Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a Haematological Malignancy Research Network report

Stuart E Lacy PhD*[1], Sharon Barrans PhD*[2], Philip A Beer PhD*[3], Daniel Painter BSc[1], Alexandra Smith PhD[1], Eve Roman PhD[1], Susanna L Cooke PhD[4], Camilo Ruiz MPhil[3], Paul Glover MSci[2], Suzan Van Hoppe MSc[2], Nichola Webster MSc[2], Peter Campbell PhD[3], Reuben Tooze PhD[5], Russell Patmore BMedSci[6], Cathy Burton MD[§2], Simon Crouch PhD[§1], and Daniel J Hodson PhD[§7].

1: Epidemiology and Cancer Statistics Group, Department of Health Sciences, University of York, York, YO10 5DD, United Kingdom.
2: Haematological Malignancy Diagnostic Service, St. James's Institute of Oncology, Leeds, LS9 7TF, United Kingdom.
3: Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom.
4: Institute of Cancer Sciences, University of Glasgow, Glasgow, United Kingdom.
5: Section of Experimental Haematology, Leeds Institute of Molecular Medicine, University of Leeds, LS2 9JT, United Kingdom.
6: Queen's Centre for Oncology and Haematology, Castle Hill Hospital, Cottingham, HU16 5JQ, United Kingdom.
7: Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, CB2 0AW, United Kingdom.

Key Points

- Robust sub-types of DLBCL are identified by model-based clustering of genetic mutations in a large (n=928) population-based cohort.
- With full follow-up data available for all sequenced patients, the prognostic significance of these sub-types is identified.

Abstract

Based on the profile of genetic alterations occurring in tumor samples from selected diffuse-large-B-cell-lymphoma (DLBCL) patients, two recent whole exome sequencing studies proposed partially overlapping classification systems. Using clustering techniques applied to targeted sequencing data derived from a large unselected population-based patient cohort with full clinical follow-up (n=928), we investigated whether molecular subtypes can be robustly identified using methods potentially applicable in routine clinical practice.

DNA extracted from DLBCL tumors diagnosed in patients residing in a catchment population of ~4 million (14 centers), were sequenced with a targeted 293-gene hematological-malignancy panel. Bernoulli mixture-model clustering was applied, and the resulting subtypes analyzed in relation to their clinical characteristics and outcomes. Five molecular subtypes were resolved, termed MYD88, BCL2, SOCS1/SGK1, TET2/SGK1 and NOTCH2, along with an unclassified group. The subtypes characterized by genetic alterations of BCL2, NOTCH2 and MYD88 respectively recapitulated recent studies showing good, intermediate and poor prognosis respectively. The SOCS1/SGK1 subtype showed biological overlap with primary mediastinal B-cell lymphoma and conferred excellent prognosis. Although not identified as a distinct cluster, NOTCH1 mutation was associated with poor prognosis. The impact of TP53 mutation varied with genomic subtypes, conferring no effect in the NOTCH2 subtype and poor prognosis in the MYD88 subtype.

Our findings confirm the existence of molecular subtypes of DLBCL, providing evidence that genomic tests have prognostic significance in non-selected DLBCL patients. The identification of both good and poor risk subtypes in R-CHOP treated patients clearly demonstrate the clinical value of the approach; confirming the need for a consensus classification.

Introduction

Diffuse-large-B-cell-lymphoma (DLBCL) is the commonest non-Hodgkin lymphoma[1–3]. Although potentially curable with immunochemotherapy, refractory or relapsed lymphoma occurs in approximately 40% of patients. Despite the substantial increase in biological understanding in recent years, attempts to improve survival by combining standard therapy with novel targeted agents have thus far yielded disappointing results, with no phase 3 trial leading to a change in the accepted standard of care since the addition of rituximab to CHOP chemotherapy in 2002[4]. One barrier to the effective use of novel therapies targeting specific pathways is the biological heterogeneity of DLBCL and the likely existence of multiple distinct disease subtypes. Hence, in order to permit more accurate targeting in clinical trials it is becoming increasingly important to define these subtypes; permitting stratification that separates patients likely to achieve cure with R-CHOP alone from high-risk groups that may benefit from emerging therapies.

Gene expression profiling distinguishes transcriptional subtypes of DLBCL including activated-B-cell-like (ABC) and germinal-center-B-cell-like (GCB) in the Cell-of-Origin (COO) classification, and more recently, molecular-high-grade (MHG)[5–8]. Genomic studies have demonstrated additional heterogeneity[9–15]; the three latest using whole exome sequencing (WES) to describe molecular subtypes based on the profile of genetic alterations within each tumour[16–18]. Two of these studies proposed partially overlapping classifications[16,17]; these converging conclusions suggesting the existence of convincing molecular subtypes with distinct biology and the potential to guide therapeutic targeting.

However, several questions remain before a consensus classification can be implemented in clinical practice. The first relates to robustness, since it is unclear how variation introduced by different sequencing platforms, variant calling algorithms, biopsy material, and methods of statistical analysis impact on the ability to resolve genetic subtypes. The second relates to the practicalities of implementing a classification in a real-world setting; the ability to resolve genetic subtypes using panel-based sequencing on DNA from both fresh and FFPE material. Finally, studies conducted to date have been largely based on specimens and data from clinical trials and specialist referral center archives; potentially limiting the generalizability of findings to the patient population as a whole.

Established with the aim of addressing such questions, the present report describes results from a genomics study embedded within a contemporary UK ‘real-world’ population-based

patient cohort, where all diagnoses in a catchment population of around 4 million (14 hospitals) are centrally made at a specialized hematological oncology reference laboratory.[19] Using surplus material archived at the time of diagnosis, 4244 lymphoid and myeloid tumor samples have now been sequenced and assessed for somatic mutations against a pan-hematological malignancy panel of 293 genes; the findings for 928 patients diagnosed with DLBCL are reported here.

Materials & Methods

Patients and Procedures

Data are from the UK's population-based Haematological Malignancy Research Network (https://www.hmrn.org)[19]. Initiated in 2004, all diagnoses within HMRN's boundaries are made and coded by clinical specialists at a single integrated hematopathology laboratory – the Haematological Malignancy Diagnostic Service (www.HMDS.info). Full details of HMRN's methods and ethical approvals are published elsewhere; ethical approval for HMRN granted under REC reference 04/Q1205/69 and for the genetic sequencing under 14/WS/0098[19,20]. In brief, covering 14 hospitals and tracking all patients with hematological malignancies (~2,400 per year) through clinical and national administrative systems (mortality and morbidity), HMRN's patient cohort operates under a legal basis that permits full treatment and outcome data to be collected from clinical records without explicit consent. The study cohort for the present report comprised 2358 subjects newly diagnosed with diffuse large B-cell lymphoma (DLBCL) between 1st September 2004 and 31st August 2012: International Classification of Diseases for Oncology, third edition (ICD-O3) 9679, 9680, 9688, 9712, 9735. Of these, 928 (39·4%) had suitable diagnostic material for genetic analysis (Figure S1). All patients were followed-up for mortality until December 31st 2018.

DNA Sequencing

DNA was extracted from surplus FFPE material archived at the time of diagnostic biopsy using the Qiagen QIAamp DNA FFPE tissue kit (Cat no. 51306). 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 hematological 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. Full details of sequencing, variant calling and annotation are provided in the supplement (section 2.1). Single nucleotide variants and copy number variants are detailed in Tables S2 & S3.

Statistical analysis

117 genetic features occurred in at least 1% of patients; these features defined binary variables denoting 105 mutations, 4 amplifications, and 8 markers indicating presence of either a homozygous deletion or amplification (Table S1). These 117 binary variables were used to identify sub-groups of the patients with similar genetic characteristics. Genetic features found in <1% of cases were not used for clustering.

In order to identify these genetic sub-groups, the data were modelled as a finite mixture of Bernoulli distributions, providing a data-driven probabilistic interpretation of group membership strength[21]. The number of identifiable clusters was selected using the Akaike Information Criterion (AIC) likelihood penalization method.

Three further techniques were employed to assess cluster strength and stability. Firstly, the Integrated Completed Likelihood (ICL)[22] was employed to determine cluster number; secondly, cluster stability under repeated resampling was investigated by consensus clustering[23]; and thirdly, the clustering process was restricted to the homogeneous group of de novo DLBCL NOS patients treated with R-CHOP (n=579). Full details of this sensitivity analysis are presented in the supplement (sections 2.2-2.4).

