Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing

Jens G. Lohr[a,b], Petar Stojanov[a,b], Michael S. Lawrence[a], Daniel Auclair[a], Bjoern Chapuy[b], Carrie Sougnez[a], Peter Cruz-Gordillo[a], Birgit Knoechel[a,b,c], Yan W. Asmann[d], Susan L. Slager[d], Anne J. Novak[d], Ahmet Dogan[d], Stephen M. Ansell[d], Brian K. Link[e], Lihua Zou[a], Joshua Gould[a], Gordon Saksena[a], Nicolas Stransky[a], Claudia Rangel-Escareño[f], Juan Carlos Fernandez-Lopez[f], Alfredo Hidalgo-Miranda[f], Jorge Melendez-Zajgla[f], Enrique Hernández-Lemus[f], Angela Schwarz-Cruz y Celis[f], Ivan Imaz-Rosshandler[f], Akinyemi I. Ojesina[a], Joonil Jung[a], Chandra S. Pedamallu[a], Eric S. Lander[a,g,h,1], Thomas M. Habermann[d], James R. Cerhan[d], Margaret A. Shipp[b], Gad Getz[a], and Todd R. Golub[a,b,g,i]

[a]Eli and Edythe Broad Institute, Cambridge, MA 02412; [b]Dana–Farber Cancer Institute, Boston, MA 02115; [d]Mayo Clinic College of Medicine, Rochester, MN 55902; [c]Children's Hospital Boston, Boston, MA 02115; [e]University of Iowa College of Medicine, Iowa City, IA 52245; [f]Instituto Nacional de Medicina Genómica, 14610 Mexico DF, Mexico; [g]Harvard Medical School, Boston, MA 02115; [h]Massachusetts Institute of Technology, Cambridge, MA 02142; and [i]Howard Hughes Medical Institute, Chevy Chase, MD 20815

Contributed by Eric S. Lander, December 29, 2011 (sent for review November 22, 2011)

To gain insight into the genomic basis of diffuse large B-cell lymphoma (DLBCL), we performed massively parallel whole-exome sequencing of 55 primary tumor samples from patients with DLBCL and matched normal tissue. We identified recurrent mutations in genes that are well known to be functionally relevant in DLBCL, including *MYD88*, *CARD11*, *EZH2*, and *CREBBP*. We also identified somatic mutations in genes for which a functional role in DLBCL has not been previously suspected. These genes include *MEF2B*, *MLL2*, *BTG1*, *GNA13*, *ACTB*, *P2RY8*, *PCLO*, and *TNFRSF14*. Further, we show that *BCL2* mutations commonly occur in patients with *BCL2*/*IgH* rearrangements as a result of somatic hypermutation normally occurring at the *IgH* locus. The *BCL2* point mutations are primarily synonymous, and likely caused by activation-induced cytidine deaminase–mediated somatic hypermutation, as shown by comprehensive analysis of enrichment of mutations in WRCY target motifs. Those nonsynonymous mutations that are observed tend to be found outside of the functionally important BH domains of the protein, suggesting that strong negative selection against *BCL2* loss-of-function mutations is at play. Last, by using an algorithm designed to identify likely functionally relevant but infrequent mutations, we identify *KRAS*, *BRAF*, and *NOTCH1* as likely drivers of DLBCL pathogenesis in some patients. Our data provide an unbiased view of the landscape of mutations in DLBCL, and this in turn may point toward new therapeutic strategies for the disease.

next-generation sequencing | human genetics | activation-induced deaminase

Diffuse large B-cell lymphoma (DLBCL) is an aggressive non-Hodgkin lymphoma that affects 30,000 new patients in the United States every year (1, 2). The standard of care for the treatment of most cases of DLBCL is the R-CHOP regimen (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) consisting of multiagent chemotherapy plus a therapeutic antibody directed against CD20, a marker of B lymphocytes. The 3-year event-free survival rate is approximately 60%, with the majority of the remaining 40% dying of their disease (3). To date, treatment strategies to improve outcome have largely included increased doses of standard agents in the context of autologous stem cell transplantation (4). Therefore, there is a great medical need to define the genetic abnormalities that are associated with DLBCL to define novel targets for therapy.

Germinal centers (GCs) in lymphoid tissues are sites of clonal expansion and editing of the Ig receptor in B lymphocytes, and this GC reaction is a physiological component of the humoral immune response. Somatic hypermutation (SHM) is part of the GC reaction, and its dysregulation contributes to the accumulation of somatic mutations in oncogenes and tumor-suppressor genes in B lymphocytes.

Traditionally, DLBCL has been classified by the morphology and immunophenotype of the malignant B-cells but more recently, molecular classifications have been reported. Specifically, gene expression-based classification of DLBCL has been proposed (5, 6), and the prognostic relevance for this has been demonstrated (7). It has been suggested that distinct signal transduction pathways are affected in the subtypes that are defined in this way, and that certain genetic defects preferentially occur in specific subtypes defined by the presumed cell of origin of the tumors (8–12).

However, comprehensive understanding of the genomic landscape of DLBCL is lacking. In particular, a key question is which mutations and pathways drive DLBCL pathogenesis. We report here the unbiased sequencing of all protein-coding exons in 55 DLBCL patients, comparing each to its patient-matched normal control. We uncover mutations that provide insights into mechanisms of lymphomagenesis.

