Supplement

1 Further Supplementary Figures

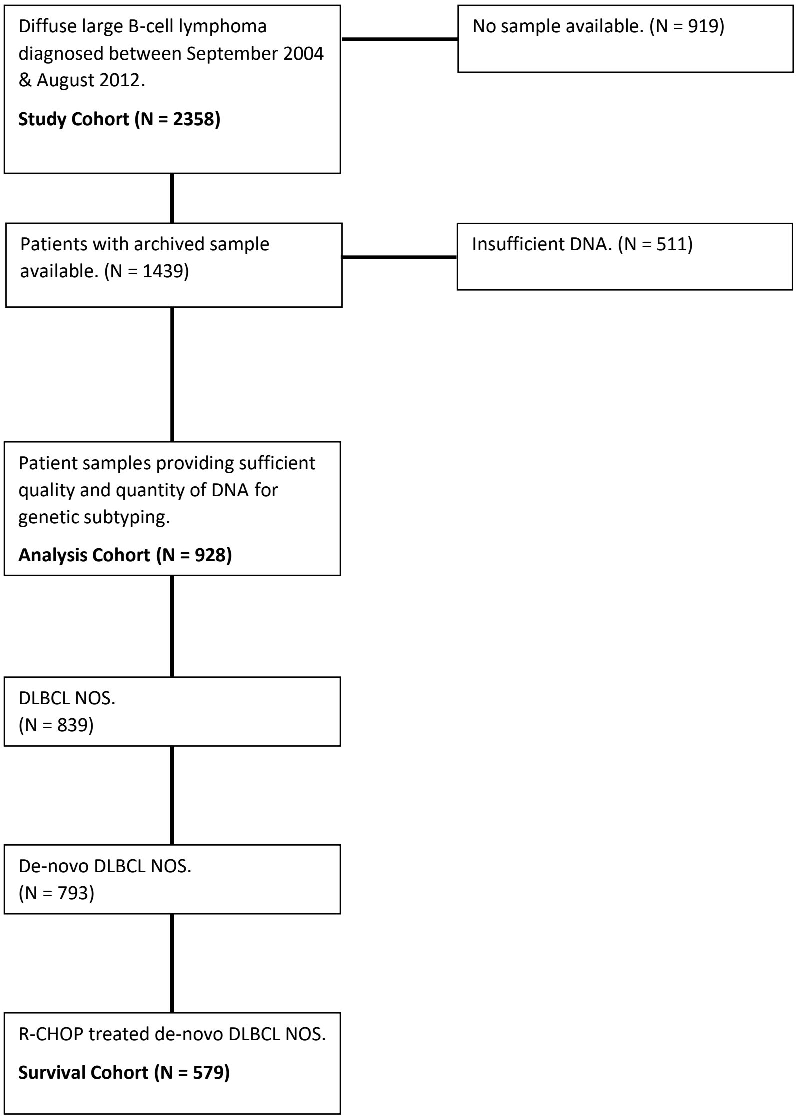
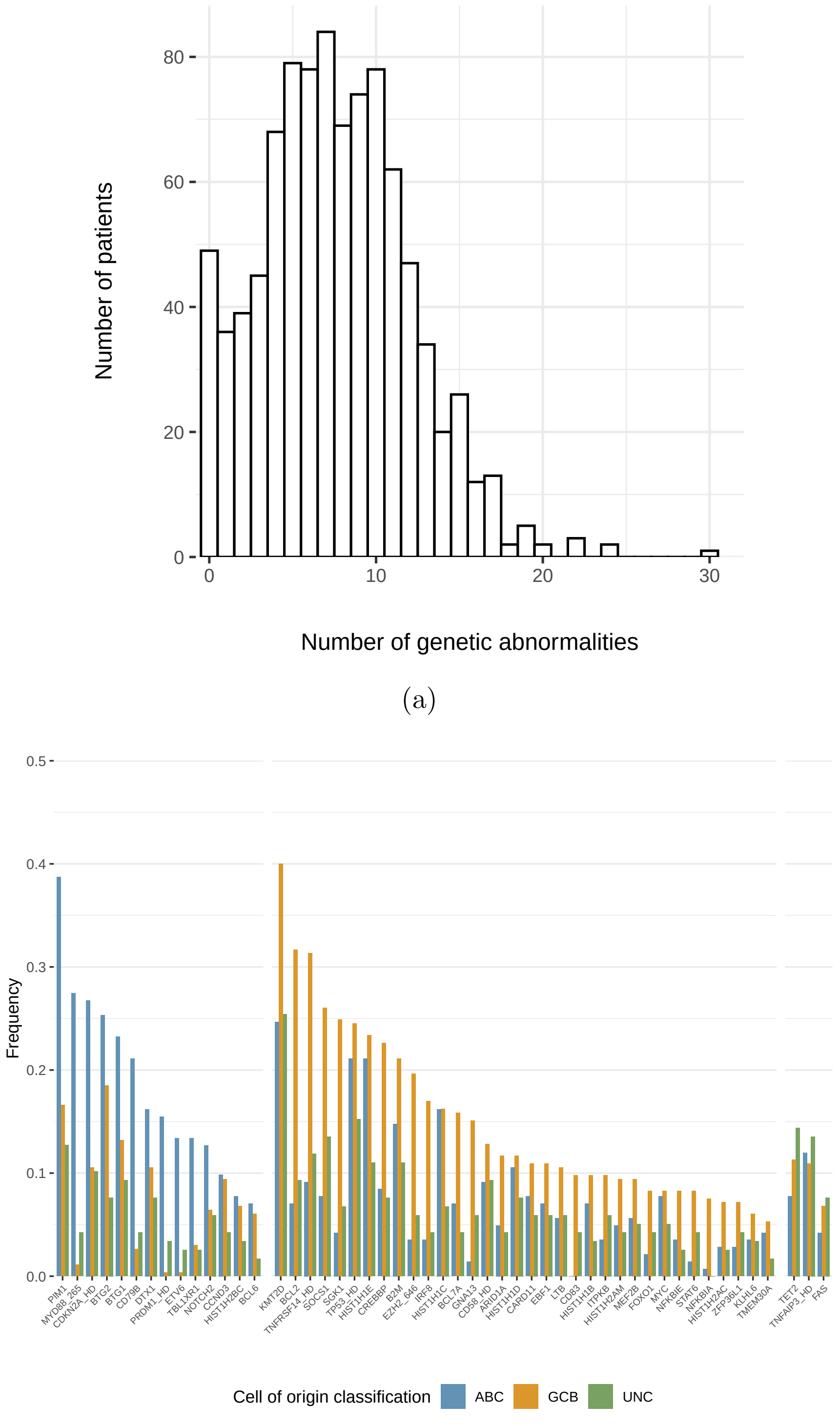

Diffuse large B-cell lymphoma diagnosed between September 2004 & August 2012.
Study Cohort (N = 2358)
No sample available. (N = 919)
Patients with archived sample available. (N = 1439)
Insufficient DNA. (N = 511)
Patient samples providing sufficient quality and quantity of DNA for genetic subtyping.
Analysis Cohort (N = 928)
DLBCL NOS. (N = 839)
De-novo DLBCL NOS. (N = 793)
R-CHOP treated de-novo DLBCL NOS.
Survival Cohort (N = 579)

Figure S1: **CONSORT diagram describing the patient cohort.** n=928 samples were available for clustering to identify subtypes of diffuse large B-cell lymphoma (DLBCL). Of these, n=579, were patients diagnosed with de-novo DLBCL-NOS, and treated with curative intent with R-CHOP therapy. This latter group provides the most heterogeneous possible subset of the patients for survival comparisons.



(a)

(b)

Figure S2: a) The distribution of the number of genetic abnormalities per patient in the Analysis Cohort (n=928). b) The distribution of mutations occurring in at least 5% of patients, stratified by cell-of-origin classification (shown by the coloring of the bars). Mutations are grouped by those that are most frequent in the ABC and GCB cell-of-origin groups respectively. The patients represented here are the n=455 with DLBCL-NOS, Primary CNS lymphoma or T-cell/histiocyte-rich large B-cell lymphoma for whom gene expression profiling data were available (n=455).

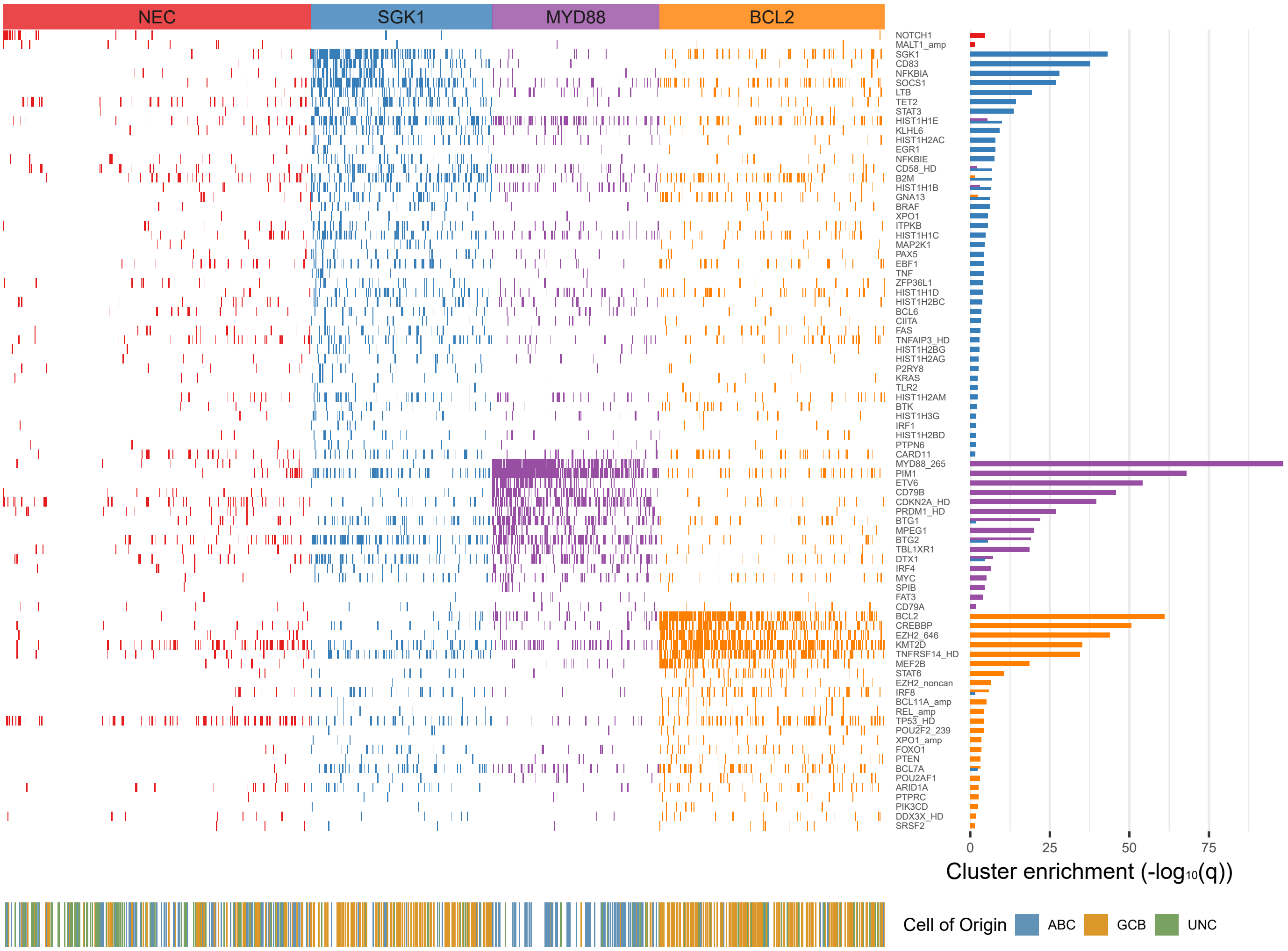


Figure S3: **Heatmap displaying the four genetic clusters that were identified using the ICL criterion in the analysis cohort (n=928).** Only those mutations are shown that are identified as significantly enriched for the given group, as determined by a Benjamini-Hochberg adjusted $q < 0.05$ from a chi-squared test of independence. "HD" stands for homozygous deletion or a mutation in this gene, "noncan" denotes a non-canonical mutation, and "amp" indicates an amplification.