Survival analysis was carried out using the Kaplan-Meier estimator and proportional-hazards regression; with all patients followed-up for mortality until December 31st 2018. Analyses were conducted in Stata 15.126[24] and R 3.5.3[25], using the libraries flexmix[21] and survival[26]. Gene expression profiling was available for a subset of cases from a previous analysis, where full details of the methods are provided.[7,8] Gene Set Enrichment Analysis (GSEA) was run using software developed by the Broad Institute[27,28]. Pearson's Chi-squared test was used for assessing differences in proportions, with the Benjamini-Hochberg adjustment for multiple-

comparisons. The data used in this study are presented in Tables S4 & S6; the R-code is available on Github, at: https://github.com/ecsg-uoy/DLBCLGenomicSubtyping

Role of the Funding Sources

Bloodwise had no role in the design of the study, data collection, data analysis, data interpretation, or writing of the report. 14M Genomics were involved in the design of the pan-hematological gene panel, but took no further part in the study.

Results

Genetic substructure of DLBCL

Demographic and clinical characteristics of the study cohort (n=2358) and analysis cohort (n=928) are summarized in Table 1. Although broadly similar to the study cohort as a whole, those with archived material of sufficient quality for NGS (analysis cohort) were, on average, slightly younger (median diagnostic age 68·5 versus 70·0 years), and were more likely to have localized disease (e.g. 33% versus 30·5% stage I/II disease), and to have been treated with curative intent (86·5% versus 81·8%).

Individual biopsies were associated with a median of 7 driver mutations (Figure S2). In 49 patients (5·3%) no genetic abnormality was detected. Genetic alterations stratified by COO, confirmed that when applied to FFPE material our sequencing platform and variant calling strategy identified mutational profiles in close agreement with previous studies[16–18] (Figure 1 & S2).

Applying the Akaike Information Criterion (AIC) to determine optimal cluster number, five genomic clusters were identified. These were named MYD88, BCL2, TET2/SGK1, SOCS1/SGK1 and NOTCH2, according to the genetic features most enriched in each cluster, leaving 27% of patients as “Not Elsewhere Classified” (NEC) (Figure 2). Information on the genetic clusters identified using the ICL criterion, and an assessment of cluster stability by consensus clustering is presented in the supplement (sections 2.2 & 2.3, Figures S3-S7, Table S5).

The MYD88 cluster (n=152) was dominated by mutation of *MYD88* (L265P), *PIM1, CD79B*, *ETV6* and frequent loss of *CDKN2A*; the available gene expression data showing that most belonged to the ABC subtype, with enrichment for signatures associated with ABC-DLBCL, IRF4 and MYC (Figure 1). In order to assess biological features of DLBCL independent of anatomical

site of presentation we included special site DLBCL in our primary clustering. However, very similar results were observed when such cases were excluded and clustering restricted to DLBCL-NOS (Figure S8). The majority of primary CNS lymphomas (PCNSL), along with primary testicular and those with breast involvement, clustered within this group (Table 2). This cluster strongly recapitulates the recently described MCD[16] and C5[17] subtypes.

The BCL2 cluster (n=176) showed frequent mutation of *EZH2, BCL2*, *CREBBP*, *TNFRSF14*, *KMT2D*, and *MEF2B*. The majority of cases for whom FISH was available (82/92) had a t(14:18) BCL2 translocation, and mutation of *BCL2* was strongly correlated with this translocation ($p < 0{\cdot}0001$). Gene expression revealed a predominance of GCB-DLBCL, but also enrichment for MHG DLBCL (Table 2, Figure 1). Most cases of transformed follicular lymphoma fell within this cluster, robustly mapping to the previously described EZB and C3 clusters[16,17] (Table 2).

The SOCS1/SGK1 group (n=111) was characterized by mutations including *SOCS1*, *CD83*, *SGK1*, *NFKBIA, HIST1H1E* and *STAT3*. Several of these genes, including SOCS1, were known targets of aberrant somatic hypermutation. This cluster appears to represent a subdivision of the recently described C4 cluster[17].

SOCS1 mutation is a finding shared with PMBCL. The inclusion of cases diagnosed as PMBCL in our clustering allowed the demonstration that 12/20 PMCBL cases clustered in this group along with 98 DLBCL-NOS cases. The latter cases in this cluster showed no pathological features of PMBCL and were not enriched for mediastinal involvement (Figure S9). Gene expression profiling showed that DLBCL-NOS cases in this group were predominantly of GCB origin and enriched for signatures associated with PMBCL and JAK/STAT signaling, suggesting a degree of biological similarity to PMBCL[29] (Figure 1). Enrichment was also seen for other genes previously noted to be enriched in PMBCL including ITPKB, NFKBIE, and CIITA[30] (Supplementary Figure 9A). These results suggest that PMBCL, a lymphoma with a unique anatomical location, shares significant molecular overlap with a subset of otherwise nodal DLBCL, supporting a previous description of non-mediastinal DLBCL tumors that share biological similarity to PMBCL.[29] This also recapitulates the findings observed for the relationship between PCNSL, primary testicular and primary breast DLCBL and the MYD88 group of DLCBL-NOS; supporting a more general overlap between special site lymphomas and specific molecular subsets of DLBCL-NOS.

The TET2/SGK1 (n=98) cluster was characterized by mutations including *TET2*, *SGK1*, *KLHL6*, *ZFP36L1*, *BRAF*, *MAP2K1* and *KRAS*. The mutation of multiple components of the ERK pathway was associated with enrichment of gene expression signatures of RAS and ERK (Figure 1). Predominantly GCB in origin, these cases appear to represent a second subdivision of the recently described C4 cluster[17].

A final NOTCH2 cluster (n=143) was characterized by mutations including *NOTCH2*, *BCL10*, *TNFAIP3*, *CCND3*, *SPEN*, *TMEM30A FAS* and *CD70*. Our targeted panel did not capture *BCL6* fusion status, but *BCL6* rearrangement by FISH, where available (317/928), confirmed a strong correlation with *NOTCH2* mutation ($p < 0{\cdot}0001$). Gene expression showed that this cluster comprised a mixture of ABC, GCB and UC DLBCL (Table 2). Enhanced NOTCH activity was not detected by gene expression, nor could we detect transcriptional evidence of increased NOTCH activity when restricting to patients with *NOTCH2* mutation. The mutational similarities suggest biological similarity to marginal zone lymphoma; a pre-existing diagnosis of marginal zone lymphoma was noted only in three patients. This group corresponds closely to the previously described BN2[16] or C1[17] subtypes.

Validation of our clustering strategy on an external dataset

To further explore the relationship of clusters defined in our data sets to those of other cohorts recently defined we took advantage of published data from Chapuy et al[17]. We considered this from two angles: firstly, applying our Bernoulli clustering approach to their total data-set; and secondly, restricting analysis to the features in common between our data set and theirs. With respect to the former, our algorithm identified six clusters; largely recapitulating those published by Chapuy et al. (Figure S10). Genetic features enriched within these clusters closely matched those enriched in the Chapuy data, as well as those enriched in clusters within our patient cohort (Table S7). However, we were not able to resolve a division of the SKG1 subgroup, which in our cohort splits into SOCS1/SGK1 and TET2/SGK1 subgroups, a finding likely to be explained by the lack of *SOCS1* and *TET2* mutation calls in the Chapuy data. Restricting the analysis to the 60 overlapping genetic features reproduced 5 of the original subgroups, but led to loss of the cluster defined by TP53 mutation and copy number alteration (CNA), which corresponded to the Chapuy C2 cluster (Figure S11). These cases were instead predominantly reassigned to the unclassified group (56%) and the BCL2 group (22.7%) (Figure S11B). Taken together these findings confirm that our clustering approach is robust across datasets and that the absence of a TP53 cluster in our own cohort relates to limited

copy number information available from our panel; the majority of cases likely to remain unclassified or allocated to the BCL2 group.

The clusters found in our main analysis are compared in summary to those of Chapuy et al[17] and Schmitz et al[16] in Table 3. A detailed comparison with Chapuy et al is presented in the supplement (section 2.5).

Genomic subtype and patient outcome.

Five-year overall survival estimates for all patients (n=928), and those treated with R-CHOP (n=648) are at the bottom of Table 2. For all patients, analyses revealed especially poor outcome for patients in the MYD88 cluster, with five-year OS of 42% (34·9-50·7). In contrast, patients within the SOCS1/SGK1, BCL2 and TET2/SGK1 clusters fared better than patients in the MYD88 group, with five-year OS of 64·9% (56·6-74·4), 62·5% (55·7-70·1) and 60·1% (51·1-70·6) respectively ($p < 10^{-3}$ for each group compared with MYD88). NEC and NOTCH2 clusters had intermediate survivals; five-year OS 53·6% (47·7-60·2) and 48·1% (40·5-57·0) respectively ($p = 0.008$, 0.063, respectively compared with MYD88).