Results

Whole-Exome Sequencing Reveals Recurrent Mutations in DLBCL. We performed solution-phase hybrid capture and whole-exome sequencing on paired tumor and germline (i.e., normal) DNA samples from 55 patients with primary DLBCL. We achieved 150-fold mean sequence coverage of targeted exonic regions, with an average of 97% of bases covered per patient (range, 91–98%). Such high coverage is important because tumor samples are often contaminated with normal cells (e.g., fibroblasts, immune cells), which can obscure the identification of somatically mutated alleles unless very deep coverage is achieved.

We excluded six samples from further analysis because of extremely low apparent mutation rates, most consistent with extensive stromal contamination. Of the remaining 49 patients,

Author contributions: J.G.L., P.S., M.S.L., E.S.L., T.M.H., J.R.C., M.A.S., G.G., and T.R.G. designed research; J.G.L., D.A., C.S., and P.C.-G. performed research; P.S., M.S.L., Y.W.A., S.L.S., A.J.N., A.D., S.M.A., B.K.L., L.Z., J.G., G.S., N.S., A.I.O., J.J., and C.S.P. contributed new reagents/analytic tools; J.G.L., P.S., M.S.L., B.C., C.S., B.K., C.R.-E., J.C.F.-L., A.H.-M., J.M.-Z., E.H.-L., A.S.-C.y.C., and I.I.-R. analyzed data; and J.G.L., P.S., M.S.L., B.K., M.A.S., G.G., and T.R.G. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the dbGAP database, www.ncbi.nlm.nih.gov/gap (accession no. phs000450.v1.p1).

[1]To whom correspondence should be addressed. E-mail: lander@broadinstitute.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1121343109/-/DCSupplemental.

the mean nonsynonymous mutation rate was 3.2 mutations per megabase, with mutation rates varying widely (0.6–8.7 mutations per megabase), which is higher than the estimated mutation rate in other hematopoietic malignancies, such as chronic lymphocytic leukemia and other leukemias (<1 per megabase) (13, 14), and multiple myeloma (1.3 per megabase) (15) (Fig. 1*A*). There was no correlation between patient-specific mutation rate and frequency of particular types of mutation (e.g., *CpG→T, Cp(A/C/T)→T, A→G, transversions, or indels; Fig. 1*C*). Similarly, there was no correlation between observed mutation rate and average allelic fraction of those mutations observed in each patient, suggesting that the variation in mutation rate was not simply a function of variability in extent of stromal contamination (Fig. S1).

To define significantly mutated genes in DLBCL, we applied the *MutSig* algorithm to identify genes harboring mutations at a higher frequency than expected by chance (15, 16). To better estimate the significance of observed mutations, *MutSig* takes into account (*i*) the sample-specific mutation rate, (*ii*) the ratio of nonsynonymous to synonymous mutations in a given gene, and (*iii*) the median expression level of each gene in DLBCL [based on gene expression profiling datasets (7)]. This approach revealed 58 statistically significant genes with a false discovery rate cutoff of 0.1 ($q_1 \leq 0.1$; Fig. 1*B*, Table 1, and Tables S1 and S2). We independently validated selected mutations by targeted resequencing in a subset of patients and obtained 97.9% validation rate (Table S5).

Among the significantly mutated genes were those for which a functional role in the pathogenesis of DLBCL has accumulated over many years. These include *CD79B*, *TP53*, *CARD11*, *MYD88*, and *EZH2* (8–11). In addition, we discovered mutations in genes for which a pathogenic role in DLBCL has been suggested recently (17, 18). These include *MLL2*, *TNFRSF14*, *BTG1*, *MEF2B*, and *GNA13*, which are discussed in greater detail later.

We discovered mutations in genes not previously recognized as drivers of cancer. For example, we found β-actin (*ACTB*) mutations in five patients (Fig. 2). Actins are highly conserved proteins of the cytoskeleton and are involved in B lymphocyte activation (19). By using the *xvar* algorithm (based on evolutionary conservation and the nature of the observed amino acid change), the predicted functional consequence of the *ACTB* mutations observed in DLBCL is high. We also note a preponderance of *ACTB* mutations toward the amino terminus of the protein, but the functional significance of this remains unknown.

Another unexpectedly recurrently mutated gene is *P2RY8*, encoding a G protein-coupled purinergic receptor whose normal function has not been extensively characterized (20). *P2RY8* is most notable for its involvement in a chromosomal translocation with *CRLF2* in 7% of patients with B-progenitor acute lymphoblastic leukemia (ALL) and 53% of individuals with ALL and Down syndrome (21). The presumed mechanism of action of such *P2RY8*/*CRLF2* fusions is activation of *CRLF2* by its coming under the control of the *P2RY8* promoter, which is highly active in B lymphoid cells (21). However, we note that *P2RY8* has itself been reported to function as an oncogene in experimental models (22). In DLBCL, we identified six patients with coding mutations in *P2RY8*, two of whom harbored two mutations (Fig. 2). In three patients, the observed allelic fraction of these mutations is greater than 0.5, suggestive of deletion of the WT allele or amplification of the mutant allele. The functional consequence of P2RY8 mutation in DLBCL remains to be determined.

