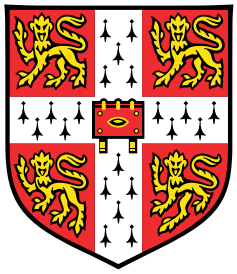
Machine Learning for Precision Oncology

Genomic Classification and Analysis of Diffuse Large B Cell Lymphoma



Camilo Ruiz

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
MPhil in Biological Sciences

King's College September 2017

I would like to dedicate this thesis to my family …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 20,000 words excluding appendices, bibliography, footnotes, tables, and equations.

Camilo Ruiz
September 2017

Contributions

This project was carried out in collaboration with the Haematological Malignancy Research Network, the Leeds Teaching Hospital, and the Sanger Institute. Patients were diagnosed and samples were collected at St James' Hospital in Leeds (Dr. Sharon Barrons, Dr. Catherine Cargo, Dr. Cathy Burton, and Dr. Jan Taylor). Clinical outcome data was collected and processed at the University of York (Dr. Eve Roman, Dr. Alexandra Smith, Dr. Russell Patmore, Dr. Simon Crouch, and Dan Painter). DNA sequencing was conducted by the sequencing core at the Wellcome Trust Sanger Institute. Genome assembly was conducted by the Sanger Institute's Cancer and Somatic Mutation Group's IT team led by Adam Butler and Jon Teague (i.e. from raw sequencing files to initial VCF files). Finally, I conducted all analysis downstream of genome assembly. I constructed a computational pipeline to identify driver mutations within the set of sequencing variants. I then manually curated these variants in collaboration with Dr. Peter Campbell and Dr. Philip Beer, a consulting Haematopathologist at Leeds Teaching Hospital. Subsequently, I used these drivers to conduct a genomic landscape analysis of the B-NHLs within our cohort. Finally, I used Bayesian Dirichlet processes to generate a preliminary classification of the disease. The project was advised by Dr. Peter Campbell at the Sanger Institute. Additionally, my thesis committee included Dr. Peter Campbell, Dr. Philip Beer, Dr. Moritz Gerstung of the European Bioinformatics Institute, and Dr. Brian Tom of the MRC Biostatistics Unit at the University of Cambridge.

Acknowledgements

To my advisor, Professor Peter Campbell: Thank you for mentoring me and guiding me on this journey. Thank you for trusting me with this project and pushing me to succeed.

To the members of my thesis committee Professor Peter Campbell, Professor Moritz Gerstung, Professor Brian Tom, and Dr. Philip Beer: Thank you for challenging me and providing guidance along this project. I appreciate your continual support and insight in helping me avoid mistakes and guiding me towards new insights.

To our collaborators Dr. Sharon Barrans, Dr. Catherine Cargo, Dr. Simon Crouch, Dr. Cathy Burton, Dr. Russell Patmore, Dr. Alexandra Smith, and Dr. Jan Taylor: Thank you for providing the clinical samples that made this study possible. Also thank you for helping integrate some of the preliminary clinical data and offering broad advice on the project direction.

To the Campbell Lab, especially Grace Collord, Dr. Daniel Leongamornlert, and Dr. Francesco Maura: Thank you for your endless help. You helped me overcome a great deal of challenges, whether graphing in R or interpreting the clinical output. A special thanks to Grace in particular for providing the graphics code necessary to visualize gene-level mutational profiles.

To the Gates Cambridge Trust: Thank you for giving me the opportunity to study at Cambridge as a Gates Scholar. I am indebted to the Trust and hope to use what I have learned to improve the world.

To my friends and family: Thank you for your constant support and love, in both the challenging and fruitful times.

Abstract

The most common lymphoma in adults, Diffuse Large B Cell Lymphoma (DLBCL) accounts for 30-35% of all non-Hodgkin lymphoma (NHL) cases. Even though DLBCL is curable in advanced stages, up to one-third of patients will not achieve a cure with their initial therapy. Today, there is no effective way to predict which patients will or will not be cured by first-line chemotherapeutic treatment. Patients who are not initially cured relapse, develop chemoresistance, and ultimately die of their disease.

Current classification and prognostication schemes do not account for much of the genetic and molecular heterogeneity of DLBCL. Indeed, the gold standard WHO classification uses clinical data, morphology, phenotype, cytogenetics, and molecular characteristics to demarcate DLBCL subtypes. However, it does not incorporate many of the genetic lesions that both cause DLBCL and make it heterogeneous. As a result, the most common WHO subtype of DLBCL – DLBCL, not otherwise specified (DLBCL NOS)–likely encapsulates multiple disease subtypes for which conventional diagnostic approaches have not yet yielded clear methods of discrimination.

The prognostication and treatment guidelines for DLBCL are similarly uniform, again not reflecting the heterogeneity inherent to DLBCL. The gold standard clinical prognostic tool, the Revised International Prognostic Index (R-IPI), sorts patients into three risk groups based on factors such as age and whether their lactate dehydrogenase level is elevated. None of the R-IPI factors, however, accounts for the genetic basis of DLBCL and cannot therefore incorporate prognostic information from genetic variability between patients within the same risk group. Virtually all DLBCL patients receive the same first-line therapy, R-CHOP, despite the probability that the genetic and biological heterogeneity will result in heterogeneous response to the potential treatments available. Up to one third of patients will not be cured by R-CHOP and their prognosis suffers significantly in the case of relapse.

In this study, we propose a novel, purely genomic classification for DLBCL and other B-cell non-Hodgkin lymphoma (B-NHLs) that incorporates the genetic heterogeneity inherent to the disease. By analysing the genetic lesions of 1607 B-NHL patients over 15 years and then performing a machine-learning based clustering, we identify seven distinct classes with characteristic genetic lesions and patterns of co-mutation. These classes aptly distinguish Follicular Lymphoma (FL) and Burkitt Lymphoma (BL) samples from DLBCL samples while simultaneously resolving the heterogeneity of DLBCL. Class 5, for example, shows hallmark mutations of Splenic Marginal Zone Lymphoma (*NOTCH2, BCL10, SPEN*),

suggesting these DLBCL patients represent transformed lymphomas. Such a conclusion could not have been drawn from histology alone and importantly, suggests these patients may respond differently to novel therapies compared to other DLBCL subtypes. We also present a genomic landscape analysis more complete and powerful than prior work since our study is nearly 10X larger than the largest prior B-NHL genetics study. We present mutation profiles at the gene level for nearly 200 genes implicated in lymphoma, identifying previously unreported mutations such as the aberrant splicing of a single exon in *SGK1*. Future work adding copy number, gene expression, and translocation data will enhance the robustness and resolution of our classification scheme and landscape analysis.

Table of Contents

List of Figures

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**.**

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.

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.

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

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*.

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.

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.

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.

Nomenclature

ABC-DLBCL: Activated B Cell-Like Diffuse Large B-Cell Lymphoma

BCL, Int.: B-cell lymphoma, intermediate between DLBCL and classical HL

BL: Burkitt Lymphoma

B-NHL: B Cell non-Hodgkin Lymphoma

DLBCL: Diffuse Large B Cell Lymphoma

FL: Follicular Lymphoma

FL-LC: Follicular lymphoma, large cell

GC: Germinal Centre

GCB-DLBCL: Germinal Centre B Cell-Like Diffuse Large B-Cell Lymphoma

GZL: B-cell lymphoma, intermediate between DLBCL and classical HL

IV-LBCL: Intravascular large B-cell lymphoma

PB-LBCL: Plasmablastic large B-cell lymphoma

SMZL: Splenic marginal zone lymphoma

THR-LBCL: T-cell/histiocyte-rich large B-cell lymphoma

1. Introduction

1.1. Classifying cancer, a deeply heterogeneous disease

Cancer is an extremely heterogeneous disease, showing distinct clinical and biological manifestations between cancer types, within subtypes, and even between patients with the same subtype. Such heterogeneity results from the pathogenesis of cancer: as somatic mutations accumulate over time, in a myriad of genes and tissues, a variety of pathways are dysregulated leading to cell proliferation. Patients of the same cancer type may carry distinct causative mutations. Indeed, different tumour cells within a patient may also carry distinct causative mutations. Overall, the myriad combinations of genetic mutations targeting distinct genes, cells, and tissues generate different clinical courses, survival likelihoods, and treatment responses between patients.

To deal with such heterogeneity, classification schemes have been developed. By grouping patients according to common characteristics, broad patterns emerge with patients sorted according to common prognoses and responses to treatments. Historically, such classification has relied on histological, morphological, and immunohistochemical examination of the patient's tumour cells. Such an approach, however, is lacking in a few respects. First, different cancer types have been shown to share similar histological, morphological, and immunohistochemical characteristics in spite of having distinct genetic causes and treatment responses. As a result, traditional classification systems often fail to resolve categories at a high enough level precisely because they do not incorporate the causative genetic changes leading to disease. Second, resulting classes are often difficult to interpret in the context of the pathways distinguishing diseases, making translation to therapy more challenging. Indeed, a distinct morphological profile does not immediately suggest a new therapeutic target. Thus, even when a new class is demarcated, it is often challenging to directly improve its clinical course. Finally, the clinical insights of some distinct classes have struggled with widespread relevance and reproducibility. For example, DLBCL was traditionally classified according to centroblastic, immunoblastic, and anaplastic subtypes with distinct clinical courses. Such clinical differences, however, have struggled with reproducibility. Additionally, the morphological subtype with the worst clinical course (anaplastic) has shown to occur in only 7.4% of cases, making widespread clinical relevance poor[1].

With the advent of more readily available patient samples and cheap sequencing, classification schemes have been shifting toward resolving cancer on the basis of molecular

and genetic differences. Throughout, blood cancers have led the way. Indeed, Chronic Myeloid Leukemia began with morphological characterization[2–4] which then gave way to the Philadelphia Chromosome and the BCR-ABL mutation as the primary classification characteristics[5]. Acute Myeloid Leukaemia then followed with the first identification of a specific genetic subtype: Acute Promyelocytic Leukaemia[6–9]. Both of these categories of disease, defined by their canonical genetic lesion, now have specific targeted therapies against this genetic change, radically improving treatment outcomes for those patients. In solid tumours, Ewing's Sarcoma was defined by a t(11;22) translocation[10]; breast cancer became defined by *ERBB2*[11,12]; and non small cell lung cancers are increasingly defined by specific kinase mutations[13].

Broadly, genetic and molecular classification approaches share a series of advantages over traditional approaches. First, these classifications rely on the causative genetic and molecular changes that underlie cancer. As a result, they are more likely to be clinically relevant, durable, and reproducible. Even as treatments change, for example, the underlying genetic structure of cancers are likely to remain the same. Second, genetic classifications group patients on the basis of pathways rather than morphology, leading to improved biological insights. By extracting the unique pathways that distinguish patient groups, the pathogenesis of distinct cancers become clearer. Finally, genetic classifications can improve clinical prognostication and suggest therapeutic targets. Targeted therapies inhibiting a specific gene that defines a genetic class can be reserved exclusively for patients of that class, improving treatment selection. Similarly, when a new patient class emerges that is resistant to traditional therapies, the pathway dysregulations allowing such resistance can be examined and new target combinations can be suggested.

1.2. A purely genetic classification for DLBCL

While an effective classification scheme could benefit all cancers, it could especially benefit DLBCL. Compared to other cancers, DLBCL exhibits a higher degree of genetic heterogeneity since it derives from Germinal Centre B cells which often have unstable genomes. Additionally, an effective classification could immediately help clinical outcomes. 30% of DLBCL patients today are not cured by R-CHOP, the front line chemotherapeutic treatment. These patients subsequently relapse upon which their prognosis suffers significantly. At present, there is no way to pre-emptively identify these patients in spite of the fact that they likely exhibit genomic differences that prevent effective R-CHOP treatment. A classification system that identifies these patients would enable physicians to move them

toward more aggressive clinical regimens such as stem cell transplantation or experimental therapies. It could also help develop more targeted clinical trial protocols, in which only those patients likely to relapse are recruited.

In this study, we propose a novel classification scheme for B-NHLs and DLBCL based purely on genetic changes. By conducting targeted deep sequencing of 1607 B-NHL patients and subsequently classifying these patients on the basis of genetics alone, we: (1) identify novel mutation patterns such as the aberrant splicing of an exon in *SGK1*, (2) produce the first ever purely genetic classification of B-NHLs broadly and DLBCL in particular, (3) unlock previously unknown patterns of co-mutation which shed light on unique pathogenesis mechanisms, (4) identify novel subclasses of DLBCL, including one with hallmark SMZL mutations, revealing new insights regarding DLBCL pathogenesis, and (5) set the stage for a follow up clinical study examining the unique lesions that give 30% of DLBCL patients poor R-CHOP responses[14], thus shedding light on the critical clinical question of DLBCL.

Our study occurs in three main stages (Figure 1a). First, we identify driver mutations in 292 genes implicated in lymphoid and myeloid malignancies across 1607 patients. Second, we conduct mutational analysis at the landscape level and at the gene-level for DLBCL, FL, and BL – the primary B-NHLs included in our study. Finally, we utilize Bayesian Dirichlet Processes – a machine learning classification approach – to classify our samples on the basis of genetics alone.

Our study draws its effectiveness from its depth and size. We sequence 1607 total patients spread across a range of B-NHL subtypes, with the largest patient populations for DLBCL and FL (Figure 1b). Our study is one of only two studies of such scope[15] and is roughly 10X larger than all other previous DLBCL and B-NHL genetic sequencing studies, allowing us to consider more B-NHL subtypes. Additionally, our targeted sequencing approach allows us to sequence at greater depth, thus identifying rarer and clinically useful variants previously missed. Combined, such scope and scale finally allows us to use Bayesian Dirichlet Processes – a machine learning approach that can effectively delineate co-mutation patterns with a sufficiently large dataset. While we apply this approach to DLBCL and B-NHLs in this study, the broad methodology should hold equally for other cancers. As a result, we see this as a foundational study for a new paradigm in cancer classification. Additionally, upon further work which will incorporate gene expression data, copy number changes, and translocation data, we will be able to (1) compare our classification robustly with the cell of

origin classification based on gene-expression profiling, potentially providing a surrogate and (2) present the most integrative classification scheme to date.

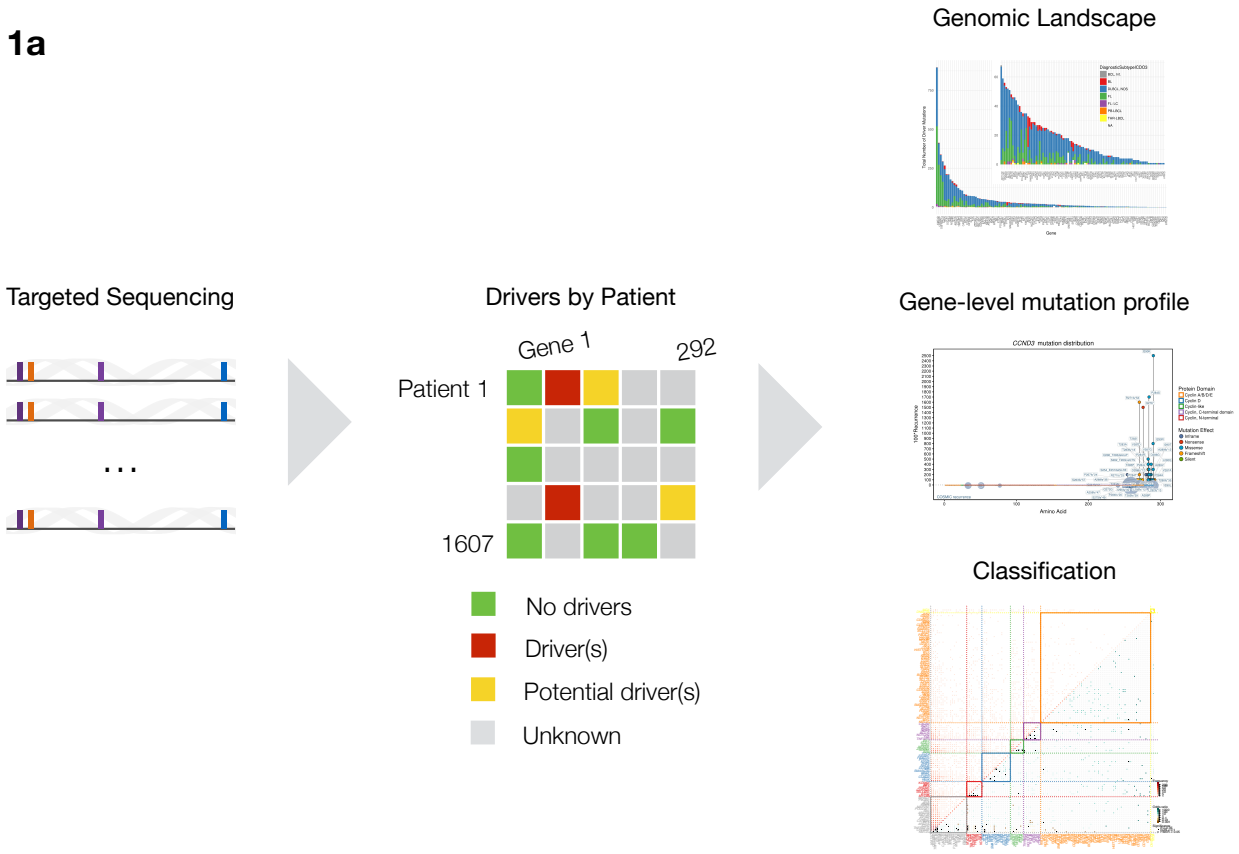


Figure 1 Overview of Study. (a) Process Overview. Targeted Sequencing of 292 genes was conducted on 1607 lymphoma samples. Subsequently, variants were called, filtered into somatic mutations, and annotated as drivers or passengers. Finally, three analyses were conducted investigating the genomic landscape of B-NHLs, examining the mutation profiles of crucial lymphoma genes, and creating the first ever purely genetic classification of B-NHLs and DLBCL in particular. **(b)** Patient Cohort Overview.

1b

	Overall	Sex		Age	OS	Treatment				Survival Status	
	n (% total)	F (% subtype)	M (% subtype)	Median (Range)	Median (Range)	Not known (% subtype)	Not treated (% subtype)	Treated (% subtype)	Watch and wait (% subtype)	0 (% subtype)	1 (% subtype)
Total	1607 (100%)	771 (47.9%)	832 (51.7%)	66 (3, 98)	2112 (-19, 4655)	69 (4.2%)	125 (7.7%)	1185 (73.7%)	224 (13.9%)	871 (54.2%)	732 (45.5%)
Diagnostic Group WHO											
DLBCL-HL intermediate	15 (0.9%)	6 (40%)	9 (60%)	54 (11, 81)	2155 (15, 2533)	0 (0%)	0 (0%)	15 (100%)	0 (0%)	12 (80%)	3 (20%)
Burkitt Lymphoma	39 (2.4%)	6 (15.3%)	33 (84.6%)	38 (3, 86)	2023 (-1, 4623)	1 (2.5%)	7 (17.9%)	31 (79.4%)	0 (0%)	24 (61.5%)	15 (38.4%)
Diffuse large B-cell lymphoma	962 (59.8%)	445 (46.2%)	517 (53.7%)	69 (8, 98)	1819 (-19, 4655)	46 (4.7%)	114 (11.8%)	802 (83.3%)	0 (0%)	439 (45.6%)	523 (54.3%)
Follicular lymphoma	587 (36.5%)	314 (53.4%)	273 (46.5%)	64 (20, 98)	2493 (1, 4655)	22 (3.7%)	4 (0.6%)	337 (57.4%)	224 (38.1%)	396 (67.4%)	191 (32.5%)
Diagnostic Subtype ICDO3											
DLBCL-HL intermediate	15 (0.9%)	6 (40%)	9 (60%)	54 (11, 81)	2155 (15, 2533)	0 (0%)	0 (0%)	15 (100%)	0 (0%)	12 (80%)	3 (20%)
Burkitt lymphoma	39 (2.4%)	6 (15.3%)	33 (84.6%)	38 (3, 86)	2023 (-1, 4263)	1 (2.5%)	7 (17.9%)	31 (79.4%)	0 (0%)	24 (61.5%)	15 (38.4%)
Diffuse large B-cell lymphoma, NOS	925 (57.5%)	430 (46.4%)	495 (53.5%)	69 (8, 98)	1824 (-19, 4655)	44 (4.7%)	107 (11.5%)	774 (83.6%)	0 (0%)	422 (45.6%)	503 (54.3%)
Follicular lymphoma	566 (35.2%)	305 (53.8%)	261 (46.1%)	64 (20, 98)	2477 (1, 4655)	22 (3.8%)	4 (0.7%)	318 (56.1%)	222 (39.2%)	378 (66.7%)	188 (33.2%)
Follicular lymphoma: large cell	21 (1.3%)	9 (42.8%)	12 (57.1%)	57 (37, 84)	3313 (88, 4589)	0 (0%)	0 (0%)	19 (90.4%)	2 (9.5%)	18 (85.7%)	3 (14.2%)
Intravascular large B-cell lymphoma	1 (0%)	0 (0%)	1 (100%)	71 (71, 71)	2232 (2232, 2232)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)
Plasmablastic large B-cell lymphoma	14 (0.8%)	5 (35.7%)	9 (64.2%)	71 (18, 95)	426.5 (2, 3379)	1 (7.1%)	4 (28.5%)	9 (64.2%)	0 (0%)	3 (21.4%)	11 (78.5%)
T-cell/histiocyte-rich large B-cell lymphoma	22 (1.3%)	10 (45.4%)	12 (54.5%)	65.5 (30, 89)	1836 (7, 2782)	1 (4.5%)	3 (13.6%)	18 (81.8%)	0 (0%)	13 (59%)	9 (40.9%)

2. Background

Before diving into our study, we explain the relevant prior work as it relates to the biological and genomic pathogenesis, clinical characteristics, and classification of B-NHLs.

2.1. Biological and Genomic Pathogenesis of B-NHLs

2.1.1. B-Cell lymphomagenesis occurs in germinal centres where transcriptional changes regulate B-cell development

The majority of B cell lymphomas originate in the Germinal Centres (GCs). The GCs are histological structures whose goal is to proliferate naïve B cells and enable their differentiation into Memory B cells and Plasma Cells [16–18](Figure 2a). Functionally, the Germinal Centre reaction takes three steps[19]. First, naïve B cells become activated upon encounter with an antigen and interaction with $CD4^+$ T cells in T cell-rich areas of secondary lymphoid organs. They subsequently aggregate into follicles to form GCs. In the dark zone of the GC, B cells proliferate rapidly, and use immunoglobulin somatic hypermutation to produce a high diversity of antibodies. Second, B cells move into the light zone where they are selected on the basis of antigen affinity. Finally, B cells either differentiate into Memory B cells, differentiate into a Plasma Cells, re-enter the dark zone, or undergo apoptosis.

Crucially, the GC reaction is regulated by a complex transcriptional network whose dysregulation produces various lymphomas [20,21](Figure 2b). The first phase of the GC reaction—initiation, B cell proliferation, and somatic hypermutation—is regulated by three major transcriptional events. First, the *MYC* gene is induced to initiate dark zone formation and encourage B cell proliferation. Although the exact molecular mechanisms are unknown, *MYC* generally stimulates proliferation by increasing DNA replication, metabolism, and telomerase activity[22]. Second, *BCL6* is induced as the master regulator of GC maintenance and formation (Figure 2c). *BCL6* encourages somatic hypermutation of immunoglobulin loci by inhibiting differentiation, B cell activation, and the DNA damage response[23–26]. Third, *EZH2*-mediated epigenetic silencing occurs to further promote proliferation and prevent differentiation[27].

The second phase and third phases of the GC reaction – selection for high affinity antigens; and differentiation, dark zone re-entry, or apoptosis – are regulated by four transcriptional events. First, the induction of *MYC* allows for dark zone re-entry[25,26]. Second,

the activation of *NF-KB* promotes selection of high affinity antibodies and differentiation of corresponding B cells[28–33]. Third, the downregulation of *BCL6* leads to GC exit and differentiation[34]. Finally, the induction of *PRDM1* allows plasma cell differentiation[35–38].

2.1.2. Dysregulation of the GC Reaction defines the characteristic genomic alterations of B-NHLs

Dysregulation of the GC reaction described above is the source of the majority of B-NHLs. Indeed, BL, FL, and DLBCL jointly comprise 80% of B-NHLs and result from dysregulation of different steps of the GC reaction[39]. These B-NHLs contain mutations standard to most tumours: deletions, amplifications, and nonsynonymous point mutations with loss-of-function or gain-of-function. More importantly, B-NHLs share a series of characteristic genomic alterations stemming from GC dysregulation. Owing to the immunoglobulin remodelling function of the GC, B-NHLs carry lesions from aberrant somatic hyper mutation and chromosomal translocation that are less common in other cancers. Moreover, translocations in B-NHLs generally pair the coding element of a gene with a heterologous promoter, leading to dysregulated expression of an oncogene[19]. By contrast, translocations in other cancers, like Acute Leukemia, generally result in fusion genes and chimeric proteins. Translocations in B-NHLs can be grouped into three categories based on the source of the error. First, translocations such as t(14;18) involving *IGH* and *BCL2* in FL result from mistakes in the RAG-mediated V(D)J recombination process. Second, translocations such as immunoglobulin-*MYC* translocations in sporadic BL result from mistakes in the AID-dependent class switch recombination process. Third, translocations such as immunoglobulin-*MYC* translocations in endemic BL result from errors in the AID-mediated somatic hypermutation mechanism which may lead to DNA breaks[19].

In addition to these characteristic genomic alterations, each B-NHL has a set of uniquely defining genetic characteristics (Figure 2a).

2.1.2.1. BL is defined by *MYC* translocation, mutations in *TCF3* and *ID3*

BL samples have gene expression patterns similar to dark zone B cells and represent aggressive malignancies[40,41]. Three main genetic changes characterize BL. Occurring in 100% of cases, the hallmark genetic lesion of BL is *MYC* translocation into the immunoglobulin locus[42,43]. Translocation causes ectopic *MYC* expression which promotes replication, causing replication stress in proliferative dark zone B cells and thus lymphomagenesis[26,40,44]. Second, 70% of BL mutations have mutations of *TCF3* or *ID3*

which promote "tonic" BCR signalling to occur in an antigen-independent way. By contrast, cells without this mutation, for example ABC-DLBCL samples (described below), rely on chronic activation of BCR[45]. Third, the Ga13-dependent pathway is dysregulated, thus causing GC B cell migration and preventing confinement[46]. This mutation also occurs in GCB-DLBCL (described below).

2.1.2.2. FL is defined by t(14;18) translocation and *KMT2D* inactivation

FL results from the clonal expansion of follicles containing GCs with high SHM activity[47]. These samples often have gene expression patterns similar to B cells arrested in the light zone[40,48]. Though an indolent disease, FL can transform into DLBCL[49,50]. Two main genetic events distinguish FL. First, 80% of FL samples have a t(14; 18) translocation, juxtaposing the *BCL2* gene with the *IGH* locus and causing ectopic expression[51,52]. The dysregulation of *BCL2* leads to an anti-apoptosis response. Second, >80% of FL cases exhibit the genetic inactivation of *KMT2D*[53,54]. The exact consequences of this inactivation are currently unknown.

2.1.2.3. DLBCL is defined by *BCL6* dysregulation, inactivation of chromatin modifiers (*EP300, CREBBP, KMT2D*), and disruption of immune surveillance

Comprising 40% of all B-NHL, DLBCL represents the most common form of B-NHL lymphoma. While some DLBCL cases arise de novo, other cases arise from transformation of less aggressive B-NHLs (chronic lymphocytic leukaemia and FL)[50,55]. DLBCL samples have gene expression profiles that map into two broad categories: activated B cell-like DLBCL (ABC-DLBCL) and GC B cell-like DLBCL (GCB-DLBCL). GCB-DLBCL samples' gene expression profiles match those of light zone B cells[40,48]. ABC-DLBCL samples' gene expression profiles match those of GC cells arrested during early stages of post-GC plasma cell differentiation (plasmablasts)[40,48]. While some mutations occur across DLBCL subtypes, each DLBCL subtype (ABC-DLBCL or GCB-DLBCL) has specific genomic lesions characterizing it.

Three broad types of genomic lesions are shared across DLBCL subtypes. First, many DLBCL patients have inactivation of *EP300* or *CREBBP* (40%) and/or *KMT2D* (30%), chromatin modifiers crucial to epigenetic regulation[53,54,56,57]. Second, 30% of DLBCL cases and 15% of FL cases exhibit *BCL6* dysregulation, thereby suppressing the DNA damage response and inhibiting differentiation[58]. *BCL6* dysregulation can occur either via disruption of *BCL6*'s autoinhibitory circuit or through chromosomal translocations with promoters of

other genes or the IGH locus[59–61]. Finally, >60% of DLBCL cases exhibit immune escape through various mechanisms. Diminished expression of MHC-I allows DLBCL cells to evade cytotoxic T lymphocytes. CD58 Inactivation or disrupted transport similarly allows evasion of natural killer cells[61]. Combined, these mutations allow DLBCL samples to evade both cytotoxic T lymphocytes and natural killer cells.

2.1.2.3.1. GCB-DLBCL, the first DLBCL subtype, is characterized by *EZH2* activation and altered GC B cell migration

GCB-DLBCL samples share some genetic overlap with BL and FL. In particular, 10% of GCB-DLBCL samples exhibit *MYC* translocation; 40% exhibit *BCL2* translocation; and samples exhibiting both (i.e. double hit cases) show worse clinical outcomes[63,64].

Beyond the similarities, two additional genetic alterations characterize GCB-DLBCL. First, 21% of GCB-DLBCL cases have a gain of function mutation in *EZH2*, thereby promoting GC proliferation and inhibiting post-GC differentiation[65,66]. Second, 30% of GCB-DLBCL cases and 15% of BL cases exhibit mutations in *S1PR2, GNA13, ARHGEF1*, or *PR2Y8* which disrupt the Ga13-dependent pathway, thus allowing B cells to migrate from the GC into lymph and blood circulation[67]. In spite of knowledge of these alterations, the precise pathogenesis of GCB-DLBCL is not well understood.

2.1.2.3.2. ABC-DLBCL, the second DLBCL subtype, is characterized by constitutive NF-KB signalling and inhibition of terminal differentiation

Two main genetic alterations characterize ABC-DLBCL. First, NF-KB is constitutively activated. Such activation can occur through multiple mechanisms. In 20% of cases, *CD79A* and/or *CD79B* mutations generate chronic BCR signalling[68]. In 10% of cases, *CARD11* activating mutations constitutively activate NF-KB. In 35% of cases, *MYD88* mutations constitutively activate MYD88 and affect JAK/STAT3 signalling[69]. In 30% of cases, *TNFAIP3* inactivating mutations inhibit the stoppage of NF-KB responses[70]. Finally, antigens or autoantigens can chronically stimulate BCR. Second, the negative regulation of *PRDM1*, the plasma cell master regulator, blocks terminal differentiation to plasma cells. This negative regulation occurs through either bi-allelic activation of *PRDM1* (30% of cases), *SPIB* gain of function which increases inhibition of *PRDM1* transcription (25% of cases), or *BCL6* translocations, which cause constitutive repression of *PRDM1*[71–75]. Combined, these genomic lesions grant ABC-DLBCL a worse clinical course and outcome than GCB-DLBCL[48,76].

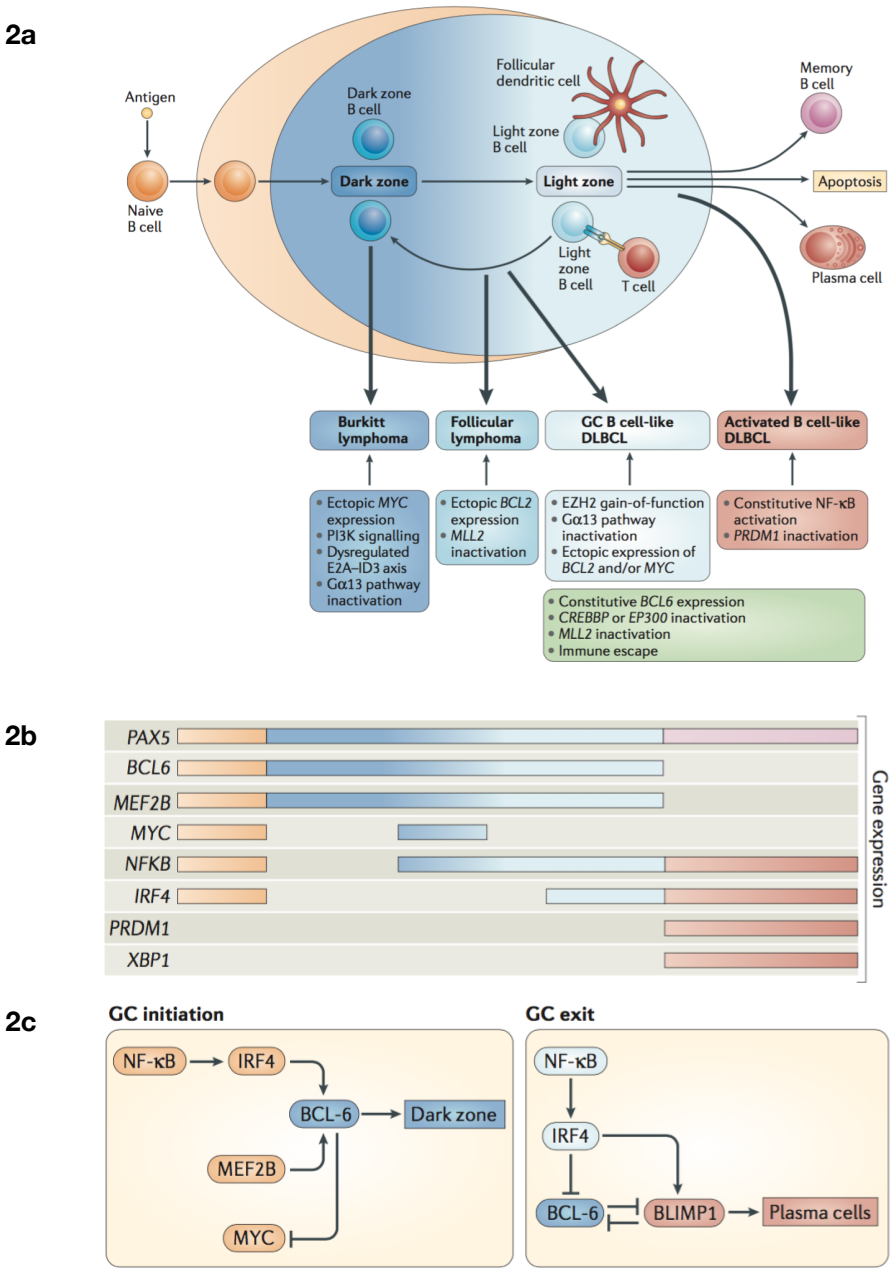


Figure 2 B-Cell Lymphomagenesis originates in the Germinal Centres. (a) B-NHLs correspond to dysregulation of different stages of B-Cell development. Each carry hallmark mutations disrupting a specific transition. **(b)** Transcriptional activity drives normal B cell development with gene expression driving transitions between stages. **(c)** Transcriptional networks work jointly to create major transitions such as GC initiation and GC exit, with *BCL6* as a master regulator. Adapted from Basso et al. 2015.

2.2. Clinical Characteristics of B-NHLs

B-NHLs share a set of clinical characteristics although they differ in their distinctive features, all reviewed below. Survival outcomes vary, but in all cases, prognostication and treatment lag behind recently acquired understanding of the molecular and genetic heterogeneity of B-NHLs. All citations from this section are taken from *Pathophysiology of Blood Disorders, Volume 2* by H. Franklin Bunn and Jon Aster[77].

2.2.1. B-NHLs share symptoms of immune dysregulation and are measured by a common staging system

Generally, B-NHLs share a set of common clinical features. B-NHLs usually present as a mass in the lymph nodes or secondary lymphoid tissues, though they can also present in virtually any organ in the body. Once presented, symptoms associated with immune dysregulation result, namely: B symptoms (weight loss, night sweats, fever), immunosuppression, and breakdown of immune tolerance. Additionally, B-NHLs generally have infectious agents as cofactors in development (*Helicobacter pylori*, HTLV-1, HHV-8, HIV, and EBV).

All B-NHLs share the same staging system: the Ann Arbor Staging for Lymphomas. Four stages exist consistent with increasing progression of the disease that are based on the number and location of nodes. Stage I corresponds to the involvement of a single lymph node group (I) or a single extralymphatic organ or site (IE). Stage II corresponds to the involvement of two or more lymph node groups on the same side of the diaphragm without (II) or with localized involvement of an extralymphatic organ or site (IIE). Stage III corresponds to the involvement of lymph node groups on both sides of the diaphragm without (III) or with localized involvement of an extralymphatic organ or site (IIIE). Stage IV corresponds to the extensive involvement of one or more extralymphatic organs or sites (i.e. bone marrow) with or without lymphatic involvement. In spite of the consistent staging system, however, the clinical course for each B-NHL is distinct: FL is indolent, DLBCL is aggressive, and BL is very aggressive. Moreover, this staging system lacks the resolution necessary to account for patient heterogeneity, particularly in DLBCL.

2.2.2. FL is an indolent lymphoma with a passive clinical course

FL is the most common indolent lymphoma, representing 20,000 new cases per year in the US. Upon presentation, FL is usually asymptomatic with painless lymphadenopathy. Patients are diagnosed via a biopsy of the lymph node and fall into two categories. The first

category of patient (20%) have spontaneous and transient remissions. The second category of patients have local symptoms due to FL progression: cytopenias from bone marrow involvement, hypersplenism, B symptoms, symptomatic extranodal disease (i.e. pleural effusions), and/or compromised organ function. Stage I patients generally are cured by local radiation. Patients in more advanced stages receive chemotherapy and rituximab whereupon >90% show excellent responses over five years with resistance being common afterwards. Finally, 2% of patients transform to more aggressive lymphomas per year. Overall, FL now shows a median survival of >10 years.

2.2.3. BL is a rare but highly aggressive lymphoma

BL is a highly aggressive lymphoma accounting for less than 2% of adult lymphomas. BL occurs in three clinical settings: (1) in subequatorial Africa where BL is latently infected with EBV and/or malaria as a cofactor, (2) in the US where BL presents in a sporadic form and 30% of cases occur with EBV, and (3) in patients with immunodeficiency, often resulting from HIV and/or EBV. Regardless of the clinical setting, BL arises in extranodal sites often in the abdomen as a rapidly growing tumour mass with a "starry sky" appearance. Immunohistochemistry shows pan-B-cell (i.e. CD20) and GC B-cell (i.e. CD10 and BCL6) markers but no BCL2. Additionally Ki-67 is seen as a marker of active growth and *MYC* rearrangements are common.

The prognosis and treatment of BL depend on stage, gender, age, and clinical setting. Endemic BL is localized and responds to chemotherapy. Sporadic and HIV-associated BL generally spreads to the Central Nervous System, thus requiring prophylactic treatment. Sporadic BL is treated with intensive combination therapies and rituximab regimens coupled with intrathecal therapy to prevent disease in the CNS.

2.2.4. DLBCL is a common and aggressive lymphoma in which 30% of patients are not cured by first line treatment

DLBCL is the most common lymphoma, accounting for 30,000 new cases per year in the US. Most DLBCL presents in older adults with a median presentation age of 65. Most DLBCL presents in lymph nodes (2/3) though some (1/3) presents in extranodal sites, generally in the gastrointestinal tract. Almost any organ can be involved in DLBCL. Regardless of subtype, DLBCL is a rapidly expanding mass with B symptoms that mark it as an aggressive disease. Diagnosis is made by tissue biopsy and immunophenotyping which reveal pan-B-cells markers (i.e. CD20), BCL6 expression, and variable expression of CD10,

BCL2, and surface immunoglobulins. Additionally, serum lactate dehydrogenase (LDH) levels are elevated in over half of DLBCL patients unlike in indolent lymphomas.

DLBCL patients are prognosticated based on the Revised International Prognostic Index (R-IPI). The R-IPI considers negative prognostic factors at the time of diagnosis (stage III/IV of the disease, age > 60 years, elevated lactate dehydrogenase (LDH) levels, Eastern Cooperative Oncology Group (ECOG) performance status >=2, and >1 extranodal sites of disease) to sort patients into three risk categories. Patients with zero risk factors have >90% chance of 4-year progression-free survival. Patients with 1 or 2 risk factors have an 80% chance of 4-year progression-free survival. Finally, patients with 3, 4, or 5 risk factors have a 50% chance of 4-year progression-free survival[78].

Regardless of prognostication, all patients today are treated with R-CHOP. Up to one-third of patients do not achieve a cure with their initial therapy. Relapse and non-responsive patients have a poor prognosis. These patients may undergo more aggressive therapies such as stem cell transplantation; however, only 25% of such patients survive > 5 years. Today, there are no effective methods to distinguish up-front which patients will not be cured by first-line chemotherapeutic treatment. Identifying these patients up-front would allow doctors to move them toward more aggressive clinical regimens sooner or potentially toward experimental therapies.

2.3. Classification of B-NHLs

Although recent studies have uncovered the genetic heterogeneity inherent to DLBCL, current classification schemes have not yet fully incorporated this heterogeneity. Similarly, these classifications have done little to change clinical practice: the same frontline treatment is given to all patients although 30% of patients are not cured by R-CHOP. Our primary goal, therefore, is to improve upon known classification systems with the hope of discovering distinctive pathogenic and clinical characteristics that can guide treatments.

The primary goals of any classification system are three-fold. First, to delineate subcategories of the disease with interpretable differences that generate biological insights related to pathogenesis. Second, harness those insights to create targeted therapies for each class. Third, to then administer the optimal treatments for patients based upon which class of the disease they express.

Consistent with these goals, classification system schemes have been increasingly shifting towards molecular and genetic classification. As an example, some high grade B-cell lymphomas are now defined on the basis of whether they exhibit *MYC* and *BCL2* and/or *BCL6* rearrrangements.

Below, we describe the three current classification systems for DLBCL and B-NHLs: the WHO classification, cell-of-origin classification, and consensus clustering.

2.3.1. WHO Classification relies primarily on morphologic, biologic, immunophenotypic, and clinical parameters

The primary classification for lymphoid neoplasms including DLBCL is the WHO classification. The WHO classification primarily uses morphologic, biologic, immunophenotypic, and clinical parameters to separate lymphoid neoplasms into subgroups. Each subtype, described below, carries unique characteristics that often translate into distinct clinical courses.

2.3.1.1. DLBCL NOS

Accounting for 25-30% of NHL, DLBCL NOS is the most common WHO subtype. Crucially, DLBCL NOS is primarily an exclusion category: rather than having positive defining characteristics, DLBCL NOS samples are defined by not fitting the characteristics of other categories.

