bjh research paper

Comparative analysis of gene expression platforms for cell-of-origin classification of diffuse large B-cell lymphoma shows high concordance

Sophia Ahmed,[1] Paul Glover,[2] Jan Taylor,[2] Chulin Sha,[1] Matthew A. Care,[3] Reuben Tooze,[2,3] Andrew Davies,[4] David R. Westhead,[1] Peter W. M. Johnson,[4] Catherine Burton[2] and Sharon L. Barrans[2]

[1]*Faculty of Biological Sciences, University of Leeds, Leeds,* [2]*Haematological Malignancy Diagnostic Service, St James' University Hospital, Leeds,* [3]*Faculty of Medicine, University of Leeds, Leeds, and* [4]*Cancer Research UK Centre, Southampton Clinical Trials Unit, University of Southampton, Southampton, UK*

Received 23 September 2020; accepted for publication 1 November 2020

Correspondence: Sharon L. Barrans, Haematological Malignancy Diagnostic Service, St James' University Hospital, Leeds LS9 7TF, UK.
E-mail: sharon.barrans@nhs.net

Summary

Cell-of-origin subclassification of diffuse large B cell lymphoma (DLBCL) into activated B cell-like (ABC), germinal centre B cell-like (GCB) and unclassified (UNC) or type III by gene expression profiling is recommended in the latest update of the World Health Organization's classification of lymphoid neoplasms. There is, however, no accepted gold standard method or dataset for this classification. Here, we compare classification results using gene expression data for 68 formalin-fixed paraffin-embedded DLBCL samples measured on four different gene expression platforms (Illumina wG-DASL™ arrays, Affymetrix PrimeView arrays, Illumina TrueSeq RNA sequencing and the HTG EdgeSeq DLBCL Cell of Origin Assay EU using an established platform agnostic classification algorithm (DAC) and the classifier native to the HTG platform, which is CE marked for *in vitro* diagnostic use (CE-IVD). Classification methods and platforms show a high level of concordance, with agreement in at least 80% of cases and rising to much higher levels for classifications of high confidence. Our results demonstrate that cell-of-origin classification by gene expression profiling on different platforms is robust, and that the use of the confidence value alongside the classification result is important in clinical applications.

Keywords: diffuse large B cell lymphoma, genetic subtyping, gene expression, cell of origin, lymphomas.

Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma and has been revealed to consist of distinct subtypes on the basis of gene expression patterns that reflect the putative cell of origin (COO).[1] The two main recognised COO subtypes within DLBCL are activated B-cell-like (ABC) and germinal centre B-cell-like (GCB), with a third category referred to as unclassified (UNC) or Type III. GCB generally has better prognosis than ABC following standard R-CHOP [rituximab, cyclophosphamide, doxorubicin hydrochloride, vincristine (Oncovin, Vincasar PFS), prednisolone] chemotherapy, and this has been consistently replicated in many studies using gene expression profiling (GEP) to assign COO groups.[2]

Gene expression profiling can now be applied to routinely processed formalin-fixed paraffin-embedded (FFPE) diagnostic tissue biopsies but, despite this, it has not been widely incorporated into routine clinical use, and the surrogate immunohistochemistry (IHC)-based Hans test remains as standard practice. This uses just three markers (CD10, BCL6 and IRF4/MUM1) to classify patient samples as either GCB or non-GCB; however, reproducibility has proved difficult and this classification does not identify significant differences in overall survival.[3] COO classification was recognised in the 2016 update of the World Health Organization classification of lymphoid neoplasms,[4] which states that COO should be defined preferably by GEP, and recommends Hans IHC only where this is not possible.

