Netboost: Statistical modeling strategies for high-dimensional data.

Dissertation zur Erlangung des Doktorgrades

vorgelegt von

Pascal Schlosser

an der Fakultät für Mathematik und Physik

der Albert-Ludwigs-Universität

Freiburg

Dean:	Prof. Dr. Gregor Herten Institute of Physics, University of Freiburg Hermann-Herder-Straße 3 79104 Freiburg, Germany
1. Reviewer:	Prof. Dr. Martin Schumacher Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg Stefan-Meier-Straße 26 79104 Freiburg, Deutschland
2. Referent:	Prof. Dr. Berthold Lausen Department of Mathematical Sciences, University of Essex, STEM 5.5, Colchester Campus, Essex, United Kingdom
Datum der Promotion:	18. November 2019

Acknowledgment

First and foremost, I thank Martin Schumacher for his great support, confidence and the opportunity to write this dissertation.

I am grateful to Michael Lübbert, who opened a window into clinical science for me and to Anna Köttgen, whose enthusiasm for research and hunger for a better understanding of our world motivated my choice to embark a scientific career myself.

I thank Edgar Brunner for constructive discussions that helped to shape the chapter on robust extensions.

I had the pleasure to work with many wonderful colleagues during my dissertation. I deeply appreciate all the discussions, that are at the hearth of our research and the support, which helped me to become a better scientist. Specifically, I would like to name: Franziska Grundner-Culemann, Gabriele Greve, Jochen Knaus and Nadja Blagitko-Dorfs.

Being blessed with some of my closest friends having an interest in Biostatistics themselves, I am very grateful to Daniela Zöller, Maja von Cube and Maren Hackenberg for proof reading my dissertation.

Lastly, I wish to thank my parents Doris and Norbert Schlosser and my wife Anna Schlosser, who are the strength behind all my endeavors in life.

Statement of Authorship

Except where reference is made in the text of this dissertation, this dissertation contains no material published elsewhere or extracted in whole or in part from a dissertation presented by me for another degree or diploma. No other person's work has been used without due acknowledgement in the main text of the dissertation. This dissertation has not been submitted for the award of any other degree or diploma in any other tertiary institution.

Some parts of this dissertation have been submitted to peer-reviewed journals:

- Parts of the article by Schlosser et al. [**1**], are stated in Chapter 1, Chapter 2, Chapter 3, Chapter 5, Chapter 6 and Chapter 8
- Parts of the article by Schlosser et al. [**2**], are stated in Chapter 4 and Chapter 8

While related and cited, none of the results in [**3**], [**4**], [**5**] and [**6**] are presented in this dissertation. I am a first author of [**4**], [**5**], [**1**] and [**2**] and responsible for concept, analysis and paper of these studies. I am a co-author of [**3**] and [**6**] and responsible for part of the analysis and paper of these studies. A complete list of accepted peer-reviewed publications in which I participated can be found at `https://orcid.org/0000-0002-8460-0462`. This list at the time of writing (January, 1th 2019) is given in the Appendix in *Complete list of peer-reviewed publications*.

Apart from peer-reviewed articles, parts of this dissertation have been presented at conferences:

- DAGStat 2016, Göttingen, Germany, oral presentation, related to Chapter 2: 2

- Statistical Computing 2016, Reisensburg, Germany, oral presentation, related to Chapter 2: 2.1 and Chapter 3
- Statistical Computing 2017, Reisensburg, Germany, oral presentation, related to Chapter 2: 2.5
- Genetic Epidemiology 2018, Grainau, Germany, oral presentation, related to Chapter 4: 1
- DGfN 2018, Berlin, Germany, invited oral presentation, related to Chapter 4: 2
- CHARGE 2018, Baltimore, USA, oral presentation (travel award 2500$), related to Chapter 4: 2.1
- DAGStat 2019, Munich, Germany, oral presentation, related to Chapter 6
- Genetic Epidemiology 2019, Grainau, Germany, oral presentation, related to Chapter 2: 2.5
- CHARGE 2019, St. Louis, USA, oral presentation (travel award 2600$), related to Chapter 4: 2

I am aware of the PhD regulations by the faculty of Mathematics and Physics of the Albert-Ludwig-University Freiburg, in particular I am aware that the right of the doctoral candidate to use the title of Doctor and to display the associated degree begins strictly with the handing over of the certificate.

Freiburg, November 20, 2019 _____

place, date signature

Abstract

Background. State-of-the art methods often fail to identify weak but cumulative effects of variables found in high-dimensional omics datasets. Nevertheless, these effects play important roles in many diseases, such as the clonal development of leukemic cells and chronic kidney disease (CKD) metabolism.

Results. We propose Netboost, a three-step dimension reduction technique. First, boosting-based filters are 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 principal components. The primary analysis is then carried out on these summary measures instead of the original data, allowing for a localized dimensionality reduction.

We demonstrate the application of the newly developed Netboost in integration with CoxBoost for survival prediction, genetic association studies to understand the human metabolism and random forests for disease classification. We applied our method in 7 independent cohorts spanning 6 diseases, a variety of high-dimensional data types (DNA methylation, metabolomics, miRNA, RNA arrays, RNA sequencing) and human as well as murine *in vivo* samples. In many of these settings, we were able to show significant advantages over state-of-the-art competitive analysis strategies with respect to prediction errors, power and mis-classification rates by cross-validation, general resampling and independent replication.

By integration of our novel method in analysis of several biomedical research projects, we were able to attain and confirm biological insights which could not have been reached by the compared state-of-the-art methods. In particular, the two biologically most insightful findings in this dissertation were both replicated in

independent datasets. First, we identified a chromatin modifying enzyme signature associated with overall survival, which separates patients into two groups with a threefold difference in median survival time. Second, we established the central concept in the human urinary metabolism to be the list of absorption, distribution, metabolism, and excretion (ADME) processes, which was originally defined in the context of pharmacological research.

Furthermore, we demonstrated in several datasets a lower sampling uncertainty of Netboost overall networks as well as individual components of the networks across Netboost, weighted gene co-expression network analysis (WGCNA) and k-means and found that method uncertainty dominated sampling uncertainty.

Finally, we integrate Netboost with robust methodology designing a Netboost adaption, which is invariant to monotone transformations of variables and thus obtain an advantageous extension in cases of non-linear relationships between variables.

Conclusion. The newly developed approach Netboost offers a versatile statistical modeling strategy for high-dimensional data, which is freely available as a Bioconductor `R` package. Via dimensionality reduction it improves accuracy, power and stability in various analysis settings, including time-to-event analysis, genome-wide association study (GWAS) and classification.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Microarray, sequencing and other high-throughput functional genomics technologies are developing rapidly, incorporating more and more measured variables. This has led to extensive advances in biomedical research. For example, it laid the foundation for the steady improvements of survival rates in many cancerous diseases over the past decades and more recently in drug development in general. When compared to traditional clinical trials, substances based on a genetic association with the disease have a more than two-fold chance of being approved [**7, 8**]. However, these achievements fuelled by advances in measurement technology call for appropriate data analysis methodology which is currently lagging behind for many applications. A major challenge for biomedical research when analyzing these high-dimensional datasets lies in the discrepancy between the potentially hundreds of thousands of variables to be investigated and a limited sample population in the range of tens to a few hundreds.

Many times, methods which were originally developed for the selection of a low number of clinical variables are now faced with the challenge of selecting from hundreds of thousands or even from millions of variables. An even greater challenge is imposed if no singular variable is expected to dominate an effect, but rather a larger group of variables acts cumulatively. This gives rise to dimensionality reduction techniques that can be partitioned into supervised and unsupervised algorithms. The former aim at simultaneously identifying the subspace with minimal dimensions and a good characterization of the outcome variable (e.g. [**9, 10, 11**]) and the unsupervised aim at identifying the subspace with minimal dimensions that still gives a good characterization of the original full dimensional dataset (e.g. [**12, 13**]). As adapted from [**14**], supervised methods can be categorized in filter, wrapper and

embedded methods. Filter methods first calculate a relevant score for each variable and then define a cut-off on these. Wrapper methods evaluate the performance of each subspace by application of the intended analysis. Embedded methods integrate model construction and variable selection in one step. Filter and wrapper methods can also be transfered to categorize unsupervised dimensionality reduction methods when the performance measures are chosen independently of the variable of interest. Furthermore, wrapper and embedded methods can be implemented as simple variable selection methods or extended by feature abstraction. In the later case, new variables are constructed via projection or compression to integrate the original variables more efficiently.

Supervised methods optimize dimensionality reduction with respect to the outcome of interest and can benefit prediction accuracy for this particular outcome. Unsupervised algorithms on the other hand are more versatile when multiple outcomes or the dimensionality reduction itself are of interest. Determining which variables are essential to the overall variation in the dataset and which are connected and form a group can often give insight into central aspects of the underlying biology. Detecting these structures independently of the outcome enables the researcher to comprehensively interpret them and gives weight to the subsequently identified association with the outcome.

Related to the topic of dimensionality reduction is the concept of network analysis ([**15, 16, 17, 18, 19, 20**]). Network analysis focuses on the identification of dependency strutures within the data. A network of variables, the nodes, is weighted if its edges, which connect the nodes, are coded on a continuous scale and unweighted if edges are binary. To achieve a form of local dimensionality reduction, we can integrate network and dimensionality reduction methods. When highly dependent structures are identified, these can be locally summarized by filter, wrapper or embedded algorithms and aggregated measures can be combined across all identified structures.

With the **Net**work analysis with **boost**ing-based filtering, which we dubbed **Netboost**, we propose an unsupervised procedure to reduce dimensions within high-dimensional datasets. In the context of the above categorization, Netboost is a hybrid network method, combining weighted and unweighted networks via the boosting-based filter, with an unsupervised wrapper methodology including the construction of new aggregate variables via local projections. We put a specific emphasis on large subgroups of variables, called modules, that show a shared effect. To this end, we aggregate subgroup information before applying the primary analysis strategy.

The integration of network and dimensionality reduction methodology facilitates a form of local dimensionality reduction. Specifically, aggregation takes place in local environments, the modules, within the network. E.g., in a simple network consisting of two modules, a global dimensionality reduction algorithm like principal component analysis (PCA) ([**12**]) would in many instances first aggregate the difference between the two modules, whereas Netboost would begin aggregation by identifying the main direction within each of the modules. Both approaches aggregate information which is relevant to many research questions, but they assemble distinct information.

In Netboost as a network-based dimensionality reduction method, we were initially inspired by the weighted gene co-expression network analysis (WGCNA) ([**21, 15, 22**]) and modified and extended the framework in various ways. Foremost is the addition of a multivariate filter and application of sparse hierarchical clustering to improve detection of relevant network edges. The general intent of the proposed extensions and modifications is to improve identification of modules by reduction of noise and pruning modules to their central variables. This becomes particularly important with an increasing number of dimensions and in integrative analysis across datatypes. For example, when we analyse the interplay of differing molecular levels, like gene expression and DNA methylation, these do contain interesting and relevant

crosslinks. However, also the overall proportion of non-null edges in the network is strongly reduced.

1. Illustration

To illustrate how Netboost can recover and use the embedded network structure to our advantage, we simulate correlated variables. Here, we can assume this to be a simplified example of gene regulation influencing overall survival in a study on the prognosis of patients with acute myeloid leukemia (AML). First, we generate 400 unrelated standardized genes by drawing 100 samples from 400 independent and identically distributed (i.i.d.) standard normal variables, which are not related to the variables of interest. Next, we draw 100 samples from a multivariate normal distribution of dimension 101 with marginal means of zero and all off-diagonal covariance entries equal to 0.8. For the first scenario, we define one of the dimensions as the variable of interest y_1, the other 100 dimensions are added as the first module to the dataset. This reflects a situation where the one underlying gene is causally associated with survival is not covered by the measurement platform, but a set of co-regulated genes is. Next, we draw 100 samples from a multivariate normal distribution of dimension 100 with marginal means of zero and all off-diagonal covariance entries equal to 0.6. For the second scenario, the variable of interest y_2 is defined as the sum across these 100 dimensions and reflects a cumulative effect of these co-regulated genes on survival.

In Figure 1 we display the pair-wise sample Pearson correlation coefficients and in Figure 2 the first and second principal components (PCs) of the 600 simulated variables. In these simplified simulated scenarios, we evaluate the strength of association directly with y_1 and y_2 and not some more complex endpoint related to y_1 or y_2, as this is sufficient if we assume conditional independence of the endpoint to the other variables given y_1 and y_2, respectively.

We evaluate univariate ordinary least square regression model fits with y_1 and y_2 as outcomes. All variables in module 1 and 2 are associated with y_1 and y_2,

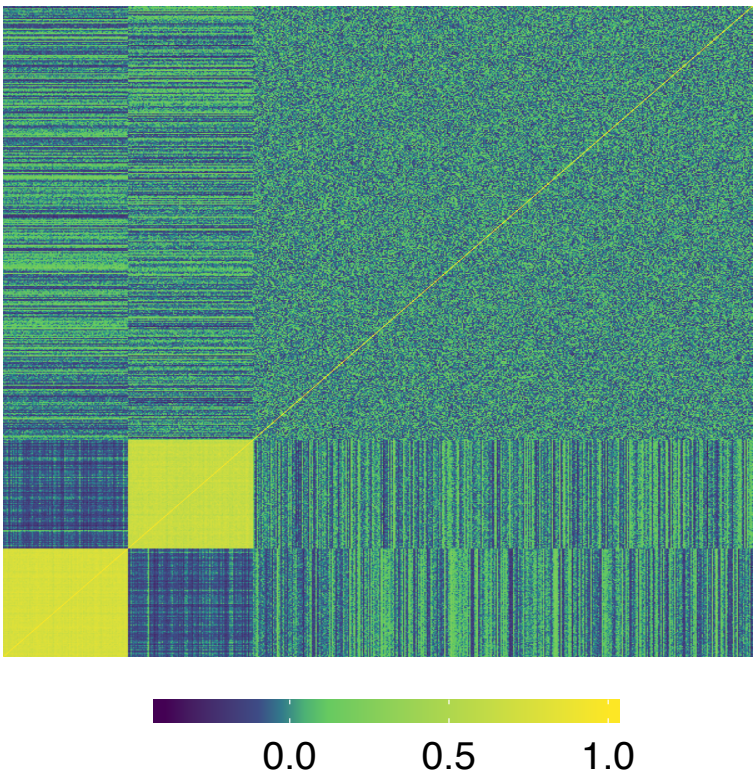

Figure 1: **Correlation matrix of simulated data.** Heatmap of the pair-wise sample Pearson correlation coefficients of 600 simulated variables.

respectively at a Bonferroni-adjusted significance level of 0.05/600. The lowest p-values were 1.0e-24 for y_1 and 1.1e-31 for y_2. While due to the strong correlation structure all univariate models already show significant associations with the respective outcome, we are able to improve this by applying Netboost. First, we identify the grouping structure in an unsupervised manner and then calculate aggregated variables for each of the identified modules. As displayed in Figure 3, Netboost correctly identifies the two existing modules in the data and sets all other variables as ungrouped. The aggregate measure for the first module combines the correlated information and exhibits a stronger association than any of the individual variables

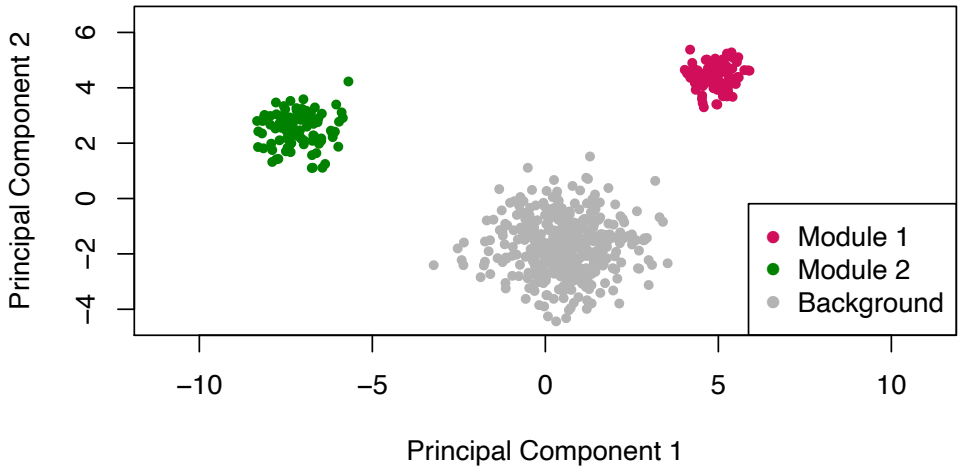

Figure 2: **First and second principal component of simulated data.**

(p-value = 9.6e-31). In case of y_2, which actually integrates information across variables, differences become even more pronounced with the p-value of the aggregated variable being 9.7e-244.

2. Structure of the presented work

As illustrated in Figure 4, this dissertation is organized as follows. Chapter 2 outlines the newly developed Netboost and describes properties and the implementation of the algorithm. In Chapter 3, Netboost is integrated with CoxBoost ([**23**]) for cross-omics time-to-event analysis, applied in 5 different cancer datasets ([**24, 25, 26**]) and evaluated against state-of-the-art alternatives ([**23, 15**]). In Chapter 4, we perform a study in the German Chronic Kidney Disease (GCKD) cohort ([**27**]) and integrate Netboost with genome-wide association studies of metabolite concentrations (mGWAS). This analysis has an inherent discovery-replication design such that all associations were replicated. We identified ADME as the major processes driving human urine metabolism and show superior power to state-of-the art single metabolite analysis.

Chapter 5 illustrates integration with random forests for classification ([**28**]) and applies this to a Huntington's Disease (HD) RNA-sequencing dataset. In Chapter 6,

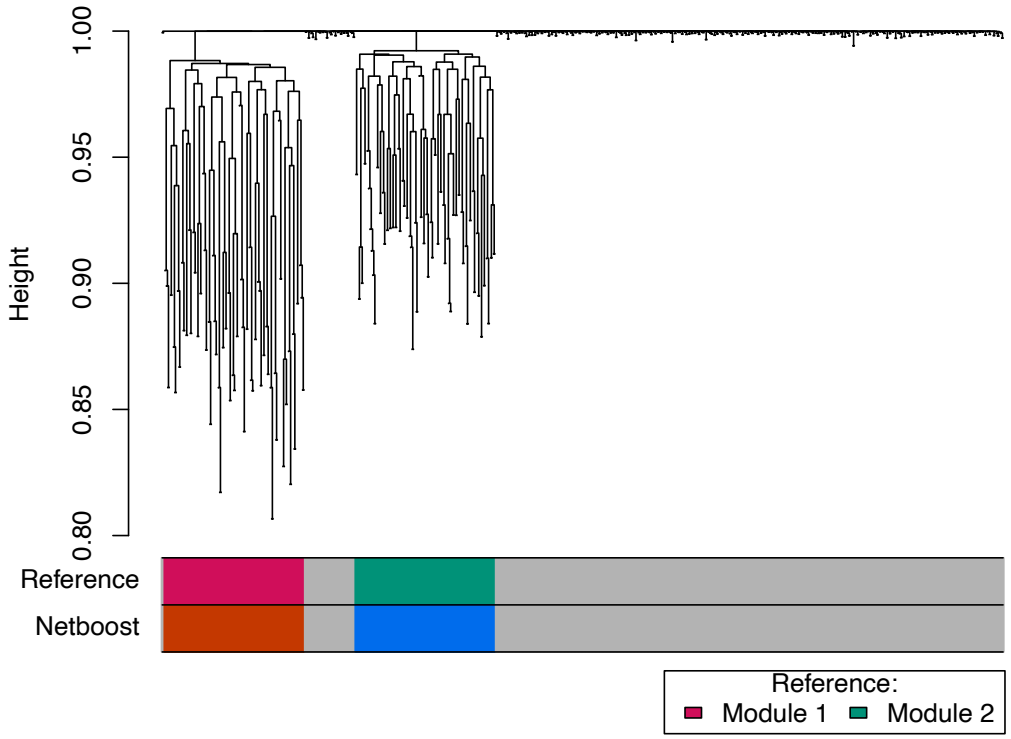

Figure 3: **Dendrogram of simulated data.** Dendrogram of 600 simulated variables. The color bands below the graph show the separation into modules with grey reflecting background variables.

preservation of the network and individual modules are studied for a variety of earlier datasets to distinguish sampling and method uncertainty and in Chapter 7, robust extensions of Netboost are introduced which are able to cover non-linear relationships between variables.

Finally, in Chapter 8, we put Netboost in context with respect to the current literature and outline advantages, limitations and future research on the proposed method to conclude this dissertation.

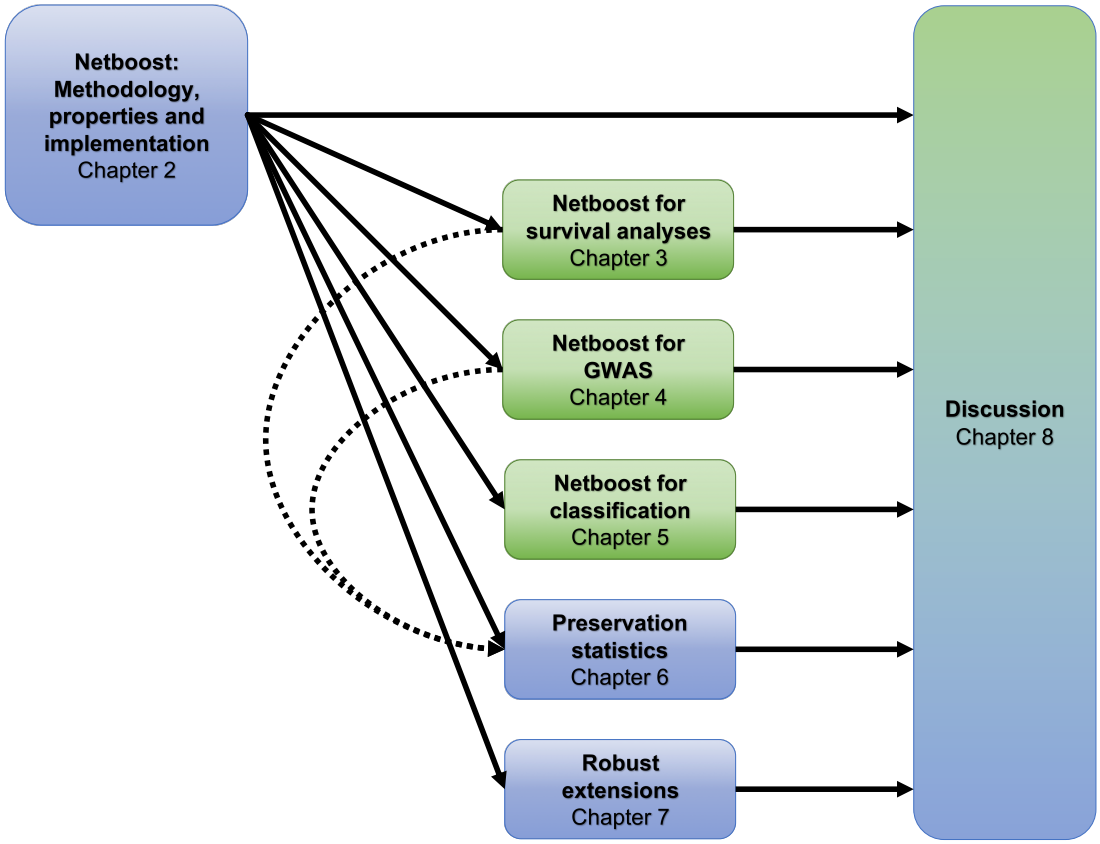

Figure 4: **Structure of the presented work.** Blue chapters focus on the methodological foundation and the properties of Netboost and green chapters lay out the integration with primary analysis strategies and its performance. Solid arrows indicate content-wise dependence and dashed arrows indicate shared data examples.

CHAPTER 2

Netboost: Network analysis with boosting-based filtering

We first introduce weighted gene co-expression network analysis (WGCNA) and its underlying theory. Based on this, we propose the methodological foundation of Netboost, show several properties of the derived measures and outline the implementation as an `R` package.

1. Weighted gene co-expression network analysis

WGCNA is a widely applied systems biology method to infer network structures, perform dimensionality reduction and study phenotypic associations of aggregated measures of variable subgroups. Development is spearheaded by the group of Steve Horvath at the University of California at Los Angeles ([**21, 29, 30, 31, 15, 22, 32**]). While the method was designed in the context of gene expression, it is suitable for most high-dimensional datasets. Prime examples of application include proteomics ([**33, 34, 35**]), metabolomics ([**3, 34**]), DNA methylation (DNAm) ([**36**]), micro ribonucleic acid (miRNA) ([**37**]) and particularly transcriptomics ([**38, 39, 40, 41, 42, 43, 44, 33**]). Additionally, this correlation-based network methodology allows for integration of complementary data types as has been done in several of the above publications ([**39, 33, 34**]).

In the following section, we describe the methodological background of WGCNA on which Netboost is based. Let X be a $n \times p$-dimensional random variable in a high-dimensional setting, where $n \ll p$ with n being the number of observations and p the number of variables. We define notation for dependent random variables for $i \in \{1, \ldots, p\}$. The vector of variable i is denoted by

$$X_i := X_{o \leq n,i}$$

and the subsetted matrix excluding variable i by

$$X_{-i} := X_{o\leq n, j\neq i}.$$

Let the image space of X_{oi} be connected and $i \in \{1, \ldots, p\}$ be a node of the network representing variable X_i. A simply network could be constructed based on the unweighted adjacency matrix defined by

$$A_{\text{unweighted}} := \begin{cases} 1 & \text{if } |\text{corr}(X_i, X_j)| > \tau, \\ 0 & \text{else,} \end{cases}$$

with corr being the Pearson correlation coefficient ([**45**]). As this hard thresholding results in a potential loss of information, we continue with a weighted design of WGCNA.

We define the adjacency matrix, A_{WGCNA}, as a measure of similarity ([**29**][1]). Let

$$A_{\text{WGCNA}} = a_{i,j\in\{1,\ldots,p\}} := |\text{corr}(X_i, X_j)|^b \tag{1}$$

with $b \in \mathbb{N}_+$. As the absolute Pearson correlation coefficient is a symmetric function and positive, it follows that A_{WGCNA} is a positive symmetric matrix. The strength of the edge between two nodes i and j can now be described by their adjacency a_{ij}. This allows for soft thresholding based on the parameter b in contrast to a hard threshold τ, which would be applied in unweighted networks.

The connectivity of node i is defined as $k_i := \sum_{j\neq i} a_{ij}$ ([**29**]). As the connectivity is a deterministic combination of random variables, it is a random variable itself. Based on the probability distribution of nodes with connectivity $k \in \mathbb{R}_+$, a network is scale-free if and only if $P(k) \propto k^{-\gamma}$ for some fixed $\gamma \in \mathbb{R}_+$. $P(k)$ is used in short for the probability of a $\mathbb{R}$-valued random variable, which assigns the connectivity to a node, taking the value k. The parameter b is then tuned to approximate a scale-free topology on the observed A_{WGCNA}. Prime examples of networks suggested to have a scale-free topology are the internet, protein-protein networks and co-authorships by mathematicians of papers which are studied in the context of Erdős numbers.

[1] While introduced in [**21**], we use the definition of [**29**] with $a_{ii} = 1$.

Given that for many datatypes we are far from measuring all nodes in the underlying biological network, this is in contrast to subnets of scale-free networks not necessarily being scale-free ([**46**]). However, the larger the mean connectivity and γ in the original network and the sampling fraction, the smaller the deviation of the random subnet will be to a scale-free network. Subsetting related to measurement technique is in many instances not random, which might lead to even more or to less severe deviations. Despite these limitations, many previous studies observe approximately scale-free topologies for genomic and other high-dimensional datatypes ([**47, 48, 49, 50, 51**]).

To evaluate the agreement of a network with the scale-free topology criterium the square correlation, $\text{corr}(\log(P(k)), \log(k))^2$, is used as means of a model fit measure for the corresponding linear regression ([**21**]). A positive slope of the regression with $\log(k)$ as the independent variable corresponds to a biologically implausible network, in the omics context, with more central nodes than peripheral nodes. Therefore, $b \in \mathbb{N}_+$ is chosen to be the minimum b for which

$$\text{corr}(\log(P(k)), \log(k))^2 > 0.85$$

and the slope is negative ([**15**]).

We define the similarity of two nodes by the topological overlap measure (TOM) ([**49, 21, 29**][2]) as

$$\text{TOM}_{ij}^{\text{WGCNA}} := \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{\min(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{uj}) + 1 - a_{ij}}.$$

We define a dissimilarity measure

$$d_{ij}^{\text{WGCNA}} := 1 - \text{TOM}_{ij}^{\text{WGCNA}}.$$

Based on d_{ij}^{WGCNA}, we use average linkage hierarchical clustering to group variables into modules. To perform this grouping, the Dynamic Tree Cut procedure is

[2]While introduced in [**49**] and corrected in [**21**] ([**49**] contained a typographical error), these publications were based on a slightly different adjacency matrix and we follow [**29**].

applied as described in [**22**]. In short, the unweighted pair group method with arithmetic mean (UPGMA) ([**52**]) is used to organize variables into a dendrogram. Next, the dendrogram is split by a static cut height. Resulting clusters are iteratively split at points where the merging height switched from increasing to a predefined sufficiently low level and the clusters are merged again if clusters are highly correlated till convergence. In opposition to a static cut height, this allows for identification of nested modules and the algorithm is suitable for automation.

An important measure of network modularity is the difference of TOM-based node connectivity based on the whole network and within modules, where TOM-based node connectivity is defined as

$$\mathrm{TOM}_i^{\mathrm{WGCNA}} := \sum_{j \in \{1,\ldots,p\}} \mathrm{TOM}_{ij}^{\mathrm{WGCNA}}.$$

The more pronounced the module separation in a network, the larger the fraction of $\mathrm{TOM}_i^{\mathrm{WGCNA}}$ is based on within-module similarities.

To summarize information of a module, PCA is used. Let $M \subseteq \mathfrak{P}(\{1,\ldots,p\})$ be a partition of nodes. For a module $m \in M$ we define

$$X^m := X_{o \leq n, j \in m}.$$

W.l.o.g. we assume each column of X^m to be standardized to mean 0 and variance 1. As the covariance matrix is symmetric and real-valued, it is diagonalizable and we can calculate orthonormal eigenvectors such that

$$(X^m)^{\mathrm{T}} X^m / (n-1) = W D W^{\mathrm{T}}$$

with W the matrix of eigenvectors and D a diagonal matrix with the eigenvalues in a decreasing order as diagonal entries. The lth principal component is given by the lth column of XW. We define the module eigengene (ME) as the first principal component

$$\mathrm{ME}_m := (XW)_{o \leq n, 1}.$$

2. Netboost

With Netboost, we propose a procedure for dimension reduction in a high-dimensional genomic context. In the general framework of WGCNA, we extend the methodology to a combination of weighted and unweighted boosting-based networks. In this semi-weighted network analysis, we integrate a filter $\mathcal{F}$ selecting relevant edges in the network which are analyzed in a weighted fashion. By this combination, we improve the noise-canceling capability of the network and increase specificity of detected modules while still utilizing the information in the underlying quantitative correlation measures.

Netboost is a three-step procedure. As shown in Figure 5, in the first step we calculate the boosting-based filter and a sparse distance matrix between variables. Therefore, we reduce the network to its essential edges. We still retain the interconnectedness and stability of complex network structures including indirect connections that occur in many omics datasets and reflect biological pathway structures. Thus, step one identifies the underlying rudimentary network (Chapter 2: 2.1).

The second step, module detection (Chapter 2: 2.2), consists of sparse hierarchical clustering ([**53**]) and the Dynamic Tree Cut procedure (Chapter 2: 1) to determine modules from the dendrogram to transfer the network into a partition of variables.

Subsequently in step three, we aggregate the information in the modules by their MEs to achieve a low-dimensional representation of the original data (Chapter 2: 2.3).

2.1. Network.

2.1.1. *Boosting-based filter.* To first identify a general structure of our network, we aggregate a filter of important network edges by linear likelihood-based boosting.

Let $Y : \mathbb{R}^{n\times(p-1)} \to \mathbb{R}^n$ be a random variable. Here, we represent the process of all variables except one in our network determining the state of this one variable for our n observations. We perform componentwise likelihood-based boosting, a forward selection-type procedure, to fit a linear approximation of Y ([**54, 55, 56**]).

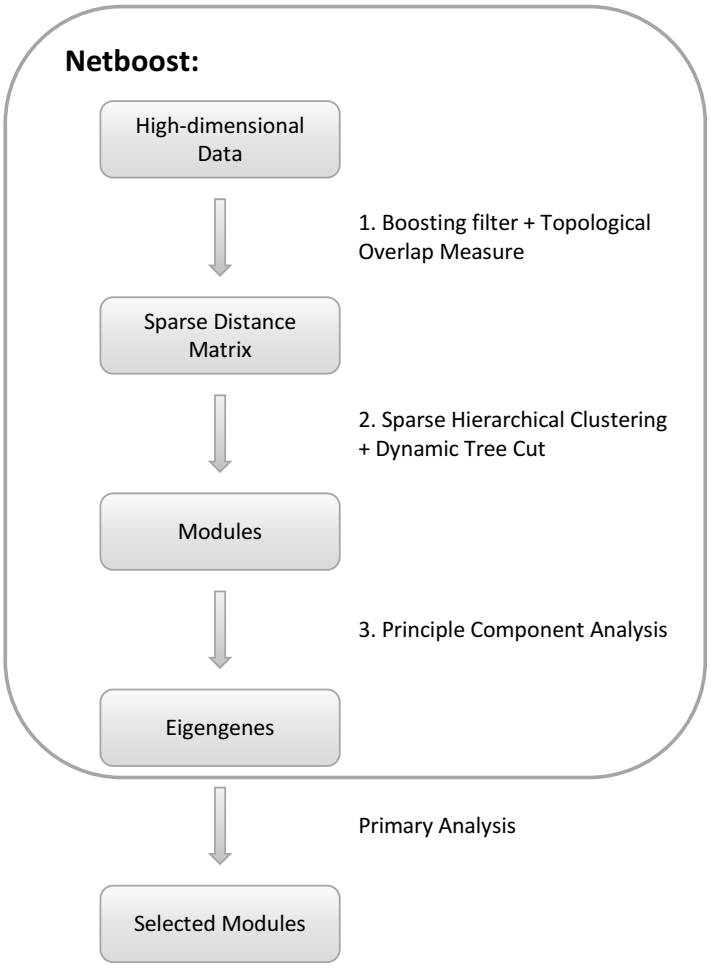

Figure 5: **Netboost concept flow chart.**

Considering the standardized X, which is assumed to be sampled from a continuous distributions, we use intercept free univariate base learners

$$h_i(X_i) := \beta_i X_i$$

with $i \in \{1, \ldots, p\} \setminus \{j\}$, $\beta_i \in \mathbb{R}$ and j indexing the dependent variable in the model.

We initialize the additive predictor with all base learners set to zero as

$$\hat{f}^{[0]} := \sum \hat{h}_i(X_i) = 0.$$

In each iteration, we fit the base learners using Fisher scoring with respect to the overall likelihood function one-by-one while keeping all other base learners fixed ([**57**]) and thus incorporating the information of the already selected variables. Based on the largest improvement of the overall likelihood, one base learner $\hat{h}_*$ is selected and the additive predictor is updated as

$$\hat{f}^{[l]} := \hat{f}^{[l-1]} + \nu \hat{h}_*$$

with $\nu = 0.1$ as a stepsize. The number of iterations taken is the main tuning parameter of the algorithm and usually set by a resampling approach. The empirically selected final model can be written as

$$\hat{f}(X_{-j}) = \sum_{i \in \{1,\dots,p\} \setminus \{j\}} \hat{\beta}_i X_i = X_{-j} \hat{\beta}$$

with $\hat{\beta} := \left((\hat{\beta}_i)_{i \in \{1,\dots,p\} \setminus \{j\}} \right)^{\mathrm{T}}$.

We perform componentwise likelihood-based boosting in turn with each of the variables being the dependent and all other the independent variables with a fixed number of steps. First we extend the $\hat{\beta}$s by a zero at the index of the dependent variable and then merge all individual $\hat{\beta}$s column-wise into a $p \times p$-matrix $\hat{B}$. Thereby, $\mathrm{diag}(\hat{B}) = \mathrm{diag}(0)$ and we can write the finally selected individual models as

$$X_i \sim X \hat{B}_{\cdot i}$$

for all $i \leq p$. By the zero extension, X_i as an independent variable is ignored in the prediction of X_i as $X \hat{B}_{\cdot i} = X_{-i} \hat{B}_{j \neq i, i}$. We fit an unsigned network and neglect the boosting coefficient sign. We define the filter by

$$\mathcal{F} := \{(i,j), (j,i) | \exists i, j \in \mathbb{N} : B_{ij} \neq 0\}.$$

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

2.1.2. *Distance calculation.* For tuples in $\mathcal{F}$, we define the adjacency of two variables by the power adjacency function. For all other tuples, the adjacency is set to 0. Hence, we have

$$a_{ij} := \begin{cases} 1 & \text{if } i = j, \\ |\text{corr}(X_i, X_j)|^b & \text{else if } (i,j) \in \mathcal{F}, \\ 0 & \text{else,} \end{cases}$$

where b is chosen data-based by the scale free topology criterion on a random subset of variables (Chapter 2: 1).