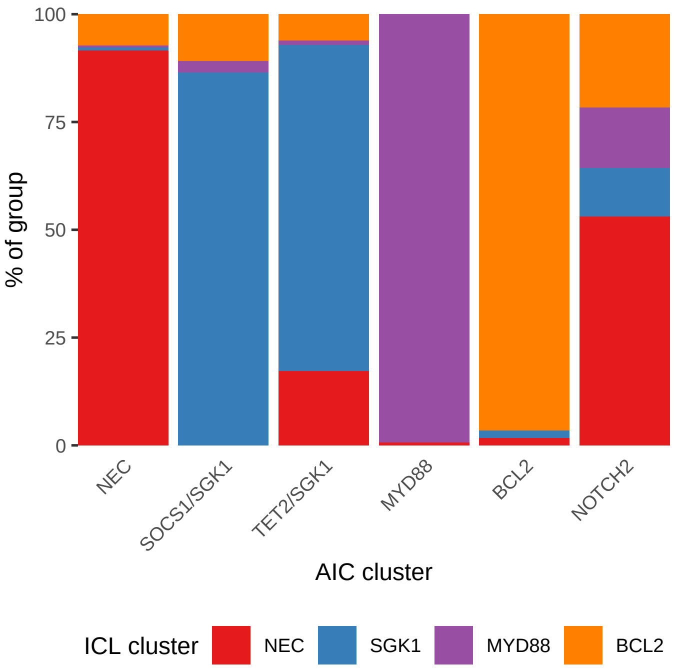


Figure S4: **Relationship between ICL and AIC criteria cluster membership**, highlighting that MYD88 and BCL2 remain largely unchanged under the more relaxed AIC penalisation, while SOCS1/SGK1 and TET2/SGK1 emerge from the SGK1 group. The bars represent the proportion of AIC cluster that were in the ICL cluster denoted by the bar colour

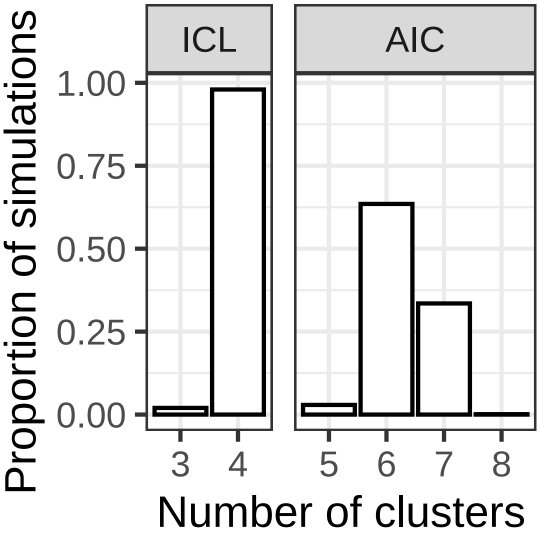


Figure S5: Number of clusters identified by ICL and AIC criteria over bootstrapping.

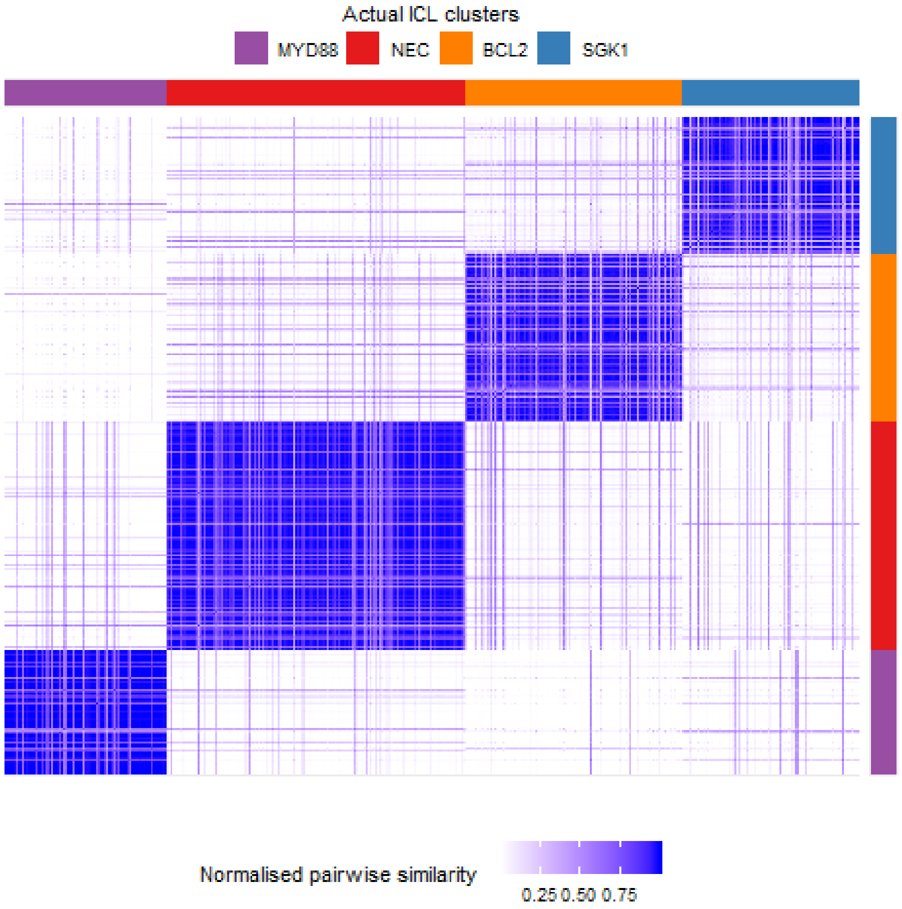


Figure S6: Consensus matrix of ICL selected clusters ordered by stability scores of genetic subtypes decreasing from left to right.

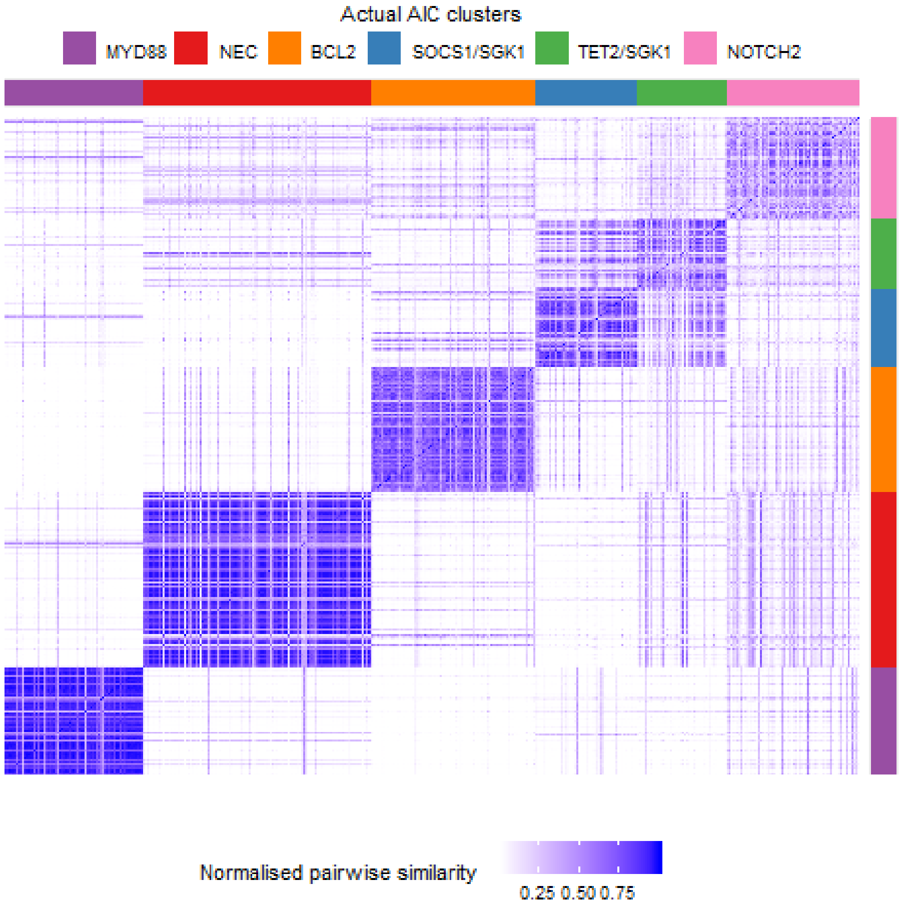


Figure S7: Consensus matrix of AIC selected clusters ordered by stability scores of genetic subtypes decreasing from left to right.

