Supporting Information

Lohr et al. 10.1073/pnas.1121343109

SI Materials and Methods

Sample Selection and Quality Assessment of DNA and Tumor Purity. This study was reviewed and approved by the human subjects review board of the Mayo Clinic, the University of Iowa, and the Broad Institute, and written informed consent was obtained from all participants. All subjects were from the Molecular Epidemiology Resource of the University of Iowa/Mayo Clinic Specialized Program of Research Excellence (1). Since 2002, consecutive newly diagnosed patients with non-Hodgkin lymphoma (within 9 mo) who were 18 y and older, residents of the United States, and had no history of HIV infection have been enrolled. All pathologic findings were reviewed by a lymphoma hematopathologist to verify the diagnosis and to classify each case according to World Health Organization classification. Paired peripheral blood and frozen tumor tissue were available on 55 patients with diffuse large B-cell lymphoma (DLBCL). DNA from the peripheral blood sample was extracted by using an automated platform (AutoGen FlexStar; Qiagen chemistries), whereas DNA from frozen tumor tissue was extracted by using the Puregene kit (Qiagen). DNA concentrations were measured using PicoGreen dsDNA Quantitation Reagent (Invitrogen). DNA sample quality was assessed by gel electrophoresis. The identities of all tumor and normal DNA samples were confirmed by MS fingerprint genotyping of 24 common SNPs (Sequenom).

Whole-Exome Capture Library Construction. For whole-exome capture library construction, we followed the procedure described (2, 3), with production-scale exome capture library construction. Exome targets were generated based on CCDS+RefSeq genes (http://www.ncbi.nlm.nih.gov/projects/CCDS/ and http://www.ncbi.nlm.nih.gov/RefSeq/), representing 188,260 exons from approximately 18,560 genes (93% of known, nonrepetitive protein coding genes and spanning ~1% of the genome). DNA oligonucleotides were amplified by PCR and subjected to in vitro transcription in the presence of biotinylated UTP to generate single-stranded RNA “baits.” Genomic DNA from primary tumor and matched blood normal was sheared and ligated to Illumina sequencing adapters including 8-bp indexes. Adaptor ligated DNA (i.e., “pond”) was then size-selected for lengths between 200 and 350 bp and hybridized with an excess of bait in solution phase as described previously (2). The “catch” was pulled down by streptavidin beads and eluted as described earlier. Barcoded exon capture libraries were then pooled into batches and sequenced on Illumina HiSeq instruments (76-bp paired-end reads) (2, 3). The 8-bp index was used to distribute sequencing reads to sample in the downstream data aggregation pipeline.

Massively Parallel Sequencing. Sequencing libraries were quantified by using a SYBR Green quantitative PCR (qPCR) protocol with specific probes complementary to adapter sequence. The qPCR assay measures the quantity of fragments properly “adapter-ligated” that are appropriate for sequencing. Based on the qPCR quantification, libraries were normalized to 2 nM and then denatured by using 0.1 N NaOH. Cluster amplification of denatured templates was performed according to manufacturer protocol (Illumina). SYBR Green dye was added to all flow cell lanes to provide a quality control checkpoint after cluster amplification and to ensure optimal cluster densities on the flow cells. Paired-end sequencing (2 × 76 bp) was carried out by using HiSeq sequencing instruments; the resulting data were analyzed with the current Illumina pipeline. Standard quality control metrics, including error rates, percentage passing filter reads, and total Gb produced, were used to characterize process performance before downstream analysis. The Illumina pipeline generates data files (BAM files) that contain the reads together with quality parameters.

Sequence Data Processing. Massively parallel sequencing data were processed using two consecutive pipelines:

First, the sequencing data processing pipeline, called Picard, developed by the Sequencing Platform at the Broad Institute, starts with the reads and qualities produced by the Illumina software for all lanes and libraries generated for a single sample (either tumor or normal) and produces, at the end of the pipeline, a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) representing the sample. The final BAM file stores all reads with well-calibrated qualities together with their alignments to the genome (for reads only that were successfully aligned).

