4. Driver Identification and Genomic Analysis

Our study first sought to understand the landscape of genomic lesions underlying B-NHLs. To accomplish this goal, we began by identifying driver variants within our list of raw sequencing variants. Subsequently, we conducted a genomic landscape analysis and gene-level mutational profiling.

4.1. The Driver Annotation Pipeline

4.1.1. Methodology

We began our analysis by extracting a list of somatic driver variants from our raw sequencing reads. Broadly, our driver identification pipeline consists of three automated steps with a final manual review step to check all variants (Figure 3). Our pipeline first removes errors from the list of all sequencing variants (VCF file) to construct a list of all real variants. Second, our pipeline identifies somatic variants by annotating polymorphisms. Third, our pipeline annotates somatic variants as drivers, passengers, or variants of unknown significance. Finally, all variants are manually curated, taking into account the flags set by the pipeline.

First, we removed errors from the list of sequencing variants. We removed errors resulting from DNA polymerase slippage by discarding variants that were (1) in homopolymeric regions of length greater than 4 and (2) in >10% of individuals. We removed variants near the noise thresholds of the CaVEMan and Pindel algorithms by discarding variants with a read depth less than 10, less than three reads, or a VAF less than 0.05. For context, our study had an average depth of 500x reads per base. Our filters are consistent with those used in prior studies[143]. Nonetheless, we also inspected both the remaining and discarded variants with GBrowse. By removing errors in this fashion, we pruned our list of sequencing variants to the set of all real variants in our study.

Second, we identified somatic mutations by flagging polymorphisms within our list of variants. Since our tumour samples lacked matched normals, we identified likely polymorphisms by flagging variants with a population frequency in ExAC non-TCGA greater than 0.001. Since ExAC non-TCGA includes some lymphoid drivers with a high population frequency, we kept a whitelist of drivers that would not be annotated as polymorphisms via this approach. No variants were removed via this step. The annotation, however, proved

helpful for manually curating drivers. Upon completion of this step, we arrived at a list of variants, some flagged as likely polymorphisms.

Third, we annotated driver mutations. We utilized a few computational approaches described below. Ultimately, however, all variants were inspected and given a final annotation manually. Three independent computational approaches were helpful in flagging potential drivers. First, we flagged all mutations that were in a whitelist of known driver mutations manually curated from COSMIC and the literature. Second, we flagged variants as potential drivers if they were highly recurrent within COSMIC (>3). Finally, we flagged variants as potential drivers if their effect in a gene of known function was likely to make them drivers. For example, a frameshift or nonsense mutation in a well-characterized tumour suppressor gene would be marked as a likely driver. Since this approach requires a functional annotation for each gene, it was only applied to a subset of the variants.

Finally, with a list of potential driver mutations we conducted an extensive manual curation to provide a final annotation to variants. In general, we annotated variants conservatively, preferring to err on the side of marking a variant as a "Variant of Unknown Significance" rather than a driver. Conservative annotation would reduce later errors in classification since the Bayesian Dirichlet Process, our classification algorithm, is more robust to false negatives (i.e. missing drivers) than to false positives (i.e. passenger mutations annotated as drivers).

4.1.2. Limitations of the Driver Annotation Pipeline and Mutations Underrepresented in DLBCL NOS

In general, the driver variants produced via our driver annotation pipeline matched expectations from the literature (Sections 4.2, 4.2.1). However, mutations in some DLBCL genes were underrepresented (*BCL2, BCL6, CIITA, CD79B, PIM1, HIST1H1E, CD58, GNA13*). Limitations of the data, the driver annotation pipeline, or the sequencing and assembly algorithms can account for these discrepancies.

First, some genes had low mutation levels based on the lack of translocation data or copy number analysis. *BCL2*, for example, was present at a lower proportion than expected (34-45% of patients in literature[144]). However, the majority of *BCL2* changes in DLBCL result from translocation; therefore, the lower prevalence of *BCL2* driver mutations in our *sans translocation* dataset can be explained. The same is true for *BCL6* and *CIITA* (33% and 38% of patients in literature, respectively[144]). The addition of translocation and copy number analysis to future versions of this study should resolve the above issues.

Second, other genes had low mutation levels due to limitations of the computational pipeline which will be improved in future iterations. Note that for all genes below, the relevant variants were indeed present within our list of real variants but were not flagged as drivers. *CD79B* had a hotspot within our list of real variants at Y197 that was not flagged as a driver. Our computational pipeline failed to annotate this hotspot because (1) it was not present within our driver whitelist and (2) our sequencing aligned to a distinct transcript of *CD79B* than that used in COSMIC; therefore, our hotspot was present at Y197 rather than COSMIC's hotspot at Y196, meaning the COSMIC recurrence flag did not call it as a hotspot. To ensure inclusion of this hotspot in the future, we plan to update the driver whitelist, ensure consistency of transcripts between our sequencing pipeline and COSMIC, and additionally flag any variants that are highly recurrent within our dataset as likely drivers.

Two other genes, *PIM1* and *HISTIH1E*, had numbers of total driver mutations lower than expected based on the literature. *HIST1H1E* has been reported to have a large number of missense mutations spread throughout the coding sequence of the gene without any obvious hotspots. *PIM1* is similar, except a few codons show recurrence > 10 in COSMIC (S97 – 14; E79 – 11; and L2 – 10). Our list of real variants indeed contained missense mutations spread throughout the coding sequence of these genes consistent with previously reported patterns. Since it is unclear, however, which of these specific missense mutations are the driver mutations and which are passenger mutations, our pipeline marked these as variants of unknown significance with the exception of the recurrently mutated codons (*PIM1* S97, E79, and L2). By comparison, other studies[15] often include these missense mutations which explains the disparity in mutation frequency. Annotating missense variants that are not in hotspots and lack biological validation as drivers remains a challenge.

Finally, our variant caller CaVEMan has a statistical limit at calling variants with VAF <5%[135] which can miss subclonal mutations. A future solution to this problem would involve utilizing DeepSNV[145], a relatively new variant caller which effectively calls variants at VAF < 5% without introducing significant errors. The variant calls resulting from both algorithms could then be manually reviewed and merged to create a more accurate set of variant calls.

Any remaining low mutation levels not due to the factors described above are likely due to other inherent limitations of our pipeline. The biological effects method requires a functional annotation (i.e. oncogene or tumour suppressor gene) which is not always present. Manual curation can be challenging, especially for missense variants with low recurrence in genes that have not had extensive previous characterization. Overall, however, since multiple

independent methods are used to annotate a driver, our results are generally accurate. With the exception of the genes described above, the genomic landscape of DLBCL NOS was consistent with expectations from the literature. We suspect that future versions of this work implementing the changes above will make the genomic landscape fully consistent.

4.1.3. Limitations of the Dataset

Before proceeding further, it is worth noting the limitations of our genomic landscape analysis and gene-level mutational profiling described below. First, the data analysed for this manuscript does not incorporate translocations fundamental to the pathogenesis of DLBCL, FL, and BL; namely translocations in *IGH/BCL2*, *BCL6*, and *MYC*[19]. Second, the data did not include any copy number analysis. As a result, amplifications and copy number gains that are well characterized and important to the pathogenesis of DLBCL were missing: iR-17~92, 2p16.1, *BCL2*, and *SPIB*[19]. While our targeted sequencing analysis was designed to detect changes in copy number, the targeted and unmatched nature of the sequencing data meant that traditional copy number analysis algorithms like Ascat[146] would not work. At present, a custom algorithm is being designed and implemented to detect copy number changes in this dataset. Finally, gene expression data was not provided for these samples. As a result, the samples could not be clustered into cell of origin clusters (i.e. ABC-DLBCL, GCB-DLBCL) which would then have enabled an analysis of genomic landscape differences between these subtypes, potentially enabling further resolution and highlighting similarities.

All of the above data are either present within or can be extracted from our collaborators' full dataset. However, it was either not received or not processed in time for this publication. A final analysis of this lymphoma dataset is currently being conducted with the aim of incorporating the translocation, copy number, and gene expression data. We expect some important changes to result from the addition of this data. For example, all BL samples should exhibit a *MYC* translocation—the hallmark genetic change of the disease[19]. Nonetheless, the broad genetic changes shown within this publication to underlie DLBCL, FL, and BL should not change and meaningful conclusions can thus still be drawn.

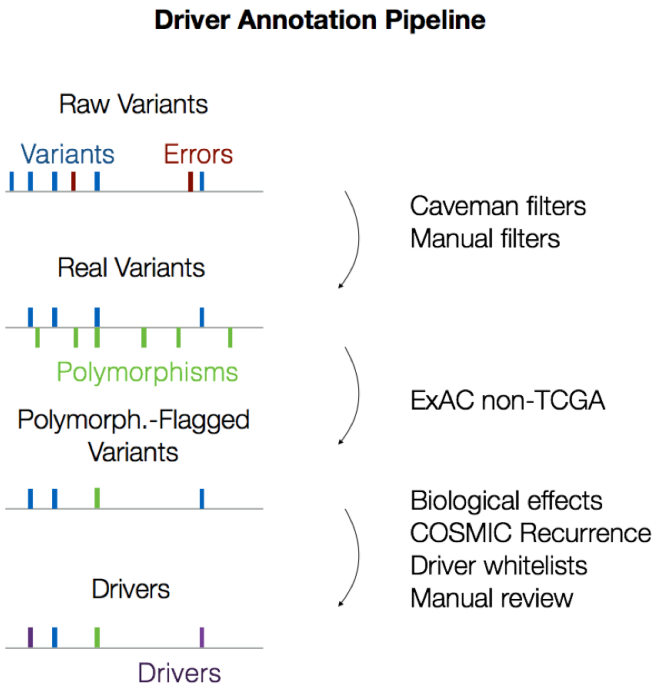


Figure 3 The driver annotation pipeline. The driver annotation pipeline annotates drivers from sequencing variants in three steps.

4.2. Genomic Landscape of Lymphoma

After identifying the driver mutations present within each dataset, we sought to gain an understanding of the genomic landscape of the B-NHLs within our dataset and of the DLBCL NOS subtype more specifically.

4.2.1. The Genomic Landscape of DLBCL NOS

Looking at the genomic landscape of drivers in just DLBCL NOS (Figure 5c), we note that driver mutations generally matched expectations consistent with the literature with a few exceptions discussed in Section 4.1.2. At a high level, the genomic landscape of DLBCL NOS exhibited a classic long tail distribution, with a small number of genes containing the majority of genetic lesions and a large number of genes more rarely mutated but collectively responsible for a large proportion of mutations.

At the gene level, the most prevalent mutations expected from DLBCL were present: chromatin modifications (*CREBBP, EP300, KMT2D*), immune escape (*B2M*), deregulated BCL6 activity (*MEF2B*), proliferation and apoptosis (*MYC*), signalling (*TNFRSF14, SGK1, PTEN*), constitutive NF-KB/BCR activity (*TNFAIP3, MYD88, CARD11*), terminal differentiation (*PRDM1*), the cell cycle checkpoint (*CDKN2A*), and JAK/STAT activation (*SOCS1*).

4.2.2. Comparative Genomic Landscapes of DLBCL NOS, FL, and BL

To understand how the genomic landscapes of DLBCL NOS, FL, and BL differed, we plotted driver mutations across all genes and highlighted which fraction of driver mutations within each gene came from which diagnostic subtype (Figure 5a).

4.2.2.1. DLBCL NOS vs. FL

Comparing the genomic landscape of DLBCL NOS with that of FL (Figure 5c, d) reveals telling differences and similarities in the genomic causes of the diseases.

First at a high level, both FL and DLBCL NOS exhibited classic long tail distributions. A small number of genes (*KMT2D, CREBBP, TNFRSF14, EZH2, ARID1A*) accounted for a large proportion of driver mutations found in patients. A high number of genes then individually had fewer drivers present yet still accounted for a large proportion of drivers when taken collectively. While the broad long-tail profile matches that of DLBCL NOS, FL had a “tighter tail”: more driver mutations concentrated in a smaller number of genes (*KMT2D, CREBBP, TNFRSF14, EZH2, ARID1A*). Collectively, these observations

point to the increased genetic heterogeneity of DLBCL compared to FL, a result consistent with expectations in the literature[19].

Second, strong similarities occur at the gene level between the DLBCL NOS and FL subtypes. Note that for both DLBCL NOS (n=925) and FL (n=566), a small number of genes contain the majority of driver mutations: *KMT2D, CREBBP, TNFRSF14, TP53, SOCS1, B2M, ARID1A, CCND3, TNFAIP3* (constitutive NF-KB activity), and *IRF8*. This strong overlap points to the strong genomic similarities present between DLBCL NOS and FL and thus similar mechanistic deregulations that enable the progression of cancer. For example, the commonalities in *KMT2D*, *CREBBP*, and *EZH2* point to the importance of epigenetic dysregulation in both FL and DLBCL NOS through similar mechanisms. Similarly, the prevalence of driver mutation in *SOCS1*, *TNFRSF14*, and *TNFAIP3* enable aberrant signalling leading to proliferation via the JAK/STAT and NF-KB pathways respectively.

Third, the prevalence of *B2M* mutations demonstrate the importance of immune escape. While at a population level, similar genes are mutated in DLBCL NOS and FL, it's worth noting that individual patients within each subtype can still have distinct combinations of mutations that distinguish the diseases. Patients of both FL and DLBCL NOS have, on average, multiple driver mutations (Figure 4). Therefore, even if two patients share a single driver mutation they may differ in the additional driver mutations they have acquired: a DLBCL NOS patient could, for example, have driver mutations in *KMT2D* and *CREBBP* while a FL patient could have driver mutations in *KMT2D* and *TNFRSF14*. Because these diseases rely on multiple driver mutations and the dysregulation of multiple pathways, substantial differences in pathogenesis and treatment response can result. Overall, this result reinforces the need for multifactorial classification. While it's unlikely that most mutations in specific genes can be assigned exclusively to DLBCL NOS or FL, it still may be the case that specific combinations of mutations occur uniquely in DLBCL NOS vs. FL. Therefore, a multifactorial classification system such as the Bayesian Dirichlet Process is needed.

