Supplementary Appendix

Table of Contents

List of Investigators

Razvan I. Panea[1*], Cassandra J. Love[1*], Jennifer R. Shingleton[1*], Anupama Reddy[1*], Jeffrey A. Bailey[2], Ann M. Moormann[3], Juliana A. Otieno[4], John Michael Ong'echa[5], Cliff I. Oduor[5], Kristin M. S. Schroeder[1,6], Nestory Masalu[6], Nelson J. Chao[1], Megan Agajanian[7], M. Ben Major[7], Yuri Fedoriw[8], Kristy L. Richards[8], Grzegorz Rymkiewicz[9], Rodney R. Miles[10], Bachir Alobeid[11], Govind Bhagat[11], Christopher R. Flowers[12], Sarah L. Ondrejka[13], Eric D. Hsi[13], William W. L. Choi[14], Rex K. H. Au-Yeung[15,16], Wolfgang Hartmann[17], Georg Lenz[18], Howard Meyerson[19], Yen-Yu Lin[20], Yuan Zhuang[20], Micah A. Luftig[21], Alexander Waldrop[1], Tushar Dave[1], Devang Thakkar[1], Harshit Sahay[1], Guojie Li[1], Brooke C. Palus[1], Vidya Seshadri[21], So Young Kim[21], Randy D. Gascoyne[22], Shawn Levy[23], Minerva Mukhopadyay[24], David B. Dunson[24], Sandeep S. Dave[1]

1. Center for Genomic and Computational Biology and Department of Medicine, Duke University, Durham, NC, USA
2. Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, USA
3. Department of Medicine, University of Massachusetts, Worcester, MA, USA
4. Jaramogi Oginga Odinga Teaching and Referral Hospital, Ministry of Health, Kisumu, Kenya
5. Center for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya
6. Bugando Medical Center, Mwanza, Tanzania
7. Department of Cell Biology and Physiology, Washington University in St. Louis, St. Louis, MO, USA

8. University of North Carolina, Chapel Hill NC, USA

9. Poland Flow Cytometry Laboratory, Department of Pathology and Laboratory Diagnostics, Maria Sklodowska-Curie Institute – Oncology Center, Warsaw, Poland

10. Department of Pathology, University of Utah, Salt Lake City, UT, USA

11. Department of Pathology and Cell Biology, Columbia University, New York, NY, USA

12. Department of Hematology and Medical Oncology, Emory University, Atlanta, GA, USA

13. Department of Laboratory Medicine, Cleveland Clinic, Cleveland, OH, USA

14. Department of Pathology, Hong Kong Sanatorium & Hospital, Hong Kong, China

15. The University of Hong Kong, Queen Mary Hospital, Hong Kong, China

16. Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany

17. Division of Translational Pathology, Gerhard-Domagk-Institute of Pathology, University Hospital Münster, Münster, Germany

18. Medical Department A, Hematology, Oncology and Pneumology, University of Münster, Münster, Germany

19. Department of Pathology, Case Western Reserve University, Cleveland, OH, USA

20. Department of Immunology, Duke University, Durham, NC, USA

21. Department of Molecular Genetics and Microbiology, Duke University, Durham, NC, USA

22. Department of Pathology and Experimental Therapeutics, BC Cancer Agency & BC Cancer Research Centre, Vancouver, BC, Canada

23. Hudson Alpha Institute for Biotechnology, Huntsville, AL, USA

24. Department of Statistical Science, Duke University, Durham, NC, USA

* Contributed equally to this work

Human Subjects

The study was conducted entirely using already collected tumor and paired-normal samples. In most cases, these tissues were left over in the pathology archive after the diagnostic needs were satisfied. In addition, the tumor and normal cases collected from Africa were consented for genomic studies. All cases were de-identified prior to sequencing. All germline cases were used solely for the purpose of identifying somatic mutations with additional restrictions on data sharing that protect patient identity. This study was approved by the Duke University Health System institutional IRB (#Pro00016490).

Data Generation

FASTQ files were obtained from the sequencer and aligned to the genome (version hg38) using BWA-MEM (version 0.7.17-r1188) after removing reads that were indeterminate reads (strings of Ns) and adapter dimers. All cases were subjected to WGS, WES and RNAseq and some samples were sequenced multiple times. Table S1 provides sequencing reads and statistics for DNA sequencing reads which includes tumor and normal pairs for each patient. Somatic variant calling was performed with HaplotypeCaller in single-sample and joint genotype mode, and Mutect2 from GATK4 software using the default parameters. The obtained variant call format files were merged and normalized using bcftools and then annotated using Annovar v.2017Jul16.8 Next, the variants were filtered by first considering only the PASS filter in at least 1 sample. We removed variants found in the repetitive and low-complexity regions reported in RepeatMasker and genomic SuperDups databases, and we eliminated variants with a high population allele frequency (>0.01) reported in ExAC, gnomAD exome, and gnomAD genome databases. Finally, we filtered out variants that had a median base quality and average median mapping quality lower than 10. Whole genome sequencing data was used for primary discovery of variants. This was followed by interrogation of whole exome and transcriptomic sequencing data for those variants to better estimate the frequency of the events in our cases.

Whole Genome DNA, Exome, and Transcriptome Library Preparation

The BL cases comprised both formalin fixed paraffin embedded (FFPE, N= 65) and freshly frozen cases (N=36). Whole genome libraries were prepared using the KAPA HyperPrep kit. RNA libraries were prepared using the KAPA Stranded RNA-Seq Library Prep kit.

We[1] and others[2] have previously established the feasibility of next generation DNA and RNA sequencing in FFPE cases. We also carefully compared the mutational rates from FFPE and frozen cases and found them to be similar.

Approximately 50-300ng of DNA was enzymatically fragmented for 10-30 minutes using dsDNA Fragmentase (NEB, Ipswich, MA). Libraries were prepared using the KAPA Hyper Prep kit (Roche, Wilmington, MA) according to the manufacturer's specifications. Subsequent libraries were PCR amplified using 6 to 12 cycles of PCR, depending on available input. Subsequent whole genome libraries were sequenced on the Illumina TenX platform, 150 paired end. Exome libraries were captured as described previously[1].

