The genetic landscape of mutations in Burkitt lymphoma

Cassandra Love[1], Zhen Sun[1], Dereje Jima[1], Guojie Li[1], Jenny Zhang[1], Rodney Miles[2], Kristy L. Richards[3], Cherie H. Dunphy[3], William W. L. Choi[4], Gopesh Srivastava[4], Patricia L. Lugar[5], David A. Rizzieri[5], Anand S. Lagoo[5], Leon Bernal-Mizrachi[6], Karen P. Mann[6], Christopher R. Flowers[6], Kikkeri N. Naresh[7], Andrew M. Evens[8], Amy Chadburn[9], Leo I. Gordon[9], Magdalena B. Czader[10], Javed I. Gill[11], Eric D. Hsi[12], Adrienne Greenough[1], Andrea B. Moffitt[1], Matthew McKinney[1], Anjishnu Banerjee[14], Vladimir Grubor[1], Shawn Levy[13], David B. Dunson[14], Sandeep S. Dave[1,5]

1. Duke Institute for Genome Sciences and Policy, Duke University, Durham, NC, USA
2. University of Utah, Salt Lake City, UT, USA
3. University of North Carolina, Chapel Hill NC, USA
4. The University of Hong Kong, Queen Mary Hospital, Hong Kong, China
5. Duke Cancer Institute, Duke University Medical Center, Durham NC, USA
6. Emory University, Atlanta GA, USA
7. Imperial College, London, UK
8. University of Massachusetts, Worcester, MA, USA
9. Northwestern University, Chicago IL, USA
10. Indiana University, Indianapolis IN, USA
11. Baylor University Medical Center, Dallas TX, USA
12. Cleveland Clinic, Cleveland, OH, USA
13. Hudson Alpha Institute for Biotechnology, Huntsville, AL, USA
14. Department of Statistical Science, Duke University, Durham, NC, USA

Nature Genetics: doi:10.1038/ng.2468

SUPPLEMENTARY NOTE

Nature Genetics: doi:10.1038/ng.2468

Fluorescence *in situ* hybridization (FISH)

FISH was used to confirm *MYC* gene translocations. Tissue was fixed and probed for LSI IGH (14q32), LSI MYC (8q24) and LSI CEP8 (8p11.1-q11.1). The test was conducted by determining the signal configuration patterns within 25 interphase nuclei of tumor cells. The percent of abnormal signal patterns (i.e. two fluorescence colors adjacently detected indicating a fusion) was calculated. The images were captured using the BioView Duet system. A representative image is shown in Supplementary Figure 1.

Publicly Available Controls

In addition to the 19 control exomes prepared in-house, raw data from 256 publicly available exomes, 1000Genome Project pilot 1(SNV calls for 179 individuals)[1] and HapMap 3[2] data were downloaded to gauge population allele frequencies.

DLBCL Exomes

Using methods similar to those used for sequencing and analysis of BL, we sequenced the gene coding regions of 94 DLBCL tumors, along with germline DNA in 34 patients. In all, we generated over 500 GB of sequence data corresponding to over 30-fold exome sequencing coverage in these cases.

Exome Coverage and Depth

Coverage and depth were determined by counting the number of bases from reads aligning to the exome and dividing that sum by the size of the complete exome. For each sample, we applied IntersectBed (BEDTools) to identify the reads from the BAM file that aligned to each of the 198,701 exons in CCDS (v36). The depth of coverage at each exon was determined by multiplying the number of reads mapping by the number of bases, and then dividing that product by the size of the exon.

We found that over 95% (189,690) of the exons had reads mapping in at least 90% of our samples; it is likely that the remaining exons (less than 5%) were not captured effectively. Of those, 93.3% had an average sequencing read-depth of greater than 10-fold (10X), 80.1% had an average depth greater than 20X, and 62.9% had an average depth greater than 30X. The average

Nature Genetics: doi:10.1038/ng.2468

depth at the 189,690 exons consistently measured was 47.2X (range 12.4-114.6). The average exonic coverage for all samples (top graph) is depicted in Supplementary Figure 2a.

We also plotted the distribution of average coverage for our samples in Supplementary Figure 2b. As shown, the vast majority of exons are covered at depths averaging 20-50 fold, with cumulatively fewer than 10% of the exons displaying highly skewed coverage.

Rate of Somatic Alterations

We compared the somatic alteration rates of Burkitt lymphomas (14 pairs) and DLBCLs (34 pairs) by counting the somatically acquired mutations, broken down by transitions, transversions, and other (alterations such as indels) in each discovery set. We found the somatic alteration rate to vary widely within each disease (Supplementary Fig. 3). The reasons for this variation, also observed in other cancers, are poorly understood. The overall somatic alteration rates are not significantly different in the two diseases.