DLBCL NOS can originate de novo or as a result of transformation from FL or CLL. The most common genetic aberrations of DLBCL NOS include *BCL6* mutations (30% of cases[29]), *MYC* translocations (10% of cases[79]), and *BCL2* mutations in GCB-DLBCL.

Historically, DLBCL was resolved on the basis of morphological features. In particular, the recognition of centroblastic, immunoblastic, and anaplastic subtypes enabled classification and corresponded with clinical differences: centroblast tumours exhibited better prognostic outcomes than immunoblast tumours[1]. Major issues exist with this approach however. First, reproducibility of clinical differences is poor. Additionally, relatively small numbers of patients show immunoblastic morphology (only 7.4% of nearly 1000 patients in a clinical trial) showing that such morphological based classification had limited clinical applicability[80]. More recently, resolution of DLBCL NOS subgroups has been accomplished through gene expression studies delineating the cell of origin, described above.

2.3.1.2. DLBCL in specific subtypes

Other subtypes of DLBCL affect specific sites of the body: intravascular large B-cell lymphoma (IV-LBCL), primary cutaneous DLBCL, leg-type, and primary CNS DLBCL. Of those subtypes, only IV-LBCL was present within our study. IV-LBCL is rare and characterized by large B-cells occurring in the lumen of small blood vessels. The majority of IV-LBCL shows a gene-expression profile consistent with ABC-DLBCL and expresses the CD5 surface marker[81]. However in the absence of definitive radiological or clinical evidence and diverse symptoms, the disease is rarely diagnosed until autopsy.

2.3.1.3. High Grade B-Cell Lymphoma, with *MYC* and *BCL2* and/or *BCL6* rearrangements

This category includes all large B cell lymphomas with *MYC* and *BCL2* and/or *BCL6* rearrangements except those that fulfil criteria corresponding to follicular or lymphoblastic lymphoma[82]. These double hit and triple hit lymphomas correspond to a set of very aggressive tumours that generally exhibit chemoimmunotherapy refractoriness and high relapse rates. Substantial research is now being conducted to improve treatment for these patients[83,84].

2.3.1.4. B-Cell Lymphoma, unclassifiable with features intermediate between DLCBL and Hodgkin Lymphoma

This category, also known as grey zone lymphoma (GZL) contains samples intermediate between classical Hodgkin's lymphoma (cHL) and DLBCL (especially PMBL) in terms of clinical, morphologic, and immunophenotypic characteristics. Defining characteristics of GZL include: mediastinal involvement[85], diversity in cytologic appearance[85], and more cytogenetic aberrations than cHL, PMBL, and GZL[85,87,88]. The gene expression profile of this subcategory has not been examined. Additionally, the optimal treatment is unknown and cHL and NHL treatments have both been ineffective[89–91]. A more refined genetic profile and understanding of the pathogenesis of GZL could therefore inform treatment approaches.

For consistency with figures, we have used BCL, Int. as the abbreviation for this class.

2.3.1.5. T-Cell/Histiocyte Rich Large B-Cell Lymphoma

THR-LBCL is characterized by tumour cells high in reactive T cell or histiocyte content. THR-LBCL has distinct clinical features from other DLBCL subtypes: it presents predominantly in males in their fourth decade; includes spleen, liver, and bone marrow involvement; and follows an aggressive clinical course[92–94]. Generally, THR-LBCL is closely pathologically related to lymphocyte predominant Hodgkin lymphoma but differs in a few respects: the absence of small B-cells, the lack of a follicular structure, and the absence of T-cell rosettes around atypical B-cells.

2.3.1.6. Plasmablastic lymphoma

PB-LBCL results when immune surveillance declines due to advanced age and/or iatrogenic immunosuppression[95]. PB-LBCL occurs primarily in males with a median age of 50, with most cases being EBV-positive. Additionally, PB-LBCL patients generally have *MYC* translocations[89,96]. In terms of treatment, PB-LCBL show early responses to therapy but a poor overall prognosis including high likelihood of relapse[90].

2.3.1.7. Additional WHO subtypes not included within our study

In addition to the subtypes in our study, described above, additional subtypes exist and are discussed in the corresponding references. Two subtypes of DLBCL relate to the presence of EBV. First, EBV^+ DLBCL, NOS has an aggressive clinical course[97]. Second, DLBCL associated with chronic inflammation[98] primarily presents in males between age 65 and 70 with an aggressive clinical course[99,100]. An additional three subtypes of DLBCL

exhibit a plasmablastic phenotype (i.e. acquisition of plasma cell markers like CD38/CD138 with loss of or weak B-cell markers and MUM-1 positivity: ALK^+ large B-cell lymphoma[101–106], plasmablastic lymphoma[89–91,95,96,107–109], and primary effusion lymphoma[110–117]). For additional rare subtypes, we refer to the official WHO classification[82].

2.3.1.8. Follicular Lymphoma, Large Cell

In addition to the WHO classification presented above, one additional subtype (Follicular Lymphoma Large Cell or FL-LC) was present within our study. Generally, FL-LC is a subset of FL that is distinct from indolent follicular lymphomas. FL-LC is an aggressive lymphoma that presents with favourable prognostic features compared to FL. Both the clinical features and treatment response in FL-LC are similar to those in DLBCL[118].

2.3.1.9. Splenic Marginal Zone Lymphoma

While we didn't have any samples explicitly diagnosed as SMZL cases, our later classification analysis uncovered patients with genetic profiles consistent with SMZL. Clinically, SMZL is a low grade B-cell lymphoma showing splenomegaly, moderate lymphocytosis, and autoimmune thrombocytopaenia or anemia[119–121]. The immunophenotype of SMZL is similar to splenic marginal zone B-cells (CD27+, IgM+, IgD+[119,120,122]), however, the cell of origin is ultimately unknown. Indeed, ~90% of cases include multiple somatic mutations at variable degrees, suggesting the possibility for multiple cells of origin.

Genetically, SMZL manifests mutations in various pathways, all affecting marginal zone B-cell development: *KLF2* (20-42%), *NOTCH2* (6.5-25%), *NF-KB* (*CARD11* ~7%, *IKBKB* ~7%, *TNFAIP3* ~7-13%, *TRAF3* ~5%, *BIRC3* 6.3%). Marginal zone B-cell development, however, is not broadly well understood making the exact pathogenesis of SMZL unclear. Additionally, most SMZL shows recurrent gains and losses (7q32 deletion in 18-44% of cases) and translocations resulting in somatic hypermutation (*IGHV1-2* in 90% of cases[123–128]). Overall, the most common changes in SMZL are 7q deletion, *KLF2* mutation, *NOTCH2* mutation, and *IGHV1-2* usage[129]. The presence of these together implies that oncogenic cooperation may occur. For example, *KLF2* and *TRAF3* mutations may work together to activate the NF-KB pathway.

2.3.2. Gene expression profiling has classified DLBCL on the basis of cell of origin, yet issues remain

More recently, DLBCL categorization has moved toward the identification of distinct genetic and epigenetic changes. Gene expression profiling has resolved the DLBCL NOS group of the WHO classification into two subcategories: ABC-DLCBL and GCB-DLBCL. These subgroups, whose genomic and pathogenetic differences are described above, are based upon a "cell of origin" interpretation of DLBCL. Additionally, ABC-DLBCL and GCB-DLBCL have been shown to follow distinct pathways toward transformation and oncogenesis.

Consistent with this, targeted therapies affecting pathways responsible in the pathogenesis of only one subtype have helped patients primarily of that subtype. As an example, Bortezomib, a protease inhibitor blocking NF-KB signalling improves survival for ABC-DLBCL but not GCB-DLBCL patients.[77] Other studies have specifically suggested downregulating the BCR pathway through inhibition of BTK, PI3K, STK, MTOR, and SRC kinases in order to improve ABC-DLBCL survival.[77]

Issues exist with the gene expression profiling based classification, however. Gene expression profiling is technically difficult to perform and has limited availability in laboratory settings. As a result, immunohistochemistry has been proposed as an alternative way to identify ABC-DLBCL and GCB-DLBCL subtypes. Immunohistochemistry, however, (1) does not correspond directly to ABC-DLBCL and GCB-DLBCL distinctions although correlations exist, (2) produces unclassifiable cases, (3) uses the Hans algorithm which shows reproducibility and reliability issues.[77] If an alternative and more reliable way to identify cell of origin could be created, for example through the identification of specific mutations that correlate with these outcomes, the ABC-DLBCL and GCB-DLBCL classification would gain substantial clinical impact. Such a question could potentially be answered by a follow-up to our present study including gene expression data.

2.3.3. Alternatively, consensus clustering strives to classify DLBCL on the basis of metabolic pathway regulation

Finally, an independent classification has arisen based on consensus clustering which separates DLCBL samples by the up and down regulation of metabolic pathways.[130] The first cluster, the OxPhos consensus cluster, expresses genes important to mitochondrial metabolism and oxidative phosphorylation. The second cluster, the BCR consensus cluster, expresses genes critical to B-cell receptor signalling, regulation of the cell cycle, DNA repair, and B-cell transcription factors. The final cluster, the host response consensus cluster, expresses genes involved in the immune inflammatory response, the classic component

pathway, and the T-cell mediated immune response. Overall, OxPhos clustering has little overlap with the gene-expression cluster subtypes (ABC-GLBCL, GBC-DLBCL) and WHO classification above. As such, it is difficult to compare with the prior classification schemes and lies largely tangential to the classification presented in this paper.

3. Methods

Two phases composed this study. First, driver variants were extracted from genetic calls and merged with relevant clinical data. Second, this joint clinical and genetic data was harnessed to create a novel genetic classification of DLBCL.

3.1. Dataset

3.1.1. Patient Cohort

Patient samples came from the Haematological Malignancy Research Network (HMRN), a UK population-based registry whose methods have been previously described[131,132]. In short, fresh frozen or formalin-fixed, paraffin-embedded (FFPE) tissue samples were collected from 1607 lymphoma patients over 15 years. All samples collected were diagnostic biopsies. DNA was subsequently extracted for sequencing. Patient characteristics are available in Figure 1b.

Since patient samples were collected over 15 years, around 90% of curatively treated patients received rituximab. At the time of this manuscript, the information relating which patients did and did not receive rituximab was not yet processed and transferred to us by our collaborators. This will primarily affect the survival analysis at the end of this study, which is marked as being preliminary and will be heavily revised in future versions of this work. The 11% of DLBCL NOS patients marked as “not treated” were treated with palliative intent.

3.1.2. Library Preparation and Sequencing

Genetic sequencing targeted the exon region of 292 genes, specific SNPs in noncoding regions to facilitate copy number analysis, and known hot spot mutations outside of exon regions. Custom RNA baits were designed according to manufacturer guidelines (Agilent). Genomic DNA (125uL, 40ng/uL) was fragmented and prepared for Illumina DNA library sequencing via a Bravo automated liquid handler. Prepared samples were then indexed to a unique DNA barcode with 6 cycles of PCR. Next, the Agilent SureSelect protocol was used to prepare and hybridize 16 equimolar pools of libraries to custom RNA baits. RNA baits were designed to target the exons of 292 genes implicated in lymphomas and myeloid cancers. Additionally, baits targeted a series of SNPs in non-coding regions to allow later extraction of copy number changes. Finally, an Illumina HiSeq machine with a 75-base pair paired-end protocol was used to sequence enriched pools of 96 cases.

3.1.3. Clinical Data

The following clinical data was collected for all patients: sex; age at diagnosis, WHO Diagnostic Group: Diffuse large B-cell lymphoma, Follicular lymphoma, Burkitt lymphoma, B-cell lymphoma (intermediate between DLBCL and classical HL; Diagnostic Subtype ICDO3: Diffuse large B-cell lymphoma (NOS); Follicular lymphoma, Burkitt lymphoma, Intravascular large B-cell lymphoma, Follicular lymphoma: large cell, Plasmablastic large B-cell lymphoma, T-cell/histiocyte-rich large B-cell lymphoma, B-cell lymphoma (intermediate between DLBCL and classical HL); overall survival: days since pathology report; survival status; and treatment: treated, not known, watch and wait. Additional clinical variables were also collected and are currently being processed by our collaborators.

3.2. Genetic Data Preparation

3.2.1. Sequencing Alignment

To align raw sequencing data to the human genome (NCBI Build 37), the BWA algorithm[133] was used. The coverage depth at each base-pair position was determined utilizing Bedtools® v2.15.0[134]. Sequencing was performed to an average target depth of 500x reads per base, although there was inevitably patient-to-patient and gene-to-gene variation around this target.

3.2.2. Variant Calling

DLBCL includes a spectrum of genetic mutations including indels, complex rearrangements, and point mutations. We utilized two approaches to call relevant variants. First, point mutations were called using a modified version of the CaVEMan [135]algorithm with a single cord blood sample designated as the normal (Cancer Variants through Expectation Maximisation, https://github.com/cancerit/CaVEMan). CaVEMan calls variants by comparing sequencing data from each tumour sample with a designated normal sample and then calculating the likelihood of a mutation at each base-pair position locus. Thereby, CaVEMan identifies point mutations. Second, indel mutations were called using a modified version of the Pindel algorithm[136]. Third, Samtools mpileup was utilized to specifically identify mutations in known hotspot regions[137] like the *TERT* promoter. Finally, we manually reviewed all remaining variants using a genome browser (Gbrowse®)[138].

3.2.3. Variant Filtering

After calling the full set of variants, we removed off-target variants and variants that were suspected errors. These variants were removed based on (1) their presence in an off-target region, (2) a set of standard CaVEMan filters, (3) a set of standard Pindel filters, (4) a manually implemented set of additional filters, and (5) manual review.

First, we removed unmapped reads, PCR duplicates, and variants in off-target regions. Off-target variants were removed using Bedtools v2.15.0[134].

Second, variants were removed based on the CaVEMan filters below:

1. DTH: Less than 1/3 of mutant alleles were >= 25 base quality
2. RP: Coverage was less than 8 and no mutant alleles were found in the first 2/3 of a read (shifted 0.08 from the start and extended 0.08 more than 2/3 of the read length)

3. MN: More than 0.03 of mutant alleles that were >= 15 base quality found in the matched normal
4. PT: mutant alleles all on one direction of read (1 read allowed on opposite strand) and in second half of the read. Second half of read contains the motif GGC[AT]G in sequenced orientation and the mean base quality of all bases after the motif was less than 20
5. MQ: Mean mapping quality of the mutant allele was <21
6. SR: Position falls within a simple repeat using the supplied bed file
7. CR: Position falls within a centromeric repeat using the supplied bed file
8. PH: Mutant reads were on one strand (permitted proportion on other strand: 0.04) and mean mutant base quality was less than 21
9. TL: More than 10 percent of reads covering this position contained an indel according to mapping
10. SRP: More than 80 percent of reads contain the mutant allele at the same read position
11. HSD: Position falls within a high sequencing depth region using the supplied bed file
12. AN: Position could not be annotated against a transcript using the supplied bed file
13. VUM: Position has >= 3 mutant alleles present in at least 1 percent unmatched normal samples in the unmatched VCF
14. SE: Coverage is >= 10 on each strand but mutant allele is only present on one strand
15. MNP: Tumour sample mutant allele proportion – normal sample mutant allele proportion < 0.2

Third, indel variants were removed based on the filters built into Pindel[136].

Fourth, we removed additional variants based on manual filters. To remove variants within the error limits of CaVEMan and Pindel, we removed: variants with a read depth less than 10, variants with less than 3 reads, and variants with a variant allele fraction less than 0.05. To remove variants due to polymerase slippage in homopolymeric regions of the genome, we removed variants with a repeat length greater than 4 that also occurred in over 10% of individuals. Finally, we removed variants with insufficient read depth (<10). For reference, the average read depth across our study was ~500x reads per base.

3.2.4. Driver Identification

In order to identify driver mutations within the set of mutant calls, we first removed suspected germline polymorphisms. Next, we executed a pipeline for automated driver annotation. Finally and crucially, we reviewed all annotations manually before marking variants as Drivers, Passengers, or Variants of Unknown Significance.

First, we removed suspected germline polymorphisms by annotating the variants according to their population frequency in ExAC non-TCGA v0.3[139]. Any variants with a population frequency in ExAC non-TCGA > 0.001 were considered likely germline polymorphisms. While ExAC non-TCGA is contaminated with some relatively common somatic driver mutations, we reduced the risk of mistakenly removing common drivers by keeping a whitelist of common driver mutations and also examining suspected somatic mutations during the manual review step.

Next, we executed a pipeline for automated driver annotation. In order to be considered a driver, a variant must:

1. Not have a Vagrent [140]annotated mutation effect of the following type: THREE_PRIME_UTR, FIVE_PRIME_UTR, FIVE_PRIME_FLANK, THREE_PRIME_FLANK, INTRONIC, SPLICE_REGION, SILENT.
2. While also fulfilling any of the four conditions below:
    a. In a whitelist of well known driver mutations;
    b. Recurrence in COSMIC v82[141] > 3;
    c. Recurrence in COSMIC subsetted to hametopoetic and lymphoid diseases > 3;
    d. Likely to be a driver mutation based on it's effect and presence in a known tumour suppressor gene or known oncogene. As an example, a truncating mutation in a tumour suppressor gene would be considered a likely driver via this process.

Finally, a manual review process triaged suspected passengers and suspected passengers into two final categories of drivers and passengers. Beyond manually reviewing all annotations, this step was particularly important for removing missense variants that were only recurrent because of Somatic Hyper Mutation and not otherwise expected to be drivers.

3.3. Classification

3.3.1. Classification Techniques

To separate DLBCL patients into maximal, non-overlapping clusters, we utilized Bayesian Dirichlet Processes[142]. Bayesian Dirichlet Processes utilize a mixture model with an infinite prior distribution for the proportion and number of clusters. A Markov chain Monte Carlo method is then used learn the number, proportion, and assignments of the clusters. Analysis relied on the R package https://github.com/nicolaroberts/hdp which implements the non-hierarchical Dirichlet process we used. To fit the data, we used 100,000 burn-in iterations and 20,000 samples at 60 iterations between samples. After fitting the data, we merged clusters more than 5% similar on a cosine similarity metric and requested that only 99% of the data require explanation. Relevant code was adapted from a prior AML study by Papaemmanuil et al.[143]

3.3.2. Statistical Analysis

R version 3.3.3 was used for all statistical analysis and visualization.

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.

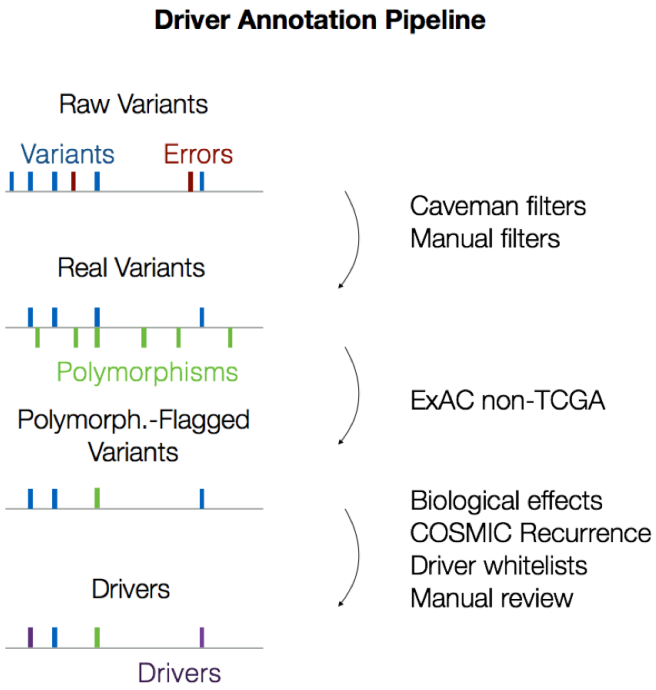


Figure 3 The driver annotation pipeline. The driver annotation pipeline annotates drivers from sequencing variants in three steps.

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.

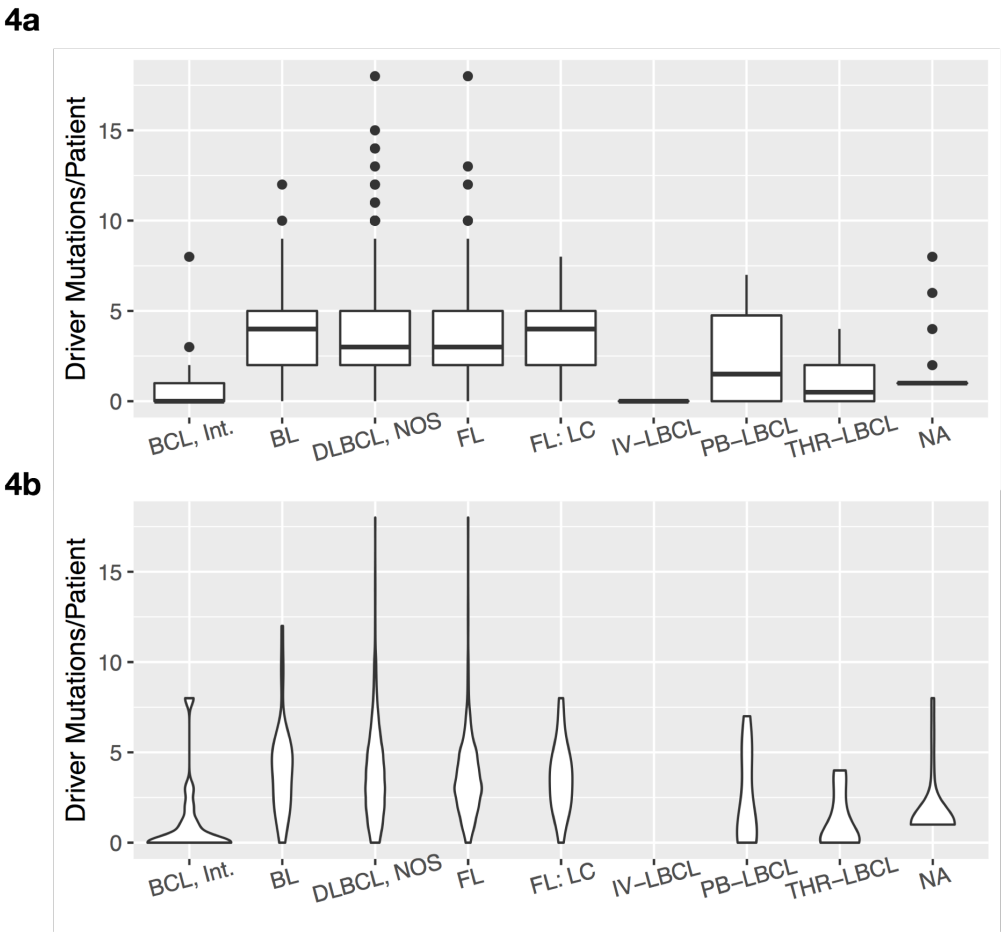
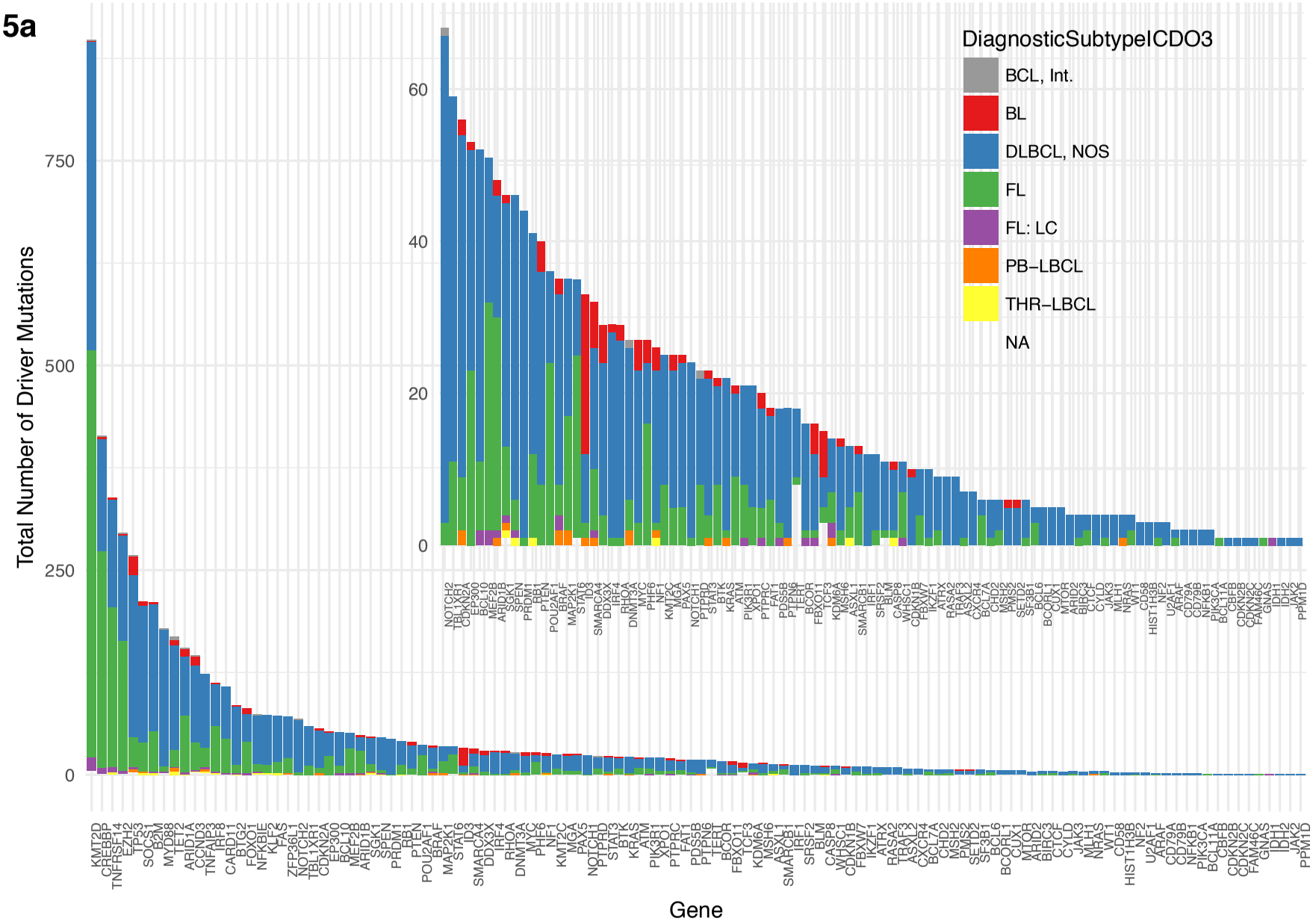


Figure 4 B-NHLs exhibit 3-4 driver mutations/patient. Average number of somatic driver mutation per patient across different diagnostic subtypes in this study. **(a)** Boxplot. Line represents median; hinges represents first and third quartile; whiskers represent furthest data point from quartile within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

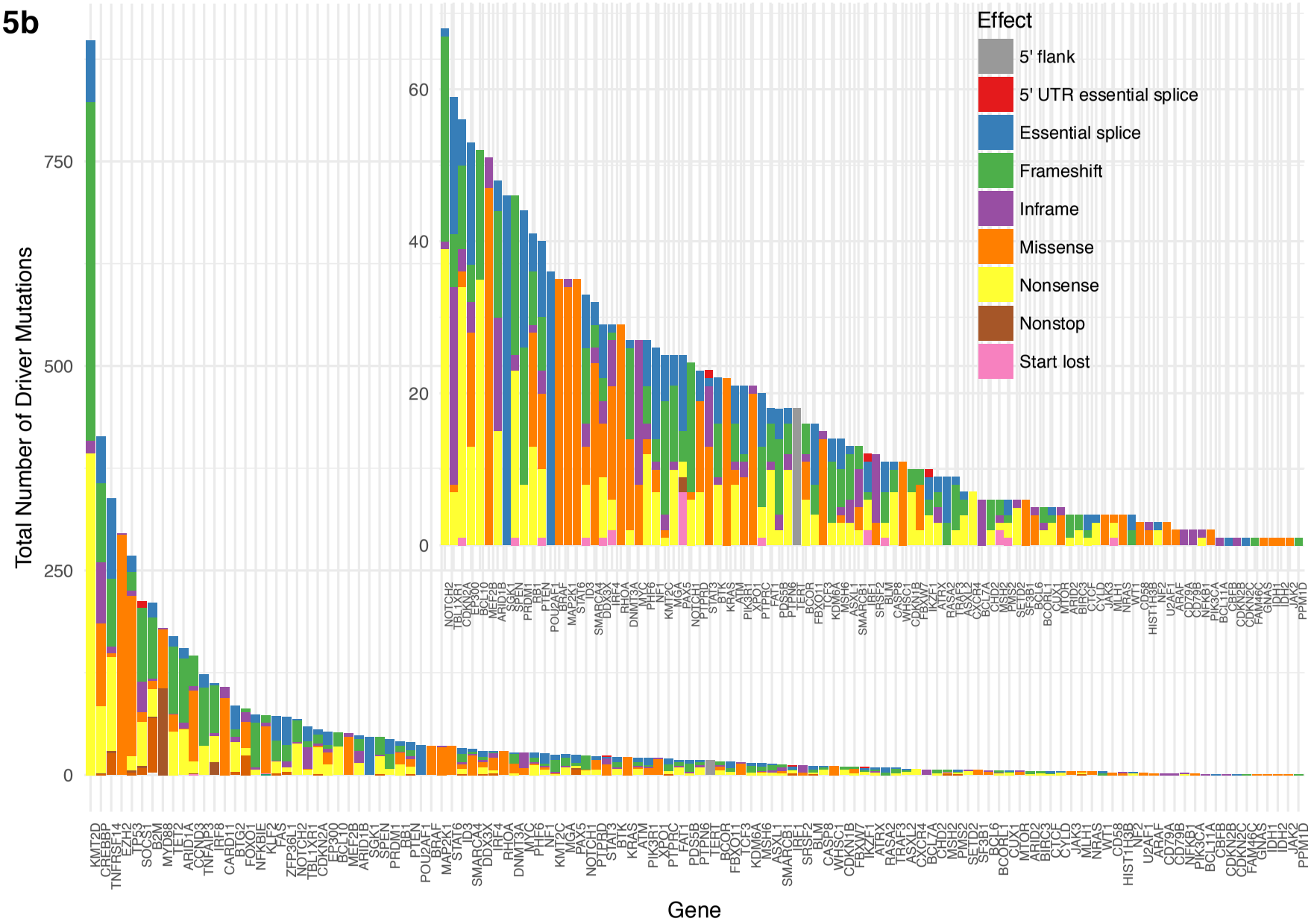
All Subtypes

5a



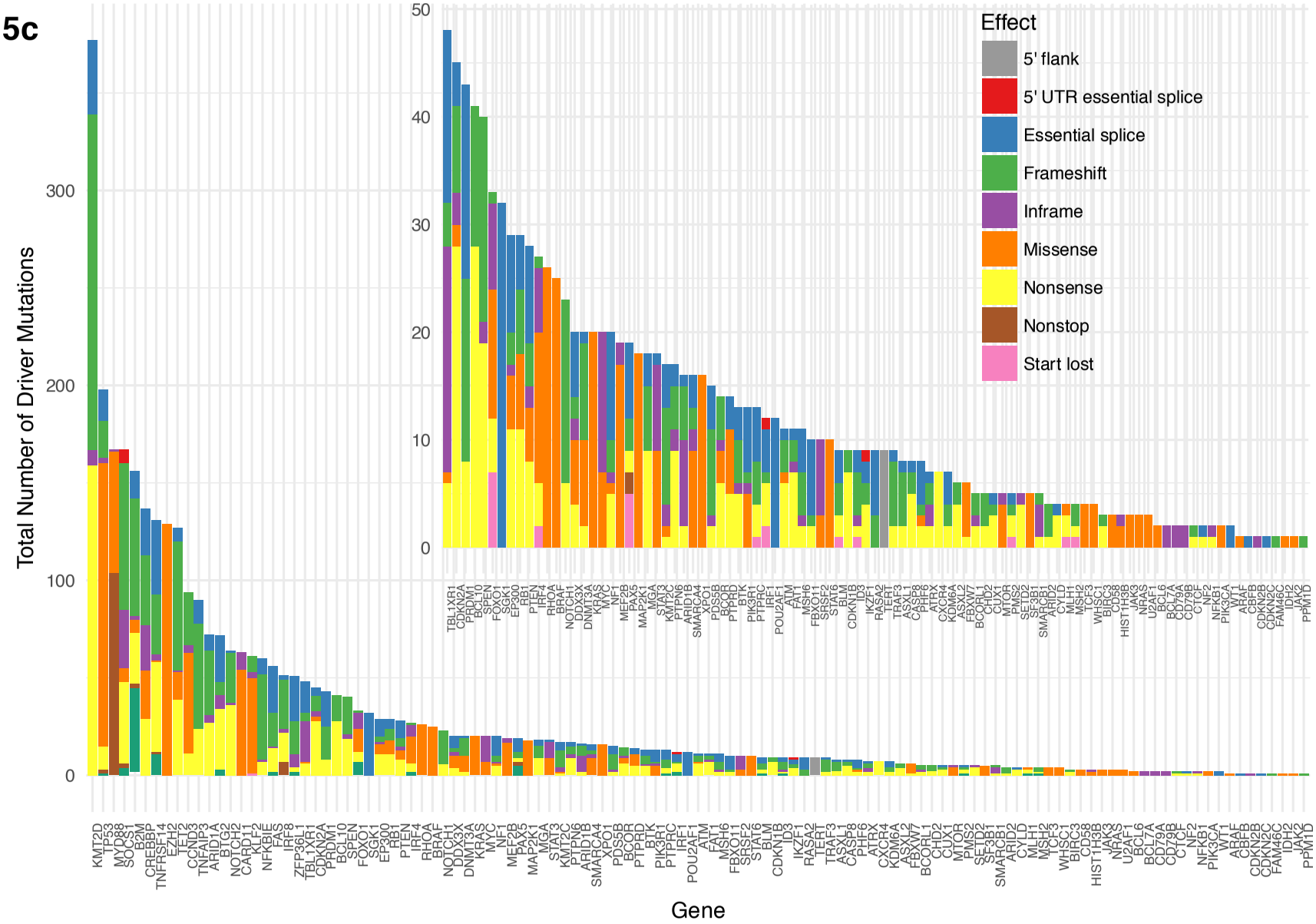
All Subtypes

5b



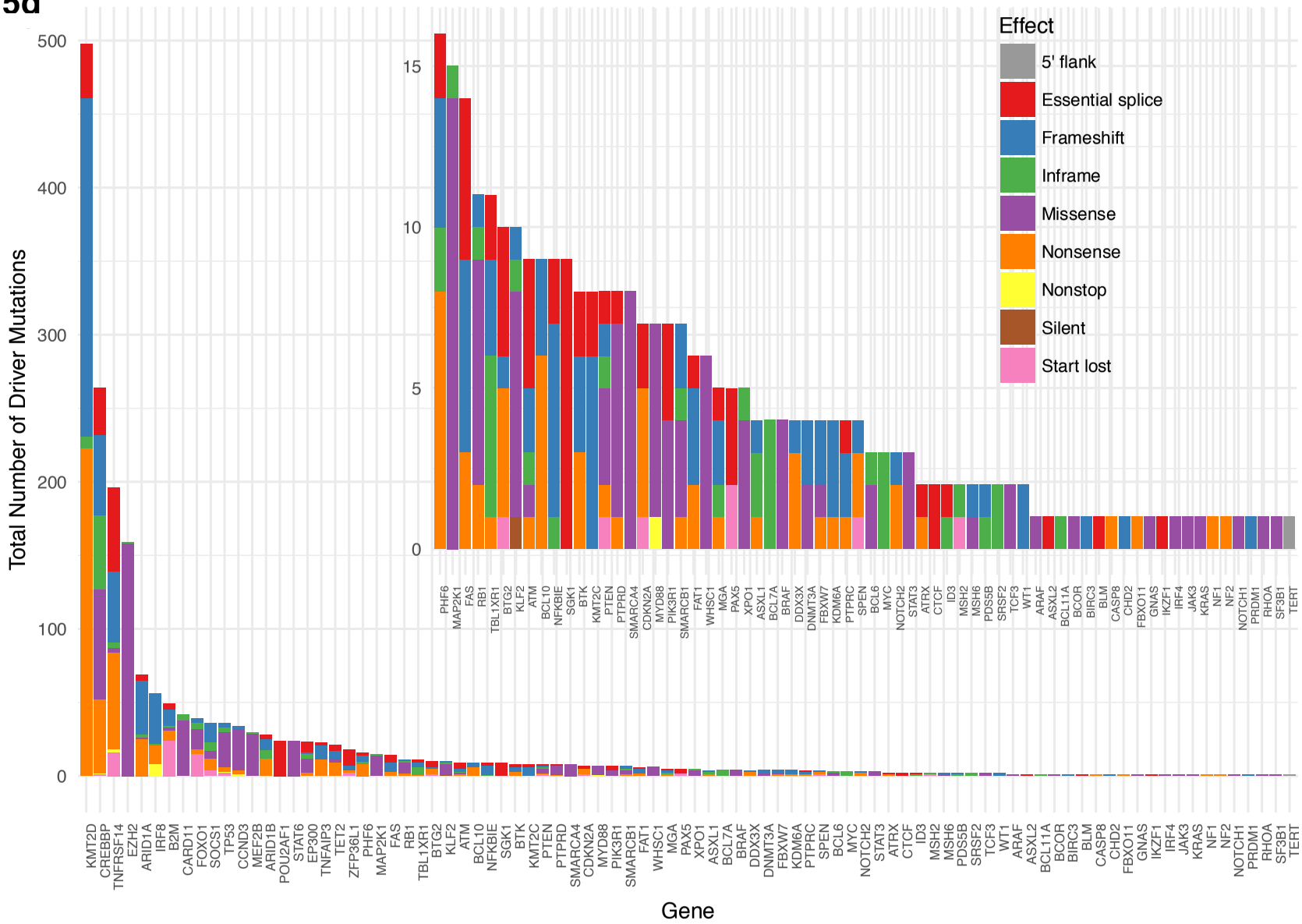
DLBCL NOS

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

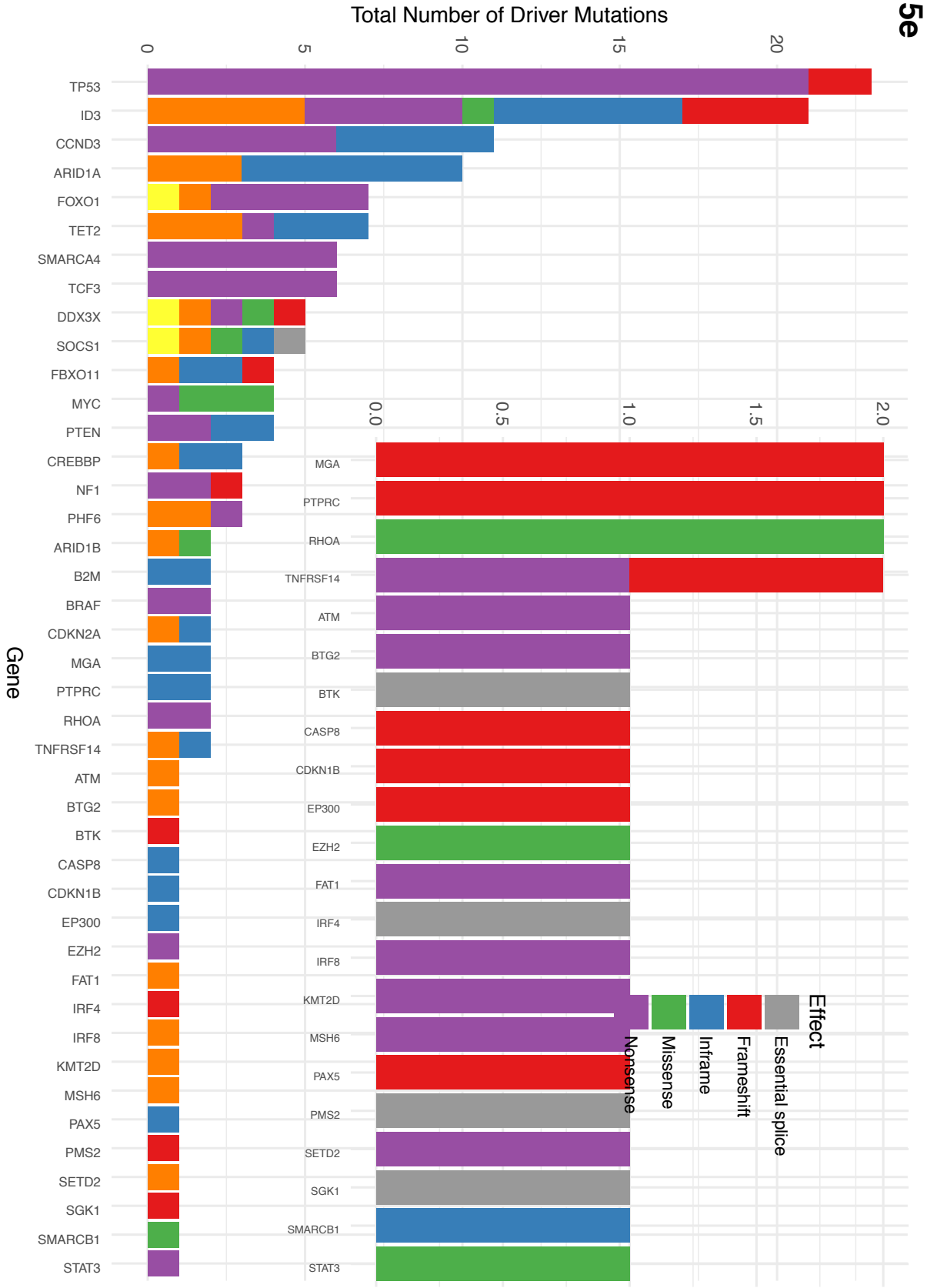

5e
BL
Total Number of Driver Mutations
0
5
10
15
20
Gene
TP53
ID3
CCND3
ARID1A
FOXO1
TET2
SMARCA4
TCF3
DDX3X
SOCS1
FBXO11
MYC
PTEN
CREBBP
NF1
PHF6
ARID1B
B2M
BRAF
CDKN2A
MGA
PTPRC
RHOA
TNFRSF14
ATM
BTG2
BTK
CASP8
CDKN1B
EP300
EZH2
FAT1
IRF4
IRF8
KMT2D
MSH6
PAX5
PMS2
SETD2
SGK1
SMARCB1
STAT3
0.0
0.5
1.0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

