Supplementary data

MetaboDiff: an R package for differential metabolomic analysis

Andreas Mock, Rolf Warta, Steffen Dettling, Benedikt Brors, Dirk Jäger and Christel Herold-Mende

Contents

1 Supplementary methods

1.1 MultiAssayExperiment class

The metabolomic data within `MetaboDiff` are stored as a `MultiAssayExperiment` class[1]. This framework enables the coordinated representation of multiple experiments on partially overlapping samples with associated metadata and integrated subsetting across experiments. In the context of metabolomic data analysis, multiple assays are needed to store raw data and imputed data which usually contain different number of metabolites due to missing values.

The core components of the `MultiAssayExperiment` class are:

- `ExperimentList` - a slot of class ExperimentList containing data for each experimental assay. Within the ExperimentList slot, the metabolomic data are stored as SummarizedExperiment objects consisting of:
  - `assay` - a matrix containing the relative metabolic measurements.
  - `rowData` - a dataframe containing the metabolite annotation.
- `colData` - a slot of class dataframe describing the sample metadata available across all experiments.
- `sampleMap` - a slot of class dataframe relating clinical data to experimental assay.
- `metadata` - a slot of class list. Within MetaboDiff, this slot contains a list of dataframes summarizing the results from the comparative data analysis.

Please refer to the `MultiAssayExperiment` vignette for more information about the class.

1.2 Metabolic correlation network analysis

The workflow was adapted from the weighted gene co-expression analysis (WGCNA) proposed by Langfelder and Horvarth[2] and makes use of the functions of the corresponding WGCNA R package[3].

Within MetaboDiff, all steps for WGCNA are performed within a set of functions connected by pipes:

- `diss_matrix` - construction of dissimilarity matrix
- `identify_modules` - identification of metabolic correlation modules
- `name_modules` - name metabolic correlation modules
- `calculate_MS` - calculation of module significance to relate sample traits to modules

The individual steps will be explained as follows. Table 1 presents the corresponding terminology.

Table 1: Glossary of WGCNA terminology.

term	definition
module	cluster of highly interconnected metabolites (high absolute correlation)
module eigenmetabolite	first principal component of a given module. It can be considered a representative of the metabolic profiles in a module
metabolite significance	minus log of p-value of hypothesis test for the individual metabolite.
module significance	average absolute metabolite significance measure for all metabolites in a given module.

[1]Sig M (2017). MultiAssayExperiment: Software for the integration of multi-omics experiments in Bioconductor. R package version 1.2.1

[2]Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4(1), Article17. http://doi.org/10.2202/1544-6115.1128

[3]Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559–559. http://doi.org/10.1186/1471-2105-9-559

1.2.1 Construction of dissimilarity matrix

The first step in constructing a metabolic correlation network is the creation of a dissimilarity matrix. Biweight midcorrelation was used as a similiarity measure as it is more robust to outliers than the absolute correlation coefficient[4]. This choice is important, as we do not expect all metabolites to be correlated in all patients.

The core concept of the so called "weighted" correlation analysis by Langfelder and Horvarth is that instead of defining a "hard" threshold (e.g. an absolute correlation coefficient > 0.8) to decide whether a node to connected to another, the adjacency `a` is defined by raising the similarity `s` to a power `beta` ("soft" threshold):

$$a_{ij} = s_{ij}^{\beta} \tag{1}$$

Lastly, the dissimilarity measure `w` is defined by

$$w_{ij} = 1 - a_{ij} \tag{2}$$

For detailed rationale of this approach, please see Zhang and Horvath[5]. For metabolic networks, we identified that a beta value of 3 was the lowest power for which the scale-free topology of the topology held true. The function `diss_matrix` creates the dissimilarity measure and saves it in the metadata slot

1.2.2 Identification of metabolic correlation modules

To identify metabolic correlation modules, metabolites are next clustered based on the dissimilarity measure where branches of the dendrogram correspond to modules. Ultimately, modules are detected by applying a branch cutting method with a minimal module size of 5 metabolites. We employed the dynamic branch cut method developed by Langfelder and colleagues[6], as constant height cutoffs exhibit suboptimal performance on complicated dendrograms. Supplementary Figure 1A shows the hierarchical clustering and corresponding modules after branch cutting.

The relation between the identified metabolic correlation modules can be visualized by a dendrogram of their *eigenmetabolite* (Suppl. Fig. 1B). The module *eigenmetabolite* is defined as the first principal component of all metabolite measurements in the respective module. It could be shown that the *eigenmetabolite* (in the case of the citation: eigengene) is highly correlated with the metabolite that has the highest intramodular connectivity[7].

