Netboost: Boosting-Supported Network Analysis Improves High-Dimensional Omics Prediction in Acute Myeloid Leukemia and Huntington's Disease

Pascal Schlosser, Jochen Knaus, Maximilian Schmutz, Konstanze Döhner, Christoph Plass, Lars Bullinger, Rainer Claus, Harald Binder, Michael Lübbert, and Martin Schumacher

Abstract—State-of-the art selection methods fail to identify weak but cumulative effects of features found in many high-dimensional omics datasets. Nevertheless, these features play an important role in certain diseases. We present Netboost, a three-step dimension reduction technique. First, a boosting-based filter is combined with the topological overlap measure to identify the essential edges of the network. Second, sparse hierarchical clustering is applied on the selected edges to identify modules and finally module information is aggregated by the first principal components. We demonstrate the application of the newly developed Netboost in combination with CoxBoost for survival prediction of DNA methylation and gene expression data from 180 acute myeloid leukemia (AML) patients and show, based on cross-validated prediction error curve estimates, its prediction superiority over variable selection on the full dataset as well as over an alternative clustering approach. The identified signature related to chromatin modifying enzymes was replicated in an independent dataset, the phase II AMLSG 12-09 study. In a second application we combine Netboost with Random Forest classification and improve the disease classification error in RNA-sequencing data of Huntington's disease mice. Netboost is a freely available Bioconductor R package for dimension reduction and hypothesis generation in high-dimensional omics applications.

Index Terms—Acute myeloid leukemia, boosting, cox proportional hazards model, dimension reduction, DNA methylation, feature selection, gene expression, hierarchical clustering, huntington's disease, random forest, survival

---

1 INTRODUCTION

Microarray, sequencing and other high-throughput functional genomics technologies are developing rapidly, incorporating more and more features. A major challenge in biomedical research is the contrast of these high-dimensional datasets and the related investigation of potentially hundreds of thousands of features with only a limited sample population in the range of tens to a few hundreds. This gives rise to dimension reduction techniques with the aim of identifying the subspace with minimal dimensions and the best characterization of the outcome variable simultaneously [1].

Many times, methods which were originally developed for the selection of tens of clinical variables are now faced with the challenge of selecting from hundreds of thousands or even from millions of variables. In cases where it is not expected that a singular feature dominates the effect but rather that a larger group of features works cumulatively, the challenge becomes even greater.

In acute myeloid leukemia (AML) part of the epigeno-type of the disease is a global increase in DNA methylation in regulatory regions [2]. Furthermore, for elderly patients the only effective drugs that counteract this effect are hypo-methylating agents [3], [4], [5]. From this it is known that the state of methylation fulfills an important role in this disease. Nevertheless it has been difficult to incorporate DNA methylation markers in patient relevant statistics like survival prediction [3], [6]. Predictive methylation sites in AML patients treated with chemotherapeutics [7] and predictive sites from chronic myelomonocytic leukemia

- P. Schlosser is with the Institute of Genetic Epidemiology, Faculty of Medicine and the Medical Center-University of Freiburg, 79106 Freiburg im Breisgau, Germany and also with the the Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, 79104 Freiburg im Breisgau, Germany.
  E-mail: pascal.schlosser@uniklinik-freiburg.de.
- J. Knaus, H. Binder, and M. Schumacher are with the Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, 79104 Freiburg im Breisgau, Germany.
  E-mail: {jo, binderh, ms}@imbi.uni-freiburg.de.
- M. Schmutz and R. Claus are with the Department of Hematology and Oncology, Augsburg University Medical Center, 86156 Augsburg, Germany. E-mail: {maximilian.schmutz, rainer.claus}@uk-augsburg.de.
- K. Döhner is with the Department of Internal Medicine III, University Hospital of Ulm, 89081 Ulm, Germany.
  E-mail: konstanze.doehner@uniklinik-ulm.de.
- C. Plass is with the Division of Cancer Epigenomics, German Cancer Research Center, 69120 Heidelberg, Germany. E-mail: c.plass@dkfz.de.
- L. Bullinger is with the Hematology, Oncology and Tumor Immunology, Campus Virchow Hospital, Charite University Medicine, 13353 Berlin, Germany. E-mail: lars.bullinger@charite.de.
- M. Lübbert is with the Department of Hematology-Oncology, Medical Center, Faculty of Medicine, University of Freiburg, 79106 Freiburg im Breisgau, Germany, and also with the German Consortium for Translational Cancer Research (DKTK), Freiburg im Breisgau, Germany.
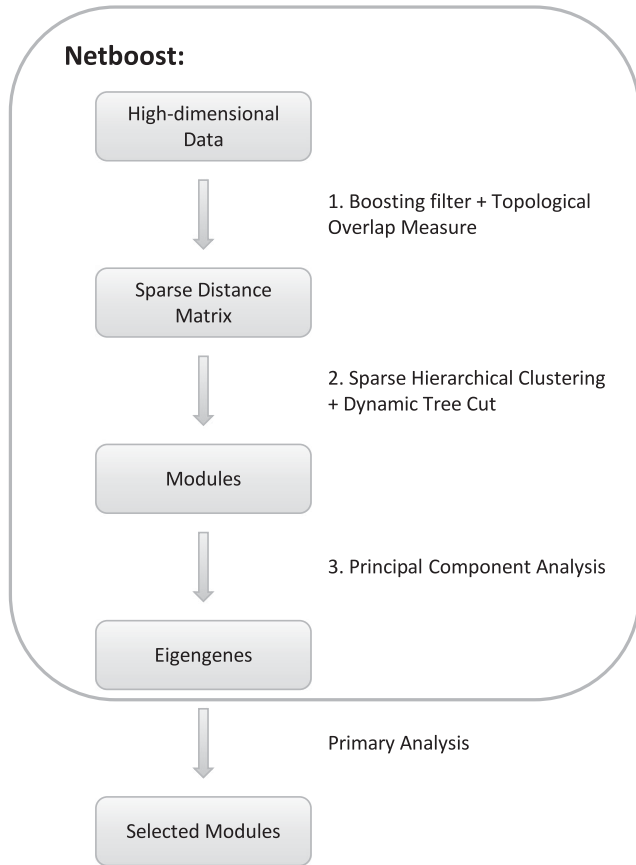 E-mail: michael.luebbert@uniklinik-freiburg.de.

Fig. 1. Netboost concept flow chart.

patients treated with hypomethylating drugs [8] could not be replicated for AML patients treated with hypomethylating drugs.

Weighted Gene Co-expression Network Analysis (WGCNA) [9] is a versatile framework to extract networks from high-dimensional data. It is able to identify biologically functional subgroups, called modules, under many differing settings [10], [11], [12]. When relating this structured information to the outcome of interest, additional challenges are faced. We are interested in the subgroups of features which are most central to the function of the modules. The method has to be even more selective in differentiating background noise from true connections to be used to explain the interplay of differing molecular levels, like gene expression and DNA methylation.

With Netboost, we propose a procedure to reduce dimensions within high-dimensional datasets. We put a specific emphasis on large subgroups of features that show a shared effect. For this we aggregate subgroup information before applying the primary analysis strategy. In public domain examples we show that we are able to extract patient relevant information from multiple high-dimensional measurement types.

The paper is organized as follows. Section 2 outlines the newly developed Netboost and describes the implementation. In Section 3 Netboost is applied in two public domain datasets and its performance is compared with two other approaches for each application. Section 4 outlines the limitations and the potential future direction of the proposed method to conclude the paper.

2 METHODS

Netboost is a three-step procedure. As shown in Fig. 1, in the first step we calculate the boosting-based filter and a sparse distance matrix between features (Sections 2.1.1 and 2.1.2). From this combination we can reduce the network to its essential edges and remove spurious connections, originating from noise. We still retain the interconnectedness and stability of complex network structures including indirect connections that occur in many omics datasets reflecting biological pathway structures.

The second step consists of sparse hierarchical clustering and the dynamic tree cut procedure to determine modules from the dendrogram (Section 2.1.3) to transfer the network into a clustering.

Subsequently in step three, we aggregate the information in the modules by their first principal components (PCs) (Section 2.2) to achieve a low-dimensional representation of the original data.

In this network-based dimension reduction method we modify the WGCNA by the addition of a multivariate filter and by the application of sparse hierarchical clustering. Netboost is then followed by the primary analysis approach with the aggregated module information instead of the original omics data. Here we present two applications, CoxBoost [13] to fit a Cox proportional hazards model integrated with a variable selection (Section 3.1) and a classification analysis by Random Forests [14] (Section 3.2).

2.1 Module Detection

Let X be a $n \times p$-dimensional data-matrix, where $n \ll p$ with n being the number of samples and p the number of features. We assume X to be continuous in each feature.

2.1.1 Boosting-Based Filter

To first identify a general structure of our network we aggregate a filter of important network edges by boosting. Let $m \in \mathrm{I\!N}$ index samples, $i, j \in \mathrm{I\!N}$ index features and

$$X_i := X_{m \leq n, i} \text{ and } X_{-i} := X_{m \leq n, j \neq i}.$$

We fit

$$X_i = X_{-i}\beta_{j \leq p-1},$$

with $\beta \in \mathrm{I\!R}^{(p-1)}$. Here, we perform component-wise likelihood-based boosting to fit a linear approximation of the outcome variable. In each iteration we fit the linear base learners using Fisher scoring with respect to the overall likelihood function one-by-one while keeping all other base learners fixed [15], [16]. Boosting is stopped after a given fixed number of steps. For all $\beta_j \neq 0$ we then add the tuple (i, j) to the filter. We estimate an undirected network, therefore the order of (i, j) is irrelevant. We define the filter by

$$\mathcal{F} := \{(i, j), (j, i) | \exists i, j \in \mathrm{I\!N} \ \beta_i \neq 0 \text{ in } X_j = X_{-j}\beta\}.$$

By pruning the network to $\mathcal{F}$ we remove uninformative edges, reduce computational load and noise in subsequent steps.

