Supplemental methods

Case accrual

Additional details relating to case accrual can be found online in the standard operating procedures (SOP; https://ocg.cancer.gov/sites/default/files/BLGSP_SOP_manual.pdf).

Cohort

The cases were accrued at the following tissue source sites: Uganda Cancer Institute (UCI, Uganda), Epidemiology of Burkitt's Lymphoma in East-African Children and Minors (EMBLEM, Uganda), Children's Oncology Group (COG, USA) who participated in a clinical trial AALL1131, and St. Jude Children's Research Hospital (USA). Contributing tissue source sites provided documentation for Institutional Review Board approval for the use of tissues submitted for molecular characterization. Clinical data was collected for each case including initial enrollment data and one year and two-year outcome data (details below). The discovery cohort consisted of 91 pediatric BL cases originating from patients aged between two and 20 years. BL subtypes within this cohort included 74 endemic and 17 sporadic pediatric cases (see Table 1 for details). Each BL case had both tumor and matched normal tissue (blood, PBMCs, lymph nodes, etc.), and the tumor was collected prior to any treatment. All cases had a standardized central pathology review by three BL pathologists and confirmed as BL diagnosis (details below). Once the diagnosis was confirmed, the tumor tissue used for molecular characterization was evaluated for tumor nuclei and necrosis (details below). The cases which did not meet the criteria of discovery, lacked matched normal tissue, normal DNA, or the RNA was degraded or essential clinical data was missing, were considered for validation. Validation cases with tumor and normal DNA were ultimately selected for targeted sequencing and

validation tumors with sufficient RNA also underwent RNA sequencing (details below).

Clinical data

The clinical data were collected by Nationwide Children's Hospital (Columbus, OH) from contributing sites after cases were accepted into the discovery or validation cohorts. Follow-up data were then collected for two subsequent years if the patient was alive or not lost to follow-up. The clinical report form, follow-up form, and treatment form can be found within the project standard operating procedures (SOP #303). The following types of clinical information were collected: demographic data (date of birth, sex, race, ethnicity, height, weight, vital status), tumor information (date of diagnosis, tumor anatomic location, tumor status (tumor free/with tumor), stage, lymph node status, history of prior cancers, synchronous cancers and subsequent cancers), HIV status (HIV antibody status, date of diagnosis, CD4 counts, HIV RNA load, CDC HIV risk group, co-infections, prior AIDS defining conditions), infectious disease status (HBV, HCV, H pylori, malaria, EBV), and treatment information (treatment type, tumor response, treatment dates, HAART treatment status). All dates and other personally indefinable information were obfuscated prior to submission to the Office of Cancer Genomics Data Coordinating Center in XML and tab-delimited formats (https://ocg.cancer.gov/programs/cgci/data-matrix).

Consensus pathology review

Unstained slides or FFPE tumor blocks were submitted to the University of Nebraska Medical Center (Omaha, NE) for processing. After review of an hematoxylin and eosin (H&E) slide qualified a case for the study, immunostains were performed on unstained slides or on slides from a tissue microarray (TMA) prepared at the University of Nebraska. The TMAs contained 3

cores from each patient block, with up to 10 patients per TMA. The following immunohistochemical (IHC) stains were performed: CD3, CD10, CD20, Ki-67, BCL2, and BCL6. Samples were also assessed for c-MYC translocation by fluorescence *in situ* hybridization (FISH) using a break-apart probe. Slides were sent to Nationwide Children's Hospital (Columbus, OH), imaged at 40X using an Aperio AT Turbo or Aperio AT2 scanner, and were independently reviewed by three hematopathologists (T.C.G., N.L.H., and E.S.J.) via digital pathology review software. The following criteria were recorded: Burkitt lymphoma, classical or atypical; other histological subtype of lymphoma; or tissue unclassifiable due to technical factors. Cases diagnosed as Burkitt lymphoma were scored further for percent tumor nuclei, percent necrosis, percent other tissue, immunopositivity of the IHC slides (CD3, CD10 > 30%, CD20, Ki67 > 90%, BCL2, and BCL6 > 30%), and positivity of the c-MYC FISH assay (SOP #309-313). Cases without a consensus diagnosis (two or three concordance) were discussed by teleconference to arrive at a final consensus diagnosis. A diagnosis of Burkitt lymphoma, classical or atypical, was required for inclusion in the data analysis.[1] More details are included in the SOP.

Consensus anatomic site classification

Anatomic site classification was performed by consensus review (C.C., T.G.G., E.S.J., and S.M.M.) based on data reported for sites of disease involvement. Many of the African cases did not have assessment of bone marrow, cerebrospinal fluid, or total body imaging. Cases were classified into the following categories: (A) Disseminated disease with no bone marrow (BM) and/or central nervous system (CNS) involvement, documented disease involvement; (B) Head-only, disease involvement of jaw with or without adjacent nodal involvement; (C) Intra-abdominal disease, disease confined to abdominal organs with or without abdominal lymph node

involvement; (D) Disseminated disease, disease involvement on both sides of diaphragm, but no documented BM or CNS involvement; (E) Unknown, insufficient data to classification anatomic involvement.

Sample processing and nucleic acid extraction

Frozen specimens were shipped to and from Nationwide Children's Hospital (Columbus, OH) using a cryoport that maintained an average temperature of less than -180°C (SOP #308). A top and bottom histologic section were cut from tumor and uninvolved tissue (if it was to be used for healthy tissue control) for pathologic quality control review. These were either stained with H&E or Wright-Giemsa and imaged at 40X using an Aperio AT Turbo or Aperio AT2 scanner. Images were reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically consistent with Burkitt lymphoma, and that uninvolved specimens contained no tumor cells. The tumor sections were required to contain a minimum of 50% tumor cell nuclei, and less than 50% necrosis for inclusion in the study. Nearly all samples had less than 20% necrosis.

RNA and DNA were extracted from frozen (SOP #305) and FFPE tumor (SOP #315-316) and normal tissue specimens (mainly blood or granulocytes) using a modification of the DNA/RNA AllPrep kit (Qiagen). Frozen samples were homogenized and applied to a Qiagen DNA column, and FFPE samples were deparaffinized and applied to a Qiagen FFPE DNA column. The flow-through from the Qiagen DNA column was processed using a mirVana miRNA Isolation Kit (Ambion) for frozen tissues, and a High Pure miRNA Kit (Roche) for FFPE tissues. This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA

was extracted from blood using the QiaAmp blood midi kit (Qiagen; SOP #307).

DNA was quantified by PicoGreen assay, and was resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpF/STR Identifiler (Applied Biosystems) was utilized to verify tumor DNA and germline DNA were derived from the same patient. One hundred nanograms of each tumor and normal DNA were sent in duplicate to Qiagen for REPLI-g whole genome amplification using a 100 µg reaction scale. RNA was quantified by measuring Abs260 with a UV spectrophotometer, and integrity was measured using the RNA6000 nano assay (Agilent) to determine the RNA Integrity Number (RIN) for frozen samples or DV200 for FFPE samples.

For inclusion in the discovery set, a tumor needed to pass pathology consensus review (University of Nebraska Medical Center, Omaha, NE) and the specimen pathology quality control review (Nationwide Children's Hospital, Columbus, OH). In addition, a primary tumor and a matched germline (blood, buccal, or uninvolved tissue) sample needed to pass the following metrics: a minimum of 0.7 µg of DNA from frozen or 0.25 µg of DNA from FFPE, and 3 µg RNA from frozen or 1 µg RNA from FFPE. The minimum RNA integrity metrics were a RIN above 7.0 or DV200 above 30. Cases that did not meet these metrics were included in the validation set if there was at least 0.7 µg of DNA from the primary tumor available for DNA sequencing. Tumor RNA sequencing was also performed for validation cases if there was sufficient RNA material.

Library construction and sequencing

Whole genome sequencing of fresh frozen samples

Whole genome sequencing (WGS) libraries were constructed from DNA provided by Nationwide Children's Hospital (Columbus, OH) using a polymerase chain reaction (PCR)-free protocol. To minimize library bias and coverage gaps associated with PCR amplification of high GC or AT-rich regions we have implemented a version of the TruSeq DNA PCR-free kit (E6875-6877B-GSC, New England Biolabs), automated on a Microlab NIMBUS liquid handling robot (Hamilton). Briefly, 500 ng of genomic DNA was arrayed in a 96-well microtitre plate and subjected to shearing by sonication (Covaris LE220). Sheared DNA was end-repaired and size selected using paramagnetic PCRClean DX beads (C-1003-450, Aline Biosciences) targeting a 300-400 bp fraction. After 3' A-tailing, full length TruSeq adapters were ligated. Libraries were purified using paramagnetic (Aline Biosciences) beads. PCR-free genome library concentrations were quantified using a qPCR Library Quantification kit (KAPA, KK4824) prior to sequencing with paired-end 150 base reads on the Illumina HiSeqX platform using V4 chemistry according to manufacturer recommendations.

Whole genome sequencing of formalin-fixed, paraffin-embedded samples

A 96-well library construction protocol was performed from formalin-fixed and paraffin-embedded (FFPE) tissue extracted genomic DNA provided by Nationwide Children's Hospital (Columbus, OH). Since DNA extracted from FFPE tissue will be damaged by the fixation process and prolonged storage in non-ideal conditions, variable DNA quality across the collection is expected with some highly degraded samples. DNA was normalized to 500 ng in a volume of 62 µL elution buffer (Qiagen) and transferred into a microTUBE plate for shearing on

an LE220 (Covaris) acoustic sonicator using the conditions: Duty Factor, 20%; Peak Incident Power, 450W; Cycle per burst, 200; Duration, 2 x 60 seconds with an intervening spin. The profile of sheared FFPE DNA extracted by the Qiagen Allprep DNA/RNA FFPE protocol has a dominant DNA peak in the size range between 300 and 400bp. To improve library quality of FFPE-derived DNA, solid phase reversible immobilization (SPRI) bead-based size selection was performed before library construction to remove smaller DNA fragments from highly degraded FFPE DNAs. If not removed early in the library construction process, these smaller fragments would otherwise dominate the final amplified library. FFPE DNA damage and end-repair and phosphorylation were combined in a single reaction using an enzymatic premix (NEB), then bead purified using a 0.8:1 (bead:sample) ratio to remove small FFPE fragments. Repaired DNA fragments were next A-tailed for ligation to paired-end, partial Illumina sequencing adapters then purified twice with SPRI beads (1:1 ratio). Full-length adaptered products were achieved by performing 8 cycles PCR with primers introducing fault-tolerant hexamer "barcodes" allowing multiplexing of libraries. Indexed PCR products were double purified with 1:1 beads. Concentration of final libraries was determined using size profiles obtained from a high sensitivity Caliper LabChip GX together with Quant-iT (Invitrogen) quantification.

Strand-specific ribosomal RNA depletion RNA sequencing

RNA sequencing (RNA-seq) libraries were constructed from RNA provided by Nationwide Children's Hospital (Columbus, OH) using a strand-specific ribosomal depletion protocol. To remove cytoplasmic and mitochondrial ribosomal RNA (rRNA) species from total RNA NEBNext rRNA Depletion Kit for Human/Mouse/Rat was used (NEB, E6310X). Enzymatic reactions were set-up in a 96-well plate (Thermo Fisher Scientific) on a Microlab NIMBUS liquid handler (Hamilton Robotics, USA). 100ng of DNase I treated total RNA in 6 µL was

hybridized to rRNA probes in a 7.5 µL reaction. Heat-sealed plates were incubated at 95°C for 2 minutes followed by incremental reduction in temperature by 0.1°C per second to 22°C (730 cycles). The rRNA in DNA hybrids were digested using RNase H in a 10 µL reaction incubated in a thermocycler at 37°C for 30 minutes. To remove excess rRNA probes (DNA) and residual genomic DNA contamination, DNase I was added in a total reaction volume of 25 µL and incubated at 37°C for 30 minutes. RNA was purified using RNA MagClean DX beads (Aline Biosciences, USA) with 15 minutes of binding time, 7 minutes clearing on a magnet followed by two 70% ethanol washes, 5 minutes to air dry the RNA pellet and elution in 36 uL DEPC water. The plate containing RNA was stored at -80°C prior to cDNA synthesis.

First-strand cDNA was synthesized from the purified RNA (minus rRNA) using the Maxima H Minus First Strand cDNA Synthesis kit (Thermo-Fisher, USA) and random hexamer primers at a concentration of 8ng/µL along with a final concentration of 0.4µg/µL Actinomycin D, followed by PCR Clean DX bead purification on a Microlab NIMBUS robot (Hamilton Robotics, USA). The second strand cDNA was synthesized following the NEBNext Ultra Directional Second Strand cDNA Synthesis protocol (NEB) that incorporates dUTP in the dNTP mix, allowing the second strand to be digested using USERTM enzyme (NEB) in the post-adapter ligation reaction and thus achieving strand specificity.

cDNA was fragmented by Covaris LE220 sonication for 130 seconds (2 x 65 seconds) at a “Duty cycle” of 30%, 450W Peak Incident Power and 200 Cycles per Burst in a 96-well microTUBE Plate (P/N: 520078) to achieve 200-250 bp average fragment lengths. The paired-end sequencing library was prepared following the BC Cancer Agency Genome Sciences Centre strand-specific, plate-based library construction protocol on a Microlab NIMBUS robot (Hamilton Robotics,

USA). Briefly, the sheared cDNA was subject to end-repair and phosphorylation in a single reaction using an enzyme premix (NEB) containing T4 DNA polymerase, Klenow DNA Polymerase and T4 polynucleotide kinase, incubated at 20°C for 30 minutes. Repaired cDNA was purified in 96-well format using PCR Clean DX beads (Aline Biosciences, USA), and 3' A-tailed (adenylation) using Klenow fragment (3' to 5' exo minus) and incubation at 37°C for 30 minutes prior to enzyme heat inactivation. Illumina PE adapters were ligated at 20°C for 15 minutes. The adapter-ligated products were purified using PCR Clean DX beads, then digested with USERTM enzyme (1 U/µL, NEB) at 37°C for 15 minutes followed immediately by 13 cycles of indexed PCR using Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) and Illumina's PE primer set. PCR parameters: 98°C for 1 minute followed by 13 cycles of 98°C 15 seconds, 65°C 30 seconds and 72°C 30 seconds, and then 72°C 5 minutes. The PCR products were purified and size-selected using a 1:1 PCR Clean DX beads-to-sample ratio (twice), and the eluted DNA quality was assessed with Caliper LabChip GX for DNA samples using the High Sensitivity Assay (PerkinElmer, Inc. USA) and quantified using a Quant-iT dsDNA High Sensitivity Assay Kit on a Qubit fluorometer (Invitrogen) prior to library pooling and size-corrected final molar concentration calculation for Illumina HiSeq2500 sequencing with paired-end 75 base reads.

miRNA sequencing

miRNA sequencing (miRNA-seq) libraries were constructed from 1 µg total RNA provided by Nationwide Children's Hospital (Columbus, OH) using a plate-based protocol developed at the British Columbia Cancer, Genome Sciences Centre (BCGSC). Negative controls were added at three stages: elution buffer was added to one well when the total RNA was loaded onto the plate, water to another well just before ligating the 3' adapter, and PCR brew mix to a final well just

before PCR amplification. A 3' adapter was ligated using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L) with an incubation at 22°C for 1 hour. This adapter is an adenylated, single-stranded DNA with the sequence 5' /5rApp/ ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligates to miRNAs. An RNA 5' adapter was then ligated, using T4 RNA ligase (Ambion USA, cat. AM2141) and ATP, and was incubated at 37°C for 1 hour. The sequence of the single strand RNA adapter is 5'GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3'.

Upon completion of adapter ligation, 1st strand cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014) and RT primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'). First-strand cDNA provided the template for the final library PCR, into which we introduce index sequences to enable libraries to be identified from a sequenced pool that contains multiple libraries. Briefly, a PCR brew mix was made with the 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat. F-540L), buffer, dNTPs and DMSO. The mix was distributed evenly into a new 96-well plate. A Microlab NIMBUS robot (Hamilton Robotics, USA) was used to transfer the PCR template (1st strand cDNA) and indexed 5' PCR primers into the brew mix plate. Each indexed 5' PCR primer, 5'-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTCCGA-3', contains a unique six-nucleotide 'index' (shown here as N's), and was added to each well of the 96-well PCR brew plate. PCR was performed at 98°C for 30 seconds, followed by 15 cycles of 98°C for 15 seconds, 62°C for 30 seconds and 72°C for 15 seconds, and finally a 5 minute incubation at 72°C. Library qualities were assessed across the whole plate using a Caliper LabChipGX DNA chip. PCR products were pooled and size selected to remove larger cDNA

fragments and smaller adapter contaminants, using a 96-channel automated size selection robot that was developed at the BCGSC. After size selection, each pool was ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool was diluted to a target concentration for cluster generation and loaded into a single lane of an Illumina HiSeq2500 flow cell. Clusters were generated, and lanes were sequenced with a 31-bp main read for the insert and a 7-bp read for the index.

Targeted sequencing by custom hybridization capture

Targeted sequencing libraries were constructed from DNA provided by Nationwide Children's Hospital (Columbus, OH) using a custom hybridization capture protocol. 50 ng from each of 20 or 21 whole genome libraries was pooled prior to custom capture using Agilent SureSelect XT Custom probes (4.8 Mbp) targeting 74,809 human and EBV features (https://cgci-data.nci.nih.gov/Public/BLGSP/targeted_capture_sequencing/DESIGN/). The features included the following: exons of recurrently mutated genes with the exception of known targets of passenger mutations (*e.g. TTN*, mucin genes); exons of several known diffuse large B-cell lymphoma (DLBCL) genes; exons of previously reported BL genes not found mutated in our data; whole gene bodies for *DDX3X* (GRCh38 chrX:41332775-41364961) and *FBXO11* (chr2:47782639-47907718); whole gene bodies and flanking regions for *ID3* (chr1:23557918-23657826) and *BCL6* (chr3:187718649-188265924); the recurrently rearranged region surrounding *MYC* (chr8:127242368-129788153); and non-coding mutation peaks (details below). The pooled libraries were hybridized to the RNA probes at 65°C for 24 hours. Following hybridization, streptavidin-coated magnetic beads (Dynal, MyOne) were used for custom capture. Post-capture material was purified on MinElute columns (Qiagen) followed by post-

capture enrichment with 10 cycles of PCR using primers that maintain the library-specific indices. Pooled libraries were sequenced on an Illumina HiSeq 2500 instruments with v4 chemistry generating 125 base paired-end reads.

Data analysis

Sequencing read alignment

Whole genome sequencing (WGS) and targeted sequencing reads were aligned to the human reference genome (GRCh38) with BWA-MEM (version 0.7.6a; parameters: -M).[2,3] The human reference genome that was used is a version of GRCh38 without alternate contigs that includes the Epstein–Barr viral genome (GenBank accession AJ507799.2), which can be downloaded at http://www.bcgsc.ca/downloads/genomes/9606/hg38_no_alt/bwa_0.7.6a_ind/genome/. Read duplicate marking was done using sambamba (version 0.5.5).[4] The WGS read alignments for the discovery tumor and matched normal had an average non-redundant depth of 82X (range 55-96) and 41X (range 30-51), respectively. The validation tumor and normal samples were sequenced to a higher average depth, namely 243X (range 158-392). RNA sequencing (RNA-seq) reads (mean 200M reads; range 100-289M) were pseudo-aligned using Salmon (version 0.8.2; details below). The RNA-seq reads were also aligned to the reference genome indicated above using the JAGuaR pipeline.[7] miRNA sequencing (miRNA-seq) reads (mean 13M reads; range 1.8-35M) were aligned to the same human reference genome with BWA-SW (version 0.5.7).[8] Tumor and matched normal WGS data for 15 cases from the International Cancer Genomic Consortium (ICGC) were obtained through a DACO-approved project using a virtual instance on the Cancer Genome Collaboratory.[5,6] The ICGC WGS reads were re-aligned using the above parameters, yielding alignments with an average depth of 40X (range 29-62).

