`hemaClass.org`: Online one-by-one microarray normalization and classification of hematological cancers for precision medicine

Steffen Falgreen*[1], Anders Ellern Bilgrau*[12], Rasmus Froberg Brøndum[1], Lasse Hjort Jakobsen[12], Jonas Have[12], Kasper Lindblad Nielsen[12], Tarec Christoffer El-Galaly[1], Julie Støve Bødker[1], Alexander Schmitz[1], Ken H. Young[3], Hans Erik Johnsen[12], Karen Dybkær[12], and Martin Bøgsted†[12]

[1]Department of Haematology, Aalborg University Hospital
[2]Department of Clinical Medicine, Aalborg University
[3]Department of Hematopathology, MD Anderson Cancer Center

August 19, 2016

Abstract

Background Dozens of omics based cancer classification systems have been introduced with prognostic, diagnostic, and predictive capabilities. However, they often employ complex algorithms and are only applicable on whole cohorts of patients, making them difficult to apply in a personalized clinical setting.
Results This prompted us to create `hemaClass.org`, an online web application providing an easy interface to one-by-one RMA normalization of microarrays and subsequent risk classifications of diffuse large B-cell lymphoma (DLBCL) and multiple myeloma (MM) into cell-of-origin and chemotherapeutic sensitivity classes. Classification results for one-by-one array pre-processing with and without a laboratory specific RMA reference dataset were compared to cohort based classifiers in 4 publicly available datasets. Classifications showed high agreement between one-by-one and whole cohort pre-processsed data when a laboratory specific reference set was supplied. The website is essentially the R-package **hemaClass** accompanied by a Shiny web application. The well-documented package can be used to run the website locally or to use the developed methods programmatically.
Conclusions The website and R-package is relevant for biological and clinical lymphoma researchers using affymetrix U-133 Plus 2 arrays, as it provides reliable and swift methods for calculation of disease subclasses. The proposed one-by-one pre-processing method is relevant for all researchers using microarrays.

Keywords: Diffuse Large B-Cell Lymphoma; Multiple Myeloma; Classification; Chemosensitivity; Microarray; Pre-processing; Robust Multichip Average; Gene expression profiling; Cell of origin; Affymetrix Human Genome U133 Plus 2 microarrays

1 Background

In addition to current clinically used risk factor scoring systems, several independent gene expression profile (GEP) based risk stratifications have been proposed, with biological and clinical significance in hematological cancers. Although drug targetable genes, which are only expressed in subtypes of e.g. DLBCL tumours have been identified, they are not readily applicable in clinical research and routine settings due to a lack of available routine diagnostic tests [34, 49].

*Shared first authorship
†To whom correspondence should be addressed (`mboegsted@dcm.aau.dk`)

Alizadeh et al. [1] developed an important example of a biological sub-classification of lymphoma. On the basis of GEP analyses, DLBCL cases were classified as activated B-cell phenotype (ABC) or germinal center B-cell phenotype (GCB) with different clinical outcomes. The validity of this classification and its prognostic importance have been confirmed in a number of later studies [27, 38, 43, 44, 46]. Recently, we have refined the ABC/GCB subclassification of DLBCL to include a B-cell Associated Gene Signature (BAGS) classifier capable of classifying DLBCL samples into 5 different B-cell subtypes: Naive (N), Centrocyte (CC), Centroblast (CB), Memory (M), and Plasmablasts (PB) [18]. The BAGS classifier stratifies the GCB phenotype into CC and CB subtypes, with superior survival in the CC subtype. Thus, different treatment regimes could be considered in subsets of the GCB class of patients. In another study we developed classification based resistance gene signatures (REGS) for the most prominent drugs used in the treatment of DLBCL patients: Cyclophosphamide (C), Doxorubicin (H), and Vincristine (O) [22]. However, these and most existing algorithms are only applicable in the presence of whole cohorts of patients, making them difficult to apply in a routine clinical setting.

The traditional lymphoma staging and risk classification systems are based on the Ann Arbor classification for staging of lymphoma (extent of disease and extranodal involvement) and simple prognostic tools such as the international prognostic index (IPI, [52]) for large cell lymphoma and the Follicular Lymphoma International Prognostic Index (FLIPI, [51]), both derived from patient age, performance status, easy available blood tests, and disease stage. Due to the simplicity of these clinical risk stratification algorithms they are still the most widely used risk scoring systems today. Risk stratification according to these algorithms has been systematized and made easily accessible for desktop, online, and even smart-phone use. Easily accessible molecular classification methods are, however, lagging behind, thereby delaying the translation of new molecular findings into clinical practice. A few methods exist for cancer types other than lymphoma, including acute myeloid leukemia (AML) [30], and for lymphoma Care et al. [9] has developed an ABC/GCB classifier, which is stable across microarray technologies and trial centres. This ABC/GCB classifier is, however, potentially biased towards classes which differentiate the prognosis instead of biological classes, since the ensemble of classifiers were chosen based on their ability to separate survival.

In clinical settings, the methods need to be applicable for a single patient sample and straightforward to use. This prompted us to develop a user-friendly and flexible web-based tool for ABC/GCB, BAGS, and REGS classification using our recently developed classifiers for microarray data based on the Affymetrix's Human Genome U-133 Plus 2 array. The classifications made by the web-based tool `hemaClass.org` are compared to the existing state-of-the-art and approved ABC/GCB classifications of DLBCL. We believe that `hemaClass.org` will provide a novel and user-friendly concept for bringing complex molecular classification of diseases more swiftly into daily clinical practice.

2 Implementation

2.1 Classification workflow

The workflow architecture of `hemaClass.org` is illustrated in Figure 1. The user is presented with a graphical interface for uploading data and adjusting settings. The user data is RMA pre-processed by the server (with an optional user one-by-one RMA reference), and subsequently processed by the classification algorithms. The results are then returned for download and inspection via the user interface.

2.2 Software availability and technical details

The interactive web application available at `hemaclass.org` was created using the statistical programming language R [14], the software package **shiny** [10], and the accompanying Linux server software. All `hemaClass.org` functionality, including the RMA normalization and classification procedures, are available through the accompanying package **hemaClass** based on a number of packages from the Comprehensive R Archive Network [14] and the Bioconductor environment [26]. The Shiny server handles the interaction between the front end web application and the back end R processing. The back end is essentially the well-

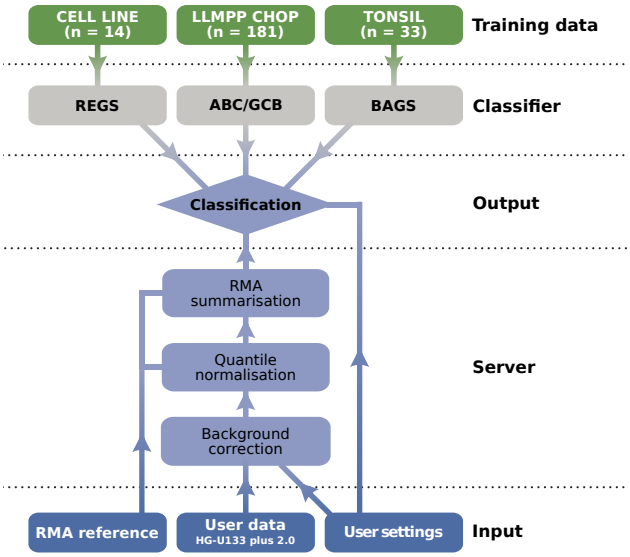

Figure 1: Diagram of the `hemaClass.org` workflow architecture.

documented **hemaClass** package which can be utilized as a programmatical interface to the functionality of the website. However, the package also allows users to run a local instance of the website if one wishes to avoid uploading large files to our server. The development and latest version of **hemaClass** is open source and freely available at `https://github.com/oncoclass/hemaclass` for sharing, modification, and redistribution. All bug-reports, suggestions, and comments on the website or package are welcome and should be posted to the github page following the link above. The regular RMA pre-processing is carried out with the Bioconductor package **affy** [25]. Core functions for the one-by-one RMA pre-processing are written in C++ and imported to R using **Rcpp** and **RcppArmadillo** [19, 20, 21, 47].

2.3 Data overview

The seven gene expression datasets used in this paper are summarized in Table 1. All GEP data are from the Affymetrix GeneChip HG-U133 Plus 2.0 array and available at the Gene Expression Omnibus (GEO) [2] website (`http://www.ncbi.nlm.nih.gov/geo/`). To establish the classifiers the following datasets are used:

1. Gene expressions from 181 CHOP treated DLBCL patients are used to establish the ABC/GCB classifier. This cohort will be referred to as the *LLMPP CHOP* (Lymphoma/Leukemia Molecular Profiling Project CHOP) cohort [38]. The cohort is also used as a default reference set throughout the paper for one-by-one RMA normalization of arrays.

2. The BAGS classifier is based on gene expression data from eight human tonsils sorted in five B-cell subsets. This dataset is also used for scaling of gene expression data for BAGS classification, and will be referred to as the *Tonsil dataset* [18].

3. The REGS classifiers are based on a panel of 12 Multiple Myeloma (MM) and 14 DLBCL cell lines. This panel will be referred to as *BCELL26*. The DLBCL part of the cell line panel is used for scaling of patient data and will be referred to as *DLBCL14* [22].

For validation the following four DLBCL cohorts are used:

4. The Aalborg OCT cohort (*CHEPRETRO*) of 89 Danish DLBCL patients undergoing first-line treatment at Aalborg University Hospital [18].

Table 1: Overview of used datasets and GEO accession numbers.

No.	Dataset	n	Usage	GEO number	Ref.
1.	LLMPP CHOP	181	Reference	GSE10846	[38]
2.	Tonsil	33	Reference	GSE56315	[18]
3.	BCELL26	26	Reference	GSE53798	[22]
4.	CHEPRETRO	89	Validation	GSE56315	[18]
5.	IDRC	470	Validation	GSE31312	[54]
6.	LLMPP R-CHOP	233	Validation	GSE10846	[38]
7.	MDFCI	90	Validation	GSE34171	[43]

5. The International DLBCL Rituximab-CHOP Consortium MD Anderson (*IDRC*) cohort of 470 DLBCL patients treated with R-CHOP first-line therapy [54]. Note, that these samples are formalin-fixed, paraffin-embedded (FFPE).

6. The Lymphoma/Leukemia Molecular Profiling Project R-CHOP (*LLMPP R-CHOP*) cohort of 233 DLBCL patients treated with R-CHOP first-line therapy [38].

7. The Mayo-Dana-Farber Cancer Institute (*MDFCI*) cohort of 90 DLBCL patients treated with R-CHOP first-line therapy [43].

The GEO datasets were downloaded using the R-package **DLBCLdata** [3].

2.4 One-by-one RMA normalization

Robust multichip average (RMA) pre-processing consists of three steps in the order: (1) Background correction, (2) quantile normalization, and (3) summarization of probes to probe-sets [32, 33]. Confer Bolstad [7] for a comprehensive account on RMA. Both the quantile normalization and summarization procedures of RMA are cohort based and hence need to be altered to facilitate a one-by-one RMA pre-processing scheme. Previous approaches similar to the one-by-one normalization approach used by `hemaClass.org` have been described by Katz et al. [35] and McCall et al. [41] As quantile normalizer, the empirical cumulative distribution function (ECDF) of the mean of the sample quantiles of an RMA background corrected reference dataset is used in place of the usually applied ECDF of the mean of the sample quantiles of the user supplied data [6]. To mimic the summarization procedure of RMA [32] the probe effects estimated by median polish for the same reference data is subtracted all probes of the user data. The RMA pre-processed expression value for each probe-set is then estimated as the median of the associated probes. For more detail on our one-by-one normalization approach see Supplementary Section S3. Finally, before classification the median of each probe-set in the RMA reference dataset is subtracted from the corresponding probe-set in the user data, since the classifiers were trained on median centered data.

The one-by-one normalization has the implicit assumption that the samples and reference follow the same distribution. This assumption might be violated by batch effects arising from differences in laboratory specific sample preparations and can cause severe bias in the normalization; to accomodate this `hemaClass.org` allows users to upload their own RMA reference dataset prepared under similar conditions. In the current study a laboratory specific RMA reference is referred to as an InLab reference. InLab references were simulated by selecting a random subset of 30 samples from each cohort. Samples were also one-by-one RMA normalized using the LLMPP CHOP dataset as an external reference; this is referred to as ExLab reference normalization.

2.5 Classification methods

2.5.1 Elastic nets

Logistic and multinomial regression were used in all classification methods available at `hemaClass.org`. However, in GEP experiments, the number of probe-sets present on the microarray always outnumbers the sample size. Collinearity present among the features further aggravates the problem of identifying genes responsible for the underlying biological mechanism. Regression under these ill-posed circumstances is typically handled by so-called regularization. Here the elastic net penalty [24, 57], which is a combination of the Lasso [53] and ridge regression [28], was used. Similar to the Lasso, this penalty ensures simultaneous variable selection and model estimation by forcing small coefficients to be zero, yielding sparse solutions, but contrary to the Lasso the elastic net penalty is capable of selecting more variables than samples.

The elastic net penalty contains two parameters α and λ. The parameter α interpolates the elastic net penalty between the ridge and the Lasso penalty which corresponds to values of 0 and 1, respectively. The parameter λ determines the amount of shrinkage of the coefficients with larger values inducing more shrinkage until no variables are contained in the model. Regularized logistic and multinomial regression were performed with the R-package **glmnet** [24].

2.5.2 ABC/GCB classification

The ABC/GCB classifier was established using logistic regression with an elastic net penalty on the LLMPP CHOP cohort. Of the 181 patients 74 were ABC, 76 were GCB, and 31 were non-classified. Using the 150 patients classified as either ABC or GCB, a dichotomous classifier capable of assigning each sample an estimate of the probability of being ABC was established.

To avoid over-fitting and limit the number of noise contributing genes, the elastic net parameters α and λ were chosen through 10 fold cross-validation. The parameter α was varied between 0.1 and 1 with step size 0.025 and $\log(\lambda)$ was varied between -10 and 2 with step size 0.06. The optimal combination of the parameters, and thereby the number of probe-sets, was found at the values minimizing the deviance. The results of the cross validations are shown in Supplementary Figure S1. The minimum deviance of 0.13 was attained at $\alpha = 0.15$ and $\log(\lambda) = -7.29$. This resulted in a gene expression classifier consisting of 381 probe-sets corresponding to 273 Ensembl Gene IDs.

When a tumour sample was classified according to the ABC/GCB classifier using cohort or InLab one-by-one normalized data the associated gene expressions are rescaled probe-set wise by the standard deviation of the LLMPP CHOP data divided by the standard deviation of the cohort or InLab reference data. For ExLab one-by-one normalization the data was used directly, since the training data for the ABC/GCB classifier was the same as the ExLab reference in the current study.

2.5.3 Wright ABC/GCB Classification

Standard GEP ABC/GCB classification is done using Wright's naive Bayes classifier. This method is not included on `hemaClass.org`, but is used in the current study for comparison of results from our elastic net classifier.