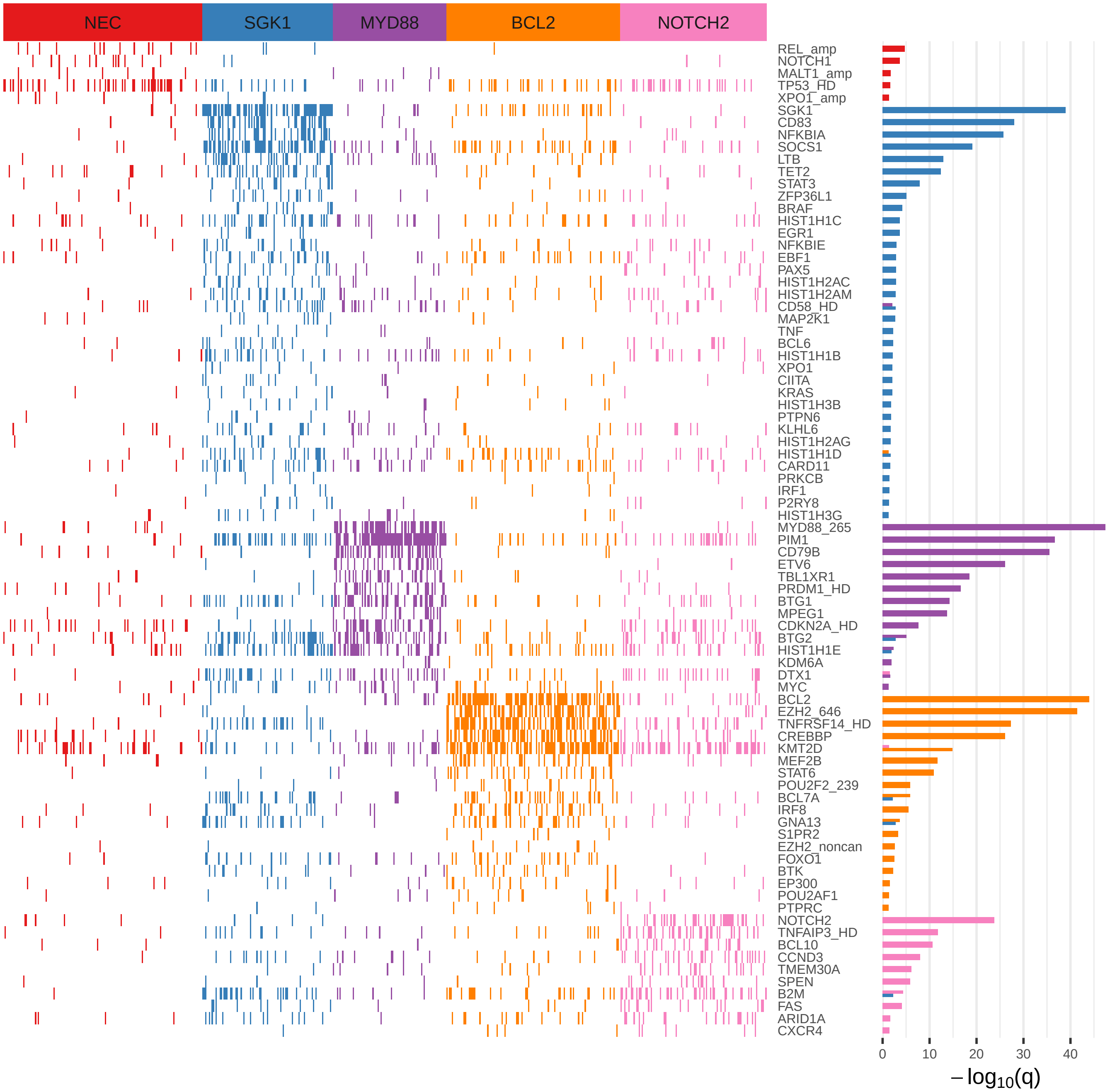
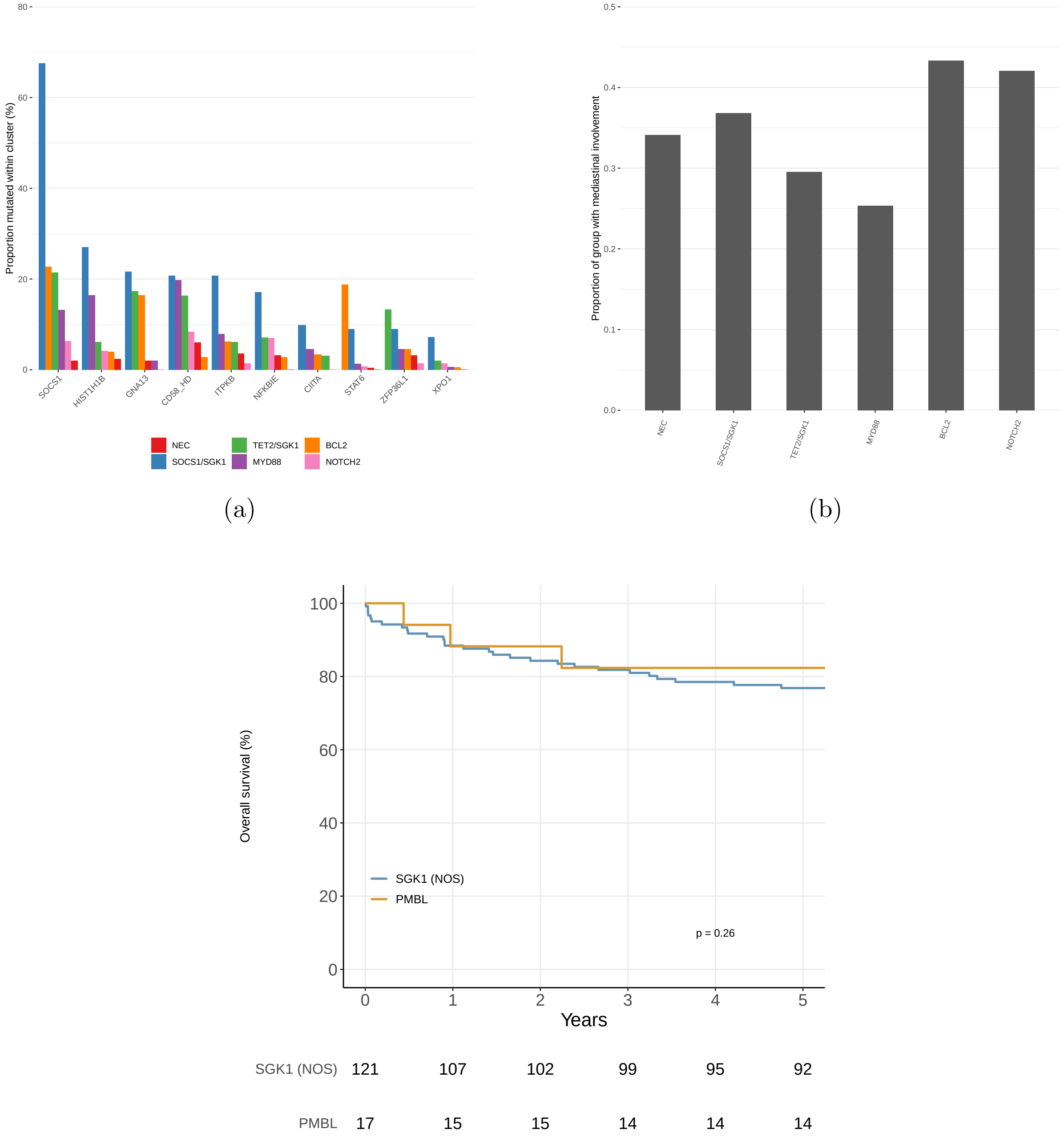


Figure S8: **Heatmap of the enriched mutations from the five clusters identified using AIC mixture modelling of the 579 de-novo DLBCL NOS who were treated with R-CHOP.**



(a)

(b)

(c)

Figure S9: **Comparison of factors related to mediastinal involvement.** a) Mutation frequency of 10 genes associated with PMBCL across the six clusters. b) Proportion of patients with mediastinal involvement, stratified by the genetic subtype identified using the AIC criterion (n=811). c) Comparison of the overall survival of R-CHOP treated patients in the SGK1 genetic subtype identified using the ICL criterion, with R-CHOP treated patients diagnosed with primary mediastinal B-cell lymphoma (PMBL).

