Application of State-of-the-Art Machine Learning Techniques to PM Data on Autism-Spectrum Disorders

Boosting with False Discovery Control

Benjamin Hofner
benjamin.hofner@fau.de

Department of Medical Informatics, Biometry and Epidemiology
Friedrich-Alexander-Universität Erlangen-Nürnberg

Conference on Phenotype MicroArray Analysis of Cells
Florence - 2015

in cooperation with
Markus Göker, DSMZ, Germany
Luigi Boccuto, GCC, USA

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Identification of Biomarkers for ASD patients (yes/no)

Autism Spectrum Disorders (ASD):

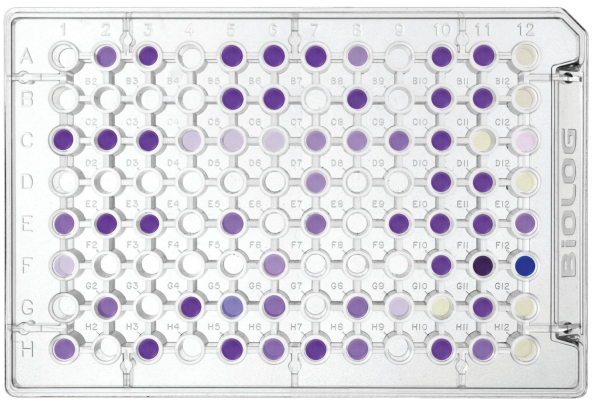
- relatively common neurodevelopmental disease
- biological basis incompletely determined
- no laboratory test for these conditions
- (relatively) hard to diagnose

Aim:

Detect differentially expressed amino acid pathways, i.e., amino acid pathways that differ between healthy subjects and ASD patients.

Identification of Biomarkers for ASD patients (yes/no)

- **Available Data:**
  - Cell lines of $n = 35$ subjects (17 ASD patients and 18 controls)
- **Data Source:**
  - Boccuto et al. [2013, Appendix]
  - Also available as `boccuto_et_al` in the R package **opm** [Vaas et al. 2013, Göker 2015]
- **Measurements:**
  - Phenotype MicroArrays (PM)
    - 96-well array per patient
    - Each well has a different carbon energy source
    - Maximum reaction (= cellular activity) per well is measured (by a color reaction)



(Source: Biolog Inc., http://www.biolog.com)

- Measurements describe metabolism of subjects (on cell basis)

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Model Differential expressions using interactions

$$\log(y) = \beta_0 + \beta_1 \text{group} + b_{\text{id}} + \beta_{2,1} I_{\text{P1}} + \beta_{2,2} I_{\text{P2}} + \ldots + \\ + X(\text{group}) \cdot \widetilde{b}_{\text{id}} + X(\text{group}) \cdot \beta_{3,1} I_{\text{P1}} + X(\text{group}) \cdot \beta_{3,2} I_{\text{P2}} + \ldots$$

with

- log of PM measurements (one value per person and well)

and

- overall group effect (ASD/healthy)
- random effects (as we have repeated measurements)
- (overall) pathways effects
- **group-specific pathway effects**

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Model Differential expressions using interactions

$$\log(y) = \beta_0 + \beta_1 \text{group} + b_{\text{id}} + \beta_{2,1} I_{\text{P1}} + \beta_{2,2} I_{\text{P2}} + \ldots +$$
$$+ X(\text{group}) \cdot \widetilde{b}_{\text{id}} + X(\text{group}) \cdot \beta_{3,1} I_{\text{P1}} + X(\text{group}) \cdot \beta_{3,2} I_{\text{P2}} + \ldots$$

with

- log of PM measurements (one value per person and well)

and

- overall group effect (ASD/healthy)
- random effects (as we have repeated measurements)
- (overall) pathways effects
- **group-specific pathway effects**

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Model Differential expressions using interactions

$$\log(y) = \beta_0 + \beta_1 \text{group} + b_{\text{id}} + \beta_{2,1} I_{\text{P1}} + \beta_{2,2} I_{\text{P2}} + \ldots +$$
$$+ X(\text{group}) \cdot \widetilde{b}_{\text{id}} + X(\text{group}) \cdot \beta_{3,1} I_{\text{P1}} + X(\text{group}) \cdot \beta_{3,2} I_{\text{P2}} + \ldots$$

with

- log of PM measurements (one value per person and well)

and

- overall group effect (ASD/healthy)
- random effects (as we have repeated measurements)
- (overall) pathways effects
- **group-specific pathway effects**