2.1.2 Distance Calculation

For tuples in $\mathcal{F}$ we define the adjacency of two features analogous to [9] by the power adjacency function. For all other tuples the adjacency is set to 0. Hence, we have

$$a_{ij} := \begin{cases} |\mathrm{cor}(X_i, X_j)|^b & \text{if} (i,j) \in \mathcal{F}, \\ 0 & \text{else,} \end{cases}$$

where b is chosen data-based by the scale free topology criterion [17] on a random subset of features. Here, the model fit of the correlation structure with the criterion is approximated by a linear regression and b is chosen to maximize this fit. cor denotes the Pearson correlation.

We combine the topological overlap measure (TOM) [18] with $\mathcal{F}$ and define

$$\mathrm{TOM}_{ij} := \begin{cases} 1 & \text{if} i = j, \\ \dfrac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{\min(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{uj}) + 1 - a_{ij}} & \text{if} (i,j) \in \mathcal{F}, \\ 0 & \text{else.} \end{cases} \quad (1)$$

As $a_{ij} \in [0,1]$ it follows that $\mathrm{TOM}_{ij} \in [0,1]$. These similarities are inverted to distances by

$$\mathrm{DistTOM}_{ij} := 1 - \mathrm{TOM}_{ij}. \quad (2)$$

2.1.3 Hierarchical Clustering and Decomposition into Modules

We apply the unweighted pair group method with arithmetic mean (UPGMA) [19] to DistTOM. Parts of the network where no path exists in $\mathcal{F}$ are clustered independently. A path between X_i and X_j exists exactly then when there is an $l \in \mathrm{IN}$ such that there are $t_{1..l} \in \mathcal{F}$ with $i = t_{11}$, $j = t_{l2}$ and $\forall s : 1 \leq s \leq l-1 \ t_{s2} = t_{(s+1)1}$. The dendrograms resulting from these hierarchical clusterings are separated into modules by the Dynamic Tree Cut procedure [20]. Thus, features which are topologically close on the filtered edges are grouped into modules.

2.2 Aggregation of Module Information

By design the first PC explains the variation in one dimensional space to the highest possible degree. Therefore, we aggregate the information in each module by its first PC, the so called eigengenes [9]. In a final step modules with highly correlated first principal components are merged to further reduce dimensionality. We define E_q as the first principal component of the qth module and

$$X_{\mathrm{modules}} := (E_1^{\mathrm{T}}, \ldots, E_m^{\mathrm{T}}),$$

where m is the number of detected modules. X_{modules} has the dimension $n \times m$ where $m \ll p$. Due to its definition a substantial part of variation in X is conserved in X_{modules}. At the same time the dimensionality is considerably reduced.

2.3 Module Selection and Evaluation

Variable selection is performed analogous to the primary analysis approach, but it is done on the set of eigengenes rather than on the set of features.

2.3.1 CoxBoost

We apply CoxBoost [13] to integrate the potentially still high-dimensional X_{modules} with clinical covariates and survival data as the primary outcome by likelihood-based boosting. Analysis is implemented with the CoxBoost R package [21]. The stopping criterion is chosen by cross-validation and a Cox proportional hazards model is fitted.

2.3.2 Prediction Errors

To evaluate the performance of CoxBoost models we used the peperr R package [22] which implements .632+ prediction errors based on subsamples without replacement as recommended in [23]. These Brier scores allow for interpretation of prediction errors accounting for censoring as well as the time-dependent nature of survival data [24], [25], [26]. In high-dimensional data constellations bootstrap samples with replacement often lead to overly complex models. Therefore, subsamples without replacement of 63.2 percent of the samples, which is equal to the expected number of unique observations in one bootstrap sample drawn with replacement, are implemented. Variability of prediction error curves is displayed by the distribution of integrated prediction error curves of the subsamples.

2.3.3 Random Forests

We apply random forests as described in [14] to classify samples based on X_{modules} to their disease severity classes. To adequately explore the space of possible trees, also for the most high-dimensional of the analyses, we grow 10,000 trees in each analysis.

2.4 Implementation

Netboost is built as an R package. It has been tested under Linux and macOS. A Windows implementation is currently not planned due to compiler dependencies.

As depicted in Algorithm 1 we first calculate $\mathcal{F}$. Under the assumption of continuous X_j and after scaling and centering each we efficiently implement the likelihood based boosting. The subsequent calculation of the adjacencies and the TOM are performed exclusively on network edges in $\mathcal{F}$. Then the sparse distance matrix is exported to *Sparse* UPGMA by [27]. Here all empty edges where the nodes are connected indirectly are assumed to have the maximal distance in the network and completely unconnected nodes of the network are processed separately in independent clusterings. This agrees with the described method as all connected nodes not in $\mathcal{F}$ have the distance of 1. By applying the filter we therefore reduce the memory load and computational burden massively as the filter is smaller than the whole network by orders of magnitudes as demonstrated in the examples in Section 3.

The algorithm is freely available as a Bioconductor R package at http://bioconductor.org/packages/release/bioc/html/netboost.html. All functionality of Netboost is available from within R whereas substantial parts of the algorithm are implemented in C++. *Sparse* UPGMA is part of the standalone MC-UPGMA software (for details see [27]). It is distributed with the Netboost R package. For extraction of modules we applied the WGCNA [9] and

dynamicTreeCut [20] R packages. As an example for the computational demand Netboost was run on a dataset with 180 samples and 413,169 features (for details see Section 3.1). Applying two Xeon E5 2690v3 at 2.6 GHz (2x12cores) and 40 GB of RAM it took Netboost 13.94 hours to compute.

Algorithm 1. Netboost

Input: X, steps, minModuleSize, MEDissThres
Result: X_{modules}
$\mathcal{F} = \emptyset$;
for $j \leftarrow 1$ **to** p **do**
    fit $X_j = X_{-j}\beta$;
    $\mathcal{F} = \mathcal{F} \cup \{(i,j)|\exists i \in \text{IN } \beta_i \neq 0\}$;
end
$\mathcal{F} = \{(i,j)|(i,j) \in \mathcal{F} \vee (j,i) \in \mathcal{F}\}$;
randomFeatures = X[,sample(n= min(10,000,ncol(X))];
scaleFreeTopologyCriterion(randomFeatures) $\rightarrow b$;
for $(i,j) \in \mathcal{F}$ **do**
$a_{ij} = |cor(X_i, X_j)|^b$;
end
for $(i,j) \in \mathcal{F}$ **do**
    compute DistTOM$_{ij}$:= Equation (2);
end
sparseUPGMA(DistTOM) $\rightarrow$ dendrogram;
cutreeDynamic(dendrogram, minModuleSize) $\rightarrow$ modules;
for $q \in$ modules **do**
    compute first principal component E_q;
end
while $\exists q, q'$ with $cor(E_q, E_{q'}) > (1 - \text{MEDissThres})$ **do**
    merge(q, q');
    compute first principal component of merged module;
end

---

3 RESULTS

We apply Netboost to two datasets. In Section 3.1 it is applied to DNA methylation and gene expression data from The Cancer Genome Atlas (TCGA) AML cohort to predict survival (see Section 2.3.1). In Section 3.2 it is applied to RNA sequencing data to classify (see Section 2.3.3) Huntington's disease severity in mice.

3.1 TCGA AML: Methylation and Gene Expression Predictive of Overall Survival

We selected the 180 AML patients in the public domain TCGA database for which overall survival data, methylome and gene expression measurements were available. TCGA data was already preprocessed and normalized. Methylation was quantified with Illumina Infinium HumanMethylation450 BeadChip arrays and gene expression by Affymetrix HG U133 Plus 2.0 arrays. Thereby 396,065 methylation and 17,104 gene expression measurements were incorporated. We compared the following three schemes each with and without a clinical score. The clinical score is a dichotomized version of the linear predictor of a Cox proportional hazards regression model [28] of age at diagnosis and cytogenetic risk group, assessed as low, intermediate or high. Baseline hazards were estimated in separate strata according to sex. The score was solely used to evaluate Netboosts performance.

In the models with the clinical score it was set as mandatory and thereby unpenalized in CoxBoost. Thereby, DNA methylation and gene expression information was only added in these models if they could improve the prediction on top of the clinical score.

1) Direct application: Application of CoxBoost on the full dataset X.
2) Blockwise modules: The same approach as in Netboost but with the module detection done by blockwise WGCNA.
3) Netboost: Module PCs are calculated as described in Section 2. CoxBoost is applied to these.

The blockwise modules approach was the initial inspiration for the Netboost method. They coincide with one another apart from Netboosts added boosting-based filter and that X has to be separated into feature subsets for WGCNA so that the whole correlation matrix on the individual subsets can be computed. This is accomplished with k-means clustering and later aggregation via correlated eigengenes [9].

CoxBoost was implemented in all analyses in R with the peperr package [22]. We used 10-fold cross validation to estimate the optimal stopping criterion on the interval from 0 to 100. We applied 200 resampling steps to estimate the 632+ prediction errors.

In models 1.-3. we integrated the analysis without the clinical score. The direct application on the full dataset, X, selected two features and the 632+ prediction error curve, depicted in Fig. 2, shows no improvement over the null model. The estimated .632+ prediction errors for days since diagnosis are given in blue for the null model and dashed blue for the clinical model. Prediction error curves based solely on DNA methylation and gene expression are presented in black: The solid line for the direct application of CoxBoost, the dotted line for the combination with weighted gene co-expression network analysis and the dashed line for the combination with Netboost. The corresponding prediction error curves additionally based on unpenalized clinical data are presented in red.

Blockwise modules identified 568 modules with a mean module size of 671 in the range of 10 to 57,548. Ten was set as the minimum module size. Henceforth, 92 percent of the features were assigned to modules. The proportion of variance explained by eigengenes ranged from 23.9 to 94.6 percent (median = 50.5 percent). In the WGCNA aggregated $X_{\text{WGCNAmodules}}$ two modules were selected by CoxBoost summarizing 26 features.