We also observed very common mutations in the gene *PCLO* (Piccolo), encoding a protein that functions as part of the presynaptic cytoskeletal matrix thought to be involved in regulating neurotransmitter release (23). A role for *PCLO* in calcium sensing has also been suggested (24), but a role in cancer has not been reported. We found a total of 23 nonsynonymous mutations in 17 patients (35%; Fig. S2), but the observed ratio of nonsynonymous to synonymous mutations (23:3) is consistent with that expected by chance, given that most of the observed mutations (12 of 23) are transversion mutations, which favor nonsynonymous outcomes by a ratio of nearly 5:1. It is thus possible that, for unknown reasons, the local rate of mutation at the

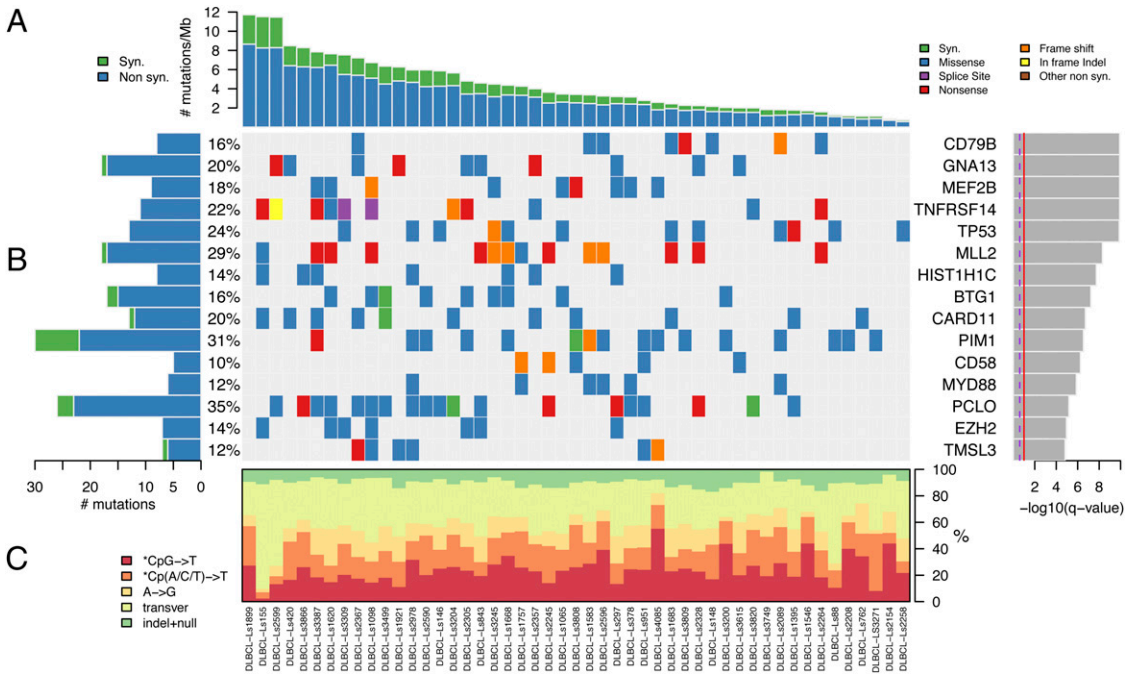

Fig. 1. Significantly mutated genes in 49 patients with DLBCL. (*A*) The rate of synonymous and nonsynonymous mutations is displayed as mutations per megabase, with individual DLBCL samples ranked by total number of mutations. (*B*) The heat map represents individual mutations in 49 patient samples, color-coded by type of mutation. Only one mutation per gene is shown if multiple mutations were found in a sample. *Left*: Histogram shows the number of mutations in each gene. Percentages represent the fraction of tumors with at least one mutation in the specified gene. *Right*: The 15 genes with the lowest q_1-value, ranked by level of significance. (*C*) Base substitution distribution of individual samples, ranked in the same order as in *A*.

Table 1. Significantly mutated genes and genes with mutations in clusters and at conserved sites

Genes	Mutations	Patients	Sites	Silent	q_1	q_2
CD79B	8	8	5	0	1.22×10^{-10}	0.025
GNA13	17	10	16	1	1.22×10^{-10}	
MEF2B	9	9	9	0	1.22×10^{-10}	
TNFRSF14	11	11	11	0	1.22×10^{-10}	
TP53	13	12	12	0	1.36×10^{-10}	0.025
MLL2	17	14	17	1	4.95×10^{-9}	
HIST1H1C	8	7	6	0	1.83×10^{-8}	
BTG1	15	8	14	2	5.95×10^{-8}	
CARD11	12	10	11	1	1.92×10^{-7}	0.035
PIM1	22	15	19	8	2.90×10^{-7}	0.077
CD58	5	5	5	0	5.51×10^{-7}	
MYD88	6	6	3	0	1.35×10^{-6}	$<1 \times 10^{-6}$
PCLO	23	17	23	3	6.65×10^{-6}	
EZH2	7	7	4	0	0.000011	0.035
TMSL3	6	6	6	1	0.000015	
CD70	5	5	5	0	0.000018	
P2RY8	8	6	8	0	0.000024	
KRTAP5-5	4	4	1	0	0.000033	0.0089
HIST1H3B	5	5	3	1	0.000062	
ACTB	5	5	4	0	0.00022	
CREBBP	9	8	9	0	0.0004	
NFKBIA	4	4	4	0	0.00048	
B2M	5	5	4	1	0.00054	
SOCS1	4	3	4	1	0.00078	
HLA-A	5	4	5	0	0.0024	
UBE2A	3	3	3	0	0.0032	
CCND3	3	3	3	0	0.0048	
POU2F2	4	4	1	0	0.0051	0.0037
OR6K3	3	3	3	0	0.0093	
LOC153328	3	3	2	0	0.0097	
UNC5D	5	5	5	0	0.0097	
PASD1	4	4	4	0	0.0099	
STAT3	5	5	5	0	0.011	
CIITA	7	5	7	0	0.015	
SYN2	3	3	1	0	0.017	0.00037
PDGFC	3	3	3	0	0.017	
TBL1XR1	3	3	3	0	0.02	
HIST1H1E	10	7	10	3	0.021	
UNC5C	5	5	5	0	0.021	
SRPX	3	3	2	1	0.032	
PCDHB6	4	4	4	0	0.033	
S1PR2	4	3	4	0	0.033	
KLHL6	5	4	5	0	0.034	
ETV6	3	3	3	0	0.042	
SLITRK6	4	4	4	0	0.042	
DUSP2	4	3	3	1	0.045	
SLC38A8	3	3	3	0	0.048	
H1FOO	2	2	1	0	0.05	
HLA-B	3	3	3	0	0.052	
CPS1	5	5	5	0	0.052	
BCR	7	5	7	1	0.066	0.041
TNF	3	3	3	1	0.066	
HIST1H2AL	2	2	2	0	0.073	
HIST1H2BC	4	3	4	1	0.081	
GABRA1	3	3	3	0	0.094	
HIST1H2BO	2	2	2	0	0.096	
LRRIQ3	3	3	3	0	0.1	
APOBEC2	2	2	2	0	0.1	
ERBB2IP	3	1	3	0		$<1 \times 10^{-6}$
STAT6	2	2	2	0		$<1 \times 10^{-6}$
MEF2C	5	2	5	0		0.0061
SGK1	9	5	9	2		0.0061
ADAM10	2	1	2	0		0.0098
PABPC1	3	3	2	0		0.011
KIF1B	2	2	1	0		0.014
ATP2C2	2	1	2	0		0.025
RGS12	2	2	2	0		0.025
TTC18	2	2	2	1		0.025
DIAPH1	2	1	2	0		0.041
ZNF830	3	2	3	0		0.041
CALR	2	1	2	0		0.050
NEB	3	3	3	0		0.051
PGAP2	2	1	2	0		0.074
SYK	2	2	2	0		0.082