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Model Differential expressions using interactions

$$\log(y) = \beta_0 + \beta_1 \text{group} + b_{\text{id}} + \beta_{2,1} I_{\text{P1}} + \beta_{2,2} I_{\text{P2}} + \ldots +$$
$$+ X(\text{group}) \cdot \widetilde{b}_{\text{id}} + X(\text{group}) \cdot \beta_{3,1} I_{\text{P1}} + X(\text{group}) \cdot \beta_{3,2} I_{\text{P2}} + \ldots$$

with

- log of PM measurements (one value per person and well)

and

- overall group effect (ASD/healthy)
- random effects (as we have repeated measurements)
- (overall) pathways effects
- **group-specific pathway effects**

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Model Differential expressions using interactions

$$\log(y) = \beta_0 + \beta_1 \text{group} + b_{\text{id}} + \beta_{2,1} I_{\text{P1}} + \beta_{2,2} I_{\text{P2}} + \ldots +$$
$$+ X(\text{group}) \cdot \widetilde{b}_{\text{id}} + X(\text{group}) \cdot \beta_{3,1} I_{\text{P1}} + X(\text{group}) \cdot \beta_{3,2} I_{\text{P2}} + \ldots$$

with

- log of PM measurements (one value per person and well)

and

- overall group effect (ASD/healthy)
- random effects (as we have repeated measurements)
- (overall) pathways effects
- **group-specific pathway effects**

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Model Differential expressions using interactions

$$\log(y) = \beta_0 + \beta_1 \text{group} + b_{\text{id}} + \beta_{2,1} I_{\text{P1}} + \beta_{2,2} I_{\text{P2}} + \ldots +$$
$$+ X(\text{group}) \cdot \widetilde{b}_{\text{id}} + X(\text{group}) \cdot \beta_{3,1} I_{\text{P1}} + X(\text{group}) \cdot \beta_{3,2} I_{\text{P2}} + \ldots$$

with

- log of PM measurements (one value per person and well)

and

- overall group effect (ASD/healthy)
- random effects (as we have repeated measurements)
- (overall) pathways effects
- **group-specific pathway effects**

Modeling Strategy

- Fit a log-linear model for the PM measurements
- Model Differential expressions using interactions

$$\log(y) = \beta_0 + \beta_1 \text{group} + b_{\text{id}} + \beta_{2,1} I_{\text{P1}} + \beta_{2,2} I_{\text{P2}} + \ldots + \\ + X(\text{group}) \cdot \widetilde{b}_{\text{id}} + X(\text{group}) \cdot \beta_{3,1} I_{\text{P1}} + X(\text{group}) \cdot \beta_{3,2} I_{\text{P2}} + \ldots$$

with

- log of PM measurements (one value per person and well)

and

- overall group effect (ASD/healthy)
- random effects (as we have repeated measurements)
- (overall) pathways effects
- **group-specific pathway effects**

- Use **boosting** methods with **stability selection** for variable selection

Boosting . . .

. . . is a versatile tool to estimate models with built-in variable selection (▶ similar to lasso, etc.)

Boosting . . .

. . . is a versatile tool to estimate models with built-in variable selection (▶ similar to lasso, etc.)

In short:

- The fitting process is iterative.
- It optimizes for example the (negative) log-likelihood.
- In each step of the algorithm only the effect $\hat{\beta}_{j*}$ of the best fitting variable is updated by adding a fraction $\nu \cdot \hat{\beta}_{j*}$ to the model (with e.g., $\nu = 0.1$).
- The main tuning parameter is the number if iterations.

Boosting . . .

. . . is a versatile tool to estimate models with built-in variable selection (▶ similar to lasso, etc.)

In short:

- The fitting process is iterative.
- It optimizes for example the (negative) log-likelihood.
- In each step of the algorithm only the effect $\hat{\beta}_{j*}$ of the best fitting variable is updated by adding a fraction $\nu \cdot \hat{\beta}_{j*}$ to the model (with e.g., $\nu = 0.1$).
- The main tuning parameter is the number if iterations.

- ▶ Use cross-validation to find the optimal stopping iteration.
- ▶ With "early stopping" variable selection is achieved.