We applied real-time COO classification, using the Illumina whole genome cDNA-mediated annealing, selection, extension and ligation (WG-DASL) gene expression profiling assay and the DAC classifier,[5] to patients enrolled in the Randomised Evaluation of Molecular guided therapy for Diffuse Large B-cell Lymphoma with Bortezomib (REMoDL-B) study (NCT01324596).[6] This aimed to evaluate the clinical efficacy of the combination of bortezomib with R-CHOP and to determine whether the COO subtypes respond

doi: 10.1111/bjh.17246

differently. This study was the first large-scale study in DLBCL to use real-time molecular characterisation for prospective stratification and randomisation and subsequent analysis of biologically distinct subgroups.

Since the initiation of this trial a number of commercially available platforms for COO assignment by GEP have emerged, including Lymph2Cx (Nanostring Technologies) and HTG EdgeSeq DLBCL Cell of origin Assay EU (HTG Molecular Diagnostics Inc., Tuscon, AZ, USA), and some authors have developed classifiers for use with RNA-seq data.[7] It is important to appreciate that there is no gold standard for COO assignment, and all studies show a spectrum of gene expression patterns with a significant 'grey zone' of intermediate cases. Intermediate cases that fall close to a classification boundary have limited biological differences to cases falling just on the other side of the boundary, and as such, perfect concordance between different COO assignment methods may not be observed, nor should they necessarily be expected. However, to validate the technique for clinical practice it is important to understand the expected degree of agreement between different GEP platforms and classification algorithms, and how this relates to measures of classification confidence.

The aim of this study was to evaluate reproducibility of COO classification in DLBCL using different GEP platforms with the platform-agnostic DAC algorithm[5] for potential implementation in routine use. We investigated an Affymetrix array based method and RNA-seq data, as well as the HTG EdgeSeq DLBCL Cell of Origin Assay EU and on board CE-IvD classifier and compared the results with those obtained using the now withdrawn Illumina WG-DASL^{TM} GEP platform classified with the DAC (herein referred to as DASL_DAC).

Materials and methods

Sample selection

DASL_DAC classification had already been performed for the >1000 samples in the ReMoDLB trial.[6] From these, a subset of representative samples was selected based on the overall distribution of COO classes and classification confidences (n = 286 in total (ABC n = 76, UNC n = 71, GCB n = 139) represented in Figure S1. Of these, n = 68 (ABC n = 20, UNC n = 10, GCB n = 38) had adequate RNA and data quality control metrics from all platforms and were used in the final analysis (see Table SI).

RNA extraction

Total RNA was extracted from 5μm paraffin sections using the Ambion RecoverAll kit (Thermo Fisher Scientific) standard protocol, with an extended 16-hour protease digestion. RNA was assessed using either a NanoDrop Spectrophotometer (Thermo Fisher Scientific) 260:280 ratio or TapeStation Genetic Analyser (Agilent) DV200 measurements for quality assessment, as per requirements of the manufacturer. Illumina WG-DASL and HTG EdgeSeq platforms did not require quality assessment of RNA. Affymetrix Primeview arrays required verification of RNA purity by 260:280 ratio. The Illumina RNA Exome kit required quantitation, which was performed using the Qubit Flourometer RNA quantitation high sensitivity kit (Thermo Fisher Scientific) and quality was assessed by Tapestation (Agilent) DV200 readings, which were required to be of 30 or above. Details of input requirements, represented genes and data output are provided in Table SII.

Gene expression profiling

Gene expression profiling methods tested in this study included four commercially available products: the now withdrawn Illumina WG-DASL Array (DASL), the Affymetrix Primeview Array, the HTG genomics EdgeSeq DLBCL cell-of-origin assay and an RNA sequencing approach, the Illumina TruSeq RNA Exome library (RNA-Seq) (see Data S1 for a description of each method and Table SII for a description of each method and the data acquisition and analysis pipelines). DASL was performed in real-time during the REMoDL-B trial recruitment (Sept 2011–May 2015). GEP on Affymetrix and RNA-seq platforms was performed respectively on the same stored RNA that was used for DASL. HTG EgdeSeq was carried out on either HTG recommended unstained sections (where available) or the same stored RNA (n = 48 vs. n = 20 respectively). As part of the platform validation, paired RNA and FFPE sections from the same sample were processed on the HTG platform. Concordant results were generated in 13 of 14 pairs and only one sample was borderline and switched between class (RNA COO = GCB, FFPE section COO = ABC both classified with HTG EdgeSeq, compared with RNA COO DASL_DAC = GCB). This sample was not included in the final set of 68 samples. Prior to downstream COO classification, expression data from DASL, Affymetrix and RNA-seq were quantile normalised using the limma package[8] implemented in R statistical software version, 4.2.[9]