Tumor EBV status and genome type

Tumor EBV status and genome type was inferred from tumor whole genome sequencing (WGS) and RNA sequencing (RNA-seq) data. To determine tumor EBV status, the fraction of reads aligning to the EBV genome was calculated using Samtools (version 1.6).[2] Tumors were considered to be EBV-positive when the EBV fraction of WGS reads was greater than 0.00006 (calculated from the fraction represented by the EBV genome in the reference genome) and the number of RNA-seq reads mapped to the *EBER1* (chrEBV:6629-6795) and *EBER2* (chrEBV:6956-7128) loci in the JAGuaR-based alignments was greater than 250. There were no cases with discordant EBV statuses inferred from the WGS and RNA-seq data. Although EBER expression was not quantified for the ICGC tumors because their RNA-seq data were not used in this project, they were all classified as EBV-negative according to their WGS data, which is consistent with the EBV status reported by the MMML-seq project. The minimum fraction of EBV reads was 0.01 for samples that underwent targeted sequencing to account for the different ratio of human and EBV genomic regions due to hybridization capture. EBV genome type was inferred for EBV-positive tumors by comparing the counts for 21-mers that are unique to either EBV type 1 (GenBank accession NC_007605.1) or type 2 (GenBank accession NC_009334.1). K-mer counting was performing on tumor WGS reads aligned to the EBV genome using Jellyfish (version 2.2.6).[9] EBV genome type was inferred to be type 1 or type 2 if the count ratio of EBV type 1–specific k-mers to EBV type 2–specific k-mers was greater than or lesser than 1, respectively.

Simple somatic mutations

Somatic single nucleotide variants (SNVs) and small insertions/deletions (indels), also known as simple somatic mutations (SSMs), were called from paired tumor-normal WGS data using the Strelka workflow (version 1.0.14).[10] The default Strelka configuration for data aligned with bwa (strelka_config_bwa_default.ini) was used with the exception of filtering SNVs with a minimum QSS of 25 (default 15). For SNVs and indels, reference and alternate allele counts were taken from the Strelka output variant call format (VCF) file.[11] SNVs and indels were annotated using vcf2maf (version 1.6.12) and Ensembl Variant Effect Predictor (release 86).[12] Transcript selection for annotation was performed by vcf2maf with the following exception. Noncanonical transcripts were selected if they had non-synonymous mutations more commonly than the canonical transcript (minimum increase of two affected cases). SNVs and indels were further filtered for a minimum alternate allele count of six and a minimum variant allele fraction (VAF) of 10% and 20% for fresh frozen (FF) and formalin-fixed, paraffin-embedded (FFPE) tumors, respectively. Tumors with a median VAF below 25% were omitted from subsequent analyses due to either excessive noise or low predicted tumor content. The same pipeline was used for detecting SNVs and indels in the targeted validation sequencing data, with the exception that depth filters were disabled for Strelka (isSkipDepthFilters = 1).

Significantly mutated genes

Considering only SNVs and indels, significantly mutated genes were identified using an ensemble approach integrating four methods: MutSigCV,[13] OncodriveFM,[14] OncodriveFML,[15] and OncodriveCLUST.[16] Mutations were lifted over from GRCh38 to GRCh37 using CrossMap (version 0.2.5) along with the "hg38ToHg19" chain file provided by the UCSC Genome

Browser.[17,18] Lifting over variants was necessary because some of the methods listed above rely on GRCh37 reference data. For consistency, the lifted-over mutations based on GRCh37 served as input for all methods. Non-synonymous mutations were defined as those with one of the following values in the MAF file Variant_Classification field, as annotated by vcf2maf: Splice_Site, Nonsense_Mutation, Frame_Shift_Del, Frame_Shift_Ins, Nonstop_Mutation, Translation_Start_Site, In_Frame_Ins, In_Frame_Del, or Missense_Mutation. To minimize noise, we only considered genes deemed significant (Q-value < 0.1) by two or more methods.

BL-associated genes

We defined BL-associated genes (BLGs) as any gene deemed significantly mutated in this study or previously described as recurrently mutated in BL with at least five affected patients in our discovery cohort. Only non-synonymous simple somatic mutations and copy number variations (minimum size 10 kbp) were considered. To avoid considering mainly large-scale events, copy number variations affecting a BLG were required to be relatively small with a median size of 10 Mbp or less. For each BLG, additional cryptic splicing variants (with support for aberrant splicing in RNA-seq data), structural variations, and copy number variations were manually curated.

McNemar's tests

Discordant cases were defined as EBV-negative endemic BL cases and EBV-positive sporadic BL cases. Differentially mutated genes and pathways (referred to here as features) were identified using the following criteria: (1) they must be mutated in at least 10% of cases, and (2) they were differentially mutated between EBV-positive endemic BL cases and EBV-negative

sporadic BL cases (Q-value < 0.1, Fisher's exact test). Discordant cases were excluded from the Fisher's exact tests to ensure that there is no reason to believe *a priori* that the mutation status of these features are preferentially associated with tumor EBV status or clinical variant status. Following that, tumor EBV status and clinical variant status were used as naive predictors of the mutation status of these differentially mutated features and determined whether or not they were correct for each case. The performance of tumor EBV status and clinical variant status as predictors were compared using McNemar's tests. Features with a significant difference according to the McNemar's test (P-value < 0.05) indicate that the "winning" predictor is more strongly associated with the mutation status of said features.

Non-coding mutation peaks

Genomic loci enriched in non-coding mutations, referred to here as "mutation peaks", were identified using the previously described Rainstorm analysis[19] with some adjustments to peak filtering and post-processing. Mutation peaks with a signal-to-noise ratio of 0 were ignored. Following this, peaks with a signal-to-noise ratio below the 95th percentile were also omitted. Mutation peaks were extended on each side by considering the inclusion of nearby mutated positions. Specifically, peaks were extended as follows: (1) for each mutated position within 10 kbp, the peak was extended up to and including said position; (2) for each of the original and extended peaks, the mutation rate (mutations/kbp) was calculated, resulting in a vector of mutation rates equal in length to the number of mutations within 10 kbp + 1 (for the original peak); (3) for each extended peak, the proportional change in mutation rate was calculated by taking the difference in mutation rate with the previous peak (i.e. the peak that omits the outermost mutated position) and dividing this difference by the mutation rate of the previous peak; (4) mutation peaks were extended up to and including the mutated position that preceded the mutated position that led to the largest proportional decrease in mutation rate. Following

peak extension, peaks were only considered if they met the following criteria: at least 5% of cases had mutations in the peak; the mutation rate was greater than or equal to 5 mutations/kbp; and the peak size was smaller than or equal to 20 kbp. Lastly, to remove mutation peaks likely caused by noise, peaks within 1 Mbp of a gap in the reference genome were ignored. Gaps were obtained from the University of California, Santa Cruz (UCSC) Table Browser, consisting of centromeres, heterochromatin regions and short chromosomal arms.[20] Each gap was expanded by 100 kbp on both sides, and gaps within 3 Mbp of one another were merged. The remaining peaks were annotated with the nearest protein-coding, long non-coding RNA or microRNA gene. Peaks that overlapped one of the immunoglobulin heavy or light chain loci, namely *IGH* (chr14:104589639-107810399), *IGK* (chr2:87999518-90599757), or *IGL* (chr22:21031465-23905532), were annotated as such. Finally, peaks were labeled according to the immunoglobulin locus they overlapped or the nearest gene and their position relative to the gene's transcription start site.

P-values were empirically determined for each peak by comparing its mutation rate with an empirical distribution produced by calculating the mutation rates of identically sized regions randomly sampled across the genome. The smallest and largest mutated position on each chromosome were used to determine the range of positions available for sampling with replacement. Positions that overlapped gaps in the reference genome such as centromeres and telomeres were excluded. A "pseudo-peak" was created from a sampled position by extending each side to create regions with the same size as the given mutation peak. The mutation rate of 100,000 such pseudo-peaks was calculated to generate the empirical null distribution of mutation rates genome-wide. The empirical P-value was calculated as the number of pseudo-peaks with a higher mutation rate than the given mutation peak divided by 100,000. Given that each mutation

peak is tested against independent null distributions, the P-values did not require multiple test correction. All peaks were significant with empirical P-values < 0.001.

Enrichment for AICDA-mediated mutations

A custom in-house Python (version 3.6.1) program was used to determine whether certain regions, such as significantly mutated genes and non-coding mutation peaks, were enriched for SNVs and indels that are presumed to be caused by AICDA-mediated mutagenesis.[21,22] Enrichment for putative AICDA-mediated mutations in a given region was measured using two binomial exact tests. First, the observed number of mutations affecting AICDA recognition sites (number of successes), defined as regions that fit the AICDA motif (RGYW), was compared to the expected number of such mutations, which was calculated from the region's mutation rate (probability of success) and the number of bases that overlap AICDA recognition sites (number of trials). Second, the observed number of mutations affecting the guanine-cytosine pair targeted by AICDA in the motif (number of successes) was compared to the expected number of such mutations, which was calculated from the region's mutation rate of guanine-cytosine pairs (probability of success) and the number of target guanine-cytosine pairs in AICDA recognition sites (number of trials). Mutation rates were calculated using the effective region size, which is equal to the product of the region size and the cohort size. The effective region size ensures that the observed number of mutations (number of successes) is never higher than the region size (number of trials). Care was taken to avoid double-counting mutations if they overlapped more than one AICDA recognition site. This process was repeated for all regions of interest. The regions for BL-associated genes were based on transcripts affected by non-synonymous SSMs; transcripts that were not mutated were not considered. The entire regions of non-coding mutation peaks were considered. The in-house program also annotated mutations based on whether they

overlapped an AICDA recognition site.

De novo mutational signatures

Mutational signatures were discovered using the previously described framework by Alexandrov *et al.*.[23] We summarized somatic SNVs based on their mutational subtype, 5' context, and 3' context. This resulted in a mutation catalog matrix of 96 SNV classes for each sample. We performed non-negative matrix factorisation on our mutation catalog to discover mutational signatures within the entire cohort. Signature stability was computed by boostrap resampling over 1000 total iterations (10 iterations in each of 100 cores). The optimal n-signature solution, n_{opt}, which simulataneously maximised signature stability and minimised the Frobenius reconstruction error, was automatically selected,

$$n_{opt} = \mathrm{argmin}_n \left(\frac{R_n - \min(R)}{\max(R) - \min(R)} - \frac{S_n - \min(S)}{\max(S) - \min(S)} \right),$$

where R and S are the vectors containing reconstruction errors and stability of each n-signature solution, and R_n and S_n are the reconstruction error and stability of the n-signature solution. This approach determined that the four-signature solution was optimal. To determine matches to known mutational signatures, cosine similarity metrics were computed against the 30 COSMIC reference mutational signatures. Where more than one signature matched to a single COSMIC signature, the highest similarity match was chosen and the remaining signatures were matched to the next most similar COSMIC signature. For each n-signature solution, the Pearson correlation was calculated between the age at diagnosis for each case and the predicted number of mutations attributable to *de novo* signatures associated with age (COSMIC reference signatures 1 and 5), taking the maximum correlation if both COSMIC signatures were paired. Similarly, for each n-signature solution, the Pearson correlation was calculated between *AICDA* expression for each

case and the predicted number of mutations attributable to the *de novo* signature associated with AICDA activity (COSMIC reference signature 9).

Somatic structural variations

Somatic structural variations (SVs) were detected using the Manta pipeline (version 1.1.0) in paired tumor-normal mode using default parameters with the exception of a minimum somatic score (SOMATICSCORE) of 45 (default 30).[24] In FFPE samples, any inversions smaller than 500 bp were considered noise and ignored. Variant allele fractions were calculated from the reference and alternate allele counts reported in the Manta output VCF file. VCF files were converted to BEDPE format using the svtools vcftobedpe tool (version 0.3.2, commit 6d7b6ec8).[25] SVs that overlapped any of the significantly mutated genes were manually curated for inclusion as non-synonymous mutations. IG-*MYC* translocations were identified as being any SV that met the following conditions: (1) one breakpoint was near *MYC* (chr8:126393182-130762146); (2) the breakpoint near *MYC* was oriented such that exons 2 and 3 are included in the rearrangement; (3) the other breakpoint was near an immunoglobulin heavy or light chain locus, namely *IGH* (chr14:104589639-107810399), *IGK* (chr2:87999518-90599757), or *IGL* (chr22:21031465-23905532); and (4) the highest-scoring translocation was selected in the event of multiple candidate SVs. Tumors where Manta failed to detect a translocation that met the above criteria were manually inspected for such events, which successfully revealed IG-*MYC* rearrangements in the remaining cases.

Somatic copy number variations

Sequenza was used to call somatic copy number variations (CNVs) in tumor-normal pairs.[26] Sequenza bam2seqz (parameters: –qlimit 30) generated the SEQZ files, which were then binned using Sequenza seqz-binning (parameters: -w 300 -s). To eliminate noise, germline heterozygous positions were filtered for any that overlap dbSNP (downloaded 2017-04-03) "common all" single nucleotide polymorphisms (SNPs). Using bedtools intersect (parameters: -wa),[27] germline heterozygous positions were removed if they overlapped gaps in the reference genome (*e.g.* centromeres) or segmental duplications, which were obtained from the UCSC Table Browser.[20] Previously, the segmental duplications were merged if they overlapped one another using bedtools merge, then filtered for a minimum size of 10 kbp, and subsequently merged again using bedtools merge (parameters: -d 10000). The Sequenza R package was used to load the binned SEQZ data, fit a model for cellularity and ploidy, and generate CNV segments.[26] Sequenza was made aware of the sex of each case to properly handle CNVs on the sex chromosomes. To simplify model fitting and avoid incorrect local optima, ploidy and cellularity options were restricted as follows. Ploidy was limited to the range between 1.8 and 2.5. Cellularity was restricted to an estimate of tumor content derived from the variant allele fraction (VAF) of SNVs and indels, defined as twice the VAF corresponding to the first local density maximum below 50%.

Gene expression quantification

The tximport Bioconductor R package was used to summarize transcript-level read counts at the gene level.[28] The DESeq2 Bioconductor R package was used to correct the read counts for library size and to perform a variance-stabilizing data transformation.[29] These variance-stabilized

expression values were used for statistical tests that require homoskedastic data.

miRNA expression profiling was performed separately on the miRNA sequencing data using Canada's Michael Smith Genome Sciences Centre miRNA processing pipeline, which was used for The Cancer Genome Atlas project.[30] The analysis was done using miRBase release 21.[31–35]

Clonal B-cell receptors

MiXCR (version 2.1.3) was used to identify immunoglobulin heavy and light chain clones from the RNA-seq and WGS data as per the standard pipeline described in their documentation.[36,37] The MiXCR pipeline was also run on 323 DLBCL tumor samples that underwent a strand-specific poly[A]-selection RNA-seq protocol.[Ennishi_undated-wu] All RNA-seq reads were aligned using "mixcr align" (parameters: -p rna-seq -OallowPartialAlignments=true) while for the WGS data, only reads originating from the immunoglobulin regions (chr2:88668078-90584447, chr14:105548159-107030529, and chr22:21897318-23046831) or unmapped reads were aligned using "mixcr align" (parameters: -p rna-seq -OallowPartialAlignments=true -OvParameters.geneFeatureToAlign=VGeneWithP). Two rounds of contig assembly was performed using "mixcr assemblePartial" followed by clone assembly using "mixcr assemble". Clones were exported using "mixcr exportClones" (parameters: -o -t) options to exclude any clones with out-of-frame sequences or stop codons. Clonal fraction was calculated for heavy and light chains separately. Dominant clones in the RNA-seq data were defined as having a clonal fraction of at least 30% with a minimum of 30 supporting reads. For the WGS analysis, dominant clones were defined as having the greatest clonal fraction with at least two supporting reads. The top-scoring V, D, J and C genes were selected for each clone when multiple genes were possible.

Data and statistical analyses

Data and statistical analyses were done using the R statistical programming language (version 3.4.2).[38] Mann–Whitney *U* tests and Fisher's exact tests were used where appropriate with the wilcox.test and fisher.test functions in R, respectively. McNemar's tests were done using the mcnemar.test function in R. Correlation between continuous variables was tested using Pearson's product-moment correlation coefficient with the cor.test function in R. Linear regressions were performed using the lm function in R and bootstrapped 10,000 times to calculate bootstrap 95% confidence intervals using the boot and boot.ci functions in R (adjusted bootstrap percentile interval). Mutual exclusivity between mutations in different genes was evaluated using the CoMEt exact test with the comet_exact_test function from the cometExactTest package.[39,40] Multiple hypothesis correction was performed using the Benjamini–Hochberg (BH) method with the p.adjust function in R. P-values below 5% and Q-values (FDR adjusted P-values) below 10% were considered significant. Significantly used R packages are listed below with their respective versions and citations.

Package	Version	Citations
argparse	1.1.1	41
bedr	1.0.4	42
biomaRt	2.32.1	43,44
bookdown	0.7	45,46
broom	0.4.3	47
circlize	0.4.1	48
cometExactTest	0.1.5	40
cowplot	0.9.3	49
data.table	1.11.4	50
DESeq2	1.16.1	51
dplyr	0.7.4	52
feather	0.3.1	53
flextable	0.4.4	54
forcats	0.2.0	55
GenomicRanges	1.28.6	56
ggbeeswarm	0.6.0	57
ggExtra	0.8	58
ggplot2	3.1.0	59
ggrepel	0.7.0	60
ggsignif	0.4.0	61
ggstance	0.3	62
Gviz	1.20.0	63
knitr	1.20	64–66
lsa	0.73.1	67
maftools	1.4.20	68
MassSpecWavelet	1.42.0	69
matrixStats	0.53.0	70
pheatmap	1.0.8	71
Publish	2018.04.17	72
purrr	0.2.5	73
RColorBrewer	1.1-2	74
readr	1.1.1	75
readxl	1.0.0	76
robustbase	0.92-7	77,78
sequenza	2.1.2	26
tidyverse	1.1.1	79
tximport	1.4.0	80
viridis	0.4.1	81

Supplemental table legends

Supplemental Table S1. Summary of clinical and molecular characteristics of the validation cohort. All cases were HIV-negative.

Supplemental Table S2. Patient metadata. Clinical and molecular characteristics of the discovery and validation cohorts. The metadata for the ICGC cases are not re-published here.

Supplemental Table S3. Simple somatic mutations in the discovery cohort. The mutations are restricted to exonic and splice regions. Unless a mutation affected a BL-associated gene and was non-synonymous, we excluded all mutations with a minor allele fraction greater than 0.0001 according to dbSNP or ExAC.[82,83] With the exception of the first two columns, this table follows The Cancer Genome Atlas (TCGA) Mutation Annotation Format (MAF).

Supplemental Table S4. Simple somatic mutations in the validation cohort. This table follows the same criteria as Supplemental Table S3.

Supplemental Table S5. Somatic copy number variations in the discovery cohort. With the exception of the first two columns, this table follows the segments output format by Sequenza.[26]

Supplemental Table S6. Somatic structural variations in the discovery cohort. With the exception of the first two columns, this table follows the BEDPE output format by the svtools vcftobedpe tool, which converted Manta VCF files.[25]

Supplemental Table S7. Non-coding mutation peaks.

Supplemental Table S8. Significantly mutated genes. This table shows the methods that identified each gene as significantly mutated (represented by 1) or not (represented by 0).

Supplemental Table S9. Mutation status for BL-associated genes (BLGs) and pathways. This table considers all mutations types as displayed in Figure 4 (minus the ICGC cases).

Supplemental Table S10. Fisher's exact tests on mutation prevalence. This table contains the underlying counts of mutated and unmutated cases that were used in comparing the mutation prevalence between disease subtypes (*i.e.* tumor EBV status, clinical variant status, and EBV genome type).