Bayesian compound covariate classification [55] with probeset list, weights and prior probabilities as described by Lenz et al. [38] was used to perform ABC/GCB classification (specific details obtained by personal communication with George Wright). In addition to this the probesets were brought to the same scale as Lenz et al.'s [38] probesets by a rescaling of the probeset-wise standard deviation.

2.5.4 REGS classification

In the paper by Falgreen et al. [22] REGS classifiers were established for prediction of resistance to the drugs C, H, and O. The classifiers were established on BCELL26 using regularized logistic regression analogous to the procedure described for the aforementioned ABC/GCB classifier. The number of microarray probes and corresponding genes for each of the REGS classifiers is shown in Table S5.

The probability of resistance to the combination therapy, p_{CHO}, was estimated based on the probabilities of drug resistance toward each of the three drugs: P_C, P_H, and P_O, respectively. This probability is calculated as the posterior probability of being resistant, given resistance towards each of the individual drugs under the assumption of conditional independence and uniform priors. The formula is also known as Graham's formula:

$$P_{CHO} = \frac{P_C P_H P_O}{P_C P_H P_O + (1 - P_C)(1 - P_H)(1 - P_O)}.$$

Derivation of the formula is shown in Supplementary Section S2. If a drug is left out in the combination therapy the drug is simply removed from the formula. This appoach to resistance to the combination therapy was used in Falgreen et al. [22].

When a tumour sample is classified according to the REGS classifiers the associated gene expressions are rescaled probe-set wise by the standard deviation of DLBCL14 divided by the standard deviation of the cohort, InLab, or Exlab RMA reference dataset.

Resistance classifiers for other chemotherapeutic drugs and diseases are also available on `hemaClass.org`, though established elsewhere [4, 5, 36]. The Rituximab sensitivity classifier of Laursen et al. [36] and Laursen et al. [37] uses an elastic net approach, as above, but with three classes. The Melphalan sensitivity classifier of Bøgsted et al. [4] uses sparse partial least squares to classify samples as either "sensitive", "intermediate" or "resistant". This classifier was developed for multiple myeloma (MM) patients and was thus based on other data [4].

2.5.5 BAGS classification

The BAGS classifier established by Dybkær et al. [18] was based on multinomial regression regularized by an elastic net penalty. The classifier was trained on the Tonsil dataset in a manner similar to the ABC/GCB classifier. The BAGS classifier uses 327 probes corresponding to 205 Ensembl Gene IDs.

When a tumour sample is classified according to the BAGS classifier the associated gene expressions are rescaled probe-set wise by the standard deviation of the Tonsil data divided by the standard deviation of the cohort, InLab, or ExLab one-by-one reference dataset. The rescaling is performed to make the data comparable to the Tonsil dataset.

2.6 Inter-method reproducibility assessments

To evaluate the reproducibility of the class probabilities obtained through cohort or reference based RMA normalization, Pearson's correlation coefficient for the logit-transformed probabilities and 95% confidence interval (CI) were calculated for each classifier and dataset. The identity and *total* least square regression lines were compared to assess bias in the estimated probabilities [11]. Total least squares regression was used as errors are present in both classification probabilities.

For each classifier the associated categories were obtained by thresholding the estimated probabilities. The ABC/GCB classifier was thresholded by 0.1 and 0.9, i.e. a tumour sample was classified as ABC when the estimated probability exceeded 0.9, GCB when it was below 0.1, and unclassified otherwise. For the BAGS classifier a tumour was classified as the class N, CB, CC, M, or PB with the highest probablity, if the associated probability exceeded 0.5 and unclassified when this threshold was not met for any subtype. For the REGS classifiers, C, H, O, and CHO combined, the thresholds were the 33% and 66% percentile of the estimated probabilities. The classifiers were applied to datasets using cohort, InLab, and ExLab one-by-one RMA normalization. Confusion matrices tabulating classifications from cohort normalized data versus InLab or Exlab normalized data were created, and from these the Accuracy (percent with similar classification to cohort), Cohen's weighted κ, and corresponding 95% CIs were computed to assess the agreement between the determined classes.

3 Results

3.1 Using `hemaClass.org`

The website is an easy-to-use, interactive interface for the **hemaClass** package with the desired RMA normalization and the classification methods selected by the user. The usage of the website is largely self-explanatory with context-dependent boxes aiding users with further information, warnings, or errors. A comprehensive tutorial and guide to both the website and package is provided on the website or by running `vignette("howto")` in R. Uploaded patient samples are normalized and classified depending on settings chosen by the user.

3.2 ABC/GCB classification

In order to classify patients as ABC/GCB based on the implemented one-by-one normalization method a classifier based on the regularised logistic regression was established. The classifications were evaluated in the four clinical cohorts CHEPRETRO, MDFCI, IDRC, and LLMPP R-CHOP, which have all been classified according to Wright's naive Bayes classifier [18, 38, 55]. The rates of agreement between the two classifiers based on cohort normalized data are shown in Table 2. Note that unclassified samples were included in the estimation of this rate i.e. a patient classified as ABC by one classifier but unclassified by the other is considered an error. The table also includes the alternative measure of agreement, Cohen's weighted κ, where misclassifications involving the unclassified group are weighted by 1/2. High agreement between the two classifiers are observed for CHEPRETRO and LLMPP R-CHOP, while the accuracy and Cohen's weighted κ are lower for IDRC and MDFCI. The accompanying confusion matrices are shown in the first rows of Supplementary Table S1.

The logit probability of being ABC estimated using the established cohort-based classifier was compared to the corresponding estimate obtained through the ExLab one-by-one normalized classification scheme in Figure 2A for CHEPRETRO. The probabilities estimated through the two methods are very similar, but values from ExLab normalization are slightly uncalibrated (or biased) and skewed downwards, indicating that different cut points might be used for the classifications. For InLab one-by-one normalization this error and bias is greatly minimized as shown in Figure 2B.

For both methods, patients are classified as ABC when the estimated probability exceeds 0.9 and GCB when it is below 0.1. In Table 3 the resulting classifications for the four validation datasets are compared in terms of accuracy and Cohen's weighted κ for cohort, using either Wrights Bayes classifier or the elastic net classifier, against InLab or ExLab RMA reference normalization. For ExLab normalization CHEPRETRO and LLMPP R-CHOP both show a high Cohen's weighted κ and accuracy considering that misclassifications involving unclassified samples count as errors, while values are moderate for MDFCI and IDRC when comparing to cohort based classifications for both the elastic net and Wrights classifier. The reduced rate of agreement and Cohen's weighted κ using ExLab one-by-one normalization in IDRC may be due to the samples being FFPE although this seems to be remedied by InLab one-by-one normalization. Accuracy and Cohen's weighted κ are very high when comparing InLab based classifications to cohort based classifications for the elastic net classifier, but values are still moderate for MDFCI when comparing against Wright classifications. The accompanying confusion matrices for the elastic net classifier are shown in the lower part of Supplementary Table S1. Note that changes in predicted classes are mainly due to shifts into NC from ABC or GCB. Direct disagreements between the classifiers are seemingly rare and only occurs in the IDRC dataset.

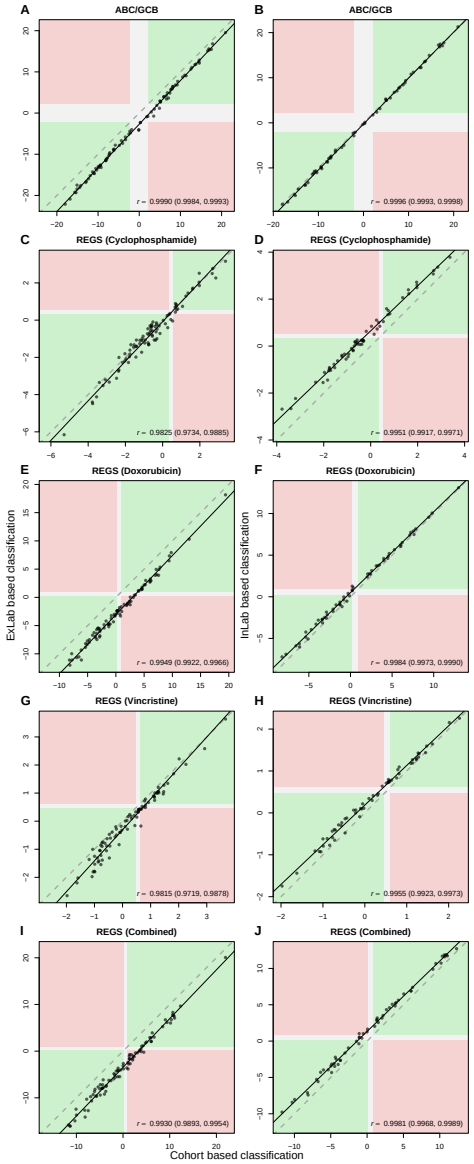

Figure 2: Comparison of logit probabilities for the ABC/GCB and REGS classifiers obtained through InLab or Exlab one-by-one normalization against cohort normalization. The areas marked with green indicate patients with similar classification between cohort based normalization and ExLab one-by-one normalization (A, C, E, G, I), or InLab one-by-one normalization (B, D, F, H, J) . The areas marked with red indicate complete misclassifications. For ABC/GCB and REGS the white areas indicate unclassified and intermediate sensitivity, respectively, in at least one of the classifiers. The dashed and solid line show the identity and total least squares line, respectively.

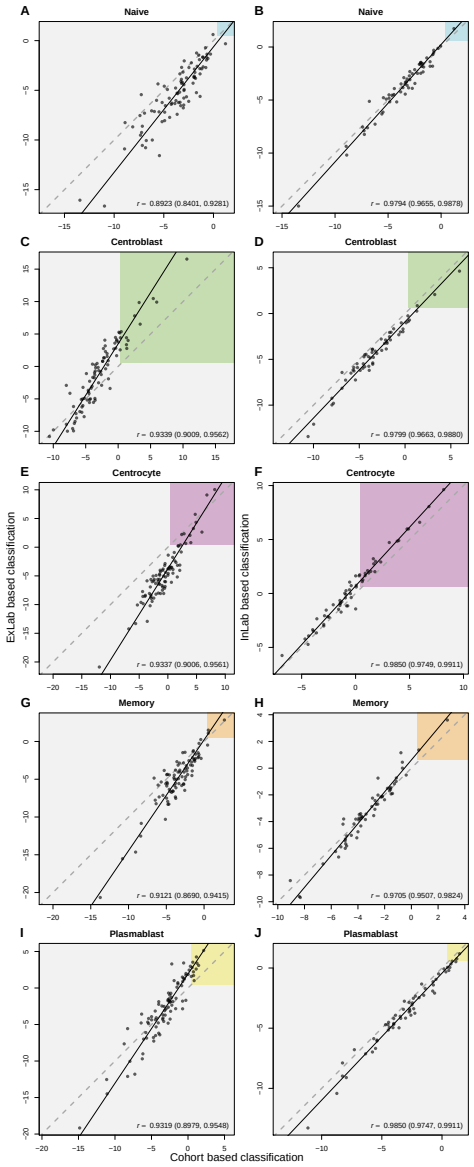

Figure 3: Comparison of logit probabilities for the BAGS classifier obtained through ExLab or InLab one-by-one normalization against cohort normalization. The coloured regions in the figure correspond to a threshold probability of 0.5. The dashed and solid line show the identity and total least squares line, respectively.

Table 2: Comparison of ABC/GCB classification performed using Wright's naive Bayes classifier [55] and the established elastic net classifier both based on cohort normalization. The second column shows the accuracy of the classifiers with 95% CI. The third column shows the Cohen's weighted κ with 95% CI.

Dataset	Accuracy	Cohen's κ
CHEPRETRO	0.94 (0.87, 0.98)	0.94 (0.88, 1.00)
MDFCI	0.78 (0.68, 0.86)	0.77 (0.66, 0.88)
IDRC	0.82 (0.79, 0.86)	0.80 (0.76, 0.85)
LLMPP R-CHOP	0.91 (0.86, 0.94)	0.90 (0.85, 0.96)

3.3 REGS classification

The probability of sensitivity towards each of the three drugs C, H, and O was estimated using `hemaClass.org` for both InLab and ExLab one-by-one normalization. The logit probabilities of sensitivity are plotted against those obtained by cohort based normalization in Figure 2 for CHEPRETRO data. Panels C, E, G, and I show the plots based on ExLab normalization and panels D, F, H, and J show the plots for the InLab based normalization. The probabilities obtained by Exlab and cohort based normalization are comparable, but similar to the other classifiers ExLab normalization leads to slightly skewed and biased probabilities, indicating that different cut points should be considered. The probabilities obtained by the InLab one-by-one normalization resembles the cohort based to a great extent, indicating that similar well-calibrated probabilities are obtainable for different laboratories by supplying an InLab reference set.

Based on the estimated probabilities, the patients were categorised as sensitive, intermediate, or resistant based on the thresholds specified in Section 2.5.4. The classes obtained by `hemaClass.org` are compared to those obtained by the cohort based approach in Table 3 in terms of rate of agreement and Cohen's weighted κ. Low to moderate rates of agreement are observed for the ExLab normalized data, and again the InLab one-by-one approach yielded higher agreement with classifications obtained from cohort based normalization The associated confusion matrices for InLab and Exlab one-by-one normalization are shown in Supplementary Table S3 and Supplementary Table S4, respectively. .

3.4 BAGS classification

The BAGS classifier was evaluated in a manner similar to the ABC/GCB classifier. The logit probability of a patient's tumour originating from one of the five subpopulations was estimated by means of cohort, InLab, and ExLab one-by-one normalization. The logit probabilities estimated by the ExLab normalization are plotted against logit probabilities from cohort based normalization in Figure 3 panels A, C, E, G, and I for CHEPRETRO. The correlations between the logit probabilities are highly significant, but also skewed and biased. The logit probabilities estimated using InLab one-by-one normalization are plotted against logit probabilities for the cohort based normalization in Figure 3 panels B, D, F, H, and J for CHEPRETRO. The InLab one-by-one normalization removes much of the aforementioned bias.

Based on the probabilities estimated by means of each of the three normalization methods the patients of the four clinical cohorts are grouped into the BAGS. The rates of agreement between the cohort based, and InLab or ExLab based classifications are shown in Table 3 along with Cohen's weighted κ.For the BAGS classifier low rates of agreement were also found when comparing results from ExLab normalization with cohort normalization, while InLab normalization again led to improved agreement. The associated confusion matrices are shown in Supplementary Table S2.

