Selection of Stable Biomarker Signature for Prediction of Metabolic Phenotypes

Jelena Čuklina[1,2,3*], Yibo Wu[1], Evan G. Williams[1], María Rodríguez-Martínez[3], Ruedi Aebersold[1,4]

[1]ETH Zurich, Institute of Molecular Systems Biology, CH-8093 Zurich, Switzerland, [2]Ph.D. Program in Systems Biology, University of Zurich and ETH Zurich, CH-8057 Zurich, Switzerland, [3]IBM Zurich Research Laboratory, Rüschlikon, CH-8803, Switzerland, [4]Faculty of Science, University of Zurich, Zurich, CH-8091, Switzerland

Motivation

In biomarker research, the goal is to construct an prediction rule on the basis of a small number of predictors. Formally, this means representing a macro-level response as a function of molecular features (DNA variants, transcript or protein abundancies) with minimal error.

$$= f(\,,\,,\,) + \varepsilon$$

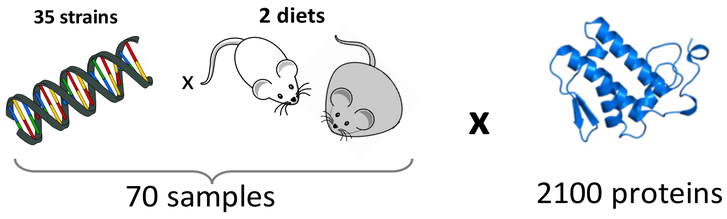
Aim

Develop a framework for selection of a composite biomarker: an ensemble of small number of predictors, that is able to predict the macro-level response.

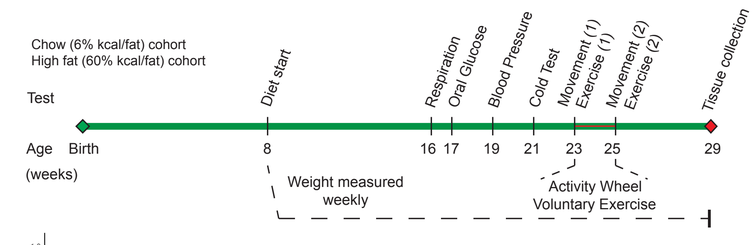
Data

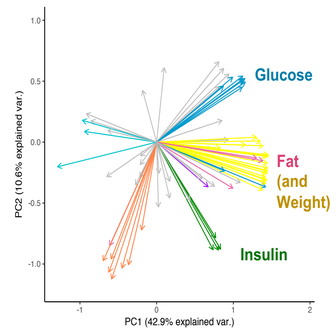
To benchmark the process of construction of the composite biomarker, we use a mouse model. Mouse model has an advantage over human samples, as many confounding factors are controlled. Here we use measurements of 35 murine strains from the BXD recombinant inbred strain panel exposed to high-fat and chow diets.

We use 2100 liver proteins measured with SWATH mass-spectrometry, used as a features to predict metabolic traits.

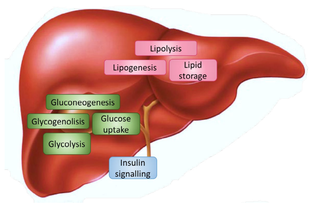


Trait selection



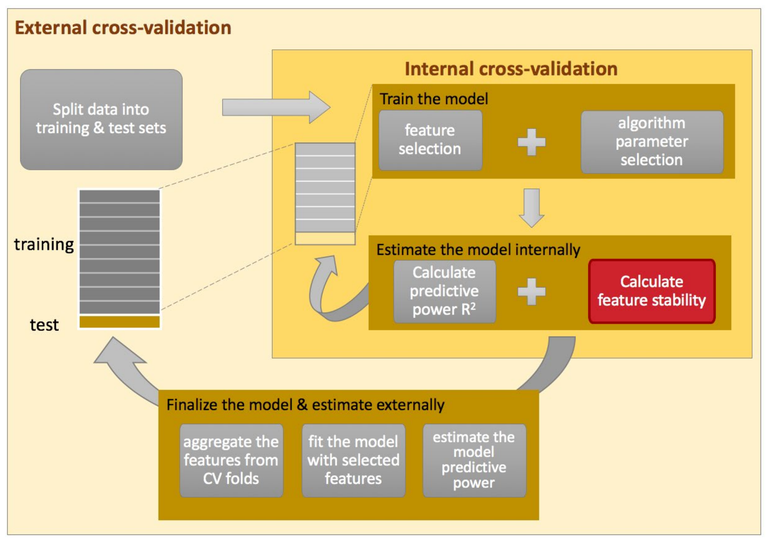


Totally, 115 metabolic traits have been measured during the phenotyping program. We selected 6 traits, representative of each highly correlated group of phenotypes, but also representative of liver physiology.