We combine the TOM with $\mathcal{F}$ and define

$$\text{TOM}_{ij} := \begin{cases} 1 & \text{if } i = j, \\ \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{\min(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{uj}) + 1 - a_{ij}} & \text{else if } (i,j) \in \mathcal{F}, \\ 0 & \text{else.} \end{cases} \tag{2}$$

We invert it to a dissimilarity measure by

$$\text{d}_{ij} := 1 - \text{TOM}_{ij} \in [0,1]. \tag{3}$$

2.2. Module detection. We apply the UPGMA ([**52**]) to the d_{ij}. Parts of the network where no path exists in $\mathcal{F}$ are clustered independently. A path between X_i and X_j exists if and only if there is an $l \in \mathbb{N}$ such that there are $t_{1\ldots l} \in \mathcal{F}$ 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 (Chapter 2: 1). Thus, features which are topologically close on the filtered edges are grouped into modules. The resulting partion into modules can be represented by some $M \subseteq \mathfrak{P}(\{1, \ldots, p\})$.

2.3. Aggregation of module information. The first PC explains the most variation possible in an one-dimensional space. By design of the modules, they consist of highly correlated variables and this first PC typically explains between 40% and 95% of the variation in X^m for an $m \in M$. This proportion of variance

explained is denoted by

$$\text{propVar}(m) := 1 - \frac{\text{residual variance of PC1}}{\text{total variance}}. \tag{4}$$

Therefore, we aggregate the information in each module by its first PC, the so called eigengenes (ME, Chapter 2: 1). In a final step, modules with highly correlated first principal components are merged to further reduce dimensionality. Given the resulting partition M, we define

$$X_{\text{modules}} := (\text{ME}_1^{\text{T}}, \ldots, \text{ME}_{|M|}^{\text{T}}).$$

X_{modules} has dimensions $n \times |M|$ where $|M| \ll p$. Due to its definition, a substantial part of variation in X is conserved in X_{modules}, while at the same time the dimensionality is considerably reduced. To enlarge the proportion of variance explained to a predefined level minVar, we optionally extend X_{modules} to the first l PCs per module. Here, for each module m l is chosen such that $l = \min(\{1, \ldots, |m|\})$ with $\text{propVarExtended}(m, l) \geq \text{minVar}$, where

$$\text{propVarExtended}(m, l) := 1 - \frac{\text{residual variance of PC}_1 \ldots \text{PC}_l}{\text{total variance}}.$$

2.4. Properties. A real-valued function is a similarity measure if and only if it is symmetric and non-negative.

LEMMA 1. *TOM_{ij} takes values in $[0, 1]$ and is symmetric. Thus, it is a similarity measure.*

PROOF. For $(i, j) \notin \mathcal{F}$ this is trivial. For $(i, j) \in \mathcal{F}$ w.l.o.g.

$$\min(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{uj}) = \sum_{u \neq i} a_{iu}$$

and therefore

$$\text{TOM}_{ij} = \frac{\sum_{u \neq i,j} a_{iu}a_{uj} + a_{ij}}{\sum_{u \neq i} a_{iu} + 1 - a_{ij}} = \frac{\sum_{u \neq i} a_{iu}a_{uj} - a_{ij}a_{jj} + a_{ij}}{\sum_{u \neq i} a_{iu} + 1 - a_{ij}} =$$
$$\frac{\sum_{u \neq i} a_{iu}a_{uj}}{\sum_{u \neq i} a_{iu} + 1 - a_{ij}} \leq \frac{\sum_{u \neq i} a_{iu}a_{uj}}{\sum_{u \neq i} a_{iu}} \leq \frac{\sum_{u \neq i} a_{iu}1}{\sum_{u \neq i} a_{iu}} = 1$$

The second to last inequality holds because if $(i,j) \in \mathcal{F}$ it follows that $\text{corr}(X_i, X_j) \neq 0$, hence $a_{ij} \neq 0$. Furthermore, as $\forall i,j\ a_{ij} \in [0,1]$ and thereby $1 - a_{ij} \geq 0$ it follows that $\text{TOM}_{ij} \geq 0$.

It follows directly from the definition (2) that TOM_{ij} is symmetric and thereby it is a similarity measure which takes values in $[0,1]$. □

Analogously, it follows that $\text{TOM}_{ij}^{\text{WGCNA}}$ is a similarity measure.

LEMMA 2. *If* $(i,j) \in \mathcal{F}$, $a_{ij} = 1$ *and for all* u *with* $a_{iu} \neq 0$ *or* $a_{uj} \neq 0$

$$a_{uj} = a_{iu} = 1,$$

then $TOM_{ij} = 1$.

PROOF.

$$\text{TOM}_{ij} = \frac{\sum_{u \neq i,j} a_{iu}a_{uj} + 1}{\min(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{uj})} \overset{\text{w.l.o.g.}}{=} \frac{\sum_{u \neq i} a_{iu}a_{uj} - a_{ij}a_{jj} + 1}{\sum_{u \neq i} a_{iu}} = 1$$

□

LEMMA 3. *If* $i \neq j$, $a_{ij} = 0$ *and* $\nexists u$ *with* $a_{iu} \neq 0 \wedge a_{uj} \neq 0$ *then* $TOM_{ij} = 0$.

PROOF. For $(i,j) \notin \mathcal{F}$, it follows from $i \neq j$ that $\text{TOM}_{ij} = 0$. Let $(i,j) \in \mathcal{F}$, then

$$\text{TOM}_{ij} \overset{\text{w.l.o.g.}}{=} \frac{\sum_{u \neq i,j} a_{iu}a_{uj} + a_{ij}}{\sum_{u \neq i} a_{iu} + 1 - a_{ij}} = \frac{\sum_{u \neq i,j} a_{iu}a_{uj}}{\sum_{u \neq i} a_{iu} + 1} = \frac{0}{\sum_{u \neq i} a_{iu} + 1} = 0$$

□

While Lemma 1 is essential for appropriate definition of d_{ij}, Lemmas 2 and 3 illustrate some basic plausibility of the introduced similarity measure. Lemma 2 shows that perfectly similar variables connected by further variables indeed receive the maximal similarity measure and Lemma 3 demonstrates that disconnected variables have the minimal similarity measure.

COROLLARY 4. $d : \mathbb{N}^2 \to \mathbb{R}$ *is a pseudosemimetric.*

PROOF. We show

(1) $d_{ij} \geq 0$,

(2) $d_{ii} = 0$ and

(3) $d_{ij} = d_{ji}$.

From Lemma 1 and its definition we have that d is symmetric and that the image space of d is $[0, 1]$.

If $i = j$ then $d_{ij} = 1 - \text{TOM}_{ij} = 1 - 1 = 0$.

□

THEOREM 5. *$d : \mathbb{N}^2 \to \mathbb{R}$ does not fulfill the triangular inequality.*

PROOF. We prove the violation of

$$d_{ij} \leq d_{iu} + d_{uj}$$

by counterexample in an computer assisted manner. First, we note that for any symmetric positive semidefinite $\Sigma \in \mathbb{R}^{d \times d}$ such that the diagonal elements are equal to 1 there exists a random variable

$$\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}} \sim \mathcal{N}(0, \Sigma).$$

As $\Sigma_{ij} := \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])]$, it follows that the population Pearson correlation coefficient of X_i and X_j is

$$\begin{aligned} \rho(X_i, X_j) &= \frac{\mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])]}{\sqrt{\text{Var}(X_i)}\sqrt{\text{Var}(X_j)}} \\ &= \frac{\Sigma_{ij}}{\sqrt{\text{Cov}(X_i, X_i)}\sqrt{\text{Cov}(X_j, X_j)}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}}\sqrt{\Sigma_{jj}}} = \Sigma_{ij}. \end{aligned}$$

Let b equal to 1 and $\mathcal{F} = \{(i, j) \mid i \neq j\}$. Let Σ further be positive in all elements. It follows that the adjacency matrix A equals Σ.

Next, we design a matrix which has a suboptimal direct but good indirect connection of node 1 and 2. For $d \geq 4$ let Σ be

$$\begin{pmatrix} 1 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & \cdots & 1 \\ 0 & 1 & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 1 & 1 & \cdots & \cdots & 1 \end{pmatrix}.$$

It follows that

$$d_{12} = 0.5 > 0 + 0 = d_{13} + d_{32}.$$

However, these matrices are not positive semidefinite. Next, we explore possible modifications of these matrices to identify a symmetric positive semidefinite matrix which still violates the triangular inequality.

For $d \in [4, 50]$ we leave the diagonal and zeros unchanged while subtracting the absolute value of $\epsilon \sim \mathcal{N}(0, 0.2)$ from all other entries. For each entry we redraw ϵ and after the subtractions for $i < j$ recursively redefine entries as

$$\Sigma_{ij} := \text{mean}(\Sigma_{ij}, \Sigma_{ji}),$$

and then

$$\Sigma_{ji} := \Sigma_{ij}.$$

Matrices with negative entries are disregarded. We round all other matrices to two digit precision and check wether the triangular inequality is still violated and if all eigenvalues are positive.

We identified multiple matrices which fulfilled all conditions. One of them is

$$\begin{pmatrix} 1.00 & 0.00 & 0.68 & 0.00 \\ 0.00 & 1.00 & 0.65 & 0.59 \\ 0.68 & 0.65 & 1.00 & 0.59 \\ 0.00 & 0.59 & 0.59 & 1.00 \end{pmatrix}.$$

We have

$$d_{12} = 0.7369048 > 0.6922642 = 0.32 + 0.3722642 = d_{13} + d_{32}.$$

□

COROLLARY 6. *$d^{WGCNA} : \mathbb{N}^2 \to \mathbb{R}$ does not fulfill the triangular inequality.*

PROOF. As we assumed $\mathcal{F} = \{(i, j) \mid i \neq j\}$ in the counterexample in the proof of Theorem 5, it follows that

$$d^{\text{WGCNA}} = d.$$

□

We define a binary relation on $\mathbb{N}$ by

$$i \sim j := \exists f : \mathbb{R} \to \mathbb{R} \text{ linear} \wedge f(X_i) = X_j,$$

where X_i and X_j represent their respective random variables. It is easily seen that this is reflexive, symmetric and transitive and thereby an equivalence relation. d is only a *pseudo*semimetric as $d_{ij} = 0$ only implies $i \sim j$ and not $i = j$.

THEOREM 7. *If $d_{ij} = 0$, then $i \sim j$.*

PROOF. Assuming we have $d_{ij} = 0$ and $(i,j) \notin \mathcal{F}$ it follows that $d_{ij} = 1$. ↯ For $d_{ij} = 0$ and $(i,j) \in \mathcal{F}$, w.l.o.g. $\min(\sum_{u\neq i} a_{iu}, \sum_{u\neq j} a_{uj}) = \sum_{u\neq i} a_{iu}$. Henceforth,

$$
\begin{aligned}
0 = d_{ij} &= 1 - \frac{\sum_{u\neq i,j} a_{iu}a_{uj} + a_{ij}}{\min(\sum_{u\neq i} a_{iu}, \sum_{u\neq j} a_{uj}) + 1 - a_{ij}} \\
&\Leftrightarrow \frac{\sum_{u\neq i,j} a_{iu}a_{uj} + a_{ij}}{\min(\sum_{u\neq i} a_{iu}, \sum_{u\neq j} a_{uj}) + 1 - a_{ij}} = 1 \\
&\Leftrightarrow \sum_{u\neq i,j} a_{iu}a_{uj} + a_{ij} = \min(\sum_{u\neq i} a_{iu}, \sum_{u\neq j} a_{uj}) + 1 - a_{ij} \\
&\Leftrightarrow a_{ij} = \frac{\sum_{u\neq i} a_{iu} + 1 - \sum_{u\neq i,j} a_{iu}a_{uj}}{2} \\
&\Leftrightarrow a_{ij} = \frac{\sum_{u\neq i} a_{iu} + 1 + a_{ij}a_{jj} - \sum_{u\neq i} a_{iu}a_{uj}}{2} \\
&\Leftrightarrow a_{ij} = \sum_{u\neq i} a_{iu} + 1 - \sum_{u\neq i} a_{iu}a_{uj} \\
&\Leftrightarrow \sum_{u\neq i} a_{iu}a_{uj} + a_{ij} = \sum_{u\neq i} a_{iu} + 1 \\
&\Leftrightarrow \sum_{u} a_{iu}a_{uj} = \sum_{u\neq i} a_{iu} + 1 \\
&\Leftrightarrow \sum_{u} a_{iu}a_{uj} = \sum_{u} a_{iu} \\
&\Rightarrow \forall u \text{ with } a_{iu} \neq 0 : a_{uj} = 1 \\
&\Rightarrow a_{ij} = |\mathrm{corr}(X_i, X_j)|^b = 1 \\
&\Rightarrow i \sim j
\end{aligned}
$$

□

In conjunction we showed that d when compared to d_{WGCNA} does not lose essential propertise due to the introduction of the filter and that d as well as d_{WGCNA} are pseudosemimetrics (Corrolary 4). Here, the *semi* restriction originates from the intended zero-distance of perfectly connected variables as showen in Lemma 2. The *pseudo* restriction originates from the integration of neighboring edges via the TOM. We so to say exchange the improved stability of the network as a whole with consistency on a local level in some rare instances.

2.5. Implementation. Netboost is built as an `R` package and was reviewed and accepted for the Bioconductor repository under Linux and macOS. A Windows implementation is currently not planned due to compiler dependencies.

As depicted in Algorithm 1, we first calculate $\mathcal{F}$. After scaling and centering each variable, we efficiently implement the likelihood-based boosting as a backend C++ routine. While the original code was designed and written by Pascal Schlosser, the conversion to C++ was done by Jochen Knaus. The subsequent calculation of adjacencies and TOM are performed exclusively on network edges in $\mathcal{F}$. Then the resulting sparse distance matrix are exported to *Sparse* UPGMA, an algorithm presented in **[53]** and implemented in C++. Here, all missing edges in the sparse matrix 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 a distance of 1.

Netboost is part of Bioconductor from release 3.9 onwards. Ongoing development of new features and adaptations, like the robust extension presented in Chapter 7, are freely available on Github at `https://github.com/PascalSchlosser/Netboost`. After thorough testing, stable releases are pushed to Bioconductor. All functionalities of Netboost are available from within R, whereas substantial parts of the algorithm are implemented in C++. *Sparse* UPGMA is part of the standalone MC-UPGMA software (**[53]**), which is distributed with the Netboost `R` package. For the cutting of dendrograms we apply the WGCNA (**[15]**) and dynamicTreeCut (**[22]**) `R` packages. As an example for the computational demand, Netboost was run on a dataset with 180 samples and 413,169 variables (for details see Chapter 3: 2). Applying two Xeon E5 2690v3 at 2.6GHz (2x12cores) and 40 GB of memory, it took Netboost 13.94 hours to compute. The package can be installed via Bioconductor or the devtools package (Listing 2.1). The vignette, a task-oriented description of package functionality, and the package manual are attached as Supplementary File S1 and Supplementary File S2.

Algorithm 1: Netboost

Input: X, steps, minModuleSize, MEDissThres

Result: X_{modules}

$\mathcal{F} = \emptyset$;

for $j \leftarrow 1$ **to** p **do**

  fit $X_j \sim X_{-j}\hat{B}_{i \neq j,j}$ by boosting;

  $\mathcal{F} = \mathcal{F} \cup \{(i,j) | \exists i \in \mathbb{N} : B_{ij} \neq 0\}$;

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))];

scaleFreeTopologyCriterium(randomFeatures) $\rightarrow b$;

for $(i,j) \in \mathcal{F}$ **do**

  $a_{ij} = |\text{corr}(X_i, X_j)|^b$;

end

for $(i,j) \in \mathcal{F}$ **do**

  compute d_{ij} according to equation (2) and (3);

end

sparseUPGMA(d) $\rightarrow$ dendrogram;

cutreeDynamic(dendrogram, minModuleSize, MEDissThres) $\rightarrow$ modules;

for $m \in$ *modules* **do**

  compute first principal component ME_m;

end

while $\exists m, m'$ *with corr(*$ME_m, ME_{m'}$*)* $> (1 -$ *MEDissThres*) **do**

  merge(m, m');

  compute first principal component of merged module;

end

```
# Bioconductor version
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
if (!requireNamespace("netboost", quietly = TRUE))
BiocManager::install("netboost", version = "3.10")

# Github version
if (!requireNamespace("devtools", quietly = TRUE))
    install.packages("devtools")
if (!requireNamespace("netboost", quietly = TRUE))
devtools::install_github("PascalSchlosser/netboost")
```

Listing 2.1: Install Netboost

What is new in Chapter 2:

- Netboot: Novel statistical modeling strategy for high-dimensional data.
- Theoretical foundation of the proposed algorithm.
- Implementation of the algorithm as a Bioconductor `R` package.

CHAPTER 3

Netboost for survival analysis

In acute myeloid leukemia (AML), part of the epigenotype of the disease is a global increase in DNAm in regulatory regions ([**58**]). For elderly patients, the only effective drugs that counteract this effect are hypomethylating agents ([**59, 60, 61**]). From this it is known that the state of methylation fulfills an important role in this disease. Nevertheless, it has been difficult to incorporate DNAm markers in patient relevant statistics like survival prediction ([**59, 62**]). Neither predictive methylation sites in AML patients treated with chemotherapeutics ([**63**]) nor predictive sites from chronic myelomonocytic leukemia patients treated with hypomethylating drugs ([**64**]) could be replicated for AML patients treated with hypomethylating drugs.

Mechanism of DNAm and its regulatory effects are only understood in an incomplete manner. Not always promoter but also gene body methylation impacts regulation and is distinctly modified by hypomethylating agents ([**5**]). A-priori schemes to structure DNAm before relating it to patient-relevant-outcomes, as survival, are unclear and variable selection procedures struggle with the high-dimensionality of DNAm. Therefore, we set out to perform an unsupervised dimension reduction via Netboost and subsequently related MEs to survival. Time-to-event associations were then replicated and to brooden the methodological significance of the demonstrated approach, the analysis design was applied to further diseases and omics data types. Throughout this Chapter, we compare results with competitive state-of-the-art approaches and find superior performance of Netboost with respect to several measures.

1. Methods

1.1. Boosting estimation of sparse high-dimensional survival models. Let T model the survival time of individuals of a population. We define T as a non-negative random variable and define the survival function

$$S(t) := P(T > t) = 1 - P(T \leq t),$$

as the probability of an event after time t. The instantaneous risk of failure at timepoint t given the individual survived up to t is the hazard function

$$\lambda(t) := \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

In the context of predictive survival models, the proportional hazards assumption is the cornerstone of Cox proportional hazards regression ([**65**]). The assumption poses that the ratio of the hazards for any two individuals is constant over time, such that it is possible to estimate the effect parameters without any consideration of the hazard function. Given this assumption, we can write the hazard rate of an observation o with a covariate vector ${}_oX := (X_{o1}, \ldots, X_{op})$ as

$$\lambda(t \mid {}_oX) = \lambda_0(t) \exp({}_oX\beta),$$

where $\lambda_0(t)$ is the non-negative baseline hazard function independent of the covariates and β a vector of regression coefficients. For estimation, the baseline hazard $\lambda_0(t)$ is left unspecified and the partial log-likelihood for estimation of β can be written as

$$l(\beta) = \sum_{o=1}^{n} \delta_o \left({}_oX\beta - \log \left(\sum_{o'=1}^{n} \mathbb{1} \left(t_{o'} \leq t_o \right) \exp \left({}_{o'}X\beta \right) \right) \right),$$

with $\mathbb{1}$ being an indicator function taking the value 1 if its argument is true. In high-dimensional settings, often a penalized version of this likelihood is optimized. Lasso-like algorithms use the L_1-norm, the sum of absolute coefficients, to achieve a sparse solution ([**66, 67, 68**]). We apply likelihood-based boosting as a forward selection method to handle the high-dimensional setting and achieve a sparse solution ([**69, 23, 70**]).

In particular, we apply CoxBoost ([**23**]) to integrate the potentially still high-dimensional X_{modules} with clinical covariates and survival data as the primary outcome by likelihood-based boosting. The algorithm is analogous to the procedure described in more detail in Chapter 2: 2.1.1. Analysis is implemented with the CoxBoost `R` package ([**70**]). The stopping criterion is chosen by cross-validation and a Cox proportional hazards model is fitted.

1.2. Prediction errors. To describe the performance of a fitted survival model we introduce prediction errors. For a prediction rule $\hat{T}$ we quantify the goodness of fit by the Brier Score ([**71**]). We use the simplified version for the binary case given as the mean square error by

$$\text{BS} := \frac{1}{n}\sum_{o=1}^{n}(T - \hat{T})^2,$$

which omits the summation over classes required for multi-category predictions. Evaluation of the Brier Score at each time point yields the prediction error curve and integration yields an aggregate measure ([**72**]). To account for censoring, we introduce time- and covariate-dependent weights within the sum as described in [**73**].

For an unbiased estimate of the prediction error curve we can use cross-validation by splitting our data into a training and a validation set, that is $\{1, \ldots, n\} = Q_n = Q_V \cup Q_T$. The split-sample error is given by

$$\widehat{\text{Err}}_{\text{split}}(t, \hat{T}, Q_V, Q_T) := \frac{1}{\#Q_V}\sum_{o \in Q_V}\left(\omega(t, {}_oX)(T(t,o) - \hat{T}(t, Q_T, {}_oX))^2\right),$$

with $\omega(t, {}_oX)$ being the aforementioned weights, $T(t,o)$ the true state of observation o at t and $\hat{T}(t, Q_T, {}_oX)$ the prediction rule trained on Q_T for covariates ${}_oX$ at t ([**74**]). While this estimate is unbiased, it exhibits a large variance when a binary outcome is studied. In k-fold cross-validation, data is split into k disjoint subsets and the prediction error is estimated as the average split-sample error on these subsets with prediction rules being trained on the respective complements. While this procedure

already reduces the variance of the estimator to some extent, we can achieve even less variance by employing bootstrapping methods as explained in the following.

Let Q_n^* be a bootstrap sample of size n drawn from Q_n with replacement. For $B_{\max} \in \mathbb{N}_+$ and independently drawn $Q_n^{*1}, \ldots, Q_n^{*B_{\max}}$, we define the leave-one-out bootstrap prediction error estimator as

$$\widehat{\mathrm{Err}}_{(1)}(t, \hat{T}) := \frac{1}{n} \sum_{o=1}^{n} \frac{\sum_{l=1}^{B_{\max}} \left(\mathbb{1}(o \notin Q_n^{*l}) \widehat{\mathrm{Err}}_{\mathrm{split}}(t, \hat{T}, \{o\}, Q_n^{*l}) \right)}{\sum_{l=1}^{B_{\max}} \mathbb{1}(o \notin Q_n^{*l})}. \tag{5}$$

Compared to the true prediction error curve of Q_n, this is biased upwards as the bootstrap samples are only expected to be based on $0.632n$ observations and smaller training data leads to more uncertainty in predictions. The factor 0.632 stems from the random draws with replacement as $1 - (1 - 1/n)^n \xrightarrow[n \to \infty]{} 1 - 1/e \approx 0.632$ where we assume equal weights for all observations.

The apparent error given by

$$\overline{\mathrm{Err}}(t, \hat{T}) := \frac{1}{n} \sum_{o=1}^{n} \widehat{\mathrm{Err}}_{\mathrm{split}}(t, \hat{T}, \{o\}, Q_n), \tag{6}$$

is biased downwards as observations are used for training and validation simultaneously. Efron suggested in **[75]** to combine the errors defined in (5) and (6) to the .632 error given by

$$\widehat{\mathrm{Err}}_{.632}(t) := (1 - \omega(t)) \overline{\mathrm{Err}}(t) + \omega(t) \widehat{\mathrm{Err}}_{(1)}(t) \tag{7}$$

with $\omega(t) = 0.632$ to achieve a less biased estimate with low variance (**[74]**). Efron and Tibshirani improved this to the nearly unbiased .632+ error in **[76]** by introducing the relative overfitting rate and adjusting the mixture according to this. First, the no information error is estimated by systematically permuting covariates with respect to the outcome and fitting the prediction rule on Q_n. Based on this reference, the relative overfitting rate is given by

$$\widehat{\mathrm{ROR}}(t) := \frac{\widehat{\mathrm{Err}}_{(1)}(t) - \overline{\mathrm{Err}}(t)}{\mathrm{NoInfErr}(t) - \overline{\mathrm{Err}}(t)}.$$

Finally, we define the .632+ error as the .632 error with

$$\omega(t) = \frac{0.632}{1 - 0.368\widehat{\mathrm{ROR}}(t)}.$$

To evaluate the performance of CoxBoost models we used the peperr `R` package ([**77**]) which implements the .632+ prediction errors based on subsamples without replacement as recommended in [**78**]. In high-dimensional data settings, bootstrap samples with replacement often lead to overly complex models ([**78**]). Therefore, subsamples without replacement of 63.2% of the samples are implemented.

1.3. Blockwise WGCNA. Due to the high dimensionality of some of the datasets in this Chapter, the implementation of WGCNA cannot be applied to the full dataset ($p > \sqrt{2^{31} - 1} \approx 46,340$). As suggested by the authors, we first split variables via k-means clustering and aggregate modules across subsets via correlated MEs ([**15**]).

2. High-dimensional survival models in acute myeloid leukemia

In this section we applied Netboost on DNAm and gene expression data from The Cancer Genome Atlas (TCGA) ([**26**]), which is a public domain project supervised by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). We use this data example to illustrate the advantages of Netboost as a dimensionality reduction technique in general and in comparison to WGCNA specifically. These advantages comprise of a reduction in prediction errors as well as the extraction of biologically meaningful units.

Data examples such as this are of particular importance for evaluating high-dimensional methods, as omics data and their complex multivariate distributions are understood to an incomplete degree, which poses further challenges to simulation studies in addition to the potentially simplifying assumptions stated explicitly in their design. To avoid some form of selection bias, all applications we studied with Netboost at the time of writing (December 25th, 2018) are reported within this dissertation. These encompass a wide variety of primary analyses (survival analyses,

classification and GWAS) and a broad range of omics data (DNAm, metabolomics, miRNA and array- and sequencing-based gene expression) such that Netboost can be evaluated in a relatively general fashion and an overfitting of the method to a specific task is avoided.

TCGA encompasses high-dimensional molecular and clinical data from 33 different cancer types. In collaboration with Prof. Dr. Michael Lübbert of the Department of Hematology-Oncology at the Medical Center Freiburg we first focussed on acute myeloid leukemia (AML) overall survival.

We selected the 180 AML patients in TCGA for which overall survival data, methylome and gene expression measurements were available ([**24**]). TCGA data was already preprocessed and normalized. DNAm was quantified with Illumina Infinium HumanMethylation450 BeadChip arrays and gene expression by Affymetrix HG U133 Plus 2.0 arrays, incorporating 396,065 methylation and 17,104 gene expression variables. We compared the following six analyses designs, the first three including and the last three excluding a clinical score. The clinical score is a dichotomized version of the linear predictor of a Cox proportional hazards regression model ([**79**]) of age at diagnosis and cytogenetic risk group, assessed as low, intermediate or high. Baseline hazards were estimated in separate strata according to sex. In the models with the clinical score it was set as mandatory, so unpenalized in CoxBoost ([**23**]). Thereby, DNAm 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 WGCNA modules: Application of CoxBoost on blockwise WGCNA module eigengenes X_{WGCNA}.
(3) Netboost modules: Application of CoxBoost on Netboost module eigengenes X_{modules}.
(4) Direct application + clinical: Application of CoxBoost on the full dataset X and the mandatory clinical score.

(5) Blockwise WGCNA modules + clinical: Application of CoxBoost on X_{WGCNA} and the mandatory clinical score.

(6) Netboost modules + clinical: Application of CoxBoost on X_{modules} and the mandatory clinical score.

In CoxBoost we used 10-fold cross validation to estimate the optimal stopping criterion on the interval from 0 to 100 and we applied 200 resampling steps to estimate the .632+ prediction errors.

The direct application on the full dataset, X, selected two variables and the .632+ prediction error curve, depicted in Figure 6, showed no improvement over the null model.

WGCNA 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 leaving all smaller modules as unclustered. Henceforth, 92% of the features were assigned to modules. The proportion of variance explained by eigengenes ranged from 23.9% to 94.6% (median = 50.5%, Figure 7). In the WGCNA aggregated $X_{\text{WGCNA modules}}$ two modules were selected by CoxBoost summarizing 26 variables.

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% of 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% of the features were assigned to modules. The dendrogram based on the sparse network is depicted in Figure 8. Netboost eigengenes generally explained a higher proportion of variance (median = 66.5%, range = [45.7%, 97.3%], Figure 7). CoxBoost selected six modules from the Netboost aggregated $X_{\text{Netboost modules}}$, summarizing 278 features. None of the features are shared by the selected Netboost modules and the selected WGCNA modules.

As depicted in Figure 6, the higher complexity indeed corresponds to a better prediction performance in the .632+ prediction errors. The blockwise WGCNA modules approach was able to extract some information but was outperformed by

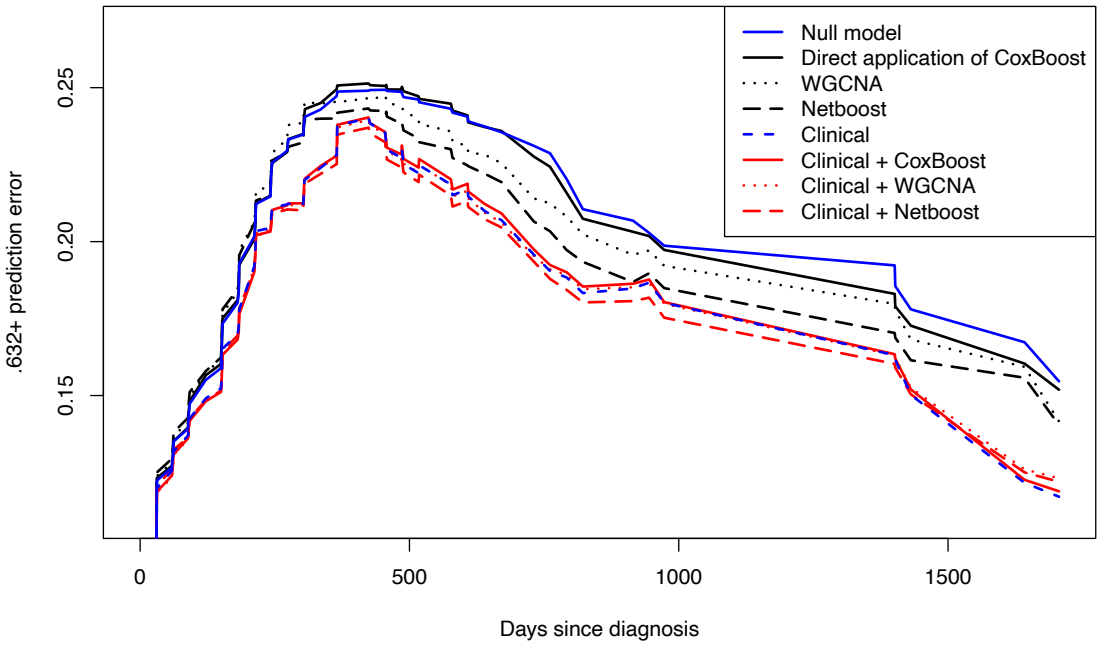

Figure 6: **.632+ prediction error curves for AML survival models.** The estimated .632+ prediction error curves for days since diagnosis are given in blue for the null model and dashed blue for the clinical model. Prediction error curves based solely on DNAm and gene expression are presented in black: The solid line for the direct application of CoxBoost, the dotted line for the combination with WGCNA and the dashed line for the combination with Netboost. The corresponding prediction error curves additionally based on unpenalized clinical data are presented in red.

Netboost. This also holds true when incorporating the variability of the individual bootstrap samples and integrating their $\widehat{\text{Err}}_{(1)}$ in Figure 9, which overestimates the true prediction error.

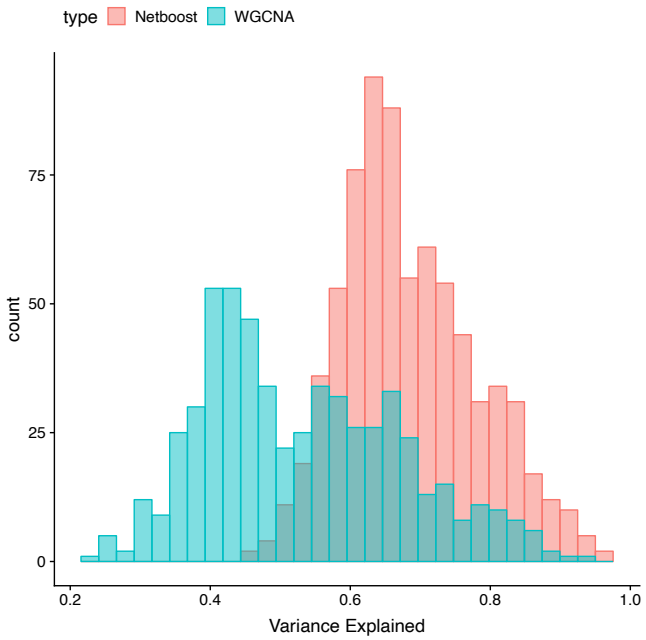

Figure 7: **Histogram of the proportion of variance explained by MEs in the TCGA AML dataset**

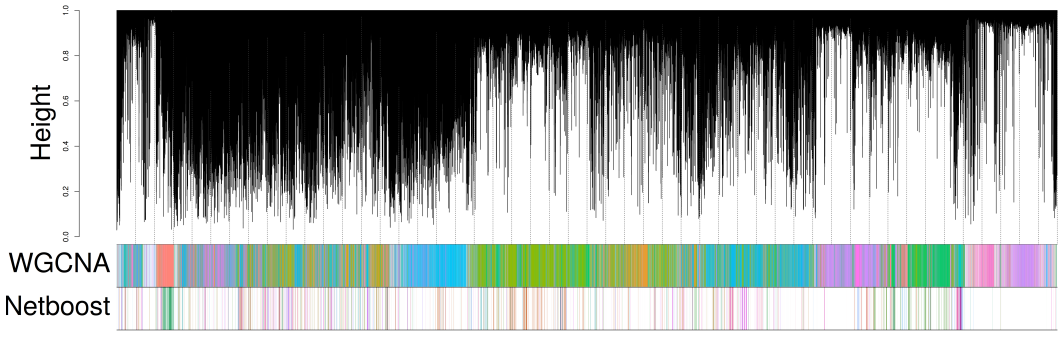

Figure 8: **Dendrogram of the TCGA AML data.** Dendrogram of the 396,065 DNAm and 17,104 gene expression variables in the TCGA AML data. The color bands below the graph show the separation into modules by blockwise WGCNA and Netboost.

As depicted in Figure 6 and Figure 9, 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 CoxBoost showed significant

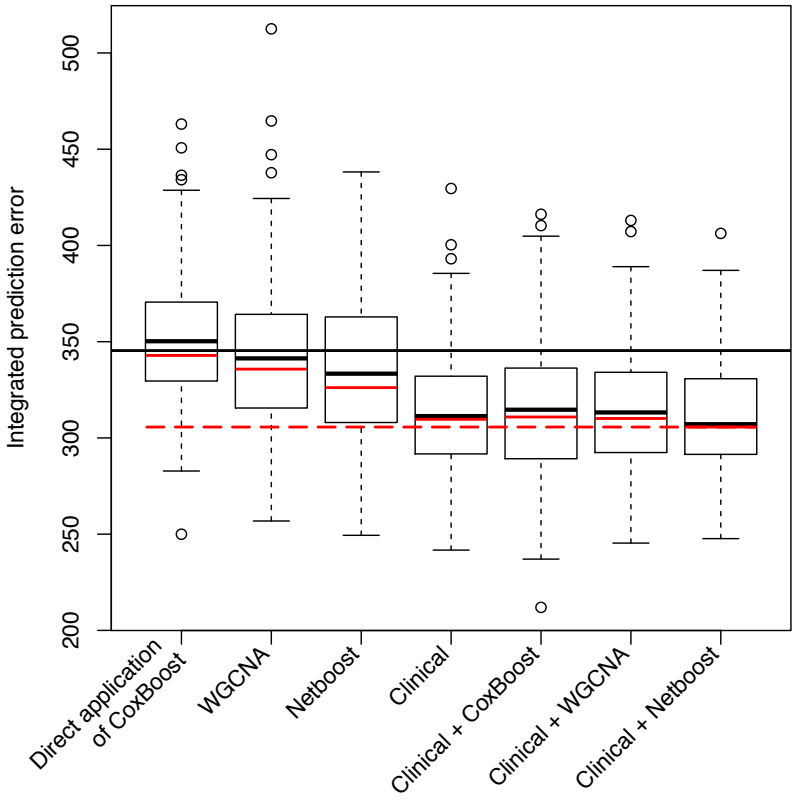

Figure 9: **Variability of the $\widehat{\text{Err}}_{(1)}$ prediction error curves in AML survival models.** Integrated $\widehat{\text{Err}}_{(1)}$ prediction error curve estimates from 200 bootstrap samples each for CoxBoost based on the different datasets. The Kaplan-Meier benchmark value based on the full dataset 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 across all datasets by a dashed line.

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 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. Between analyses including the clinical score, no significant differences were observed (two-samples Student's t-test, p-value < 0.05). The order of approaches is consistent with and without the clinical score indicating smallest prediction errors for Netboost.

2.1. Biological relevance of Netboost network. The central dogma of molecular biology as published in *Nature* **[80]** states:

> The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.