Cell-of-origin classification

Cell-of-origin classification was performed using the DLBCL automated classifier (DAC)[5] with expression data from the DASL, Affymetrix and RNA-seq platforms. As well as providing a COO output, the classification is also associated with a classification probability (P) for each class (P_{ABC}, P_{UNC} and P_{GCB} with $P_{ABC} + P_{UNC} + P_{GCB} = 1$) and quality control metrics. In each case, the assigned class is the one with highest probability, and here we define a log-odds measure of classification confidence $L = \log_2(P_x/(1-P_x))$ where P_x is the probability of the assigned class. For example, a case with P_{ABC} = 0·4 P_{UNC} = 0·3 and P_{GCB} = 0·3 would be assigned as

ABC with $L = \log_2(0.4/0.6) = -0.6$, and considered a 'weak' assignment (defined as $L < 0$) with the other two classes collectively more probable than the assigned class. Alternatively, $P_{ABC} = 0{\cdot}8$ $P_{UNC} = 0{\cdot}1$ and $P_{GCB} = 0{\cdot}1$ has $L = \log_2(0.8/0.2) = 2.0$ and is a 'confident' assignment (defined as $L \geq 1$: assigned class more than twice as probable as the other two collectively). Other moderate confidence assignments have $0 \leq L < 1$.

The HTG EdgeSeq DLBCL panel measures the expression of 92 genes, but with only 13 of 20 DAC genes represented. While DAC can be used with a subset of its classifier genes, such an analysis is not presented here since it is not possible to quantify the effect of missing genes on the final assignments. Instead for these data we used the COO classification algorithm native to the HTG platform, for comparison with the DAC assignments above. The HTG EdgeSeq DLBCL Cell of Origin Assay EU on board classifier is CE marked for *in vitro* diagnostic use (CE-IvD). We note that the HTG native algorithm is trained to minimise the UNC class compared with other methods (https://www.freepatentsonline.com/y2018/0340231.html#google_vignette), preferring classification as either ABC or GCB: in this dataset the HTG classifier calls four samples UNC, compared to an average of nine UNC for DAC.

Results

Full COO results can be found in Fig 1 and Table SIII. The heatmap in Fig 1 shows that the same pattern of COO-related gene expression over the gene set used by the DAC classifier is detected by each of the four gene expression measurement technologies (limited in the HTG data to those genes available on the platform). Fig 1 also reveals the broad trend that classifications that are discordant between two or more platforms tend to be ABC/UNC or GCB/UNC disagreements, rather than ABC/GCB, and tend to occur in cases with lower confidence on the DASL_DAC platform. The number of concordant and discordant assignments are further elucidated by the confusion matrices provided in Table SIV.

First, we compared classifications from the DAC algorithm with the CE-IvD HTG platform and its on-board classifier. Using DAC with gene expression data from the DASL platform (where DAC is most extensively validated, particularly in the REMoDL-B trial) we found 58 of 68 ($85 \pm 7\%$)[1] samples in agreement on the COO class (Fig 1). However, 7 of 10 samples where classifications did not agree were classed as UNC by DASL_DAC and either ABC or GCB by HTG. We commented above that the HTG classifier is trained to minimise the UNC class, so these disagreements were not unexpected. In view of this, comparing only on cases not classed as UNC by DASL_DAC gives agreement of 55 of 58 ($95 \pm 5\%$), and of the three discordant cases two were DASL_DAC weak confidence assignments ($L < 0$) and one was moderate confidence ($0 \leq L < 1$). Comparison of HTG with DAC derived from Affymetrix and RNA-seq gene expression platforms showed similar results with, respectively, $82\% \pm 7$ and $79\% \pm 8$ overall concordance, increasing to $88\% \pm 7$ and $86\% \pm 8$ when discounting cases assigned as UNC by DAC. Overall, therefore, there is a good level of concordance between HTG and the combination of DAC with any gene expression platform, with discordance generally associated with lower confidence cases and dominated by cases involving the UNC class.