4 Discussion

Despite the enormous amount of resources spent on developing molecular based cancer classification systems, most of these are still not available in daily clinical practice. To allow for fast validation of our recent findings

Table 3: Comparison of classifications obtained using cohort based normalization against Exlab and InLab reference based normalization. The classifications are compared in terms of accuracy, Cohen's weighted κ, and Pearson's correlation coefficient r all supplied with 95% CIs. The comparisons in the first and last three columns are based on the ExLab and InLab reference based normalization method, respectively. For ABC/GCB classification, results from InLab or Exlab classification with the elasitic net classifier is compared against ABC/GCB classes for cohort normalized data obtained using both Wrights Bayes classifier and the elastic net classifier

	ExLab RMA pre-processing			**InLab RMA pre-processing**		
	Accuracy	Cohen's κ	Pearson's r	Accuracy	Cohen's κ	Pearson's r
ABC/GCB (Wright)						
CHEPRETRO	.89 (.80, .94)	.89 (.79, .98)	-	.97 (.88, 1.)	.97 (.90, 1.)	-
MDFCI	.63 (.52, .73)	.52 (.40, .64)	-	.72 (.59, .83)	.71 (.55, .86)	-
IDRC	.67 (.63, .71)	.62 (.56, .67)	-	.84 (.80, .87)	.82 (.77, .86)	-
LLMPP R-CHOP	.83 (.77, .87)	.82 (.74, .89)	-	.88 (.83, .92)	.88 (.82, .93)	-
ABC/GCB						
CHEPRETRO	.88 (.79, .94)	.87 (.78, .97)	.999 (.998, .999)	.98 (.91, 1.)	.98 (.93, 1.)	1. (.999, 1.)
MDFCI	.69 (.59, .78)	.68 (.53, .82)	.998 (.998, .999)	.98 (.91, 1.)	.98 (.85, 1.)	1. (.999, 1.)
IDRC	.65 (.61, .69)	.62 (.57, .68)	.986 (.983, .988)	.93 (.91, .95)	.93 (.90, .96)	.993 (.991, .994)
LLMPP R-CHOP	.82 (.77, .87)	.82 (.74, .89)	.999 (.999, .999)	.94 (.90, .97)	.94 (.90, .98)	.991 (.988, .993)
BAGS						
CHEPRETRO	.58 (.47, .69)	.56 (.28, .84)	-	.78 (.65, .88)	.74 (.33, 1.)	-
MDFCI	.54 (.43, .64)	.48 (.17, .79)	-	.80 (.68, .89)	.83 (.30, 1.)	-
IDRC	.52 (.47, .56)	.41 (.32, .50)	-	.79 (.75, .83)	.79 (.62, .96)	-
LLMPP R-CHOP	.56 (.49, .62)	.53 (.36, .70)	-	.88 (.82, .92)	.88 (.60, 1.)	-
REGS						
CHEPRETRO	.73 (.68, .78)	.71 (.64, .77)	.934 (.920, .946)	.84 (.79, .88)	.83 (.76, .89)	.992 (.990, .994)
MDFCI	.60 (.55, .65)	.55 (.48, .61)	.824 (.788, .855)	.90 (.86, .94)	.89 (.83, .96)	.997 (.996, .997)
IDRC	.52 (.49, .54)	.33 (.30, .36)	.660 (.635, .685)	.85 (.84, .87)	.84 (.81, .86)	.981 (.979, .983)
LLMPP R-CHOP	.58 (.54, .61)	.50 (.46, .54)	.810 (.786, .831)	.90 (.87, .92)	.89 (.85, .92)	.992 (.990, .993)

[18, 22], we have developed an easily accessible web application that permits other users to apply ABC/GCB, BAGS, and drug resistance classification to their own datasets. Since GEP classifiers rely on RMA normalized data, we also implemented a reference based RMA normalization that allows samples to be pre-processed one-by-one instead of in entire cohorts.

One-by-one normalization was done using both an external RMA reference (ExLab) and a mimicked laboratory specific reference (InLab). Classifications obtained through one-by-one pre-processing performed by `hemaClass.org` were then compared to those obtained using cohort based normalization in four clinical cohorts. The results showed that a one-by-one array analysis approach is feasible and performs comparably with the whole cohort based method when an InLab reference is used, while differences between ExLab and cohort normalized classifications were too large to be satisfactory. Users are thus encouraged to supply their own reference for RMA pre-processing. It seems that this approach allows for a realistic application of microarray based lymphoma classification for research projects, and after suitable standardisation and calibration, even for clinical use.

The poor results for ExLab one-by-one normalization and classication is likely caused by bias in the normalization process. To check for bias in the normalization `hemaClass.org` calculates the inter-quartile range (IQR) of the relative log expression (RLE) [8] across probes of one-by-one normalized arrays. Large values of the RLE IQR indicates bad arrays resulting from e.g. incorrect laboratory procedures for cohort normalized data, but can also indicate normalization against an improper reference for one-by-one normalized data, and can thus be used in cases with uncertainty of the validity of a user supplied reference. Based on the validation in Supplementary Section S6, we recommend removing samples from the analysis if the RLE IQR exceeds a value of 0.6. Users are, however, still encouraged to take proper precautions when selecting an RMA reference.

The present treatment algorithms for DLBCL are based on disease stage and clinical risk stratification without accounting for the underlying tumour-biology [48] and does not routinely account for the enormous variations in tumor biology between patients. The CHOP combination therapy (cyclophosphamide, doxorubicin, vincristine, and prednisone) has been the backbone of DLBCL therapy for decades with the

only significant improvement being the addition of monoclonal CD20 antibodies (Rituximab) [13]. Despite the addition of antibody therapy to conventional chemotherapy only 55% of patients with poor risk disease achieve durable remission [56]. Thus, the need for new therapeutic options in DLBCL is obvious. Currently a number of new drugs have shown promising activity in DLBCL, but their role outside clinical trials have not been defined. These drugs are different from conventional chemotherapeutic compounds by targeting specific deregulated cell-cycle pathways [23]. An important example is inhibition of the NF-κB pathway by proteasome inhibitors (i.e. bortezomib). Interestingly, the constitutive activation of the NF-κB pathway is characteristic for the ABC subtype of DLBCL which consequently enhances the effect of bortezomib in this subtype [17].

With the increasing number of new drugs likely to become available over the next years and the fact that their efficacy may vary between subsets of patients defined by gene expression profiles, the current treatment of patients based on disease stage and clinical information alone will not be sufficient. `hemaClass.org` provides an example of fast processing of complex molecular information in a way that is simple and readily at hand for clinicians.

An immediate limitation of the study is the need for a reference dataset for one-by-one RMA pre-processing established under the same conditions as the samples one wishes to classify. In these types of analyses one has to trust that the right tissue has been extracted and handled correctly through all the steps in the laboratory ending up with a reference array data set of sufficient quality, since in these types of analyses it is not possible to expect that the validation data (i.e. GEP from patient data) represent the training data (i.e. GEP from sorted normal tissue), as the training data is based on very different tissue as well as subpopulations sorted and profiled under very well controlled conditions. However, calibration of laboratory equipment is a well-known issue for many experimental techniques used in molecular biology like qPCR, mass spectrometry, immunohistochemistry, and flow cytometry. An important part of the calibration is that samples should be calibrated towards a dataset consisting of a representative set of tissue samples. The latter could be ensured by having a central tissue bank with officially approved data by e.g the international medical consortia for the specific diseases. Our reference data have for instance been controlled by looking at the frequency of ABC/GCBs, survival curves, and through tissue control by experienced pathologists

Alternatively the frozen RMA (fRMA) approach suggested by McCall et al. [41] could be used. This approach allows samples to be RMA normalized one-by-one against a frozen reference established across many different tissues and laboratories, taking the variation across laboratories and tissues for single probes into account. The current implementation of fRMA does not center and scale the data, so it cannot be used with the classifiers implemented in `hemaClass.org`, but had the training data for the classifiers been normalized with fRMA this might eliminate the need for centering and scaling sample data.

Another limitation of the current web application is that `hemaClass.org` only works with Affymetrix HG-U133 Plus 2.0 microarrays. This can, however, be circumvented by either re-annotation to HUGO Gene Nomenclature Committee (HGNC) approved symbols as suggested by Care et al. [9] or by re-annotated chip definition files as suggested by Dai et al. [16]. At the moment we are working on extending the web application to other array types along these lines. In the future, gene expressions will likely be measured using RNA-seq technology instead of microarrays. By summarizing the expression levels at gene-annotations rather than affymetrix probes and scaling the data it might also be possible to use microarray based classifiers with RNA-seq data. Alternatively the transcriptome, for the training data used for establishing the BAGS and REGS classifiers, would have to be measured with RNA-seq and the classifiers retrained.

Traditionally ABC/GCB classification has been achieved using the naive Bayes classifier of Wright et al. [55] which is based on cohorts, MAS5.0 normalized arrays, and a Bayesian approach assuming an equal amount of ABC and GCB patients. However, a classifier based on logistic regression regularized by an elastic net penalty was implemented to make the classification more adaptable to RMA normalization and one-by-one processing. This classifier proved to be quite comparable with the naive Bayes classifier over the four studied datasets confirming the strong and stable signal of the ABC/GBC subclasses of DLBCL.

Under the validation of the one-by-one method one should notice that the unclassified is treated as a class in its own right. This implies a lower accuracy compared to an approach where the unclassified are left out of the validations. The latter approach seems reasonable as changing classifications to unclassified is less

serious than changing real classes. Despite the disputed properties of Cohen's κ the conservative approach is retained and the issue is addressed using a Cohen's weighted κ approach. Given that an idealised approach is problematic to formulate, readers are encouraged to consider the confusion matrices in the supplementary material to make an overall evaluation of the performance.

ABC/GCB, BAGS, and REGS are only a part of the GEP-based armamentarium of methods for stratifying lymphoma patients into risk groups [39, 40, 50] and it would be interesting to extend the tool to include other classification systems. For a comprehensive review see [15]. To our knowledge only a few other classification methods have been made easily accessible as either web or desktop applications. Hopefully, this research will inspire bioinformaticians and statisticians to make their cancer classification methods easily accessible for usage, speedy validation, critical reviews, and mutual inspiration.

5 Conclusion

Although high throughput technologies in molecular biology have been around for almost two decades, only a few of the numerous biomarkers identified have undergone extensive validation and made it into the clinic [12]. It is our hope that making our own findings publicly available in this way will speed up validation and testing of BAGS and REGS by other researchers without having to delve into extensive bioinformatics implementations. Although `hemaClass.org` is still separated from the clinic we believe that a web based tool and suggestion for a clinical reference sample will bring cancer classification closer to the clinic. Hopefully, this work can also spawn interesting discussions on the clinical requirements of GEP based diagnostic and prognostic tools.

All material for reproducing this paper and its results is found at `https://github.com/oncoclass/hemaclass-paper`. Comments, suggestions, bug reports, and other issues are warmly welcome at `https://github.com/oncoclass/hemaclass/issues` or by mail to the corresponding author.

Acknowledgements

We thank Mads Boye and Bo Nygaard Bai at IT Services, Aalborg University, for their assistance on deploying the public server. SF is supported by a Mobility PhD fellowship at the Graduate School of Health, Faculty of Health Sciences, Aarhus University. The research is supported by MSCNET, a translational programme studying cancer stem cells in multiple myeloma supported by the EU FP6; CHEPRE, a programme studying chemo sensitivity in malignant lymphoma by genomic signatures supported by the Danish Agency for Science, Technology, and Innovation, and the National Experimental Therapy Partnership (NEXT), which is financed by a grant from Innovation Fund Denmark, as well as the Karen Elise Jensen Foundation. The technical assistance from Ann-Maria Jensen, Louise Hvilshøj Madsen, and Helle Høholt is greatly appreciated.
Conflict of interest statement: None declared.

References

[1] Ash A Alizadeh, M B Eisen, R E Davis, C Ma, I S Lossos, A Rosenwald, J C Boldrick, H Sabet, T Tran, X Yu, J I Powell, L Yang, G E Marti, T Moore, J Hudson, L Lu, D B Lewis, R Tibshirani, G Sherlock, W C Chan, T C Greiner, D D Weisenburger, J O Armitage, R Warnke, R Levy, W Wilson, M R Grever, J C Byrd, D Botstein, P O Brown, and L M Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000. doi: 10.1038/35000501.

[2] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, A Yefanov, H Lee, N Zhang, CL Robertson, N Serova, S Davis, and A Soboleva. NCBI GEO: Archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1):D991–D995, 2013. doi: 10.1093/nar/gks1193.

[3] Anders Ellern Bilgrau and Steffen Falgreen Larsen. *DLBCLdata: Automated and reproducible download and preprocessing of DLBCL data*, 2015. URL `http://github.com/AEBilgrau/DLBCLdata`. R package version 1.0.

[4] Martin Bøgsted, Johanne M Holst, Kirsten Fogd, Steffen Falgreen, Suzette Sørensen, Alexander Schmitz, Anne Bukh, Hans E Johnsen, Mette Nyegaard, and Karen Dybkær. Generation of a predictive melphalan resistance index by drug screen of B-cell cancer cell lines. *PLOS ONE*, 6(4):e19322, 2011. doi: 10.1371/journal.pone.0019322.

[5] Martin Bøgsted, Anders E Bilgrau, Christopher P Wardell, Uta Bertsch, Alexander Schmitz, Julie S Bødker, Malene K Kjeldsen, Hartmut Goldschmidt, Gareth J Morgan, Karen Dybkær, et al. Proof of the concept to use a malignant B cell line drug screen strategy for identification and weight of melphalan resistance genes in multiple myeloma. *PLOS ONE*, 8(12):e83252, 2013. doi: 10.1371/journal.pone.0083252.

[6] Benjamin M Bolstad, Rafael A. Irizarry, M Astrand, and Terence P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003. doi: 10.1093/bioinformatics/19.2.185,.

[7] Benjamin Milo Bolstad. *Low-level Analysis of High-Density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkeley, 2004.

[8] B.M. Bolstad, F. Collin, K.M. Simpson, R.A. Irizarry, and T.P. Speed. Experimental Design and Low-Level Analysis of Microarray Data. *International Review of Neurobiology*, 60:25–58, 2004.

[9] Matthew A Care, Sharon Barrans, Lisa Worrillow, Andrew Jack, David R Westhead, and Reuben M Tooze. 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*, 8(2):e55895, January 2013. doi: 10.1371/journal.pone.0055895.

[10] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. ***shiny**: Web Application Framework for R*, 2015. URL `http://CRAN.R-project.org/package=shiny`. R package version 0.12.2.

[11] S Chen, S A Billings, and W Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, November 1989. doi: 10.1080/00207178908953472.

[12] Xiaoming Chen, Roland Andersson, William Cs Cho, David Christiani, Richard Coico, Jeffery Drazen, Markus Ege, Thomas Fehniger, Hongwei Gao, Kunlin Jin, Michael N Liebman, Elena Lopez, Giuseppe Marraro, Gyorgy Marko-Varga, Francesco M Marincola, Laurentiu M Popescu, Claudio Spada, Aamir Shahzad, Ena Wang, Wei Wang, Xiangdong Wang, Yong-Xiao Wang, Jinglin Xia, and Jia Qu. The international effort: Building the bridge for translational medicine: Report of the 1st international conference of translational medicine (ICTM). *Clinical and Translational Medicine*, 1(1):1–15, January 2012. doi: 10.1186/2001-1326-1-15.