For Netboost the multivariate filter was stopped after 20 steps and resulted in a filter of 4,956,518 network edges. This represents approximately 0.003 percent of the edges. Based on this Netboost identified 739 modules with an average module size of 52 in the range of 10 to 4,251. Accordingly 9 percent of the features were assigned to modules. The dendrogram based on the sparse network is depicted in Fig. 3. Netboost eigengenes generally explained a higher proportion of variance (median = 66.5%, range = [45.7%, 97.3%]). CoxBoost selected six modules from the Netboost aggregated $X_{\text{Netboostmodules}}$, summarizing 278 features. The minimum p-value of individual tests regarding a violation of the proportional hazards assumption is 0.13 and the mean is 0.46. The global p-value is 0.53 [29]. For situations were the proportional hazards assumption can not be made, we refer
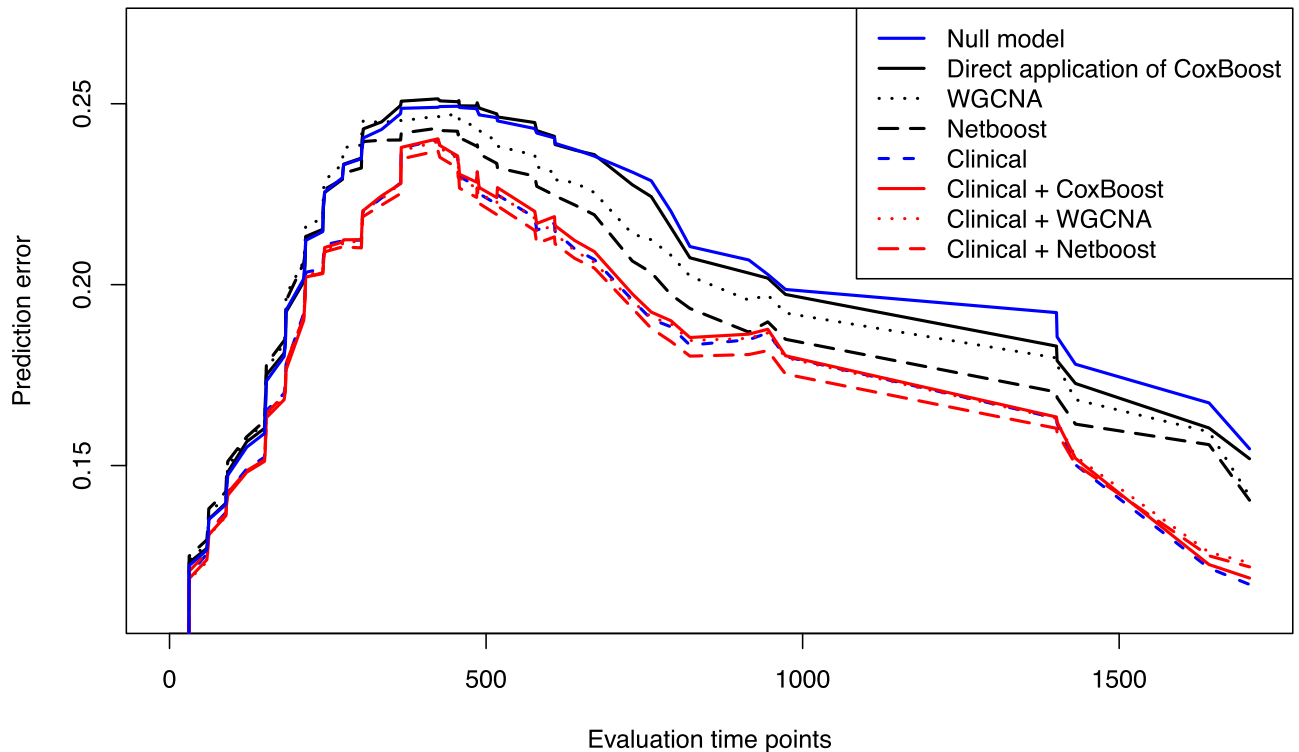
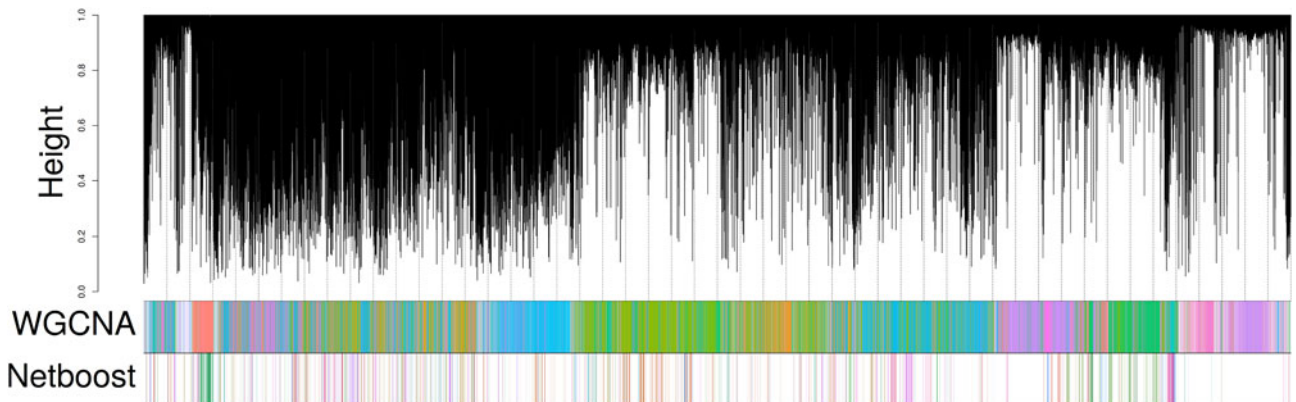Fig. 2. .632+ prediction error estimates for AML survival models.

Fig. 3. *Dendrogram of the TCGA AML data.* Dendrogram of the DNA methylation and gene expression features in the TCGA AML data. The two rows below show the separation into modules by blockwise Weighted Gene Co-expression Network Analysis (WGCNA) and Netboost.

to [30], where an integration of landmarking and CoxBoost is proposed. None of the features are shared by the selected Netboost modules and the selected WGCNA modules.

As shown in Fig. 2 the higher complexity indeed corresponds to a better prediction performance in the 632+ prediction errors. The blockwise modules approach was able to extract some information but was outperformed by Netboost. This also holds true when incorporating the variability of the individual 632+ resampling steps in Fig. 4.

As depicted in Figs. 2 and 4, once we added the clinical score as a mandatory covariate, none of the three approaches was able to extract substantial additional information from the molecular data. Overall, when comparing integrated prediction errors all analyses but the direct application of Cox-Boost showed significant improvements over the null model (one-samples Student's t-test, p-value < 0.05). Netboost

including the clinical score had the lowest p-value (p-value = 1.3e-27). When comparing analyses with each other the integration with WGCNA and the Netboost significantly improved CoxBoost (p-value = 0.0437 and p-value = 0.0002, respectively) and Netboost improved the accuracy of survival prediction on top of WGCNA (p-value = 0.0413). Furthermore, all analyses including the clinical score significantly improved prediction when compared with any analysis without the clinical score. In between analyses including the clinical score no significant differences were observed (two-sample Student's t-test, p-value < 0.05).

To investigate the possibility of the molecular information extracted by Netboost being a surrogate for the clinical score, we fitted logistic regression models for the module eigengenes to the clinical score. We compared random selections of features out of all DNA methylation and gene
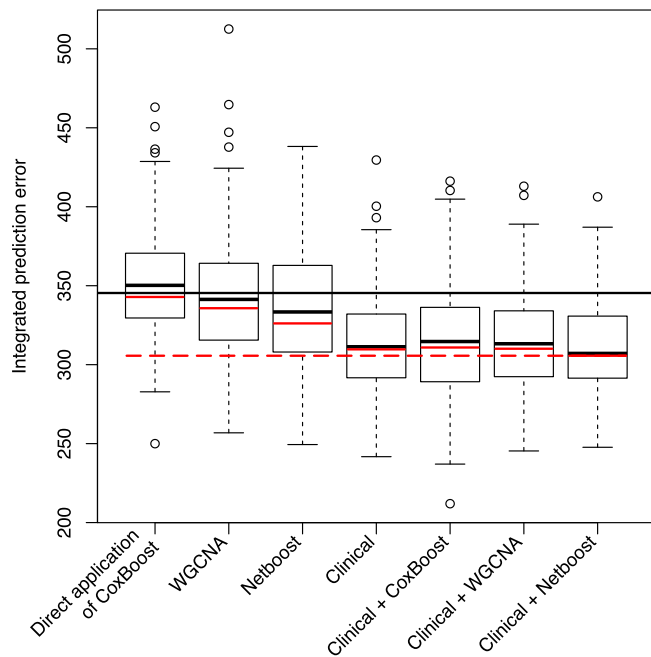
Fig. 4. *Variability of the .632+ prediction error estimates in AML survival models.* Integrated prediction error curve estimates from single subsamples for CoxBoost on the full dataset, CoxBoost on X_{WGCNA} and CoxBoost on X_{Netboost}. The Kaplan-Meier benchmark value is indicated by a horizontal line. Red lines indicate the integrated .632+ prediction error estimates with the line for the Clinical + Netboost model (lowest error) being extended by a dashed line.

Fig. 5. *Mis-classification rate for logistic regression models of the clinical score in AML.* We compare randomly selected features of the raw data with randomly selected modules and the modules selected for survival prediction performance. The complexity of models is fixed to two and six to match the final survival models for Netboost and WGCNA respectively. The horizontal line indicates the expected mis-classification rate at random. Asterisks indicate significance of unpaired two-samples Student's t-test (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, NS. $p \geq 0.05$). Only neighbouring columns were tested.

expression features and modules, WGCNA and Netboost respectively, of similar size to the modules selected by WGCNA and Netboost and the modules selected for survival prediction. We fitted 500 models on subsamples of size 100 and evaluated the misclassification-rate on the remaining samples. For the random selections, features were reselected with each fit. As shown in Fig. 5, the selected Netboost modules approximated the clinical score best.