To examine survival in a more homogeneous patient cohort we restricted our analysis to patients with de novo DLBCL NOS treated with curative intent (n=690). Similar survival trends were observed, poorest outcome seen amongst the MYD88 subgroup (Figures 3A,B, S12). Since some of these patients were treated with attenuated regimens that might be considered "R-CHOP-like", we repeated the analysis on the 579 patients with de novo DLBCL treated with R-CHOP (Figures 3,C,D, S12). Especially good survival was seen in SOCS1/SGK1 (n=84), where five-year OS was 80·2% (71·5-90·1). However, the apparently negative prognostic impact of MYD88 cluster membership (n=78) was reduced, increasing the OS estimate of this subgroup to 62·8% (53·0-74·5); although not statistically significant (p=0.1), this intriguing result suggests a specific sensitivity of the MYD88 subtype to either patient-intrinsic or treatment related factors that may partially explain previous controversy around its prognostic implication[31].

Inspection of clinical risk factors showed variation across molecular cluster; the NOTCH2 group being associated with the greatest number of patients with high-intermediate and high IPIs (Figure S13). The IPI adjusted hazard ratios for each molecular subtype confirming inferior

survival estimates in the MYD88 subtype, neutral effect of NOTCH2 and enhanced survival in the TET2/SGK1*,* SOCS1/SGK1 and BCL2 groups are shown in Figure 3.

Although our clustering algorithm did not resolve a distinct *NOTCH1* cluster, the majority of *NOTCH1* mutant cases were found within the unclassified NEC group, suggesting they may represent a distinct subtype too small to be detected by our modelling strategy. Indeed, the presence of *NOTCH1* mutation (n=16) conferred an especially poor outcome with a 5-year OS of just 39% (Figure 4A, p=0.004 NOTCH1 mutant vs. wildtype).

MYC rearrangement was most enriched within the BCL2 cluster (Figure S13), conferring poor prognosis in R-CHOP treated patients, 5-year OS 50·0%, (31·5-79·4); in comparison to patients without MYC rearrangement who had 71% OS (62·4-81·8) (Figure 4B). Similarly, most cases of double hit lymphoma and MHG were found in the BCL2 cluster, where both conferred a poor prognosis in R-CHOP treated patients (5-year OS 43·8% (25·1, 76·3) and 44·4% (21·4, 92·3) respectively). There were insufficient numbers of events to determine the effect of MYC rearrangement, DHL, or MHG in other clusters. (Figure S13). Also strongly enriched within the BCL2 cluster were cases of transformed follicular lymphoma and cases of DLBCL with a concurrent diagnosis of follicular lymphoma detected on the diagnostic biopsy. Whilst patients with transformed FL had an inferior survival estimates, in the BCL2 cluster no differences in survival between DLBCL with/without a concurrent diagnosis of follicular lymphoma was observed. This finding agrees with a recent study by Wang et al.[32] (Figure S14).

As noted above we did not identify a distinct TP53/CNA cluster in our cohort. The potential for other clusters, in particular the BCL2 subgroup, to contain cases that might otherwise have been classified to a TP53 cluster prompted us to examine survival of each cluster stratified by TP53 mutation status. Remarkably the prognostic impact of TP53 mutation varied considerably across the different molecular subtypes (Figure 4D). *TP53* mutation was associated with a worse prognosis in the NEC and BCL2 subtypes and, although uncommon, *TP53* mutation conferred an extremely poor prognosis in the MYD88 cluster. In contrast, there was no evidence of a prognostic effect in patients within the NOTCH2 or SOCS1/SGK1 clusters, and was rarely detected in the TET2/SGK1 cluster. This suggests that in addition to its presence in a TP53-mutant cluster, TP53 mutation may also be acquired by tumors belonging to other subgroups and that in this scenario its prognostic effect is subtype-dependent.

Discussion

Based on the genomic profile of individual tumors, our large population-based study defined five molecular DLBCL subtypes, named MYD88, BCL2, SOCS1/SGK1, TET2/SGK1 and NOTCH2, with each having distinct features in terms of both their biology and clinical outcome. With prospective tracking of treatment and outcome in 928 DLBCL patients (trial and non-trial) diagnosed by clinical staff at a single center, we escape the biases that may result from sequencing archived biopsy collections established from the selective referral of clinically or diagnostically challenging cases to specialist pathology centers. Indeed, the only determinant of whether patients were included in the analysis cohort was the availability of sufficient sample for the extraction of adequate DNA. This unselected cohort therefore provides a unique representation of real-world DLBCL that may not have been captured in previous molecular subtyping studies. We have not restricted the cluster analysis to any particular subtype of DLBCL, considering an a-priori approach to the evolving classification of these possibly heterogeneous disease types a particular strength of this study.

Considerable variability in genomic subtype identification can be introduced by differences in biopsy type, sequencing strategy, variant calling pipeline, driver annotation algorithms and the statistical method used to identify genomic clusters. It remained uncertain whether the genomic clusters described in recent studies represented sufficiently robust entities to be resolved against variation in each of the above factors. We utilized alternative, but widely accepted, pipelines for variant and driver annotation and different approaches to the identification of genomic clusters. Despite these considerable differences, we have independently converged on very similar genomic subtypes. Indeed, when comparing enriched mutations between equivalent clusters across studies we observe a very high degree of overlap. Therefore, our data strongly support the conclusion that DLBCL exists as molecularly distinct subtypes that are resolvable through genomic analysis. Importantly, we were able to identify these subtypes using a targeted sequencing panel applied to biopsy material, a strategy that would be applicable to a non-academic diagnostic laboratory.

The lack of BCL6 fusion data is a limitation of our study. However, available FISH data confirmed the association between *NOTCH2* mutation and rearrangement of *BCL6*. Furthermore, even in the absence of *BCL6* fusion data we were able to identify a *NOTCH2* subtype that shows similar mutation enrichment to the previously identified BN2/C1[16,17] clusters. Our study was also limited by the extent of copy number data available from our

panel. As such we did not resolve a subgroup mapping to the C2 cluster of Chapuy et al[17]. The application of our clustering algorithm to Chapuy's data, applying all or just overlapping genetic features, supports the existence of this subtype and the requirement for adequate copy number information for its detection. Future sequencing panels designed to assign DLBCL molecular subtype will likely include probes that target regions of recurrent copy number variability. The existence of a TP53/CNA cluster is further supported by the finding that the majority (56%) of these cases were not alternatively classified but rather assigned to an unclassified NEC group by our cluster algorithm (Figure S11). However, 23% of the TP53/CNA cases carried a mutational profile that, in the absence of copy number data, reassigns them to the BCL2 cluster. This leads us to hypothesize that TP53 mutation and associated copy number alterations may arise in two different contexts: either as the primary determinant of a distinct disease subtype or alternatively as a secondary event in a tumor arising from one of the other subgroup, most commonly the BCL2 cluster.

A novel finding in our study was the division of the previously identified SKG1 cluster[17] into SOCS1/SGK1 and TET2/SGK1 subgroups; the biological validity of which was supported by the enrichment of JAK/STAT and ERK gene expression signatures respectively. Driver mutations of *TET2* and *SOCS1* were not reported in the earlier study[17], which impaired the ability our clustering algorithm to resolve this split within the Chapuy data. These differences in mutation calling, likely due to differences in variant and driver annotation strategies, highlight some of the challenges to be overcome when implementing a consensus classification system. Interestingly our SOCS1/SGK1 subtype shared both mutation and GEP similarities with PMBCL, supporting a previous description of non-mediastinal DLBCL tumors that share biological similarity to PMBCL[29]. This analysis also has parallels with other subsets of special site lymphomas including PCNSL and those occurring de novo in testis and breast, which overlap with the MYD88 subgroup, highlighting a recurrent link between special site and nodal DLBCL-NOS across several molecular subsets.

Prospective clinical data on all cases (Table S4) provided greater insight into the prognostic impact of clusters, revealing how of survival of MYD88 clustered patients was especially sensitive either patient-intrinsic or treatment-related characteristics. This difference in outcome when analysis of MYD88 patients is restricted to those treated with full R-CHOP versus other R-CHOP-like regimens may explain some of the controversy around the prognostic impact of this group[31]. Multivariate analysis allowed us to quantify the relative risk

of individual subtypes independent of IPI factors, an analysis that reinforced the importance of clinical risk factors in determining prognosis. Most notably, we have revealed the extremely favorable outcome of patients in the SOCS1/SGK1 subtype, suggesting this group might be the subject of future trials examining de-escalation of therapy.