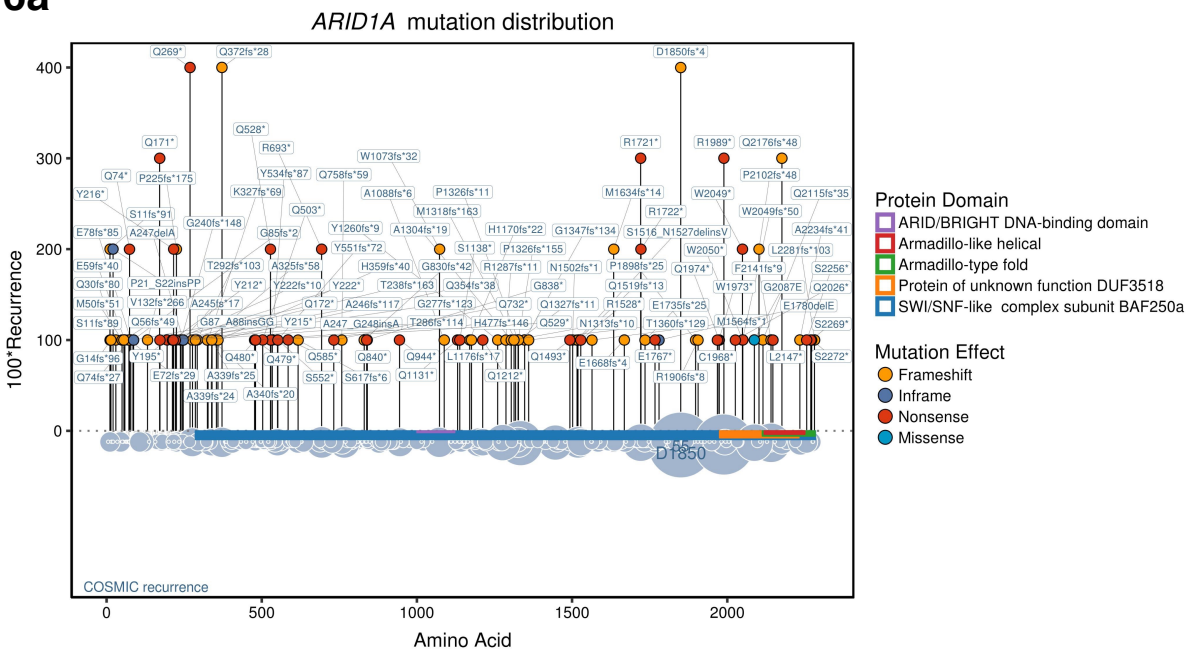
We identified the following tumour suppressor genes within in our cohort: *EP300, ARID1A, KTM2D, MGA, PTEN, PTPN6, PTPRC, PTPRD, RB1, TET2, TNFAIP3, ZFP36L1*. All have been previously characterized as tumour suppressor genes, either in lymphoma or in other cancer types. Therefore, our ability to reproduce 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

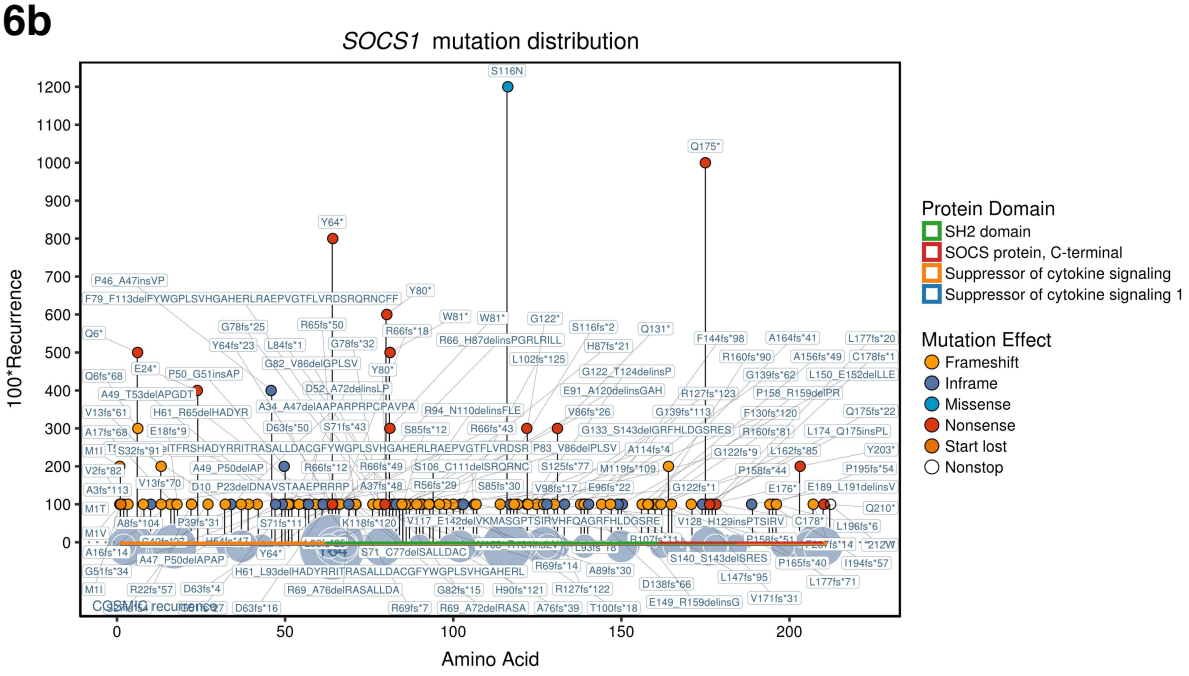
missense mutations in its helicase, superfamily 1/2, ATP-binding domain (T910, P913) and a recurrent missense mutation in its helicase, C-terminal domain (R1192). None had been previously reported in DLCBL although alternate mutations had been reported in small cell carcinoma of the ovary[152]. SMARCA4 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

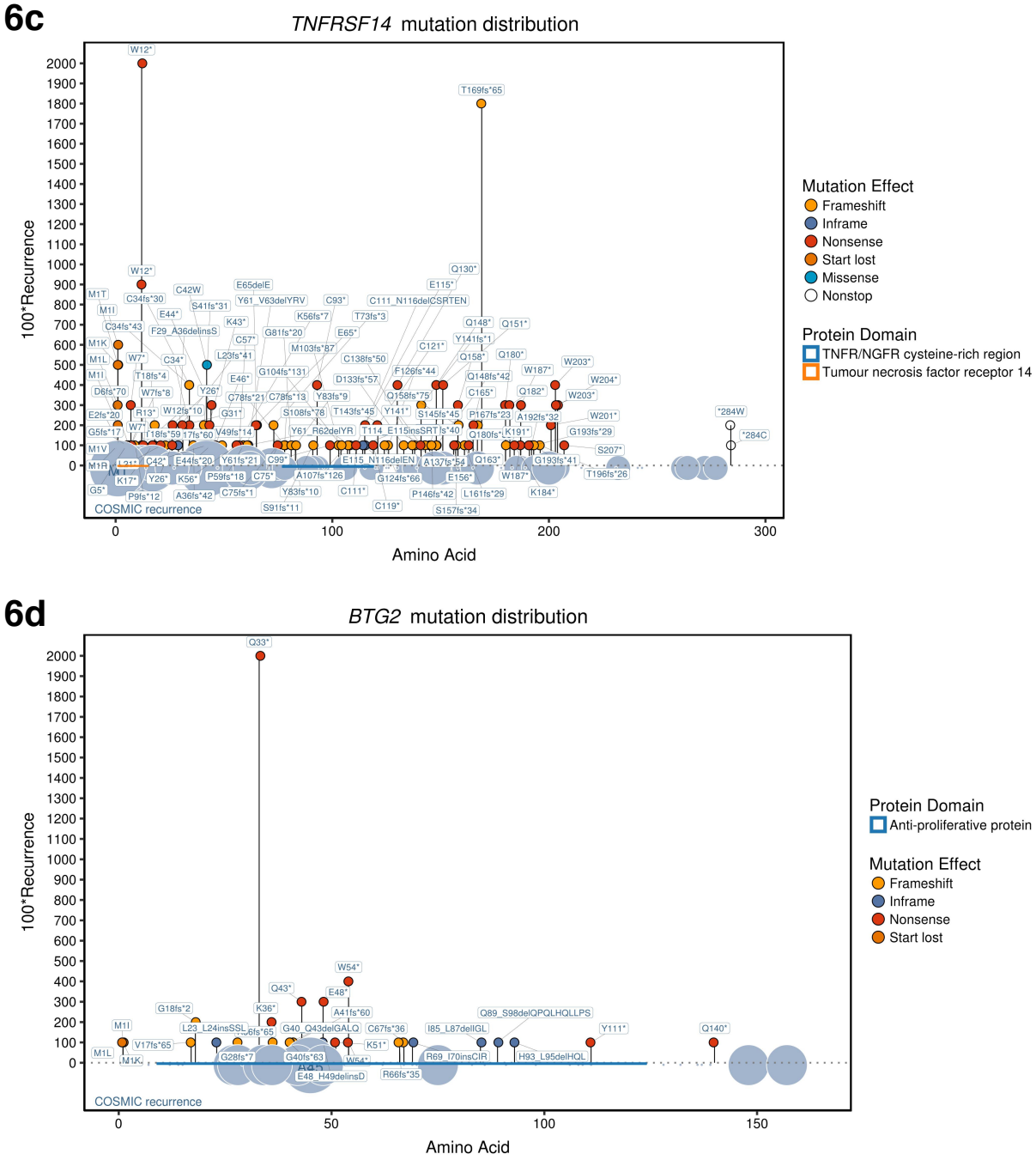


Figure 6 Gene-level analysis demonstrates tumour suppressor gene mutational profiles and reveals recurrent disruptive mutations. Each gene plot shows driver mutations found in the coding sequence, (2) protein domains from UniProtKB, and (3) bubbles. Bottom half of plots show bubbles sized according to 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.

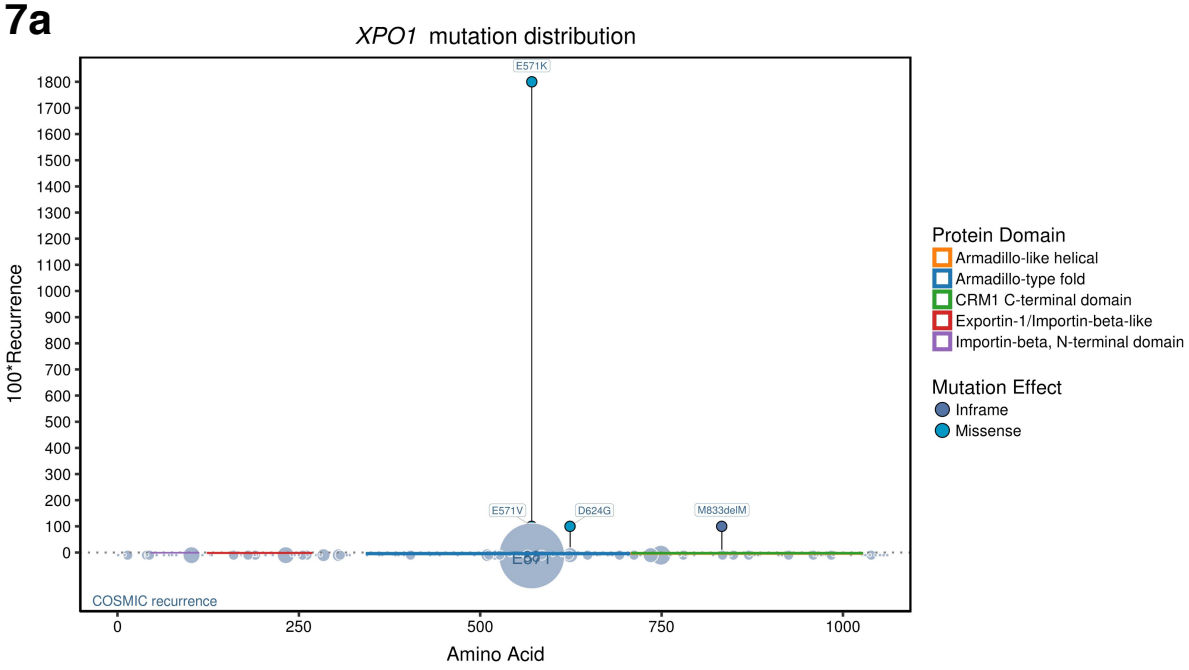
The second category of oncogenes exhibited genetic profiles that, in addition to recreating known hotspots, also revealed new hotspots (Figure 7b). First, the *CARD11* gene recreated known hotspots at D230, D357, D401, and L251[151] while also exhibiting a new mutation at Q249. The *CARD11* mutations 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]

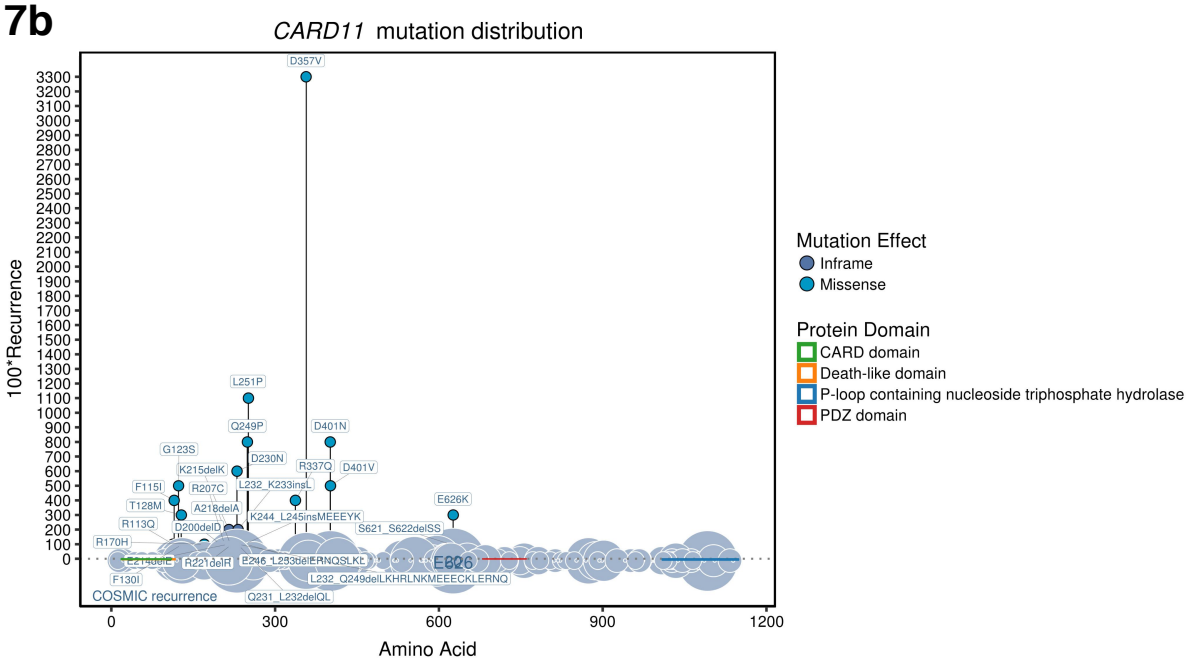
Finally, the third category of oncogenes exhibited genetic profiles that had previously been undescribed. One oncogene, *STAT3*, was present within this category (Figure 7c). STAT3 is a transcription factor, shown to be constitutively activated in many cancers, with a variety of downstream targets which 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

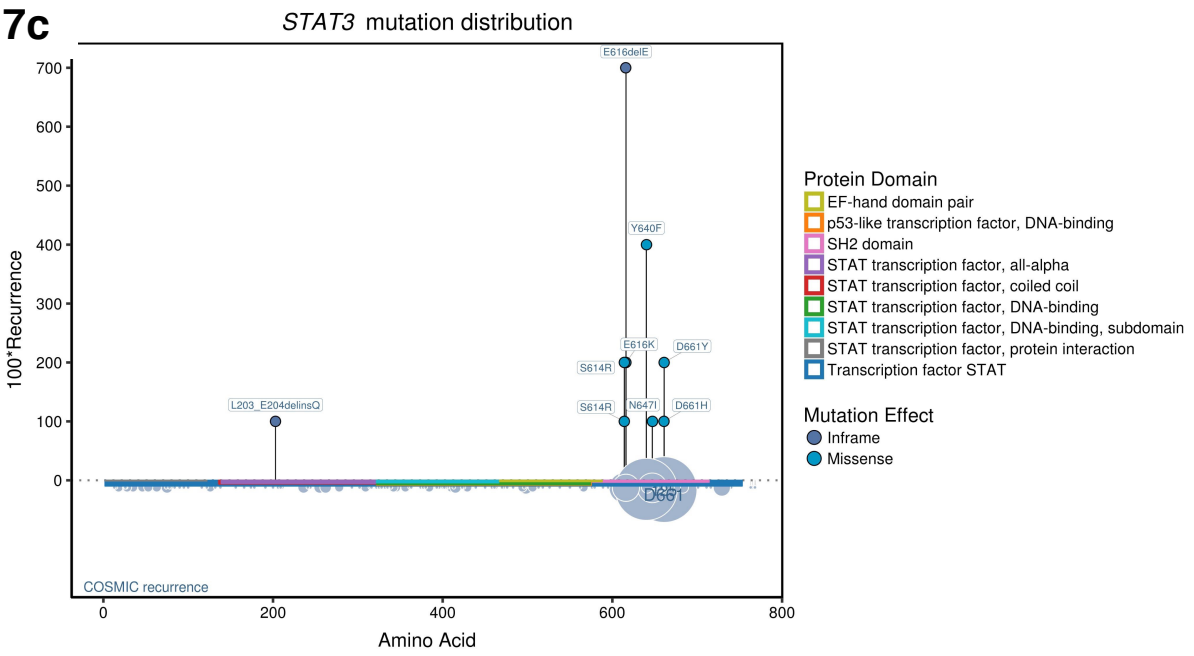


Figure 7 Gene-level analysis demonstrates known and novel oncogene hot spots. (a) Oncogenes exhibit missense hot spots. *XPO1* is shown as a representative example. **(b)** We additionally identified novel hotspots in known oncogenes. *CARD11* is shown as a representative example. **(c)** We created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

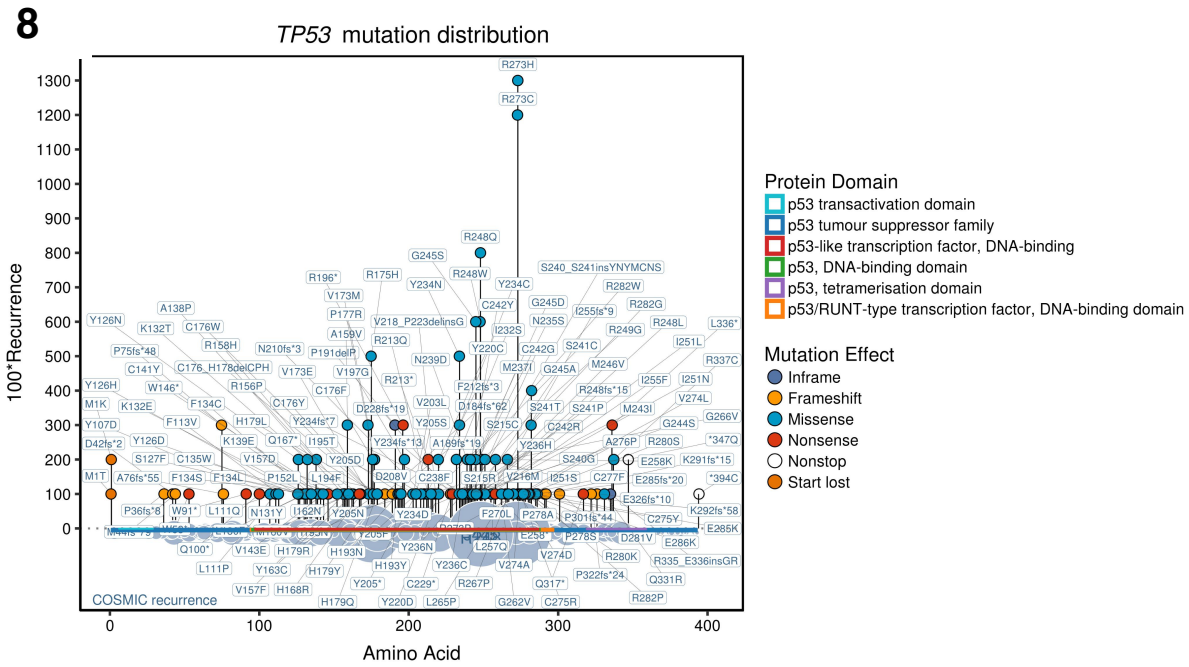


Figure 8 Gene-level analysis shows the potential for genes to serve as both tumour suppressors and oncogenes. *TP53* is shown as a representative example.

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.

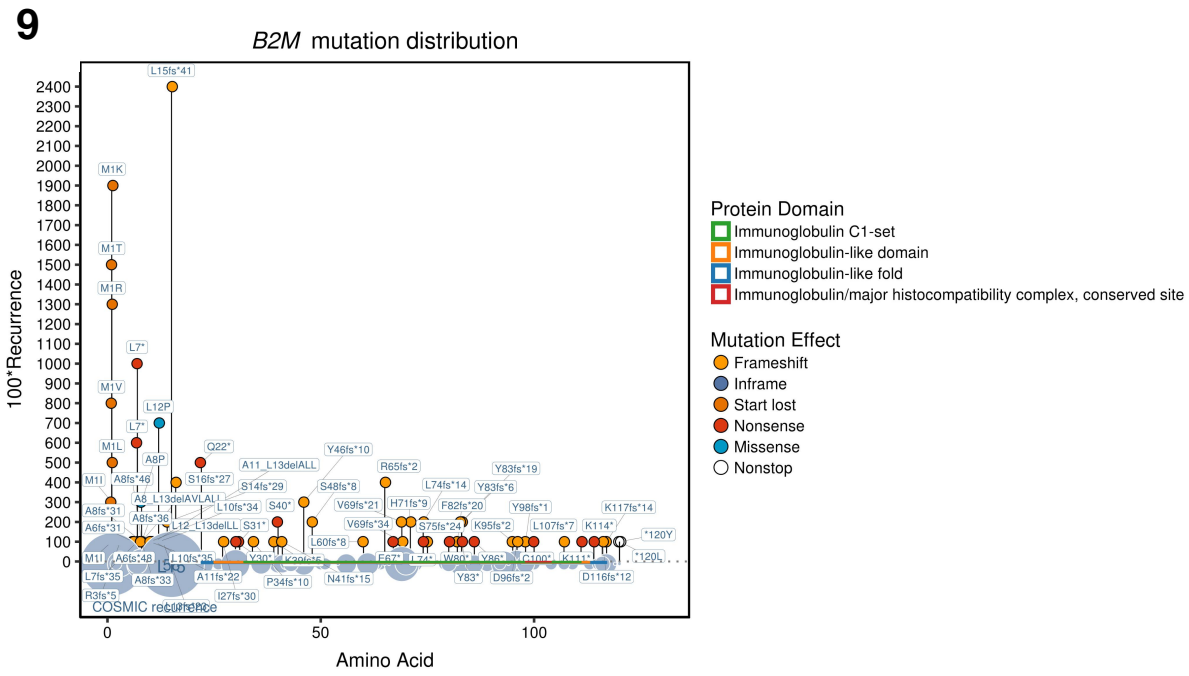


Figure 9 Gene-level analysis shows patterns of aberrant somatic hypermutation. *B2M* is shown as a representative example.

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.

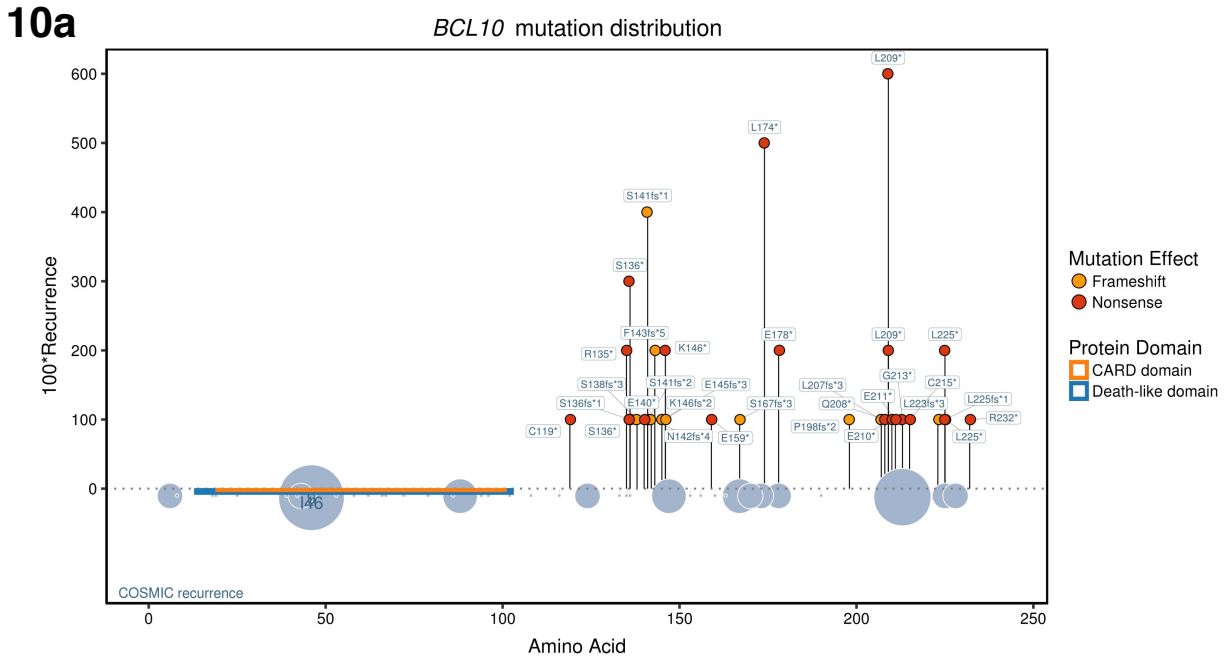


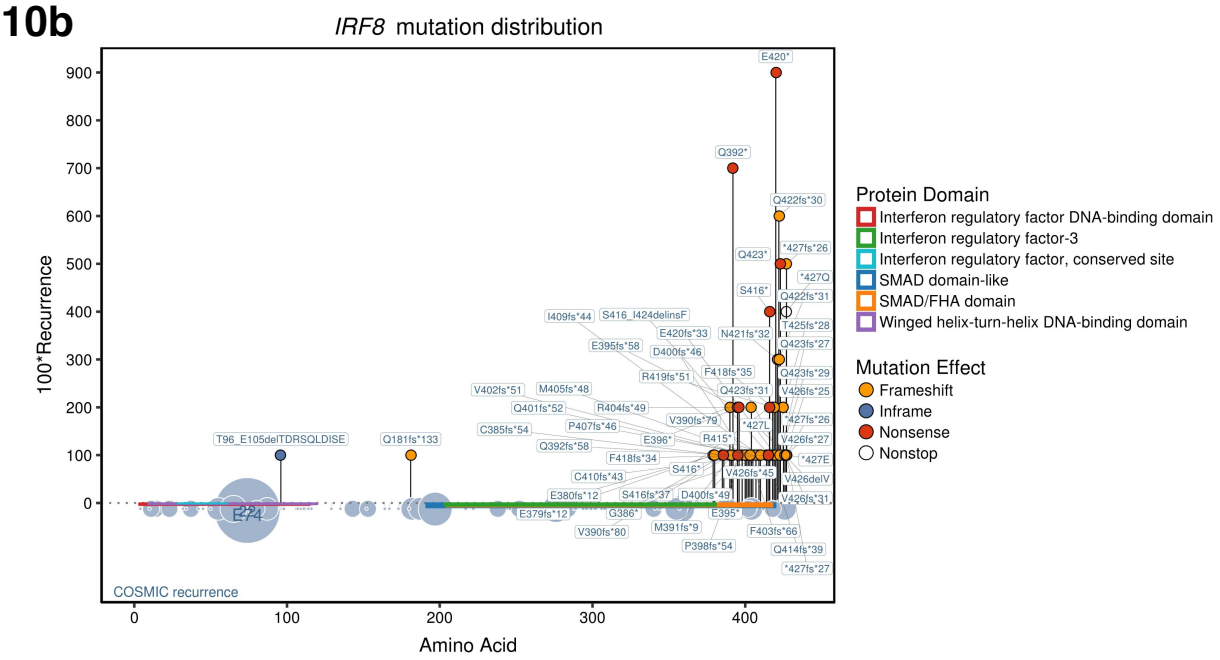
Figure 10 Gene-level analysis reveals disrupting mutations clustered in highly specific domains. (a) *BCL10*, **(b)** *IRF8*, **(c)** *FAS*, **(d)** *ARID1B*, **(e)** *NOTCH1*, **(f)** *NOTCH2*, **(g)** *KLF2*, **(h)** *TCF3*, **(i)** *SMARCB1*.

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.



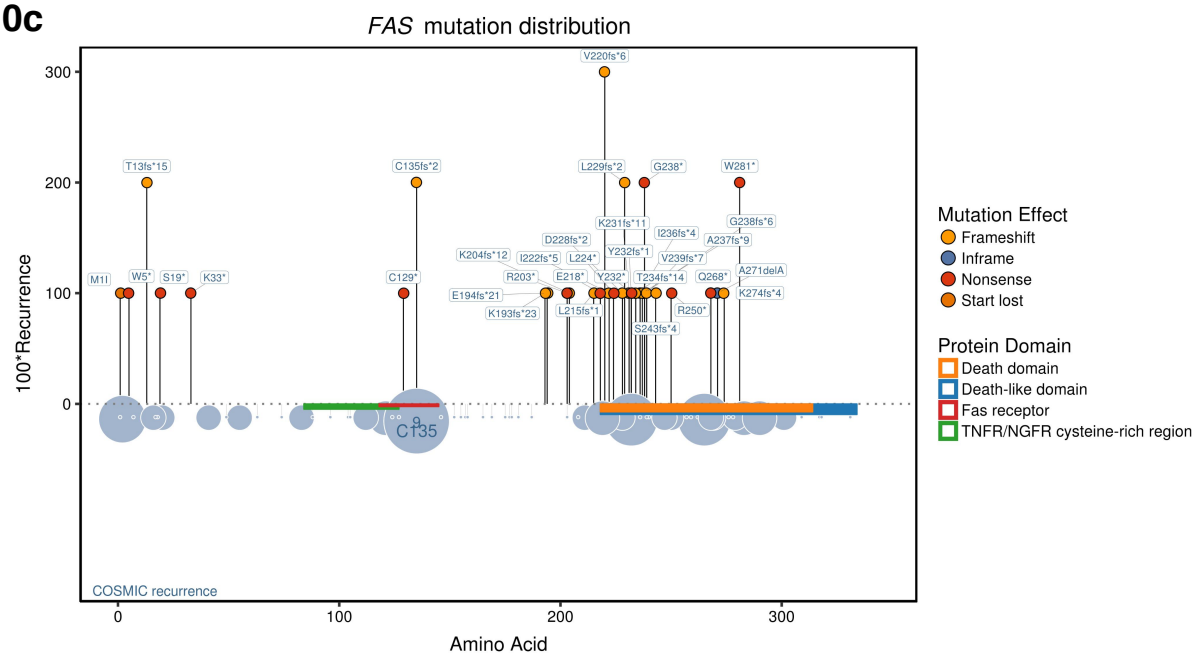
4.3.2.2.3. *FAS*

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

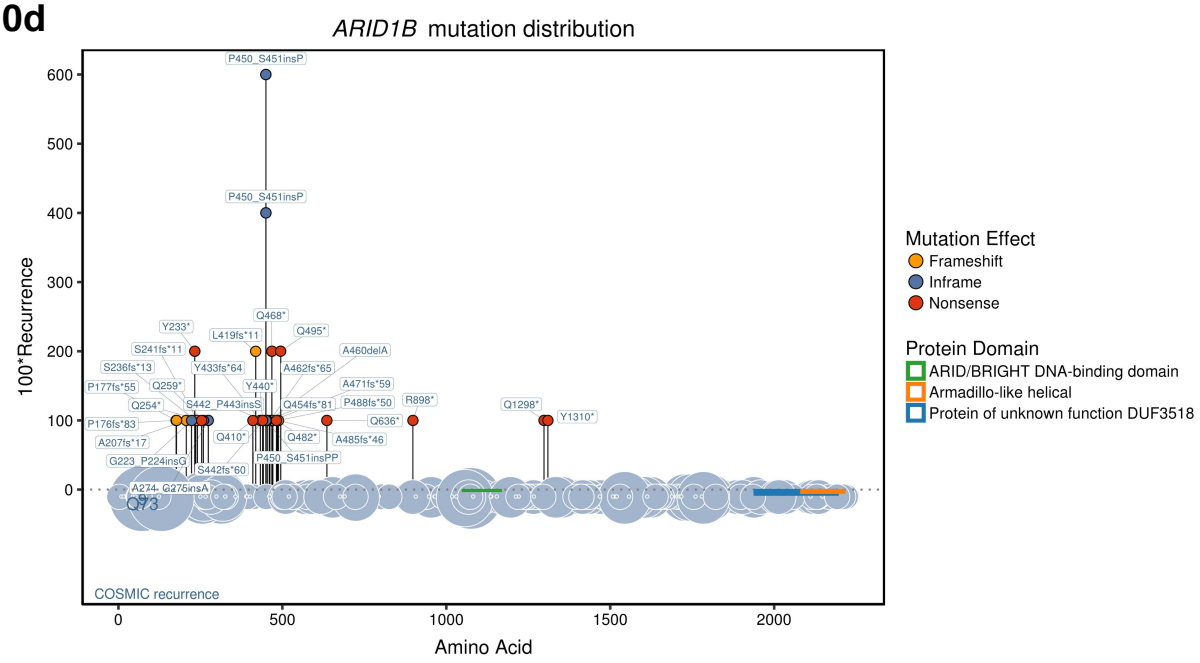


4.3.2.2.4. *ARID1B*

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.

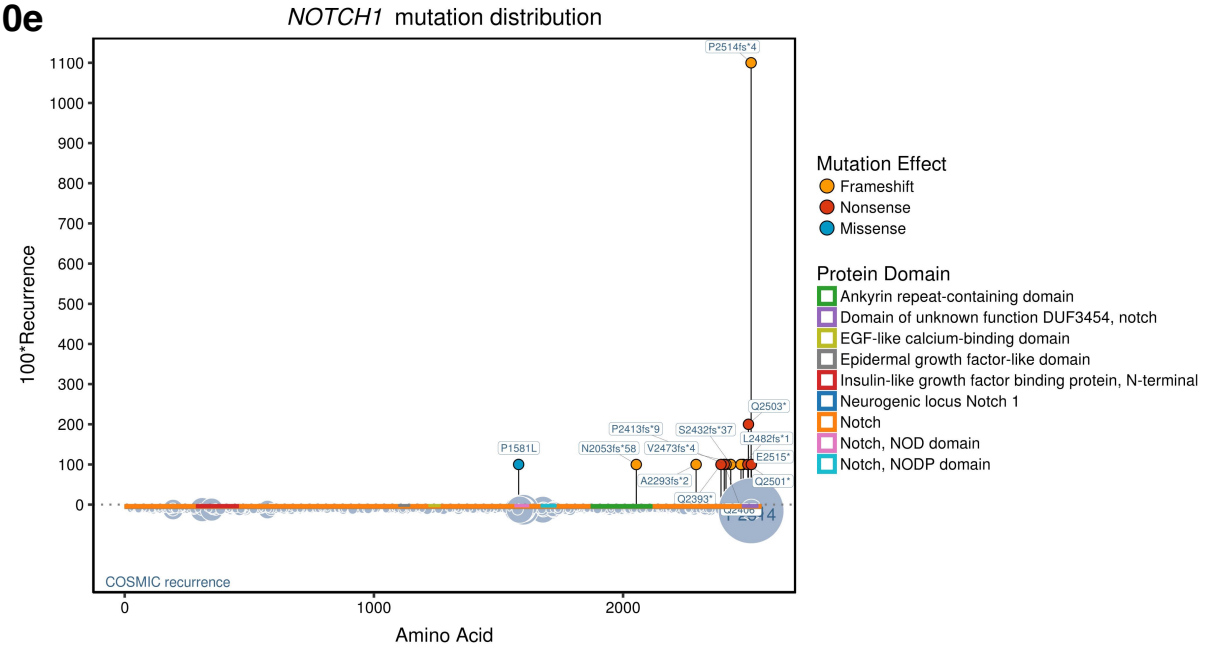
10d



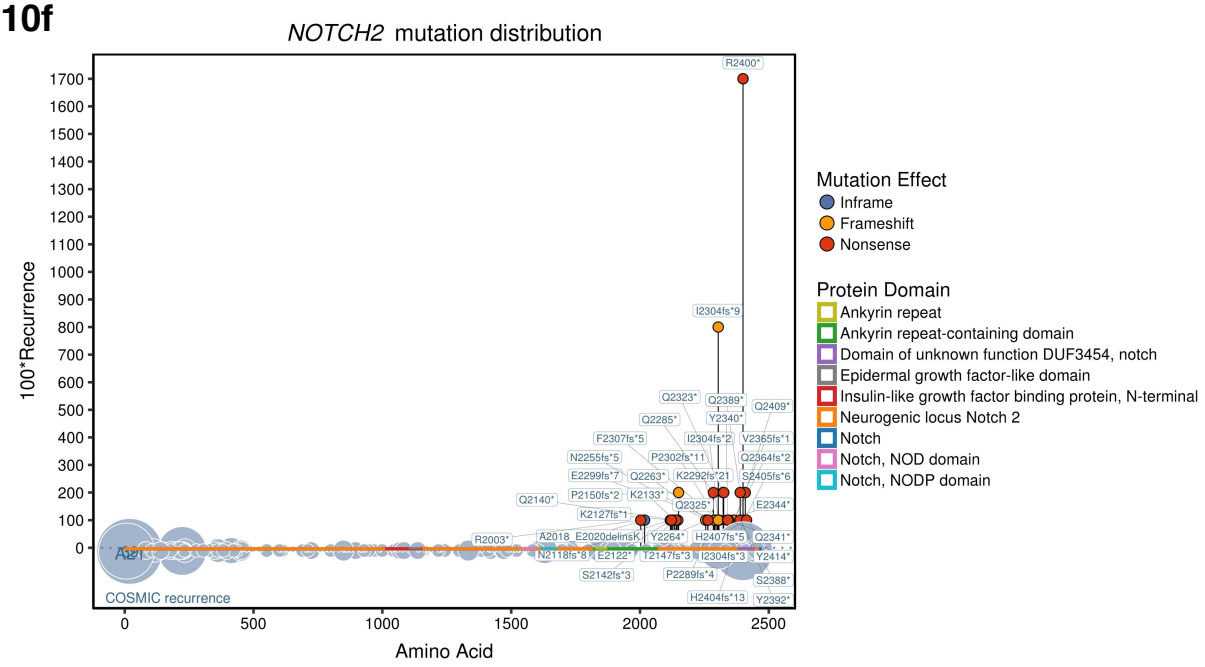
4.3.2.2.5. *NOTCH1/NOTCH2*

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

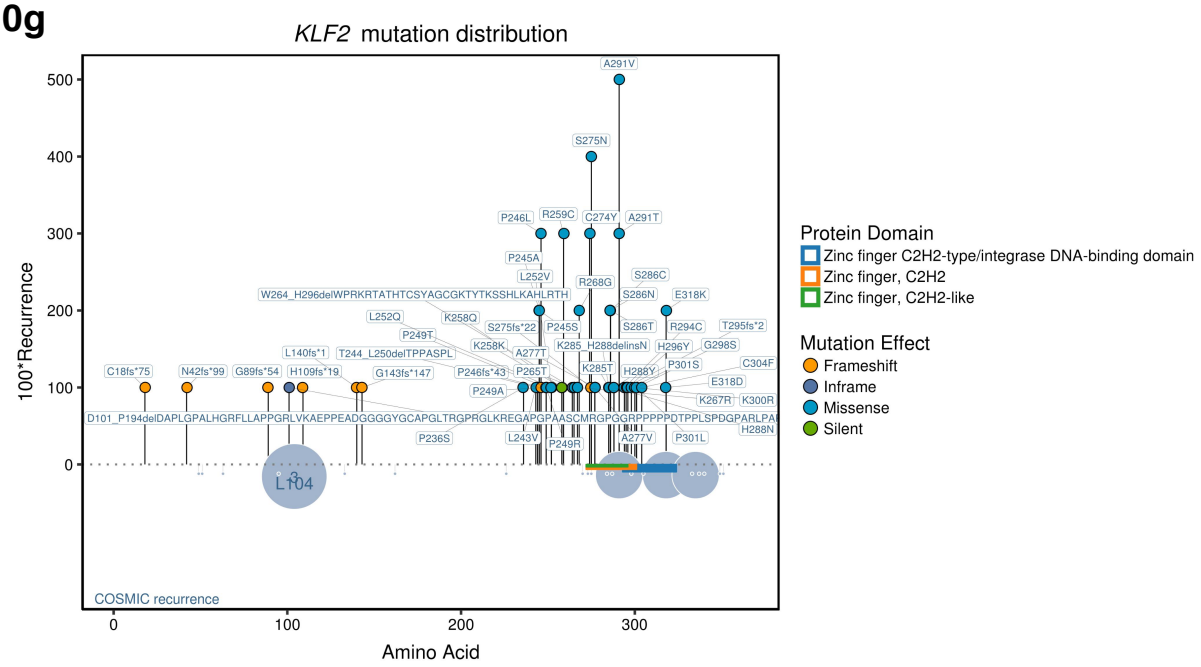


4.3.2.2.6. *KLF2*

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.

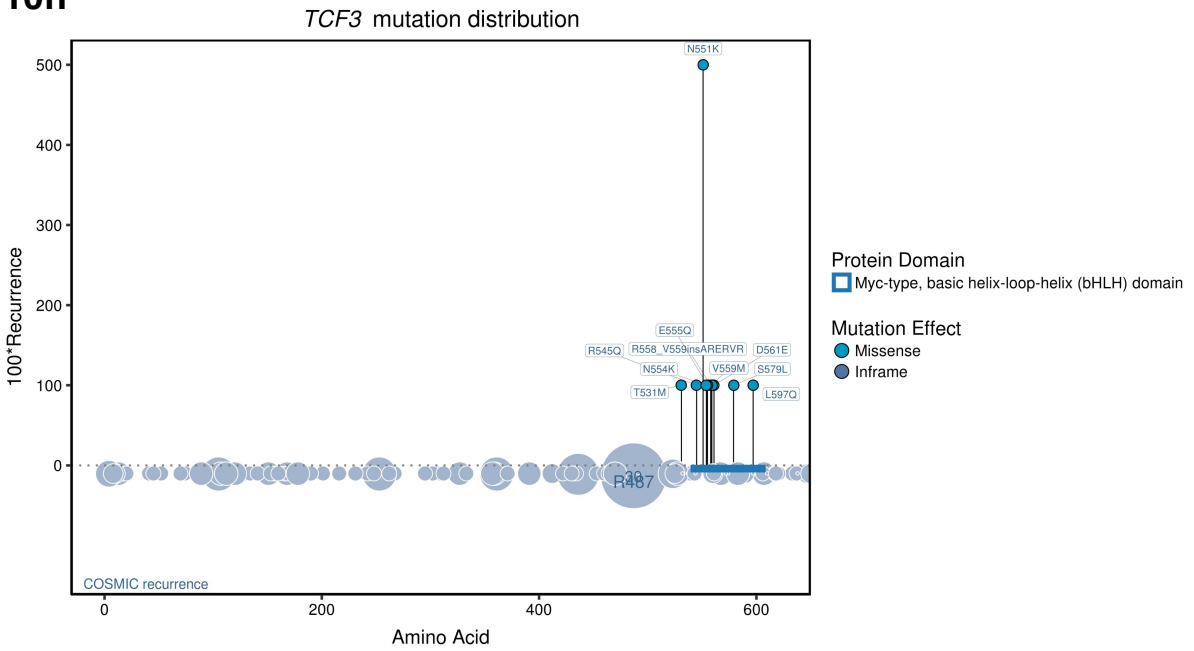
10g



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].

