5. Classification Analysis

With all drivers identified, we then proceeded to classify our dataset by identifying patterns of co-mutation within the set of drivers. We classified samples of all diagnostic subtypes together with the aims of (1) ensuring we could successfully differentiate known diagnostic subtypes and (2) utilizing the known classifications to generate a granular and accurate classification for DLBCL samples. In particular, we strived to produce a genetic classification that could add granularity and accuracy to the classifications already built by the WHO and the gene expression based, cell of origin classification for DLBCL.

We chose to classify all samples at once as opposed to dividing them by subtype and then classifying them as such an approach would increase our ability to differentiate between DLBCL subtypes. Crucially, DLBCL can either arise *de novo* or as the transformation of various indolent lymphomas. Therefore, the genetic patterns present within a given DLBCL cohort are a mixture of the patterns which underlie DLBCL *de novo* and the patterns which underlie various indolent lymphoma. By including both DLBCL samples and samples of other lymphomas in the same classification, the Bayesian Dirichlet processes were able to robustly extract the genomic patterns of FL and BL more effectively based on those samples and then apply those patterns to differentiate among samples marked as DLBCL samples. Had DLBCL samples been including in isolation, it would have been substantially more difficult to differentiate the genomic patterns of DLBCL samples that had transformed from other types.

Compared to prior classification studies, our project primarily derives its power from its scope. First, 1607 B-NHL lymphoma patients were analysed. By comparison, only one prior DLBCL study had 1,001 DLBCL samples whereas other prior B-NHL studies were about 10X smaller[15]. Similarly, the depth of our targeted coverage (~500x) substantially exceeded that of prior studies, enabling the identification of rarer variants. Combined, such scope and power enable the use of powerful classification technologies that would otherwise be ineffective.

Two important features distinguish a genetic classification of DLBCL NOS and cancer more broadly. First, while the treatments and clinical course of DLBCL and B-NHL patients will change over time as new therapies are introduced, we suspect that the underlying genomic patterns that contribute to the pathogenesis of these diseases will remain the same. Thereby, a genetic classification is likely to be stable and lasting, simply gaining refinement as more driver variants and genetic datasets are added. Second because genomic changes

have been well characterized as the cause of various cancer types, classifying cancers on a genetic basis reveals the co-mutation patterns that fundamentally cause pathogenesis. Thereby, genetic classifications grant unique insight into the mechanistic onset and progression of disease which can then ideally be utilized to design new treatments. Overall, therefore, we believe that a genetic based classification for DLBCL NOS, and for other cancers more generally, is both causal and stable.

As with the genomic landscape section, this classification section will similarly be substantially improved over the next few months via the addition of copy number and translocation data. Given the well-characterized importance of copy number alterations and translocations in various types of B-NHL lymphoma, we suspect the classification may change substantially. While the underlying driver mutations will not change, we suspect class defining lesions may be present in the copy number alteration and translocation data that will substantially change the grouping. For example, the *MYC* translocation is a well-known hallmark lesion for BL that will likely become class defining once added to our dataset. Similarly, *BCL2* and *MYC* double hit patients are known to have a substantially more aggressive clinical course[182] and we suspect these patients may also form their own cluster. In the absence of this data, however, initial conclusions about mutation patterns underlying DLBCL and B-NHLs can be drawn.

5.1. Bayesian Dirichlet Processes

In order to classify the dataset, we used Bayesian Dirichlet Processes, a nonparametric and hierarchical clustering approach[142]. Bayesian Dirichlet Processes work in a fashion similar to Mixture Models. Mixture Models operate by creating a fixed set *n* of multivariate distributions, seeing how well these distributions explain the data at present, modifying the distributions to explain the data more effectively, and repeating until convergence is met. Bayesian Dirichlet Processes function similarly except the number *n* of multivariate distributions is not fixed. In other words, in Bayesian Dirichlet Processes the algorithm must learn both the optimal shape and parameters of each distribution as well as the optimal number *n* of distributions that can describe the dataset overall. Bayesian Dirichlet Processes accomplish this task by cycling each data point and either assigning the data point to (1) an existing cluster or (2) a newly created cluster. The probability of being assigned to an existing cluster scales with the number of data points already assigned to that cluster. Thereby, the algorithm prevents overfitting: if too many clusters are created that have too few points, then in subsequent iterations, the data points in small clusters are likely to be

reassigned to larger clusters, thus eliminating the smaller clusters and reducing the number of overall clusters.

By utilizing this nonparametric clustering approach, we can remove bias inherent to the classification methodology. Had we instead use a parametric approach, such as the mixture models mentioned above, we would have had to define the number of clusters which would have artificially biased the classification. By instead leaving the optimal number of clusters to be learned, we can produce a classification more representative of the underlying dataset.

5.2. Classification on All Subtypes

Overall, our classification yielded 8 distinct classes within our cohort of B-NHLs. (Figure 11). All eight classes within our classification are well defined and meaningfully distinct from each other. The genes which denote each class are strongly co-mutated with each other but mutually exclusive with mutations in driver genes that define other classes. Statistically, this appears as strong patterns of correlation between genes in a given genomic class and anti-correlation between genes in different genomic classes. The strength and distinctness of these co-mutation patterns give us confidence in the accuracy of our classification, even in the absence of incorporating translocation data and copy number analysis.