Sanger Sequence Validation

Variants of interest were chosen for Sanger validation from genes of interest which included known and novel cancer genes, as well as the most frequently mutated genes, including *MYC* and *ID3*. We also randomly chose 15 single variants that were observed in only one case for Sanger. Also validated were all of the confirmed variants from available paired normal samples. Single nucleotide variants and small insertions/deletions (indels) were visualized using the Integrated Genomics Viewer (IGV)[3] and subjected to Sanger sequencing.

We performed Sanger sequencing for 124 distinct variant/sample combinations and observed that high quality Sanger sequencing variant calls were identical in 80% of cases with exome sequencing results (Supplementary Table 2). We also did 154 additional Sanger sequencing for alleles in cases which were expected to be wild type. In all, we tested a total of 278 individual variants/cases with an overall validation rate of over 90%.

Nature Genetics: doi:10.1038/ng.2468

Identification of Burkitt Lymphoma Mutated Genes

We began our analysis of recurrently mutated genes in Burkitt lymphoma by designating the 14 Burkitt lymphoma cases with paired germ-line DNA as our discovery set. The remaining 45 Burkitt lymphoma cases were designated as the validation set.

In all, we identified 1241 somatically mutated variants in 1104 unique genes in the discovery set. We then identified additional genetic variants in those 1104 genes in the validation set as those rare, non synonymous variants that were not present in databases of normal variation including dbSNP135[4], publicly available data from the thousand genomes project[5], publicly available exomes from healthy individuals[6-8] (N=256) or in the 19 additional exomes that we sequenced from control patients without lymphoma. In all, we identified a total of 2318 such variants in those 1104 genes from the cases that comprised the discovery and validation sets.

These 2318 variants became the starting point for our identification of driver mutations and genes that were recurrently mutated in Burkitt lymphoma. From this set of variants, we eliminated all variants that originated from genes that were in the 90th percentile or higher for non-synonymous variation in our normal controls (e.g olfactory receptor genes). We also excluded any variants that possibly arose from mapping issues with pseudogenes. We identified the potential contribution of pseudogenes by examining the pseudogene.org database, maintained by the Gerstein Lab at Yale University (http://www.pseudogene.org/cgi-bin/db-gen.cgi?type-Eukaryote) and HGNC (The HUGO Gene Nomenclature Committee)**.** All genes with variants mapping to these pseudogenes (N=11) were excluded from further analysis.

From the remaining variants, we retained all that were already in COSMIC or that were in the same protein domain as a COSMIC variant (N=216). We also retained variants that were predicted to be "functional" using three separate algorithms: SIFT ("damaging"), Polyphen2 ("possibly damaging" or "probably damaging") and mutation assessor (predicted impact "medium" or "high"). All nonsense and frameshift mutations were automatically classified as functional. The remaining set consisted of 462 variants in 269 genes.

Nature Genetics: doi:10.1038/ng.2468

For these 269 genes, we excluded those that had only events that came from recurrent SNVs and likely represented polymorphisms, or had only one event in the 462 variants. 88 genes and 432 variants remained. We examined all 432 variants and eliminated those that were not in either the vicinity of a COSMIC variant nor targeted the same protein domain as another variant in the same gene. We eliminated all genes with fewer than two events, resulting in 70 recurrently mutated Burkitt lymphoma genes (listed in Supplementary Table 3) with a total of 305 variants (Supplementary Table 4). The schema for the identification of mutated genes is depicted in Supplementary Figure 4.

Among the BL cases, we observed a preponderance of missense mutations predicted to alter the encoded amino acid. Small insertions and deletions (indels) accounted for fewer than 5% of the genetic alterations. The majority of these indels occurred in sizes that were multiples of three, predicted to preserve reading frames.

Calculation of Association between Individual Genes

In order to compare the distance between mutational patterns of each gene to all other genes recurrently mutated in Burkitt lymphoma or DLBCL, we began by selecting all genes that were identified as being mutated in our study (for Burkitt lymphoma and DLBCL), as well as the other published studies in DLBCL[9-11]. We exported the identified variants from our study for all of the genes-mutations that affected at least 10% of DLBCL or Burkitt lymphoma cases.

We identified the seven genes from the publications above that did not have identified variants in our DLBCL cases. All of these genes had excellent coverage in our Burkitt lymphoma cases and therefore can be excluded as commonly mutated genes in Burkitt lymphoma. These genes include *MLL2, CMYA5, ETS1, FOXO1, P2RY8, PCLO* and *RAPGEF1*.