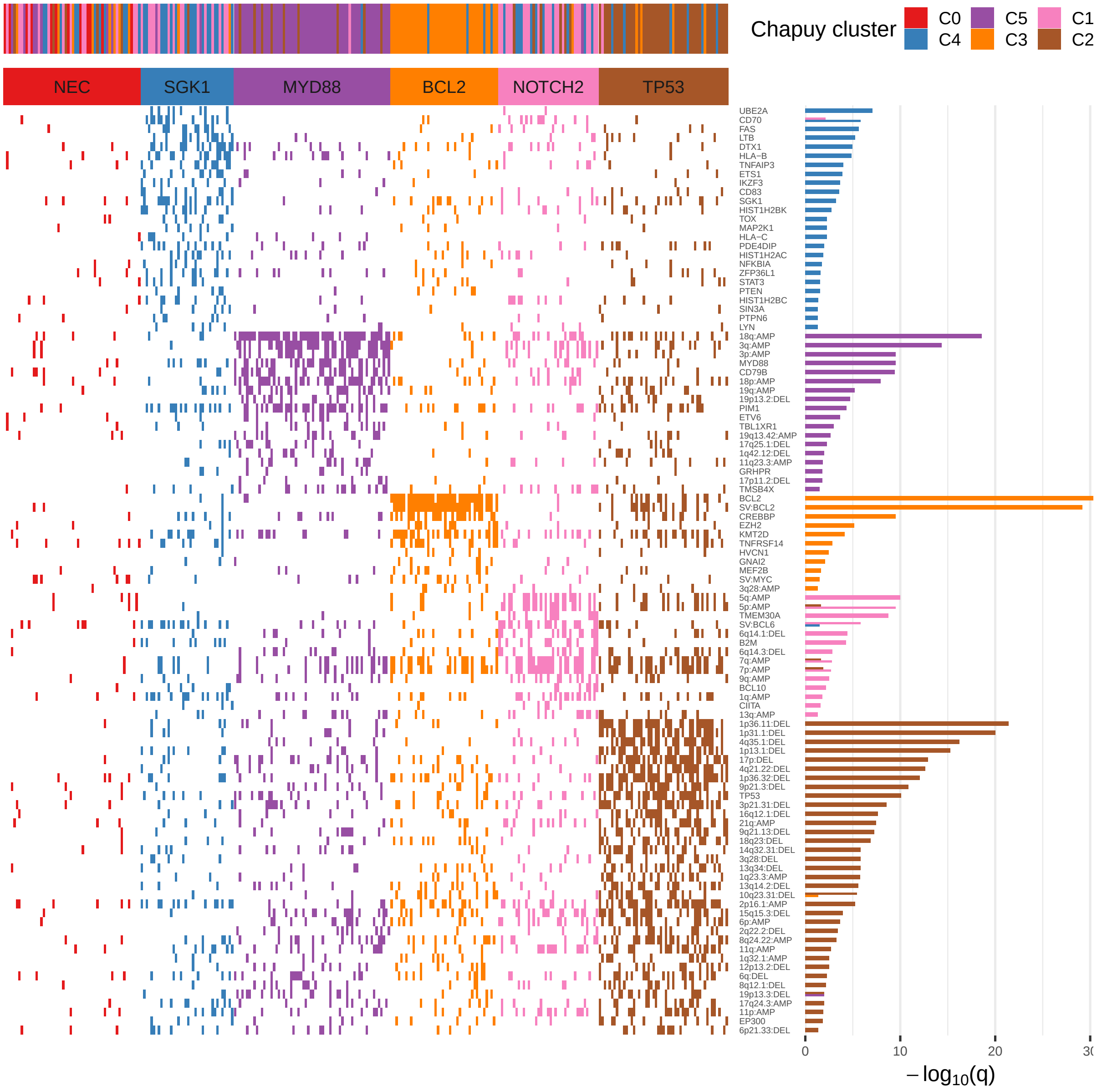
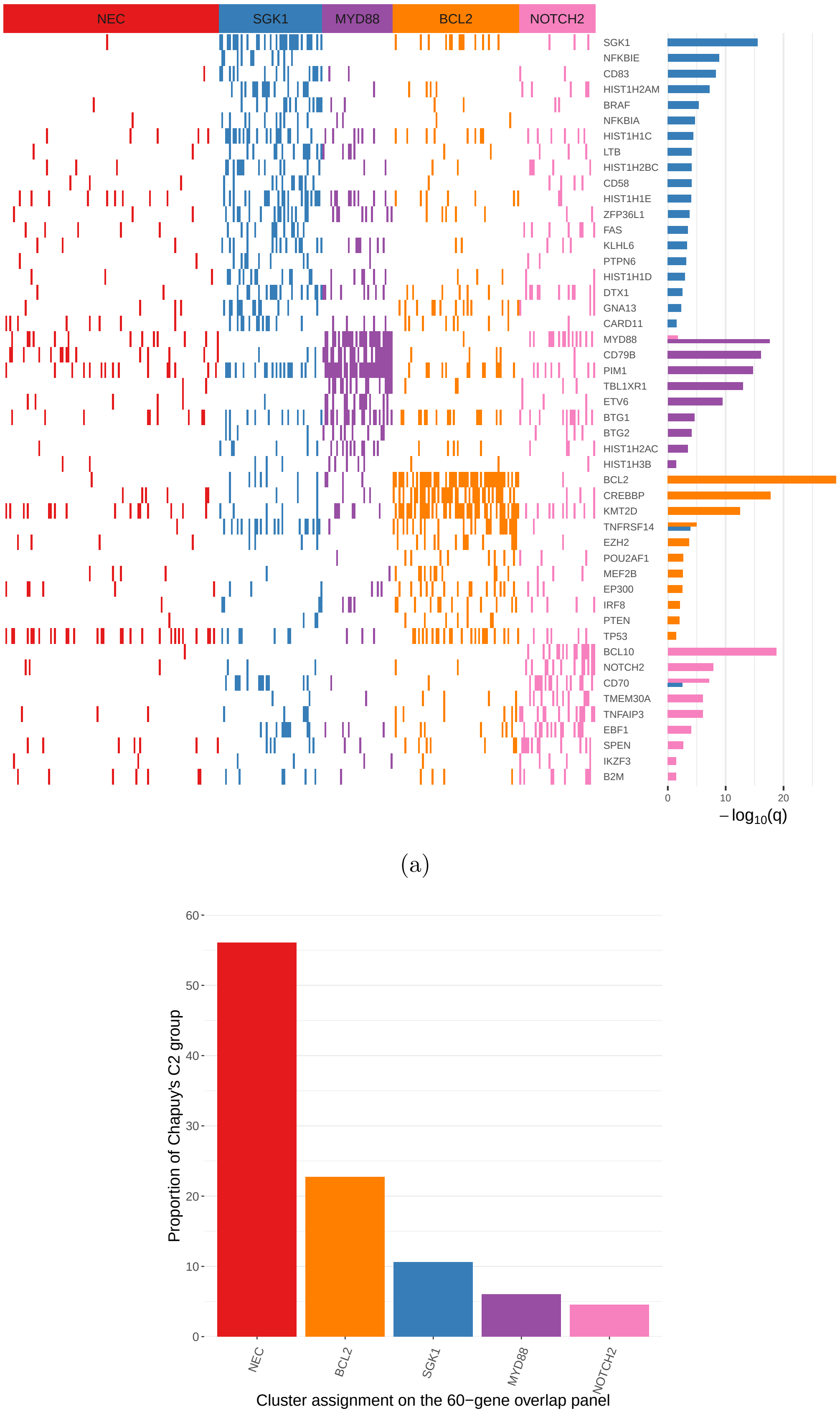


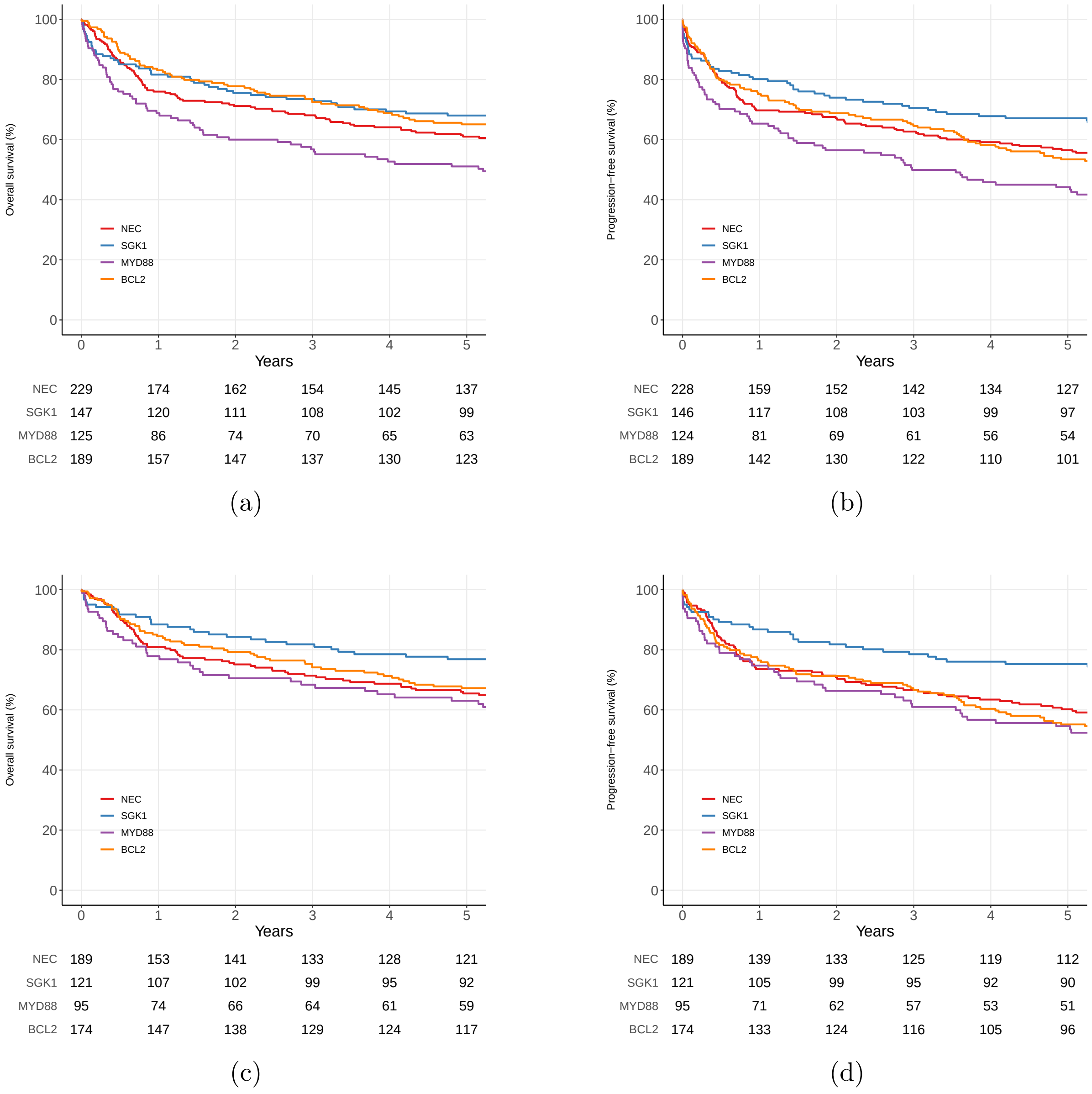
Figure S10: **Resultant heatmap of enriched genetic features from AIC mixture model clustering of Chapuy et al's dataset using their 158 panel.** The top bar denotes the original assigned subtype as published in Chapuy et al.



(a)

(b)

Figure S11: **Results from AIC mixture model clustering of Chapuy et al's dataset restricted to the 60 mutations that are shared in common with our panel.** a) Heatmap of enriched mutations. b) Reassignments of the members of Chapuy et al's C2 group.



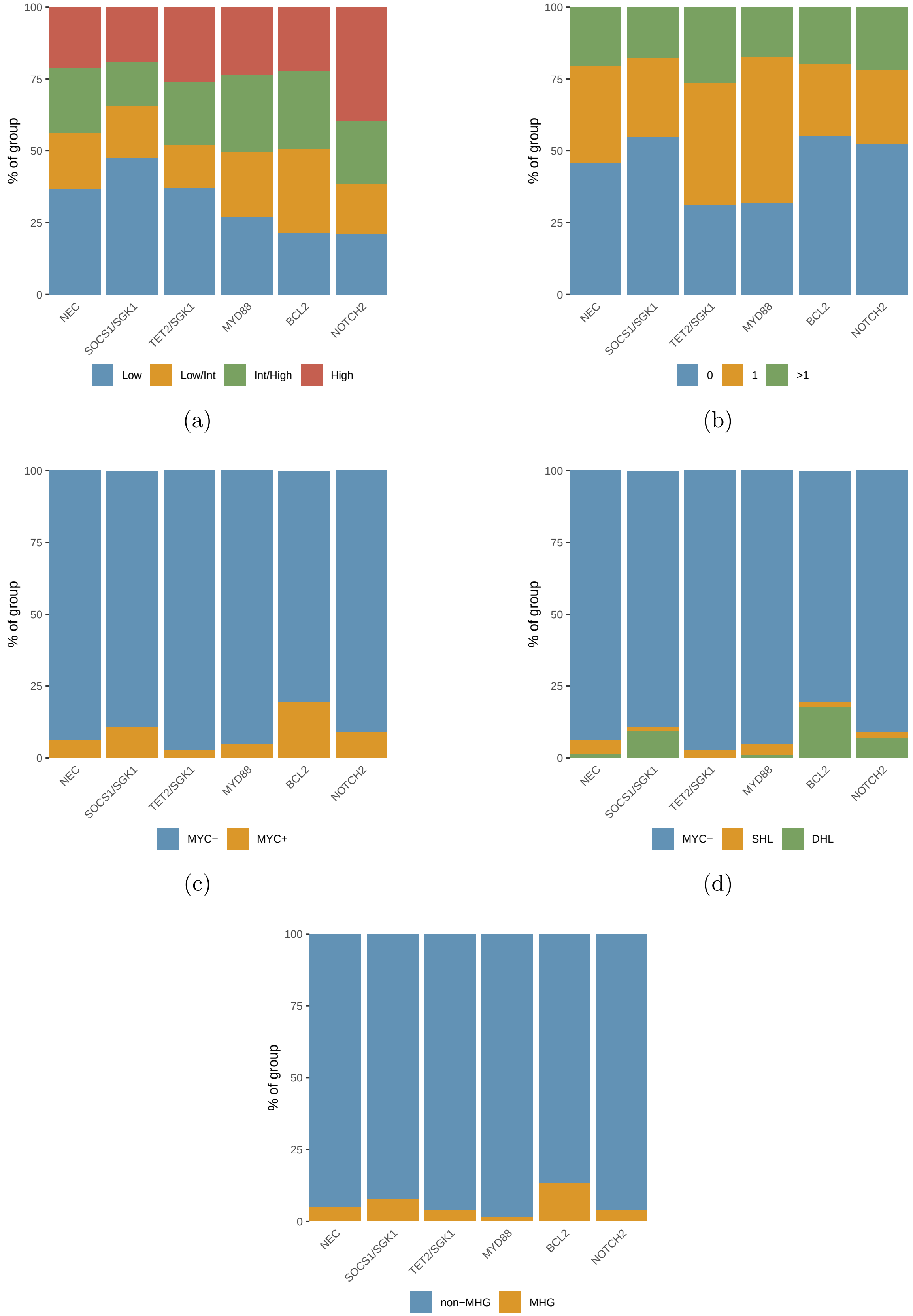
(a)

(b)

(c)

(d)

Figure S12: **Survival in clusters identified by the ICL criterion.** (a): Overall survival of the n=690 patients treated with curative intent stratified by cluster, (b): overall survival of the n=579 de-novo DLBCL NOS patients treated using R-CHOP, stratified by cluster. (c): Progression-free survival of the n=690 patients treated with curative intent stratified by cluster, (d): Progression-free survival of the n=579 de-novo DLBCL NOS patients treated using R-CHOP, stratified by cluster.