[13] Bertrand Coiffier, Eric Lepage, Josette Briere, Raoul Herbrecht, Hervé Tilly, Reda Bouabdallah, Pierre Morel, Eric van den Neste, Gilles Salles, Philippe Gaulard, Felix Reyes, Pierre Lederlin, and Christian Gisselbrecht. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *New England Journal of Medicine*, 346(4):235–242, January 2002. doi: 10.1056/NEJMoa011795.

[14] R Core Team. R: A language and environment for statistical computing, 2015. URL `http://www.R-project.org/`.

[15] Rita Coutinho and John Gribben. Biomarkers of diffuse large B-cell lymphoma: Impact on diagnosis, treatment, and prognosis. *Current Biomarker Finding*, 3:17–34, 2013.

[16] Manhong Dai, Pinglang Wang, Andrew D Boyd, Georgi Kostov, Brian Athey, Edward G Jones, William E Bunney, Richard M Myers, Terry P Speed, Huda Akil, Stanley J Watson, and Fan Meng. Evolving Gene/Transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, 33(20):e175, January 2005. doi: 10.1093/nar/gni179.

[17] Kieron Dunleavy, Stefania Pittaluga, Myron S Czuczman, Sandeep S Dave, George Wright, Nicole Grant, Margaret Shovlin, Elaine S Jaffe, John E Janik, Louis M Staudt, et al. Differential efficacy of bortezomib plus chemotherapy within molecular subtypes of diffuse large B-cell lymphoma. *Blood*, 113 (24):6069–6076, 2009. doi: http://dx.doi.org/10.1182/blood-2009-01-199679.

[18] K. Dybkær, M. Bøgsted, S. Falgreen, J. S. Bødker, M. K. Kjeldsen, A. Schmitz, A. E. Bilgrau, Z. Y. Xu-Monette, L. Li, K. S. Bergkvist, M. B. Laursen, M. Rodrigo-Domingo, S. C. Marques, S. B. Rasmussen, M. Nyegaard, M. Gaihede, M. B. Møller, R. J. Samworth, R. D. Shah, P. Johansen, T. C. El-Galaly, K. H. Young, and H. E. Johnsen. Diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. *Journal Of Clinical Oncology*, 33(12):1379–1388, 2015. doi: 10.1200/JCO.2014.57.7080.

[19] Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer-Verlag, New York, 1st edition, 2013. doi: 10.1007/978-1-4614-6868-4.

[20] Dirk Eddelbuettel and Romain François. `Rcpp`: Seamless R and `C++` integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08.

[21] Dirk Eddelbuettel and Conrad Sanderson. `RcppArmadillo`: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, March 2014. doi: http://dx.doi.org/10.1016/j.csda.2013.02.005.

[22] Steffen Falgreen, Karen Dybkær, Ken H Young, Zijun Y Xu-Monette, Tarec C El-Galaly, Maria B Laursen, Julie S Bødker, Malene K Kjeldsen, Alexander Schmitz, Mette Nyegaard, Hans Erik Johnsen, and Martin Bøgsted. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC Cancer*, 15(235):1–15, 2015. doi: 10.1186/s12885-015-1237-6.

[23] Jonathan W Friedberg. New strategies in diffuse large B-cell lymphoma: Translating findings from gene expression analyses into clinical practice. *Clinical Cancer Research*, 17(19):6112–6117, October 2011. doi: 10.1158/1078-0432.CCR-11-1073.

[24] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal Statistical Software*, 33(1):1–24, 2010. doi: 10.18637/jss.v033.i01.

[25] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. **affy**–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004. doi: 10.1093/bioinformatics/btg405.

[26] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, January 2004. doi: 10.1186/gb-2004-5-10-r80.

[27] Christine P Hans, Dennis D Weisenburger, Timothy C Greiner, Randy D Gascoyne, Jan Delabie, German Ott, H Konrad Müller-Hermelink, Elias Campo, Rita M Braziel, Elaine S Jaffe, Zenggang Pan, Pedro Farinha, Lynette M Smith, Brunangelo Falini, Alison H Banham, Andreas Rosenwald, Louis M Staudt, Joseph M Connors, James O Armitage, and Wing C Chan. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, 103(1):275–282, January 2004. doi: 10.1182/blood-2003-05-1545.

[28] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

[29] Daniel Holder, Richard F Raubertas, V Bill Pikounis, Vladimir Svetnik, and Keith Soper. Statistical analysis of high density oligonucleotide arrays: A safer approach. In *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data*, 2001.

[30] Liang-Tsung Huang. An integrated method for cancer classification and rule extraction from microarray data. *Journal of Biomedical Science*, 16(1):1–25, January 2009. doi: 10.1186/1423-0127-16-25.

[31] Rob J Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.

[32] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003. doi: 10.1093/nar/gng015.

[33] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, U Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003. doi: 10.1093/biostatistics/4.2.249.

[34] Elaine S Jaffe. The 2008 WHO classification of lymphomas: Implications for clinical practice and translational research. *Hematology*, 2009(1):523–531, January 2009. doi: 10.1182/asheducation-2009.1.523.

[35] Simon Katz, Rafael A Irizarry, Xue Lin, Mark Tripputi, and Mark W Porter. A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC bioinformatics*, 7:464, 2006.

[36] Maria Bach Laursen, Steffen Falgreen, Julie Støve Bødker, Alexander Schmitz, Malene Krag Kjeldsen, Suzette Sørensen, Jakob Madsen, Tarec Christoffer El-Galaly, Martin Bøgsted, Karen Dybkær, et al. Human B-cell cancer cell lines as a preclinical model for studies of drug effect in diffuse large B-cell lymphoma and multiple myeloma. *Experimental Hematology*, 42(11):927–938, 2014. doi: http://dx.doi.org/10.1016/j.exphem.2014.07.263.

[37] Maria Bach Laursen, Linn Rønlev Reinholt, Suzette Sørensen, Steffen Falgreen, Alexander Schmitz, Julie Støve Bødker, Tarec El-Galaly, Malene Krag Kjeldsen, Martin Bøgsted, Hans Erik Johnsen, and Karen Dybkær. Studies of anti-CD20 antibody mediated complement dependent cytotoxicity in a preclinical cell line model of diffuse large B-cell lymphoma. *In preparation*, 2016.

[38] G Lenz, G Wright, S S Dave, W Xiao, J Powell, H Zhao, W Xu, B Tan, N Goldschmidt, J Iqbal, J Vose, M Bast, K Fu, D D Weisenburger, T C Greiner, J O Armitage, A Kyle, L May, R D Gascoyne, J M Connors, G Troen, H Holte, S Kvaloy, D Dierickx, G Verhoef, J Delabie, E B Smeland, P Jares, A Martinez, A Lopez-Guillermo, E Montserrat, E Campo, R M Braziel, T P Miller, L M Rimsza, J R Cook, B Pohlman, J Sweetenham, R R Tubbs, R I Fisher, E Hartmann, A Rosenwald, G Ott, H-K Muller-Hermelink, D Wrench, T A Lister, E S Jaffe, W H Wilson, W C Chan, and L M Staudt. Stromal gene signatures in large-B-cell lymphomas. *New England Journal of Medicine*, 359(22):2313–23, November 2008. doi: 10.1056/NEJMoa0802885.

[39] Izidore S Lossos, Debra K Czerwinski, Ash A Alizadeh, Mark A Wechser, Rob Tibshirani, David Botstein, and Ronald Levy. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New England Journal of Medicine*, 350(18):1828–1837, April 2004. doi: 10.1056/NEJMoa032520.

[40] Raquel Malumbres, Jun Chen, Rob Tibshirani, Nathalie A Johnson, Laurie H Sehn, Yaso Natkunam, Javier Briones, Ranjana Advani, Joseph M Connors, Gerald E Byrne, Ronald Levy, Randy D Gascoyne, and Izidore S Lossos. Paraffin-based 6-gene model predicts outcome in diffuse large B-cell lymphoma patients treated with R-CHOP. *Blood*, 111(12):5509–5514, June 2008. doi: 10.1182/blood-2008-02-136374.

[41] Matthew N. McCall, Benjamin M. Bolstad, and Rafael A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, 2010.

[42] Matthew N McCall, Peter N Murakami, Margus Lukk, Wolfgang Huber, and Rafael a Irizarry. Assessing affymetrix GeneChip microarray quality. *BMC bioinformatics*, 12(1):137, 2011.

[43] Stefano Monti, Bjoern Chapuy, Kunihiko Takeyama, Scott J Rodig, Yansheng Hao, Kelly T Yeda, Haig Inguilizian, Craig Mermel, Treeve Currie, Ahmet Dogan, Jeffery L Kutok, Rameen Beroukhim, Donna Neuberg, Thomas M Habermann, Gad Getz, Andrew L Kung, Todd R Golub, and Margaret a Shipp. Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B-cell lymphoma. *Cancer Cell*, 22(3):359–372, September 2012. doi: http://dx.doi.org/10.1016/j.ccr.2012.07.014.

[44] Christian Bjørn Poulsen, Rehannah Borup, Finn Cilius Nielsen, Niels Borregaard, Mads Hansen, Kirsten Grønbæk, Michael Boe Møller, and Elisabeth Ralfkiaer. Microarray-based classification of diffuse large B-cell lymphoma. *European Journal of Haematology*, 74(6):453–465, 2005. doi: 10.1111/j.1600-0609.2005.00429.x.

[45] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frdrique Lisacek, Jean-Charles Sanchez, and Markus Mller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

[46] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, Elaine M Hurt, Hong Zhao, Lauren Averett, Liming Yang, Wyndham H Wilson, Elaine S Jaffe, Richard Simon, Richard D Klausner, John Powell, Patricia L Duffey, Dan L Longo, Timothy C Greiner, Dennis D Weisenburger, Warren G Sanger, Bhavana J Dave, James C Lynch, Julie Vose, James O Armitage, Emilio Montserrat, Armando López-Guillermo, Thomas M Grogan, Thomas P Miller, Michel LeBlanc, German Ott, Stein Kvaloy, Jan Delabie, Harald Holte, Peter Krajci, Trond Stokke, and Louis M Staudt. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, June 2002. doi: 10.1056/NEJMoa012914.

[47] Conrad Sanderson. *`Armadillo`: An Open Source `C++` Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments.* Technical Report, NICTA, 2010. URL `http://arma.sourceforge.net`.

[48] H J Schmoll, E Van Cutsem, A Stein, V Valentini, B Glimelius, K Haustermans, B Nordlinger, C J van de Velde, J Balmana, J Regula, I D Nagtegaal, R G Beets-Tan, D Arnold, F Ciardiello, P Hoff, D Kerr, C H Köhne, R Labianca, T Price, W Scheithauer, A Sobrero, J Tabernero, D Aderka, S Barroso, G Bodoky, J Y Douillard, H El Ghazaly, J Gallardo, A Garin, R Glynne-Jones, K Jordan, A Meshcheryakov, D Papamichail, P Pfeiffer, I Souglakos, S Turhal, and A Cervantes. ESMO consensus guidelines for management of patients with colon and rectal cancer. A personalized approach to clinical decision making. *Annals of Oncology*, 23(10):2479–2516, October 2012. doi: 10.1093/annonc/mds236.

[49] Laurie H. Sehn and Randy D. Gascoyne. Diffuse large B-cell lymphoma: Optimizing outcome in the context of clinical and biologic heterogeneity. *Blood*, 125(1):22–32, 2014. doi: 10.1182/blood-2014-05-577189.

[50] Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo C T Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, Tane S Ray, Margaret a

Koval, Kim W Last, Andrew Norton, T Andrew Lister, Jill Mesirov, Donna S Neuberg, Eric S Lander, Jon C Aster, and Todd R Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, January 2002. doi: 10.1038/nm0102-68.

[51] Philippe Solal-Céligny, Pascal Roy, Philippe Colombat, Josephine White, Jim O Armitage, Reyes Arranz-Saez, Wing Y Au, Monica Bellei, Pauline Brice, Dolores Caballero, et al. Follicular lymphoma international prognostic index. *Blood*, 104(5):1258–1265, 2004. doi: http://dx.doi.org/10.1182/blood-2003-12-4434.

[52] The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-hodgkin's lymphoma. *New England Journal of Medicine*, 329(14):987–994, 1993. doi: 10.1056/NEJM199309303291402.

[53] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

[54] Carlo Visco, Yan Li, Z Y Xu-Monette, R N Miranda, T M Green, Y Li, A Tzankov, W Wen, W-m Liu, B S Kahl, E S G d'Amore, S Montes-Moreno, K Dybkær, A Chiu, W Tam, A Orazi, Y Zu, G Bhagat, J N Winter, H-Y Wang, S O'Neill, C H Dunphy, E D Hsi, X F Zhao, R S Go, W W L Choi, F Zhou, M Czader, J Tong, X Zhao, J H van Kriken, Q Huang, W Ai, J Etzell, M Ponzoni, A J M Ferreri, M A Piris, M B Møller, C E Bueso-Ramos, L J Medeiros, L Wu, and K H Young. Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: A report from the international DLBCL rituximab-CHOP consortium. *Leukemia*, 26(9):2103–2113, 2012. doi: 10.1038/leu.2012.83.

[55] George Wright, Bruce Tan, Andreas Rosenwald, Elaine H Hurt, Adrian Wiestner, and Louis M Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *PNAS*, 100(17):9991–9996, August 2003. doi: 10.1073/pnas.1732008100.

[56] Marita Ziepert, Dirk Hasenclever, Evelyn Kuhnt, Bertram Glass, Norbert Schmitz, Michael Pfreundschuh, and Markus Loeffler. Standard international prognostic index remains a valid predictor of outcome for patients with aggressive CD20+ B-cell lymphoma in the rituximab era. *Journal of Clinical Oncology*, 28(14):2373–2380, May 2010. doi: 10.1200/JCO.2009.26.2493.

[57] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, April 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

SUPPLEMENTARY MATERIAL

`hemaClass.org`: Online one-by-one microarray normalization and classification of hematological cancers for precision medicine

S1 Supplementary figures and tables

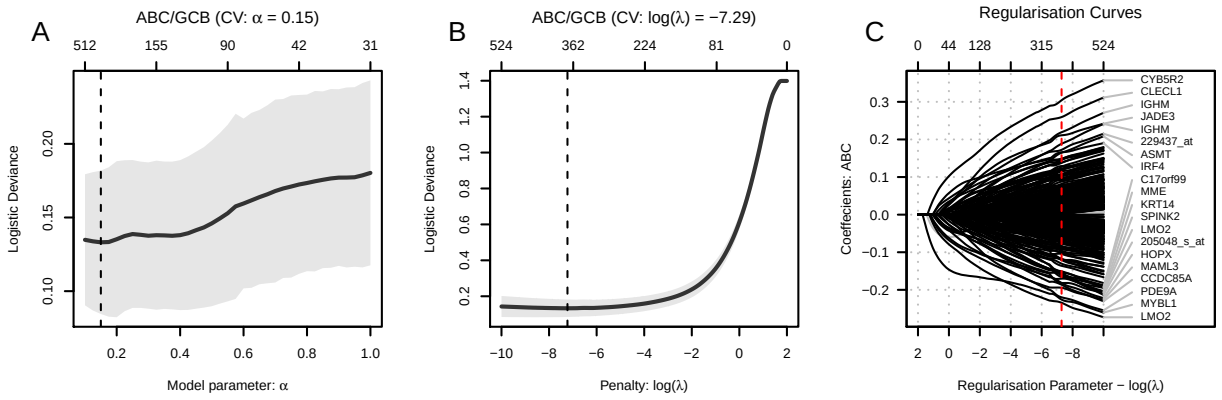

Figure S1: Ten fold cross validation for the parameters α and λ in a logistic regression regularized by elastic net. In panels A and B the deviance is plotted against the model parameter α and regularization parameter λ, respectively. In Panel C the regularization curves are shown. Black and grey curves represent selected and non-selected probe-sets, respectively. Positive and negative coefficients indicate that high expression values for the associated gene are related to ABC and GCB, respectively. The red line indicates the model chosen through 10 fold cross validation. The gene symbols for the 20 probe-sets associated with the largest absolute coefficients in the chosen gene expression predictors are displayed in Panel C.