100-1000 ng of total RNA was subjected to ribosomal depletion by hybridization, RNase H and DNase I digestion (NEB, Ipswich, MA) and bead purification. Library preparation was performed using the KAPA Stranded RNA-Seq Library Prep kit according to the manufacturer's specifications (Roche, Wilmington, MA). Multiplexed capture was performed on up to 24 RNA libraries using the SureSelectXT All Exon V6 + UTR bait set, according to the manufacturer's specifications (Agilent, Santa Clara, CA). Subsequent libraries were sequenced on the Illumina Hiseq 2500 V4 platform, 50 paired end.

EBV Identification

We obtained the unaligned read pairs from the tumor and normal samples and performed Diamond BLASTX on the Non-redundant (NR) protein sequences database. The output format of Diamond alignment was set for taxonomic classification ("-f 102"). Samples were considered EBV-positive if at least 5,000 reads or >10% of unmapped reads were classified as *Lymphocryptovirus* at the genus level (NCBI tax ids: 10375, 10376, 10377). EBV status calls were not made for samples with fewer than 10,000 input unmapped reads.

EBV subtype calls were made for EBV+ samples by determining the proportion of Type-1 (B95-8) vs. type-2 EBNA-2 (AG876) sequences among unmapped reads. Briefly, representative nucleotide sequences were downloaded in FASTA format from UniProt (Type 1: P12978, Type 2: Q69022). As above, reads were aligned to a custom diamond database containing only representative sequences from each EBV subtype (diamond makedb -in <ebna2.fa>) with the option –max-target-seqs set to 1 to return only the top alignment. Alignments were filtered to include only significant hits which were defined as >35 amino acids in length, >50% amino acid identity, and with a maximum e-value of 1.0e-5. After filtering, a custom R-script was then used to determine the number of reads in each sample mapping to each EBNA-2 gene variant. A subtype call was made if >99% of reads mapping to EBNA-2 mapped to a single subtype. All cases were verified using in situ hybridization for EBER where material was available.

CRISPR Screening

Lentiviral particles were transduced into each cell line in triplicate, and harvested 3 days post transduction (early) and 3 weeks post transduction (late). In each case, DNA was isolated and targeted sequencing of the guide RNA sequences was performed.

High throughput Illumina sequencing of sgRNA libraries amplified from plasmid sequences was used to determine sgRNA abundance for populations at each time-point. The median raw sequencing depth across sequencing libraries was 33,951,761 reads corresponding to 282 reads/sgRNA for a given library. sgRNA preprocessing, QC, quantification and computing gene knockout scores was performed similar to our previous publication[1]. Subsequent sgRNA abundance was assessed and compared between the two time points. Highly abundant sgRNAs were predicted to target tumor suppressor genes, whereas less abundant sgRNAs were predicted oncogenic.

Identification of essential BL genes

The effects of gene knockout were considered significant for a cell line if the probability of observing a larger difference in the mean sgRNA counts between early and late time points was <0.05 (after correcting for multiple comparisons) and the CRISPR gene score for that gene was >1 or <-1. For each gene in each cell line, a paired Mann-Whitney U test was used to determine whether the normalized sgRNA counts observed in early and late time points was significantly larger than would be expected from chance alone ($\alpha = 0.05$). Early and late sgRNA counts of the same sgRNA from the same replicate population were considered as pairs. Normalized sgRNA counts were log-transformed (log2 [1 + sgRNA count]) prior to significance testing to minimize the effects of outliers. Additionally, p-values were corrected for multiple comparisons using the

FDR method as implemented in the R statistical environment (p.adjust). A similar method was used to determine whether gene knockout resulted in a significant effect at the disease level. Genes were considered significant at the pan-Burkitt level if a significant effect was observed when replicates from multiple cell lines (BJAB, BL41, Jijoye) were considered as a single cell line and assessed for significance using paired Mann-Whitney U tests as above. A similar method was used to identify pan-DLBCL genes from DLBCL cell lines (HBL1, Ly3, SUDHL4). BL- and DLBCL-specific essential genes were identified by removing genes which had a significant effect in at least one cell line from each disease.

ID3 Affinity Purification, Mass Spectrometry and Protein Identification

ID3-FLAG was constructed by cloning *ID3* into the HindIII and BamHI sites in pLFLAG-N1; FLAG-ID3 was constructed by cloning *ID3* into the BglII and HindIII sites in pLFLAG-C1. Constructs were confirmed by Sanger sequencing. Subsequent constructs were packaged into retrovirus and transduced into cell lines. Overexpression of ID3-Flag in BJAB cells was confirmed by Western blot. Cells stably expressing FLAG-ID3 were lysed in 1% Triton X-100 lysis buffer, then incubated with FLAG resin, rotating at 4 °C for 1 hour. An on-bead digest was performed using trypsin (Promega, Madison, WI) and the FASP Protein Digestion Kit (Protein Discovery) at 37 °C overnight (16 hours). Eluted peptides were cleaned up with a C18 spin column (Peirce, Waltham, MA) and 3 ethyl acetate washes. To perform LC-MS, a reverse phase nano-HPLC using a nanoAquity UPLC system (Waters Corp., Milfor, MA) and Orbitrap fusion lumos mass spectrometer was utilized to perform mass spectral analysis. The raw mass spectrometry data was searched with MaxQuant.