To further comprehend the differences in the clustering we took random subsets of size 100 and compared the resulting Netboost and WGCNA clusterings using pairwise adjusted Rand Indices and Jaccard Indices [31], [32]. Additionally, we calculated kmeans clusterings with the number of clusters fixed to the median number of clusters in Netboost clusterings (646) and WGCNA clusterings (533) and generated random clusterings with the respective number of clusters. The Rand Index is defined as the proportion of consistently clustered features between the two clusterings so tuples of features that are in both clusterings either in a common cluster or in both clusterings in differing clusters. The adjusted Rand Index corrects this for the expected number of consistent tuples given that the number of features and the number of clusters such that $E[\text{adjusted Rand Index}] = 0$. The Jaccard Index is similar to the Rand Index, however it disregards tuples for features that are in different clusters for both clusterings. Both indices are less than or equal 1 and exactly 1 for identical clusterings. As seen in Fig. 6 both random clusterings had consistently pairwise indices of 0 and both kmeans clusterings were outperformed by WGCNA and Netboost with respect to both metrics. With respect to the adjusted Rand Index Netboosts median was below WGCNAs median while the order of minima was vice versa.
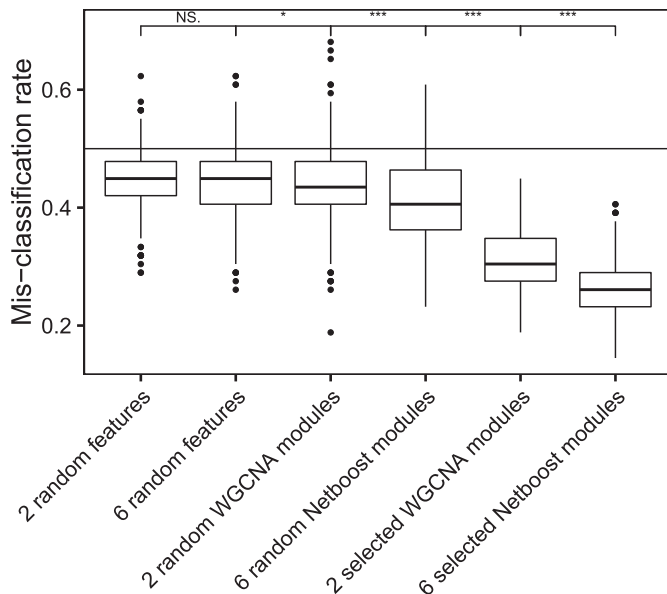
When comparing the Jaccard Indices, Netboost outperforms WGCNA and shows a higher similarity for all pairwise comparisons with respect to this measure.

Netboost modules reflected known biology. 206 of the 739 Netboost modules consist of CpGs within 1,000 base pairs demonstrating the strength of local dependency in DNA methylation data. Netboost re-identified data-driven the association of CpGs in close proximity and *cis* association of gene methylation and expression. In total, six different modules were selected that were variable in size and composition: 4 of the 6 modules consisted only of CpGs, one predominantly of CpGs and, in addition, 2 RNAs, one module only of 14 RNAs. The total number of CpGs varied from 10 to 88. The largest module (88 CpGs) contained numerous genes associated with hematopoiesis, such as *WT1* and *CXCL2*. The 2nd largest module (80 CpGs, 2 RNAs) represented several genes encoding chromatin-modifying enzymes such as the H3K9 histone methyltransferase EHMT1 and the DNA demethylase TET3. To illustrate the strong association of this chromatin associated module alone we plotted stratified Kaplan-Meier curves according to its bimodal distribution (Fig. 7 A,B). The p-value of the likelihood ratio test of the dichotomised module levels (p-value = 7.0e-7) surpassed the one of the continuous module levels (p-value = 4.0e-6); indicating that there might indeed be two states of these genes. Several of these have already been implicated in AML pathogenesis and appear very promising for future predictive scores. Specifically, 4 CpGs mapped to the gene encoding EHMT1, also represented in the 4-gene methylation signature described by [7].

To validate the Netboost signature, we transferred it to DNA methylation data generated on pre-treatment patient samples from the phase II AMLSG 12-09 study [33]. In this
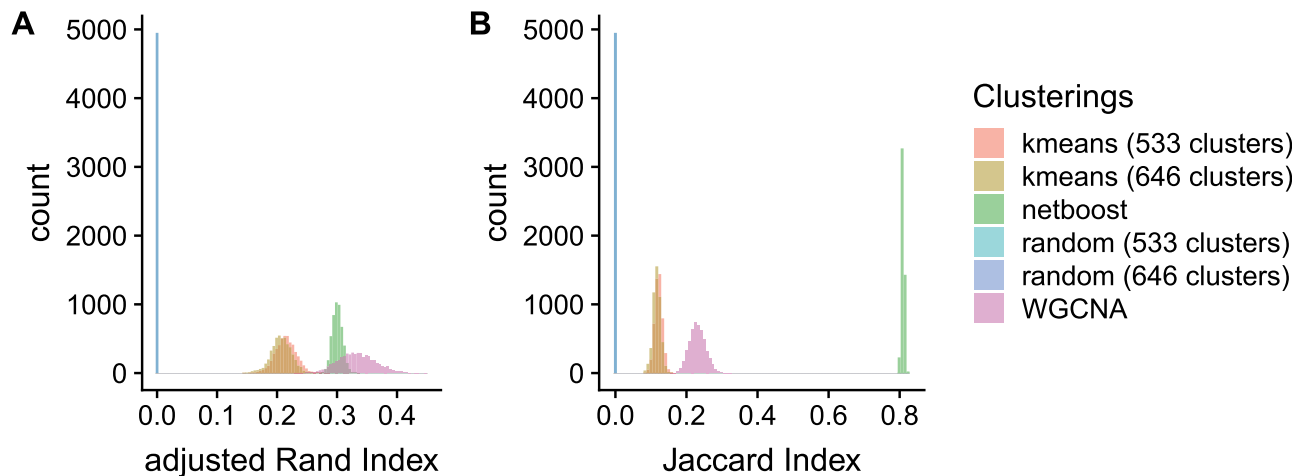
Fig. 6. *Clustering indices of the TCGA AML data.* Histogram of cluster indices of 100 clusterings on random subsets of 100 samples applying Netboost, WGCNA, kmeans with n classes and random selection of n labels, where n was set to the median number of modules in the Netboost runs (646) and WGCNA runs (533) . A) shows the adjusted Rand Index and B) the Jaccard Index.

study, DNA methylation based on the same Illumina Infinium 450k array and overall survival was available for 55 AML patients. For processing and quality control of the raw methylation data, a customized version of the CPACOR pipeline [34] was used for data normalization and calculation of beta values. The complete preprocessing pipeline is available on Github (https://github.com/genepi-freiburg/Infinium-preprocessing). As no data on gene expression was available one of the six modules could not be studied at all, while 2 were partially available (79 of 82 and 64 of 67 features) and 3 modules were available with all features. While the Cox proportional hazards model of these five modules was not significant in this smaller dataset (p-value = 0.4) the above mentioned chromatin associated module alone did replicate (p-value = 0.04). Furthermore, this module exhibited a similar bimodal pattern as in TCGA and again, dichotomization led to a smaller p-value (p-value = 0.01, Fig. 7 C,D).

After the detailed analysis of the TCGA-AML DNA methylation and gene expression dataset we downloaded three more TCGA datasets; DNA methylation data of 774 breast invasive carcinoma (TCGA-BRCA) and 315 kidney renal clear cell carcinoma (TCGA-KIRC) patients and miRNA data of 464 ovarian serous cystadenocarcinoma (TCGA-OV) patients with available overall survival information. The 1,422 TCGA-OV miRNAs without missing observations and the 20,000 CpG sites with the largest variance for TCGA-BRCA and TCGA-KIRC respectively were selected for analysis. From a descriptive perspective the KIRC dataset differed from the BRCA dataset as there was a group of highly correlated features being consistently grouped into the largest module by WGCNA, Netboost and k-means, respectively, whereas the BRCA dataset exhibited overall lower pairwise correlations and more even module sizes. The OV dataset exhibited an even simpler network structure across methods with the lower dimensionality co-occuring with approximately a dozen modules with one of them subsuming more than half the variable. For each dataset we performed the same three analyses as for AML and calculated the 632+ prediction error estimates. Boxplots of the integrated prediction errors on the test set of the individual subsamplings are depicted in Fig. 8. For clear cell carcinoma we observed

similar performance as in AML. The integration with WGCNA significantly improved CoxBoost (p-value = 0.0013) and integration with Netboost improved the accuracy of survival prediction on top of WGCNA (p-value = 0.0006). For the other two datasets none of the three approaches was able to improve overall survival prediction.

3.2 Huntington's Disease: Gene Expression and CAG Repeats

Huntington's Disease (HD) is driven by the number of CAG repeats in the huntingtin gene. In [12] WGCNA revealed 13 striatal gene expression modules that correlated with CAG length and age in a HD knock-in mouse model. Further it was shown that several of these effects translate to other HD models and patients. Recently, the analysis was extended to miRNA from the same mice in [35].

To evaluate the performance of Netboost we used the mRNA dataset in an inverse setup and determined the prediction errors in a classification task. We selected the 48 mRNA-sequencing samples from mouse striatum consisting of six genetically engineered disease severities (20, 80, 92, 111, 140 and 175 CAG repeats) with four female and four male mice all harvested at 6 months. We downloaded the preprocessed mRNA-sequences from the Gene Expression Omnibus. After removal of invariant transcripts, data consisted of 28,010 transcripts.

We compared three setups:

1) Direct application: Random forest (RF) on the full dataset X.
2) Blockwise modules: Blockwise WGCNA + RF on module PCs
3) Netboost: RF on module PCs determined by Netboost

RF was implemented in all analyses in R with the randomForest package [14]. We used 200 iterations of leave-one-out cross-validation.

The direct application on the full dataset, X, resulted in a mean prediction error of 30.8 percent.