Having confirmed the high level of agreement between DAC classifications and those from HTG, we next moved to a like-for-like comparison of the DAC algorithm applied to gene expression data from DASL arrays, Affymetrix arrays and RNA-seq. In this three-way comparison, overall, 53 of 68 ($78 \pm 8\%$) cases had the same class on all three platforms. The remaining 15 cases all agreed on two platforms, with the single discordant result approximately evenly spread between the platforms (five RNA-seq, seven DASL, three Affymetrix). The distributions of classification confidences are shown in Fig 2, where it is notable that the distribution of log-odds values is different, depending on both platform and assigned class, and that GCB classifications are consistently of higher confidence. Of the 15 discordant cases (highlighted on Fig 2), 12 were ABC/UNC or GCB/UNC disagreements and only three were ABC/GCB disagreements. The average log-odds confidence of the discordant assignments was 0·6 compared to 1·6 for assignments from DAC overall, showing that discordant calls are significantly lower in confidence ($P < 0{\cdot}01$, *t*-test). Viewed another way, compared to a $78 \pm 8\%$ 3-way concordance overall, eliminating 'weaker' assignments and considering only moderately confident and confident assignments yields 3-way concordance of 39 of 48 ($81 \pm 9\%$) cases and considering only confident assignments yields 25 of 25 (100%) 3-way concordance. These results are consistent with discordant calls being mostly associated with lower confidence on the discordant platform as well as lying near the classification boundaries between UNC and the main ABC/GCB classes.

Discussion

The results reported above show a good degree of agreement between the DAC algorithm applied to DASL data and the native COO classifier with the HTG Edgeseq gene expression platform. Most disagreements are cases in the UNC class from DAC, a class which the training of the HTG classifier aims to minimise and which are preferentially classified as ABC or GCB by that algorithm. We prefer an approach that does not minimise the UNC class, since there is evidence that it is more than a group of intermediate cases, in

[1]Uncertainty estimates here and in the text that follows are 90% confidence intervals on the binomial parameter (p), calculated using the Wald method.