Supplemental Table S11. McNemar's test results. This table compares tumor EBV status and clinical variant status in their ability to predict the mutation status of genes/pathways that are differentially mutated between EBV-positive eBLs and EBV-negative sBLs (*i.e.* excluding discordant cases). The McNemar's test P-value indicates whether there is a significant difference in the predictive performance of tumor EBV status and clinical variant status.

Supplemental Table S12. Linear regression of mutational signatures. Linear regression of the estimated number of mutations per signature as a function of various covariates. Tumor EBV status and clinical variant status were used as covariates in all models, age was used as a covariate for BL signature A given its association with age, and *AICDA* expression was used as a

covariate for BL signatures B, C, and D. The linear models were also bootstrapped 10,000 times to calculate bootstrap 95% confidence intervals (CI).

Supplemental figure legends

Supplemental Figure S1. Features of EBV-positive BL tumors. (A) Fraction of mapped reads from whole genome sequencing data that aligned to the EBV genome (log scale). The minimum threshold for calling EBV-positive samples was 0.006, indicated by the dashed line. (B) RNA-seq read counts for *EBER1* and *EBER2* (log scale). The minimum count for calling EBV-positive samples was 250 reads, indicated by the dashed line. A pseudocount of 1 was added to all values prior to log transformation. This excludes the ICGC cases whose RNA-seq data were not analyzed. (C) Ratio between the counts for 21-mers that are unique to EBV type 1 and type 2, respectively, calculated from whole genome sequencing reads aligned to the EBV genome. The minimum ratio for calling EBV type 1 samples was 1, indicated by the dashed line. (D) Variance-stabilized expression values for EBV latency genes in EBV-positive and EBV-negative tumors.

Supplemental Figure S2. Clonal B-cell receptors in BL. (A) Clonal fraction estimates of immunoglobulin heavy and light chain clones. Clonal rearrangements (shown in red) must have a minimum clonal fraction of 30% (indicated by dashed line) and at least 30 supporting reads. The samples are ordered consistently along the X axis for each immunoglobulin chain. (B) Percentage of EBV-positive and EBV-negative BL tumors (N = 91) that use the given immunoglobulin V genes to encode their most clonal B-cell receptors (*i.e.* with the highest clonal fraction). Only showing V genes appearing in Figure 1B. (C) Total read count per sample supporting heavy and light IG chain clones according to whether a clonal BCR was detected. Significance brackets: **, P-value < 0.001; ***, P-value < 0.00001 (Mann–Whitney *U* test).

Supplemental Figure S3. Non-coding mutation peaks in BL. (A) Size distribution of non-coding mutation peaks (or simply, "peaks"). (B) Distance between peaks and the respective nearest transcription start sites (TSS). TSS-proximal and TSS-distal peaks are defined as those within and beyond 3000 bp of a TSS, respectively, and are shown separately here. Peaks overlapping immunoglobulin loci are omitted. (C) Number of mutations in peaks according to tumor EBV status and clinical variant status. P-values were calculated using Mann–Whitney *U* tests. (D) Enrichment or depletion of mutations affecting AICDA recognition sites (RGYW) in peaks altered in at least 15 cases. The X-axis displays the odds ratio between the observed and expected mutation rates of all bases in AICDA recognition sites. The Y-axis shows the odds ratio between the observed and expected mutation rates of guanine-cytosine pairs in AICDA recognition sites. Peaks with a significant enrichment according to either metric are displayed in red (Q-values < 0.1, binomial test). (E) Percentile of peak target gene expression (TPM). miRNA genes are not shown. (F) Variance-stabilized expression values of genes associated with TSS-proximal peaks according to the mutation status of each peak. Only protein-coding genes are displayed. For each gene, expression values were normalized by the median expression in unmutated tumors. Significance brackets: *, Q-value < 0.1; **, Q-value < 0.001 (Mann–Whitney *U* test). (G) Pearson's product-moment correlation between variance-stabilized *AICDA* expression and the number of mutations in non-coding mutation peaks. (H) Variance-stabilized *AICDA* expression in sporadic and endemic BL according to tumor EBV status. Significance brackets (panels C and H): *, P-value < 0.05; ***, P-value < 0.00001 (Mann–Whitney *U* test).

Supplemental Figure S4. Non-synonymous mutations in BL-associated genes. (A) Enrichment or depletion of mutations affecting AICDA recognition sites (RGYW) in BL-associated genes (BLGs). The X-axis displays the odds ratio between the observed and expected

mutation rates of all bases in AICDA recognition sites. The Y-axis shows the odds ratio between the observed and expected mutation rates of guanine-cytosine pairs in AICDA recognition sites. BLGs with a significant enrichment or depletion according to either metric are displayed in red and blue, respectively (Q-values < 0.1, binomial test). (B) Mutual exclusivity of mutations affecting BLGs associated with each pathway (Comet exact test followed by Benjamini–Hochberg correction). The dashed line represents the minimum Q-value threshold (0.1). (C) Differential incidence of mutations in BLGs between BL subtypes (Fisher's exact test followed by Benjamini-Hochberg correction). Significant differences are highlighted in red (Q-values < 0.1, indicated by dashed lines).

Supplemental Figure S5. Patterns of non-synonymous mutations in BL-associated genes in BL and DLBCL. Lolliplot plots displaying non-synonymous simple somatic mutations in BL-associated genes. Mutations detected in BL (N = 106 cases) and DLBCL (N = 153 cases) genomes are shown above and below the gene model, respectively.

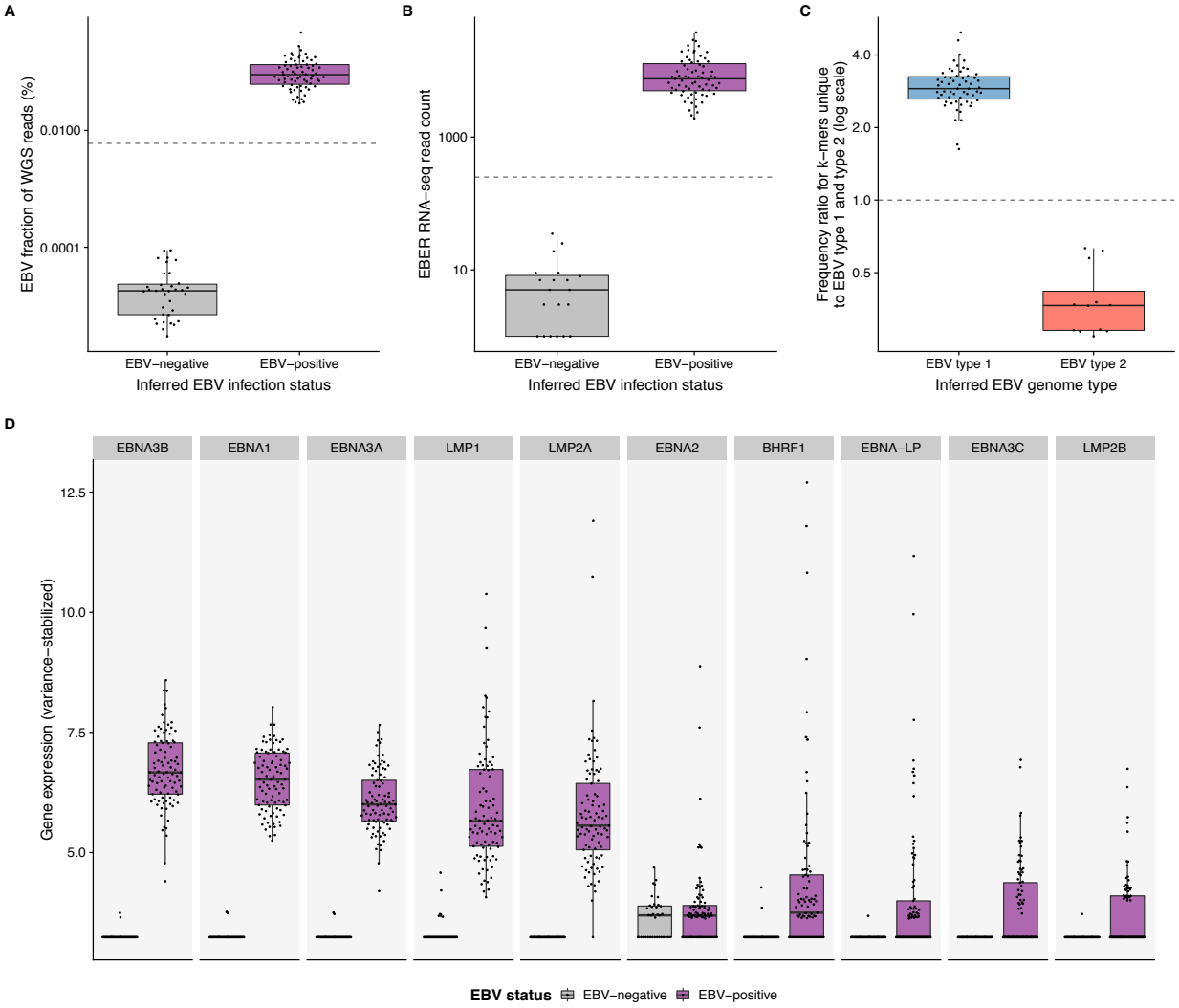
Supplemental Figure S6. Landscape of copy number variations in BL. Proportion of cohort affected by copy number gains and losses are shown in red and blue, respectively. CNVs that are smaller than 100 kbp are not displayed.

Supplemental Figure S7. *De novo* mutational signatures in BL. (A) Selecting the optimal number of *de novo* mutational signatures (shown in red) by minimizing reconstruction error and maximizing stability. (B) After pairing each *de novo* mutational signature ("BL signature") with COSMIC reference signatures, Pearson's product-moment correlations were calculated between the exposure of BL signatures associated with COSMIC signatures 1 or 5 and age at diagnosis,

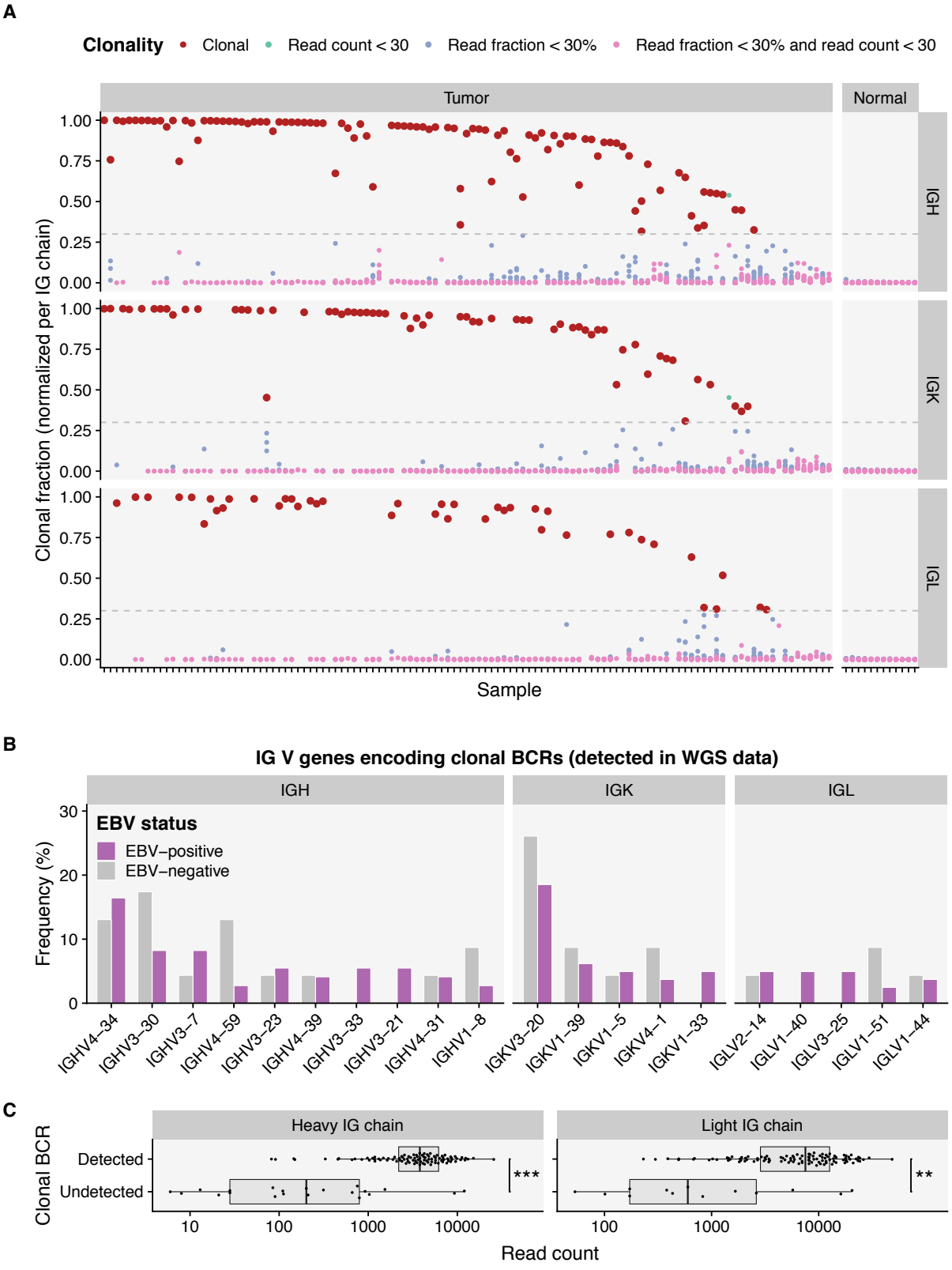
and between the exposure of BL signatures associated with COSMIC signature 9 and tumor *AICDA* expression. (C) Pearson's product-moment correlation between age at diagnosis and the estimated number of mutations per signature from the optimal solution. The dashed line represents the minimum Q-value threshold (0.1). (D) Pearson's product-moment correlation between *AICDA* expression and the estimated number of mutations per signature from the optimal solution. The dashed line represents the minimum Q-value threshold (0.1). (E) Cosine similarity between BL signatures from the optimal solution and COSMIC reference signatures. Pairs between BL signatures and COSMIC signatures are outlined in red. (F) Composition of each BL signature per base change and trinucleotide context.