Second, the Broad Cancer Genome Analysis pipeline, also known as Firehose, starts with the BAM files for each DLBCL sample and matched normal sample from peripheral blood (hg19), and performs various analyses, including quality control, local realignment, mutation calling, small insertion and deletion identification, rearrangement detection, coverage calculations, and others. The details of our sequencing data processing have been described elsewhere (2, 3).

Calculation of Sequence Coverage, Mutation Calling, and Significance Analysis. Somatic single-nucleotide variations were detected using *MuTect* (2), and we evaluated the fraction of all bases suitable for mutation calling whereby a base is defined as covered if at least 14 and eight reads overlapped the base in the tumor and in the germline sequencing, respectively. Passing single nucleotide variants found within coding areas of the genome were annotated for the chromosomal location, the type of the variant, the codon change and the change in the protein sequence (Ramos et al., unpublished work). Insertions and deletions in coding areas (both frameshift and in-frame) were detected by using the algorithm Indelocator (refs. 2, 3 and Sivachenko et al., unpublished work).

The ranking of genes in terms of estimated conferred selective advantage was performed by using the mutation statistical analysis algorithm *MutSig* (Lawrence et al., unpublished work). The *MutSig* algorithm works with an aggregated list of mutations across the entire patient set, and estimates the background mutation rate. The P and q_1-values for a certain gene are determined for the mutation rate observed in that gene in relation to the background model. *MutSig* uses various factors to accurately estimate the background mutation rate, taking into account the background mutation rates of different mutation categories (i.e., transitions or transversions in different sequence contexts), as well as the fact that different samples have different background mutation rates. It then uses convolutions of binomial distributions to calculate the P and q_1-values for each gene, which represents the probability that we obtain the observed or a more significant set of mutations in a gene by chance, given the background model. For the complete aggregated set of somatic mutations across all patients, we ran two *MutSig* analyses.

The first analysis took into account the observed number of nonsilent mutations per patient per gene, the nonsynonymous to synonymous mutation ratio for each gene, as well as the expression level, to identify genes having a large number of nonsynonymous events compared with the number of synonymous events, as well as compared with the number expected from the background mutation rate estimated from genes of similar expression level, in order to account for the decreased levels of transcription-coupled repair and resultant increase in mutation rate, in unexpressed genes. We used expression data from a previous study for this analysis (4).

The second part of the analysis determined the significance of mutations by their positional clustering, their conservation relative to other sites in the gene, as well as the significance of the joint effect of these two factors (conservation and clustering), expressed by a joint q_2-value. The final joint q_2-value expresses the probability by chance that the mutations in a gene are with the observed positional configuration and conservation values, or a more significant outcome than the observed one. The aim of this analysis is to discover novel mutational hotspots that are important to carcinogenesis.

The *BCL2* mutation statistics [(*i*), nonuniform distribution of silent mutations; (*ii*), enrichment of silent mutations near the 5′ end of the gene; and (*iii*), enrichment of nonsilent mutations outside the BH domains] were all calculated by performing permutations and comparing the observed value for each metric to the resulting null distribution. The permutations were performed taking into account the base composition of the gene and the categories of the mutations observed.

The *xvar* algorithm at http://mutationassessor.org/ was used.

Validation of Selected Mutations by Resequencing. Validation of selected mutations was performed by targeted resequencing using microfluidic PCR (Access array system, Fluidigm) and the MiSeq sequencing system (Illumina). Tumor and matched normal samples were selected based on the presence of the indicated mutations by whole exome sequencing. Target specific primers were designed to flank sites of interest and produce amplicons of 200 bp ± 20 bp. Molecularly barcoded, Illumina-compatible specific oligos, containing sequences complementary to the primer tails were added to the access array chip in the same well as the genomic DNA samples (20–50 ng of input) such that all amplicons for a given genomic sample share the same index. PCR was performed on the Fluidigm access array according to the manufacturer's instructions. Indexed libraries were recovered for each sample in a single collection well from the Fluidigm chip, quantified using picogreen, and then normalized for uniformity across libraries. Resulting normalized libraries were loaded on the MiSeq instrument and sequenced using paired end 150 bp sequencing reads.