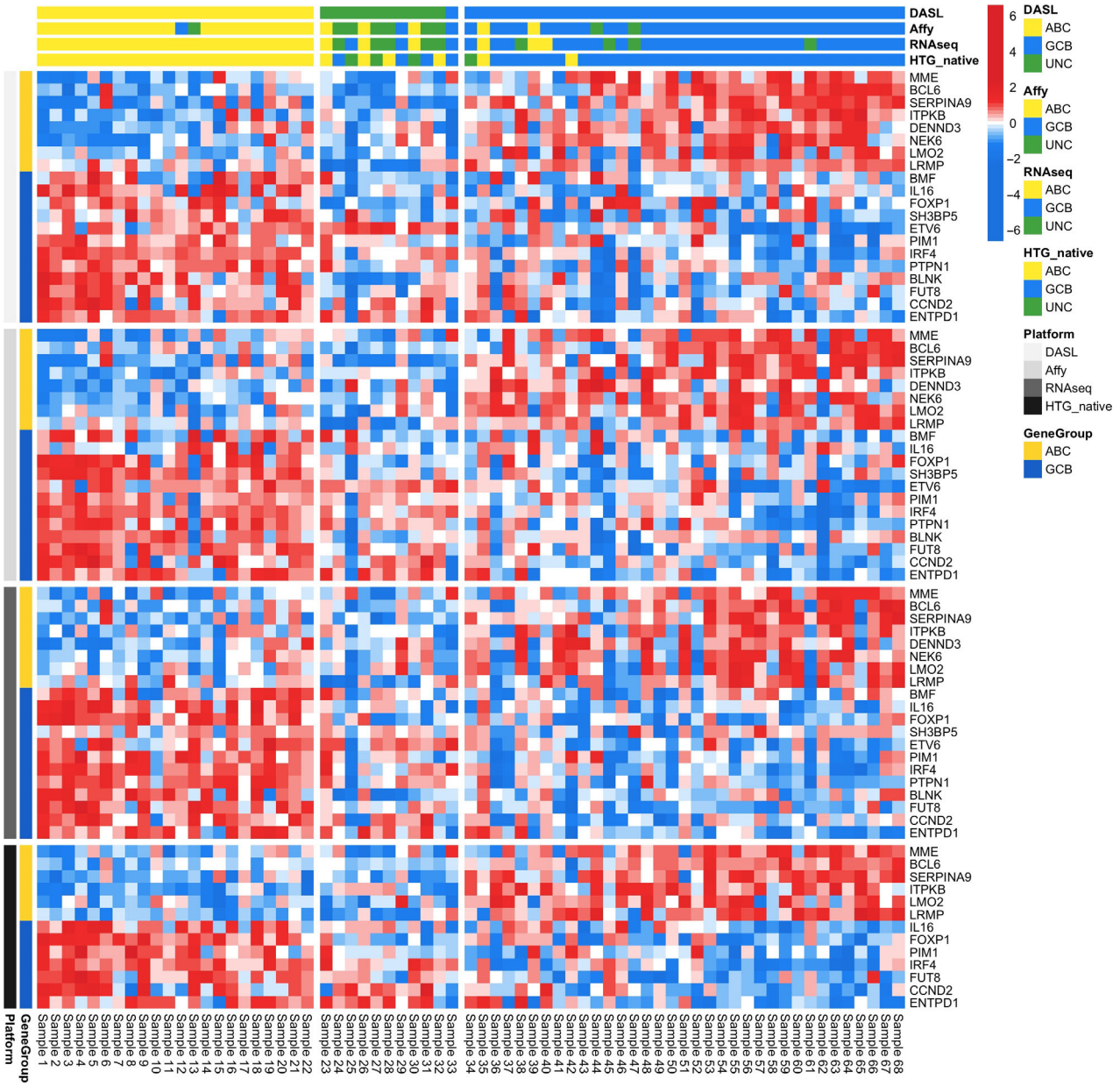

Fig 1. Gene expression patterns and classification results for 68 diffuse large B cell lymphoma samples (columns). The heat map (bottom) shows gene expression levels (blue, low; red, high) for the 20 gene signature used by the agnostic classification algorithm (DAC) cell-of-origin classifier, as measured on four different platforms [Illumina DASL, Affymetrix, Illumina RNA-seq and HTG molecular]. For each platform, the genes are divided in two groups, those up-regulated in germinal centre B cell (GCB) (top, 8 genes) and those up-regulated in the activated B cell (ABC) (bottom, 12 genes). Only the 13 DAC classifier genes that are present on the HTG EdgeSeq DLBCL panel are shown. The classification results (top) are from the DAC classifier with expression data from the DASL, Affymetrix and RNA-seq platforms, compared to the results from the HTG COO classifier applied to HTG platform expression data. Samples are sorted by the class on the DASL platform (yellow, ABC; green, unclassified (UNC); blue, GCB) and ordered by classification probability on the same platform.

particular containing cases with a T-cell-dominated immune response.[10] Otherwise, disagreements tend to have low classification confidence from the DAC algorithm, indicating that they may lie close to classification boundaries and reflect biological heterogeneity and/or ongoing differentiation within the tumour.