Supplemental figures

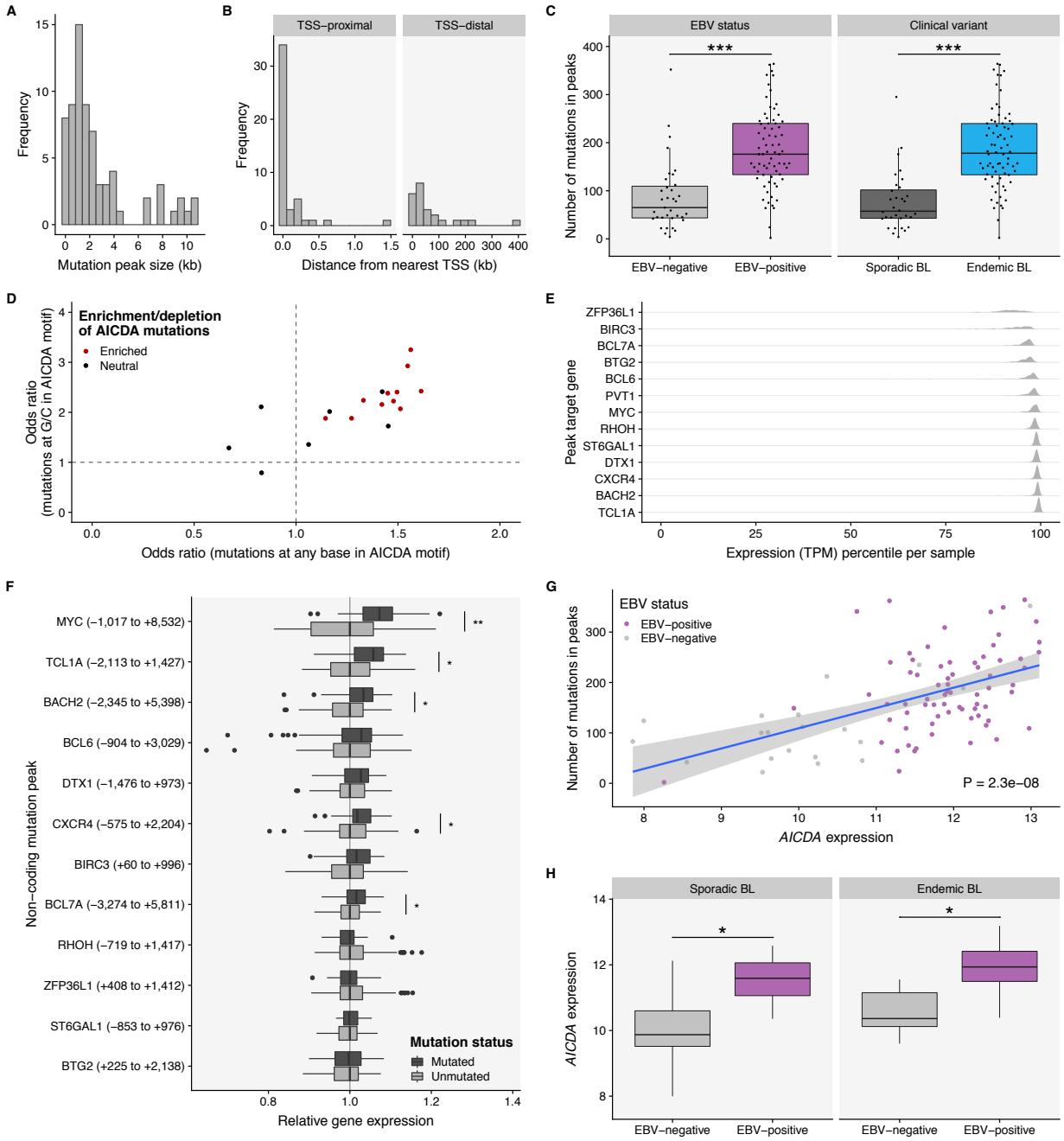
Supplemental Figure S1



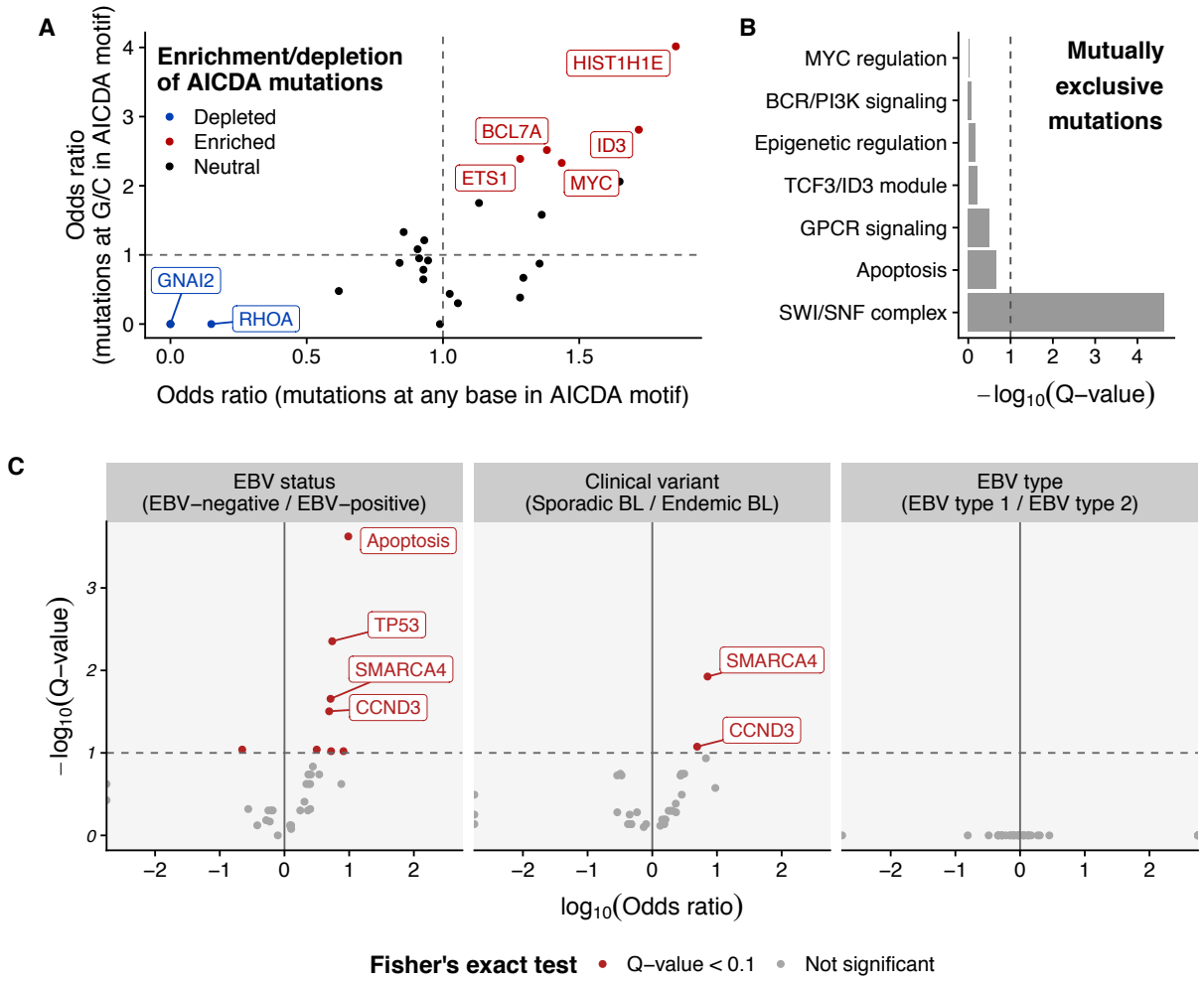
Supplemental Figure S2



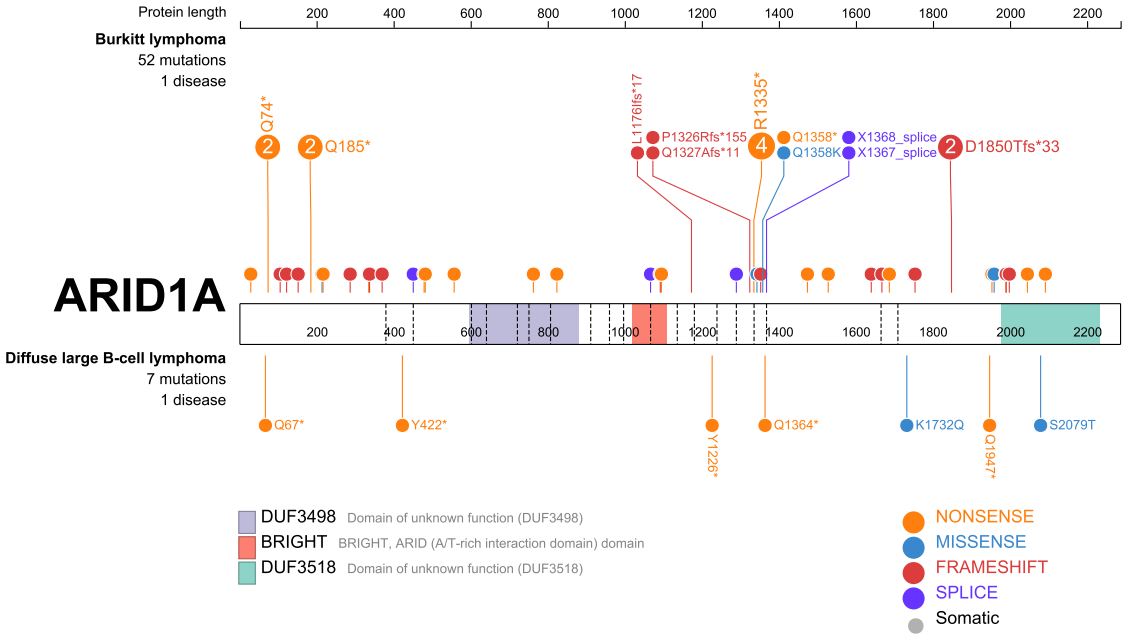
Supplemental Figure S3

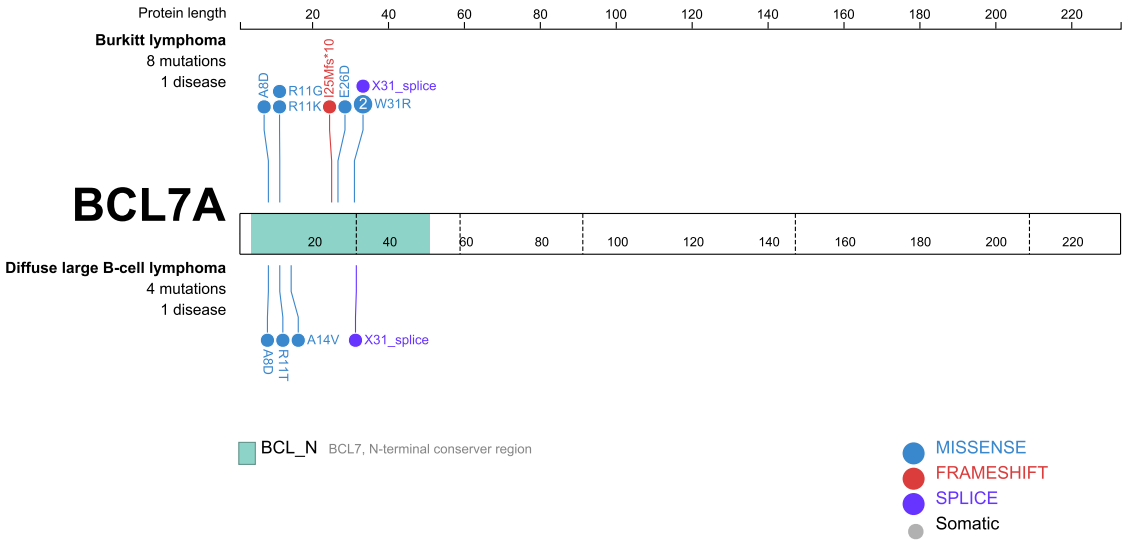


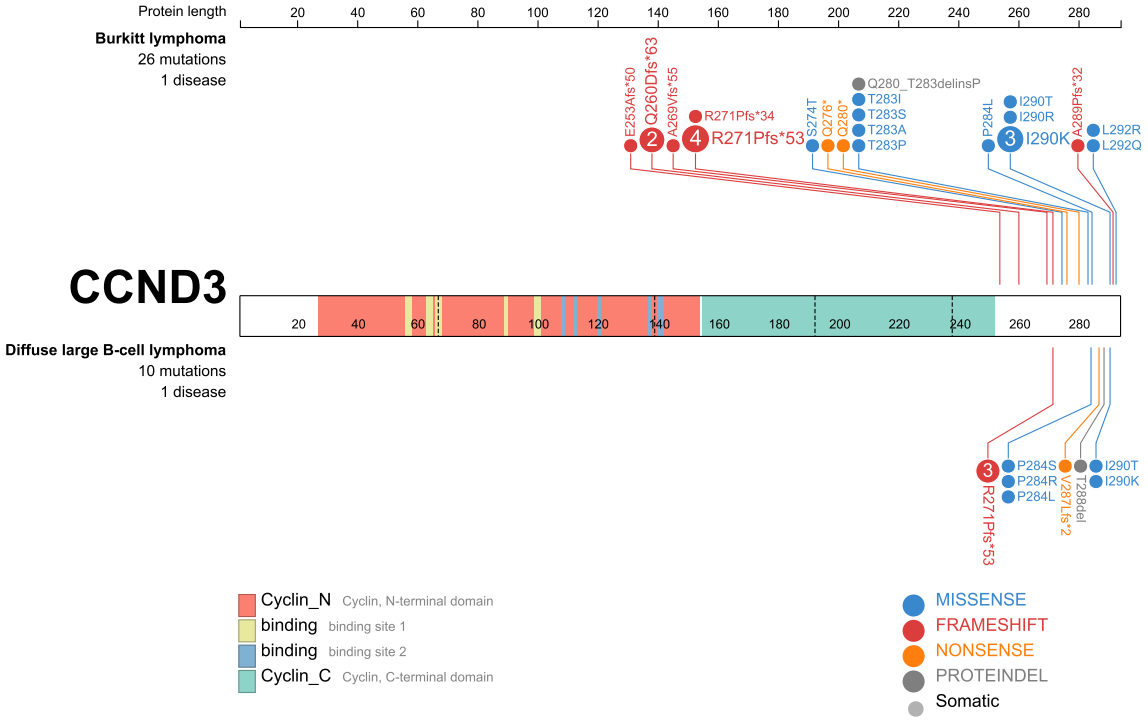
Supplemental Figure S4

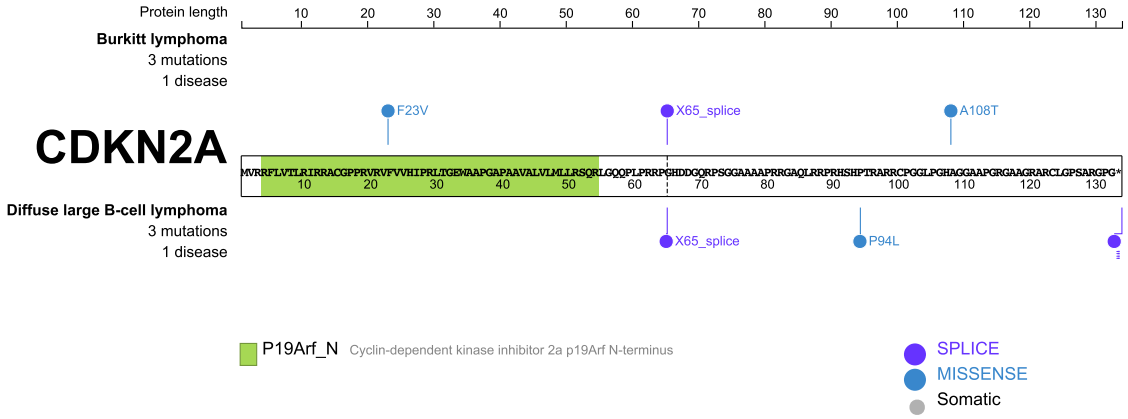


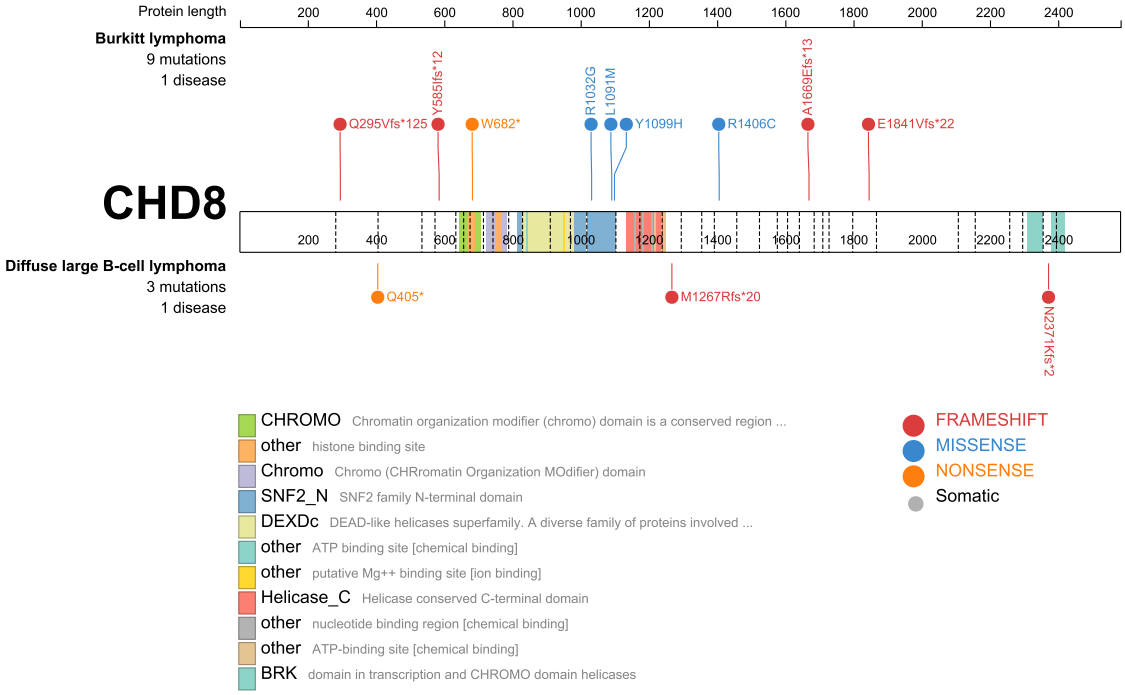
Supplemental Figure S5

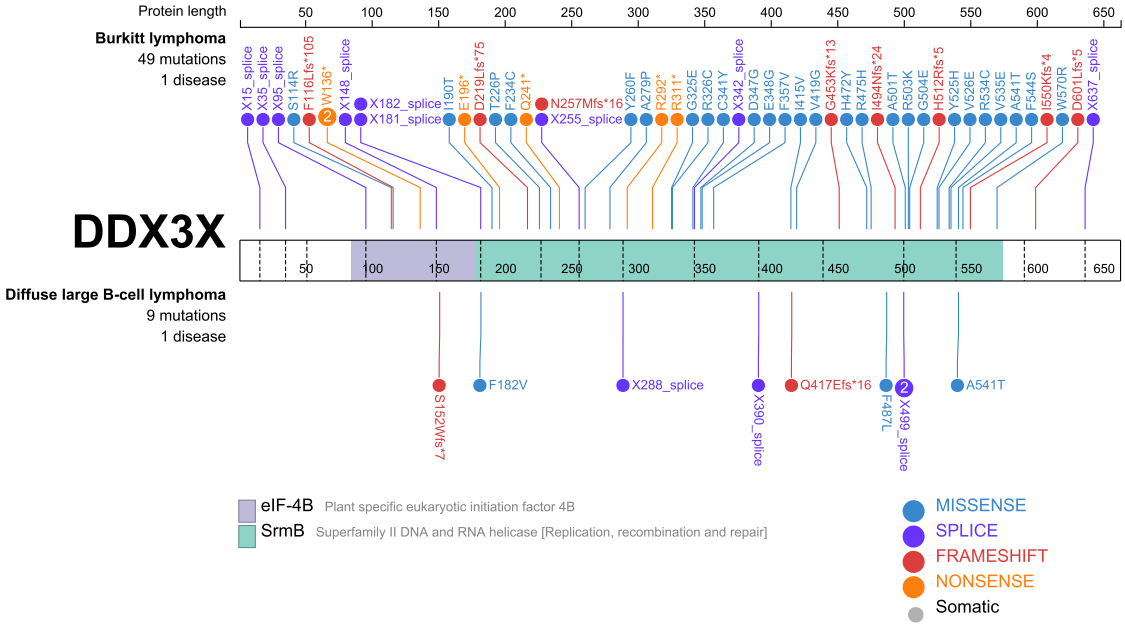


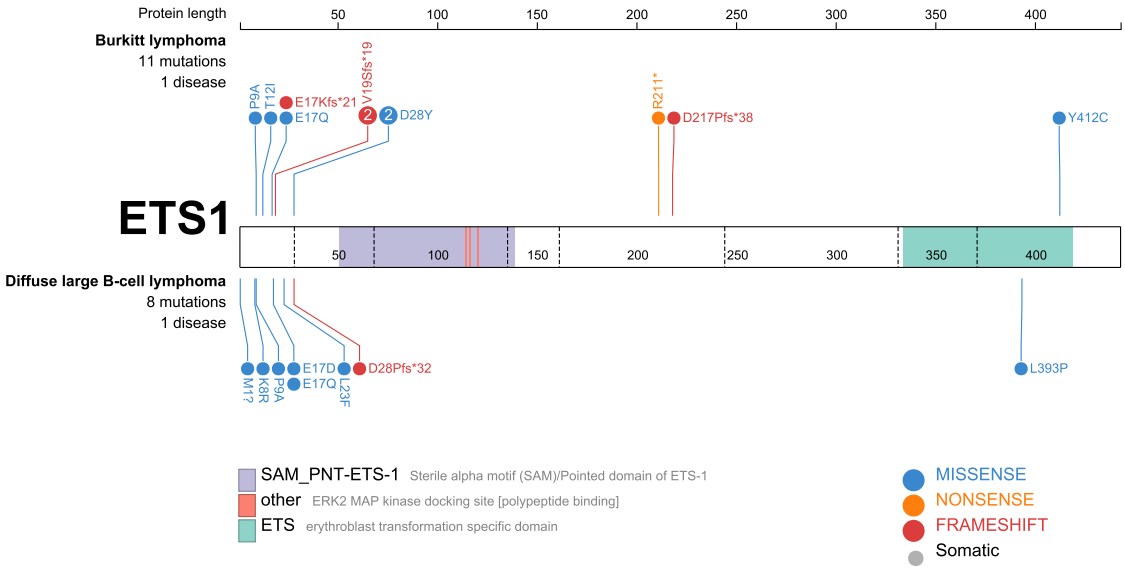


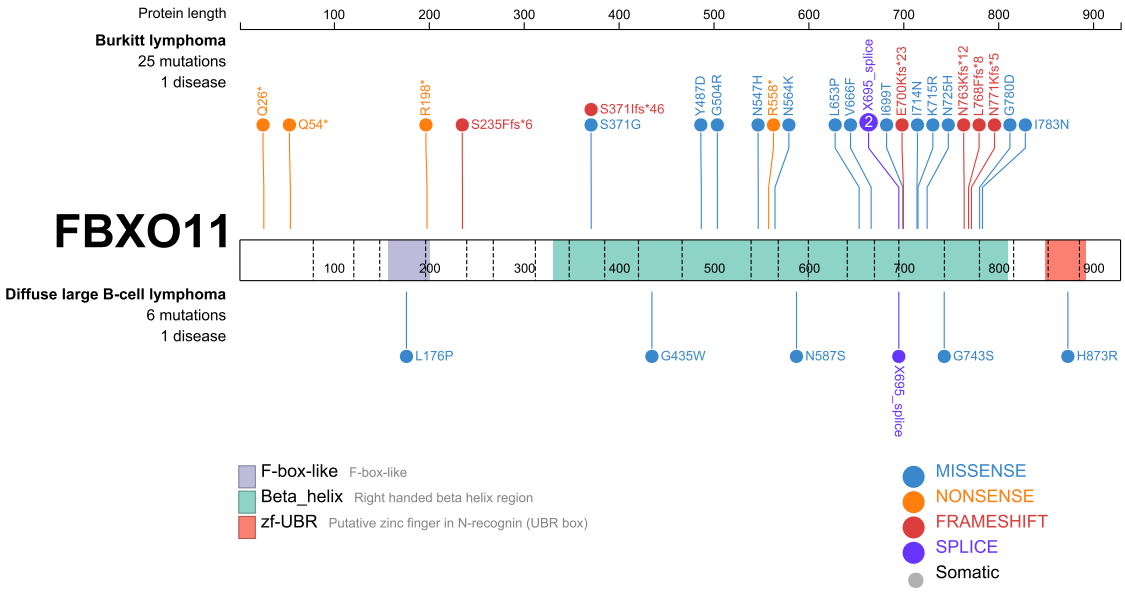

Protein length
20 40 60 80 100 120 140 160 180 200 220 240 260 280
Burkitt lymphoma
26 mutations
1 disease
E253Afs*50
Q260Dfs*63
A269Vfs*55
R271Pfs*34
2
4
R271Pfs*53
S274T
Q276*
Q280*
Q280_T283delinsP
T283I
T283S
T283A
T283P
P284L
I290T
I290R
3
I290K
A289Pfs*32
L292R
L292Q
CCND3
Diffuse large B-cell lymphoma
10 mutations
1 disease
3
R271Pfs*53
P284S
P284R
P284L
V287Ffs*2
T288del
I290T
I290K
Cyclin_N Cyclin, N-terminal domain
binding binding site 1
binding binding site 2
Cyclin_C Cyclin, C-terminal domain
MISSENSE
FRAMESHIFT
NONSENSE
PROTEINDEL
Somatic