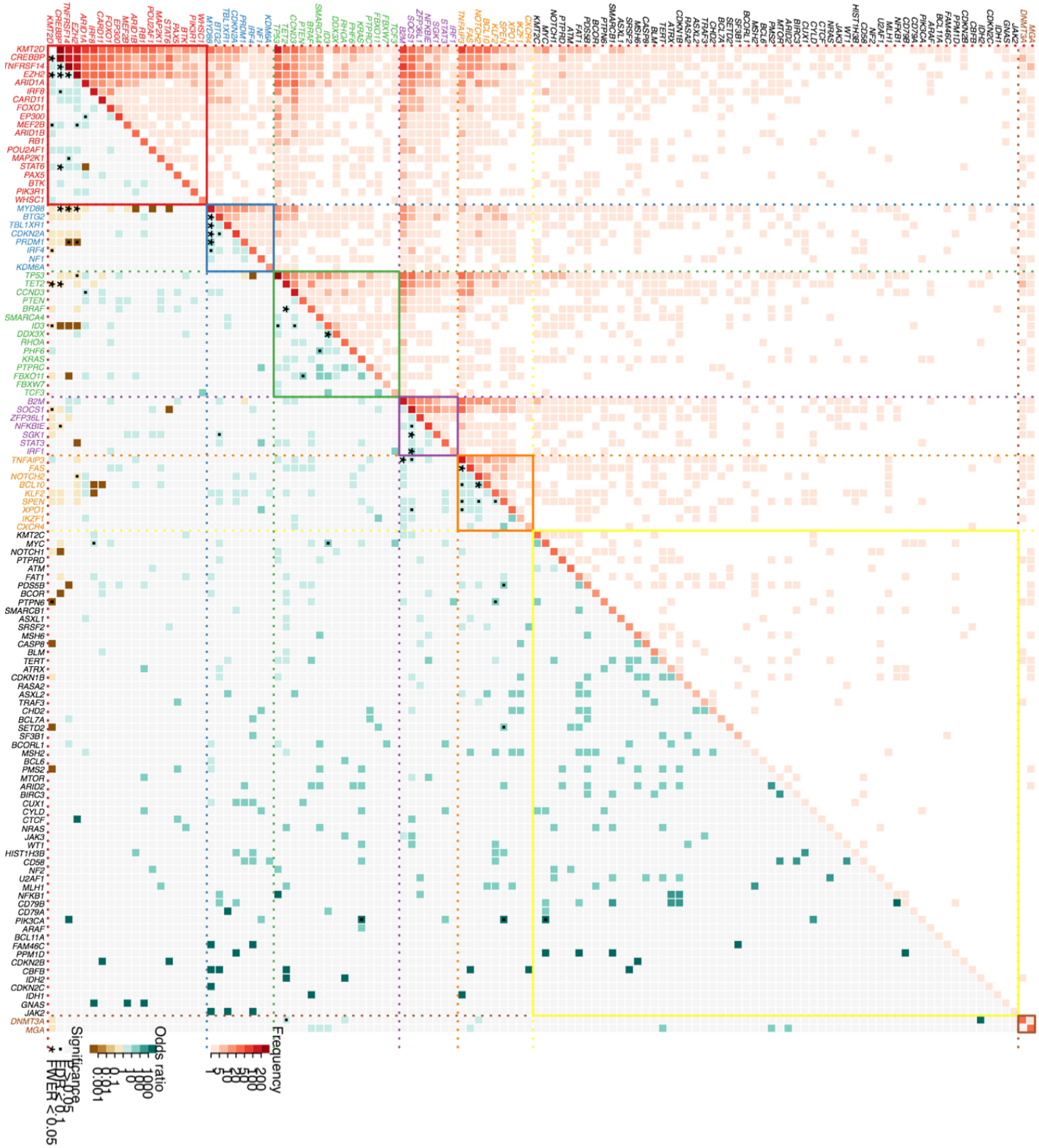


Figure 11 Co-mutation and mutual exclusivity patterns generate eight distinct classes in FL, BL, and DLBCL. Lower triangle depicts pairwise association between lesions in genetic classes. The colour of each tile corresponds to the odds ratio for each pair, with brown representing mutual exclusivity and blue indicating co-mutation. Odds ratios are computed by observed co-mutation rates compared to expected co-mutation based on each lesion's gene frequency. Coloured tiles represent significant relationships ($p < 0.05$), asterisks show significant family wise error rates (FWER < 0.05), boxes show false discovery rates < 0.1 (FDR < 0.1). Upper triangle depicts absolute occurrences of co-mutation for each pair, coloured on a gradient.

5.2.1. Class 0 (*TET2*, *TP53*)

Class 0 (*TET2*; *TP53*) is an "error" class designated by the Bayesian Dirichlet Classification algorithm for outliers (Figure 12b). This class contained 8% of patients, emphasizing the heterogeneity of B-NHLs and DLBCL and the challenge that heterogeneity poses to effective classification methods.

5.2.2. Class 1 (*KMT2D*, *CREBBP*, *TNFRSF14*, *EZH2*, *ARID1A*)

Class 1 (*KMT2D, CREBBP, TNFRSF14, EZH2*, and *ARID1A*) showed a mutational pattern consistent with FL, reinforcing the distinctness of the FL genomic landscape and the capacity for our Bayesian Dirichlet clustering to extract distinct genomic patterns (Figure 12c). Most FL lymphoma patients clustered into Class 1 (Figure 13a), and indeed upon examination, the predominant lesions defining Class 1 are hallmark lesions of FL. The role of *KMT2D*, *CREBBP*, *EZH2*, and *EP300* in chromatin remodelling and the pathogenesis of FL have been well-described and are present in significant proportions of the Class 1 patient population. Some hallmark mutations of FL were indeed missing, namely the t(14;18) translocation leading to ectopic expression of *BCL2*[19]. However, this lesion was missing simply because translocation data was not incorporated within the classification analysis rather than due to a flaw in analysis or a discrepancy within the dataset.