Finally, important differences between DLBCL NOS and FL nonetheless persist. For DLBCL NOS patients, mutations in *MYD88, TET2, BTG2, NOTCH2, IRF4*, and *RHOA* appear to happen at a higher proportion than for patients with any another subtype. For FL patients, mutations in *MEF2B* and *STAT6* appear to happen at a higher proportion than for patients with any another subtype. The high prevalence of these mutations within their corresponding subtypes point to the importance of those mutations to the unique pathogenesis mechanisms inherent to that particular subtype. *MYD88*, for example, has a well known L265P hotspot unique to DLBCL although the precise clinical and pathological significance

is unknown[147]. Similarly, activating mutations in the *STAT6* transcription factor are known to improve B-cell survival in FL[148]. From a classification perspective, therefore, we expect mutations in these genes to become "class defining" lesions that enable us to distinguish such subtypes.

4.2.2.2. DLBCL NOS vs. BL

While DLBCL NOS and FL are largely similar with a few distinct class defining lesions, BL (Figure 5e) appears to have strong genetic differences with the DLBCL NOS and FL subtypes. Note that the genes which contained a high proportion of the driver mutations in FL and DLBCL NOS (*KMT2D, CREBBP, TNFRSF14, EZH2, TP53, SOCS1, B2M, ARID1A, CCND3, TNFAIP3, IRF8*) contain a far lower proportion of driver mutations in BL. Conversely, individual genes that were rarely mutated in FL and DLBCL NOS such as *ID3* and *TCF3*, now contain high proportions of the driver mutations in BL. From a mechanistic level, *ID3* and *TCF3* are well known mutations specific to the pathogenesis of BL that often work in conjunction with the *MYC* translocation – the hallmark of BL[149,150]. Combined, these observations point to a substantially distinct genetic landscape of BL as compared to DLBCL NOS and FL. Therefore, we expect the classification to draw a distinct and separate category for BL as separate from DLBCL NOS and FL that is more easily distinguishable than the categories drawn between DLBCL NOS and BL.

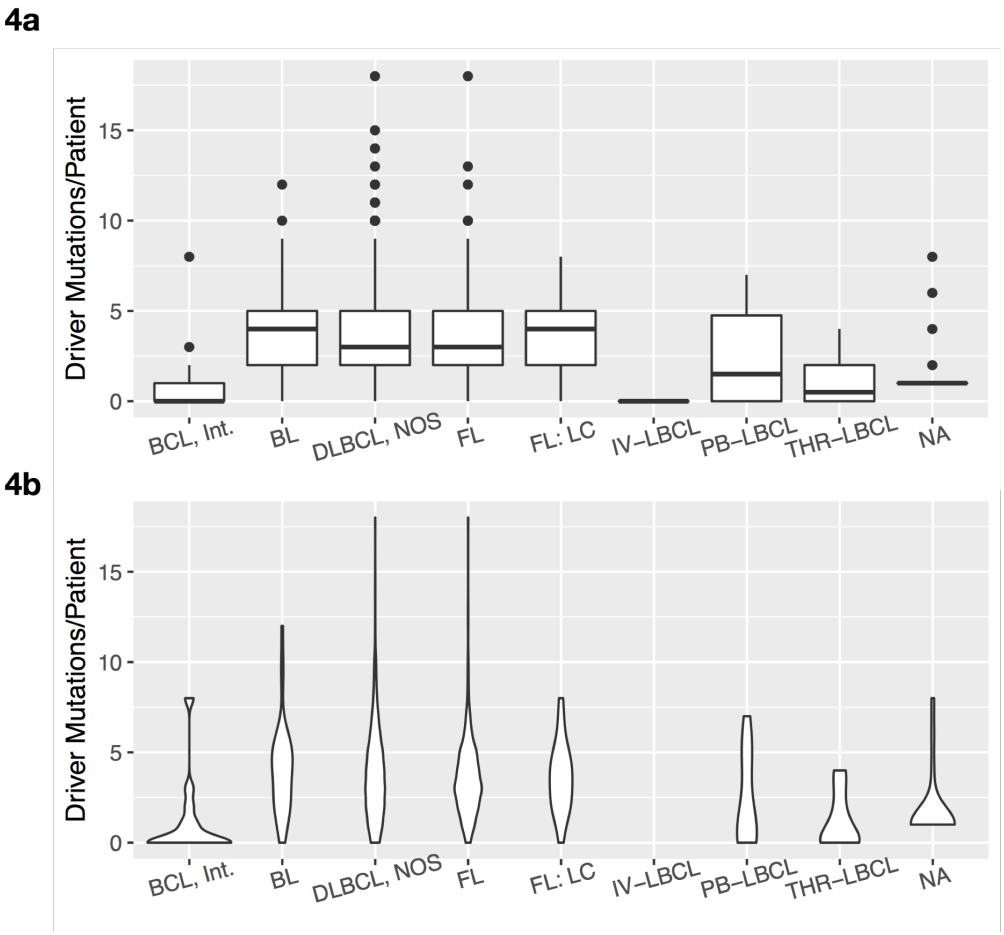
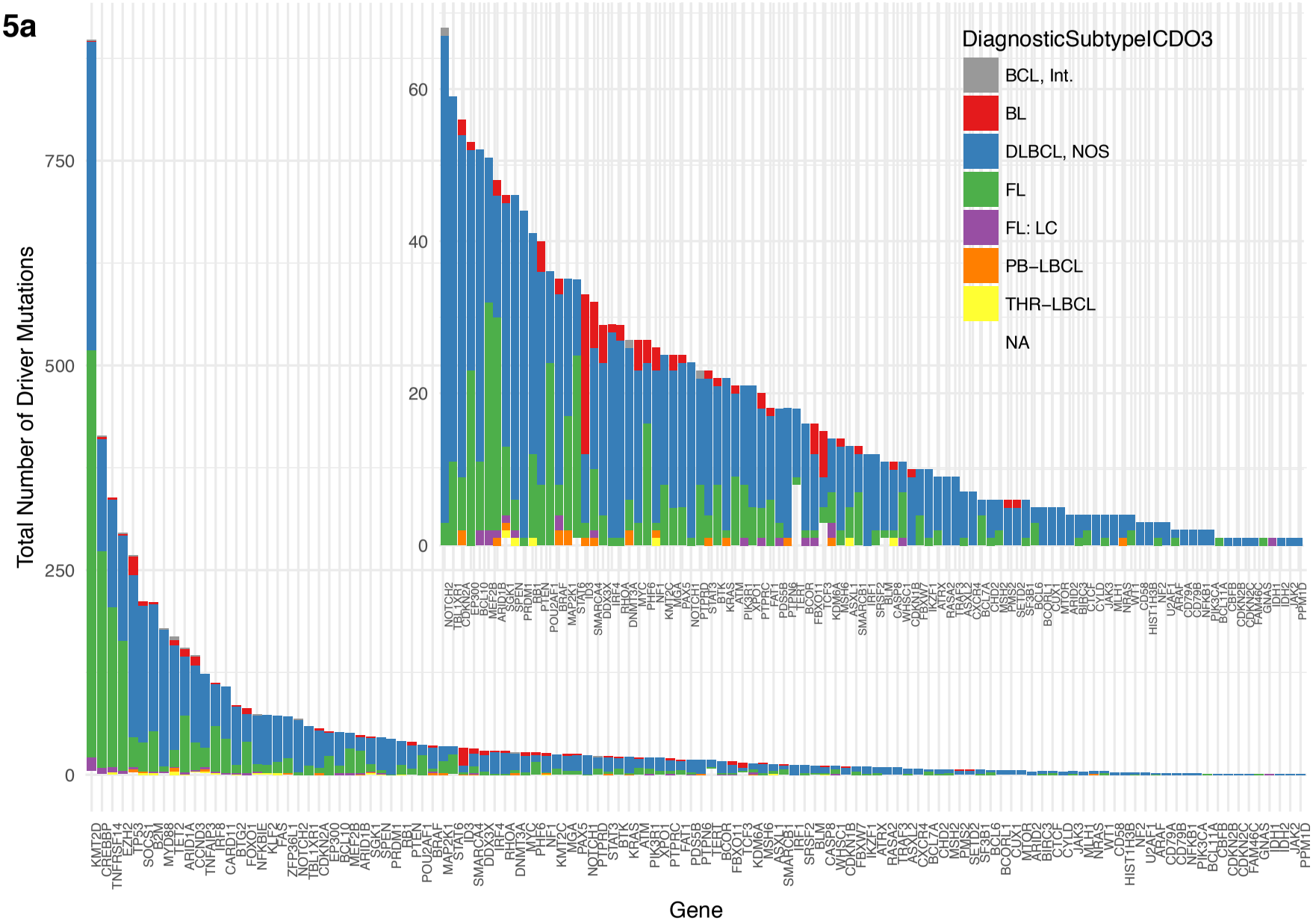


Figure 4 B-NHLs exhibit 3-4 driver mutations/patient. Average number of somatic driver mutation per patient across different diagnostic subtypes in this study. **(a)** Boxplot. Line represents median; hinges represents first and third quartile; whiskers represent furthest data point from quartile within 1.5X the interquartile range. Individual points represent outliers beyond that range. **(b)** Violin plot.

