq data as input to the FusionCatcher algorithm.[25] In addition, we developed an in-house script to detect

gene-gene fusions involving the *BCL6* gene. In short, the in-house script searched for anomalous reads that aligned to the *BCL6* locus but had either been soft-clipped by BWA-MEM (indicating a region of non-alignment) or had a paired-end read that did not map to the *BCL6* locus. Soft clipping is a feature of BWA-MEM that allows for the mapping of reads that have sequencing artifacts near their ends. However, this feature also prevents the discovery of reads that represent fusion transcripts between two genomic loci. Soft clipped reads are flagged by BWA-MEM, allowing us to obtain the full-length sequences of these reads along with their paired-end counterparts. BLAT[26] was used to align these sequences against human genome build NCBI build 37. Any pair of reads that mapped both to *BCL6* and to another chromosomal location was declared as evidence of a *BCL6* gene-gene fusion. For analysis, we used *BCL6* gene-gene fusions detected either by FusionCatcher or by our in-house method. All told, we detected 112 *BCL6* fusions. Fusion partners include *IgH* (69.6%), *IgK* (0.89%), *IgL* (6.25%), non-immunoglobulin genes (23.2%).

To develop a *BCL2* translocation genetic feature, we utilized evidence from FusionCatcher data, *BCL2* mutations, and *BCL2* fluorescence in situ hybridization (FISH) breakapart probe studies. *BCL2* FISH data was available from 196 cases, of which 26 (13%) were translocated. *BCL2* FISH-positive (translocated) cases were most common in GCB DLBCL (29.7% of cases), less common in Unclassified (10.8%), and absent in ABC (0%), as expected[27]. FusionCatcher detected *BCL2* fusions involving the IgH locus in 19 cases. Among 6 such cases for which we had *BCL2* FISH data, 5 were FISH-positive and 1 was FISH-negative, suggesting that a *BCL2-IGH* fusion from FusionCatcher can function as a surrogate for a *BCL2* translocation. In addition, we identified *BCL2* mutations in 56 cases, which were more common in GCB DLBCL (21.3%), than in Unclassified (6.1% of cases), or ABC (8.9% of cases). Among the 15 *BCL2* mutant GCB cases for which we had *BCL2* FISH data, all 15 were FISH-positive, demonstrating that *BCL2* mutation in GCB DLBCL is a useful surrogate for *BCL2* translocation. For ABC and Unclassified, we only had one *BCL2* mutant case each that had *BCL2* FISH data, which was insufficient to ascertain whether *BCL2* mutations are associated with *BCL2* translocations in those subgroups. Given the above, we defined a composite *BCL2* translocation genetic feature that included cases that were *BCL2* FISH-positive, had a *BCL2-IGH* fusion from FusionCatcher, or were GCB with a *BCL2* mutation.

Analysis of DNA copy number alterations

Among CGH samples from 569 donors, 9 had segmentation results that were too noisy to be usable and were excluded from analysis. Additionally, thirteen samples appeared to be strongly over-segmented but appeared to still contain usable signal. For these, the following statistic was used to estimate the extent to which a division between two segments could be explained by noise:

$$\text{Difference Statistic} = \frac{(M_1 - M_2)^2}{\frac{1}{n_1} + \frac{1}{n_1}}$$

where M_1 and M_2 are the means of the adjacent segments, and n_1 and n_2 are the number of probes in adjacent segments. Starting from the breakpoint with the smallest Difference Statistic, adjacent segments were merged until no segment had a Difference Statistic less than 30.

We observed that there appeared to be great variability from sample to sample in terms of the association between signal value and copy number. One possible explanation for this are differences between samples in tumor content i.e. the ratio of malignant tumor cells to normal infiltrating cells in the biopsy. If probe *i* on sample *j* is at a location with copy number c_{ij} and if the proportion of the sample that is tumor (as opposed to normal infiltrating cells) is τ_j, then assuming that the sample has a normal copy number of 2 at that location, the signal value for that probe should theoretically be:

$$v_{ij} = log_2\left(\frac{c_{ij}\tau_j + 2(1-\tau_j)}{2}\right) + \delta_j + \varepsilon_{\mathrm{ij}}$$

where δ_j represents a sample normalization adjustment, and $\varepsilon_{\mathrm{ij}}$ represents random noise.

Over a long segment containing many probes with the same copy number, the noise should be heavily reduced according to the law of large numbers, and so the segment mean should depend entirely on the tumor content, the normalization factor, and a discrete copy number. By plotting the segment mean versus the number of probes in a segment (Fig. S7, henceforth termed an MvsN plot), we observe tall peaks that represent different integer copy numbers.

From this, the association between copy number and segment mean for a particular sample can be deduced. Interestingly, a plot of the relationship between the average signal values of segments with single copy gains on a given sample versus the average signal values of segments with single copy losses on that sample indicated that our theoretical equation for the association between copy number and value did not match the data well. However, simply changing the base of the logarithm from 2 to 3 made for an excellent fit (Fig. S8). Average signal values for long segments representing copy numbers of 0 and 4 also seemed well-matched to this formulation. We therefore used instead the following equation for the theoretical mean value associated with a given copy number:

$$M(c,\tau,\delta) = log_3\left(\frac{c\tau + 2(1-\tau)}{2}\right) + \delta$$

For each sample, we applied the method of least squares to estimate the values of τ and δ that provided the best match between the observed segment means and their purported copy numbers based on the MvsN plot of that sample:

$$\langle \tau, \delta \rangle = \mathrm{Argmin}\left(\sum n_i (m_i - M(c_i,\tau,\delta))^2\right)$$

where the sum is over all segments for which the copy number could be identified from the MvsN plot, n_i is the number of probes in a segment, and c_i is the purported copy number. Based on these values, we calculated an estimated copy number for each segment (including those that appeared ambiguous in the MvsN plot), which were then rounded to the nearest integer value to give a copy number estimate for every segment. Those with a copy number of 0 were designated as homozygous losses. Those with copy number of 1 were designated as heterozygous losses. Those with copy number equal to 2 were designated wild type. Those with copy number equal to 3 were designated single copy gain. Finally, those with copy numbers of 4 and above were designated amplifications.

We recognized that there exist regions that frequently have copy numbers unequal to wild type in normal tissue in the human population, and that there is also the possibility that there might be regions in which the Affymetrix platform gives erroneous results. We therefore developed a blacklist of

genomic regions which if abnormal are unlikely to represent somatic copy number change. This list included 1,309 regions identified in the literature[28] as natural CNVs, plus an additional 254 short regions that were frequently found to exist as abnormal copy numbers in our data but which didn't appear to be associated with change in gene expression or exon probe density. The total area covered by these regions was 111 Mb. For the TCGA samples, which also had matched normal SNP 6.0 data, we also considered any abnormal segment discovered on the normal sample to be part of the blacklist for the corresponding tumor sample.

Segments for which the overlap with one of the blacklisted regions was greater than 25% of the segment length and also greater than 25% of the region length were flagged for exclusion. 30% of segments were flagged in this manner. Generally, the overlap of flagged regions was much higher than the 25% mutual overlap required, with 58% of the flagged segments having a mutual overlap greater than 90%. Each flagged segment was eliminated by setting its copy number designation equal to the copy number of the longer of the two segments that border it.

Once the copy number designations of all segments had been determined, adjacent segments of equal copy number were merged to form longer, combined segments. Segments of length less than 30Mb were reported as focal copy number changes. A chromosomal arm was declared to be amplified if segments covering more than 70% of the area were designated as amplifications. An arm was declared to have a single copy gain if it was not amplified but more than 70% of its length was covered by segments designated as either single copy gains or amplifications. A chromosomal arm was declared to be homozygously deleted if segments covering more than 70% of the area were designated as homozygous deletions. An arm was declared to have a heterozygous loss if it was not homozygously deleted but more than 70% of its length was covered by segments designated as either heterozygous losses or homozygous deletions. A chromosome was declared to be amplified if both arms were declared to be amplified. A chromosome was declared to be homozygously deleted if both arms were declared to be homozygously deleted. A chromosome was declared to be trisomy if either both arms indicated a single copy gain, or one arm was declared to have a single copy gain and the other was declared amplified. A chromosome was declared to have a heterozygous loss if either both arms indicated a heterozygous loss, or one arm was declared to have a heterozygous loss and the other was declared homozygously deleted.

Creation and testing a predictive model for somatic mutations in DLBCL

The majority of cases lacked matched normal DNA, requiring us to develop and test a tumor-only mutation calling pipeline. In brief, we analyzed WES data from TCGA DLBCL tumor and matched normal samples using the MuTect2 algorithm[29], thereby generating a "gold standard" set of somatic mutations. On a training set of 23 TCGA cases, we developed a Random Forest-based predictor of the MuTect2-derived somatic mutations, using a set of 25 annotation attributes of the MuTect2 somatic mutations as input for the Random Forest algorithm. We assessed this model on an independent validation set of 23 TCGA cases, and then applied it to variants that were generated by our tumor-only mutation calling pipeline, as detailed below.

Random Forest model for detecting somatic mutations in tumor without matched normal

Due to the complexity of cancer genome rearrangement as well as sample impurity and sub-clonal mutations, regular SNV callers such as GATK's HaplotypeCaller, which rely on a ploidy assumption, do

not work well on whole exome cancer samples. MuTect2[29], which was specifically developed for such a situation, albeit with paired-normal samples, is regarded as one of the most reliable cancer SNV callers[30]. However, cancer genome studies without paired normal samples cannot take advantage of tools such as MuTect2. In this supplement, we hypothesize that a machine-learning model based on variant toxicity scores can be trained using MuTect2 variants called from tumor-normal samples and applied to predict somatic mutations variants from a tumor-only pipeline.

Many available tools attempt to predict the pathogenicity of missense variants; most of these tools consider evolutionary factors such as the degree of conservation of the affected residue[31]. The premise of the proposed model is that a feature space consisting of toxicity scores provided by multiple models (CADD, MutationTaster, MutationAssessor, GERP++, Polyphen2, SiPhy, SIFT, LRT, PhyloP, VEST3, and FATHMM) along with a few intrinsic sequencing features can be used to classify a variant as somatic.

A random forest is an instantiation of an ensemble learning method in which a multitude of decision trees are constructed by sampling the feature space[32]. The training algorithm employs a form of bootstrap aggregation, or 'bagging,' which has been shown to improve classification performance over that of single classifiers operating on the full feature space. The model was trained using the R package 'randomForest'[33], which is an R implementation of the original work of Breiman[34]. The number of trees was set to 500; all other hyperparameters were set to their default values.

To develop and test the RF model, we downloaded 46 cancer-normal pairs of DLBCL exon-seq data from TCGA[7], including 40 DLBCL datasets utilized in previous analysis. The fastq-format sequencing files were first aligned to the human genome Hg19 via BWA in accordance with TCGA guidelines. MuTect2 was used to make both somatic and germline SNV calls. Those calls that passed statistical filters and fell into coding regions were used to train and test the classifier. The RF model was initially trained on 23 of the 46 samples consisting of a total of 2609 unique calls. The feature space comprises intrinsic data such as mutation rate, read depth, and duplication status as well as scores from 16 toxicity assessors. Table S14 provides a summary of the data from the 46 TCGA DLBCL samples.

Preprocessing

The toxicity scores that comprise most of the feature space are missing in many calls for a variety of reasons. Although simple linear regression can be used to impute single missing variables, missing data that occur in more than one variable (feature) present a challenge to classification algorithms including the random forest. Furthermore, the toxicity scores are a heterogeneous mix of numerical and categorical data. Multiple imputation is the method of choice for complex incomplete data problems. We used a method called *Multivariate Imputation by Chained Equations* (MICE)[35], an algorithm that has become popular in econometrics. We adapted MICE to the current problem of imputing the feature space of a somatic mutation caller.