Blockwise modules identified 61 modules with a mean module size of 423 in the range of 11 to 6221. Ten was set as
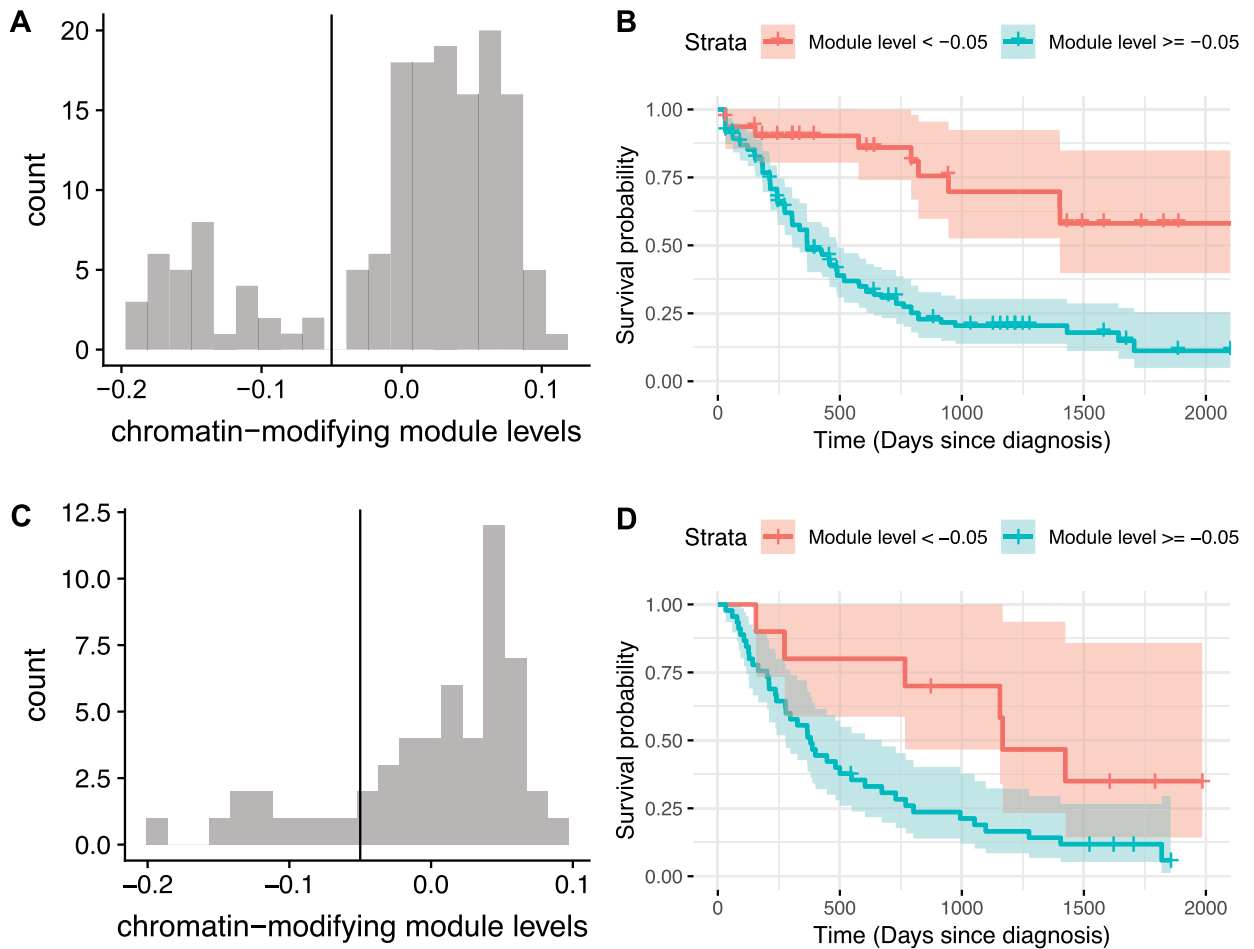
Fig. 7. *Chromatin modifying module eigengene distribution and Kaplan Meier plot in discovery and replication data.* A) shows the bimodal eigengene distribution of the TCGA AML DNA methylation and gene expression module associated with chromatin modifying enzymes. The vertical line indicates at which point patients were stratified. B) depicts the Kaplan Meier curves stratified by the modules eigengene for TCGA patients. C) shows the bimodal eigengene distribution of the transferred module in AMLSG 12-09 DNA methylation data. The vertical line indicates at which point patients were stratified. D) depicts the Kaplan Meier curves stratified by the modules eigengene for AMLSG 12-09 patients.
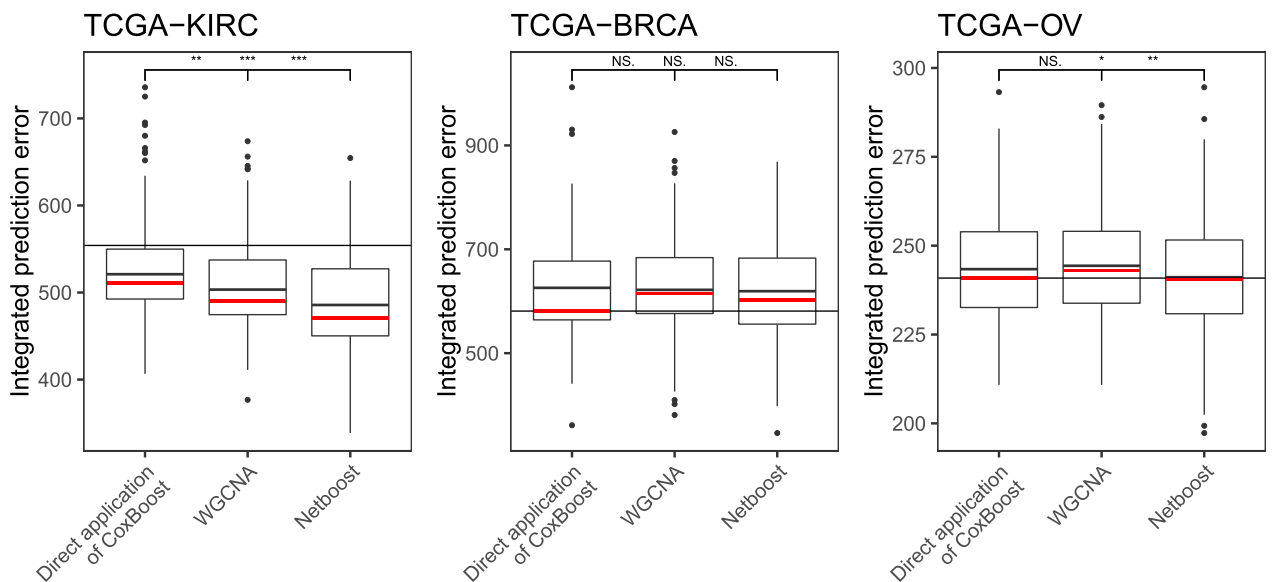

Fig. 8. *Variability of the .632+ prediction error estimates in TCGA Kidney Renal Clear Cell Carcinoma (KIRC), Breast Invasive Carcinoma (BRCA) and Ovarian Serous Cystadenocarcinoma (OV) survival models.* Integrated prediction error curve estimates from single subsamples for CoxBoost on the full dataset, CoxBoost on X_{WGCNA} and CoxBoost on X_{Netboost}. The Kaplan-Meier benchmark value is indicated by a horizontal line. Red lines show the integrated .632+ prediction error estimates. Asterisks indicate significance of unpaired two-samples Student's t-test (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, NS. $p \geq 0.05$).
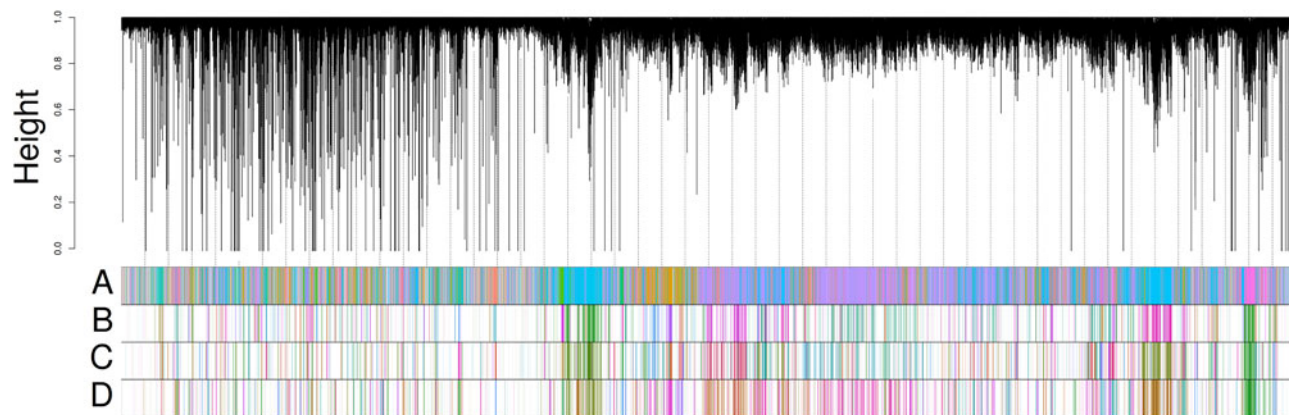
Fig. 9. *Dendrogram of Huntington's disease data.* Dendrogram of the gene expression features in the Huntington's disease data. A) shows the separation into modules by blockwise Weighted Gene Co-expression Network Analysis (WGCNA) and B), C) and D) show Netboost modules with 25, 20 and 15 boosting steps respectively.

the minimum module size. Henceforth, 92 percent of the features were assigned to modules. In the HD application the proportion of variance explained by eigengenes was lower than in the AML data (median = 42.1%, range = [29.3%, 63.4%]). On the WGCNA aggregated $X_{\text{WGCNAmodules}}$ the mean prediction error was 37.1 percent.