This is usually interpreted as deoxyribonucleic acid (DNA) coding ribonucleic acid (RNA), which codes for proteins, and proteins being the active substances of our body. While DNA regulates itself, there is a regulatory impact of RNA on DNA and RNA has some self-regulating function there is no backward information flow once it reached the protein stage. Much of the regulation on the DNA-DNA, DNA-RNA and RNA-DNA level happens in *cis*, so in a physically short distance along the nucleotide code, and DNAm plays a central role in it. Netboost modules reflected this known biology. Netboost re-identified the association of cytosine–phosphate–guanines (CpGs) in close proximity and the *cis* association of gene methylation and expression in a data-driven manner. Of the 739 Netboost modules 206, consisted of CpGs within 1,000 base pairs demonstrating the strength of local dependency in DNAm data.

The six selected Netboost modules were variable in size and composition. Four consisted only of CpGs, one predominantly of CpGs, supplemented by two RNAs, and one module only of 14 RNAs. The total number of CpGs varied from 10 to 88. The largest selected module (88 CpGs) contained numerous genes associated with hematopoiesis, such as *WT1* and *CXCL2*. The second 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 (Figure 10). The p-value of the likelihood ratio test of the dichotomized 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 [**63**]. Additionally, *WT1* was suggested to regulate *TET2* methylation ([**81**]), which supports the relevance of the selected modules.

Overall, the general network structure detected by Netboost is in line with known biology lending validity to the novel connections suggested by Netboost. In addition many of the DNAm and gene expression variabels incorporated in the survival associated signature have been implicated in AML and other cancerous diseases, while the strong molecular differentiating power with respect to survival was yet undescribed.

2.2. Molecular surrogate information for clinical covariates. 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 MEs to the clinical score. We compared random selections of variables out of all DNAm and gene expression variables and modules, WGCNA and Netboost respectively, of similar size to the modules selected from WGCNA and Netboost modules 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. As shown in Figure 11, the selected Netboost modules approximated the clinical score best.

2.3. Replication of the ME71 survival association. To validate the Netboost module structure, we transfered it to DNAm data generated on pre-treatment patient samples from the phase II acute myeloid leukemia study group (AMLSG)

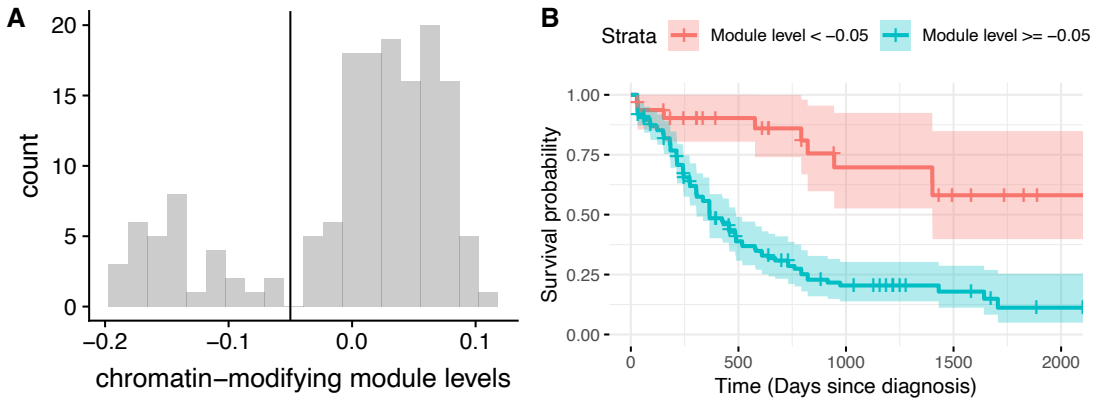

Figure 10: **TCGA AML DNA methylation module associated with chromatin modifying enzymes.** A) shows the bimodal distribution of the eigengene. The vertical line indicates at which point patients were stratified. B) depicts the Kaplan Meier curves stratified by the modules eigengene.

12-09 study ([**25**]). In this study, DNAm based on the same Illumina Infinium 450k array and overall survival was available for 55 AML patients with a maximum follow up of 1,985 days after diagnosis. For processing and quality control of the raw DNAm data, a customized version of the incorporating Control Probe Adjustment and reduction of global CORrelation (CPACOR) pipeline ([**82**]) 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. As replication data was incomplete with respect to variables, we transfered the grouping of modules and refitted PCs. 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, Figure 12).

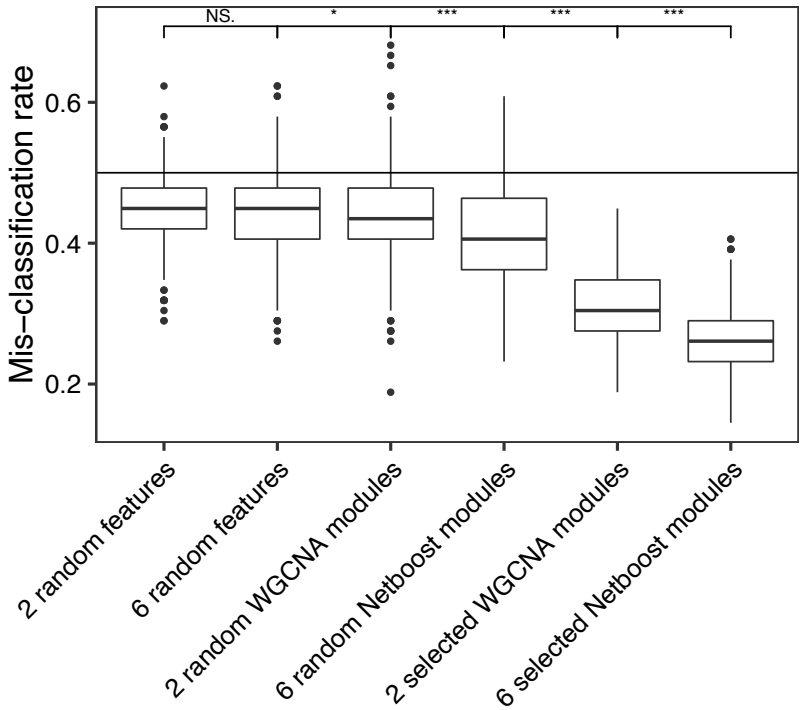

Figure 11: **Mis-classification rate for logistic regression models of the clinical score in AML.** We compare randomly selected variables of the raw data with randomly selected modules and the modules selected for survival prediction performance. The complexity of models is fixed to two and six to match the final survival models for Netboost- and WGCNA-based approaches respectively. The horizontal line indicates the expected mis-classification rate at random. Asterisks indicate significance of unpaired two-samples Wilcoxon tests (***<0.001, **<0.01, *<0.05, NS.≥ 0.05). Only neighbouring columns were tested.

It is of interest that validation of the chromatin associated module was successful in this independent AML patient DNAm dataset 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 and that no gene expression measurements were available. Particularly, patients with core-binding factor AML, AML with

mutated *NPM1*, and AML with *FLT3* internal tandem duplication were excluded in this trial.

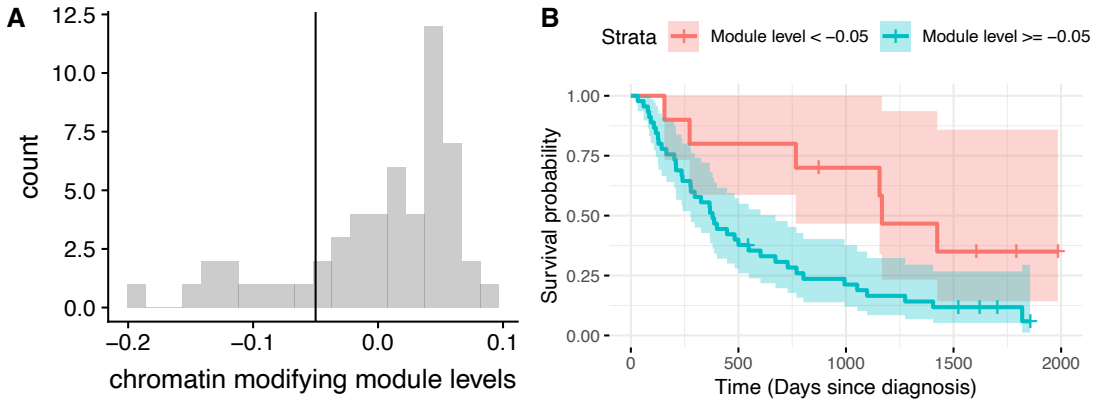

Figure 12: **Replication analysis in the AMLSG study.** A) shows the bimodal distribution of the transferred eigengene. The vertical line indicates at which point patients were stratified a-priori. B) depicts the Kaplan Meier curves stratified by the modules eigengene.

3. High-dimensional survival models in other entities

We transfered the analysis presented for AML in Section 2 to further TCGA entities, namely DNAm data of 774 breast invasive carcinoma (BRCA) and 315 kidney renal clear cell carcinoma (KIRC) patients and miRNA data of 464 ovarian serous cystadenocarcinoma (OV) patients with available overall survival information. The 1,422 TCGA-OV miRNAs without missings and the 20,000 CpG sites with the largest variance for TCGA-BRCA and TCGA-KIRC respectively were selected for analysis. For each dataset we performed the same three analyses as for AML without the clinical score and calculated the .632+ prediction error estimates.

Boxplots of the integrated prediction errors on the test set of the individual subsamplings are depicted in Figure 13. For KIRC we observed similar performance as in AML. The integration with WGCNA significantly improved CoxBoost (p-value = 0.0013) and the 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 beyond the Kaplan-Meier reference estimate.

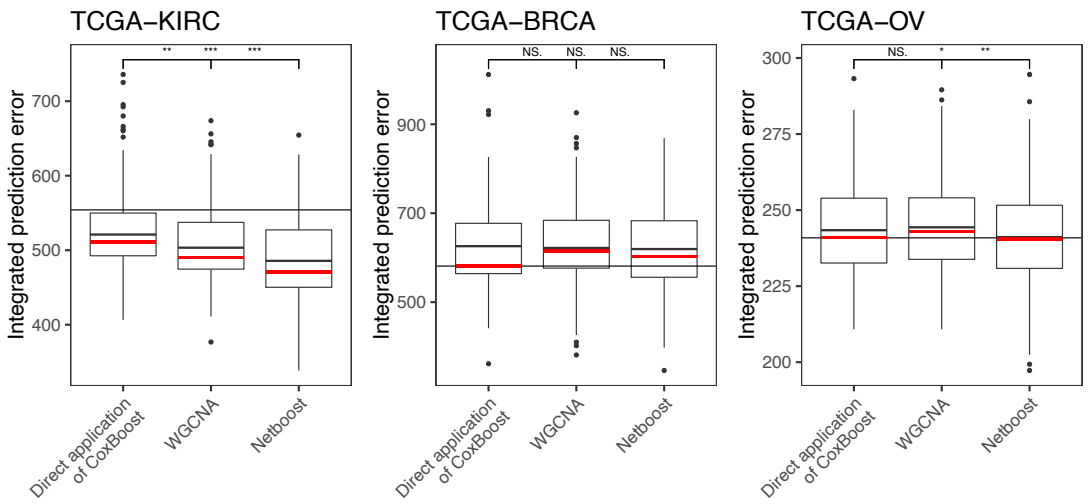

Figure 13: **Variability of the .632+ prediction error estimates in TCGA KIRC, BRCA and OV survival models.** Integrated prediction error curve estimates from single subsamples for CoxBoost on the full dataset, CoxBoost on X_{WGCNA} and CoxBoost on X_{Netboost}. The Kaplan-Meier benchmark value is indicated by a horizontal line. Red lines show the integrated .632+ prediction error estimates. Asterisks indicate significance of unpaired two-samples Wilcoxon tests (***$<$0.001, **$<$0.01, *$<$0.05, NS.$\geq$0.05).

What is new in Chapter 3:

- Integration of Netboost with Coxboost as a modeling strategy for high-dimensional survival prediction.
- Significantly better prediction errors than WGCNA and Coxboost alone in AML and KIRC datasets and non-inferiority for BRCA and OV.
- Replication of ME survival association in an independent dataset (AMLSG 12-09 study).
- Modules provide biological context for interpretation and provide evidence for the importance of chromatin-modifying enzymes in AML.

CHAPTER 4

Netboost for multi-trait genome-wide association study

In the following chapter we applied Netboost to improve our understanding of the human metabolism. We study this by identifying single nucleotide–polymorphisms (SNPs), and thereby their respective genes, associated with controlling metabolite levels in urine. We applied linear regression modeling to test gene-metabolite relations. In a genome-wide association study (GWAS), for each of several million SNPs the same theoretical model,

$$\text{outcome} \sim \beta_{\text{SNP}} \cdot \text{SNP} + \beta_{\text{covariates}} \cdot \text{covariates}, + \beta_0$$

with β_0 being the intercept, is tested independently with respect to the null hypothesis $\beta_{\text{SNP}} = 0$. Here, we were able to extend the genome-wide perspective of the SNPs with a metabolome-wide perspective by measuring 1,172 different metabolites and analyze them as the outcomes, the dependent variables. Given these genome-wide association studies of metabolite concentrations (mGWAS) as a reference we extended it by applying Netboost to the metabolite concentrations and analyzing MEs subsequently in a GWAS-fashion.

To understand the metabolism, we have to study the absorption, distribution, metabolism, and excretion (ADME) processes, which describe the handling of a compound within an organism ([**83**]). While often used in the context of pharmaceutical research and development, the governing principles of ADME are generally applicable and also influence the concentrations of naturally occurring compounds and their metabolites, the intermediates and end products of metabolism. Organs and tissues that strongly influence each of the respective ADME components are the intestinal tract, the blood, the liver, and the kidneys. As a major excretory organ, the kidneys integrate information over continuously ongoing systemic ADME

processes by determining the amount and concentrations of metabolites that are excreted in urine ([**84**]). While the concentrations of many metabolites in blood are tightly controlled, metabolite concentrations in urine can have a much wider range and serve as a read-out of metabolic capacities not detected through the study of blood ([**85**]). We therefore hypothesized that the study of metabolite concentrations in urine can be particularly informative of ADME processes in humans.

In addition to filtering metabolites from blood, the kidneys have an important role in the generation, breakdown, as well as active reabsorption and secretion of metabolites, which further determines their presence and concentrations in urine ([**86**]). In particular, many metabolites are excreted through active detoxification and transport processes in the epithelial cells of proximal tubules, where specialized enzymes and transport proteins coordinate their breakdown and clearance ([**87**]). The identity of these transporters and enzymes as well as their substrates *in vivo* are not yet completely understood. We hypothesized that the presence of CKD may represent a particularly informative "challenge" state that might trigger an epigenetic responds ([**4**]). CKD may not only lead to the up-regulation of ADME transporters and enzymes to compensate for reduced filtration function, but patient data also carries information about metabolism of uremic toxins and drugs commonly prescribed to CKD patients.

Furthermore, we hypothesized that transporters and enzymes, being active on multiple metabolites, lead to higher order genetic associations with one genetic variant being associated with a group of metabolites. As our current understanding of reaction partners is incomplete and a large proportion of measured metabolites are of unknown identity, we applied Netboost to first identify metabolite modules to then test their genetic associations in an unbiased and genome-wide manner.

mGWAS can provide novel insights into human metabolism in health, inborn errors of metabolism, and complex traits and diseases ([**88, 89, 90, 91**]). Most

previous mGWAS have been carried out based on blood samples from population-based studies or mostly healthy individuals ([**84**]). By performing unbiased, genome-wide searches for genetic variants that are associated with the concentrations of a metabolite, a metabolite quantitative trait locus (mQTL), mGWAS can implicate the enzymes and transporters influencing its uptake, generation, transport and distribution, breakdown or excretion ([**88, 91, 92**]).

Here, we focused on the study of metabolites quantified from urine of CKD patients, a unique setting, to generate a better understanding of ADME processes in humans. We carried out mGWAS of 1,172 endogenous and xenobiotic metabolites and applied Netboost on the same metabolites and carried out GWAS on the MEs. This allowed for detection of underlying processes affecting more than one metabolite, effects on metabolites not measured via their correlates and deorphanization of unnamed but measured compounds. We use complementary approaches to map genes onto pathways and show that mGWAS in urine of CKD patients are indeed highly informative of ADME processes, mechanisms of excretion and detoxification, and small molecule metabolism. The comprehensive list of target genes, their corresponding substrates and cell types in humans is relevant to basic science, clinical medicine, and pharmaceutical research.

A flowchart summarizing the design of the Netboost application is shown in Figure 14 and for the GWAS of single metabolites in Figure 15. Here, we used the analyses depicted in Figure 15 as a benchmark to which we can compare the Netboost screening approach presented in Figure 14.

In the paper related to this work ([**2**]) we also studied the following aspects which are not presented here:

- We co-localized health relevant traits of detected associations in 450,000 individuals illuminating mediated molecular mechanisms. We used an adapted version of Giambartolomei's method for co-localization ([**93**]) and identified a multitude of known and novel modes of action which can now be validated experimentally.

- We provided evidence for the generalizability of mQTLs in CKD patients to the healthy SHIP-trend cohort with 977 participants.
- Associations of metabolite ratios to directly represent bivariate enzymatic and transport processes. This helped for example in the identification of a genetic variant associated with slower degradation of metoprolol, a hypertension medication prescribed to more than 20% of GCKD participants. Carriers keep metoprolol in its active form in their blood. This leads to a strong unintended pulse lowering effect.
- Variant prioritization via bayesian statistical fine-mapping to resolve linkage disequilibrium.
- Gene, tissue and cell type prioritization via gene expression analysis of 57,979 single murine kidney cells ([**94**]), 4,524 nuclei of adult human kidney cells ([**95**]) and tissue specific samples for 38 tissues from 714 donors.

1. Methods

1.1. Study design and participants. The GCKD study is an ongoing prospective observational cohort study of CKD patients. Between 2010 and 2012, 5,217 adult CKD patients under nephrological care, provided written informed consent, were enrolled into the study and are followed for clinical endpoints over ten years ([**27**]). Patients were included into the study if they had an estimated Glomerular filtration rate (eGFR) between 30–60 mL/min per 1.73 m^2 or an eGFR >60 mL/min per 1.73 m^2 and overt albuminuria/proteinuria, where overt albuminuria/proteinuria was defined by either urinary albumin-to-creatinine ratio (UACR) >300 mg/g, albuminuria >300 mg/day, a urinary protein-to-creatinine ratio >500 mg/g or proteinuria >500 mg/day([**96**]). A more detailed description of the study design and the recruited study population can be found in previous publications [**97, 27**]. The GCKD Study was registered in the national registry for clinical studies (Deutsches Register Klinischer Studien (DRKS) 00003971) . For this project, urine specimens collected at baseline were selected for metabolite measurements. The analyzed discovery cohort consisted of 1,221 patients with an eGFR <45 mL/min per 1.73 m^2

and the replication cohort of 406 patients with an eGFR between 45-50 mL/min per 1.73 m^2. All patients were selected to have neither micro- nor macroalbuminuria (UACR <30 mg/g) in order to minimize the influence of urinary albumin on metabolite concentrations.

Table 1: **Study sample characteristics GCKD.** For categorical variables % (n) is presented and for continuous variables mean (standard deviation) except for urinary albumin-to-creatinine ratio for which median (interquartile range) is shown. * Characteristics with few missing values (rate of missing values: max 2%)

Characteristic	Discovery (N=1,221)	Replication (N=406)	Overall (N=1,627)
Male sex	56 (683)	52 (213)	55 (896)
Age (years)	63.2 (10.42)	64.8 (8.31)	63.6 (9.96)
BMI (kg/m^2)*	30.2 (6.06)	30.2 (5.46)	30.2 (5.91)
Systolic blood pressure (mm Hg)*	136.8 (19.61)	138.8 (20.3)	137.3 (19.8)
Diabetes	38 (464)	35 (144)	37 (608)
eGFR ($ml/min/1.73m^2$)	40 (13.79)	47.6 (2.62)	41.9 (12.46)
Urinary albumin-to-creatinine ratio (mg/g)	14.2 (6.37-40.44)	8.3 (4.25-15.91)	12.3 (5.71-30.58)

1.2. Genotyping and imputation. Detailed information on genotyping and data cleaning has been described previously in [**6**]. In brief, genomic DNA of 5,123 GCKD participants was extracted and genotyped at 2,612,357 variants using Illumina Omni2.5Exome BeadChip arrays (Illumina, GenomeStudio, Genotyping Module Version 1.9.4). Genotype imputation was performed using minimac3 at the Michigan Imputation Server ([**98**]). The Haplotype Reference Consortium (HRC) haplotypes version r1.1 were used as the reference panel, and Eagle 2.3 was used for phasing. After filtering the imputed genotypes to retain only bi-allelic variants of good or acceptable imputation quality and of minor allele frequency (MAF) ≥1%, 7,750,367 high-quality autosomal variants were available for genome-wide association studies.

1.3. Quality control and data cleaning of quantified metabolites. Sample preparation was carried out as described previously in [**99**] at Metabolon, Inc.

Overall process variability was determined by calculating the median relative standard deviation (RSD)s for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the pooled human urine samples (median RSD = 7-9%; n >1,000 metabolites). All RSDs for metabolites present in at least 90% of the pooled human urine samples are reported in Supplementary Table S1 along other information for each metabolite, including biochemical name, super- and sub-pathway.

After receipt of the quantified metabolites from Metabolon, Inc, an in-house pipeline was set up for data quality control, filtering of metabolites and samples, and for normalizing concentrations to account for urine dilution. No sample had to be excluded for a high proportion of missing data (>50%). On the level of the non-xenobiotic metabolites, 74 (discovery) and 42 (replication) metabolites were excluded because of a high proportion (>80%) of missing values. To account for urine dilution, concentrations of each metabolite were normalized using the probabilistic quotient based on endogenous metabolites with <1% missing values ([**100**]). Subsequently, metabolites were minimum imputed and median scaled ([**101**]). Xenobiotic metabolites were analyzed without imputation because missing values are likely to reflect true absence. None of the remaining metabolites was excluded due to low variance (<0.01) or many outliers (>5% of samples outlying >5 standard deviation (SD)) based on $\log_2$-transformed data. Likewise, no sample represented an outlier >5 SD along any of the first 10 principal components based on metabolites with complete information. Outlying values (>5 SD) for each metabolite were set to missing. Finally, 62 (discovery) and 51 xenobiotic metabolites (replication) with <50 measurements were excluded. Metabolite annotation was aligned between the two batches and yielded a dataset of 1172 metabolites quantified in both discovery and replication. After removal of 27 samples with missing genotypes, the final dataset consisted of 1,221 discovery and 406 replication samples.

1.4. Definition of additional variables. Serum creatinine was measured using an IDMS traceable enzymatic assay (Creatinine plus, Roche). The glomerular

filtration rate was estimated using the Chronic Kidney Disease Epidemiology Collaboration (CKD EPI) formula ([**102**]). The UACR was based on creatinine measured using the same assay as in serum and albumin with the ALBU-XS assay (Roche/Hitachi Diagnostics GmbH, Mannheim, Germany).

1.5. Genome-wide association studies of urinary metabolite concentrations. Prior to GWAS, metabolite concentrations were $\log_2$-transformed to reduce the observed skewness and generate approximately normally distributed data for analysis while maintaining an easy interpretability of units. Similar to previous GWAS of metabolite concentrations ([**88, 89, 6**]), residuals adjusted for age, sex, log(eGFR), log(UACR) and the first three genetic principal components were generated. GWAS were then performed on these residuals separately for discovery and replication as described previously in [**6**] using imputed genotype dosages and assuming an additive genetic model. While not ideally powered for recessive variants, this is generally preferred to incorporating the additional tests and increasing the multiple testing penalty. Summary statistics from discovery and replication were subjected to quality control using GWAtoolbox ([**103**]) and subsequently meta-analyzed assuming a fixed effects inverse variance model as implemented in METAL ([**104**]), retaining only metabolites with a minimum sample size of 300 in the meta-analysis. Statistical significance was defined as genome-wide significant (p-value <5e-8) in the discovery cohort, a one-sided p-value of <0.05 in the replication cohort, and significant in the meta-analysis after correcting for testing of 1,172 metabolites by a Bonferroni procedure (p-value <4.3e-11 = 5e-8/1172). The established genome-wide significance threshold of 5e-8 originates from a Bonferroni adjustment for one million independent SNPs with a minor allel frequency great than 1%.

Significantly associated SNPs were assigned to loci by selecting, for each metabolite, the SNP with the lowest p-value across the genome as the index SNP, defining the corresponding locus as a 1-Mb interval centered on the index SNP, and repeating the procedure until no further genome-wide significant SNP remained. For each

metabolite, overlapping windows were combined into "loci" and clipped at chromosomal borders. The extended MHC region (chromosome 6, 25.5-34 Mb) was considered as one region. For each metabolite and each significantly associated locus, a regional association plot centered on the index SNP was generated using the stand-alone version of LocusZoom (v1.3) ([**105**]) (Supplementary Figure S1). Loci were further merged across metabolites into genetic regions if the index SNPs of the different metabolites were in linkage disequilibrium (r^2>0.8, as defined in [**106**]). A circular plot of associations with metabolites was created using Circos version 0.69-6 (Figure S4). A sensitivity analysis was conducted without adjustment for log(eGFR) and log(UACR) and yielded very similar results (231/232 detected mQTLs overlapped with results from the main analysis; data not shown).

1.6. Annotation. Annotation of SNPs was performed by querying the single nucleotide polymorphisms annotator (SNiPA) database v3.3 (released June 25th, 2018) ([**107**]) and genomic positions correspond to build 37 (GRCh37). SNiPA was used to collect the following annotations for each index SNPs and its proxies ($r^2 \geq 0.8$): gene hit or close-by, regulated genes, Combined Annotation Dependent Depletion (CADD) score, SnpEff effect impact (exonic and noncoding), mQTL, protein quantitative trait locus (pQTL), GWAS Catalog, cis eQTL, disease genes (based on Clinvar, OMIM, HGMD and Drugbank). Novelty of loci and regions identified in our screen were assigned, based on the presence of SNiPA entries in the "mQTL" and "GWAS Catalog" categories for the respective index SNP and its proxies, as "confirmed for urine", "novel for urine" (but identified in another body fluid) and "novel". The asterisk assignment for genetic regions containing novel substrates was based on string matching of biochemical names within these "mQTL" and "GWAS catalog" entries. For index SNPs that were missing in SNiPA, ldlink ([**108**]) was used to identify the best proxy that was part of the SNiPA database, and proxy information was used instead.

To select the most likely causal gene for each index SNP, we first compiled the SNiPA "genes" and "evidence" information. The evidence codes h, r, e, p, m and

c correspond to gene hit or close-by, regulated genes, cis-eQTL, pQTL, missense variants, and disease genes based on specific variants known to cause monogenic diseases, respectively. Evidence codes were summed for each gene. Additionally, index SNPs were queried for association with differential expression of a nearby gene in tubulo-interstitial kidney portions (cis-expression quantitative trait locus (eQTL)) using the NephQTL browser, a gene expression resource based on kidney biopsies from 187 patients with CKD ([**109**]). When one or more eQTL associations with p-value $<0.05/159$ were identified within ± 100 kb of each index SNP, co-localization analyses of the respective metabolite(s) mQTL and each of the eQTL association(s) were performed, with the region for each co-localization test defined as the eQTL cis window in the underlying study (± 500 kb). We used an adapted version of Giambartolomei's method for co-localization ([**93**]) as implemented in the 'coloc.fast' function from the `R` package gtx (`https://github.com/tobyjohnson/gtx`) and used default parameters and prior definitions. Each gene with evidence for co-localization also received a scoring. The gene with the highest sum of scores within each locus was assigned as the most likely causal gene. In the case of ties, genes with evidence for co-localization were prioritized, followed by genes for which an inborn error of metabolism with the corresponding metabolite is known. In all other cases, ties were resolved by prioritizing the closest gene.

1.7. Genome-wide association studies of Netboost modules. Metabolites were $\log_2$-transformed and imputed using a k-nearest neighbor (knn) algorithm with k=10 to impute missing values ([**101**]). Netboost was applied as implemented in the Bioconductor `R` package Netboost v1.0.0. MEs were dubbed eigenmetabolites.

Netboost was applied to data from the discovery cohort, and the resulting clustering was transferred to the replication cohort. Replication data in the AML application in Chapter 3: 2.3 did not consist of the full set of variables and we refitted MEs after transferring the grouping. Here, we had identical sets of variables and combined the rotation matrixes from module-wise PCAs and transfered MEs including their directions.

Analogous to the single metabolite screen, GWAS were conducted separately for discovery and replication summary measures and results were meta analyzed using METAL ([**104**]) assuming a fixed effects inverse variance model. Statistical significance was defined as a genome-wide significant (p-value <5e-8) in the discovery cohort, a one-sided p-value of <0.05 in the replication cohort, and a significant (p-value <2.3e-10) in the meta-analysis after correcting for testing of 212 eigenmetabolites by a Bonferroni procedure (5e-8/212). Significantly associated 1-Mb intervals were merged by overlap, and loci were merged into genetic regions across eigenmetabolites if their index SNPs were correlated (r^2>0.8, as defined in [**106**]). The extended MHC region (chromosome 6, 25.5-34 Mb) was considered as one region.

For each eigenmetabolite and each significantly associated locus, a regional association plot centered on the index SNP was generated using the stand-alone version of LocusZoom (v1.3) ([**105**]) (see Supplementary Figure S2). Loci were further merged across metabolites into genetic regions if the index SNPs of the different metabolites were in linkage disequilibrium (r^2>0.8). A circular plot of associations with eigenmetabolites were created using Circos version 0.69-6 (see Figure 17).

1.8. Curation of genes involved in ADME processes. The curation of genes involved in ADME processes was done by Franziska Grundner-Culemann and Anna Köttgen. The list of 298 ADME core genes was obtained from `http://www.pharmaadme.org/`. In addition, an extended list was manually curated by the identification of additional members of gene families known to be involved in phase I, II and III biotransformation reactions, starting out with the list of families included at `https://en.wikipedia.org/wiki/ADME` on November 23rd, 2018. Lastly, all 86 unique genes identified in the GWAS of urinary metabolite concentrations were evaluated for the presence of publications on their involvement in phase I, II and III biotransformation reactions in a PubMed search in December of 2018.

1.9. ADME, GO and KEGG enrichment analyses. Using PLINK v1.9063, we computed the number of independent SNPs per gene based on GCKD genotypes. With the Bioconductor `R` database org.Hs.eg.db v3.8.2 we extended this to a database of Entrez gene identifiers additionally annotated for gene length, ADME genes, Gene Ontology (GO) terms ([**110**]) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways ([**111**]). To test enrichment with respect to ADME, GO terms and KEGG pathways we performed random draws which were matched with respect to deciles of gene length and deciles of the number of independent SNPs (GO and KEGG: 1e7 draws; ADME: 1e8 draws). When none of the random draws matched or exceeded the selected genes, we indicated the p-value with its upper limit (e.g. p-value <1.0e-7). P-values were adjusted using the Benjamini–Hochberg procedure ([**112**]).

2. Metabolome-wide GWAS in chronic kidney disease patients

We performed GWAS of the concentrations of 1,172 metabolites and 212 eigenmetabolites in urine from a discovery sample of 1,221 and a replication sample of 406 independent CKD patients, followed by meta-analysis and downstream characterization of replicated findings. These patients were selected from the GCKD study ([**27, 97**]) as participants with eGFR of <50 ml/min/1.73m^2 and with normoalbuminuria (UACR <30 mg/g).

2.1. Netboost provides biological context for yet unnamed metabolites and improves power. Metabolites are intermediates of homeostatic reactions and as such inter-connected beyond pair-wise relationships. Groups of correlated metabolites may reflect shared biochemical pathways or be co-regulated. We used Netboost to construct 212 metabolite modules (Figure 16) and their respective eigenmetabolites. GWAS of these eigenmetabolites identified and replicated 46 genomic intervals ("loci", Chapter 4: 1.7) that contained at least one SNP significantly associated with at least one of 38 unique urinary eigenmetabolite levels (p-value<2.3e-10, Figure 17).

For each of the loci, regression estimates were annotated with module composition as well as exonic and intronic effects of the SNP and the most likely underlying gene (Supplementary Table S2, Chapter 4: 1.6). The regional association plots (RAPs) illustrate the local correlation structure of SNPs (Supplementary Figure S2, Chapter 4: 1.7). Out of the 38 modules which had at least one significant association, 32 consisted of metabolites from one super pathway or combined metabolites from one super pathway with yet unnamed metabolites ("unknowns"). In this manner, Netboost aided in the identification of yet unknown metabolites. Guided by the inferred network and shared association, we were able to identify formerly unknown metabolites together with Metabolon, Inc. E.g., a module of known vitamin E (tocopherol)-related metabolites also contained the two unknowns X-13689 and X-24359 (Supplementary Table S2) and was associated with rs55744319, which is in high linkage disequilibrium (LD) with a missense variant in *CYP4F2*, encoding p.Val433Met. This variant has previously been identified in response to vitamin E supplementation ([**113**]), vitamin E levels ([**114**]), and warfarin maintenance dose ([**114, 115**]). Investigation of the unknown metabolites based on their mass, retention time, spectral information and genetic evidence nominated the unknown molecules as structurally related to Vitamin E, with the glucuronide of alpha-CMBHC as a candidate for X-13689. We experimentally verified this prediction through the examination and comparison of retention times from ion chromatograms and the locations and intensities of the MS/MS fragmentation spectra between a standard of glucuronide of alpha-CMBHC and X-13689 (Supplementary Figure S3). Thus, knowledge of a yet uncharacterized metabolite's module membership and its genetic association can provide information beyond mass and retention time by restricting the search space of their possible identity for experimental verification.

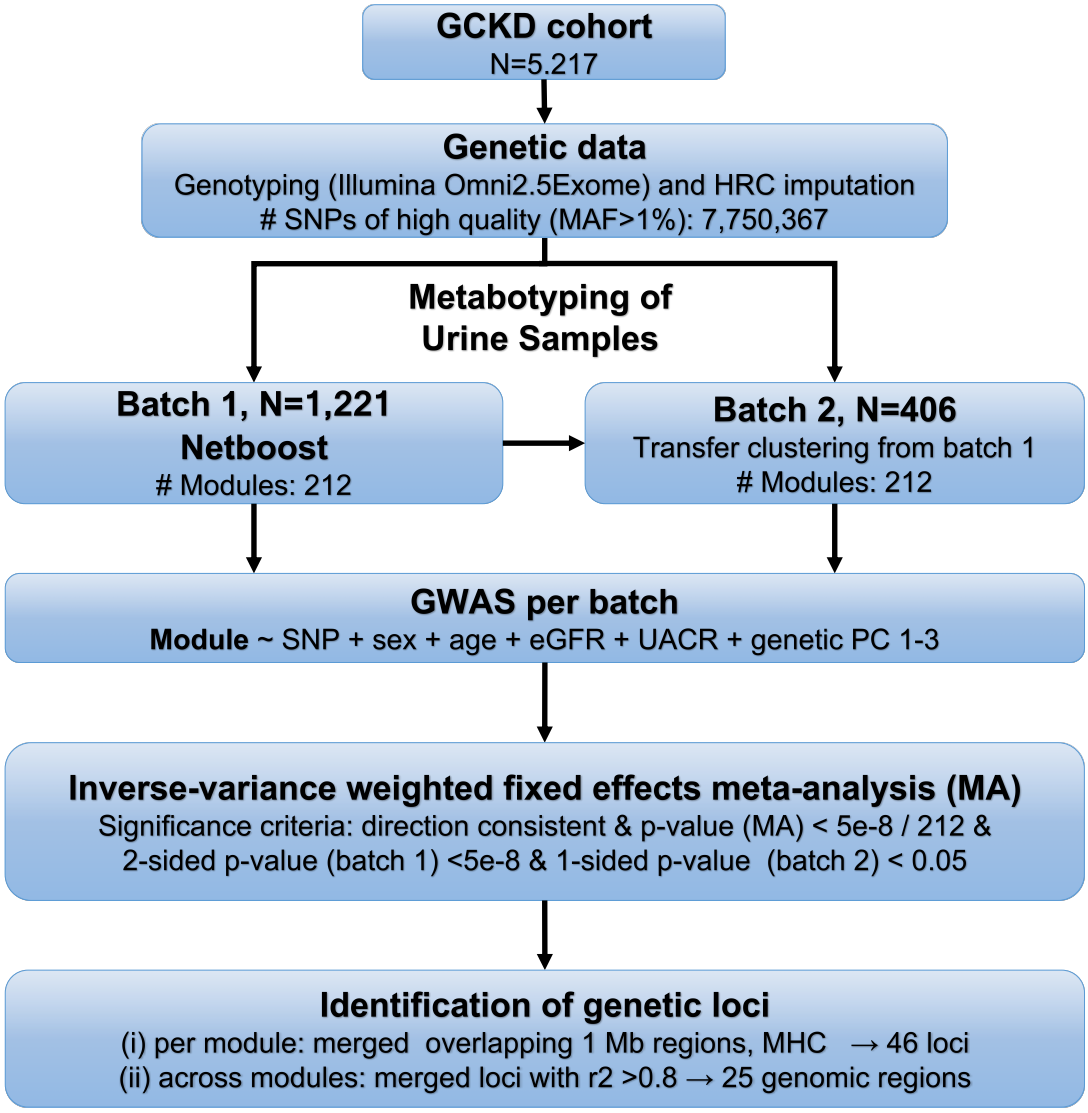

Figure 14: **Overview of the Netboost application for GCKD metabolomics and genetics.** Schematic representation of the GWAS for eigenmetabolites. Top to bottom: Genetic data of the 5,217 patients in the GCKD study was preprocessed and combined with the urinary metabolite concentrations of 1,221 (batch 1) and 406 (batch 2) of these patients. Netboost was applied to batch 1 and the network structure transfered to batch 2. Next, GWAS for eigenmetabolites were performed in each batch and then meta analyzed. Finally, significant associations were merged into genetic loci.