PCR. A PCR assay was used for detection of the *t*(14;18) translocation, which targets the joining region of the *IgH* gene, and distinct regions of the *BCL2* (InVivoScribe Technologies). Three individual PCR reactions were used to detect breakpoints in the major breakpoint region and minor cluster region of the *BCL2* *t*(14;18) translocations. The fourth PCR, the specimen control size ladder, targets multiple genes and generates a series of amplicons of 100, 200, 300, 400, and 600 bp to ensure that the quality and quantity of input DNA is adequate to yield a valid result. This PCR approach captures 80–90% of all *BCL2*/*IgH* rearrangements. PCR conditions were used as reported previously (5).

1. Drake MT, et al. (2010) Vitamin D insufficiency and prognosis in non-Hodgkin's lymphoma. *J Clin Oncol* 28:4191–4198.
2. Chapman MA, et al. (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature* 471:467–472.
3. Stransky N, et al. (2011) The mutational landscape of head and neck squamous cell carcinoma. *Science* 333:1157–1160.
4. Lenz G, et al.; Lymphoma/Leukemia Molecular Profiling Project (2008) Stromal gene signatures in large-B-cell lymphomas. *N Engl J Med* 359:2313–2323.
5. van Dongen JJ, et al. (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* 17:2257–2317.

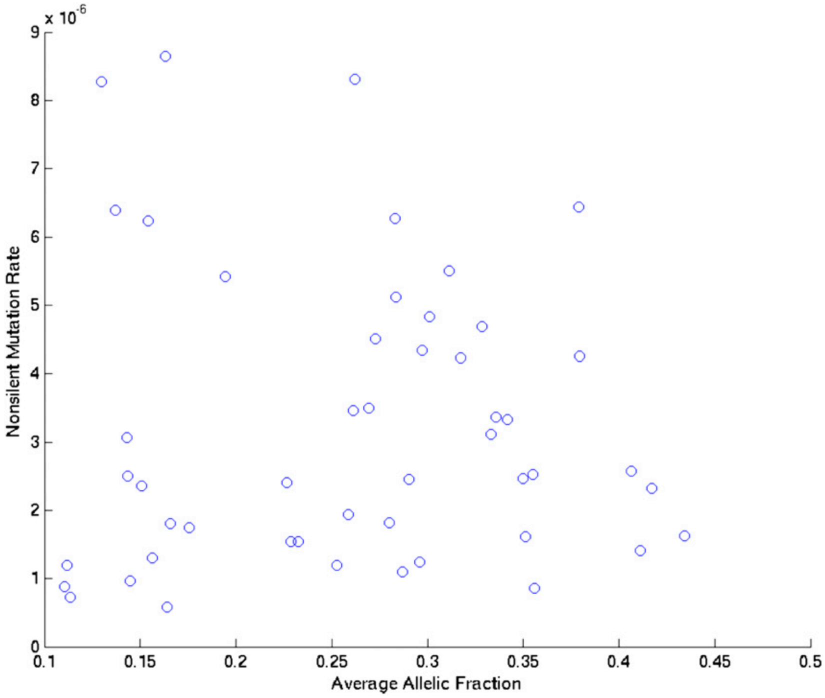

Fig. S1. Number of nonsilent mutations plotted over the average allelic fraction of 49 DLBCL samples used for analysis. Each circle represents one individual patient sample.

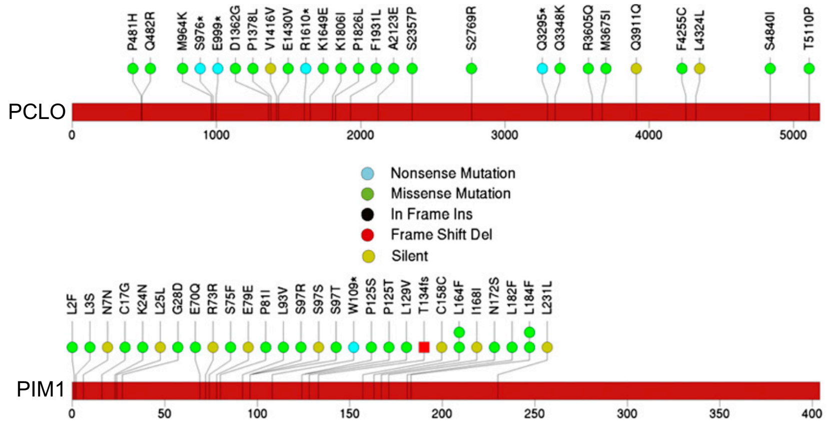

Fig. S2. Sites of somatic mutations in significantly mutated genes. A diagram of the relative positions of somatic mutations is shown for *PCLO* and *PIM1*. The type of the mutation is indicated in the legend.