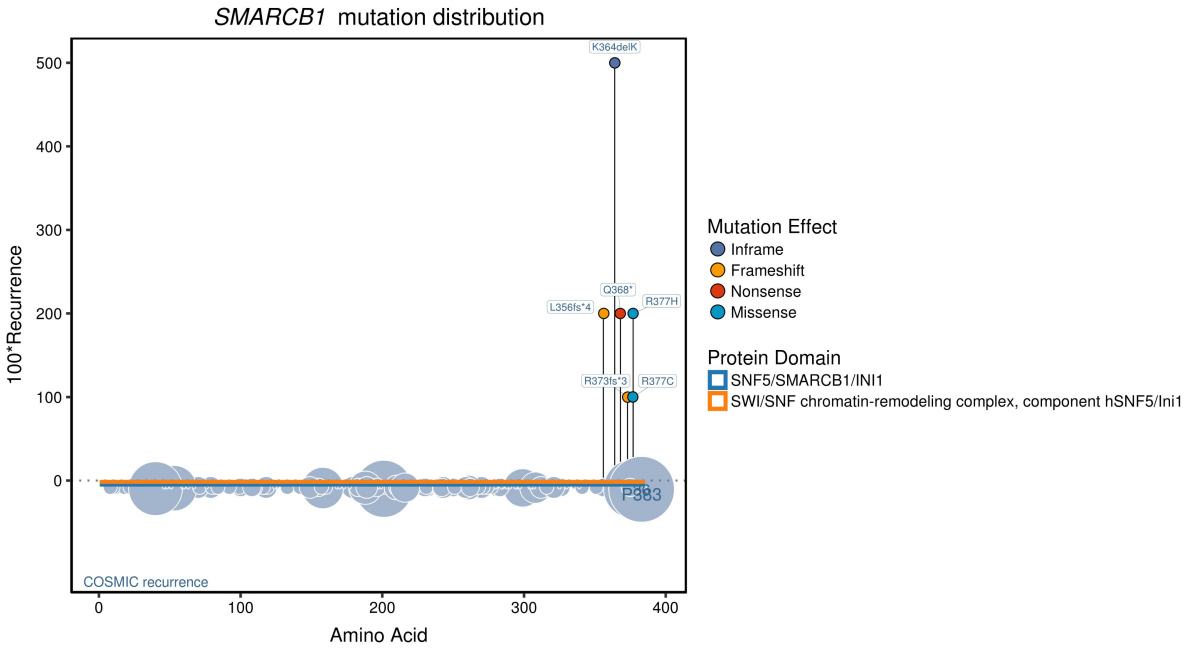
10h



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.

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.

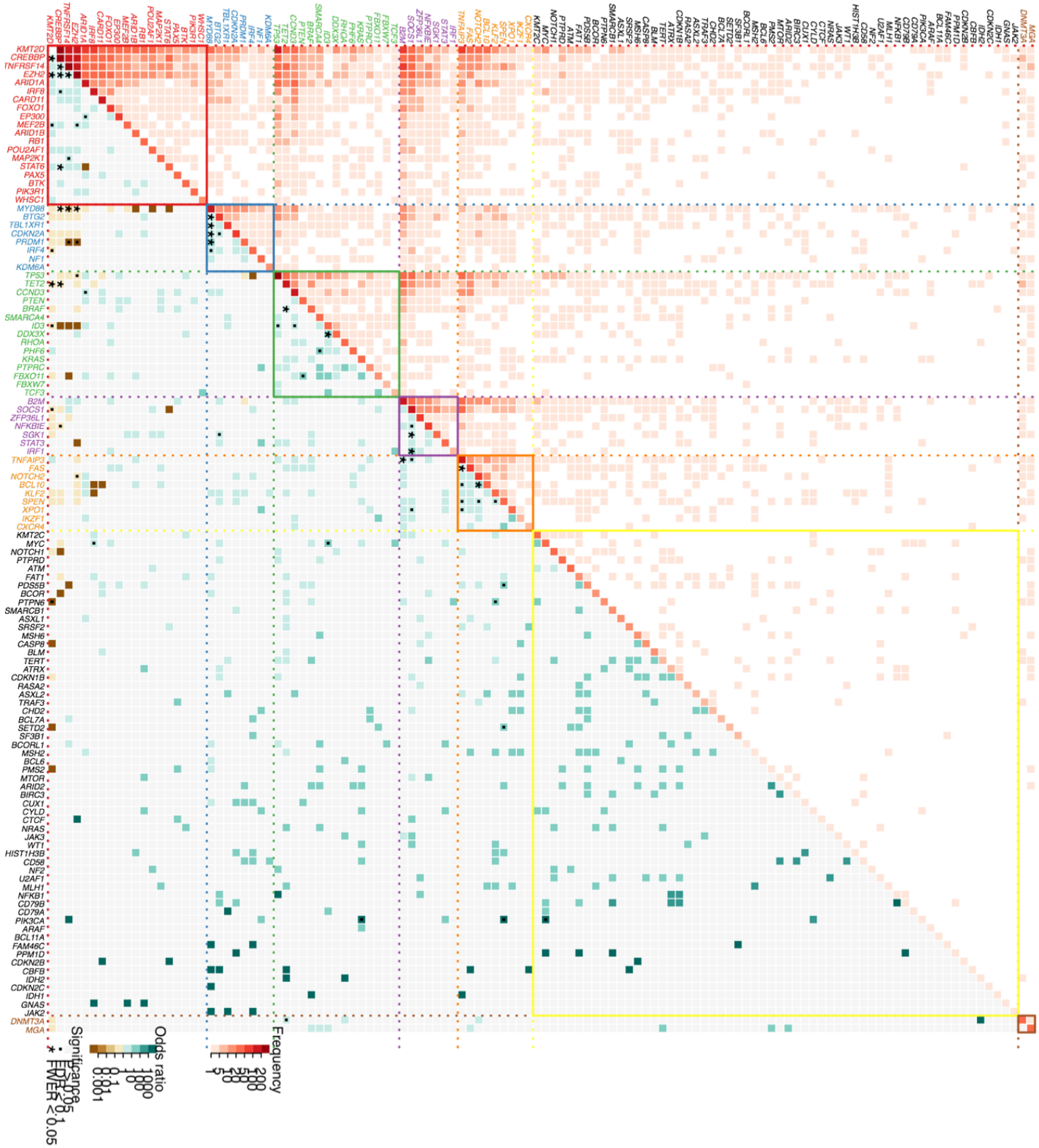


Figure 11 Co-mutation and mutual exclusivity patterns generate eight distinct classes in FL, BL, and DLBCL. Lower triangle depicts pairwise association between lesions in genetic classes. The colour of each tile corresponds to the odds ratio for each pair, with brown representing mutual exclusivity and 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.

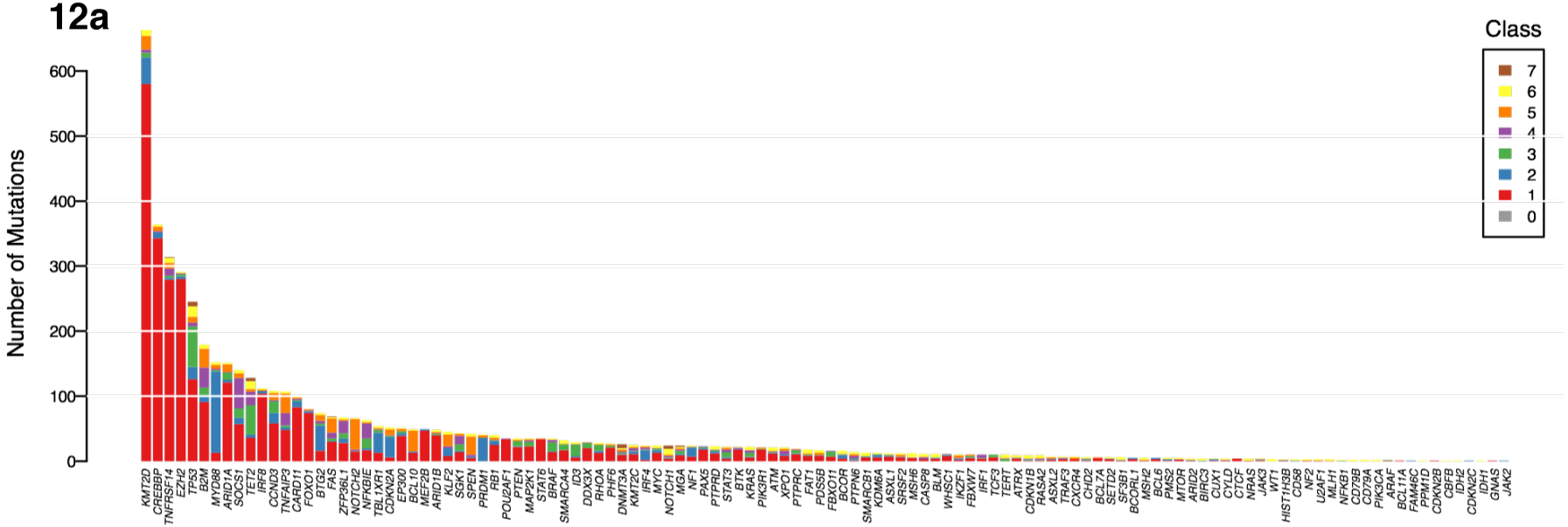
Class 6 samples came from BL, DLBCL, and FL lymphoma subtypes. We suspect these samples, in practice, reflect a variety of rare mechanisms that can cause the pathogenesis of each disease. Importantly, the distinct genome profiles of Class 6 DLBCL and Class 6 BL patients compared to DLCBL patients in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

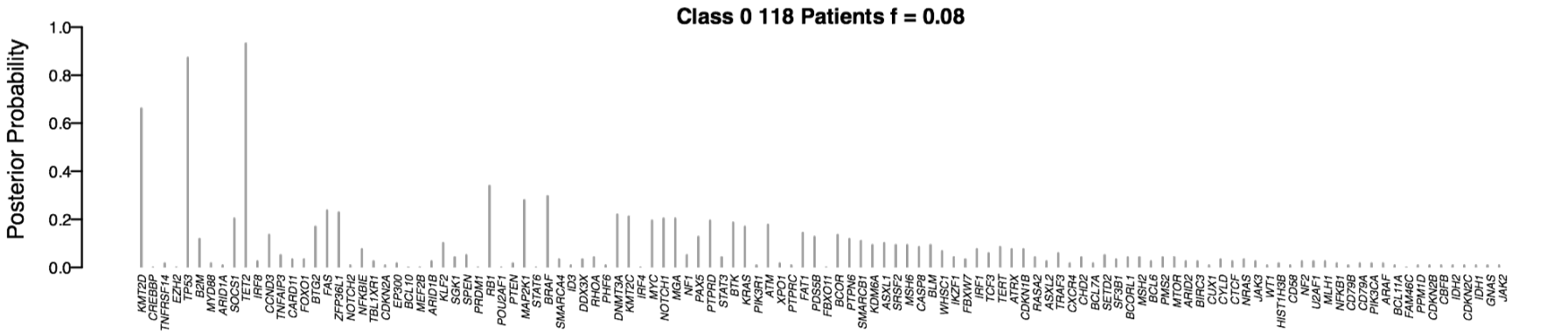
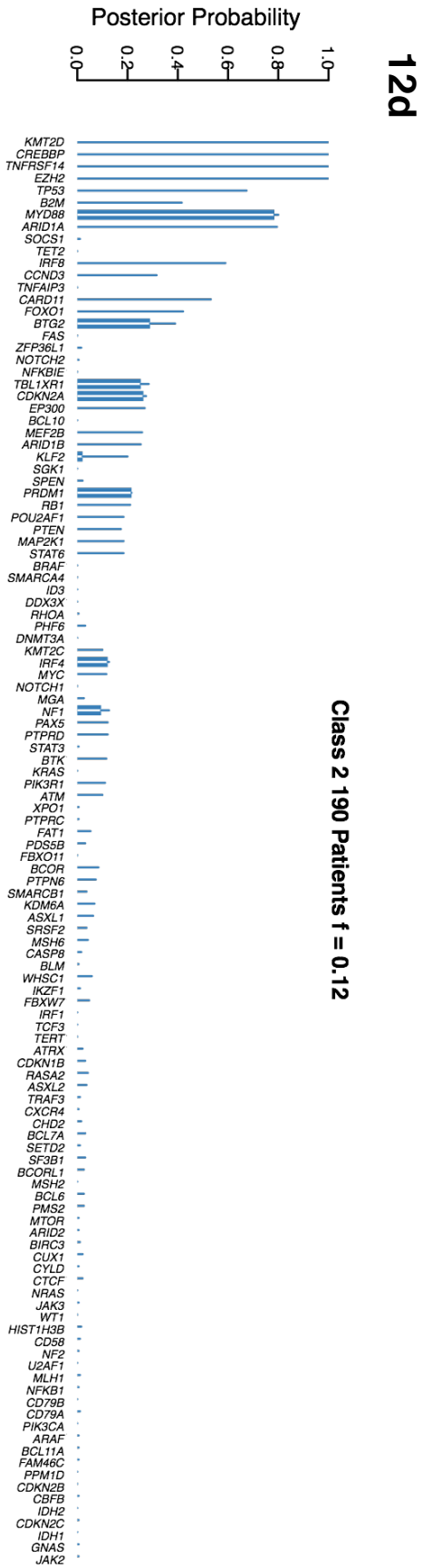
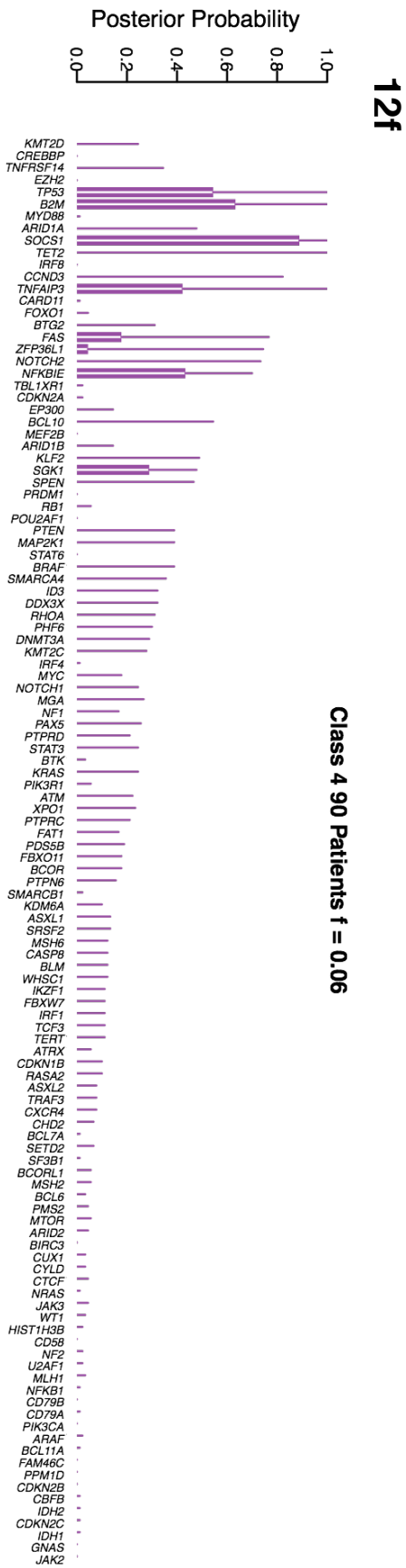
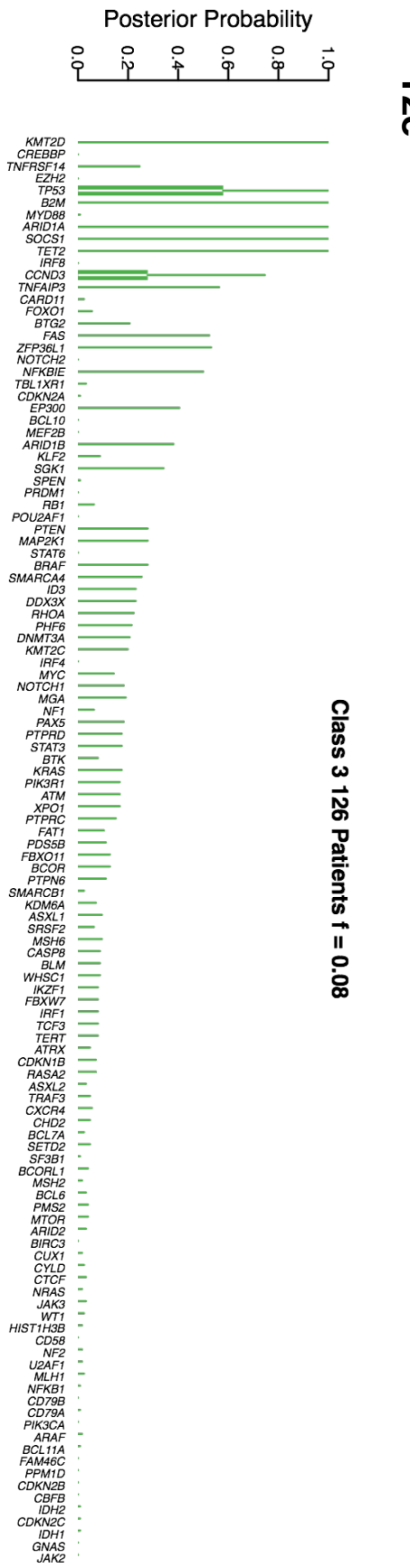
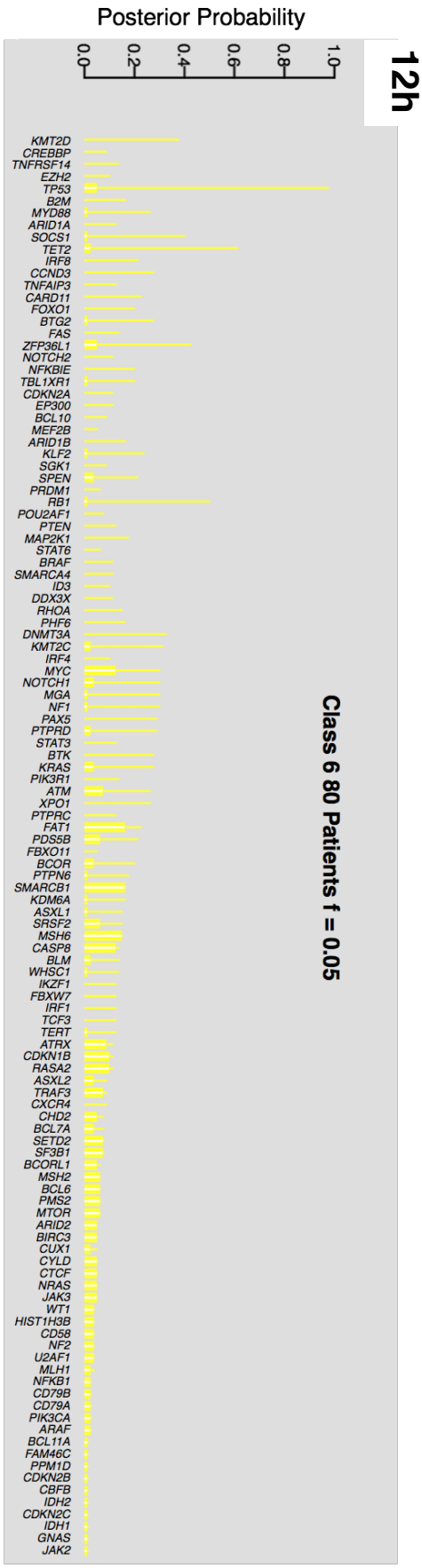

Class 0 118 Patients f = 0.08
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
KMT2D
CREBBP
TNFRSF14
EZH2
TP53
B2M
MYD88
ARID1A
SOCS1
TET2
IRF8
CCND3
TNFAIP3
CARD11
FOXO1
BTG2
FAS
ZFP36L1
NOTCH2
NFKBIE
TBL1XR1
CDKN2A
EP300
BCL10
MEF2B
ARID1B
KLF2
SGK1
SPEN
PRDM1
RB1
POU2AF1
PTEN
MAP2K1
STAT6
BRAF
SMARCA4
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

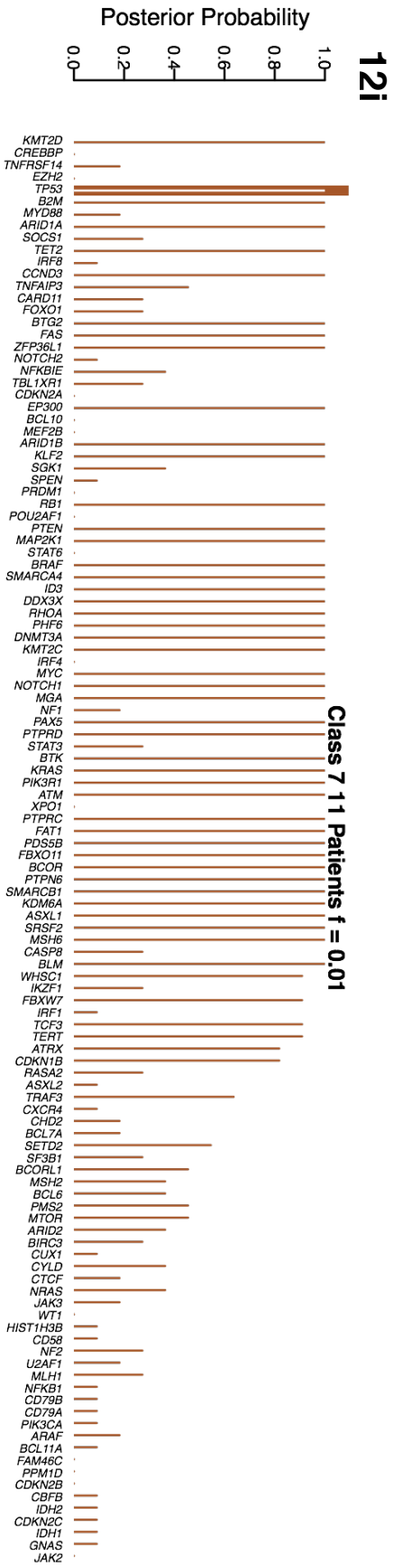
Figure 12 Each class shows a distinct mutational signature profile. (a) Number of driver mutations across all classes, coloured by proposed class assignment for patient with that mutation. **(b-i)** Mutational signature of each class. Numbers next to class show number and fraction of patients assigned 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.

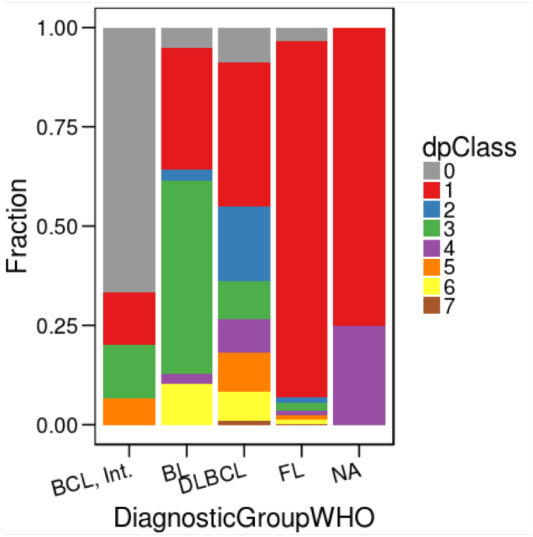
5.5. Preliminary Survival Analysis

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

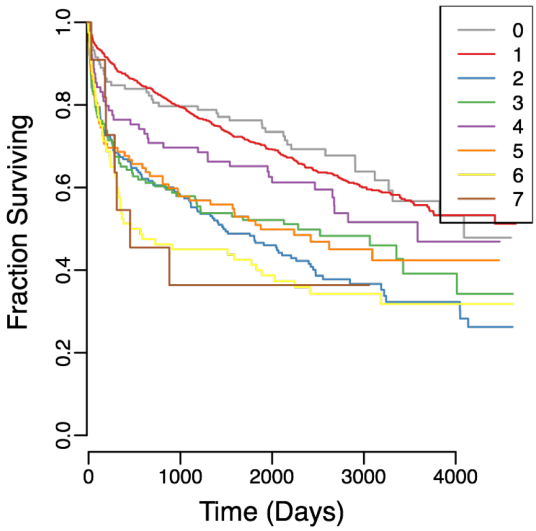


Figure 13 Classes show distinct subtype compositions and survival outlooks. (a, b) Patient assignment to WHO diagnostic groups or subtypes compared to patient assignment to proposed classes. **(c)** Kaplan-Meier plot for proposed classes.

6. Discussion

Here, we have provided the largest sequencing study on B-NHLs to date, proposing a novel genetics-based classification and profiling the mutational landscapes of FL, BL, and DLBCL with greater resolution than previously described.

6.1. Genomic Landscape and Gene Level Analysis

Our genomic landscape analysis for DLBCL NOS, FL, and BL was largely consistent with literature expectations but provided additional resolution due to the size and depth of our study. DLBCL NOS, FL, and BL all exhibited classic long tail distributions although DLBCL NOS in particular showed the greatest heterogeneity: the most recurrently mutated genes in DLBCL NOS accounted for a lower fraction of the overall mutations than those in FL and BL. Such a result was consistent with our later classification finding in which Class 5 contained 85 distinguishing genes all rarely mutated, indicating high heterogeneity. Because of the scope of our study, we also identified a variety of novel driver mutations, some rare, occurring across the 292 genes in our study.

Additionally, our landscape analysis found a small number of genes that showed a high mutation frequency across DLBCL NOS, FL, and BL (i.e. *KMT2D, CREBBP, TNFRSF14, TP53, SOCS1, B2M, ARID1A, CCND3, TNFAIP3, IRF8*). These mutational similarities initially pointed to the need for similar pathway dysregulations for B-NHLs to progress. By contrast, however, the only gene that was commonly mutated across all classes in our classification analysis was *TP53*. The difference in these results demonstrates that genetic classification can more accurately distinguish classes than histology; and importantly, can resolve pathway differences that demarcate patients into classes that have consistent pathway mutations that are largely non-overlapping.

Our mutation analysis demonstrated that patients, regardless of B-NHL condition, generally have 3-4 driver mutations. This insight, combined with the later classification description of co-mutation within classes, shows that multiple pathways tend to be dysregulated within B-NHLs and DLBCL. As a result, oncogenic cooperation may be occurring to, for example, increase proliferation while also evading the immune system. The presence of multiple driver mutations increases the complexity of pathogenesis and also classification. Rather than single genes demarcating novel classes, combinations of genetic mutations distinguish patients. As a result, far more possibilities exist and heterogeneity similarly increases.

At the gene level, we found genes broadly falling into oncogenic and tumour suppressor mutation profiles as expected and identified the presence of expected mutational processes such as aberrant somatic hypermutation. More interestingly, we identified clusters of disrupting mutations in specific gene domains that we suspect caused gains in function and thus allowed oncogenic activity. The specific mechanisms and mutations had not, to our knowledge, been previously reported for B-NHLs. For example, we observed a high proportion of frameshift and missense mutations in the death domain of the *FAS* gene, which generally initiates a caspase cascade leading to apoptosis. We suspect the inactivation of the *FAS* domain improves tumour cell survival. Similarly, we found a high number of frameshift and nonsense mutations in the SMAD/FHA domain of *IRF8* which we suspect could cause a gain in function that prevents apoptosis. In *SGK1*, we found a series of essential splice site mutations affecting a single exon, causing a likely gain in function and flagging that exon's importance in *SGK1* regulation. None of the above mechanisms, to our knowledge, had been previously reported in the context of DLBCL or B-NHLs.

6.2. Classification

Our classification system resolved seven distinct categories of B-NHLs, successfully separating FL, BL, and DLBCL while simultaneously highlighting the inherent heterogeneity of DLBCL. Compared to the WHO classification, we demonstrated significant heterogeneity and potential for further resolution within given subtypes. Indeed, patients marked as DLBCL NOS patients by the WHO classification were present in all seven classes identified here, indicating the necessity for further resolution.

We cannot directly compare our work to the cell of origin classification due to the absence of gene expression data from our dataset, however, Class 2 shared genetic characteristics largely consistent with ABC-DLBCL. The future addition of gene expression data to our study will allow us to directly compare our classification with the cell of origin classification. Crucially, we will be able to answer whether or not cell of origin can be distinguished on the basis of genetic mutations alone. If so, our approach could become an important surrogate for gene expression profiling as a way of determining cell of origin, which has already shown clinical relevance with the ABC-DLBCL group responding differently to targeted treatments than the GCB-DLBCL group.

Overall, DLBCL shows a high heterogeneity compared to other cancers. Unlike similar genetic classification schemes, such as that for AML, DLBCL presented a category with a larger number of rarely mutated genes (Class 6). The separation of these rarely

mutated genes into their own class rather than their presence within other classes points to the increased heterogeneity of DLBCL compared to other cancers. Indeed, the large number of potential driver lesions that can cause cancer within this category point to the potential for pathogenesis in a variety of different ways. Each likely follow distinct mechanisms and effective resolution of this class would require substantially higher sample sizes in order to create additional subcategories. Such heterogeneity reinforces the distinct clinical responses to treatments and the need for classification to resolve such differences.

Our classification approach additionally demonstrated its ability to resolve patients who had likely transformed. The first example was the identification of Class 1 patients, a class with hallmark mutations for FL, that were diagnosed by our clinicians as having DLBCL. Since the transformation of FL into DLBCL is well documented, such a result was expected and consistent with the literature. More surprising, however, was the fact that Class 5, consisting primarily of DLBCL and BCL, Int. patients demonstrated hallmark mutations of SMZL, likely corresponding to patients that had transformed from SMZL. Crucially, only a genetic classification approach of this sort – not histology alone – could identify the root disease from which DLBCL had transformed. Biologically, our result reinforces the possibility of SMZL to transform into DLBCL, which had been previously reported but rarely[185]. Clinically, it could suggest that Class 5 Patients have a distinct pathogenesis and thus may respond differently to novel treatments compared to other DLBCL subtypes.

Overall for aggressive diseases such as DLBCL which often transform from indolent cancers, the ability to distinguish the original genetic mutations that led to cancer could substantially affect patient outcomes. We expect our approach, therefore, to generalize across other cancers, identify additional indolent diseases and their transformation pathways, and flag patients which may respond more effectively to distinct regimens.

Our classification is based on causal genetic changes, and as a result, is likely to be durable, reproducible, and clinically relevant. We note that while treatments and clinical practices may change over time, improving the survival of DLBCL and B-NHL patients, the underlying genomic changes causing B-NHLs will remain consistent. Therefore, our classification represents fundamentally different pathogenesis mechanisms inherent to DLBCL and captures lasting biological information. With the addition of translocation, copy number, and gene expression data in a follow up study, this classification will additionally gain resolution, accuracy, reproducibility, and clinical relevance.

6.3. Comparison to Recent Large Scale DLBCL Genomics Study

Recently, Reddy *et al.* published an integrative analysis of 1,001 DLBCL samples that complements the results of this manuscript[15]. Whole exome sequencing, transcriptomics, copy number analysis, and FISH tests were conducted . Additionally, 400 of the samples had paired normals. In comparison, our study conducted targeted sequencing, transcriptomics, copy number analysis, and FISH tests on 962 DLBCL samples without paired normals. The targeted sequencing has been completed, and the outcomes of the remaining analyses are being processed by collaborators. The complementarities between our studies enable synergies to refine genomic analysis and classification of DLBCL.

First, Reddy *et al.*'s genomic analysis is generally consistent with this work. The genes in our study with the highest number of driver mutations were generally consistent with Reddy *et al.*'s list of frequently mutated genes with a few exceptions discussed in Section 4.1.2. A few other notable differences exist. Reddy *et al.* conducted whole exome sequencing rather than targeted sequencing of genes. Whole Exome Sequencing allows Reddy *et al.* to identify driver genes with previously unreported pattern of mutations, something not possible through our targeted study. Indeed, a few of the 150 genes identified as drivers are not present within our bait set (*DUSP2, ZNF608*, and *BIRC6*) and we thus do not report variants in these genes. Conversely, our targeted sequencing study also uncovered genes and specific mutations not present in Reddy et al.'s study. For example, we found splicing errors in *SGK1* which were not reported by Reddy et al.'s work. Therefore, we see these studies as complementary. A meta-analysis involving both sets of variants would prove helpful to fully understanding the genomic changes underlying DLBCL.

Second, Reddy *et al.* take a distinct approach to DLBCL classification. Reddy *et al.* classify patients on the basis of gene expression patterns. As a result, they can identify functional signatures based on gene expression such as the Monti Host Response signature. Conversely, our study classifies DLBCL on the basis of genetic lesions. Therefore, we can identify patterns at the genetic level such as our Class 5 which is suspected to contain SMZL patients. Ultimately, future work could seek to simultaneously incorporate both gene expression patterns and genetic lesions as the basis for classification. Therefore, both types of findings could be drawn out from the clusters. Note that this is distinct from Reddy *et al.*'s work which first generated a classification based on gene expression and subsequently identified the genetic alterations associated with each cluster.

In spite of the distinct approach to classification, some commonalities were observed. Namely, our genetic classification identified *MYD88* and *CDKN2A* as defining Class 2 which we suspect to be primarily composed of ABC-DLBCL. Both of these genes had more

genomic alterations in the ABC-DLBCL expression cluster of Reddy *et al*.'s work than in other clusters.

Third, Reddy *et al.*'s study also conducted a functional CRISPR screen and created a prognostication model with implications for our study. First, the CRISPR screen only identified 35 of the 150 driver genes Reddy *et al.* had initially flagged as having functional relevant to DLBCL cell lines. This result reinforces the need to biologically validate the driver variants we have discovered. Second, Reddy *et al.* created a prognostication model that outperformed the R-IPI by using only genetic and molecular features. The prognostication first enumerates all combinations involving up to 4 distinct genetic and molecular features and affecting at least 20 patients. These 313 combinatorial features are then fed into an Elastic regression. We hope to make two improvements when developing a similar prognostication model for our dataset. First, we hope to use a more robust feature selection method such as bootstrapping or stepwise regression. Second, we hope to include additional clinical characteristics into the set of regression features. Indeed prior work for AML[143] has shown that clinical variables often have even more predictive power than genetics[143,186]. A regression model incorporating both may provide more accurate classification.

Finally, the union of these works could provide validation for both studies. Comparison of genomic variants could validate pipelines and drivers in both studies. Testing whether Reddy *et al.*'s cohort classifies into similar genetic clusters as ours could validate our genetic classification. Finally, testing Reddy *et al*.'s prognostication tool on our cohort could validate its generalizability.

6.4. Future Work

While the aforementioned project describes the genetic landscape and provides a genetic classification of various B-NHL malignancies, substantial additional potential exists.

6.4.1. Incorporating Copy Number Analysis, Gene Expression, and Translocation Data

First, the incorporation of copy number analysis, gene expression, and translocation information will add to both the pathogenesis insights derived from this project as well as the resolution of classification. Crucially, both copy number amplifications/deletions and translocations are well known to affect progression of B-NHLs while also providing subtype-differentiating lesions. Current work is underway implementing a custom algorithm to extract copy number from this targeted, unmatched dataset. Similarly, translocation data from collaborators is currently being processed and will be added. Once incorporated, our study

will be one of the two largest and most complete genetic analyses of B-NHL, and DLBCL in particular, ever conducted, thereby enabling new insights regarding causality, molecular progression, and differentiating feature of each disease. Moreover because specific copy number and translocation changes are known to predominantly present in specific subtypes (i.e. *MYC* translocation in BL), the incorporation of such data will draw sharper divisions between classes of our classification and potentially define entirely new classes.

Second, the incorporation of gene expression data in particular will allow us to compare our classification to the cell-of-origin classification currently leading the literature. By providing additional differentiating information (i.e. genetic mutations, copy number changes, and translocations), our dataset will be able to refine the cell-of-origin categories currently based purely on gene expression. Importantly, our study may also be able to define whether ABC and GCB DLBCL are indeed distinct entities or whether information inherent to genetic mutations rather than gene expression provide more convincing differentiation among DLBCL subtypes. Finally, by adding additional genetic information to the samples classified via the cell-of-origin classification, our analysis will provide mechanistic insight into the pathogenesis of GCB-DLBCL in particular whose pathogenesis is presently unknown[19].

6.4.2. Survival Analysis for Classification

Only a preliminary survival analysis was conducted to understand the distinct clinical courses of the identified classes within this study. A full survival analysis would additionally correct here for age, date of diagnosis, centre, treatment, and a variety of other variables. Such corrections are especially critical because our study incorporates samples taken over 15 years. The introduction of CHOP and subsequently R-CHOP therefore occurred within the time window of our study and the substantially improved outcomes for patients receiving these treatments versus previous ones must be accounted for. Similarly, improvement in general clinical treatment must also be accounted for.

Such a survival analysis could generate crucial clinical insights. By distinguishing which subclasses of DLBCL and the other B-NHL malignancies presented here both (1) exhibit the worst clinical course and (2) are the least likely to respond to treatment, we may be able to identify the subset of patients which should be moved toward more aggressive treatments such as stem cell transplantations and considered for experimental therapies. Moreover, by specifically conducting this analysis on the subset of DLBCL patients that respond poorly to an R-CHOP regimen vs. those that respond well to an R-CHOP regimen,

we will hopefully be able to delineate the causative genetic and molecular differences that prevent cure in 30% of DLBCL cases. If we are able to sufficiently distinguish these patients, additional studies could then fully characterize their distinct pathogenesis, leading to suggestions for new treatments and therapies that will help them. Additionally, such a survival analysis could be coupled with a survival analysis for the specific genetic lesions that are most deleterious. By identifying such lesions, both within given classes and across all classes, we would be able to more effectively identify the patients with the most aggressive clinical course and subsequently shift them onto more intensive therapies and potentially experimental clinical trials.

6.4.3. Validation of M7-FLIPI Prognostication Tool for FL

Our dataset could validate the M7-FLIPI prognostication tool for FL. M7-FLIPI seeks to risk stratify FL patients receiving first-line immunochemotherapy by considering their mutations in seven genes (*EZH2, ARID1A, MEF2B, EP300, CREBBP,* and *CARD11*), their Follicular Lymphoma International Prognostic Index (FLIPI), and their Eastern Cooperative Oncology Group performance status (ECOG)[187,188]. Our dataset contains 337 FL patients which were treated and 222 which were placed under a "watch and wait" regimen (Figure 1b). All samples were diagnostic biopsies, and all of these FL patients have the relevant genetic, clinical, and survival data required to utilize the M7-FLIPI prognostication tool. Once the appropriate clinical data is processed to subset treated patients based on the treatments they receive, we believe our dataset will be sufficiently large to validate the M7-FLIPI prognostication tool.

6.4.4. Prediction of Treatment Outcomes Based on Genetics

Finally, future work will focus on providing a machine learning based approach to improve the prognostication of DLBCL patients. The gold standard clinical prognostic tool, the Revised International Prognostic Index (R-IPI), sorts patients into three risk groups based on factors such as age and whether their lactate dehydrogenase level is elevated.[14] None of the R-IPI factors, however, account for the genetic basis of DLBCL and cannot therefore incorporate prognostic information from genetic variability between patients within the same risk group. Virtually all DLBCL patients receive the same first-line therapy, R-CHOP, despite the probability that the genetic and biological heterogeneity will result in heterogeneous response to the potential treatments available.[189] By utilizing a machine learning based approach that considers all possible lesions as well as clinical variables, we

may be able to more effectively predict which patients are likely to respond well to R-CHOP and which are not. If such an identification is possible, the patients at greater risk may be moved toward more aggressive treatments or experimental therapies.

7. References

1. Engelhard, M. *et al.* Subclassification of Diffuse Large B-Cell Lymphomas According to the Kiel Classification: Distinction of Centroblastic and Immunoblastic Lymphomas Is a Significant Prognostic Risk Factor. *Blood* **89,** 2291–2297 (1997).
2. Shepherd, P. C. A., Ganesan, T. S. & Galton, D. A. G. Haematological classification of the chronic myeloid leukaemias. *Baillières Clin. Haematol.* **1,** 887–906 (1987).
3. Bennett, J. M. *et al.* The chronic myeloid leukaemias: guidelines for distinguishing chronic granulocytic, atypical chronic myeloid, and chronic myelomonocytic leukaemia: Proposals by the French - American - British Cooperative Leukaemia Group. *Br. J. Haematol.* **87,** 746–754 (1994).
4. Vardiman, J. W., Harris, N. L. & Brunning, R. D. The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood* **100,** 2292–2302 (2002).
5. Druker, B. J. *et al.* Activity of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in the Blast Crisis of Chronic Myeloid Leukemia and Acute Lymphoblastic Leukemia with the Philadelphia Chromosome. *N. Engl. J. Med.* **344,** 1038–1042 (2001).
6. Longo, L. *et al.* Rearrangements and aberrant expression of the retinoic acid receptor alpha gene in acute promyelocytic leukemias. *J. Exp. Med.* **172,** 1571–1575 (1990).
7. Tong, J. H. *et al.* Molecular rearrangements of the MYL gene in acute promyelocytic leukemia (APL, M3) define a breakpoint cluster region as well as some molecular variants. *Oncogene* **7,** 311–316 (1992).
8. Hillestad, L. K. Acute promyelocytic leukemia. *Acta Med. Scand.* **159,** 189–194 (1957).
9. Stone, R. M. & Mayer, R. J. The unique aspects of acute promyelocytic leukemia. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **8,** 1913–1921 (1990).
10. Jeon, I. S. *et al.* A variant Ewing's sarcoma translocation (7;22) fuses the EWS gene to the ETS gene ETV1. *Oncogene* **10,** 1229–1234 (1995).

11. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235,** 177–182 (1987).

12. Slamon, D. J. *et al.* Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* **244,** 707–712 (1989).

13. Ansari, J., Palmer, D. H., Rea, D. W. & Hussain, S. A. Role of tyrosine kinase inhibitors in lung cancer. *Anticancer Agents Med. Chem.* **9,** 569–575 (2009).