The large size of our cohort allowed us to examine the impact of individual factors within clusters. Of particular interest we were able to comment on the survival of patients in the BCL2 cluster with and without concurrent follicular lymphoma in the diagnostic biopsy. The similarity of survival in response to R-CHOP mirrors that recently reported[32] and leads us to hypothesize that DLBCL with concurrent follicular lymphoma may represent an entity that is biologically and clinically indistinguishable from other DLBCL patients clustered into the BCL2 subgroup, with the finding of concurrent follicular lymphoma dependent solely on the region of tumor captured by the biopsy. Conversely, transformed follicular lymphoma was associated with inferior survival, consistent with transformed follicular lymphoma representing a distinct clinical entity.

Individual genetic features allowed us to identify other high-risk groups. Although we did not identify a distinct *NOTCH1* cluster, the predominance of *NOTCH1* mutant cases within the NEC group and their especially poor outcome supports the concept that these patients should be considered separately in clinical practice. Many (27/66) translocated, and the majority (21/38) of double hit, lymphomas were classified into the BCL2 cluster, allowing very good and poor risk subgroups to be resolved from this cluster. Finally, we observe intriguing variation of the impact of TP53 mutation across clusters with no detectable effect in the NOTCH2 group, contrasting with the dismal prognosis conferred in the MYD88 cluster. Overall this suggests that individual prognostic information may ultimately be best tailored to patients through the combined use of genomic cluster and individual gene data.

This is the first study to analyze the genetic structure of DLBCL using a large unselected population-based register of patients with full clinical follow-up. Our findings substantiate the conclusions of recent studies by confirming the existence of reproducible molecular subtypes of DLBCL defined by their profile of genomic alterations. We show that genetic subtypes can be resolved using a targeted sequencing panel applied to biopsy material acquired in routine clinical practice. We refine the molecular classification further to identify a new very good risk subtype that shares biological features of PMBCL. We also provide greater insight into the prognostic impact of the genomic subtypes and their interaction with other genetic and

clinical factors. Together with previous studies, these findings suggest that the field is ready for a concerted effort to standardize the molecular sub-classification of DLBCL. Stratification by molecular subtype will guide the design and interpretation of future clinical trials in DLBCL.

Acknowledgements

Bloodwise (grant number 15037) funded the majority of this study. Genetic sequencing was funded by 14M Genomics, a start-up company that ceased trading February 2016.
DJH was funded by a clinician scientist fellowship from the MRC and receives core funding from Wellcome and MRC to the Wellcome-MRC Cambridge Stem Cell Institute.
Some of the analysis in this study was performed on the "Viking" high performance computing cluster at the University of York.12

Authorship Contributions

Conception and design: SB, PB, DP, AS, ER, PC, RT, RP, CB, SC.
Provision of study material or patients: SB, PB, DP, AS, ER, SLC, PG, SvH, NW, PC, RP, CB, SC.
Collection and assembly of data: All authors.
Analysis and interpretation: SL, SB, PB, DP, AS, ER, SLC, CR, PC, RT, RP, CB, SC, DH.
Manuscript writing: SL, SB, PB, AS, ER, CB, SC, DH.
Final approval of manuscript: All authors.
Accountable for all aspects of the work: All authors.

Declaration of interests

PAB: consultancy for Karus Therapeutics (Oxford, UK), OncoDNA (Gosselies, Belgium) and Everything Genetic (London, UK).
DJH; Research funding from Gilead International, consultancy for Karus.

The current affiliation for CR is Department of Bioengineering, Stanford University, Palo Alto, 94305, United States.

Correspondence:

Simon Crouch, Epidemiology and Cancer Statistics Group, Department of Health Sciences, University of York, York, YO10 5DD, United Kingdom. simon.crouch@york.ac.uk; tel: +44 (0)1904 321938.

References

1. Smith A, Crouch S, Lax S, et al. Lymphoma incidence, survival and prevalence 2004-2014: sub-type analyses from the UK's Haematological Malignancy Research Network. *Br. J. Cancer*. 2015;112(9):1575–1584.
2. Swerdlow S, Campo E, Harris N, et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. Lyon, France: World Health Organization; 2017.
3. Teras LR, DeSantis CE, Cerhan JR, et al. 2016 US lymphoid malignancy statistics by World Health Organization subtypes. *CA. Cancer J. Clin.* 2016;
4. Coiffier B, Lepage E, Briere J, et al. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N. Engl. J. Med.* 2002;346(4):235–42.
5. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403(6769):503–511.
6. Wright G, Tan B, Rosenwald A, et al. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* 2003;100(17):9991–9996.
7. Sha C, Barrans S, Cucco F, et al. Molecular High-Grade B-Cell Lymphoma: Defining a Poor-Risk Group That Requires Different Approaches to Therapy. *J. Clin. Oncol.* 2018;37(3):202–212.
8. Painter D, Barrans S, Lacy S, et al. Cell-of-origin in diffuse large B-cell lymphoma: findings from the UK's population-based Haematological Malignancy Research Network. *Br. J. Haematol.* 2019;185(4):781–784.
9. Pasqualucci L, Trifonov V, Fabbri G, et al. Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.* 2011;43(9):830–837.
10. Morin RD, Mendez-Lago M, Mungall AJ, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*. 2011;476(7360):298–303.
11. Lohr JG, Stojanov P, Lawrence MS, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci.* 2012;109(10):3879–3884.
12. Monti S, Chapuy B, Takeyama K, et al. Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell*. 2012;22(3):359–372.
13. Morin RD, Mungall K, Pleasance E, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood*. 2013;122(7):1256–1265.
14. de Miranda NFCC, Georgiou K, Chen L, et al. Exome sequencing reveals novel mutation targets in diffuse large B-cell lymphomas derived from Chinese patients. *Blood*. 2014;124(16):2544–2553.
15. Karube K, Enjuanes A, Dlouhy I, et al. Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia*. 2018;32(3):675–684.
16. Schmitz R, Wright GW, Huang DW, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 2018;378(15):1396–1407.
17. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* 2018;24(5):679–690.
18. Reddy A, Zhang J, Davis NS, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell*. 2017;171(2):481-494.e15.

19. Smith A, Howell D, Crouch S, et al. Cohort Profile: The Haematological Malignancy Research Network (HMRN); a UK population-based patient cohort. *Int. J. Epidemiol.* 2018;
20. Smith A, Roman E, Howell D, et al. The Haematological Malignancy Research Network (HMRN): a new information strategy for population based epidemiology and health service research. *Br J Haematol*. 2010;148(5):739–753.
21. Leisch F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *J. Stat. Softw.* 2004;11(1):1–18.
22. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* 2000;22(7):719–725.
23. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach. Learn.* 2003;52(1):91–118.
24. Stata-Corp. Stata Statistical Software: Release 15. Stata-Corp, College Station, TX, USA; 2017.
25. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2019.
26. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer-Verlag; 2000.
27. Mootha VK, Lindgren CM, Eriksson K-F, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 2003;34(3):267–273.
28. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 2005;102(43):15545–15550.
29. Yuan J, Wright G, Rosenwald A, et al. Identification of Primary Mediastinal Large B-cell Lymphoma at Nonmediastinal Sites by Gene Expression Profiling. *Am. J. Surg. Pathol.* 2015;39(10):1322–1330.
30. Mottok A, Hung SS, Chavez EA, et al. Integrative genomic analysis identifies key pathogenic mechanisms in primary mediastinal large B-cell lymphoma. *Blood*. 2019;134(10):802–813.
31. Genetics of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 2018;379(5):493–494.
32. Wang Y, Link BK, Witzig TE, et al. Impact of concurrent indolent lymphoma on the clinical outcome of newly diagnosed diffuse large B-cell lymphoma. *Blood*. 2019;134(16):1289–1297.

Table 1: Patient and tumour characteristics: diffuse large B-cell lymphomas (DLBCL) newly diagnosed within HMRN, September 2004 to August 2012.