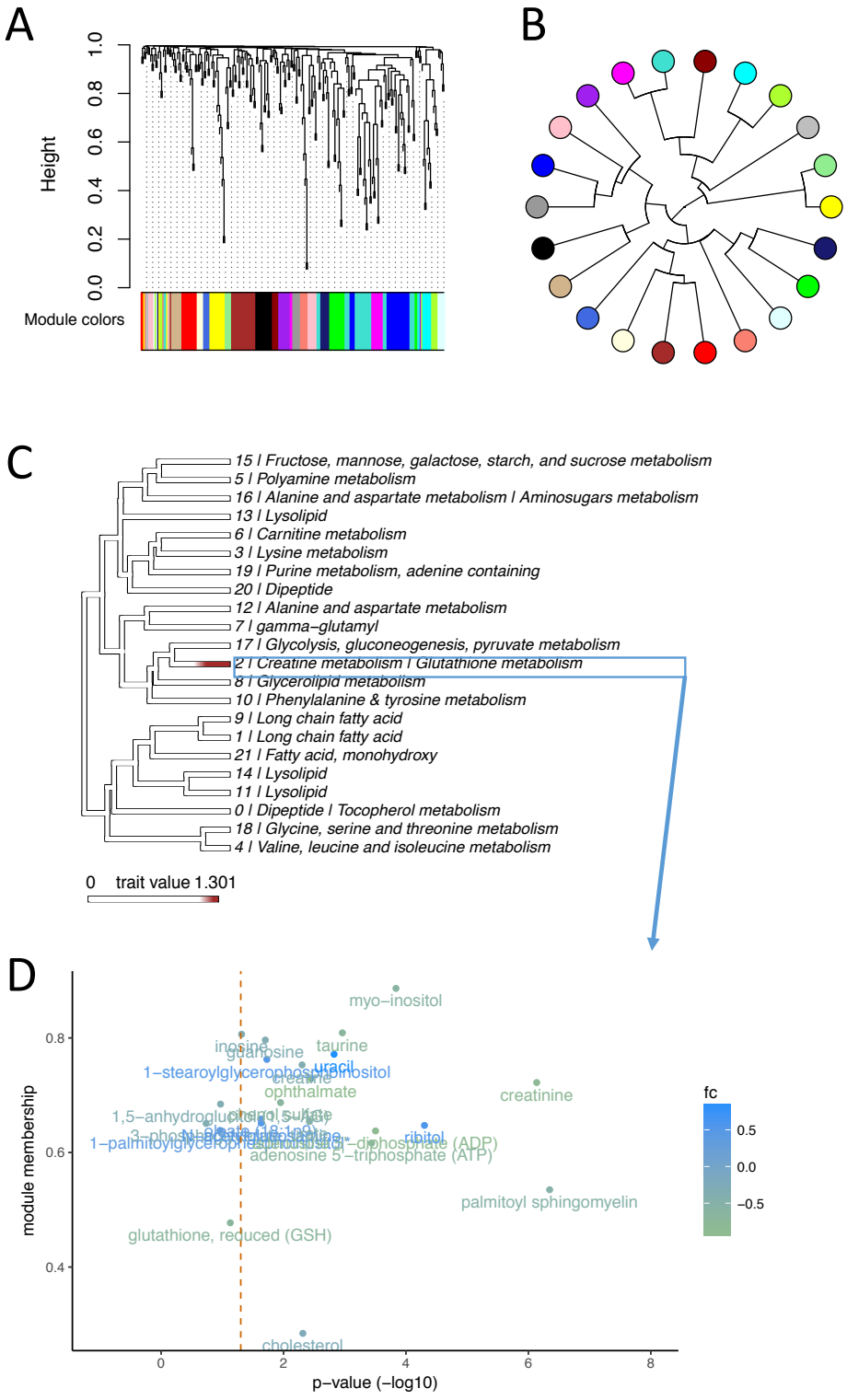
To enable a better interpretation of metabolic correlation modules, modules are named according to the most abundant pathway annotation in a module.

[4]Zheng, C.-H., Yuan, L., Sha, W., & Sun, Z.-L. (2014). Gene differential coexpression analysis based on biweight correlation and maximum clique. BMC Bioinformatics, 15 Suppl 15(Suppl 15), S3. http://doi.org/10.1186/1471-2105-15-S15-S3

[5]Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4(1), Article17. http://doi.org/10.2202/1544-6115.1128

[6]Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics, 24(5), 719–720. http://doi.org/10.1093/bioinformatics/btm563.

[7]Horvath, S., & Dong, J. (2008). Geometric Interpretation of Gene Coexpression Network Analysis. PLoS Computational Biology (PLOSCB) 4(8), 4(8), e1000117–e1000117. http://doi.org/10.1371/journal.pcbi.1000117



Supplementary Figure 1 (A) Hierarchical clustering of metabolites. The different colors represent the modules identified by the dynamic branch cutting method. (B) Hierarchy of metabolic correlation modules as revealed by the clustering of module eigengenes. Each node represents a metabolic correlation module. (C) Association of tumor vs normal trait with metabolic correlation modules. Module 2 (Creatine metabolism / Glutathione metabolism) showed a significant module significance. (D) Exploration of indidivual metabolites in module 2 by plotting the module membership over the p-value.

1.2.3 Calculation of module signficance to relate sample traits to modules

An advantage of correlation network analysis is the possibility to integrate external information. At the lowest hierarchical level, *metabolite significance* (MetS) measures can be defined as the statistical significance (i.e. p-value, p_i) between the i-th node profile (metabolite) x_i and the sample trait T

$$MetS_i = -log\ p_i \tag{3}$$

Module significance (MS) in turn can be determined as the average absolute metabolite significance measure. This conceptual framework can be adapted to any research question.