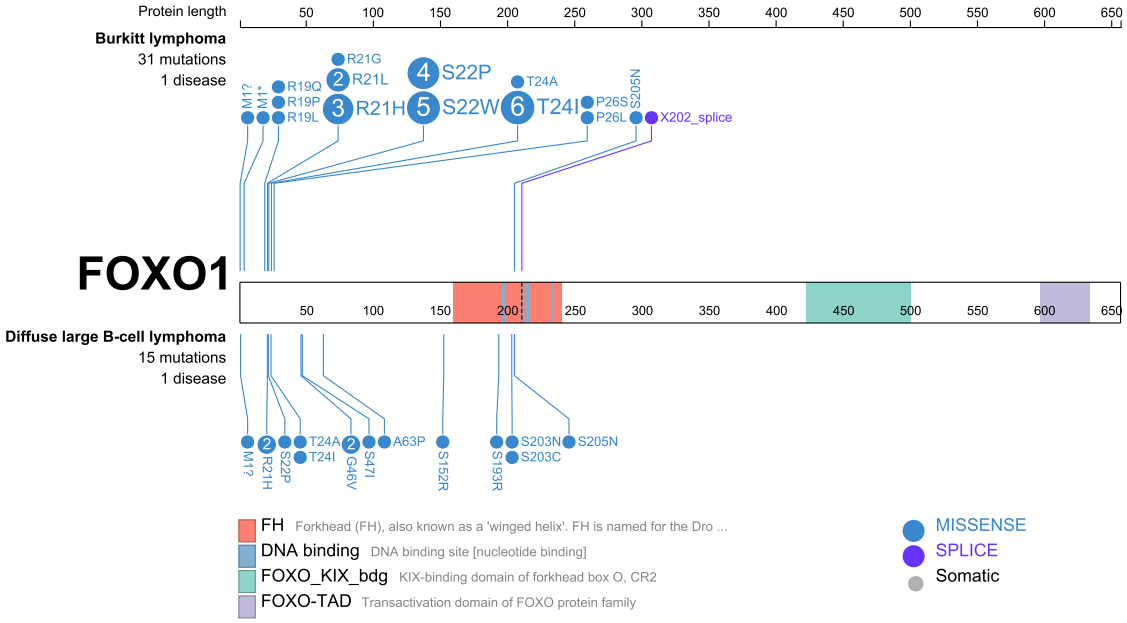

Protein length
10 20 30 40 50 60 70 80 90 100 110 120 130
Burkitt lymphoma
3 mutations
1 disease
F23V
X65_splice
A108T
CDKN2A
Diffuse large B-cell lymphoma
3 mutations
1 disease
X65_splice
P94L
P19Arf_N Cyclin-dependent kinase inhibitor 2a p19Arf N-terminus
SPLICE
MISSENSE
Somatic

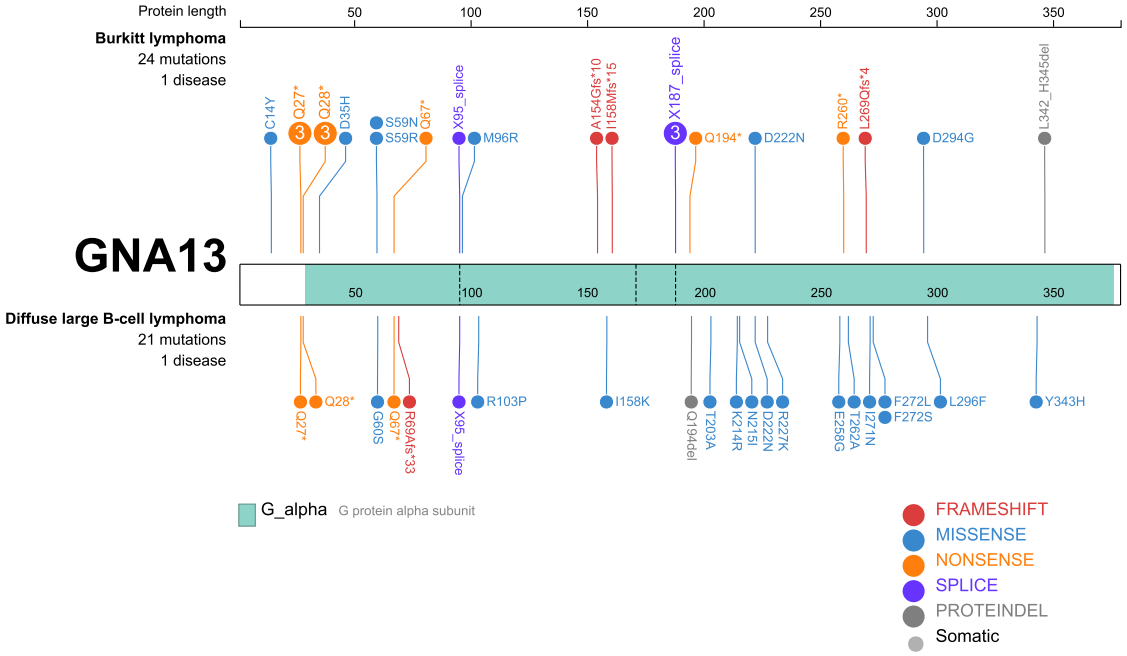

Protein length
200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400
Burkitt lymphoma
9 mutations
1 disease
Q295Vfs*125
Y585fs*12
W682*
R1032G
L1091M
Y1099H
R1406C
A1669Efs*13
E1841Vfs*22
CHD8
Diffuse large B-cell lymphoma
3 mutations
1 disease
Q405*
M1267Rfs*20
N2371Kfs*2
CHROMO Chromatin organization modifier (chromo) domain is a conserved region ...
other histone binding site
Chromo Chromo (CHRromatin Organization MOdifier) domain
SNF2_N SNF2 family N-terminal domain
DEXDc DEAD-like helicases superfamily. A diverse family of proteins involved ...
other ATP binding site [chemical binding]
other putative Mg++ binding site [ion binding]
Helicase_C Helicase conserved C-terminal domain
other nucleotide binding region [chemical binding]
other ATP-binding site [chemical binding]
BRK domain in transcription and CHROMO domain helicases
FRAMESHIFT
MISSENSE
NONSENSE
Somatic

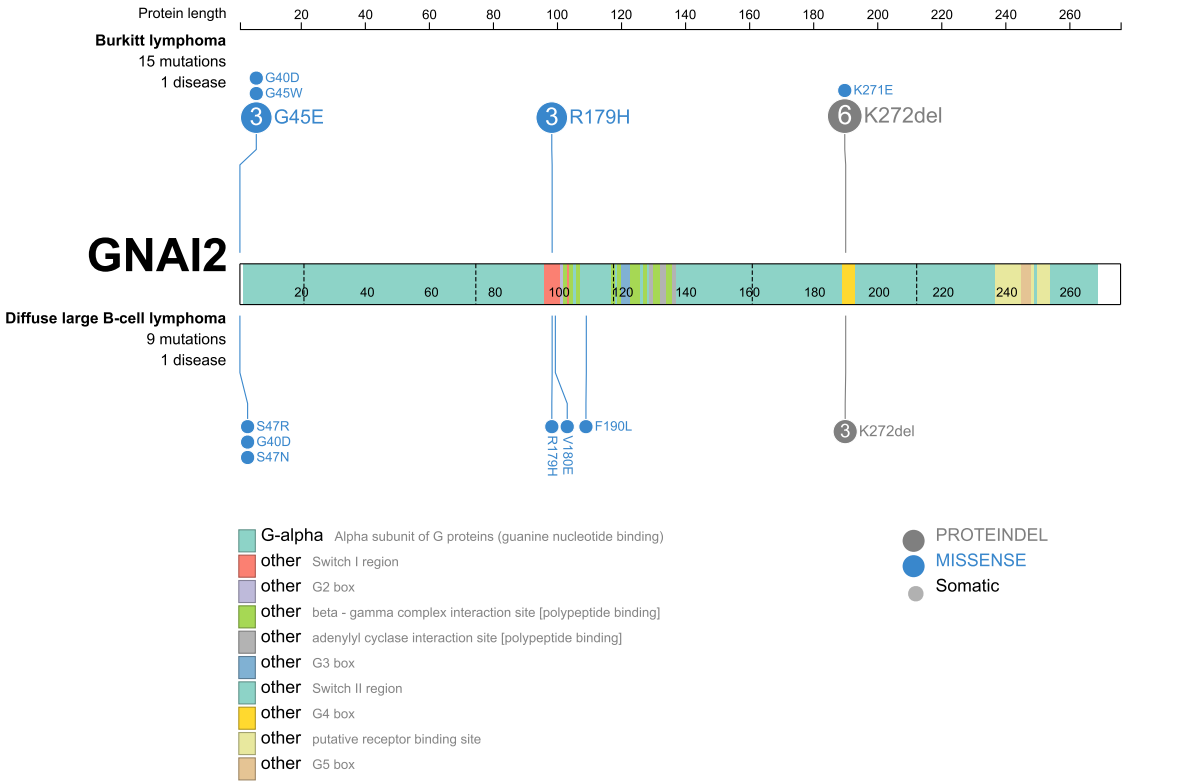

Protein length
50 100 150 200 250 300 350 400 450 500 550 600 650
Burkitt lymphoma
49 mutations
1 disease
X15_splice
X35_splice
X95_splice
S114R
F116Lfs*105
W136*
X148_splice
X182_splice
X181_splice
I190T
E196*
D219Lfs*75
T226P
F234C
Q241*
N257Mfs*16
X255_splice
Y260F
A279P
R292*
R311*
G325E
R326C
C341Y
X342_splice
D347G
E348G
F357V
I415V
V419G
G453Kfs*13
H472Y
R475H
I494Nfs*24
A501T
R503K
G504E
H512Rfs*5
Y525H
V526E
R534C
V535E
A541T
F544S
S550Kfs*4
W570R
D601Lfs*5
X637_splice
DDX3X
Diffuse large B-cell lymphoma
9 mutations
1 disease
S152Wfs*7
F182V
X288_splice
X390_splice
Q417Efs*16
F487L
X499_splice
A541T
eIF-4B Plant specific eukaryotic initiation factor 4B
SrmB Superfamily II DNA and RNA helicase [Replication, recombination and repair]
MISSENSE
SPLICE
FRAMESHIFT
NONSENSE
Somatic

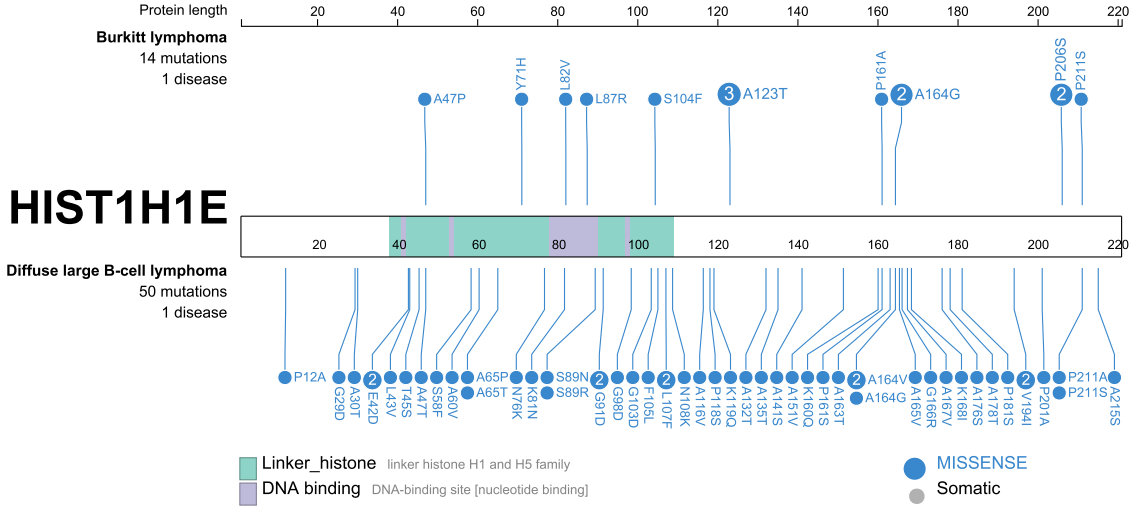

Protein length
50
100
150
200
250
300
350
400
Burkitt lymphoma
11 mutations
1 disease
P9A
T12I
E17Kfs*21
E17Q
V195fs*19
D28Y
R211*
D217Pfs*38
Y412C
ETS1
Diffuse large B-cell lymphoma
8 mutations
1 disease
M1?
K8R
P9A
E17D
E17Q
L23F
D28Pfs*32
L393P
SAM_PNT-ETS-1 Sterile alpha motif (SAM)/Pointed domain of ETS-1
other ERK2 MAP kinase docking site [polypeptide binding]
ETS erythroblast transformation specific domain
MISSENSE
NONSENSE
FRAMESHIFT
Somatic

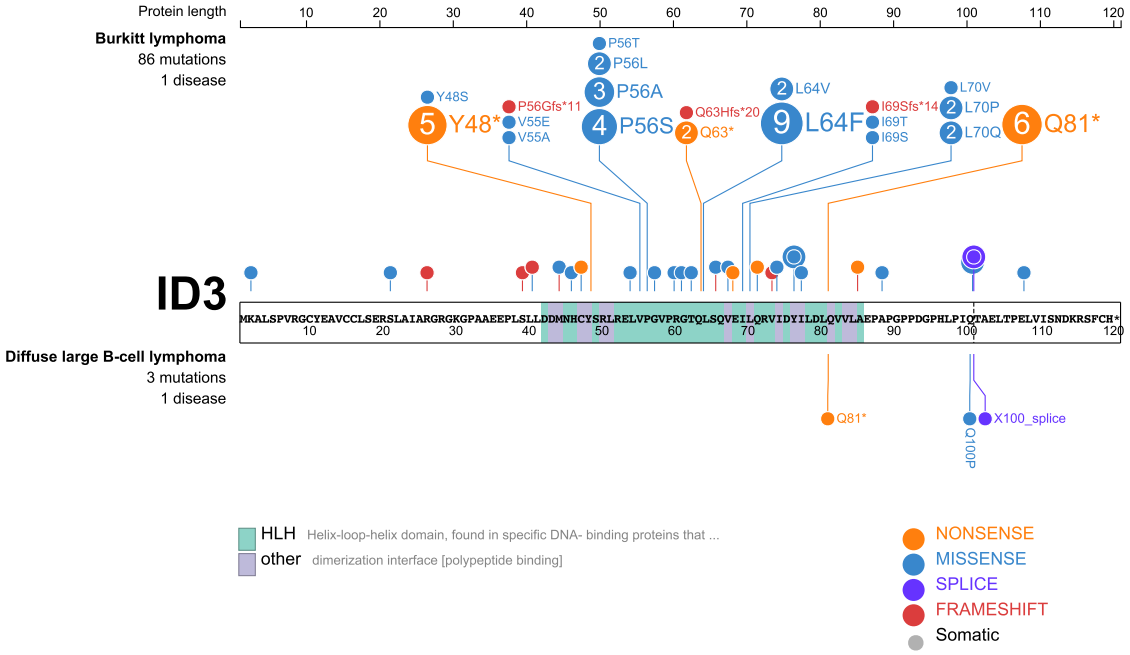

Protein length
100
200
300
400
500
600
700
800
900
Burkitt lymphoma
25 mutations
1 disease
Q26*
Q54*
R198*
S235Ffs*6
S371Ifs*46
S371G
Y487D
G504R
N547H
R558*
N564K
L653P
V666F
X695_splice
I699T
E700Kfs*23
I714N
K715R
N725H
N763Kfs*12
L768Ffs*8
N771Kfs*5
G780D
I783N
FBXO11
Diffuse large B-cell lymphoma
6 mutations
1 disease
L176P
G435W
N587S
X695_splice
G743S
H873R
F-box-like F-box-like
Beta_helix Right handed beta helix region
zf-UBR Putative zinc finger in N-recognin (UBR box)
MISSENSE
NONSENSE
FRAMESHIFT
SPLICE
Somatic

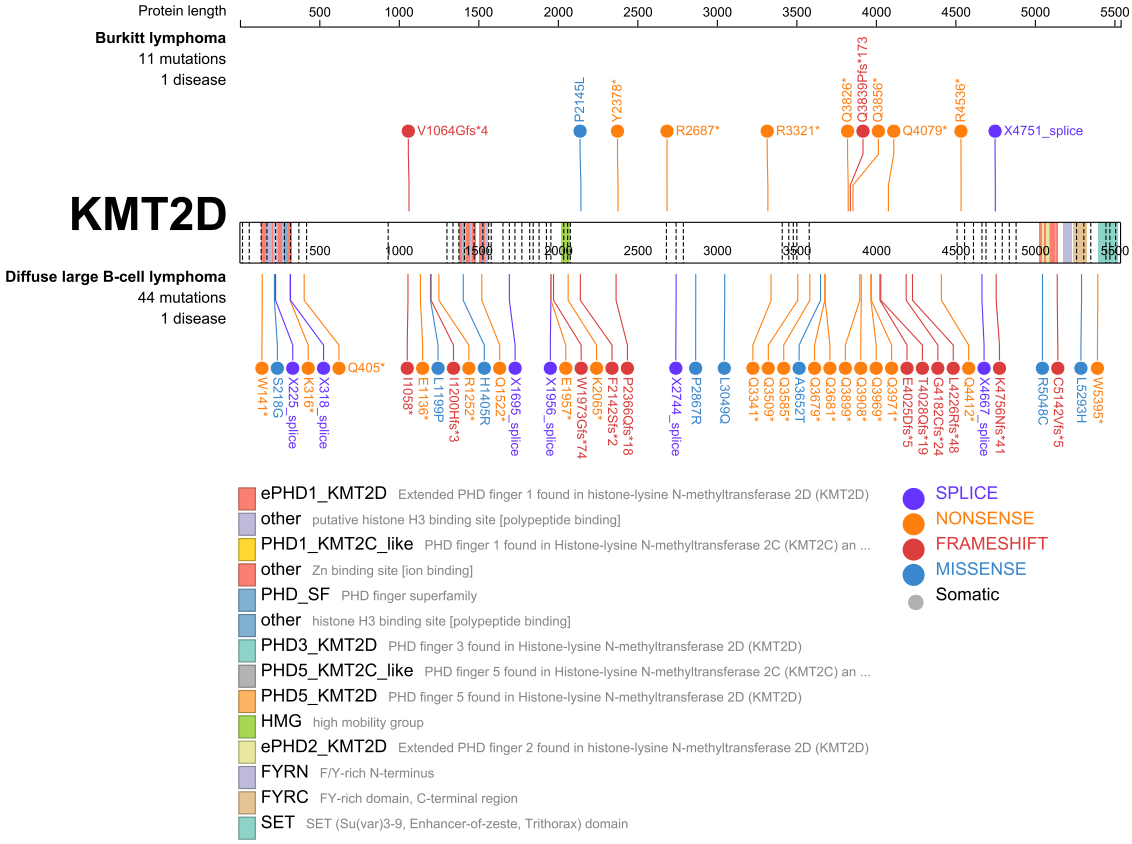

Protein length
50 100 150 200 250 300 350 400 450 500 550 600 650
Burkitt lymphoma
31 mutations
1 disease
M1?
M1I
R19Q
R19P
R19L
R21G
2 R21L
3 R21H
4 S22P
5 S22W
T24A
6 T24I
P26S
P26L
S205N
X202_splice
FOXO1
Diffuse large B-cell lymphoma
15 mutations
1 disease
M1?
2 R21H
S22P
T24A
T24I
2 G46V
S47I
A63P
S152R
S193R
S203N
S203C
S205N
FH Forkhead (FH), also known as a 'winged helix'. FH is named for the Dro ...
DNA binding DNA binding site [nucleotide binding]
FOXO_KIX_bdg KIX-binding domain of forkhead box O, CR2
FOXO-TAD Transactivation domain of FOXO protein family
MISSENSE
SPLICE
Somatic