		Study cohort	Analysis cohort			
			Total	DLBCL NOS		
				All patients	R-CHOP treated	De-novo R-CHOP treated
All subtypes combined		**2358 (100)**	**928 (100)**	**839 (100)**	**609 (100)**	**579**
DLBCL NOS		2170 (92·0)	839 (90·4)	··	··	··
Primary CNS		74 (3·1)	33 (3·6)	··	··	··
Primary mediastinal		54 (2·3)	20 (2·2)	··	··	··
T-cell/histiocyte-rich		31 (1·3)	21 (2·3)	··	··	··
Plasmablastic		24 (1·0)	14 (1·5)	··	··	··
Intravascular		5 (0·2)	1 (0·1)	··	··	··
Age at diagnosis (years, range)		70·0 (1·6 - 97·8)	68·5 (8·5 - 97·8)	68·8 (8·5 - 97·8)	66·3 (22·8 - 89·0)	66·1 (22·8 - 89·0)
Males n (%)		1231 (52·2)	500 (53·9)	451 (53·8)	326 (53·5)	309 (53·4)
Performance status (ECOG)						
0-1		1423 (60·3)	588 (63·4)	540 (64·4))	448 (73·6)	448 (77·4)
≥2		680 (28·8)	256 (27·6)	219 (26·1))	120 (19·7)	120 (20·7)
Not known[1]		255 (10·8)	84 (9·1)	80 (9·5)	41 (6·7)	11 (1·9)
Stage (Ann Arbor)						
I		337 (14·3)	145 (15·6)	137 (16·3)	111 (18·2)	111 (19·2)
II		383 (16·2)	161 (17·3)	149 (17·8)	139 (22·8)	139 (24·0)
III		313 (13·3)	157 (16·9)	148 (17·6)	124 (20·4)	124 (21·4)
IV		900 (38·2)	321 (34·6)	269 (32·1)	184 (30·2)	184 (31·8)
Not fully staged/not known[1]		425 (18·0)	144 (15·5)	136 (16·2)	51 (8·4)	21 (3·6)
International Prognostic Index						
Low		445 (20·3)	213 (23·0)	192 (22·9)	178 (29·2)	178 (30·8)
Low/Intermediate		365 (16·7)	143 (15·4)	128 (15·3)	110 (18·1)	110 (19·0)
Intermediate/High		394 (18·0)	156 (16·8)	146 (17·4)	116 (19·0)	116 (20·1)
High		462 (19·6)	167 (18·0)	150 (17·9)	95 (15·6)	95 (16·4)
Not known[1]		692 (29·3)	249 (26·8)	223 (26·6)	110 (18·1)	80 (13·8)
Cell of origin (COO)						
Classic[2]	GCB	410 (49·7)	265 (50·5)	252 (52·7)	181 (52·6)	172 (53·6)
	ABC	233 (28·2)	142 (27·0)	141 (29·5)	99 (28·8)	91 (28·3)
	Unclassified	182 (22·1)	118 (22·5)	85 (17·8)	64 (18·6)	58 (18·1)
Refined[2]	GCB	359 (43·5)	234 (44·6)	223 (46·7)	163 (47·4)	155 (48·3)
	ABC	228 (27·6)	140 (26·7)	139 (29·1)	98 (28·5)	90 (28·0)
	MHG	60 (7·3)	34 (6·5)	32 (6·7)	20 (5·8)	19 (5·9)
	Unclassified	178 (21·6)	117 (22·3)	84 (17·6)	63 (18·3)	57 (17·8)
Treated curatively		1929 (81·8)	803 (86·5)	730 (87·0)	609 (100·0)	579 (100·0)
R-CHOP treated		1536 (65·1)	648 (69·8)	609 (72·6)		
Median Survival (yrs, 95%CI)		4·1 (3·4 - 4·9)	6·2 (5·5 - 7·2)	6·4 (5·6 -7·5)	10·2 (8·3-11·9)	10·6 (8·4 - 12·5)
R-CHOP treated		9·3 (8·3-10·6)	10·4 (8·4 - 12·5)	10·2 (8·3-11·9)		
5-year overall survival (95% CI)		47·7 (45·7 - 49·7)	54·6 (51·5 - 57·9)	55·4 (52·1 - 58·8)	67·3 (63·6 - 71·1)	68·0 (64·3 - 71·9)
R-CHOP treated		62·8 (60·4 - 65·2)	67·5 (64·0 - 71·2)	67·3 (63·6 - 71·1)		

[1]Includes transformed follicular lymphoma (study cohort n=169, analysis cohort n=55) where baseline information is taken at the time first diagnosis, but not at transformation; [2] Percent of those with material available.

Table 2: Characteristics by AIC cluster: diffuse large B-cell lymphomas (DLBCL) diagnosed within HMRN, September 2004 to August 2012

		Genetic subtype n (%)					
		NEC	**SOCS1/SGK1**	**TET2/SGK1**	**MYD88**	**BCL2**	**NOTCH2**
All subtypes combined		**248 (100)**	**111 (100)**	**98 (100)**	**152 (100)**	**176 (100)**	**143 (100)**
Diffuse large B-cell lymphoma, NOS		211 (85·1)	98 (88·3)	91 (92·9)	130 (85·5)	173 (98·3)	136 (95·1)
Primary CNS		9 (3·6)	0 (0·0)	1 (1·0)	22 (14·5)	0 (0·0)	1 (0·7)
Primary mediastinal		2 (0·8)	12 (10·8)	0 (0·0)	0 (0·0)	3 (1·7)	3 (2·1)
T-cell/histiocyte-rich		16 (6·5)	1 (0·9)	2 (2·0)	0 (0·0)	0 (0·0)	2 (1·4)
Plasmablastic		9 (3·6)	0 (0·0)	4 (4·1)	0 (0·0)	0 (0·0)	1 (0·7)
Intravascular		1 (0·4)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)
De novo/transformed							
De novo		237 (95·6)	110 (99·1)	93 (94·9)	150 (98·7)	155 (88·1)	135 (94·4)
Transformed[1]		11 (4·4)	1 (0·9)	5 (5·1)	2 (1·3)	21 (11·9)	8 (5·6)
DLBCL with Concurrent FL[2]		9 (3·6)	10 (9·0)	2 (2·0)	2 (1·3)	47 (26·7)	15 (10·5)
Testicular Involvement[3]		4 (2·8)	1 (1·8)	0 (0·0)	21 (26·9)	0 (0·0)	2 (3·1)
Breast Involvement[3]		2 (2·5)	0 (0·0)	1 (2·0)	9 (12·9)	1 (1·1)	0 (0·0)
Cell of origin[4]							
Classic	GCB	43 (30·3)	56 (86·2)	33 (66·0)	5 (8·2)	92 (82·1)	36 (37·9)
	ABC	38 (26·8)	4 (6·2)	5 (10·0)	50 (82·0)	3 (2·7)	42 (44·2)
	Unclassified	61 (43·0)	5 (7·7)	12 (24·0)	6 (9·8)	17 (15·2)	17 (17·9)
Refined	GCB	36 (25·4)	51 (78·5)	31 (62·0)	5 (8·2)	78 (69·6)	33 (34·7)
	ABC	38 (26·8)	4 (6·2)	5 (10·0)	49 (80·3)	3 (2·7)	41 (43·2)
	MHG	7 (4·9)	5 (7·7)	2 (4·0)	1 (1·6)	15 (13·4)	4 (4·2)
	Unclassified	61 (43·0)	5 (7·7)	12 (24·0)	6 (9·8)	16 (14·3)	17 (17·9)
Age at diagnosis (years, range)		67·0 (12·1 - 95·2)	66·8 (13·5 - 97·8)	72·7 (22·8 - 92·0)	70·2 (35·1 - 97·4)	66·4 (24·1 - 88·8)	70·2 (8·5 - 95·8)
R-CHOP treated (%)		172 (69·4)	85 (76·6)	70 (71·4)	80 (52·6)	141 (80·1)	100 (69·9)
5-year overall survival (%, 95% CI)		53·6 (47·7 - 60·2)	64·9 (56·6 - 74·4)	60·1 (51·1 - 70·6)	42·1 (34·9 - 50·7)	62·5 (55·7 - 70·1)	48·1 (40·5 - 57·0)
R-CHOP treated		65·6 (58·9 - 73·1)	80·0 (71·9 - 89·0)	69·8 (59·8 - 81·5)	63·8 (54·0 - 75·2)	69·5 (62·3 - 77·5)	58·8 (49·8 - 69·3)

[1] Prior diagnosis: 22 follicular lymphoma, 14 marginal zone lymphoma, 7 chronic lymphocytic leukaemia, 2 hairy cell leukaemias, 3 Hodgkin's lymphomas; [2] Includes FL discovered at time of DLBCL diagnosis, excluding those with prior FL diagnosis; [3] sex-specific proportions; [4] Percent of those with material available.