Supplementary Figure 1C shows that metabolic correlation module 2 (Creatine metabolism / Glutathione metabolism) was significantly associated with the tumor vs. normal comparison in the example data.

1.2.4 Exploration of individual metabolites within correlation module

Assessing the module significance for different sample traits facilitates an understanding of individual metabolic correlation modules. As for metabolomics, we are next interested in the role of the individual metabolite within module. To this end, Langfelder and Horvath suggest a 'fuzzy' measure of *module membership* defined as

$$K^q = |cor(x_i, E^q)| \tag{4}$$

where x_i is the profile of metabolite i and E^q is the eigenmetabolite of module q. Based on this definition, K describes how closely related metabolite i is to module q. A meaningful visualization is consequently plotting the module membership over the p-value of the respective $MetS$ measure (Suppl. Fig. 1D). As a third dimension, the color is scaled according to the effect size (i.e. fold-change).

2 Supplementary results

2.1 Case study

To showcase and benchmark the functionality of MetaboDiff, we performed a case study using three datasets of the work by Priolo and coworkers[8].

Table 2: Sample characteristics and number of metabolites (n) measured for three study sets of the case study.

Study set	control samples	AKT1 samples	MYC samples	n (metab.)	n (metab.) after imputation
cells	5	5	5	133	118
mice	6	6	6	170	142
human	25	21	9	307	236

We deliberately chose these data sets as (i) they are derived from three different organisms (cells, mice and human), (ii) they comprise different number of metabolites (133-307 metabolites), (iii) they relate a research question across data sets and (iv) they compare three groups (Control samples vs. AKT1-driven tumors vs. MYC-driven tumors).

The findings of the case study will be presented as follows. To ensure reproducibility, the data as well as the code to generate all plots is part of the MetaboDiff package markdown vignette `Case_study`.

Research question. AKT1 and MYC are two of the most common prostate cancer oncogenes. Priolo and colleagues used untargeted metabolomics to compare the metabolic profiles of AKT1- and MYC-driven prostate cancers to normal control. They spanned their research question across three different organisms (transfected cells, transgenic mice and human prostate (cancer) tissues).