Gene Editing by CRISPR

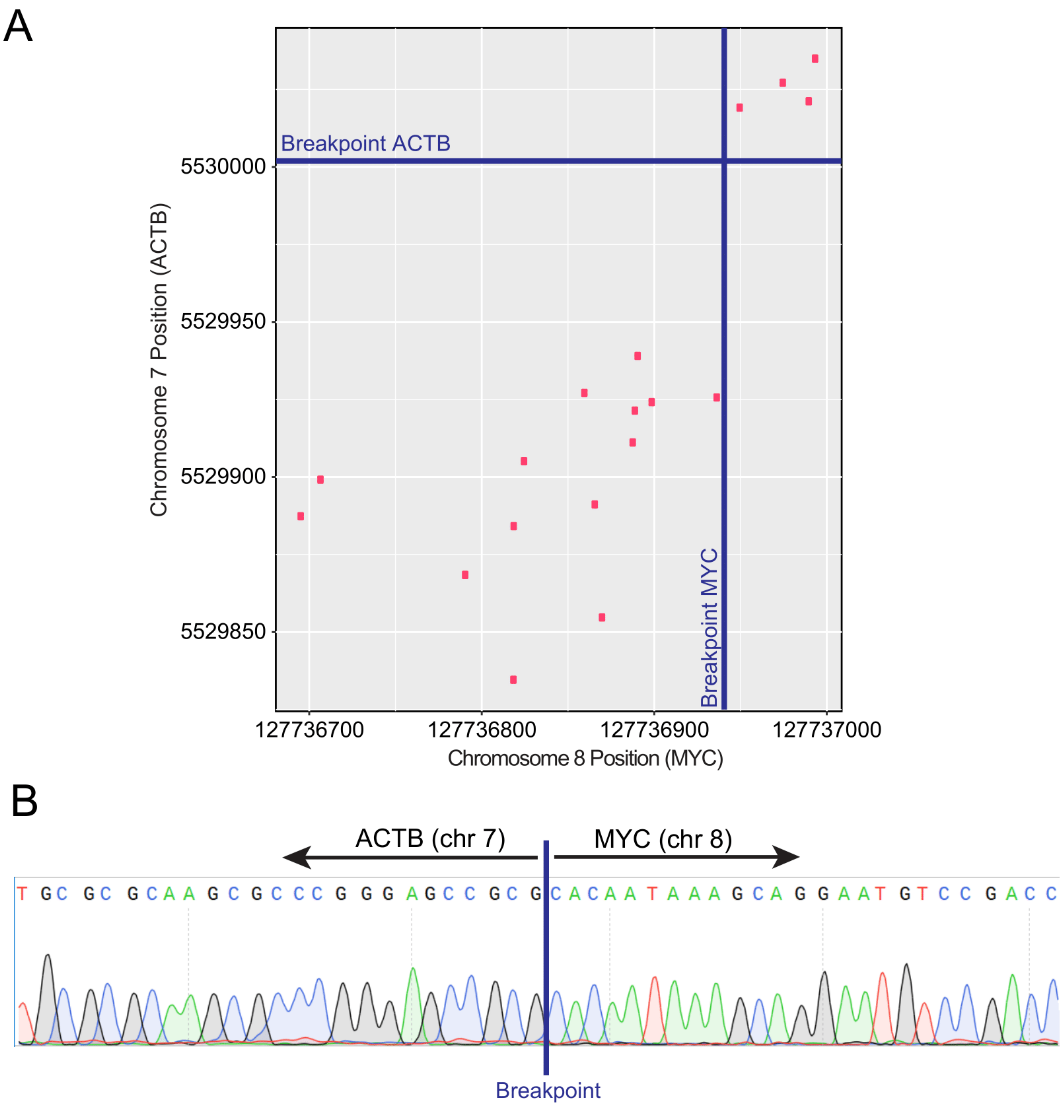
Guide RNA sequences were chosen targeting the helix loop helix region of the *ID3* and *TCF4* genes using the MIT CRISPR Design tool (http://crispr.mit.edu/) and designed for BbsI digestion. sgRNAs were annealed using T4 PNK, digested alongside pSpCas9 (BB)-2A-GFP (pX458) (Addgene, Plasmid #48138) using the BbsI, and ligated using T4 DNA ligase (NEB, Ipswich, MA). Constructs were transformed into ElectroMAX™ DH5α-E™ Cells (ThermoFisher Scientific, 11319019) and colonies confirmed for vector + insert by Sanger. Constructs were nucleofected into cell lines using the Amaxa Electroporator (Lonza, Basel, Switzerland). Cells were harvested and mixed with Nucleofector solution, 22% Supplement 1 (Cell Line Nucleofector® Kit V) containing 1 µg of vector. Cells single cell sorted based on GFP positivity. Clones were expanded for 3 weeks, and subsequent deletions in the *ID3* and *TCF4* genes were confirmed by Sanger sequencing.

BrdU Staining and Flow Cytometry Analysis

Approximately 5 x 10^5 cells were harvested and processed using reagents from the FITC BrdU Kit (Biolegend, San Diego, CA) according to the manufacturer's guidelines. Briefly, cells were subject to BrdU incorporation for 6 hours. Cells were in Cytofix/Cytoperm and DNase treated for 30 minutes followed by stained with FITC-anti-BrdU and 7-AAD. Samples were analyzed using a Sony flow cytometer machine (Sony Biotechnologies, SH800) with proper compensation for dual color analysis. Flowjo software was used to analyze the final data.