Machine learning methodology

The goal of training a machine-learning model on MuTect2 calls from paired normal DLBCL samples is to mimic the MuTect2 assessment in the absence of paired normal samples. The model is trained with an aggregation of toxicity scores derived from ANNOVAR annotation (Table S15), some of which are aggregations of other scores, as well as intrinsic features such as mutation allele frequency, depth of reference reads, and duplication status of that variant across the training samples.

Random forests are manifestations of a popular ensemble learning method for classification tasks that operates by constructing a multitude of decision trees by sampling combinations of the feature space, thereby adding power to the prediction through fusion of the consensus prediction of the ensemble. In addition to the bagging of the feature space which is intrinsic to the random forest, we have employed a sample bagging scheme to boost prediction accuracy. In this scheme, 10 random forests are trained and their results are averaged. Since far more somatic calls were available for training than germline, we performed 10-fold subsampling of the somatic calls to provide balanced somatic/germline training sets to each member of the sample-bagging ensemble. Note that in so doing we effectively sample a grid of both samples and features since the random forest is itself an ensemble.

Cross-validation and holdout testing on DLBCL data

The random forest model was trained and cross-validated on 23 of the 46 available TCGA DLBCL samples. The cross-validation was a 23-fold procedure whereby for each fold, the variant calls from 22 of the 23 samples were used for training and the variants from the remaining sample were used for testing. The cross-validation performance is shown in Table S16**.** In the performance measures, somatic calls are regarded as predicted positives and germline calls are regarded as predicted negatives. The MuTect2 calls on the paired TCGA data are regarded as ground truth for this evaluation.

The receiver operating curve (ROC) for the 23-fold cross validation is depicted in Figure S9 along with the performance of the individual toxicity callers (shown as dashed curves) that constitute the feature space of the model.

Since the callers will typically perform well in one performance measure but not others, the ROC curves formed by the toxicity callers appears to be clearly inferior to that of the random forest, demonstrating the efficacy of the aggregation of toxicity scores being performed by the random forest model. In the current application of identifying somatic variants in the absence of paired normal samples, it is of paramount concern that positive (somatic) predictions be made with a high degree of confidence. For this reason, we looked at the measure of positive predictive value (PPV), which measures the likelihood that positive predictions are true positives, i.e., PPV = TP / (TP + FP).

Holdout testing was performed on the remaining 23 TCGA DLBCL samples not used in training or cross-validation, as summarized in Table S17. As the samples used for training and testing were selected randomly, the holdout set had a lower proportion of germline variants than the training set, leading to the performance discrepancies shown. The ROC and sensitivity-PPV curves for the holdout test are presented in Figure S10, which also shows that the ROC curves of the toxicity assessors are significantly worse than that of the aggregated model, in a manner consistent with the cross-validation result.

Somatic prediction on variants in the present study

The random forest model was applied to the filtered list of DLBCL variants featured in the present study. Although we do not have matched normal samples in this study, we can evaluate the rate of predicted somatic variants. Table S18 provides the results of this experiment on all variants (A) as well as those variants from TCGA donor samples (B). The predicted rate varies slightly between the subtypes of DLBCL. Since there was no significant difference in this estimated somatic mutation rate on the samples from TCGA versus those from other sources (p=0.36), we expect that the agreement between this model and MuTect2 on our entire data set would be similar to what we observed on the validation set. We

therefore estimate that if we had normal samples available and were able to apply MuTect2 to our samples, approximately 91.8% (95% CI 91.1 - 92.4) would be somatic. Given this high rate we decided for simplicity to assume that all the filtered mutations were somatic.

Prediction of AID-dependent somatic hypermutation target genes

The variant collection for hypermutation analysis was based on the output of BWA alignment and VarScan calls described in Exon-Seq analysis section. For hypermutation analysis, the variant selection, including silent coding region variants, variants in intron/UTR regions as well as typical somatic mutations, was much broader than that of somatic mutation analysis. To mitigate bioinformatic and experimental noise, each variant needed to fulfill the following criteria: 1) ≥5 mutation reads which constitute ≥20% of total reads; 2) ≤0.0001 variant frequency in the normal population as judged using the ExAC exome sequencing database, unless variant is present as a somatic mutation in the COSMIC database of cancer mutations; 3) variant not observed in our exome sequencing data of normal B cell control samples.

To create a predictor of somatic hypermutation, we collected mutations from our DLBCL samples in 44 genes previously described as targets of AID-dependent hypermutation[36]. This previous study developed a hypermutation predictor starting with 12 genes that are canonical targets of AID-dependent hypermutation in DLBCL. These investigators predicted 28 additional AID targets in DLBCL based on their relative proportions of: 1) transition versus transversion mutations; 2) mutations within the AID hotspot motif (WRCY|RGYW); 3) mutations within 2 kbp of the transcription start site; 4) A/T variants; 5) C/G variants; 6) silent versus disruptive variants[36]. We extended this mutation feature space based on concepts developed to comprehensively classify somatic mutations in cancer[37,38]. Specifically, we considered the sequence context of a mutation, defined as a 3-base window beginning 1 bp 5' of the mutation and ending 1 bp 3' of the mutation, and determined these proportion of mutations in a gene assigned to each of these triplet bins. We collected the above proportional and frequency feature variables for all genes aggregated from all the DLBCL datasets in this study. We trained hypermutation models based on the characteristics of the 44 published AID target genes within this feature space, and used the models to predict additional genes as AID targets in our DLBCL samples.

Two supervised classification algorithms were used to train models and compare the respective prediction results: a support vector machine (SVM) with a radial basis function kernel[39], and a random forest. Both models were trained on the same set of 44 candidate AID-dependent hypermutation genes described above. A similar number of genes were randomly sampled as pseudo-negatives for training the model. The rarity of somatic hypermutation genes justifies the random sampling strategy; fortunately, none of the sampled negatives co-appeared in the main cluster of hypermutations in dimensionality reduction experiments (as explained in the caption of Fig. S11).

The cross-validation performance of the models was tested in leave-one-out experiments. These experiments revealed that to achieve strong performance, the training set must be filtered to include only genes with a sufficient number of total mutations across all samples; we chose 40 total mutations as this minimum threshold through empirical testing. While the 40-mutation threshold filtered out only twelve positives from training, it greatly improved positive predictive value while preserving acceptable

sensitivity. Table S19 provides the cross-validation results for both models, which are strong in specificity and positive predictive value.

In Figure S11, standard dimensionality-reduction plots (from the R package “dimRed”) are compared between the 44 previously reported hypermutation genes and the hypermutation genes predicted by the SVM model. These plots revealed that that the spatial pattern of the hypermutation genes predicted by our methods overlaps with the spatial pattern of the previously reported hypermutation genes. Although the cross-validation results suggest slightly stronger performance of the random forest, we have adopted a cautious approach of requiring confirmation from both model predictions in designating a newly identified hypermutation gene. Table S20 lists the genes for which the posterior probability score of both the SVM and random forest is greater than 0.5. Complete results are available in Table S12.

Multi-platform analysis

Genomic feature definition

We adopted a gene-centric analysis strategy in which we defined exonic or splice junction mutations, copy number alterations (amplification (Amp); single copy gain (Gain); heterozygous loss (HL); homozygous deletion (HD)) and digital gene expression values for each protein-coding gene. Chromosomal rearrangements were predicted based on evidence of fusion transcripts from RNA-seq data (Fusion), supplemented with fluorescence in situ hybridization (FISH) data in the case of *BCL2* and *BCL6* (see Methods).

We use the word "feature" to indicate a set of samples which share a specified set of abnormalities on a specified gene. For each gene, we considered features in each of the following categories:

1) **Mutation:** Includes all samples that have a verified mutation in that gene with more than 10% estimated allele frequency.
2) **Subclonal Mutation:** Includes all samples that have a verified mutation in that gene with more than 2% allele frequency. This is a superset of the Mutation feature.
3) **Truncation:** Includes all samples that have a verified truncation mutation in that gene with more than 10% estimated allele frequency. This is a subset of the Mutation feature.
4) **Subclonal Truncation:** Includes all samples that have a verified truncation mutation in that gene with more than 2% allele frequency. This is a superset of the Truncation feature and a subset of the Subclonal Mutation feature.
5) **Focal Amplifications:** Samples which, according to the copy number analysis, have an Amplification segment of length less than 30Mb covering the specified gene.
6) **Focal Homozygous Deletions**: Samples which, according to the copy number analysis, have a Homozygous Deletion segment of length less than 30Mb covering the specified gene.

We also generated the following feature to be included in combination with other features (see below), but we did not include it in any analyses on its own.

7) **Focal Losses:** Samples which, according to the copy number analysis, have a Heterozygous Loss or Homozygous Deletion segment of length less than 30Mb covering the specified gene.

Those sample/feature combinations for which we lacked the data to make an assessment (e.g., copy number features on samples for which there was not a good SNP6.0 array available) were indicated as unavailable.

Additionally, we considered twelve features made of a combination of a mutational feature (Features 1-4) with a copy number feature (Features 5-7). For example, for each gene we considered "Truncation/Amplification" features which included all samples that had either a truncation or amplification for that gene. If for a particular sample, one or more of the features used to make the combination was unavailable, then the combination was indicated as unavailable, unless the other feature used to make the combination was positive. In that case, the combination was indicated as positive.

Finally, we included five specialized features related to our model subtyping:

a) *MYD88*^{L265P} mutations
b) Mutations of *MYD88* that were not L265P
c) *BCL6* fusions
d) *BCL2* translocations
e) *CD274* (PD-L1) or *PDCD1LG2* (PD-L2) fusions

To eliminate noise and concentrate on those features frequent enough to be of biological interest, we excluded all features that were found in fewer than four samples. For the combination features, we required that there be at least as many samples that were included in the feature due to their mutation as were included due to their copy number, and that at least four samples were included due to their copy number.

The GenClass iterative genetic subtype algorithm

In what follows, we will use the term "classification" to denote a mapping between a set of samples and one of five groups (N1, MCD, BN2, EZB and Other). Our goal was to start with an initial "seed" classification and evolve it in such a way as to maximize its association with our set of features while still maintaining the biology suggested by the initial classification. This is done through an iterative algorithm, termed "GenClass", that slowly adjusts the sample classification so as to maximize the strength of association between the classes and the set of features. The model development, and core group selection was done blinded to the clinical data which was unblinded only after a locked down model had been developed. Briefly, the algorithm followed the outline below:

1) Begin with an initial seed classification of samples.
2) Identify the list of features that are most highly associated with the current classification.
3) Based on these features, calculate an association statistic for the current classification and for all alternative classifications that differ from it by a change of at most one sample, and identify the classification with the highest prediction score.

4a) If the best classification identified in step 3 is not the current classification, return to step 2 using this best alternative classification in place of the current classification.

4b) If the best classification identified in step 3 is the current one, halt the iteration and report it as the final result.

These steps are spelled out in greater detail below:

Step 1) Initial sample classification

We started with the following initial seed classification:

1) Those samples with a NOTCH1 mutation were initially classified as N1.
2) Those samples with both a MYD88^{L265P} mutation and a CD79B mutation were initially classified as MCD.
3) Those samples with either a NOTCH2 mutation or a BCL6 fusion were initially classified as BN2.
4) Those samples which had either a EZH2 mutation (clonal or subclonal) or a BCL2 translocation were initially classified as EZB.
5) All other samples were initially classified as Other.

Step 2) Identification of features associated with current classification

Given a set of samples S and a feature F, we define the significance of the association between S and F by the chi-squared statistic including a Yates continuity correction:

$$\chi(S,F) = \frac{(|\sum I_S I_F \sum(1-I_S)(1-I_F) - \sum(1-I_S)I_F \sum I_S(1-I_F)| - N/2)^2}{\sum I_S \sum I_F \sum(1-I_S) \sum(1-I_S)}$$

where N is the total number of samples, I_S, and I_F are the indicators of the sample being in set S or having feature F, and the summation takes place over the set of samples for which feature F was available.