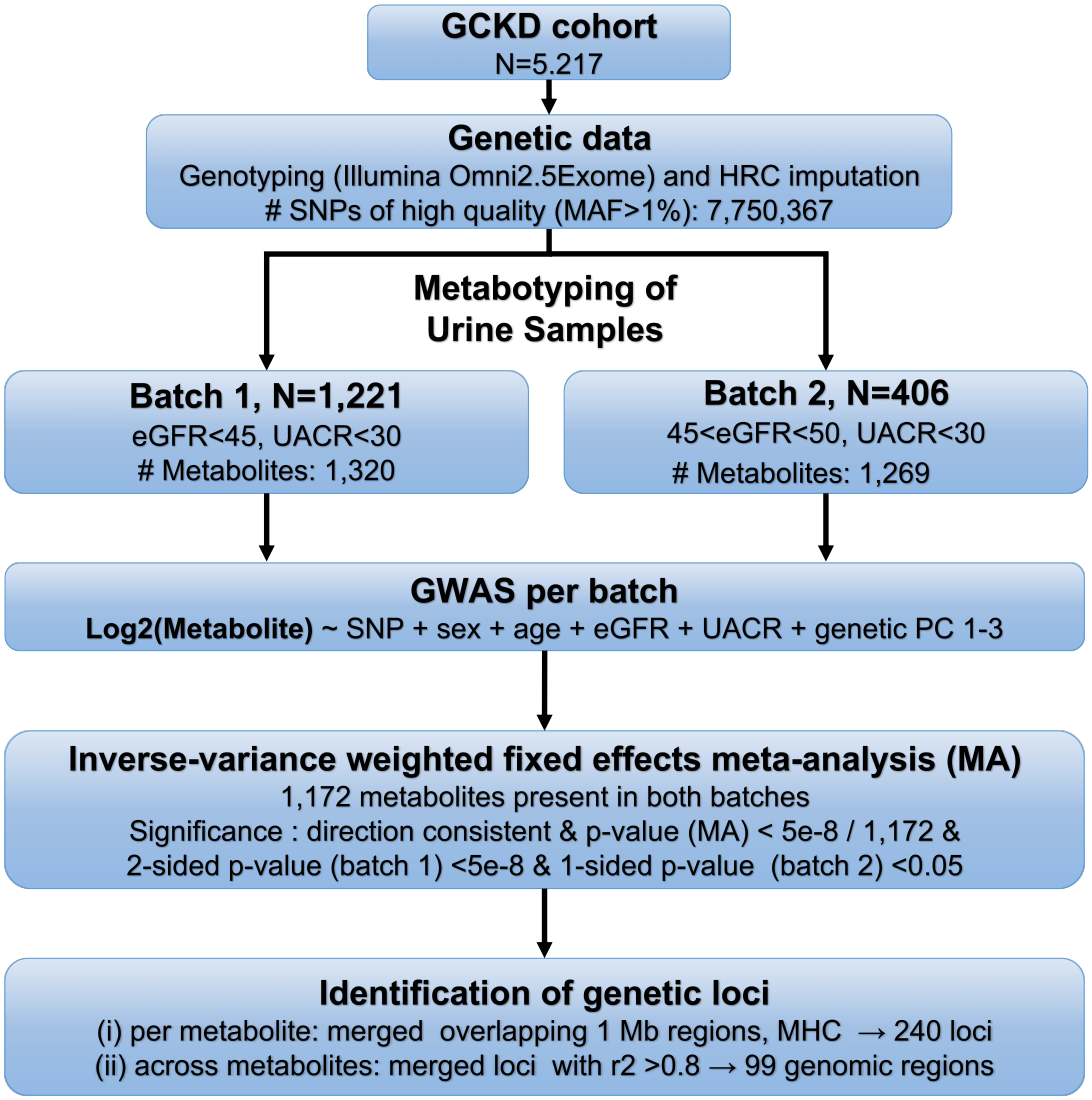

Figure 15: **Overview of the reference GWAS for GCKD metabolomics and genetics.** Schematic representation of the GWAS for single metabolites. Top to bottom: Genetic data of the 5,217 patients in the GCKD study was preprocessed and combined with the urinary metabolite concentrations of 1,221 (batch 1) and 406 (batch 2) of these patients. Next, GWAS for metabolite concentrations were performed in each batch and then meta analyzed. Finally, significant associations were merged into genetic loci.

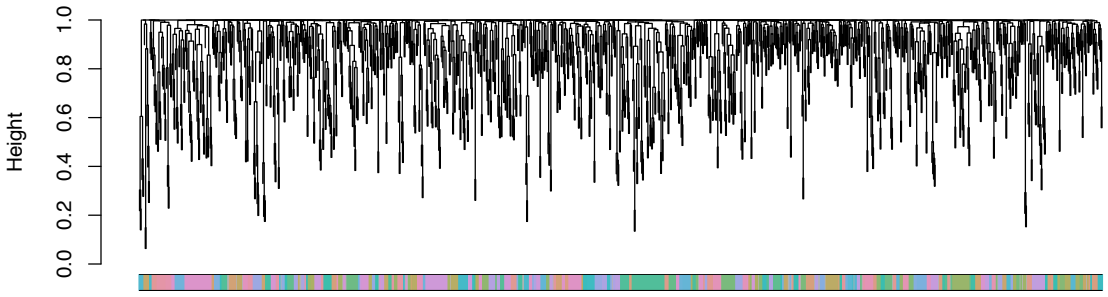

Figure 16: **Netboost dendrogram of GCKD metabolomics.** The band of color indicates membership of each of the 1172 metabolites in one of 212 Netboost modules.

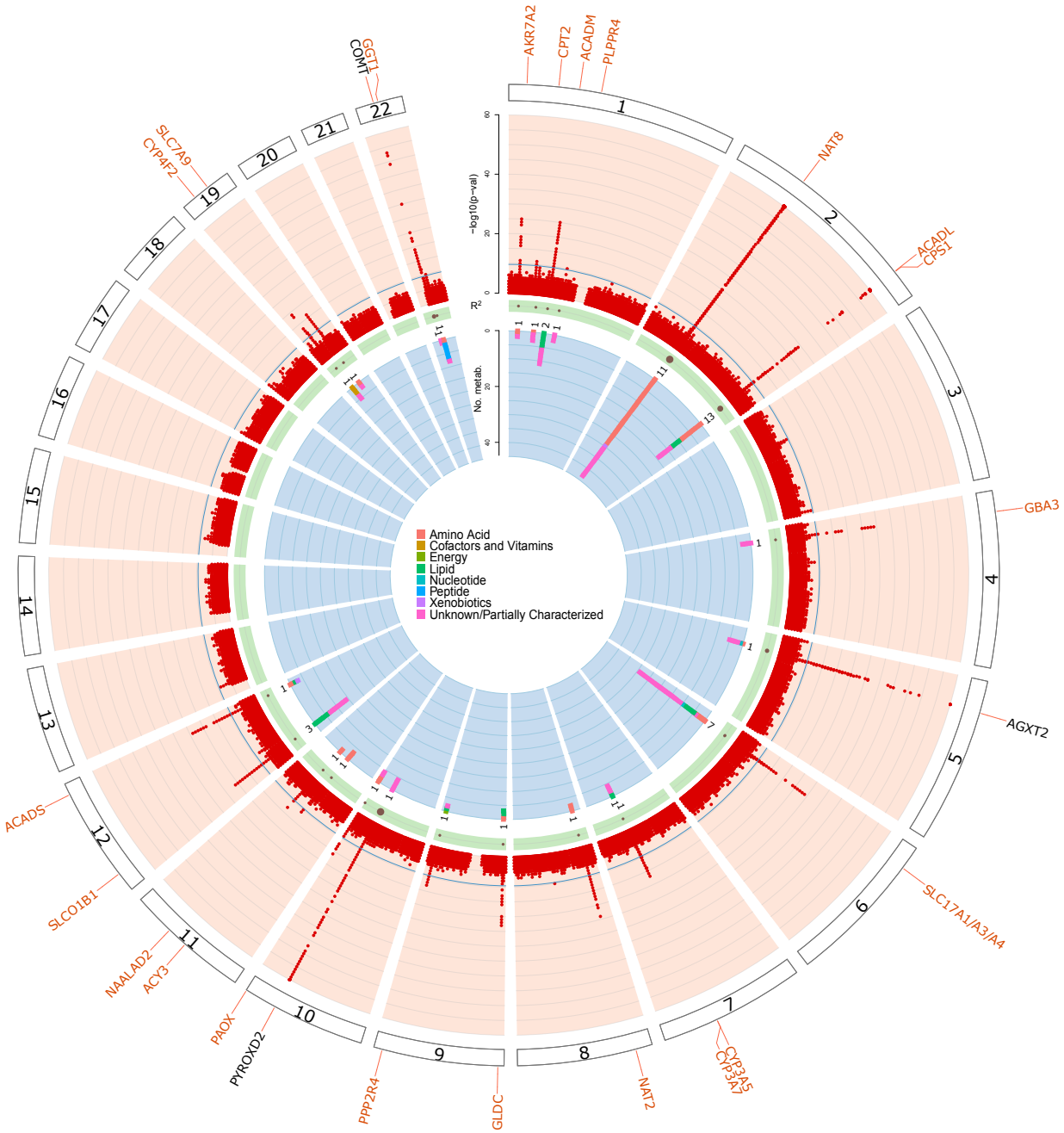
AKR7A2
CPT2
ACADM
PLPPR4
NAT8
ACADL
CPS1
GBA3
AGXT2
SLC17A1/A3/A4
CYP3A5
CYP3A7
NAT2
GLDC
PPP2R4
PYROXD2
PAOX
NAALAD2
ACY3
SLCO1B1
ACADS
SLC7A9
CYP4F2
COMT
GGT1
−log10(p−val)
R²
No. metab.
Amino Acid
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Peptide
Xenobiotics
Unknown/Partially Characterized

Figure 17: **Genetic associations with eigenmetabolites.** The light red band shows the $-\log_{10}$(p-value) for genetic associations with eigenmetabolite concentrations, representing their respective module, by chromosomal position. Associations of all 212 eigenmetabolites are overlaid in the red band, and are capped at p-value=1e-60. The blue line indicates genome-wide significance (p-value=2.4e-10). Black gene labels indicate genetic regions in which all members of a given module were also identified in the single metabolite mGWAS, orange labels indicate genetic regions were additional metabolites were implicated as members of a module. The light green band shows the maximum variance in eigenmetabolite levels explained by the index SNP at each genetic region, with dot sizes corresponding to ([0,0.1],(0.1,0.25],(0.25,0.5],(0.5,1]) of explained variance. The inner blue band shows a stacked representation of the number of implicated metabolites in each genetic region, is colored according to the super-pathways to which they belong, and the number of modules in the genetic region is given next to it. Color keys of metabolite super-pathways are presented in the middle.

Across the 1,172 GWAS of metabolite concentrations, we identified and replicated 240 loci (Chapter 4: 1.5) associated with at least one of 211 unique urinary metabolite concentrations (Supplementary Table S3, Supplementary Figure S1+S4). Of the 240 loci, there were 26 in which the index SNP or a good proxy (r^2>0.8) had previously been described in association with the same urinary metabolite ([**116, 88, 117**]), and we additionally identified 54 newly associated metabolites in these 26 known regions. At the time of annotation, the remaining 160 loci represented novel loci for urinary metabolite concentrations (Supplementary Table S3), some of which had been previously detected in blood ([**118, 89, 107, 119**]). The variance in metabolite levels explained by the index SNP at each locus ranged from 2.0% to 63.1% (Supplementary Table S3), and was generally very high in comparison to that of commonly studied complex traits and diseases ([**120**]), highlighting the close and specific link between the genome and the metabolome. Notably, the

variance in eigenmetabolite levels explained by the index SNP exceeded even this and reached up to 72.0% (Supplementary Table S2).

The overwhelming majority of associations was biologically plausible and confirmed that mGWAS in urine can capture ongoing intracellular enzymatic reactions and transport across membranes. As an example of intracellular reactions that balance metabolite concentrations, urinary N-acetyl-tyrosine, N-acetyl-phenylalanine and eigenmetabolite ME169 levels were each associated with SNPs mapping into *ACY3* on chromosome 11, encoding an enzyme important in de-acetylating N-acetylated aromatic amino acids in kidney proximal tubules ([**121**]), and with SNPs mapping into *NAT8* on chromosome 2, important in the N-acetylation of metabolites in renal proximal tubule and liver cells ([**122**]). Another example highlights the hypothesis-generating potential with respect to transport processes: concentrations of 3-aminoisobutyrate in urine were associated with SNPs mapping into *AGXT2*, encoding the enzyme responsible for its metabolization ([**123**]), and with SNPs mapping into *SLC6A13*, which is known to encode a transporter of gamma-aminobutyric acid (GABA) that is highly expressed in the kidney ([**124**]) but has not been shown to transport 3-aminoisobutyrate. GABA is a structural analog of 3-aminoisobutyrate, nominating 3-aminoisobutyrate as a novel candidate substrate for renal *SLC6A13*, which can now be tested experimentally.

The three loci with the strongest associations among metabolite concentrations were detected for yet unnamed metabolites at *PYROXD2* (X-24809, p-value=3.6e-574), *NAT8* (X-12125, p-value=2.4e-570), and *AKRD7A2* (X-24462, p-value= 2.3e-412). While all three genetic regions were also identified with eigenmetabolites, two showed particularly strong genetic associations. Eigenmetabolite ME193 originating from a module of five unknown metabolites (X-12093, X-12112, X-23776, X-24809 and X-24983) and was associated with missense rs2147896 in *PYROXD2* (p-value=2.5e-917; Figure 18). ME161, a module composed of N2-acetyllysine, N-alpha-acetylornithine, X-12124, X-12125, and X-15666, was associated with missense rs13538 in *NAT8* (p-value=4.3e-635; Supplementary Figure S5). Such associations

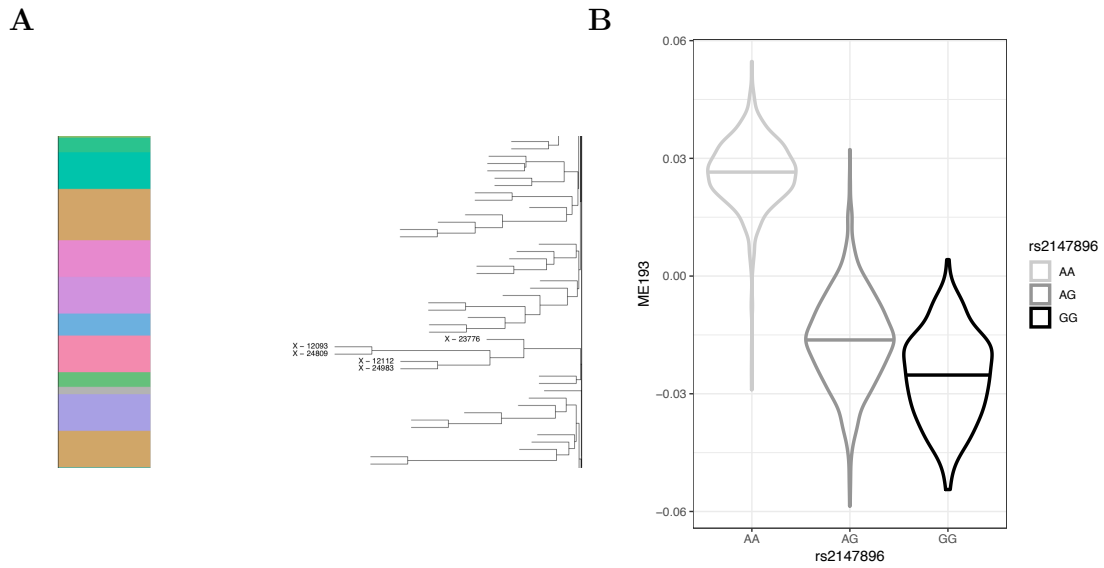

Figure 18: **Eigenmetabolite ME193 composition and genetic association with *PYROXD2* variants.** A) shows module ME193, for which metabolites are labeled, within the dendrogram for GCKD metabolites (Figure 16). B) displays the distribution of the eigenmetabolite of ME193 (Y-axis) with genotype at rs2147896 in *PYROXD2* (X-axis).

are suggestive of a common function of the enzyme on metabolites in the module, implicating the unknown molecules in the *NAT8*-associated module as additional N-acetylated compounds or their precursors.

When comparing the RAPs of ME193, the strongest association signal in the study, and X-24809, which is part of the module, the similarity of the association signal becomes apparent. Simultaneously, the difference in the y-axis illustrates the stronger link to the Netboost derived eigenmetabolite levels (Figure 19). In total, there were 13 module associations which had a stronger association signal than any of their individual module members. Additionally, screening of eigenmetabolites provided the important advantage of permitting a complete screen of higher order genetic associations. The assessment of all pair-wise metabolite ratios, another strategy we partially explored in **[2]**, would already have accumulated to 686,206 GWAS.

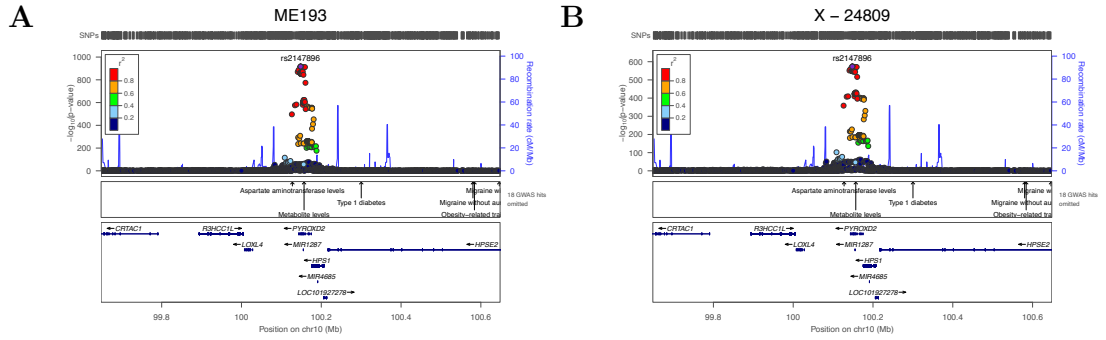

Figure 19: **Regional association plots of *PYROXD2*.** P-values (y-axis) surrounding *PYROXD2* are showen along the chromosome (x-axis). The index SNP with the lowest p-value, rs2147896, is indicated. LD (r^2) used to color-code correlation with the index SNP was based on the analyzed subsample of the GCKD study. A) displays the RAP of rs2147896 in *PYROXD2* for ME193, the module with the strongest association signal. B) displays the RAP of rs2147896 in *PYROXD2* for X-24809, the metabolite with the strongest association signal.

Furthermore, 35 of the 46 replicated eigenmetabolite associations implicate additional metabolites that were not identified individually in mGWAS after correction for multiple testing (Figure 17, Supplementary Table S2). For example, GWAS of acisoga concentrations would not have detected an association at the *PAOX* locus (p-value=9.7e-7), but acisoga was part of the replicated module ME97 (eigenmetabolite p-value=9.0e-17). This links acisoga to *PAOX* function and illustrates that modules can places metabolites into their biological context, given that acisoga is a catabolic product of acetylspermidine, that acetylspermidine was another member of module ME97, and that the *PAOX*-encoded polyamine oxidase acts on acetylspermidine as a substrate. In total, Netboost implicated 191 such mQTLs with the 46 eigenmetabolites. Out of these, 75 did not reach the meta-analysis multiple testing threshold on their own but only as part of the module (Supplementary Table S2).

2.2. Identified genes illuminate ADME processes, handling of uremic toxins and amino acid metabolism in humans. The most likely gene underlying the association signal in each locus was annotated based on proximity, functional consequence, and regulatory potential of the index SNP and its association with gene expression (4: 1.6), resulting in 86 unique genes across eigenmetabolites and metabolites.

Next, we aimed to identify common processes and pathways in which these 86 unique genes may be involved. These genes were strongly enriched among 298 genes known to participate in ADME processes in humans (p-value<1.0e-8, Chapter 4: 1.9, Supplementary Table S4), with 23 of the 86 genes annotated to phase I, II, and III detoxification and excretion reactions known to be important in drug metabolism. Additional consideration of a manually curated, extended list of 544 human ADME genes and targeted literature review allowed the placing of 14 additional genes into the ADME context. This strongly supports the idea that studies of metabolite concentrations in urine provide an integrative read-out of ongoing ADME processes in humans.

We systematically tested for enrichment of the 86 genes in pathways, molecular functions, cellular components and biological processes as contained in the GO ([**110**]) and KEGG ([**111**]) databases. Altogether, there were 153 significantly enriched GO terms, and 22 KEGG pathways (Chapter 4: 1.9, Supplementary Table S4). Both GO biological functions and KEGG pathways implicated several processes related to detoxification and drug metabolism as strongly enriched (e.g., "xenobiotic metabolic process", p-value<1.0e-7; "chemical carcinogenesis", p-value<1.0e-7), consistent with the role of ADME processes. Additionally, enrichment was observed for processes related to metabolism and catabolism of small molecules, including organic, carboxylic, amino and fatty acids. This is consistent with the prominent role of the proximal tubule in the metabolism as well as active reabsorption and secretion of amino, organic and carboxylic acids, and with the importance of fatty acid metabolism to satisfy its high metabolic need. Highly enriched molecular functions

further supported the importance of enzymes mediating phase I and II ADME biotransformation reactions (e.g., "cofactor binding", p-value<1.0e-7; "monooxygenase activity", p-value<1.0e-7). The most strongly enriched cellular components ("mitochondrial matrix", p-value=2.6e-6; "mitochondrion", p-value=8.1e-6) were consistent with the localization where many detoxification reactions, fatty acid metabolism, as well as amino acid metabolism and transport are known to occur. This study of CKD patients was also informative of the metabolism of uremic toxins, as reflected in the enrichment of the biological processes "kynurenine metabolic process" (p-value=1.2e-6) and "amine metabolic process" (p-value=3.9e-6). The genes *AFMID*, *GOT2*, *KMO* and *KYAT3* encode enzymes operating on kynurenine and its metabolites, which rise in the setting of uremia ([**125, 126**]). Similarly, the polyamines spermidine, spermine and putrescine accumulate with declining kidney function ([**125**]), which probably aided in the identification of an enriched proportion of genes involved in amine metabolism (*AGMAT*, *KMO*, *HDAC10*, *PAOX*, *SULT1A2*, *AFMID* and *COMT*). Together, the enriched processes, pathways, molecular functions and cellular components support the notion that studies of a specific biosample in a selected study population can highlight processes of special importance in a given organ and clinical setting.

In summary, this study generates a catalog of genes, causal variants, molecular mechanisms and metabolome modules that constitutes a comprehensive resource to guide experimental studies in physiology, basic science, clinical medicine and the pharmaceutical industry.

What is new in Chapter 4:

- Novel methodology to explore higher order genetic associations in an unbiased unsupervised manner by integration of Netboost with multi-trait GWAS.
- Identification and replication of 46 genetic metabolite module associations. Often module associations are stronger than any of the individual metabolites and additional metabolites can be implicated.
- Nearly 4 times the number of genes associated with metabolism in urine when compared to previous studies (86 genes vs 22 genes).
- Modules provide biological context for interpretation, which e.g. aids in deorphanization of unknown metabolites.

CHAPTER 5

Netboost for classification

Huntington's Disease (HD) is driven by the number of cytosine, adenine and guanine (CAG) trinucleotide repeats in the huntingtin gene. Langfelder et al. [**33**] used WGCNA to reveal 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 and recently the analysis was extended to miRNA from the same mice in [**37**].

To evaluate the performance of Netboost, we used the messenger ribonucleic acid (mRNA) dataset in an inverse setup and determined the prediction errors in a classification task for the known disease severities of the mice. As the primary classifier, random forests were used as these had for the required versatility to be applied with and without the dimension reduction.

1. Methods

1.1. Gene expression data. We downloaded the 48 mRNA-sequencing samples from mouse striatum from the Gene Expression Omnibus (GEO) (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65774` last accessed January 5th, 2019). Data originated from six genetically engineered mouse strains with different disease severities (20, 80, 92, 111, 140 and 175 CAG trinucleotide repeats) with four female and fourres[i male mice each. Aged 6 months mice were harvested and processed. The mRNA-sequences were already preprocessed and after removal of invariant transcripts, data consisted of 28,010 transcripts.

1.2. Random forests. Random forests are extensions of tree-based classifiers. These algorithms are non-parametric and combine stratification techniques for classification and prediction. They learn in a supervised manner and one of the most

well known tree-based approaches is the classification and regression tree (CART) procedure ([**127**]). In [**28**] Breiman extended this concept to random forests by growing many trees and then ensembling them into one classifier. Randomness is introduced into the procedure as each tree is based on a randomly drawn bootstrap sample of the original data and as at each splitting point of each tree only a random subset of the variables is used as candidates for splitting. The number of trees is not prone to overfitting and can be set to a large value ([**128**]). The main tuning parameter is the number of considered variables at each splitting point, which can be determined using cross-validation.

We applied random forests as described in [**28**] to classify samples based on X and 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 grew 10,000 trees in each analysis. The implementation in the randomForest `R` package with 200 iterations of leave-one-out cross-validation was used ([**28**]).

2. Classification of disease severity for Huntington's disease

For the HD-mouse-models we compared three setups based on the RNA-sequencing data for prediction of the underlying CAG trinucleotide repeats:

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

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

Blockwise WGCNA was applied as described in Chapter 3: 1.3 and identified 61 modules with a mean module size of 423 in the range of 11 to 6221. Ten was set as the minimum module size. Henceforth, 92% of the features were assigned to modules. The proportion of variance explained by MEs was comparably low and

reached a median of 42.1% (range = [29.3%, 63.4%], Figure 20). On the WGCNA aggregated $X_{\text{WGCNA modules}}$ the mean prediction error was 37.1%.

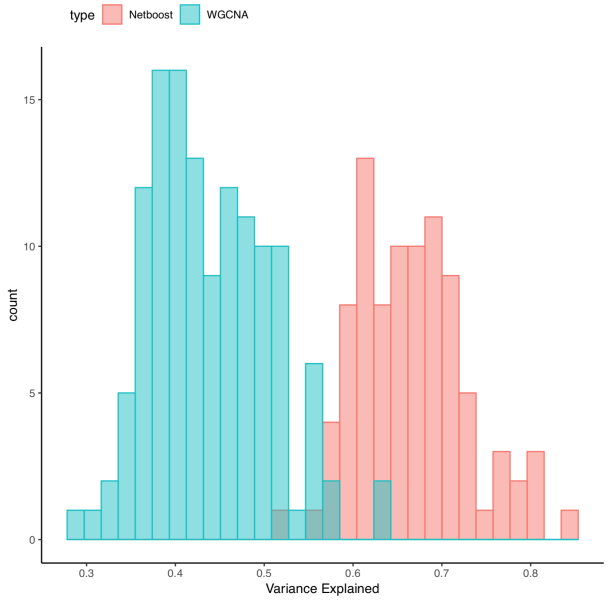

Figure 20: **Histogram of the proportion of variance explained by MEs in the HD dataset**

Netboost was applied as described in Chapter 2: 2 and the multivariate filter was stopped after 20 steps, resulting in $|\mathcal{F}| = 247,497$. This represents approximately 0.06% 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% of the features were assigned to modules and MEs of the Netboost modules explained a higher proportion of variance (median = 66.2%, range = [52.3%, 84.9%], Figure 20). On the Netboost aggregated $X_{\text{Netboost modules}}$ the mean prediction error was 28.2%. The dendrogram based on the sparse network is depicted in Figure 21. 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).

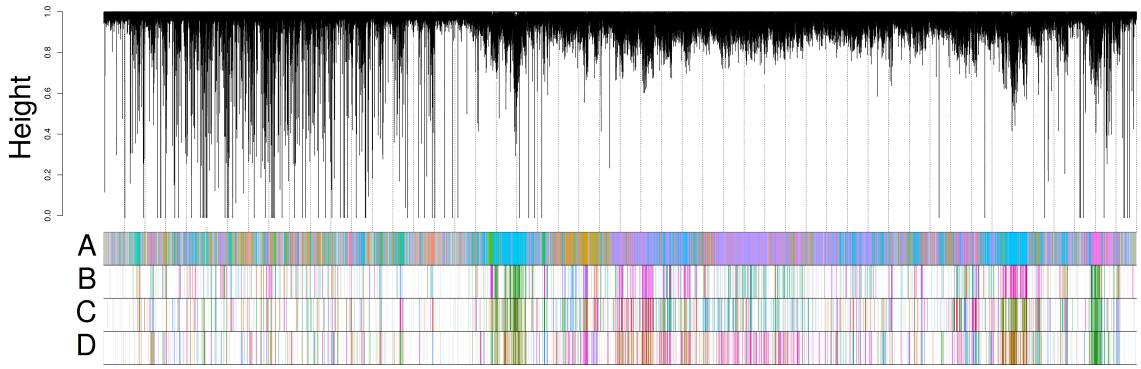

Figure 21: **Dendrogram of Huntington's disease data.** Dendrogram of the gene expression features in the Huntington's disease data. A) shows the separation into modules by blockwise WGCNA and B), C) and D) show Netboost modules with 25, 20 and 15 boosting steps respectively.

What is new in Chapter 5:

- Integration of Netboost with random forests as a modeling strategy for high-dimensional classification.
- Significantly improved prediction errors for the HD application.

CHAPTER 6

Network preservation

In high-dimensional analyses, stability of an algorithm is a crucial factor. Due to the challenges of $n \ll p$ and the diversity of data types, results can be unstable ([**129**]). In the following Chapter, we study the preservation of networks as predicted by Netboost. To do this, we first use resampling-based methodology to estimate the sampling uncertainty ([**130**]) and compare this to other clustering methods. Subsequently, we set this sampling uncertainty in relation to the method uncertainty and then study module-wise preservation in more detail.

1. Methods

1.1. Cluster indices. We define a clustering $M = \{M_1, \ldots, M_l\} \subseteq \mathfrak{P}(\{1, \ldots, p\})$ as a partition of variables, i.e. $M_1, \ldots, M_l$ are non-empty, disjoint and $M_1 \cup \ldots \cup M_l = \{1, \ldots, p\}$. For two clusterings M, M', the contingency table is given by

$$C = (c_{rs})_{r \leq l, s \leq l'} := (|M_r \cap M'_s|))_{r \leq l, s \leq l'}.$$

When evaluating the similarity of two clusterings, many measures are based on counts of unordered pairs of variables, which can be split into four groups

$$S_{11} := \{\{i,j\} \in \{1,\ldots,p\}^2 \mid i \neq j \wedge \exists r \leq l : i,j \in M_r \wedge \exists s \leq l' : i,j \in M'_s\},$$

$$S_{00} := \{\{i,j\} \in \{1,\ldots,p\}^2 \mid i \neq j \wedge \nexists r \leq l : i,j \in M_r \wedge \nexists s \leq l' : i,j \in M'_s\},$$

$$S_{10} := \{\{i,j\} \in \{1,\ldots,p\}^2 \mid i \neq j \wedge \exists r \leq l : i,j \in M_r \wedge \nexists s \leq l' : i,j \in M'_s\},$$

$$S_{01} := \{\{i,j\} \in \{1,\ldots,p\}^2 \mid i \neq j \wedge \nexists r \leq l : i,j \in M_r \wedge \exists s \leq l' : i,j \in M'_s\}.$$

Their cardinalities are denoted by n_{11}, n_{00}, n_{10} and n_{01} ([**131**]). In Figure 22 four groups of nodes are displayed with two illustrated clusterings, coded by proximity. For example the edges between turquoise nodes belong to S_{11} as in both panels all

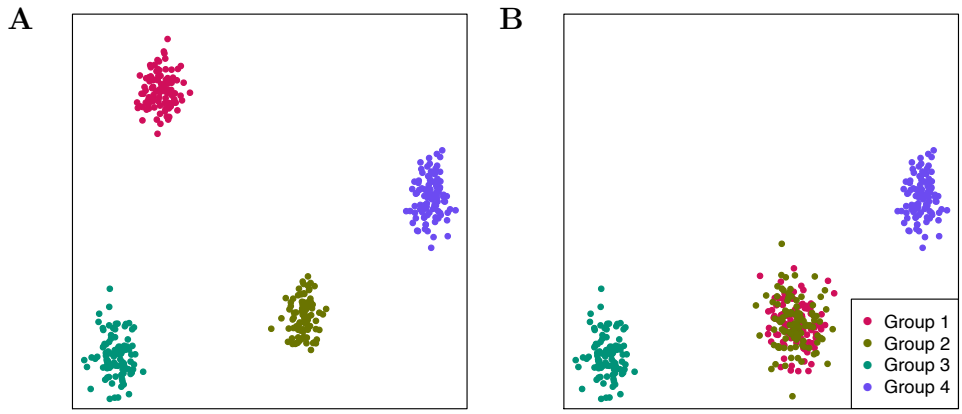

Figure 22: **Illustration of clustering edge counts.** Two network representations based on the same set of nodes.

turquoise nodes are grouped together, whereas any turquoise-violet edge belongs to S_{00} and the red-olive edges belong to S_{01}.

The General Rand Index is defined by the proportion of pairs being identically clustered,

$$\text{Rand}(M, M') := \frac{2(n_{11} + n_{00})}{p(p-1)}$$

and ranges from 0 to 1. This measure depends on both, the number of clusters and the number of elements. Furthermore, the expected value of random partitions is not necessarily equal to zero. Therefore, Hubert and Arabie proposed in [**132**] an adjusted version. The adjusted Rand Index assumes a generalized hypergeometric distribution as the null hypothesis. That is given $M_r \in M$ we assume the probability distribution of the random variable Y describing the overlap size with an $M'_s \in M'$ to be

$$P(Y = a) = \frac{\binom{|M_r|}{a}\binom{p-|M_r|}{|M'_s|-a}}{\binom{p}{|M'_s|}}.$$

The adjusted Rand Index then corrects the expected value to zero and is given by

$$\text{adjRand}(M, M') := \frac{\sum_{r=1}^{l}\sum_{s=1}^{l'}\binom{c_{rs}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}$$

where $t_1 = \sum_{r=1}^{l} \binom{|M_r|}{2}, t_2 = \sum_{s=1}^{l'} \binom{|M'_s|}{2}$ and $t_3 = \frac{2t_1t_2}{p(p-1)}$ ([**131**]). Due to the adjustment, the adjRand can also take negative values and cannot be interpreted as a proportion.

The Jaccard Index is related to the Rand Index but disregards the pairs of elements that are separated in both clusterings. It is given by

$$\text{Jaccard}(M, M') := \frac{n_{11}}{n_{11} + n_{10} + n_{01}}.$$

Thereby, this similarity measure focuses on apparent pairings rather than separation. It favors the specificity of pair detection over the sensitivity of clustering methods when comparing random draws to evaluate performance of methods.

1.2. Modulewise preservation statistics. To measure preservation of correlation-based network structures, we define several statistics as summarized in [**32**].

To be able to compare preservation in Netboost and WGCNA networks, we define the mean adjacency, cluster coefficient and maximum adjacency ratio of a module m based on A_{WGCNA} (equation (1), Chapter 2: 1). The mean adjacency is a measure of similarity within a module and is given by

$$\text{meanAdj}(m) := \overline{a_{ij|i,j\in m}}. \tag{8}$$

The cluster coefficient of a node i was introduced for general networks in [**133**] and extended for weighted networks in [**21**]. It is given by

$$\text{clusterCoef}(i) := \frac{\sum_{j\neq i}\sum_{m\neq i,j} a_{ij}a_{jm}a_{mi}}{\left(\sum_{j\neq i} a_{ij}\right)^2 - \sum_{j\neq i} a_{ij}^2}.$$

For modules we again use the mean and define

$$\text{meanClCoef}(m) := \overline{\text{clusterCoef}(i)_{i\in m}}. \tag{9}$$

meanClCoef(m) is a measure of mutual relation of neighbors of a node.

Finally, the maximum adjacency ratio is given by

$$\text{MAR}(i) := \frac{\sum_{j\neq i} a_{ij}^2}{\sum_{j\neq i} a_{ij}}$$

and for modules

$$\text{meanMAR}(m) := \overline{\text{MAR}(i)_{i \in m}}. \tag{10}$$

The MAR is a measure of distinction. A high MAR indicates that a node has a relatively bimodal distribution of adjacency, so it has some highly correlated neighbors and all other nodes are by comparison uncorrelated. The clusterCoef and the MAR of a node are only defined if the respective denominators are non-zero.

All measures introduces above can be calculated for the original data X as well as different data X', e.g. a subsample or replication data. To denote calculation on separate data, we write $\text{meanAdj}(m, X'), \text{meanClCoef}(m, X')$ and $\text{meanMAR}(m, X')$.