Validation of novel translocation t(8,7) – *MYC* and *ACTB*

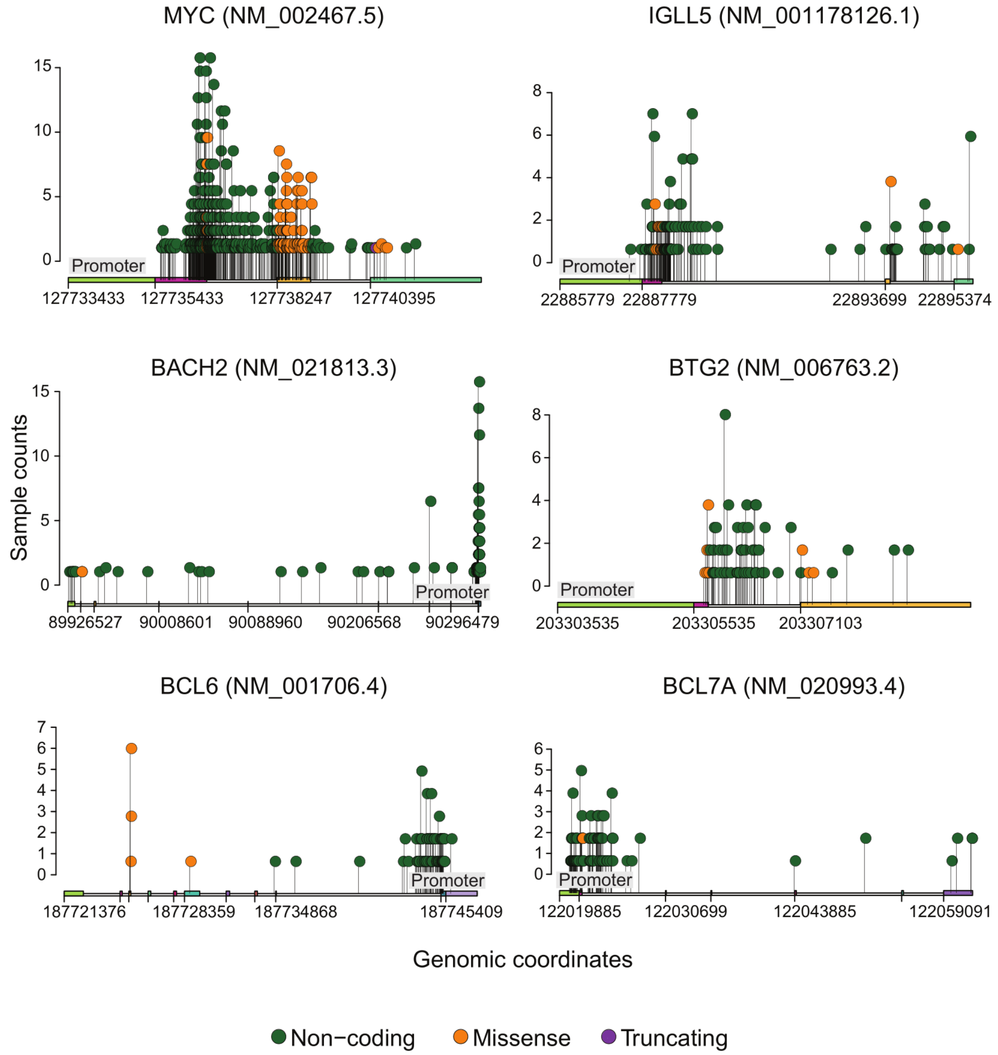
Through clustering discordant reads (i.e. read pairs whose mates map on different chromosomes), we identified a novel translocation between *MYC* (chr8) and *ACTB* (chr7) (**Figure S1A**). We used density based spectral clustering[3] to cluster the discordant reads using a Euclidean distance metric (max distance of 300 bases). The breakpoints were identified using chimeric reads in *MYC* and *ACTB* genes. We then validated the presence of this translocation using Sanger sequencing around the breakpoint region (**Figure S1B**).



Supplementary Figure S1: Validation of translocation t(8,7) – MYC and ACTB
(A) Scatter plot shows discordant reads that support the translocation, with breakpoints shown as a blue line. **(B)** Sanger sequencing plot shows ±25 base pairs around the breakpoint region (shown as a blue line).

Distribution of somatic variants in driver genes

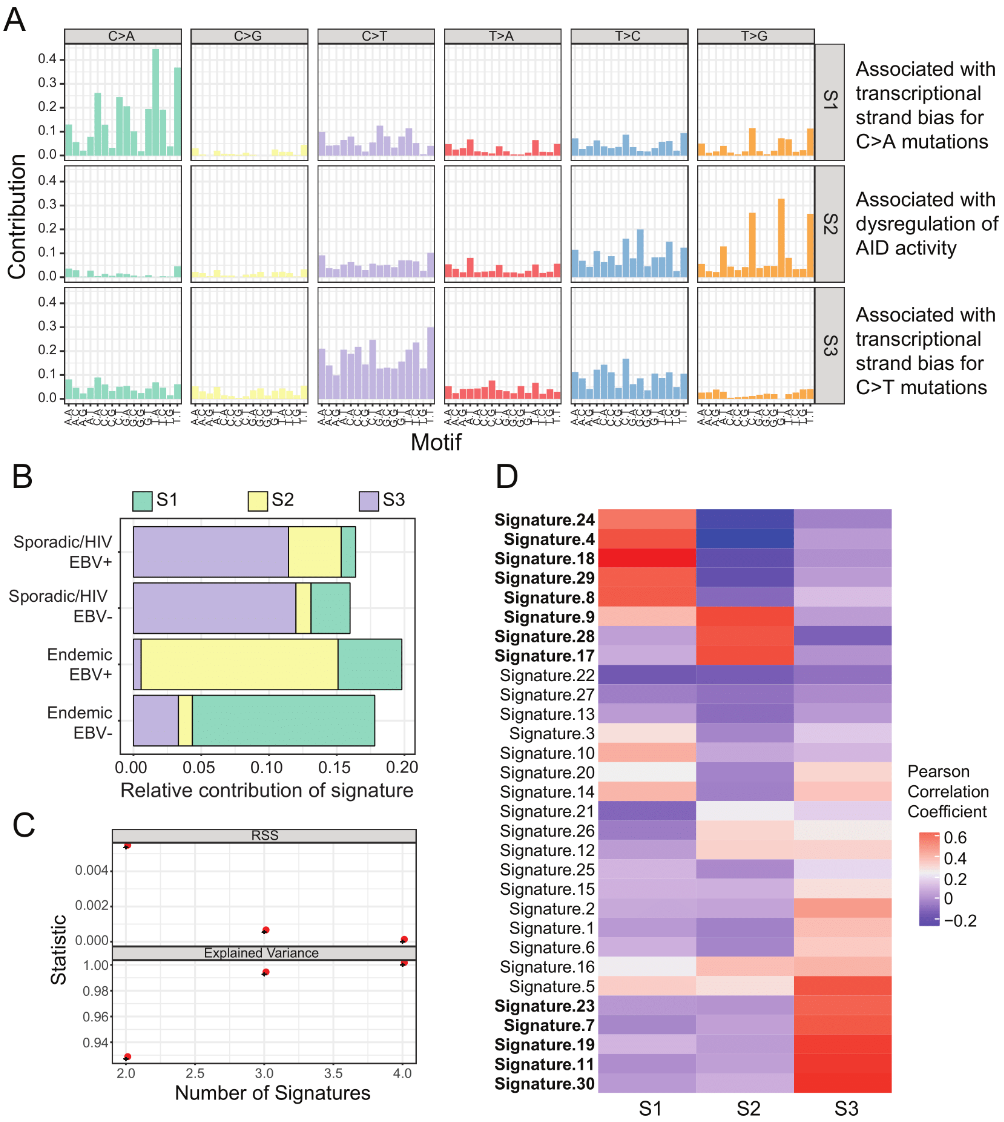
Variant stem plots for the top mutated driver genes with a high proportion of non-coding variants are shown in **Figure S2**. These plots show an enrichment of non-coding variants in the driver genes and highlight the importance of studying whole genomes in Burkitt lymphoma. Variants are plotted across the genomic coordinates (x-axis) with their height showing the number of mutated Burkitt lymphoma samples[4]. The variants are colored based on their functional annotation (missense, truncating or non-coding). The colored bars along the x-axis show the promoter and the exons.