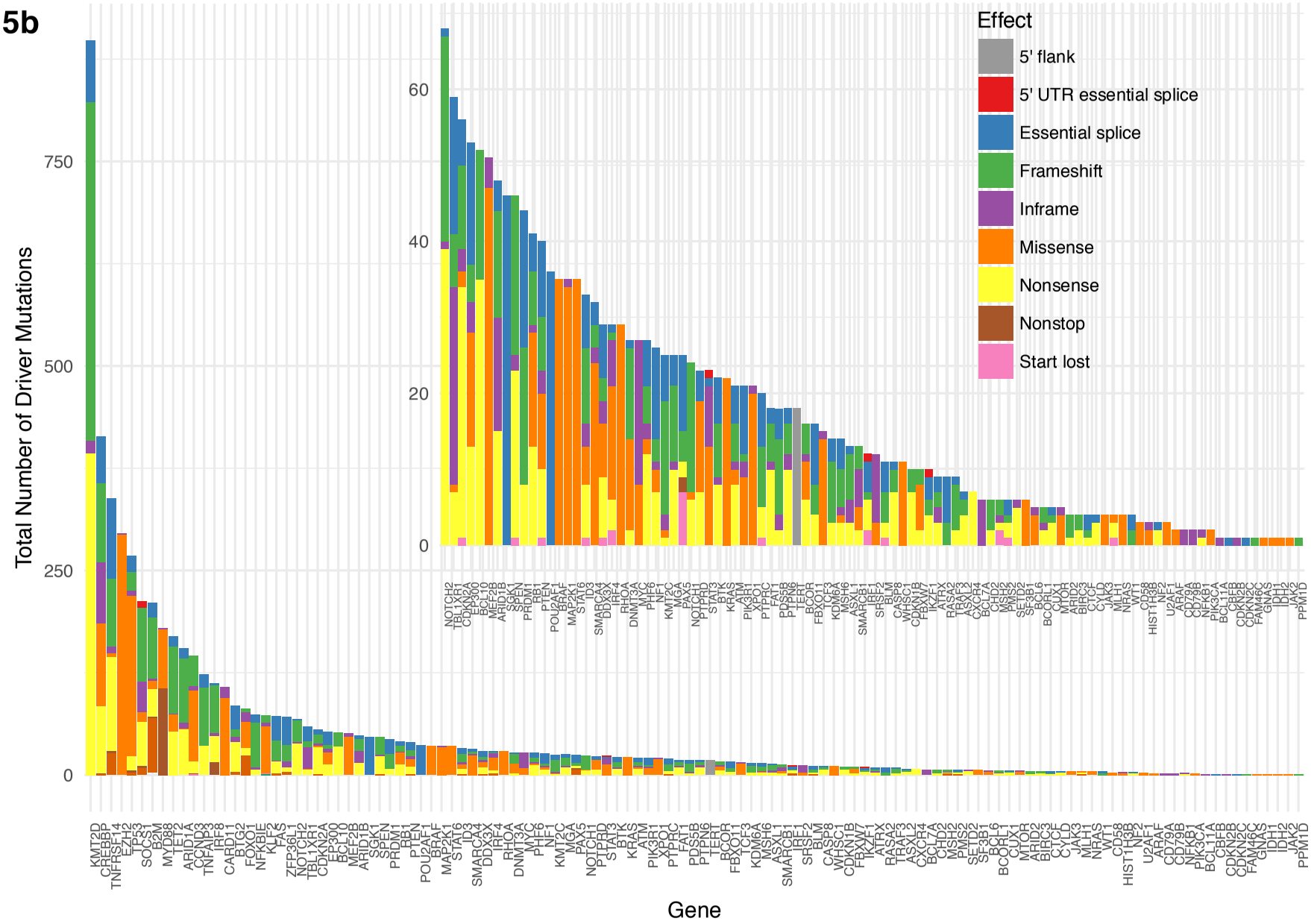
All Subtypes

5a



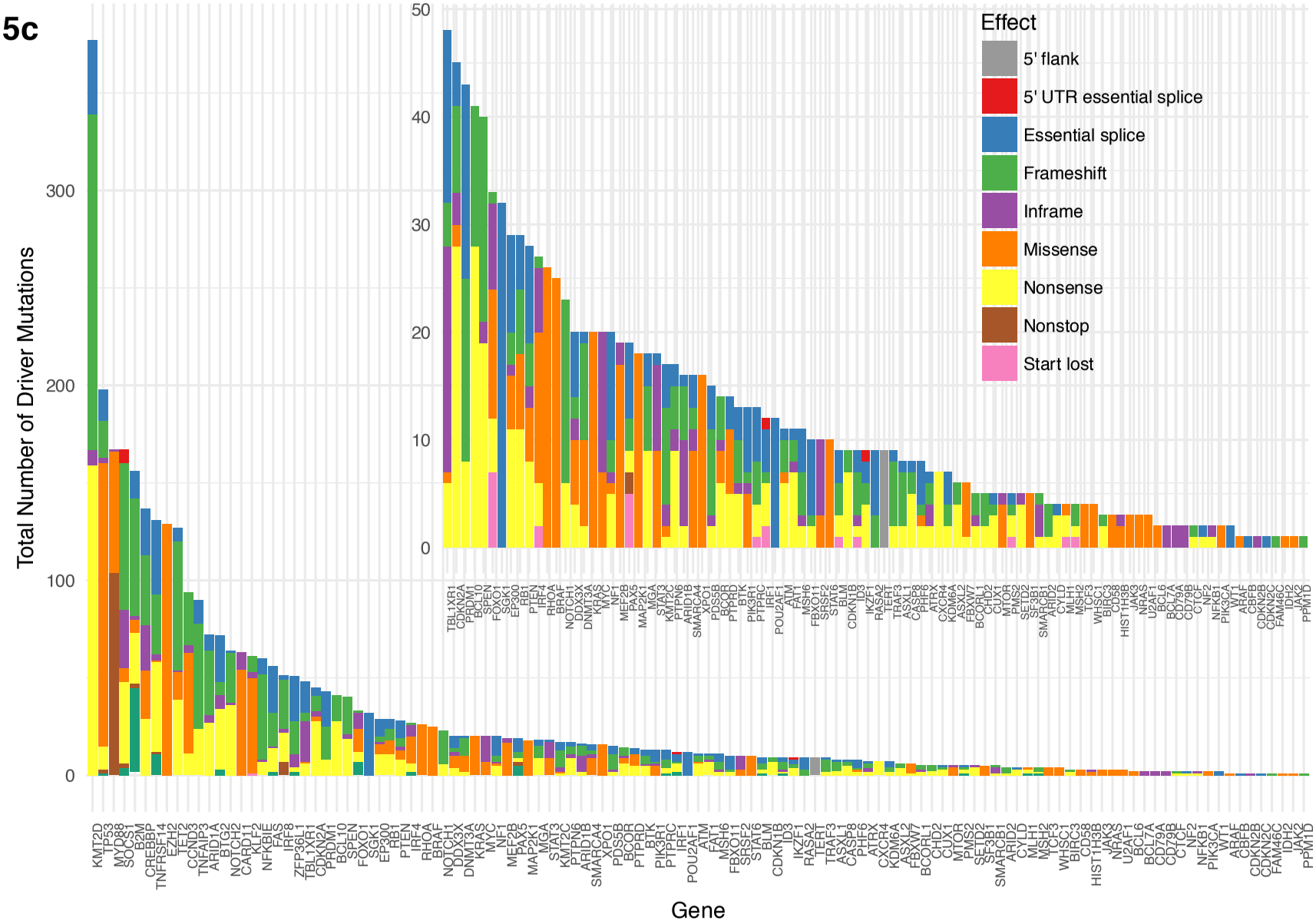
All Subtypes

5b



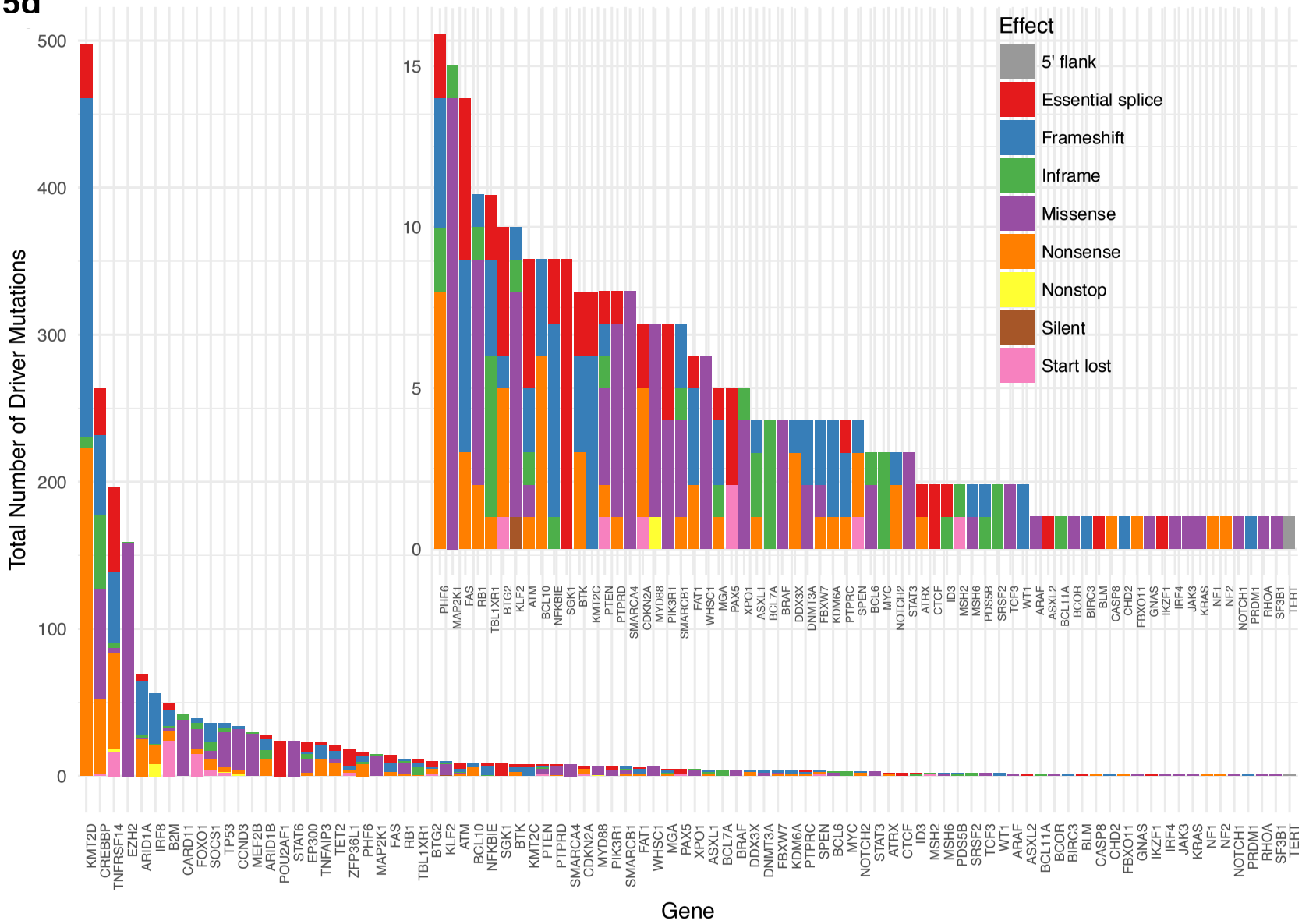
DLBCL NOS

5c


Total Number of Driver Mutations
0
100
200
300
50
40
30
20
10
0
Effect
5' flank
5' UTR essential splice
Essential splice
Frameshift
Inframe
Missense
Nonsense
Nonstop
Start lost
KMT2D TP53 MYD88 SOCS1 B2M CREBBP TNFRSF14 EZH2 TET2 CCND3 TNFAIP3 ARID1A BTG2 NOTCH2 CARD11 KLF2 NFKBIE FAS IRF8 ZFP36L1 TBL1XR1 CDKN2A PRDM1 BCL10 SPEN FOXO1 SGK1 EP300 RB1 PTEN IRF4 RHOA BRAF NOTCH1 DDX3X DNMT3A KRAS MYC NF1 MEF2B PAX5 MAP2K1 MGA STAT3 KMT2C PTPN6 ARID1B SMARCA4 XPO1 PDS5B BCOR PTPRD BTK PIK3R1 PTPRC IRF1 POU2AF1 ATM FAT1 MSH6 FBXO11 SRSF2 STAT6 BLM CDKN1B ID3 IKZF1 RASA2 TERT TRAF3 ASXL1 CASP8 PHF6 ATRX CXCR4 KDM6A ASXL2 FBXW7 BCORL1 CHD2 CUX1 MTOR PMS2 SETD2 SF3B1 SMARCB1 ARID2 CYLD MLH1 MSH2 TCF3 WHSC1 BIRC3 CD58 HIST1H3B JAK3 NRAS U2AF1 BCL6 BCL7A CD79A CD79B CTCF NF2 NFKB1 PIK3CA WT1 ARAF CBFB CDKN2B CDKN2C FAM46C IDH2 JAK2 PPM1D
Gene

FL

5d


Effect
5' flank
Essential splice
Frameshift
Inframe
Missense
Nonsense
Nonstop
Silent
Start lost
Total Number of Driver Mutations
0
100
200
300
400
500
0
5
10
15
Gene
KMT2D CREBBP TNFRSF14 EZH2 ARID1A IRF8 B2M CARD11 FOXO1 SOCS1 TP53 CCND3 MEF2B ARID1B POU2AF1 STAT6 EP300 TNFAIP3 TET2 ZFP36L1 PHF6 MAP2K1 FAS RB1 TBL1XR1 BTG2 KLF2 ATM BCL10 NFKBIE SGK1 BTK KMT2C PTEN PTPRD SMARCA4 CDKN2A MYD88 PIK3R1 SMARCB1 FAT1 WHSC1 MGA PAX5 XPO1 ASXL1 BCL7A BRAF DDX3X DNMT3A FBXW7 KDM6A PTPRC SPEN BCL6 MYC NOTCH2 STAT3 ATRX CTCF ID3 MSH2 MSH6 PDS5B SRSF2 TCF3 WT1 ARAF ASXL2 BCL11A BCOR BIRC3 BLM CASP8 CHD2 FBXO11 GNAS IKZF1 IRF4 JAK3 KRAS NF1 NF2 NOTCH1 PRDM1 RHOA SF3B1 TERT

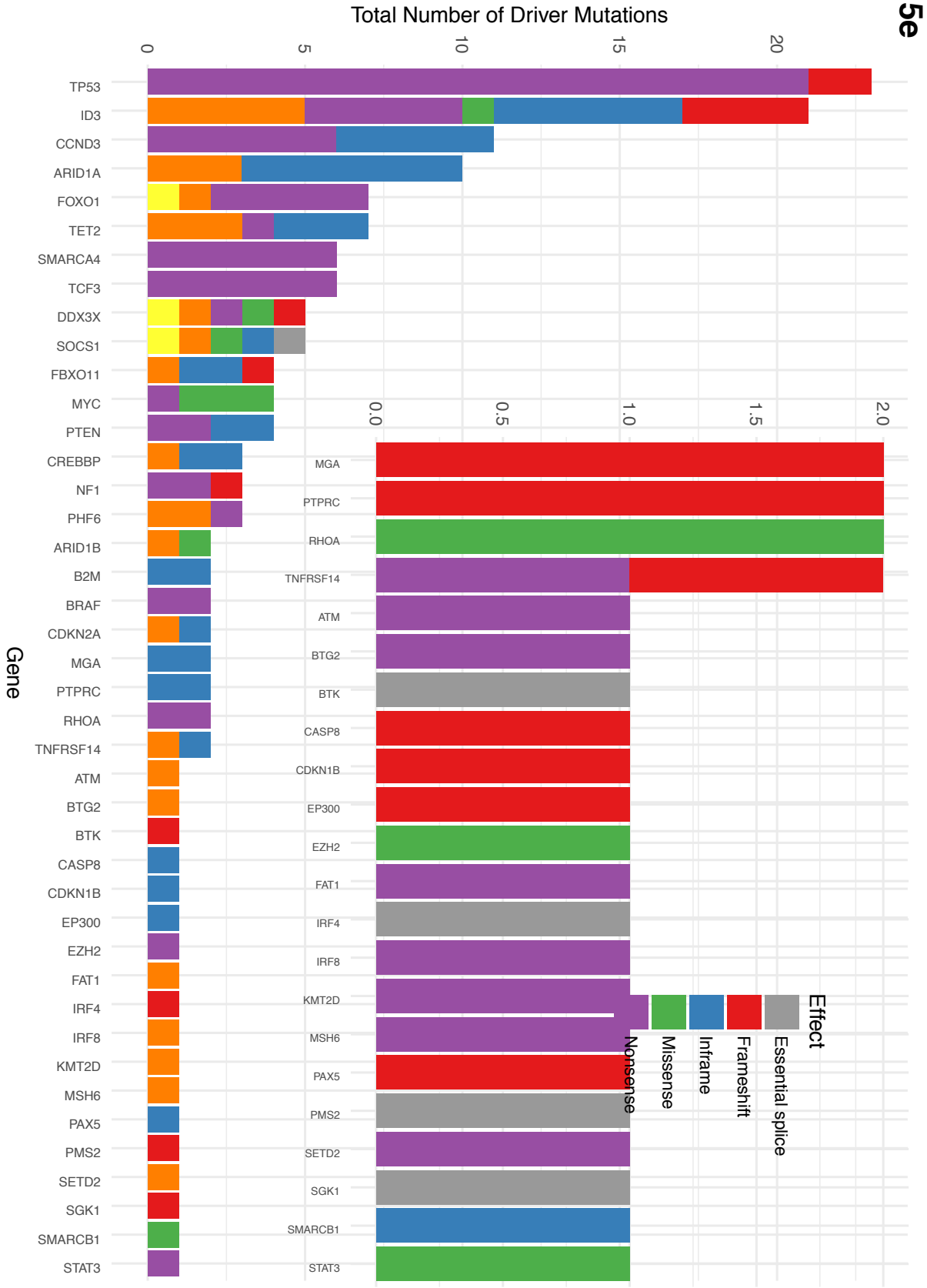

5e
BL
Total Number of Driver Mutations
0
5
10
15
20
Gene
TP53
ID3
CCND3
ARID1A
FOXO1
TET2
SMARCA4
TCF3
DDX3X
SOCS1
FBXO11
MYC
PTEN
CREBBP
NF1
PHF6
ARID1B
B2M
BRAF
CDKN2A
MGA
PTPRC
RHOA
TNFRSF14
ATM
BTG2
BTK
CASP8
CDKN1B
EP300
EZH2
FAT1
IRF4
IRF8
KMT2D
MSH6
PAX5
PMS2
SETD2
SGK1
SMARCB1
STAT3
0.0
0.5
1.0
1.5
2.0
Effect
Nonsense
Missense
Inframe
Frameshift
Essential splice

Figure 5 B-NHL Diagnostic subtypes comprise distinct genomic landscapes. (a) Driver mutations identified in all B-NHL subtypes, coloured by diagnostic subtype in which they are identified. **(b)** Driver mutations identified in all B-NHL subtypes, coloured by effect of mutation. **(c)** Driver mutations identified in DLBCL NOS, coloured by effect of mutation. **(d)** Driver mutations identified in FL, coloured by effect of mutation. **(e)** Driver mutations identified in BL, coloured by effect of mutation.