Not all Class 1 patients were diagnosed as FL patients, however. Indeed, substantial proportions of BL patients and DLBCL patients were also assigned to Class 1 (Figure 13a). First, we suspect that the DLBCL patients assigned to Class 1 are likely DLBCL whose lymphoma initiated as a FL and subsequently transformed to the more aggressive DLBCL. Similarly, we suspect that the BL patients within Class 1 may similarly have transformed from FL. Although FL generally transforms into DLBCL, cases of transformation into BL have also been reported[183]. Such an explanation is supported by the class composition of BL. Indeed the majority of BL samples in our study classified into Class 3 (*TP53*; *CCND3*) which, as described below, contained the hallmark mutations of BL and could thus represent *de novo* BL. The second major proportion of BL samples classified into Class 1, which may have resulted from FL transformation. Future work incorporating *MYC* translocation data will likely resolve this question.

For both Class 1 DLBCL patients and Class 1 BL patients, the benefits of a genetic classification approach are clear: even though these patients have histological characteristics consistent with DLBCL and BL, the underlying genetics driving their pathogenesis are similar to FL. As a result, these patients may respond differently to current and novel

treatments compared to other DLBCL and BL patients. We hope to investigate these treatment responses moving forward in the hope of generating novel clinical insights.

5.2.3. Class 2 (*MYD88*, *BTG2*, *TBL1XR1*, *CDKN2A*, *PRDM1*, *IRF4*, *NF1*, and *KDM6A*)

Class 2 (*MYD88; BTG2; TBL1XR1; CDKN2A; PRDM1; IRF4; NF1;* and *KDM6A*) showed a genomic profile broadly consistent with ABC-DLBCL (Figure 12d). *MYD88* (constitutive NF-KB/BCR activity), *CDKN2A* (cell cycle checkpoint), and *PRDM1* (terminal differentiation block) are mutations with well-known pathogenetic functions specific to ABC-DLBCL. The clustering of these mutations within Class 2 thereby make it likely to contain the majority of ABC-DLBCL cases. Importantly, such a clustering was accomplished with mutation data alone. Thereby, both epigenetic and genetic causes could differentiate ABC-DLBCL and GCB-DLBCL classes within the cell of origin classification, which up until now has predominantly relied on epigenetics to distinguish cell types via gene expression patterns.

The remaining genes mutated within Class 2, though numerous, were mutated in substantially smaller proportions than the aforementioned genes. Driver mutations in these genes could yield additional heterogeneity within the ABC-DLBCL category, although the broad causative drivers remain equivalent.

Some mutations which define the ABC-DLBCL category were found within other classes. Namely, *TNFAIP3* (Class 5), *CD79A and CD79B* (Class 6), and *CARD11* (Class 1). However these genes, though important to ABC-DLBCL pathogenesis may similarly be important to the pathogenesis of other classes. Therefore, although prevalent, they may not be class-defining in the same way as *MYD88*, *CDKN2A*, and *PRDM1*. Indeed, these mutations provide the unique elements of ABC-DLBCL pathogenesis as distinct from the pathogenesis of other subtypes.

Consistent with the explanation of Class 2 as ABC-DLBCL, the majority of Class 2 patients were DLBCL patients (Figure 13a).

5.2.4. Class 3 (*TP53, CCND3, ID3, TCF3*)

Class 3 (*TP53, CCND3, ID3, TCF3, PTEN*) displayed a genomic profile largely consistent with BL (Figure 12e). The *ID3, TCF3,* and *PTEN* mutations in BL are well characterized hallmarks which prevent effective regulation of PI3K, thus leading to cell proliferation[19]. The presence of these mutations in Class 3, therefore, indicate a genomic landscape consistent with BL. Note, the most important hallmark mutation of BL, the *MYC*

translocation, was missing simply because translocation data was not present within our dataset. However, it is also worth noting that the most two prevalently mutated driver genes of Class 3 (*TP53* and *CCND3*) have, in the literature, been indicated in lymphomas beyond just BL (FL and DLBCL). *TP53* is prevalent among various classes (3, 4, 7) and is thus discussed below. *CCND3*, however, is predominantly expressed only in Class 3. In contrast with literature which denotes the importance of *CCND3* across FL, BL, and DLBCL – and similarly in contrast with Figure 13a which points to *CCND3* mutations being distributed across all three histologies, our classification shows the unique contribution of *CCND3* to this classification. Class 3 also includes a range of other genes mutated at substantially lower rates; these genes could add additional heterogeneity.

Consistent with the explanation of Class 3 as characteristic of BL, the majority of BL patients were classified into Class 3. The second largest proportion of patients were classified into Class 1 (Figure 13a); we suspect these patients initially manifested FL which then transformed into BL. While their histology would be consistent with BL, their genomic landscape would be more similar to FL, thus classifying them into Class 2.

5.2.5. Class 4 (*B2M, SOCS1, ZFP36L1, NFKBIE, SGK1, STAT3, IRF1*)

Class 4 (*B2M, SOCS1, ZFP36L1, NFKBIE, SGK1, STAT3*, and *IRF1*) denotes a class of mutations not previously described (Figure 12f). Indeed, each gene has been independently implicated in a variety of lymphoma diseases, however no patterns arise that are consistent with any of the subtypes mentioned previously. Interestingly, some of the most prevalent mutations within Class 4 are also prevalent in other classes (*TP53, TNFAIP3*) whereas others are prevalent primarily within Class 4 (*B2M, SOCS1, NFKBIE*, and *KLF2*). *TP53* and *TNFAIP3* could thus be mutations fundamental to the initiation and progression of various lymphomas while the *B2M, SOCS1, NFKBIE*, and *KLF2* mutations could be the mutations driving the unique pathogenesis of Class 4. Overall, Class 4 is a relatively rare class, accounting for only 6% of the patients, primarily those who did not receive a WHO histological classification (Figure 13a). Nonetheless, it's strong patterns of co-mutation of genes within Class 4 and mutual exclusivity between genes of Class 4 and genes of other classes mark it as a separate category.