The association between a classification and a specific feature was equal to the maximum of this statistic over the first four subsets (N1, MCD, BN2, EZB):

$$V(C,F) = \max_{\substack{S\in\{N1,MCD,BN2,EZB\} \\ \text{according to classification C}}} \left(\chi(S,F)\right)$$

Based on these statistics, the list of features associated with the current classification, C_{current}, were defined as all those features F that met the following criteria:

1) According to the current classification, the prevalence of F was greater than 10% in at least one the four main subtypes (N1, MCD, BN2, EZB).
2) There are no other features F' that are associated with the same gene as F for which $V(C_{\text{current}}, F') > V(C_{\text{current}}, F)$. If two or more features for the same gene are tied for highest V score, then one of the features is chosen at random for inclusion and the remaining are excluded.
3) If F is a copy number or combination feature, there is no other copy number or combination feature F' such that $V(C_{\text{current}}, F') > V(C_{\text{current}}, F)$ and the gene associated with F' is within 15Mb of the gene associated with F. If two or more such features within 15Mb are tied for highest V score, one of the features is chosen at random for inclusion and the remaining are excluded.
4) $V(C_{\text{current}}, F) > 10.85$ (the equivalent of significance p<0.001 according to a chi-squared test)

Step 3) Calculate association statistic for current classification and alternative classifications

Let $\Omega(C_{\text{current}})$ be the set of classifications that differ from the current classification by the change of at most one class label. We identify the best classification according to the formula,

$$C_{\text{best}} = \underset{C\epsilon\Omega(C_{\text{current}})}{\text{argmax}} \left(\sum V(C,F)\right)$$

where the sum is taken over all features identified in step 2 under the current classification. Note that in this step, the set of features remains fixed and does not change for different alternative features in $\Omega(C_{\text{current}})$. If multiple classifications are tied for best, one is chosen at random.

Step 4) Halt procedure or continue to the next iteration

If $C_{\text{best}} \neq C_{\text{current}t}$ then the reclassification of that sample made a step towards the improvement of the classification. We then use C_{best} as our current classification and return to step 2, identify a new set of features associated with this classification and look to see which next change of class label most improves the association with the set of features. If $C_{\text{best}} = C_{\text{current}t}$, then there is no single change in class label that can improve the association, and so we consider the current classification to be optimal and report it as the final answer. Since the set of features over which the optimization takes place changes from step to step, it is theoretically possible that the iteration falls into a loop and could fail to converge. To prevent this, we actually expanded the halting criteria to stop if the C_{best} has ever been used as $C_{\text{current}t}$ in any of the previous iterations.

Application of the GenClass method to classify a single sample

As defined above, the GenClass algorithm was developed as an evolving classifier acting on a study set of cases. However, to be clinically useful, it is necessary to develop a method that can take a fixed, previously developed classification and apply it to a novel sample. We can do this by first identifying a set of features based on the current classification as in step 2 of the modeling procedure, and then, in a similar manner to steps 3-4 above, calculate the sum of V-scores for the 5 classifications consisting of the current classification plus the addition of the novel sample classified in the 5 possible ways (BN2, MCD, EZB, N1, Other). We then choose the classification for the novel sample which results in the greatest sum of V scores.

We recognize that in a clinical setting it is unreasonable to expect a molecular characterization of patient samples as complete as what was available for the samples in our study. We expect it be feasible to obtain information regarding BCL6 fusions and BCL2 translocations as well as mutation data from a limited panel of genes. We also consider it likely that we could also identify cases with homozygous deletions or high-level amplifications of a small set of selected genes, but that it is unlikely that information about heterozygous losses would be available.

As a proof of principle, we selected 58 features based on BCL6 fusions and BCL2 translocations, mutational information on 52 selected genes and either amplification or homozygous deletion information on 12 selected genes, and tested models based on these features using 10-fold cross validation (Table S13). This model matched our original classification in 97.5% of the samples. In a second pilot model in which we selected only BCL6 fusions and BCL2 translocations and mutation information from 50 selected genes, we obtained a 10-fold cross validated agreement in 94.8% of the samples.

These results are preliminary, do not take into account the technical variability of any platform that might be used in a clinical setting, and will need to be independently validated in future studies, but they do strongly suggest that it should be possible to develop a clinical test for this distinction.

Permutation testing

We wished to demonstrate that the groups we identified were more strongly associated with the feature space of our data than would be expected by chance. To do this, we used the criteria described in step 2 of the class prediction algorithm to associate a classification with a set of genes, with the exception that we chose a chi-squared cut-off of 6.635 (equivalent to $p<0.01$) rather than 10.85. We

then counted the number of associated genes under the original classification and compared this value to the number of associated genes when the class labels were randomly permuted. This was done for the original seed classification defined in step 1 of the prediction algorithm, and for the final genetic subtype classification produced by the prediction algorithm. All features from genes used to define the original seed groups were excluded from the list of associated genes. For the original seed classification, 6/1000 permutations had a larger number of associated genes than the number of genes associated with the unpermuted class labels, so we reported a p-value for significance of p=0.006. For the final genetic subtypes, the number of genes found to be associated with the unpermuted class labels was substantially higher than for any of the permuted results, and so we conservatively reported the p-values as p<0.001. A similar analysis for the ABC/GCB subgroup distinction (Fig. 1A, Fig. S1B) yielded a permutation probability of p<0.001.

Random forest model for genetic subtype prediction

In addition to the GenClass iterative method for DLBCL genetic subtype prediction described above, we created an independent random forest method to define genetic subtypes. The genetic feature space that we used for the random forest prediction was identical that used by the GenClass method. To create a 5-category random forest, we used the same 4 "seed" subsets of DLBCL samples as were used as the starting point for the GenClass method. In addition, the random forest method requires a neutral fifth subset of DLBCL, which we term "5CAT", that was constituted using an algorithm that finds 'negatives' in the data, as follows. Initially, 80 DLBCL samples (more than needed for the final model) were randomly selected for membership in 5CAT, excluding samples in the 4 seed subsets. A random forest model was trained on the seed subsets plus 5CAT, and leave-one-out cross-validated prediction was applied to the 5CAT samples only. Any of the initial 5CAT members that were predicted to belong to one of the other subsets was removed from the 5CAT set. The random forest was retrained again and cross-validation performed to remove additional samples from 5CAT that were predicted to belong to one of the 4 seed subsets. This process was repeated until the 5CAT set was stable at 39 samples. This final 5CAT subset plus the 4 seed subsets were used to train a final 5-category random forest model, which was then used to predict membership of each remaining sample in one of the 4 subsets, or to declare the sample unassigned.

Figure S12 depicts Venn diagrams comparing genomic subtype membership predicted by the GenClass predictor and the random forest predictor. As mentioned above, this result was based on using the same genetic feature space for the two methods, namely the set of genetic features that was statistically differential between each starting seed subset and all other cases (see Genclass methods for details). When the random forest method used independently selected features selected from the entire genetic feature space, the random forest model predictions overlapped with those of the GenClass algorithm to a similar extent (data not shown).

Additional statistical analysis

Unless otherwise specified, all p-values relating discrete variables to each other are calculated using a Fisher exact test. All p-values for survival are calculated from a Cox proportional hazard score test. IPI score was treated as a categorical variable taking on three values: Low for IPI=(0,1), Intermediate for IPI = (2,3), High for IPI =(4,5). Gene set enrichment p-values are calculated as previously described[22]. All p-values reported are two-sided.

Figure S1

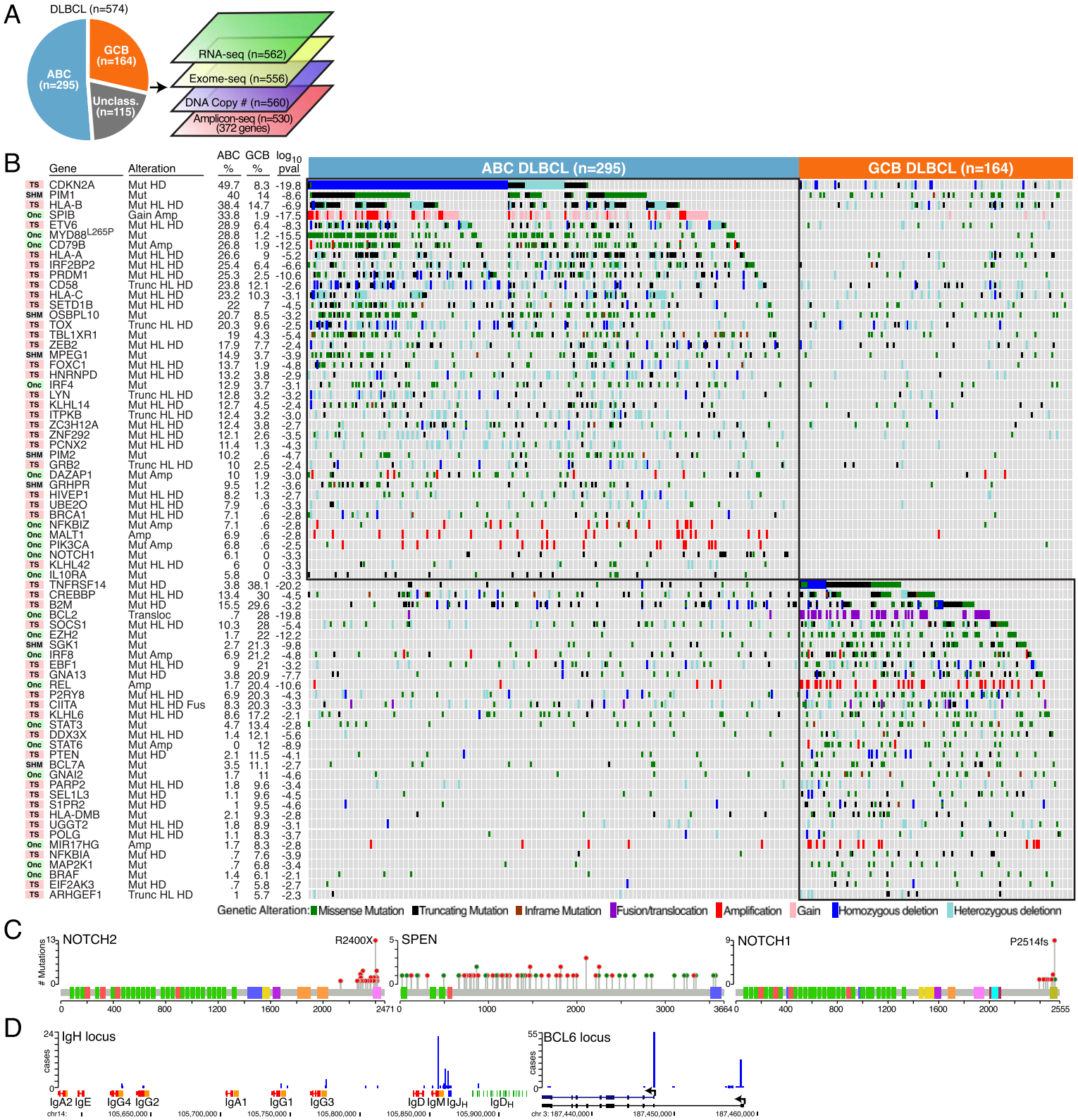
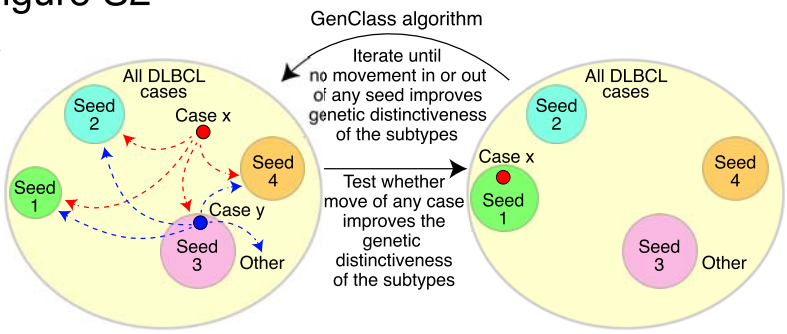


