Supplementary Material

This supplemental material has been provided by the authors to give readers additional information about their work.

Supplemental Appendix

I. Materials and Methods
II. Supplemental Figures 1-3

I. Materials and Methods

Patients

We studied 475 patients with de novo adult DLBCL cases that had been diagnosed between January 2002 and October 2009, as part of the International DLBCL Rituximab-CHOP Consortium Program Study. Cases were selected on the basis of the availability of GEP analysis results and of sufficient clinical data. All cases were reviewed by a group of hematopathologists (all primary participation center pathologists, SMM, MAP, MBM, AT, and KHY), and the diagnoses were confirmed on the basis of WHO classification criteria. Patients with transformation from low-grade lymphoma, primary mediastinal large B-cell lymphoma, primary DLBCL of the central nervous system, or primary cutaneous DLBCL were excluded from the analysis due to differences in the biological features and clinical behavior of these tumor types. Patients positive for human immunodeficiency virus were also excluded.

Treatment consisted of R-CHOP (432, 91%) or R-CHOP-like regimens (43, 9%; epirubicin or mitoxantrone based combination chemotherapy). Second-line chemotherapy was based on most common cytarabine-based regimens (DHAP, ICE, or ESHAP with or without rituximab), followed by

hematopoietic stem cell transplant in 25% of patients with recurrent disease. Approximately 71% of the cases occurred in patients from North America, 6% from Asia, and the remainder were from Western Europe (Italy, Spain, Switzerland, Denmark, and The Netherlands).

The International DLBCL Rituximab-CHOP Consortium Program has the principal investigation center at The University of Texas MD Anderson Cancer Center, Houston, TX, USA, and includes following participation centers for the collaboration study: San Bortolo Hospital, Vicenza, Italy; Odense University Hospital, Odense, Denmark; University of Louisville School of Medicine, Comprehensive Cancer Center, Louisville, KY, USA; University Hospital, Basel, Switzerland; Hospital Universitario Marques de Valdecilla, Santander, Spain; Aalborg Hospital, Aarhus University Hospital, Aalborg, Denmark; Brigham and Women Hospital, Harvard Medical School, Boston, MA, USA; Weill Medical College of Cornell University, New York, NY, USA; The Methodist Hospital, Houston, TX, USA; Columbia University School of Medicine, New York, NY, USA; Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; University of California San Diego School of Medicine, San Diego, CA, USA; University of North Carolina School of Medicine, Chapel Hill, NC, USA; Cleveland Clinic, Cleveland, OH, USA; University of Maryland School of Medicine, Baltimore, MD, USA; Gundersen Lutheran Health System, La Crosse, WI, USA; University of Hong Kong Li Ka Shing Faculty of Medicine, Hong Kong, China; Southwest Washington Medical Center, Vancouver, WA, USA; University of Indiana School of Medicine, Indianapolis, IN, USA; Zhejiang University School of Medicine, Second University Hospital, Hangzhou, China; Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands; City of Hope National Medical Center, Los Angeles, CA, USA; University of California San Francisco School of Medicine, San Francisco, CA, USA; San Raffaele H. Scientific Institute, Milan, Italy; St. Vincent Hospital, Worcester, MA; Ulsan University College of Medicine, Asan Medical Center, Seoul, Korea; Ohio State University, Comprehensive Cancer Center,

Columbus, OH; Momerial Sloan-Kettering Cancer Center, New York, NY, and Switzerland Oncology Institute, Bellinzona, Switzerland. The current study was reviewed and approved as being of minimal to no risk or as exempt by each of the participating center Institutional Review Boards, and the overall collaborative study was approved by the Institutional Review Board at The University of Texas MD Anderson Cancer Center in Houston, Texas. A Material Transfer Agreement was established and approved by each of the participating centers joining this collaborative project for the International DLBCL Rituximab-CHOP Consortium Program.