(a)

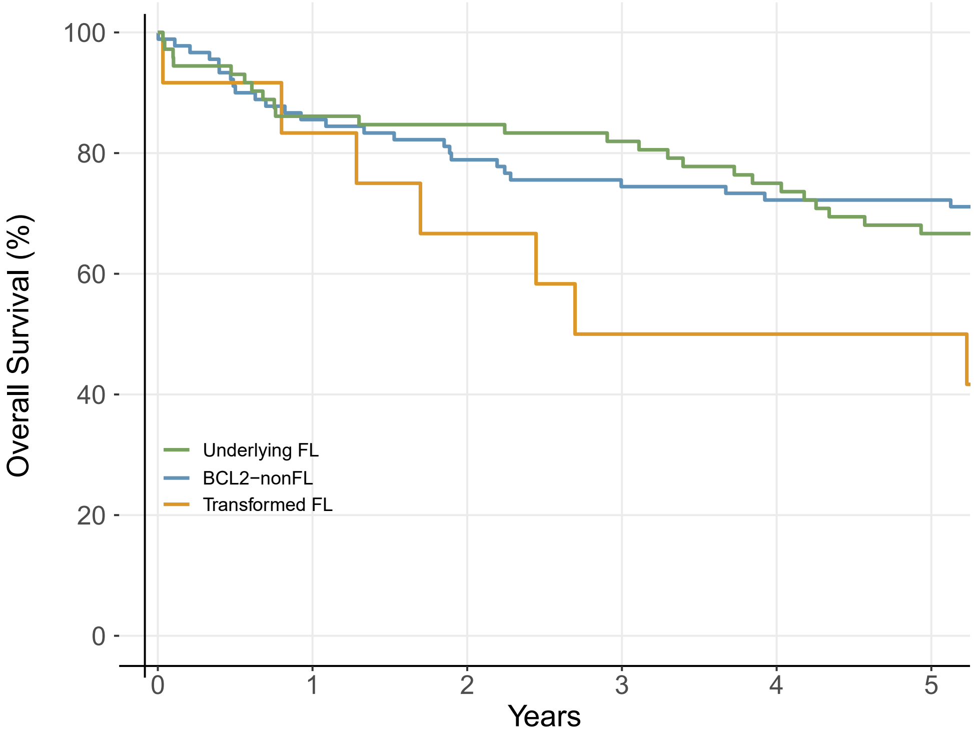
(b)

(c)

(d)

(e)

Figure S13: **Characteristics of the genetic subtypes identified by the AIC criterion.** a) The distribution of International Prognostic Index classifications (n=611). b) The distribution of the number of extra-nodal sites (n=712). c) MYC rearrangement (MYC+) (n=602). d) Double-hit lymphoma (DHL = MYC rearrangement plus either BCL2 or BCL6, SHL = MYC+ only, MYC- = MYC- and any, or no other, rearrangement)(n=602). e) Molecular high grade (MHG) according to the revised cell-of-origin classification (n=455).



Underlying FL	72	62	61	59	54	48
BCL2−nonFL	90	77	71	67	65	65
Transformed FL	12	10	8	6	6	6

Figure S14: **Comparison between the survival of R-CHOP treated patients with transformed follicular lymphoma, non-transformed underlying follicular lymphoma, and other BCL2-group DLBCL patients with no evidence of follicular lymphoma.**

2 Supplementary Methods and Results

2.1 Sequencing, Variant Calling, and Annotation

Sequencing

For each sample, 50-200ng of genomic DNA was sheared using a Covaris LE220 focused ultrasonicator (Covaris) to produce 100-200bp fragments. Indexed libraries were generated using a modified version of the SureSelect XT protocol (Agilent Technologies), pooled (16-plex) and captured with a bespoke set of 120nt biotinylated RNA baits (Agilent Technologies) covering 293 genes implicated in haematological malignancy (Table S1) using the SureDesign interface (Agilent Technologies) on default parameters (all coding exons of all genes targeted with 10 flanking bases at 3 and 5 end of each exon). Capture-libraries were quantified, assessed for size distribution and quality, and sequenced on Illumina HiSeq 2500 instruments using 75 base paired-end sequencing according to the manufacturer's instructions. The average read depth across all samples was 500x reads per base.

Variant calling

Short insert paired-end reads were aligned to the reference human genome (GRCh37) using the Burrows-Wheeler Aligner. The variant calling pipeline of the Cancer Genome Project, Wellcome Trust Sanger Institute, was used to call substitutions (CaVEMan: Cancer Variants Through Expectation Maximization) and indels (Pindel). Copy number analysis was performed using GeneCN.

For substitutions, CaVEMan was run using a composite normal control (as only tumour samples were sequenced in this study). Unmapped reads, PCR duplicates and off-target variants were removed. Post-processing was then performed to remove likely artefact, which involved the removal of variants meeting the following criteria:

- variant base position supported by < 10 total reads
- variant supported by < 3 reads reporting the variant
- variant with an allele fraction < 0.05
- variant with a repeat length > 4 in a region present in $> 10\%$ of normal individuals
- less than one third of variant alleles at minimum base quality of 25
- composite normal control harbours 3% or more reads reporting the variant allele at minimum base quality of 15
- variant alleles lacking bidirectional support (< 2 supporting reads in each direction)
- variants lacking bidirectional support where mean variant base quality was less than 21 ($< 4\%$ supporting reads in each direction)
- variants falling within the second half of a read containing a GGC[AT]G motif in sequenced orientation where the mean base quality after the motif was less than 20
- mean mapping quality of the variant allele < 21
- variant falls within a simple or centromeric repeat
- more than 10% of reads reporting the variant contain an indel

- more than 80% of reads contain the variant allele at the same read position
- variant falls within a blacklisted region (based on coordinates) known to generate artefactual results
- variant is reported by ≥ 3 reads in $\geq 1\%$ of samples in composite normal control sample

For indels, Pindel was run using a composite normal control (as only tumour samples were sequenced in this study). Unmapped reads, PCR duplicates and off target variants were removed. Post-processing was then performed to remove likely artefact using the filters built into Pindel136. Variants meeting the following criteria were also removed.

- variant base position supported by < 10 total reads
- variant supported by < 3 reads reporting the variant
- variant with an allele fraction < 0.05
- variant with a repeat length > 4 in a region present in $> 10\%$ of normal individuals

Variant annotation

Germline variants were filtered by referencing against the ExAC database using non-TCGA samples (alleles present 10 or more times were excluded). Each variant was then annotated according to likely biological effect, as either a driver event, a passenger event, or a reflection of somatic hypermutation. Any gene disrupting events targeting tumour suppressor genes were classed as driver variants (frameshift, nonsense, essential splice [-2, -1, +1, +2, +5], loss of start, in-frame indel ≥ 2 codons). Missense variants and in-frame indels in either tumour suppressor genes or oncogenes were classed as driver alleles based on codon-level recurrence if any of the following conditions were met: (i) ≥ 5 samples with a variant at the codon in a cohort of 1,529 lymphoma cases, (ii) ≥ 10 cases reported across published datasets: COSMIC, AACR-GENIE v1.0 and PMID 28985567, or (iii) known driver event with biological support from published literature. Splice variants in oncogenes were annotated as driver events using the same rules as missense variants. For a small number of genes with high frequency aberrant somatic hypermutation the variant annotation strategy was modified to also include all bona fide variants within 2kb of the transcriptional start site. These genes are indicated in Table S1. Only alleles annotated as either driver events or somatic hypermutation were included in the downstream analysis.

Gene-level copy number calls were corrected for tumour cellularity using mean allele fractions of missense variants (driver or passenger; variants in somatic hypermutation genes and other subclonal outliers were excluded). Amplifications were classed as driver events if they targeted a known oncogene and resulted in predicted copy number of ≥ 6. Deletions were classed as driver events if they targeted a known tumour suppressor gene and resulted in heterozygous or homozygous loss (as defined by the log2 parameter for the gene).

2.2 Cluster structure under alternative information criteria

As described in the main paper, in order to identify groups with similar genetic characteristics, data were modelled as a finite mixture of Bernoulli distributions, providing a data-driven probabilistic interpretation of group membership strength. In the main paper, the number of identifiable clusters was selected using the Akaike Information Criterion (AIC). In order to investigate the sensitivity of our analysis to the criterion used to choose the number of clusters, an alternative information criterion, the Integrated Completed Likelihood (ICL) was employed. For the sample size used in this study, the ICL is comparatively conservative, favouring a smaller number of broader groups, whereas the AIC is less stringent, producing a larger number of more narrowly defined groups. The use of the Bayesian Information Criterion (BIC) was also investigated, but since it selected the same clusters as the ICL, only the results from using the AIC and the ICL are reported.