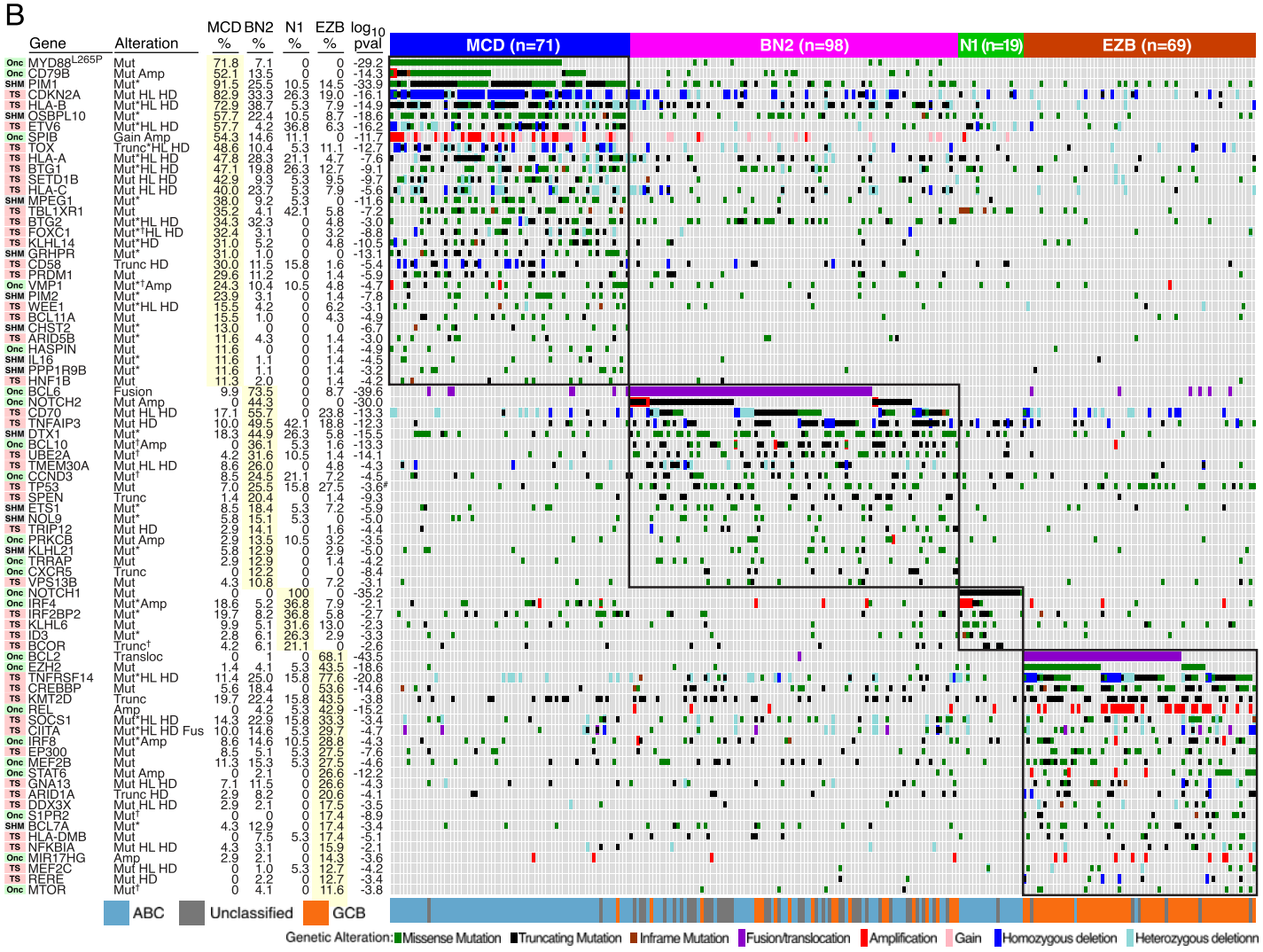
Figure S1. **A.** Study design illustrating assignments of the cases to gene expression subgroups (left) and the number of cases that were studied using each genomic platform (right). **B.** Genetic aberrations that distinguish ABC and GCB DLBCL. For each gene, the constellation of genetic aberrations that best distinguish the ABC and GCB subgroups is shown, together a $-\log_{10}$ P-value from a Fisher's exact test for this distinction. Shown are aberrations in subtype distinction genes with $\geq$5% prevalence within the subtype and P<0.01 for the distinction of the subtype from all other DLBCL. Putative assignment as an oncogene (Onc), tumor suppressor (TS) or target of aberrant somatic hypermutation (SHM) is shown. **C.** Position of mutational alterations in the protein structures of NOTCH2, SPEN, and NOTCH1 (see Supplemental Methods for selection of candidate somatic mutations). Green: missense; Red: truncation (nonsense, frameshift). **D.** Genomic position of mRNA fusions involving the immunoglobulin heavy chain (IgH) locus and the *BCL6* locus.

Figure S2

A



B



	Gene	Alteration	MCD %	BN2 %	N1 %	EZB %	$\log_{10}$ pval
Onc	MYD88^{L265P}	Mut	71.8	7.1	0	0	-29.2
Onc	CD79B	Mut* Amp	52.1	13.5	0	0	-14.3
SHM	PIM1	Mut*	91.5	25.5	10.5	14.5	-33.9
TS	CDKN2A	Mut HL HD	82.9	33.3	26.3	19.0	-16.1
TS	HLA-B	Mut*HL HD	72.9	38.7	5.3	7.9	-14.9
SHM	OSBPL10	Mut*	57.7	22.4	10.5	8.7	-18.6
TS	ETV6	Mut*HL HD	57.7	4.2	36.8	6.3	-16.2
Onc	SPIB	Gain Amp	54.3	14.6	11.1	0	-11.7
TS	TOX	Trunc*HL HD	48.6	10.4	5.3	11.1	-12.7
TS	HLA-A	Mut*HL HD	47.8	28.3	21.1	4.7	-7.6
TS	BTG1	Mut*HL HD	47.1	19.8	26.3	12.7	-9.1
TS	SETD1B	Mut HL HD	42.9	9.3	5.3	9.5	-9.7
TS	HLA-C	Mut HL HD	40.0	23.7	5.3	7.9	-5.6
SHM	MPEG1	Mut*	38.0	9.2	5.3	0	-11.6
TS	TBL1XR1	Mut	35.2	4.1	42.1	5.8	-7.2
TS	BTG2	Mut*HL HD	34.3	32.3	0	4.8	-3.0
TS	FOXC1	Mut*†HL HD	32.4	3.1	0	3.2	-8.8
TS	KLHL14	Mut*HD	31.0	5.2	0	4.8	-10.5
SHM	GRHPR	Mut*	31.0	1.0	0	0	-13.1
TS	CD58	Trunc HD	30.0	11.5	15.8	1.6	-5.4
TS	PRDM1	Mut	29.6	11.2	0	1.4	-5.9
Onc	VMP1	Mut*†Amp	24.3	10.4	10.5	4.8	-4.7
SHM	PIM2	Mut*	23.9	3.1	0	1.4	-7.8
TS	WEE1	Mut*HL HD	15.5	4.2	0	6.2	-3.1
TS	BCL11A	Mut	15.5	1.0	0	4.3	-4.9
SHM	CHST2	Mut*	13.0	0	0	0	-6.7
TS	ARID5B	Mut*	11.6	4.3	0	1.4	-3.0
Onc	HASPIN	Mut	11.6	0	0	1.4	-4.9
SHM	IL16	Mut*	11.6	1.1	0	1.4	-4.5
SHM	PPP1R9B	Mut*	11.6	1.1	0	1.4	-3.2
TS	HNF1B	Mut	11.3	2.0	0	1.4	-4.2
Onc	BCL6	Fusion	9.9	73.5	0	8.7	-39.6
Onc	NOTCH2	Mut Amp	0	44.3	0	0	-30.0
TS	CD70	Mut HL HD	17.1	55.7	0	23.8	-13.3
TS	TNFAIP3	Mut HD	10.0	49.5	42.1	18.8	-12.3
SHM	DTX1	Mut*	18.3	44.9	26.3	5.8	-15.5
Onc	BCL10	Mut†Amp	0	36.1	5.3	1.6	-13.3
TS	UBE2A	Mut†	4.2	31.6	10.5	1.4	-14.1
TS	TMEM30A	Mut HL HD	8.6	26.0	0	4.8	-4.3
Onc	CCND3	Mut†	8.5	24.5	21.1	7.2	-4.5
TS	TP53	Mut	7.0	25.5	15.8	27.5	-3.6#
TS	SPEN	Trunc	1.4	20.4	0	1.4	-9.3
SHM	ETS1	Mut*	8.5	1.8	5.3	7.2	-5.9
SHM	NOL9	Mut*	5.8	15.1	5.3	0	-5.0
TS	TRIP12	Mut HD	2.9	14.1	0	1.6	-4.4
Onc	PRKCB	Mut Amp	2.9	13.5	10.5	3.2	-3.5
SHM	KLHL21	Mut*	5.8	12.9	0	2.9	-5.0
Onc	TRRAP	Mut	2.9	12.9	0	1.4	-4.2
Onc	CXCR5	Trunc	0	12.2	0	0	-8.4
TS	VPS13B	Mut	4.3	10.8	0	7.2	-3.1
Onc	NOTCH1	Mut	0	0	100	0	-35.2
Onc	IRF4	Mut*Amp	18.6	5.2	36.8	7.9	-2.1
TS	IRF2BP2	Mut*	19.7	8.2	36.8	5.8	-2.7
TS	KLHL6	Mut	9.9	5.1	31.6	13.0	-2.3
TS	ID3	Mut*	2.8	6.1	26.3	2.9	-3.3
TS	BCOR	Trunc†	4.2	6.1	21.1	0	-2.6
Onc	BCL2	Transloc	0	2.1	0	68.1	-43.5
Onc	EZH2	Mut	1.4	4.1	5.3	43.5	-18.6
TS	TNFRSF14	Mut*HL HD	11.4	25.0	15.8	77.6	-20.8
TS	CREBBP	Mut	5.6	18.4	0	53.6	-14.6
TS	KMT2D	Trunc	19.7	22.4	15.8	43.5	-3.8
Onc	REL	Amp	0	4.2	5.3	42.9	-15.2
TS	SOCS1	Mut*HL HD	14.3	22.9	15.8	33.3	-3.4
TS	CIITA	Mut*HL HD Fus	10.0	14.6	5.3	29.7	-4.7
Onc	IRF8	Mut*Amp	8.6	14.6	10.5	28.8	-4.3
TS	EP300	Mut	8.5	5.1	5.3	27.5	-7.6
Onc	MEF2B	Mut	11.3	15.3	5.3	27.5	-4.6
Onc	STAT6	Mut Amp	0	2.1	0	26.6	-12.2
TS	GNA13	Mut HL HD	7.1	11.5	0	26.6	-4.3
TS	ARID1A	Trunc HD	2.9	8.2	0	20.6	-4.1
TS	DDX3X	Mut HL HD	2.9	2.1	0	17.5	-3.5
Onc	S1PR2	Mut†	0	0	0	17.4	-8.9
SHM	BCL7A	Mut*	4.3	12.9	0	17.4	-3.4
TS	HLA-DMB	Mut	0	7.5	5.3	17.4	-5.1
TS	NFKBIA	Mut HL HD	4.3	3.1	0	15.9	-2.1
Onc	MIR17HG	Amp	2.9	2.1	0	14.3	-3.6
TS	MEF2C	Mut HL HD	0	1.0	5.3	12.7	-4.2
TS	RERE	Mut HD	0	2.2	0	12.7	-3.4
Onc	MTOR	Mut†	0	4.1	0	11.6	-3.8

Figure S2. **A.** Schematic of the GenClass iterative genetic classifier used to create the DLBCL genetic subtypes. The method starts with a set of seed classes, four shown here, and an "Other" category. At each step of the iteration, all possible moves of a single sample into or out of a seed class are considered, and the change in a Chi-square-based genetic distinctiveness metric is assessed. This metric assesses all differences between the prevalence of genetic alterations in a class versus the prevalence among cases not in that class, and aggregates these differences across all classes to provide an overall measure of genetic distinctiveness of a particular assignment of samples to the classes (see Supplemental Methods for details). The single move that makes the biggest improvement in this genetic distinctiveness metric is chosen, and this process is iterated to the point at which no movement improves the metric. **B.** Genetic alterations that distinguish the DLBCL genetic subtypes. For each subtype, the constellation of genetic aberrations that best distinguishes subtype cases from all other DLBCLs is shown, together a $\log_{10}$ P-value from a Fisher's exact test for this distinction. Shown are aberrations in subtype distinction genes with $\geq$10% prevalence within the subtype ($\geq$20% for N1 genes) and P<0.001 for the distinction of the subtype from all other cases (P<0.01 for N1). Asterisks indicate genes that are predicted to be targets of aberrant somatic hypermutation mediated by AID (see Supplemental Methods). (†) genes for which subclonal mutations were included. (#) P-value refers to MCD vs. other cases. Putative assignment as an oncogene (Onc), tumor suppressor (TS) or target of aberrant somatic hypermutation (SHM) is shown. Mut: mutation; Amp: amplification; Gain: single copy gain; HL: heterozygous loss; HD: homozygous deletion; Fus: gene fusion; Transloc: translocation.