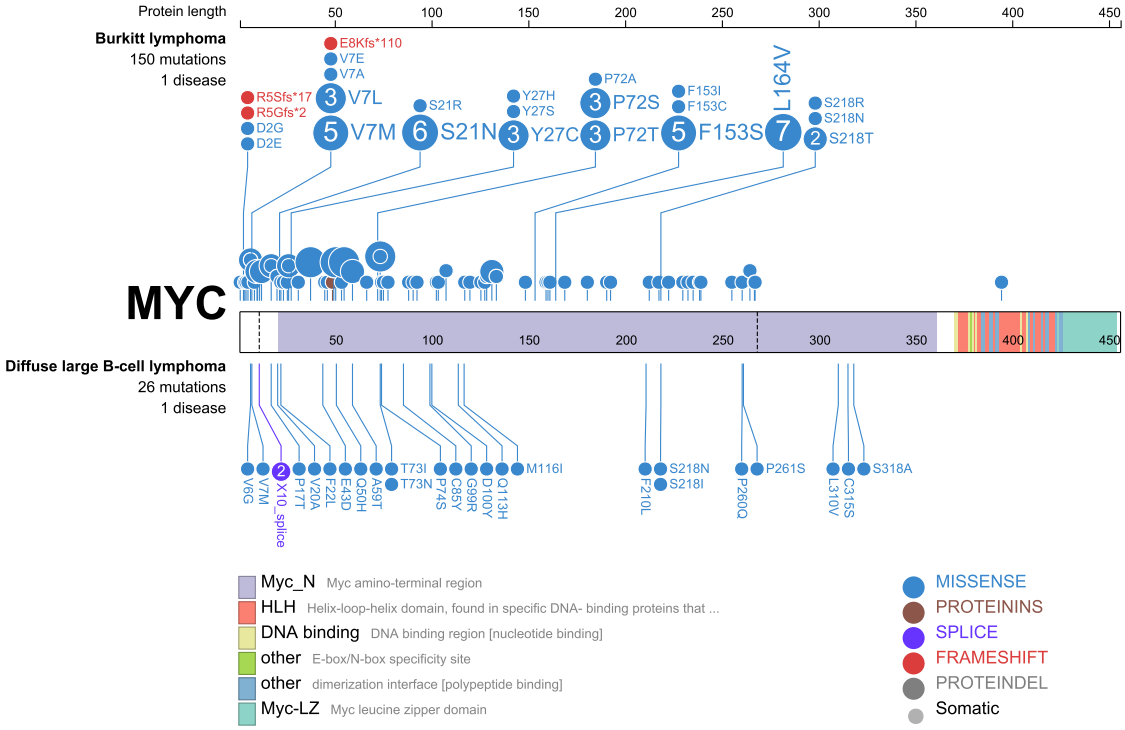

Protein length
50 100 150 200 250 300 350
Burkitt lymphoma
24 mutations
1 disease
C14Y
3 Q27*
3 Q28*
D35H
S59N
S59R
Q67*
X95_splice
M96R
A154Gfs*10
I158Mfs*15
3 X187_splice
Q194*
D222N
R260*
L269Qfs*4
D294G
L342_H345del
GNA13
Diffuse large B-cell lymphoma
21 mutations
1 disease
Q27*
Q28*
G60S
Q67*
R69Afs*33
X95_splice
R103P
I158K
Q194del
T203A
K214R
N215I
D222N
R227K
E258G
T262A
I271N
F272L
F272S
L296F
Y343H
G_alpha G protein alpha subunit
FRAMESHIFT
MISSENSE
NONSENSE
SPLICE
PROTEINDEL
Somatic

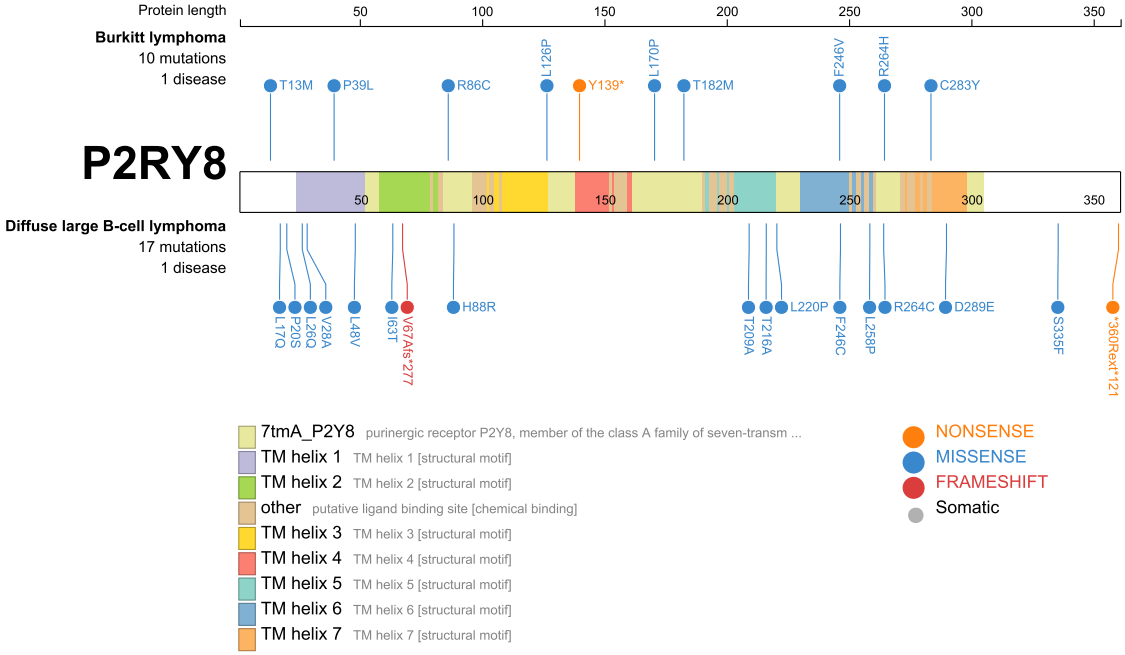

Protein length
20 40 60 80 100 120 140 160 180 200 220 260
Burkitt lymphoma
15 mutations
1 disease
G40D
G45W
3 G45E
3 R179H
K271E
6 K272del
GNAI2
20 40 60 80 100 120 140 160 180 200 220 240 260
Diffuse large B-cell lymphoma
9 mutations
1 disease
S47R
G40D
S47N
R179H
V180E
F190L
3 K272del
G-alpha Alpha subunit of G proteins (guanine nucleotide binding)
other Switch I region
other G2 box
other beta - gamma complex interaction site [polypeptide binding]
other adenylyl cyclase interaction site [polypeptide binding]
other G3 box
other Switch II region
other G4 box
other putative receptor binding site
other G5 box
PROTEINDEL
MISSENSE
Somatic

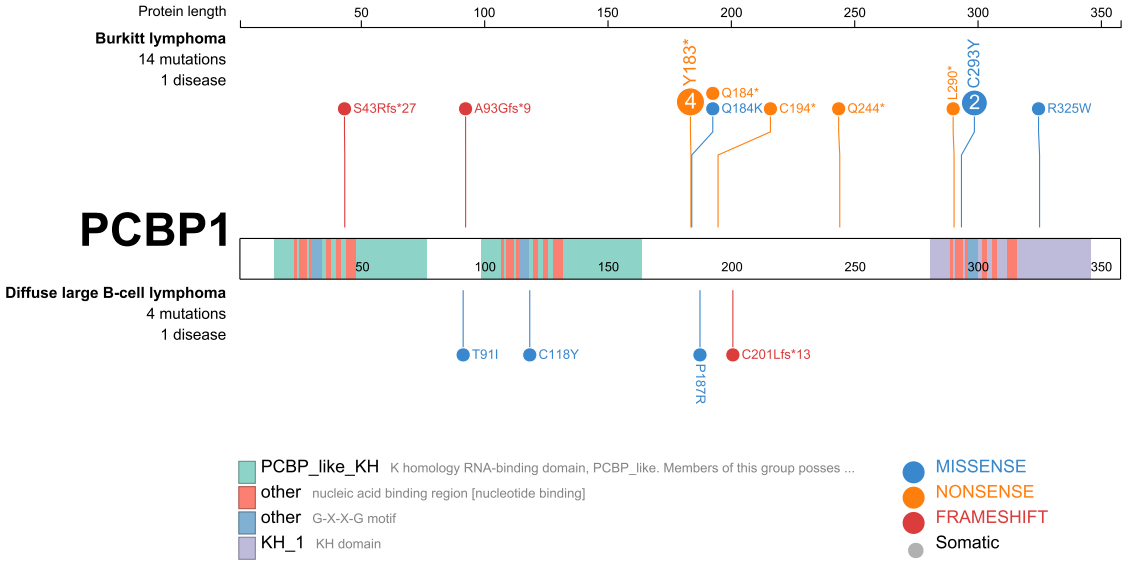

Protein length
20 40 60 80 100 120 140 160 180 200 220
Burkitt lymphoma
14 mutations
1 disease
A47P
Y71H
L82V
L87R
S104F
3 A123T
P161A
2 A164G
2 P206S
P211S
HIST1H1E
20 40 60 80 100 120 140 160 180 200 220
Diffuse large B-cell lymphoma
50 mutations
1 disease
P12A
G29D
A30T
2 E42D
L43V
T45S
A47T
S58F
A60V
A65P
A65T
N76K
K81N
S89N
S89R
2 G91D
G98D
G103D
F105L
2 L107F
N108K
A116V
P118S
K119Q
A132T
A135T
A141S
A151V
K160Q
P161S
A163T
2 A164V
A164G
A165V
G166R
A167V
K168I
A176S
A178T
P181S
2 V194I
P201A
P211A
P211S
A215S
Linker_histone linker histone H1 and H5 family
DNA binding DNA-binding site [nucleotide binding]
MISSENSE
Somatic

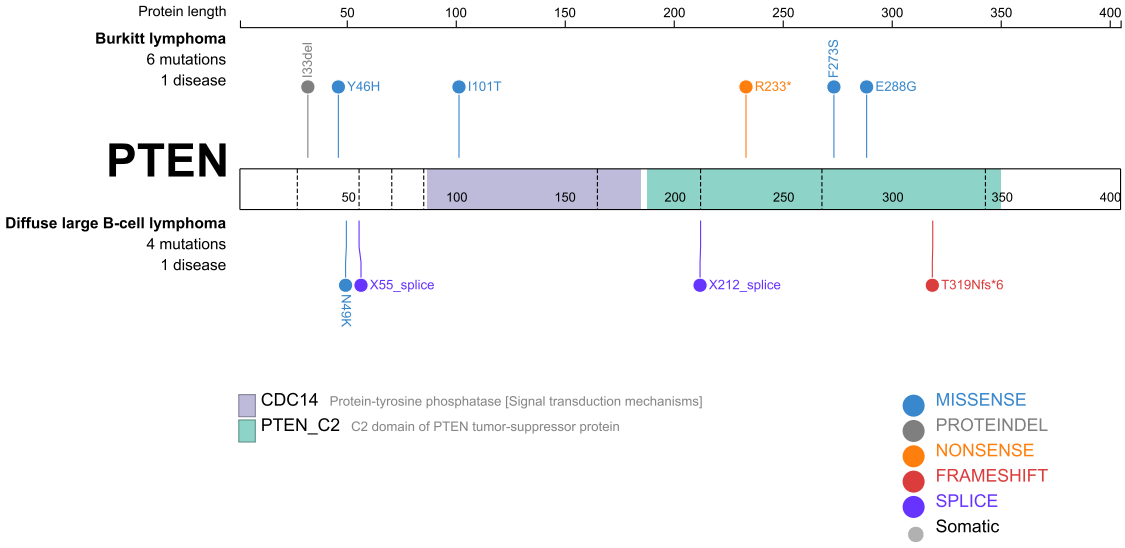

Protein length
10 20 30 40 50 60 70 80 90 100 110 120
Burkitt lymphoma
86 mutations
1 disease
P56T
P56L
P56A
P56S
Y48S
Y48*
P56Gfs*11
V55E
V55A
Q63Hfs*20
Q63*
L64V
L64F
I69Sfs*14
I69T
I69S
L70V
L70P
L70Q
Q81*
ID3
MKALSPVRGCYEAVCCLSERSLAIARGRGKGPAAEEPLSLLDDMNHCYSRLRELVPGVPRGTQLSQVEILQRVIDYILDLQVVLAEPAPGPPDGPHLPIQTAELTPELVISNDKRSFCH*
10 20 30 40 50 60 70 80 90 100 110 120
Diffuse large B-cell lymphoma
3 mutations
1 disease
Q81*
X100_splice
Q100P
HLH Helix-loop-helix domain, found in specific DNA- binding proteins that ...
other dimerization interface [polypeptide binding]
NONSENSE
MISSENSE
SPLICE
FRAMESHIFT
Somatic

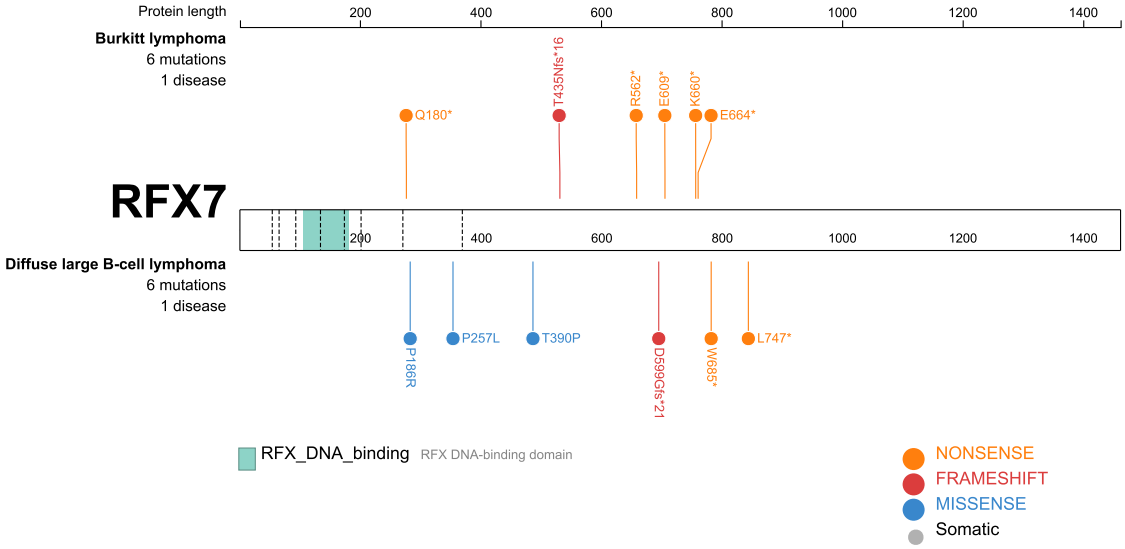

Protein length
500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500
Burkitt lymphoma
11 mutations
1 disease
V1064Gfs*4
P2145L
Y2378*
R2687*
R3321*
Q3826*
Q3839Pfs*173
Q3856*
Q4079*
R4536*
X4751_splice
KMT2D
500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500
Diffuse large B-cell lymphoma
44 mutations
1 disease
W141*
S218G
X225_splice
K316*
X318_splice
Q405*
I1058*
E1136*
L1199P
I1200Hfs*3
R1252*
H1405R
Q1522*
X1695_splice
X1956_splice
E1957*
W1973Gfs*74
K2065*
F2142Sfs*2
P2366Qfs*18
X2744_splice
P2867R
L3049Q
Q3341*
Q3509*
Q3585*
A3652T
Q3679*
Q3681*
Q3889*
Q3908*
Q3969*
Q3971*
E4025Dfs*5
T4028Qfs*19
G4182Cfs*24
L4226Rfs*48
Q4412*
X4667_splice
K4756Nfs*41
R5048C
C5142Vfs*5
L5293H
W5395*
ePHD1_KMT2D Extended PHD finger 1 found in histone-lysine N-methyltransferase 2D (KMT2D)
other putative histone H3 binding site [polypeptide binding]
PHD1_KMT2C_like PHD finger 1 found in Histone-lysine N-methyltransferase 2C (KMT2C) an ...
other Zn binding site [ion binding]
PHD_SF PHD finger superfamily
other histone H3 binding site [polypeptide binding]
PHD3_KMT2D PHD finger 3 found in Histone-lysine N-methyltransferase 2D (KMT2D)
PHD5_KMT2C_like PHD finger 5 found in Histone-lysine N-methyltransferase 2C (KMT2C) an ...
PHD5_KMT2D PHD finger 5 found in Histone-lysine N-methyltransferase 2D (KMT2D)
HMG high mobility group
ePHD2_KMT2D Extended PHD finger 2 found in histone-lysine N-methyltransferase 2D (KMT2D)
FYRN F/Y-rich N-terminus
FYRC FY-rich domain, C-terminal region
SET SET (Su(var)3-9, Enhancer-of-zeste, Trithorax) domain
SPLICE
NONSENSE
FRAMESHIFT
MISSENSE
Somatic

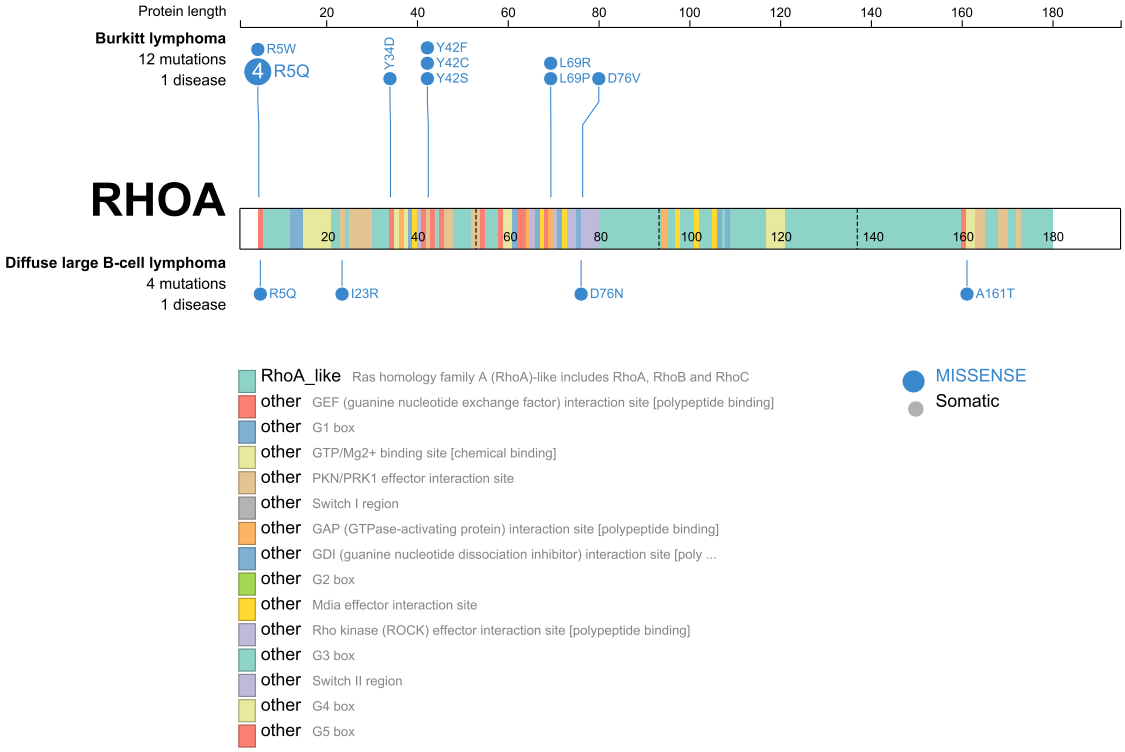

Protein length
50 100 150 200 250 300 350 400 450
Burkitt lymphoma
150 mutations
1 disease
E8Kfs*110
V7E
V7A
R5Sfs*17
R5Gfs*2
D2G
D2E
V7L
V7M
S21R
S21N
Y27H
Y27S
Y27C
P72A
P72S
P72T
F153I
F153C
F153S
L164V
S218R
S218N
S218T
MYC
50 100 150 200 250 300 350 400 450
Diffuse large B-cell lymphoma
26 mutations
1 disease
V6G
V7M
X10_splice
P17T
V20A
F22L
E43D
Q50H
A59T
T73I
T73N
P74S
C85Y
G99R
D100Y
Q113H
M116I
F210L
S218N
S218I
P260Q
P261S
L310V
C315S
S318A
Myc_N Myc amino-terminal region
HLH Helix-loop-helix domain, found in specific DNA- binding proteins that ...
DNA binding DNA binding region [nucleotide binding]
other E-box/N-box specificity site
other dimerization interface [polypeptide binding]
Myc-LZ Myc leucine zipper domain
MISSENSE
PROTEININS
SPLICE
FRAMESHIFT
PROTEINDEL
Somatic