2. Network preservation in applications

2.1. Sampling uncertainty in AML methylation and gene expression data. We applied a resampling-based approach to the AML data from TCGA analyzed in Chapter 3: 2. To further comprehend the differences in the clusterings, we took 100 random subsets of 100 patients and compared the resulting Netboost and WGCNA clusterings using pair-wise adjusted Rand Indices and Jaccard Indices. Additionally, we calculated k-means 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. Both indices are less than or equal to 1 and exactly 1 for identical clusterings. As seen in Figure 23, both random clusterings consistently had pair-wise indices very close to 0 and both k-means clusterings were outperformed by WGCNA and Netboost with respect to both metrics. With respect to the adjusted Rand Index, the median for the Netboost clustering was below the median for the WGCNA clustering while the order of minima was vice versa. This implicates that while the median similarity was lower for Netboost, there were more outlying clusterings for WGCNA. When comparing the Jaccard Indices, Netboost outperformed WGCNA and showed a higher similarity for all pair-wise comparisons.

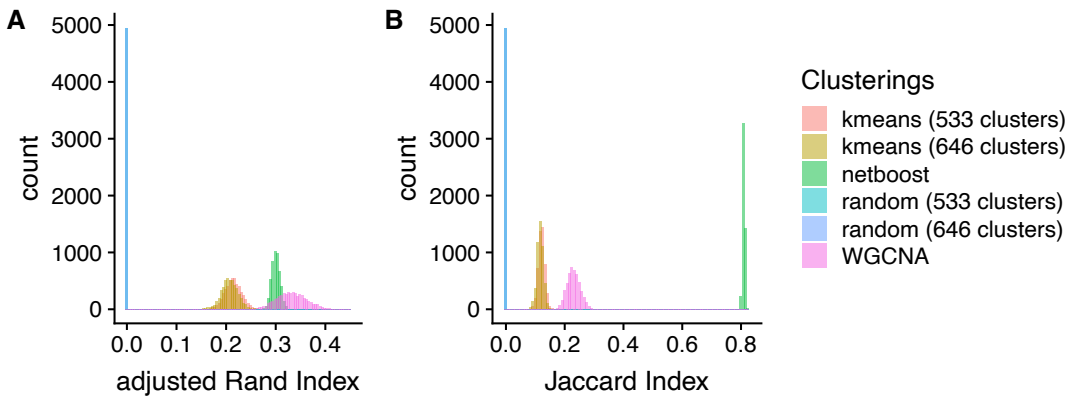

Figure 23: **Clustering indices of the TCGA AML data.** Histogram of pair-wise cluster indices of 100 random subsets of 100 samples applying Netboost, WGCNA, k-means with k clusters and random selection of k labels, where k was set to the median number of modules in the Netboost runs (646) and WGCNA runs (533). A) shows the adjusted Rand Index and B) the Jaccard Index.

2.2. Sampling uncertainty in BRCA, KIRC and OV TCGA data. The other three TCGA datasets illustrate a variety of networks. For each of the datasets we sampled 100 times 63.2% of the patients (analogous to Chapter 3: 1.2) and applied Netboost, WGCNA and k-means with k set to the median number of clusters for Netboost and WGCNA. In addition, we created random clusterings as a reference. Random clusters were equally sized and split across the median number of Netboost and WGCNA clusters respectively. Finally, pair-wise adjusted Rand Indices and Jaccard Indices were computed.

In the BRCA setting, none of the survival analyses improved the analysis excluding omics data altogether (Chapter 3: 3). Regarding the reproducibility of clusterings, when incorporating node separation via the adjusted Rand Index Netboost performed unstably, while WGCNA showed a wider range of similarities and the performance of k-means was in between the two. When focusing on linkage via the Jaccard Index, WGCNA and k-means performed slightly worse and Netboost

improved. As in the AML application, Netboost had the largest minimum similarity while WGCNA performed best with respect to the median (Figure 24 A).

The survival analyses for the KIRC setting proved informative and achieved improvements by application of Netboost were comparable to the AML application (Chapter 3: 3). When comparing adjusted Rand indices, WGCNA and k-means performed similarly while results from Netboost performed nearly comparable but also included some outliers.

When we focused on linkage and compared the Jaccard indices, Netboost outperformed the other algorithms and only minor differences between WGCNA k-means (Figure 24 B).

For OV, adjusted Rand and Jaccard indices showed a similar behavior. WGCNA exhibited a visible double peak, indicating two separate networks being identified dependent on subsampling selection, Netboost was slightly more stable and k-means resulted in nearly identical clusterings independent of the respective sample (Figure 24 C). This is due to the k-means algorithm always generating one very large module consisting of most of the features, which in turn was then nearly identical every time.

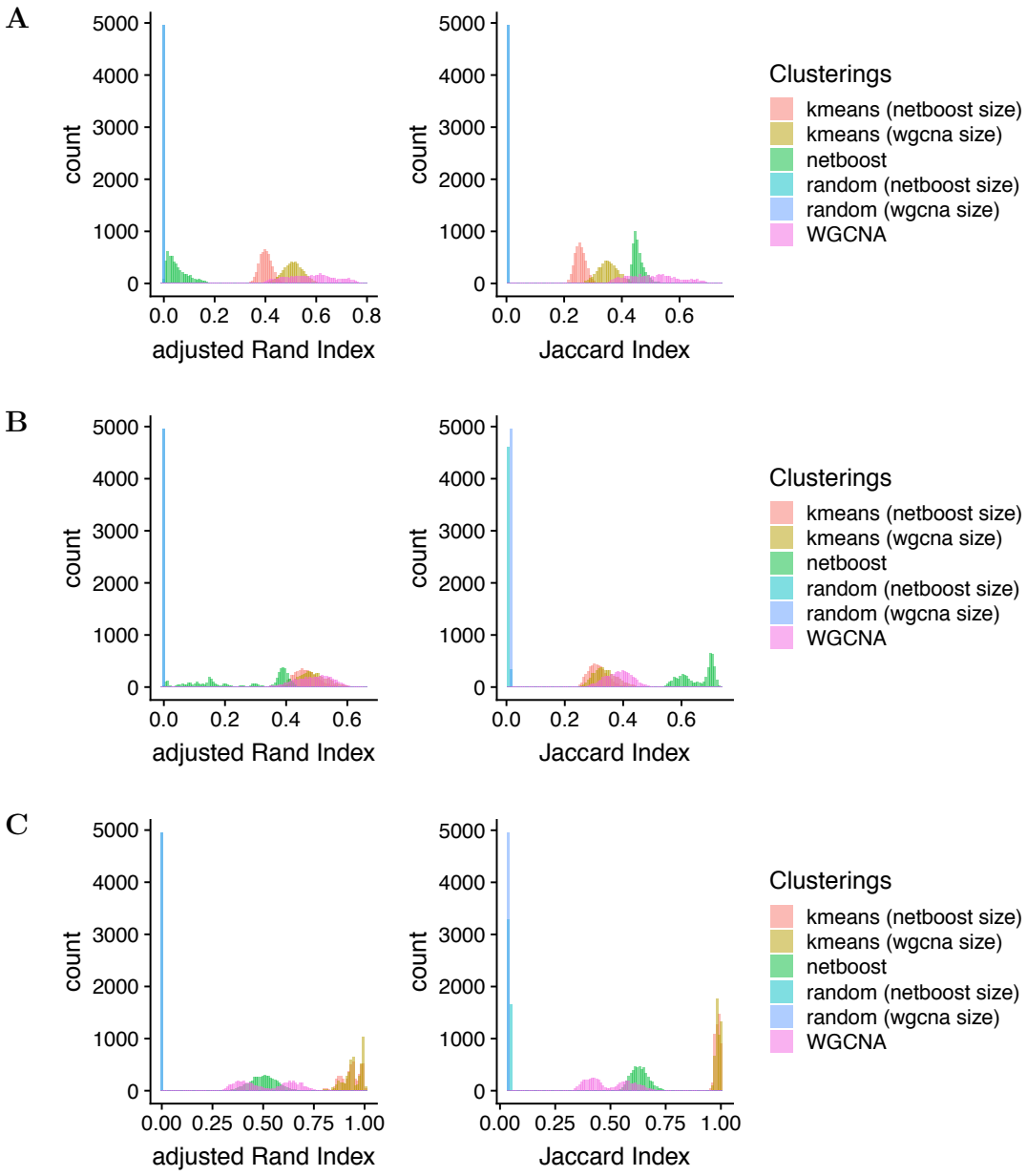

Figure 24: **Clustering indices of TCGA BRCA, KIRC and OV data.** Histograms of cluster indices of 100 clusterings on random .632 subsets applying Netboost, WGCNA, k-means with k clusters and random selection of k labels, where k was set to the median number of modules in Netboost and WGCNA runs. A) BRCA data with subset size of 489, a median of 100 Netboost modules and a median of 46

WGCNA modules. B) KIRC data with subset size of 199, a median of 46 Netboost modules and a median of 34 WGCNA modules. C) OV data with subset size of 293, a median of 12 Netboost modules and a median of 13 WGCNA modules.

2.3. Sampling uncertainty in CKD metabolome data. We examined a superset of the CKD metabolome data analyzed in Chapter 4. This dataset only became available at a later stage, when the analyses presented in Chapter 4 had already been completed. For the resampling-based preservation statistics presented in this Chapter we chose to use the full dataset of 5,088 patients and 1,487 metabolites. We applied the same design as in the TCGA datasets; sampling of 63.2% of the patients was followed by application of clustering algorithms and calculation of pair-wise cluster indices. Again, Netboost and WGCNA showed similar indices and both outperformed k-means.

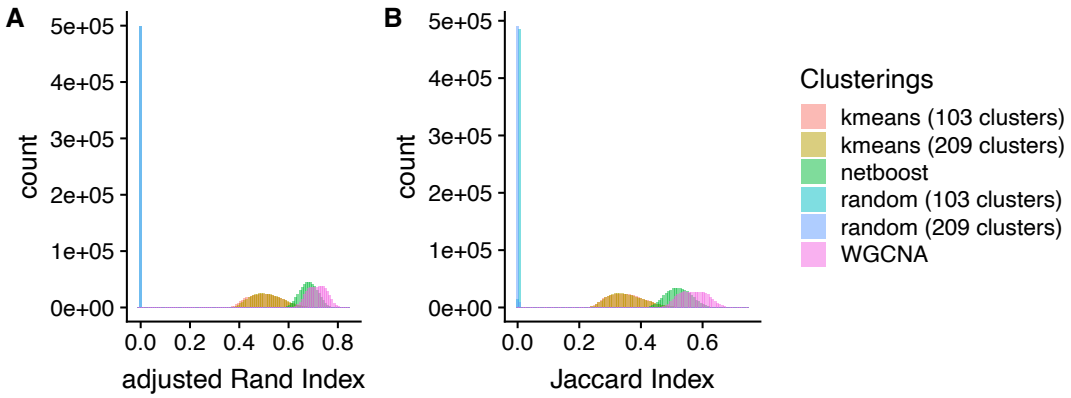

Figure 25: **Clustering indices of the GCKD metabolomics data.** Histogram of cluster indices of 1000 clusterings on random subsets of 3215 samples applying Netboost, WGCNA, k-means with k clusters and random selection of k labels, where k was set to the median number of modules in the Netboost runs (209) and WGCNA runs (103). A) shows the adjusted Rand Index and B) the Jaccard Index.

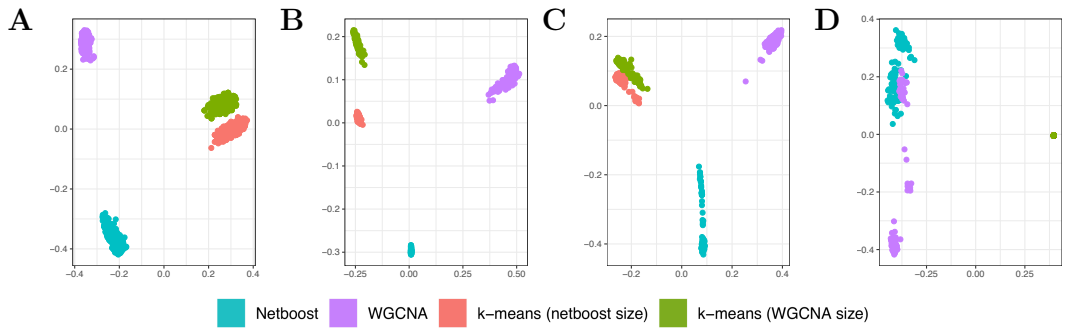

Figure 26: **PCA scatterplot of inverted Jaccard Indices.** First (X-axis) and second PC (Y-axis) of pair-wise inverted Jaccard Indices for CKD (A), BRCA (B), KIRC (C) and OV (D) datasets.

2.4. Method uncertainty. Next, we relate above estimated sampling uncertainties to the respective method uncertainties. In the CKD, BRCA, KIRC and OV datasets we additionally computed all pair-wise inter-method Jaccard Indices and performed a PCA on inverted indices (1-x) for each. Figure 26 displays the first two PCs and shows that method uncertainty clearly exceeds sampling uncertainty for CKD, BRCA and KIRC. While k-means analyses with differing k were closely related, the differences between Netboost, WGCNA and k-means dominated the first two PCs. For OV we can now see from where the WGCNA double peak in Figure 24 originates. For some samples, WGCNA and Netboost converge to similar networks, thus displaying their methodological relatedness. Moreover, k-means stability in this scenario becomes apparent once more; with all points overlapping.

3. Module preservation in applications

While above analyses illustrated the stability of the whole network, it might also be of interest how stable specific parts of the network, namely modules, are. We evaluated this by a combination of a discovery-replication-split and a permutation-based standard score. Specifically, for each dataset we drew 63.2% of observations without replacement and applied Netboost, WGCNA and k-means with k set to the median number of clusters for Netboost and WGCNA to identify the studied

modules. In the remaining observations, we computed propVar(m), meanAdj(m), meanClCoef(m) and meanMAR(m) for each module (equation (4), (8), (9), (10)). Then we computed the same statistics in the same datasets with permuted module labels as described in [**32**]. After repeating this procedure 100 times, we subtracted the mean of the permuted statistics from the observed and divided it by the standard deviation of the permuted statistics to determine standard scores.

First, we applied this methodology to the full GCKD metabolome data of 5088 patients. As k-means results were very similar for $k = 103$ and $k = 209$, we displayed only one of them. From here, on we only report on k-means with k set to the medium number of Netboost clusters as there were only minor inter-k-means-differences for different k.

All three methods identify highly preserved structures with some standard scores, assumed to follow a standard normal distribution, being greater than 150.

Additionally, this analysis serves to identify modules that excel for some statistics while scoring lower on others, i.e., showing desired properties with respect to certain qualities only. For example, the largest k-means module has standard scores greater ten for meanClCoef and meanMAR (Figure 28). Assuming a standard normal distribution, this corresponds to a p-value below 7.7e-24 and highlights that neighbors of variables are in turn closely connected (meanClCoef) and that variables in the module can be well differentiated from variables outside the module (meanMAR). At the same time, as meanAdj is close to zero, these variables do not exhibit a high correlation when compared to a random grouping of variables of the same size (Figure 27).

When comparing clustering algorithms, Netboost and WGCNA performed similarly and achieved higher median and maximum statistics across modules when compared with k-means. For meanAdj, k-means realizes the highest standard score. However, this module constitutes of only two highly correlated variables, as opposed to the Netboost module of size >50 achieving almost the same score.

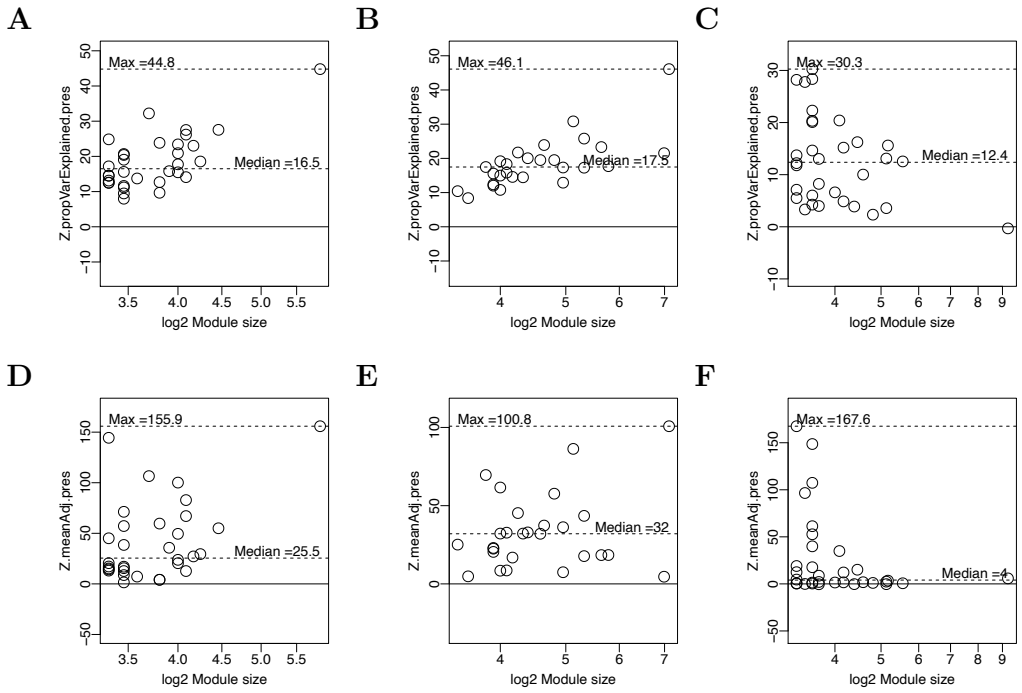

Figure 27: **GCKD preservation statistics: explained variance and adjacency.** The top row shows propVar(m) for GCKD metabolites for Netboost (A), WGCNA (B) and k-means with $k = 209$ (C). The lower row displays meanAdj(m) for GCKD metabolites for Netboost (D), WGCNA (E) and k-means with $k = 209$ (F). Dashed lines indicate maximum and median statistics across modules.

Similar graphics for TCGA BRCA, TCGA KIRC and TCGA OV are given in Supplementary Figures S6-S11. To summarize these additional analyses, we give the maximum and median statistics in Table 2. When equally weighing the maximum across modules and median across modules and equally weighing the four types of statistics, Netboost performs best for three out of the four datasets. Netboost achieved four, five and five times the highest statistic out of eight for GCKD, TCGA KIRC and TCGA OV, respectively.

For TCGA KIRC, we noted a strong difference between the largest module und all other modules for all algorithms. These largest modules consistently scored

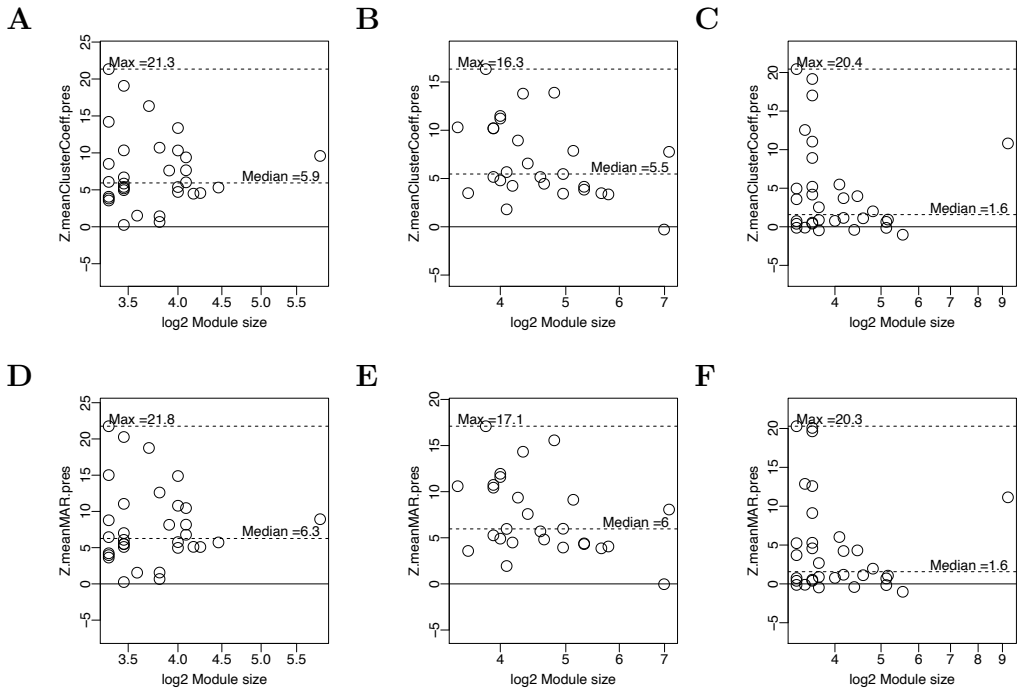

Figure 28: **GCKD preservation statistics: cluster coefficient and maximum adjacency ratio.** The top row shows meanClCoef(m) for GCKD metabolites for Netboost (A), WGCNA (B) and k-means with $k = 209$ (C). The lower row displays meanMAR(m) for GCKD metabolites for Netboost (D), WGCNA (E) and k-means with $k = 209$ (F). Dashed lines indicate maximum and median statistics across modules.

magnitude higher than the second best modules with respect to any statistic (Supplementary Figures S8-S9) and achieved exceptionally high standard scores of up to 620.0 (Netboost meanAdj). When considering only this apparently most essential network structure, Netboost outperforms WGCNA and k-means in its detection with respect to all four module-wise statistics.

Only for TCGA BRCA Netboost did not extract structures with high scores for the four statistics (Table 2), while we did see above that it extracted reproducible structures according to the Jaccard Index (Figure 24). For this dataset, WGCNA performed particularly well.

Table 2: **Preservation statistics overview.** Maximum and median preservation statistics for Netboost, WGCNA and k-means in the four studied datasets. Best scoring algorithm per setting and statistic is highlighted. Netboost is abbreviated by NB.

Disease:	CKD			BRCA			KIRC			OV		
Study:	GCKD			TCGA			TCGA			TCGA		
Training size:	3,215			489			199			293		
Test size:	1,873			285			116			171		
Variables:	1,487			20,000			20,000			1,422		
Variable type:	metabolites			methylation			methylation			miRNA		
Algorithm:	NB	WGCNA	k-means	NB	WGCNA	k-means	NB	WGCNA	k-means	NB	WGCNA	k-means
Max. propVar:	44.8	46.1	30.3	5.0	68.5	61.1	209.3	192.0	140.7	40.4	29.2	14.1
Med. propVar:	16.5	17.5	12.4	0.9	13.2	18.0	10.8	16.4	13.3	12.0	10.7	3.8
Max. meanAdj:	155.9	100.8	167.6	6.8	322.6	608.0	620.0	413.4	601.8	228.3	101.7	124.5
Med. meanAdj:	25.5	32.0	4.0	0.3	121.0	17.1	4.0	2.7	1.9	16.4	19.5	5.1
Max. meanClCoef:	21.3	16.3	20.4	2.6	86.2	69.3	121.6	100.2	95.6	17.9	21.4	62.2
Med. meanClCoef:	5.9	5.5	1.6	-0.2	33.5	7.2	1.1	0.2	3.6	5.6	4.9	3.5
Max. meanMAR:	21.8	17.1	20.3	2.5	85.7	64.7	119.3	83.1	103.0	18.5	22.9	13.4
Med. meanMAR:	6.3	6.0	1.6	-0.2	32.4	8.7	1.2	0.7	2.8	6.2	5.5	3.5
# highest statistics:	4	3	1	0	6	2	5	1	2	5	2	1

What is new in Chapter 6:

- Introduction of a framework for evaluation of sampling uncertainty and method uncertainty.
- Adaptation of module-wise preservation statistics from [**32**].
- Verification of low sampling uncertainty and high module preservation for Netboost in a multitude of omics settings.
- Affirmation of the data dependency of detected network structures and method stability.

CHAPTER 7

Robust extensions of the Netboost concept

One inherent feature of Netboost is the assumption of linear dependencies between variables. At several steps this becomes apparent. When constructing the filter, the boosting algorithm fits linear regression models, the subsequent similarity measure is based on the Pearson correlation coefficient and the final principal components extract linear components. However, for many datatypes linear dependencies are not a close approximation of reality. To achieve some generalizability and explore potentially unknown complex dependencies, more robust methods are required.

1. Simulation setting with non-linear dependencies

To illustrate this, we reintroduce and extend the simulated network from Chapter 1. We simulate 100 samples from 400 i.i.d. standardnormal variables and two modules with 100 multivariate normal variables each. We add a third modules such that the off-diagonal covariance entries for the three modules are 0.8, 0.6 and 0.4, respectively. To introduce non-linear dependencies, we transform each variable by $f(x) := x^7$ according to a Bernoulli trial with probability 0.5, i.e., the transformation is applied with a 0.5 probability. We explored random assignment from larger sets of monotone positive transformations with polynomials of different exponents and additionally incorporating log- and root-based functions, but the larger variety of transformations allowed the algorithm to bridge between them and reduced their impact (data not shown). The presented design, which includes only one transformation, led to the strongest observed disturbance of the algorithm.

2. Spearman- and Kendall-based extensions

To address this additional challenge, we replace the underlying Pearson correlation coefficient with Spearman's rank correlation coefficient and with Kendall's rank correlation coefficient. Both rank-based coefficients are invariant under monotone positive transformations and can detect general monotone relationships between variables ([**134**]).

For X and Y random variables, Spearman's rank correlation coefficient is given by Pearsons correlation coefficient between the rank variables,

$$\operatorname{corr}(\operatorname{rank}(X), \operatorname{rank}(Y)) = \frac{\operatorname{cov}(\operatorname{rank}(X), \operatorname{rank}(Y))}{\sigma_{\operatorname{rank}(X)} \sigma_{\operatorname{rank}(Y)}}.$$

Kendall's rank correlation coefficient is given by the concordant minus the discordant pairs divided by the total number of pairs of samples,

$$\frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}.$$

As shown in Figure 29, we applied Netboost in all three constellations to the original and the transformed data. Independent of the added transformations, the failed detection of the third module emphasizes a general requirement of the methodology. For a fixed sample size, a minimum degree of correlation is required to separate modules. To further explore this, we drew different sample sizes from the 400 i.i.d. standard normal variables and plotted the pair-wise sample Pearson correlation coefficients in Figure 30. As the variables were i.i.d., the population correlation coefficients are zero and all observed sample correlation coefficients not equal to zero can be classified as noise. Only if this noise is sufficiently lower than the intra-module correlation, we can reliably detect the network structure.

The dependence on linear relationships outlined above becomes apparent when Netboost based on the Pearsons correlation coefficient is applied to the partially transformed data. The algorithm incorrectly declares independent variables to form modules and fails to completely assemble existing modules, i.e. produces many false positive modules as well as false negative ones.

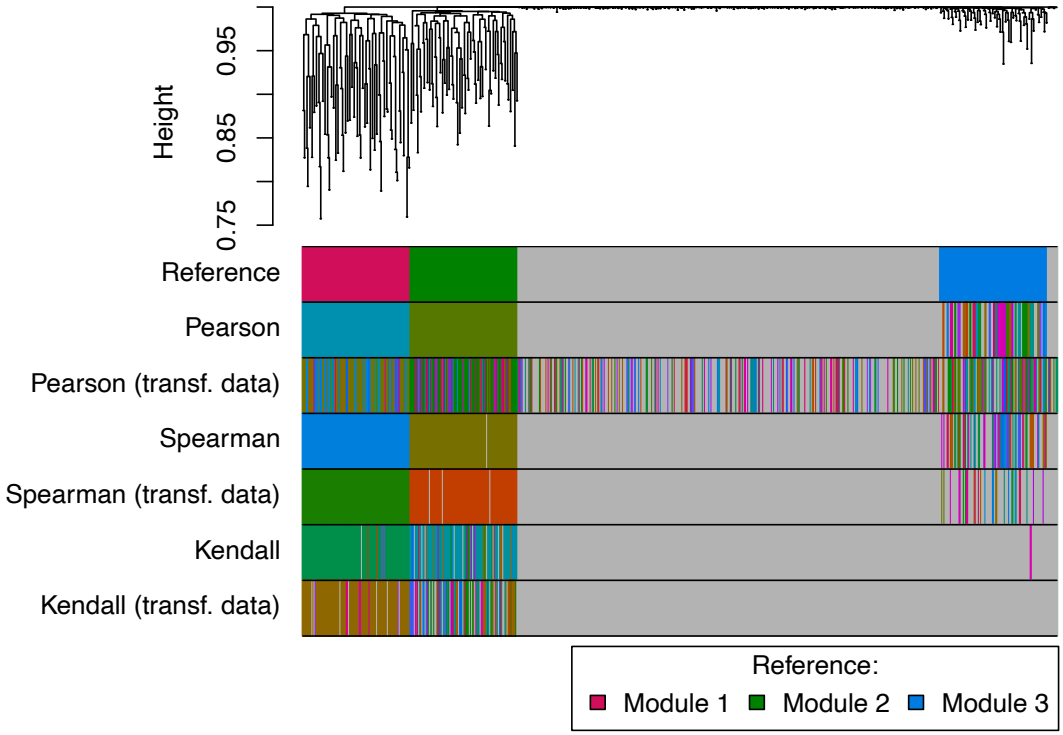

Figure 29: **Dendrogram of extended simulated data.** Dendrogram based on Pearson correlation coefficients of the 700 untransformed variables. The color bands below the graph depict the separation into modules with grey reflecting background variables. Row names indicate the underlying correlation coefficient and if the procedure was applied on the transformed data.

While we only addressed one of the three dependencies on linearity, once we use Spearman's correlation coefficient or Kendall's correlation coefficient the algorithm is far more robust under the transformations. However, especially in the case of Kendall's correlation coefficient this comes with the cost of less power to detect the three modules already on the untransformed data.

To address the second dependency on linearity, we introduce alternative filters $\mathcal{F}_s$ and $\mathcal{F}_k$. These include all edges which are nominally significant (p-value < 0.05) for distribution-free tests of independence based on Spearman's rank correlation coefficient ($\mathcal{F}_s$) and Kendall's rank correlation coefficient ($\mathcal{F}_k$) ([**135**]). To illustrate

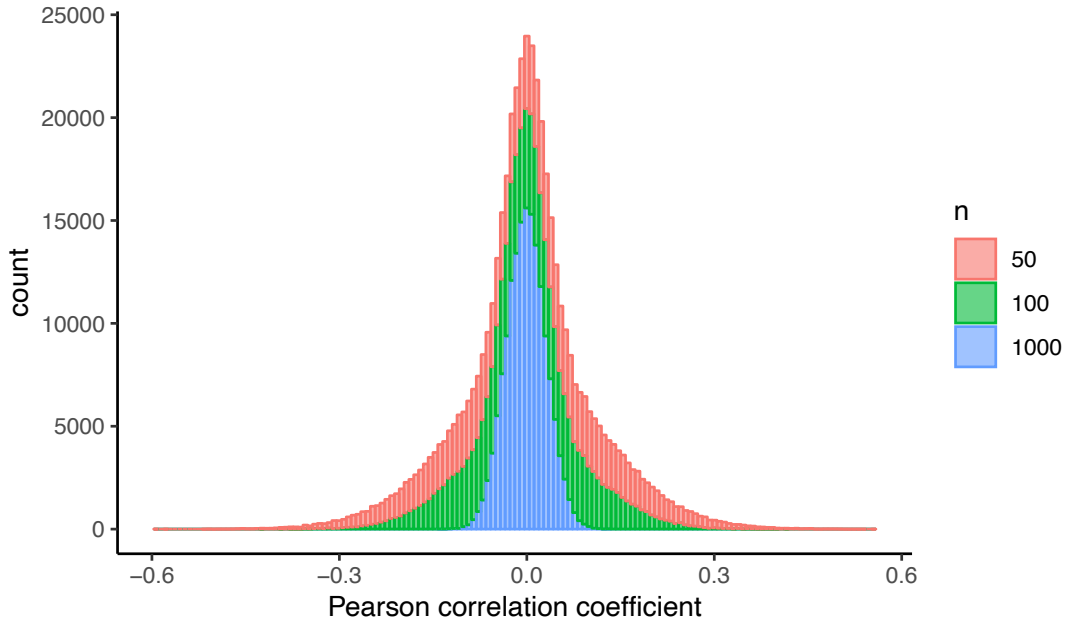

Figure 30: **Sample Pearson correlation coefficients.** Histogram of the sample Pearson correlation coefficients of 400 i.i.d. standardnormal variables based on 50, 100 or 1000 samples respectively.

their impact, we apply them in combination with their corresponding correlation coefficients and display the resulting color codes in Figure 31.

These variations of Netboost are fully rank-based and thereby invariant under positive monotone transformations up to the point where base modules are detected (Chapter 2: 2.2) and principal components are computed. As we use the absolute correlation coefficient in the computation of the similarity measure (Chapter 2: 2.1.2), these adaptations are more generally invariant under monotone transformations. Consequently, the detected modules on the original and on the transformed data are identical. Only in the final step of merging modules, with correlated PCs, the transformation impacts calculations and leads to differing networks.

To address the final dependency on linearity, we integrate robust principal component analysis (rPCA), which identifies the components of greatest variation based on a transformation of the Spearman instead of the Pearson correlation matrix

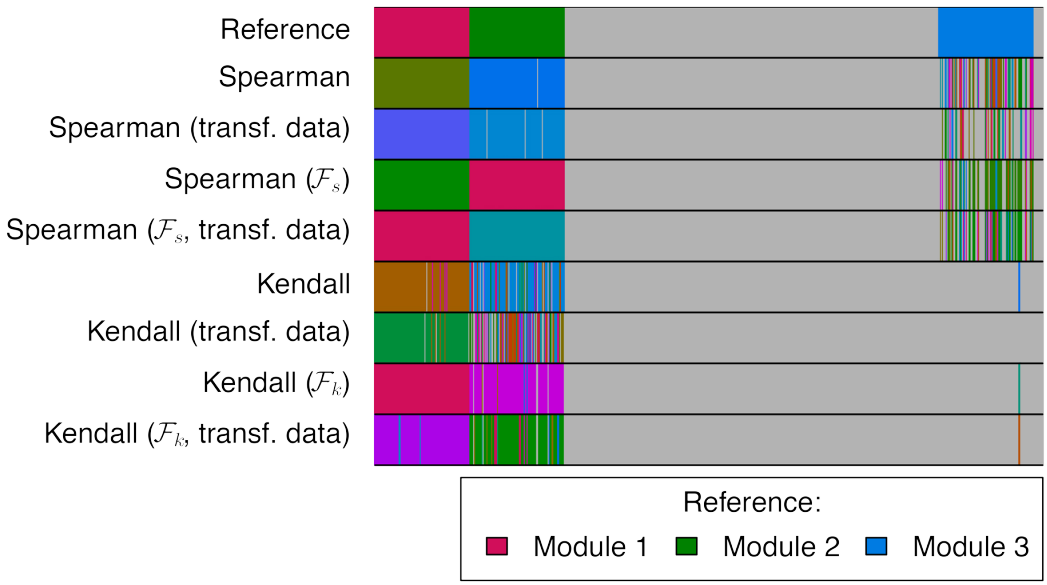

Figure 31: **Network structure under robust filters.** Color bands of the 700 variables indicating the separation into modules by robust extensions of Netboost with grey reflecting background variables. Row names indicate the underlying correlation coefficient and if the procedure used an alternate filter and if it was applied on the transformed data.

([**136**]), with Netboost. In combination, the Spearman-based similarities, the Spearman-based filtering and the Spearman-based PCA result in an algorithm which is completely invariant under monotone transformations and performs very similar to the original Pearson-based design in our simulation (Figure 32). The only apparent cost in this specific application is a slightly higher computational demand as $|\mathcal{F}| = 3189 (\sim 1.3\%)$ and $|\mathcal{F}_s| = 26080 (\sim 10.7\%)$.

In conclusion, both the numerical experiments on the synthetic data and the theoretical foundation of the robust approaches look encouraging and we will investigate their theoretical properties and their pratical performance on real data in future studies.

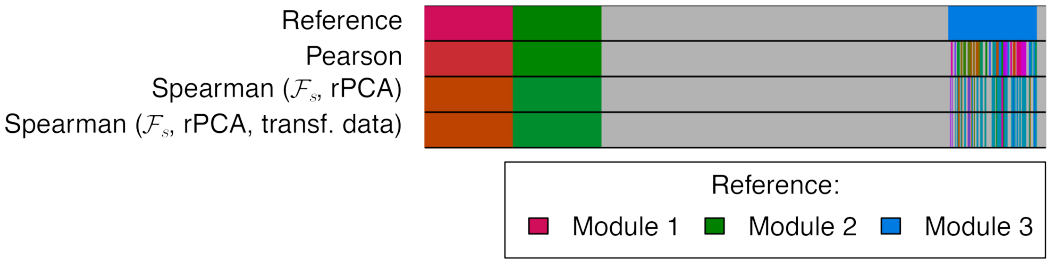

Figure 32: **Network structure in a fully robust design.** Color bands of the 700 variables indicating the separation into modules by robust extensions of Netboost with grey reflecting background variables. Row names indicate the underlying correlation coefficient, filter and PCA method and if it was applied on the transformed data.

What is new in Chapter 7:

- Introduction of extensions of Netboost replacing all three dependencies on linearity with robust alternatives.
- Exploration of their performance in a simulated setting.