Figure S3

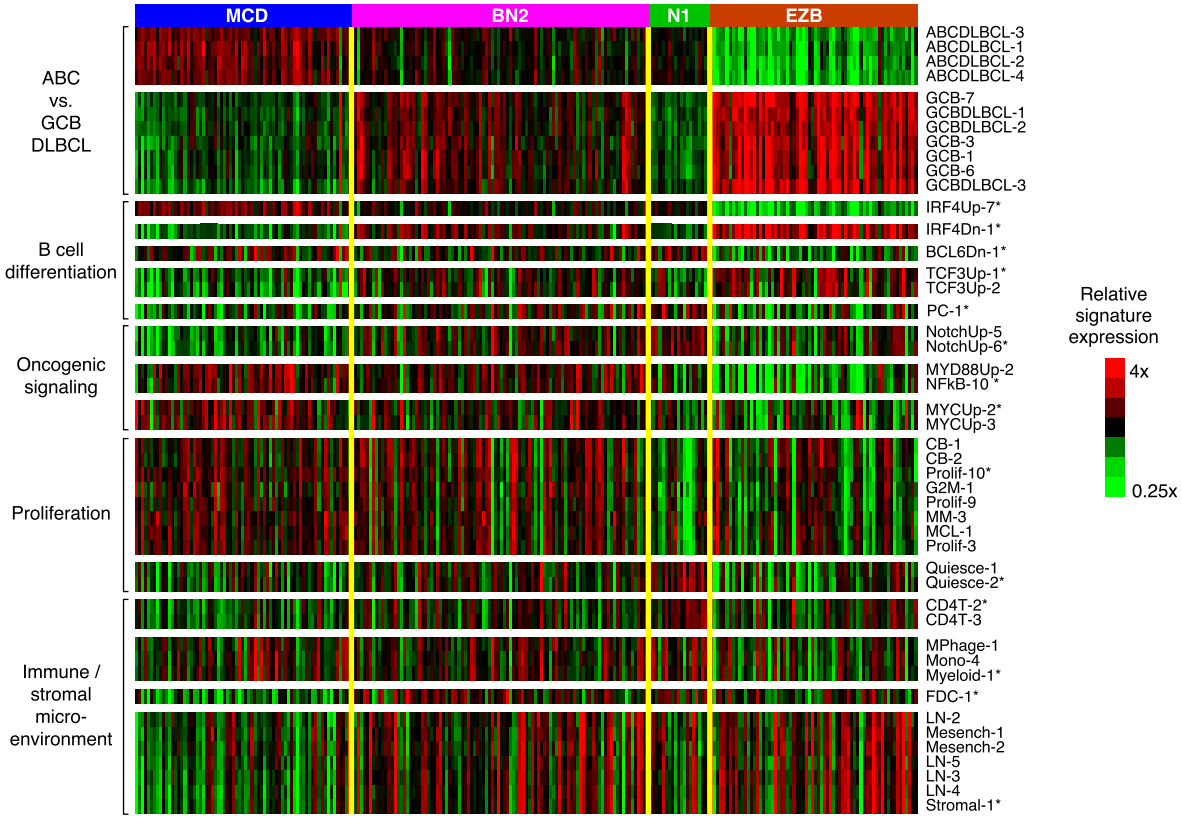


Figure S3. Gene expression signatures that distinguish the DLBCL genetic subtypes.
Each row represents the signature average value (see Supplemental Methods) for each patient in the indicated DLBCL genetic subtypes, according to the color scale shown. The signatures were selected from the SignatureDB database[24] and were significantly differential between the genetic subtype ($P<0.05$, F-test). Signature averages with correlated values across DLBCL cases are grouped together, revealing biological distinctions between the genetic subtypes. A full annotation of these signatures is available at https://lymphochip.nih.gov/signaturedb/. Asterisks indicate signatures presented in Figure 3. Briefly, their annotations are: IRF4Up-7: direct IRF4-activated genes that are highly expressed in ABC DLBCL; IRF4Dn-1: Direct IRF-repressed genes that are expressed at low levels in ABC DLBCL; BCL6Dn-1: genes that are downregulated by BCL6; TCF3Up-1: genes that are upregulated by TCF3 in Burkitt lymphoma; PC-1: genes more highly expressed in normal bone marrow-derived plasma cells than in mature B cells; Notch1Up-6: Direct NOTCH1-transactivated genes in chronic lymphocytic leukemia; NFkB-10: genes upregulated by IκB kinase-induced NF-κB activity in ABC DLBCL; MycUp-2: genes upregulated by Myc overexpression; Prolif-10: cell cycle-regulated genes that are upregulated in proliferating cells; Quiesce-2: genes upregulated in non-proliferating, quiescent cells; CD4T-2: genes characteristically expressed in naïve CD4 T cells; Myeloid-1: genes characteristically in normal myeloid blood subpopulations; FDC-1: genes characteristically expressed in normal follicular dendritic cells; Stromal-1: genes expressed in DLBCL tumors with a high content of mesenchymal cells and extracellular matrix, which is associated with favorable survival in response to R-CHOP chemotherapy.[5]

Figure S4

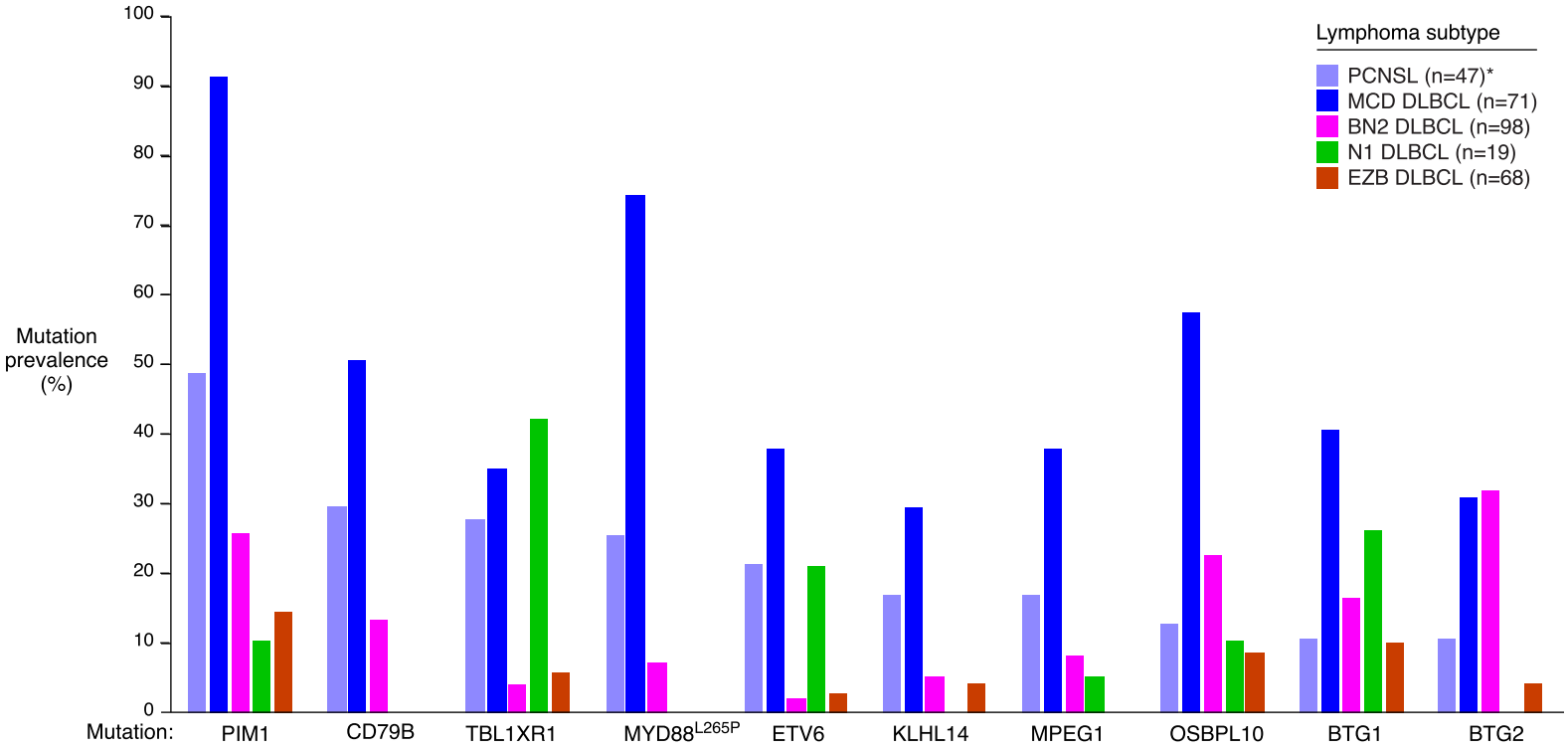


Figure S4. Mutations shared by primary central nervous system lymphoma (PCNSL) and DLBCL genetic subtypes. PCNSL mutational data was curated from 4 whole exome sequencing studies of PCNSL biopsies from 47 donors.[1-4] Shown are genes present in >10% of PCNSL tumors that are characteristically altered in MCD DLBCL.