For Netboost the multivariate filter was stopped after 20 steps, resulting in a filter of 247,497 network edges. This represents approximately 0.06 percent of the edges. Based on this Netboost identified 106 modules with an average module size of 46 in the range of 10 to 561. Accordingly 17 percent of the features were assigned to modules. Eigengenes of the Netboost modules explained a higher proportion of variance (median = 66.2%, range = [52.3%, 84.9%]). On the Netboost aggregated $X_{\text{Netboostmodules}}$ the mean prediction error was 28.2 percent. The dendrogram based on the sparse network is depicted in Fig. 9 and illustrates that many Netboost modules correspond to the central parts of WGCNA modules. As shown for 25, 20 and 15 steps the clustering is stable under the choice of boosting steps.

Two-sample tests for equality of proportions with continuity correction showed significant differences in means of prediction errors with Netboost errors being smaller than direct application (p-value = 0.019) and WGCNA (p-value < 2.2e-16) and direct application errors being smaller than WGCNA (p-value < 2.2e-16).

4 DISCUSSION

Netboost is designed in an unsupervised manner. While a supervised approach to module detection would possibly further improve accuracy, these approaches are complementary.

In [36] another approach for data with correlated features is proposed. The first step is to cluster the features, and then choose a cluster representative based on prediction performance. The second step is to apply either lasso or marginal significance testing on these representatives. Apart from differences in the applied clustering technique as with other supervised clustering techniques this might lead to improved predictive performance but hinder the interpretation of the selected clusters. Here, the primary aim is to maximize predictive power and thereby optimize the algorithm for biomarker detection. This is done at the cost of potentially

introducing some form of bias. Dependent on size and connectivity of the module the supervised selection might pick up peripheral features - voiding their function as a representative.

Thereby, we are hindered if we are interested in the biological interpretation of identified biomarkers in context of the network. Keeping outcome and network detection separate allows for unbiased interpretation of any potential connections between subsequently in the primary analysis selected modules and the outcome.

In [37] two extensions to sparse canonical correlation analysis (CCA) [38] are introduced. First they propose a supervised form of sparse CCA and second they generalize the framework from two to multiple datasets. With this they offer a framework for identification of sparse linear combinations of the multiple sets of features that are highly correlated with each other and associated with the outcome. While Netboost can also identify cross-omics correlations associated with the outcome, [37] omit within datatype connections and optimize their algorithm solely for cross-dataset combinations.

Starting with the WGCNA methodology our original design is the introduction of the filtering step before constructing the correlation-based network. Here, we chose a boosting based edge detection to allow for efficient selection of essential edges. By introduction of this sparsity to the network we modified the TOM based distance and replaced UPGMA with the sparse UPGMA by Loewenstein *et al.* [27]. As Netboost is still based on the Pearson correlation coefficient and PCA based dimension reduction Netboost and WGCNA share many of the same advantages and drawbacks due to their similarity in design.

Having a single representative for each cluster might be an advantage for biomarker identification. Another approach to consider for this purpose are hub genes replacing the eigengenes we applied, as discussed in [39] and [40]. A hub gene is the most central node with the highest connectivity of the module as opposed to a summary measure and therefore allows cost efficient replication and application as a biomarker [41]. Eigengenes might be superior in mechanistic studies, exploratory studies and the identification of previously unknown biological features.

In general for dimension reduction of modules, eigengenes optimize explained variance with respect to the predefined

dimensionality. In our applications we fixed this to one to achieve comparability to WGCNA. However, our R package is more flexible than WGCNA and allows for the optional export of the first i PCs with a fixed i or for each module the first j PCs which cumulatively explain at least $xy\%$ of variance. Principal component based dimension reduction works particularly well if features have linear relationships. If features in modules would have non-linear relationships other more flexible dimension reduction techniques such as autoencoders [42], [43] might be more suited for calculation of aggregate measures. As the proposed module detection is ultimately based on the Pearson correlation coefficient which measures the strength of the linear relationship PCA based aggregation of modules is well defined as is observed by the high proportion of variance explained. If non-linear relationships between the features are of interest a complementary method would be required. A first step towards higher flexibility of dependencies is implemented in the R package, a non-parametric adaptation of Netboost using Spearman's ρ or Kendall's τ as the underlying correlation coefficient, the respective tests of independence to define the filter and robust principal components [44] as aggregate measures.

In Netboost feature-wise distances are defined based on Pearson correlation coefficients, e.g., [39] constructs networks based on partial correlations. In the form of Gaussian graphical models (GGM) partial correlations are frequently applied for network construction [45], [46]. In [47] GGMs are combined with a filtering step to exclude insignificant edges from the network much like Netboost. Partial correlations adjust for other variables in the network and identify the independent connections between features. In contrast, in Netboost we integrate indirect connections even further by the TOM. This is done to identify interacting subgroups irrespective of whether this interaction is direct or indirect. The focus lies on modules rather than on the individual edges, and the incorporation of indirect connections further stabilizes module detection.

As with GGMs a prime area of application beyond gene expression and DNA methylation is metabolome and proteome data. Due to their inherent co-regulation structures they offer themself to a network based analysis as was recently successfully demonstrated by combination of WGCNA on proteome data and subsequent genome-wide association studies in [48] and by combination of Netboost on metabolome data and subsequent genome-wide association studies in [49].

Datatype specific features, like dependency of CpG sites in close proximity, are not incorporated as a-priori information in Netboost. While this could inform the network, we prefer a universal design for omics data in general. Therefore, a known biological nexus can be used for module evaluation as was done in Section 3.1.

In [50] the authors introduce Net-Cox which also introduces network theory to improve survival prediction in a high-dimensional context. In contrast to our combination of Netboost and Coxboost, they introduce the estimated gene co-expression structure directly to the penalty term of the Cox model. Net-Cox is thereby inherently designed for survival analysis, whereas Netboost is more flexible in its application.

In Section 3.2 WGCNA might be improved by tailoring parameters to this applications. We choose this dataset as

WGCNA was already successfully applied to it [12] and applied both, Netboost and WGCNA, with its standard settings as recommended by the authors [9] to achieve a fair comparison. While we know from [12] that WGCNA extracts relevant information of this dataset, for the task of subtype classification the WGCNA clustering superimposes the signals and Netboost kept a more compartmentalized and detailed network with the standard setting due to the applied filtering step.

In the shown applications we prefer specificity over sensitivity with respect to the clusterings. While it might be acceptable to miss an additional feature being part of a module we want to be confident regarding the selected features. Consistent with this, we deem the Jaccard Index as more meaningful to our applications as most features are unrelated. As shown in Fig. 6 Netboost is more robust than any of the competing clustering algorithms, when compared to WGCNA by the adjusted Rand Index and the Jaccard Index. To an extent it trades some sensitivity for specificity as expected by the concept of integrating a filter on the network edges. With the adjusted Rand Index and Jaccard Index we chose basic measures of stability, which are especially reliable as they are used in a comparative fashion in identical resampling setting with different methodologies. When trying to assess stability of multi-level clustering structures [51] or specific modules [52] other measures might be more suitable.

The applied clustering technique on the dendrogram in WGCNA and Netboost was introduced and compared to other methods in [20]. We kept the same cutting method to improve comparability. Primary applications of WGCNA are related to identification of the network structure itself. With respect to this we regard Netboost as a complementary approach designed to improve analysis including feature selection.

Netboost introduces the number of boosting steps as a parameter. This number can be chosen high as overfitting in filter estimation would only result in a less stringent filter rather than bias. Possible extensions include a probing based stopping criterion in the boosting step. Boosting could be stopped by inserting uninformative features without the need to perform cross validation or a fixed number of boosting steps during generation of the filter. This would automate the choice of boosting steps, while circumventing the often extensive additional computational load of cross-validation. Another direction to extend Netboost is inclusion of unclustered features which are currently ignored in the primary analysis. This implies that isolated singular features can not achieve a significant impact on the outcome. This is not true for all settings. In the primary analysis X_{modules} could be combined with a filtering method on the unclustered features.

5 CONCLUSION

With Netboost we present an efficient dimension reduction technique based on boosting and weighted gene co-expression networks distributed as a Bioconductor R package. By introducing the boosting-based filter integrated with the TOM and sparse hierarchical clustering combined with the dynamic tree cut procedure we were able to extend efficiency and predictive performance simultaneously. Additionally,

in contrast to black-box algorithms the extracted network structures in form of modules compress biological interactions to relative small and homogenous groups of features that exhibit a comprehendible complexity.

In the gene expression and DNA methylation examples this resulted in a 559- (TCGA AML) and 264-fold (HD) reduction of features for the primary analyses. Choosing eigengenes as summary measures we maximized the explained variance within each modules without an assumptions-based extension to keep Netboost applicable to a diverse set of biological experiments and primary analysis strategies. Here, we displayed applications to *in vivo* DNA methylation array, RNA array and RNA-seq measurements from patient and mouse samples. Paired with the clustering reflecting biological structures this leads to improvements in highdimensional survival analysis as well as in highdimensional classification.

In the Section 3.1 the molecular prediction was improved by identification of a surrogate for clinical information within the molecular data and by the identification of hematopoietic genes and genes encoding chromatin-modifying enzymes. In this application we were able to first abstract new features from the highdimensional data (modules), demonstrate a higher robustness than state-of-the-art alternative methods (cross-validation prediction errors) and validate the discovered correlates in an independent dataset (phase II AMLSG 12-09 study). Numerous of these genes have been suspected to play a role in AML pathogenesis before [7], [53]. Overall, the identified signature is promising for future research regarding AML pathogenesis and as a prognostic/predictive marker. Furthermore, the association with chromatin-modifying enzymes could be replicated in an independent DNA methylation data set from the phase II AMLSG 12-09 clinical trial [33] despite no available gene expression measurements. In the AMLSG 12-09 study, the effect of substituting cytarabine by the DNA methyltransferase inhibitor 5-azacitidine in AML induction therapy was studied. This trial tested the hypothesis that 5-azacytidine might reduce failure rates of intensive induction therapy particularly in AML patients with unfavorable genetic features. It is of interest that validation of the chromatin associated module was successful in this independent AML patient DNA methylation data set although the distribution of genetic aberrations in patients treated within the AMLSG 12-09 trial differed considerably from AML patients of the TCGA data set. Particularly, patients with core-binding factor AML, AML with mutated NPM1, and AML with FLT3 internal tandem duplication were excluded in this trial.