5.2.6. Class 5 (*TNFAIP3, FAS, NOTCH2, BCL10, KLF2, SPEN, XPO1, 1KZF1, CXCR4*)

Class 5 (*TNFAIP3, FAS, NOTCH2, BCL10, KLF2, SPEN, XPO1, 1KZF1, CXCR4*) shows a genomic profile consistent with Splenic Marginal Zone Lymphoma (SMZL) (Figure

12g). In particular, three hallmark mutations of SMZL (*NOTCH2, BCL10, SPEN*) were all present in Class 5, marking it as a SMZL class[184]. Conversely, three common SMZL mutations were either in different classes or not present within our analysis. *NOTCH1* was present primarily in Class 6, *NFKBIE* was present primarily in Class 4, and *KLF2* was present primarily in Class 2. All three of these lesions, though prevalent in other classes, were not the defining or most prevalent genetic lesions of those classes. Moreover, the total number of samples attributed to Class 5 (n = 102) was relatively small. Combined, therefore, we believe the *NOTCH1, KLF2, and NFKBIE* mutations are still important to the pathogenesis of SMZL and a higher sample size of SMZL patients may have shifted those mutations into Class 5.

The majority of Class 5 patients were considered either DLBCL or BCL Int. patients on the basis of histology (Figure 13a). Therefore, we suspect that these patients likely originated with undiagnosed SMZL that had transformed into DLBCL by the time of histological diagnosis. Crucially, SMZL has both a distinct clinical course and distinct treatment options than DLBCL. A substantial proportion of SMZL patients display few symptoms and are thus handled as "watch and wait cases" at a higher proportion than the more aggressive DLBCL counterpart[184]. Similarly, SMZL offers a wider variety of treatment options (splenectomy, &c.) than DLBCL[184]. We suspect, therefore, that Class 5 patients may respond to different types of novel therapeutic compared to other DLBCL subtypes.

5.2.7. Class 6 (58 distinguishing genes)

Class 6 contains 58 distinguishing genes, all mutated in a relatively low proportion of the patients (Figure 12h). Additionally, Class 6 had the weakest co-mutation and mutual exclusivity patterns among all classes in our classification analysis. Finally, the 58 genes that compose Class 6 are among the rarest genes mutated in lymphomas. Overall, the weak patterns of co-mutation and large size of Class 6 indicate that it is likely composed of multiple classes that could not be resolved by our study. However, resolution of these classes would likely require a substantially higher sample size due to the rare nature of mutations within these genes and also the rare assignments of patients to this class.

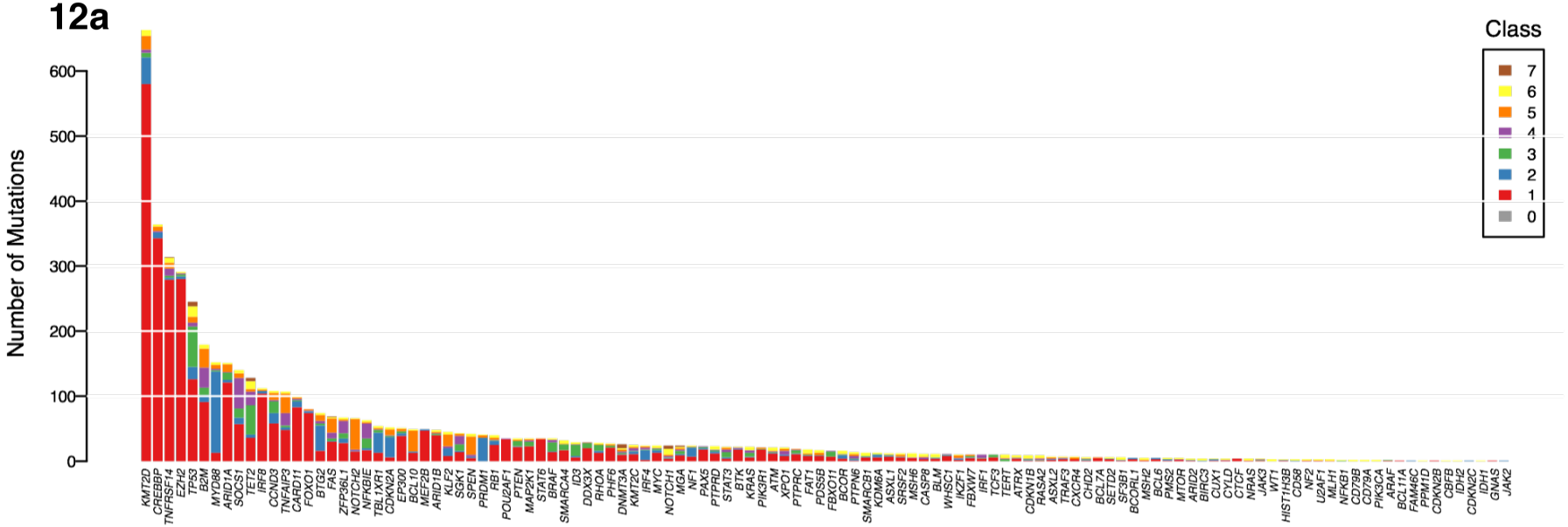
Class 6 samples came from BL, DLBCL, and FL lymphoma subtypes. We suspect these samples, in practice, reflect a variety of rare mechanisms that can cause the pathogenesis of each disease. Importantly, the distinct genome profiles of Class 6 DLBCL and Class 6 BL patients compared to DLCBL patients in other classes and Class 3 BL

patients suggest that Class 6 patients could have their lymphoma arise *de novo* as opposed to resulting from the transformation of an indolent lymphoma.