Figure S5

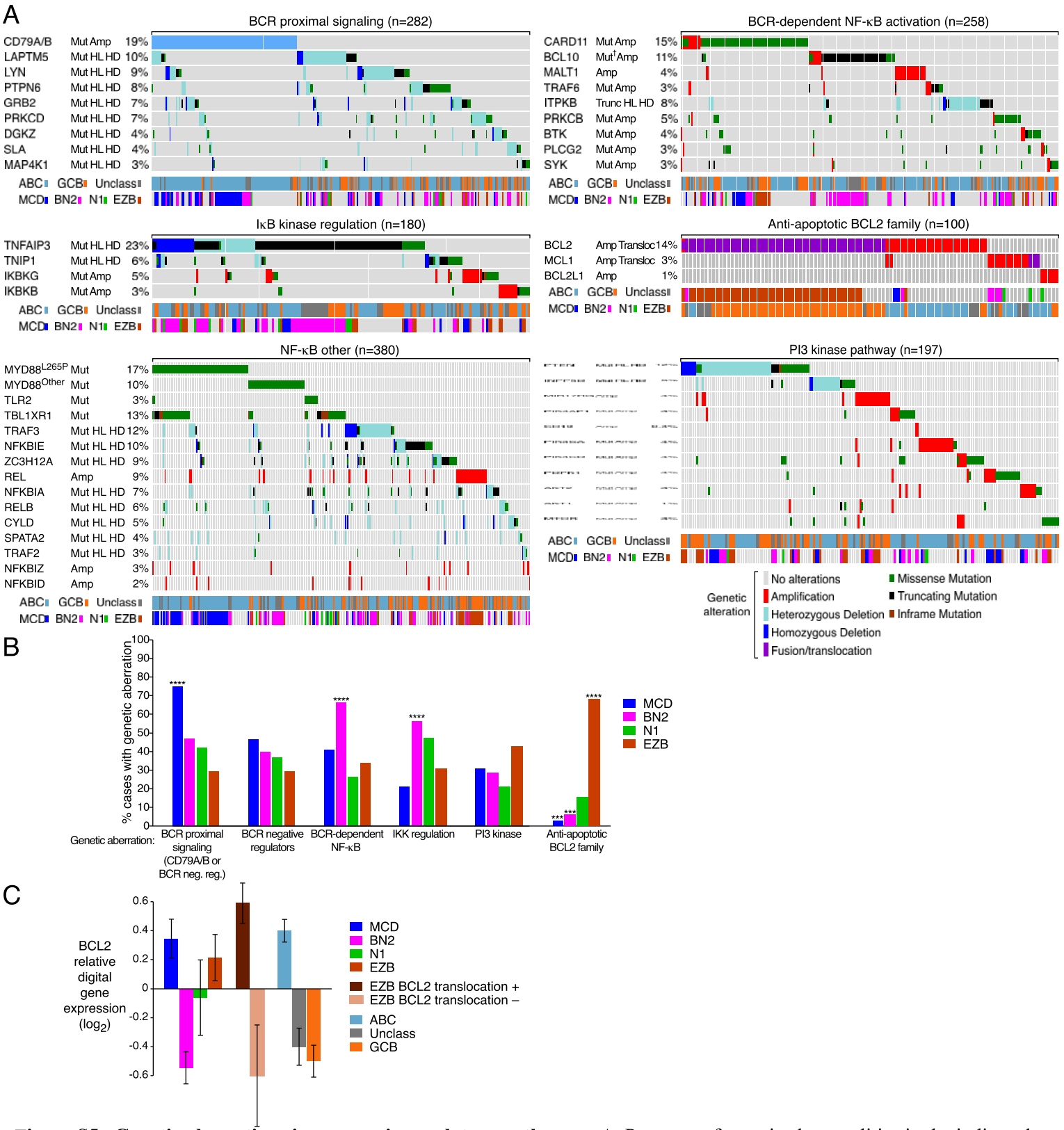


Figure S5. Genetic aberrations in oncogenic regulatory pathways. A. Presence of genetic abnormalities in the indicated oncogenic signaling pathways. For each gene, the genetic aberrations that are displayed is give along with the percentage of cases with these aberrations. Cases lacking these aberrations are not shown. Assignment to the gene expression subgroups and genetic subtypes is shown. Mut: mutation; Trunc: truncating mutation; HL: heterozygous loss; HD: homozygous deletion; Amp: amplification; Transloc: translocation. †Subclonal mutations included. **B.** Prevalence of genetic aberrations targeting oncogenic signaling pathways in the DLBCL genetic subtypes. The genetic aberrations included in each of the indicated oncogenic signaling pathways are those shown in Fig. S5A ****P<0.0001; ***P<0.001. **C.** Relative digital gene expression for BCL2 in the indicated DLBCL subsets. Error bars: SEM.

Figure S6

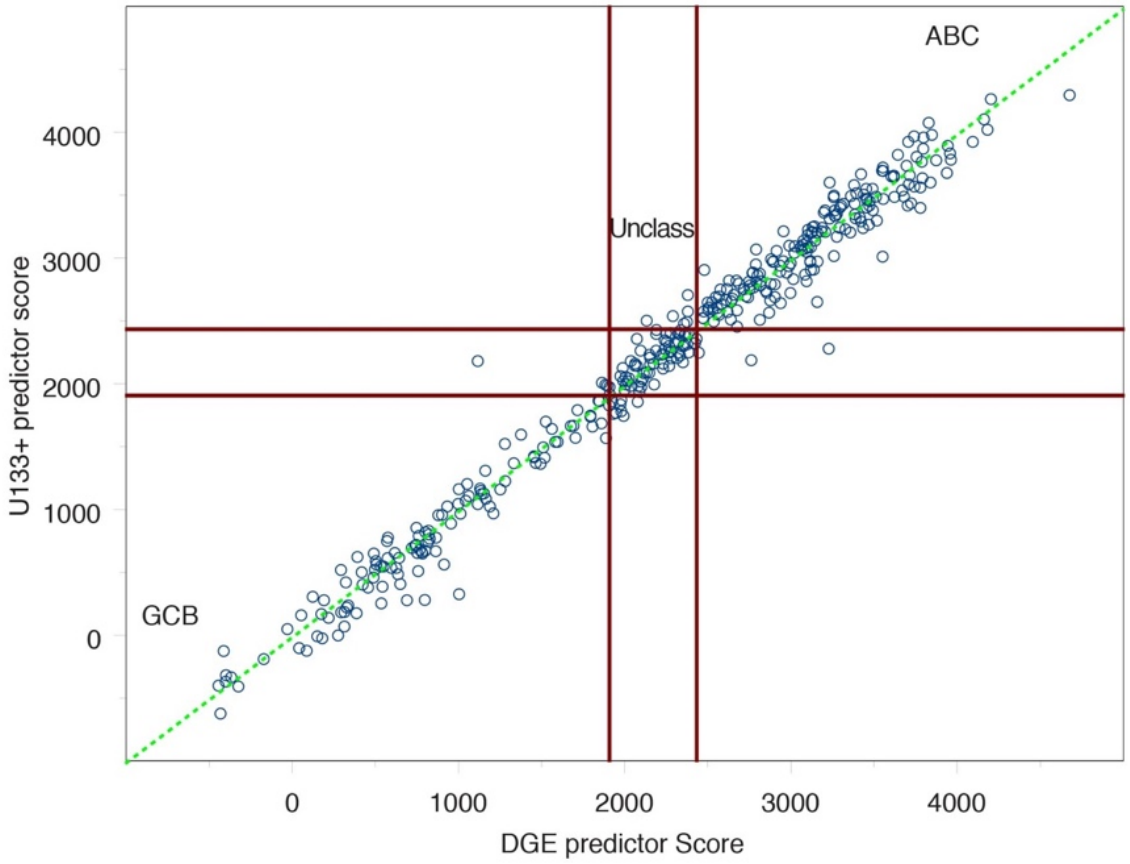


Figure S6. Comparison of Affymetrix U133+ based predictor score to RNAseq based DGE predictor score
381 samples with matched U133+ and RNAseq data were available. Vertical and horizontal lines represent cut-points between class calls.

Figure S7

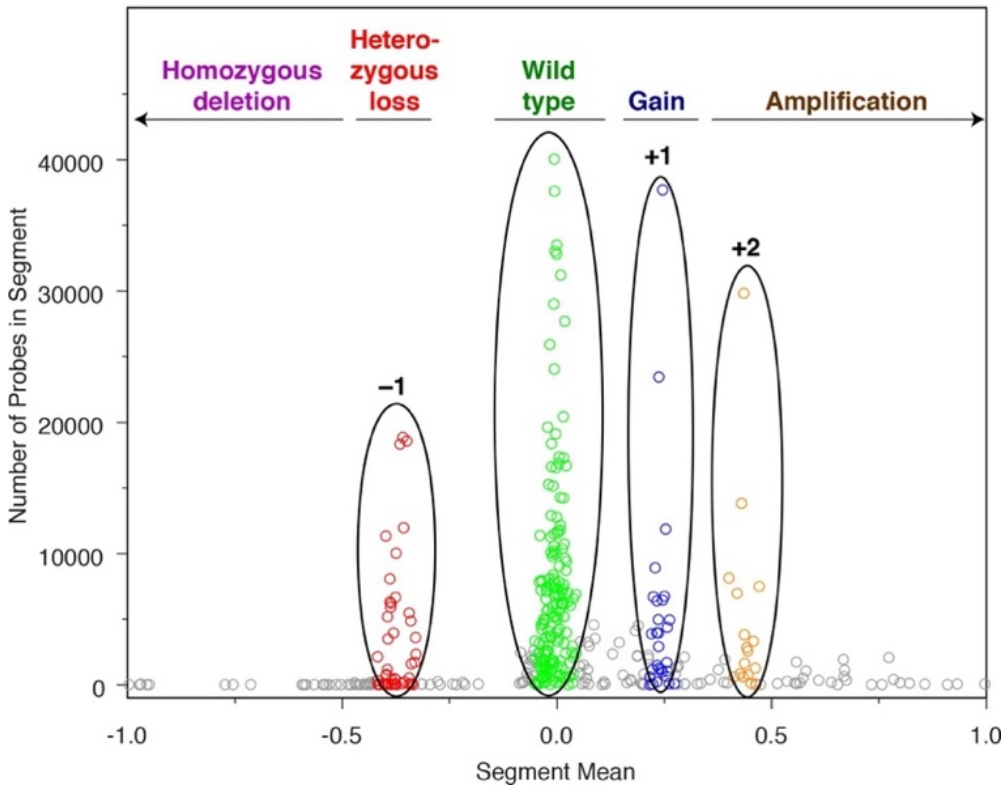


Figure S7. MvsN plot of CGH segmentation for a typical sample.
Each dot represents a segment. X-axis indicates the average signal value for probes in segment, the Y-axis represents the number of probes available in segment.

Figure S8

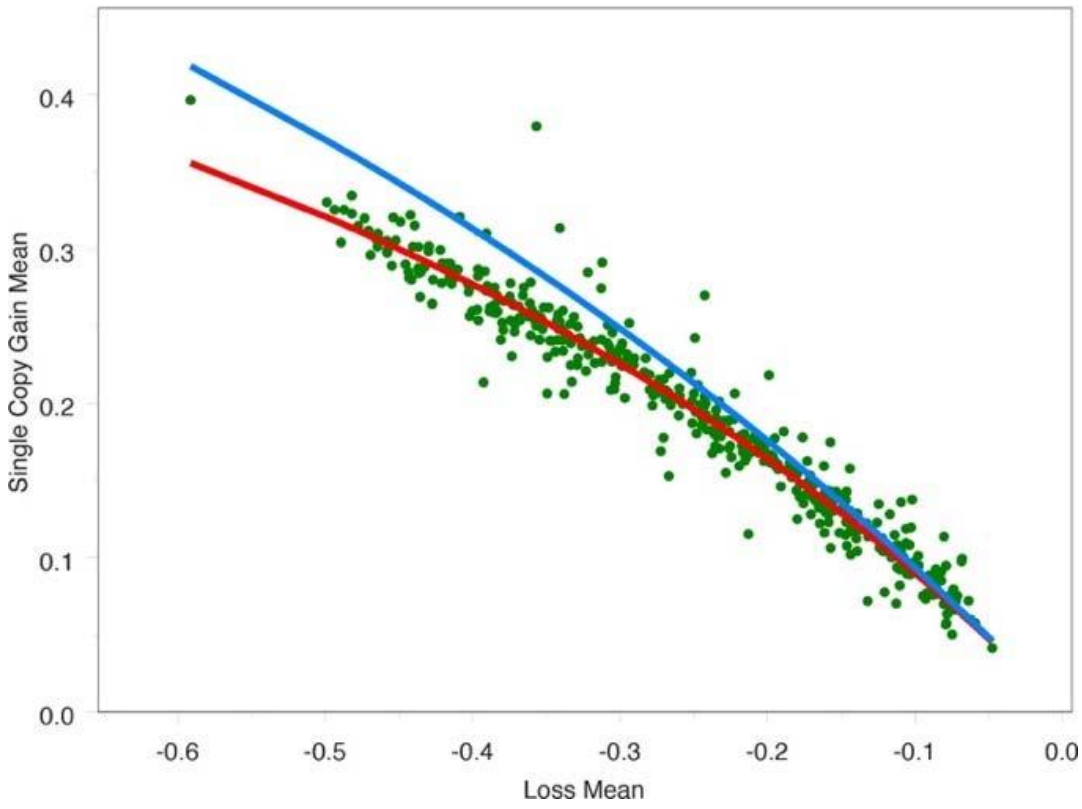


Figure S8. Relationship between signal values of single copy gains and single copy losses
Each dot represents a sample for which there were at least 10,000 probes that were identified as being associated with segments representing single copy losses and at least 10,000 probes representing single copy gains according to their MvsN plot. The Y-axis represents the average signal of the single copy gains. The X-axis represents the average signal of the single copy losses. The blue line represents the theoretical value for a model based on Log_2. The red line represents the theoretical value for a sample for a model based on Log_3.