4.3. Gene-Level Mutational Profiling

After analysing the genomic landscape of BL, FL, and DLBCL at a population level, we analysed the genetic lesions incurred on each gene within our bait set. Overall, we were able to reproduce expected mutation patterns in well-characterized oncogenes and tumour suppressor genes. Additionally, we identified new patterns of recurrence and novel driver mutations of biological interest.

4.3.1. Recreation of Expected Mutational Profiles

First, we accurately reproduced expected genetic mutation profiles for key genes in DLBCL, FL, and BL.

4.3.1.1. Well-Characterized Tumour Suppressor Genes

As expected, well-characterized tumour suppressor genes exhibit a range of disrupting mutations (frameshift, missense, and nonsense) spread throughout the coding sequence of a given gene (Figure 6). The diversity in both type of disrupting mutation and residue targeted result from the fact that truncating a protein along its primary sequence, shifting the frame of large regions, or even disrupting an amino acid can cause a loss-of-function, regardless of the specific residue within which such a change occurs (Figure 6a). Broadly therefore, these patterns of disrupting mutation spread throughout the coding sequence of a gene correspond to tumour suppressor genes and were identified within our study.

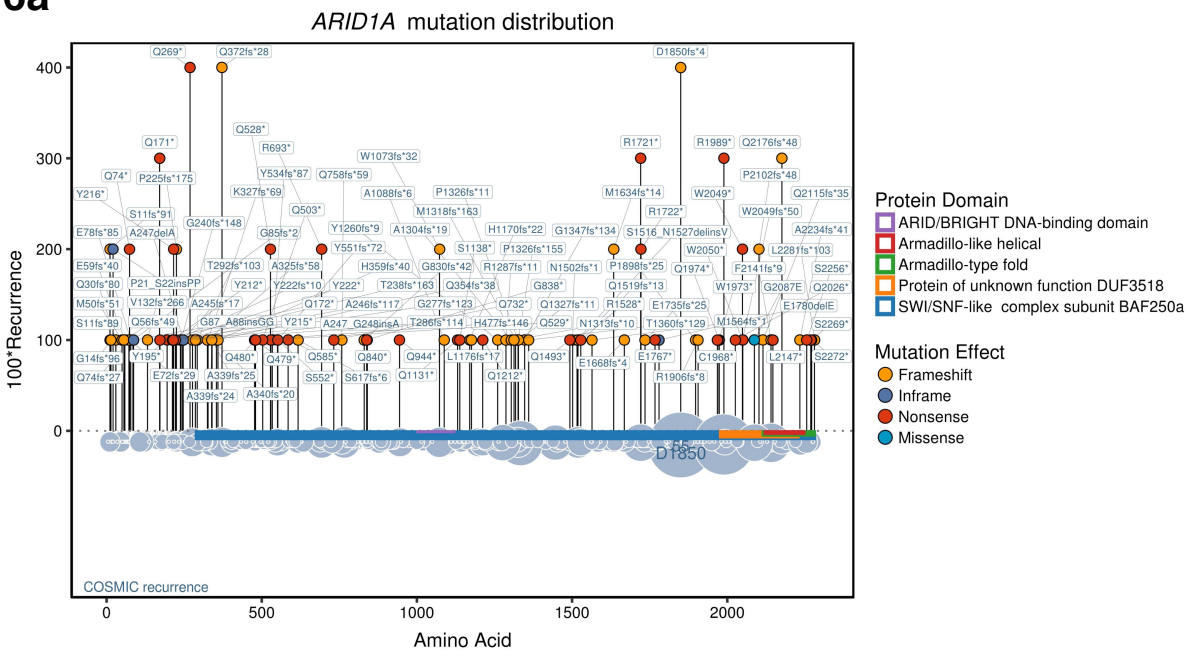
We identified the following tumour suppressor genes within in our cohort: *EP300, ARID1A, KTM2D, MGA, PTEN, PTPN6, PTPRC, PTPRD, RB1, TET2, TNFAIP3, ZFP36L1*. All have been previously characterized as tumour suppressor genes, either in lymphoma or in other cancer types. Therefore, our ability to reproduce the genetic mutation profiles for these tumour suppressor genes provided a partial validation of the effectiveness of our variant calling methodology.

Additionally, a few tumour suppressor genes demonstrated a small number of highly recurrent mutations (Figure 6b). These mutations are likely disrupting critical residues, consistent with tumour suppressor activity. First, *TBLXR1* exhibited an in-frame deletion (S324delS) whose function is unclear. A follow up study determining the function of this specific residue could illuminate *TBLXR1* activity. Second, *SOCS1* exhibited a missense mutation at S116 in its SH2 domain which binds JAKs and inhibits their catalytic activity, a critical function of the SOCS1 protein[151]. Finally, *SMARCA4* exhibited various recurrent

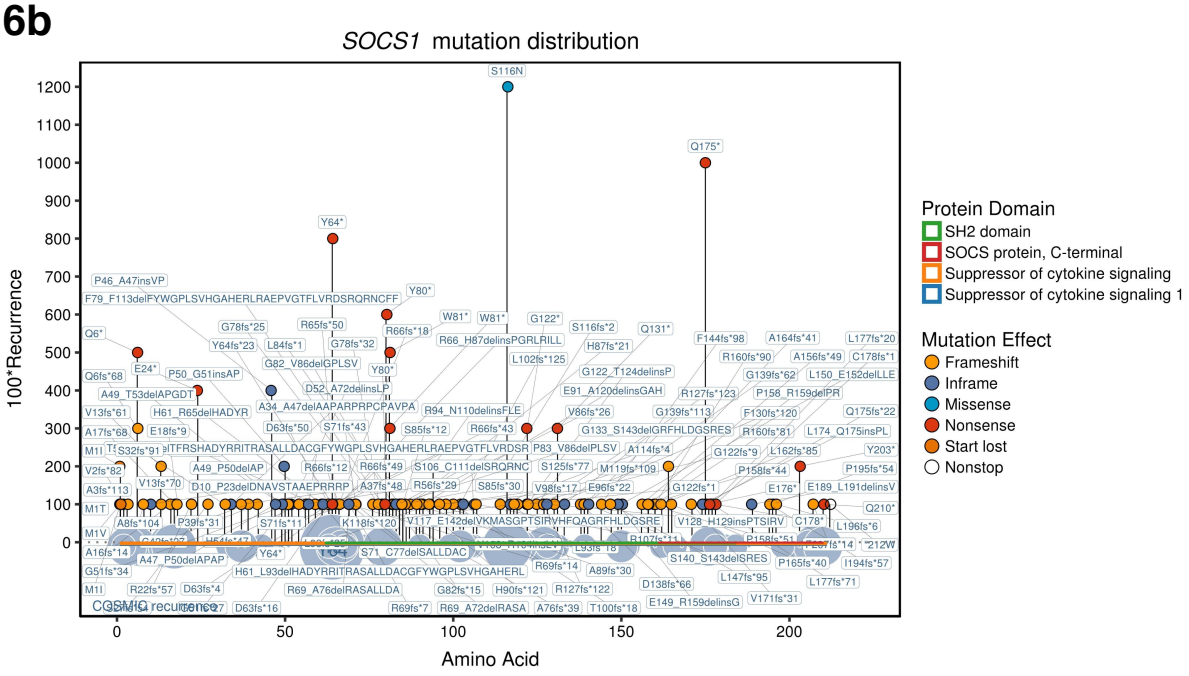
missense mutations in its helicase, superfamily 1/2, ATP-binding domain (T910, P913) and a recurrent missense mutation in its helicase, C-terminal domain (R1192). None had been previously reported in DLCBL although alternate mutations had been reported in small cell carcinoma of the ovary[152]. SMARCA4 is an ATP-dependent transcriptional activator that often acts through the SWI/SNF nucleosome remodelling complex[153]. Therefore, we suspect the T910 and P913 mutations are interfering with phosphorylation/dephosphorylation while the R1192 mutations are interfering with specific binding to the transcriptional targets of SMARCA4.

Finally, two tumour suppressor genes (*TNFRSF14* and *BTG2*) exhibited highly recurrent frameshift, nonsense, and nonstop mutations of interest. In addition to showing a general genomic landscape of frameshift and nonsense mutations spread throughout the coding sequence of the genome, *TNFRSF14* exhibited a highly recurrent nonstop mutation at W12 and a highly recurrent frameshift mutation at T169fs*65 (Figure 6c). Similarly, *BTG2* displayed a highly recurrent nonsense mutation at Q33 (Figure 6d). While these mutations align with the broad theme of disrupting the tumour suppressor activity of *TNFRSF14* and *BTG2*, their high recurrence sets them apart from other similar disrupting mutations. We suspect the high recurrence of these mutations could either point to regions of the coding sequence that are more exposed to mutation generally or these mutations could result from unique mutational processes that disproportionately target them. The exact function of both of these recurrent mutations, however, is unknown.

6a


ARID1A mutation distribution
100*Recurrence
Amino Acid
COSMIC recurrence
D1850
Protein Domain
ARID/BRIGHT DNA-binding domain
Armadillo-like helical
Armadillo-type fold
Protein of unknown function DUF3518
SWI/SNF-like complex subunit BAF250a
Mutation Effect
Frameshift
Inframe
Nonsense
Missense

6b


SOCS1 mutation distribution
100*Recurrence
Amino Acid
COSMIC recurrence
Protein Domain
SH2 domain
SOCS protein, C-terminal
Suppressor of cytokine signaling
Suppressor of cytokine signaling 1
Mutation Effect
Frameshift
Inframe
Missense
Nonsense
Start lost
Nonstop

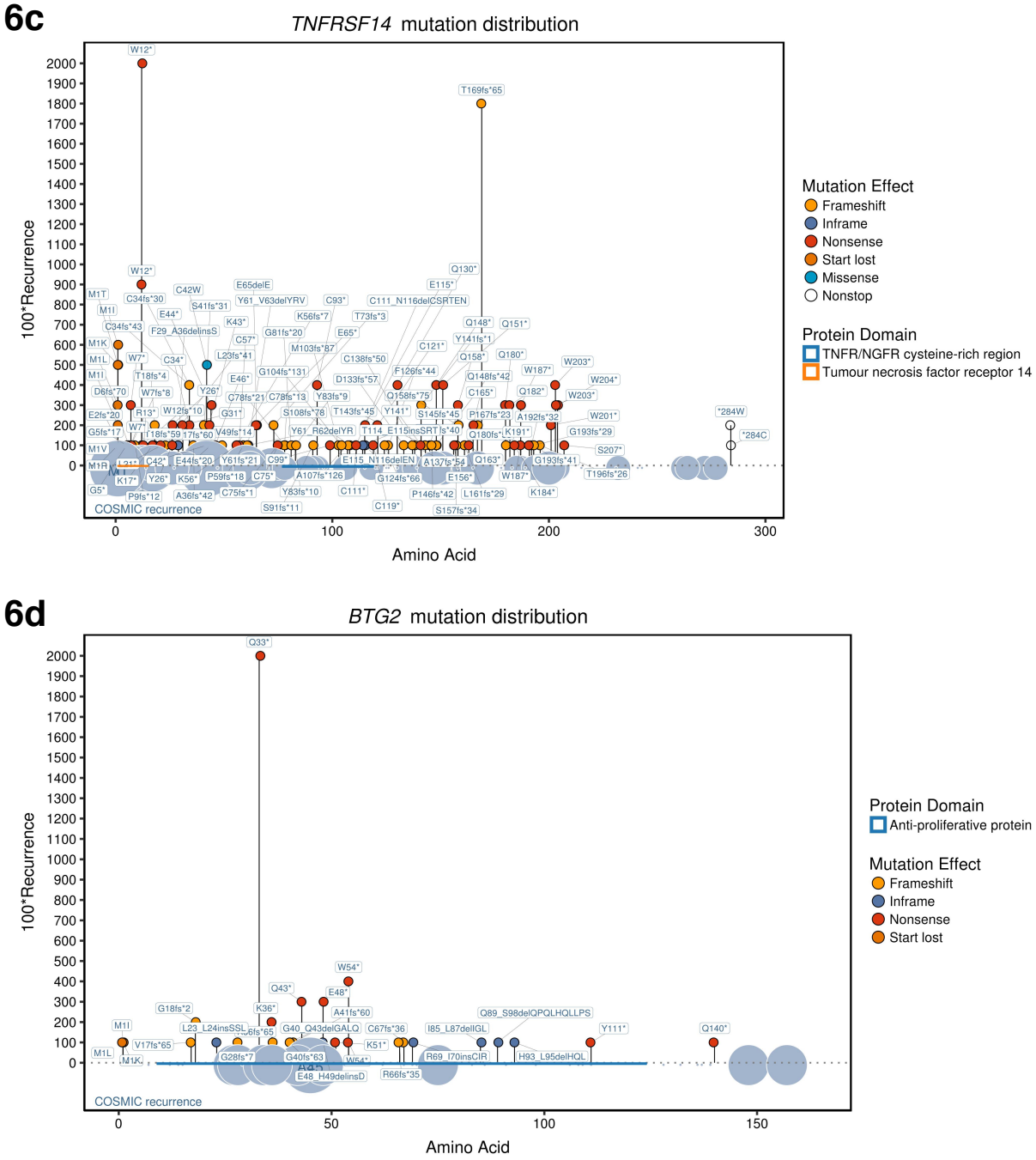


Figure 6 Gene-level analysis demonstrates tumour suppressor gene mutational profiles and reveals recurrent disruptive mutations. Each gene plot shows driver mutations found in the coding sequence, (2) protein domains from UniProtKB, and (3) bubbles. Bottom half of plots show bubbles sized according to the number of mutations found in COSMIC. **(a)** Tumour suppressor genes exhibit disrupting mutations spread throughout the coding sequence of the gene. *ARID1A* is shown as a representative example. **(b)** Highly recurrent missense mutations may disrupt a key residue. *SOCS1* is shown as a representative example. **(c, d)** *TNFRSF14* and *BTG2* exhibited recurrent nonsense, frameshift, and nonstop mutations**.**

4.3.1.2. Well-Characterized Oncogenes

Similarly, we were able to recreate expected genomic profiles for well-characterized oncogenes: strong hotspots of missense mutations that likely cause a gain in function (Figure 7). Unlike disrupting mutations in tumour suppressor genes, gain of function mutations in oncogenes often require more specificity: inactivating a specific self-regulatory domain for example or increasing the affinity of a protein for its target, causing constitutive binding. Therefore, activating mutations in oncogenes generally occur at specific residues, appearing as "hotspots" with significant mutational recurrence within genes. Within our dataset, we successfully recreated major hotspots within DLBCL, FL, and BL.