For the remaining 55 genes meeting the above described criteria in our data, we considered all the Burkitt lymphoma and DLBCL rare variants together, and recoded the presence of a variant as 1 and the absence of the variant as 0. The variant frequencies were tabulated and are depicted

Nature Genetics: doi:10.1038/ng.2468

in Figures 3a and 3b. We calculated the distances between the patterns of variation for each gene as follows: $\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$

Where n represents the total number of samples,

where X_i represents the measurement in gene 1 in sample i

$\bar{X}$ represents the average across all samples.

Y_i represents measurement for the (remaining) gene i

$\bar{Y}$ represents the average measurement for the remaining genes across all samples.

All instances where the distance between the two genes was one standard deviation above or below the mean were retained and plotted in Figure 3c. For example, in Figure 3c, one would observe that *MYC* and *ID3*, a pair with positive association, must have many samples either mutated in both genes or not mutated in both genes, whereas *TLN2* and *ID3* have negative association, indicating poor concordance in mutational patterns, or a high fraction of samples that are mutated in one gene but not the other. This analysis allowed us to identify alterations in the SWI/SNF subunits encoded by *ARID1A* and *SMARCA4* as mutually exclusive events.

MYC translocated DLBCL

We tested 54 cases of DLBCL for the presence of translocations of the MYC gene. We identified four cases (7%) with such translocations. We compared these DLBCL cases with the MYC translocation to DLBCL cases without a *MYC* translocation (N=48), as well as BL cases, all of which had the translocation (N=59). We found that there were no clear differences in gene-coding mutations that distinguished these MYC-translocated cases of DLBCL, suggesting considerable heterogeneity in their biology, similar to that observed previously through gene expression profiling[12,13].

Gene Expression of Recurrently Mutated Genes

We examined the gene expression for the 70 recurrently mutated Burkitt lymphoma genes and performed gene expression analysis across samples in mutated genes. All of these genes were found to be measurably expressed in normal B cells, Burkitt lymphomas and DLBCLs [12-14]. For

Nature Genetics: doi:10.1038/ng.2468

instance, *RET* (shown in Supplementary Fig. 5) was found to be expressed 2-fold higher in germinal center and plasma cells compared to naïve and memory B cells.

ID3 expression in Burkitt lymphoma and DLBCL

ID3 gene expression in Burkitt lymphoma and DLBCL was measured using microarrays. We found that *ID3,* which was never mutated in DLBCLs, was expressed at over two-fold higher levels in Burkitt lymphoma compared to DLBCL (p=0.002). We further examined *ID3* expression between *ID3* wild type and mutant Burkitt lymphoma samples and found that *ID3* is higher in those patient samples with *ID3* mutation (p=0.003). These expression values are plotted as a bar graph in Supplementary Figure 6.

Allele specific expression of *ID3*

We also investigated allele specific expression of *ID3* in Burkitt lymphoma by ligating RNA adapters to RNA from five cases with ID3 mutations. RNA was then reverse transcribed in a strand-specific fashion and subjected to PCR, followed by sequencing. We found that both alleles were measurably expressed in all of the five cases. There was no discernible difference in the expression of the mutated and wild type alleles in these cases (Supplementary Fig. 7).

Gene Set Enrichment Analysis

To discern the differences in gene expression between the *ID3*-mutated and wild type cases, we divided the Burkitt lymphoma cases into two classes, *ID3* mutated (N=6) and *ID3* wild type (N=15). Gene set enrichment analysis[15] (GSEA) was applied between *ID3* wild type and mutant samples using 1000 permutations. We found significant enrichment of 5 different cell cycle gene-sets in the *ID3* mutated group as compared to *ID3* wild type (FDR<0.05). The enrichment plots are depicted in Supplementary Figure 8. In particular, G1-S phase was the only cell-cycle stage-specific gene set to be enriched in the ID3-mutant samples.

The oncogenic role of ID3 mutations in Burkitt lymphoma

The high prevalence of *ID3* mutations in Burkitt lymphoma is striking given that these mutations have not been identified in other malignancies. Given the central role of MYC in Burkitt lymphoma, we reasoned that *ID3* mutations might serve to potentiate the pro-proliferative role of

Nature Genetics: doi:10.1038/ng.2468

MYC. Consistent with this notion, we found that the vast majority of *ID3* mutations occurred in the setting of additional *MYC* mutations. ID-proteins including *ID3*, are important in repressing basic helix-loop-helix proteins such as *MYC*, suggesting that *ID3* may play a role in the repression of *MYC*. Consistent with that notion, a number of mutations in *ID3* were clearly silencing mutations (i.e. nonsense and frameshift).