Implementation

- A very flexible implementation of boosting is available in the *R* package **mboost** [Hothorn et al. 2015]
- A hands-on tutorial is given in [Hofner et al. 2014]

Practical notes

- We get an **interpretable model**, similar to models from maximum likelihood estimation or least squares estimation.
- Additionally, regularization is achieved via **variable selection** and shrinkage of effects.
- This makes boosting viable in situation with many variables and also with highly correlated variables.

Practical notes

- We get an **interpretable model**, similar to models from maximum likelihood estimation or least squares estimation.
- Additionally, regularization is achieved via **variable selection** and shrinkage of effects.
- This makes boosting viable in situation with many variables and also with highly correlated variables.

Yet,

- in high-dimensional settings, i.e., with many predictors, we might erronously select a lot of uninformative variables.
- In many situations a **formal selection procedure with error control** seems advisable.

- Use **stability selection**

Stability Selection

[Meinshausen and Bühlmann 2010]

... is a versatile approach, which can be combined with (all) high-dimensional variable selection approaches.

... is based on subsampling:

- draw simple random samples of the observations
- each sample contains 50% of the original data
- derive impact of variation in the data on the results

... controls the per-family error rate

$$\text{PFER} = \mathbb{E}(V),$$

where V is the number of false positives.

Insertion: Overview of Error Rates

see e.g. Dudoit et al. [2003]

per-family error rate (PFER): $\mathbb{E}(V)$

per-comparison error rate (PCER): $\mathbb{E}(V)/m$

standard testing procedure, no multiplicity correction

family-wise error rate (FWER): $\mathbb{P}(V \geq 1)$

false discovery rate (FDR): $\mathbb{E}(\frac{V}{R})$

	Keep H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	T	S	$m - m_0$
	$m - R$	R	m

Insertion: Overview of Error Rates

see e.g. Dudoit et al. [2003]

per-family error rate (PFER): $\mathbb{E}(V)$

per-comparison error rate (PCER): $\mathbb{E}(V)/m$

standard testing procedure, no multiplicity correction

family-wise error rate (FWER): $\mathbb{P}(V \geq 1)$

false discovery rate (FDR): $\mathbb{E}(\frac{V}{R})$

	Keep H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	T	S	$m - m_0$
	$m - R$	R	m

Note: The PFER is the most conservative error control.

Stability Selection

Algorithm (simplified)

1. Select a random subset of size $\lfloor n/2 \rfloor$ of the data.
2. Fit boosting model until q variables are selected (out of p).
3. Record which variables were selected.
4. Repeat $B = 100$ times.

5. Compute selection frequency per variable.
6. Select variables with frequency $\geq \pi_{\text{thr}}$.

Stability Selection

Algorithm (simplified)

1. Select a random subset of size $\lfloor n/2 \rfloor$ of the data.
2. Fit boosting model until q variables are selected (out of p).
3. Record which variables were selected.
4. Repeat $B = 100$ times.

5. Compute selection frequency per variable.
6. Select variables with frequency $\geq \pi_{\text{thr}}$.

- Conservative upper bound for the per-family error rate (PFER):

$$\text{PFER} = \mathbb{E}(V) \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

(if exchangeability assumption holds for all noise variables)

Improved Stability Selection

[Shah and Samworth 2013]

- Tighter, i.e., less conservative error bounds can be derived under certain conditions.

  a) If distribution of (simultaneous) selection probabilities is unimodal:

$$\mathbb{E}(V) \leq \frac{q^2}{c(\pi_{\text{thr}}, B) \cdot p} \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

  b) If distribution of (simultaneous) selection probabilities is r-concave:

$$\begin{aligned} \mathbb{E}(V) &\leq \min\left\{ D\left(2\pi_{\text{thr}} - 1; \frac{q^2}{p^2}, B, -\frac{1}{2}\right), D\left(\pi_{\text{thr}}; \frac{q}{p}, 2B, -\frac{1}{4}\right)\right\} p \\ &\leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p} \end{aligned}$$

Implementation

- Stability selection is implemented in the *R* package **stabs** [Hofner and Hothorn 2015] in the function `stabsel()`
- `stabsel()` can be used on fitted boosting models.
- `stabsel()` can also be used with arbitrary variable selection approaches (some are already implemented, others can be easily added).

Implementation

- Stability selection is implemented in the *R* package **stabs** [Hofner and Hothorn 2015] in the function `stabsel()`
- `stabsel()` can be used on fitted boosting models.
- `stabsel()` can also be used with arbitrary variable selection approaches (some are already implemented, others can be easily added).

