Comparison of different variable selection strategies to formulate predictive models in medicine

Anita Windhorst & Jörn Pons-Kühnemann

Institute of Medical Informatics, Medical Statistics,
Justus-Liebig-University, Giessen, Germany

Background & Aim

Aim: Compare transcript selections strategies PAM and sPlSDA in regard to class separation and associated biological functions in order to understand mechanisms of mild Bronchopulmonary dysplasia (BPD) better.

But:

Transcriptome analysis:

- High number of differentially regulated transcripts
- Highly correlated transcripts

Clinical data:

- Correlated diagnostic markers

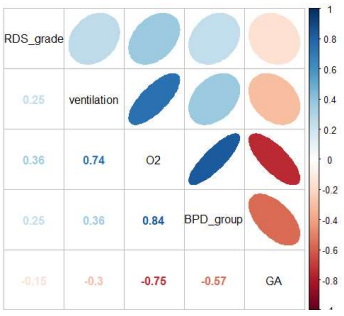


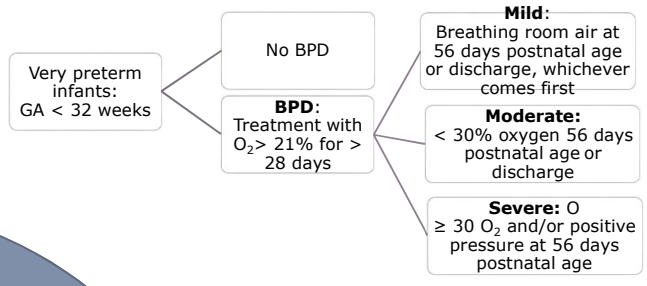
Fig (above): High association between BPD grades and Gestational age and RDS grades. The duration of ventilation and oxygenation is used to diagnose BPD

Best variable selection strategy for predictive models ?

Patients & Material

Source of photo: http://www.breathecalifornia.org/images/health-img/PrematureInfant.jpg

Umbilical blood is taken at birth of the preterm infants and analyzed using Codelink Human I10k Bioarrays, gene expression was compared in regard to grade of BPD. As incidences of moderate and severe were low, these grades were analyzed as one group.



PAM
Predictive Analysis of Microarrays

PAM is a method based on nearest shrunken centroids and so identifies subsets of transcripts that best describe each BPD group.

Implementation in R: pamr

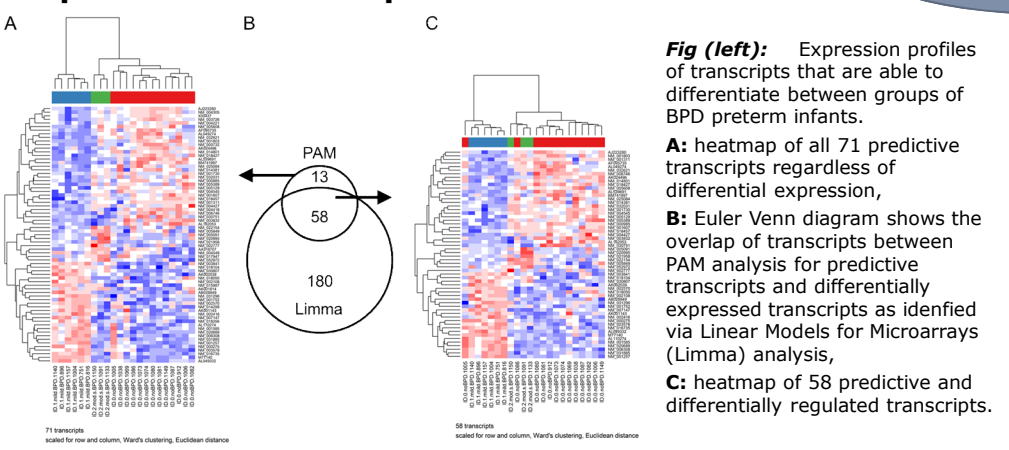


Fig (left): Expression profiles of transcripts that are able to differentiate between groups of BPD preterm infants.

A: heatmap of all 71 predictive transcripts regardless of differential expression,

B: Euler Venn diagram shows the overlap of transcripts between PAM analysis for predictive transcripts and differentially expressed transcripts as idenfied via Linear Models for Microarrays (Limma) analysis,

C: heatmap of 58 predictive and differentially regulated transcripts.

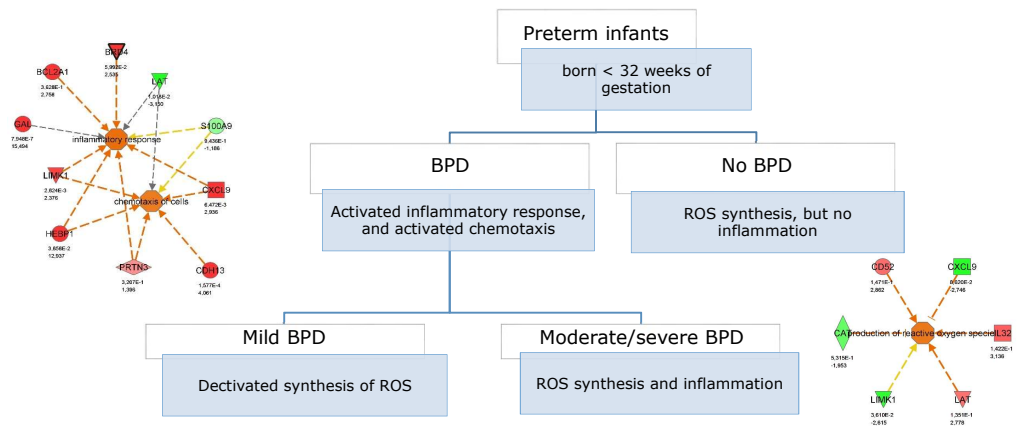


Fig (above): In BPD are processes in inflammatory response and chemotaxis activated, but the deactivation of ROS synthesis leads to an only mild form of BPD. Networks show which transcripts are involved in the respective biological process