Table S1: Confusion tables for the ABC/GCB classifiers. The columns represent cohort based normalisztion using the ABC/GCB classifier based on elastic net. The first part of the table compares Wright's method for ABC/GCB classification with the elastic net based. In the second and third part ExLab and InLab reference based normalization is compared to cohort based normalization using the ABC/GCB classifier based on elastic net.

	CHEPRETRO			**MDFCI**			**IDRC**			**LLMPP R-CHOP**		
	ABC	NC	GCB	ABC	NC	GCB	ABC	NC	GCB	ABC	NC	GCB
Wright's method												
ABC	38	2	0	28	14	0	188	24	1	90	3	0
NC	1	4	0	1	11	3	6	27	14	6	19	8
GCB	0	2	42	0	1	29	7	35	193	0	5	102
ExLab Normalization												
ABC	34	0	0	24	0	0	95	0	0	76	0	0
NC	5	2	0	6	4	0	102	19	0	20	6	0
GCB	0	6	42	0	22	35	4	67	208	0	21	110
InLab Normalization												
ABC	26	0	0	19	0	0	183	9	0	86	6	0
NC	0	5	0	0	19	0	6	63	7	0	13	4
GCB	0	1	27	0	1	22	0	9	188	0	2	92

Table S2: Confusion tables for the BAGS classifier. ExLab and InLab reference based normalization are shown in the columns and cohort normalization in the rows.

	ExLab normalization						**InLab normalization**					
	N	CB	CC	M	PB	UC	N	CB	CC	M	PB	UC
CHEPRETRO												
Naive	1	0	0	0	0	1	2	0	0	0	0	0
Centroblast	0	18	0	0	0	0	0	4	4	0	0	1
Centrocyte	0	10	11	0	5	9	0	0	25	1	0	0
Memory	0	0	0	3	0	1	0	0	0	2	0	0
Plasmablast	0	0	0	0	16	0	0	0	0	0	8	3
Unclassified	0	7	0	0	4	3	0	0	2	2	0	5
MDFCI												
Naive	1	1	0	1	2	3	3	0	0	1	0	2
Centroblast	0	18	0	0	1	0	0	9	0	0	0	3
Centrocyte	0	7	8	2	10	6	0	0	22	1	0	1
Memory	0	0	0	6	0	0	0	0	0	5	0	0
Plasmablast	0	0	0	0	11	0	0	0	0	0	7	0
Unclassified	0	1	0	1	7	5	1	0	0	2	1	3
IDRC												
Naive	0	0	3	0	6	4	12	0	0	0	1	0
Centroblast	0	16	27	0	27	22	2	62	2	0	5	12
Centrocyte	0	0	140	0	39	18	1	2	146	7	8	18
Memory	0	0	1	8	20	10	0	0	0	35	3	1
Plasmablast	0	0	1	0	74	4	0	0	0	1	76	1
Unclassified	0	0	14	0	44	17	7	0	0	9	16	38
LLMPP R-CHOP												
Naive	1	2	0	0	5	6	8	0	0	1	0	1
Centroblast	0	32	0	0	2	7	0	37	1	0	0	1
Centrocyte	0	5	54	0	18	12	0	1	66	1	1	4
Memory	0	0	0	5	13	4	0	0	0	22	0	0
Plasmablast	0	0	0	0	32	0	0	0	0	1	24	4
Unclassified	0	2	0	0	27	6	7	1	0	1	0	21

Table S3: Confusion tables for the REGS classifiers. ExLab normalization is shown in the rows and cohort normalization in the columns.

	CHEPRETRO			**MDFCI**			**IDRC**			**LLMPP R-CHOP**		
	Sen	Int	Res	Sen	Int	Res	Sen	Int	Res	Sen	Int	Res
Cyclophosphamide												
Sensitive	40	1	0	34	5	0	178	0	0	108	2	0
Intermediate	6	17	0	0	15	6	114	0	0	8	32	1
Resistant	0	3	22	0	1	30	203	0	0	0	15	67
Doxorubicin												
Sensitive	30	0	0	29	0	0	25	86	39	77	0	0
Intermediate	21	6	0	32	0	0	0	6	170	78	1	0
Resistant	0	14	18	6	12	12	0	0	169	13	43	21
Vincristine												
Sensitive	36	0	0	33	0	0	42	90	33	78	0	0
Intermediate	7	9	0	24	2	0	1	17	136	59	15	0
Resistant	1	10	26	1	15	16	1	3	172	11	36	34
Combined												
Sensitive	32	0	0	33	0	0	135	14	1	87	0	0
Intermediate	19	9	0	27	1	0	19	143	21	70	0	0
Resistant	0	13	16	3	13	14	0	27	135	16	42	18

Table S4: Confusion tables for the REGS classifiers. InLab normalization is shown in the rows and cohort normalization in the columns. Note, 30 samples were used as reference data and hence not present in this table.

	CHEPRETRO			**MDFCI**			**IDRC**			**LLMPP R-CHOP**		
	Sen	Int	Res	Sen	Int	Res	Sen	Int	Res	Sen	Int	Res
Cyclophosphamide												
Sensitive	13	10	0	26	4	0	134	32	0	89	5	0
Intermediate	0	7	10	0	10	1	3	77	29	0	27	9
Resistant	0	0	19	0	0	20	0	9	181	0	2	71
Doxorubicin												
Sensitive	18	2	0	19	0	0	132	7	0	50	15	0
Intermediate	0	14	2	0	21	0	24	143	3	0	55	13
Resistant	0	0	23	0	3	18	0	16	140	0	0	70
Vincristine												
Sensitive	18	5	0	16	6	0	127	32	0	71	0	0
Intermediate	0	10	1	0	8	9	12	83	46	9	49	0
Resistant	0	0	25	0	0	22	1	10	154	0	10	64
Combined												
Sensitive	19	3	0	23	1	0	125	14	0	64	12	0
Intermediate	0	11	5	0	16	0	12	148	14	0	46	10
Resistant	0	0	21	0	0	21	0	6	146	0	0	71

Classifier	nProbes	nHGNC	nEnsembl
ABC/GCB	381	291	273
BAGS	327	224	205
Vincristine Classifier	33	32	29
Vincristine Predictor	22	21	18
Cyclophosphamide Classifier	74	73	66
Cyclophosphamide Predictor	28	27	25
Doxorubicine Classifier	119	118	112
Doxorubicine Predictor	53	52	48
Combined Classifier	203	202	185
Combined Predictor	90	88	80

Table S5: Number of probes used in the classifiers and the number of corresponding HGNC and Ensembl gene IDs

S2 Graham's formula

This section derives Graham's formula which, in our context, yields the posterior probability of resistance to the combination of multiple drugs, given resistance to the individual drugs. For simplicity, the formula is derived for two drugs. The formula straightforwardly generalizes to three or more drugs.

Let C, H, and B be Bernoulli distributed random variables with probability parameter $1/2$, where $C = 1$ indicates resistance to Cyclophosphamide C, $H = 1$ indicates resistance to Doxorubicin H, and $B = 1$ indicates resistance to the combination of H and C. Conversely, C, H, and $B = 0$ indicate sensitivity towards C, H, and B, respectively. Under an assumption of conditional drug independence

$$\begin{aligned} P(C = 1, H = 1|B = 1) &= P(C = 1|B = 1)P(H = 1|B = 1), \text{ and} \\ P(C = 1, H = 1|B = 0) &= P(C = 1|B = 0)P(H = 1|B = 0) \end{aligned}$$

we have that

$$\begin{aligned} &P(B = 1|H = 1, C = 1) \\ &\quad = \frac{P(C = 1, H = 1, B = 1)}{P(C = 1, H = 1)} \\ &\quad = \frac{P(C = 1, H = 1|B = 1)P(B = 1)}{P(C = 1, H = 1, B = 1) + P(C = 1, H = 1, B = 0)} \\ &\quad = \frac{P(C = 1|B = 1)P(H = 1|B = 1)P(B = 1)}{P(C = 1, H = 1|B = 1)P(B = 1) + P(H = 1, C = 1|B = 0)P(B = 0)}, \end{aligned}$$

by the definition of conditional probabilities, the law of total probability, and the assumptions. From the distributional assumption on B, $P(B = 0) = P(B = 1) = 1/2$, and the above then simplifies to:

$$P(B = 1|H = 1, C = 1) = \frac{P(C = 1|B = 1)P(H = 1|B = 1)}{P(C = 1, H = 1|B = 1) + P(H = 1, C = 1|B = 0)}.$$

For notational convenience, we abbreviate $P_C = P(C = 1|B = 1)$, $P_H = P(H = 1|B = 1)$, $P_{CH} = P(B = 1|H = 1, C = 1)$. The distributional assumptions then imply:

$$\begin{aligned} P_{CH} &= \frac{P_C P_H}{P_C P_H + P(C = 1|B = 0)P(H = 1|B = 0)} \\ &= \frac{P_C P_H}{P_C P_H + P(B = 0|C = 1)P(B = 0|H = 1)} \\ &= \frac{P_C P_H}{P_C P_H + \big(1 - P(B = 1|C = 1)\big)\big(1 - P(B = 1|H = 1)\big)} \\ &= \frac{P_C P_H}{P_C P_H + (1 - P_C)(1 - P_H)}, \end{aligned}$$

which is the two-drug equivalent to the used formula.

S3 RMA normalization

Recall that ordinary robust multichip average (RMA) pre-processing consists of three steps: (1) Background adjustment, (2) quantile normalization, and (3) summarization of probes to probe-sets, see e.g. [32, 33]. For completeness we review ordinary cohort based RMA normalization.

S3.1 Background correction

In order to produce background adjusted probe intensities we will use the within array normal-exponential de-convolution scheme as implemented by the `rma.background.correct` command in the Bioconductor package `preprocessCores`, see [7, 32].

S3.2 Quantile normalization

Let x_{ijk} be the $\log_2$-transformed and background adjusted cohort data, where $i = 1, \ldots, I$ index the arrays of the cohort data, $j = 1, \ldots, J$ index the probe-sets, and $k = 1, \ldots, K_j$ index the probes nested within probe-sets.

Furthermore, let G_i denote the empirical cumulative distribution function (ECDF) of the probes $\{x_{ijk}\}_{jk}$ on the i'th cohort array and F the ECDF of the across array averaged sample quantiles $\{\bar{x}_{\cdot(jk)}\}_{ij}$, where $\{x_{i(jk)}\}_{jk}$ is the order statistic of all probes on the i'th cohort array based on the lexicographic ordering of the indices $\{jk\}$. Then each data point is quantile normalized in the following way

$$\tilde{x}_{ijk} = F^{-1}(G_i(x_{ijk})),$$

where F^{-1} is calculated as the quantiles of type 2 [31]. This step is performed by the `RMA_norm` function with option `generateQuan` equal to one in the **hemaClass** package.

S3.3 Summarization

For each probe-set j we let μ_{ij} represent the $\log_2$-scale expression level for array i and probeset j, α_{jk} the probe affinity effect, and the ϵ_{ijk}'s are independent identically distributed error terms with mean 0 and formulate the following linear additive model

$$\tilde{x}_{ijk} = \mu_{ij} + \alpha_{jk} + \epsilon_{ijk},$$

where $\sum_{k=0}^{n_j} \alpha_{jk} = 0$ for all probe-sets. The parameters are estimated by median polish [29]. The probe affinity estimates are denoted by $\hat{\alpha}_{jk}$.

The RMA normalized cohort data are then given by

$$\hat{x}_{ij} = \hat{\mu}_{ij}.$$

This step is performed by the `RMA_sum` function in the `hemaClass.org` package.

S4 One-by-one RMA normalization of user supplied data

S4.1 Background correction

The background correction in one-by-one RMA normalization is unaltered as it is already works in a one-by-one fashion.

S4.2 Quantile normalization

Let x_{ijk} be the $\log_2$-transformed and background corrected reference data, where $i = 1, \ldots, I_R$ index the arrays of the reference data, $j = 1, \ldots, J$ index the probe-sets, and $k = 1, \ldots, K_j$ index the probes. Assume x_{ijk} has been RMA normalized as described above. Similarly, let y_{ijk} be the log2-transformed and background corrected user supplied data, where $i = 1, \ldots, I_U$ index the arrays of the user supplied data, $j = 1, \ldots, J$ index the probe-sets, and $k = 1, \ldots, K_j$ index the probes. Furthermore, let H_i denote the ECDF of the user supplied data $\{y_{ijk}\}_{jk}$.

As quantile normalizer the ECDF of the background corrected reference data is used in place of the ususally applied ECDF of the mean of the sample quantiles

$$\tilde{y}_{ijk} = F^{-1}(H_i(y_{ijk})).$$

This step is performed by the `RMA_norm` function with options `generateQuan` equal to zero and `quantile` equal to the quantiles of the reference data in the **hemaClass** package.

S4.3 Summarization

To mimic the RMA summarization the probe effects estimated by median polish on the reference data is subtracted all probes of the user data

$$\hat{y}_{ijk} = \tilde{y}_{ijk} - \hat{\alpha}_{jk}.$$

The pre-processed expression value for each probe-set is then estimated as the median of the associated probes.

$$\hat{y}_{ij} = \underset{k \in \{1,\ldots,n_j\}}{\text{median}} \{\hat{y}_{ijk}\}.$$

S5 Classification

To ensure identical classification probabilities whether data is supplied as a cohort or one-by-one, we finally subtract the median of each probe-set in the reference from the corresponding probe-set and scale by the standard deviation of each probe-set in the reference data

$$(\hat{y}_{ij} - \hat{x}_{\cdot j})/s_{\cdot j},$$

where $\hat{x}_{\cdot j} = \text{median}_{i \in \{1,\ldots,I_R\}}\{\hat{x}_{ij}\}$ and $s_{\cdot j} = \text{sd}_{i \in \{1,\ldots,I_R\}}\{\hat{x}_{ij}\}$.

S6 Model control of one-by-one RMA normalization

Given the large difference in misclassifications between InLab and ExLab one-by-one normalized samples, we would like to be able to distinguish between the two in a setting where we are unsure how closely the RMA reference resembles the samples.