Table 3: Comparison of main clusters

This study	Chapuy et al	Schmitz et al	Notes
MYD88 MYD88, PIM1, CD79B, ETV6, CDKN2A	**C5** CD79B, MYD88, ETV6, PIM1, TBL1XR1	**MCD** MYD88, CD79B	Strongly associated with ABC-type DLBCL. The most robust group, occurring in all reports. Contains the majority of cases with primary CNS lymphoma and primary testicular lymphoma. Associated with a poor prognosis.
BCL2 EZH2, BCL2, CREBBP, TNFRSF14, KMT2D	**C3** BCL2, CREBBP2, EZH2, KMT2D, TNFRSF14	**EZB** BCL2 translocation, EZH2	Strongly associated with GCB-type DLBCL. Mutational profile is shared with Follicular lymphoma. Contains the majority of cases of transformed FL and cases with a concurrent diagnosis of FL. Generally favorable prognosis, although enriched for cases of Double Hit lymphoma and MHG.
SOCS1/SGK1 SOCS1, CD83, SGK1, NFKBIA, HIST1H1E	**C4** SGK1, HIST1H1E, NFKBIE, BRAF, CD83		Predominantly GCB-type DLBCL. Shares genetic and gene expression features of PMBL. Associated with the most favorable prognosis.
TET2/SGK1 TET2, BRAF, SGK1, KLHL6, ID3			A less strongly identifiable subtype emerging from SGK1 when applying the AIC criterion (see supplement). Has very strong similarity to SOCS1/SGK1, but differentiated by the addition of TET2 and BRAF and the lack of SOCS1 and CD83. Associated with a favorable prognosis.
NOTCH2 NOTCH2, BCL10, TNFAIP3, CCND3, SPEN	**C1** BCL6 translocation, BCL10, TNFAIP3, UBE2A, CD70	**BN2** BCL6 translocation, NOTCH2	Not associated with any COO. Shares mutational similarity to marginal zone lymphoma but not enriched for cases of transformed MZL. Less strongly defined than other subgroups (see supplement).
NEC NOTCH1, REL amplification, TP53		**Other**	A default category, containing cases that could not be classified elsewhere. Contains cases with no detected mutation. Likely to also contain cases belonging to both NOTCH1 and TP53/CNA subgroups. Even though 3 abnormalities are significantly enriched in this group, their q-values are far less extreme than those of characteristic mutations from the other subtypes.
	C2 TP53, frequent deletions		Characterized by TP53 mutation and widespread copy number changes. Due to limited CNA in our study these cases were predominantly allocated to the NEC group.
	C0 No detected abnormalities		Cases with no detectable mutation were allocated to the NEC group.
		N1 **NOTCH1**	Characterized by NOTCH1 mutation, this was significantly elevated in our NEC group, but only mutated in 2·5% of samples. Associated with poor outcome.

Figure 1: Relationships between gene expression profiling and the six genetic clusters that were identified using the AIC criterion, for the n=519 subjects with available gene expression profiling data. Panel A shows a heatmap of a selection of signatures and Panel B highlights a small number of these that demonstrate the strongest trends.

Figure 2: Heatmap of characteristic mutations from each of the six genetic clusters that were identified using the AIC criterion in the analysis cohort (n=928). Distinct clusters are identified by colour. Along the top of the figure the colour strip shows the corresponding cell-of-origin classification for each patient. The panel on the right-hand side shows the enrichment for mutations within each cluster, with a logarithmic q-value scale. 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 3: **Survival in clusters identified by the AIC 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. E: crude hazard ratios of cluster membership for the n=579 de-novo DLBCL NOS R-CHOP treated patients, and F: adjusted hazard ratios of the subset of the 579 group that have IPI data available (n=499).

Figure 4: **Overall survival of the de-novo DLBCL NOS R-CHOP treated patients (n=579), stratified by a selection of genetic features**. A: NOTCH1 mutation. B: MYC rearrangement status in the subset of patients belonging to the BCL2 AIC subtype. C: TP53 mutation or homozygous deletion. D: The effect of TP53 mutation or homozygous deletion in each genetic subtype.

Figure 1

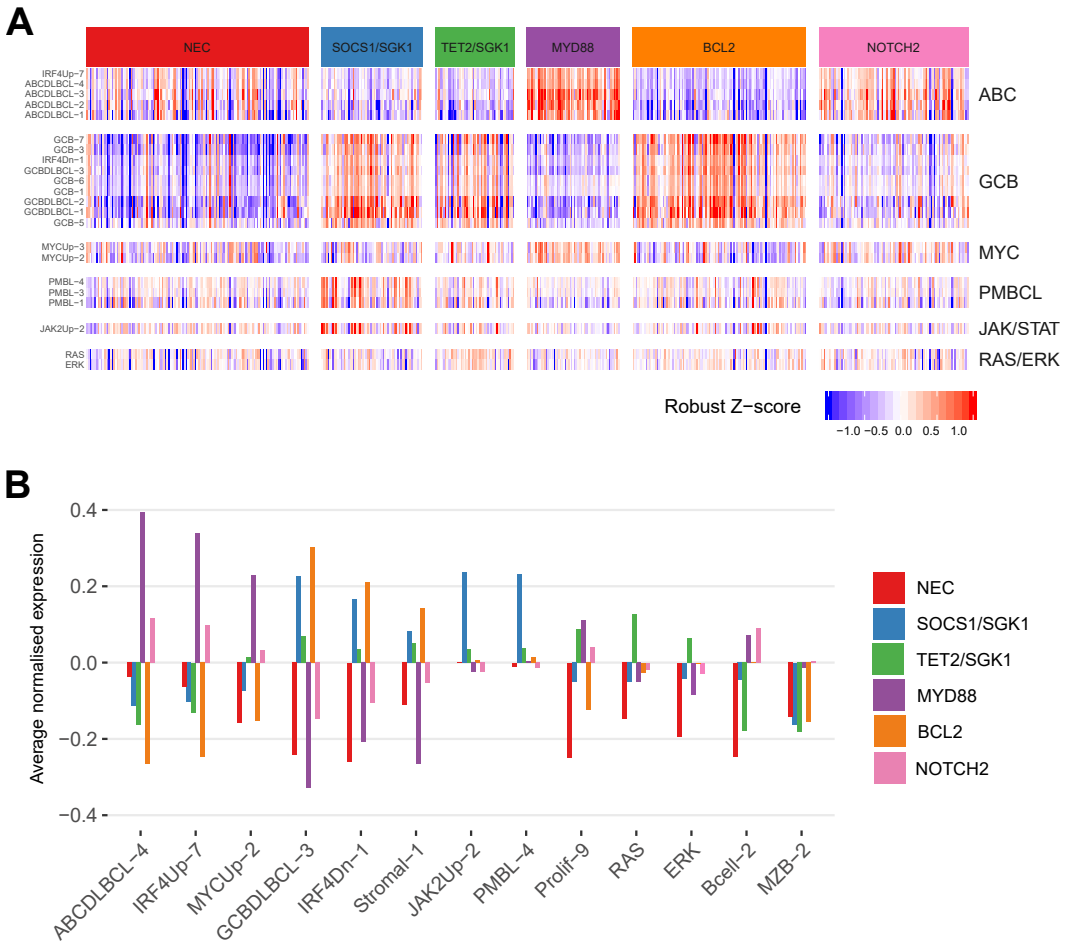

A
NEC
SOCS1/SGK1
TET2/SGK1
MYD88
BCL2
NOTCH2
IRF4Up−7
ABCDLBCL−4
ABCDLBCL−3
ABCDLBCL−2
ABCDLBCL−1
ABC
GCB−7
GCB−3
IRF4Dn−1
GCBDLBCL−3
GCB−6
GCB−1
GCBDLBCL−2
GCBDLBCL−1
GCB−5
GCB
MYCUp−3
MYCUp−2
MYC
PMBL−4
PMBL−3
PMBL−1
PMBCL
JAK2Up−2
JAK/STAT
RAS
ERK
RAS/ERK
Robust Z−score
−1.0 −0.5 0.0 0.5 1.0
B
Average normalised expression
0.4
0.2
0.0
−0.2
−0.4
ABCDLBCL−4
IRF4Up−7
MYCUp−2
GCBDLBCL−3
IRF4Dn−1
Stromal−1
JAK2Up−2
PMBL−4
Prolif−9
RAS
ERK
Bcell−2
MZB−2
NEC
SOCS1/SGK1
TET2/SGK1
MYD88
BCL2
NOTCH2