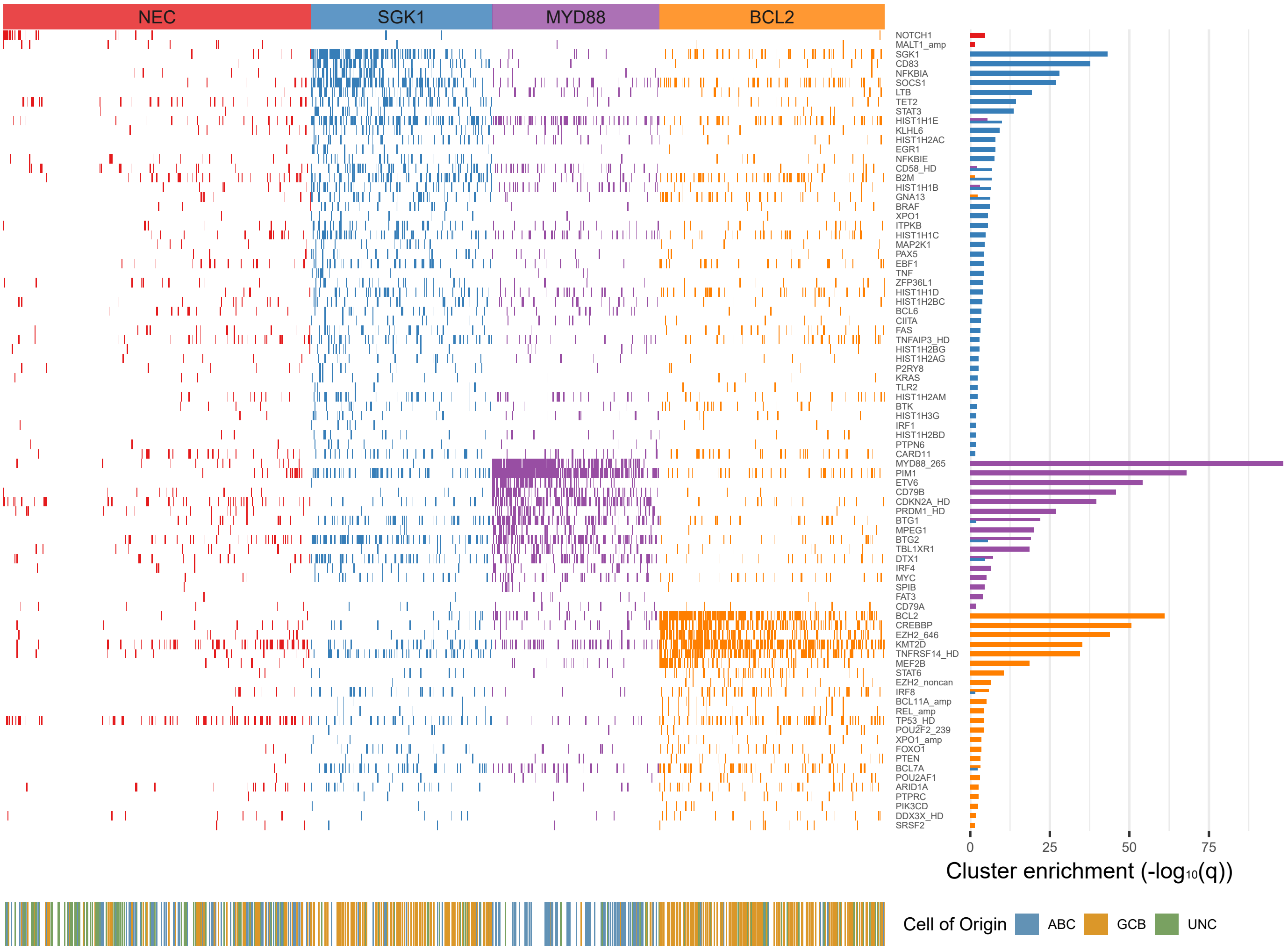


Figure S3: **Heatmap displaying the four genetic clusters that were identified using the ICL criterion in the analysis cohort (n=928).** Only those mutations are shown that are identified as significantly enriched for the given group, as determined by a Benjamini-Hochberg adjusted $q < 0.05$ from a chi-squared test of independence. "HD" stands for homozygous deletion or a mutation in this gene, "noncan" denotes a non-canonical mutation, and "amp" indicates an amplification.

The heatmap for the ICL-based cluster analysis, together with the cluster enrichments, are shown in Figure S3, as a parallel to Figure 2 in the main paper. Patient characteristics in each cluster are described in Table S5, the parallel to Table 2 in the main paper. Survival outcomes are shown in Figure S12, the parallel to Figure 2 in the main paper. Figure S4 shows the relationship between the ICL-based and the AIC-based cluster assignments. The ICL-based MYD88 & BCL2 clusters remain largely unchanged using the more relaxed AIC, while the SOCS1/SGK1 & TET2/SGK1 groups emerge from the SGK1 group. In addition, the NOTCH2 group emerges largely from ICL-uncategorized patients.

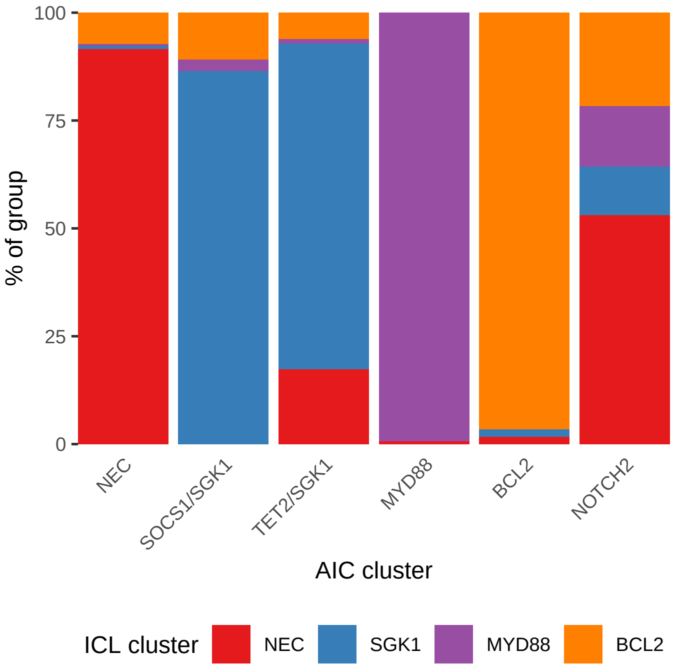


Figure S4: **Relationship between ICL and AIC criteria cluster membership**, highlighting that MYD88 and BCL2 remain largely unchanged under the more relaxed AIC penalisation, while SOCS1/SGK1 and TET2/SGK1 emerge from the SGK1 group. The bars represent the proportion of AIC cluster that were in the ICL cluster denoted by the bar colour

This graded approach to cluster identification, using both the ICL and the AIC criteria, allowed us to demonstrate the robustness of the MYD88, SGK1 and BCL2 ICL-based clusters, closely recapitulating those already described in the literature. AIC analysis allowed us to divide the SGK1 ICL-based cluster (that maps to the C4 cluster proposed by Chapuy et al.) into the TET2/SGK1 and SOCS1/SGK1 groups; the distinction between these two subtypes being supported by gene expression signatures of the RAS and STAT pathways respectively.

Using both information criteria in tandem, together with the assessment of cluster stability (described in the next supplementary section), provides a more complete assessment of cluster structure and strength. The most robustly identified groups being present under both information criteria, and replicating well under resampling.

2.3 Assessing cluster stability through consensus clustering

Consensus clustering

Consensus clustering provides a framework to describe the stability of statistical clusters with respect to changes in the dataset (Monti et al. 2003). A large number of perturbed datasets are resampled from the original dataset and the same clustering algorithm is run on each of these to generate resampled clusters. The stability is then measured by how frequently pairs of individuals are assigned to the same group. It is a simple algorithm but flexible and agnostic to the clustering method, the resampling method, and even allows for perturbation of the features themselves, although this latter aspect has not been included in this analysis as it is a reasonable assumption that all relevant mutations will be included on a diagnostic panel.

The procedure is briefly described below:

- For $i = 1, \dots, N$ (suitably large number, 1,000 used here)
  - Draw a bootstrap sample of the original dataset
  - Save an inclusion matrix I_i where each cell is 1 if both those individuals were included in this bootstrap sample or 0 if not (928 x 928 of binary 1/0)
  - Fit the mixture model as before using the same list of variables on the bootstrapped sample
  - For each of AIC and ICL:
    - Calculate a connectivity matrix C_i where each cell measures the similarity of two observations' cluster assignments through the dot product of their assigned probabilities, i.e. for observation j, k: $Pr(z_j = \text{NEC}) * Pr(z_k = \text{NEC}) + Pr(z_j = \text{MYD88}) * Pr(z_k = \text{MYD88}) + \dots)$ where z_j is the assigned cluster for individual j (another 928 x 928 matrix with values in [0,1])
- Sum up an overall inclusion matrix $I = \sum_{i=1}^{N} I_i$
- For each of AIC and ICL:
  - Sum up an overall connectivity matrix $C = \sum_{i=1}^{N} C_i$
  - The consensus matrix M is the connectivity matrix normalised by the number of runs each pairwise combination of individuals are sampled together $M = \frac{P}{I}$

Number of clusters chosen by ICL and AIC

Over the 1,000 bootstrap samples the ICL is extremely consistent in the number of clusters that are identified, identifying the optimal $k = 4$ in 98% of runs (below). As expected, the AIC is more variable, owing to the more relaxed penalty and thus placing a greater emphasis on the model likelihood (that increases with k) than ICL. Under the ICL, 6 clusters are identified in 64% of the samples, with the next most common k being 7, which occurs in 34% of runs.

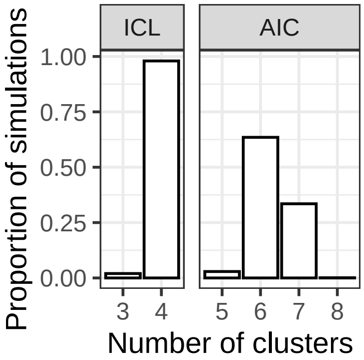


Figure S5: Number of clusters identified by ICL and AIC criteria over bootstrapping.

Cluster stability

The 928 x 928 consensus matrix M for the ICL is displayed below, ordered by the assigned genetic subtype from the complete dataset. A value of 1 for a pair of observations denotes that the pair had the exact same cluster assignment probabilities on every bootstrap sample where both observations were sampled, while a zero means they didn't have any overlap in probabilities no matter how small.