We investigated the effect of *ID3* mutations on *MYC* target genes by examining the expression of known *MYC* target genes in Burkitt lymphoma cases that either harbored or lacked *ID3* mutations. Using a previously described set of known *MYC* target genes[12], we compared the average *MYC* target expression for these *ID3*-mutated cases and *ID3* wild type cases (Supplementary Fig. 9a). We found that the cases with *ID3* mutation had the highest expression of *MYC*-target genes (P = 3.8 E-7).

ID3 has been described previously as a direct transcriptional target of MYC[16,17]. Thus, if *ID3* played a role in repressing MYC activity, we would expect that silencing mutations in *ID3* would cause increased *ID3* expression. Our data indicate that to be the case (Supplementary Fig. 9a).

Thus, our results are consistent with the hypothesis that silencing *ID3* mutations result in de-repression of *MYC* (Supplementary Fig. 9b), hence increased transcriptional activity of MYC, and consequently increased proliferation. This hypothesis will need to be investigated further.

Expression of ID3 Mutant and Wildtype Proteins

Constructs were made to tag mutant and wild type versions of ID3 to green fluorescent protein (GFP) and overexpressed in cell lines. Their overexpression in cell lines was validated using western blot analysis against ID3 protein. Clear expression of the fusion protein is observed in Supplementary Figure 10. No band was observed in the first column depicting only GFP, indicating the specificity of the antibody. Overexpressing cells were also observed using fluorescence microscopy (FITC filter), indicating fusion protein expression.

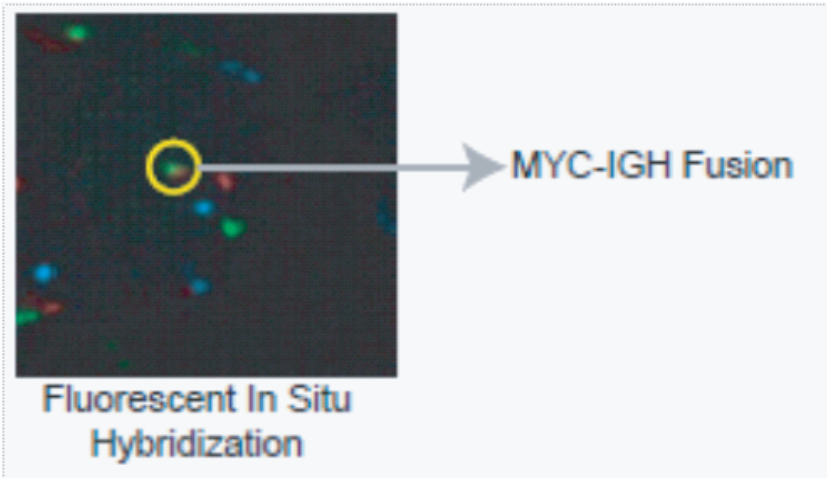
Nature Genetics: doi:10.1038/ng.2468

Supplementary References

1. Kamai, T. et al. Increased Rac1 activity and Pak1 overexpression are associated with lymphovascular invasion and lymph node metastasis of upper urinary tract cancer. *BMC Cancer* **10**, 164 (2010).
2. Altshuler, D.M. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
3. Robinson, J.T. et al. Integrative genomics viewer. *Nat Biotechnol* **29**, 24-6 (2011).
4. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
5. Siva, N. 1000 Genomes project. *Nat Biotechnol* **26**, 256 (2008).
6. Ng, S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-6 (2009).
7. Li, Y. et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42**, 969-72 (2010).
8. Yi, X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-8 (2010).
9. Lohr, J.G. et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci U S A* **109**, 3879-84.
10. Morin, R.D. et al. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* **42**, 181-5.
11. Pasqualucci, L. et al. Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat Genet* **43**, 830-7.
12. Dave, S.S. et al. Molecular diagnosis of Burkitt's lymphoma. *N Engl J Med* **354**, 2431-42 (2006).
13. Hummel, M. et al. A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N Engl J Med* **354**, 2419-30 (2006).
14. Lenz, G. et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci U S A* **105**, 13520-5 (2008).
15. 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).
16. Basso, K. et al. Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**, 382-90 (2005).
17. Seitz, V. et al. Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma. *PLoS One* **6**, e26837 (2011).