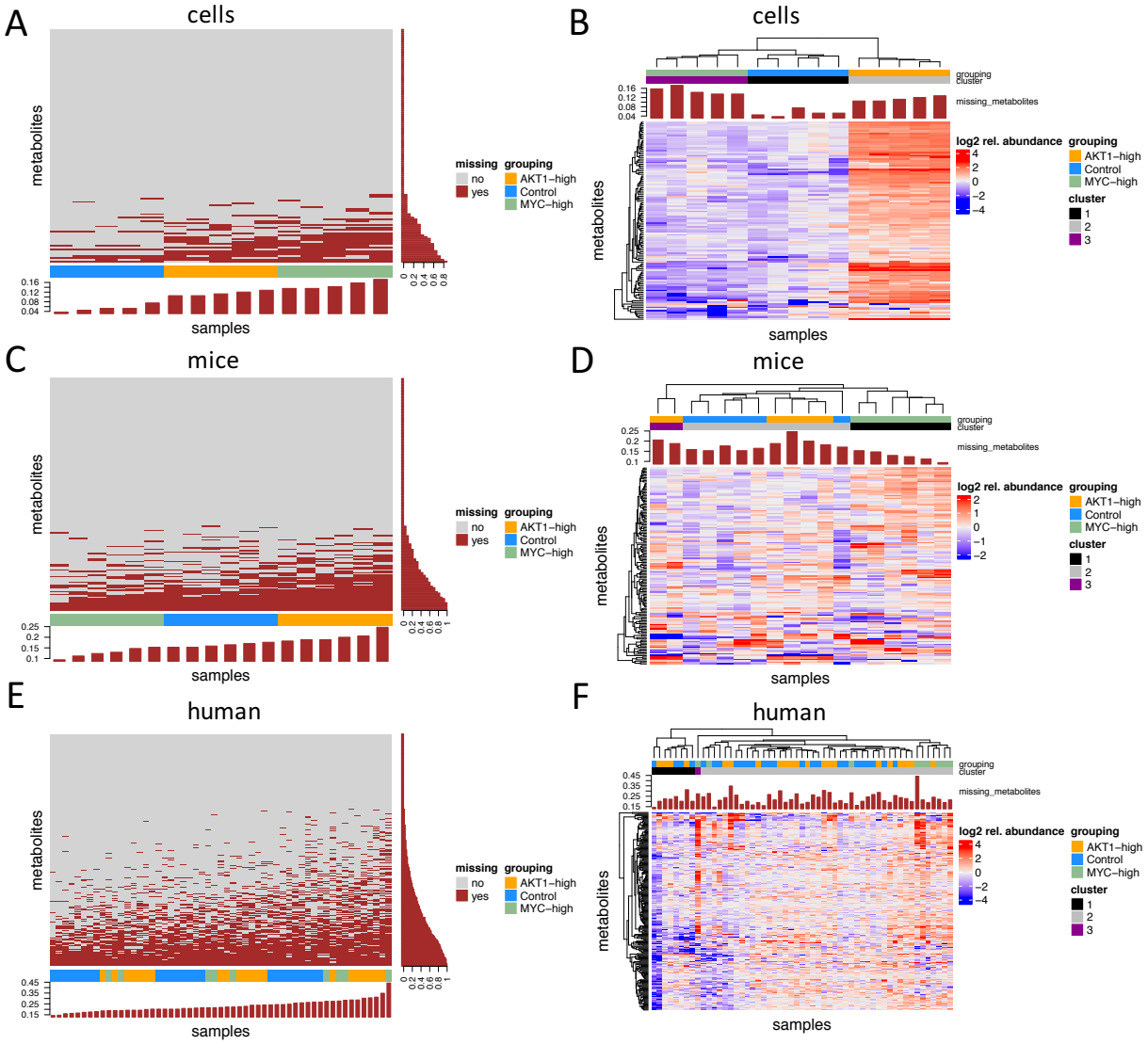
2.2 Data processing

The raw relative metabolic measurements of the data sets were obtained from the supplemental material of the publication. The number of named metabolites ranged from 133 (cells) to 307 (human). The authors did not comment on reasons for this difference. As for every untargated metabolomic data set, a substantial number of metabolite measurements were missing (Suppl. Fig. 2A,C,E). K-nearest neighbor imputation was used to impute all metabolites with raw measurements in more than 60% of cases. Metabolites with less than 60% measurements across the data set were excluded. Table 2 summarizes the number of metabolites before and after imputation. In relative terms, most metabolites were excluded in the human data set (> 23 %).

To identify outliers and to assess the impact of imputation the imputed metabolite measurements were clustered by means of a hierarchical clustering and plotted as a heatmap (Suppl. Fig. 2B,D,F). The column annotation also displays the results of a k-means clustering with k=3. As expected, the transfected cell lines showed the most distinct clustering without a sign for outliers, seperating all three groups. It is intriguing, that nearly all measured metabolites seem to be more abundant in AKT1-driven cells. In addition, fewer metabolites were missing in the control cells. In the transgenic mice and the human samples the clustering was not that distinct and putative outliers were present. This finding is line with the clustering of metabolic subpathways in figure 2 of the Priolo manuscript.

[8]Priolo, C., Pyne, S., Rose, J., Regan, E. R., Zadra, G., Photopoulos, C., et al. (2014). AKT1 and MYC Induce Distinctive Metabolic Fingerprints in Human Prostate Cancer. Cancer Research, 74(24), 7198–7204. http://doi.org/10.1158/0008-5472.CAN-14-1490

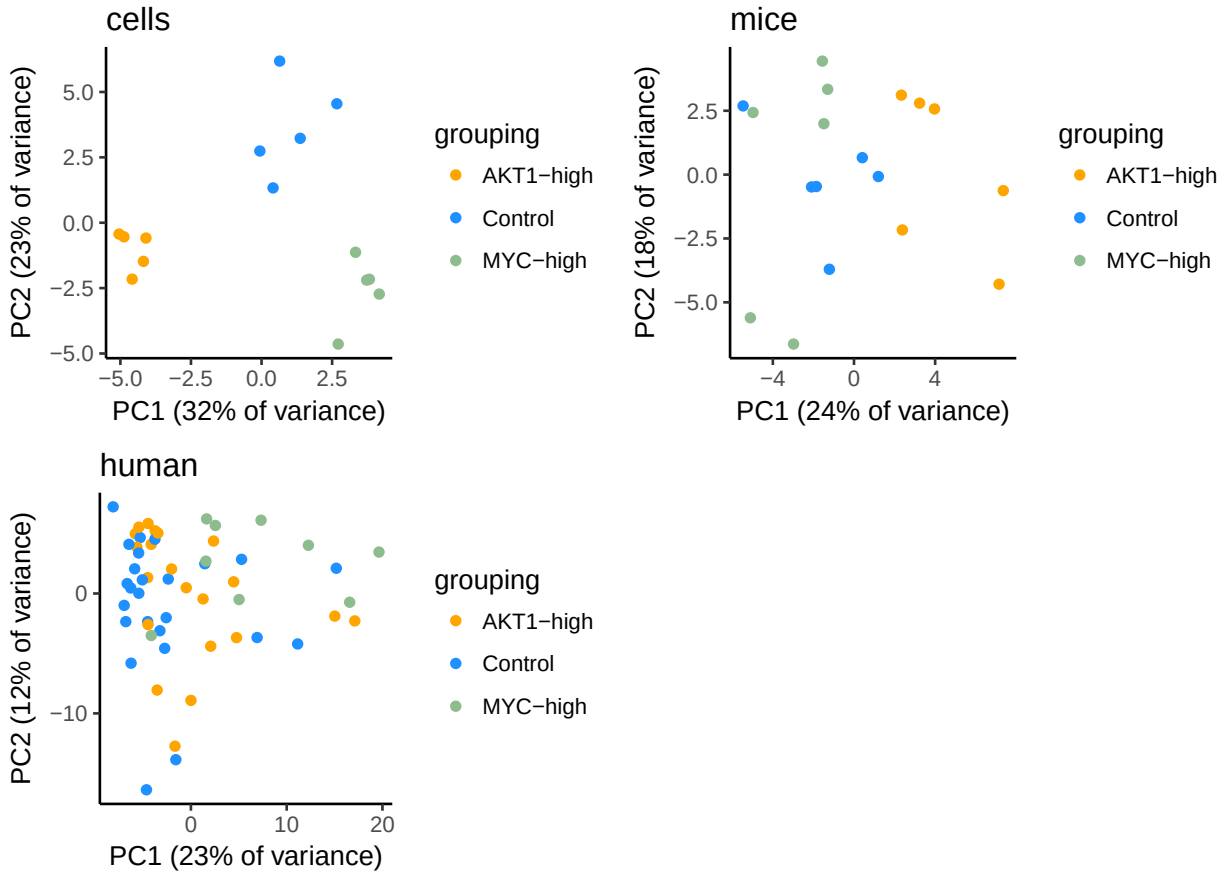
However, due to the missing batch information and additional sample traits, a removal of samples was not performed. As the last step of data processing the imputed metabolite measurements were normalized by variance stabilizing normalization (vsn).