Supplementary Figure S2: Distribution of somatic variants in driver genes
Variant-stem plots for selected driver genes show the distribution of somatic variants across the genomic coordinates on the x-axis and sample counts for each of the variants on the y-axis. The colored bars on the x-axis show the promoter and the exons. The promoter regions are also labeled. Variants are colored based on their function.

Mutational signatures analysis

We have analyzed the somatic variants identified in the 101 Burkitt lymphoma samples to infer mutational signatures[5] and to highlight processes that lead to tumorigenesis. We initially inferred somatic signatures in 6 groups (endemic, sporadic and HIV with their EBV status) and observed that sporadic and HIV cases shared the same signature profiles, and thus combined them to increase statistical power. We reanalyzed mutation signatures from the 4 groups (Endemic_EBV+, Endemic_EBV-, Sporadic/HIV_EBV+, Sporadic/HIV_EBV-) and identified 3 signatures which explained >99% of the variance (**Figure S3A, B**). These 3 inferred signatures (S1, S2 and S3) were correlated to the COSMIC database of known signatures[5], and are associated with distinct mechanisms. S1 is associated with transcriptional strand bias for C>A mutations, S2 is associated with dysregulation of AID activity, and S3 is associated with transcriptional strand bias for C>T mutations.

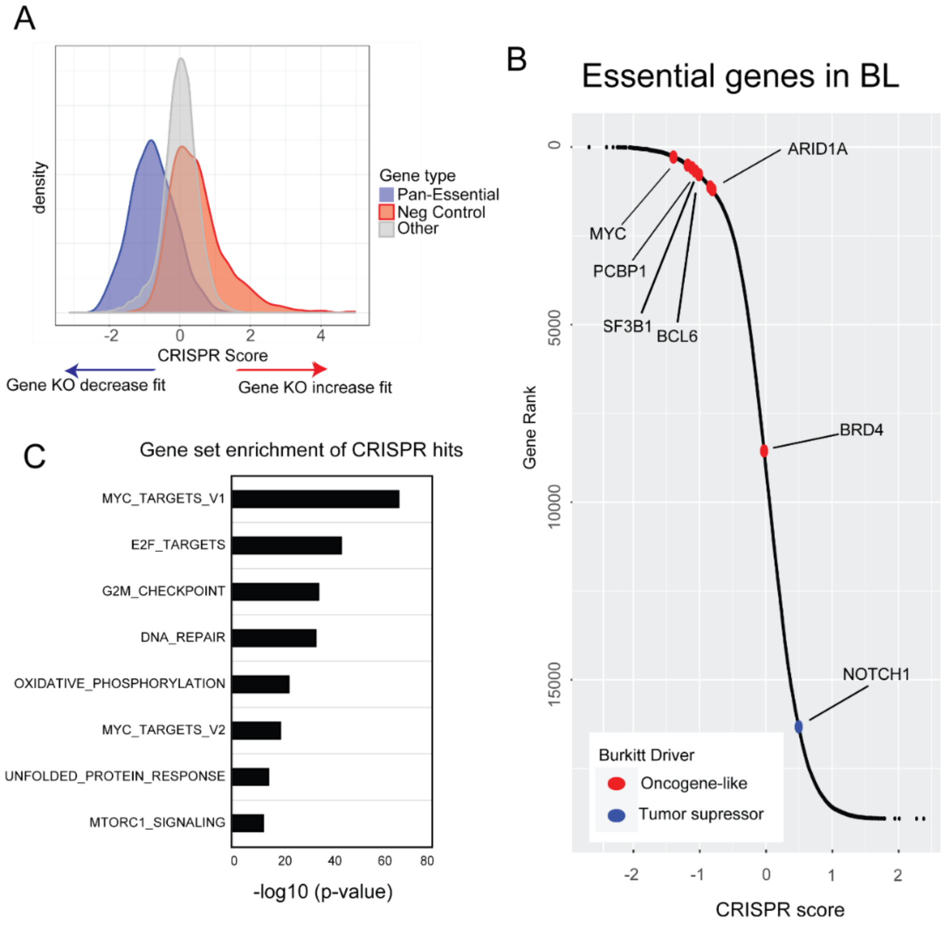


Supplementary Figure S3: Mutational signatures analysis **(A)** Mutational spectrum of the three mutational signatures computed from the somatic variants. (**B)** Barplots representing the contribution of the somatic mutations signatures. **(C)** Scatter plots presenting the residual sum of squares and the explained variance of NMF decomposition with different numbers of signatures. **(D)** Heatmap shows Pearson correlation between the three signatures and the annotated mutation signatures from COSMIC database.