S6.1 Relative Log Expression

The relative log expression (RLE) is a quality measure for microarrays introduced by Bolstad et al. [8]. Using the notation from above we define the RLE for probe j on array i following RMA normalization as:

$$\text{RLE}(\hat{y}_{ij}) = \hat{y}_{ij} - \hat{x}_{\cdot j}$$

i.e. the difference between the estimated expression for probe j on array i and the median expression for probe j in the cohort. A non-zero median RLE across probes for an array thus indicates differences in the number of up- and downregulated genes, while a large interquartile range (IQR) indicates that most genes on a given array are differentially expressed [42]. Extreme values of these measures may be used to identify arrays with low-quality data. We propose that they may also be used to evaluate how well a sample resembles a given RMA reference following one-by-one normalization (how well it has been normalized), by substituting the cohort median with the RMA reference median.

S6.2 RLE for separation of InLab and ExLab RMA references

For each of the five datasets used in the current study a random subset of 30 samples were extracted and set as an InLab reference. The remaining samples for each dataset where then RMA normalized as a cohort, or one-by-one against the randomly selected InLab reference or against the four other datasets (ExLab). RLE values were calculated in all six scenarios and summarized as the absolute value of the median and IQR for each sample. These values are shown in panel A and B in Figure S2 to Figure S6. We find that the RLE values in the InLab one-by-one scenarios more closely resemble the values calculated from cohort normalization than the values from ExLab one-by-one normalization. Furthermore differences between InLab and ExLab RLE values are more pronounced for the IQR of RLE than for the median. ROC curves showing the ability of the RLE median or IQR to distinguish between an InLab or ExLab reference were calculated and plotted

using the pROC package version 1.8 [45] in R, as shown in panels C and D in Figure S2 to Figure S6. The area under the curve (AUC) for the ROC curves confirm the superiority of the RLE IQR for distinguishing between a "correct" reference and a "wrong" reference, i.e. in most cases we observe a higher AUC when using the IQR instead of the median RLE.

Using Youden's index, which maximizes the sum of sensitivity and specificity, we calculated the optimal threshold for each ROC curve based on the RLE IQR as shown in Table S6. Calculating the median RLE IQR across all datasets and one-by-one RMA references gives a value of 0.62. By rounding of we set a conservative threshold of 0.6 for reference normalized samples, i.e. if the RLE IQR exceeds this value the array is expected to have been normalized incorrectly.

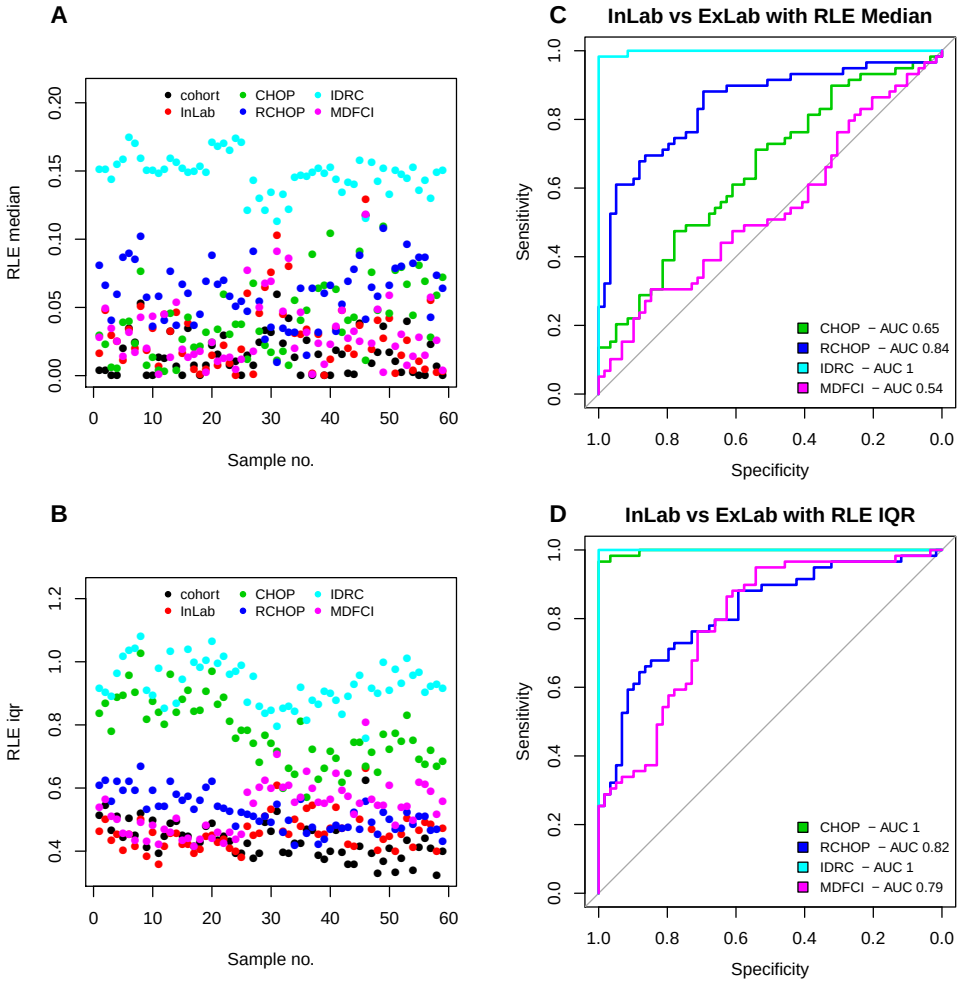

Figure S2: Absolute value of the median (A) and IQR (B) RLE values for different RMA normalizations of the CHEPRETRO dataset and ROC curves for using these values to separate between an InLab and Exlab RMA reference (C,D)

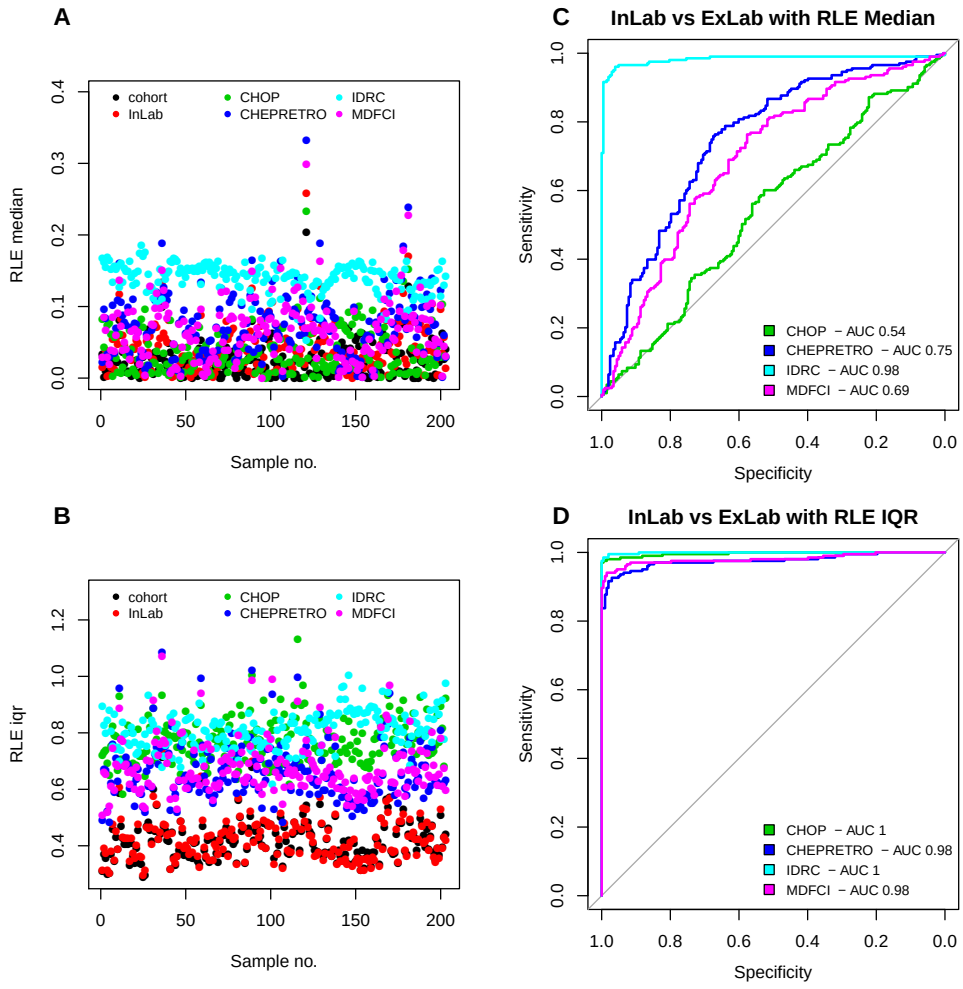

Figure S3: Absolute value of the median (A) and IQR (B) RLE values for different RMA normalizations of the RCHOP dataset and ROC curves for using these values to separate between an InLab and Exlab RMA reference (C,D)

S6.3 RLE IQR vs Classification accuracy

The results in Supplementary Section S6.2 showed that the RLE IQR can be used to determine if samples have been normalized against an Inlab or ExLab reference. In this section we compare the RLE IQR to the classification accuracy. The CHEPRETRO dataset was one-by-one normalized against an InLab reference, the LLMPP CHOP reference, and the LLMPP RCHOP reference, and the LLMPP RCHOP dataset was one-by-one normalized against an InLab reference, the LLMPP CHOP reference, and the CHEPRETREO reference. RLE IQR values were calculated and the proportion of samples below a given threshold and the accuracy (proportion of samples with similar classification in cohort normalized data) were calculated for increasing values of the RLE IQR. This was done for ABC/GCB, BAGS and REGS (CHO) classification. Results are shown in Figure S7 to Figure S12.

For CHEPRETRO we saw that most samples were retained at the suggested RLE IQR value of 0.6 when samples were normalized against the InLab reference, and a tendency towards higher classification accuracy for samples with low RLE IQR. When normalizing CHEPRETRO against RCHOP the RLE IQR value of

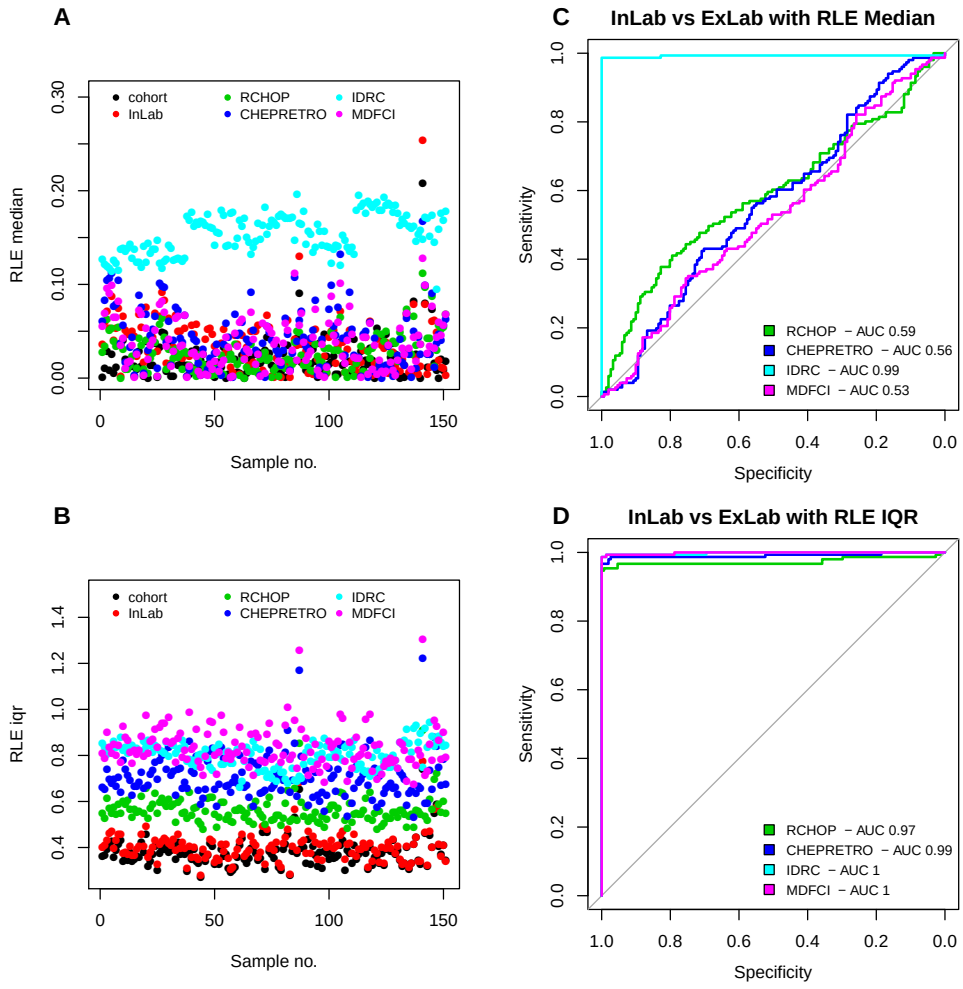

Figure S4: Absolute value of the median (A) and IQR (B) RLE values for different RMA normalizations of the CHOP dataset and ROC curves for using these values to separate between an InLab and Exlab RMA reference (C,D)

0.6 only excludes a small proportion of the samples and a low accuracy is observed for BAGS and REGS classification while higher accuracies are seen for ABC/GCB. For CHOP normalization most samples are removed at the suggested value, and a high accuracy for the few remaining samples are seen for BAGS and ABC/GCB classification while an accuracy of zero is seen for REGS.

For the RCHOP dataset most InLab reference normalized samples were retained at the suggested value of 0.6, but lower RLE IQR values did not give higher classification accuracies. Most samples normalized against the CHEPRETRO or CHOP reference are excluded at an RLE IQR of 0.6, but there is no clear indication of higher accuracies for samples with RLE IQR below the threshold.

Looking at results as a whole, a low RLE IQR in itself does not guarantee a higher classification accuracy, but it does exclude most ExLab normalized samples, and accordingly in most cases should exclude samples with low classification accuracy from biased RMA normalization.