Tissue microarray (TMA) immunohistochemistry

Immunohistochemical staining was performed on samples from all 475 cases. Tissue biopsy specimens were fixed in 4% buffered formalin, routinely processed, and embedded in paraffin, and 4-μm sections were cut and stained with hematoxylin-eosin. The hematoxylin-eosin stained slides from each tumor block were reviewed, and representative areas with the highest percentage of tumor cells were selected for TMA construction. TMA blocks were constructed with a tissue microarrayer (Beecher Instrument, Silver Spring, MD) at one laboratory, similar to all immunohistochemical stainings. For each DLBCL primary tissue block, 1-2 tissue cores that were 1 or 1.5 mm in diameter were arrayed into a recipient TMA block, and 3 TMA blocks were prepared for each of the cases. Immunohistochemical analysis was performed on 4-μm TMA sections using a streptavidin-biotin complex technique, and antibodies reactive against the following antigens were utilized: CD3, CD5, CD10, CD20, CD30, CD79a, CD138, ALK-1, BCL2, BCL6, FOXP1, GCET1, GCET2, and MUM1. Immunoreactivity was determined without any knowledge of patient survival or other clinical data and blinded to the GEP data. Antigen expression was evaluated by assessing the percentage of immunoreactive tumor cells in 5% increments. Internal positive controls for each core were required for interpretation. The samples were

analyzed independently by a group of six hematopathologists/pathologists in addition to each of the contributing center hematopathologists), and disagreements were resolved by joint review on a multiheaded microscope. To test the reliability of the immunohistochemistry, each of 4 hematopathologists counted 300-400 cells in 4-5 fields on average in the TMA cores on the same cases. Average 300-400 cells in 4-5 fields are counted in the TMA cores. We observed inter and intra variations at 2-6% ranges, however, no significant difference was observed in majority of the cases among different hematopathologists. In cases that showed greater than 10% variation, the debate was solved by a group discussion. Such event occurred in less than 5% of the cases. The source of antibodies used in this project was Ventana (Tucson, AR) for CD3 (clone PS1 at 1:200 dilution), Novocastra (Leica Microsystems Inc., Buffalo Grobe, IL) for CD10 (clone 56C6, at 1:100) and BCL6 (clone LN22 at 1:100), DAKO (Carpinteria, CA) for CD20 (clone L26 at 1:200) and MUM1 (clone MUMp1 at 1:20), and Abcam (Cambridge, MA) for GCET1 (clone RAM at 1:4) and FOXP1 (clone JC12 at 1:16,000).

Gene expression profiling analysis

RNA was extracted from 475 formalin-fixed, paraffin-embedded tissue (FFPET) samples using HighPure Paraffin RNA Extraction Kit (Roche Applied Science). Fifty ng of FFPET total RNA was transcribed into cDNA and linearly amplified using the WT-Ovation™ FFPE System (Nugen) SPIA method[21]. The cDNA was fragmented and biotin labeled using FL-Ovation™ cDNA Biotin Module V2 (Nugen) in all the cases. For GeneChip hybridization, 5 μg of WT-Ovation amplified cDNA was applied to HG-U133 Plus 2.0 GeneChips (Affymetrix) and hybridized overnight per the manufacturer's recommendations. GeneChips were washed, stained, and scanned using the Fluidic Station 450 and GeneChip Scanner 3000 (Affymetrix) according to the manufacturer's recommendations. For data analysis and classification, the microarray DQN (trimmed mean of differences of perfect match and

mismatch intensities with Quantile Normalization[22]) signals were generated and normalized to the quantiles of beta distribution with parameters p=1.2 and q=3. A support vector machine (SVM) was used for classification. A Bayesian model[23] was also utilized to determine the class probability. The classification model was built on the 47 paired FFPET-fresh frozen sample dataset previously generated with confidence rate of 90.2-100% in fresh frozen tissue and 91.3-100% in FFPE tissue.[24] The same methodology developed during this pilot study has been validated and demonstrated to be applicable by using the LLMPP dataset in the Gene Expression Omnibus (GEO) database GSE#10846 that have 181 CHOP-treated and 233 R-CHOP-treated DLBCL patients with FF samples.[25] We obtained 80% concordance rate of classification for 3 classes (GCB, ABC and unclassified), while the concordance is over 97% for 2 classes (GCB and ABC) excluding the unclassified. During these series analysis and cross-validation studies, we do not see case misclassification as a concern with the method established for FFPET application. This model was then applied to the 475 DLBCL cases on this study.