14. Martelli, M. *et al.* Diffuse large B-cell lymphoma. *Crit. Rev. Oncol. Hematol.* **87,** 146–171 (2013).

15. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171,** 481–494.e15 (2017).

16. Klein, U. & Dalla-Favera, R. Germinal centres: role in B-cell physiology and malignancy. *Nat. Rev. Immunol.* **8,** 22–33 (2008).

17. MacLennan, I. C. Germinal centers. *Annu. Rev. Immunol.* **12,** 117–139 (1994).

18. Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annu. Rev. Immunol.* **30,** 429–457 (2012).

19. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nat. Rev. Immunol.* **15,** 172–184 (2015).

20. Schmitz, R. *et al.* Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **490,** 116–20 (2012).

21. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* **15,** 137–148 (2015).

22. Calado, D. P. *et al.* MYC is essential for the formation and maintenance of germinal centers. *Nat. Immunol.* **13,** 1092–1100 (2012).

23. Basso, K. *et al.* Integrated biochemical and computational approach identifies BCL6 direct target genes controlling multiple pathways in normal germinal center B cells. *Blood* **115,** 975–984 (2010).

24. Ci, W. *et al.* The BCL6 transcriptional program features repression of multiple oncogenes in primary B cells and is deregulated in DLBCL. *Blood* **113,** 5536–5548 (2009).

25. Calado, D. P. *et al.* The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers. *Nat. Immunol.* **13,** 1092–1100 (2012).

26. Dominguez-Sola, D. *et al.* The proto-oncogene MYC is required for selection in the germinal center and cyclic reentry. *Nat. Immunol.* **13,** 1083–1091 (2012).

27. EZH2-mediated epigenetic silencing in germinal center B cells contributes to proliferation and lymphomagenesis | Blood Journal. Available at: http://www.bloodjournal.org/content/116/24/5247. (Accessed: 14th September 2017)

28. Heise, N. *et al.* Germinal center B cell maintenance and differentiation are controlled by distinct NF-κB transcription factor subunits. *J. Exp. Med.* **211,** 2103–2118 (2014).

29. Saito, M. *et al.* A Signaling Pathway Mediating Downregulation of BCL6 in Germinal Center B Cells Is Blocked by BCL6 Gene Alterations in B Cell Lymphoma. *Cancer Cell* **12,** 280–292 (2007).

30. Chu, Y. *et al.* B cells lacking the tumor suppressor TNFAIP3/A20 display impaired differentiation and hyperactivation and cause inflammation and autoimmunity in aged mice. *Blood* **117,** 2227–2236 (2011).

31. Tavares, R. M. *et al.* The Ubiquitin Modifying Enzyme A20 Restricts B Cell Survival and Prevents Autoimmunity. *Immunity* **33,** 181–191 (2010).

32. Klein, U. *et al.* Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. *Nat. Immunol.* **7,** 773–782 (2006).

33. Sciammas, R. *et al.* Graded Expression of Interferon Regulatory Factor-4 Coordinates Isotype Switching with Plasma Cell Differentiation. *Immunity* **25,** 225–236 (2006).

34. Tunyaplin, C. *et al.* Direct Repression of prdm1 by Bcl-6 Inhibits Plasmacytic Differentiation. *J. Immunol.* **173,** 1158–1165 (2004).

35. Cobaleda, C., Schebesta, A., Delogu, A. & Busslinger, M. Pax5: the guardian of B cell identity and function. *Nat. Immunol.* **8,** 463–470 (2007).

36. Delogu, A. *et al.* Gene Repression by Pax5 in B Cells Is Essential for Blood Cell Homeostasis and Is Reversed in Plasma Cells. *Immunity* **24,** 269–281 (2006).

37. Nera, K.-P. *et al.* Loss of Pax5 Promotes Plasma Cell Differentiation. *Immunity* **24,** 283–293 (2006).

38. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat. Rev. Immunol.* **15,** 160–171 (2015).

39. Stevenson, F. K. *et al.* The occurrence and significance of V gene mutations in B cell-derived human malignancy. *Adv. Cancer Res.* **83,** 81–116 (2001).

40. Victora, G. D. *et al.* Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas. *Blood* **120,** 2240–2248 (2012).

41. Dave, S. S. *et al.* Molecular Diagnosis of Burkitt's Lymphoma. *N. Engl. J. Med.* **354,** 2431–2442 (2006).

42. Schmitz, R., Ceribelli, M., Pittaluga, S., Wright, G. & Staudt, L. M. Oncogenic Mechanisms in Burkitt Lymphoma. *Cold Spring Harb. Perspect. Med.* **4,** a014282 (2014).

43. Nussenzweig, A. & Nussenzweig, M. C. Origin of Chromosomal Translocations in Lymphoid Cancer. *Cell* **141,** 27–38 (2010).

44. Dominguez-Sola, D. *et al.* Non-transcriptional control of DNA replication by c-Myc. *Nature* **448,** 445–451 (2007).

45. Corso, J. *et al.* Elucidation of tonic and activated B-cell receptor signaling in Burkitt's lymphoma provides insights into regulation of cell survival. *Proc. Natl. Acad. Sci.* **113,** 5688–5693 (2016).

46. Muppidi, J. R. *et al.* Loss of signalling via Gα13 in germinal centre B-cell-derived lymphoma. *Nature* **516,** 254–258 (2014).

47. McCann, K. J., Johnson, P. W. M., Stevenson, F. K. & Ottensmeier, C. H. Universal N-glycosylation sites introduced into the B-cell receptor of follicular lymphoma by somatic mutation: a second tumorigenic event? *Leukemia* **20,** 530–534 (2006).

48. Alizadeh, a a *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–11 (2000).

49. Kridel, R., Sehn, L. H. & Gascoyne, R. D. Pathogenesis of follicular lymphoma. *J. Clin. Invest.* **122,** 3424–3431 (2012).

50. Montoto, S. & Fitzgibbon, J. Transformation of Indolent B-Cell Lymphomas. *J. Clin. Oncol.* **29,** 1827–1834 (2011).

51. Küppers, R. Mechanisms of B-cell lymphoma pathogenesis. *Nat. Rev. Cancer* **5,** 251–62 (2005).

52. Mechanisms of chromosomal translocations in B cell lymphomas. *Publ. Online 10 Sept. 2001 Doi101038sjonc1204640* **20,** (2001).

53. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).

54. Pasqualucci, L. *et al.* Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.* **43,** 830–837 (2011).

55. Jamroziak, K., Tadmor, T., Robak, T. & Polliack, A. Richter syndrome in chronic lymphocytic leukemia: updates on biology, clinical features and therapy. *Leuk. Lymphoma* **56,** 1949–1958 (2015).

56. 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,** 3879–3884 (2012).

57. Pasqualucci, L. *et al.* Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* **471,** 189–195 (2011).

58. Basso, K. & Dalla-Favera, R. Roles of BCL6 in normal and transformed germinal center B cells. *Immunol. Rev.* **247,** 172–183 (2012).

59. Pasqualucci, L. *et al.* Mutations of the BCL6 proto-oncogene disrupt its negative autoregulation in diffuse large B-cell lymphoma. *Blood* **101,** 2914–2923 (2003).

60. Ye, B. H. *et al.* Chromosomal translocations cause deregulated BCL6 expression by promoter substitution in B cell lymphoma. *EMBO J.* **14,** 6209–6217 (1995).

61. Challa-Malladi, M. *et al.* Combined Genetic Inactivation of β2-Microglobulin and CD58 Reveals Frequent Escape from Immune Recognition in Diffuse Large B Cell Lymphoma. *Cancer Cell* **20,** 728–740 (2011).

62. Boknäs, N. *et al.* Response: Platelets do not generate activated factor XII—how inappropriate experimental models have led to misleading conclusions. *Blood* **124,** 1692–1694 (2014).

63. Aukema, S. M. *et al.* Double-hit B-cell lymphomas. *Blood* **117,** 2319–2331 (2011).

64. Li, S. *et al.* MYC/BCL2 double-hit high-grade B-cell lymphoma. *Adv. Anat. Pathol.* **20,** 315–326 (2013).

65. Béguelin, W. *et al.* EZH2 Is Required for Germinal Center Formation and Somatic EZH2 Mutations Promote Lymphoid Transformation. *Cancer Cell* **23,** 677–692 (2013).

66. Caganova, M. *et al.* Germinal center dysregulation by histone methyltransferase EZH2 promotes lymphomagenesis. *J. Clin. Invest.* **123,** 5009–5022 (2013).

67. Green, J. A. *et al.* The sphingosine 1-phosphate receptor S1P2 maintains the homeostasis of germinal center B cells and promotes niche confinement. *Nat. Immunol.* **12,** 672–680 (2011).

68. Davis, R. E. *et al.* Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature* **463,** 88–92 (2010).

69. Ngo, V. N. *et al.* Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470,** 115–119 (2011).

70. Compagno, M. *et al.* Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **459,** 717–721 (2009).

71. Lenz, G. *et al.* Aberrant immunoglobulin class switch recombination and switch translocations in activated B cell-like diffuse large B cell lymphoma. *J. Exp. Med.* **204,** 633–643 (2007).

72. Lenz, G. *et al.* Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci USA* **105,** 13520–13525 (2008).

73. Care, M. A. *et al.* SPIB and BATF provide alternate determinants of IRF4 occupancy in diffuse large B-cell lymphoma linked to disease heterogeneity. *Nucleic Acids Res.* **42,** 7591–7610 (2014).

74. Schmidlin, H. *et al.* Spi-B inhibits human plasma cell differentiation by repressing BLIMP1 and XBP-1 expression. *Blood* **112,** 1804–1812 (2008).

75. Yang, Y. *et al.* Exploiting synthetic lethality for the therapy of ABC diffuse large B cell lymphoma. *Cancer Cell* **21,** 723–737 (2012).

76. Rosenwald, A., Wright, G., Chan, W. et al, WC, C. & Al., E. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *N. Engl. J. Med.* **346,** 1937–1947 (2002).

77. Bunn, H. F. & Aster, J. C. *Pathophysiology of Blood Disorders*. (McGraw-Hill Education / Medical, 2011).

78. Sehn, L. H. *et al.* The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. *Blood* **109,** 1857–1862 (2015).

79. Savage, K. J. *et al.* MYC gene rearrangements are associated with a poor prognosis in diffuse large B-cell lymphoma patients treated with R-CHOP chemotherapy. *Blood* **114,** 3533–3537 (2009).

80. Ott, G. *et al.* Immunoblastic morphology but not the immunohistochemical GCB/nonGCB classifier predicts outcome in diffuse large B-cell lymphoma in the RICOVER-60 trial of the DSHNHL. *Blood* **116,** 4916–4925 (2010).

81. Murase, T. *et al.* Intravascular large B-cell lymphoma (IVLBCL): a clinicopathologic study of 96 cases with special reference to the immunophenotypic heterogeneity of CD5. *Blood* **109,** 478–485 (2007).

82. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127,** 2375–2390 (2016).

83. Rosenthal, A. & Younes, A. High grade B-cell lymphoma with rearrangements of MYC and BCL2 and/or BCL6: Double hit and triple hit lymphomas and double expressing lymphoma. *Blood Rev.* **31,** 37–42 (2017).

84. Sesques, P. & Johnson, N. A. Approach to the diagnosis and treatment of high-grade B-cell lymphomas with MYC and BCL2 and/or BCL6 rearrangements. *Blood* **129,** 280–288 (2017).

85. Eberle, F. C. *et al.* Gray zone lymphoma: chromosomal aberrations with immunophenotypic and clinical correlations. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **24,** 1586–1597 (2011).

86. Traverse-Glehen, A. *et al.* Mediastinal gray zone lymphoma: the missing link between classic Hodgkin's lymphoma and mediastinal large B-cell lymphoma. *Am. J. Surg. Pathol.* **29,** 1411–1421 (2005).

87. Zinzani, P. L. *et al.* Anaplastic large-cell lymphoma: clinical and prognostic evaluation of 90 adult patients. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **14,** 955–962 (1996).

88. Dunleavy, K. & Wilson, W. H. Primary mediastinal B-cell lymphoma and mediastinal gray zone lymphoma: do they require a unique therapeutic approach? *Blood* **125,** 33–39 (2015).

89. Taddesse-Heath, L., Meloni-Ehrig, A., Scheerle, J., Kelly, J. C. & Jaffe, E. S. Plasmablastic lymphoma with MYC translocation: evidence for a common pathway in the generation of plasmablastic features. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **23,** 991–999 (2010).

90. Castillo, J. J. Plasmablastic lymphoma: are more intensive regimens needed? *Leuk. Res.* **35,** 1547–1548 (2011).

91. Castillo, J. J. *et al.* Human immunodeficiency virus-associated plasmablastic lymphoma: poor prognosis in the era of highly active antiretroviral therapy. *Cancer* **118,** 5270–5277 (2012).

92. Achten, R., Verhoef, G., Vanuytsel, L. & De Wolf-Peeters, C. Histiocyte-rich, T-cell-rich B-cell lymphoma: a distinct diffuse large B-cell lymphoma subtype showing characteristic morphologic and immunophenotypic features. *Histopathology* **40,** 31–45 (2002).

93. Abramson, J. S. T-cell/histiocyte-rich B-cell lymphoma: biology, diagnosis, and management. *The Oncologist* **11,** 384–392 (2006).

94. Achten, R., Verhoef, G., Vanuytsel, L. & De Wolf-Peeters, C. T-cell/histiocyte-rich large B-cell lymphoma: a distinct clinicopathologic entity. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **20,** 1269–1277 (2002).

95. Colomo, L. *et al.* Diffuse large B-cell lymphomas with plasmablastic differentiation represent a heterogeneous group of disease entities. *Am. J. Surg. Pathol.* **28,** 736–747 (2004).

96. Valera, A. *et al.* IG/MYC rearrangements are the main cytogenetic alteration in plasmablastic lymphomas. *Am. J. Surg. Pathol.* **34,** 1686–1694 (2010).

97. Oyama, T. *et al.* Age-related EBV-associated B-cell lymphoproliferative disorders constitute a distinct clinicopathologic group: a study of 96 patients. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **13,** 5124–5132 (2007).

98. Aozasa, K., Takakuwa, T. & Nakatsuka, S. Pyothorax-associated lymphoma: a lymphoma developing in chronic inflammation. *Adv. Anat. Pathol.* **12,** 324–331 (2005).

99. Tomita, N. *et al.* Clinicopathological features of lymphoma/leukemia patients carrying both BCL2 and MYC translocations. *Haematologica* **94,** 935–943 (2009).

100. Nakatsuka, S.-I. *et al.* Pyothorax-associated lymphoma: a review of 106 cases. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **20,** 4255–4260 (2002).

101. Delsol, G. *et al.* A new subtype of large B-cell lymphoma expressing the ALK kinase and lacking the 2; 5 translocation. *Blood* **89,** 1483–1490 (1997).

102. Gesk, S. *et al.* ALK-positive diffuse large B-cell lymphoma with ALK-Clathrin fusion belongs to the spectrum of pediatric lymphomas. *Leukemia* **19,** 1839–1840 (2005).

103. Laurent, C. *et al.* Anaplastic lymphoma kinase-positive diffuse large B-cell lymphoma: a rare clinicopathologic entity with poor prognosis. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **27,** 4211–4216 (2009).

104. Morgan, E. A. & Nascimento, A. F. Anaplastic lymphoma kinase-positive large B-cell lymphoma: an underrecognized aggressive lymphoma. *Adv. Hematol.* **2012,** 529572 (2012).

105. Bedwell, C. *et al.* Cytogenetically complex SEC31A-ALK fusions are recurrent in ALK-positive large B-cell lymphomas. *Haematologica* **96,** 343–346 (2011).

106. Chabner, B. A. Early accelerated approval for highly targeted cancer drugs. *N. Engl. J. Med.* **364,** 1087–1089 (2011).

107. Delecluse, H. J. *et al.* Plasmablastic lymphomas of the oral cavity: a new entity associated with the human immunodeficiency virus infection. *Blood* **89,** 1413–1420 (1997).

108. Vega, F. *et al.* Plasmablastic lymphomas and plasmablastic plasma cell myelomas have nearly identical immunophenotypic profiles. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **18,** 806–815 (2005).

109. Gabrea, A. *et al.* Secondary genomic rearrangements involving immunoglobulin or MYC loci show similar prevalences in hyperdiploid and nonhyperdiploid myeloma tumors. *Genes. Chromosomes Cancer* **47,** 573–590 (2008).

110. Cesarman, E., Chang, Y., Moore, P. S., Said, J. W. & Knowles, D. M. Kaposi's sarcoma-associated herpesvirus-like DNA sequences in AIDS-related body-cavity-based lymphomas. *N. Engl. J. Med.* **332,** 1186–1191 (1995).

111. Cobo, F. *et al.* Expression of potentially oncogenic HHV-8 genes in an EBV-negative primary effusion lymphoma occurring in an HIV-seronegative patient. *J. Pathol.* **189,** 288–293 (1999).

112. Teruya-Feldstein, J. *et al.* Expression of human herpesvirus-8 oncogene and cytokine homologues in an HIV-seronegative patient with multicentric Castleman's disease and

primary effusion lymphoma. *Lab. Investig. J. Tech. Methods Pathol.* **78,** 1637–1642 (1998).

113. Chadburn, A. *et al.* KSHV-positive solid lymphomas represent an extra-cavitary variant of primary effusion lymphoma. *Am. J. Surg. Pathol.* **28,** 1401–1416 (2004).

114. Petitjean, B. *et al.* Pyothorax-associated lymphoma: a peculiar clinicopathologic entity derived from B cells at late stage of differentiation and with occasional aberrant dual B- and T-cell phenotype. *Am. J. Surg. Pathol.* **26,** 724–732 (2002).

115. Beaty, M. W. *et al.* A biophenotypic human herpesvirus 8--associated primary bowel lymphoma. *Am. J. Surg. Pathol.* **23,** 992–994 (1999).

116. Ahmad, A. *et al.* Kaposi sarcoma-associated herpesvirus-encoded viral FLICE inhibitory protein (vFLIP) K13 cooperates with Myc to promote lymphoma in mice. *Cancer Biol. Ther.* **10,** 1033–1040 (2010).

117. Bubman, D., Guasparri, I. & Cesarman, E. Deregulation of c-Myc in primary effusion lymphoma by Kaposi's sarcoma herpesvirus latency-associated nuclear antigen. *Oncogene* **26,** 4979–4986 (2007).

118. Rodriguez, J. *et al.* Follicular Large Cell Lymphoma: An Aggressive Lymphoma That Often Presents With Favorable Prognostic Features. *Blood* **93,** 2202–2207 (1999).

119. Campo, E. *et al.* The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood* **117,** 5019–5032 (2011).

120. Matutes, E. *et al.* Splenic marginal zone lymphoma proposals for a revision of diagnostic, staging and therapeutic criteria. *Leukemia* **22,** 487–495 (2008).

121. Thieblemont, C. *et al.* Splenic marginal-zone lymphoma: A distinct clinical and pathological entity. *Lancet Oncol.* **4,** 95–103 (2003).

122. Weill, J.-C., Weller, S. & Reynaud, C.-A. Human marginal zone B cells. *Annu. Rev. Immunol.* **27,** 267–285 (2009).

123. Zibellini, S. *et al.* Stereotyped patterns of B-cell receptor in splenic marginal zone lymphoma. *Haematologica* **95,** 1792–1796 (2010).

124. Over 30% of Patients With Splenic Marginal Zone Lymphoma Express the Same Immunoglobulin Heavy Variable Gene: Ontogenetic Implications. *PubMed Journals* Available at: https://ncbi.nlm.nih.gov/labs/articles/22222599/. (Accessed: 14th September 2017)

125. Brisou, G. *et al.* A restricted IGHV gene repertoire in splenic marginal zone lymphoma is associated with autoimmune disorders. *Haematologica* **99,** e197–e198 (2014).

126. Analysis of mutations in immunoglobulin heavy chain variable region genes of microdissected marginal zone (MGZ) B cells suggests that the MGZ of human spleen is a reservoir of memory B cells. *J. Exp. Med.* **182,** 559–566 (1995).

127. Tierens, A., Delabie, J., Michiels, L., Vandenberghe, P. & De Wolf-Peeters, C. Marginal-zone B cells in the human lymph node and spleen show somatic hypermutations and display clonal expansion. *Blood* **93,** 226–234 (1999).

128. Colombo, M. *et al.* Expression of Immunoglobulin Receptors with Distinctive Features Indicating Antigen Selection by Marginal Zone B Cells from Human Spleen. *Mol. Med.* **19,** 294–302 (2013).

129. Du, M.-Q. Pathogenesis of splenic marginal zone lymphoma. *Pathogenesis* **2,** 11–20 (2015).

130. Caro, P. *et al.* Metabolic Signatures Uncover Distinct Targets in Molecular Subsets of Diffuse Large B-Cell Lymphoma. *Cancer Cell* **22,** 547–560 (2012).

131. Smith, A. *et al.* The Haematological Malignancy Research Network (HMRN): a new information strategy for population based epidemiology and health service research. *Br. J. Haematol.* **148,** 739–753 (2010).

132. Wang, H.-I. *et al.* Long-term medical costs and life expectancy of acute myeloid leukemia: a probabilistic decision model. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* **17,** 205–214 (2014).

133. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

134. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

135. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. in *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2002). doi:10.1002/cpbi.20

136. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).

137. Ramirez-Gonzalez, R. H., Bonnal, R., Caccamo, M. & MacLean, D. Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. *Source Code Biol. Med.* **7,** 6 (2012).

138. Donlin, M. J. Using the Generic Genome Browser (GBrowse). in *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2002). doi:10.1002/0471250953.bi0909s28

139. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

140. Menzies, A. *et al.* VAGrENT: Variation Annotation Generator. in *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2002). doi:10.1002/0471250953.bi1508s52

141. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2015).

142. Whye, Y. *et al.* Hierarchical Dirichlet. **101,** 1566–1581 (2012).

143. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374,** 2209–2221 (2016).

144. Pasqualucci, L. & Dalla-Favera, R. SnapShot: Diffuse Large B Cell Lymphoma. *Cancer Cell* **25,** 132–132.e1 (2014).

145. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3,** 811 (2012).

146. Loo, P. V. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107,** 16910–16915 (2010).

147. Lee, J.-H., Jeong, H., Choi, J.-W., Oh, H. & Kim, Y.-S. Clinicopathologic significance of MYD88 L265P mutation in diffuse large B-cell lymphoma: a meta-analysis. *Sci. Rep.* **7,** 1785 (2017).

148. Bösl, M. W. *et al.* STAT6 Is Recurrently and Significantly Mutated in Follicular Lymphoma and Enhances the IL-4 Induced Expression of Membrane-Bound and Soluble CD23. *Blood* **126,** 3923–3923 (2015).

149. Project, the I. M.-S. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.* **44,** 1316–1320 (2012).

150. Campo, E. New pathogenic mechanisms in Burkitt lymphoma. *Nat. Genet.* **44,** 1288–1289 (2012).

151. Yoshimura, A., Naka, T. & Kubo, M. SOCS proteins, cytokine signalling and immune regulation. *Nat. Rev. Immunol.* **7,** 454–465 (2007).

152. Jelinic, P. *et al.* Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. *Nat. Genet.* **46,** 424–426 (2014).

153. Shain, A. H. & Pollack, J. R. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. *PloS One* **8,** e55119 (2013).

154. Beà, S. *et al.* Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18250–18255 (2013).

155. Camus, V. *et al.* Detection and prognostic value of recurrent XPO1 mutations in tumor and cell-free circulating DNA of patients with classical Hodgkin Lymphoma. *Haematologica* haematol.2016.145102 (2016). doi:10.3324/haematol.2016.145102

156. Pon, J. R. *et al.* MEF2B mutations in non-Hodgkin lymphoma dysregulate cell migration by decreasing MEF2B target gene activation. *Nat. Commun.* **6,** 7953 (2015).

157. Jeelall, Y. S. *et al.* Human lymphoma mutations reveal CARD11 as the switch between self-antigen–induced B cell death or proliferation and autoantibody production. *J. Exp. Med.* **209,** 1907–1917 (2012).

158. Lenz, G. *et al.* Oncogenic CARD11 Mutations in Human Diffuse Large B Cell Lymphoma. *Science* **319,** 1676–1679 (2008).

159. Nikolaev, S. I. *et al.* Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat. Genet.* **44,** 133–139 (2011).

160. Schmidt, J. *et al.* Mutations of MAP2K1 are frequent in pediatric-type follicular lymphoma and result in ERK pathway activation. *Blood* blood-2017-03-776278 (2017). doi:10.1182/blood-2017-03-776278

161. Levy, D. E. & Lee, C. What does Stat3 do? *J. Clin. Invest.* **109,** 1143–1148 (2002).

162. Xie, T.-X. *et al.* Stat3 activation regulates the expression of matrix metalloproteinase-2 and tumor invasion and metastasis. *Oncogene* **23,** 3550–3560 (2004).

163. Khodabakhshi, A. H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3,** 1308–19 (2012).

164. BCL10 gene mutations rarely occur in lymphoid malignancies. *Publ. Online 17 April 2000 Doi101038sjleu2401747* **14,** (2000).

165. Tsushita, K. *et al.* Mutation study of the BCL10 gene in lymphoma with both RNA and DNA. *Leukemia* **15,** 1139–1140 (2001).

166. Turvey, S. E. *et al.* The CARD11-BCL10-MALT1 (CBM) signalosome complex: Stepping into the limelight of human primary immunodeficiency. *J. Allergy Clin. Immunol.* **134,** 276–284 (2014).

167. Yang, C., David, L., Qiao, Q., Damko, E. & Wu, H. The CBM signalosome: Potential therapeutic target for aggressive lymphoma? *Cytokine Growth Factor Rev.* **25,** 175–183 (2014).

168. Rebeaud, F. *et al.* The proteolytic activity of the paracaspase MALT1 is key in T cell activation. *Nat. Immunol.* **9,** 272–281 (2008).

169. Xu, Y. *et al.* Loss of IRF8 Inhibits the Growth of Diffuse Large B-cell Lymphoma. *J. Cancer* **6,** 953–961 (2015).

170. Waight, J. D., Banik, D., Griffiths, E. A., Nemeth, M. J. & Abrams, S. I. Regulation of the interferon regulatory factor-8 (IRF-8) tumor suppressor gene by the signal transducer and activator of transcription 5 (STAT5) transcription factor in chronic myeloid leukemia. *J. Biol. Chem.* **289,** 15642–15652 (2014).

171. Müschen, M., Warskulat, U. & Beckmann, M. W. Defining CD95 as a tumor suppressor gene. *J. Mol. Med. Berl. Ger.* **78,** 312–325 (2000).

172. Villa-Morales, M. *et al.* FAS system deregulation in T-cell lymphoblastic lymphoma. *Cell Death Dis.* **5,** e1110 (2014).

173. Helming, K. C. *et al.* ARID1B is a specific vulnerability in ARID1A-mutant cancers. *Nat. Med.* **20,** 251–254 (2014).

174. Santen, G. W. E. *et al.* Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. *Nat. Genet.* **44,** 379–380 (2012).

175. Sausen, M. *et al.* Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat. Genet.* **45,** 12–17 (2013).

176. Cajuso, T. *et al.* Exome sequencing reveals frequent inactivating mutations in ARID1A, ARID1B, ARID2 and ARID4A in microsatellite unstable colorectal cancer. *Int. J. Cancer* **135,** 611–623 (2014).

177. Khursheed, M. *et al.* ARID1B, a member of the human SWI/SNF chromatin remodeling complex, exhibits tumour-suppressor activities in pancreatic cancer cell lines. *Br. J. Cancer* **108,** 2056–2062 (2013).

178. Kridel, R. *et al.* Whole transcriptome sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma. *Blood* **119,** 1963–1971 (2012).

179. Clipson, A. *et al.* KLF2 mutation is the most frequent somatic change in splenic marginal zone lymphoma and identifies a subset with distinct genotype. *Leukemia* **29,** 1177–1185 (2015).

180. Piva, R. *et al.* The Krüppel-like factor 2 transcription factor gene is recurrently mutated in splenic marginal zone lymphoma. *Leukemia* **29,** 503–507 (2015).

181. Christiaans, I. *et al.* Germline SMARCB1 mutation and somatic NF2 mutations in familial multiple meningiomas. *J. Med. Genet.* **48,** 93–97 (2011).

182. Friedberg, J. W. Double-Hit Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol.* **30,** 3439–3443 (2012).

183. Hwang, Y. Y., Loong, F., Chung, L. P. & Chim, C. S. Atypical burkitt's lymphoma transforming from follicular lymphoma. *Diagn. Pathol.* **6,** 63 (2011).

184. Arcaini, L., Rossi, D. & Paulli, M. Splenic marginal zone lymphoma : from genetics to management. **127,** 2072–2082 (2016).

185. Gao, X. *et al.* High-Grade Transformation in a Splenic Marginal Zone Lymphoma with a Cerebral Manifestation. *Am. J. Case Rep.* **18,** 611–616 (2017).

186. Gerstung, M. *et al.* Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat. Genet.* **49,** 332–340 (2017).

187. Pastore, A. *et al.* Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. *Lancet Oncol.* **16,** 1111–1122 (2015).

188. Jurinovic, V. *et al.* Clinicogenetic risk models predict early progression of follicular lymphoma after first-line immunochemotherapy. *Blood* **128,** 1112–1120 (2016).

189. Roschewski, M., Staudt, L. M. & Wilson, W. H. Diffuse large B-cell lymphoma-treatment approaches in the molecular era. *Nat. Rev. Clin. Oncol.* **11,** 12–23 (2014).

8. Appendix 1: Classification Code

tmp.html

AML classification using Dirichlet Processes

```
# /lustre/scratch117/casm/team154/cr8/DLBCL_study/annovar_transfer/reference/bsub_farm yeste
# /lustre/scratch117/casm/team154/cr8/DLBCL_study/annovar_transfer/reference/bsub_farm yeste
```

Code run on

```
options(markdown.HTML.header = "tmp.html")
system("hostname -f", intern=TRUE)
```

```
## [1] "bc-29-2-08.internal.sanger.ac.uk"
```

```
Sys.time()
```

```
## [1] "2017-08-27 17:03:27 BST"
```

```
getwd()
```

```
## [1] "/lustre/scratch117/casm/team154/cr8/DLBCL_study/annovar_transfer/full_study/synthesiz
```

using

```
library(knitr)
```

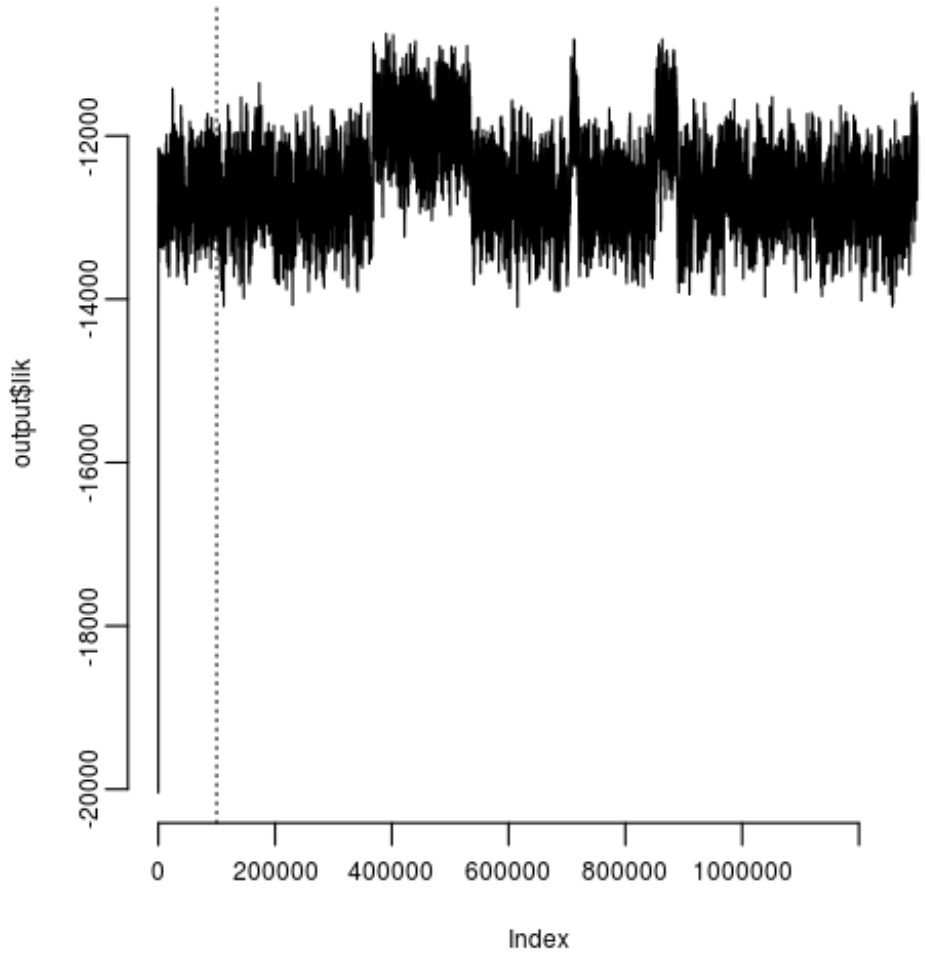
Libraries and data

```
source("/lustre/scratch117/casm/team154/cr8/DLBCL_study/cleaning_and_annotation/code/global_s
load_global_packages()

library(CoxHD) # library(devtools); install_github("mg14/CoxHD/CoxHD")
library(mg14) # library(devtools); install_github("mg14/mg14")
library(hdp)
library(lattice)
set1 <- brewer.pal(8, "Set1")

# If running from classification_workspace.Rdata instead of from scratch
load("classification_workspace.Rdata")
spin("/lustre/scratch117/casm/team154/cr8/DLBCL_study/cleaning_and_annotation/code/visualize_
```

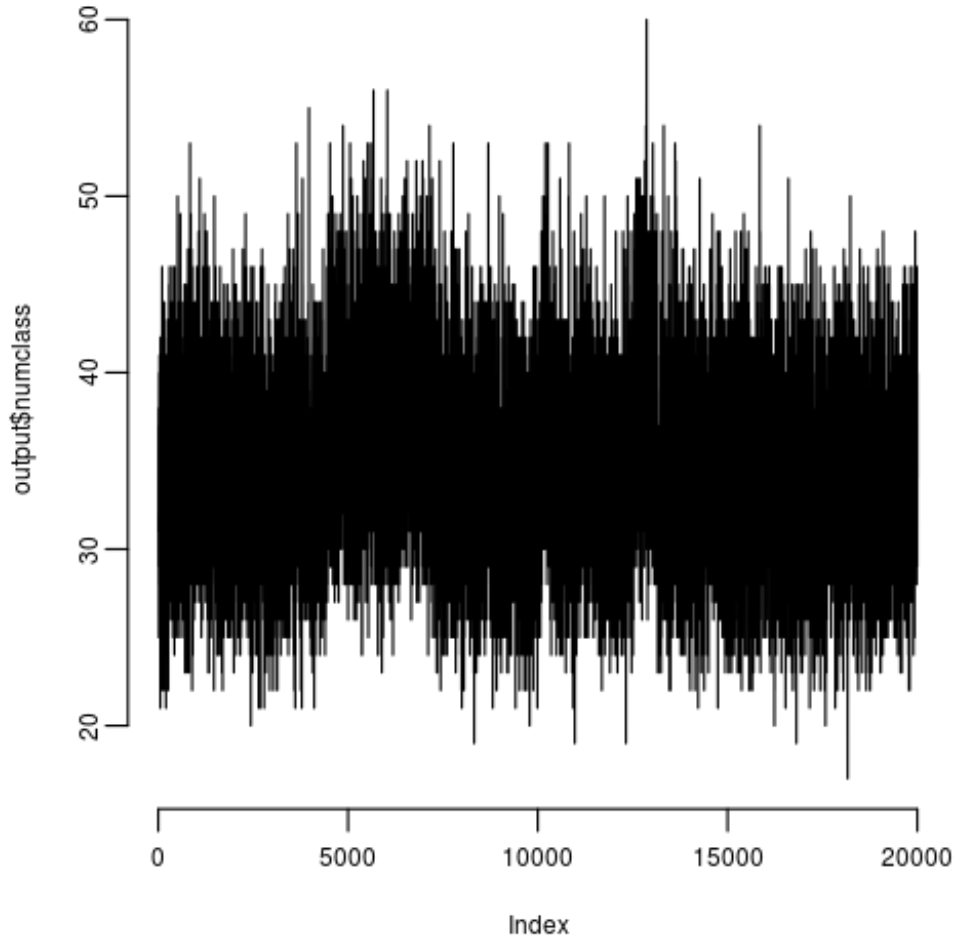
```
##
##
## processing file: /lustre/scratch117/casm/team154/cr8/DLBCL_study/cleaning_and_annotation/c
```

```
## Error in parse_block(g[-1], g[1], params.src): duplicate label 'run'
```

```
plot(output$lik, type='l'); abline(v=burnin, lty=3)
```




```
plot(output$numclass, type='l')
```



AML classes

```
posteriorMerged <- hdp_extract_signatures(output, prop.explained=0.99, cos.merge=0.95)
```

```
#posteriorMeans <- Reduce("+",posteriorMerged$sigs_qq)/length(posteriorMerged$sigs_qq)
posteriorSamples <- array(unlist(posteriorMerged$sigs_qq), dim=c(dim(posteriorMerged$sigs_qq[
rownames(posteriorSamples) <- colnames(genotypesImputed)
colnames(posteriorSamples) <- 1:ncol(posteriorSamples) -1
posteriorMeans <- rowMeans(posteriorSamples, dim=2)
posteriorQuantiles <- apply(posteriorSamples, 1:2, quantile, c(0.025,.5,0.975))
posteriorMode <- apply(posteriorSamples, 1:2, function(x) {t <- table(x); as.numeric(names(t)
kable(posteriorQuantiles[2,,], "html", table.attr = 'id="posteriorMedian"') # Posterior media
```

	0	1	2	3	4	5	6	7
SOCS1	0	0	0	0	80	0	1	0
FOXO1	0	77	0	0	0	0	0	0
BTG2	0	0	55	0	0	0	1	0
FAS	0	0	0	0	16	31	0	0
ARID1A	0	115	0	0	0	0	0	0
KMT2D	0	607	0	0	0	0	0	0
CREBBP	0	363	0	0	0	0	0	0
TNFRSF14	0	312	0	0	0	0	0	0
ID3	0	0	0	0	0	0	0	0

	0	1	2	3	4	5	6	7
ZFP36L1	0	0	0	0	4	0	4	0
ASXL2	0	0	0	0	0	0	3	0
NOTCH2	0	0	0	0	0	65	0	0
WT1	0	0	0	0	0	0	3	0
IRF4	0	0	23	0	0	0	0	0
CDKN2A	0	0	50	0	0	0	0	0
PTPRC	0	0	0	0	0	0	0	0
MYC	0	0	0	0	0	0	10	0
B2M	0	2	0	0	57	47	0	0
TP53	0	37	0	73	49	0	4	30
TBL1XR1	0	0	48	0	0	0	1	0
TET2	0	0	0	0	0	0	2	0
TNFAIP3	0	0	0	0	38	56	0	0
ASXL1	0	0	0	0	0	0	1	0
NFKBIE	0	0	0	0	39	0	0	0
NOTCH1	0	0	0	0	0	0	3	0
STAT3	0	0	0	0	0	0	0	0
IRF8	0	110	0	0	0	0	0	0
SMARCB1	0	0	0	0	0	0	13	0
BCL10	0	0	0	0	0	48	0	0
SRSF2	0	0	0	0	0	0	5	0
PDS5B	0	0	0	0	0	0	5	0
MEF2B	0	49	0	0	0	0	0	0
BTK	0	0	0	0	0	0	0	0
POU2AF1	0	34	0	0	0	0	0	0
ARID1B	0	42	0	0	0	0	0	0
MGA	0	0	0	0	0	0	1	0
CARD11	0	97	0	0	0	0	0	0
EP300	0	46	0	0	0	0	0	0
SPEN	0	0	0	0	0	35	3	0
ATRX	0	0	0	0	0	0	7	0
DNMT3A	0	0	0	0	0	0	0	0
NF1	0	0	18	0	0	0	1	0
PIK3R1	0	19	0	0	0	0	0	0
FBXO11	0	0	0	0	0	0	0	0
PRDM1	0	0	41	0	0	0	0	0
IKZF1	0	0	0	0	0	0	0	0
TRAF3	0	0	0	0	0	0	6	0
SMARCA4	0	0	0	0	0	0	0	0
PTEN	0	0	0	0	0	0	0	0
CDKN1B	0	0	0	0	0	0	8	0
BCORL1	0	0	0	0	0	0	4	0

MSH2	0	0	0	0	0	0	5	0
	0	1	2	3	4	5	6	7
MYD88	0	0	149	0	0	0	1	0
RB1	0	0	0	0	0	0	1	0
EZH2	0	290	0	0	0	0	0	0
ATM	0	0	0	0	0	0	6	0
BCL11A	0	0	0	0	0	0	1	0
BCOR	0	0	0	0	0	0	3	0
KMT2C	0	0	0	0	0	0	2	0
RASA2	0	0	0	0	0	0	8	0
CHD2	0	0	0	0	0	0	4	0
NFKB1	0	0	0	0	0	0	2	0
PHF6	0	0	0	0	0	0	0	0
MAP2K1	0	0	0	0	0	0	0	0
CASP8	0	0	0	0	0	0	10	0
CCND3	0	0	0	35	0	0	0	0
BCL7A	0	0	0	0	0	0	3	0
KDM6A	0	0	0	0	0	0	1	0
XPO1	0	0	0	0	0	0	0	0
PAX5	0	0	0	0	0	0	0	0
DDX3X	0	0	0	0	0	0	0	0
FAM46C	0	0	0	0	0	0	1	0
ARID2	0	0	0	0	0	0	4	0
CD79B	0	0	0	0	0	0	2	0
FBXW7	0	0	0	0	0	0	0	0
PTPN6	0	0	0	0	0	0	1	0
FAT1	0	0	0	0	0	0	13	0
MSH6	0	0	0	0	0	0	12	0
BIRC3	0	0	0	0	0	0	4	0
BLM	0	0	0	0	0	0	2	0
CD79A	0	0	0	0	0	0	2	0
PPM1D	0	0	0	0	0	0	1	0
BCL6	0	0	0	0	0	0	5	0
SETD2	0	0	0	0	0	0	6	0
HIST1H3B	0	0	0	0	0	0	3	0
CUX1	0	0	0	0	0	0	2	0
PMS2	0	0	0	0	0	0	5	0
CYLD	0	0	0	0	0	0	4	0
IRF1	0	0	0	0	0	0	0	0
CDKN2B	0	0	0	0	0	0	1	0
CBFB	0	0	0	0	0	0	1	0
CTCF	0	0	0	0	0	0	4	0
MTOR	0	0	0	0	0	0	5	0