Broadly, oncogenes within our cohort generated genetic mutation profiles that either (1) matched known hotspots and offered no new hotspots, (2) matched known hotspots and offered new hotspots, or (3) elucidated mutation profiles not previously described. We discuss each sequentially.

The first category of oncogenes exhibited genetic profiles that recreated their known hotspots and did not reveal any new hotspots (Figure 7a): *EZH2* (Y646); *BRAF* (G466, G469, N581, D594, L597, V600, K601); *WHSC1* (E1099, TT1150)[154]; *XPO1* (E571)[155]; *MEF2B* (D83)[156]; *STAT6* (D419)[148]. Broadly, these genes tend to be among the most well characterized and in some cases, the most frequently mutated genes in lymphoma. As a result, it was unlikely that a study with a larger patient sample size and more coverage depth would be likely to uncover new additional hotspots. Regardless, our ability to recreate the genomic profiles for these known genes largely validate our approach.

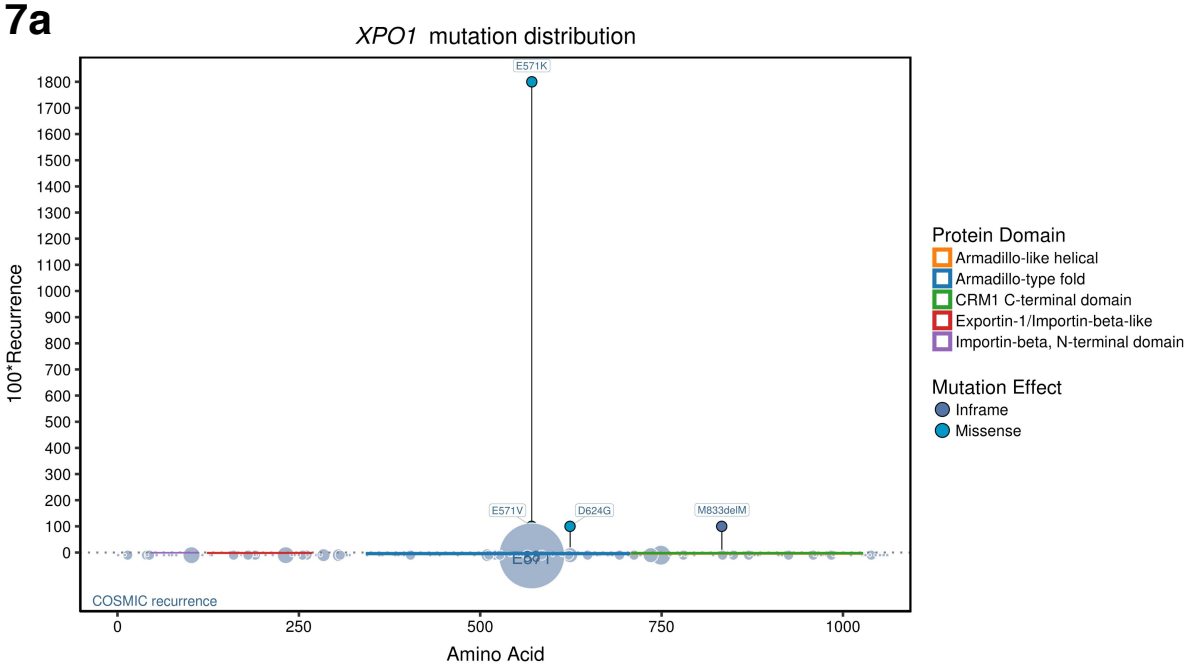
The second category of oncogenes exhibited genetic profiles that, in addition to recreating known hotspots, also revealed new hotspots (Figure 7b). First, the *CARD11* gene recreated known hotspots at D230, D357, D401, and L251[151] while also exhibiting a new mutation at Q249. The *CARD11* mutations shown above all occur within the coiled domain of the protein, the disruption of which is known to cause constitutive NF-KB activation and enhanced NF-KB activity, hallmarks of DLBCL[158]. Second, the *MAP2K1* gene recreated known hotspots at G203, P124, F53, C121[160], while revealing a new recurrent mutation at D67. While the above mutations had been reported for melanoma[159] and pediatric type follicular lymphoma[160], we show their presence here in B-NHL samples, previously unreported. We suspect the D67 mutation functions through the same mechanism: causing constitutive ERK phosphorylation and activity. Third, the *MYD88* gene recreated known hotspots at L265P, S219C, and V217F while also revealing a new recurrent mutation at S251N.[69] All mutations are believed to cause constitutive NF-KB and JAK signalling although the exact mechanism for such dysregulation is unknown. Fourth, *CCND3*,

previously reported as an oncogene, exhibited missense hot spots at I290 and P284 and recurrent frameshift/nonsense mutations at R271 and Q276. While these recurrent mutations had been reported before, the degree of recurrence had not been analysed at scale and these mutations had not yet been considered strong hot spots. All mutations appear to disrupt the Cyclin D domain at the end of the CCND3 protein. Such mutations have been previously reported to increase the stability of the CCND3 protein and lead to CCND3 accumulation within the cell.[20]

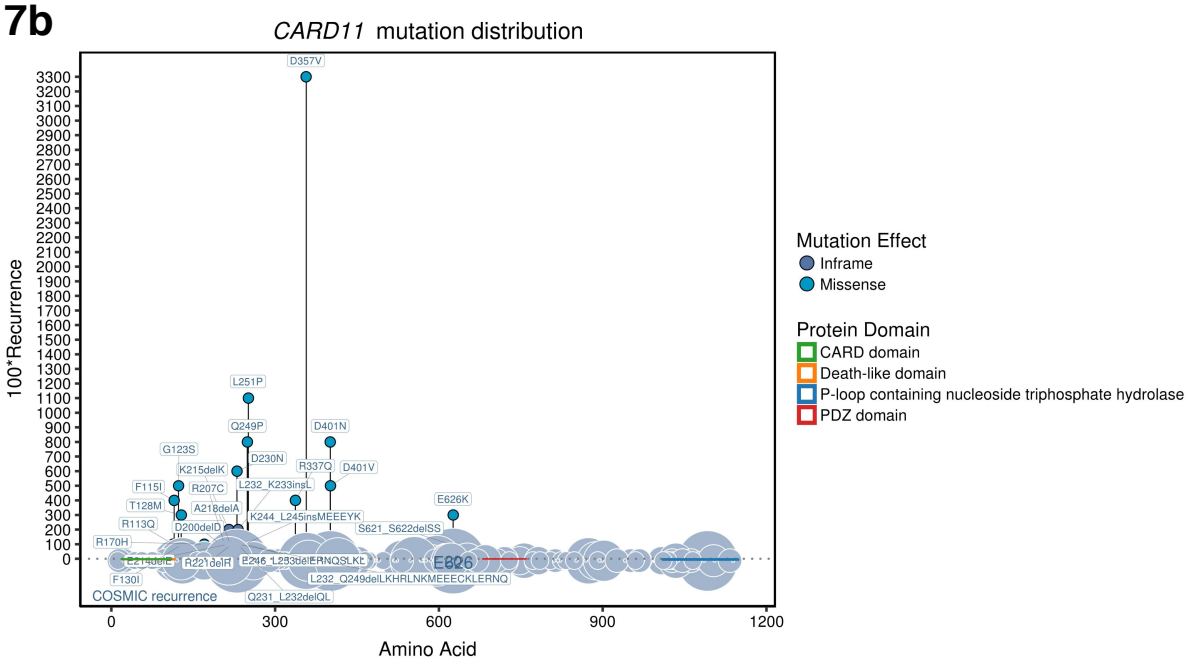
Finally, the third category of oncogenes exhibited genetic profiles that had previously been undescribed. One oncogene, *STAT3*, was present within this category (Figure 7c). STAT3 is a transcription factor, shown to be constitutively activated in many cancers, with a variety of downstream targets which regulate cell proliferation. Crucially, the activation of STAT3 relies on phosphorylation of Y705 which in turn requires docking with tyrosine kinases which is modulated by the SH2 domain[161]. This SH2 domain similarly affects the interaction of STAT3 with its transcriptional targets, thus affecting its ability to effectively regulate their expression. We found two recurrent mutations in *STAT3*: a E616 in-frame deletion and a Y640 missense mutation, both within the SH2 domain. We believe that by modulating the activation of STAT3 and the ability of STAT3 to repress or activate its transcriptional targets, these mutations are generating a cancerous phenotype. As an example, STAT3 has also been shown to activate the expression of matrix metalloproteinase-2 (MMP2), a crucial protein which shows elevated levels in cases of tumour invasion, angiogenesis, and metastasis[162]. The E616 and Y640 mutations therefore could either be keeping STAT3 in a constitutively activated form or within STAT3 proteins that are transiently activated, activating MMP2 transcription more effectively.

Crucially, the above mechanisms are new within the context of B-NHL and DLBCL in particular. Indeed, the only reported mechanism for STAT3-based pathogenesis in ABC-DLBCL involves the dysregulation of STAT3 by BCL6 which directly represses STAT3. In this scenario, dysregulation of the BCL6 pathway leads to elevated STAT3 levels. The reported mechanism here, if biologically validated, would provide an alternative mechanism for STAT3-based pathogenesis.

7a


XPO1 mutation distribution
100*Recurrence
Amino Acid
COSMIC recurrence
E571K
E571V
D624G
M833delM
E691
Protein Domain
Armadillo-like helical
Armadillo-type fold
CRM1 C-terminal domain
Exportin-1/Importin-beta-like
Importin-beta, N-terminal domain
Mutation Effect
Inframe
Missense

7b


CARD11 mutation distribution
100*Recurrence
Amino Acid
COSMIC recurrence
D357V
L251P
Q249P
D401N
G123S
D230N
R337Q
D401V
K215delK
F115I
R207C
L232_K233insL
E626K
T128M
A218delA
K244_L245insMEEEYK
R113Q
D200delD
S621_S622delSS
R170H
E626
F130I
L232_Q249delLKHRLNKMEEECKLERNQ
Q231_L232delQL
Mutation Effect
Inframe
Missense
Protein Domain
CARD domain
Death-like domain
P-loop containing nucleoside triphosphate hydrolase
PDZ domain

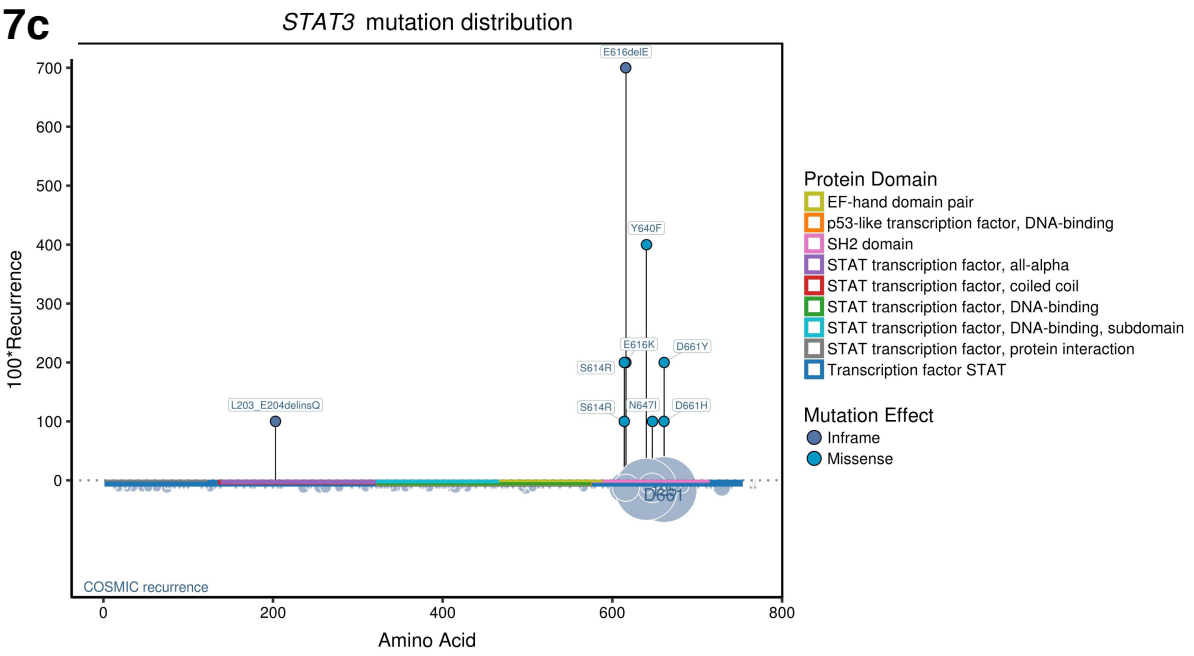


Figure 7 Gene-level analysis demonstrates known and novel oncogene hot spots. (a) Oncogenes exhibit missense hot spots. *XPO1* is shown as a representative example. **(b)** We additionally identified novel hotspots in known oncogenes. *CARD11* is shown as a representative example. **(c)** We created the mutational profile for *STAT3*, a known but uncharacterized oncogene.