Figure 2

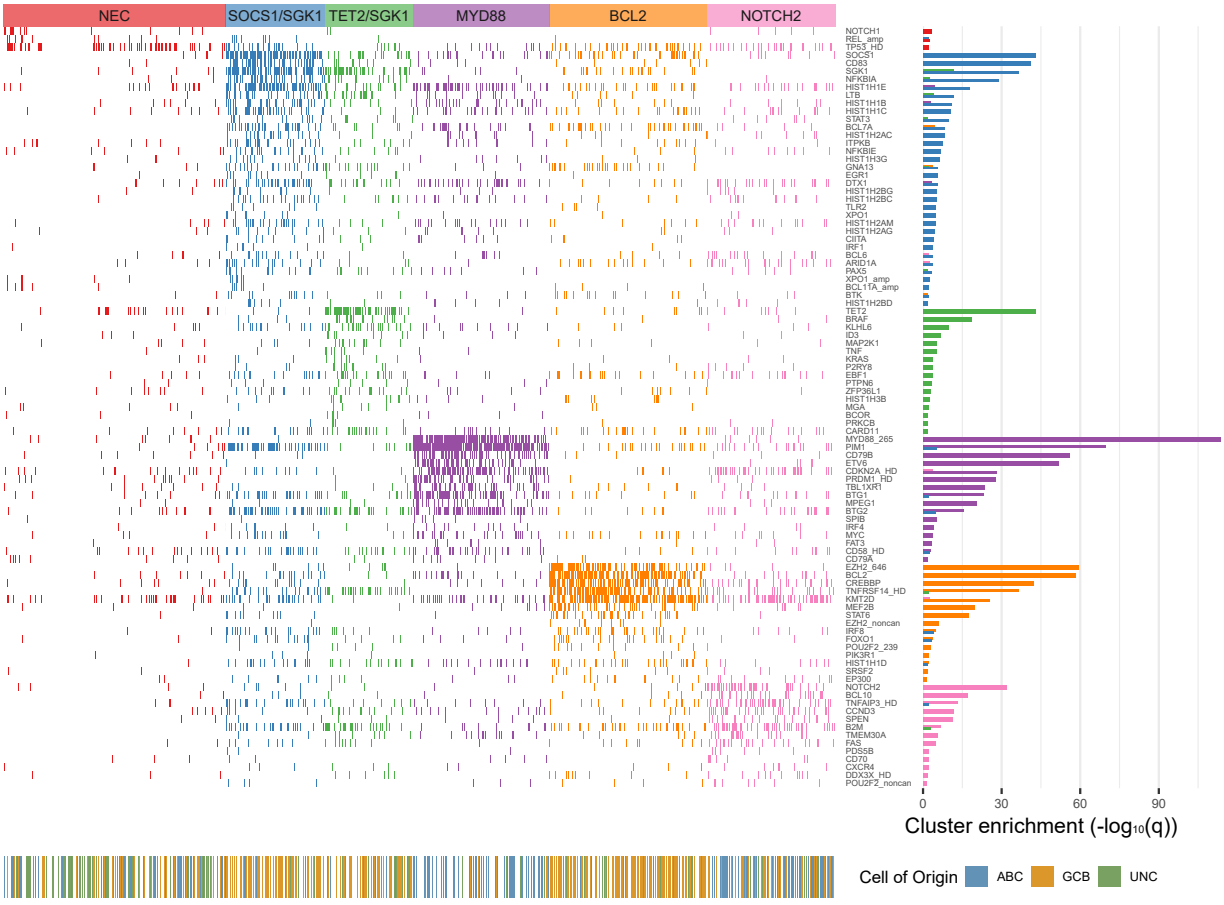

NEC
SOCS1/SGK1
TET2/SGK1
MYD88
BCL2
NOTCH2
NOTCH1
REL_amp
TP53_HD
SOCS1
CD83
SGK1
NFKBIA
HIST1H1E
LTB
HIST1H1B
HIST1H1C
STAT3
BCL7A
HIST1H2AC
ITPKB
NFKBIE
HIST1H3G
GNA13
EGR1
DTX1
HIST1H2BG
HIST1H2BC
TLR2
XPO1
HIST1H2AM
HIST1H2AG
CIITA
IRF1
BCL6
ARID1A
PAX5
XPO1_amp
BCL11A_amp
BTK
HIST1H2BD
TET2
BRAF
KLHL6
ID3
MAP2K1
TNF
KRAS
P2RY8
EBF1
PTPN6
ZFP36L1
HIST1H3B
MGA
BCOR
PRKCB
CARD11
MYD88_265
PIM1
CD79B
ETV6
CDKN2A_HD
PRDM1_RD
TBL1XR1
BTG1
MPEG1
B2M
SPIB
IRF4
MYC
FAT3
CD58_HD
CD70A
EZH2_646
BCL2
CREBBP
TNFRSF14_HD
KMT2D
MEF2B
STAT6
EZH2_noncan
IRF8
FOXO1
POU2F2_239
PIK3R1
HIST1H1D
SRSF2
EP300
NOTCH2
BCL10
TNFAIP3_HD
CCND3
SPEN
B2M
TMEM30A
FAS
PDSS2
CD70
CXCR4
DDX3X_HD
POU2F2_noncan
0
30
60
90
Cluster enrichment (-log10(q))
Cell of Origin
ABC
GCB
UNC