Two algorithms were used to prioritize significantly mutated genes. The top 58 genes, reflected by *q*-values (q_1) of 0.1 or lower, are determined by analyzing the frequency of somatic mutations across samples, corrected for gene length, the sample-specific mutation rate, the nonsynonymous/synonymous mutation ratio, and expression of these genes in independent DLBCL datasets. The remainder of the list represents genes that have a q_1-value greater than 0.1, but are significant by an independent analysis that identifies genes that may be functionally relevant based on clustering of mutations and evolutionary sequence conservation ($q_2 \leq 0.1$). Genes for which q_1- and q_2-value are listed are prioritized by both algorithms. Mutations, number of nonsynonymous mutations in this gene across the individual set; patients, number of patients with at least one nonsynonymous mutation; sites, number of unique sites with a nonsynonymous mutation; silent, number of silent (i.e., synonymous) mutations in this gene across the individual set.

PCLO locus is unusually high, giving rise to passenger mutations of no functional consequence in DLBCL. Additional work is clearly needed to resolve the role, if any, of *PCLO* mutations in DLBCL and other cancers.

Histone 1 (H1) family proteins are linker histone proteins that bind to the DNA entering and exiting the nucleosomal cores. Different forms are expressed at different stages of the cell cycle in different tissue types and are involved in chromatin compaction and possibly transcription (25). We observed a striking accumulation of mutations in H1 family proteins, with 59 nonsynonymous and 35 synonymous mutations among 31 histone H1 proteins in 34 patients (69%; Table S2). The functional significance of these mutations remains to be explored, but hotspot analysis as outlined later suggests that *HIST1H3B* and possibly other core histone proteins are subject to activation-induced cytidine deaminase (AID)–mediated SHM.

We also identified mutations in genes that were recently reported to be significantly mutated (17, 18). *MLL2* is a histone methyltransferase of the SET1 family that is responsible for histone H3-lysine 4 trimethylation (H3K4me3) during oogenesis and early development (26). Inactivating mutations have been reported in medulloblastoma (27) and multiple myeloma (15), and chromosomal translocations involving the MLL family member *MLL1* are well described in acute leukemias (28). The *MLL2* mutations we observed in DLBCL are highly biased toward truncating events, the large majority being nonsense mutations and frameshift-inducing insertions and deletions (Fig. 2). As has been suggested previously, our data suggest that *MLL2* may function as an important tumor suppressor in DLBCL (17, 29).

TNFRSF14, also known as LIGHT Receptor, belongs to the TNF-receptor superfamily most extensively studied in T cells. Interestingly, it can convey opposing signals based on its specificity for diverse ligands. In our data, five of nine mutations suggest loss of function ($n = 4$ nonsense mutations and $n = 1$ frameshift deletion), with an additional in-frame insertion and three missense mutations. No synonymous mutations were seen (Fig. 2). As has been suggested, these results strongly suggest a tumor-suppressive role of *TNFRSF14* in DLBCL (17). It has been proposed that LIGHT-mediated triggering of TNFRSF14 renders B-cell lymphomas more immunogenic and sensitive to