Supplementary Figure 2 (A,C,E) Binary heatmaps of missing values in raw metabolite measurements. (B,D,F) Hierarchical clustering of the imputed metabolite measurements. The results of a k-means clustering (k=3) are included in the column annotatation.

2.3 Unsupervised analysis

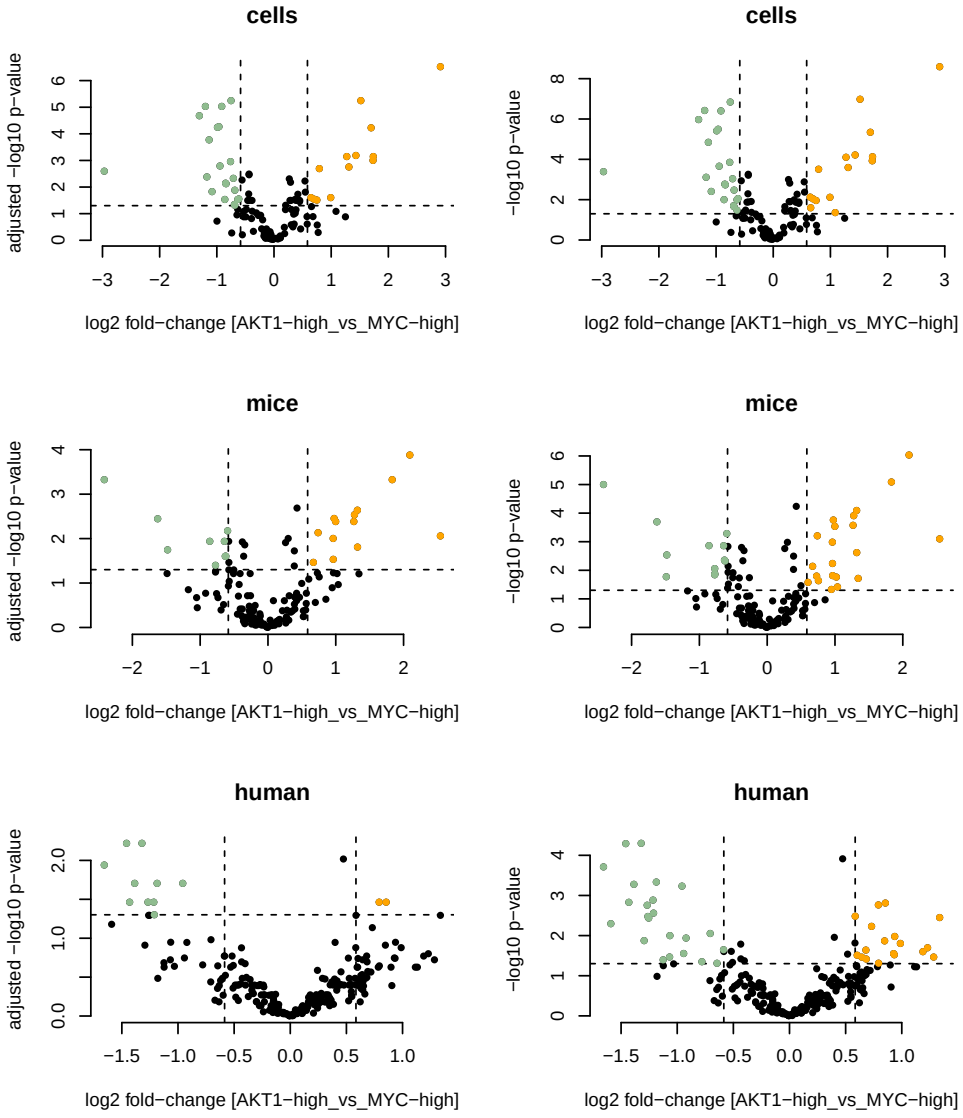
To explore group-wide metabolomic differences within the data sets in an unsupervised fashion, principal component analysis (PCA) was performed on the imputed and normalized metabolomic data. The PCA revealed a very similar grouping as already observed in the hierarchical clustering. Within the cell line data, all three group are show a distinct seperation by principal component 1 (PC1) and principal component 2 (PC2). In mice, PC1 mainly seperated AKT1-high vs. Control and MYC-high samples, whereas in human tissues no clear seperation was present. Hence, at least in the human tissue samples, AKT1 or MYC overexpression did not result in a global metabolomic change. The same conclusions were drawn by Priolo et al.