From M, a measure of genetic subtype stability can be expressed as the average consensus between all pairs of observations assigned to the same 'ground-truth' genetic subtype on the complete dataset. Formally, where L_k denotes the indices of observations belonging to ground-truth subtype k, the expression for cluster stability S_k is given by Eq 3 in Monti et al. (2003).

$$S_k = \frac{1}{N_k(N_k - 1)/2} \sum_{i,j \in L_k, i<j} M(i,j)$$

The stability scores are shown below in descending order, providing quantitative evidence that MYD88 is the most stable genetic subtype, followed by BCL2 (NEC is not counted as a distinct subtype, rather it is a catch-else group). The subtypes in the heatmap are ordered by these scores as well.

```
##     MYD88       NEC      BCL2      SGK1
## 0.8739051 0.8477668 0.7691424 0.7351941
```

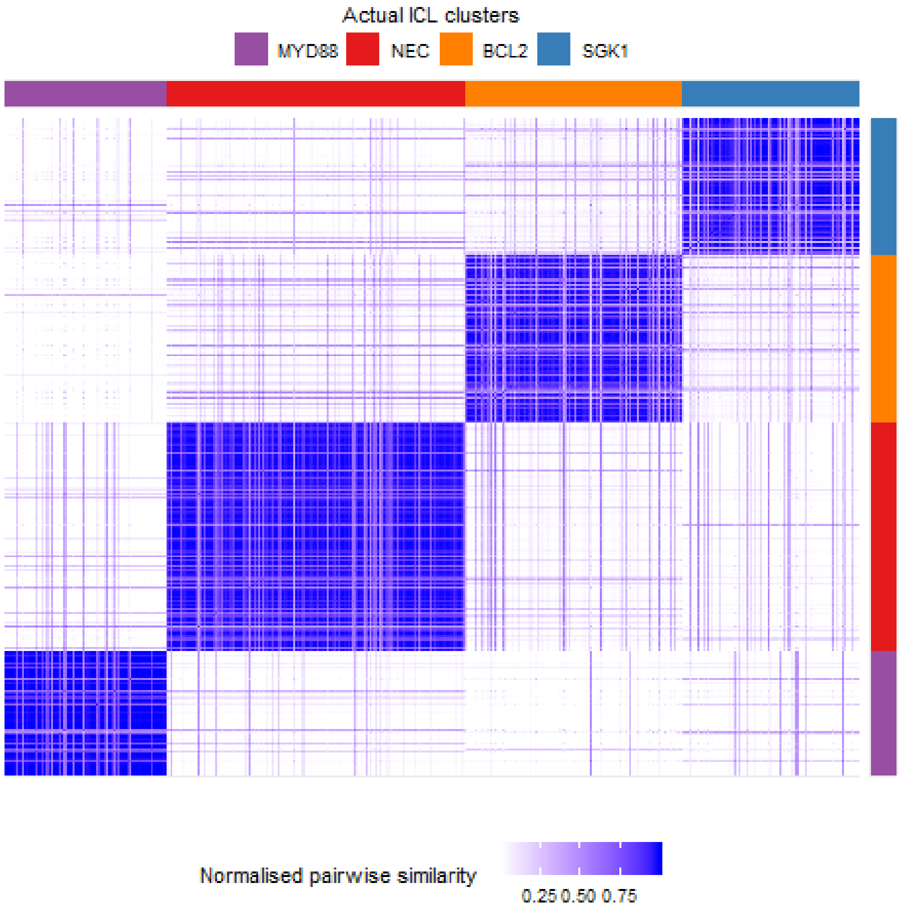


Figure S6: Consensus matrix of ICL selected clusters ordered by stability scores of genetic subtypes decreasing from left to right.

The connectivity matrix for the AIC selected clusters and the subtype stability scores are displayed below. MYD88 and BCL2 are once again the most robust subtypes, followed by the three new groups. It is unsurprising that SOCS1/SGK1 has a higher stability score than TET2/SGK1 since it is closest in constitution to the ICL SGK1 group. The NOTCH2 group are the least well defined of all seen here.

Furthermore, note that the stability scores are lower for the AIC than ICL groups across the board, again providing statistical evidence that the AIC identifies a greater number of more weakly evidenced groups.

```
##      MYD88        NEC       BCL2 SOCS1/SGK1  TET2/SGK1     NOTCH2
## 0.8244861  0.6782058  0.6025819  0.5733697  0.4416366  0.3775761
```

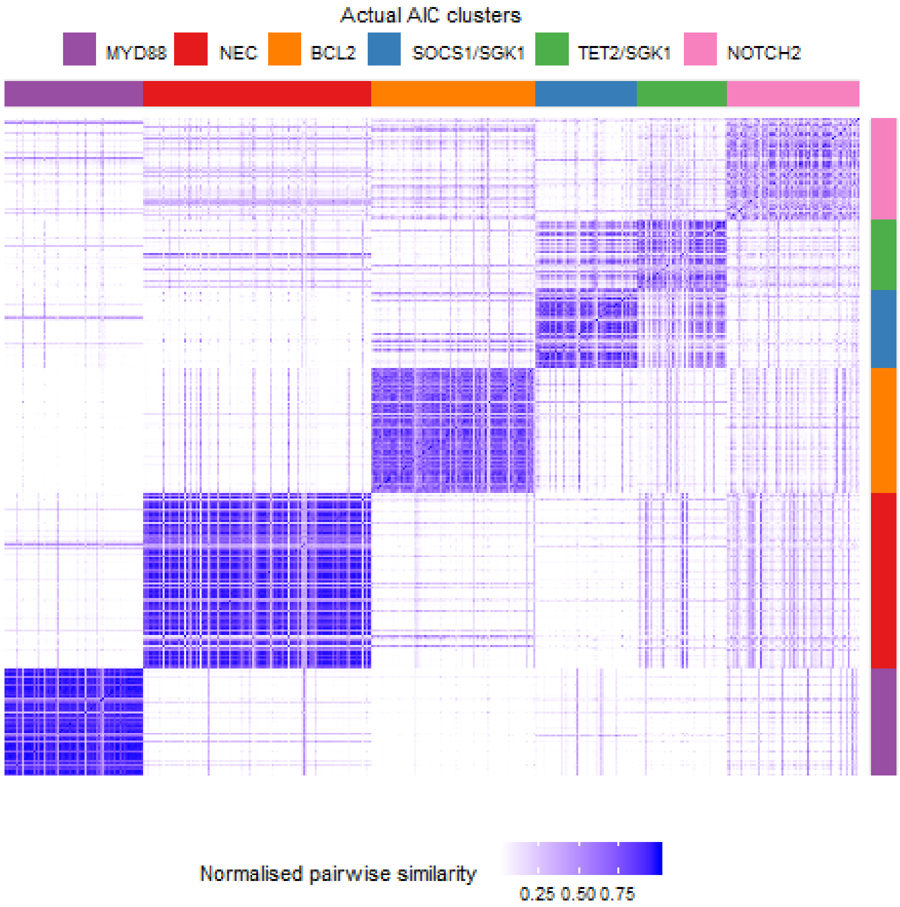


Figure S7: Consensus matrix of AIC selected clusters ordered by stability scores of genetic subtypes decreasing from left to right.

References

Monti, Stefano, Pablo Tamayo, Jill Mesirov, and Todd Golub. 2003. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data." *Machine Learning* 52 (July): 91–118.

`https://doi.org/10.1023/A:1023949509487`.

2.4 Clustering of the homogeneous survival cohort

A further analysis investigating the stability of the clusters reported in the main paper (AIC-based clusters) and in this supplement (ICL-based clusters) repeats our clustering procedure on the subset of patients comprising the de novo R-CHOP treated DLBCL patients of the survival cohort.

The clustering from this homogeneous dataset (n=579) is consistent with the groupings found in the full dataset (n=928), with the only exception being the recombination of the TET2/SGK1 and SOCS1/SGK1 clusters (Figure S8). Given the fact that these AIC-defined clusters emerge from a single ICL-defined SGK1 cluster, their recombination is consistent, and not surprising given the reduction in sample size.

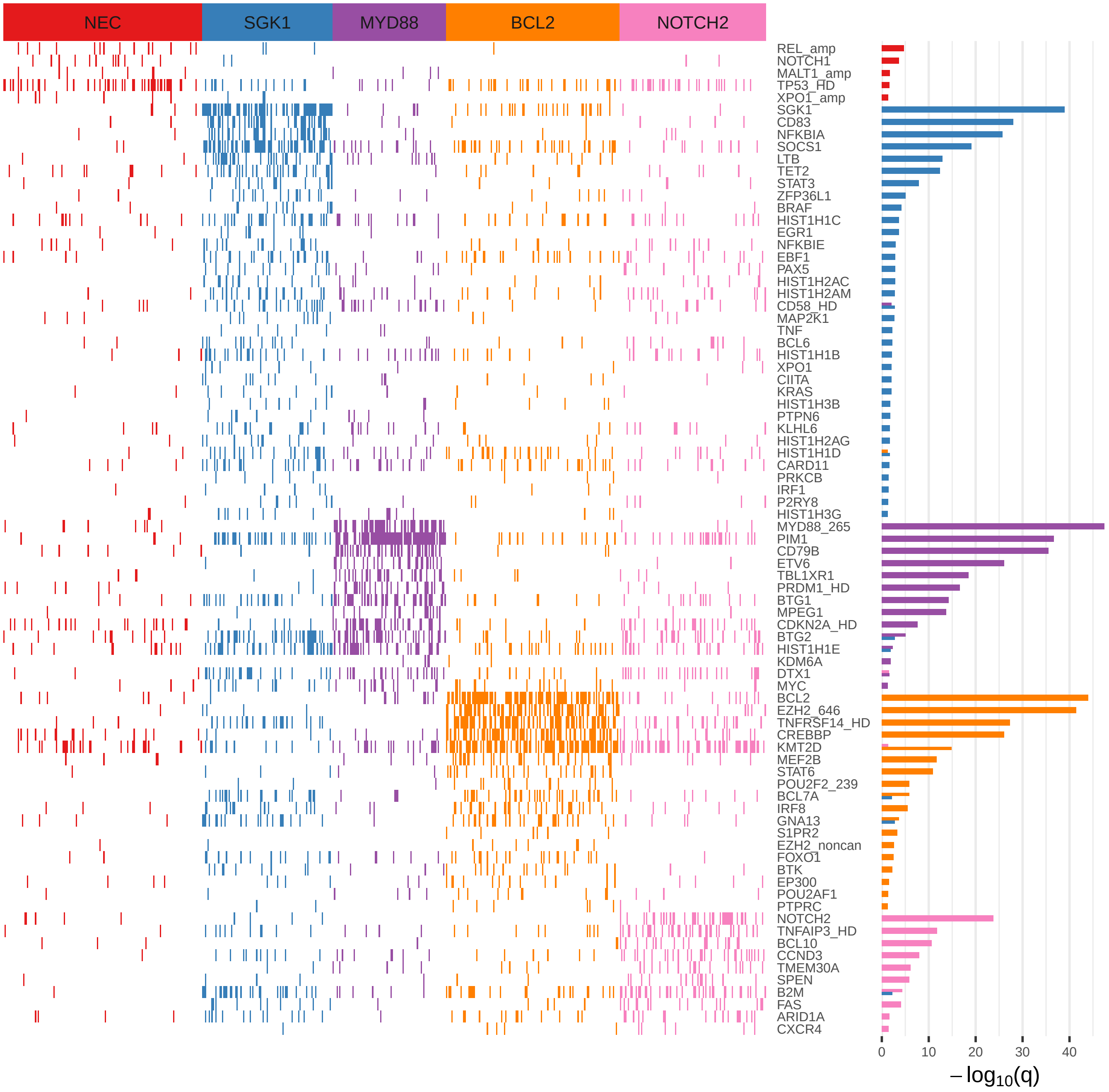


Figure S8: **Heatmap of the enriched mutations from the five clusters identified using AIC mixture modelling of the 579 de-novo DLBCL NOS who were treated with R-CHOP.**

2.5 Application of Bernoulli clustering to the dataset of Chapuy et al.

In order to assess how our analysis corresponds to previous published analyses, we ran the individual patient level data available from the study of Chapuy et al. through our analysis pipeline. The genomic panel used by Chapuy et al. uses 158 features (85, mutations, 32, gains, 33 losses and 8 SNV). We performed two analyses: in the first we used all of the 158 genetic features employed by Chapuy; in the second we used the 60 genetic features shared in common with our panel. (NB: we have no CNA or SNV in common).