Practical recommendation:

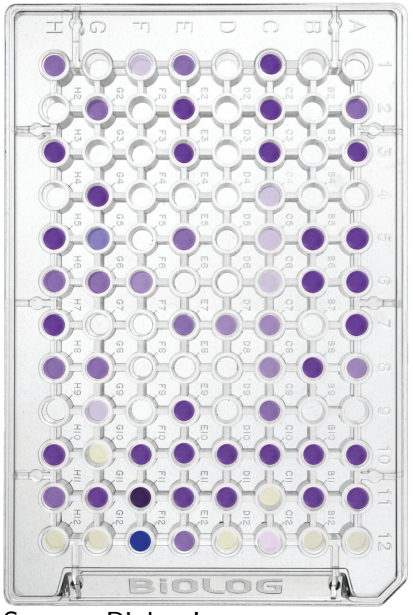
- Choose an upper bound for the PFER and specify *either* q or π_{thr}.
- Check that the computed value is sensible
  (e.g., that is q large enough if π_{thr} and PFER were specified).

Identification of Biomarkers for ASD patients (yes/no)

1) Obtain amino acid pathway annotation for each well
   using the R package **opm** [Vaas et al. 2013]
2) Fit main effects model for maximum reaction (`y`),
   given disease status (`group`), pathway annotation (`pathway`), and patient (`id`)

```
log(y) ~ group + pathway + (1 | id)
```

- Each well constitutes one observation!
- Each well can belong to multiple pathways!



Source: Biolog Inc.

3) Use model 2) as offset model and add group specific effects:

```
log(y) ~ ... + (group | id) + pathway:group
```

4) **Which of the group-specific pathway effects is selected additionally to the offset model (with PFER ≤ 1)?**