(Red/ Orange: up-regulation, green: down-regulation, below transcripts, adjusted p-values and fold changes are given)

sPLS
sparse Partial Least Squares Analyses

Sparse partial least squares (sPLS) discriminant and regression analysis are methods to simultaneously reduces dimensions in dependent variables and of the independent variables.

Implementation in R: mixOmics

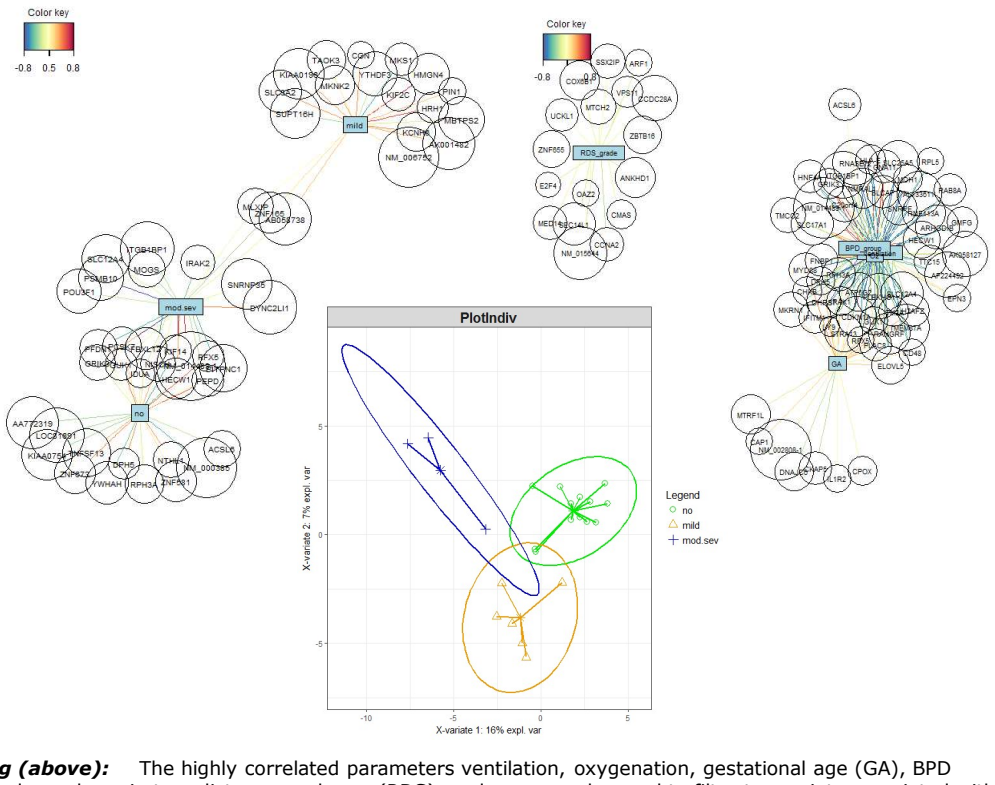


Fig (above): The highly correlated parameters ventilation, oxygenation, gestational age (GA), BPD grade, and respiratory distress syndrome (RDS) grade now can be used to filter transcripts associated with one or more of the clinical variables.
Networks show the associated transcripts with different BPD groups (left) as a result of sPLSDA, where differentiation between classes works very well (middle), and association between gene expression and ventilation, oxygenation, GA, and RDS, as well as BPD (right).

Outlook

- Comparison of downstream processes and upstream regulators
- Extending the range of transcript selection strategies
- Test the selection strategies for other conditions and diseases

Contact information:

Anita.C.Windhorst@informatik.med.uni-giessen.de
Joern.Pons@informatik.med.uni-giessen.de
Institute of Medical Informatics, Medical Statistics
Faculty of Medicine, Justus-Liebig-University
Rudolf-Buchheim-Str. 6, D-35392 Giessen, Germany

References

1. Jobe, Alan H., und Eduardo Bancalari. „Bronchopulmonary Dysplasia". *American Journal of Respiratory and Critical Care Medicine* 163, Nr. 7 (Juni 2001): 1723–29. doi:10.1164/ajrccm.163.7.2011060.
2. T. Hastie, R. Tibshirani, Balasubramanian Narasimhan and Gil Chu (2014). pamr: Pam: prediction analysis for microarrays. R package version 1.55. https://CRAN.R-project.org/package=pamr
3. Kim-Anh Le Cao, Florian Rohart, Ignacio Gonzalez, Sebastien Dejean with key contributors Benoit Gautier, Francois Bartolo, contributions from Pierre Monget, Jeff Coquery, FangZou Yao and Benoit Liquet. (2017). mixOmics: Omics Data Integration Project. R package version 6.1.2. https://CRAN.R-project.org/package=mixOmics