In the Section 3.2 Netboost outperformed the two other approaches and achieved the lowest prediction error. In direct comparison to WGCNA, Netboost kept more compartmentalized networks with eigengenes better reflecting their respective module and these eigengenes exhibiting stronger associations with variables of interest. Additionally, separation between, network detection and association with the trait of interest, allows for unbiased analysis and interpretation of the obtained structural information. Due to this, our biological understanding of these complex diseases and experiments might benefit from the increase in prediction accuracy and added information via the extracted network.

Netboost offers a versatile statistical modeling strategy for high-dimensional data. Due to its unsupervised design the returned lower-dimensional data matrix can effortlessly be integrated with a wide variety of analyses as shown for time-to-event analyses, classification and genetic associations studies [49] and across a wide variety of data types, from DNA methylation over transcriptomics to metabolomics.

List of Abbreviations

AML: Acute Myeloid Leukemia
CAG: Trinucleotide of Cytosine, Adenine and Guanine
CpG: 5'—Cytosine—phosphate—Guanine—3'
DNA: Deoxyribonucleic Acid
GGM: Gaussian Graphical Models
HD: Huntington's Disease
miRNA: micro RNA
mRNA: messenger RNA
PC: Principal Component
RF: Random Forest
RNA: Ribonucleic Acid
TCGA: The Cancer Genome Atlas
TOM: Topological overlap measure
UPGMA: Unweighted Pair Group Method with Arithmetic Mean
WGCNA: Weighted Gene Co-expression Network Analysis

Authors' Contributions

PS, RC, ML wrote the manuscript, PS, MaS, CP, RC, ML analyzed data, PS, HB, MS developed the method, PS, JK developed the R package, PS, RC, ML, MS interpreted results, KD, LB supervised the AMLSG 12-09 study, MS supervised the research project, and all authors read and approved the final manuscript.

Availability of Data and Materials

The datasets supporting the conclusions of this article are available in the GDC Data Portal repository, TCGA-LAML data [https://portal.gdc.cancer.gov/projects/TCGA-LAML] and the Gene Expression Omnibus [GSE65776].

REFERENCES

[1]   Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm344">+http://dx.doi.org/10.1093/bioinformatics/btm344

[2] D. Sobieszkoda et al., "MGMT promoter methylation as a potential prognostic marker for acute leukemia," *Archives Med. Sci.*, vol. 13, no. 6, pp. 1433–1441, Oct. 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5701700/

[3] C. Gardin and H. Dombret, "Hypomethylating agents as a therapy for AML," *Curr. Hematol. Malignancy Rep.*, vol. 12, no. 1, pp. 1–10, Feb. 2017. [Online]. Available: https://doi.org/10.1007/s11899–017-0363-4

[4] X. Thomas and C. Le Jeune, "Treatment of elderly patients with acute myeloid leukemia," *Curr. Treatment Options Oncol.*, vol. 18, no. 1, Jan. 2017, Art. no. 2. [Online]. Available: https://doi.org/10.1007/s11864–017-0445-5

[5] E. Papaemmanuil et al.,"Genomic classification and prognosis in acute myeloid leukemia," *N. Engl. J. Med.*, vol. 374, no. 23, pp. 2209–2221, Jun. 2016.

[6] J. Prada-Arismendy, J. C. Arroyave, and S. Röthlisberger, "Molecular biomarkers in acute myeloid leukemia," *Blood Rev.*, vol. 31, no. 1, pp. 63–76, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268960X16300601

[7] J. Yamazaki et al., "Hypomethylation of TET2 target genes identifies a curable subset of acute myeloid leukemia," *J. Nat. Cancer Inst.*, vol. 108, no. 2, Feb. 2016, Art. no. djv323.

[8] K. Meldi et al., "Specific molecular signatures predict decitabine response in chronic myelomonocytic leukemia," *J. Clin. Invest.*, vol. 125, no. 5, pp. 1857–1872, May 2015.

[9] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, 2008, Art. no. 559.

[10] K. R. van Eijk et al., "Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects," *BMC Genomics*, vol. 13, 2012, Art. no. 636.

[11] S. Horvath et al., "Aging effects on DNA methylation modules in human brain and blood tissue," *Genome Biol.*, vol. 13, no. 10, 2012, Art. no. R97.

[12] P. Langfelder et al., "Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice," *Nat. Neurosci.*, vol. 19, no. 4, pp. 623–633, Apr. 2016.

[13] H. Binder and M. Schumacher, "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models," *BMC Bioinformatics*, vol. 9, 2008, Art. no. 14.

[14] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[15] G. Tutz and H. Binder, "Generalized additive modeling with implicit variable selection by likelihood-based boosting," *Biometrics*, vol. 62, no. 4, pp. 961–971, Dec. 2006.

[16] T. Hastie and R. Tibshirani, *Generalized Additive Models*. New York, NY, USA: Taylor & Francis, 1990. [Online]. Available: https://books.google.de/books?id=qa29r1Ze1coC

[17] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statist. Appl. Genetics Mol. Biol.*, vol. 4, 2005, Art. no. 17.

[18] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002.

[19] R. Sokal, C. Michener, and U. of Kansas, *A Statistical Method for Evaluating Systematic Relationships*. Lawrence, KS, USA: Univ. Kansas, 1958. [Online]. Available: https://books.google.de/books?id=o1BlHAAACAAJ

[20] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R," *Bioinformatics*, vol. 24, pp. 719–720, 2008.

[21] H. Binder, "CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks," 2013. [Online]. Available: https://CRAN.R-project.org/package=CoxBoost

[22] C. Porzelius, H. Binder, and M. Schumacher, "Parallelized prediction error estimation for evaluation of high-dimensional models," *Bioinformatics*, vol. 25, no. 6, pp. 827–829, Mar. 2009.

[23] H. Binder and M. Schumacher, "Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples," *Statist. Appl. Genetics Mol. Biol.*, vol. 7, no. 1, Jan. 2008, Art. no. 12.

[24] M. Schumacher, H. Binder, and T. Gerds, "Assessment of survival prediction models based on microarray data," *Bioinformatics*, vol. 23, no. 14, pp. 1768–1774, Jul. 2007.

[25] T. A. Gerds, T. Cai, and M. Schumacher, "The performance of risk prediction models," *Biometrical J.*, vol. 50, no. 4, pp. 457–479, Aug. 2008.

[26] T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu, "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring," *Statist. Med.*, vol. 32, no. 13, pp. 2173–2184, Jun. 2013.

[27] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial, "Efficient algorithms for accurate hierarchical clustering of huge datasets: Tackling the entire protein space," *Bioinformatics*, vol. 24, no. 13, pp. 41–49, 2008. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/24/13/i41.abstract

[28] T. Therneau and P. Grambsch, *Modeling Survival Data: Extending the Cox Model*. Berlin, Germany: Springer, 2000. [Online]. Available: https://books.google.de/books?id=9kY4XRuUMUsC

[29] P. M. Grambsch and T. M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, vol. 81, no. 3, pp. 515–526, Sep. 1994. [Online]. Available: https://doi.org/10.1093/biomet/81.3.515

[30] S. B. Brant, "Dynamic survival prediction for high-dimensional data," Master's thesis, Dept. Math., Univ. Oslo, Oslo, Norway, 2018. [Online]. Available: https://www.duo.uio.no/bitstream/handle/10852/63417/simon_boge_brant_thesis.pdf

[31] S. Wagner and D. Wagner, "Comparing clusterings - An overview," *Karlsruhe, Tech. Rep. 4*, 2007. [Online]. Available: https://publikationen.bibliothek.kit.edu/1000011477

[32] U. von Luxburg, "Clustering stability: An overview," *Found. Trends Mach. Learn.*, vol. 2, no. 3, pp. 235–274, 2010. [Online]. Available: http://dx.doi.org/10.1561/2200000008

[33] R. F. Schlenk et al.., "Randomized phase-II trial evaluating induction therapy with idarubicin and etoposide plus sequential or concurrent azacitidine and maintenance therapy with azacitidine," *Leukemia*, vol. 33, pp. 1923–1933, Feb. 2019.

[34] B. Lehne et al., "A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies," *Genome Biol.*, vol. 16, Feb. 2015, Art. no. 37.

[35] P. Langfelder et al., "MicroRNA signatures of endogenous huntingtin CAG repeat expansion in mice," *PLoS One*, vol. 13, no. 1, pp. 1–20, Jan. 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0190550

[36] S. Reid and R. Tibshirani, "Sparse regression and marginal testing using cluster prototypes," *Biostatistics*, vol. 17, no. 2, pp. 364–376, 2016. [Online]. Available: http://dx.doi.org/10.1093/biostatistics/kxv049

[37] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Statist. Appl. Genetics Mol. Biol.*, vol. 8, 2009, Art. no. 28.

[38] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, Jul. 2009.

[39] D. Yu, J. Lim, X. Wang, F. Liang, and G. Xiao, "Enhanced construction of gene regulatory networks using hub gene information," *BMC Bioinformatics*, vol. 18, no. 1, p. 186, Mar. 2017. [Online]. Available: https://doi.org/10.1186/s12859–017-1576-1

[40] S. Das, P. K. Meher, A. Rai, L. M. Bhar, and B. N. Mandal, "Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: An application to aluminum stress in soybean (Glycine max L.)," *PLoS One*, vol. 12, no. 1, pp. 1–24, Jan. 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0169605

[41] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS Comput. Biol.*, vol. 4, no. 8, pp. 1–27, Aug. 2008. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1000117

[42] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biol. Cybern.*, vol. 59, no. 4/5, pp. 291–294, 1988.