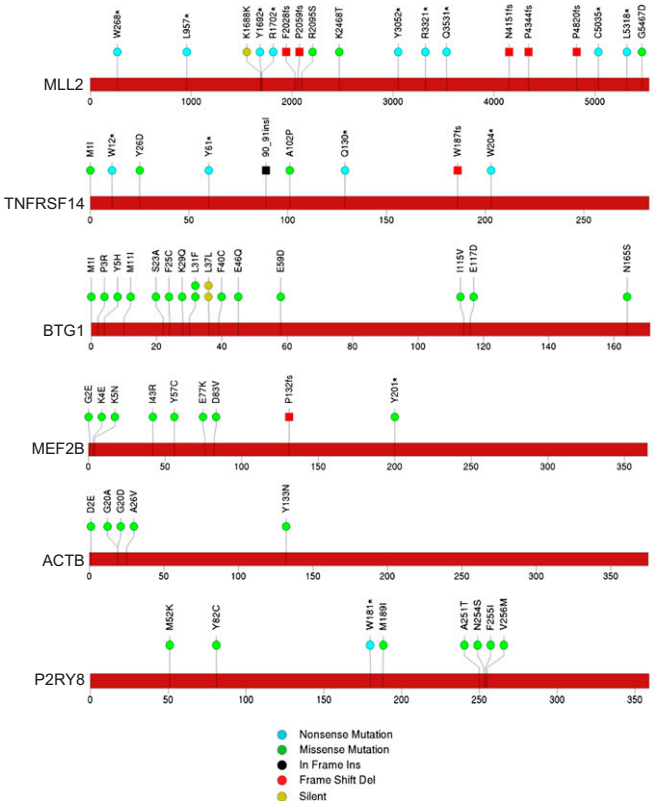

Fig. 2. Somatic mutations in DLBCL affect genes of various classes. Sites of somatic mutations in significantly mutated genes called by analysis pipeline and passing manual review. A diagram of the relative positions of somatic mutations is shown for *MLL2*, *TNFRSF14*, *BTG1*, *MEF2B*, *ACTB*, and *P2RY8*. The type of the mutation is indicated in the key (*Bottom*). The overall validation rate of mutation calls was 97.9% [of 47 selected mutations tested, only one (ACTB G20A) failed to validate; see Table S5].

FAS-induced apoptosis (30)—a potential mechanism by which *TNFRSF14* can act as a tumor suppressor.

Mutations in *BTG1* were also observed to be relatively common (15 nonsynonymous and two synonymous mutations). BTG1 belongs to the BTG/Tob family of proteins that regulate cell cycle progression in a variety of cells. BTG1 is thought to confer DNA binding of sequence-specific transcription factors (31). Whether mutations in *BTG1* result in aberrations in chromatin structure (as with MLL-family mutations) as opposed to nonhistone targets remains to be determined. Interestingly, we observed several patients with more than one mutation in *BTG1*, including two patients with three mutations and one patient with four mutations. In the latter patient, two sets of adjacent mutations (L37L plus L31F and P3R plus M1I) were never found in the same sequencing read, suggesting that they occurred on different alleles, i.e., in *trans*. We note that two patients had the identical silent mutation at codon 31, suggesting that this mutation, although it did not affect protein coding sequence, may alter codon use or mRNA stability. Overall, the nature of *BTG1* mutations does not clearly point toward a gain-of-function or loss-of-function mechanism, although the biallelic involvement seen in some patients tends to favor loss of function (Fig. 2).

MEF2B mutations were observed in 18% of patients with DLBCL, similar to data reported recently (17, 18), predominantly in the MADS box or MEF2 domains. *MEF2B* belongs to a family of calcium-regulated transcription factors that recruit histone-modifying enzymes. Further supporting an oncogenic role for MEF2 proteins, *MEF2C* has been identified as a T-cell acute lymphoblastic leukemia oncogene that is activated by chromosomal rearrangement (32).

Negative Selection Against Deleterious BCL2 Mutations. A hallmark of cancer genes (i.e., genes that harbor "driver" mutations, which contribute to the formation and progression of cancer) is the preponderance of nonsynonymous mutations compared with synonymous mutations (typically with an expected ratio of ~2.8:1). Curiously, we observed a striking opposite effect in the *BCL2* gene—a known driver in some DLBCLs. We observed a very high mutation rate in *BCL2*, but with a depletion of nonsynonymous mutations, with 18 nonsynonymous mutations and 28 synonymous mutations, for a ratio of 0.64, far below that expected by chance ($P = 8.8 \times 10^{-6}$; Fig. 3). We hypothesized that this phenomenon might be caused by two processes: first, a locally high mutation rate at the *BCL2* locus occurring via SHM, and second, negative selection against functionally deleterious mutations.

We first explored the possibility that the high rate of *BCL2* mutations observed in our patients might be related to SHM. This physiologic process in B lymphocytes has been best studied at Ig gene loci, where it facilitates the development of antibody diversity (33), but other genes have been shown to be subject to aberrant SHM (including the *PIM1* gene, in which we also observed hypermutation with 30 mutations in 16 patients; Fig. S2) (34). Chromosomal translocations of the *BCL2* gene into the Ig heavy chain (*IgH*) locus occur in approximately 20% of DLBCL tumors, and the resulting *BCL2* overexpression provides an antiapoptotic signal to the lymphoma cells (35). We hypothesized that *BCL2* hypermutation in our patients might be explained by the *BCL2* gene adopting the *IgH* locus's normal process of SHM as a result of the translocation (36). If this hypothesis were correct, those tumors with elevated *BCL2* mutation rates would be expected to also harbor *BCL2*/*IgH* translocations. We tested for *BCL2*/*IgH* translocation in 26 patients with DLBCL (13 with *BCL2* mutation and 13 others selected randomly). As predicted, the vast majority of patients with *BCL2* hypermutation (10 of 13, 77%) had *BCL2*/*IgH* translocation, whereas only one of 13 patients (8%) lacking *BCL2* mutation had the translocation ($P = 0.0005$, one-sided Fisher exact test; Fig. S3). To determine whether *BCL2* or other genes may be targets of SHM, which is mediated by AID, we asked whether mutations occur preferentially in the context of WRCY motifs, which are known target sequences of AID. This analysis indeed revealed that *BCL2*, as well as other genes, including *PIM1*, have significant enrichment of mutations in WRCY motifs (Table S3).