CHAPTER 8

Discussion

With Netboost, we introduce an efficient dimension reduction technique based on a combination of unweighted and weighted networks. To accomplish this, we integrated boosting- and rank-based filters with the TOM, sparse hierarchical clustering and the dynamic tree cut procedure. In consequence, Netboost identifies highly correlated variables and aggregates them. Beside the inherent information in this identified network we can subsequently use the aggregated data, which is of lower dimensionality, for the analysis of primary interest, such as time-to-event or genetic association analyses. Additionally, we developed a corresponding Bioconductor `R` package to make the approach easily accesible to a wider community of researchers. With the resulting modeling method we were able to show several theoretical advantages and practical advantages on the computational side as well as demonstrate successful applications.

In this approach, we utilize the modular organization of many organisms. We reflect the often interacting systems with functional units found for example in epigenetic regulation of gene expression and the metabolism of endogenous metabolites and xenobiotics, by designing a hierarchical correlation-based network, which allows for organization on multiple levels. Efficiently recovering these structures in a data-driven manner and aggregating representations of lower dimensionality along them enabled us to find strong associations of these modules and patient relevant outcomes. Furthermore, the modular structure often supplemented existing biological knowledge and permitted comprehensive interpretation of the results.

1. Summary of the presented results

Our proposed two-stage design, Netboost (Chapter 2), is extremely versatile as the different applications presented in this dissertation (Chapter 3, Chapter 4 and Chapter 5) illustrate. While integration of the primary analysis, variable selection and dimension reduction has the potential for further improvements in prediction errors and power, this inherently depends on said primary analysis strategy. Furthermore, an integration needs to be carried out carefully in order not to void our potential for biological interpretation of the network structures themselves. This interpretability, that is achieved by the unsupervised two-stage design, proved vital for the scientific advance when identifying the chromatin-modifying module related to AML survival (Chapter 3: 2), deorphanizing the metabolite X-13689 as the glucuronide of alpha-CMBHC and associating acisoga and *PAOX* (Chapter 4: 2.1).

Beyond the illustrated applications to DNAm, RNA array, RNA-seq, miRNA and metabolome data from *in vivo* human and murine samples, there are many more data types which are believed to follow scale-free topologies or reflect correlation-based networks in general and are thereby well-suited for Netboost. An intuitive reasoning behind that is that the evolutionary pressure which is driving many of these networks is some form of random non-targeted attack ([**21**]). Consequently, organisms must be particularly stable under random attacks. In relation to this, a broad range of scale-free networks, unlike many other networks, have been shown in [**137**] to have a percolation threshold of zero, meaning that removing randomly any fraction of nodes will not destroy the network. For example, in [**35**] proteome data was analyzed by a combination of WGCNA with subsequent GWAS of MEs. Analogously to Chapter 4, the authors identify several associations of MEs with SNPs but in contrast to Chapter 4, they lack the comparison to the single variable associations as a reference to evaluate benefits of the added dimension reduction.

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 sparse UPGMA ([**53**]). In the

AML application (Chapter 3: 2), this resulted in a 559-fold and in the HD application (Chapter 5: 2) in a 264-fold reduction of variables for the primary analyses. This not only reduced the burden on subsequent variable selection procedures but simultaneously decreased the risk of overfitting present in high-dimensional analyses ([**73**]). At the same time, the proportion of variance explained is high such that most information is kept in the low-dimensional representation (Figure 7, Figure 20). In the mGWAS application (Chapter 4), this same dimensionality reduction enabled the detection of biological determinants of metabolite modules and aided the deorphanization of unknowns. GWAS of MEs represent an efficient way to perform hypothesis-generating genome-wide screens on combinatorial information from metabolites. The p-gain values in Supplementary Table S2 display that the p-values obtained from GWAS of MEs are up to 1.5e343 times lower than the lowest p-value for any of the metabolites in the module. Moreover, we only need to account for 212 modules instead of all 1,172 single metabolites when adjusting the threshold for multiple testing. When comparing this to testing all 686,206 pair-wise metabolite ratios, which we explored as an alternate approach to capture combinatorial information in [**2**], this discrepancy increases even more.

Apart from the reduction in the multiple testing burden, the maximal computational load is also reduced. As long as the maximum number of steps in the boosting filter is set below p, the filter is necessarily strictly smaller than the whole network. Even if the parameter was set to such an unreasonably large value, networks leading to selection of a substantial proportion of edges would indicate measurement of very few independent variables. Thus, it would contradict the initial assumption on which the Netboost methodology is build. Namely, that we are in a high-dimensional setting and though there is a certain structure to the data, there are also independent processes being measured within this data. In practice, the filter is smaller than the whole network by orders of magnitudes. Accordingly, we reduce the memory load and computational burden massively.

Often methodological advances lack a replication analysis when they are initially proposed. This is especially critical in the high-dimensional setting where overfitting is imminent ([**73**]). In Chapter 4, we were able to plan a discovery/replication design from the beginning and all 46 reported ME-associations were replicated. In the public domain TCGA AML application, we lacked a sufficient number of observations to split the dataset into two. However, we were able to establish a collaboration with the AMLSG study group and in Chapter 3: 2.3, the association with the chromatin-modifying enzymes could be replicated in an independent DNAm data set from the phase II AMLSG 12-09 clinical trial ([**25**]) despite no available gene expression measurements. The AMLSG 12-09 study tested the hypothesis that 5-azacytidine might reduce failure rates of intensive induction therapy particularly in AML patients with unfavorable genetic variables. Interestingly, validation was successful in this independent dataset, not only despite the missing gene expression measures but also despite distinct distributions of genetic aberrations in patients within the AMLSG 12-09 trial and the TCGA dataset, pointing to a more general mechanism applicable to a wide variety of AML patients.

In the shown applications, we prefer specificity over sensitivity with respect to the clusterings. While it might be acceptable to miss an additional variable being part of a module, we want to be sure about the selected variables. Consistent with this, we regard the Jaccard Index as more important to our applications (Chapter 6), as most variables are assumed to be independent of each other. A similar strategy for assessment of sampling and method uncertainty was recently presented in [**130**]. In the context of variable selection, the authors suggested to randomly split the given data set in two independent data set halves and then apply the proposed analysis strategies on each of these. For the subsequent comparison between methods and data set halves, they as well used the Jaccard Index.

Even though we do not perform variable selection but network detection in our analysis, we can directly transfer methodology when we abstract the network detection to a variable selection problem with selected variables being network edges. In

general, 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 an identical resampling setting with different methodologies.

In Chapter 6, we used this resampling-based clustering indices to asses the sampling uncertainty of Netboost, WGCNA and k-means. As expected, it depended on the dataset which methodology offered the lowest sampling uncertainty and generally method uncertainty was more pronounced than sampling uncertainty ([**138**]). In contrast to the procedure presented in [**130**], we did not evaluate the full analysis design but only the general part which is shared by all presented applications. Thus, we estimated uncertainty up to the partitioning into modules and excluded the time-to-event-, GWAS- and random forest-specific parts. To incorporate these aspects, new measures of relatedness between the results would be required, as Rand Index and Jaccard Index do not translate to the more complex ME-survival and ME-SNP associations, which is again related to the MEs not being identical in between resamplings.

Moreover, the module-wise preservation statistics helped to reveal how different detected networks can be. For example, the KIRC dataset was dominated by one large and extraordinary preserved module with a standard score of 620.0 (Chapter 6, Supplementary Figures S8-S9). On the other hand, most networks exhibited more complex structures with many highly preserved modules (Chapter 6, Supplementary Figures S6-S11). Particularly the OV example was interesting as despite the methodological differences Netboost and WGCNA resulted in similar networks (Figure 26) for a subset of samplings.

In Chapter 7, we extended the Netboost concept with robust methodology. Both the Spearman and the Kendall correlation coefficient underlying these extensions are rank based, broaden the concept to a wider range of applications and make the approach robust against outlying measurements. While the original design assumed linear relationships between variables at several steps, the adaptations fully generalize it to monotone relationships. Even though Spearman and Kendall statistics

share many properties, there are some theoretical and practical differences. The Kendall correlation coefficient is theoretically superior as it leads to an unbiased estimator of the difference between the probability of concordance and the probability of discordance in the full population ([**139**]), whereas the Spearman coefficient is not an unbiased estimator of the population correlation ([**140**]). In practice, Spearman's statistic is widely applied; for instance as it corresponds to the Pearson correlation coefficient of the rank matrix and thereby offers an intuitive interpretation. Furthermore, in our simulation it proved more powerful as modules were detected based on smaller sample sizes. Based on these initial observations, we favor the Spearman-based adaptation of Netboost. In [**141**], datasets including non-linear relationships and their impact on network analysis are discussed. While Spearman correlation coefficients are briefly mentioned, only Pearson correlation coefficients are used in the comparison. The authors suggest model fit parameters for polynomials of degree three and model fit parameters for splines as similarity measures which are robust to their simulated quadratic relationships. A similar but more comprehensive future study covering a larger set of non-linear relationships, both in simulated and real data, and a wider range of robust similarity measures would be required to firmly establish a rank-based extension of Netboost.

In addition to the theoretical and practical advantages of Netboost, we offer a convenient and comprehensive implementation of the methodology as a Bioconductor `R` package (Chapter 2: 2.5). As configuration and installation of sparse UPGMA ([**53**]) was demanding, we automated this and integrated installation and execution within our `R` package. All individual steps of the Netboost algorithm are accessible and documented as their own functions and additional functions for plotting and transfer of a network to another dataset are provided. The centerpiece of the package is the one-stop-function which integrates all steps, from filtering over hierarchical clustering to module detection and integration to reporting central features of the network (netboost(...), Supplementary File S1 and Supplementary File S2) into one function.

2. Netboost in the context of literature

WGCNA is an established analysis method for high-dimensional data, especially in the context of biomedical research, with the main paper ([**15**]) being cited more than 3,063 times up to now (Web of knowledge accessed September 1th, 2019. URL: `http://tiny.cc/pq01bz`). Development continues and adaptations ([**142**]) and new fields of application ([**35**]) evolve to date. WGCNA revealed higher order organization of data, allowed for interpretable high-dimensional analysis and uncovered associations with patient-relevant endpoints, e.g. overall survival, in many setting where more traditional analysis strategies failed due to the high dimensionality and the general cross-omics complexity. By extending the WGCNA approach to the Netboost modeling strategy, we were able to demonstrate improved prediction errors for AML and KIRC time-to-event analyses in Chapter 3 and improved random forest mis-classification rates for HD in Chapter 5.

One drawback of the implementation of WGCNA is that due to limitations of indexing capabilities in `R` , applications with $p > \sqrt{2^{31} - 1} \approx 46,340$ first need to be split into parts of smaller size to be processed independently. This is implemented via k-means clustering and later aggregation via correlated MEs ([**15**]). This limitation does not transfer to the implementation of Netboost (Chapter 2: 2.5) and we already applied the full methodology, without need for modification, in a setting with more than 400,000 variables (Chapter 3: 2).

Focusing on the core components of the network allows us to be more selective and thereby more specific in edge detection. As demonstrated in Chapter 6, this results in more finely grained networks with smaller and more modules than WGCNA. Additionally, these are more stable under subsampling.

In all applications we refrained from extensive parameter tuning and usually applied algorithms in their standard settings to allow for impartial comparisons between approaches. In Chapter 5, the application of WGCNA superimposed the disease classifying signals and the direct random forest application on the high-dimensional dataset achieved better misclassification rates. As we did not adjust

the WGCNA parameter settings, we cannot eliminate the possibility that parameters simply did not fit the application context. Netboost kept a more compartmentalized and detailed network without the need for parameter tuning due to the applied filtering step and was able to improve classification in comparison to both - WGCNA and the direct classification based on the full dataset.

An even more widely applied dimension reduction technique is PCA ([**12**]). When PCA is applied to the full dataset rather than to the highly correlated module members, the extracted aggregate measures represent something decidedly different. Specifically, single Netboost MEs do not need to dominate the full dataset to be extracted and can be relevant to only a smaller part of the network, i.e., to a smaller part of the underlying biology, which in turn might be related to the studied research question. In that manner, Netboost is a local dimensionality reduction technique, as opposed to ordinary PCA being a global dimensionality reduction method (Chapter 1). Furthermore, while global PCs are relevant to many studies, they lack the biologically informative network that comes along with Netboost. Thus, the two approaches present themself as complementary strategies.

In [**9**], another approach for data with correlated variables is proposed. The first step is to cluster the variables, 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. 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 variables - voiding their function as representatives, which interfers with the assignment of biological meaning of identified biomarkers in the context of the network. Keeping outcome and network detection separate, as done with Netboost, allows for an

unbiased interpretation of any potential connections between subsequently selected modules and the outcome.

In **[10]**, the supervised algorithm Net-Cox is introduced, which applies network theory to improve survival prediction in a high-dimensional context. In contrast to our combination of Netboost and Coxboost, they employ the estimated gene co-expression structure directly to the penalty term of the Cox model. In addition to the advantages and limitations of supervised algorithms discussed above, Net-Cox is thereby inherently designed for survival analysis, whereas Netboost is more flexible in its application.

In **[17]**, two extension to sparse canonical correlation analysis (CCA) ([**16**]) are introduced. First, the authors propose a supervised form of sparse CCA and secondly, they generalize the framework from two to multiple datasets. With this approach, they offer a framework for identification of sparse linear combinations of the multiple sets of variables that are highly correlated with each other and associated with the outcome. While Netboost can also identify cross-omics correlations associated with the outcome, the algorithms presented in **[17]** omit within-datatype connections and are optimized solely for cross-dataset combinations.

In the light of correlation-based networks in general, Netboost defines variable-wise distances based on pair-wise Pearson, Spearman or Kendall correlation coefficients, whereas, e.g., the approach of **[19]** constructs networks based on partial correlations. In the form of gaussian graphical models (GGMs) partial correlations are frequently applied for network construction ([**143, 144**]). In **[20]**, a GGM is 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 variables. This is often done to identify the "true underlying / causal" connections in the network ([**143, 145**]). In contrast, in Netboost we integrate indirect connections even further by the TOM (Chapter 2: 2.1.2) in order to identify interacting subgroups irrespective of whether these interaction are direct or indirect. Hence, the focus lies on modules rather

than on the individual edges, and the incorporation of indirect connections further stabilizes module detection. GGMs have also been extended with supervised module construction ([**11**]), leading to improved predictive performance but similar limitations as the supervised choice of cluster representatives in [**9**].

3. Limitations and future work

We introduced the number of boosting steps as a parameter. This number can be chosen high, as overfitting in the filter estimation would only result in a less stringent filter rather than bias. Nevertheless, a possible extension is a probing-based stopping criterion in the boosting step, e.g. by inserting shadow variables ([**56**]). Here, permutated variants of the original variables would be introduced, which are independent of the other variables. The algorithm would then be stopped once it starts to select these shadow variables ([**146**]). This would automate the choice of boosting steps, while circumventing the often extensive additional computational load of cross-validation. However, it is a non-trivial task to determine a suitable proportion of shadow variables and the specific computational load introduced would need to be explored.

Especially in the context of network analyses, integration of a-priori information is a frequently studied choice. In [**18**], the KeyPathwayMinerWeb is introduced, which allows for detection of differentially regulated pathways in a known network and in the yet unpublished Grand Forest methodology by the same group (`https://grandforest.compbio.sdu.dk/`), random forest methodology is altered by subsetting the space of possible trees to a known network. Such an integration of a-priori knowledge could also be considered to aid network construction in the Netboost approach. While there is a an ever-improving body of research on the assemble of generalizable networks for certain data types, e.g. the interactome ([**147**]), GO ([**110**]) and KEGG ([**111**]), much of this is based on studies of cell lines or model organism like yeast. It remains unclear what can truly be transfered to a human *in vivo* setting or more generally to the studied setting in an organism. For example in our manuscript related to Chapter 4 ([**2**]), one of the aims was to determine whether

associations identified in CKD patients can be generalized to a healthy population. This includes the metabolome structure identified by Netboost and transferability could not have been confirmed if we already assumed it to begin with. Additionally, a generic drawback of incorporating a-priori knowledge is its specificity to data type as opposed to Netboost currently being a broad analysis strategy for general high-dimensional data. However, dependent on the research question, such a-priori knowledge might be informative in order to aid network construction and should be explored in future studies.

Another option to extend Netboost is the inclusion of unclustered variables which are currently being ignored in primary analysis. This implies that isolated singular variables can not achieve a significant impact on the outcome, which is of course not true for all settings. For example, in the primary analysis X_{modules} could be combined with a filtering method on the unclustered variables based on the proportional overlapping score (POS) ([**148**]).

Likewise, the methodology could be extended to signed networks. The sign of correlation coefficients could be integrated as suggested in [**149**] to more explicitly model the often biologically meaningful inversion of directions.

Similarly, we could also modify the module aggregation method. Netboost integrates modules via their leading PCs (Chapter 2: 2.3). However, for purposes such as biomarker identification, a single representative for each module might be an advantage. Another approach to consider for this are hub genes replacing the MEs we applied, as discussed in [**19**] and [**150**]. A hub gene is the most central node with the highest connectivity of the module as opposed to a summary measure, thus allowing for cost-efficient replication and application as a biomarker ([**31**]). MEs, on the other hand, might be superior in mechanistic studies, exploratory studies and the identification of previously unknown biological features.

Another possibility for the aggregation of individual modules to cover non-linear relationships could be autoencoders ([**151, 13**]) to allow for even more flexibility in the extraction of summary measures.

Generally speaking, regarding the dimension reduction of modules, MEs optimize the explained variance with respect to a predefined number of dimensions. In our applications we fixed this dimensionality to one to achieve comparability to WGCNA. The observed proportions of explained variance were particularly high for Netboost modules (Chapter 3: 2, Chapter 5: 2, Chapter 6: 3). However, in case of more complex module substructures where additional PCs might be needed to explain an adequate proportion of variance, we implemented the optional export of the first i PCs for a fixed i (Chapter 2: 2.3). Furthermore, for each module the first j PCs which cumulatively explain at least a certain predefined proportion of variance can also be exported automatically.

Independent of the number of PCs, these aggregate information best if variables have linear relationships. As demonstrated in Chapter 7, the algorithm was disturbed when non-linear relations were introduced. We offer a robust adaptation of Netboost, ultimately transferring the full methodology to a rank-based one. Accordingly, this results in an algorithm invariant under monotone transformations of the variables, allowing for an even wider set of data types for which Netboost is now suited. Recently, a similar but much more reduced approach was suggested on bioRxiv ([**152**]), where hard thresholding was applied on Spearman correlation coefficients in a DNAm setting without further extensions. While the authors also highlight the benefits of a rank-based measure, the filter and subsequent calculation of TOM adds further robustness to our approach and integrates indirect connections between variables.

4. Conclusion

Netboost offers a versatile statistical modeling strategy for high-dimensional data. We introduced boosting-based and rank-based filters, combined these with sparse hierarchical clustering, module aggregation and incorporate these to a full dimension reduction methodology. We then integrated Netboost with various exemplary analysis strategies and provided evidence for its statistical advantages in terms of prediction errors and power and analytical advantages in terms of biological

interpretability. Finally, we investigated its theoretical properties and examined the preservation of detected network structures.

Remarkably, in every analysis setting we encountered an application where Netboost led to a significant improvement in prediction errors or power. This observation holds true across a wide variety of research questions, from time-to-event analyses over classification to genetic associations, and across a wide variety of data types, from DNAm over transcriptomics to metabolomics. Furthermore, we can exclude selection bias as a source of this as all analysis settings and datasets, where we applied Netboost, are reported.

With this perspective we are looking forward to extending and building on the methodology presented in this dissertation in our future work.

Bibliography

[1] P. Schlosser, J. Knaus, M. Schmutz, K. Döhner, C. Plass, L. Bullinger, R. Claus, H. Binder, M. Lübbert, and M. Schumacher. Netboost: Boosting-supported network analysis improves high-dimensional omics prediction in acute myeloid leukemia and huntington's disease. *In revision.*

[2] P. Schlosser, Y. Li, P. Sekula, J. Raffler, F. Grundner-Culemann, M. Pietzner, Y. Cheng, M. Wuttke, I. Steinbrenner, U. T. Schultheiss, F. Kotsis, T. Kacprowski, L. Forer, B. Hausknecht, A. B. Ekici, M. Nauck, U. Völker, GCKD Investigators, G. Walz, P. J. Oefner, F. Kronenberg, R. P. Mohney, M. Köttgen, K. Suhre, K.-U. Eckardt, G. Kastenmüller, and A. Köttgen. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *In revision.*

[3] D. H. Heiland, I. Mader, P. Schlosser, D. Pfeifer, M. S. Carro, T. Lange, R. Schwarzwald, I. Vasilikos, H. Urbach, and A. Weyerbrock. Integrative network-based analysis of magnetic resonance spectroscopy and genome wide expression in glioblastoma multiforme. *Sci Rep*, 6, 2016.

[4] A. Y. Chu, A. Tin, P. Schlosser, Y.-A. Ko, C. Qiu, C. Yao, R. Joehanes, M. E. Grams, L. Liang, C. A. Gluck, C. Liu, J. Coresh, S.-J. Hwang, D. Levy, E. Boerwinkle, J. S. Pankow, Q. Yang, M. Fornage, C. S. Fox, K. Susztak, and A. Köttgen. Epigenome-wide association studies identify DNA methylation associated with kidney function. *Nat Commun*, 8(1), 2017.

[5] N. Blagitko-Dorfs, P. Schlosser, G. Greve, D. Pfeifer, R. Meier, A. Baude, D. Brocks, C. Plass, and M. Lübbert. Combination treatment of acute myeloid leukemia cells with DNMT and HDAC inhibitors: predominant synergistic gene downregulation associated with gene body demethylation. *Leukemia*, Nov 2018.

[6] Y. Li, P. Sekula, M. Wuttke, J. Wahrheit, B. Hausknecht, U. T. Schultheiss, W. Gronwald, P. Schlosser, S. Tucci, A. B. Ekici, U. Spiekerkoetter, F. Kronenberg, K.-U. Eckardt, P. J. Oefner, and A. Köttgen. Genome-wide association studies of metabolites in patients with CKD identify multiple loci and illuminate tubular transport mechanisms. *J Am Soc Nephrol*, 29(5), 2018.

[7] M. R. Nelson, H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, Y. Shen, A. Floratos, P. C. Sham, M. J. Li, J. Wang, L. R. Cardon, J. C. Whittaker, and P. Sanseau. The support of human genetic evidence for approved drug indications. *Nat. Genet.*, 47(8):856–860, Aug 2015.

[8] H. Fang, H. De Wolf, B. Knezevic, K. L. Burnham, J. Osgood, A. Sanniti, A. Lledo Lara, S. Kasela, S. De Cesco, J. K. Wegner, L. Handunnetthi, F. E. McCann, L. Chen, T. Sekine, P. E. Brennan, B. D. Marsden, D. Damerell, C. A. O'Callaghan, C. Bountra, P. Bowness, Y. Sundstrom, L. Milani, L. Berg, H. W. Gohlmann, P. J. Peeters, B. P. Fairfax, M. Sundstrom, J. C. Knight, G. Beckmann, C. Bountra, P. Bowness, N. Burgess-Brown, L. Carpenter, L. Chen, D. Damerell, U. Egner, H. Fang, R. Fujii, T. Howe, P. J. Jakobsson, A. Katopodis, J. C. Knight, B. D. Marsden, J. De Martino, G. Matthias, G. McVean, A. Mueller-Fahrnow, A. Malarstig, C. A. O'Callaghan, N. Ostermann, J. R. Paez-Cortez, P. J. Peeters, F. Prinz, P. Soulard, M. Sundstrom, C. Yabuki, and J. Vlach. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.*, 51(7):1082–1091, Jul 2019.

[9] S. Reid and R. Tibshirani. Sparse regression and marginal testing using cluster prototypes. *Biostatistics*, 17(2):364–376, 2016.

[10] 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 Computational Biology*, 9(3):1–16, 03 2013.

[11] K. T. Do, D. J. N. Rasp, G. Kastenmüller, K. Suhre, and J. Krumsiek. MoDentify: phenotype-driven module identification in metabolomics networks at different resolutions. *Bioinformatics*, 35(3):532–534, Feb 2019.

[12] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[13] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[15] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, page 559, 2008.

[16] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, Jul 2009.

[17] D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, 8, 2009.

[18] M. List, N. Alcaraz, M. Dissing-Hansen, H. J. Ditzel, J. Mollenhauer, and J. Baumbach. KeyPathwayMinerWeb: online multi-omics network enrichment. *Nucleic Acids Research*, 44(W1):W98–W104, 05 2016.

[19] D. Yu, J. Lim, X. Wang, F. Liang, and G. Xiao. Enhanced construction of gene regulatory networks using hub gene information. *BMC Bioinformatics*, 18(1):186, Mar 2017.

[20] S. Lee, F. Liang, L. Cai, and G. Xiao. Integrative analysis of gene networks and their application to lung adenocarcinoma studies. *Cancer Informatics*, 16:1176935117690778, 2017.

[21] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4, 2005.

[22] P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24:719–720, 2008.

[23] H. Binder and M. Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9:14, 2008.

[24] T. J. Ley, C. Miller, L. Ding, B. J. Raphael, A. J. Mungall, A. Robertson, K. Hoadley, T. J. Triche, P. W. Laird, J. D. Baty, L. L. Fulton, R. Fulton, S. E. Heath, J. Kalicki-Veizer, C. Kandoth, J. M. Klco, D. C. Koboldt, K. L. Kanchi, S. Kulkarni, T. L. Lamprecht, D. E. Larson, L. Lin, C. Lu, M. D. McLellan, J. F. McMichael, J. Payton, H. Schmidt, D. H. Spencer, M. H. Tomasson, J. W. Wallis, L. D. Wartman, M. A. Watson, J. Welch, M. C. Wendl, A. Ally, M. Balasundaram, I. Birol, Y. Butterfield, R. Chiu, A. Chu, E. Chuah, H. J. Chun, R. Corbett, N. Dhalla, R. Guin, A. He, C. Hirst, M. Hirst, R. A. Holt, S. Jones, A. Karsan, D. Lee, H. I. Li, M. A. Marra, M. Mayo, R. A. Moore, K. Mungall, J. Parker, E. Pleasance, P. Plettner, J. Schein, D. Stoll, L. Swanson, A. Tam, N. Thiessen, R. Varhol, N. Wye, Y. Zhao, S. Gabriel, G. Getz, C. Sougnez, L. Zou, M. D. Leiserson, F. Vandin, H. T. Wu, F. Applebaum, S. B. Baylin, R. Akbani, B. M. Broom, K. Chen, T. C. Motter, K. Nguyen, J. N. Weinstein, N. Zhang, M. L. Ferguson, C. Adams, A. Black, J. Bowen, J. Gastier-Foster, T. Grossman, T. Lichtenberg, L. Wise, T. Davidsen, J. A. Demchok, K. R. Shaw, M. Sheth, H. J. Sofia, L. Yang, J. R. Downing, G. Eley, S. Alonso, B. Ayala, J. Baboud, M. Backus, S. P. Barletta, D. L. Berton, A. L. Chu, S. Girshik, M. A. Jensen, A. Kahn, P. Kothiyal, M. C. Nicholls, T. D. Pihl, D. A. Pot, R. Raman, R. N. Sanbhadti, E. E. Snyder, D. Srinivasan, J. Walton, Y. Wan, Z. Wang, J. P. Issa, M. Le Beau, M. Carroll, H. Kantarjian, S. Kornblau, M. S. Bootwalla, P. H. Lai, H. Shen, D. J. Van Den Berg, D. J. Weisenberger, D. C. Link, M. J. Walter, B. A. Ozenberger, E. R. Mardis, P. Westervelt, T. A. Graubert, J. F. DiPersio, and R. K. Wilson. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, 368(22):2059–2074, 05 2013.

[25] R. F. Schlenk, D. Weber, W. Herr, G. Wulf, H. R. Salih, H. G. Derigs, A. Kuendgen, M. Ringhoffer, B. Hertenstein, U. M. Martens, M. Griesshammer, H. Bernhard, J. Krauter, M. Girschikofsky, D. Wolf, E. Lange, J. Westermann, E. Koller, S. Kremers, M. Wattad, M. Heuser, F. Thol, G. Gohring, D. Haase, V. Teleanu, V. Gaidzik, A. Benner, K. Döhner, A. Ganser, P. Paschka, and H. Döhner. Randomized phase-II trial evaluating induction therapy with idarubicin and etoposide plus sequential or concurrent azacitidine and maintenance therapy with azacitidine. *Leukemia*, Feb 2019.

[26] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, K. Chang, C. J. Creighton, C. Davis, L. Donehower, J. Drummond, D. Wheeler, A. Ally, M. Balasundaram, I. Birol, S. N. Butterfield, A. Chu, E. Chuah, H. J. Chun, N. Dhalla, R. Guin, M. Hirst, C. Hirst, R. A. Holt, S. J. Jones, D. Lee, H. I. Li, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, A. G. Robertson, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, R. J. Varhol, R. Beroukhim, A. S. Bhatt, A. N. Brooks, A. D. Cherniack, S. S. Freeman, S. B. Gabriel, E. Helman, J. Jung, M. Meyerson, A. I. Ojesina, C. S. Pedamallu, G. Saksena, S. E. Schumacher, B. Tabak, T. Zack, E. S. Lander, C. A. Bristow, A. Hadjipanayis, P. Haseley, R. Kucherlapati, S. Lee, E. Lee, L. J. Luquette, H. S. Mahadeshwar, A. Pantazi, M. Parfenov, P. J. Park, A. Protopopov, X. Ren, N. Santoso, J. Seidman, S. Seth, X. Song, J. Tang, R. Xi, A. W. Xu, L. Yang, D. Zeng, J. T. Auman, S. Balu, E. Buda, C. Fan, K. A. Hoadley, C. D. Jones, S. Meng, P. A. Mieczkowski, J. S. Parker, C. M. Perou, J. Roach, Y. Shi, G. O. Silva, D. Tan, U. Veluvolu, S. Waring, M. D. Wilkerson, J. Wu, W. Zhao, T. Bodenheimer, D. N. Hayes, A. P. Hoyle, S. R. Jeffreys, L. E. Mose, J. V. Simons, M. G. Soloway, S. B. Baylin, B. P. Berman, M. S. Bootwalla, L. Danilova, J. G. Herman, T. Hinoue, P. W. Laird, S. K. Rhie, H. Shen, T. Triche, D. J. Weisenberger, S. L. Carter, K. Cibulskis, L. Chin, J. Zhang, G. Getz, C. Sougnez, M. Wang, G. Saksena, S. L. Carter, K. Cibulskis, L. Chin, J. Zhang, G. Getz, H. Dinh, H. V. Doddapaneni, R. Gibbs, P. Gunaratne, Y. Han, D. Kalra, C. Kovar, L. Lewis, M. Morgan, D. Morton, D. Muzny, J. Reid, L. Xi, J. Cho, D. DiCara, S. Frazer, N. Gehlenborg, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, Y. Liu, M. S. Noble, P. Stojanov, D. Voet, H. Zhang, L. Zou, C. Stewart, B. Bernard, R. Bressler, A. Eakin, L. Iype, T. Knijnenburg, R. Kramer, R. Kreisberg, K. Leinonen, J. Lin, Y. Liu, M. Miller, S. M. Reynolds, H. Rovira, I. Shmulevich, V. Thorsson, D. Yang, W. Zhang, S. Amin, C. J. Wu, C. C. Wu, R. Akbani, K. Aldape, K. A. Baggerly, B. Broom, T. D. Casasent, J. Cleland, C. Creighton, D. Dodda, M. Edgerton, L. Han, S. M. Herbrich, Z. Ju, H. Kim, S. Lerner, J. Li, H. Liang, W. Liu, P. L. Lorenzi, Y. Lu, J. Melott, G. B. Mills, L. Nguyen, X. Su, R. Verhaak, W. Wang, J. N. Weinstein, A. Wong, Y. Yang, J. Yao, R. Yao,

K. Yoshihara, Y. Yuan, A. K. Yung, N. Zhang, S. Zheng, M. Ryan, D. W. Kane, B. A. Aksoy, G. Ciriello, G. Dresdner, J. Gao, B. Gross, A. Jacobsen, A. Kahles, M. Ladanyi, W. Lee, K. V. Lehmann, M. L. Miller, R. Ramirez, G. Ratsch, B. Reva, C. Sander, N. Schultz, Y. Senbabaoglu, R. Shen, R. Sinha, S. O. Sumer, Y. Sun, B. S. Taylor, N. Weinhold, S. Fei, P. Spellman, C. Benz, D. Carlin, M. Cline, B. Craft, K. Ellrott, M. Goldman, D. Haussler, S. Ma, S. Ng, E. Paull, A. Radenbaugh, S. Salama, A. Sokolov, J. M. Stuart, T. Swatloski, V. Uzunangelov, P. Waltman, C. Yau, J. Zhu, S. R. Hamilton, G. Getz, C. Sougnez, S. Abbott, R. Abbott, N. D. Dees, K. Delehaunty, L. Ding, D. J. Dooling, J. M. Eldred, C. C. Fronick, R. Fulton, L. L. Fulton, J. Kalicki-Veizer, K. L. Kanchi, C. Kandoth, D. C. Koboldt, D. E. Larson, T. J. Ley, L. Lin, C. Lu, V. J. Magrini, E. R. Mardis, M. D. McLellan, J. F. McMichael, C. A. Miller, M. O'Laughlin, C. Pohl, H. Schmidt, S. M. Smith, J. Walker, J. W. Wallis, M. C. Wendl, R. K. Wilson, T. Wylie, Q. Zhang, R. Burton, M. A. Jensen, A. Kahn, T. Pihl, D. Pot, Y. Wan, D. A. Levine, A. D. Black, J. Bowen, J. Frick, J. M. Gastier-Foster, H. A. Harper, C. Helsel, K. M. Leraas, T. M. Lichtenberg, C. McAllister, N. C. Ramirez, S. Sharpe, L. Wise, E. Zmuda, S. J. Chanock, T. Davidsen, J. A. Demchok, G. Eley, I. Felau, B. A. Ozenberger, M. Sheth, H. Sofia, L. Staudt, R. Tarnuzzer, Z. Wang, L. Yang, J. Zhang, L. Omberg, A. Margolin, B. J. Raphael, F. Vandin, H. T. Wu, M. D. Leiserson, S. C. Benz, C. J. Vaske, H. Noushmehr, T. Knijnenburg, D. Wolf, L. Van 't Veer, E. A. Collisson, D. Anastassiou, T. H. Ou Yang, N. Lopez-Bigas, A. Gonzalez-Perez, D. Tamborero, Z. Xia, W. Li, D. Y. Cho, T. Przytycka, M. Hamilton, S. McGuire, S. Nelander, P. Johansson, R. Jornsten, T. Kling, and J. Sanchez. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, Oct 2013.

[27] K. U. Eckardt, B. Barthlein, S. Baid-Agrawal, A. Beck, M. Busch, F. Eitner, A. B. Ekici, J. Floege, O. Gefeller, H. Haller, R. Hilge, K. F. Hilgers, J. T. Kielstein, V. Krane, A. Köttgen, F. Kronenberg, P. Oefner, H. U. Prokosch, A. Reis, M. Schmid, E. Schaeffner, U. T. Schultheiss, S. A. Seuchter, T. Sitter, C. Sommerer, G. Walz, C. Wanner, G. Wolf, M. Zeier, and S. Titze. The German Chronic Kidney Disease (GCKD) study: design and methods. *Nephrol. Dial. Transplant.*, 27(4):1454–1460, Apr 2012.

[28] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[29] J. Dong and S. Horvath. Understanding network concepts in modules. *BMC Syst Biol*, 1:24, Jun 2007.

[30] P. Langfelder and S. Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol*, 1:54, Nov 2007.

[31] S. Horvath and J. Dong. Geometric Interpretation of Gene Coexpression Network Analysis. *PLOS Computational Biology*, 4(8):1–27, 08 2008.

[32] P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath. Is my network module preserved and reproducible? *PLoS Comput. Biol.*, 7(1):e1001057, Jan 2011.

[33] P. Langfelder, J. P. Cantle, D. Chatzopoulou, N. Wang, F. Gao, I. Al-Ramahi, X. H. Lu, E. M. Ramos, K. El-Zein, Y. Zhao, S. Deverasetty, A. Tebbe, C. Schaab, D. J. Lavery, D. Howland, S. Kwak, J. Botas, J. S. Aaronson, J. Rosinski, G. Coppola, S. Horvath, and X. W. Yang. Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat. Neurosci.*, 19(4):623–633, Apr 2016.

[34] G. Pei, L. Chen, and W. Zhang. WGCNA Application to Proteomic and Metabolomic Data Analysis. *Meth. Enzymol.*, 585:135–158, 2017.

[35] V. Emilsson, M. Ilkov, J. R. Lamb, N. Finkel, E. F. Gudmundsson, R. Pitts, H. Hoover, V. Gudmundsdottir, S. R. Horman, T. Aspelund, L. Shu, V. Trifonov, S. Sigurdsson, A. Manolescu, J. Zhu, O. Olafsson, J. Jakobsdottir, S. A. Lesley, J. To, J. Zhang, T. B. Harris, L. J. Launer, B. Zhang, G. Eiriksdottir, X. Yang, A. P. Orth, L. L. Jennings, and V. Gudnason. Co-regulatory networks of human serum proteins link genetics to disease. *Science*, 361(6404):769–773, 08 2018.

[36] S. Horvath, Y. Zhang, P. Langfelder, R. S. Kahn, M. P. Boks, K. van Eijk, L. H. van den Berg, and R. A. Ophoff. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.*, 13(10):R97, Oct 2012.