Our mixture modelling strategy applied to the data of Chapuy et al. (using the AIC for consistency with our main analysis) identifies six clusters from the 158-feature set, which are largely concordant with Chapuy;s published subtype (Figure S10). When restricted to the 60 features that are common with our panel, five clusters were identified; losing the group defined by TP53 mutation, the majority of whom were reassigned as NEC (Figure S11).

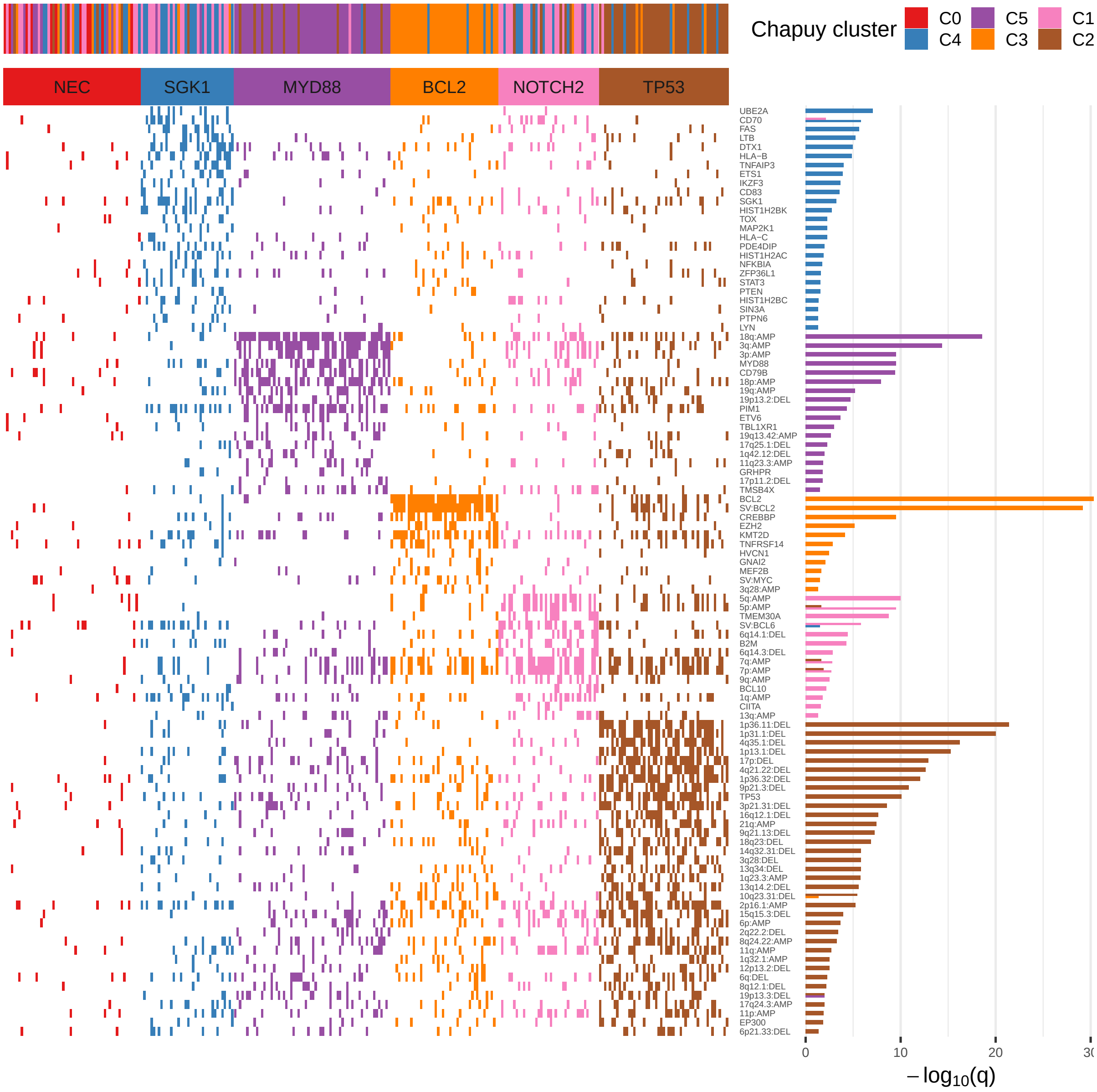
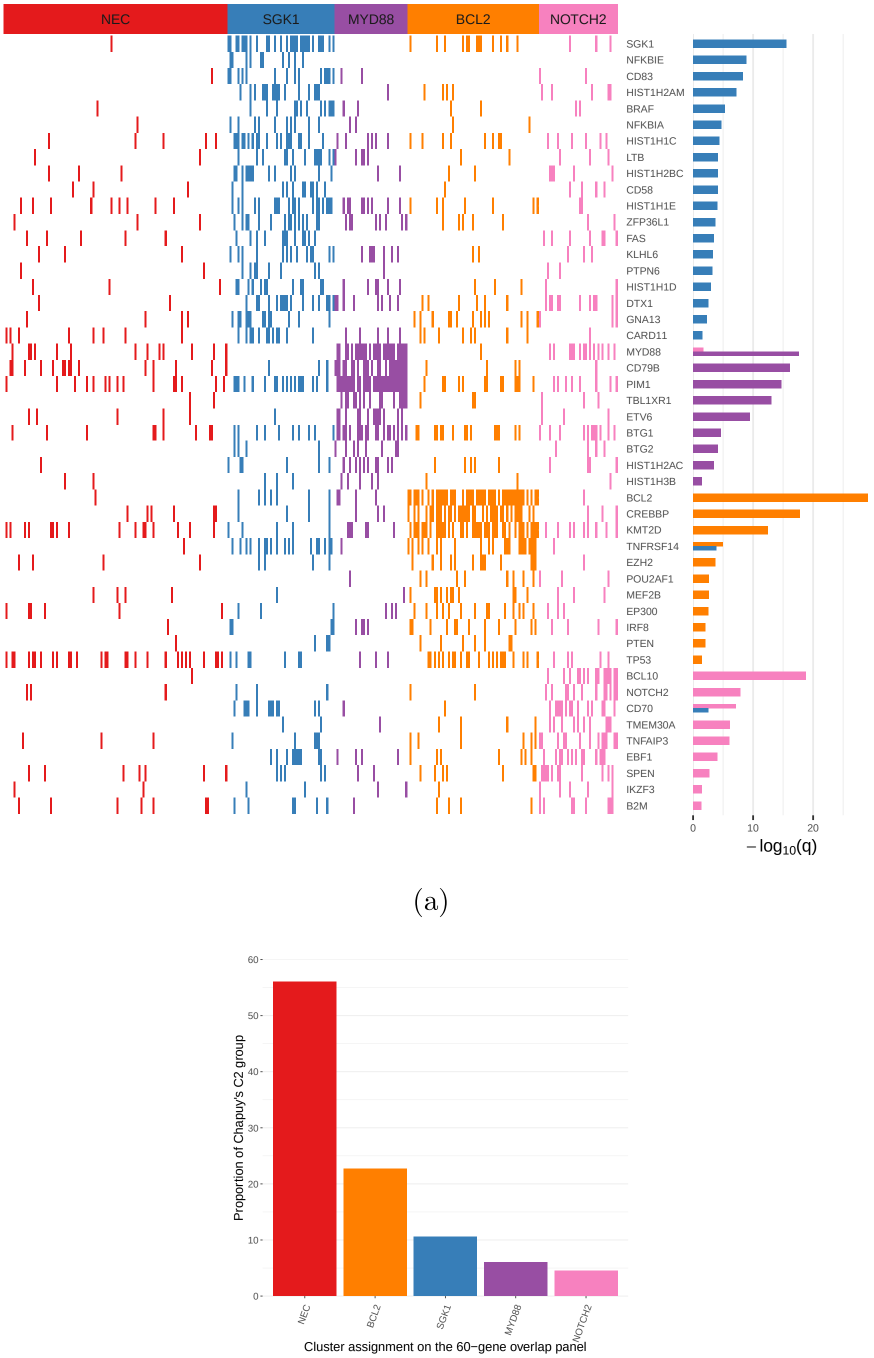


Figure S10: **Resultant heatmap of enriched genetic features from AIC mixture model clustering of Chapuy et al's dataset using their 158 panel.** The top bar denotes the original assigned subtype as published in Chapuy et al.



(a)

(b)

Figure S11: **Results from AIC mixture model clustering of Chapuy et al's dataset restricted to the 60 mutations that are shared in common with our panel.** a) Heatmap of enriched mutations. b) Reassignments of the members of Chapuy et al's C2 group.

Table S7 summarises the three groupings found by analysis of the dataset of Chapuy et al. (i) Those originally found by Chapuy et al, (ii) the six groups found when using mixture modelling on the full dataset, and (iii) the 5 groups found by mixture modelling using the 60 mutations in common. The subtype name in the left-most column has been determined by inspection of the most characteristic features of each cluster, determined by q-value. The five features with the highest q-values per cluster are displayed in the table.

Table S7: Comparison between the enriched mutations of each study

Primary subtype identifier	NMF	Mixture modelling (158)	Mixture modelling (60)
NM/NEC			
NOTCH2	BCL6 translocation	5q	BCL10
	BCL10	5p	NOTCH2
	TNFAIP3	TMEM30A	CD70
	UBE2A	BCL6 translocation	TMEM30A
	CD70	6q14.1	TNFAIP3
TP53	TP53	1p36.11	NA
	17p	1p31.1	
	21q	4q35.1	
	9p21.3	1p13.1	
	9q21.13	17p	
BCL2	BCL2	BCL2	BCL2
	BCL2 translocation	BCL2 translocation	CREBBP
	CREBBP	CREBBP	KMT2D
	EZH2	EZH2	TNFRSF14
	KMT2D	KMT2D	EZH2
SKG1	SGK1	UBE2A	SGK1
	HIST1H1E	CD70	NFKBIE
	NFKBIE	FAS	CD83
	BRAF	LTB	HIST1H2AM
	CD83	DTX1	BRAF
MYD88	18q	18q	MYD88
	3q	3q	CD79B
	CD79B	3p	PIM1
	3p	MYD88	TBL1XR1
	MYD88	CD79B	ETV6

The lack of complete concordance between Chapuy et als reported subtypes and those identified using the 60-mutation panel can be explained by the absence of certain mutations that were key in our analysis, such as: SOCS1, TET2 (both SGK1 enriched), and CDKN2A, MPEG1, IRF4, MYC (all MYD88 enriched).

Given that the sole difference between the two mixture modelling experimental setups is the exclusion of CNA and other features not available in our dataset, we can conclude that the primary reason why our study did not identify a TP53 group is the lack of CNA predictor variables, which correlate strongly with TP53 mutations. Our re-analysis of the data from Chapuy et al. supports the existence of a TP53 group.