Although these results likely explain the hypermutation at the *BCL2* locus, they do not explain the preponderance of synonymous mutations observed in these patients. We hypothesized that they represent a vestige of negative selection against deleterious mutations in *BCL2*, on which the tumor cells are dependent for survival. If this hypothesis were correct, we would expect that any nonsynonymous mutations would be confined to functionally nonessential domains of the protein, whereas the synonymous mutations would not. To address this, we focused on the *BCL2* BH domains that mediate interactions with proapoptotic proteins (37). We observed 18 nonsynonymous mutations, but only three of these fell within BH domains, whereas 15 of 28 synonymous mutations fell within BH domains (Fig. 3). We used permutations to determine the probability of nonsynonymous mutations being located within versus outside BH domains, corrected for domain size. To increase the power of this analysis, we included the *BCL2* mutations observed in the present study and those recently reported (17). This analysis indicated a significant enrichment of nonsynonymous mutations falling outside of BH domains ($P = 0.041$). These results argue that the nonsynonymous mutations preserve *BCL2* function by avoiding critical BH domains. On the contrary, we found the synonymous mutations to be preferentially located within BH domains ($P = 0.02$). This can be explained by nonuniform distribution of these mutations along the gene ($P = 0.045$) with preferential clustering at the 5′ end. This is consistent with previous demonstrations of aberrant SHM preferentially occurring closer to the 5′ end (34).

Our results are most consistent with purifying selection (i.e., negative selection) against mutations that inhibit or decrease *BCL2* function. Such mutations likely result in loss of or

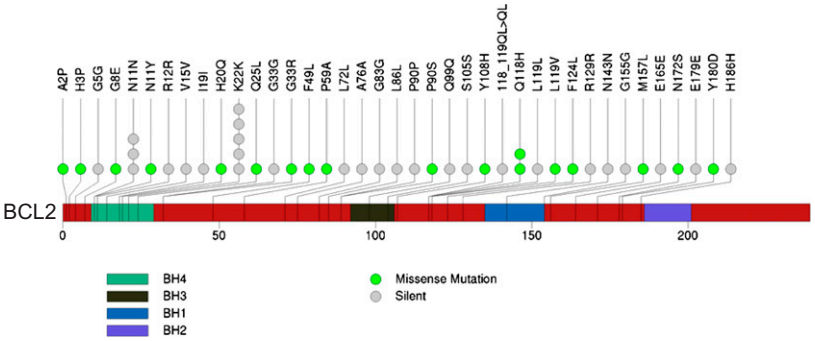

Fig. 3. Selection for functionally inconsequential mutations in *BCL2* and sites of somatic mutations in *BCL2*. Nonsynonymous mutations in *BCL2* are preferentially located outside of BH-domains.

impaired tumor cell viability, leaving behind only those mutations that are synonymous or that affect nonrelevant domains. Interestingly, we note the presence of recurrent synonymous mutations at codons N11 and K22 (Fig. 3), suggesting a preference for certain sites to acquire mutations. Whether there is a functional consequence of these synonymous mutations, for example by affecting mRNA stability, remains to be determined.

Identifying Functionally Relevant but Rare Mutations. The preceding sections focused on the identification of cancer genes based on their mutation frequency across the whole genome. However, additional genes may harbor functionally important mutations that fall short of statistical significance (in the standard mutation frequency test) given the modest size of our dataset. Some of these genes are recognizable based on their known importance in the pathogenesis of other malignancies. For example, we observed two mutations in *KRAS* (G13D), four mutations in *NOTCH1*, and two mutations in *BRAF*, often involving specific amino acids documented to be sites of recurrent mutation in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. It therefore seems likely that these mutations similarly play a causal role in DLBCL, albeit at a low frequency. To perform this search in a rigorous manner, we also tested each gene whether it has an increased mutation rate only in sites previously reported in the COSMIC database. Table S4 shows a list of the most significant genes identified by this test.

We also sought to prioritize rare mutations based on the clustering of mutations within particular regions of the coding sequence. Mutations occurring by chance (i.e., passenger mutations) would be expected to be scattered randomly across the protein, whereas functionally consequential mutations may cluster within critical domains. First, we assumed that mutations have a higher likelihood of functional impact if they are found at the same site or in close proximity to each other. For example, highly clustered mutations are observed in *MYD88*, *CARD11*, *CD79B*, and *EZH2* (Fig. S4). Second, we assumed that mutated genes are more likely to be functionally relevant if the affected amino acid residue is more conserved across species. Our algorithm thus returns a list of mutated genes that is rank-ordered by a composite score reflecting sequence conservation and clustering of mutation sites (Table 1). This analysis prioritized genes known to be mutated in DLBCL and to play a role in its pathogenesis and also novel genes that would not otherwise be considered as likely drivers. For example, the tyrosine kinase SYK was identified as a likely driver based on these criteria, and it has recently been demonstrated that SYK inhibitors have significant clinical activity in non-Hodgkin lymphoma, including DLBCL (38, 39). Similarly, the serine/threonine kinase SGK1 was identified by this analysis despite its not reaching statistical significance based on frequency alone ($q_1 > 0.1$), but a recent study confirms the recurrent nature of SGK1 mutation in DLBCL (17).