4.3.1.3. Oncogene/Tumour Suppressor Genes

While most genes exhibited mutation profiles consistent with oncogenes and tumour suppressor genes, a set of genes (*TP53, CREBBP,* and *FOXO1*) exhibited mutational profiles with characteristics of both: disrupting mutations spread across the coding sequence of the genome with a few missense hotspots (Figure 8). We suspect that these genes are acting as tumour suppressor genes in a subset of the patients shown here but oncogenes in another subset of patients. The ability of these genes to function as both oncogenes and tumour suppressors had been previously described for other malignancies but not for B-NHLs.

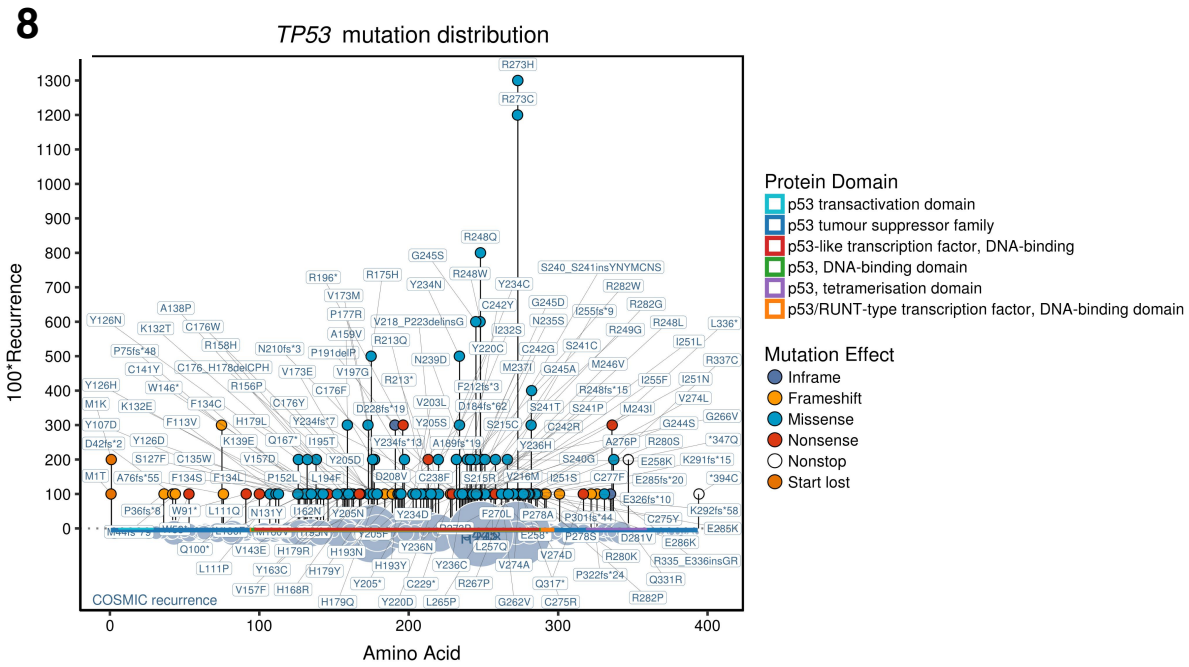


Figure 8 Gene-level analysis shows the potential for genes to serve as both tumour suppressors and oncogenes. *TP53* is shown as a representative example.

4.3.2. Mutational Patterns

4.3.2.1. Targets of Aberrant Somatic Hypermutation

The role of aberrant somatic hypermutation (SHM) is well documented as contributing to DLBCL pathogenesis by either causing gain of function mutations in oncogenes or contributing to genome instability[163]. Crucially, SHM generally targets a 2kb region downstream of the transcriptional start site[163]. Therefore, genes targeted by SHM tend to display a high proportion of mutations near the N-terminal end of the gene's coding sequence. Other criteria also exist to identify SHM within a gene, namely considering the percentage of single nucleotide variants (SNVs) within specific hot spots and the ratio of C:G mutations to A:T mutations[163]. Based on these rules, roughly 44 genes have been identified as SHM targets. While we have not yet applied this full rule set to identify all SHM-targeted genes within our cohort and thus characterize a more extensive set of SHM targets, we did indeed find evidence of SHM causing mutation within our study.

B2M, *RHOA*, and *MYC* all demonstrated a proclivity toward missense mutations near the N-terminal end of the gene's coding sequence (Figure 9). Additionally, these missense

mutations showed great variety in the residue targeted and the resulting change. While the mechanism of SHM in *MYC* is well-defined as resulting from translocation of *MYC* with the *IGH* locus, the mechanism of SHM in *B2M* and *RHOA* may result from either translocation or simply aberrant targeting of non-IGV loci. The specific mechanism is currently unknown.

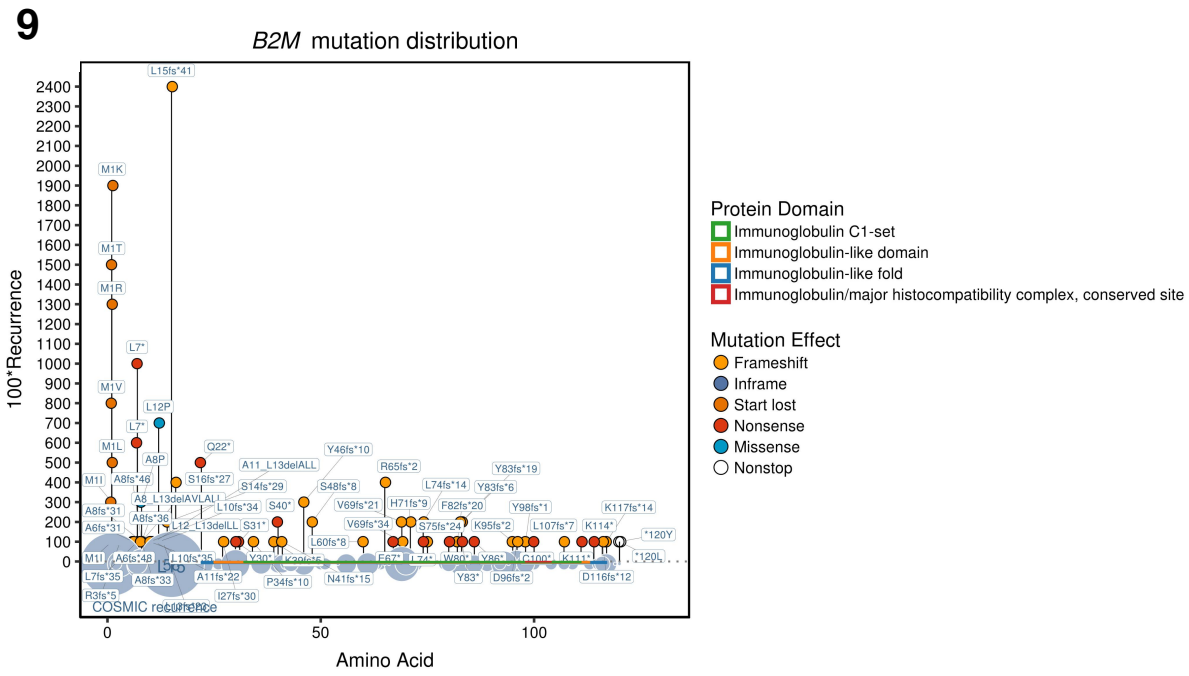


Figure 9 Gene-level analysis shows patterns of aberrant somatic hypermutation. *B2M* is shown as a representative example.

4.3.2.2. Disrupting Mutations Clustered in Specific Domains

Finally, we observed a set of genes with disrupting mutations clustered in specific domains (Figure 10). We suspect such mutations may be working to inactivate specific domains, such as regulatory or binding domains, that thereby cause a gain of function of the gene.

4.3.2.2.1. *BCL10*

BCL10 is a well-characterized oncogene primarily prevalent in SMZL and FL[164,165]. Rather than presenting a standard oncogene genomic profile, however, with a hotspot of missense mutations, *BCL10* instead exhibits a cluster of frameshift and nonsense mutations primarily toward the C-terminal end of the gene (Figure 10a). In previous studies, in-frame

deletions near the C-terminal end of the *BCL10* gene had been previously reported in a small subset of FL and DLBCL patients and postulated to contribute to the function of *BCL10* in lymphomagenesis[165]. Our cohort, however, did not replicate these in-frame deletions. The specific pattern of frameshift and nonsense deletions clusters we present here have not been previously reported.

We suspect these mutations are causing lymphomagenesis by leading to an activation of the NF-KB pathway by dysregulation of the CARD11-MALT1-BCL10 signalling complex. Generally, BCL10 forms a complex with CARD11, and MALT1 in order to activate NF-KB as a result of either an upstream CD40 or BCR stimulus[166]. An upstream stimulus is thought to phosphorylate CARD11, causing a conformational change which allows recruitment of BCL10-MALT1 which are believed to be constitutively associated[166,167]. Subsequently, CARD11 is thought to cause BCL10 to oligomerize into helical filamentous structures, and BCL10 and MALT1 are then ubiquitinated, ultimately allowing the translocation of NF-KB dimers from the cytoplasm to the nucleosome where they induce transcription. The BCL10 mutations reported here near the C-terminal end of the gene could therefore either (1) increase the affinity of BCL10-MALT1 for CARD11, bypassing the CARD11 conformational change usually necessary for association and thus activation of the NF-KB pathway, (2) cause BCL10 to oligomerize in the absence of CARD11, thus encouraging ubiquitination of the BCL10-MALT1 complex and allowing for NF-KB translocation to the nucleus in the absence of a stimulus, or (3) interfere with de-phosphorylation and de-ubiquitination events necessary to reduce the response inherent to the prior pathways.

We also suspect an independent mechanism could be acting. In particular, the C-terminal end of BCL10 is also thought to enable the interaction between BCL10 and MALT1. Disruption of the C-terminal end of BCL10 could therefore lead to a CARD11-BCL10 complex assembling without MALT1. It is additionally known that MALT1 is a caspase which generally cleaves BCL10. Therefore, these mutations could prevent effective cleavage of BCL10. The downstream pathogenetic effects of such a chain are uncertain; BCL10 cleavage by MALT1 has not been shown to activate KF-KB though it has been shown to allow T-cells to adhere to fibronectin[168]. Ultimately, the effect of such a change on the

pathogenesis of FL and SMZL is unclear.