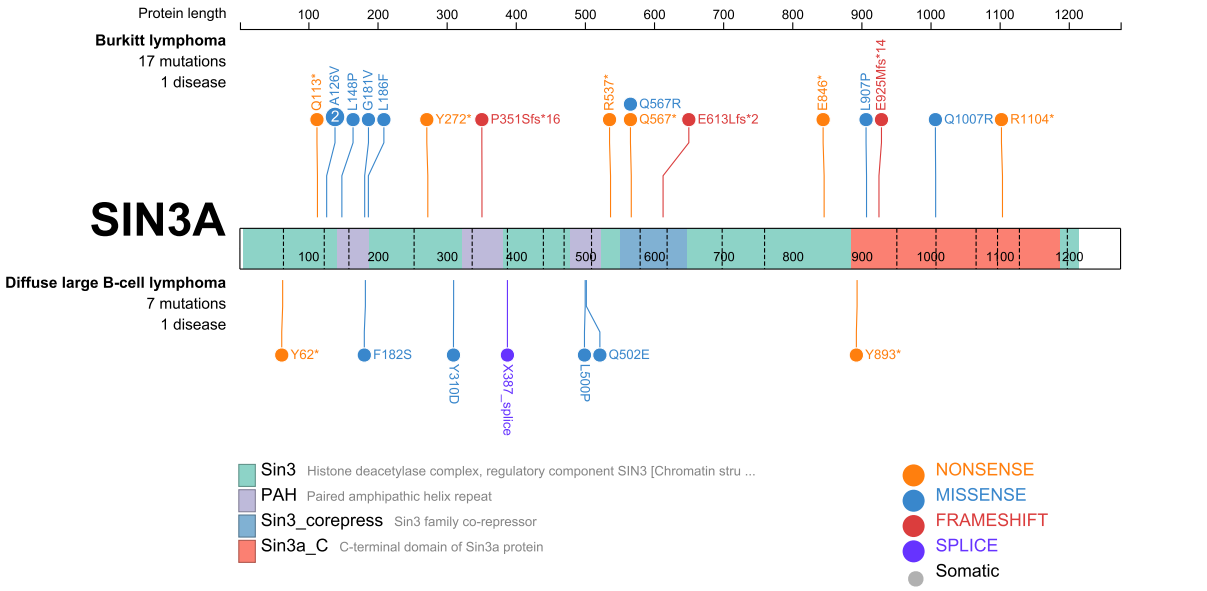

Protein length
50 100 150 200 250 300 350
Burkitt lymphoma
10 mutations
1 disease
T13M
P39L
R86C
L126P
Y139*
L170P
T182M
F249V
R264H
C283Y
P2RY8
50 100 150 200 250 300 350
Diffuse large B-cell lymphoma
17 mutations
1 disease
L17Q
P20S
L26Q
V28A
L48V
I63T
V67Afs*277
H88R
T209A
T216A
L220P
F246C
L258P
R264C
D289E
S335F
*360Rext*121
7tmA_P2Y8 purinergic receptor P2Y8, member of the class A family of seven-transm ...
TM helix 1 TM helix 1 [structural motif]
TM helix 2 TM helix 2 [structural motif]
other putative ligand binding site [chemical binding]
TM helix 3 TM helix 3 [structural motif]
TM helix 4 TM helix 4 [structural motif]
TM helix 5 TM helix 5 [structural motif]
TM helix 6 TM helix 6 [structural motif]
TM helix 7 TM helix 7 [structural motif]
NONSENSE
MISSENSE
FRAMESHIFT
Somatic

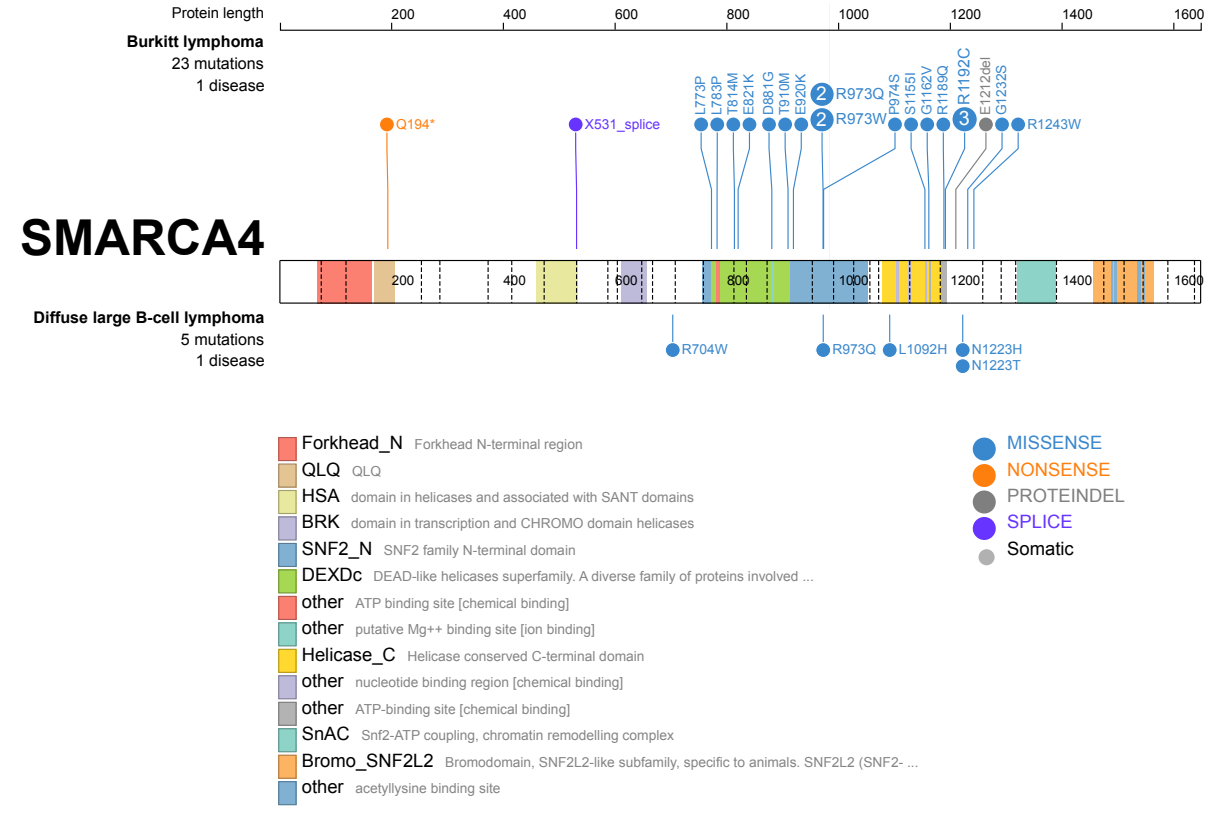

Protein length
50
100
150
200
250
300
350
Burkitt lymphoma
14 mutations
1 disease
S43Rfs*27
A93Gfs*9
Y183*
4
Q184*
Q184K
C194*
Q244*
L290*
C293Y
2
R325W
PCBP1
Diffuse large B-cell lymphoma
4 mutations
1 disease
T91I
C118Y
P187R
C201Lfs*13
PCBP_like_KH K homology RNA-binding domain, PCBP_like. Members of this group posses ...
other nucleic acid binding region [nucleotide binding]
other G-X-X-G motif
KH_1 KH domain
MISSENSE
NONSENSE
FRAMESHIFT
Somatic

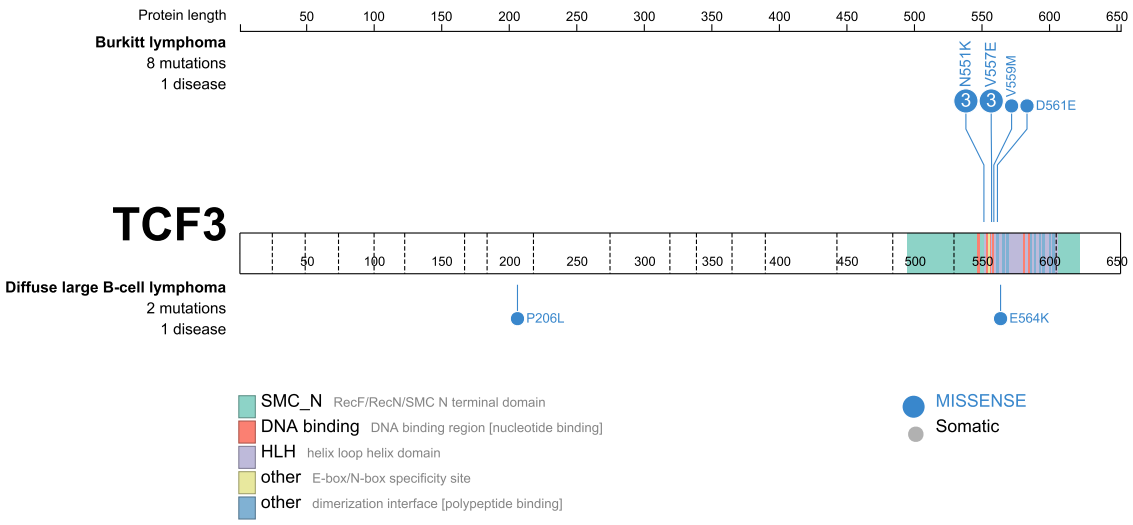

Protein length
50
100
150
200
250
300
350
400
Burkitt lymphoma
6 mutations
1 disease
I33del
Y46H
I101T
R233*
F273S
E288G
PTEN
Diffuse large B-cell lymphoma
4 mutations
1 disease
N49K
X55_splice
X212_splice
T319Nfs*6
CDC14 Protein-tyrosine phosphatase [Signal transduction mechanisms]
PTEN_C2 C2 domain of PTEN tumor-suppressor protein
MISSENSE
PROTEINDEL
NONSENSE
FRAMESHIFT
SPLICE
Somatic

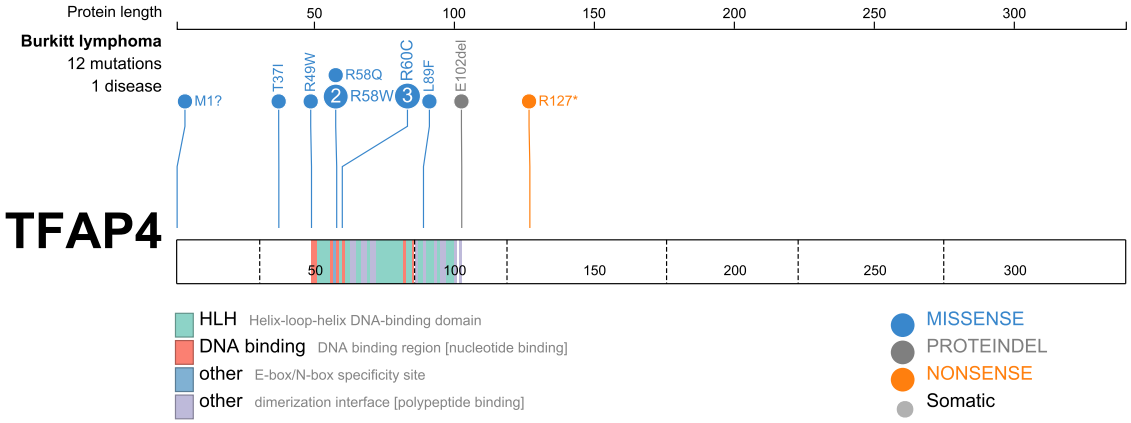

Protein length
200
400
600
800
1000
1200
1400
Burkitt lymphoma
6 mutations
1 disease
Q180*
T435Nfs*16
R562*
E609*
K960*
E664*
RFX7
Diffuse large B-cell lymphoma
6 mutations
1 disease
P186R
P257L
T390P
D599Gfs*21
W685*
L747*
RFX_DNA_binding
RFX DNA-binding domain
NONSENSE
FRAMESHIFT
MISSENSE
Somatic

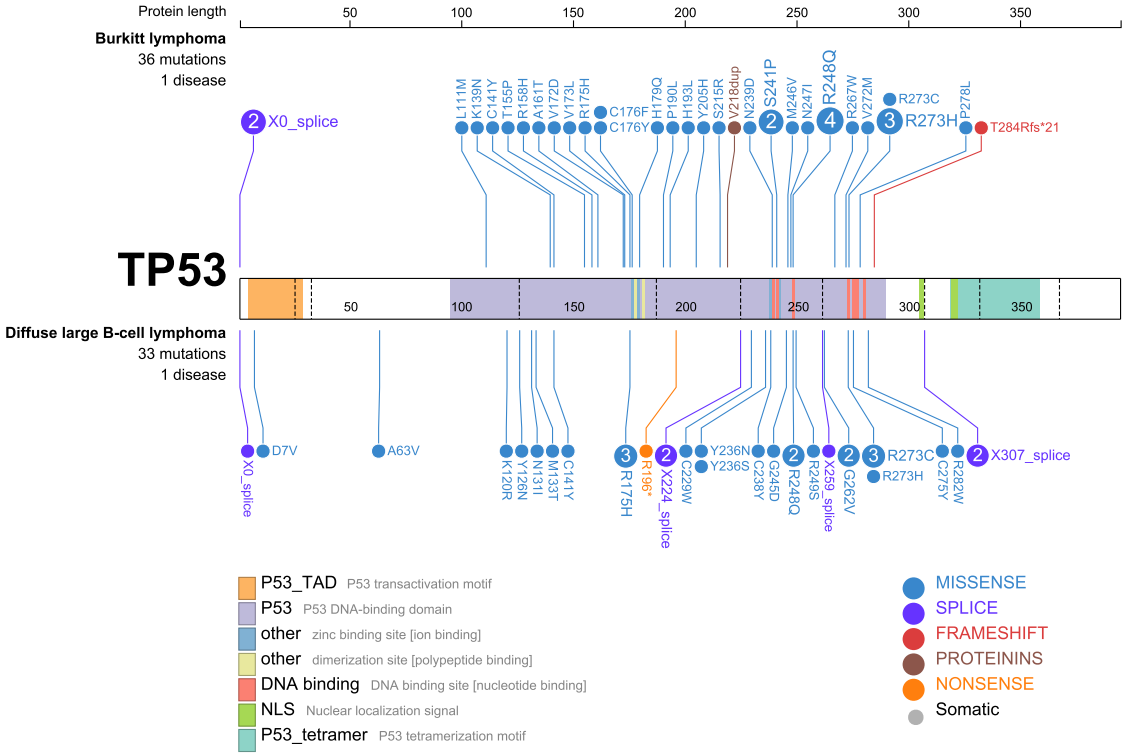

Protein length
20
40
60
80
100
120
140
160
180
Burkitt lymphoma
12 mutations
1 disease
R5W
4
R5Q
Y34D
Y42F
Y42C
Y42S
L69R
L69P
D76V
RHOA
Diffuse large B-cell lymphoma
4 mutations
1 disease
R5Q
I23R
D76N
A161T
RhoA_like
Ras homology family A (RhoA)-like includes RhoA, RhoB and RhoC
other
GEF (guanine nucleotide exchange factor) interaction site [polypeptide binding]
other
G1 box
other
GTP/Mg2+ binding site [chemical binding]
other
PKN/PRK1 effector interaction site
other
Switch I region
other
GAP (GTPase-activating protein) interaction site [polypeptide binding]
other
GDI (guanine nucleotide dissociation inhibitor) interaction site [poly ...
other
G2 box
other
Mdia effector interaction site
other
Rho kinase (ROCK) effector interaction site [polypeptide binding]
other
G3 box
other
Switch II region
other
G4 box
other
G5 box
MISSENSE
Somatic

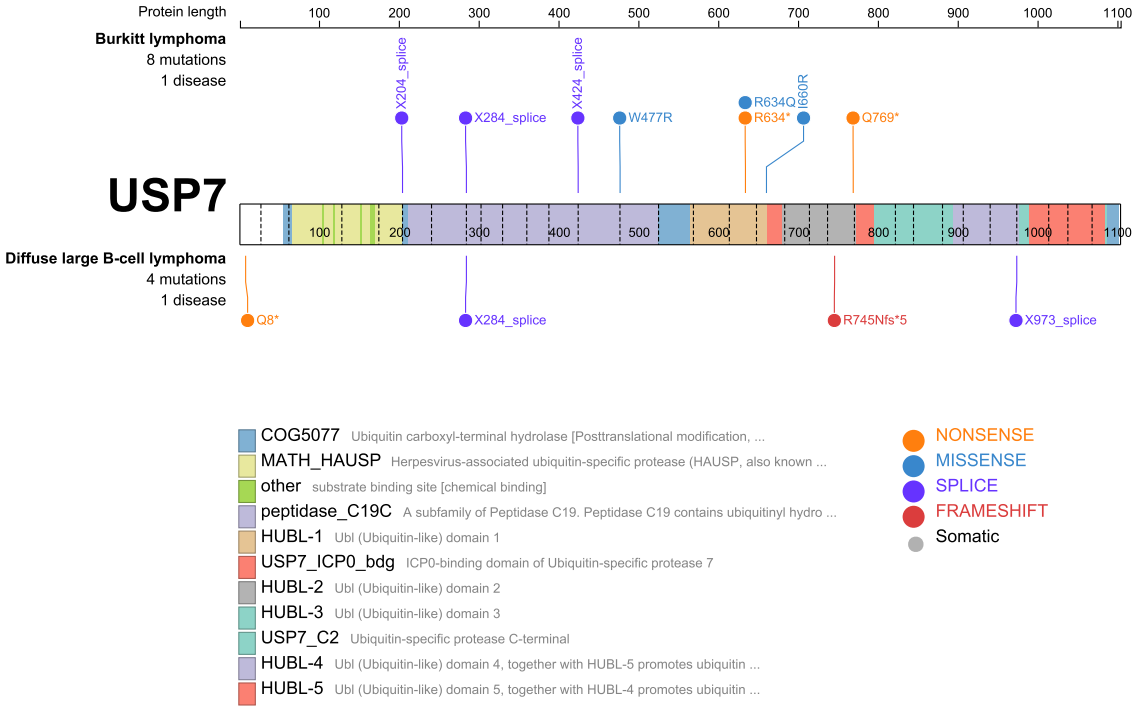

Protein length
100 200 300 400 500 600 700 800 900 1000 1100 1200
Burkitt lymphoma
17 mutations
1 disease
Q113*
A126V
L148P
G181V
L186F
Y272*
P351Sfs*16
R537*
Q567R
Q567*
E613Lfs*2
E846*
L907P
E925Mfs*14
Q1007R
R1104*
SIN3A
Diffuse large B-cell lymphoma
7 mutations
1 disease
Y62*
F182S
Y310D
X387_splice
L500P
Q502E
Y893*
Sin3 Histone deacetylase complex, regulatory component SIN3 [Chromatin stru ...
PAH Paired amphipathic helix repeat
Sin3_corepress Sin3 family co-repressor
Sin3a_C C-terminal domain of Sin3a protein
NONSENSE
MISSENSE
FRAMESHIFT
SPLICE
Somatic