Receiver operating characteristic (ROC) curve analysis to assess discriminatory accuracy of each marker

The ROC curves allowed us to visualize the specificity and sensitivity of each marker (CD10, GCET1, FOXP1, MUM1, and BCL6) in assigning cases to GCB or ABC classification prior to further categorization.[26] The performance of each marker could be quantified by the area under the ROC (AUROC) curve (Supplemental Figure 1). Based on these parameters, CD10 was the best marker for GCB (AUROC 0.800; 95% CI, 0.750–0.849; $P < 0.00001$), followed by GCET1 (AUROC 0.683, 95% CI, 0.625–0.741; $P < 0.0001$) and BCL6 (AUROC 0.658; 95% CI, 0.600–0.717; $P < 0.001$). In contrast, MUM1 (AUROC 0.799; 95% CI, 0.753–0.846; $P < 0.00001$) was more specific but less sensitive than FOXP1 (AUROC 0.643; 95% CI, 0.689–0.796; $P < 0.0001$) in recognizing ABC-DLBCL. To test the

reliability of the cut-off scores, the 475 cases included in this study were randomly divided into two (first and second) data sets. Cut-off scores for all the immunohistochemical markers were determined on the first (test) set and applied to the second (validation) set and vice versa to assess concordance between the groups. The cut off scores obtained on the first and on the second set were 27% and 22%, respectively, for CD10 (35% on the whole collective), 35% and 28% for BCL6 (33% on the whole collective), 37% and 42% for GCET (45% on the whole collective), 65% and 65% for FOXP1 (75% on the whole collective) and 55% and 62% for MUM1 (58% on the whole collective). All cases were classified separately as GCB or non-GCB based on the cut-off scores form both datasets and the proposed three marker algorithm. Except for 8 cases (1.7%), which were classified as GCB according to the cut-off scores from set 2 but not from set 1 due to the lower cut-off for BCL6, all other 467 cases were completely matched between both groups (Kappa=0.978), demonstrating the validity and reliability of our model. Of these 8 cases, 6 (all 6 being CD10-negative) were GCB and 2 ABC according to GEP, indicating that the lower (30%) cut-off score for BCL6 is more sensitive and useful to identify those, especially CD10-negative GCB cases.

Rationale for the structure of the algorithm

In designing the algorithm, we emphasized the importance of CD10 expression (step 1), which is usually part of the initial diagnostic staining panel for hematopathologists, and its staining has shown the best concordance in different studies between different laboratories.[27] We then analyzed *GCET1*, *FOXP1*, and *MUM1*[19,28] expression in this order (step 2-4), following our rationale that will be discussed below. Finally, we left to BCL6 a minor role in recognizing patients with GCB-DLBCL (step 5) due to the variability and reliability of its staining.[26] The five steps of the global algorithm are shown in Figure 1.

Cutoff establishment

We avoided cutoff values based on mean and median expression, since our protein marker expression had a non-Gaussian distribution (Table 1). Instead, by calculating the Youden index[29] from our ROC curves, we identified the point on the curve corresponding to the maximum sensitivity and specificity for each marker to classify a DLBCL as either of GCB or ABC type according to GEP analysis. The Youden index pointed to optimal cutoff scores of 35% for CD10, 33% for BCL6, 45% for GCET, 75% for FOXP1, and 58% for MUM1. For CD10 and BCL6, the cutoffs were very close to 30%, which is the accepted cutoff for these two molecules.[9] In order to avoid too many different cutoffs in the final algorithm, we compared the optimal cut-off of GCET1 and FOXP1 to 60% and found no change occurred in their sensitivity and specificity. Therefore, we modified the cutoff scores for both GCET1 and FOXP1 to 60%, thus maintaining the optimal cutoff for MUM1.

Refining the global algorithm

The initial algorithm with the established cutoffs exhibited a straightforward concordance with GEP analysis (Figure 1). This concordance could be further improved by removing unnecessary passages or redundant decisional points. We removed all the subsequent steps for $CD10^+$ patients and we eliminated step 4 (MUM1), obtaining a four-marker algorithm, which is shown in Figure 2A. Furthermore, after removing step 2 (GCET1) for $CD10^-$ patients, we obtained a three-marker algorithm, shown in Figure 2B. By simplifying the algorithm, we increased the number of concordant patients.