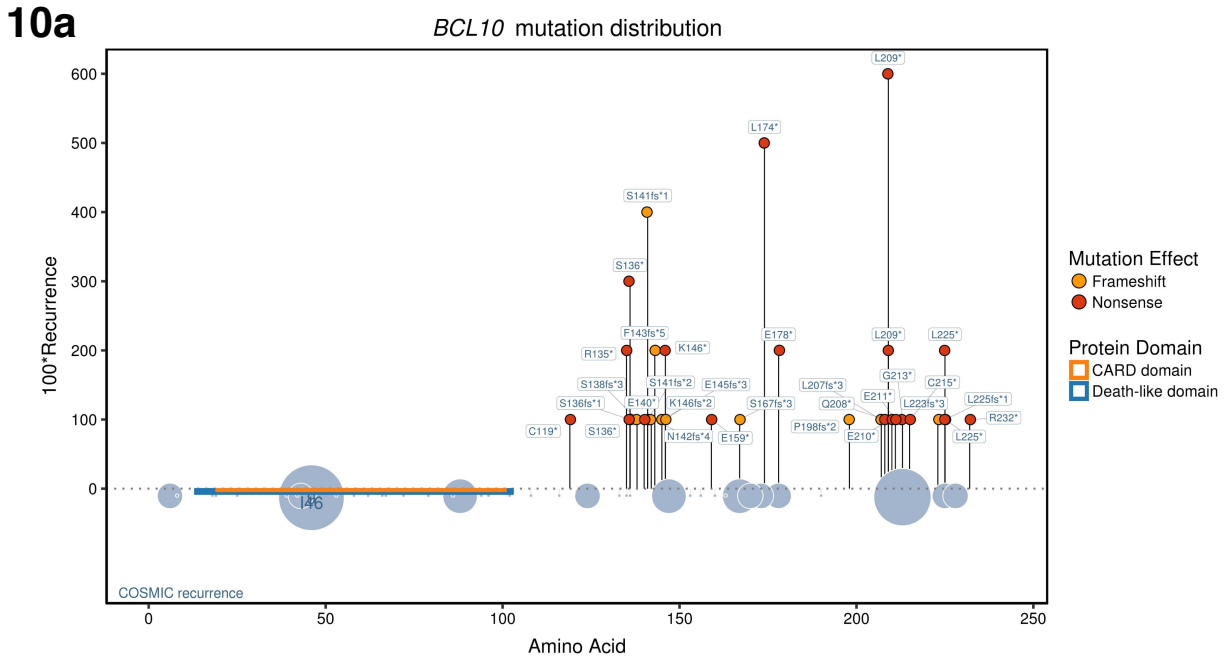


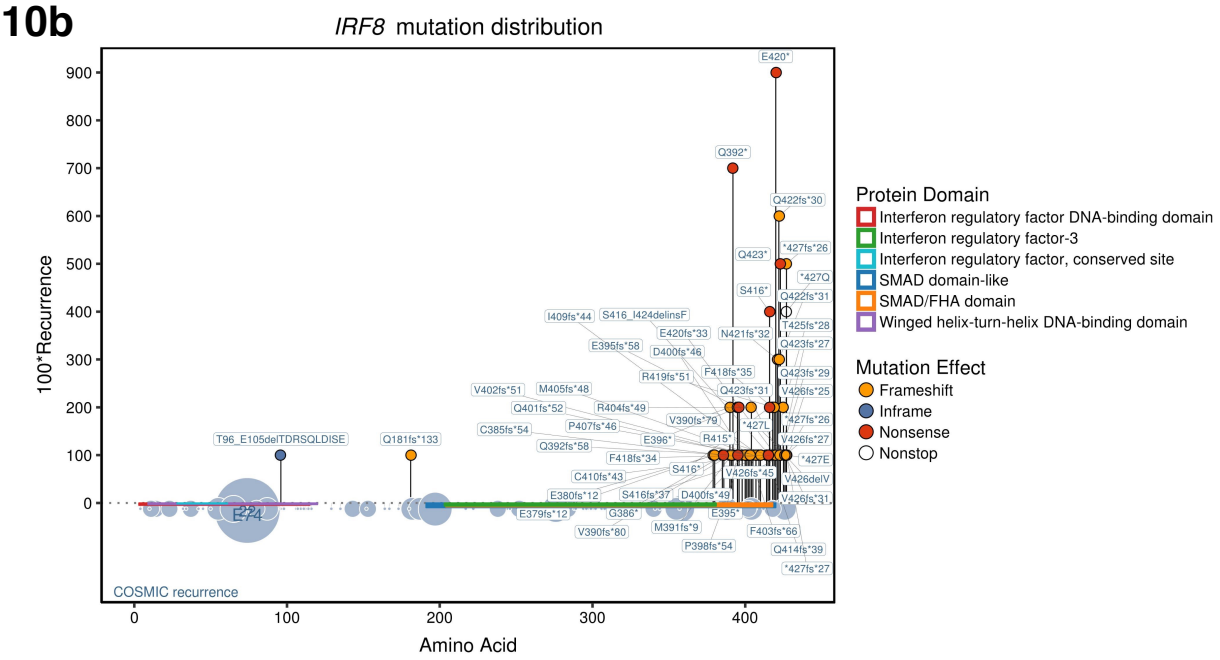
Figure 10 Gene-level analysis reveals disrupting mutations clustered in highly specific domains. (a) *BCL10*, **(b)** *IRF8*, **(c)** *FAS*, **(d)** *ARID1B*, **(e)** *NOTCH1*, **(f)** *NOTCH2*, **(g)** *KLF2*, **(h)** *TCF3*, **(i)** *SMARCB1*.

4.3.2.2.2. *IRF8*

IRF8 exhibits a high number of frameshift and nonsense mutations at the C-terminal end of the gene, primarily in the SMAD/FHA domain (Figure 10b). Previous studies have postulated that overexpression of IRF8 in lymphoma via an IGH-IRF8 gene fusion could lead to oncogenesis through various pathways[169]. However, to our knowledge, we are the first to report specific frameshift and nonsense mutations in the C-terminal end of the *IRF8* gene which potentially confer gain of function. This independent mechanism for oncogenic activity of *IRF8* could provide an alternative target for therapies.

Historically, *IRF8* has been considered a tumour suppressor gene in both DLBCL and FL[170] however more recent studies have considered it an oncogene[169]. Based on our results, the high clustering of disrupting mutations in the SMAD/FHA domain suggests that *IRF8* is an oncogene in which the disruption of the SMAD/FHA domain confers a gain of function. In DLBCL, knockdown of IRF8 has been shown to decrease phosphorylation of p38 and ERK MAP, proteins critical to B lymphocyte proliferation[169]. Therefore, a gain of function in IRF8 via these mutations may instead stimulate B lymphocyte proliferation. Additionally, IRF8 has been shown to regulate MDM2 and TP53 in germinal center B cells, thus

preventing apoptosis[169]. Therefore, gain of function in IRF8 could additionally allow DLBCL and FL to evade apoptosis.



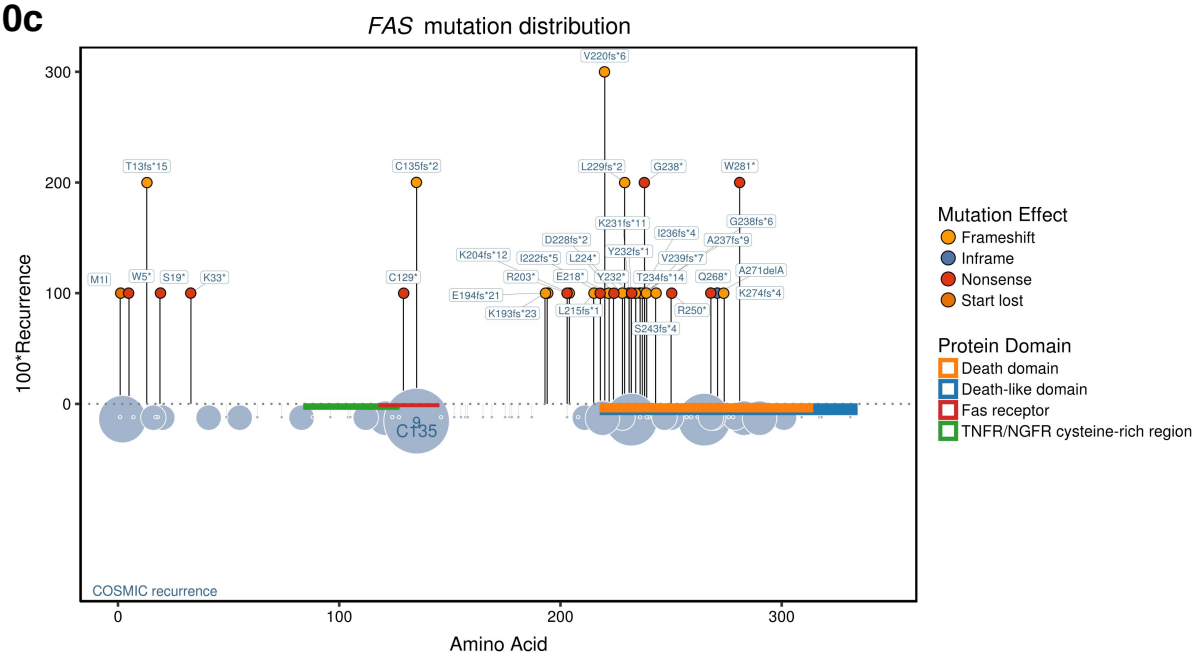
4.3.2.2.3. *FAS*

FAS exhibits a pattern of frameshift and missense mutations once again clustered near the C-terminal end of the gene, in the Death and Death-like domains (Figure 10c). *FAS* has been identified as a tumour suppressor gene in FL, DLBCL, and BL[171]. Biologically, FAS serves as a membrane receptor in the tumour necrosis factor receptor (TNFR) super family. FAS molecules on the cell surface spontaneously preassociate into homotrimers. Upon activation via ligand binding, interaction between the death domains of FAS lead to the recruitment of CASP8, a procaspase which activates the caspase cascade eventually leading to apoptosis[172]. The high proportion of frameshift and missense mutations in the death domain of FAS therefore are likely preventing homotypic interaction between death domains in the FAS homotrimer. Thereby, CD95-based apoptosis of B cells via FAS is being inhibited and cells with these mutations are allowed to proliferate.

Overall, while hot spot mutations in the intracellular signalling domains of FAS have been identified previously[172], frameshift and missense mutations affecting the death domains have not been previously identified. Specifically, mutations involving the SP, CRD1, CRD2, CRD3, and TM domains of FAS have been identified as important to the pathogenesis of T-

cell lymphoblastic lymphoma[172]. However, to our knowledge, the specific disrupting mutations in the death domain for FL, BL, and DLBCL patients in our cohort have not been identified. Moreover, the absence of the SP, CRD1, CRD2, CRD3, and TM mutations identified for T-cell lymphoblastic lymphoma in our cohort suggest that the *FAS* gene could be functioning via distinct oncogenic mechanisms depending on the condition. Overall, our mutational profile suggests an independent and previously unreported mechanism for *FAS* mutations to induce cancerous proliferation in B-NHL.

10c

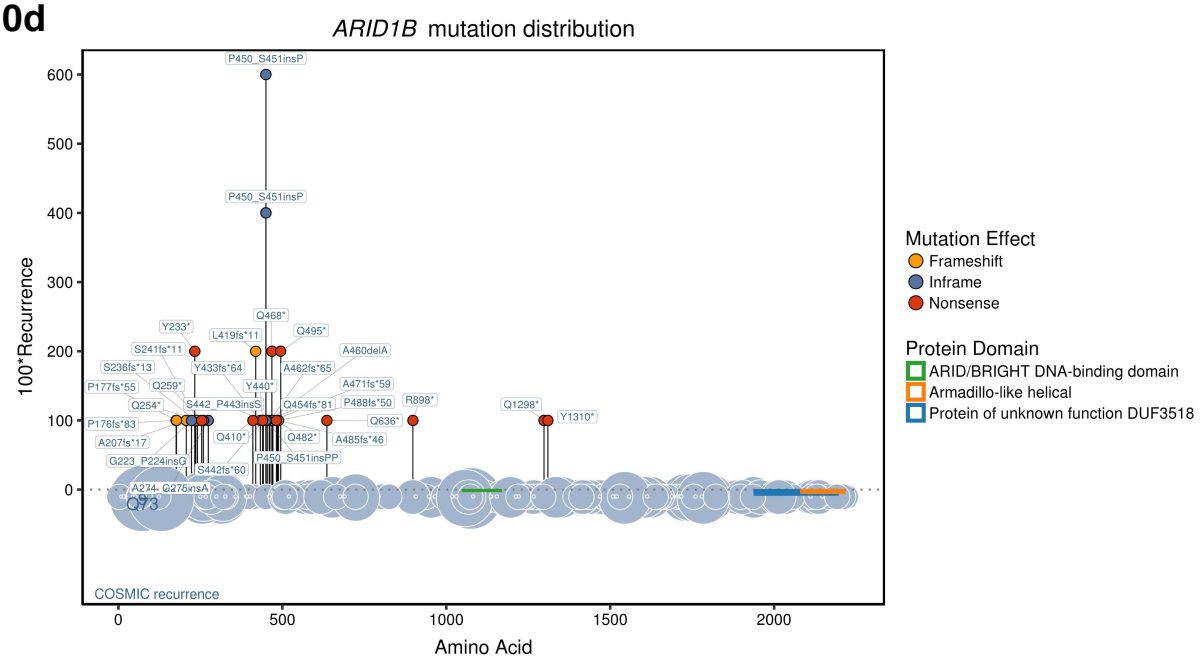


4.3.2.2.4. *ARID1B*

ARID1B is a member of the SWI/SNF chromatin remodelling complex and is involved in cell cycle regulation. Broadly, *ARID1B* mutations in B-NHLs have not been previously characterized though mutations distinct from those mentioned here have been found for other diseases[173–177]. In our study, *ARID1B* exhibited a tight cluster of disrupting mutations (frameshift mutations, nonsense mutations, and proline insertion mutations) between amino acids 176-274 and 410-488 (Figure 10d). The clustering of these mutations near the N-terminal end of the coding sequence implies aberrant somatic hypermutation as a potential mechanism for the introduction of these mutations. The exact functions of these regions are currently unknown for *ARID1B*, however, they are likely breaking the alpha-

helices crucial to ARID1B folding and thus disrupting overall activity.