Figure 3

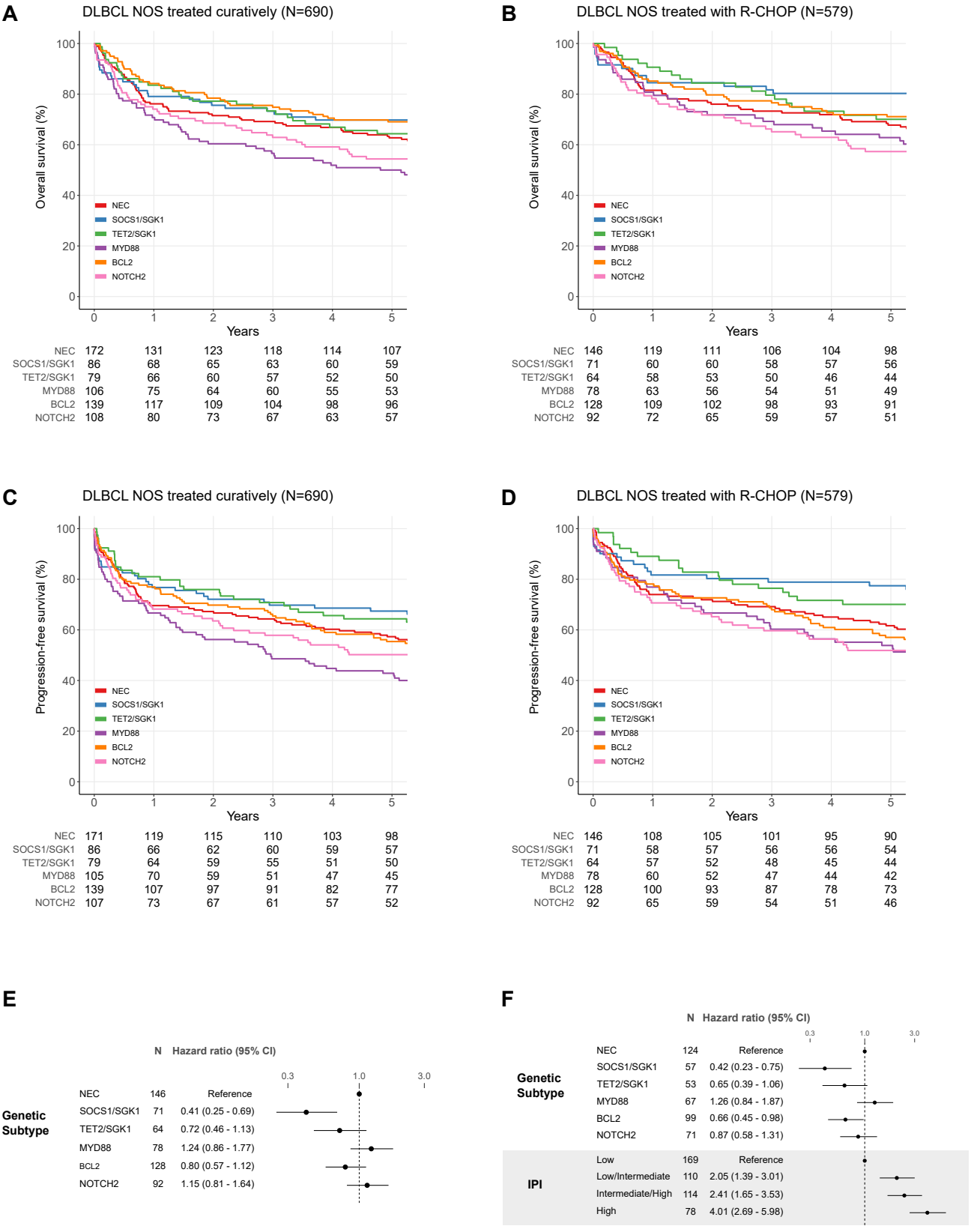
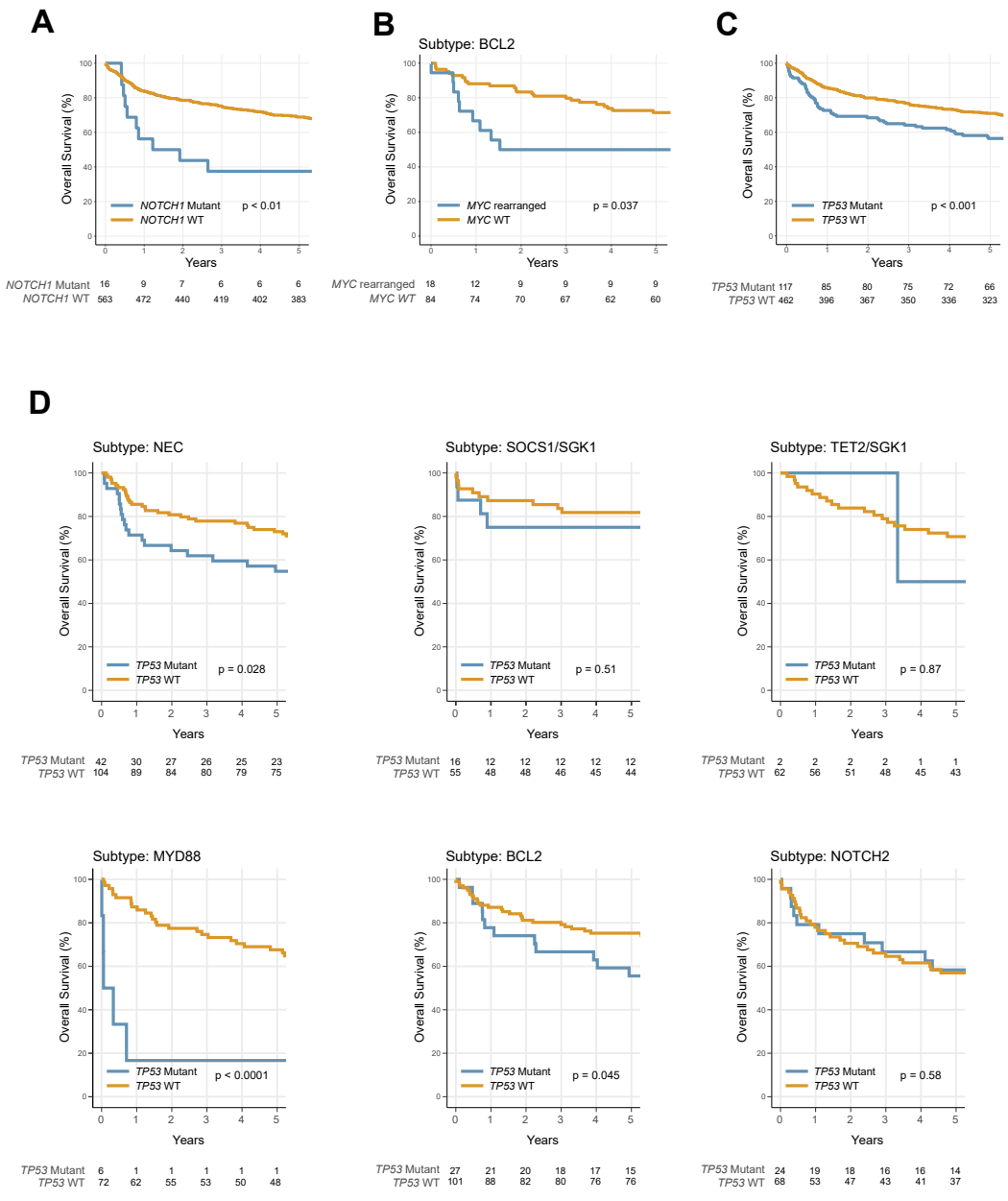

A
DLBCL NOS treated curatively (N=690)
Overall survival (%)
Years
NEC
SOCS1/SGK1
TET2/SGK1
MYD88
BCL2
NOTCH2
NEC 172 131 123 118 114 107
SOCS1/SGK1 86 68 65 63 60 59
TET2/SGK1 79 66 60 57 52 50
MYD88 106 75 64 60 55 53
BCL2 139 117 109 104 98 96
NOTCH2 108 80 73 67 63 57
B
DLBCL NOS treated with R-CHOP (N=579)
Overall survival (%)
Years
NEC 146 119 111 106 104 98
SOCS1/SGK1 71 60 60 58 57 56
TET2/SGK1 64 58 53 50 46 44
MYD88 78 63 56 54 51 49
BCL2 128 109 102 98 93 91
NOTCH2 92 72 65 59 57 51
C
DLBCL NOS treated curatively (N=690)
Progression-free survival (%)
Years
NEC 171 119 115 110 103 98
SOCS1/SGK1 86 66 62 60 59 57
TET2/SGK1 79 64 59 55 51 50
MYD88 105 70 59 51 47 45
BCL2 139 107 97 91 82 77
NOTCH2 107 73 67 61 57 52
D
DLBCL NOS treated with R-CHOP (N=579)
Progression-free survival (%)
Years
NEC 146 108 105 101 95 90
SOCS1/SGK1 71 58 57 56 56 54
TET2/SGK1 64 57 52 48 45 44
MYD88 78 60 52 47 44 42
BCL2 128 100 93 87 78 73
NOTCH2 92 65 59 54 51 46
E
N Hazard ratio (95% CI)
Genetic Subtype
NEC 146 Reference
SOCS1/SGK1 71 0.41 (0.25 - 0.69)
TET2/SGK1 64 0.72 (0.46 - 1.13)
MYD88 78 1.24 (0.86 - 1.77)
BCL2 128 0.80 (0.57 - 1.12)
NOTCH2 92 1.15 (0.81 - 1.64)
F
N Hazard ratio (95% CI)
Genetic Subtype
NEC 124 Reference
SOCS1/SGK1 57 0.42 (0.23 - 0.75)
TET2/SGK1 53 0.65 (0.39 - 1.06)
MYD88 67 1.26 (0.84 - 1.87)
BCL2 99 0.66 (0.45 - 0.98)
NOTCH2 71 0.87 (0.58 - 1.31)
IPI
Low 169 Reference
Low/Intermediate 110 2.05 (1.39 - 3.01)
Intermediate/High 114 2.41 (1.65 - 3.53)
High 78 4.01 (2.69 - 5.98)

Figure 4


A
Overall Survival (%)
NOTCH1 Mutant
NOTCH1 WT
p < 0.01
Years
NOTCH1 Mutant 16 9 7 6 6 6
NOTCH1 WT 563 472 440 419 402 383
B
Subtype: BCL2
Overall Survival (%)
MYC rearranged
MYC WT
p = 0.037
Years
MYC rearranged 18 12 9 9 9 9
MYC WT 84 74 70 67 62 60
C
Overall Survival (%)
TP53 Mutant
TP53 WT
p < 0.001
Years
TP53 Mutant 117 85 80 75 72 66
TP53 WT 462 396 367 350 336 323
D
Subtype: NEC
Overall Survival (%)
TP53 Mutant
TP53 WT
p = 0.028
Years
TP53 Mutant 42 30 27 26 25 23
TP53 WT 104 89 84 80 79 75
Subtype: SOCS1/SGK1
Overall Survival (%)
TP53 Mutant
TP53 WT
p = 0.51
Years
TP53 Mutant 16 12 12 12 12 12
TP53 WT 55 48 48 46 45 44
Subtype: TET2/SGK1
Overall Survival (%)
TP53 Mutant
TP53 WT
p = 0.87
Years
TP53 Mutant 2 2 2 2 1 1
TP53 WT 62 56 51 48 45 43
Subtype: MYD88
Overall Survival (%)
TP53 Mutant
TP53 WT
p < 0.0001
Years
TP53 Mutant 6 1 1 1 1 1
TP53 WT 72 62 55 53 50 48
Subtype: BCL2
Overall Survival (%)
TP53 Mutant
TP53 WT
p = 0.045
Years
TP53 Mutant 27 21 20 18 17 15
TP53 WT 101 88 82 80 76 76
Subtype: NOTCH2
Overall Survival (%)
TP53 Mutant
TP53 WT
p = 0.58
Years
TP53 Mutant 24 19 18 16 16 14
TP53 WT 68 53 47 43 41 37