Validation set

To test the efficacy of the new algorithm in predicting survival in an independent series of cases, we applied the algorithm to a second group of 574 archival DLBCL cases studied using TMAs similarly to the first cohort but for which no GEP analysis was available. Of these, 237 patients had been treated with R-CHOP and 337 with CHOP without rituximab. The same selection criteria as those for the first cohort were applied to these patients. Clinical characteristics at presentation for the validation set were similar to the test set in terms of median gender (female in 45%, p = 0.37), LDH (elevated in 34%, p = 0.66), AAS (III-IV in 49%, p = 0.28), presence of B-symptoms (32%, p = 0.77), or IPI (0-2 in 64%, p = 0.69), except for age. Patients of the validation set were significantly younger than patients of the test set (median age 58 years, p=0.007).

Fluorescence in situ hybridization for *C-MYC* gene rearrangement

Fluorescence *in situ* hybridization (FISH) was performed on paraffin-embedded tissue sections with a locus-specific identifier *IGH/MYC/CEP 8* tri-color, dual fusion probes (DFP, 05J75-001 from Vysis, Downers Grove, Illinois, USA) and, due to shortcommings of the former in identifying alternative (*non-IGH*) *C-MYC* rearrangement partners, a locus-specific identifier *C-MYC* dual-color, break-apart probe (BP, 05J91-001 from Vysis,). FISH signals were scored with a Zeiss fluorescence microscope. Cases on the TMA were considered for evaluation if at least 200 tumor cell nuclei per core displayed positive signals. Abnormal FISH signals were recorded as percentage of cells showing an abnormality.

Response definitions and statistical analysis

Response assessment was standardized among different Institutions following the criteria based on CT-scan and bone marrow biopsy.[30] Progression-free survival (PFS) was measured from the time of

diagnosis to the time of progression or death from any cause. Late deaths not related to the underlying lymphoma or its treatment were not considered treatment failures.[30] Overall survival (OS) was measured from the time of diagnosis to last follow-up or to death from any cause. The actuarial probability of PFS and OS was determined using the Kaplan–Meier method,[31] and differences were compared using the log-rank test. A Cox proportional-hazards model was used for multivariate analysis.[32] All variables with $P < 0.05$ were considered to be statistically significant. The comparison of clinical and laboratory features at presentation was carried out with the χ^2 test or the Spearman rank correlation. All statistical calculations, except for ROC which was performed with SPSS 18.0 (SPSS Inc., Chicago, IL), were performed using StatView (Abacus Concepts, Berkeley, CA).