5.2.8. Class 7 (*DNMT3A, MGA*)

Class 7 (*DNMT3A, MGA*) exhibits a genomic profile not previously described (Figure 12i). Drivers in the *DNMT3A* gene have been implicated in AML, AITL, and T-ALL. Drivers in the *MGA* gene have been implicated in CLL. No immediate pattern emerges tying these two genes together, however, the high comutation between these genes and mutual exclusivity with mutations in other genes renders them an important. Overall, however, this class is extremely rare (1% of patients).

12a


Class
7
6
5
4
3
2
1
0
Number of Mutations
0
100
200
300
400
500
600
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDSS2
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2

12b

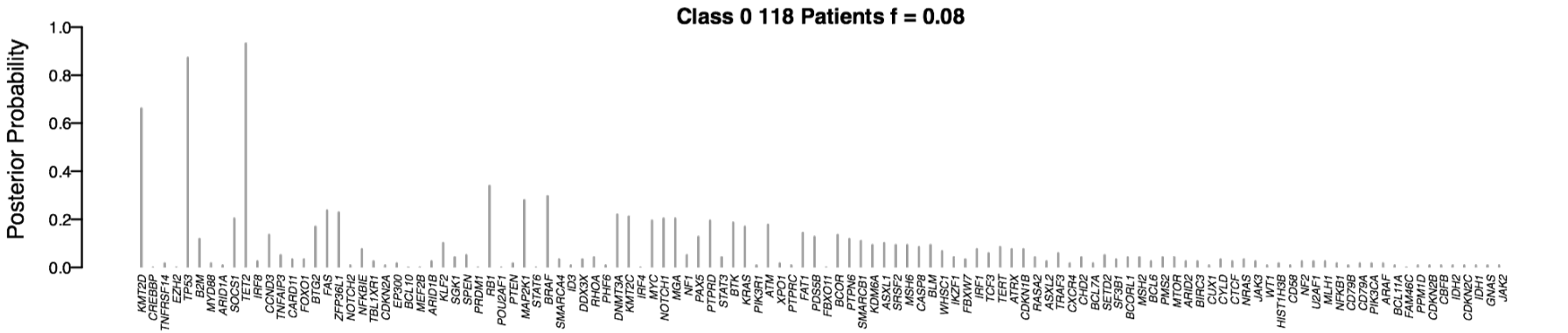
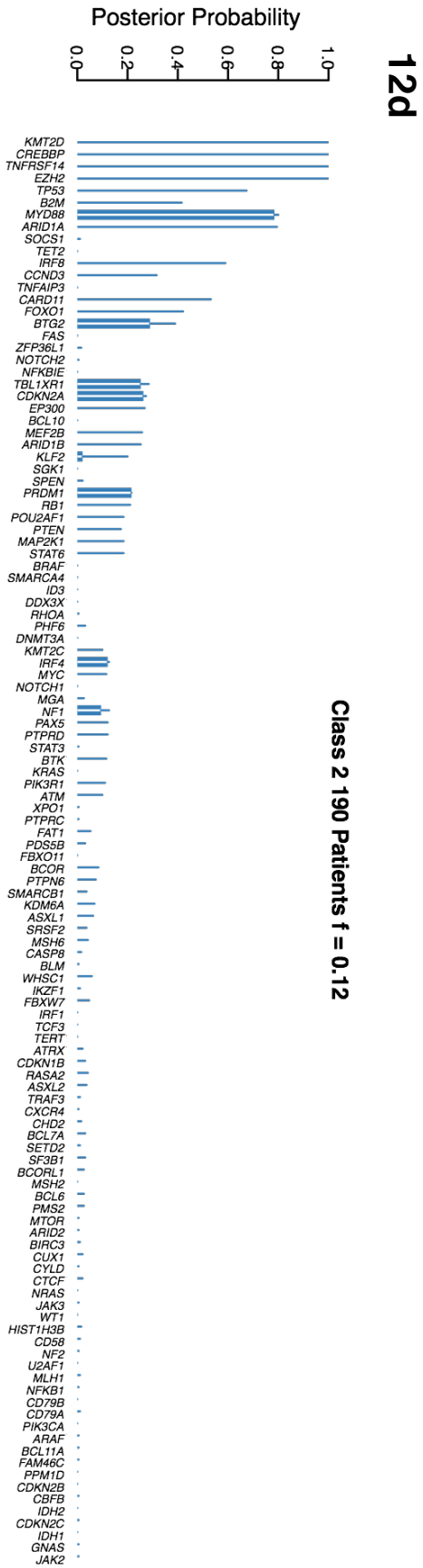
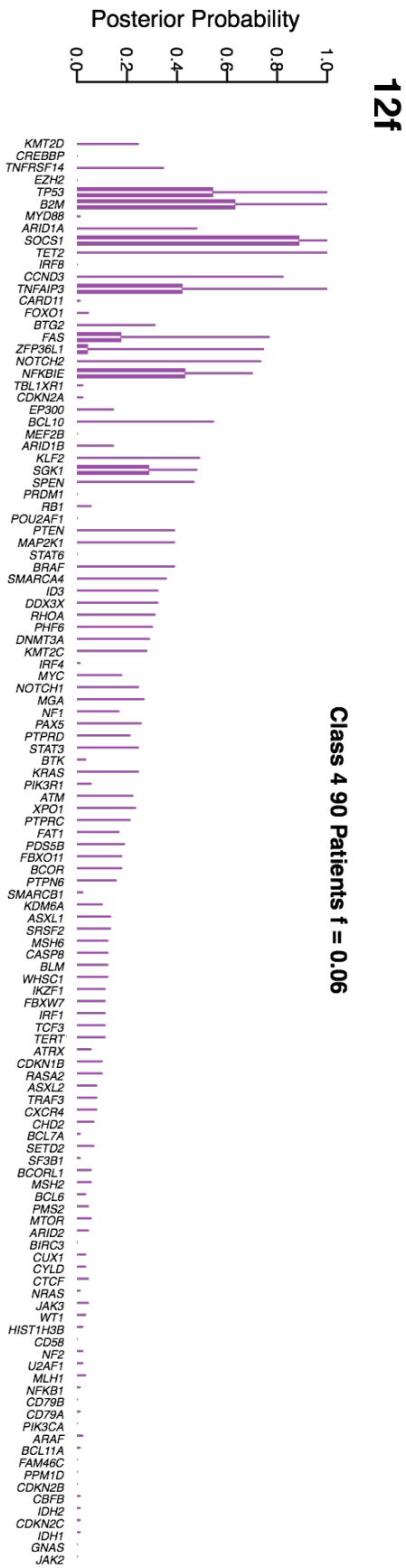
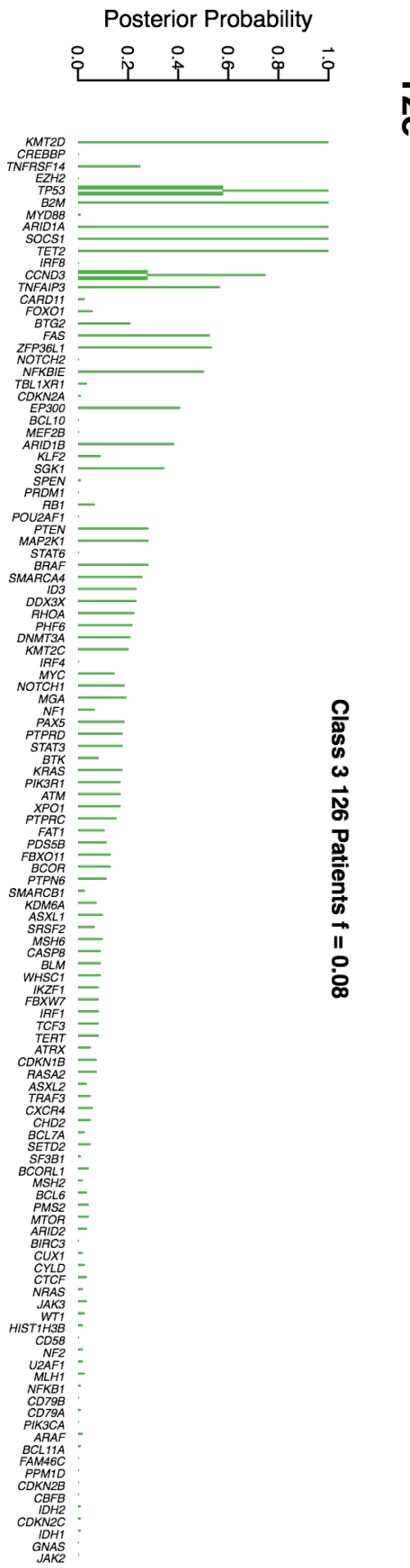
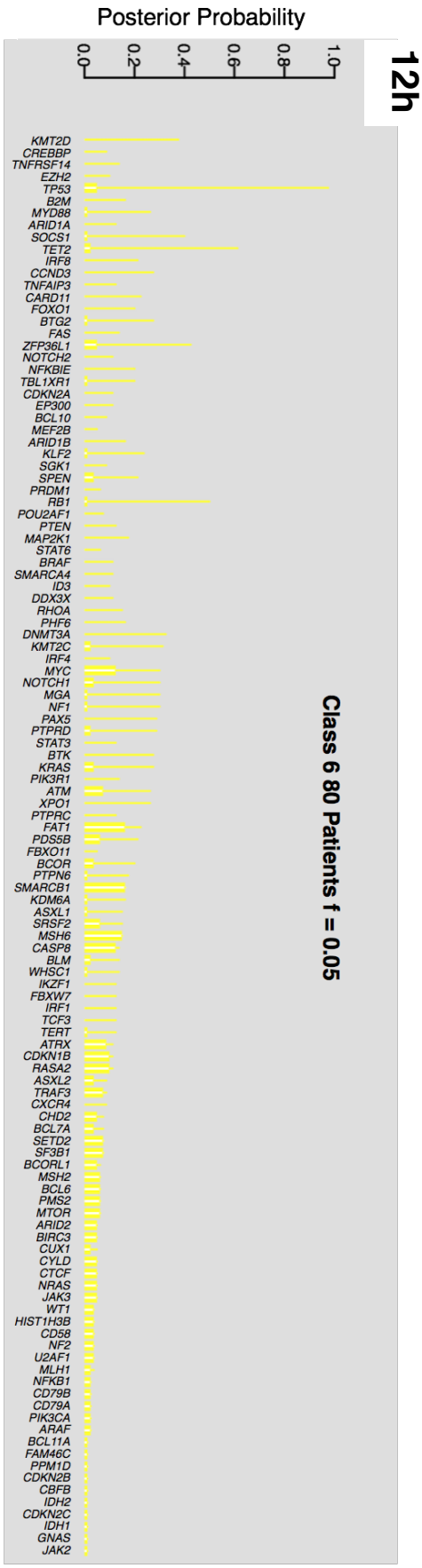