[43] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, *arXiv:1312.6114*. [Online]. Available: 00tlt00-tran-2989333

[44] I. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer, 2011, pp. 1094–1096. [Online]. Available: https://doi.org/10.1007/978–3-642-04898-2_455

[45] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis, "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data," *BMC Syst. Biol.*, vol. 5, no. 1, Jan. 2011, Art. no. 21. [Online]. Available: https://doi.org/10.1186/1752-0509-5-21

[46] Y. Xie, Y. Liu, and W. Valdar, "Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics," *Biometrika*, vol. 103, no. 3, pp. 493–511, 2016. [Online]. Available: http://dx.doi.org/10.1093/biomet/asw035

[47] S. Lee, F. Liang, L. Cai, and G. Xiao, "Integrative analysis of gene networks and their application to lung adenocarcinoma studies," *Cancer Informat.*, vol. 16, 2017, Art. no. 1176935117690778. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5392014/

[48] V. Emilsson *et al.*, "Co-regulatory networks of human serum proteins link genetics to disease," *Science*, vol. 361, no. 6404, pp. 769–773, Aug. 2018.

[49] P. Schlosser *et al.*, "Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans," *Nat. Genetics*, vol. 52, pp. 167–176, Jan. 2020. [Online]. Available: https://www.nature.com/articles/s41588-019-0567-8

[50] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang, "Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment," *PLoS Comput. Biol.*, vol. 9, no. 3, pp. 1–16, Mar. 2013. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002975

[51] A. Bertoni and G. Valentini, "Discovering multi-level structures in bio-molecular data through the Bernstein inequality," *BMC Bioinformatics*, vol. 9 Suppl 2, Mar. 2008, Art. no. S4.

[52] P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath, "Is my network module preserved and reproducible?" *PLoS Comput. Biol.*, vol. 7, no. 1, Jan. 2011, Art. no. e1001057.

[53] R. Rampal *et al.*, "DNA hydroxymethylation profiling reveals that WT1 mutations result in loss of TET2 function in acute myeloid leukemia," *Cell Rep.*, vol. 9, no. 5, pp. 1841–1855, Dec. 2014.

Maximilian Schmutz received the MD degree in medicine from the University of Heidelberg, Germany, in 2014. He received training in academic research at the German Cancer Research Center, Heidelberg, Germany, from 2011 to 2012 and was awarded the Heinrich F.C. Behr research scholarship. At present, he is a physician at the Department of Hematology and Oncology, Augsburg University Medical Center, Germany. His research interests include epigenetics and clonal evolution in leukemogenesis, development of epigenomic biomarkers in hematologic neoplasms, liquid biopsy as tool for therapy monitoring in solid neoplasias and development of personalized therapy strategies in oncology.

Konstanze Döhner received the MD degree in medicine at the University Medical School in Heidelberg, Germany, in 1992. From 1995 to 1997 she spent a postdoctoral fellowship at the German Cancer Research Center in Heidelberg, Germany, and the Hospital for Sick Children, the Department of Genetics Toronto, Canada. In 2000 she moved to the Department of Hematology/Oncology at the University Hospital Ulm, Germany. She became head of the laboratory for Cytogenetic and Molecular Diagnostics in myeloid leukemias. In 2003 she received her board certification as Hematologist and Oncologist and in 2005 she became professor. In 2018 she was elected as a board member of the European Hematology Association (EHA); in 2019 as a member of the EHA Education committee, and as chair of the EHA Scientific Working Group for AML. In addition, she is member of the European LeukemiaNet MRD working party. Beside AML she has clinical and scientific interest in myeloproliferative neoplasms and cofounded the German Study Group on Myeloproliferative Neoplasms (GSG-MPN). She is involved in a number of national and international scientific cooperations which is also reflected by numerous highly ranked publications. Her main scientific interest is the molecular characterization of AML and the translation of her findings into clinical studies.

Christoph Plass received the PhD degree in biology from the Institut für Biologie der Medizinischen Universität, Lübeck, Germany, in 1993. He was a postdoctoral researcher and cancer research scientist with the Roswell Park Cancer Institute, Buffalo, New York from 1993 to 1997, and an assistant, associate and full professor with the Ohio State University, Columbus, Ohio from 1997 until 2007. Currently, he is division head of the Department of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. His main research interests include the area of epigenetics, genetics, prostate cancer, CLL, and AML.

Pascal Schlosser received the BSc and MSc degrees in mathematics from the University of Freiburg, Germany, in 2011 and 2014, respectively, and the PhD degree from the Institute of Medical Biometry and Statistics, Freiburg, Germany, in 2019. For extraordinary achievements during his studies he was awarded a scholarship by the German Academic Scholarship Foundation and the Young Statisticians Award by the German Region of the International Biometrical Society, in 2010 and 2015, respectively. At present, he is a postdoctoral fellow at the Institute of Genetic Epidemiology, Freiburg, Germany and his research interests include large-scale network analysis, survival analysis, variable selection, and the statistical modeling and analysis of complex medical systems.

Jochen Knaus received the Dipl degree in the history of modern German literature, computer science and psychology, in 2003 and worked as research assistant at the University of Freiburg, Germany until 2006. Currently, he is working as a computer scientist at the Institute of Medical Biometry and Statistics, Freiburg, Germany and as the head of research data management of the Faculty of Medicine - University of Freiburg, Germany, since 2006 and 2019, respectively.

Lars Bullinger received the MD degree from the Heidelberg Medical School, Germany, in 1999. He was appointed as professor in Ulm and Berlin, in 2011, and 2017, respectively. During his residency/fellowship within the Department of Internal Medicine, University of Heidelberg, Heidelberg, Germany and the University Hospital Ulm, Ulm, Germany, he stayed at the Stanford University, Stanford, California, and the Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts for a postdoctoral fellowship and a sabbatical, respectively. He returned to the University Hospital Ulm, Ulm, Germany, in 2011, where he continued his physician scientist career as an attending physician, principal investigator of clinical trials and group leader of a translational molecular genomic research group. He received different awards and honors including the Franziska Kolb Award for leukemia research and the Heisenberg scholarship, in 2005 and 2011, respectively. Currently, he is holding a position as the director of the Department of Hematology, Oncology, and Tumor- immunology (CVK), Charité University Medicine, Berlin, Germany. His research interests include the molecular mechanisms underlying leukemia biology and the delineation of genetic and epigenetic alterations of prognostic and predictive relevance for leukemia management.

Rainer Claus received the MD degree from the Medical Center - University of Freiburg, Germany, in 2003. After his residency within the Department of Hematology and Oncology, Medical Center - University of Freiburg, Germany, he stayed at The Ohio State University, Columbus, Ohio and the German Cancer Research Center (DKFZ), Heidelberg, Germany for a postdoctoral fellowship (DFG scholarship) until 2012. He completed his board certification for internal medicine and hematology/oncology, in 2015. He coauthored more than 80 scientific publications in peer-reviewed journals with a research focus on epigenomics in leukemogenesis and CLL pathogenesis, specifically. He received different awards and honors including the Artur-Pappenheim-Award of the DGHO, in 2012. Currently, he is holding a position as attending physician and research group leader at the Department of Hematology/Oncology, Augsburg University Medical Center, Germany. His research interests include epigenetic mechanisms of gene (dys-)regulation in leukemogenesis and the implementation of comprehensive genomic/epigenomic profiling in tumor specimens and body fluids aiming at improvement of molecularly guided precision medicine.

Harald Binder received the Dipl degree from the University of Regensburg, Germany, the MA degree from the University of California, Irvine, and the PhD degree from the Ludwig Maximilian University of Munich, Germany, in 2000, 2001, and 2006, respectively. He was appointed as professor in Mainz and Freiburg, in 2011 and 2017, respectively. He coauthored more than 150 scientific publications. He currently is incoming president of the German Association for Medical Informatics, Biometry, and Epidemiology (GMDS e.V.). At present, he is director of the Institute of Medical Biometry and Statistics, Freiburg, Germany and his research interests include integrative modeling of molecular and clinical data, and machine learning techniques in particular deep learning with limited sample sizes.

Michael Lübbert received the MD degree from the University of Hamburg, Germany, in 1985, then joined the group of professor H. Phillip Koeffler, the UCLA School of Medicine, Division of Hematology/ Oncology, Los Angeles, from 1986 to 1988. He received the venia legendi, in 1995, his Board Certifications for Internal Medicine and Hematology/ Oncology, in 1996 and 1998, respectively, was appointed attending physician, in 1996 and applied professor, in 2003 at the Medical Center - University of Freiburg, Germany. He has coordinated investigator-initiated clinical trials with epigenetically active agents in AML and myelodysplastic syndromes since 2003. He served as secretary from 2006-2012, and chairman since 2018, of the EORTC Leukemia Cooperative Group, and since 2008 as coordinator, together with Prof. C. Plass, of two German Research Foundation-funded multicenter Consortia with a focus on epigenetics and myeloid neoplasias. He has coauthored almost 300 scientific publications and is editor of several books including "*Epigenetic Therapy of Cancer*" (Springer, 2014). His research interests include mechanisms of myeloid differentiation, genetic and epigenetic alterations in leukemias and lung cancer, and in vitro and in vivo mechanisms of action of chromatin-modifying agents.

Martin Schumacher received the Dipl and PhD degrees from the University of Dortmund, Germany, in 1974 and 1977, respectively. He was appointed as professor in Dortmund, Seattle and Freiburg, in 1983, 1984 and 1986, respectively. From 1986 until 2017 he was head of the Institute of Medical Biometry and Statistics, Freiburg, Germany and cofounded the Freiburg Center for Data Analysis, in 1994 and one of the first coordination centers for clinical studies in Germany, in 1999. He coauthored more than 250 scientific publications. Among the many honors he received are the Bradford Hill Memorial Lecture, the honorary membership of the German Region of the International Biometrical Society and the honorary membership of the International Society for Clinical Biostatistics, in 2012, 2014, and 2015, respectively. At present, he is a professor emeritus in Medical Biometry and Statistics at the Institute of Medical Biometry and Statistics, Freiburg, Germany and his research interests include survival analysis, statistical modeling of clinical and epidemiological studies, nonparametric methodology and statistical classification.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.