Supplementary Figure 3 Plot of the first two principal components for the three data sets including the percent of variance that is explained by PC1 and PC2. The points are color-coded according to the grouping.

2.4 Supervised analysis

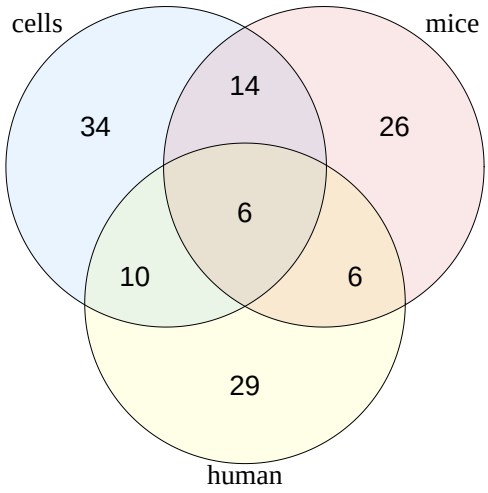
In their supervised analysis, Priolo et al. directly compared the metabolic phenotype of AKT1- and MYC-driven tumors across data sets (cells, mice, human). With multiple testing correction, 46% of metabolites were different in cells, 23% in mice and only 6% of metabolites in human cancer samples. These numbers are in line with the global differences observed in the unsupervised analysis. Supplemental figure 4 depicts the respective comparisons in form of volcano plots.



Supplementary Figure 4 Volcano plot of comparative analyis between AKT1-high and MYC-high tumors. Significantly different metabolites ($p<0.05$ and absolute non-log fold-change >1.5) are color coded. A higher abundance in AKT1-high tumors is encoded by an orange color. The left volcano plots of the respective comparison are adjusted for multiple testing.

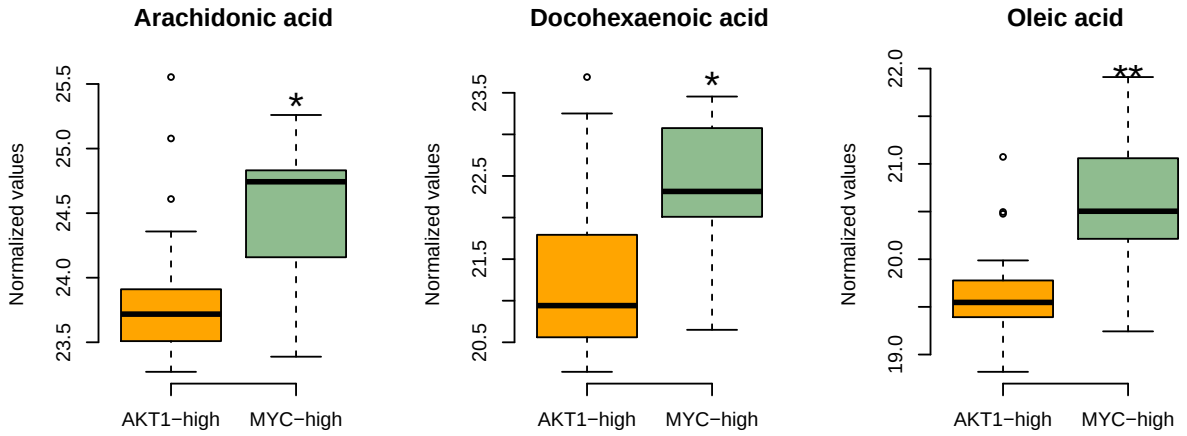
To our knowledge, Priolo and colleagues did not perform multiple testing correction in their enrichment analysis, hence, the following analysis will be conducted with the raw p-values.

The venn diagram of the significantly different metabolites between AKT1- and MYC-driven tumors in the three datasets (Suppl. Fig. 5) shows that only 6 metabolites differed in all data sets, including the top metabolites docosahexaenoic acid and arachidonic acid presented by Priolo et al. in Figure 3B of their manuscript. The third metabolite presented in Figure 3B, oleic acid, was only found to be common between the cell and the human data set, but not the mice data set.



Supplementary Figure 5 Venn diagram of metabolites differing significantly (without multiple testing correction) between AKT1- and MYC-driven tumors for the three different data sets (cells, mice, human).

To assess comparability of our data preprocessing steps with the steps of Priolo and coworkers, we sucessfully reproduced figure 3B of the manuscript (Suppl. Fig. 6).



Supplementary Figure 6 Normalized measurements of arachidonic acid, docosahexaenoic acid, and oleic acid in AKT1-high vs. MYC-high tumor samples as presented in Priolo et al. (Figure 3B).

2.5 Metabolic correlation networks

After identifying individual metabolites that were significantly associated with a sample trait, we aimed to obtain meaningful (sub)pathway-wide changes. Priolo et al. performed single-sample Gene Set Enrichment Analysis (GSEA) on 50 metabolite sets (KEGG annotation) with the majority of metabolites containing only 2-5 metabolites per set (Priolo et al., Supplemental Dataset 1). In Figure 3A of their manuscript, the authors present a table summarizing the GSEA, however, no significant measure was associated with the presented enrichment.