II. Supplemental Figures 1-3

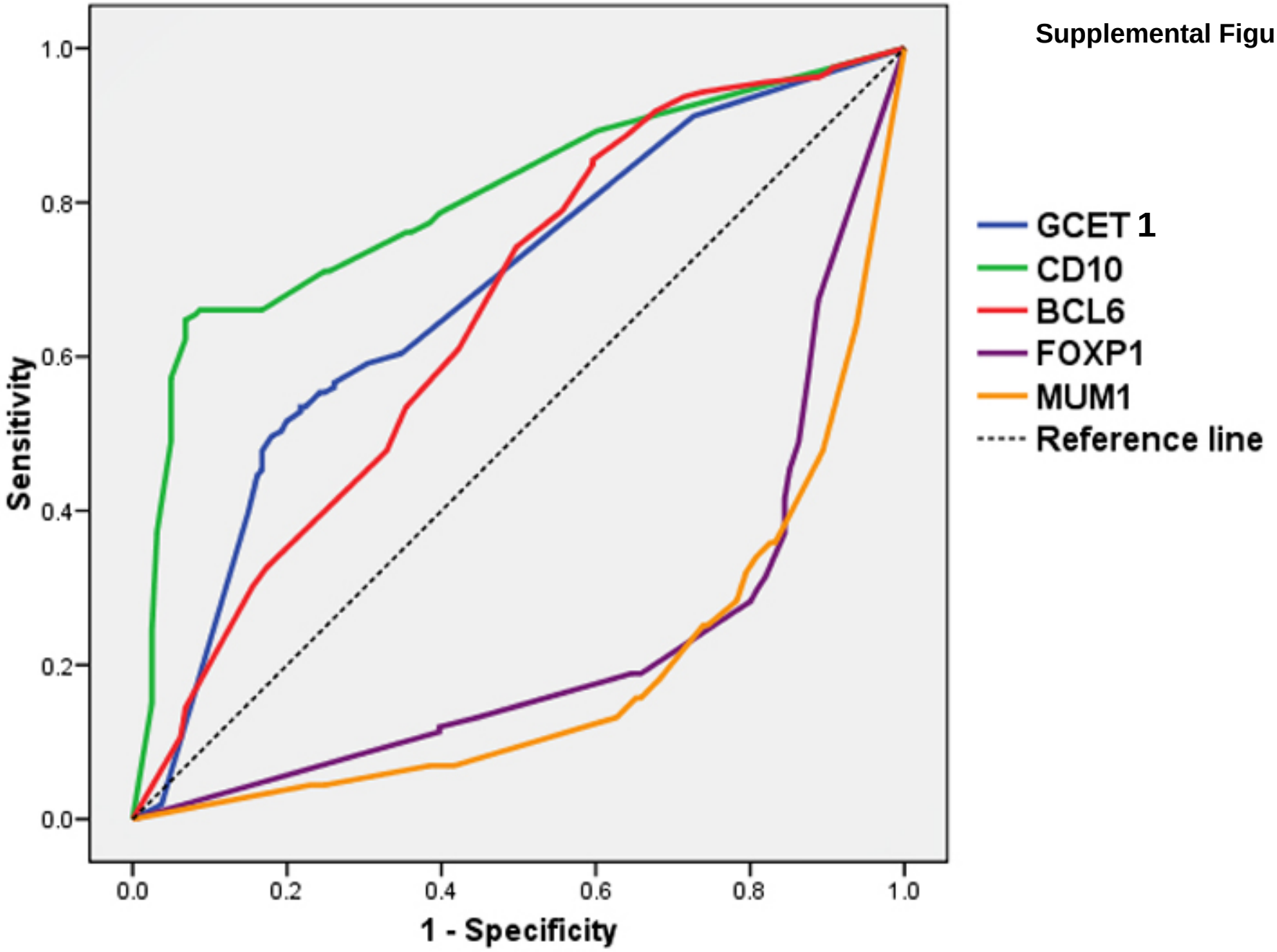
Supplemental Figure 1. Receiver operating characteristic (ROC) curves. The antibodies in our series were associated with either gene expression of GCB-DLBCL (green, blue, and red lines) or with ABC-DLBCL (orange and purple lines). Deviation from the ROC curve of the markers associated with GCB-DLBCL is inversed to that of the markers associated with ABC-DLBCL. The diagonal dotted line corresponds to a hypothetic variable marker equally expressed in both GCB- and ABC-DLBCL. Based on these parameters, CD10 was the best marker for GCB (AUROC 0.800; 95% CI, 0.750–0.849; $P < 0.00001$), followed by GCET1 (AUROC 0.683, 95% CI, 0.625–0.741; $P < 0.0001$) and BCL6 (AUROC 0.658; 95% CI, 0.600–0.717; $P < 0.001$). In contrast, MUM1 (AUROC 0.799; 95% CI, 0.753–0.846; $P < 0.00001$) was more specific but less sensitive than FOXP1 (AUROC 0.643; 95% CI, 0.689–0.796; $P < 0.0001$) in recognizing ABC-DLBCL. The source of antibodies used in this project was Ventana (Tucson, AR) for CD3 (clone PS1 at 1:200 dilution), Novocastra (Leica Microsystems Inc., Buffalo

Grobe, IL) for CD10 (clone 56C6, at 1:100) and BCL6 (clone LN22 at 1:100), DAKO (Carpinteria, CA) for CD20 (clone L26 at 1:200) and MUM1 (clone MUMp1 at 1:20), and Abcam (Cambridge, MA) for GCET1 (clone RAM at 1:4) and FOXP1 (clone JC12 at 1:16,000).