Figure S9

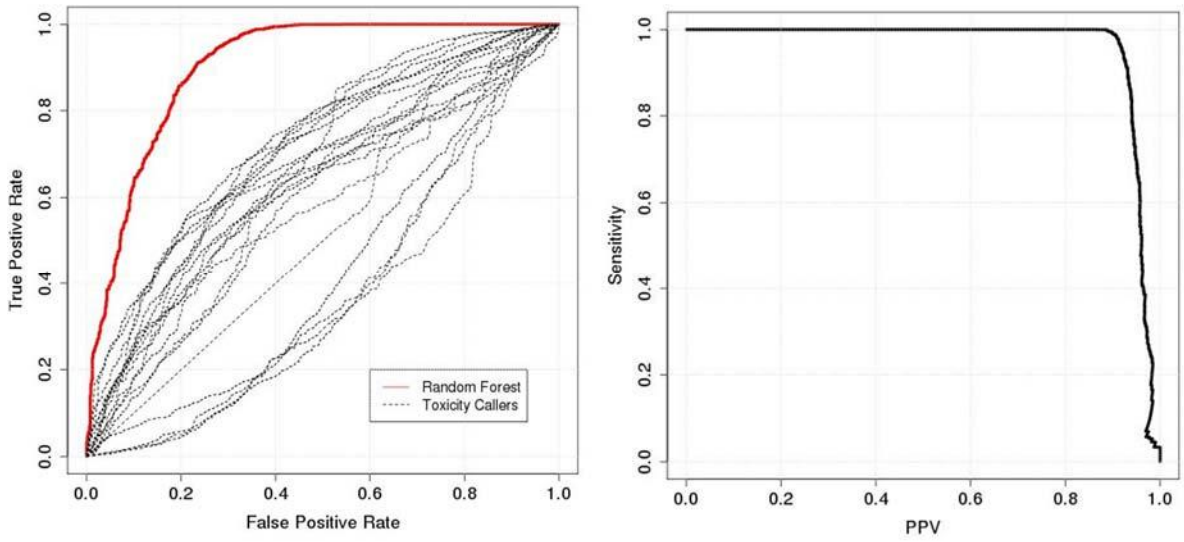


Figure S9. ROC and PPV-sensitivity curves of the RF model in cross-validation tests
The ROC (left) and PPV-sensitivity curve (right) of the RF model in 23-fold cross-validation tests with all 44 available features (solid curve) as well as performance of individual toxicity assessors (dotted curves).

Figure S10

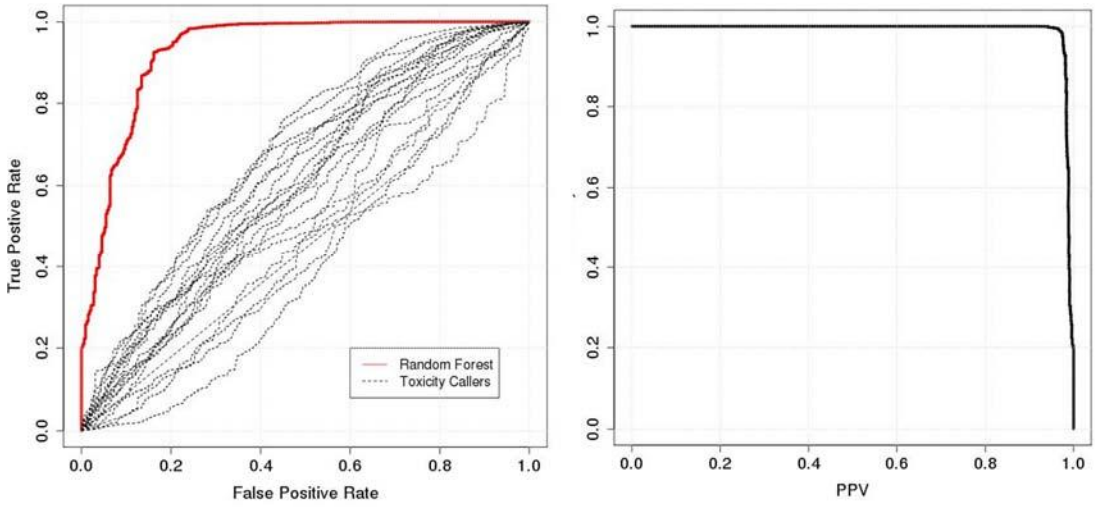


Figure S10. ROC and PPV-sensitivity curves of the RF model on holdout samples
ROC (left) and PPV-sensitivity curve (right) of the RF model on the 23 holdout TCGA DLBCL samples (solid curve). The dotted curves present the ROCs of the toxicity assessors on the same data.

Figure S11

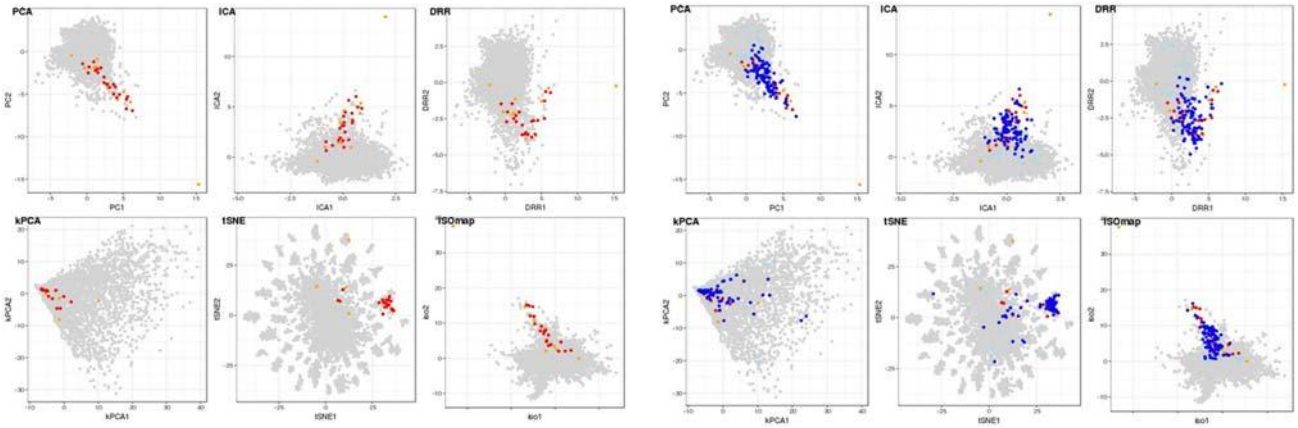


Figure S11. Dimensionality-reduction plots of published and predicted somatic hypermutation genes.
Dimensionality-reduction plots of previously published (left) as well as predicted (right) somatic hypermutation genes. Red dots indicate hypermutations from 32 genes predicted by Khodabakhshi et al.[36]; orange dots are hypermutations in 12 canonical AID target genes; dark blue dots are hypermutations predicted by our models; light blue dots are pseudo-negatives. Dimensionality reduction methods are labeled on their respective plots. None of the pseudo-negative genes landed on the t-SNE "island" that occupies most of the hypermutations.

Figure S12

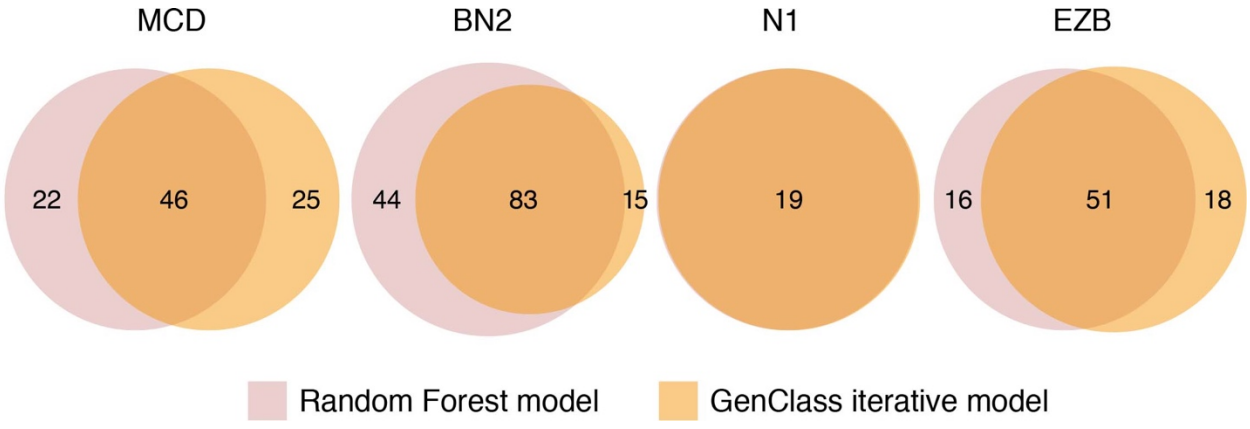


Figure S12: Overlap between predicted membership of the four genomic DLBCL subtypes
Venn diagrams depicting overlap between predicted membership of the four genomic DLBCL subtypes, as predicted by two different algorithms, the random forest and iterative predictor.

Supplemental Tables

Table S1: Mutation frequency of genes mutated in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing percentage of DLBCL with mutations (> 10% allele frequency) in 13,743 genes. DLBCL were sub-grouped by gene expression in ABC, Unclassified, and GCB or by genetic sub-setting in MCB, BN2, N1, EZB, and Other (i.e. genetically unclassifiable). Other ABC, other GCB, and other Unclassified denotes mutation frequency of genetically unclassifiable cases within the gene expression subgroups. Total indicates mutation frequency in all DLBCL. Cases were counted maximally once per gene.

Table S2: Mutation frequency of genes including subclonal mutations in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing percentage of DLBCL with mutations and subclonal mutations (> 2% allele frequency) in 13,743 genes. DLBCL were grouped as described for Table S1.

Table S3: Frequency of chromosomal amplifications of genes in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing percentage of DLBCL with chromosomal amplifications (i.e. copy numbers of 4 and above) in 18,130 genes. DLBCL were grouped as described for Table S1.

Table S4: Frequency of chromosomal gains of genes in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing percentage of DLBCL with single copy gains (copy number of 3) in 21,301 genes. DLBCL were grouped as described for Table S1.

Table S5: Frequency of heterozygous losses of genes in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing percentage of DLBCL with loss of one allele in 21,124 genes. DLBCL were grouped as described for Table S1.

Table S6: Frequency of homozygous losses of genes in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing percentage of DLBCL with loss of both alleles in 13,343 genes. DLBCL were grouped as described for Table S1.

Table S7: Frequency of individual mutations in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing mutation frequency of 54,182 individual mutations in DLBCL subgroups. Mutation type indicates predicted mutation categories (missense, truncating (TRUNC), or in-frame deletions/insertions (INFRAME). Study specific Mutation IDs were assigned to individual mutations. DLBCL were grouped as described for Table S1.