Class 0 118 Patients f = 0.08
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDSS2
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2

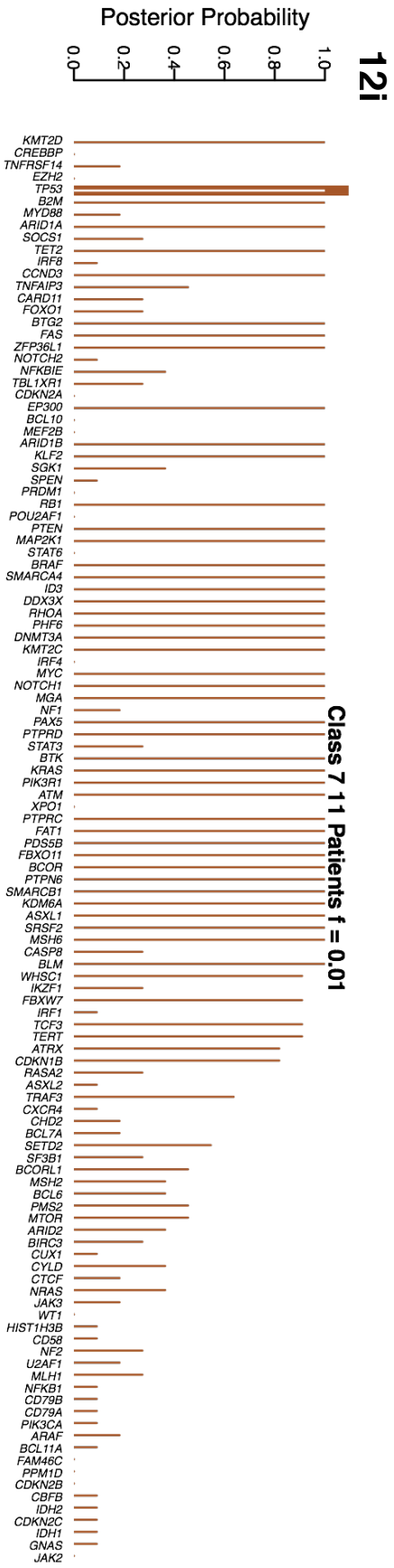
Figure 12 Each class shows a distinct mutational signature profile. (a) Number of driver mutations across all classes, coloured by proposed class assignment for patient with that mutation. **(b-i)** Mutational signature of each class. Numbers next to class show number and fraction of patients assigned to that class. Each bar shows the median posterior probability of a given lesion with error bars corresponding to the 2.5 and 97.5 quantiles.


12c
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDS5B
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2
Class 1 890 Patients f = 0.58
12d
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDS5B
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2
Class 2 190 Patients f = 0.12


12f
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDS5B
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2
Class 4 90 Patients f = 0.06


12e
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDS5B
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2
Class 3 126 Patients f = 0.08


12g
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDS5B
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2
Class 5 102 Patients f = 0.07
12h
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDS5B
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2
Class 6 80 Patients f = 0.05


12i
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
ID3
DDX3X
RHOA
PHF6
DNMT3A
KMT2C
IRF4
MYC
NOTCH1
MGA
NF1
PAX5
PTPRD
STAT3
BTK
KRAS
PIK3R1
ATM
XPO1
PTPRC
FAT1
PDS5B
FBXO11
BCOR
PTPN6
SMARCB1
KDM6A
ASXL1
SRSF2
MSH6
CASP8
BLM
WHSC1
IKZF1
FBXW7
IRF1
TCF3
TERT
ATRX
CDKN1B
RASA2
ASXL2
TRAF3
CXCR4
CHD2
BCL7A
SETD2
SF3B1
BCORL1
MSH2
BCL6
PMS2
MTOR
ARID2
BIRC3
CUX1
CYLD
CTCF
NRAS
JAK3
WT1
HIST1H3B
CD58
NF2
U2AF1
MLH1
NFKB1
CD79B
CD79A
PIK3CA
ARAF
BCL11A
FAM46C
PPM1D
CDKN2B
CBFB
IDH2
CDKN2C
IDH1
GNAS
JAK2
Class 7 11 Patients f = 0.01

5.3. Classification of Histological Subtypes

Concurrent with the co-mutation based classification analysis, we analysed what proportion of samples from each histological subtype were assigned to each class (Figure 13a). While FL was primarily assigned to Class 1, BL was assigned primarily to Class 1 and Class 3. Interpretations for both of these are discussed in the Class 1 and Class 3 sections above. DLBCL had patients split across all seven classes. Crucially, this result highlights the heterogeneity inherent to DLBCL demonstrating that even within the established WHO histological classification, substantially more granularity can be resolved which represents unique and distinct pathogenesis mechanisms. Similarly, this analysis sheds light on the mechanisms that likely cause DLBCL pathogenesis de novo rather than as a result of transformation from an indolent lymphoma. While DLBCL patients assigned to Classes 1, 3, and 5 may have DLBCL that transformed from FL, BL, and SMZL respectively, DLBCL patients assigned to classes 2, 4, 6, and 7 may have either *de novo* DLBCL or DLBCL transforming from indolent lymphomas whose genomic landscapes have either not been adequately characterized or were not identified within this study.