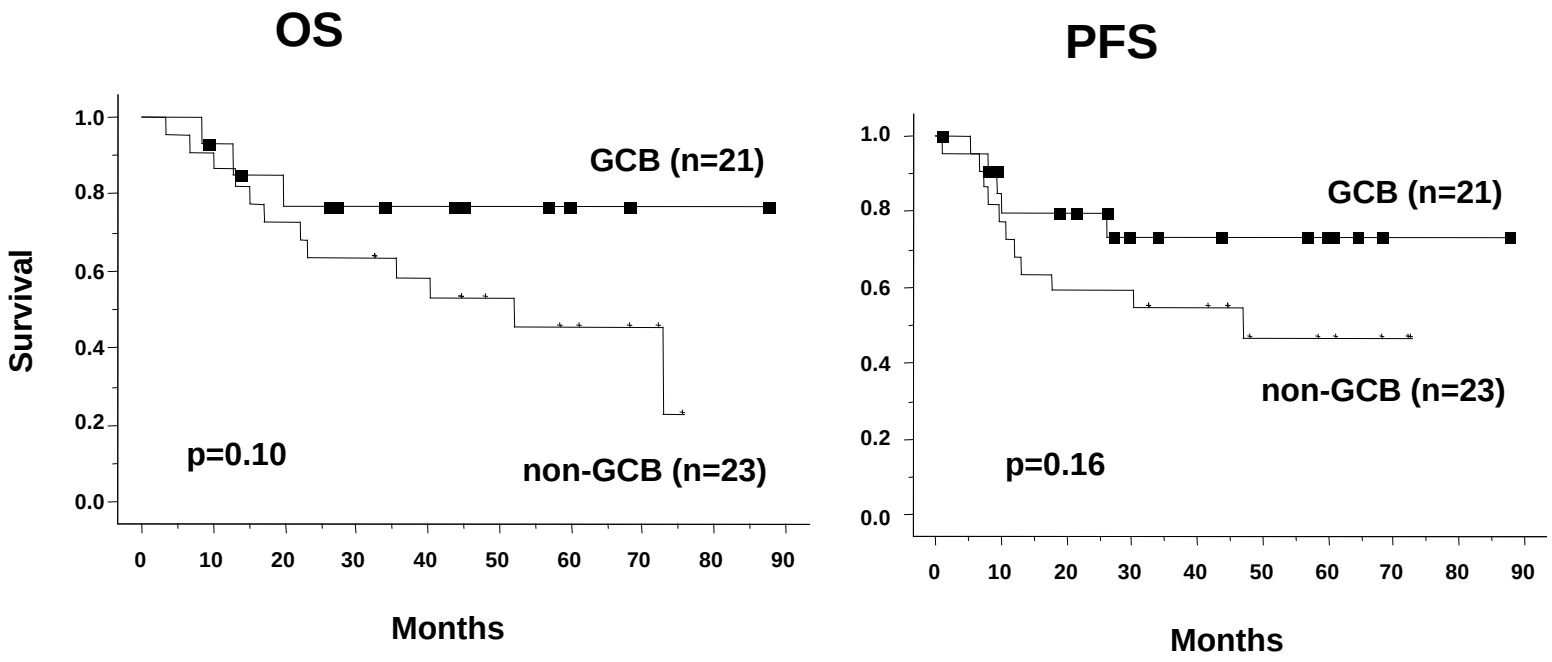
Supplemental Figure 2. Overall survival and PFS analyses of the 44 unclassified DLBCL cases according to GEP analysis. These patients were stratified in two groups with different OS and PFS by our TMA and immunohistochemistry algorithm. Five-year OS was 77 ± 2% for GCB vs. 45 ± 5% for non-GCB (p=0.10), while 5-year PFS was 73 ± 3% for GCB vs. 44 ± 6% for non-GCB (p=0.16). Most of 44 unclassified cases by GEP show preserved morphology with greater than 60% tumor content. The unclassifiable mechanism is unclear, but maybe related to altered epigenetic regulatory mechanisms. Derivations of signaling pathway, autophage, cell cycle regulation, and apoptosis may be also important to this unclassifiable mechanism.

Supplemental Figure 3. Relationship between our TMA immunohistochemistry algorithm and International Prognostic Index (IPI) scores. Survival analysis of 431 R-CHOP-treated patients with available IPI score. PFS curves of GCB (A) and non-GCB (B) patients divided according to IPI score 0–1 (low) vs. 2–3 (intermediate) vs. 4–5 (high). A statistically significant difference was observed between all curves, except for IPI 2–3 vs. 4–5 in GCB patients ($P = 0.12$).

Supplemental Figure 1



Supplement Figure 2



Supplemental Figure 3