Table S8: Frequency of individual subclonal mutations in DLBCL subtypes (Included in a separate Excel file)
Excel spreadsheet showing mutation frequency of 1314 subclonal (> 2% allele frequency) individual mutations in DLBCL subgroups. DLBCL were grouped as described for Table S1.

Table S9: Characteristics of DLBCL patients (Included in a separate Excel file)
Excel spreadsheet showing the patient characteristics of DLBCL included in the study. NA: not available IPI range groups: 0-1 (low); 2-3 (intermediate); 4-5 (high). If some IPI components are missing and the range of the IPI would fall between these groups, this parameter is left blank and the case was not used for the IPI analysis. Chemoimmunotherapy includes CHOP or CHOP-like chemotherapy plus Rituximab.

Table S10: Statistical analysis addition DLBCL subtype distinction to International Prognostic Index (IPI) model (Included in a separate Excel file)
Excel spreadsheet showing results of statistical analysis adding genetic subtype distinction to the IPI groups. IPI score was treated as a categorical variable taking on three values: Low for IPI=(0,1), Intermediate for IPI = (2,3), High for IPI =(4,5).

Table S11: HaloPlex Design (Included in a separate Excel file)
Excel spreadsheet showing HaloPlex Design for deep amplicon sequencing. Features of Haloplex design were as follows: H. sapiens, hg19, GRCh37, 602 targets comprising 5643 regions, region size: 2.549 Mbp, 137864 total amplicons, total target bases analyzable: 2.48 Mbp, total sequence design: 4.10 Mbp, target coverage: 97.25 %, databases: RefSeq, Gencode, regions including coding exons, 5' UTR, and 3' UTR, region extension: 50 bases from 3' end and 50 bases from 5' end. (TargetID: Gene symbol or chromosomal region. Interval: The genomic interval of the target. Regions: The number of regions within this target. Size: The total size (in base pairs) of the regions. Database(s): The databases in which this target was found. High Coverage: Number of regions where analyzable amplicon overlap >= 90%. Low Coverage: Number of regions where analyzable amplicon overlap < 90%).

Table S12: Prediction values aberrant somatic hypermutation in DLBCL (Included in a separate Excel file)
Excel spreadsheet showing probability scores from support vector machine (SVM) and random forest (RF) models predicting aberrant somatic hypermutations (aSHM). Only those genes with at least 40 mutations across all ExonSeq sample are listed in this worksheet. Genes are ranked by decreasing SVM score. Training genes are listed without a score (N/A). Previously defined aSHM genes were used as training positives; genes that were not previously defined as targets of aSHM used in training were pseudo-negatives. Probability scores highlighted in yellow predicting aSHM. Only genes with ≥ 30 mutations are shown.

Table S13: Genetic features included in subtype predictors (Included in a separate Excel file)
Excel spreadsheet listing the genetic aberrations that were included in two genetic classification methods. One method utilizes translocations of BCL2 and BCL6, mutations, amplifications and homozygous deletions. The second utilizes only translocations of BCL2 and BCL6, and mutations. Feature descriptions are presented above in the “Genomic feature definition” section (p. 19). Genomic coordinates according to hg19.

Table S14. Summary characteristics of the samples used for predictive model for somatic mutations in DLBCL

Characteristics of the 46 DLBCL TCGA samples used in the training/holdout test as well as the 23 samples used in the cross-validation tests.

Characteristic	Cross-validation data	All DLBCL data
# somatic calls (MuTect2)	2263 (78%)	5156 (82%)
# germline calls (MuTect2)	650 (22%)	1146 (18%)
Mean reference depth	124.8	118.6
Std. dev., reference depth	117.3	109.6
Duplicates across samples	451	888
Unique calls containing duplicates	147	236
Total unique calls	2609	5650
# samples	23	46

Table S15. List of annotation features used to create a Random Forest model of somatic mutations in DLBCL.

1 SIFT_score	9 MutationTaster_score	17 LR_score
2 SIFT_pred	10 MutationTaster_pred	18 LR_pred
3 Polyphen2_HDIV_score	11 MutationAssessor_score	19 VEST3_score
4 Polyphen2_HDIV_pred	12 MutationAssessor_pred	20 CADD_raw
5 Polyphen2_HVAR_score	13 FATHMM_score	21 CADD_phred
6 Polyphen2_HVAR_pred	14 FATHMM_pred	22 GERP++_RS
7 LRT_score	15 RadialSVM_score	23 phyloP46way_placental
8 LRT_pred	16 RadialSVM_pred	24 phyloP100way_vertebrate
		25 SiPhy_29way_logOdds

Table S16. Performance of the leave-one out cross validation samples used in training the RF model

Classification performance of the 23-fold, leave-one (sample) out cross validation on the 23 samples used in training the RF model.

		MuTect2 call	
		Somatic	Germline
RF model prediction	Somatic	1957	134
	Germline	306	516

AUC	Accuracy	Sensitivity	Specificity	PPV	F1-score
0.90	0.85	0.87	0.79	0.94	0.90

Table S17. Holdout testing performance of the RF model on 23 TCGA DLBCL samples not used in training.

		MuTect2 call	
		Somatic	Germline
RF Prediction	Somatic	2751	96
	Germline	142	400

AUC	Accuracy	Sensitivity	Specificity	Pos. Pred. Val.	F1-score
0.93	0.93	0.95	0.81	0.97	0.96

Table S18. Application of the RF model to predict somatic variants on the full set of filtered DLBCL variants used in the current study

(A) – Prediction on all variants; (B) – prediction on those donor samples downloaded from TCGA.

(A)	RF prediction, all missense variants		
DLBCL subtype	Somatic	Germline	% Somatic
ABC	1291	16869	93
GCB	381	9135	96
Unclassified	592	7122	92
Overall	2264	33126	94

(B)	RF prediction, TCGA missense variants		
DLBCL subtype	Somatic	Germline	% Somatic
ABC	49	643	93
GCB	69	1296	95
Unclassified	23	306	93
Overall	141	2245	94

Table S19. Leave-one-out cross validation results of models for predicting somatic hypermutations

Model	Accuracy	PPV	Sensitivity	Specificity	F1
RF	0.89	0.96	0.81	0.97	0.88
SVM	0.83	0.92	0.72	0.94	0.81

Table S20. Known (shaded and in italic) and predicted hypermutation genes in DLBCL.

BACH2	*LTB*	ACTB	HIST1H2AB	HIST1H3H	KLHL21	TAS1R1
BCL2	*MYC*	ACTG1	HIST1H2AD	HIST2H2AA4	LIMD2	TNF
BCL6	*NCOA3*	ARID5B	HIST1H2AE	HIST2H2AB	LST1	TNFRSF14
BCL7A	*P2RY8*	ATXN2	HIST1H2AG	HIST2H2AC	LTA	UBE3C
BTG1	*PAX5*	C1orf167	HIST1H2AH	HIST2H2BE	MAP3K3	VMP1
BTG2	*PIM1*	CD44	HIST1H2AI	HIST2H2BF	MCL1	WEE1
CD74	*POU2AF1*	CDKN1B	HIST1H2AL	HIST2H3A	MPEG1	ZFP36L2
CD83	*SGK1*	DTX4	HIST1H2AM	HIST2H3C	NOL9	ZNF608
CIITA	*SOCS1*	EGR1	HIST1H2BC	HIST2H3D	OSBPL10	ZNF804A
CXCR4	*TCL1A*	EHD1	HIST1H2BD	HIST4H4	PIM2	ZNF860
DMD	*TMSB4X*	EIF4A2	HIST1H2BF	HLA-B	PPP1R9B	
DTX1	*ZFP36L1*	ETV6	HIST1H2BG	ID3	PRAMEF25	
DUSP2	*RHOH*	FAM102A	HIST1H2BJ	IGLL5-RSPH14	PRAMEF26	
ETS1	*BIRC3*	FCRL3	HIST1H2BK	IL10RA	PRAMEF7	
GADD45B	*SERPINA9*	FOXC1	HIST1H2BL	IL16	PRAMEF8	
GRHPR	*MS4A1*	FOXO1	HIST1H2BO	IRF2BP2	PRRT2	
HIST1H2AC	*S1PR2*	H2AFJ	HIST1H3B	ITPKB	RCC1	
IRF4	*ST6GAL1*	HIST1H1B	HIST1H3D	KLF2	RFTN1	
IRF8	*SPRED2*	HIST1H1C		KLHL14	RNF144B	
LRMP	*UBE2J1*	HIST1H1E				

References

1. Bruno A, Boisselier B, Labreche K, et al. Mutational analysis of primary central nervous system lymphoma. Oncotarget 2014;5:5065-75.
2. Braggio E, Van Wier S, Ojha J, et al. Genome-wide analysis uncovers novel recurrent alterations in primary central nervous system lymphomas. Clin Cancer Res 2015.
3. Vater I, Montesinos-Rongen M, Schlesner M, et al. The mutational pattern of primary lymphoma of the central nervous system determined by whole-exome sequencing. Leukemia 2015;29:677-85.
4. Chapuy B, Roemer MG, Stewart C, et al. Targetable genetic features of primary testicular and primary central nervous system lymphomas. Blood 2016;127:869-81.
5. Lenz G, Wright G, Dave SS, et al. Stromal gene signatures in large-B-cell lymphomas. N Engl J Med 2008;359:2313-23.
6. Wilson WH, sin-Ho J, Pitcher BN, et al. Phase III randomized Ssudy of R-CHOP versus DA-EPOCH-R and molecular analysis of untreated diffuse large B-Cell lymphoma: CALGB/Alliance 50303. Blood 2016;128:469.
7. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113-20.
8. Scott DW, Wright GW, Williams PM, et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. Blood 2014;123:1214-7.
9. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589-95.
10. 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;20:1297-303.
11. 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;22:568-76.
12. Gao J, Chang MT, Johnsen HC, et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med 2017;9:4.
13. Chang MT, Asthana S, Gao SP, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol 2016;34:155-63.
14. Davis RE, Ngo VN, Lenz G, et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature 2010;463:88-92.
15. Kopan R, Ilagan MX. The canonical Notch signaling pathway: unfolding the activation mechanism. Cell 2009;137:216-33.
16. Puente XS, Bea S, Valdes-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 2015.
17. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.
18. GDC. NCI Genomic Data Commons RNA-Seq pipeline. NCI GDC documentaton 2018; https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/.
19. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166-9.
20. Affymetrix. Copy Number Algorithm with Built-in GC Waviness Correction in Genotyping Console™ Software. Genotyping Console documentation 2009;http://tools.thermofisher.com/content/sfs/brochures/genotyping_console_copynumber_whitepaper.pdf.
21. Seshan VE, Olshen A. DNAcopy: DNA copy number data analysis. R package version 1501 2017.

22. Lam LT, Wright G, Davis RE, et al. Cooperative signaling through the signal transducer and activator of transcription 3 and nuclear factor-{kappa}B pathways in subtypes of diffuse large B-cell lymphoma. Blood 2008;111:3701-13.
23. Lenz G, Wright GW, Emre NC, et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. Proc Natl Acad Sci U S A 2008;105:13520-5.
24. Shaffer AL, Wright G, Yang L, et al. A library of gene expression signatures to illuminate normal and pathological lymphoid biology. Immunol Rev 2006;210:67-85.
25. Nicorici D, Satalan M, Edgren H, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv 2014;011650.
26. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res 2002;12:656-64.
27. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937-47.
28. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 2008;40:1166-74.
29. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213-9.
30. Wang Q, Jia P, Li F, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med 2013;5:91.
31. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers 2010;14:533-7.
32. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Analysis and Machine Intelligence 1998;20:832-44.
33. Liaw A. Documentation for R package randomForest. R package version 1501 2015;https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.
34. Brieman L. Machine learning. Machine Learning 2001;45:5-32.
35. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20:40-9.
36. Khodabakhshi AH, Morin RD, Fejes AP, et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. Oncotarget 2012;3:1308-19.
37. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 2016;534:47-54.
38. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013;3:246-59.
39. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20:273-97.