PTPRD	0	0	0	0	0	0	2	0
STAT6	0	35	0	0	0	0	0	0
SGK1	0	0	0	0	26	0	0	0
BRAF	0	0	0	0	0	0	0	0
WHSC1	0	0	0	0	0	0	1	0
RHOA	0	0	0	0	0	0	0	0
KRAS	0	0	0	0	0	0	3	0
CXCR4	0	0	0	0	0	0	0	0
SF3B1	0	0	0	0	0	0	6	0
CD58	0	0	0	0	0	0	3	0
NF2	0	0	0	0	0	0	3	0
PIK3CA	0	0	0	0	0	0	2	0
U2AF1	0	0	0	0	0	0	3	0
IDH2	0	0	0	0	0	0	1	0
NRAS	0	0	0	0	0	0	4	0
ARAF	0	0	0	0	0	0	2	0
JAK3	0	0	0	0	0	0	4	0
CDKN2C	0	0	0	0	0	0	1	0
IDH1	0	0	0	0	0	0	1	0
MLH1	0	0	0	0	0	0	2	0
GNAS	0	0	0	0	0	0	1	0
JAK2	0	0	0	0	0	0	1	0
TCF3	0	0	0	0	0	0	0	0
KLF2	0	0	4	0	0	23	1	0
TERT	0	0	0	0	0	0	1	0

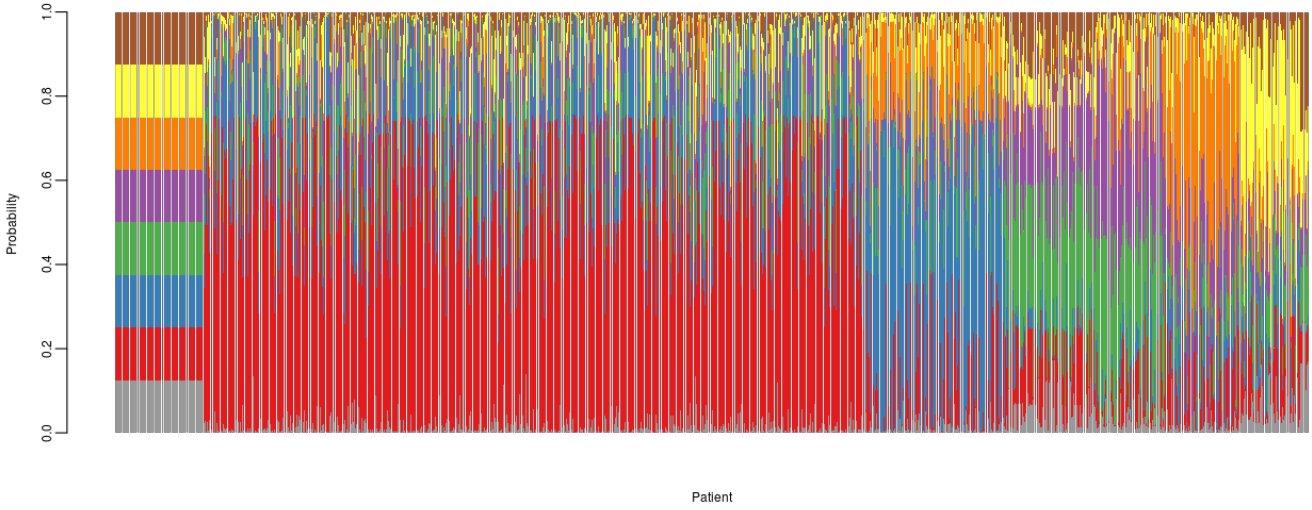
Most prevalent lesions

```
genes <- apply(posteriorMeans, 2, function(x) paste(ifelse(x>10,rownames(posteriorMeans),"")[c
genes <- gsub(";+$","",genes)
genes
```

```
##                                     0                                     1
##                          "TET2;TP53" "KMT2D;CREBBP;TNFRSF14;EZH2;ARID1A"
##                                     2                                     3
##  "KMT2D;MYD88;CREBBP;TNFRSF14;EZH2"          "TP53;SOCS1;TET2;B2M;CCND3"
##                                     4                                     5
##       "SOCS1;B2M;TP53;TET2;TNFAIP3"    "TNFAIP3;NOTCH2;B2M;BCL10;MYD88"
##                                     6                                     7
##                     "TP53;MYC;FAT1"                   "TP53;TET2;KMT2D"
```

Assignment from posterior samples

```
library(RColorBrewer)
col <- c(brewer.pal(9,"Set1")[c(9,1:8)], brewer.pal(8,"Dark2"))
posteriorProbability <- apply(sapply(posteriorMerged$sigs_nd_by_dp, colMeans)[,-1],2,function
o <- order(apply(posteriorProbability,2,which.max))
barplot(posteriorProbability[,o], col=col, border=NA, ylab="Probability", xlab="Patient")
```




```
data.frame(Prob=rowMeans(posteriorProbability), genes)
```

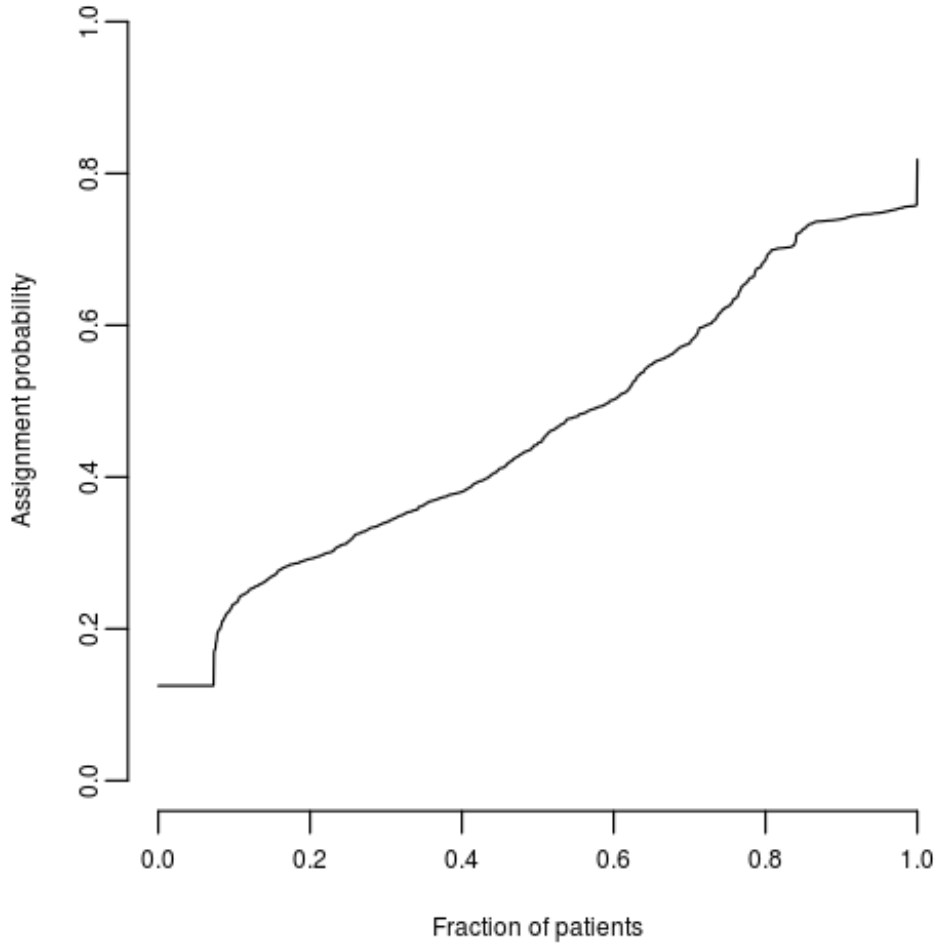
```
##         Prob                              genes
## 0 0.03245091                          TET2;TP53
## 1 0.35612484 KMT2D;CREBBP;TNFRSF14;EZH2;ARID1A
## 2 0.18731913  KMT2D;MYD88;CREBBP;TNFRSF14;EZH2
## 3 0.11346064         TP53;SOCS1;TET2;B2M;CCND3
## 4 0.10614709       SOCS1;B2M;TP53;TET2;TNFAIP3
## 5 0.08139741    TNFAIP3;NOTCH2;B2M;BCL10;MYD88
## 6 0.07549752                     TP53;MYC;FAT1
## 7 0.04760247                  TP53;TET2;KMT2D
```

Classes

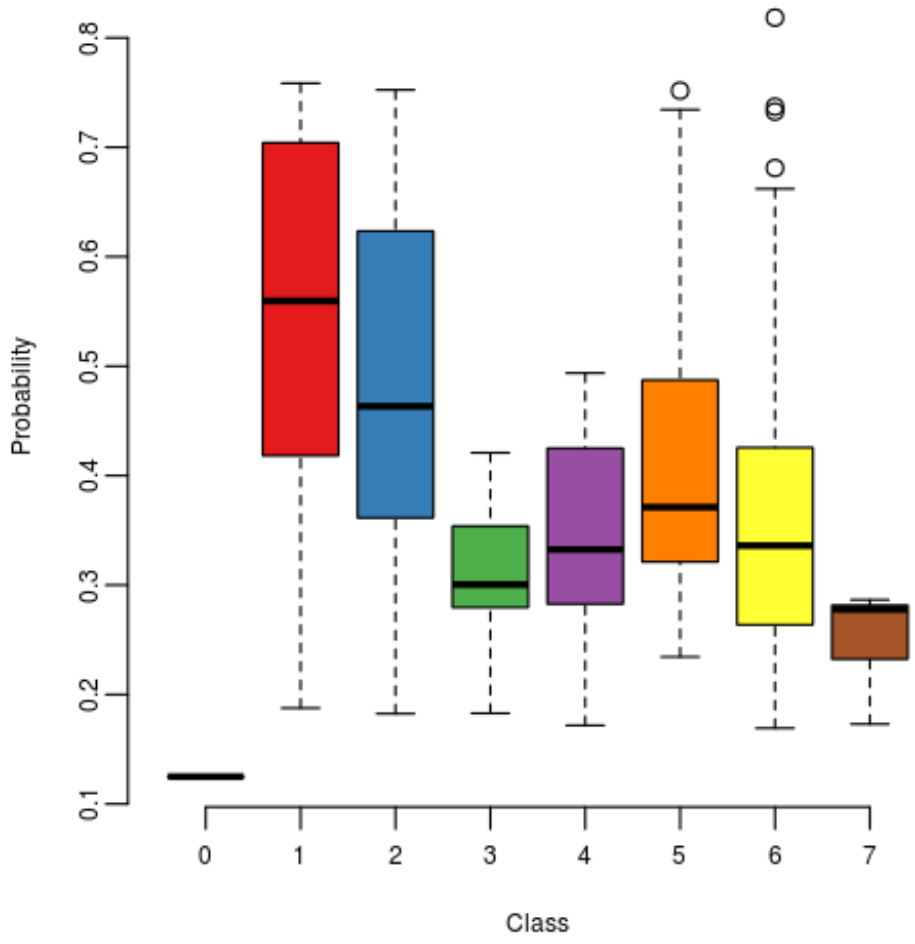
```
dpClass <- factor(apply(posteriorProbability, 2, which.max)-1)
table(dpClass)
```

```
## dpClass
##   0   1   2   3   4   5   6   7
## 118 890 190 126  90 102  80  11
```

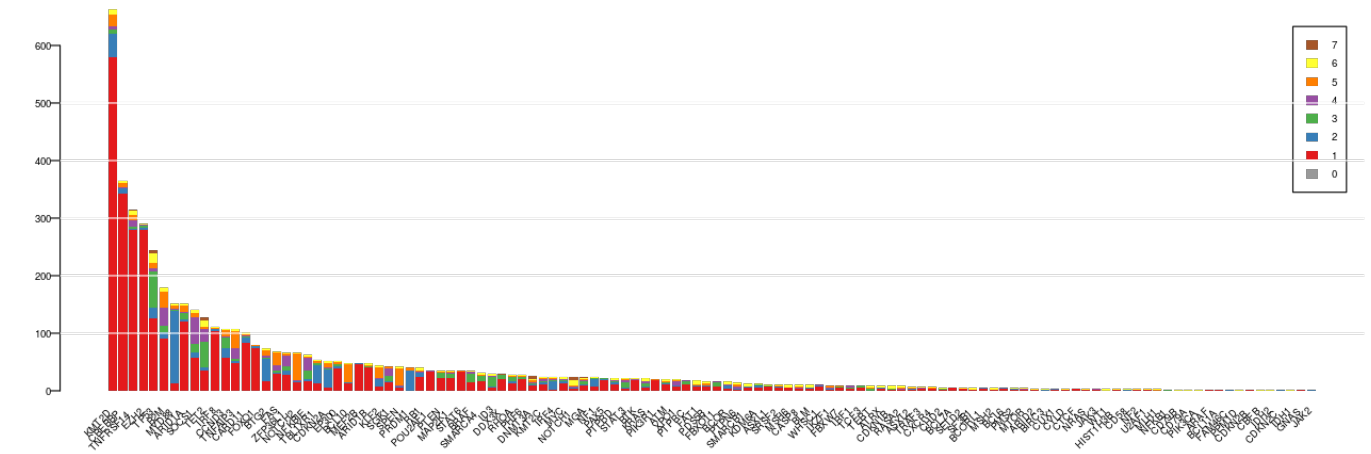
```
plot(seq(0,1,l=ncol(posteriorProbability)),sort(apply(posteriorProbability,2,max)), type='l',
```




```
boxplot(apply(posteriorProbability,2,max) ~ dpClass, col=col, ylab="Probability", xlab="Class
```

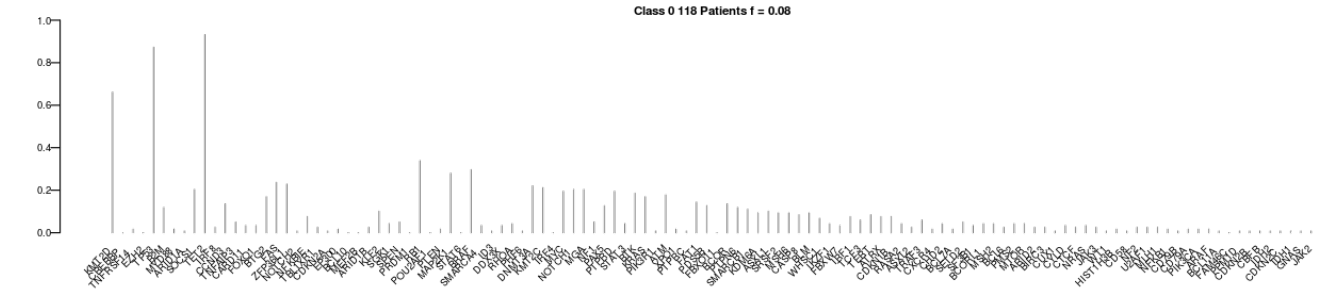


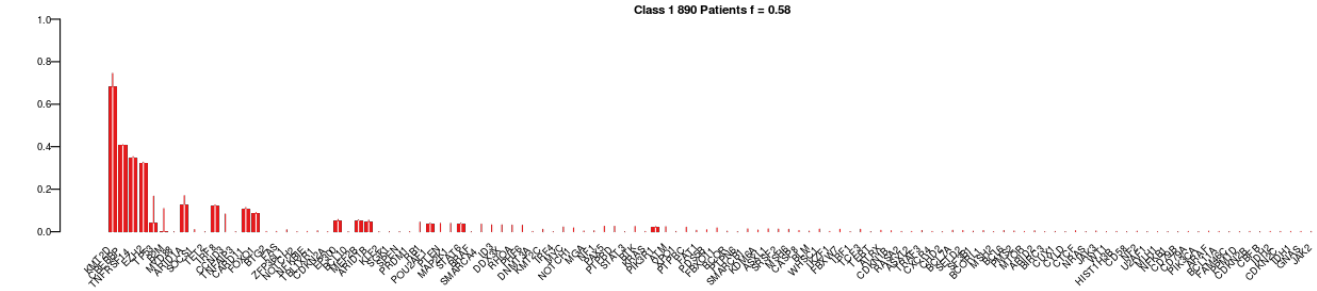

```
par(mar=c(6,3,1,1)+.1, cex=.8)
o <- order(colSums(genotypesImputed), decreasing=TRUE)
driverPrevalence <- t(sapply(split(as.data.frame(as.matrix(genotypesImputed)), dpClass), colS
b <- barplot(driverPrevalence, col=col, las=2, legend=TRUE, border=NA, args.legend=list(borde
abline(h=seq(100,500,100), col="white")
rotatedLabel(b, labels=colnames(genotypesImputed)[o])
```

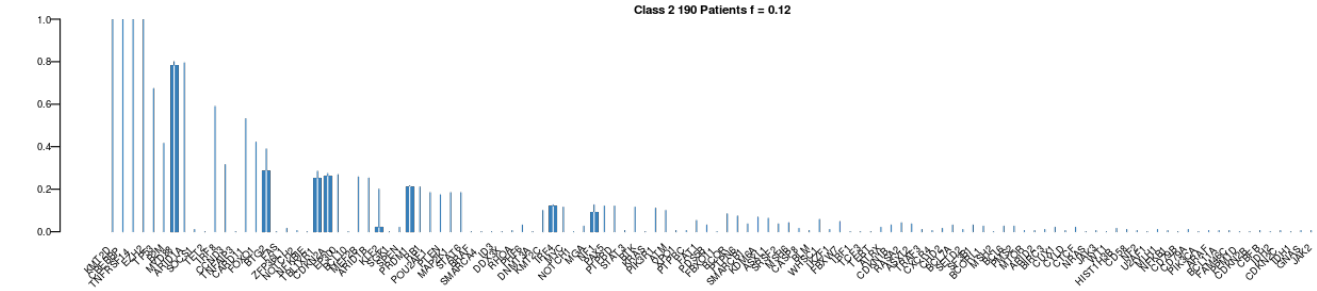


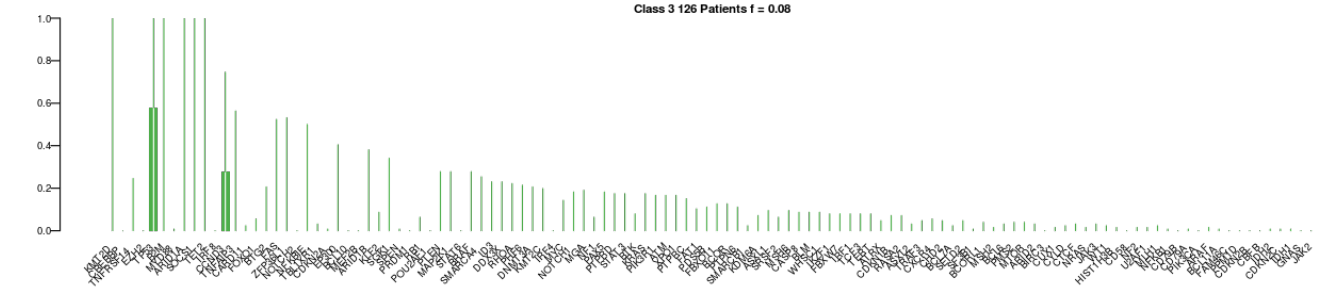
Driver signatures

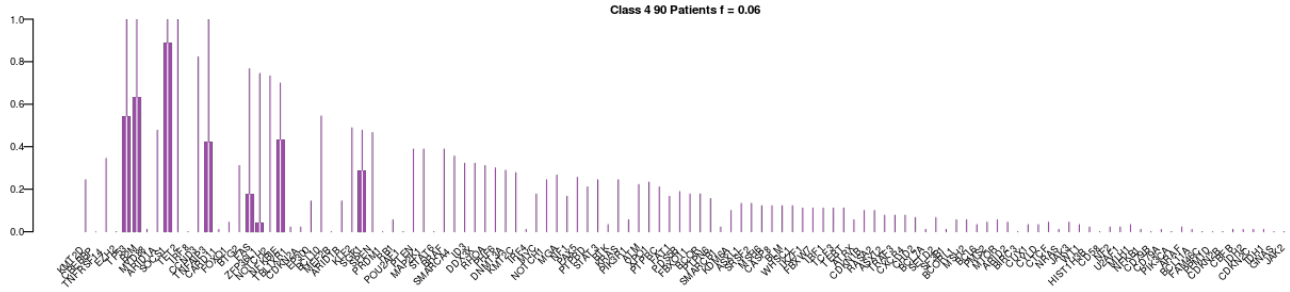
```
par(mar=c(6,3,1,1)+.1, cex=.8)
t <- table(dpClass)
i <- 0; for(c in levels(dpClass)){i <- 1+i
b <- barplot(posteriorQuantiles[2,o,c]/t[i], col=col[i], las=2, legend=FALSE, border=NA,  name
segments(b, posteriorQuantiles[1,o,c]/t[i], b, posteriorQuantiles[2,o,c]/t[i], col="white")
segments(b, posteriorQuantiles[2,o,c]/t[i], b, posteriorQuantiles[3,o,c]/t[i], col=col[i])
rotatedLabel(b, labels=colnames(genotypesImputed)[o])
}
```

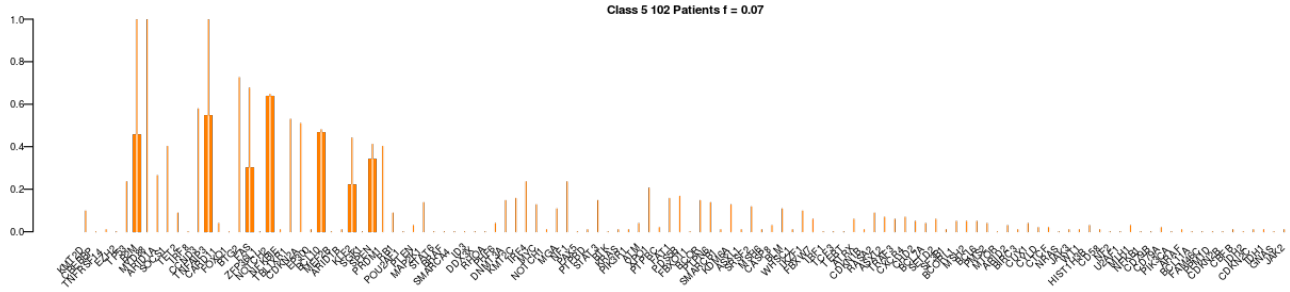


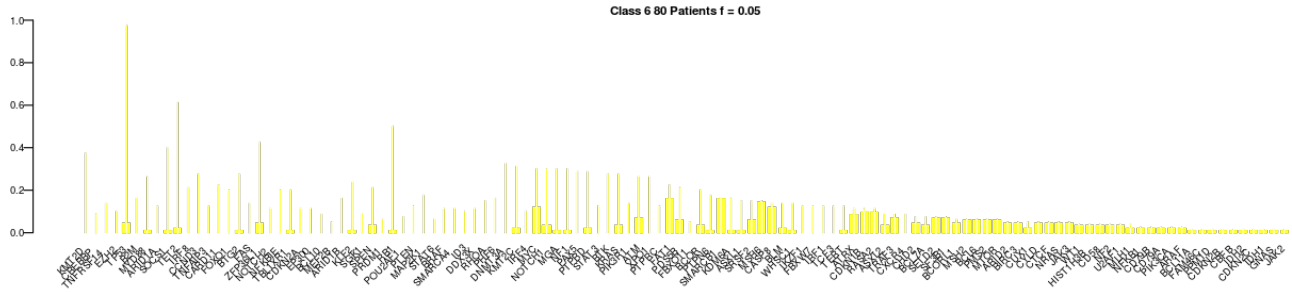


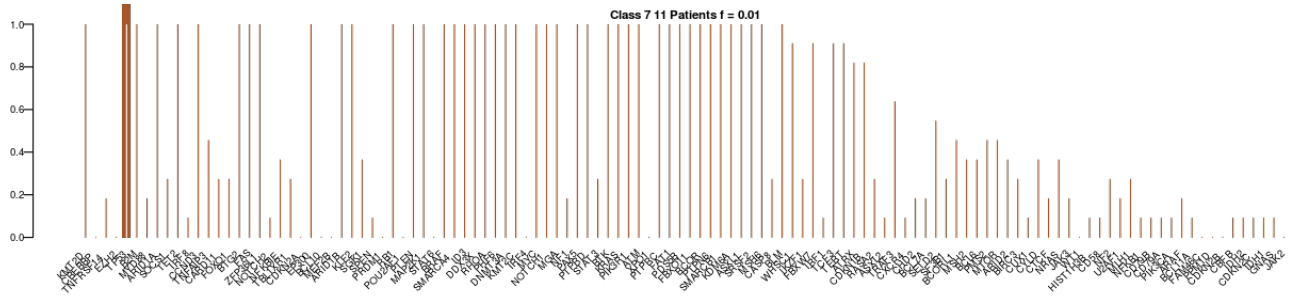






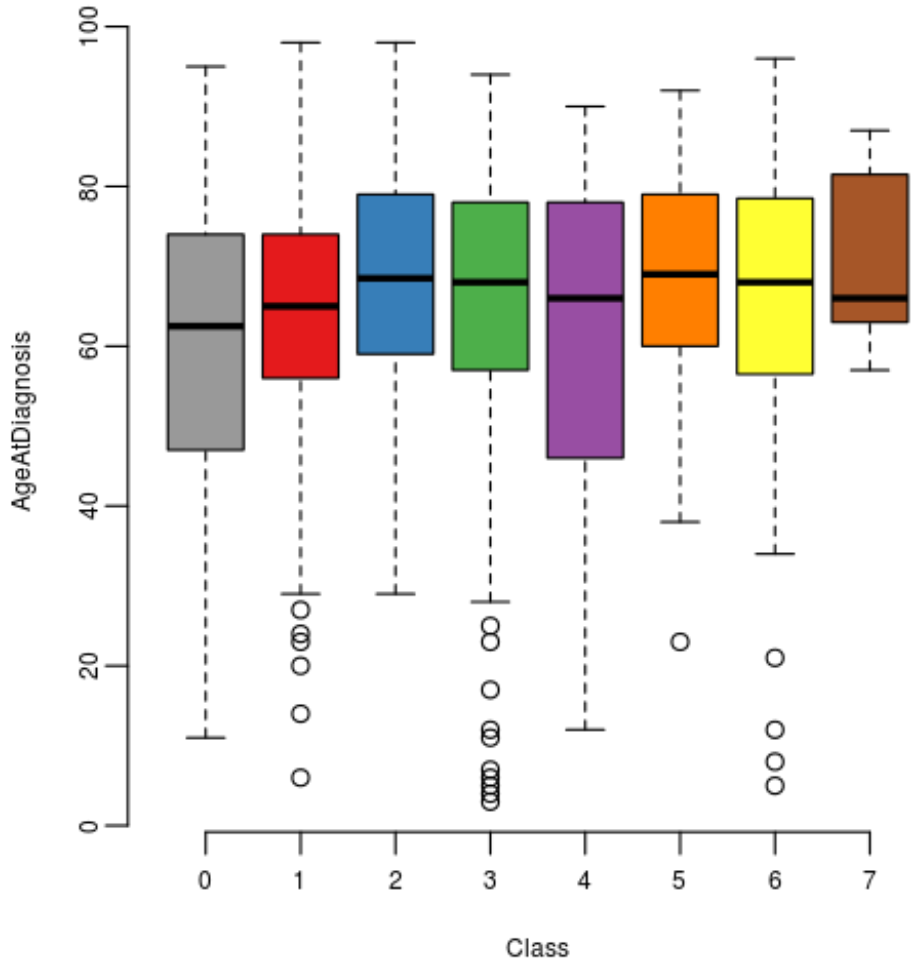




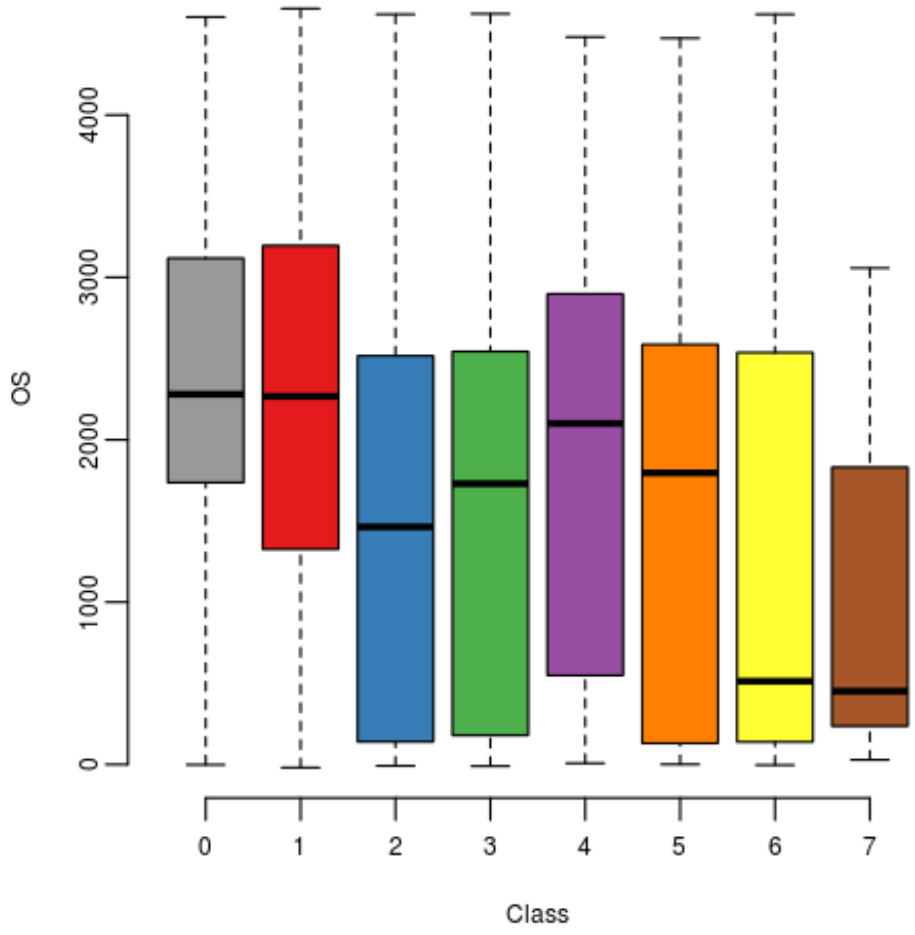


Clinical associations

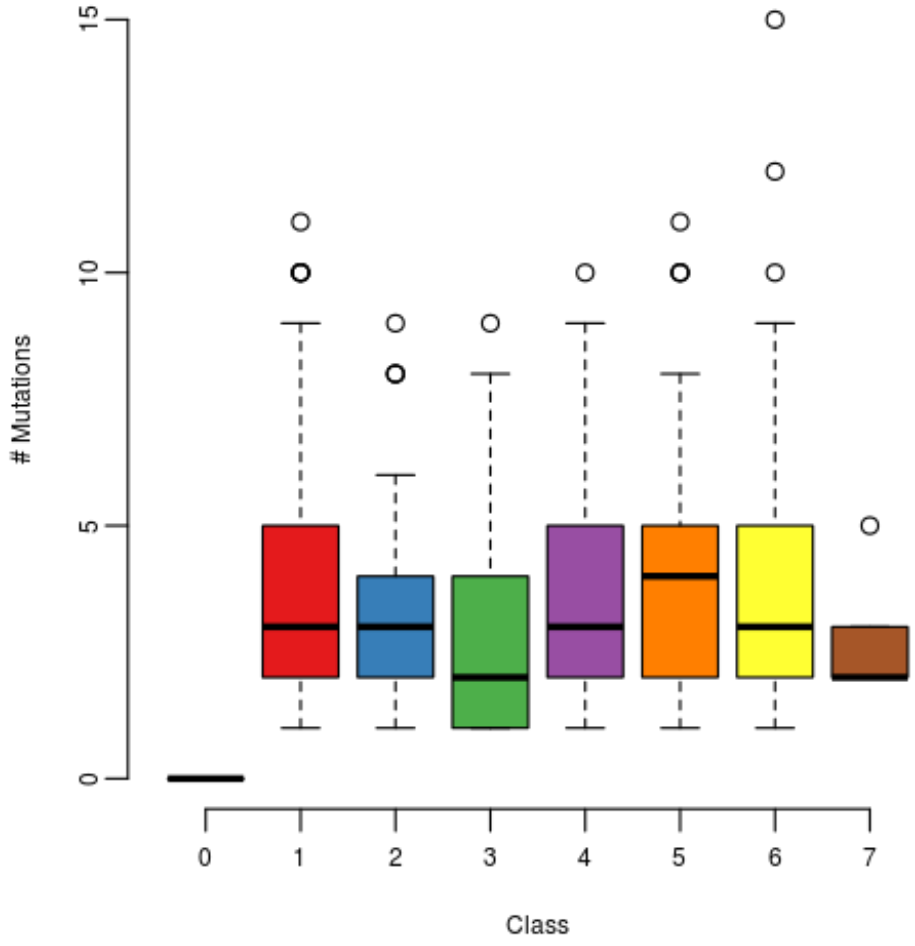
```
# Numerical
boxplot(clinical_information$AgeAtDiagnosis ~ factor(dpClass), xlab = "Class", ylab = "AgeAtD
```



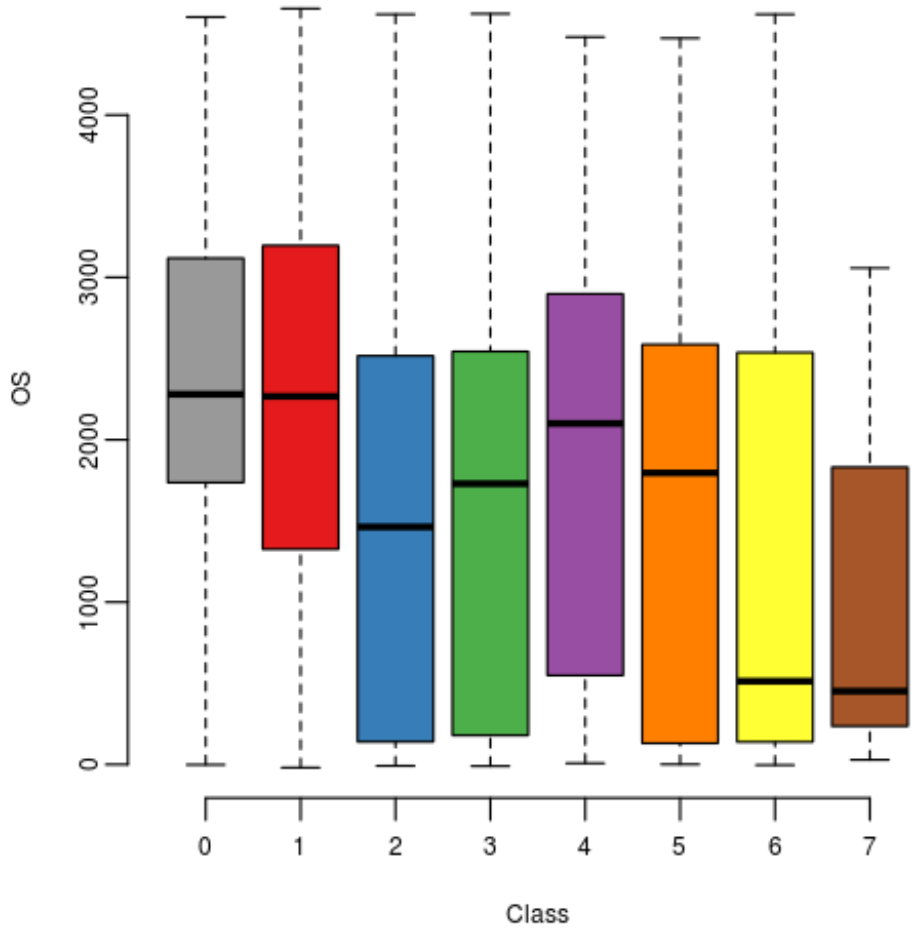

```
boxplot(clinical_information$OS ~ factor(dpClass), xlab = "Class", ylab = "OS", col=col)
```



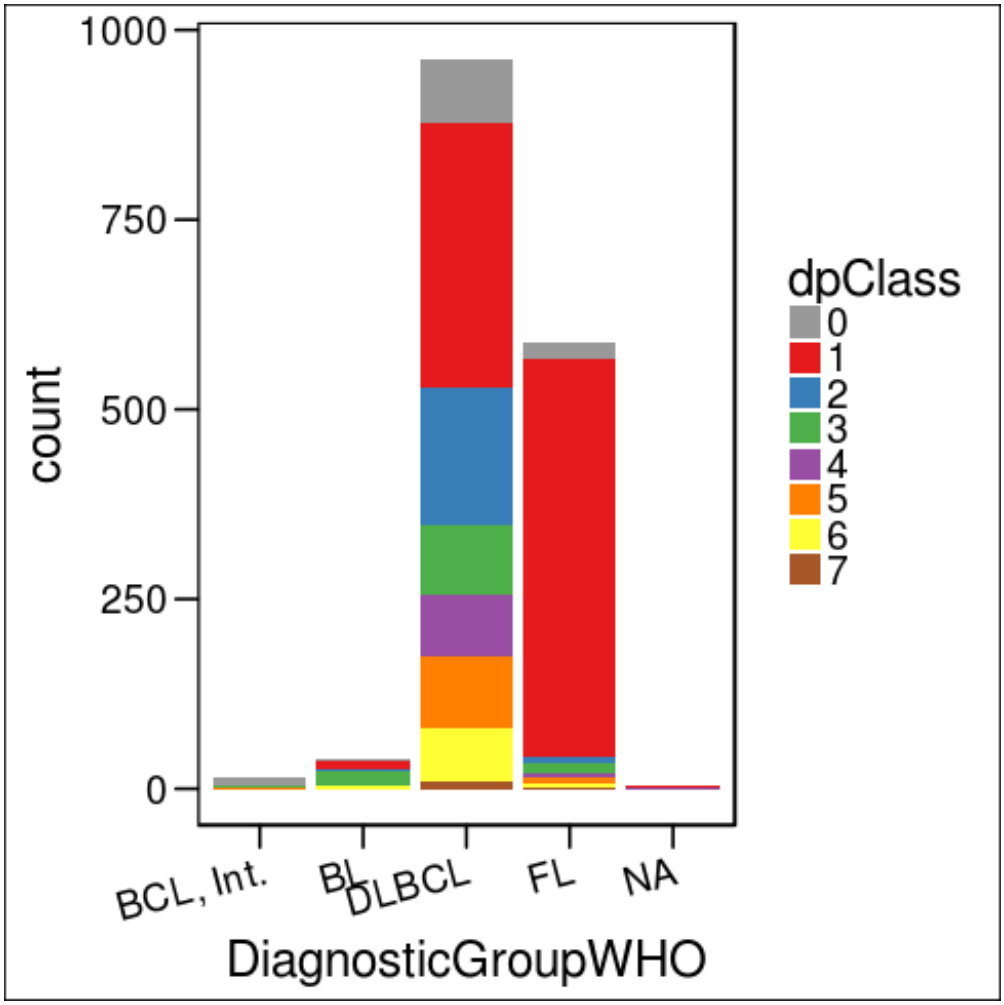

```
boxplot(rowSums(genotypesImputed) ~ factor(dpClass), xlab="Class",ylab="# Mutations", col=col
```



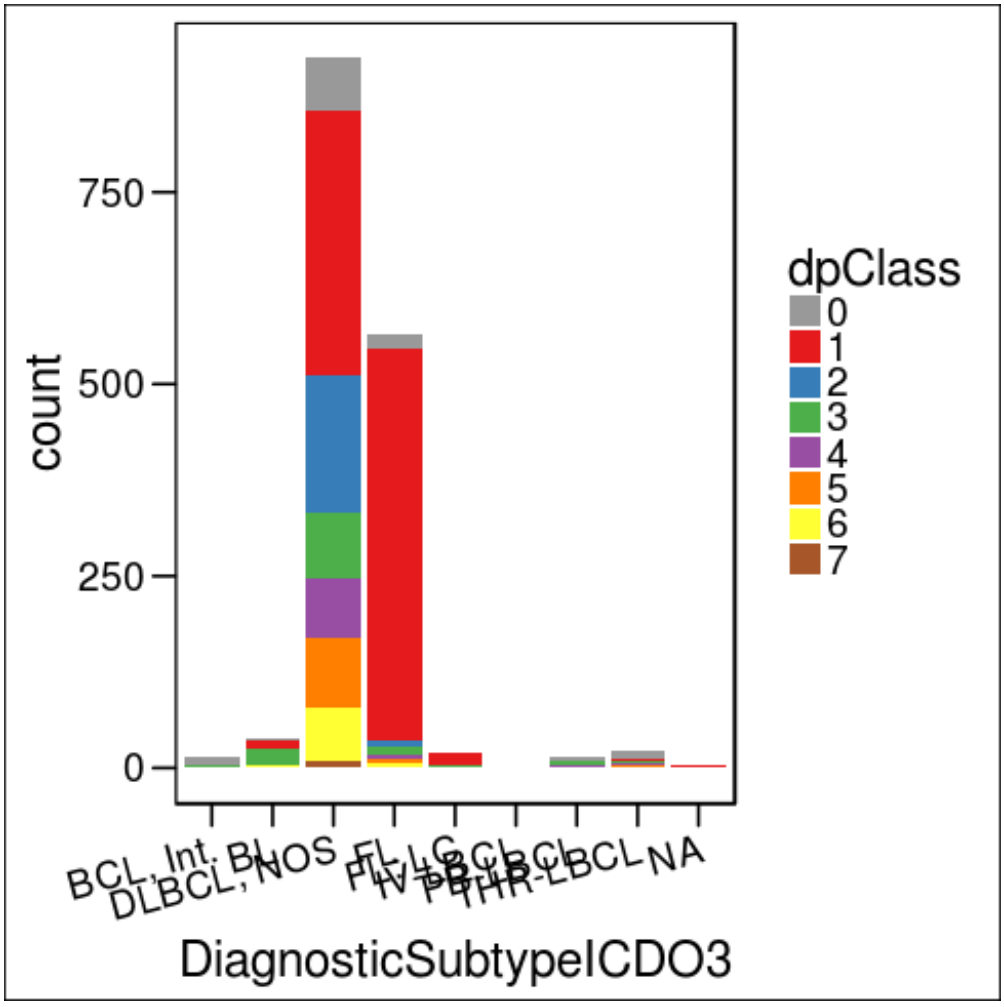

```
boxplot(clinical_information$OS ~ factor(dpClass), xlab = "Class", ylab = "OS", col=col)
```