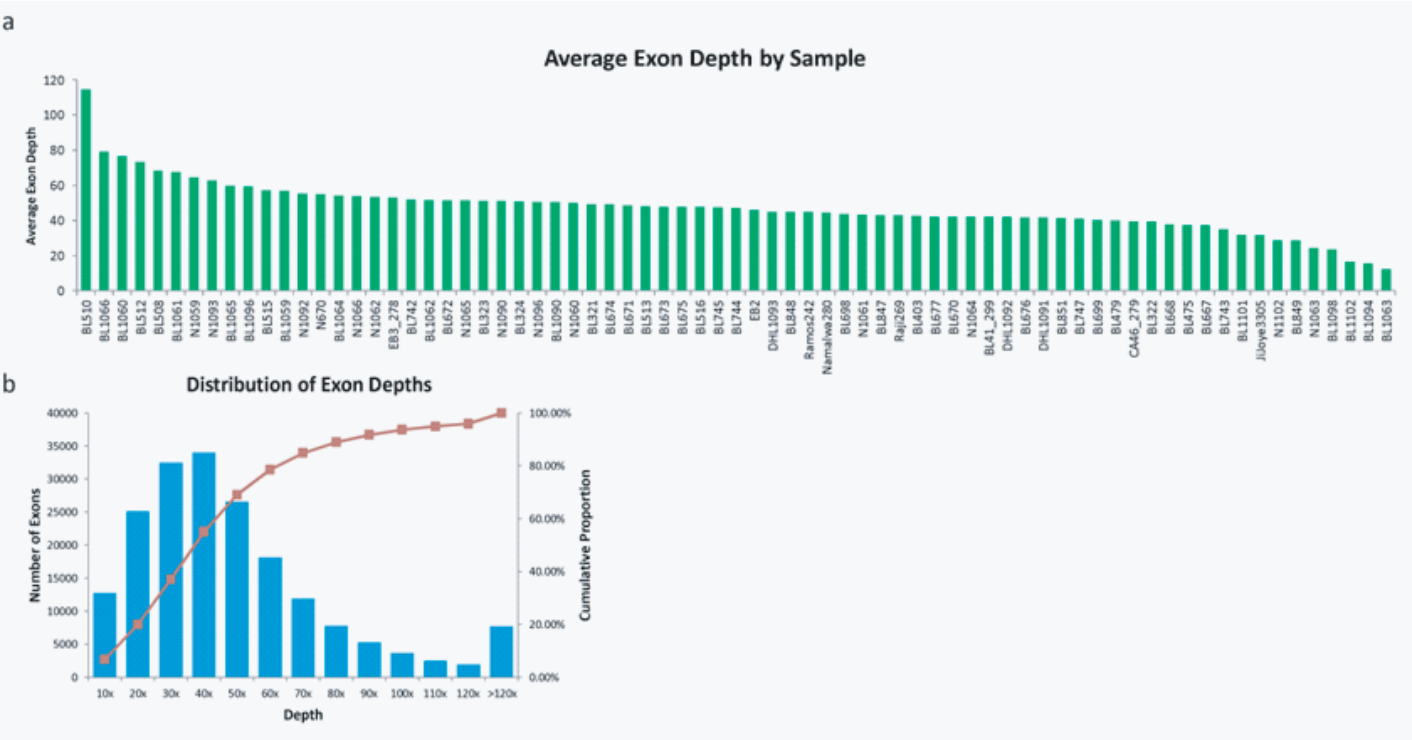
Nature Genetics: doi:10.1038/ng.2468

Supplementary Figure 1: Fluorescence *in-situ* Hybridization
FISH analysis indicates presence of t(8;14) translocation and MYC-IGH fusion in a Burkitt lymphoma sample.