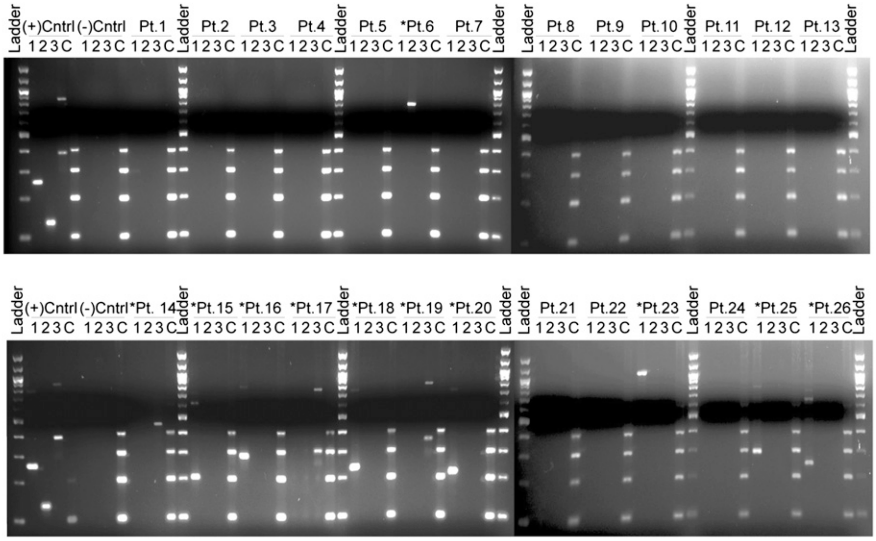

Fig. S3. PCR results for *BCL2 t*(14;18) translocation. *BCL2/IgH* rearrangement was tested by PCR in all 13 patients of our cohort with at least one *BCL2* mutation (patients 14–26; *Lower*) and 13 randomly selected patients without *BCL2* mutations (patients 1–13; *Upper*). Four individual PCR reactions were performed for every patient. The first three lanes for every patient represent three different PCR assays to detect different breakpoints, which result in differing sizes of the bands. The fourth lane is a control PCR for DNA integrity. The data shows that 10 of 13 patients with *BCL2* mutations also have *BCL2/IgH* rearrangements, but only one of 13 patients without *BCL2* mutations has a *BCL2/IgH* rearrangement (*$P < 0.005$, Fisher exact test).