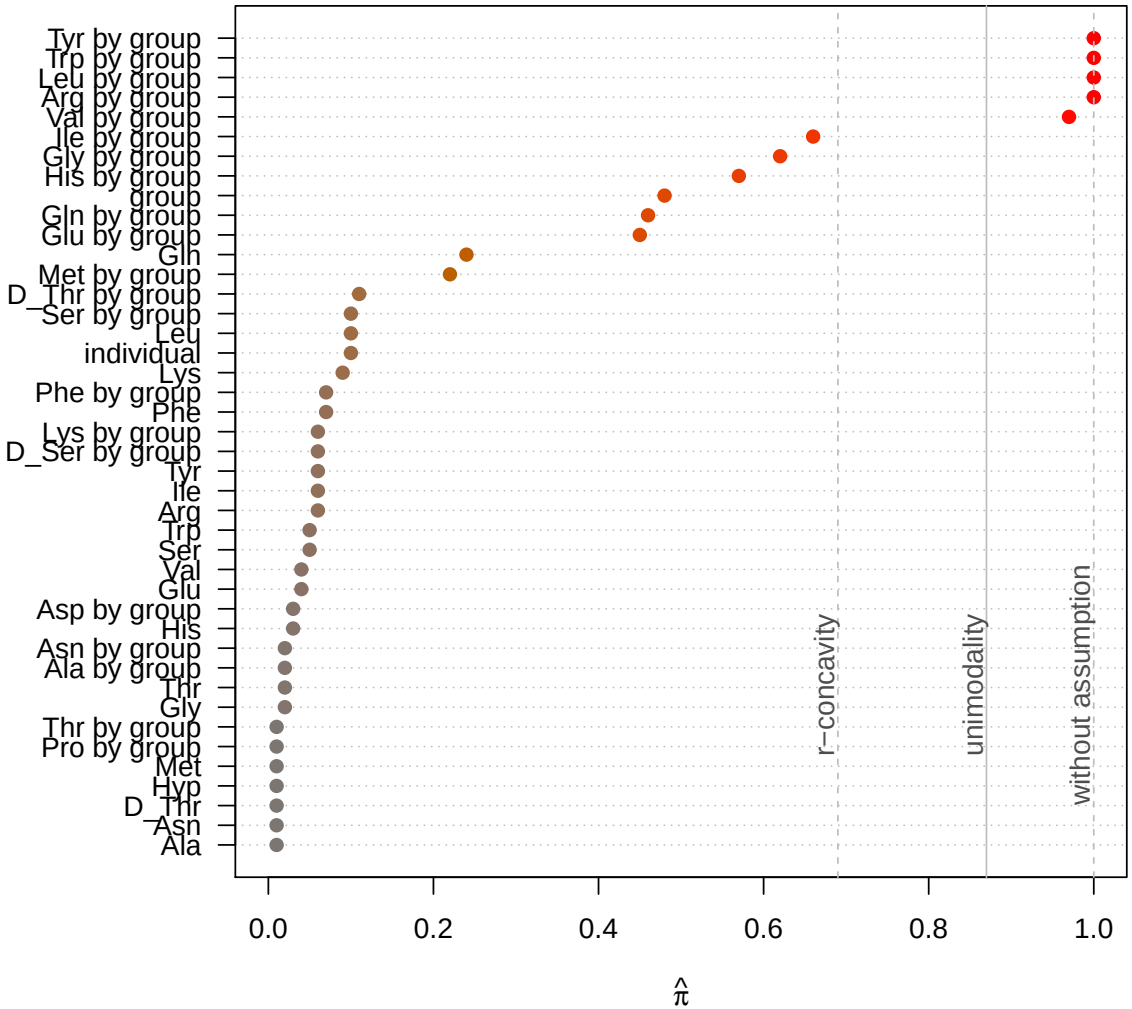
Code and Details

- **Code:** See Additional File 2 on www.biomedcentral.com/1471-2105/16/144/additional
- **Details:** See Hofner et al. [2015]