The results for the DAC algorithm, when used with different gene expression measurement technologies, reveals complete concordance across all three platforms approaching 80%, with disagreements similarly dominated by those involving the UNC class and having low confidence on the discordant platform. Performance of the three gene

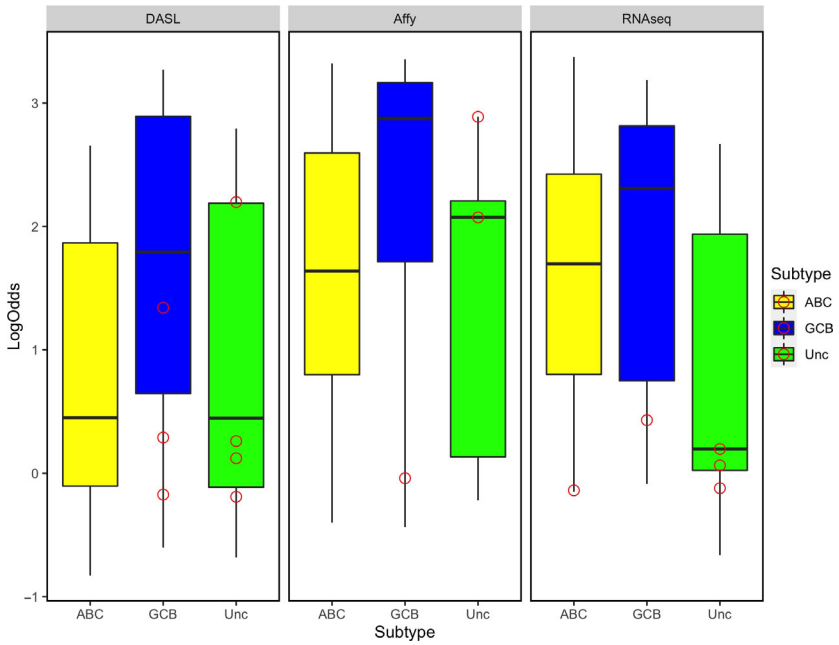

Fig 2. Box and whisker plots of log-odds confidence measures from the agnostic classification algorithm classifier applied to gene expression data from three different platforms (left, Illumina WG-DASL™; middle, Affymetrix; right, Illumina RNA-seq) split according to the assigned class (yellow, activated B cell (ABC); blue, germinal centre B cell (GCB); green, unclassified UNC). Confidence values for samples that are discordant on the platform in question are shown as red circles.

expression platforms was similar, with no platform dominating the discordance statistics. Since these studies were carried out at significantly different times, it is possible that even this relatively small level of disagreement is influenced to some extent by RNA deterioration with time.

The agnostic classification algorithm produces a classification confidence measure expressed as a probability, which we have here converted to a log-odds score for convenience. The results indicate that this is a useful measure of classification confidence, where high confidence is associated with clear cases and agreement between platforms. In view of the lack of a gold standard for COO classification, and the observed spectrum of gene expression patterns between the clear cases of ABC and GCB, where methods tend to agree, we suggest that any clinical use of COO classes should consider the confidence measure alongside the assigned class.

While COO classification remains relevant in clinical practice, further GEP subgroups, molecular high grade (MHG)[11] and DHitSig[12] have been recently defined by our group and others. These groups are largely consistent between studies and add further subgroups beyond the COO, with both studies identifying a poor prognostic group within GCB. The results of this study provide confidence in classification across different gene expression technologies, both in the classification of COO and, moving forward, with more recently described classification schema (MHG/DHitSig).

Furthermore, we have previously shown that biologically relevant mutations differentiate appropriately between COO classes, both in REMoDL-B[6] and in our population dataset,[13] providing further confidence that the COO calls are not only reproducible, but also biologically important. Large-scale sequencing studies in DLBCL have identified mutation clusters as an alternative approach to subclassification of DLBCL.[14,15] There is significant overlap between COO and the mutational clusters, providing additional knowledge of the underlying biology of these disorders and, if used in combination, the vision is towards providing a more personalised medicine approach for individual patients.