Discussion

We performed whole-exome sequencing on tumor samples and matched normal samples from 55 patients with DLBCL to identify the spectrum of mutations associated with the disease. This analysis provides a rich description of the DLBCL genome and forms the basis for future discovery and therapeutic target identification. In this single study, we rediscovered the genes previously discovered to be important drivers of DLBCL, and identified candidates deserving of functional follow-up.

Near the completion of our study, two groups reported their analysis of the DLBCL genome (17, 18). Remarkably, the recurrently mutated genes identified in these studies are highly overlapping [of our 58 significantly mutated genes, 20 were also reported as frequently mutated by Morin et al. (17), and 14 were reported by Pasqualucci et al. (18)]. Such overlap provides strong evidence that the mutations identified by our significance thresholds are indeed recurrent in DLBCL. The fact that these studies identified largely the same genes at similar frequency suggests that these are the most common targets of somatic mutation in DLBCL. Although our results and interpretation are largely concordant with those recently published, we differ in our interpretation of the frequent mutations seen in the *BCL2* gene. Whereas Morin et al. (17) suggest that these are likely indicative of positive selection for nonsynonymous variants, we suggest the mutations are in fact passenger mutations and observe a depletion of damaging mutations in *BCL2*. Thus, whereas *BCL2*/*IgH* translocation is likely indeed a driver of DLBCL, it also increases the overall mutation rate in *BCL2* (via AID), and we believe that the resulting point mutations are likely noncontributory to the pathogenesis of the disease. However, it is interesting to speculate that the striking recurrence of silent mutations at two distinct residues may indicate a selective advantage that is conferred by silent mutations.

More generally, a unique feature of B-cell malignancies is the role of SHM in increasing mutation frequency. Under normal conditions, such SHM promotes affinity maturation of antibodies, but the enzymes that mediate this effect, such as AID, may also erroneously cause somatic mutations and structural abnormalities in oncogenes and tumor-suppressor genes. High mutation rates observed at particular loci in the genome may therefore indicate the presence of dysfunctional SHM, rather than indicating positive selective growth advantage conferred by such mutations. Our analyses suggest that *BCL2*, as well as *PIM1* and other genes, are subject to enrichment of mutations in WRCY hotspots, and therefore likely to represent AID targets. Although, in the case of *BCL2*, SHM can be explained as a result of translocation to the *IgH* locus (40), this is less clear for other genes. Whole-genome sequencing approaches may reveal whether hypermutation and chromosome translocation events are related more generally. A complete elucidation of those frequently mutated genes that are contributory to the malignant phenotype, versus those that are simply frequent passenger mutations caused by local SHM will require future studies.

Our list of significantly mutated genes is based on the frequency of the occurrence of mutations, corrected for the size of the gene and its expression level, the sample-specific mutation rate, and the ratio of nonsynonymous to synonymous mutations. However, some mutations may be functionally important despite not meeting statistical significance based on these criteria alone. As an additional tool to discover these mutations, we analyzed the clustering of mutations in hotspots of individual genes, evolutionary conservation, and overlap with mutated genes reported in

the COSMIC database. These analyses reveal several genes that are rarely mutated, but that may nevertheless play an important role in the pathophysiology of DLBCL. Future studies may benefit from an even more thorough computational assessment of the likely functional consequence of observed mutations.

With a first draft of the genomic landscape of DLBCL now defined, the next step for the field should be to establish the functional consequence of the observed mutations. In many cases, the consequence is already known (e.g., activating the B-cell receptor pathway or activating the NF-κB pathway). In others, however (particularly in the case of rare mutations), there is no current insight into the role of those gene products in cancer pathogenesis. Although the traditional approach to this problem has been to attack such candidate genes one by one, we propose that new, systematic approaches to the functional characterization of candidate oncogenes and tumor suppressors are needed. Systematic studies to connect such genes to known pathways and/or processes will help to extend the utility of cancer genome studies and accelerate the pace at which genetic findings are translated into therapeutic impact.

Materials and Methods

Sample Selection and Massively Parallel Sequencing. A total of 55 patients with DLBCL provided DNA for this study. 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. DNA was extracted from lymph node samples (tumor) and blood (normal) as previously described and processed as detailed in *SI Materials and Methods*.

Calculation of Sequence Coverage, Mutation Calling, and Significance Analysis. Massively parallel sequencing data were processed by using two consecutive pipelines developed at the Broad Institute: "Picard" generates a single BAM file representing the sample; "Firehose" starts with the BAM files for each DLBCL sample and matched normal sample from peripheral blood (hg19) and performs various analyses (15, 16). 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. Subsequent analysis is described in more detail in *SI Materials and Methods*.

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* locus (InVivoScribe Technologies). Details are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank the members of the T.R.G. laboratory, Maria Cortez, Jadwiga Grabarek, Ami S. Bhatt, Niall J. Lennon, and all members of the Broad Institute's Biological Samples Platform; Genetic Analysis Platform; and Genome Sequencing Platform, without whom this work would not have been possible. This work was conducted as part of the Slim Initiative for Genomic Medicine (SIGMA), a joint US–Mexico project funded by the Carlos Slim Health Institute, and was also supported by National Cancer Institute Grants 5P01 CA092625-07 and P50 CA97274.