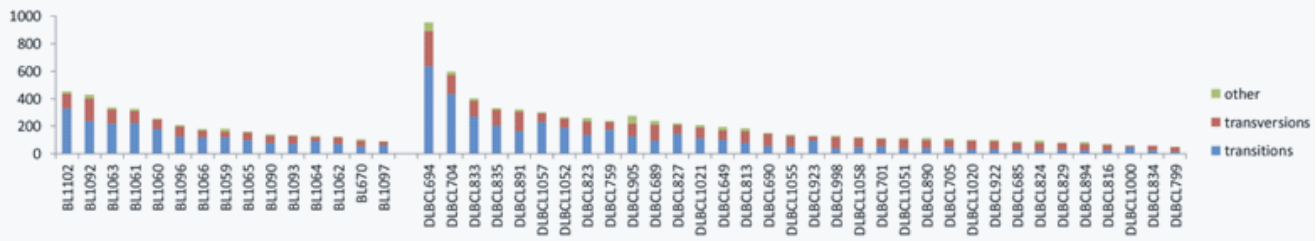


Nature Genetics: doi:10.1038/ng.2468

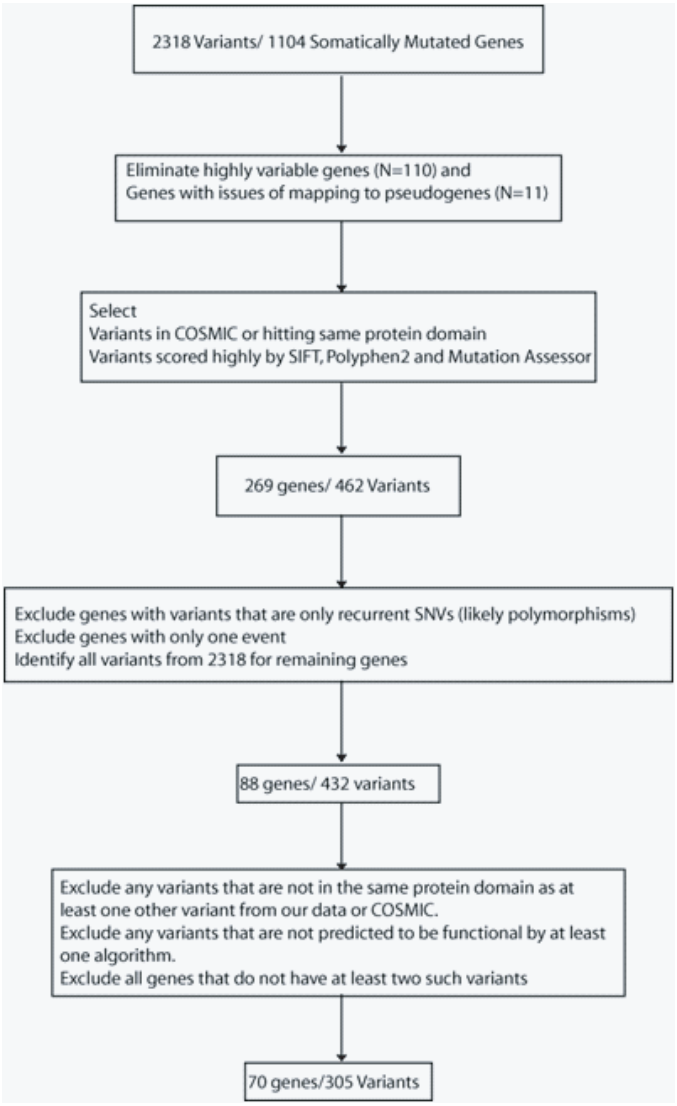
Supplementary Figure 2: a)The average exon read depth by sample is shown for Burkitt lymphoma tumor and paired normal cases. b) The distribution of exon depths is shown in bins of 10. (For example, about 25,000 exons were measured at an average depth between 10x and 20x across all samples, shown for 20X.)



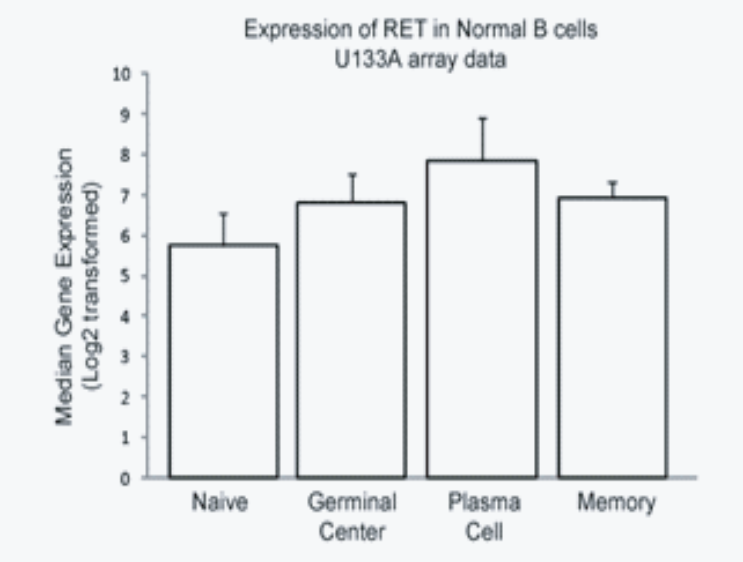
Supplementary Figure 3: Comparison of somatic alteration compositions in BL samples (left) and DLBCL samples (right).