[37] P. Langfelder, F. Gao, N. Wang, D. Howland, S. Kwak, T. F. Vogt, J. S. Aaronson, J. Rosinski, G. Coppola, S. Horvath, and X. W. Yang. MicroRNA signatures of endogenous Huntingtin CAG repeat expansion in mice. *PLOS ONE*, 13(1):1–20, 01 2018.

[38] S. Horvath, B. Zhang, M. Carlson, K. V. Lu, S. Zhu, R. M. Felciano, M. F. Laurance, W. Zhao, S. Qi, Z. Chen, Y. Lee, A. C. Scheck, L. M. Liau, H. Wu, D. H. Geschwind, P. G. Febbo, H. I. Kornblum, T. F. Cloughesy, S. F. Nelson, and P. S. Mischel. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl. Acad. Sci. U.S.A.*, 103(46):17402–17407, Nov 2006.

[39] Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, A. Leonardson, L. W. Castellini, S. Wang, M. F. Champy, B. Zhang, V. Emilsson, S. Doss, A. Ghazalpour, S. Horvath, T. A. Drake, A. J. Lusis, and E. E. Schadt. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, Mar 2008.

[40] C. L. Plaisier, S. Horvath, A. Huertas-Vazquez, I. Cruz-Bautista, M. F. Herrera, T. Tusie-Luna, C. Aguilar-Salinas, and P. Pajukanta. A systems genetics approach implicates USF1,

FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet.*, 5(9):e1000642, Sep 2009.

[41] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351):380–384, May 2011.

[42] M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley, B. David Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. N. Grant, and A. R. Jones. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416):391–399, Sep 2012.

[43] P. Langfelder, P. S. Mischel, and S. Horvath. When is hub gene selection better than standard meta-analysis? *PLoS ONE*, 8(4):e61505, 2013.

[44] Z. Xue, K. Huang, C. Cai, L. Cai, C. Y. Jiang, Y. Feng, Z. Liu, Q. Zeng, L. Cheng, Y. E. Sun, J. Y. Liu, S. Horvath, and G. Fan. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593–597, Aug 2013.

[45] K. Pearson. VII. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

[46] M. P. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. U.S.A.*, 102(12):4221–4224, Mar 2005.

[47] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.

[48] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.

[49] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, Aug 2002.

[50] S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, 2(1):E9, Jan 2004.

[51] A. L. Barabasi. Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413, Jul 2009.

[52] R. Sokal, C. Michener, and U. of Kansas. *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas, 1958.

[53] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics*, 24(13):41–49, 2008.

[54] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990.

[55] G. Tutz and H. Binder. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62(4):961–971, Dec 2006.

[56] A. Mayr, B. Hofner, E. Waldmann, T. Hepp, S. Meyer, and O. Gefeller. An Update on Statistical Boosting in Biomedicine. *Comput Math Methods Med*, 2017:6083072, 2017.

[57] N. T. Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827, 1987.

[58] D. Sobieszkoda, J. Czech, N. Gablo, M. Kopanska, J. Tabarkiewicz, A. Kolacinska, T. Robak, and I. Zawlik. MGMT promoter methylation as a potential prognostic marker for acute leukemia. *Archives of Medical Science: AMS*, 13(6):1433–1441, 10 2017.

[59] C. Gardin and H. Dombret. Hypomethylating Agents as a Therapy for AML. *Current Hematologic Malignancy Reports*, 12(1):1–10, Feb 2017.

[60] X. Thomas and C. Le Jeune. Treatment of Elderly Patients With Acute Myeloid Leukemia. *Current Treatment Options in Oncology*, 18(1):2, Jan 2017.

[61] E. Papaemmanuil, M. Gerstung, L. Bullinger, V. I. Gaidzik, P. Paschka, N. D. Roberts, N. E. Potter, M. Heuser, F. Thol, N. Bolli, G. Gundem, P. Van Loo, I. Martincorena, P. Ganly, L. Mudie, S. McLaren, S. O'Meara, K. Raine, D. R. Jones, J. W. Teague, A. P. Butler, M. F. Greaves, A. Ganser, K. Döhner, R. F. Schlenk, H. Döhner, and P. J. Campbell. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.*, 374(23):2209–2221, Jun 2016.

[62] J. Prada-Arismendy, J. C. Arroyave, and S. Röthlisberger. Molecular biomarkers in acute myeloid leukemia. *Blood Reviews*, 31(1):63 – 76, 2017.

[63] J. Yamazaki, R. Taby, J. Jelinek, N. J. Raynal, M. Cesaroni, S. A. Pierce, S. M. Kornblau, C. E. Bueso-Ramos, F. Ravandi, H. M. Kantarjian, and J. P. Issa. Hypomethylation of TET2 Target Genes Identifies a Curable Subset of Acute Myeloid Leukemia. *J. Natl. Cancer Inst.*, 108(2), Feb 2016.

[64] K. Meldi, T. Qin, F. Buchi, N. Droin, J. Sotzen, J. B. Micol, D. Selimoglu-Buet, E. Masala, B. Allione, D. Gioia, A. Poloni, M. Lunghi, E. Solary, O. Abdel-Wahab, V. Santini, and M. E. Figueroa. Specific molecular signatures predict decitabine response in chronic myelomonocytic leukemia. *J. Clin. Invest.*, 125(5):1857–1872, May 2015.

[65] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[66] R. Tibshirani. The lasso method for variable selection in the Cox model. *Stat Med*, 16(4):385–395, Feb 1997.

[67] M. Y. Park and T. Hastie. L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(4):659–677, 2007.

[68] J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, 33(1):1–22, 2010.

[69] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.*, 28(2):337–407, 04 2000.

[70] H. Binder. *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*, 2013. R package version 1.4.

[71] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[72] J. F. Lawless and Y. Yuan. Estimation of prediction error for survival models. *Stat Med*, 29(2):262–274, Jan 2010.

[73] M. Schumacher, H. Binder, and T. Gerds. Assessment of survival prediction models based on microarray data. *Bioinformatics*, 23(14):1768–1774, Jul 2007.

[74] T. A. Gerds and M. Schumacher. Efron-type measures of prediction error for survival analysis. *Biometrics*, 63(4):1283–1287, Dec 2007.

[75] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

[76] B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

[77] C. Porzelius, H. Binder, and M. Schumacher. Parallelized prediction error estimation for evaluation of high-dimensional models. *Bioinformatics*, 25(6):827–829, Mar 2009.

[78] H. Binder and M. Schumacher. Adapting Prediction Error Estimates for Biased Complexity Selection in High-Dimensional Bootstrap Samples. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 01 2008.

[79] T. Therneau and P. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer, 2000.

[80] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[81] R. Rampal, A. Alkalin, J. Madzo, A. Vasanthakumar, E. Pronier, J. Patel, Y. Li, J. Ahn, O. Abdel-Wahab, A. Shih, C. Lu, P. S. Ward, J. J. Tsai, T. Hricik, V. Tosello, J. E. Tallman, X. Zhao, D. Daniels, Q. Dai, L. Ciminio, I. Aifantis, C. He, F. Fuks, M. S. Tallman, A. Ferrando, S. Nimer, E. Paietta, C. B. Thompson, J. D. Licht, C. E. Mason, L. A. Godley, A. Melnick, M. E. Figueroa, and R. L. Levine. DNA hydroxymethylation profiling reveals that WT1 mutations result in loss of TET2 function in acute myeloid leukemia. *Cell Rep*, 9(5):1841–1855, Dec 2014.

[82] B. Lehne, A. W. Drong, M. Loh, W. Zhang, W. R. Scott, S. T. Tan, U. Afzal, J. Scott, M. R. Jarvelin, P. Elliott, M. I. McCarthy, J. S. Kooner, and J. C. Chambers. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol.*, 16:37, Feb 2015.

[83] J. Caldwell, I. Gardner, and N. Swales. An introduction to drug disposition: the basic principles of absorption, distribution, metabolism, and excretion. *Toxicol Pathol*, 23(2):102–114, 1995.

[84] A. Köttgen, J. Raffler, P. Sekula, and G. Kastenmüller. Genome-Wide Association Studies of Metabolite Concentrations (mGWAS): Relevance for Nephrology. *Semin. Nephrol.*, 38(2):151–174, 03 2018.

[85] G. Homuth, A. Teumer, U. Völker, and M. Nauck. A description of large-scale metabolomics studies: increasing value by combining metabolomics with genome-wide SNP genotyping and transcriptional profiling. *J. Endocrinol.*, 215(1):17–28, Oct 2012.

[86] S. Kalim and E. P. Rhee. An overview of renal metabolomics. *Kidney Int.*, 91(1):61–69, 01 2017.

[87] S. K. Nigam, W. Wu, K. T. Bush, M. P. Hoenig, R. C. Blantz, and V. Bhatnagar. Handling of drugs, metabolites, and uremic toxins by kidney proximal tubule drug transporters. *Clin J Am Soc Nephrol*, 10(11):2039–49, 2015.

[88] K. Suhre, H. Wallaschofski, J. Raffler, N. Friedrich, R. Haring, K. Michael, C. Wasner, A. Krebs, F. Kronenberg, D. Chang, C. Meisinger, H. E. Wichmann, W. Hoffmann, H. Volzke, U. Völker, A. Teumer, R. Biffar, T. Kocher, S. B. Felix, T. Illig, H. K. Kroemer, C. Gieger, W. Romisch-Margl, and M. Nauck. A genome-wide association study of metabolic traits in human urine. *Nat Genet*, 43(6):565–9, 2011.

[89] S. Y. Shin, E. B. Fauman, A. K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T. P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A. M. Valdes,

C. L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, M. Waldenberger, J. B. Richards, R. P. Mohney, M. V. Milburn, S. L. John, J. Trimmer, F. J. Theis, J. P. Overington, K. Suhre, M. J. Brosnan, C. Gieger, G. Kastenmüller, T. D. Spector, and N. Soranzo. An atlas of genetic influences on human blood metabolites. *Nat. Genet.*, 46(6):543–550, Jun 2014.

[90] T. Long, M. Hicks, H. C. Yu, W. H. Biggs, E. F. Kirkness, C. Menni, J. Zierer, K. S. Small, M. Mangino, H. Messier, S. Brewerton, Y. Turpaz, B. A. Perkins, A. M. Evans, L. A. Miller, L. Guo, C. T. Caskey, N. J. Schork, C. Garner, T. D. Spector, J. C. Venter, and A. Telenti. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.*, 49(4):568–578, Apr 2017.

[91] K. Suhre, J. Raffler, and G. Kastenmüller. Biochemical insights from population studies with genetics and metabolomics. *Arch. Biochem. Biophys.*, 589:168–176, Jan 2016.

[92] C. Gieger, L. Geistlinger, E. Altmaier, M. Hrabe de Angelis, F. Kronenberg, T. Meitinger, H. W. Mewes, H. E. Wichmann, K. M. Weinberger, J. Adamski, T. Illig, and K. Suhre. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.*, 4(11):e1000282, Nov 2008.

[93] C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, 10(5):e1004383, May 2014.

[94] J. Park, R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch, and K. Susztak. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*, 360(6390):758–763, 05 2018.

[95] H. Wu, K. Uchimura, E. L. Donnelly, Y. Kirita, S. A. Morris, and B. D. Humphreys. Comparative Analysis and Refinement of Human PSC-Derived Kidney Organoid Differentiation with Single-Cell Transcriptomics. *Cell Stem Cell*, 23(6):869–881, Dec 2018.

[96] H. U. Prokosch, S. Mate, J. Christoph, A. Beck, F. Kopcke, S. Stephan, M. W. Beckmann, T. Rau, A. Hartmann, B. Wullich, B. Breil, K. U. Eckardt, S. Titze, J. K. Habermann, J. Ingenerf, M. Hackmann, M. Ries, T. Burkle, and T. Ganslandt. Designing and implementing a biobanking IT framework for multiple research scenarios. *Stud Health Technol Inform*, 180:559–563, 2012.

[97] S. Titze, M. Schmid, A. Köttgen, M. Busch, J. Floege, C. Wanner, F. Kronenberg, K. U. Eckardt, K. U. Eckardt, S. Titze, H. U. Prokosch, B. Barthlein, A. Beck, T. Ganslandt, O. Gefeller, M. Schmid, J. Koster, M. Malzer, G. Schlieper, F. Eitner, S. Meisen, K. Kehl, E. Arweiler, J. Floege, E. Schaeffner, S. Baid-Agrawal, R. Schindler, S. Titze, S. Hubner,

T. Dienemann, K. F. Hilgers, K. U. Eckardt, A. Köttgen, U. Schultheiss, G. Walz, J. T. Kielstein, J. Lorenzen, H. Haller, C. Sommerer, M. Zeier, M. Busch, K. Paul, G. Wolf, R. Hilge, T. Sitter, V. Krane, D. Schmiedeke, S. Toncar, C. Wanner, A. B. Ekici, A. Reis, L. Forer, S. Schonherr, H. Weissensteiner, B. Kollertits, J. Raschenberger, F. Kronenberg, W. Gronwald, H. Zacharias, and P. Oefner. Disease burden and risk profile in referred patients with moderate chronic kidney disease: composition of the German Chronic Kidney Disease (GCKD) cohort. *Nephrol. Dial. Transplant.*, 30(3):441–451, Mar 2015.

[98] S. Das, L. Forer, S. Schonherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. Next-generation genotype imputation service and methods. *Nat. Genet.*, 48(10):1284–1287, 10 2016.

[99] A. M. Evans, B. R. Bridgewater, Q. Liu, M. W. Mitchell, R. J. Robinson, H. Dai, S. J. Stewart, C. D. DeHaven, and L. A. D. Miller. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. *Metabolomics*, 4(2):7, 2014.

[100] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.*, 78(13):4281–4290, Jul 2006.

[101] K. T. Do, S. Wahl, J. Raffler, S. Molnos, M. Laimighofer, J. Adamski, K. Suhre, K. Strauch, A. Peters, C. Gieger, C. Langenberg, I. D. Stewart, F. J. Theis, H. Grallert, G. Kastenmüller, and J. J. M. Krumsiek. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics*, 14(10):128, 2018.

[102] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene, and J. Coresh. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.*, 150(9):604–612, May 2009.

[103] C. Fuchsberger, D. Taliun, P. P. Pramstaller, and C. Pattaro. GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. *Bioinformatics*, 28(3):444–445, Feb 2012.

[104] C. J. Willer, Y. Li, and G. R. Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, Sep 2010.

[105] R. J. Pruim, R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337, Sep 2010.

[106] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, 38(6):226–231, Jun 1968.

[107] M. Arnold, J. Raffler, A. Pfeufer, K. Suhre, and G. Kastenmüller. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics*, 31(8):1334–1336, Apr 2015.

[108] M. J. Machiela and S. J. Chanock. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31(21):3555–3557, Nov 2015.

[109] C. E. Gillies, R. Putler, R. Menon, E. Otto, K. Yasutake, V. Nair, P. Hoover, D. Lieb, S. Li, S. Eddy, D. Fermin, M. T. McNulty, N. Hacohen, K. Kiryluk, M. Kretzler, X. Wen, M. G. Sampson, J. Sedor, K. Dell, M. Schachere, K. Lemley, L. Whitted, T. Srivastava, C. Haney, C. Sethna, K. Grammatikopoulos, G. Appel, M. Toledo, L. Greenbaum, C. S. Wang, B. Lee, S. Adler, C. Nast, J. LaPage, A. Athavale, A. Neu, S. Boynton, F. Fervenza, M. Hogan, J. C. Lieske, V. Chernitskiy, F. Kaskel, N. Kumar, P. Flynn, J. Kopp, E. Castro-Rubio, J. Blake, H. Trachtman, O. Zhdanova, F. Modersitzki, S. Vento, R. Lafayette, K. Mehta, C. Gadegbeku, D. Johnstone, D. Cattran, M. Hladunewich, H. Reich, P. Ling, M. Romano, A. Fornoni, L. Barisoni, C. Bidot, M. Kretzler, D. Gipson, A. Williams, R. Pitter, P. Nachman, K. Gibson, S. Grubbs, A. Froment, L. Holzman, K. Meyers, K. Kallem, F. Cerecino, K. Sambandam, E. Brown, N. Johnson, A. Jefferson, S. Hingorani, K. Tuttle, L. Curtin, S. Dismuke, A. Cooper, B. Freedman, J. J. Lin, S. Gray, M. Kretzler, L. Barisoni, C. Gadegbeku, B. Gillespie, D. Gipson, L. Holzman, L. Mariani, M. G. Sampson, P. Song, J. Troost, J. Zee, E. Herreshoff, C. Kincaid, C. Lienczewski, T. Mainieri, A. Williams, K. Abbott, C. Roy, T. Urv, and J. Brooks. An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. *Am. J. Hum. Genet.*, 103(2):232–244, Aug 2018.

[110] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[111] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, Jan 2000.

[112] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

[113] J. M. Major, K. Yu, C. C. Chung, S. J. Weinstein, M. Yeager, W. Wheeler, K. Snyder, M. E. Wright, J. Virtamo, S. Chanock, and D. Albanes. Genome-wide association study identifies three common variants associated with serologic response to vitamin e supplementation in men. *J Nutr*, 142(5):866–71, 2012.

[114] J. M. Major, K. Yu, W. Wheeler, H. Zhang, M. C. Cornelis, M. E. Wright, M. Yeager, K. Snyder, S. J. Weinstein, A. Mondul, H. Eliassen, M. Purdue, A. Hazra, C. A. McCarty, S. Hendrickson, J. Virtamo, D. Hunter, S. Chanock, P. Kraft, and D. Albanes. Genome-wide association study identifies common variants associated with circulating vitamin e levels. *Hum Mol Genet*, 20(19):3876–83, 2011.

[115] F. Takeuchi, R. McGinnis, S. Bourgeois, C. Barnes, N. Eriksson, N. Soranzo, P. Whittaker, V. Ranganath, V. Kumanduri, W. McLaren, L. Holm, J. Lindh, A. Rane, M. Wadelius, and P. Deloukas. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet*, 5(3):e1000433, 2009.

[116] J. Raffler, N. Friedrich, M. Arnold, T. Kacprowski, R. Rueedi, E. Altmaier, S. Bergmann, K. Budde, C. Gieger, G. Homuth, M. Pietzner, W. Romisch-Margl, K. Strauch, H. Volzke, M. Waldenberger, H. Wallaschofski, M. Nauck, U. Völker, G. Kastenmüller, and K. Suhre. Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. *PLoS Genet.*, 11(9):e1005487, Sep 2015.

[117] R. Rueedi, M. Ledda, A. W. Nicholls, R. M. Salek, P. Marques-Vidal, E. Morya, K. Sameshima, I. Montoliu, L. Da Silva, S. Collino, F. P. Martin, S. Rezzi, C. Steinbeck, D. M. Waterworth, G. Waeber, P. Vollenweider, J. S. Beckmann, J. Le Coutre, V. Mooser, S. Bergmann, U. K. Genick, and Z. Kutalik. Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS Genet*, 10(2):e1004132, 2014.

[118] K. Suhre, S. Y. Shin, A. K. Petersen, R. P. Mohney, D. Meredith, B. Wagele, E. Altmaier, P. Deloukas, J. Erdmann, E. Grundberg, C. J. Hammond, M. H. de Angelis, G. Kastenmüller, A. Köttgen, F. Kronenberg, M. Mangino, C. Meisinger, T. Meitinger, H. W. Mewes, M. V. Milburn, C. Prehn, J. Raffler, J. S. Ried, W. Romisch-Margl, N. J. Samani, K. S. Small, H. E. Wichmann, G. Zhai, T. Illig, T. D. Spector, J. Adamski, N. Soranzo, C. Gieger, S. Kathiresan, M. P. Reilly, N. J. Samani, H. Schunkert, J. Erdmann, T. L. Assimes, E. Boerwinkle, J. Erdmann, A. Hall, C. Hengstenberg, S. Kathiresan, I. R. Konig, R. Laaksonen, R. McPherson, M. P. Reilly, N. J. Samani, H. Schunkert, J. R. Thompson, U. Thorsteinsdottir, A. Ziegler, I. R. Konig, J. R. Thompson, D. Absher, L. Chen, L. A. Cupples, E. Halperin, M. Li, K. Musunuru, M. Preuss, A. Schillert, G. Thorleifsson, B. F. Voight, G. A. Wells, D. Absher, T. L. Assimes, P. Deloukas,

J. Erdmann, H. Holm, S. Kathiresan, I. R. Konig, R. McPherson, M. P. Reilly, R. Roberts, N. J. Samani, H. Schunkert, A. F. Stewart, S. Fortmann, A. Go, M. Hlatky, C. Iribarren, J. Knowles, R. Myers, T. Quertermous, S. Sidney, N. Risch, H. Tang, S. Blankenberg, T. Zeller, A. Schillert, P. Wild, A. Ziegler, R. Schnabel, C. Sinning, K. Lackner, L. Tiret, V. Nicaud, F. Cambien, C. Bickel, H. J. Rupprecht, C. Perret, C. Proust, T. Munzel, M. Barbalic, J. Bis, E. Boerwinkle, I. Y. Chen, L. A. Cupples, A. Dehghan, S. Demissie-Banjaw, A. Folsom, N. Glazer, V. Gudnason, T. Harris, S. Heckbert, D. Levy, T. Lumley, K. Marciante, A. Morrison, C. J. O'Donnell, B. M. Psaty, K. Rice, J. I. Rotter, D. S. Siscovick, N. Smith, A. Smith, K. D. Taylor, C. van Duijn, K. Volcik, J. Whitteman, V. Ramachandran, A. Hofman, A. Uitterlinden, S. Gretarsdottir, J. R. Gulcher, H. Holm, A. Kong, K. Stefansson, G. Thorgeirsson, K. Andersen, G. Thorleifsson, U. Thorsteinsdottir, J. Erdmann, M. Fischer, A. Grosshennig, C. Hengstenberg, I. R. Konig, W. Lieb, P. Linsel-Nitschke, M. Preuss, K. Stark, S. Schreiber, H. E. Wichmann, A. Ziegler, H. Schunkert, Z. Aherrahrou, P. Bruse, A. Doering, J. Erdmann, C. Hengstenberg, T. Illig, N. Klopp, I. R. Konig, P. Linsel-Nitschke, C. Loley, A. Medack, C. Meisinger, T. Meitinger, J. Nahrstedt, A. Peters, M. Preuss, K. Stark, A. K. Wagner, H. E. Wichmann, C. Willenborg, A. Ziegler, H. Schunkert, B. O. Bohm, H. Dobnig, T. B. Grammer, E. Halperin, M. M. Hoffmann, M. Kleber, R. Laaksonen, W. Marz, A. Meinitzer, B. R. Winkelmann, S. Pilz, W. Renner, H. Scharnagl, T. Stojakovic, A. Tomaschitz, K. Winkler, B. F. Voight, K. Musunuru, C. Guiducci, N. Burtt, S. B. Gabriel, D. S. Siscovick, C. J. O'Donnell, R. Elosua, L. Peltonen, V. Salomaa, S. M. Schwartz, O. Melander, D. Altshuler, S. Kathiresan, A. F. Stewart, L. Chen, S. Dandona, G. A. Wells, O. Jarinova, R. McPherson, R. Roberts, M. P. Reilly, M. Li, L. Qu, R. Wilensky, W. Matthai, H. H. Hakonarson, J. Devaney, M. S. Burnett, A. D. Pichard, K. M. Kent, L. Satler, J. M. Lindsay, R. Waksman, C. W. Knouff, D. M. Waterworth, M. C. Walker, V. Mooser, S. E. Epstein, D. J. Rader, N. J. Samani, J. R. Thompson, P. S. Braund, C. P. Nelson, B. J. Wright, A. J. Balmforth, S. G. Ball, and A. S. Hall. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60, Sep 2011.

[119] H. H. M. Draisma, R. Pool, M. Kobl, R. Jansen, A. K. Petersen, A. A. M. Vaarhorst, I. Yet, T. Haller, A. Demirkan, T. Esko, G. Zhu, S. Bohringer, M. Beekman, J. B. van Klinken, W. Romisch-Margl, C. Prehn, J. Adamski, A. J. M. de Craen, E. M. van Leeuwen, N. Amin, H. Dharuri, H. J. Westra, L. Franke, E. J. C. de Geus, J. J. Hottenga, G. Willemsen, A. K. Henders, G. W. Montgomery, D. R. Nyholt, J. B. Whitfield, B. W. Penninx, T. D. Spector, A. Metspalu, P. E. Slagboom, K. W. van Dijk, P. A. C. t Hoen, K. Strauch, N. G. Martin, G. B. van Ommen, T. Illig, J. T. Bell, M. Mangino, K. Suhre, M. I. McCarthy, C. Gieger, A. Isaacs,

C. M. van Duijn, and D. I. Boomsma. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun*, 6:7208, 2015.

[120] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: Biology, function, and translation. *Am J Hum Genet*, 101(1):5–22, 2017.

[121] A. Pushkin, G. Carpenito, N. Abuladze, D. Newman, V. Tsuprun, S. Ryazantsev, S. Motemoturu, P. Sassani, N. Solovieva, R. Dukkipati, and I. Kurtz. Structural characterization, tissue distribution, and functional expression of murine aminoacylase iii. *Am J Physiol Cell Physiol*, 286(4):C848–56, 2004.

[122] M. Veiga-da Cunha, D. Tyteca, V. Stroobant, P. J. Courtoy, F. R. Opperdoes, and E. Van Schaftingen. Molecular identification of nat8 as the enzyme that acetylates cysteine s-conjugates to mercapturic acids. *J Biol Chem*, 285(24):18888–98, 2010.

[123] Y. Kakimoto, K. Taniguchi, and I. Sano. D-beta-aminoisobutyrate:pyruvate aminotransferase in mammalian liver and excretion of beta-aminoisobutyrate by man. *J Biol Chem*, 244(2):335–40, 1969.

[124] B. Christiansen, A. K. Meinild, A. A. Jensen, and H. Brauner-Osborne. Cloning and characterization of a functional human gamma-aminobutyric acid (GABA) transporter, human GAT-2. *J Biol Chem*, 282(27):19331–41, 2007.

[125] R. Vanholder, R. De Smet, G. Glorieux, A. Argiles, U. Baurmeister, P. Brunet, W. Clark, G. Cohen, P. P. De Deyn, R. Deppisch, B. Descamps-Latscha, T. Henle, A. Jorres, H. D. Lemke, Z. A. Massy, J. Passlick-Deetjen, M. Rodriguez, B. Stegmayr, P. Stenvinkel, C. Tetta, C. Wanner, and W. Zidek. Review on uremic toxins: classification, concentration, and interindividual variability. *Kidney Int.*, 63(5):1934–1943, May 2003.

[126] E. P. Rhee and R. Thadhani. New insights into uremia-induced alterations in metabolic pathways. *Curr. Opin. Nephrol. Hypertens.*, 20(6):593–598, Nov 2011.

[127] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[128] C. Porzelius, M. Schumacher, and H. Binder. The benefit of data-based model complexity selection via prediction error curves in time-to-event data. *Computational Statistics*, 26(2):293–302, Jun 2011.

[129] W. Sauerbrei, A. L. Boulesteix, and H. Binder. Stability investigations of multivariable regression models derived from low- and high-dimensional data. *J Biopharm Stat*, 21(6):1206–1231, Nov 2011.

[130] S. Klau, M. L. Martin-Magniette, A. L. Boulesteix, and S. Hoffmann. Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. *Biom J*, May 2019.

[131] S. Wagner and D. Wagner. Comparing clusterings - an overview. Technical Report 4, Karlsruhe, 2007.

[132] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec 1985.

[133] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998.

[134] E. Brunner and U. Munzel. *Nichtparametrische Datenanalyse*. Springer, 2002.

[135] M. Hollander, D. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics. Wiley, 2013.

[136] I. Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[137] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85(21):4626–4628, Nov 2000.

[138] A.-L. Boulesteix, R. Hornung, and W. Sauerbrei. *On Fishing for Significance and Statistician's Degree of Freedom in the Era of Big Molecular Data*, pages 155–170. Springer Fachmedien Wiesbaden, Wiesbaden, 2017.

[139] J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*, pages 977–979. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[140] M. A. Fligner and S. W. Rust. On the independence problem and kendall's tau. *Communications in Statistics - Theory and Methods*, 12(14):1597–1607, 1983.

[141] L. Song, P. Langfelder, and S. Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13:328, Dec 2012.

[142] J. Li, D. Zhou, W. Qiu, Y. Shi, J. J. Yang, S. Chen, Q. Wang, and H. Pan. Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Sci Rep*, 8(1):622, 01 2018.

[143] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, 5(1):21, Jan 2011.

[144] Y. Xie, Y. Liu, and W. Valdar. Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. *Biometrika*, 103(3):493–511, 2016.

[145] J. Krumsiek, K. Suhre, A. M. Evans, M. W. Mitchell, R. P. Mohney, M. V. Milburn, B. Wagele, W. Romisch-Margl, T. Illig, J. Adamski, C. Gieger, F. J. Theis, and G. Kastenmüller. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.*, 8(10):e1003005, 2012.

[146] J. Thomas, T. Hepp, A. Mayr, and B. Bischl. Probing for Sparse and Fast Variable Selection with Model-Based Boosting. *Comput Math Methods Med*, 2017:1421409, 2017.

[147] K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, May 2005.

[148] O. Mahmoud, A. Harrison, A. Perperoglou, A. Gul, Z. Khan, M. V. Metodiev, and B. Lausen. A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. *BMC Bioinformatics*, 15:274, 2014.

[149] D. M. Gysi, A. Voigt, T. M. Fragoso, E. Almaas, and K. Nowick. wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics*, 19(1):392, Oct 2018.

[150] 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*, 12(1):1–24, 01 2017.

[151] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biol Cybern*, 59(4-5):291–294, 1988.

[152] I. Mallona, S. Aussó, A. Díez-Villanueva, V. Moreno, and M. A. Peinado. Modular dynamics of DNA co-methylation networks exposes the functional organization of colon cancer cells' genome. *bioRxiv*, 2018.

[153] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009.

[154] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4:7, 2015.

[155] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[156] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[157] H. Pagès, M. Carlson, S. Falcon, and N. Li. *AnnotationDbi: Annotation Database Interface*, 2018. R package version 1.44.0.

[158] T. Wei and V. Simko. *R package "corrplot": Visualization of a Correlation Matrix*, 2017. (Version 0.84).

[159] C. O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2019. R package version 0.9.4.

[160] H. Wickham, J. Hester, and W. Chang. *devtools: Tools to Make Developing R Packages Easier*, 2018. R package version 2.0.1.

[161] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2019. R package version 0.8.0.1.

[162] A. Kassambara and F. Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2017. R package version 1.0.5.999.

[163] S. Le, J. Josse, and F. Husson. FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[164] D. Wuertz, T. Setz, and Y. Chalabi. *fBasics: Rmetrics - Markets and Basic Statistics*, 2017. R package version 3042.89.

[165] G. project developers. *GenABEL: genome-wide SNP association analysis*, 2013. R package version 1.8-0.

[166] R. Gentleman and Biocore. *geneplotter: Graphics related functions for Bioconductor*, 2018. R package version 1.60.0.

[167] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[168] M. Carlson. *GO.db: A set of annotation maps describing the entire Gene Ontology*, 2018. R package version 3.7.0.

[169] M. Carlson. *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2)*, 2016. R package version 3.2.3.

[170] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[171] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[172] W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.

[173] M. C. Team, G. Blanchard, T. Dickhaus, N. Hack, F. Konietschke, K. Rohmeyer, J. Rosenblatt, M. Scheer, and W. Werft. *mutoss: Unified Multiple Testing Procedures*, 2017. R package version 0.1-12.

[174] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2018. R package version 3.1-137.

[175] A. Walker. *openxlsx: Read, Write and Edit XLSX Files*, 2018. R package version 4.1.0.

[176] S. Turner. *qqman: Q-Q and Manhattan Plots for GWAS Data*, 2017. R package version 0.1.4.

[177] J. D. Storey, A. J. Bass, A. Dabney, and D. Robinson. *qvalue: Q-value estimation for false discovery rate control*, 2019. R package version 2.14.1.

[178] D. Eddelbuettel and J. J. Balamuta. Extending *R* with *C++*: A Brief Introduction to *Rcpp*. *PeerJ Preprints*, 5:e3188v1, aug 2017.

[179] C. J. Miller. *simpleaffy: Very simple high level analysis of Affymetrix data*, 2019. http://www.bioconductor.org, http://bioinformatics.picr.man.ac.uk/simpleaffy/.

[180] R. Barter and B. Yu. *superheat: A Graphical Tool for Exploring Complex Datasets Using Heatmaps*, 2019. R package version 1.0.0.

[181] T. M. Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

[182] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr. Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, 2015.

[183] H. Wickham and L. Henry. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2018. R package version 0.8.2.

[184] H. Chen. *VennDiagram: Generate High-Resolution Venn and Euler Plots*, 2018. R package version 1.6.20.

[185] A. A. Dragulescu and C. Arendt. *xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files*, 2018. R package version 0.6.1.

[186] M. A. Jette, A. B. Yoo, and M. Grondona. Slurm: Simple linux utility for resource management. In *In Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP) 2003*, pages 44–60. Springer-Verlag, 2002.

APPENDIX A

Supplemental Material

File S1. ***Netboost package vignette.*** The vignette is attached as the separate file SupplementaryFile1.html.

File S2. ***Netboost package manual.*** The manual is attached as the separate file SupplementaryFile2.pdf.

Tables

Table S1: ***Metabolite annotation.*** Metabolite annotation is attached as the separate file SupplementaryTable1.xlsx.

Table S2: ***Statistics for the 46 eigenmetabolite-associated index SNPs.*** This table is attached as the separate file SupplementaryTable2.xlsx.

Table S3: ***Statistics for the 240 metabolite-associated index SNPs.*** This table is attached as the separate file SupplementaryTable3.xlsx.

Table S4: ***Results from ADME, KEGG pathway and GO term enrichment analysis for the 86 unique, implicated genes.*** This table is attached as the separate file SupplementaryTable4.xlsx.

Figures

Figure S1: ***Regional association plots for loci identified in GWAS of urinary metabolite concentrations.*** Regional association plots are attached as the separate file SupplementaryFigure2.pdf. For each of the 240 loci, the region for plotting was selected as the outer borders of merged overlapping 1-Mb windows. The extended MHC region was treated as one region. The index SNP with the lowest p-value is indicated. The metabolite giving rise to the association is included in the title. LD used to color-code correlation with the index SNP was based on the analyzed subsample of the GCKD study.

Figure S2: ***Regional association plots for loci identified in GWAS of eigenmetabolites.*** Regional association plots are attached as the separate file SupplementaryFigure1.pdf. For each of the 46 loci, the region for plotting was selected as the outer borders of merged overlapping 1-Mb windows. The extended MHC region was treated as one region. The index SNP with the lowest p-value is indicated. The eigenmetabolite giving rise to the association is included in the title. LD used to color-code correlation with the index SNP was based on the analyzed subsample of the GCKD study.

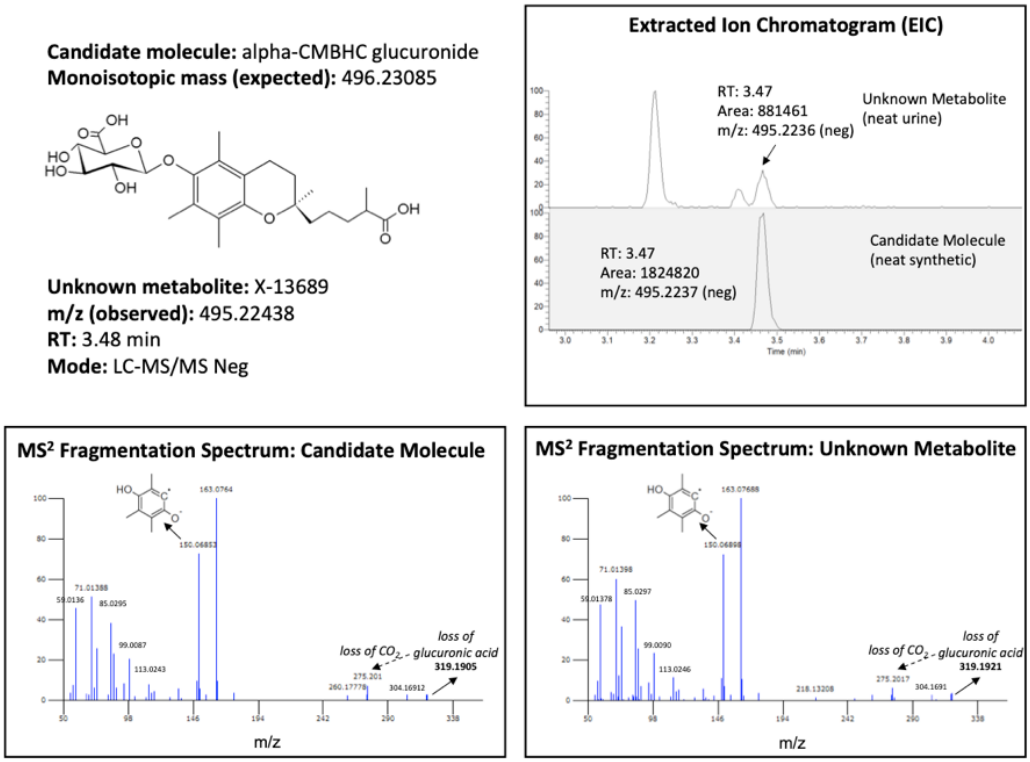

Figure S3: ***Identification of the unknown metabolite X-13689 as the glucuronide of alpha-CMBHC.*** The extracted ion chromatograms (upper right) show the same retention time for both the unknown metabolite in a reference urine matrix ("neat urine") and the candidate molecule in a neat solution ("neat synthetic"). The MS/MS fragmentation spectra of the candidate molecule (lower left) and of the unknown metabolite (lower right) show the same fragments with equal relative intensities; consequently, the candidate molecule is verified. The m/z (observed) for X-13689 is 495.22438, and the m/z (predicted) for alpha-CMBHC glucuronide is 495.22357, representing a 1.6 ppm error. The 319.1921 fragment peak represents the loss of glucuronic acid (a loss of 176), from which a loss of CO2 (-43.9898) yields the 275.201 fragment peak.