Acknowledgements

This work was supported by a grant from Blood Cancer UK (formerly Bloodwise), grant number 15002.

Conflict of interest

The authors declare no competing financial interests.

Author contributions

The study was designed by SB, RT, DRW, AD, PWMJ and CB. Laboratory work was carried out by SA, SB. Data were analysed by SA, PG, JT, CS, MAC and DRW. The paper was written by SA, SB, JT and DRW.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Supplementary materials and methods.

Table SI. Samples grid for case selection.

Table SII. Technical comparison of gene expression platforms

Table SIII. COO classification of $n = 68$ DLBCL samples on four GEP platforms.

Table SIV. Confusion matrices for three COO classes across the four gene expression platforms.

Figure S1. Confidence values for the samples analysed by all 4 methods (red points, $n = 68$) compared to the total number of samples in the study (black points, $n = 286$).

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;**403**(6769):503–11.
2. Lenz G, Wright GW, Emre NCT, Kohlhammer H, Dave SS, Davis RE, et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci USA*. 2008;**105**(36):13520–5.
3. Read JA, Koff JL, Nastoupil LJ, Williams JN, Cohen JB, Flowers CR. Evaluating cell-of-origin subtype methods for predicting diffuse large B-cell lymphoma survival: a meta-analysis of gene expression profiling and immunohistochemistry algorithms. *Clin Lymphoma Myeloma Leuk*. 2014;**14**(6):460–467 e2.
4. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;**127**(20):2375–90.
5. Care MA, Barrans S, Worrillow L, Jack A, Westhead DR, Tooze RM. A microarray platform-independent classification tool for cell of origin class allows comparative analysis of gene expression in diffuse large B-cell lymphoma. *PLoS One*. 2013;**8**(2):e55895.
6. Davies A, Cummin TE, Barrans S, Maishman T, Mamot C, Novak U, et al. Gene-expression profiling of bortezomib added to standard chemoimmunotherapy for diffuse large B-cell lymphoma (REMoDL-B): an open-label, randomised, phase 3 trial. *Lancet Oncol*. 2019;**20**(5):649–62.
7. Reddy A, Zhang J, Davis NS, Moffitt AB, Love CL, Waldrop A, et al. Genetic and functional drivers of diffuse large B cell lymphoma. *Cell*. 2017;**171**(2):481–494 e15.
8. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;**43**(7):e47.
9. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
10. Care MA, Westhead DR, Tooze RM. Gene expression meta-analysis reveals immune response convergence on the IFNgamma-STAT1-IRF1 axis and adaptive immune resistance mechanisms in lymphoma. *Genome Med*. 2015;**7**:96.
11. Sha C, Barrans S, Cucco F, Bentley MA, Care MA, Cummin T, et al. Molecular high-grade b-cell lymphoma: defining a poor-risk group that requires different approaches to therapy. *J Clin Oncol*. 2019;**37**(3):202–12.
12. Ennishi D, Jiang A, Boyle M, Collinge B, Grande BM, Ben-Neriah S, et al. Double-hit gene expression signature defines a distinct subgroup of germinal center b-cell-like diffuse large b-cell lymphoma. *J Clin Oncol*. 2019;**37**(3):190–201.
13. Lacy SE, Barrans SL, Beer PA, Painter D, Smith AG, Roman E, et al. Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a Haematological Malignancy Research Network report. *Blood*. 2020;**135**(20):1759–71.
14. Chapuy B, Stewart C, Dunford AJ, Kim J, Kamburov A, Redd RA, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med*. 2018;**24**(5):679–90.
15. Schmitz R, Wright GW, Huang DW, Johnson CA, Phelan JD, Wang JQ, et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. *N Engl J Med*. 2018;**378**(15):1396–407.