CRISPR screen analysis

We performed an unbiased CRISPR screen targeting 19,050 genes to identify essential genes in three BL cell lines (BJAB, BL41, Jijoye). Gene knockout effects were consistent with biological expectation and results from previous CRISPR screens[6,7]. As shown in **Figure S4A**, sgRNAs targeting genes that had been previously demonstrated to be essential across multiple human cell types (shown in blue) resulted in significantly decreased cell fitness relative to randomly chosen sgRNAs (shown in gray). By contrast, non-targeting control sgRNAs (shown in red) showed a significant increase in cell fitness relative to randomly chosen sgRNAs.

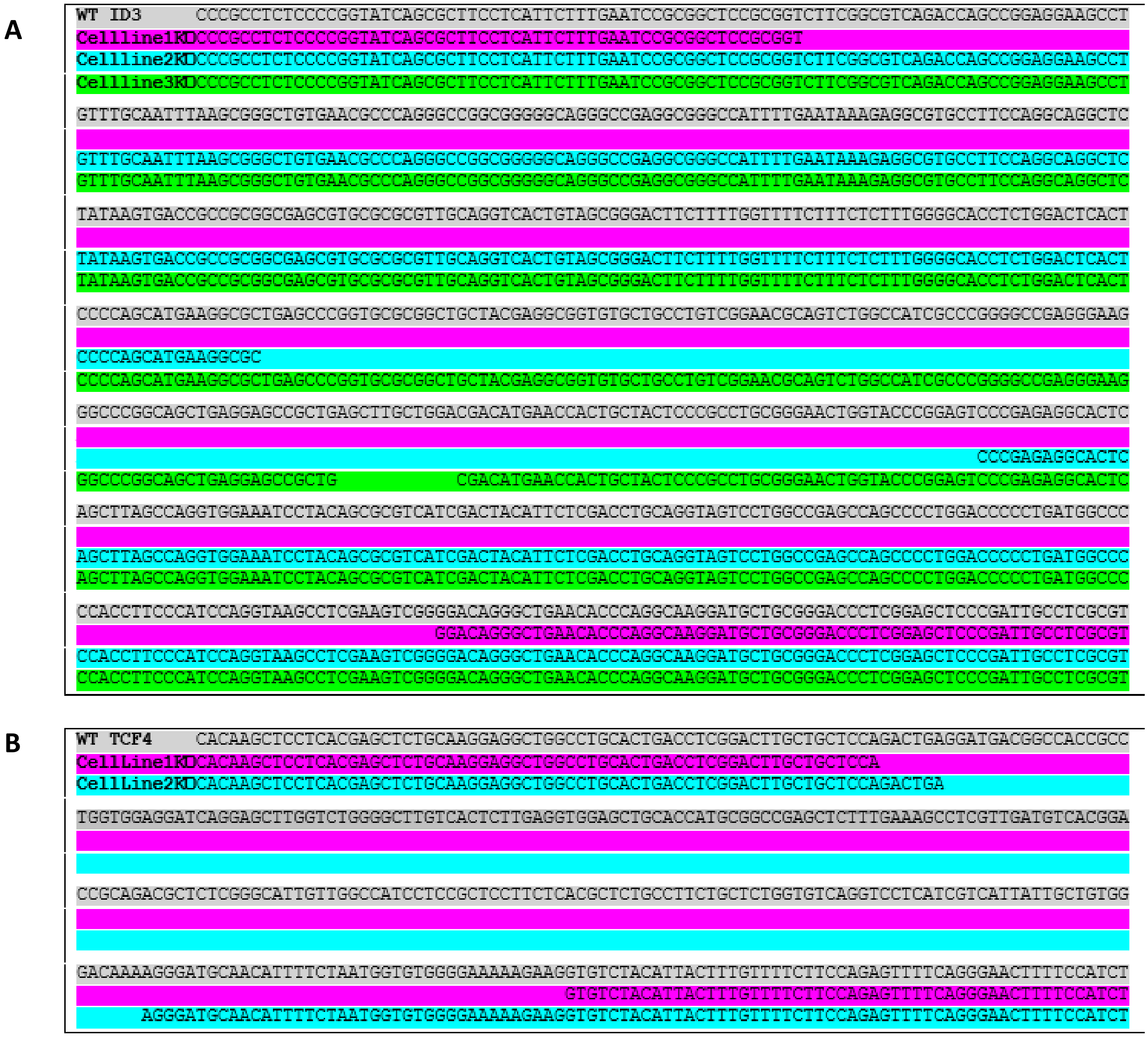
We identified 889 essential genes whose silencing resulted in significantly decreased proliferation in the BL cell lines (**Figure S4B**). Driver genes are highlighted in the plot. We then performed a gene set enrichment analysis[8] on these 889 essential genes using Msigdb Hallmarks gene sets and identified *MYC*, E2F targets, cell cycle and DNA repair as the top results (**Figure S4C**).



Supplementary Figure S4: Essential genes identified using CRISPR screens in BL. **(A)** Histogram of CRISPR scores shows that pan-essential genes have lower scores compared to negative controls. **(B)** Waterfall plot showing genes ranked by CRISPR score. Essential genes which are also BL drivers are highlighted. **(C)** Gene set enrichment performed on CRISPR hits using Msigdb Hallmarks gene sets are shown as a barplot.

Confirmatory data of *ID3* and *TCF4* CRISPR engineered deletions

We used CRISPR methods to introduce deletion events in *ID3* and *TCF4* in order to disrupt function. We successfully engineered 3 cell lines to have large deletions (ranging from 11 to 548 base pairs) in *ID3* (**Figure S5A**). Similarly, we engineered 2 cell lines to have large deletions (ranging from 217 to 261 base pairs) in *TCF4* (**Figure S5B).** Western blot analysis confirms loss of protein in *ID3* CRISPR engineered cells (**Figure S5C**).