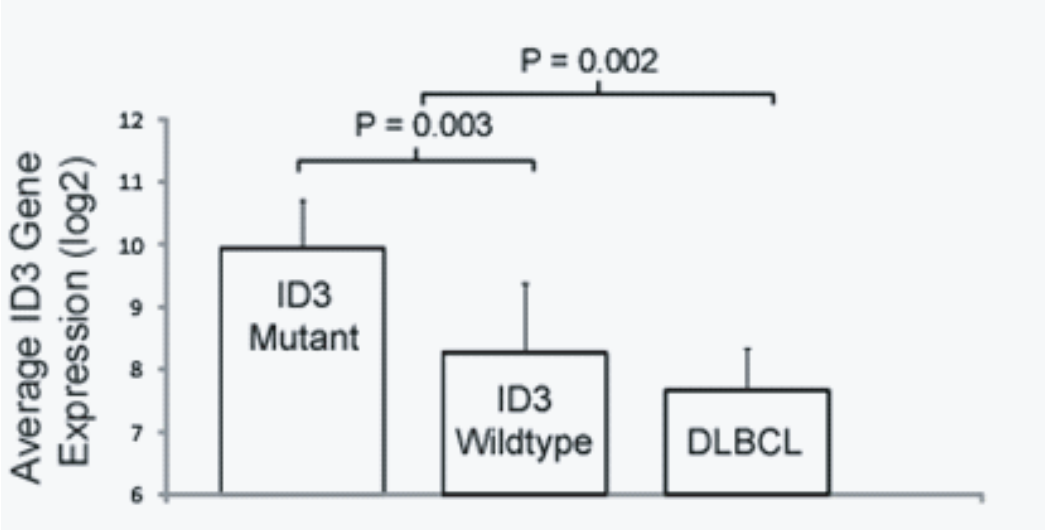
Supplementary Figure 4: Schema for identifying genes recurrently mutated in Burkitt Lymphoma.



Supplementary Figure 5: RET Expression in normal B cells.

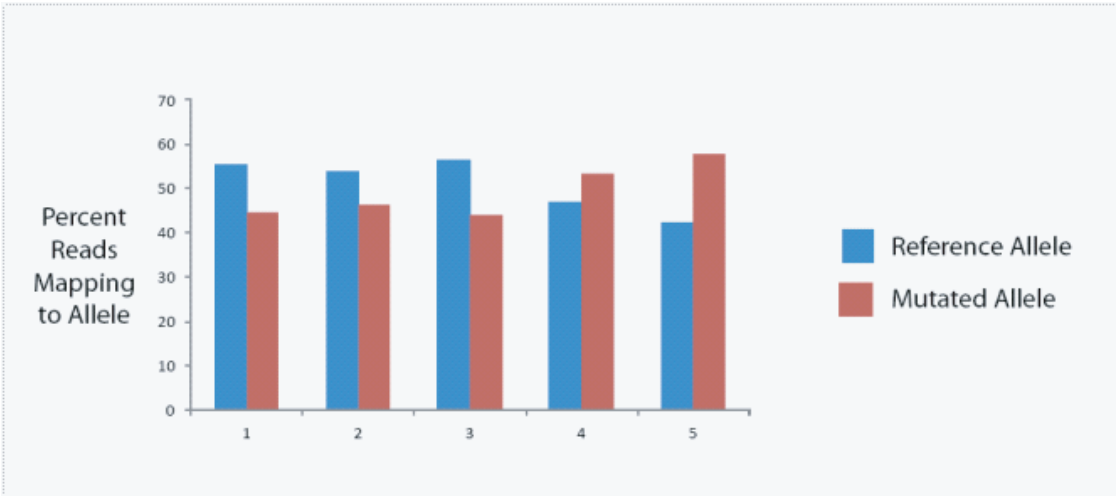


Supplementary Figure 6: *ID3* Expression in ID3 mutant and ID3 wild type Burkitt Lymphoma and DLBCL.