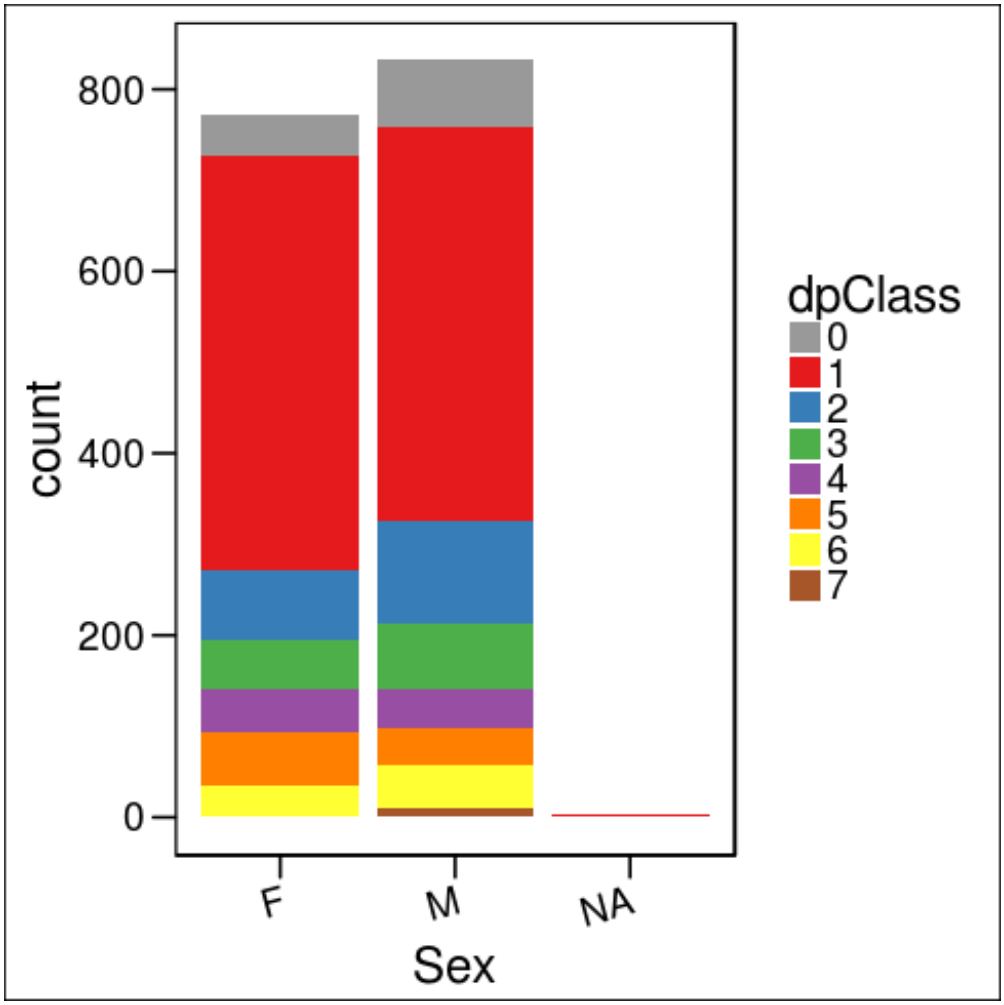

```
# Categorical
categorical_df <- cbind(clinical_information, dpClass = factor(dpClass))
# First, see results incorporating total counts
ggplot(categorical_df, aes(x = DiagnosticGroupWHO, fill = dpClass)) + geom_bar() + scale_fill
```



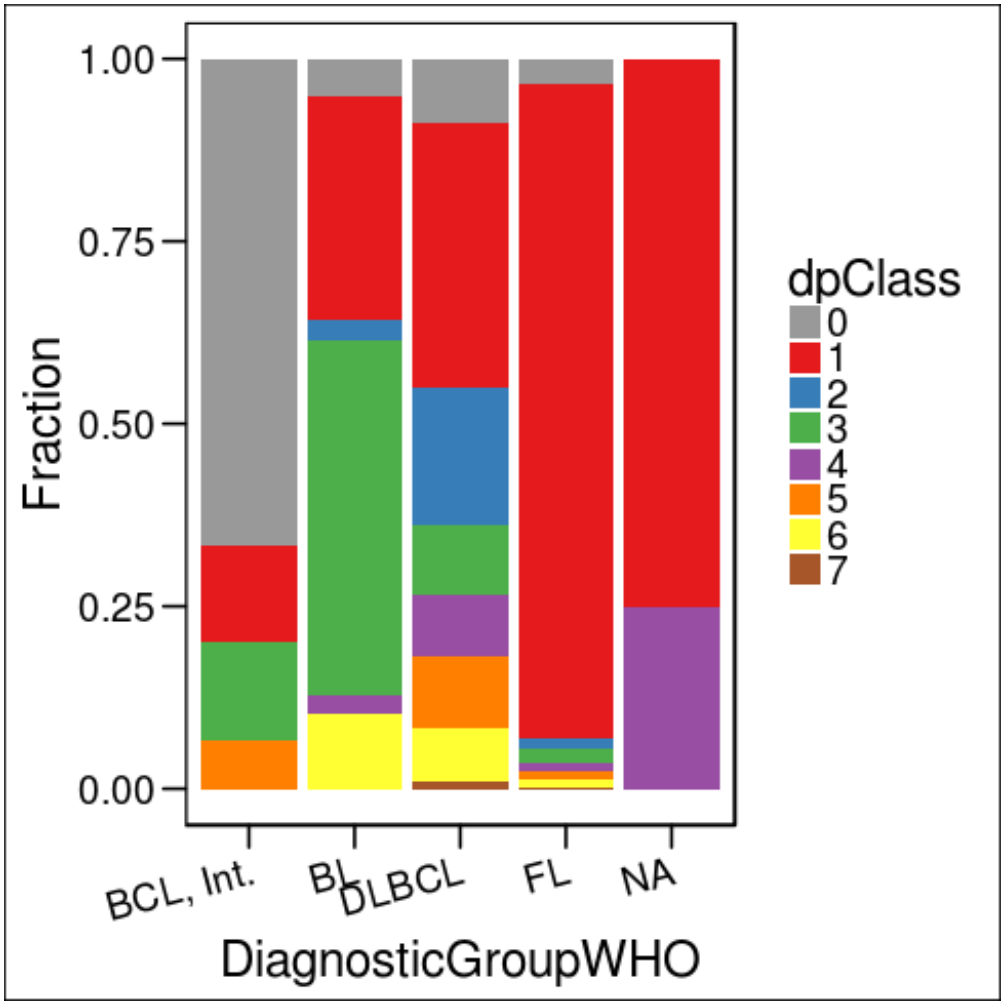

```
ggplot(categorical_df, aes(x = DiagnosticSubtypeICDO3, fill = dpClass)) + geom_bar() + scale_
```



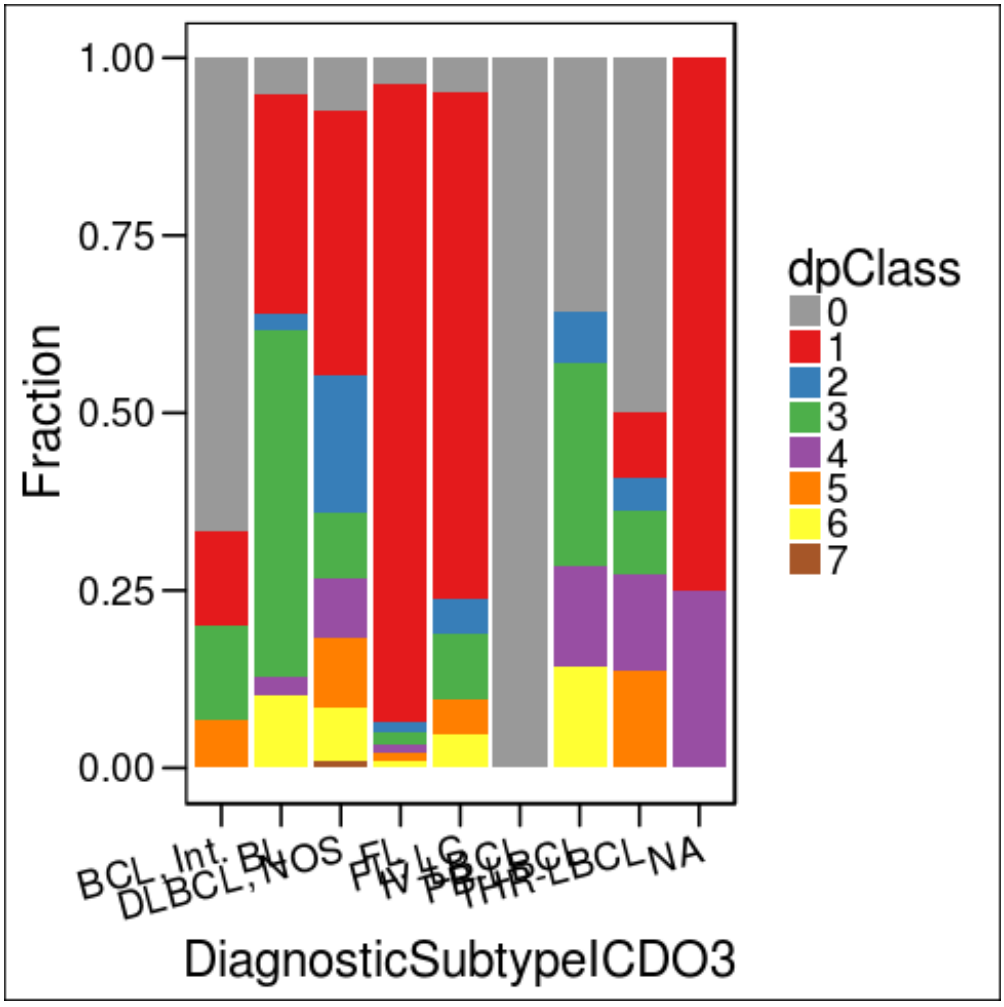

```
ggplot(categorical_df, aes(x = Sex, fill = dpClass)) + geom_bar() + scale_fill_manual(values
```



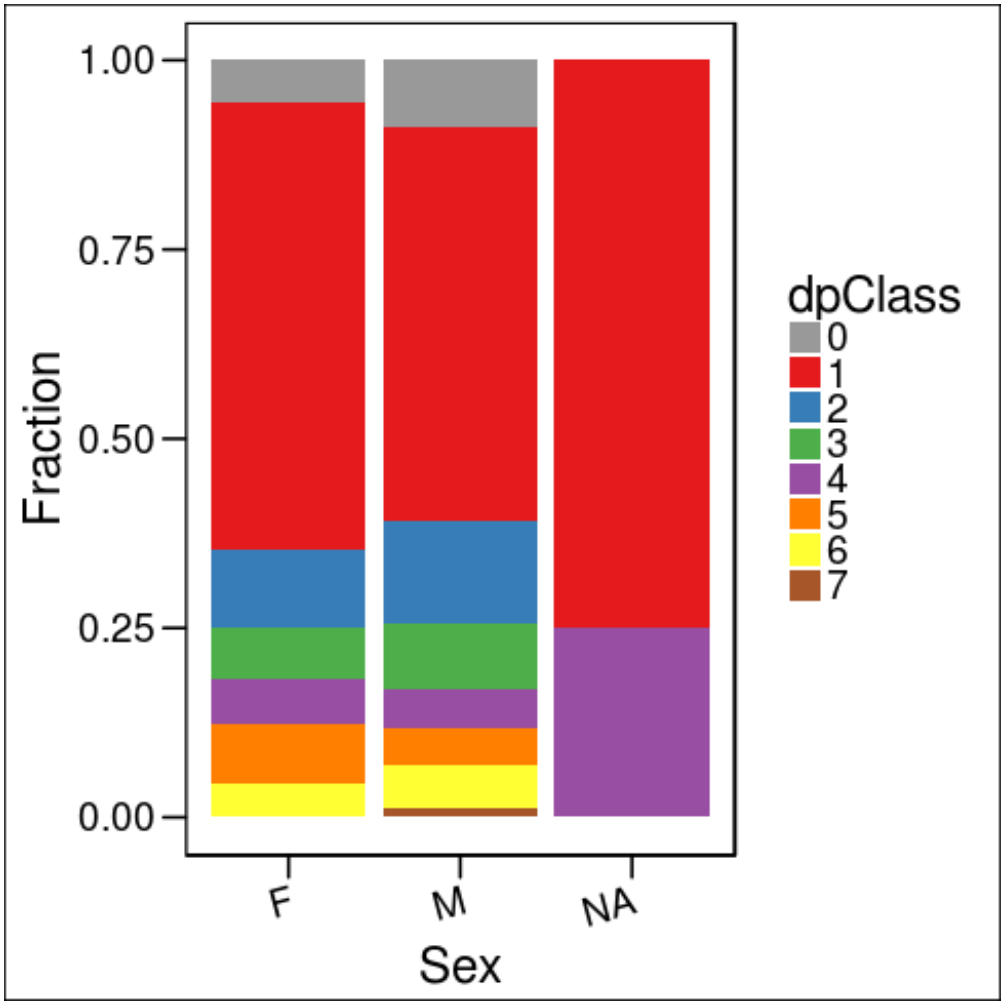

```
# Next, see results just looking at percentages
x_labels_rotation_angle <- 15
ggplot(categorical_df, aes(x = DiagnosticGroupWHO, fill = dpClass)) + geom_bar(position="fill
```




```
ggplot(categorical_df, aes(x = DiagnosticSubtypeICDO3, fill = dpClass)) + geom_bar(position="
```



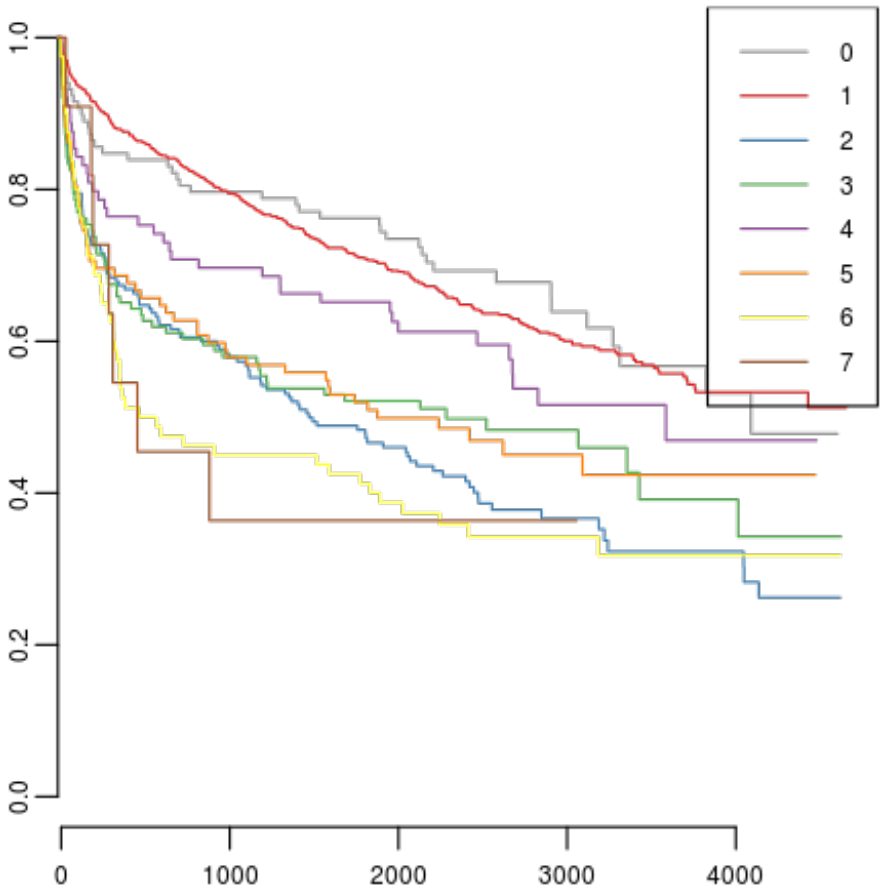

```
ggplot(categorical_df, aes(x = Sex, fill = dpClass)) + geom_bar(position="fill") + scale_fill
```



Survival

Simple coxph

```
# Actual survival code
os <- Surv(time = clinical_information$OS, event = clinical_information$SurvivalStatus)

plot(survfit(os ~ dpClass), col=col)
legend("topright", legend = levels(dpClass), col=col, lty=1)
```




```
kable(summary(survfit(os ~ dpClass))$table)
```

	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
dpClass=0	118	118	118	43	3179.909	173.27050	4091.0	3271	NA
dpClass=1	887	887	887	346	3136.705	62.45958	NA	3756	NA
dpClass=2	190	190	190	121	2079.030	141.87390	1470.0	1108	2399
dpClass=3	126	126	126	67	2327.578	189.45547	2283.0	957	NA
dpClass=4	89	89	89	40	2789.888	215.21273	3585.0	2466	NA
dpClass=5	102	102	102	55	2378.981	207.44751	1876.0	982	NA
dpClass=6	80	80	80	53	1870.553	225.21687	513.5	318	2023
dpClass=7	11	11	11	7	1887.727	623.81602	452.0	283	NA

```
summary(coxph(os ~ dpClass))
```

```
## Call:
## coxph(formula = os ~ dpClass)
##
##   n= 1603, number of events= 732
##    (4 observations deleted due to missingness)
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## dpClass1 0.05994   1.06178  0.16171 0.371 0.710867
## dpClass2 0.82397   2.27952  0.17764 4.638 3.51e-06 ***
## dpClass3 0.64791   1.91154  0.19548 3.314 0.000918 ***
## dpClass4 0.31127   1.36515  0.21969 1.417 0.156533
## dpClass5 0.64184   1.89998  0.20364 3.152 0.001623 **
## dpClass6 0.94815   2.58094  0.20542 4.616 3.92e-06 ***
## dpClass7 1.01213   2.75146  0.40796 2.481 0.013102 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## dpClass1     1.062     0.9418    0.7734     1.458
## dpClass2     2.280     0.4387    1.6093     3.229
## dpClass3     1.912     0.5231    1.3031     2.804
## dpClass4     1.365     0.7325    0.8875     2.100
## dpClass5     1.900     0.5263    1.2747     2.832
## dpClass6     2.581     0.3875    1.7255     3.860
## dpClass7     2.751     0.3634    1.2368     6.121
##
## Concordance= 0.598  (se = 0.01 )
## Rsquare= 0.053   (max possible= 0.998 )
## Likelihood ratio test= 87.52  on 7 df,   p=4.441e-16
## Wald test            = 93.49  on 7 df,   p=0
## Score (logrank) test = 97.93  on 7 df,   p=0
```

#' Risk variance #+ RFX, cache=TRUE library(CoxHD) dataFrameOsTD <- dataFrame[tplSplitOs,] dataFrameOsTD[which(tplIndexOs), grep("TPL", colnames(dataFrameOsTD), value=TRUE)] <- 0 ## Set pre-tpl variables to zero mainGroups <- grep("[A-Z][a-z]+[A-Z]",levels(groups), invert=TRUE, value=TRUE) mainIdx <- groups %in% mainGroups osTDIdx <- !grepl("TPL_efs", colnames(dataFrame)) mainIdxOsTD <- mainIdx & osTDIdx whichRFXOsTDGG <- which((colSums(dataFrame)>=8 | mainIdxOsTD) & osTDIdx & groups %in% c(mainGroups,"GeneGene")) # ie, > 0.5%

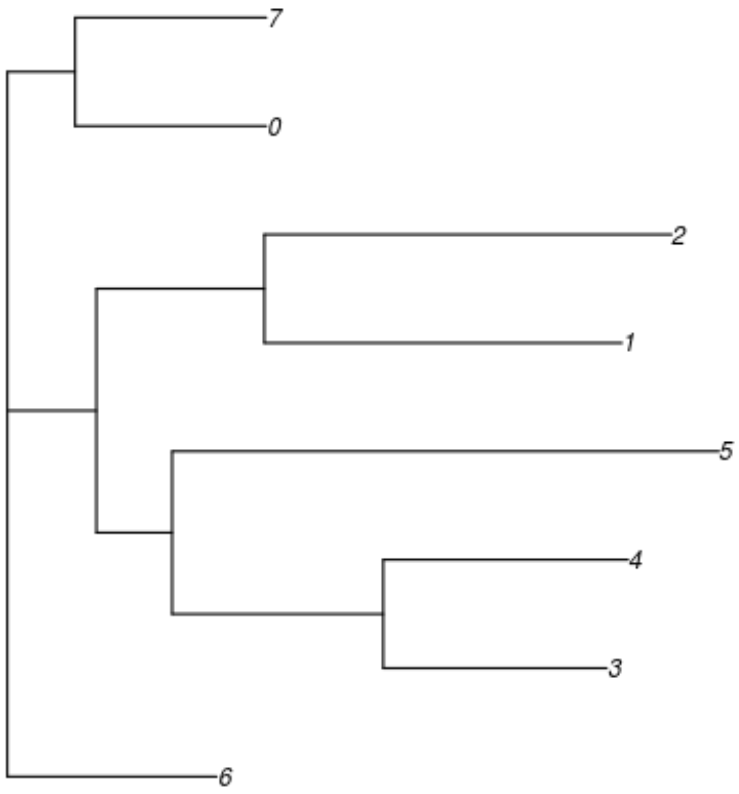
coxRFXFitOsTDGGc <- CoxRFX(dataFrameOsTD[,whichRFXOsTDGG], osTD, groups[whichRFXOsTDGG], which.mu=mainGroups) ## allow only the main groups to have mean different from zero.. coxRFXFitOsTDGGc

d <- cbind(dataFrameOsTD[,whichRFXOsTDGG],DP=t(posteriorProbability)[tplSplitOs,-1]) coxRFXFitOsTDGGcDP <- CoxRFX(d, osTD, c(as.character(groups[whichRFXOsTDGG]),rep("DP", nlevels(dpClass)-1)), which.mu=mainGroups) ## allow only the main groups to have mean different from zero.. coxRFXFitOsTDGGcDP

PlotVarianceComponents(coxRFXFitOsTDGGcDP, col=col) round(cov(PartialRisk(coxRFXFitOsTDGGcDP)),2)

Phylogeny

```
library(ape)
plot(nj(dist(t(posteriorMeans/(rep(rowSums(posteriorProbability), each=nrow(posteriorMeans)))
```

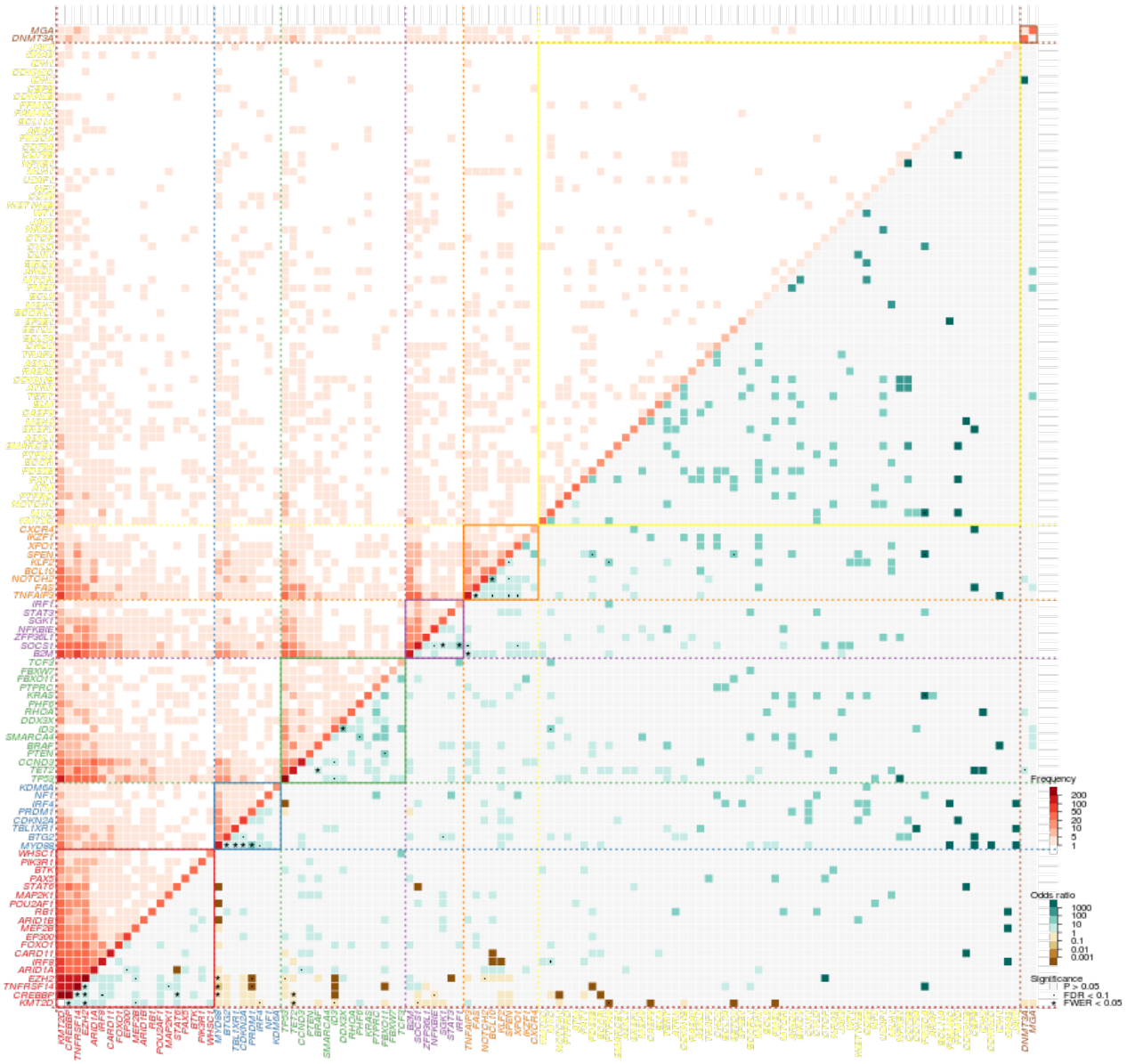


Gene:Gene interactions

Population based

```
geneToClass <- factor(apply(posteriorMeans, 1,which.max) -1, levels = as.numeric(colnames(pos
getOdds <- function(x) {
  f <- sapply(1:ncol(x),
              function(i) sapply(1:ncol(x),
                                 function(j) {
                                   if(j <= i) return(c(NA,NA))
                                   f<- try(fisher.test(x[,i], x[,j]), silent=TRUE)
                                   if(class(f)=="try-error") c(0,NA)
                                   else if(f$estimate>1) c(-log10(f$p.val),f$estimate)
                                   else c(log10(f$p.val), f$estimate)}
              ),
              simplify="array")
  for(i in 1:2)
    f[i,,][upper.tri(f[i,,])] <- t(f[i,,])[upper.tri(f[i,,])]
  return(f)
}
f <- getOdds(genotypesImputed)
logPInt <- f[1,,]
odds <- f[2,,]
pairs <- sapply(1:ncol(genotypesImputed), function(i) colMeans(genotypesImputed * genotypesIm
diag(logPInt) <- 0
diag(odds) <- 1
colnames(odds) <- rownames(odds) <- colnames(logPInt) <- rownames(logPInt) <- colnames(genoty
odds[odds<1e-3] = 1e-4
odds[odds>1e3] = 1e4
```

```
odds[10^-abs(logPInt) > 0.1] = 1
logOdds=log10(odds)
diag(logPInt) <- NA
par(bty="n", mgp = c(2,.5,0), mar=c(4,4,4,4)+.1, las=2, tcl=-.33)
ix = TRUE#colnames(interactions) %in% colnames(all_genotypes)
o = order(geneToClass, -colSums(genotypesImputed))
M <-  matrix( NA, ncol=ncol(odds), nrow=nrow(odds))
M[lower.tri(M)] <- cut(logOdds[o,o][lower.tri(M)], breaks = c(-4:0-.Machine$double.eps,0:4), i
M[upper.tri(M, diag=TRUE)] <- as.numeric(cut(pairs[o,o][upper.tri(M, diag=TRUE)]*nrow(genotype
image(x=1:ncol(logPInt), y=1:nrow(logPInt), M, col=c(brewer.pal(9,"BrBG"), c("white",brewer.pa
l <- colnames(logPInt)[o]
mtext(side=1, at=1:ncol(logPInt), l, cex=.66, font=ifelse(grepl("^[A-Z]",l),3,1), col=col[gene
mtext(side=2, at=1:ncol(logPInt), l, cex=.66, font=ifelse(grepl("^[A-Z]",l),3,1), col=col[gene
abline(h=0:ncol(logPInt)+.5, col="white", lwd=.5)
abline(v=0:ncol(logPInt)+.5, col="white", lwd=.5)
P <- 10^-abs(logPInt[o,o])
P[upper.tri(P)] <- NA
w = arrayInd(which(p.adjust(P, method="BH") < .1), rep(nrow(logPInt),2))
points(w, pch=".", col="black")
w = arrayInd(which(p.adjust(P) < .05), rep(nrow(logPInt),2))
points(w, pch="*", col="black")
image(y = 1:9 +18, x=rep(ncol(logPInt),2)+c(2,3), z=matrix(c(1:8), nrow=1), col=c("white",brew
axis(side = 4, at = seq(1,7) + 19, cex.axis=.66, tcl=-.15, label=c(1,5,10,20,50,100,200), las=
mtext(side=4, at=28, "Frequency", las=2, line=-1,cex=.66)
image(y = 1:8 +5, x=rep(ncol(logPInt),2)+c(2,3), z=matrix(c(1:8), nrow=1), col=brewer.pal(8,"B
axis(side = 4, at = seq(1,7) + 5.5, cex.axis=.66, tcl=-.15, label=10^seq(-3,3), las=1, lwd=.5)
mtext(side=4, at=14, "Odds ratio", las=2, line=-1,cex=.66)
mtext(side=4, at=4, "Significance", las=2, line=-1,cex=.66)
points(x=rep(ncol(logPInt),2)+2.5, y=1:2, pch=c("*","."))
image(x=rep(ncol(logPInt),2)+c(2,3), y=(2:3) +0.5, z=matrix(1), col=brewer.pal(3,"BrBG"), add=
mtext(side=4, at=3:1, c("P > 0.05", "FDR < 0.1", "FWER < 0.05"), cex=.66, line=0.2)
t <- c(0,table(geneToClass))
s <- cumsum(t)+.5
abline(h=s[-length(s)], col=col, lty=3)
abline(v=s[-length(s)], col=col, lty=3)
rect(s[-1],s[-1], s[-length(s)], s[-length(s)], border=col)
```



Expected heatmap

```
set.seed(42)
t <- table(dpClass)
pp <- t(t(posteriorMeans)/as.numeric(t))
expectedOdds <- sapply(colnames(genotypesImputed), function(j){
  sapply(colnames(genotypesImputed), function(i){
    if(i==j) return(0)
    P <- Reduce("+",lapply(seq_along(t), function(k) {t[k] * (pp[i,k] * c(1,-1) + c(0,1)) %o%
    #res <- round(log10(M[1,1]*M[2,2]/M[1,2]/M[2,1]))
    #if( sum(M[,1]) * sum(M[1,])/sum(M) < 5 & res < 0) res <- 0
    #return(res)
    M <- matrix(rmultinom(1,sum(t), P), ncol=2)
    f <- fisher.test(round(M))
    res <- pmin(pmax(round(log10(f$estimate)),-4),4)
    if(f$p.value > 0.05)
      res <- 0
    return(res)
  })
})
```

```
## Error in rmultinom(1, sum(t), P): negative probability
```

```
par(bty="n", mgp = c(2,.5,0), mar=c(4,4,4,4)+.1, las=2, tcl=-.33)
o = order(geneToClass, -colSums(genotypesImputed))
image(x=1:ncol(expectedOdds), y=1:nrow(expectedOdds), expectedOdds[o,o], col=brewer.pal(9,"Br
```

```
## Error in ncol(expectedOdds): object 'expectedOdds' not found
```

```
l <- colnames(expectedOdds)[o]
```

```
## Error in is.data.frame(x): object 'expectedOdds' not found
```

```
mtext(side=1, at=1:ncol(expectedOdds), l, cex=.66, font=ifelse(grepl("^[A-Z]",l),3,1), col=co
```

```
## Error in ncol(expectedOdds): object 'expectedOdds' not found
```

```
mtext(side=2, at=1:ncol(expectedOdds), l, cex=.66, font=ifelse(grepl("^[A-Z]",l),3,1), col=co
```

```
## Error in ncol(expectedOdds): object 'expectedOdds' not found
```

```
abline(h=0:ncol(expectedOdds)+.5, col="white", lwd=.5)
```

```
## Error in ncol(expectedOdds): object 'expectedOdds' not found
```

```
abline(v=0:ncol(expectedOdds)+.5, col="white", lwd=.5)
```

```
## Error in ncol(expectedOdds): object 'expectedOdds' not found
```

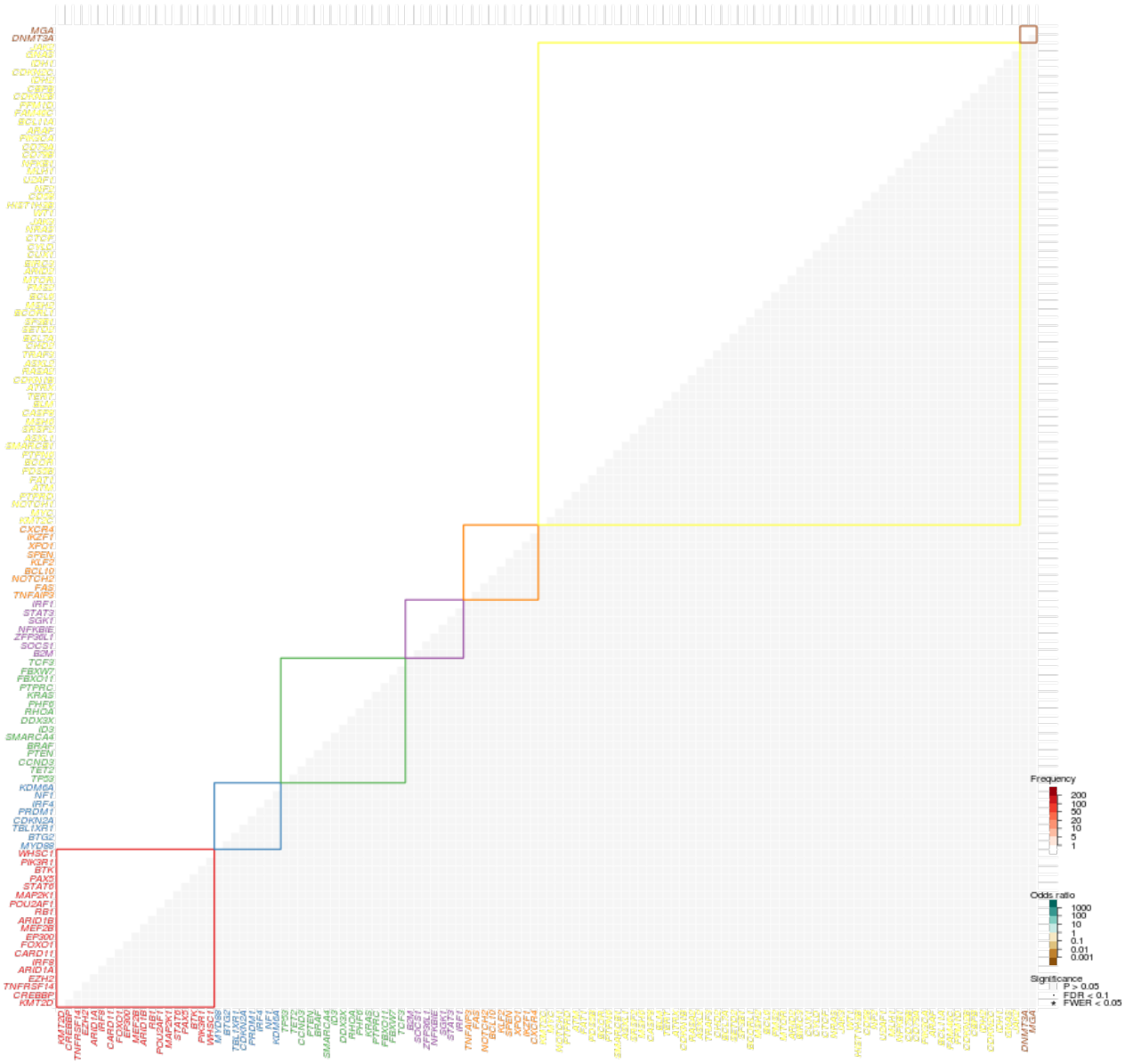
```
t <- c(0,table(geneToClass))
s <- cumsum(t)+.5
rect(s[-1],s[-1], s[-length(s)], s[-length(s)], border=col)
```

```
## Error in rect(s[-1], s[-1], s[-length(s)], s[-length(s)], border = col): plot.new has not
```

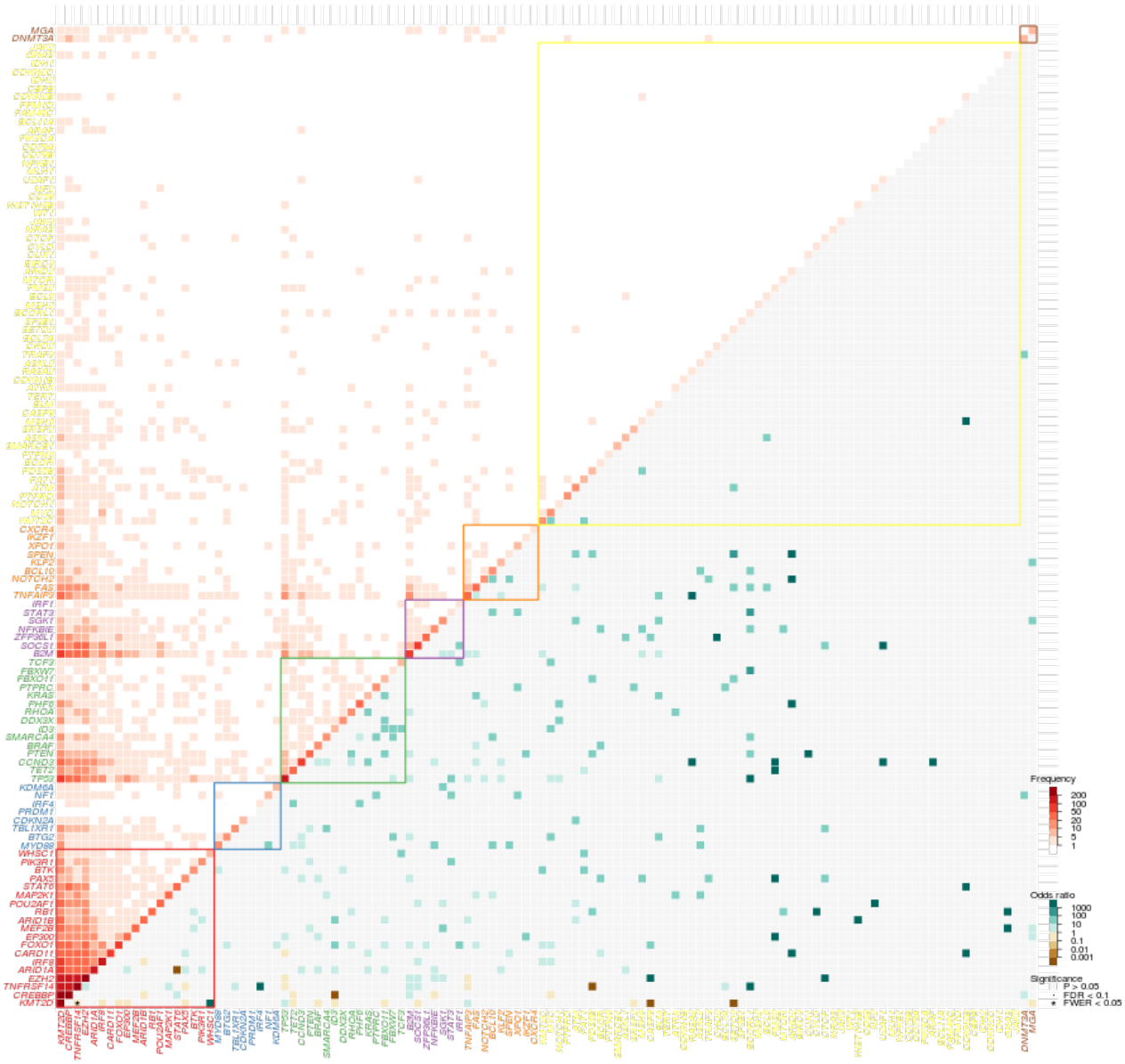
Per class

```r
for(cls in levels(dpClass)){
  w <- dpClass == cls
  f <- getOdds(genotypesImputed[w,])
  logPInt <- f[1,,]
  odds <- f[2,,]
  pairs <- sapply(1:ncol(genotypesImputed), function(i) colMeans(genotypesImputed[w,] * genot
  diag(logPInt) <- 0
  diag(odds) <- 1
  colnames(odds) <- rownames(odds) <- colnames(logPInt) <- rownames(logPInt) <- colnames(geno
  odds[odds<1e-3] = 1e-4
  odds[odds>1e3] = 1e4

  odds[10^-abs(logPInt) > 0.1] = 1
  logOdds=log10(odds)
  diag(logPInt) <- NA
  par(bty="n", mgp = c(2,.5,0), mar=c(4,4,4,4)+.1, las=2, tcl=-.33)
  ix = TRUE#colnames(interactions) %in% colnames(all_genotypes)
  o = order(geneToClass, -colSums(genotypesImputed))
  M <-  matrix( NA, ncol=ncol(odds), nrow=nrow(odds))
  M[lower.tri(M)] <- cut(logOdds[o,o][lower.tri(M)], breaks = c(-4:0-.Machine$double.eps,0:4),
  M[upper.tri(M, diag=TRUE)] <- as.numeric(cut(pairs[o,o][upper.tri(M, diag=TRUE)]*nrow(genoty
  image(x=1:ncol(logPInt), y=1:nrow(logPInt), M, col=c(brewer.pal(9,"BrBG"), c("white",brewer
  l <- colnames(logPInt)[o]
  mtext(side=1, at=1:ncol(logPInt), l, cex=.66, font=ifelse(grepl("^[A-Z]",l),3,1), col=col[ge
  mtext(side=2, at=1:ncol(logPInt), l, cex=.66, font=ifelse(grepl("^[A-Z]",l),3,1), col=col[ge
  abline(h=0:ncol(logPInt)+.5, col="white", lwd=.5)
  abline(v=0:ncol(logPInt)+.5, col="white", lwd=.5)
  P <- 10^-abs(logPInt[o,o])
  P[upper.tri(P)] <- NA
  w = arrayInd(which(p.adjust(P, method="BH") < .1), rep(nrow(logPInt),2))
  points(w, pch=".", col="black")
  w = arrayInd(which(p.adjust(P) < .05), rep(nrow(logPInt),2))
  points(w, pch="*", col="black")
  image(y = 1:9 +18, x=rep(ncol(logPInt),2)+c(2,3), z=matrix(c(1:8), nrow=1), col=c("white",br
  axis(side = 4, at = seq(1,7) + 19, cex.axis=.66, tcl=-.15, label=c(1,5,10,20,50,100,200), la
  mtext(side=4, at=28, "Frequency", las=2, line=-1,cex=.66)
  image(y = 1:8 +5, x=rep(ncol(logPInt),2)+c(2,3), z=matrix(c(1:8), nrow=1), col=brewer.pal(8,
  axis(side = 4, at = seq(1,7) + 5.5, cex.axis=.66, tcl=-.15, label=10^seq(-3,3), las=1, lwd=.
  mtext(side=4, at=14, "Odds ratio", las=2, line=-1,cex=.66)
  mtext(side=4, at=4, "Significance", las=2, line=-1,cex=.66)
  points(x=rep(ncol(logPInt),2)+2.5, y=1:2, pch=c("*","."))
  image(x=rep(ncol(logPInt),2)+c(2,3), y=(2:3) +0.5, z=matrix(1), col=brewer.pal(3,"BrBG"), ad
  mtext(side=4, at=3:1, c("P > 0.05", "FDR < 0.1", "FWER < 0.05"), cex=.66, line=0.2)
  t <- c(0,table(geneToClass))
  s <- cumsum(t)+.5
  rect(s[-1],s[-1], s[-length(s)], s[-length(s)], border=col)
  title(main=paste0("Class ",cls,": ", genes[cls]))
}
```