Supplemental Figure 7 depicts the metabolic correlation networks generated form the three data sets. Reflecting the different number of metabolites the networks for cells, mice and human contain 12, 13 and 21 correlation modules, respectively. In line with the analysis carried out before, the difference between the MYC- and AKT1-phenotype was most pronounced in cells (5 modules), followed by mice (3 modules) and human (2 modules) as derived from the number of significant module enrichments (red coloring).

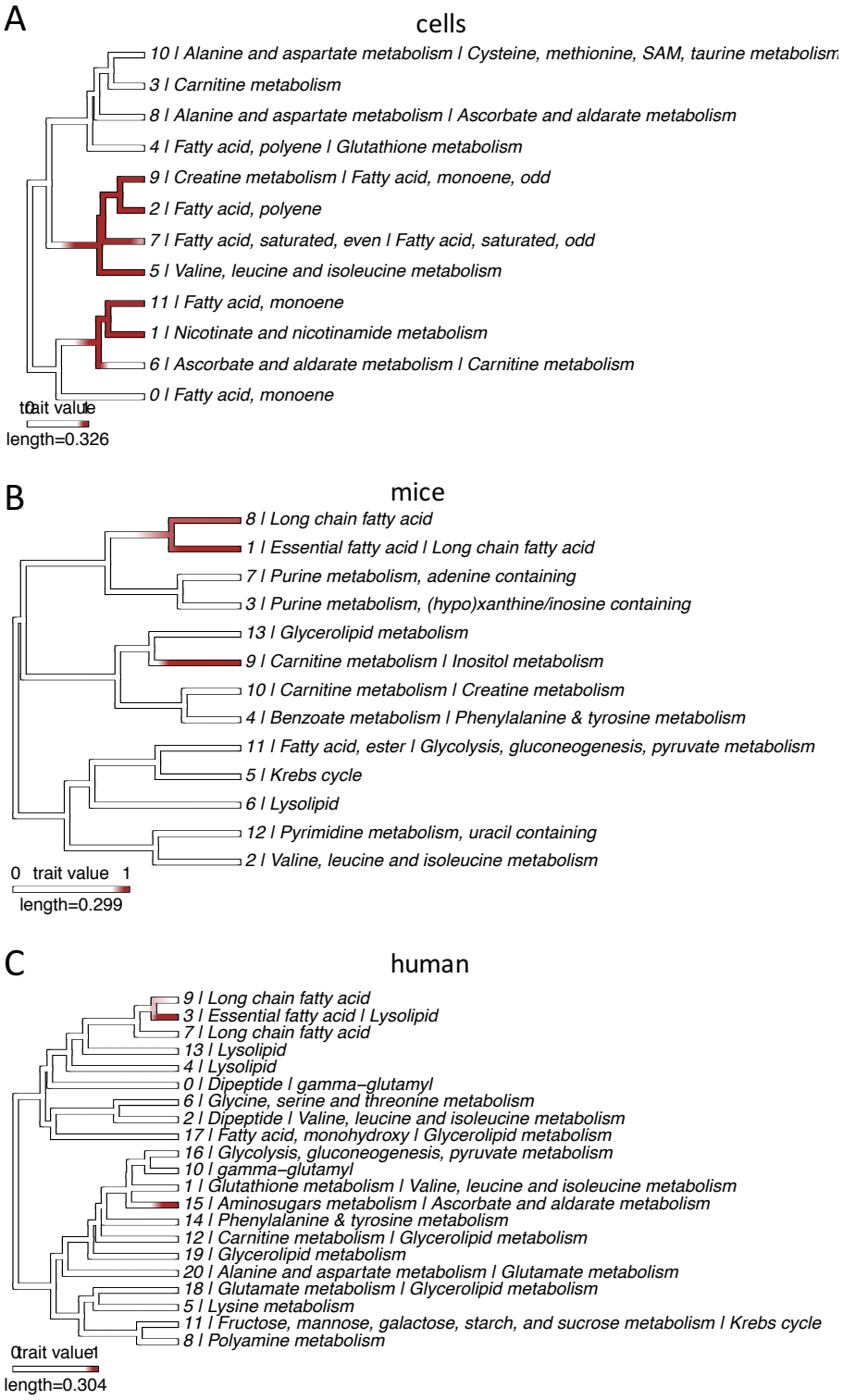
As revealed by the Venn diagram (Suppl. Fig. 5), there was only a partial overlap in significant metabolites between the three data sets. In line, there was only one module with a module significance across the three data sets. The module was associated with the fatty acid metabolism (module 2 in cells, module 1 in mice and module 3 in human). Throughout the data sets, the metabolites in this module showed a higher abundance in MYC-driven samples.

The validity and biological significance of the module enrichment is underpinned by the exploration of this fatty acid module (module 2 in cells, module 1 in mice and module 3 in human; Suppl. Fig 8). Intriguingly, the metabolites most closely related to the respective module (i.e. high module membership) were omega-3 and omega-6 fatty acids and all showed a higher abundance in MYC-driven tumors. Unsaturated fatty acids could be shown to be a prime energy source during early tumorigenesis (via fatty acid oxidation)[9].