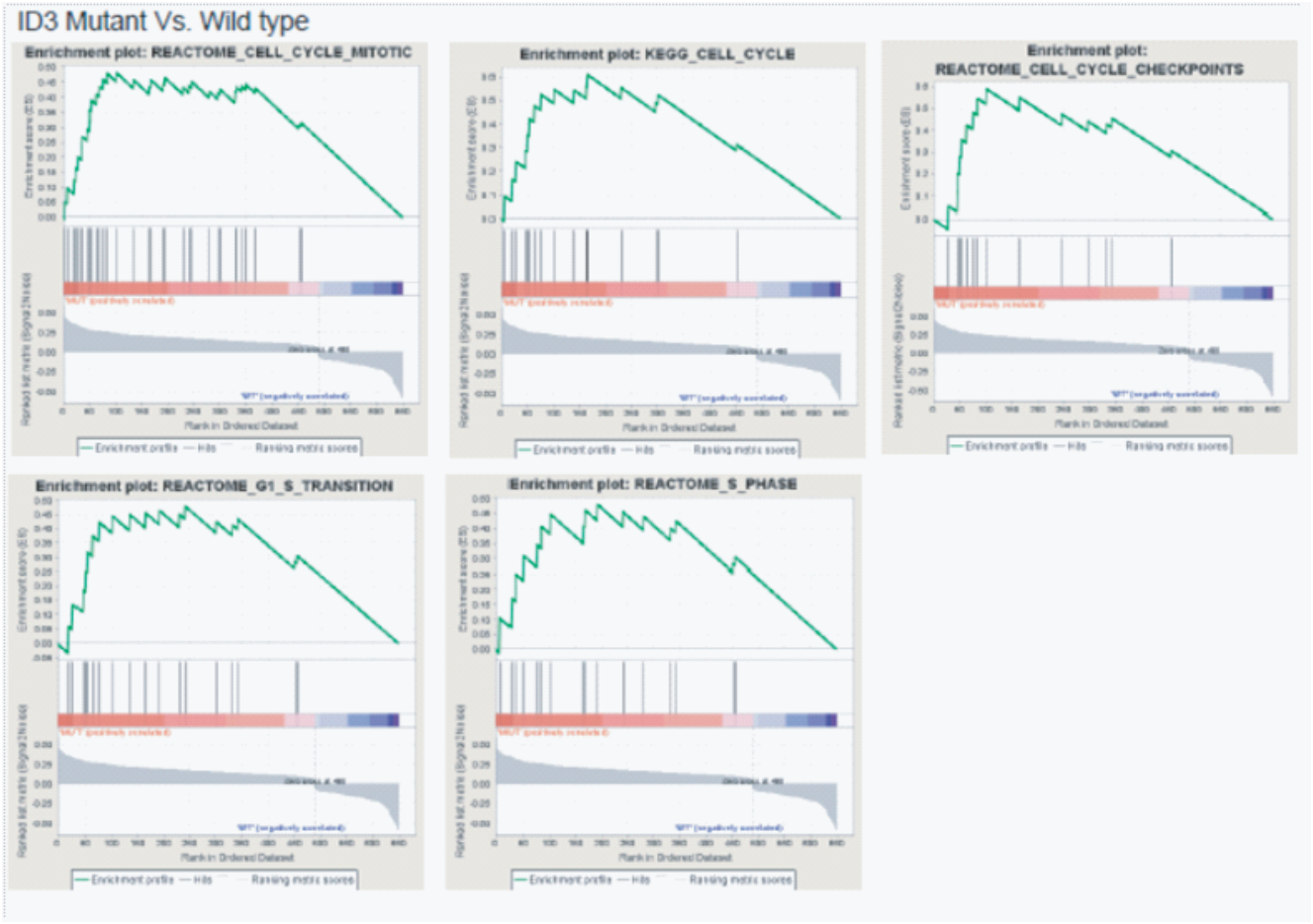


Nature Genetics: doi:10.1038/ng.2468

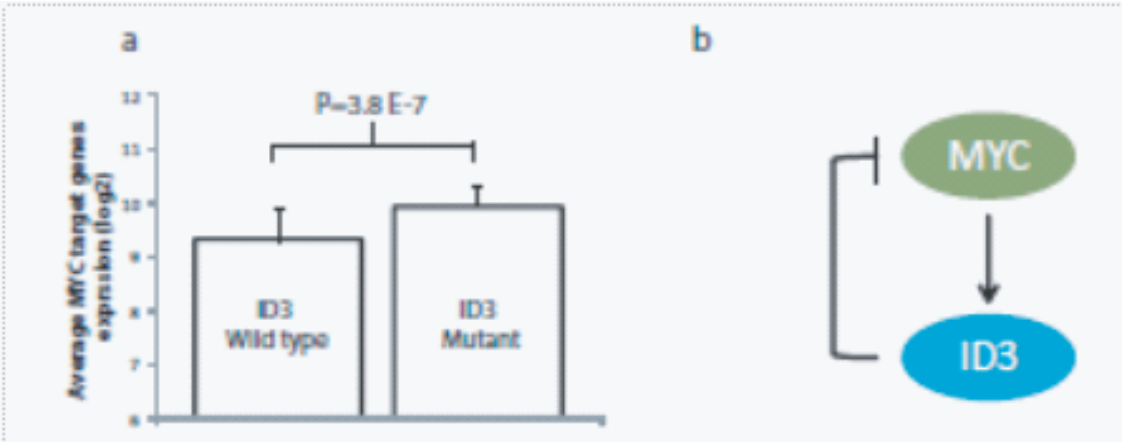
Supplementary Figure 7: Allele Specific Expression of *ID3* in five Burkitt lymphoma cases with ID3 mutations.



Supplementary Figure 8: Gene Set Enrichment Analysis Plots for gene sets enriched in ID3-mutated Burkitt lymphoma cases.



Supplementary Figure 9: a) MYC target gene expression in *ID3*-mutated is higher compared to *ID3*-wild type Burkitt lymphoma cases. b) Proposed mechanism of ID3 and MYC interactions. Silencing mutations in ID3 would serve to de-repress MYC and its target-gene expression.



Supplementary Figure 10: Western blot analysis and GFP microscopy showing expression of ID3-GFP fusion protein and mutant constructs in cell lines.