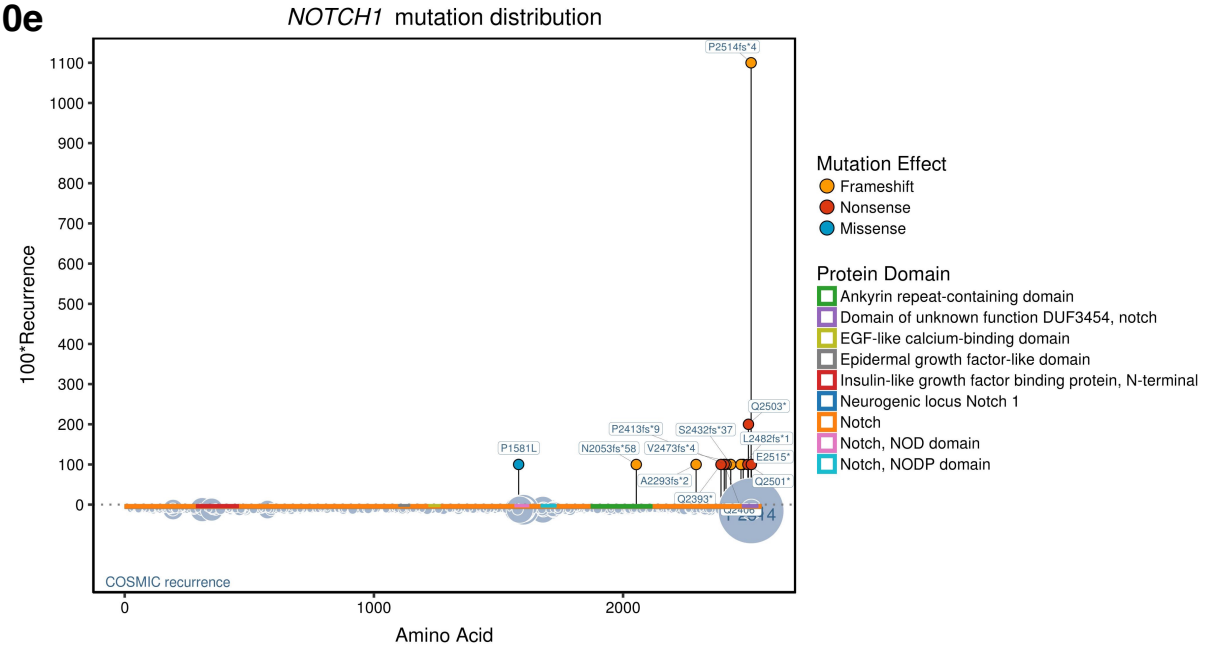
10d



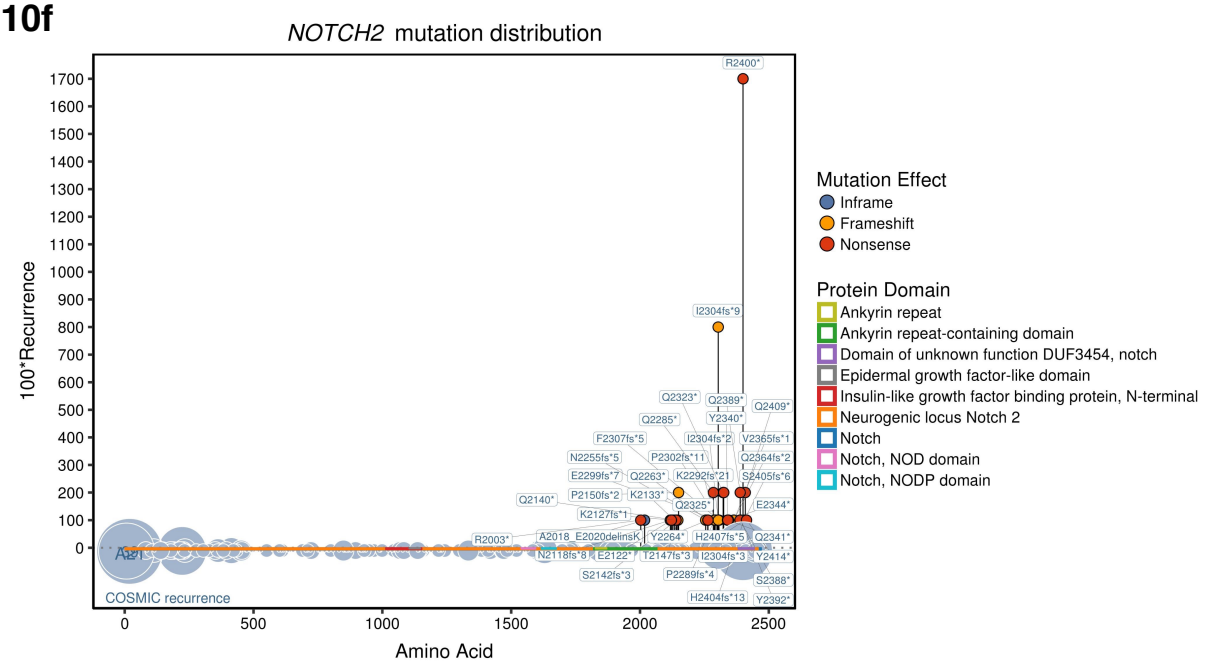
4.3.2.2.5. *NOTCH1/NOTCH2*

NOTCH1 and NOTCH2 are Type I transmembrane proteins that transduce signals across the cellular membrane. Both *NOTCH1* and *NOTCH2* exhibit clusters of frameshift and nonsense mutations at the C-terminal end of their gene in the same domain (DUF3545) (Figure 10e, f). Both mutations imply loss of function in the DUF3545 domain, which is an intracellular domain. While the exact effects of these losses on NOTCH-based signalling are unclear, we suspect they are removing the site of recognition for the E3 ligase FBW7 that targets NOTCH1 for ubiquitin-mediated proteasomal degradation[178]. Indeed in mantle cell lymphoma, disrupting and truncating mutations near the C-terminal end of the *NOTCH* gene have been shown to dysregulate NOTCH signalling through such a mechanism.

10e



10f

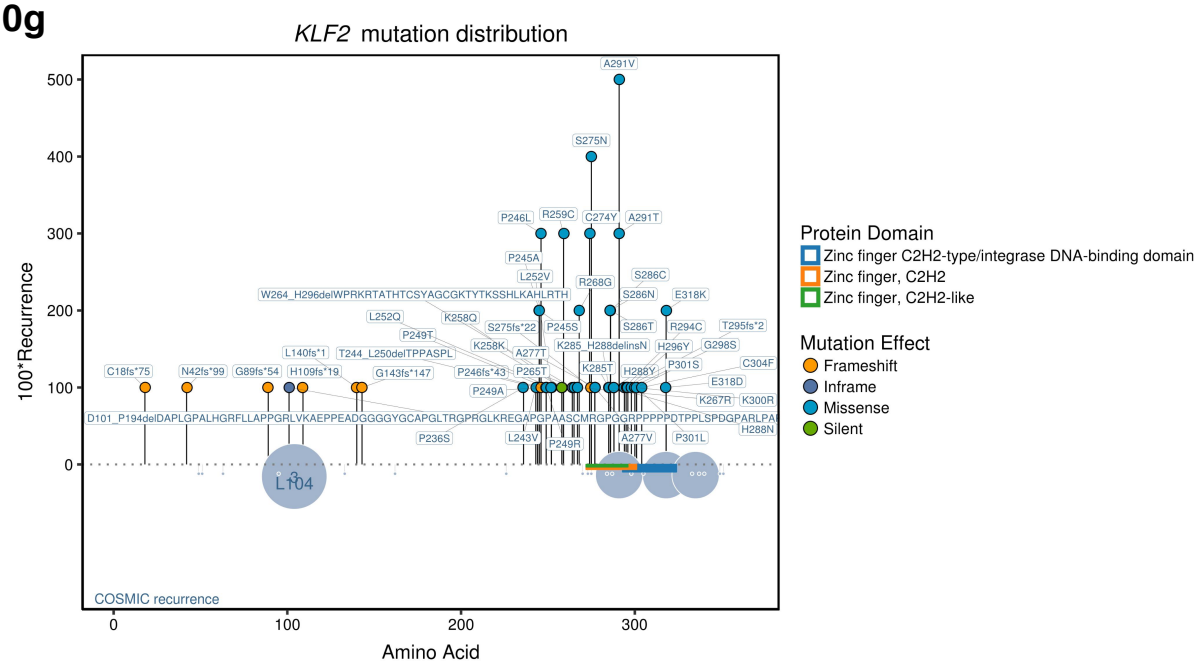


4.3.2.2.6. *KLF2*

KLF2 is a zinc finger protein that plays a transcriptional activation role. Additionally, *KLF2* mutation is the most frequent somatic change in splenic marginal zone lymphoma[179]. *KLF2* exhibited a series of missense mutations near the C-terminal end of its gene in or near its zinc finger domains (Figure 10g). Such mutations are likely inhibiting the ability of KLF2

to accurately recognize its transcriptional targets and are therefore disrupting mutations. Such inactivating mutations likely have a pathogenic role: in SMZL, for example, KLF2 deficiency causes follicular B cells to migrate to the splenic marginal zone[180]. For DLBCL, however, the exact pathogenesis mechanism of KLF2 is unknown.

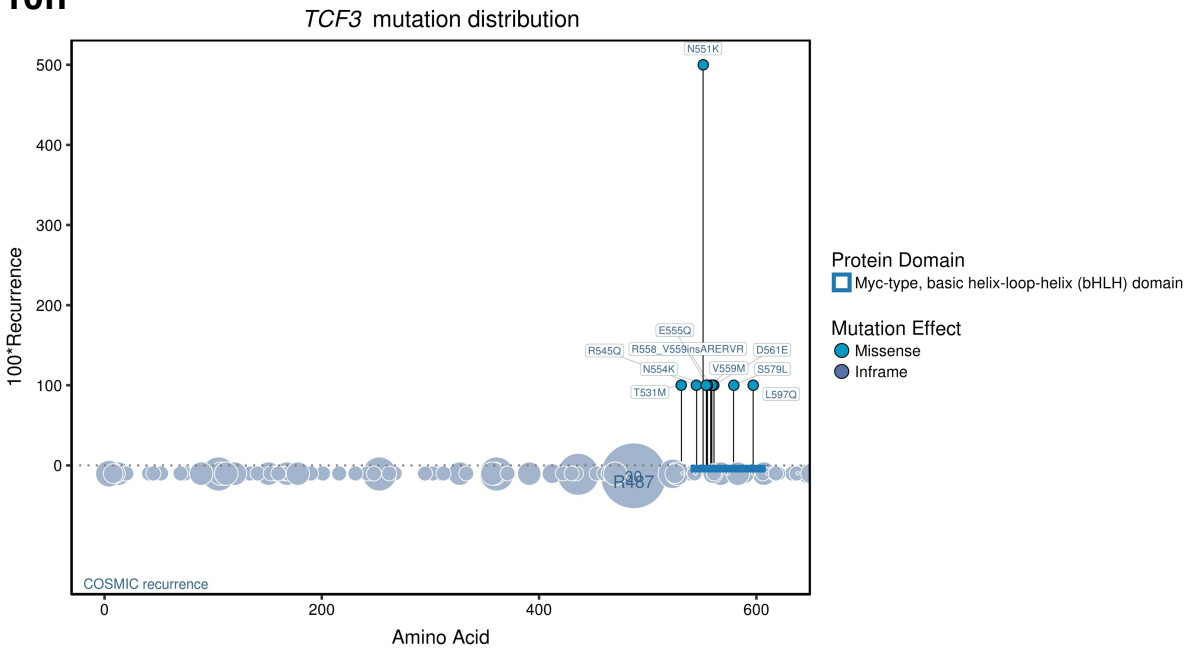
10g



4.3.2.2.7. *TCF3*

TCF3 is a helix-loop-helix transcription factor critical to B cell development whose dysregulation is implicated in BL pathogenesis. In our study, *TCF3* exhibited missense mutations clustered in the Myc-type, basic helix-loop-helix (bHLH) domain, replicating those seen previously in BL samples[20] (Figure 10h). Here, as in the previously reported BL cases, we suspect these mutations are disrupting the bHLH domain and thereby disrupting TCF3 function and tonic B-cell receptor signalling more broadly[20].

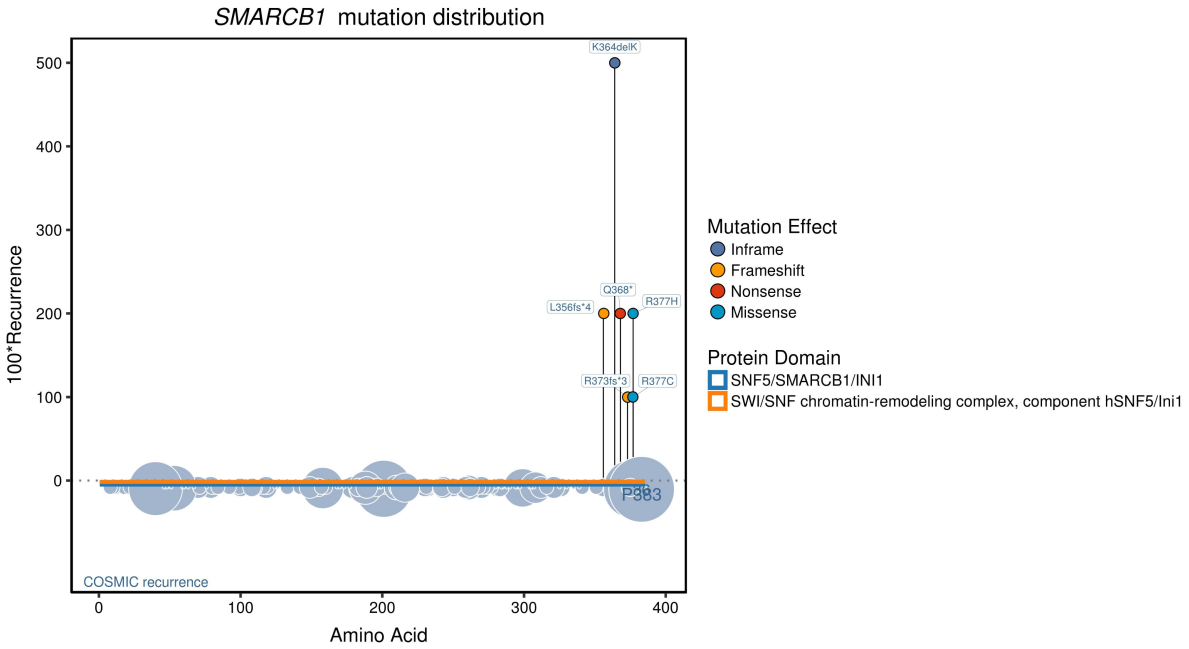
10h



4.3.2.2.8. *SMARCB1*

SMARCB1 is part of the SWI/SNF complex, enabling transcriptional machinery to access its targets. In our B-NHL cohort, we found a cluster of frameshift, nonsense, and missense mutations near the C-terminal end of the *SMARCB1* gene (Figure 10i). *SMARCB1* mutations have been primarily found in multiple meningiomas[181] and epitheloid sarcomas[153], where the gene is present as a tumour suppressor gene. Indeed, knockouts have been shown to generate tumour growth[153]. Unfortunately, it is unclear whether these mutations are ultimately activating or disruptive. However if they are indeed disruptive, then a key question arises surrounding why disrupting mutations are found only in the C-terminal end of the gene but not in earlier parts of the coding sequence.

10i



4.3.2.2.9. *SGK1*

SGK1 carried a very specific set of mutations that affected essential splice sites. Twelve essential splice site mutations were found at Chr6:134495648 and thirty-four essential splice site mutations were found at Chr6:134495725. These two mutations flanked the 5' and 3' end of a single exon within *SGK1* and thus likely cause aberrant splicing of that exon. Previous studies have suggested *SGK1* is a tumour suppressor gene on the basis of the splice site mutations[53], but the high degree of clustering of these at a single exon (not previously evident due to the small numbers of patients), coupled with the absence of nonsense and frameshift mutations, suggests these might be gain-of-function mutations.