Results: Biomarkers for ASD

Stability selection with PFER ≤ 1 and $q = 10$

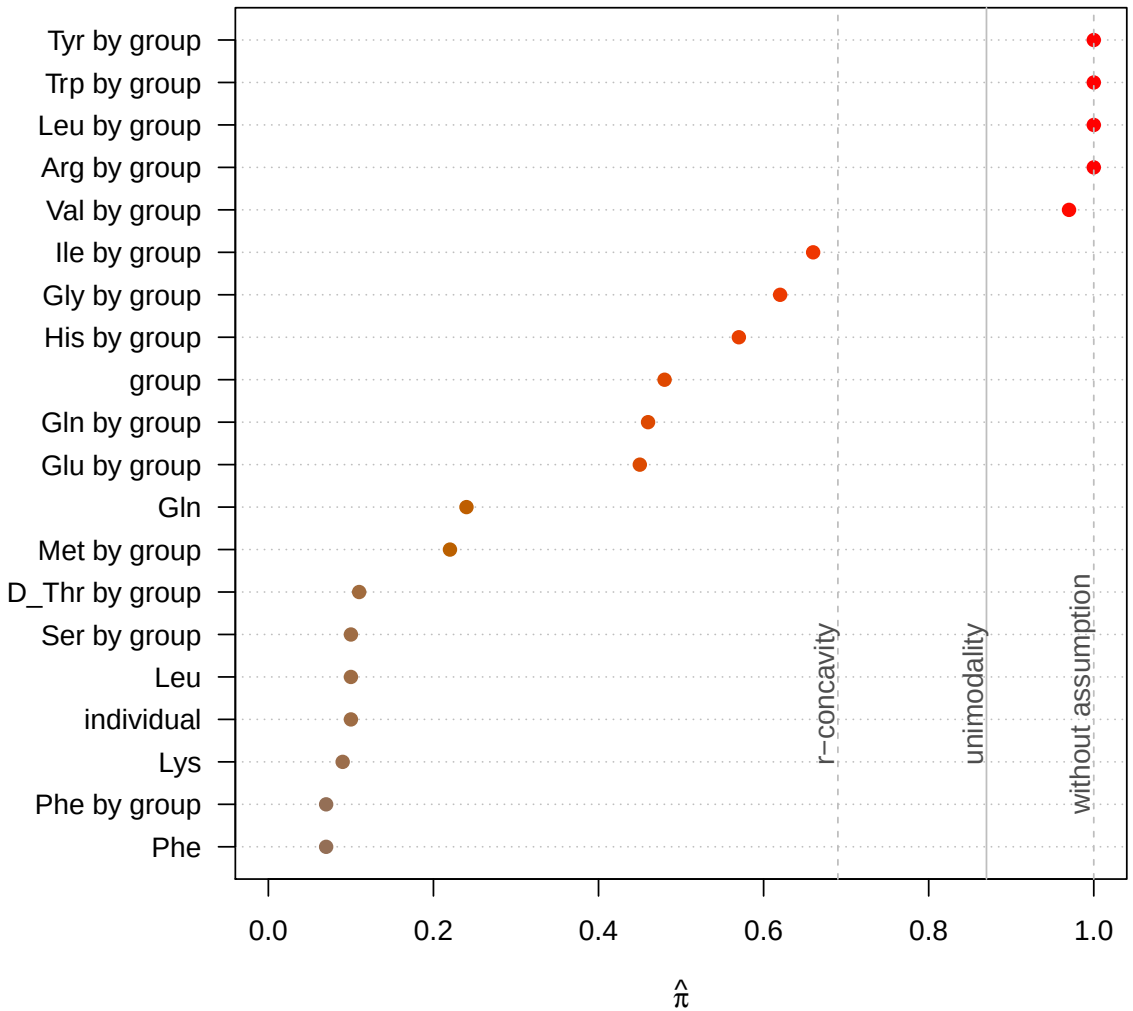
Selection Frequency



Results: Biomarkers for ASD

Stability selection with PFER ≤ 1 and $q = 10$

Selection Frequency (TOP 20)



Results: Biomarkers for ASD

Differentially Expressed Amino Acids

- tyrosine (`Tyr`), tryptophan (`Trp`), leucine (`Leu`), arginine (`Arg`)
  - selection frequency $\hat{\pi} = 100\%$
- valine (`Val`)
  - selection frequency $\hat{\pi} = 96\%$
- glycine (`Gly`)
  - selection frequency $\hat{\pi} = 71\%$

Results: Biomarkers for ASD

Differentially Expressed Amino Acids

- tyrosine (`Tyr`), tryptophan (`Trp`), leucine (`Leu`), arginine (`Arg`)
  - selection frequency $\hat{\pi} = 100\%$
- valine (`Val`)
  - selection frequency $\hat{\pi} = 96\%$
- glycine (`Gly`)
  - selection frequency $\hat{\pi} = 71\%$

Biomedical Conclusion

- Confirms abnormal metabolism of tryptophan in ASD cells [see Boccuto et al. 2013]
- \+ Additional amino acids seem to be affected, although on a milder level
- Suggest an abnormal metabolism of large amino acids

Summary and Outlook

- Stability selection works well in conjunction with boosting.
- It controls the PFER and is especially useful in sparse, high-dimensional settings.
- Boosting with stability selection can produce novel insights into PM data by finding differentially expressed amino acid pathways.

- Yet, stability selection is quite conservative
  - as it controls the PFER
  - and as even this control seems to be conservative (at least for the standard error bound).
- Higher selection numbers (i.e. higher TPR) can be obtained by tighter error bounds, yet, sometimes error bounds do not hold any more.

- Details and a simulation study can be found in Hofner et al. [2015]

Slides and further information available from
`http://benjaminhofner.de`

References I

L Boccuto, C-F Chen, A Pittman, C Skinner, H McCartney, K Jones, B Bochner, R Stevenson, and C Schwartz. Decreased tryptophan metabolism in patients with autism spectrum disorders. *Molecular Autism*, 4(1):16, 2013.

S Dudoit, J Popper Shaffer, and JC Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.

M Göker with contributions by B Hofner, LAI Vaas, J Sikorski, N Buddruhs and A Fiebig. *opm: Analysing Phenotype Microarray and Growth Curve Data*, 2015. URL http://opm.dsmz.de. R package version 1.2.14.

B Hofner and T Hothorn. *stabs: Stability Selection with Error Control*, 2015. URL http://CRAN.R-project.org/package=stabs. R package version 0.5-1.

B Hofner, A Mayr, N Robinzonov, and M Schmid. Model-based boosting in R – A hands-on tutorial using the R package mboost. *Computational Statistics*, 29:3–35, 2014.

B Hofner, L Boccuto, and M Göker. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, 16:144, 2015. URL http://dx.doi.org/10.1186/s12859-015-0575-3.

T Hothorn, P Bühlmann, T Kneib, M Schmid, and B Hofner. *mboost: Model-Based Boosting*, 2015. URL http://CRAN.R-project.org/package=mboost. R package version 2.5-0.

N Meinshausen and P Bühlmann. Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:417–473, 2010.

References II

RD Shah and RJ Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:55–80, 2013.

LAI Vaas, J Sikorski, B Hofner, N Buddruhs, A Fiebig, H-P Klenk, and M Göker. opm: An R package for analysing OmniLog®phenotype microarray data. *Bioinformatics*, 29(14):1823–1824, 2013.