1. Abramson JS, Shipp MA (2005) Advances in the biology and therapy of diffuse large B-cell lymphoma: moving toward a molecularly targeted approach. *Blood* 106:1164–1174.
2. Lenz G, Staudt LM (2010) Aggressive lymphomas. *N Engl J Med* 362:1417–1429.
3. Pfreundschuh M, et al.; German High-Grade Non-Hodgkin Lymphoma Study Group (DSHNHL) (2008) Six versus eight cycles of bi-weekly CHOP-14 with or without rituximab in elderly patients with aggressive CD20+ B-cell lymphomas: A randomised controlled trial (RICOVER-60). *Lancet Oncol* 9:105–116.
4. Glass B, et al.; German High-Grade Non-Hodgkin Lymphoma Study Group (DSHNHL) (2010) High-dose therapy followed by autologous stem-cell transplantation with and without rituximab for primary treatment of high-risk diffuse large B-cell lymphoma. *Ann Oncol* 21:2255–2261.
5. Alizadeh AA, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511.
6. Monti S, et al. (2005) Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* 105:1851–1861.
7. 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.
8. Davis RE, et al. (2010) Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature* 463:88–92.
9. Ngo VN, et al. (2011) Oncogenically active MYD88 mutations in human lymphoma. *Nature* 470:115–119.
10. Lenz G, et al. (2008) Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* 319:1676–1679.
11. Morin RD, et al. (2010) Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* 42:181–185.
12. Pasqualucci L, et al. (2011) Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* 471:189–195.
13. Puente XS, et al. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 475:101–105.
14. Greenman C, et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446:153–158.
15. Chapman MA, et al. (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature* 471:467–472.
16. Stransky N, et al. (2011) The mutational landscape of head and neck squamous cell carcinoma. *Science* 333:1157–1160.
17. Morin RD, et al. (2011) Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* 476:298–303.
18. Pasqualucci L, et al. (2011) Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat Genet* 43:830–837.
19. Harwood NE, Batista FD (2011) The cytoskeleton coordinates the early events of B-cell activation. *Cold Spring Harb Perspect Biol* 3.
20. Cantagrel V, et al. (2004) Disruption of a new X linked gene highly expressed in brain in a family with two mentally retarded males. *J Med Genet* 41:736–742.
21. Mullighan CG, et al. (2009) Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. *Nat Genet* 41:1243–1246.
22. Fujiwara S, et al. (2007) Transforming activity of purinergic receptor P2Y, G protein coupled, 8 revealed by retroviral expression screening. *Leuk Lymphoma* 48:978–986.
23. Leal-Ortiz S, et al. (2008) Piccolo modulation of Synapsin1a dynamics regulates synaptic vesicle exocytosis. *J Cell Biol* 181:831–846.
24. Fujimoto K, et al. (2002) Piccolo, a Ca2+ sensor in pancreatic beta-cells. Involvement of cAMP-GEFII.Rim2. Piccolo complex in cAMP-dependent exocytosis. *J Biol Chem* 277:50497–50502.
25. Izzo A, Kamieniarz K, Schneider R (2008) The histone H1 family: Specific members, specific functions? *Biol Chem* 389:333–343.
26. Andreu-Vieyra CV, et al. (2010) MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. *PLoS Biol* 8.
27. Parsons DW, et al. (2011) The genetic landscape of the childhood cancer medulloblastoma. *Science* 331:435–439.
28. Coenen EA, et al. (2011) Prognostic significance of additional cytogenetic aberrations in 733 de novo pediatric 11q23/MLL-rearranged AML patients: Results of an international study. *Blood* 117:7102–7111.
29. Ng SB, et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 42(9):790–793.
30. Costello RT, et al. (2003) Stimulation of non-Hodgkin's lymphoma via HVEM: an alternate and safe way to increase Fas-induced apoptosis and improve tumor immunogenicity. *Leukemia* 17:2500–2507.
31. Winkler GS (2010) The mammalian anti-proliferative BTG/Tob protein family. *J Cell Physiol* 222:66–72.
32. Homminga I, et al. (2011) Integrated transcript and genome analyses reveal NKX2-1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia. *Cancer Cell* 19:484–497.
33. Klein U, Dalla-Favera R (2008) Germinal centres: Role in B-cell physiology and malignancy. *Nat Rev Immunol* 8:22–33.
34. Pasqualucci L, et al. (2001) Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* 412:341–346.
35. Iqbal J, et al. (2011) BCL2 predicts survival in germinal center B-cell-like diffuse large B-cell lymphoma treated with CHOP-like therapy and rituximab. *Clin Cancer Res* 17(24):7785–7795.
36. Saito M, et al. (2009) BCL6 suppression of BCL2 via Miz1 and its disruption in diffuse large B cell lymphoma. *Proc Natl Acad Sci USA* 106:11294–11299.
37. Letai AG (2008) Diagnosing and exploiting cancer's addiction to blocks in apoptosis. *Nat Rev Cancer* 8:121–132.
38. Friedberg JW, et al. (2010) Inhibition of Syk with fostamatinib disodium has significant clinical activity in non-Hodgkin lymphoma and chronic lymphocytic leukemia. *Blood* 115:2578–2585.
39. Chen L, et al. (2008) SYK-dependent tonic B-cell receptor signaling is a rational treatment target in diffuse large B-cell lymphoma. *Blood* 111:2230–2237.
40. Tanaka S, Louie DC, Kant JA, Reed JC (1992) Frequent incidence of somatic mutations in translocated BCL2 oncogenes of non-Hodgkin's lymphomas. *Blood* 79:229–237.