Protein length
200 400 600 800 1000 1200 1400 1600
Burkitt lymphoma
23 mutations
1 disease
Q194*
X531_splice
L773P
L783P
T814M
E821K
D881G
T910M
E920K
R973Q
R973W
P974S
S1155I
G1162V
R1189Q
R1192C
E1212del
G1232S
R1243W
SMARCA4
Diffuse large B-cell lymphoma
5 mutations
1 disease
R704W
R973Q
L1092H
N1223H
N1223T
Forkhead_N Forkhead N-terminal region
QLQ QLQ
HSA domain in helicases and associated with SANT domains
BRK domain in transcription and CHROMO domain helicases
SNF2_N SNF2 family N-terminal domain
DEXDc DEAD-like helicases superfamily. A diverse family of proteins involved ...
other ATP binding site [chemical binding]
other putative Mg++ binding site [ion binding]
Helicase_C Helicase conserved C-terminal domain
other nucleotide binding region [chemical binding]
other ATP-binding site [chemical binding]
SnAC Snf2-ATP coupling, chromatin remodelling complex
Bromo_SNF2L2 Bromodomain, SNF2L2-like subfamily, specific to animals. SNF2L2 (SNF2- ...
other acetyllysine binding site
MISSENSE
NONSENSE
PROTEINDEL
SPLICE
Somatic

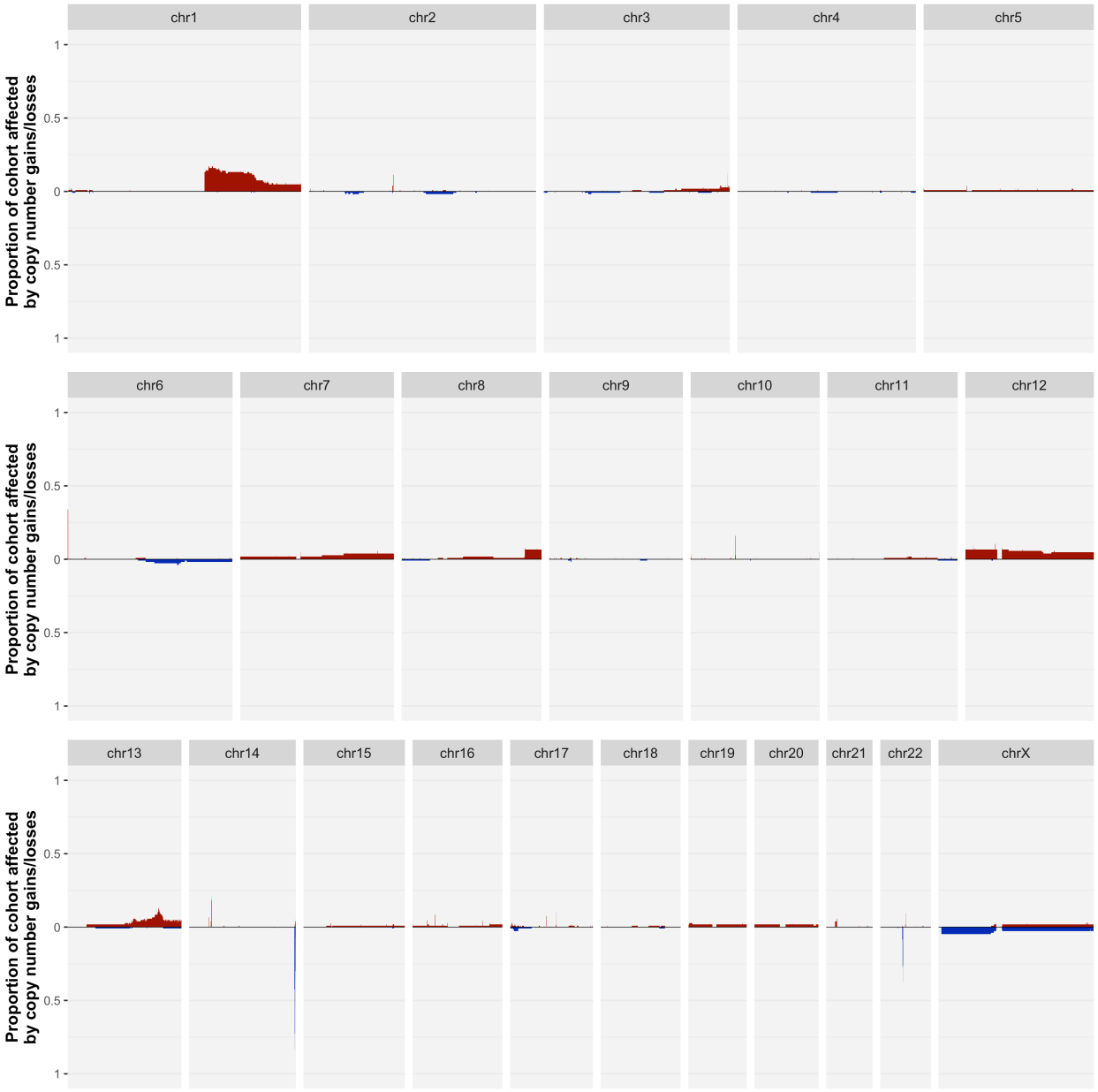

Protein length
50 100 150 200 250 300 350 400 450 500 550 600 650
Burkitt lymphoma
8 mutations
1 disease
N551K
V557E
V559M
D561E
TCF3
Diffuse large B-cell lymphoma
2 mutations
1 disease
P206L
E564K
SMC_N RecF/RecN/SMC N terminal domain
DNA binding DNA binding region [nucleotide binding]
HLH helix loop helix domain
other E-box/N-box specificity site
other dimerization interface [polypeptide binding]
MISSENSE
Somatic


Protein length
50 100 150 200 250 300
Burkitt lymphoma
12 mutations
1 disease
M1?
T37I
R49W
R58Q
R58W
R60C
L89F
E102del
R127*
TFAP4
HLH Helix-loop-helix DNA-binding domain
DNA binding DNA binding region [nucleotide binding]
other E-box/N-box specificity site
other dimerization interface [polypeptide binding]
MISSENSE
PROTEINDEL
NONSENSE
Somatic

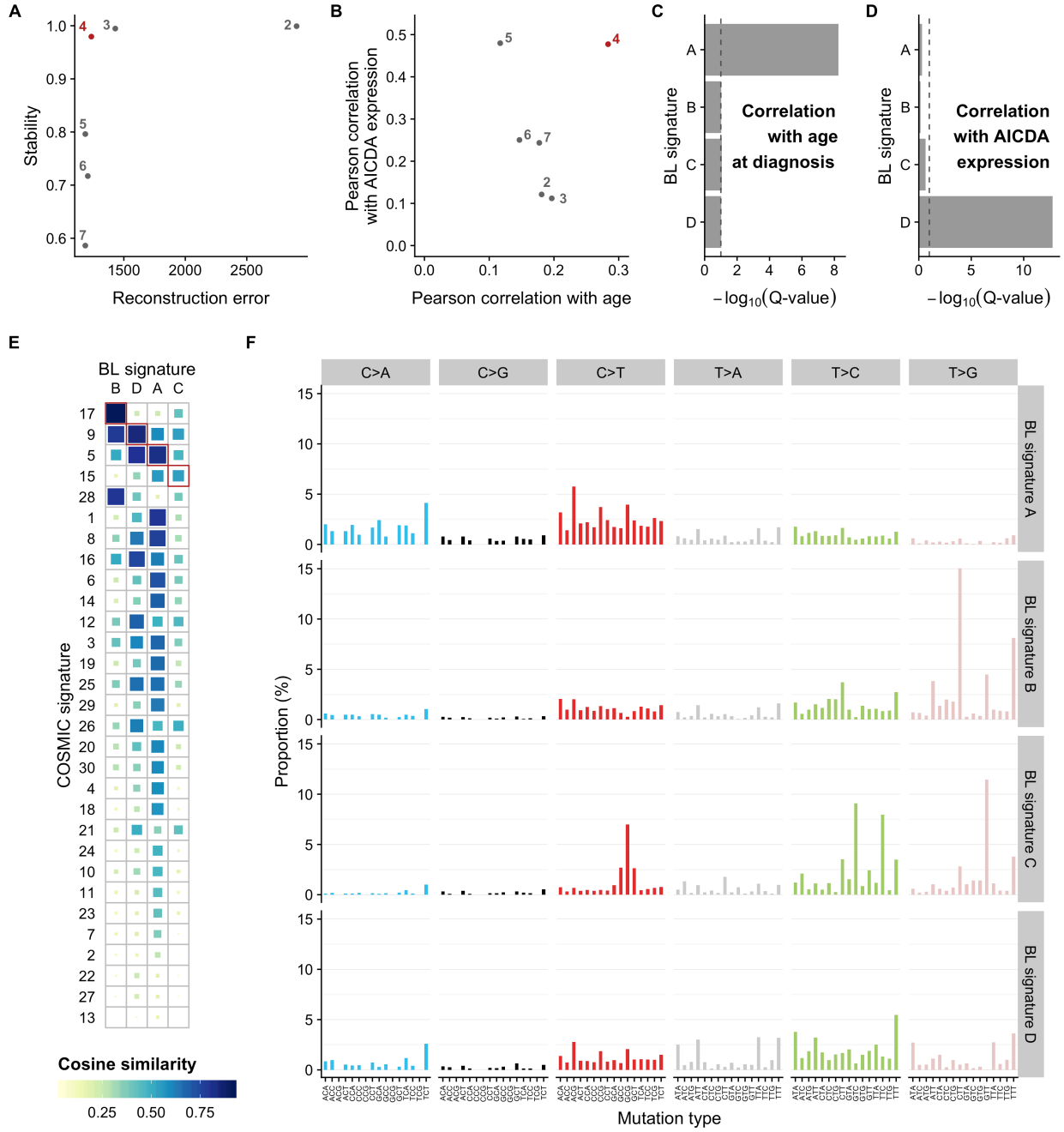

Protein length
50
100
150
200
250
300
350
Burkitt lymphoma
36 mutations
1 disease
X0_splice
L111M
K139N
C141Y
T155P
R158H
A161T
V172D
V173L
R175H
C176F
C176Y
H179Q
P190L
H193L
Y205H
S215R
V218dup
N239D
S241P
M246V
N247I
R248Q
R267W
V272M
R273C
R273H
P278L
T284Rfs*21
TP53
Diffuse large B-cell lymphoma
33 mutations
1 disease
X0_splice
D7V
A63V
K120R
Y126N
N131I
M133T
C141Y
R175H
R196*
X224_splice
C229W
Y236N
Y236S
C238Y
G245D
R248Q
R249S
X259_splice
G262V
R273C
R273H
C275Y
R282W
X307_splice
P53_TAD P53 transactivation motif
P53 P53 DNA-binding domain
other zinc binding site [ion binding]
other dimerization site [polypeptide binding]
DNA binding DNA binding site [nucleotide binding]
NLS Nuclear localization signal
P53_tetramer P53 tetramerization motif
MISSENSE
SPLICE
FRAMESHIFT
PROTEININS
NONSENSE
Somatic


Protein length
100
200
300
400
500
600
700
800
900
1000
1100
Burkitt lymphoma
8 mutations
1 disease
X204_splice
X284_splice
X424_splice
W477R
R634Q
R634*
I660R
Q769*
USP7
Diffuse large B-cell lymphoma
4 mutations
1 disease
Q8*
X284_splice
R745Nfs*5
X973_splice
COG5077 Ubiquitin carboxyl-terminal hydrolase [Posttranslational modification, ...
MATH_HAUSP Herpesvirus-associated ubiquitin-specific protease (HAUSP, also known ...
other substrate binding site [chemical binding]
peptidase_C19C A subfamily of Peptidase C19. Peptidase C19 contains ubiquitinyl hydro ...
HUBL-1 Ubl (Ubiquitin-like) domain 1
USP7_ICP0_bdg ICP0-binding domain of Ubiquitin-specific protease 7
HUBL-2 Ubl (Ubiquitin-like) domain 2
HUBL-3 Ubl (Ubiquitin-like) domain 3
USP7_C2 Ubiquitin-specific protease C-terminal
HUBL-4 Ubl (Ubiquitin-like) domain 4, together with HUBL-5 promotes ubiquitin ...
HUBL-5 Ubl (Ubiquitin-like) domain 5, together with HUBL-4 promotes ubiquitin ...
NONSENSE
MISSENSE
SPLICE
FRAMESHIFT
Somatic

Supplemental Figure S6



Supplemental Figure S7



References

1. Swerdlow SH, Campo E, Harris NL, et al. WHO classification of tumours of haematopoietic and lymphoid tissues. Lyon, France: International Agency for Research on Cancer; 2008.

2. Li H, Handsaker B, Wysoker A, et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.

3. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;

4. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*. 2015;31(12):2032–2034.

5. Richter J, Schlesner M, Hoffmann S, et al. Recurrent mutation of the ID3 gene in burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.* 2012;44(12):1316–1320.

6. Hezaveh K, Kloetgen A, Bernhart SH, et al. Alterations of microRNA and microRNA-regulated messenger RNA expression in germinal center b-cell lymphomas determined by integrative sequencing analysis. *Haematologica*. 2016;101(11):1380–1389.

7. Butterfield YS, Kreitzman M, Thiessen N, et al. JAGuaR: Junction alignments to genome for RNA-seq reads. *PLoS One*. 2014;9(7):e102398.

8. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589–595.

9. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764–770.

10. Saunders CT, Wong WSW, Swamy S, et al. Strelka: Accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*. 2012;28(14):1811–1817.

11. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156–2158.

12. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol.* 2016;17(1):122.

13. 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.

14. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* 2012;40(21):e169.

15. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* 2016;17(1):128.

16. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*. 2013;29(18):2238–2244.

17. Zhao H, Sun Z, Wang J, et al. CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. 2014;30(7):1006–1007.

18. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996–1006.

19. Arthur S, Jiang A, Grande B, et al. Genome-wide discovery of somatic coding and regulatory variants in diffuse large b-cell lymphoma. *bioRxiv*. 2017;225870.

20. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* 2004;32(Database issue):D493–6.

21. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for python. 2001;

22. Shirley MD, Ma Z, Pedersen BS, Wheelan SJ. Efficient “pythonic” access to FASTA files using pyfaidx. PeerJ PrePrints; PeerJ Inc. 2015.

23. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* 2013;3(1):246–259.

24. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220–1222.

25. Larson D, abelhj, Chiang C, et al. Hall-lab/svtools: Svtools v0.3.2. 2017;

26. Favero F, Joshi T, Marquard AM, et al. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* 2015;26:64–70.

27. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842.

28. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Res.* 2015;4:1521.

29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

30. Chu A, Robertson G, Brooks D, et al. Large-scale profiling of microRNAs for the cancer genome atlas. *Nucleic Acids Res.* 2016;44(1):e3.

31. Kozomara A, Griffiths-Jones S. miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 2014;42(Database issue):D68–73.

32. Kozomara A, Griffiths-Jones S. miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 2011;39(Database issue):D152–7.

33. Griffiths-Jones S, Saini HK, Dongen S van, Enright AJ. miRBase: Tools for microRNA genomics. *Nucleic Acids Res.* 2008;36(Database issue):D154–8.

34. Griffiths-Jones S, Grocock RJ, Dongen S van, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 2006;34(Database issue):D140–4.

35. Griffiths-Jones S. The microRNA registry. *Nucleic Acids Res.* 2004;32(Database issue):D109–11.

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

37. Bolotin DA, Poslavsky S, Davydov AN, et al. Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* 2017;35(10):908–911.

38. R Core Team. R: A language and environment for statistical computing. 2017;

39. Leiserson MDM, Wu H-T, Vandin F, Raphael BJ. CoMEt: A statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* 2015;16:160.

40. Leiserson M, Wu H-T, Vandin F, Raphael B. CoMEt: A statistical approach to identify combinations of mutually exclusive alterations in cancer. 2015.

41. Davis TL. Argparse: Command line optional and positional argument parser. 2018;

42. Waggott D, Haider S, C. Boutros P. Bedr: Genomic region processing using tools such as 'BEDTools', 'BEDOPS' and 'tabix'. 2017;

43. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 2009;4:1184–1191.

44. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*. 2005;21:3439–3440.

45. Xie Y. Bookdown: Authoring books and technical documents with R markdown. 2018.

46. Xie Y. Bookdown: Authoring books and technical documents with R markdown. Boca Raton, Florida: Chapman; Hall/CRC; 2016.

47. Robinson D. Broom: Convert statistical analysis objects into tidy data frames. 2017.

48. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014;30:2811–2812.

49. Wilke CO. Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. 2017.

50. Dowle M, Srinivasan A. Data.table: Extension of 'data.frame'. 2018;

51. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.

52. Wickham H, Francois R, Henry L, Müller K. Dplyr: A grammar of data manipulation. 2017.

53. Wickham H. Feather: R bindings to the feather 'API'. 2016.

54. Gohel D. Flextable: Functions for tabular reporting. 2018.

55. Wickham H. Forcats: Tools for working with categorical variables (factors). 2017.

56. Lawrence M, Huber W, Pagès H, et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 2013;9:

57. Clarke E, Sherrill-Mix S. Ggbeeswarm: Categorical scatter (violin point) plots. 2017.

58. Attali D, Baker C. GgExtra: Add marginal histograms to 'ggplot2', and more 'ggplot2' enhancements. 2018.

59. Wickham H. Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York; 2009.

60. Slowikowski K. Ggrepel: Repulsive text and label geoms for 'ggplot2'. 2017.

61. Ahlmann-Eltze C. Ggsignif: Significance brackets for 'ggplot2'. 2017.

62. Henry L, Wickham H, Chang W. Ggstance: Horizontal 'ggplot2' components. 2016.

63. Hahne F, Ivanek R. Statistical genomics: Methods and protocols. 2016;335–351.

64. Xie Y. Knitr: A General-Purpose package for dynamic report generation in R. 2018.

65. Xie Y. Dynamic documents with R and knitr. Boca Raton, Florida: Chapman; Hall/CRC; 2015.

66. Xie Y. Knitr: A comprehensive tool for reproducible research in R. *Implementing reproducible computational research*. 2014;

67. Wild F. Lsa: Latent semantic analysis. 2015.

68. Mayakonda A, Koeffler PH. Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. *BioRxiv*. 2016;

69. Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*. 2006;22:2059–2065.

70. Bengtsson H. MatrixStats: Functions that apply to rows and columns of matrices (and to vectors). 2018.

71. Kolde R. Pheatmap: Pretty heatmaps. 2015.

72. Gerds TA, Ozenne B. Publish: Format output of various routines in a suitable way for reports and publication. 2018;

73. Henry L, Wickham H. Purrr: Functional programming tools. 2018.

74. Neuwirth E. RColorBrewer: ColorBrewer palettes. 2014.

75. Wickham H, Hester J, Francois R. Readr: Read rectangular text data. 2017.

76. Wickham H, Bryan J. Readxl: Read excel files. 2017.

77. Maechler M, Rousseeuw P, Croux C, et al. Robustbase: Basic robust statistics. 2016.

78. Todorov V, Filzmoser P. An Object-Oriented framework for robust multivariate analysis. *J. Stat. Softw.* 2009;32(3):1–47.

79. Wickham H. Tidyverse: Easily install and load 'tidyverse' packages. 2017.

80. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Res.* 2015;4:

81. Garnier S. Viridis: Default color maps from 'matplotlib'. 2018.

82. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308–311.

83. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285–291.