5.4. Comparison with Gene Expression, Cell of Origin Classification

While we lack the gene expression data to definitively assign patient samples according to the cell of origin classification and then compare those assignments with our classification, we can nonetheless draw conclusions about the genomic characteristics of suspected ABC-DLBCL and GCB-DLBCL patients.

First, note that Class 2 shared genetic characteristics largely consistent with those expected from ABC-DLBCL. Upon incorporation of gene expression data, therefore, we will hopefully be able to – on the basis of genetic mutation alone – identify the cell of origin of these lymphomas.

Second, the genetic lesions that characterize GCB-DLBCL were spread across multiple classes, suggesting that GCB-DLBCL can likely be broken into further subcategories with distinct pathogenesis mechanisms. Lesions common to GCB-DLBCL were found in Class 1 (*TNFRSF14, EZH2*), Class 3 (*PTEN*), Class 4 (*SGK1*), and Class 6 (*GNAS*). While the mutations in Class 1 and 3 (*TNFRSF14, EZH2*, and *PTEN*) are common across a range of lymphomas, the mutations in Class 4 and Class 6 (*SGK1* and *GNAS*) are found with less prevalence. We suspect therefore, that GCB-DLCBL patients may have been split across Classes 4 and 6 which would then form subclasses of the GCB-DLBCL category.

Ultimately, however, gene expression and translocation data will need to be incorporated to generate a definite cell of origin classification that can then be superimposed on this classification to understand the patterns inherent to ABC-DLBCL and GCB-DLBCL. Such an analysis would yield valuable insights into the precise pathogenesis of GCB-DLBCL which is, at present, not well-understood.

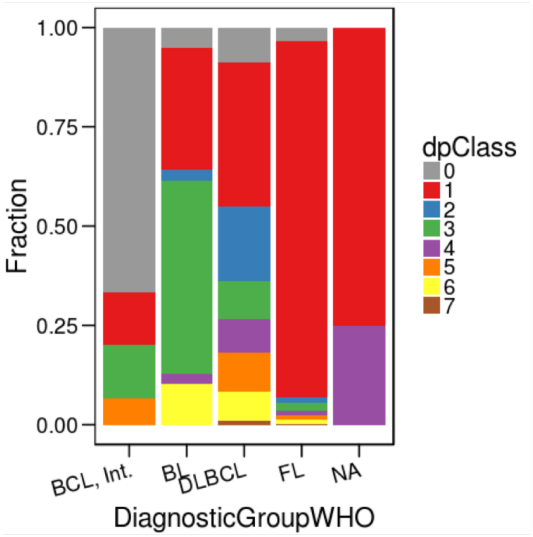
5.5. Preliminary Survival Analysis

After classifying patients according to their genetic profiles, we also conducted a preliminary survival analysis (Figure 13b). Due to time constraints, this analysis is incomplete and has not accounted for confounding factors. In particular, the contributions of age, treatment, date of diagnosis, and centre of treatment to overall survival have not been accounted for. Individually, each of these factors could skew the survival curves of any class. For example, if Class 1 had a disproportionately younger set of patients compared to the other classes, we would expect an improved survival outlook. A full survival analysis accounting for the above factors will be completed after submission of this publication. Nonetheless, preliminary results are presented here.

Overall, the survival analysis generated survival outlooks consistent with our prior interpretations of the genetic classes. As expected, Class 1 which is primarily composed of FL showed the most favourable survival outlook. FL is generally an indolent disease and has the least aggressive clinical course[19] of the subtypes represented; therefore, the result was consistent with expectation. Conversely, Class 2 suffered the worst overall survival outlook. As discussed above, we suspect Class 2 is primarily composed of ABC-DLBCL samples which are known to have a more aggressive clinical course than GCB-DLBCL samples[19]. Therefore, this result was also consistent with expectation. Finally, BL showed a survival outlook intermediate between DLBCL and FL, again consistent with expectation.

Upon completion of a more robust survival analysis, accounting for the confounding factors above, additional insights will be drawn about the categories specified above. In particular if any class shows a particularly aggressive clinical course that is previously unknown or a lack of response to R-CHOP, patients within this class could potentially be put on an experimental clinical trial with more aggressive treatments. Similarly, discovery of such a class would then allow us to identify the specific pathogenesis mechanisms unique to that class which made it more aggressive than other classes. Thereby, meaningful biological insight into the progression of lymphoma would result. Additionally, novel targets for potential drugs could be discovered.

13a



13b

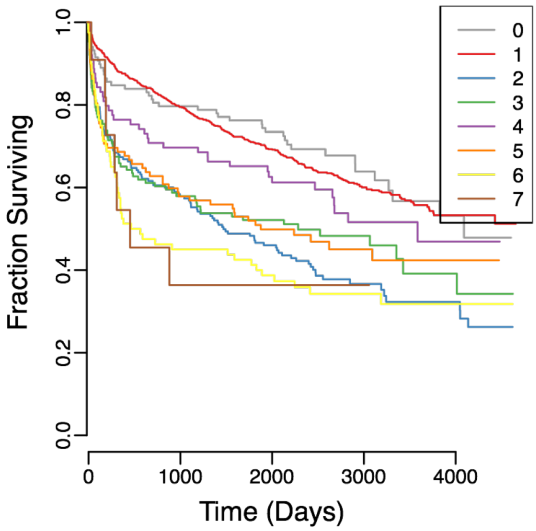


Figure 13 Classes show distinct subtype compositions and survival outlooks. (a, b) Patient assignment to WHO diagnostic groups or subtypes compared to patient assignment to proposed classes. **(c)** Kaplan-Meier plot for proposed classes.