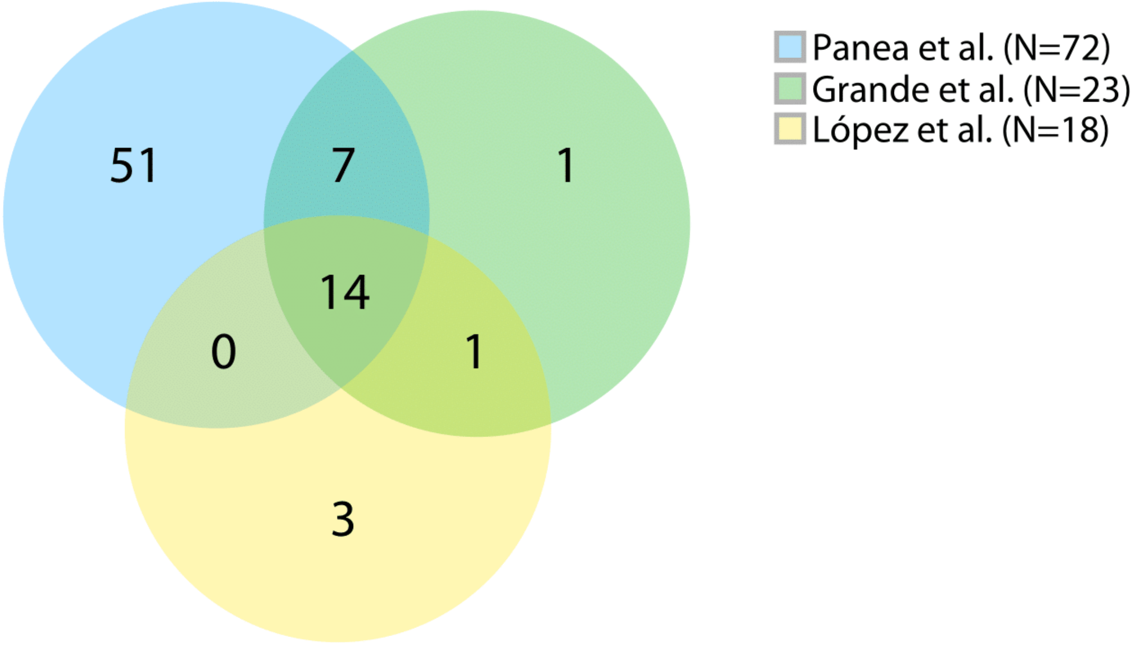
```
WT ID3     CCCGCCTCTCCCCGGTATCAGCGCTTCCTCATTCTTTGAATCCGCGGCTCCGCGGTCTTCGGCGTCAGACCAGCCGGAGGAAGCCT
Cellline1KOCCCGCCTCTCCCCGGTATCAGCGCTTCCTCATTCTTTGAATCCGCGGCTCCGCGGT
Cellline2KOCCCGCCTCTCCCCGGTATCAGCGCTTCCTCATTCTTTGAATCCGCGGCTCCGCGGTCTTCGGCGTCAGACCAGCCGGAGGAAGCCT
Cellline3KOCCCGCCTCTCCCCGGTATCAGCGCTTCCTCATTCTTTGAATCCGCGGCTCCGCGGTCTTCGGCGTCAGACCAGCCGGAGGAAGCCT

GTTTGCAATTTAAGCGGGCTGTGAACGCCCAGGGCCGGCGGGGGCAGGGCCGAGGCGGGCCATTTTGAATAAAGAGGCGTGCCTTCCAGGCAGGCTC

GTTTGCAATTTAAGCGGGCTGTGAACGCCCAGGGCCGGCGGGGGCAGGGCCGAGGCGGGCCATTTTGAATAAAGAGGCGTGCCTTCCAGGCAGGCTC
GTTTGCAATTTAAGCGGGCTGTGAACGCCCAGGGCCGGCGGGGGCAGGGCCGAGGCGGGCCATTTTGAATAAAGAGGCGTGCCTTCCAGGCAGGCTC

TATAAGTGACCGCCGCGGCGAGCGTGCGCGCGTTGCAGGTCACTGTAGCGGGACTTCTTTTGGTTTTCTTTCTCTTTGGGGCACCTCTGGACTCACT

TATAAGTGACCGCCGCGGCGAGCGTGCGCGCGTTGCAGGTCACTGTAGCGGGACTTCTTTTGGTTTTCTTTCTCTTTGGGGCACCTCTGGACTCACT
TATAAGTGACCGCCGCGGCGAGCGTGCGCGCGTTGCAGGTCACTGTAGCGGGACTTCTTTTGGTTTTCTTTCTCTTTGGGGCACCTCTGGACTCACT

CCCCAGCATGAAGGCGCTGAGCCCGGTGCGCGGCTGCTACGAGGCGGTGTGCTGCCTGTCGGAACGCAGTCTGGCCATCGCCCGGGGCCGAGGGAAG

CCCCAGCATGAAGGCGC
CCCCAGCATGAAGGCGCTGAGCCCGGTGCGCGGCTGCTACGAGGCGGTGTGCTGCCTGTCGGAACGCAGTCTGGCCATCGCCCGGGGCCGAGGGAAG

GGCCCGGCAGCTGAGGAGCCGCTGAGCTTGCTGGACGACATGAACCACTGCTACTCCCGCCTGCGGGAACTGGTACCCGGAGTCCCGAGAGGCACTC

                                                                                    CCCGAGAGGCACTC
GGCCCGGCAGCTGAGGAGCCGCTG           CGACATGAACCACTGCTACTCCCGCCTGCGGGAACTGGTACCCGGAGTCCCGAGAGGCACTC

AGCTTAGCCAGGTGGAAATCCTACAGCGCGTCATCGACTACATTCTCGACCTGCAGGTAGTCCTGGCCGAGCCAGCCCCTGGACCCCCTGATGGCCC

AGCTTAGCCAGGTGGAAATCCTACAGCGCGTCATCGACTACATTCTCGACCTGCAGGTAGTCCTGGCCGAGCCAGCCCCTGGACCCCCTGATGGCCC
AGCTTAGCCAGGTGGAAATCCTACAGCGCGTCATCGACTACATTCTCGACCTGCAGGTAGTCCTGGCCGAGCCAGCCCCTGGACCCCCTGATGGCCC

CCACCTTCCCATCCAGGTAAGCCTCGAAGTCGGGGACAGGGCTGAACACCCAGGCAAGGATGCTGCGGGACCCTCGGAGCTCCCGATTGCCTCGCGT
                                 GGACAGGGCTGAACACCCAGGCAAGGATGCTGCGGGACCCTCGGAGCTCCCGATTGCCTCGCGT
CCACCTTCCCATCCAGGTAAGCCTCGAAGTCGGGGACAGGGCTGAACACCCAGGCAAGGATGCTGCGGGACCCTCGGAGCTCCCGATTGCCTCGCGT
CCACCTTCCCATCCAGGTAAGCCTCGAAGTCGGGGACAGGGCTGAACACCCAGGCAAGGATGCTGCGGGACCCTCGGAGCTCCCGATTGCCTCGCGT
```

```
WT TCF4    CACAAGCTCCTCACGAGCTCTGCAAGGAGGCTGGCCTGCACTGACCTCGGACTTGCTGCTCCAGACTGAGGATGACGGCCACCGCC
CellLine1KOCACAAGCTCCTCACGAGCTCTGCAAGGAGGCTGGCCTGCACTGACCTCGGACTTGCTGCTCCA
CellLine2KOCACAAGCTCCTCACGAGCTCTGCAAGGAGGCTGGCCTGCACTGACCTCGGACTTGCTGCTCCAGACTGA

TGGTGGAGGATCAGGAGCTTGGTCTGGGGCTTGTCACTCTTGAGGTGGAGCTGCACCATGCGGCCGAGCTCTTTGAAAGCCTCGTTGATGTCACGGA



CCGCAGACGCTCTCGGGCATTGTTGGCCATCCTCCGCTCCTTCTCACGCTCTGCCTTCTGCTCTGGTGTCAGGTCCTCATCGTCATTATTGCTGTGG



GACAAAAGGGATGCAACATTTTCTAATGGTGTGGGGAAAAAGAAGGTGTCTACATTACTTTGTTTTCTTCCAGAGTTTTCAGGGAACTTTTCCATCT
                                             GTGTCTACATTACTTTGTTTTCTTCCAGAGTTTTCAGGGAACTTTTCCATCT
      AGGGATGCAACATTTTCTAATGGTGTGGGGAAAAAGAAGGTGTCTACATTACTTTGTTTTCTTCCAGAGTTTTCAGGGAACTTTTCCATCT
```