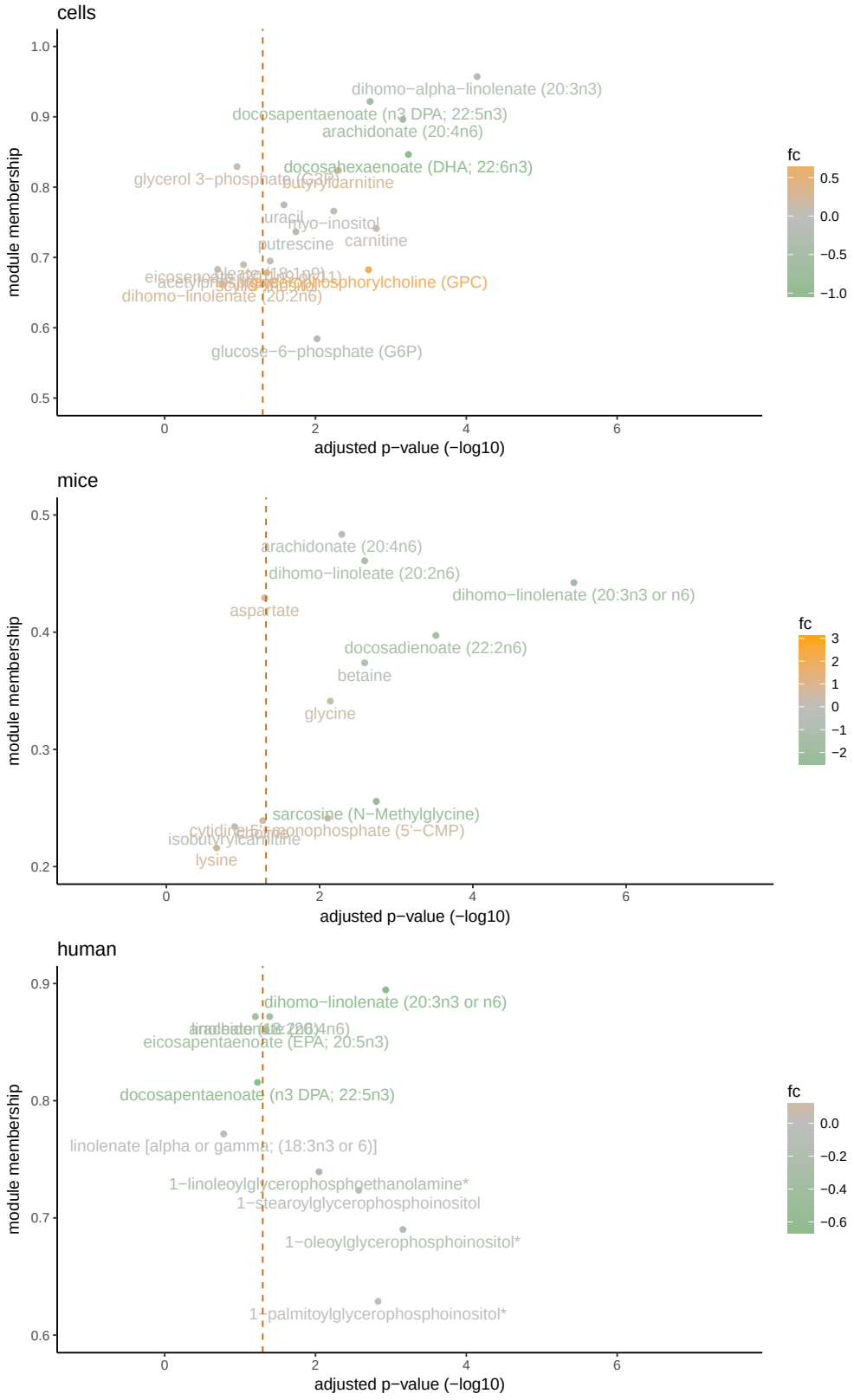
The importance of this finding could recently been confirmed by demonstrating that MYC-overexpressing tripple-negative breast cancer display an increased bioenergetic reliance on fatty acid oxidation and that inhibition of fatty acid oxidation is a potential therapeutic strategy[10].

[9] Carracedo, A., Cantley, L. C., & Pandolfi, P. P. (2013). Cancer metabolism: fatty acid oxidation in the limelight. Nature Reviews. Cancer, 13(4), 227–232. http://doi.org/10.1038/nrc3483

[10] Camarda, R., Zhou, A. Y., Kohnz, R. A., Balakrishnan, S., Mahieu, C., Anderton, B., et al. (2016). Inhibition of fatty acid oxidation as a therapy for MYC-overexpressing triple-negative breast cancer. Nature Medicine, 22(4), 427–432. http://doi.org/10.1038/nm.4055



Supplementary Figure 7 Metabolic correlation networks for the three data sets. Module significance for the comparison AKT1-high vs. MYC-high is color-coded ($p<0.1$; red).



Supplementary Figure 8 Exploration of metabolic correlation modules. Module membership is plotted over the adjusted p-values. Cells: module 2. Mice: module 1. Human module 3.