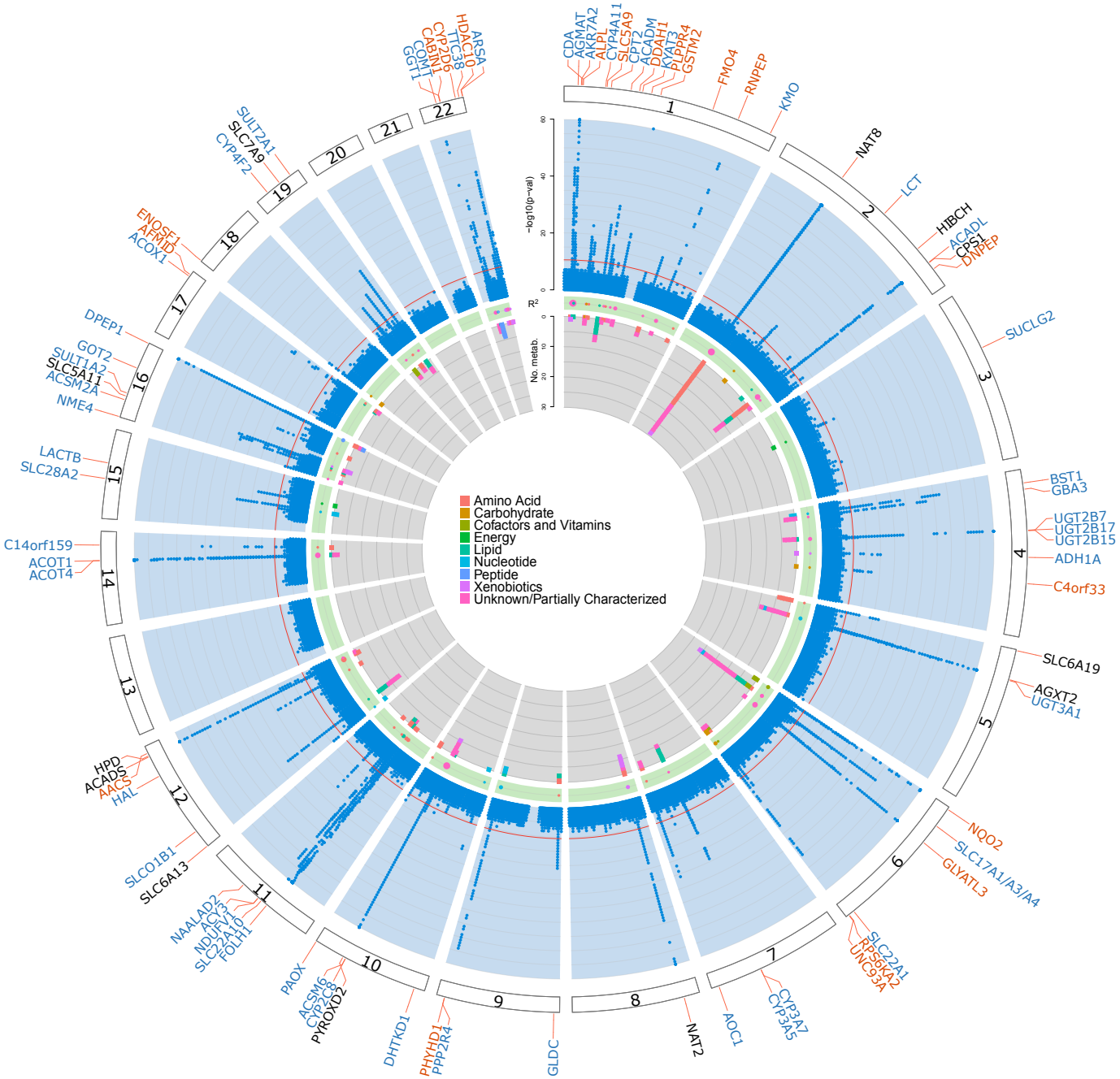
CDA
AGMAT
KCNJ16
AQP7
CYP4A11
SLC5A9
CPS1
ACADM
DDAH1
KYAT3
PLPP3
PLPHR4
GSTM2
FMO4
RNPEP
KMO
NAT8
LCT
HIBCH
ACADL
CPS1
DNPEP
SUCLG2
BST1
GBA3
UGT2B7
UGT2B17
UGT2B15
ADH1A
C4orf33
SLC6A19
AGXT2
UGT3A1
NQO2
SLC17A1/A3/A4
GLYATL3
SLC22A1
RPS6KA2
UNC93A
CYP3A7
CYP3A5
AOC1
NAT2
GLDC
PHYKPL
PPP2R4
DHTKD1
PAOX
ACSM6
CYP2C8
PYROXD2
NAALAD2
ACY3
NDUFV1
SLC2A10
FOLH1
SLCO1B1
SLC6A13
HPD
ACADS
AACS
HAL
C14orf159
ACOT1
ACOT4
LACTB
SLC28A2
DPEP1
GOT2
SULT1A2
SLC5A11
ACSM2A
NME4
ENOSF1
AMID
ACOX1
SULT2A1
SLC7A9
CYP4F2
ARSA
HDAC10
TTC38
CYP2D6
CABIN1
COMT
GGT1
−log10(p−val)
R²
No. metab.
Amino Acid
Carbohydrate
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Peptide
Xenobiotics
Unknown/Partially Characterized

Figure S4: **Genetic associations with metabolite concentrations in urine.** The light blue band shows the $-\log_{10}$(p-value) for genetic association with metabolite concentrations by chromosomal position. Associations of all 1,172 metabolites are overlaid in the light blue band, and are capped at 1e-60. The red line indicates genome-wide significance (p-value=4.3e-11). Black gene labels indicate genetic regions identified in previous mGWAS of urine, blue labels indicate genetic regions not identified in previous mGWAS of urine, and orange labels indicate genetic regions not yet identify in any mGWAS. The light green band shows the maximum variance in metabolite levels explained by the index SNP at each genetic region, with dot sizes corresponding to ([0,0.1),[0.1,0.25),[0.25,0.5),[0.5,1]) of explained variance, and dot colors reflecting the super-pathway of the metabolite with maximum variance explained. The inner gray band shows a stacked representation of the number of associated metabolites in each genetic region colored according to the super-pathways to which they belong. Color keys of metabolite super-pathways are presented in the middle.

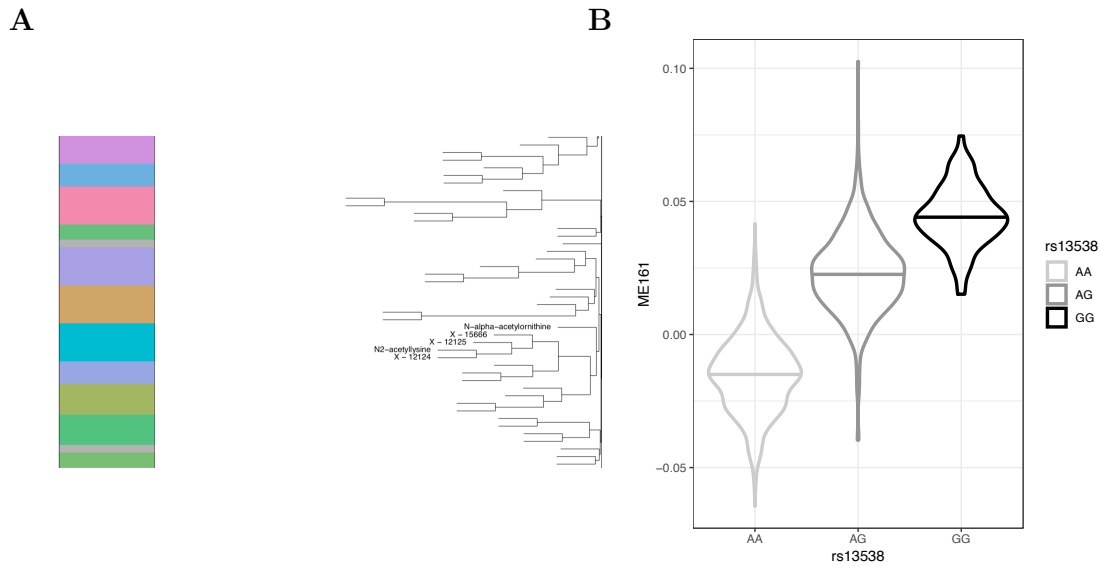

Figure S5: **Eigenmetabolite ME161 composition and genetic association with *NAT8*.** A) shows module ME161, for which metabolites are labeled, within the dendrogram for GCKD metabolites (Figure 16). B) displays the distribution of the eigenmetabolite of ME161 (Y-axis) with genotype at rs13538 in *NAT8* (X-axis).

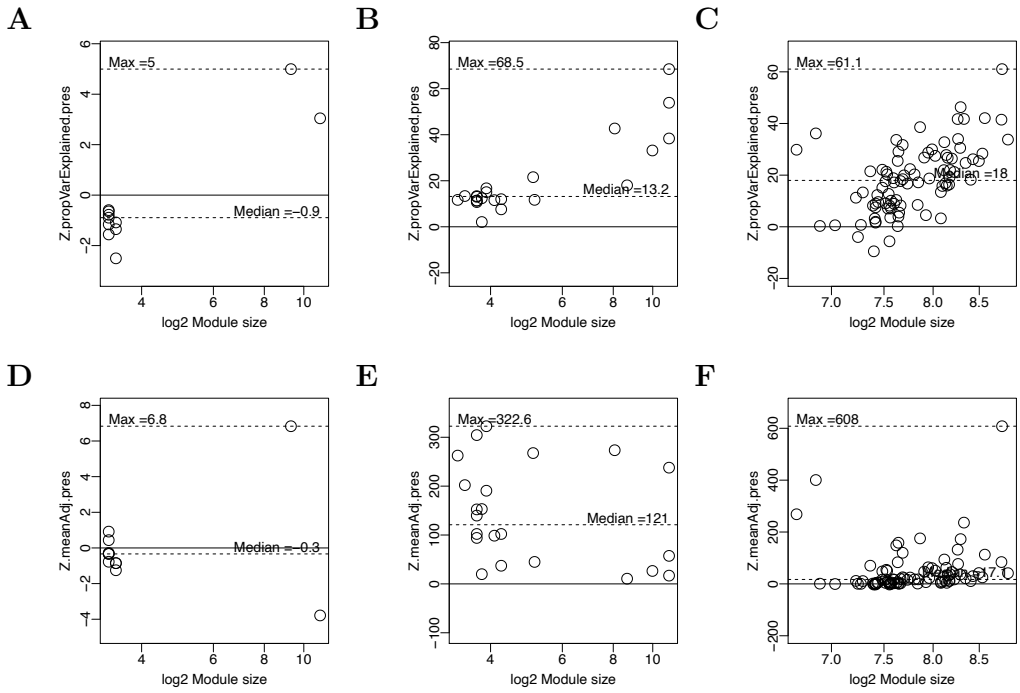

Figure S6: ***TCGA BRCA preservation statistics: explained variance and adjacency.*** The top row shows propVar(m) for TCGA BRCA DNAm for Netboost (A), WGCNA (B) and k-means with $k = 86$ (C). The lower row displays meanAdj(m) for TCGA BRCA DNAm for Netboost (D), WGCNA (E) and k-means with $k = 86$ (F). Dashed lines indicate maximum and median statistics across modules.

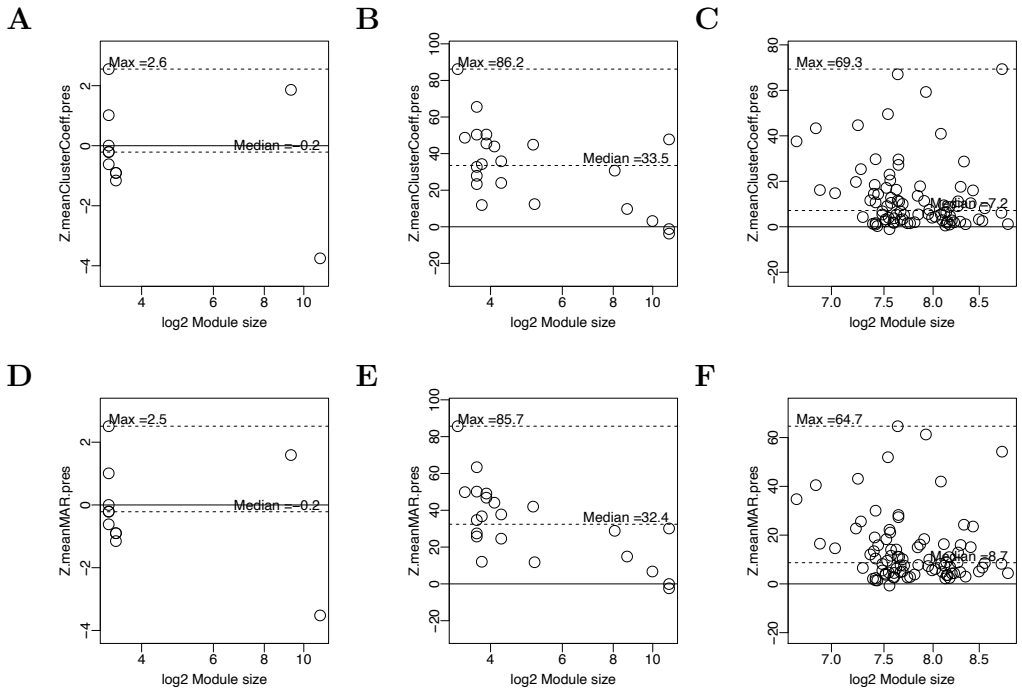

Figure S7: ***TCGA BRCA preservation statistics: cluster coefficient and maximum adjacency ratio.*** The top row shows meanClCoef(m) for TCGA BRCA DNAm for Netboost (A), WGCNA (B) and k-means with $k = 86$ (C). The lower row displays meanMAR(m) for TCGA BRCA DNAm for Netboost (D), WGCNA (E) and k-means with $k = 86$ (F). Dashed lines indicate maximum and median statistics across modules.

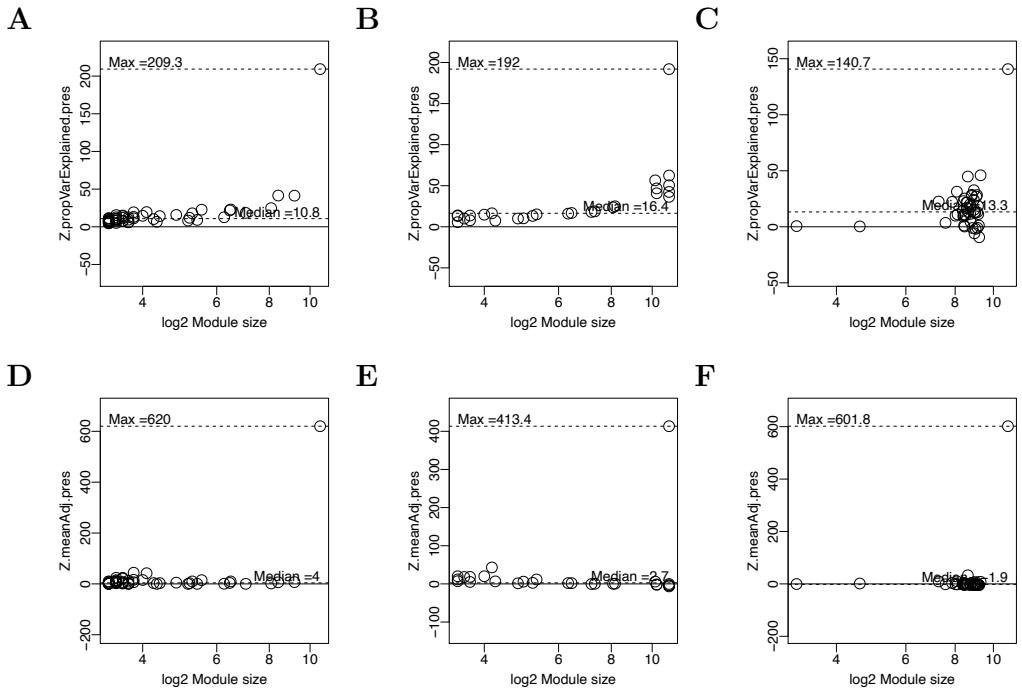

Figure S8: ***TCGA KIRC preservation statistics: explained variance and adjacency.*** The top row shows propVar(m) for TCGA KIRC DNAm for Netboost (A), WGCNA (B) and k-means with $k = 46$ (C). The lower row displays meanAdj(m) for TCGA KIRC DNAm for Netboost (D), WGCNA (E) and k-means with $k = 46$ (F). Dashed lines indicate maximum and median statistics across modules.

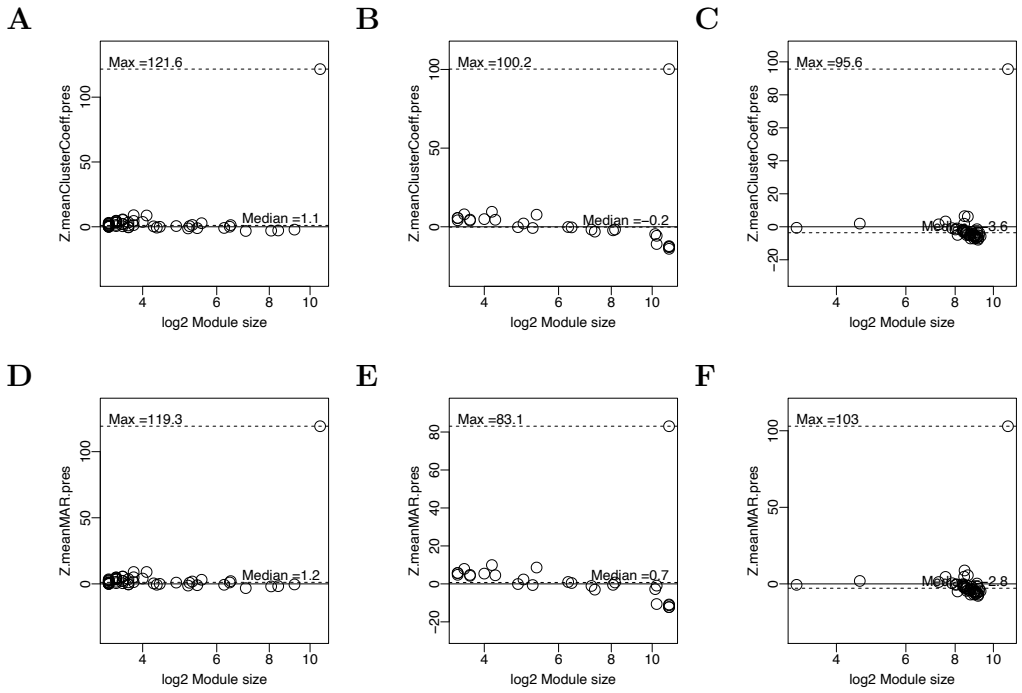

Figure S9: ***TCGA KIRC preservation statistics: cluster coefficient and maximum adjacency ratio.*** The top row shows meanClCoef(m) for TCGA KIRC DNAm for Netboost (A), WGCNA (B) and k-means with $k = 46$ (C). The lower row displays meanMAR(m) for TCGA KIRC DNAm for Netboost (D), WGCNA (E) and k-means with $k = 46$ (F). Dashed lines indicate maximum and median statistics across modules.

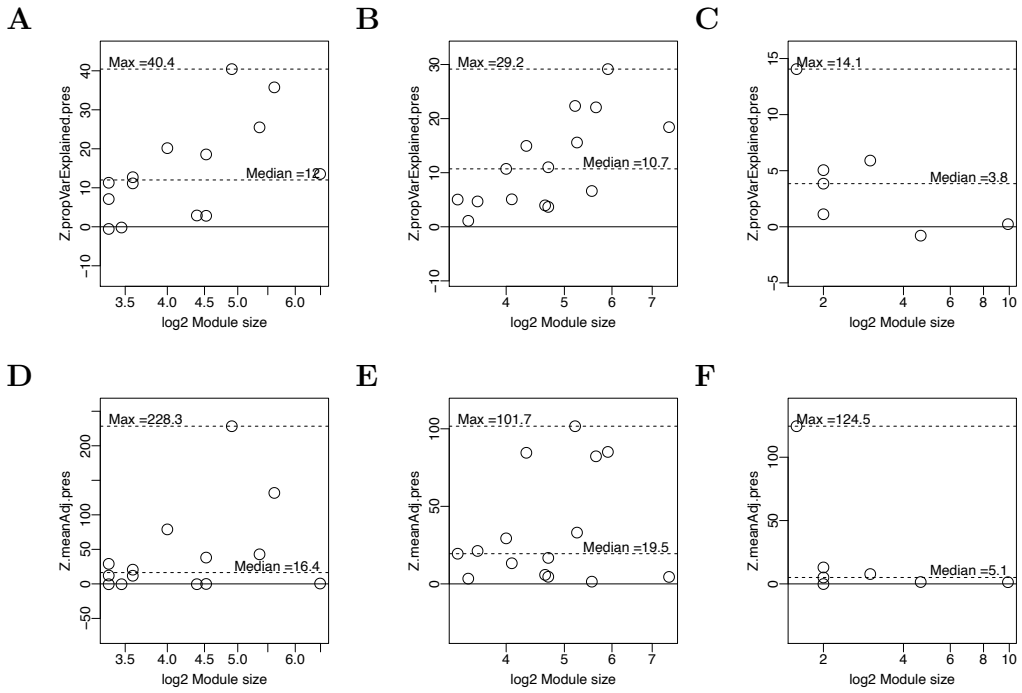

Figure S10: ***TCGA OV preservation statistics: explained variance and adjacency.*** The top row shows propVar(m) for TCGA OV DNAm for Netboost (A), WGCNA (B) and k-means with $k = 12$ (C). The lower row displays meanAdj(m) for TCGA OV DNAm for Netboost (D), WGCNA (E) and k-means with $k = 12$ (F). Dashed lines indicate maximum and median statistics across modules.

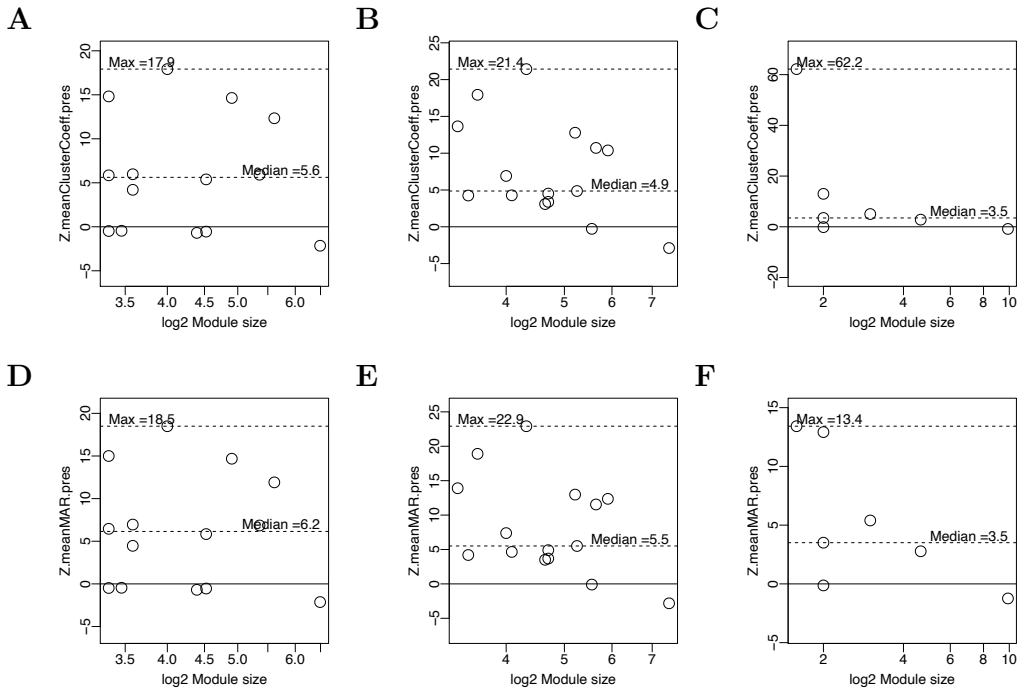

Figure S11: ***TCGA OV preservation statistics: cluster coefficient and maximum adjacency ratio.*** The top row shows meanClCoef(m) for TCGA OV DNAm for Netboost (A), WGCNA (B) and k-means with $k = 12$ (C). The lower row displays meanMAR(m) for TCGA OV DNAm for Netboost (D), WGCNA (E) and k-means with $k = 12$ (F). Dashed lines indicate maximum and median statistics across modules.

Software

The central applied software used in analyses is explicitly mentioned and referenced in the respective chapters. It follows a compilation of all software used:

- Circos [**153**]
- METAL [**104**]
- LDlink [**108**]
- LocusZoom [**105**]
- PLINK [**154**]
- `R` [**155**]
- `R` packages from CRAN and Bioconductor:
  `affy` [**156**], `AnnotationDbi` [**157**], `corrplot` [**158**], `cowplot` [**159**], `devtools` [**160**], `dplyr` [**161**], `DynamicTreeCut` [**22**], `factoextra` [**162**], `FactoMineR` [**163**], `fBasics` [**164**], `GAMBoost` [**55**], `GenABEL` [**165**], `geneplotter` [**166**], `ggplot2` [**167**], `GO.db` [**168**], `gtx` [**93**], `GWAtoolbox` [**103**], `hgu133plus2.db` [**169**], `limma` [**170**], `MASS` [**171**], `metafor` [**172**], `mutoss` [**173**], `nlme` [**174**], `openxlsx` [**175**], `parallel` [**155**], `peperr` [**77**], `qqman` [**176**], `qvalue` [**177**], `Rcpp` [**178**], `simpleaffy` [**179**], `superheat` [**180**], `survival` [**181**], `TCGAbiolinks` [**182**], `tidyr` [**183**], `VennDiagram` [**184**], `WGCNA` [**15**], and `xlsx` [**185**]
- Slurm [**186**]
- Sparse UPGMA [**53**]

Notation

$\mathbb{1}$	Indicator function (Chapter 3: 1.1)
$A = a_{i,j \in \{1,\dots,p\}}$	Adjacency Matrix (Chapter 2: 1, Chapter 2: 2.1.2)
$\text{adjRand}(M, M')$	Adjusted Rand index (Chapter 6: 1.1)
$b \in \mathbb{N}_+$	soft thresholding parameter (Chapter 2: 1)
β	Vector of regression coefficients (Chapter 3: 1.1)
B	Matrix of regression coefficients (Chapter 2: 2.1.1)
BS	Brier Score (Chapter 3: 1.2)
C	Contingency table (Chapter 6: 1.1)
$\text{clusterCoef}(i, X)$	Cluster Coefficient of a node (Chapter 6: 1.2)
corr	Pearson correlation coefficient
d_{ij}	Distance between node i and j (Chapter 2: 1, Chapter 2: 2.1.2)
exp	Exponential function
$\widehat{\text{Err}}_{(1)}$	Leave-one-out bootstrap error (Chapter 3: 1.2)
$\widehat{\text{Err}}_{\text{split}}$	Split-sample error (Chapter 3: 1.2)
$\overline{\text{Err}}$	Apparent error (Chapter 3: 1.2)
$\widehat{\text{Err}}_{.632}$	.632 error (Chapter 3: 1.2)
$\widehat{\text{Err}}_{.632+}$	.632+ error (Chapter 3: 1.2)
$\gamma \in \mathbb{R}_+$	Exponent of a scale-free topology (Chapter 2: 1)
$f : \mathbb{R} \to \mathbb{R}$	Function (Chapter 2: 2.4)
$\mathcal{F}$	Set of selected edges (Chapter 2: 2.1.1)
$h_i(\cdot)$	Base learner (Chapter 2: 2.1.1)
$i, j, u \in \{1, \dots, p\}$	Indices of variables / node
$\text{Jaccard}(M, M')$	Jaccard index (Chapter 6: 1.1)

$k \in \mathbb{R}_+$	Connectivity of a node (Chapter 2: 1).
	There are three different usages of k in this dissertation.
$k \in \mathbb{N}_+$	Number of nodes considered in a k-nearest neighbor
	algorithm (Chapter 4: 1.7).
	There are three different usages of k in this dissertation.
$k \in \mathbb{N}_+$	Number of clusters considered in a k-means
	algorithm (Chapter 6: 2).
	There are three different usages of k in this dissertation.
$l, r, s \in \mathbb{N}$	Indices
$l(\beta)$	Partial log-likelihood (Chapter 3: 1.1)
$L(\cdot)$	Loss function (Chapter 3: 1.1)
log	Natural logarithm
$\lambda_0(t)$	Baseline hazard (Chapter 3: 1.1)
$\lambda(t \mid Z_i)$	Conditional hazard, given covariate vector Z_i (Chapter 3: 1.1)
$m \in M$	Module (Chapter 2: 1)
$M \subseteq \mathfrak{P}(\{1, \ldots, p\})$	Partition of nodes (Chapter 6: 1.1)
ME_m	Module eigengene of module m
	(first principal component, Chapter 2: 1)
$\mathrm{meanAdj}(m, X)$	Mean adjacency of a module (Chapter 6: 1.2)
$\mathrm{meanClCoef}(m, X)$	Mean cluster coefficient of a module (Chapter 6: 1.2)
$\mathrm{meanMAR}(m, X)$	Mean maximum adjacency ratio of a module (Chapter 6: 1.2)
$\mathrm{MAR}(i, X)$	Maximum adjacency ratio of a node (Chapter 6: 1.2)
n	Number of observations
$n_{ab} := \lvert S_{ab} \rvert$	Number of pairs of variables that are clustered identically
	or not ($a, b \in \{0, 1\}$) (Chapter 6: 1.1)
NoInfErr	No information error (Chapter 3: 1.2)
$o \in \{1, \ldots, n\}$	Index of an observation
p	Number of variables

$P(T > t)$	Probability of T being greater than the value t
$\mathfrak{P}()$	Power set
Q	Set of observation indices (Chapter 3: 1.2)
R^2	Explained variation in a linear regression model
r^2	Measure of linkage disequilibrium (Chapter 4: 1.5)
ρ	Population correlation coefficient (Chapter 2: 2.4)
$\mathrm{Rand}(M, M')$	Rand index (Chapter 6: 1.1)
$\widehat{\mathrm{ROR}}$	Relative overfitting rate (Chapter 3: 1.2)
S_{ab}	Unordered pairs of variables that are clustered identically or not ($a, b \in \{0, 1\}$) (Chapter 6: 1.1)
$S(t)$	Survival function (Chapter 3: 1.1)
Σ	Covariance matrix (Chapter 2: 2.4)
t	Timepoint (Chapter 3: 1.1)
T	$\mathbb{R}_0^+$-valued random variable modeling survival time (Chapter 3: 1.1)
T_o	Realization of T for observation o (Chapter 3: 1.1)
TOM_{ij}	Topological overlap measure of node i and j (Chapter 2: 1, Chapter 2: 2.1.2)
$X \in \mathbb{R}^{n \times p}$	Data-matrix
X_i	$X_{o \leq n, i}$
X_{-i}	$X_{o \leq n, j \neq i}$
X^m	$X_{o \leq n, j \in m}$
${}_oX$	$X_{o, i \leq p}$
$Y : \mathbb{R}^{n \times (p-1)} \to \mathbb{R}^n$	Random variable

Acronyms

ADME absorption, distribution, metabolism, and excretion

AML acute myeloid leukemia

AMLSG acute myeloid leukemia study group

BRCA breast invasive carcinoma

CADD Combined Annotation Dependent Depletion

CAG cytosine, adenine and guanine

CART classification and regression tree

CCA canonical correlation analysis

CKD chronic kidney disease

CKD EPI Chronic Kidney Disease Epidemiology Collaboration

CPACOR incorporating Control Probe Adjustment and reduction of global CORrelation

CpG cytosine–phosphate–guanine

DNA deoxyribonucleic acid

DNAm DNA methylation

DRKS Deutsches Register Klinischer Studien

eGFR estimated Glomerular filtration rate

eQTL expression quantitative trait locus

GABA gamma-aminobutyric acid

GCKD German Chronic Kidney Disease

GEO Gene Expression Omnibus

GGM gaussian graphical model

GO Gene Ontology

GWAS genome-wide association study

HD Huntington's Disease

HGMD Human Gene Mutation Database

HRC Haplotype Reference Consortium

i.i.d. independent and identically distributed

KIRC kidney renal clear cell carcinoma

KEGG Kyoto Encyclopedia of Genes and Genomes

knn k-nearest neighbor

LD linkage disequilibrium

MAF minor allele frequency

mGWAS genome-wide association studies of metabolite concentrations

MHC major histocompatibility complex

NCI National Cancer Institute

NHGRI National Human Genome Research Institute

ME module eigengene

mRNA messenger ribonucleic acid

miRNA micro ribonucleic acid

mQTL metabolite quantitative trait locus

OMIM Online Mendelian Inheritance in Man

OV ovarian serous cystadenocarcinoma

PC principal component

PCA principal component analysis

POS proportional overlapping score

pQTL protein quantitative trait locus

RAP regional association plot

RNA ribonucleic acid

rPCA robust principal component analysis

RSD relative standard deviation

SD standard deviation

SNiPA single nucleotide polymorphisms annotator

SNP single nucleotide–polymorphism

TCGA The Cancer Genome Atlas

TOM topological overlap measure

UACR urinary albumin-to-creatinine ratio

UPGMA unweighted pair group method with arithmetic mean

WGCNA weighted gene co-expression network analysis

w.l.o.g. without loss of generality

Complete list of peer-reviewed publications

[1] Nadja Blagitko-Dorfs*, **Pascal Schlosser***, Gabriele Greve*, Dietmar Pfeifer, Ruth Meier, Annika Baude, David Brocks, Christoph Plass, and Michael Lübbert. Combination treatment of acute myeloid leukemia cells with DNMT and HDAC inhibitors: predominant synergistic gene downregulation associated with gene body demethylation. *Leukemia*, Nov 2018.

[2] Yong Li*, Peggy Sekula*, Matthias Wuttke, Judith Wahrheit, Birgit Hausknecht, Ulla T. Schultheiss, Wolfram Gronwald, **Pascal Schlosser**, Sara Tucci, Arif B. Ekici, Ute Spiekerkoetter, Florian Kronenberg, Kai-Uwe Eckardt, Peter J. Oefner, and Anna Köttgen. Genome-wide association studies of metabolites in patients with CKD identify multiple loci and illuminate tubular transport mechanisms. *J Am Soc Nephrol*, 29(5), 2018.

[3] Michael Jeserich, Bela Merkely, **Pascal Schlosser**, Simone Kimmel, Gabor Pavlik, Stephan Achenbach, and Jürgen Biermann. Concurrent exercise-associated ventricular complexes and a prolonged QT interval are associated with evidence of myocarditis. *J Cardiovasc Dis Diagn*, 6(1), 2018.

[4] Audrey Y. Chu*, Adrienne Tin*, **Pascal Schlosser***, Yi-An Ko, Chengxiang Qiu, Chen Yao, Roby Joehanes, Morgan E. Grams, Liming Liang, Caroline A. Gluck, Chunyu Liu, Josef Coresh, Shih-Jen Hwang, Daniel Levy, Eric Boerwinkle, James S. Pankow, Qiong Yang, Myriam Fornage, Caroline S. Fox, Katalin Susztak, and Anna Köttgen. Epigenome-wide association studies identify DNA methylation associated with kidney function. *Nat Commun*, 8(1), 2017.

[5] Michael Jeserich, Bela Merkely, **Pascal Schlosser**, Simone Kimmel, Gabor Pavlik, and Jürgen Biermann. Early diastolic septal movement in patients with myocarditis. *Clin Radiol*, 73(2), 2017.

[6] Michael Jeserich, Bela Merkely, **Pascal Schlosser**, Simone Kimmel, Gabor Pavlik, and Jürgen Biermann. Assessment of edema using STIR+ via 3D cardiovascular magnetic resonance imaging in patients with suspected myocarditis. *MAGMA*, 30(3), 2017.

[7] Anselm Hoppmann, **Pascal Schlosser**, Rolf Backofen, Ekkehart Lausch, and Anna Köttgen. GenToS: Use of orthologous gene information to prioritize signals from human GWAS. *PLoS One*, 11(9), 2016.

[8] Dieter Henrik Heiland, Irina Mader, **Pascal Schlosser**, Dietmar Pfeifer, Maria Stella Carro, Thomas Lange, Ralf Schwarzwald, Ioannis Vasilikos, Horst Urbach, and Astrid Weyerbrock. Integrative network-based analysis of magnetic resonance spectroscopy and genome wide expression in glioblastoma multiforme. *Sci Rep*, 6, 2016.

* indicates that these authors contributed equally to the respective publication.