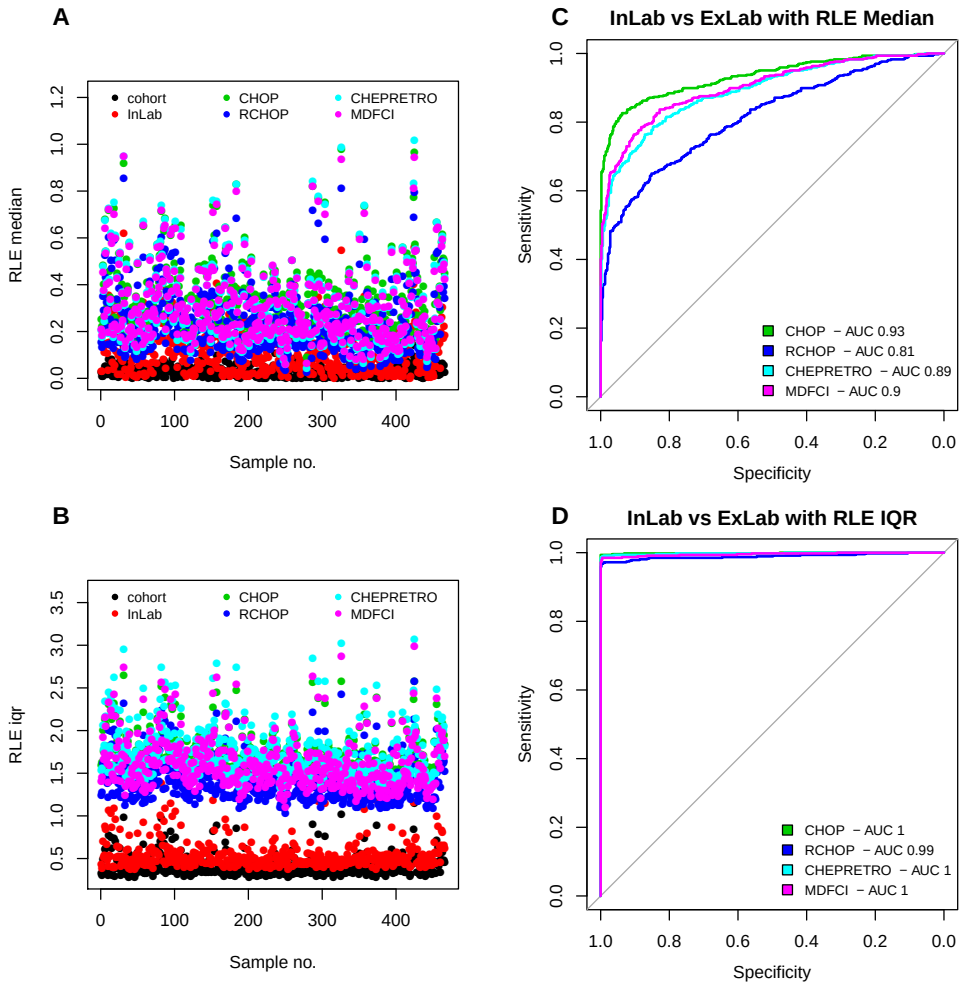

Figure S5: Absolute value of the median (A) and IQR (B) RLE values for different RMA normalizations of the IDRC dataset and ROC curves for using these values to separate between an InLab and Exlab RMA reference (C,D)

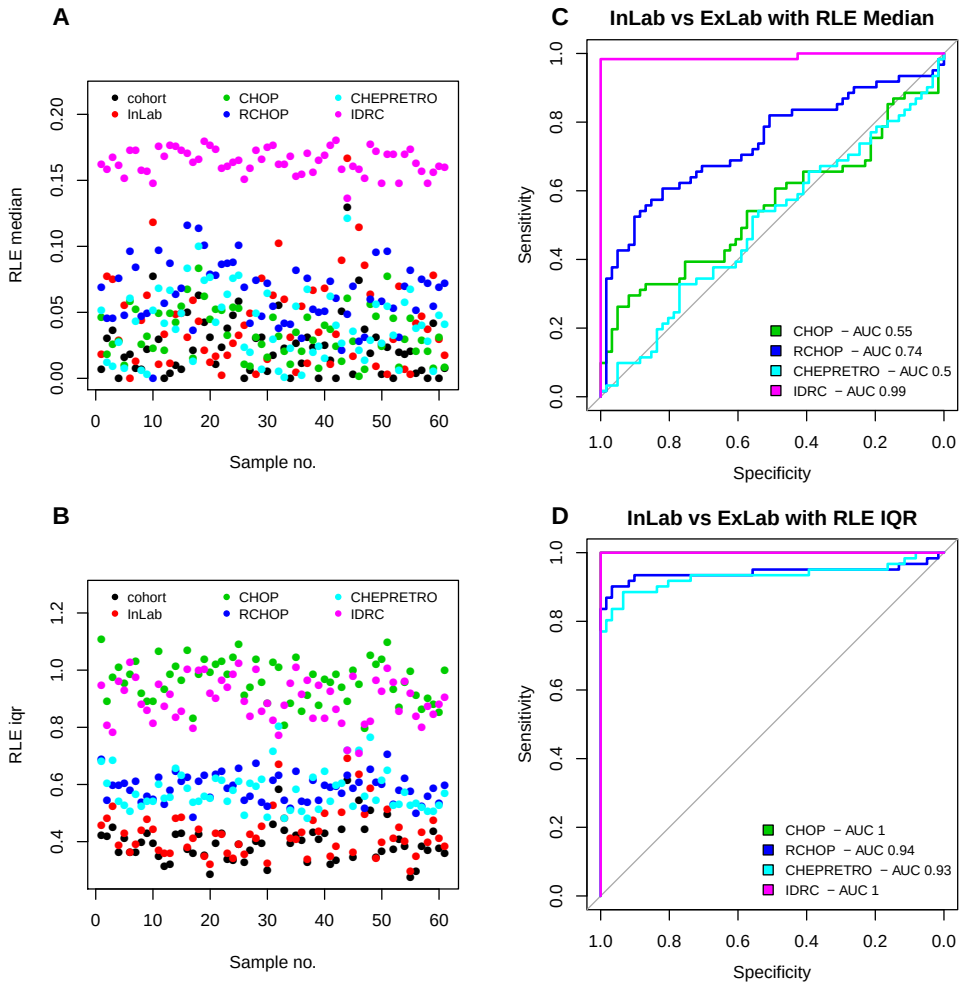

Figure S6: Absolute value of the median (A) and IQR (B) RLE values for different RMA normalizations of the MDFCI dataset and ROC curves for using these values to separate between an InLab and Exlab RMA reference (C,D)

Dataset	RMA reference	Threshold	Sensitivity	Specificity
CHEPRETRO	CHOP	0.56	0.97	1.00
CHEPRETRO	RCHOP	0.47	0.68	0.85
CHEPRETRO	IDRC	0.71	1.00	1.00
CHEPRETRO	MDFCI	0.54	0.95	0.54
RCHOP	CHOP	0.61	0.98	1.00
RCHOP	CHEPRETRORETRO	0.52	0.93	0.97
RCHOP	IDRC	0.68	0.99	1.00
RCHOP	MDFCI	0.53	0.94	0.99
CHOP	RCHOP	0.48	0.95	0.99
CHOP	CHEPRETRORETRO	0.51	0.97	1.00
CHOP	IDRC	0.62	0.99	1.00
CHOP	MDFCI	0.62	0.99	1.00
IDRC	CHOP	1.39	0.99	1.00
IDRC	RCHOP	1.08	0.97	1.00
IDRC	CHEPRETRORETRO	1.32	0.99	1.00
IDRC	MDFCI	1.19	0.98	1.00
MDFCI	CHOP	0.74	1.00	1.00
MDFCI	RCHOP	0.51	0.90	0.97
MDFCI	CHEPRETRORETRO	0.50	0.89	0.93
MDFCI	IDRC	0.70	1.00	1.00
Median	-	0.62	0.97	1.00

Table S6: Optimal thresholds for RLE IQR

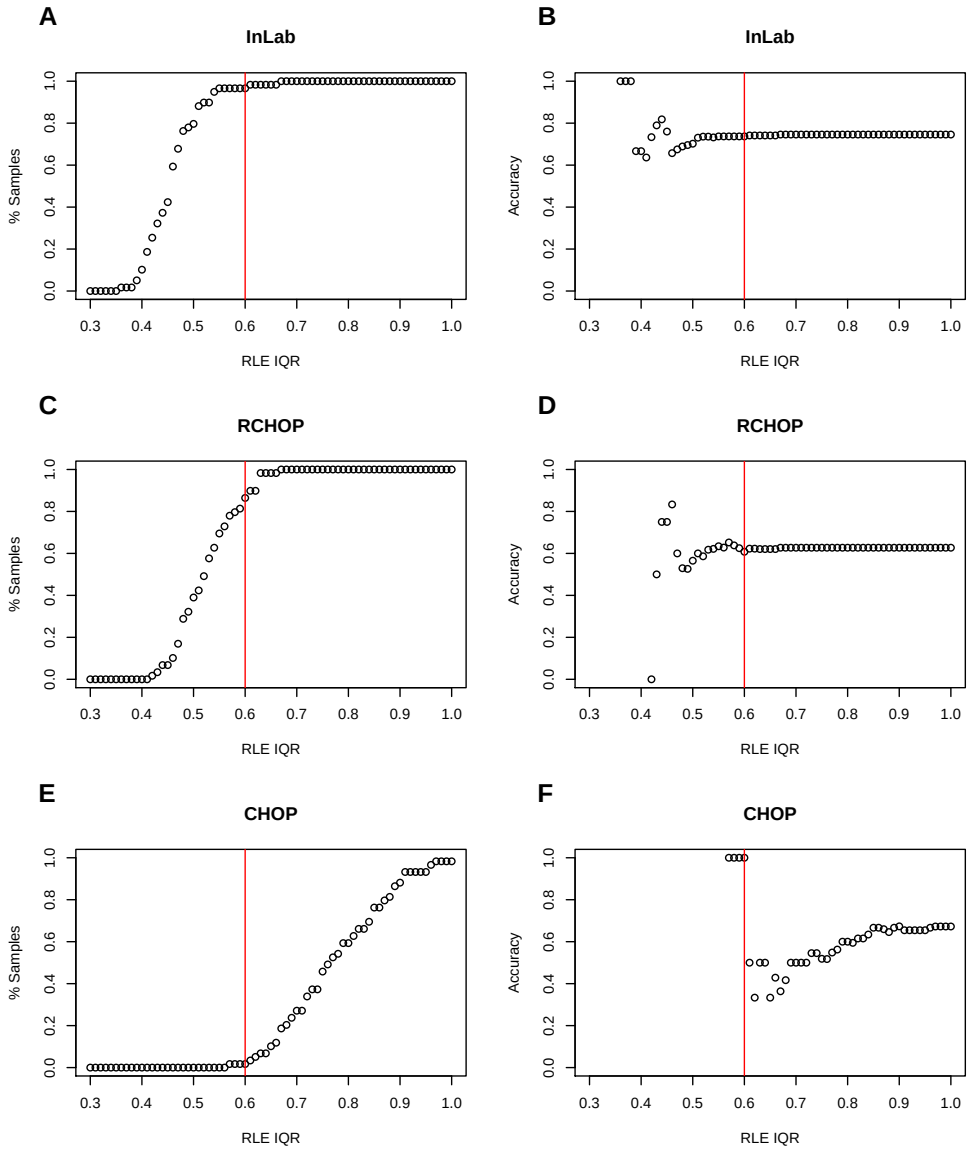

Figure S7: Proportion samples retained and accuracy of BAGS classification (percent similar with cohort based) against increasing RLE IQR thresholds for different references in CHEPRETRO. The vertical line marks the suggested threshold of 0.6

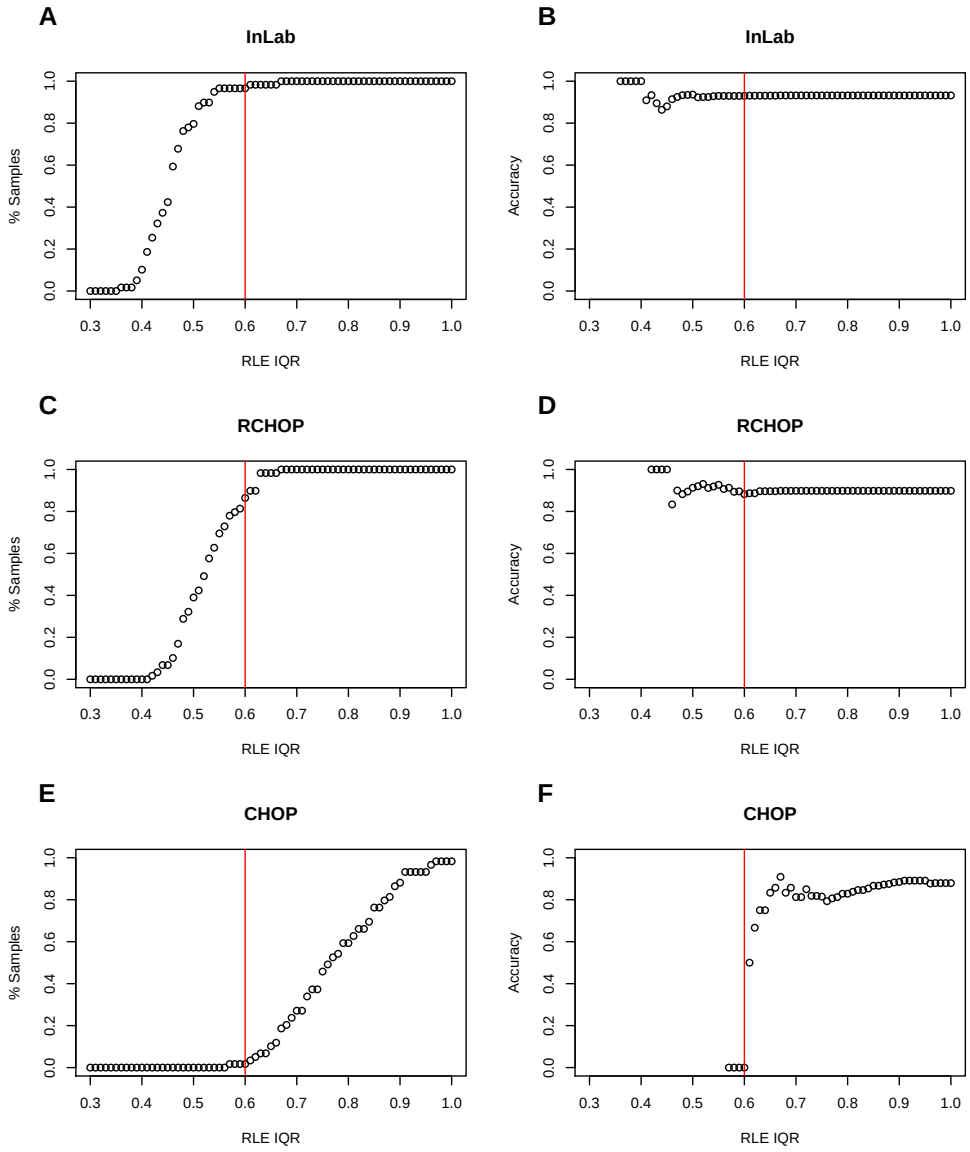

Figure S8: Proportion samples retained and accuracy of ABC/GCB classification (percent similar with cohort based) against increasing RLE IQR thresholds for different references in CHEPRETRO. The vertical line marks the suggested threshold of 0.6