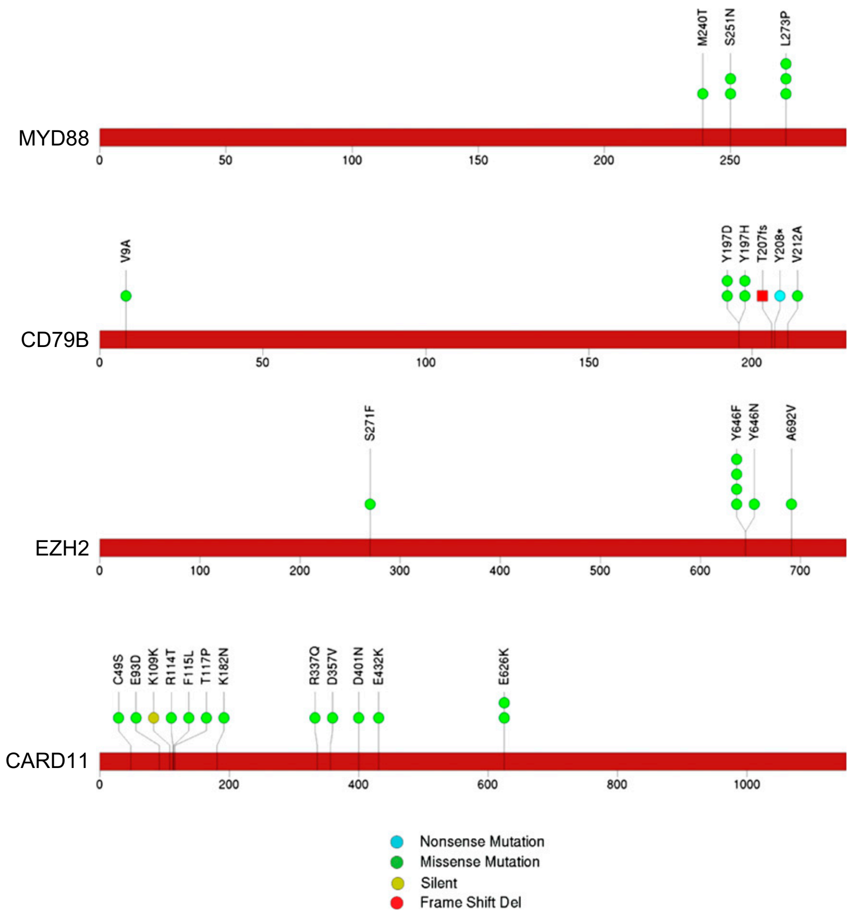

Fig. S4. Clustering of mutations in genes. The relative positions in the protein of somatic mutations are shown for *MYD88*, *CD79B*, *EZH2*, and *CARD11*. Each symbol represents a single point mutation in an individual tumor of the type indicated in the key (*Bottom*).

Table S1. All significantly mutated genes identified in 49 patients rank-ordered by decreasing significance

Table S1 (XLS)

The description of column annotations is shown next to the list.

Table S2. All somatic events identified in 49 patients with DLBCL by whole-exome sequencing

Table S2 (XLSX)

The following events were detected by the mutation caller: Missense mutations, nonsense mutations, synonymous mutations, splice site SNPs, frameshift deletions, splice site DNP, frameshift insertions, in-frame insertions, nonstop mutations, in-frame deletions, de novo Start outofframe mutations, splice site deletions, missense mutations, splice site insertions. For each variant, the change in the genomic DNA, cDNA, and protein position is shown.

Table S3. Genes enriched with mutations in AID target motifs

Table S3 (XLSX)

We searched for genes that are enriched with mutations in AID target motifs in an unbiased fashion. For each gene, we calculated the ratio between the number of C mutations in WRCY hotspots (or G mutations in the complement) and the total number of C (or G) mutations. Then, we compared the observed ratio to a null distribution representing no enrichment of WRCY mutations, generated by permuting the observed mutations while maintaining their type and context (e.g., C > T transitions in the CpG context could move to a different C in a CpG context). The *P* value was estimated by the fraction of permutations which yielded a ratio that was at least as high as the observed one. We performed a maximum of 1,000,000 permutations. Next, we corrected for multiple hypothesis testing and calculated a *q*-value using the Benjamini–Hochberg False Discovery Rate procedure (1). As shown, we identified four genes with *q*-values <0.05. This suggests that these genes may be subject to somatic hypermutation mediated by AID.

1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser A Stat Soc* 57:289–300.

Table S4. Overlap between mutated genes identified in our cohort by whole-exome sequencing with mutations reported in the COSMIC database rank-ordered according to significance

Table S4 (XLSX)

COSMIC, Catalogue of Somatic Mutations in Cancer.

Table S5. Validation of selected mutations by resequencing

Table S5 (XLS)

We performed independent validation of mutations in selected genes. We focused on significantly mutated genes that have not been implicated as cancer genes previously, including *TMSL3*, *PCLO*, *P2RY8*, and *ACTB*. We also validated mutations in *NOTCH1*, a potential cancer gene in DLBCL. *MYD88* mutations were included as positive controls. We queried a total of 48 mutations by targeted resequencing as described in *Materials and Methods*, with one assay failure due to sample handling. Forty-six out of 47 mutations were validated with >90 reads harboring the mutation, accounting for a validation rate of 97.9%.