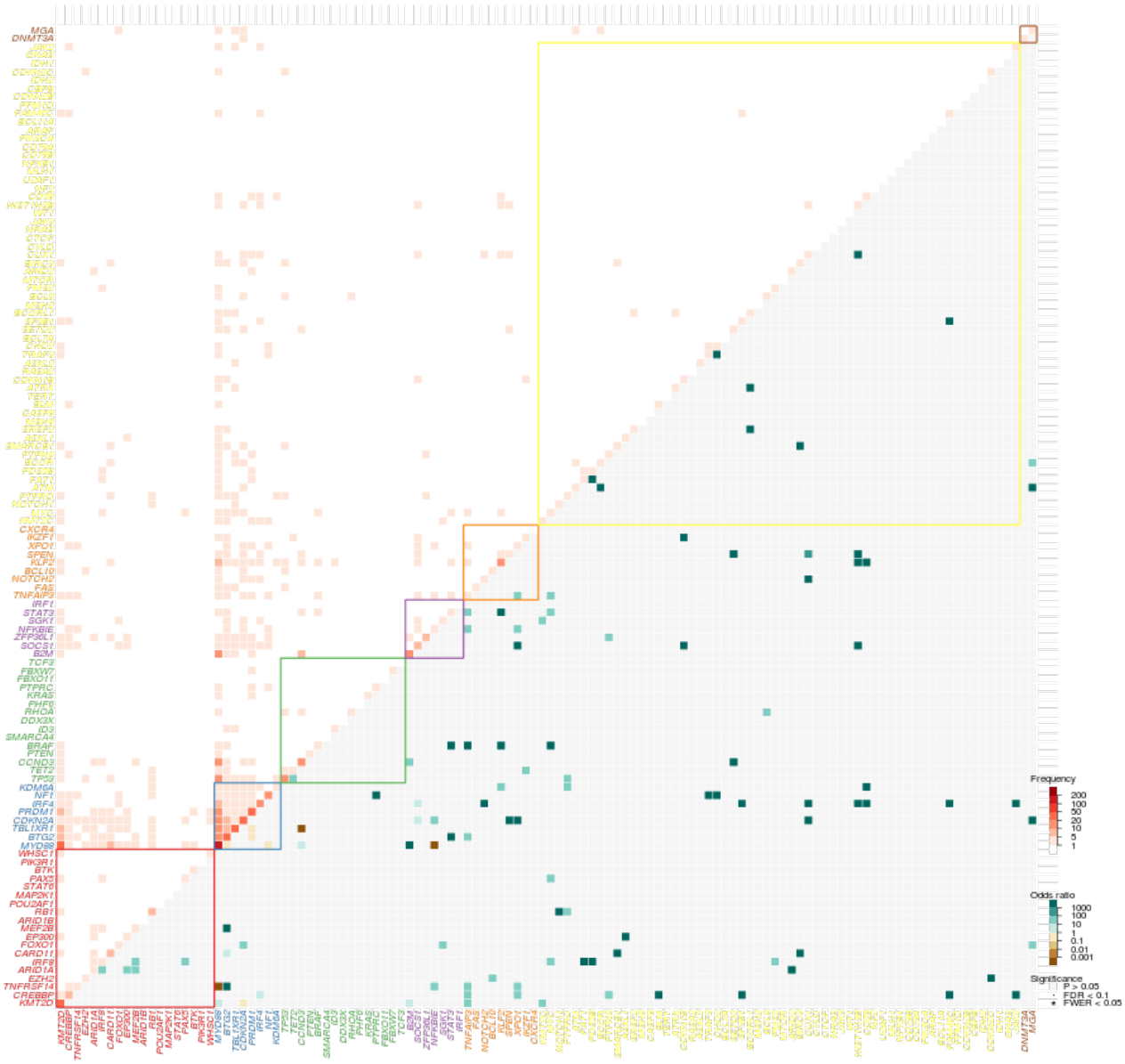
Class 0: TET2;TP53



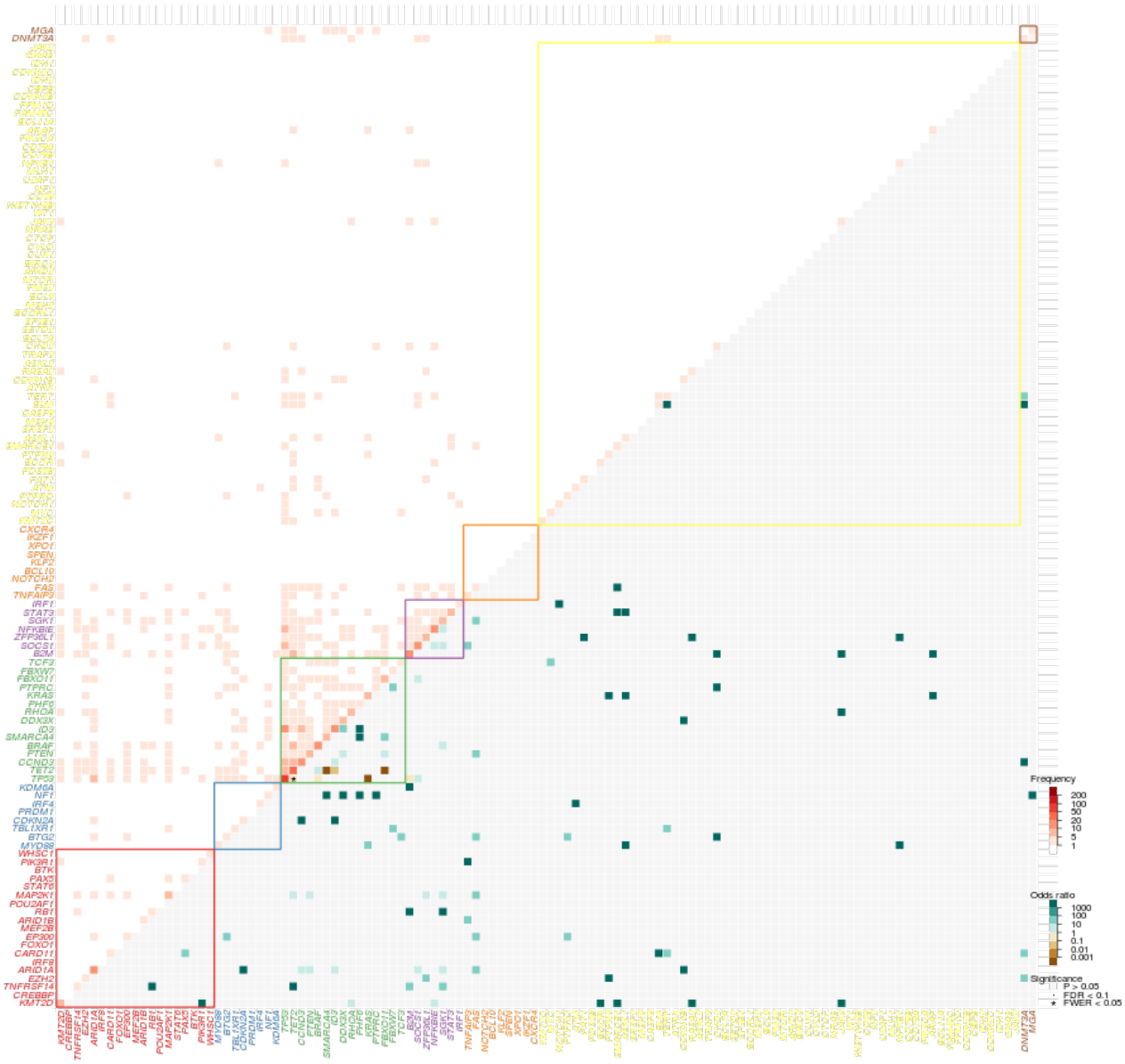
Class 1: KMT2D;CREBBP;TNFRSF14;EZH2;ARID1A



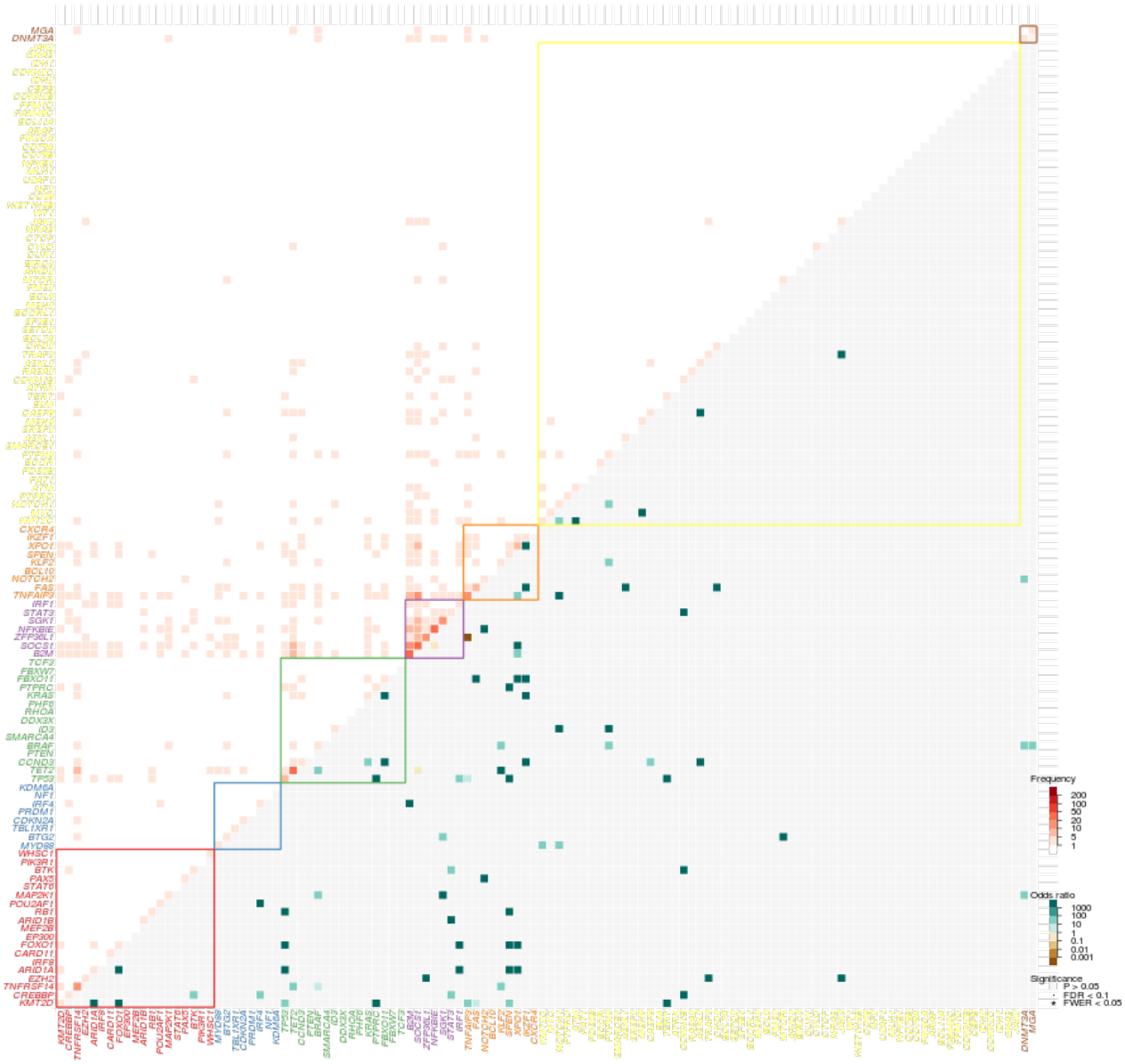
Class 2: KMT2D;MYD88;CREBBP;TNFRSF14;EZH2



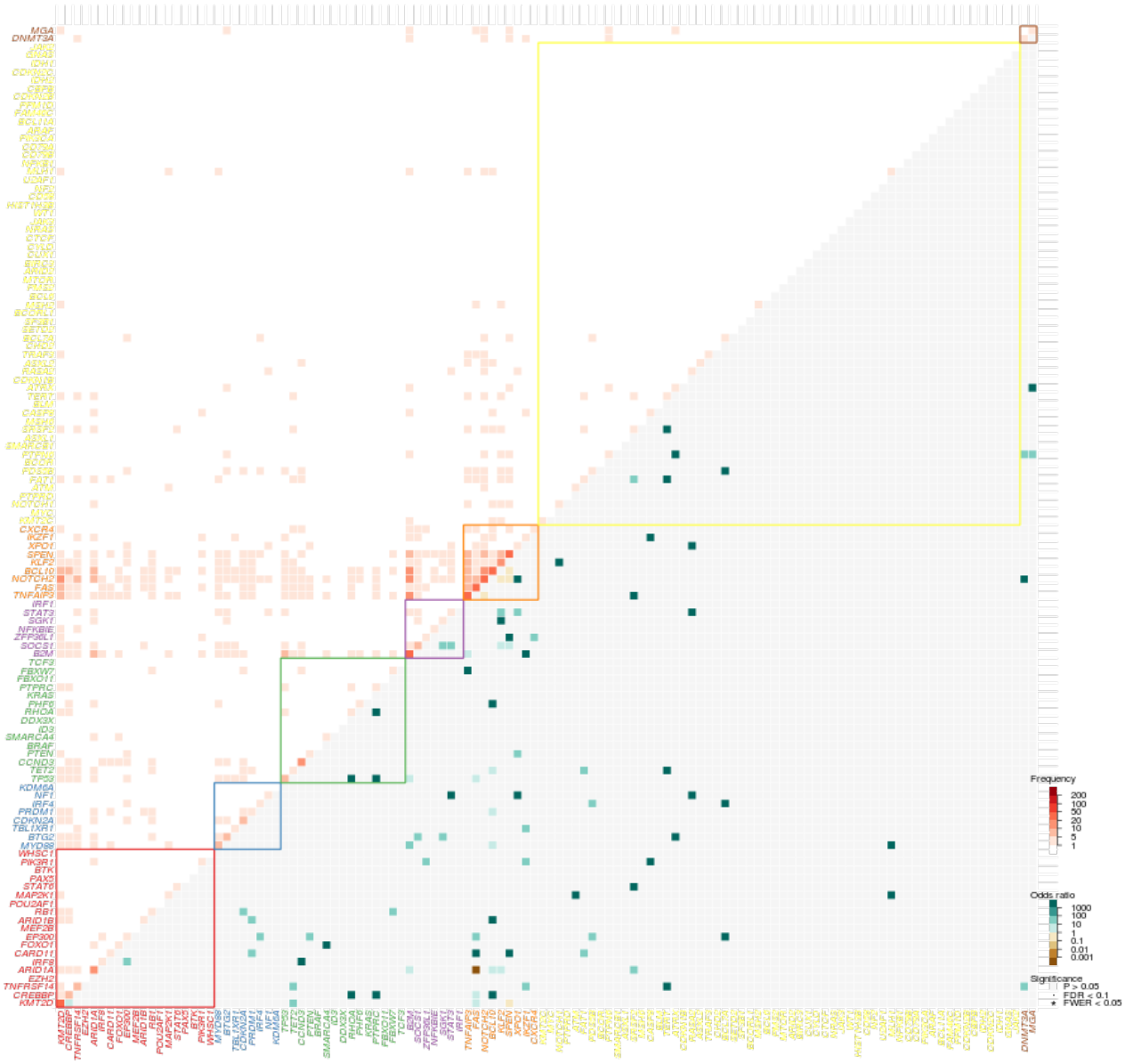
Class 3: TP53;SOCS1;TET2;B2M;CCND3



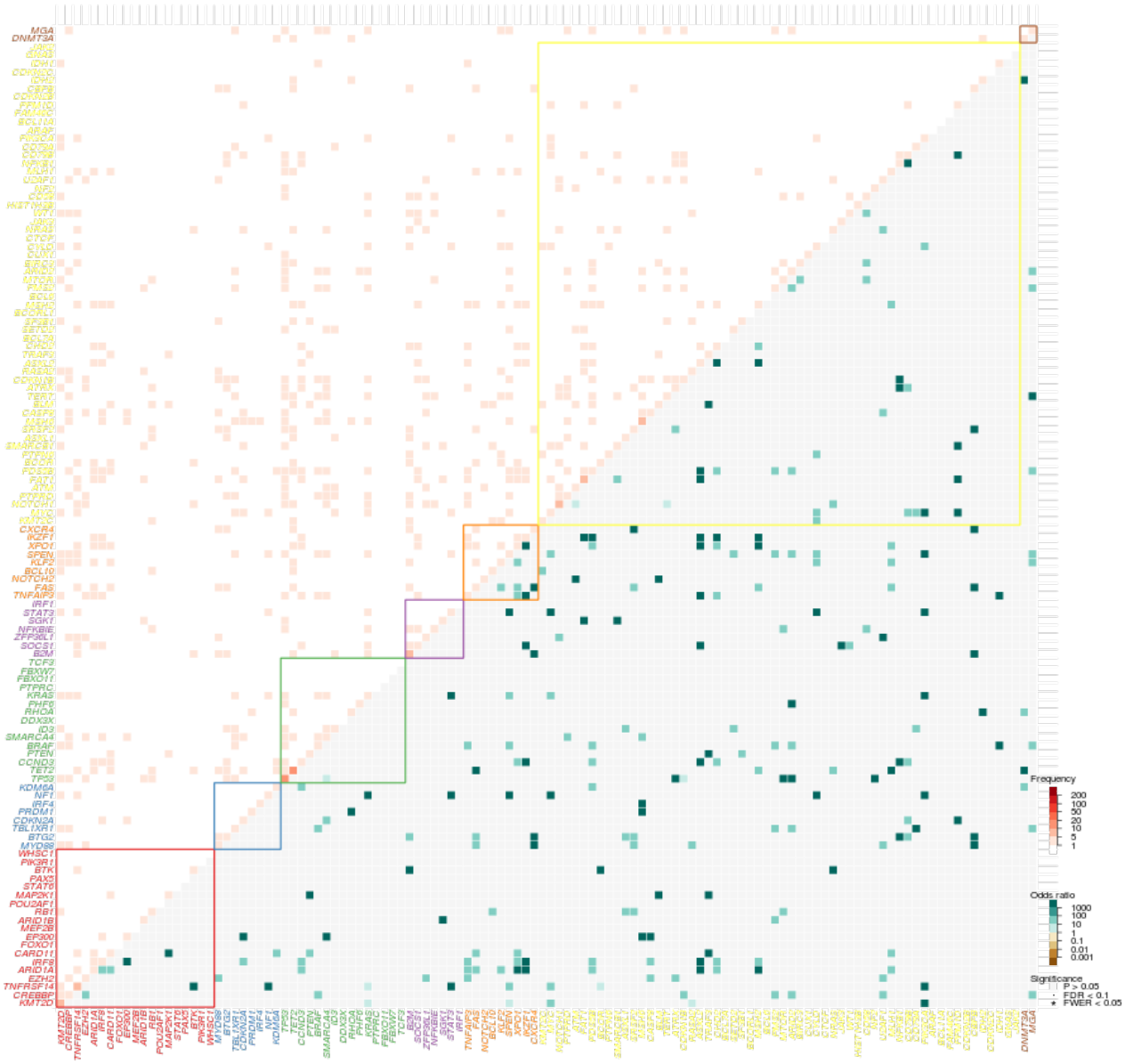
Class 4: SOCS1;B2M;TP53;TET2;TNFAIP3

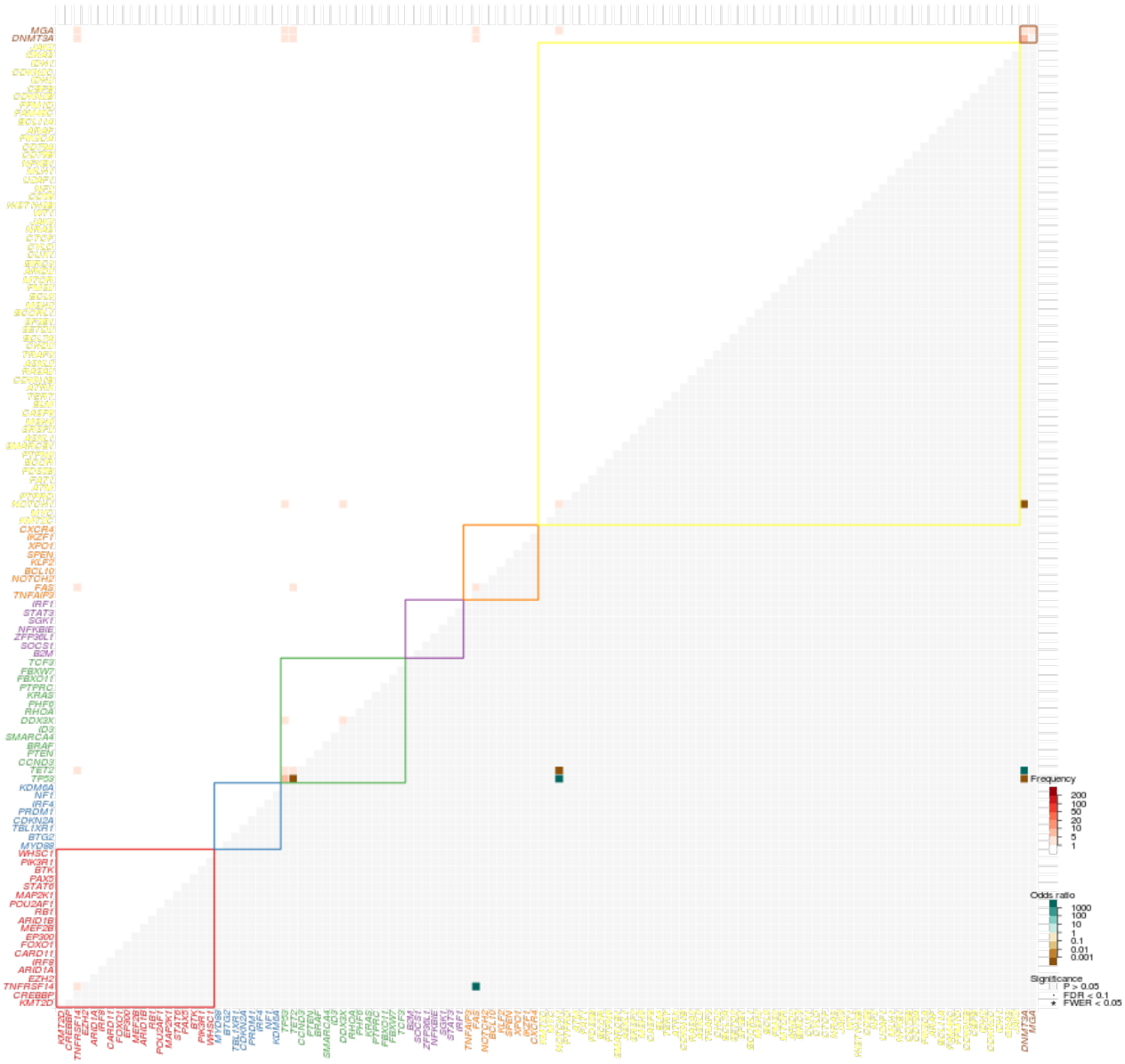


Class 5: TNFAIP3;NOTCH2;B2M;BCL10;MYD88



Class 6: TP53;MYC;FAT1





Alternatively: Naive Bayes assignment

```
sigBayes <- function(genotype, sigs){
  dSig <- function(sig,genotype){
    dmultinom(genotype, prob=sig/sum(sig))
  }
  lik <- apply(sigs,2, dSig, genotype)
  lik/sum(lik)
}

naiveBayes <- t(apply(genotypesImputed,1,sigBayes,posteriorMeans))
```

Save

```
# names(dpClass) <- clinicalData$PDID
# save(dpClass, posteriorMeans, posteriorQuantiles, posteriorProbability, file='dpClass.RData
```

#' ## Curated classification c <- read.table("../data/reduced_classes.txt", header=TRUE, sep="\t") curatedClass <- c$ReductionClass names(curatedClass) <- c$Sample rm©

library(clue) t <- table(dpClass, curatedClass) s <- solve_LSAP(t, maximum=TRUE) t[,c(s, setdiff(1:ncol(t),s))]

#' ### Associations

X <- as.matrix(MakeInteger(curatedClass)) Z <- as.matrix(genotypesImputed) Y <- as.matrix(dataFrame[groups %in% c("Clinical","Demographics")])

cv.glm <- function(x,y, fold=5, family="gaussian"){ cvIdx <- sample(1:nrow(x)%% fold +1) p <- numeric(length(y)) for(i in 1:fold){ p[cvIdx==i] <- predict(glm(y ~ ., data=x, subset=cvIdx!=i, family=family), newdata=x[cvIdx==i,]) } if(family=="gaussian"){ m <- mean((p-y)$^{2)}$ s <- sd(sapply(1:fold, function(i) mean((p-y)[cvIdx==i]^{2})))/sqrt(fold) return(c(avg=m, sd=s)) } else if(family=="binomial"){ m <- performance(prediction(p, y), "auc")@y.values[[1]] s <- sd(sapply(1:fold, function(i) performance(prediction(p[cvIdx==i], y[cvIdx==i]), "auc")@y.values[[1]]))/sqrt(fold) return(c(avg=m, sd=s)) } }

#' #### White counts #+ wbc_box y <- log(clinicalData$wbc) boxplot(y ~ factor(curatedClass), ylab="log wbc", las=2) #+ wbc_bar, fig.width=1.5 set.seed(42) mse0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), na.omit(y)) g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), type="mse", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) mse1 <- c(min(g$cvm), g$cvsd[which.min(g$cvm)]) v <- var(y, use='c') a <- (v-c(subtypes=mse0[1], `subtypes+genomics`=min(mse1[1])))/v*100 barplot(rbind(a,100-a), ylab="Explained variance (%)", names=rep("",2), ylim=c(0,100)) -> b segments(b,(v-c(mse0[1]-mse0[2],mse1[1]-mse1[2]))*100/v,b,(v-c(mse0[1]+mse0[2], mse1[1]+mse1[2]))*100/v) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="White counts")

#' #### Bone marrow blasts #+ BM_box y <- car::logit(clinicalData$BM_Blasts/100) boxplot(y ~ factor(curatedClass), ylab="logit BM blasts", las=2) #+ BM_bar, fig.width=1.5 set.seed(42) mse0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), na.omit(y)) g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), type="mse", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) mse1 <- c(min(g$cvm), g$cvsd[which.min(g$cvm)]) v <- var(y, use='c') a <- (v-c(subtypes=mse0[1], `subtypes+genomics`=min(mse1[1])))/v*100 barplot(rbind(a,100-a), ylab="Explained variance (%)", names=rep("",2), ylim=c(0,100)) -> b segments(b,(v-c(mse0[1]-mse0[2],mse1[1]-mse1[2]))*100/v,b,(v-c(mse0[1]+mse0[2], mse1[1]+mse1[2]))*100/v) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="BM blasts")

#' #### PB blasts #+ PB_box y <- car::logit(clinicalData$PB_Blasts/100) boxplot(y ~ factor(curatedClass), ylab="logit PB blasts", las=2) #+ PB_bar, fig.width=1.5 set.seed(42) mse0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), na.omit(y)) g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), type="mse", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) mse1 <- c(min(g$cvm), g$cvsd[which.min(g$cvm)]) v <- var(y, use='c') a <- (v-c(subtypes=mse0[1], `subtypes+genomics`=min(mse1[1])))/v*100 barplot(rbind(a,100-a), ylab="Explained variance (%)", names=rep("",2), ylim=c(0,100)) -> b segments(b,(v-c(mse0[1]-mse0[2],mse1[1]-mse1[2]))*100/v,b,(v-c(mse0[1]+mse0[2], mse1[1]+mse1[2]))*100/v) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="PB blasts")

#' #### Age #+ Age_box y <- clinicalData$AOD boxplot(y ~ factor(curatedClass), ylab="Age", las=2) #+ Age_bar, fig.width=1.5 set.seed(42) mse0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), na.omit(y)) g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), type="mse", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) mse1 <- c(min(g$cvm), g$cvsd[which.min(g$cvm)]) v <- var(y, use='c') a <- (v-c(subtypes=mse0[1], `subtypes+genomics`=min(mse1[1])))/v*100 barplot(rbind(a,100-a), ylab="Explained variance (%)", names=rep("",2), ylim=c(0,100)) -> b segments(b,(v-c(mse0[1]-mse0[2],mse1[1]-mse1[2]))*100/v,b,(v-c(mse0[1]+mse0[2], mse1[1]+mse1[2]))*100/v) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="Age")

#' #### LDH #+ LDH_box y <- log(clinicalData$LDH) boxplot(y ~ factor(curatedClass), ylab="log LDH", las=2) #+

LDH_bar, fig.width=1.5 set.seed(42) mse0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), na.omit(y)) g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), type="mse", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) mse1 <- c(min(g$cvm), g$cvsd[which.min(g$cvm)]) v <- var(y, use='c') a <- (v-c(subtypes=mse0[1], `subtypes+genomics`=min(mse1[1])))/v*100 barplot(rbind(a,100-a), ylab="Explained variance (%)", names=rep("",2), ylim=c(0,100)) -> b segments(b,(v-c(mse0[1]-mse0[2],mse1[1]-mse1[2]))*100/v,b,(v-c(mse0[1]+mse0[2], mse1[1]+mse1[2]))*100/v) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="LDH")

#' #### Platelets #+ platelets_box y <- log(clinicalData$platelet) boxplot(y ~ factor(curatedClass), ylab="log platelets", las=2) #+ platelets_bar, fig.width=1.5 set.seed(42) mse0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), na.omit(y)) g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), type="mse", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) mse1 <- c(min(g$cvm), g$cvsd[which.min(g$cvm)]) v <- var(y, use='c') a <- (v-c(subtypes=mse0[1], `subtypes+genomics`=min(mse1[1])))/v*100 barplot(rbind(a,100-a), ylab="Explained variance (%)", names=rep("",2), ylim=c(0,100)) -> b segments(b,(v-c(mse0[1]-mse0[2],mse1[1]-mse1[2]))*100/v,b,(v-c(mse0[1]+mse0[2], mse1[1]+mse1[2]))*100/v) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="Platelets")

#' #### HB #+ HB_box y <- log(clinicalData$HB) boxplot(y ~ factor(curatedClass), ylab="log HB", las=2) #+ HB_bar, fig.width=1.5 set.seed(42) mse0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), na.omit(y)) g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), type="mse", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) mse1 <- c(min(g$cvm), g$cvsd[which.min(g$cvm)]) v <- var(y, use='c') a <- (v-c(subtypes=mse0[1], `subtypes+genomics`=min(mse1[1])))/v*100 barplot(rbind(a,100-a), ylab="Explained variance (%)", names=rep("",2), ylim=c(0,100)) -> b segments(b,(v-c(mse0[1]-mse0[2],mse1[1]-mse1[2]))*100/v,b,(v-c(mse0[1]+mse0[2], mse1[1]+mse1[2]))*100/v) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="HB")

oddsPlot <- function(t){ plot(x=rep(1:2, ncol(t)), t[], ylab="Number of cases", xlab= names(dimnames(t))[1], xaxt="n", log='y') mtext(side=1,at=c(1,2), text=paste0(rownames(t), ", n=", rowSums(t)), las=1, pch=16) segments(1,t[1,],2, t[2,]) p <- sapply(1:ncol(t), function(i) {f <- fisher.test(cbind(rowSums(t[,-i]), t[,i])); c(p.value=f$p.value, OR=f$estimate[1], f$conf.int)}) mtext(side=4, at=t[2,], text=paste0(colnames(t), ", OR=", format(p[2,],digits=1), "(", apply(round(p[3:4,],2),2,paste, collapse="-"),")", sig2star(p[1,]))) }

#' #### Splenomegaly #+ Splenomegaly_bar, fig.width=1.5 library(ROCR) set.seed(42) y <- clinicalData$Splenomegaly table(Splenomegaly=y,factor(curatedClass)) auc0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), y[!is.na(y)], family="binomial") g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), family='binomial',type="auc", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) auc1 <- c(max(g$cvm), g$cvsd[which.max(g$cvm)]) a <- c(subtypes=auc0[1], `subtypes+genomics`=auc1[1])*100 barplot(rbind(a,150-a)-50, ylab="AUC (%)", main="Splenomegaly", offset=50, names=rep("",2), ylim=c(50,100)) segments(b,c(auc0[1]-auc0[2],auc1[1]-auc1[2])*100,b,c(auc0[1]+auc0[2], auc1[1]+auc1[2])*100) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="Splenomegaly")

#' #### Gender #+ Gender_bar, fig.width=1.5 set.seed(42) y <- clinicalData$gender -1 table(Gender=factor(y, labels=c('male','female')),factor(curatedClass)) auc0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), y[!is.na(y)], family="binomial") g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), family='binomial',type="auc", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) auc1 <- c(max(g$cvm), g$cvsd[which.max(g$cvm)]) a <- c(subtypes=auc0[1], `subtypes+genomics`=auc1[1])*100 barplot(rbind(a,150-a)-50, ylab="AUC (%)", offset=50, names=rep("",2), ylim=c(50,100)) segments(b,c(auc0[1]-auc0[2],auc1[1]-auc1[2])*100,b,c(auc0[1]+auc0[2], auc1[1]+auc1[2])*100) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="Gender")

#' #### CR #+ CR_bar, fig.width=1.5 set.seed(42) y <- !is.na(clinicalData$CR_date) y[is.na(clinicalData$CR_date) & clinicalData$OS==0] <- NA table(CR=y,factor(curatedClass)) auc0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), y[!is.na(y)], family="binomial") g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), family='binomial',type="auc", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) auc1 <- c(max(g$cvm), g$cvsd[which.max(g$cvm)]) a <- c(subtypes=auc0[1], `subtypes+genomics`=auc1[1])*100 barplot(rbind(a,150-a)-50, ylab="AUC (%)", offset=50, names=rep("",2), ylim=c(50,100)) segments(b,c(auc0[1]-auc0[2],auc1[1]-auc1[2])*100,b,c(auc0[1]+auc0[2], auc1[1]+auc1[2])*100) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="Complete remission")

#' #### OS #+ OS_bar, fig.width=1.5 set.seed(42) y <- os[1:1540,2] y[os[1:1540,1] < 3 * 365 & os[1:1540,2]==0] <- NA table(OS=factor(y,labels=c("alive","dead")),factor(curatedClass)) auc0 <- cv.glm(as.data.frame(X[!is.na(y),-1]), y[!is.na(y)], family="binomial") g <- cv.glmnet(cbind(X,Z)[!is.na(y),], na.omit(y), family='binomial',type="auc", alpha=1, penalty.factor=c(rep(0,ncol(X)), rep(1,ncol(Z)))) auc1 <- c(max(g$cvm), g$cvsd[which.max(g$cvm)]) a <- c(subtypes=auc0[1], `subtypes+genomics`=auc1[1])*100 barplot(rbind(a,150-a)-50, ylab="AUC (%)", offset=50, names=rep("",2), ylim=c(50,100)) segments(b,c(auc0[1]-auc0[2],auc1[1]-auc1[2])*100,b,c(auc0[1]+auc0[2], auc1[1]+auc1[2])*100) rotatedLabel(b, labels=c("subtypes","subtypes+genomics")) title(main="Overall survival at 3yr")

Session

```
devtools::session_info()
```

```
## Session info -------------------------------------------------------------
```

```
##  setting  value
##  version  R version 3.3.3 (2017-03-06)
##  system   x86_64, linux-gnu
##  ui       X11
##  language (EN)
##  collate  en_GB.UTF-8
##  tz       Europe/London
##  date     2017-08-29
```

```
## Packages -----------------------------------------------------------------
```

```
##  package       * version  date       source
##  AnnotationDbi   1.36.2   2017-05-15 Bioconductor
##  ape           * 4.1      2017-02-14 CRAN (R 3.3.3)
##  assertthat      0.2.0    2017-04-11 CRAN (R 3.3.3)
##  base          * 3.3.3    2017-03-15 local
##  Biobase         2.34.0   2017-05-15 Bioconductor
##  BiocGenerics    0.20.0   2017-04-24 Bioconductor
##  biomaRt       * 2.30.0   2017-05-15 Bioconductor
##  bitops          1.0-6    2013-08-17 CRAN (R 3.3.3)
##  cellranger      1.1.0    2016-07-27 CRAN (R 3.3.3)
##  codetools       0.2-15   2016-10-05 CRAN (R 3.3.3)
##  colorspace      1.3-2    2016-12-14 CRAN (R 3.3.3)
##  CoxHD         * 0.0.61   2017-07-04 Github (mg14/CoxHD@d295566)
##  data.table    * 1.10.4   2017-02-01 CRAN (R 3.3.3)
##  datasets      * 3.3.3    2017-03-15 local
##  DBI             0.6-1    2017-04-01 CRAN (R 3.3.3)
##  devtools        1.13.2   2017-06-02 CRAN (R 3.3.3)
##  digest          0.6.12   2017-01-27 CRAN (R 3.3.3)
##  dplyr         * 0.5.0    2016-06-24 CRAN (R 3.3.3)
##  dtplyr        * 0.0.2    2017-04-21 CRAN (R 3.3.3)
##  evaluate        0.10.1   2017-06-24 cran (@0.10.1)
##  foreach         1.4.3    2015-10-13 cran (@1.4.3)
##  ggplot2       * 2.2.1    2016-12-30 CRAN (R 3.3.3)
##  ggrepel       * 0.6.5    2016-11-24 CRAN (R 3.3.3)
##  ggthemes      * 3.4.0    2017-02-19 CRAN (R 3.3.3)
```

```
##  glmnet          2.0-10   2017-05-06 cran (@2.0-10)
##  graphics      * 3.3.3    2017-03-15 local
##  grDevices     * 3.3.3    2017-03-15 local
##  grid          * 3.3.3    2017-03-15 local
##  gridExtra     * 2.2.1    2016-02-29 CRAN (R 3.3.3)
##  gsubfn        * 0.6-7    2017-04-13 Github (ggrothendieck/gsubfn@d2ef6c4)
##  gtable          0.2.0    2016-02-26 CRAN (R 3.3.3)
##  hdp           * 0.0.1    2017-07-19 Github (nicolaroberts/hdp@506f381)
##  highr           0.6      2016-05-09 cran (@0.6)
##  hms             0.3      2016-11-22 CRAN (R 3.3.3)
##  IRanges         2.8.2    2017-04-24 Bioconductor
##  iterators       1.0.8    2015-10-13 cran (@1.0.8)
##  knitr         * 1.16     2017-05-18 cran (@1.16)
##  labeling        0.3      2014-08-23 CRAN (R 3.3.3)
##  lattice       * 0.20-35  2017-03-25 CRAN (R 3.3.3)
##  lazyeval        0.2.0    2016-06-12 CRAN (R 3.3.3)
##  lsa             0.73.1   2015-05-08 cran (@0.73.1)
##  magrittr        1.5      2014-11-22 CRAN (R 3.3.3)
##  MASS            7.3-47   2017-04-21 CRAN (R 3.3.3)
##  Matrix          1.2-10   2017-04-28 CRAN (R 3.3.3)
##  memoise         1.1.0    2017-04-21 CRAN (R 3.3.3)
##  methods       * 3.3.3    2017-03-15 local
##  mg14          * 0.0.5    2017-07-04 Github (mg14/mg14@a8b4ba8)
##  mice            2.30     2017-02-18 cran (@2.30)
##  munsell         0.4.3    2016-02-13 CRAN (R 3.3.3)
##  mvtnorm         1.0-6    2017-03-02 cran (@1.0-6)
##  nlme            3.1-131  2017-02-06 CRAN (R 3.3.3)
##  nnet            7.3-12   2016-02-02 CRAN (R 3.3.3)
##  parallel      * 3.3.3    2017-03-15 local
##  plyr            1.8.4    2016-06-08 CRAN (R 3.3.3)
##  proto         * 1.0.0    2016-10-29 cran (@1.0.0)
##  R6              2.2.2    2017-06-17 cran (@2.2.2)
##  RColorBrewer  * 1.1-2    2014-12-07 CRAN (R 3.3.3)
##  Rcpp            0.12.11  2017-05-22 cran (@0.12.11)
##  RCurl           1.95-4.8 2016-03-01 CRAN (R 3.3.3)
##  readr         * 1.1.0    2017-03-22 CRAN (R 3.3.3)
##  readxl        * 1.0.0    2017-04-18 CRAN (R 3.3.3)
##  rpart           4.1-11   2017-04-21 CRAN (R 3.3.3)
##  RSQLite         1.1-2    2017-01-08 CRAN (R 3.3.3)
##  S4Vectors       0.12.2   2017-04-24 Bioconductor
##  scales        * 0.4.1    2016-11-09 CRAN (R 3.3.3)
##  SnowballC       0.5.1    2014-08-09 cran (@0.5.1)
##  splines         3.3.3    2017-03-15 local
##  stats         * 3.3.3    2017-03-15 local
##  stats4          3.3.3    2017-03-15 local
##  stringi         1.1.5    2017-04-07 CRAN (R 3.3.3)
##  stringr       * 1.2.0    2017-02-18 CRAN (R 3.3.3)
##  survival      * 2.41-3   2017-04-04 CRAN (R 3.3.3)
##  tcltk           3.3.3    2017-03-15 local
##  tibble          1.3.0    2017-04-01 CRAN (R 3.3.3)
##  tidyr         * 0.6.1    2017-01-10 CRAN (R 3.3.3)
```

```
##  tools           3.3.3    2017-03-15 local
##  utils         * 3.3.3    2017-03-15 local
##  withr           1.0.2    2016-06-20 CRAN (R 3.3.3)
##  XML             3.98-1.7 2017-05-03 CRAN (R 3.3.3)
```