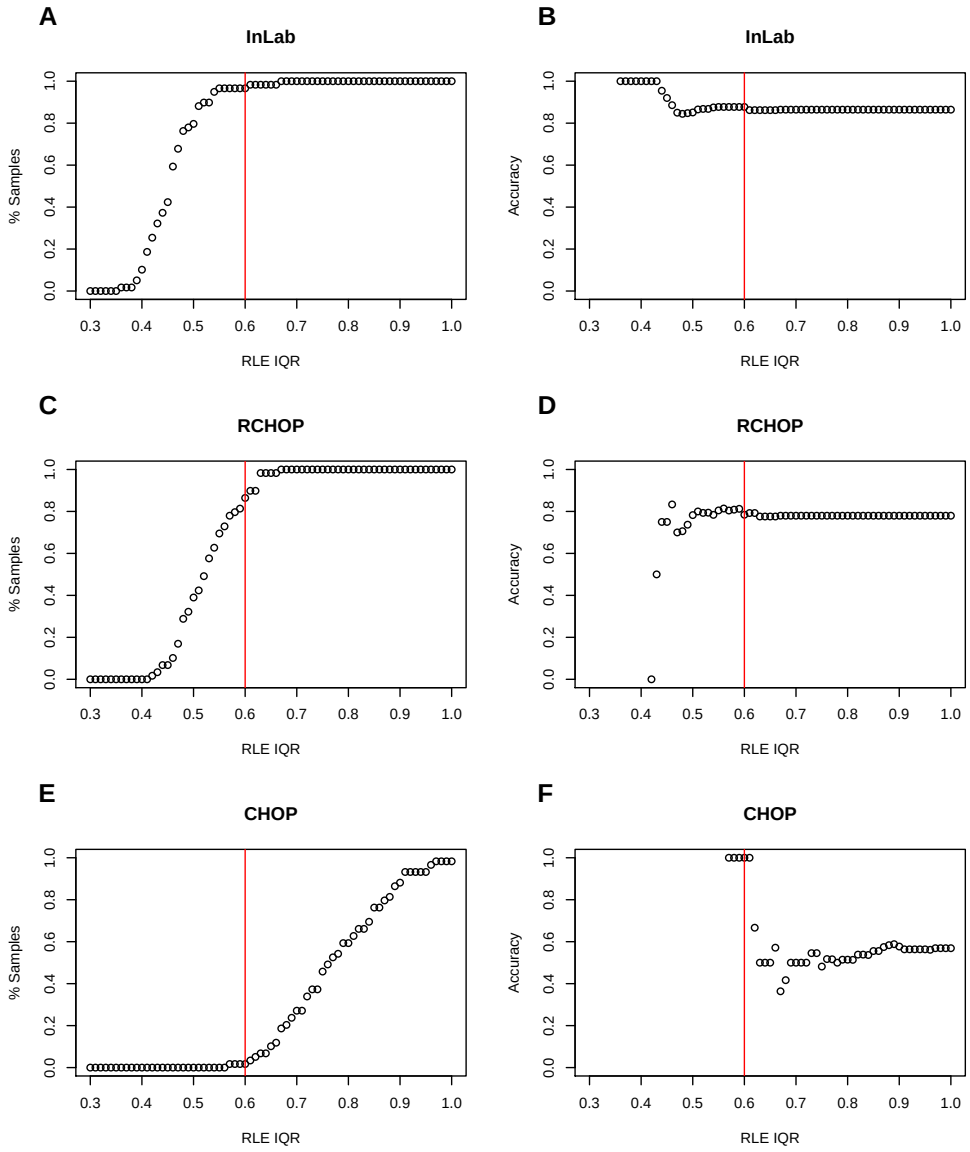

Figure S9: Proportion samples retained and accuracy of REGS(combined) classification (percent similar with cohort based) against increasing RLE IQR thresholds for different references in CHEPRETRO. The vertical line marks the suggested threshold of 0.6

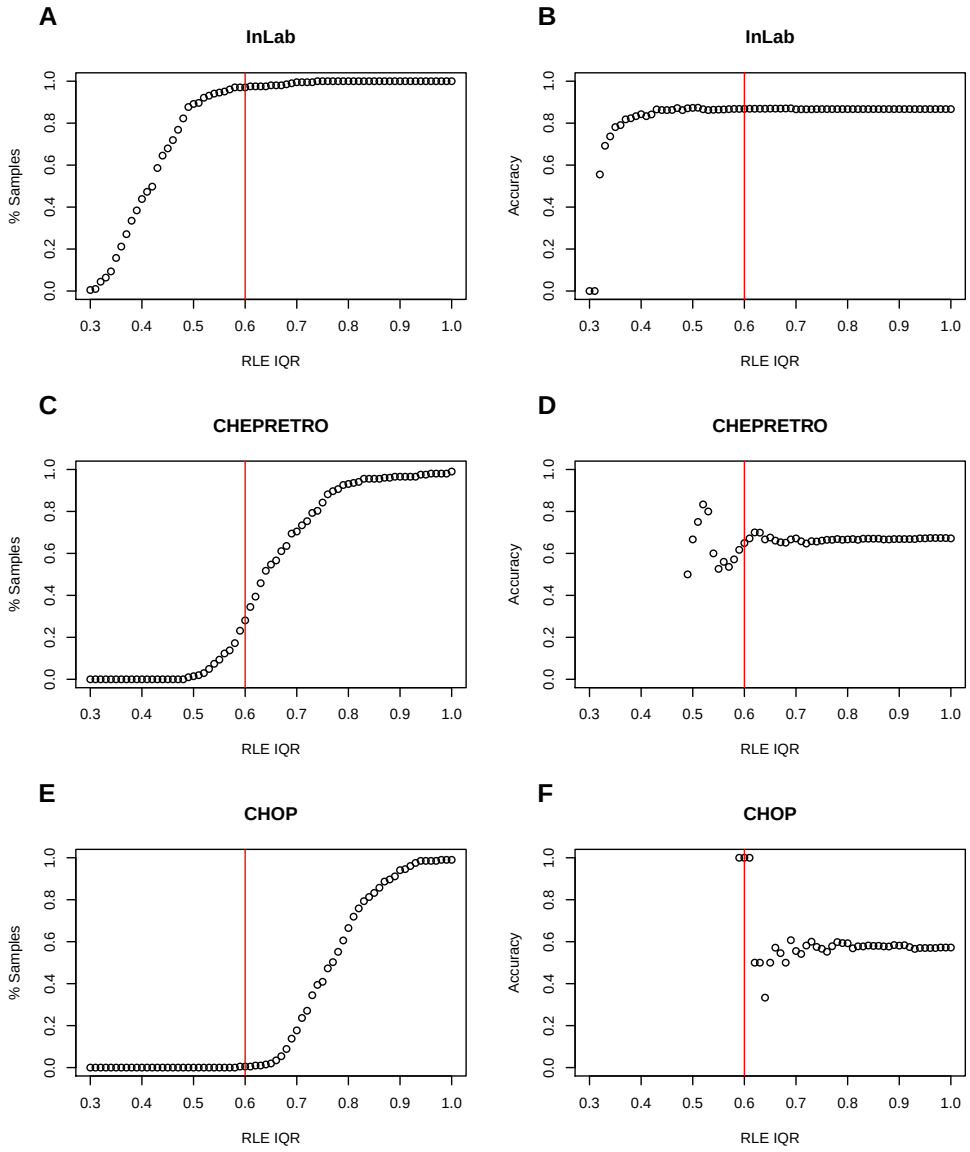

Figure S10: Proportion samples retained and accuracy of BAGS classification (percent similar with cohort based) against increasing RLE IQR thresholds for different references in RCHOP. The vertical line marks the suggested threshold of 0.6

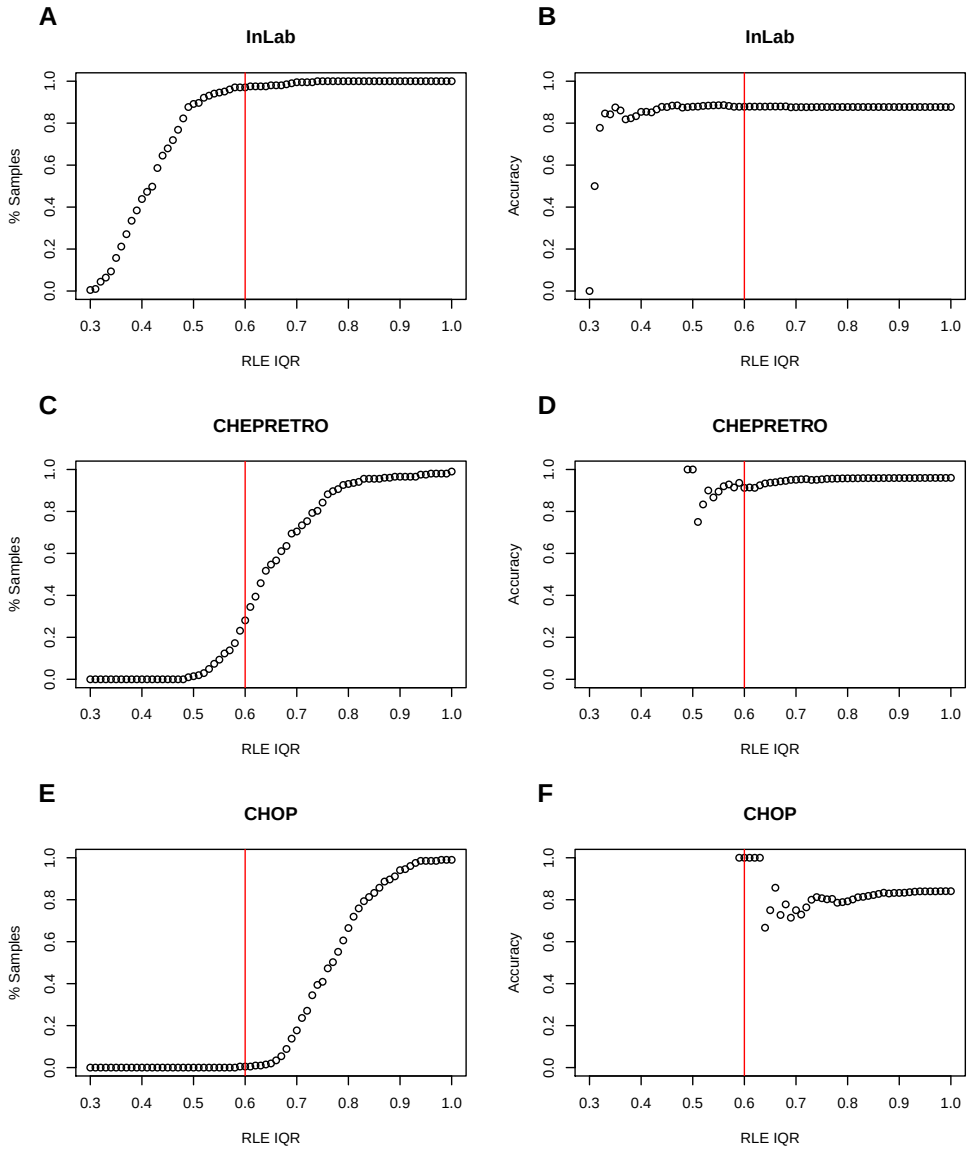

Figure S11: Proportion samples retained and accuracy of ABC/GCB classification (percent similar with cohort based) against increasing RLE IQR thresholds for different references in RCHOP. The vertical line marks the suggested threshold of 0.6

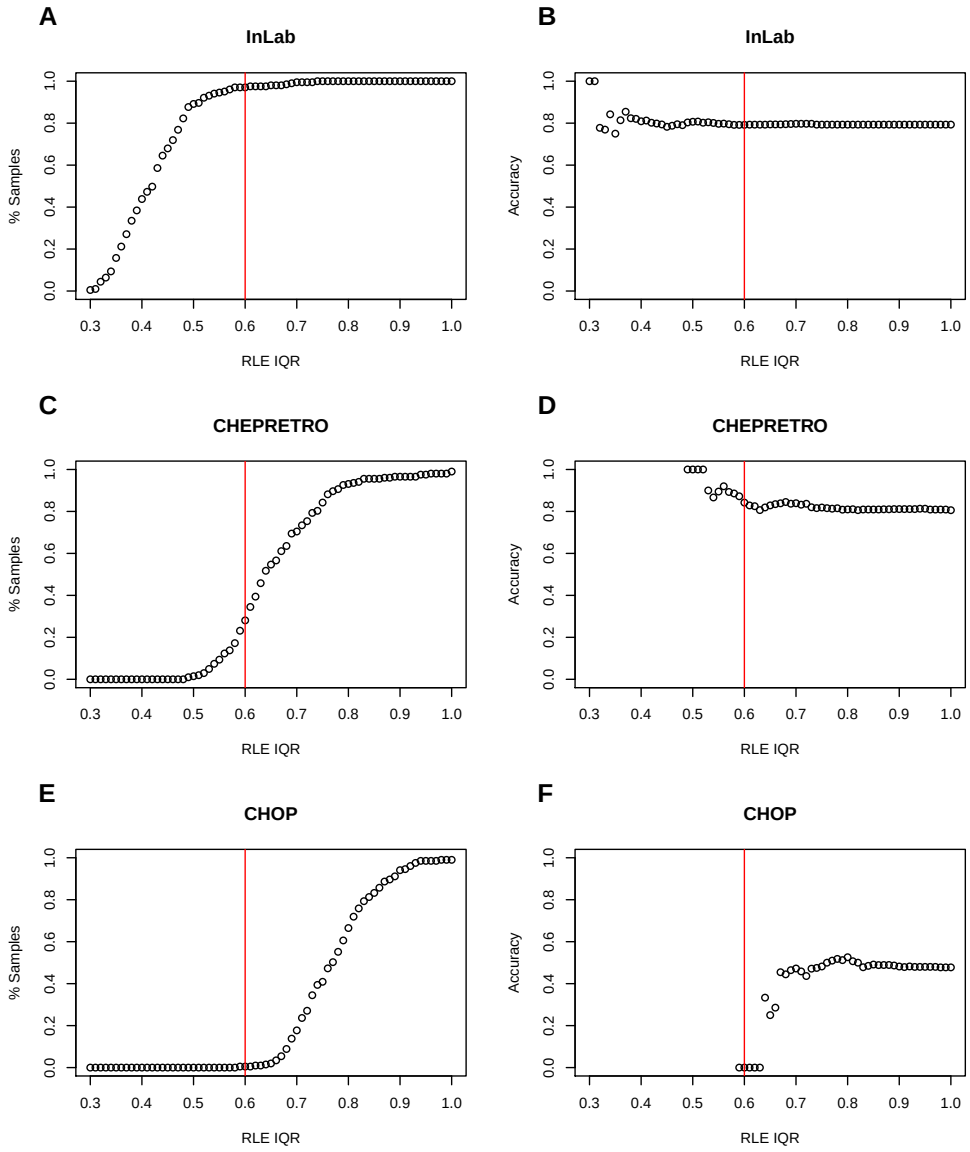

Figure S12: Proportion samples retained and accuracy of REGS(combined) classification (percent similar with cohort based) against increasing RLE IQR thresholds for different references in RCHOP. The vertical line marks the suggested threshold of 0.6