Supplementary Figure S5: Confirmation of CRISPR engineered deletions in ID3 and TCF4 in cell lines. Alignment of wild type and CRISPR engineered knock out in **(A)** *ID3* and **(B)** *TCF4*.

Comparison of driver genes in Burkitt lymphoma whole genome studies

We compared our driver genes list with recent whole genome studies[9,10] and we found an overlap of 21 out of the 23 significant genes from *Grande et al*. and 14 out of 18 significant genes from *López et al*. Our study did not identify as driver genes 2 genes (*PCBP1*, *P2RY8*) from *Grande* et al and 4 genes (*PCBP1, E2F2, ADNP, HNRNDP*) from *López et al*, while we identified an additional set of 51 mutated driver genes, including *MYC*, *IGLL5* and *BACH2*. **Figure S6** shows the comparison as a Venn diagram.



Supplementary Figure S6: Driver gene comparison with previous studies. Venn diagram shows the overlap between the identified driver genes by the current article, *Grande et al*. and *López et al*.

Identification of 11q gains/losses in BL samples

We have identified one sporadic case to have 11q24.2-qter loss, but no copy number gains for 11q23.2-q23.3. Additionally, we have identified an HIV-positive case to have 11q23.2-q23.3 copy number gains, but no telomeric losses of 11q24.1-qter.

Mutational distributions for Frozen vs. FFPE samples

We have extensively compared the application of next generation sequencing in frozen and paraffin cases in our previous work with diffuse large B cell lymphoma[1]. Here, we compared the number of driver variants in FFPE vs. Frozen tumors to evaluate potential biases introduced by sequencing FFPE cases. In all, we had 65 FFPE cases and 36 frozen cases. The mean number of somatic and driver events in both groups were highly similar (P>0.4, Wilcoxon rank sum test).

Supplemental Figures

Supplemental Tables

Table S1: Sample annotations for the 101 Burkitts patients in our study. Mapping rate and PCR duplication rate are calculated after trimming and alignment.

Table S2: Filtered somatic variants for the 101 Burkitt lymphoma samples.

Table S3: Genomic clusters, of the 228,010 somatic variants, that contain at least 4 variants and 3 unique samples.

Table S4: Sanger validation for driver gene variants.

Table S5: Gene-level genetic alterations for the 101 samples.

Table S6: Gene-level mutational associations.

Table S7: Differential gene expression associated with driver gene mutations.

Table S8: Genesets associated with driver gene mutations.

Table S9: Comparison of differential gene expression for Burkitts vs. DLBCL.

Table S10: Comparison of differential genesets for Burkitts vs. DLBCL.

Table S11: CRISPR score for each gene by cellline.

Table S12: Oligonucleotides used in the study.

References

1. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481-494. e15 (2017).
2. Chapuy, B. *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature medicine* (2018).
3. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Institute for Computer Science, University of Munich. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).* (1996).
4. Ou, J., Wang, Y. & Zhu, L. trackViewer: A R/Bioconductor package for drawing elegant interactive tracks or lollipop plot to facilitate integrated analysis of multi-omics data. . *R package version 1.16.0.* (2018).
5. Alexandrov, L.B. *et al.* Clock-like mutational processes in human somatic cells. *Nature Genetics* **47**, 1402 (2015).
6. Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular Systems Biology* **10**, 733-733 (2014).
7. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3: Genes|Genomes|Genetics* **7**, 2719 (2017).
8. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
9. Grande, B.M. *et al.* Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood* **133**, 1313 (2019).
10. López, C. *et al.* Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nature Communications* **10**, 1459 (2019).