Trait	Alias
Organ_pWAT_.g.	Fat
CLAMS_Lean_.g.	Lean_mass
ColdTest_5h..grad_C.	Cold_test
OGTT_AUC_Insulin_.AUC.	Insulin
OGTT_AUC_Glucose	Glucose
VO2_Sprint_PostWheel_.m.	Respiration

Pipeline



Biomarker identification procedure can be formulated as a feature selection problem. As omics datasets are inherently multivariate with both independent and multicollinear variables and only a few samples are typically available, this task is very challenging. However, feature selection can also introduce optimistic bias into statistical inference.

To realistically assess algorithm performance, cross-validation approach is commonly used. If data-specific optimization parameters cannot be determined beforehand, internal cross-validation is also required. Each fold of cross-validation delivers a different signature. To be able to translate the combined signature to the clinic, we suggest to assess the stability of the signature identified.

Therefore, to determine which algorithms perform optimally, both feature selection stability and overall performance must be evaluated together.

We benchmark this pipeline on the Random Forest algorithm [1].

Results

It is well known that algorithm parameters affect the performance, however, this aspect is much less explored for the stability studies. We show that the algorithm parameters (ntree and mtry for Random Forest heavily affect the stability of the top features, see **Fig.1**).

The model performance and the stability of the signature heavily depend on the relevance of the molecular profile to the phenotype it aims to predict. For our dataset, biomarker signature derived from liver proteome is stable for fat mass and glucose level, which are traits related to liver metabolism (**Fig.2**).

Algorithm parameter choice

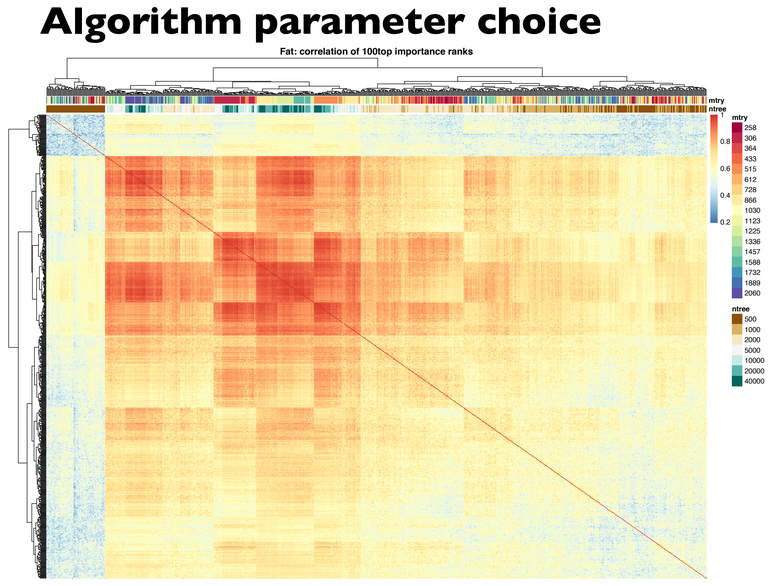
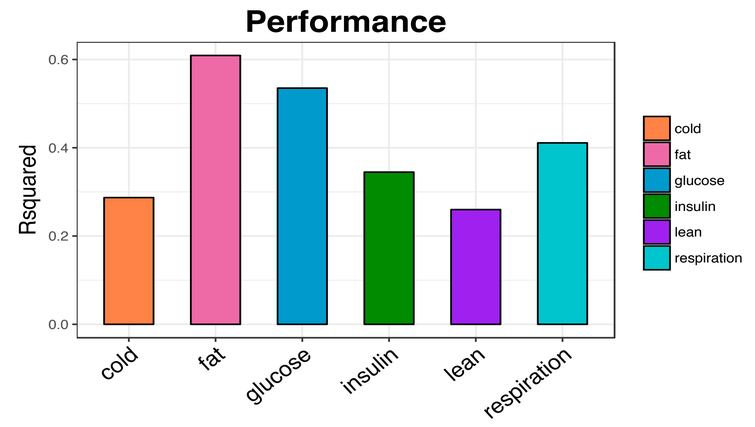
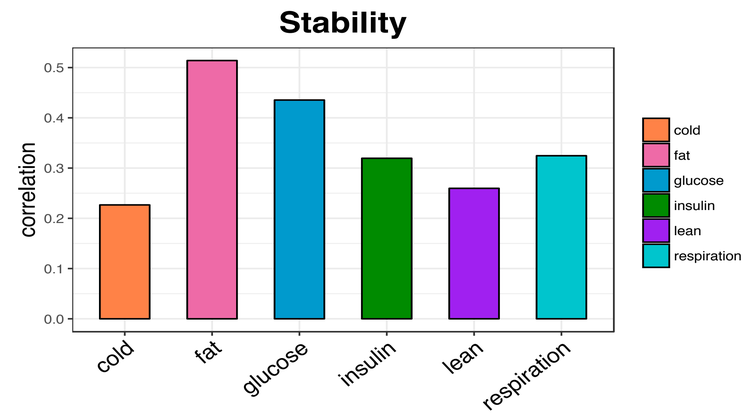


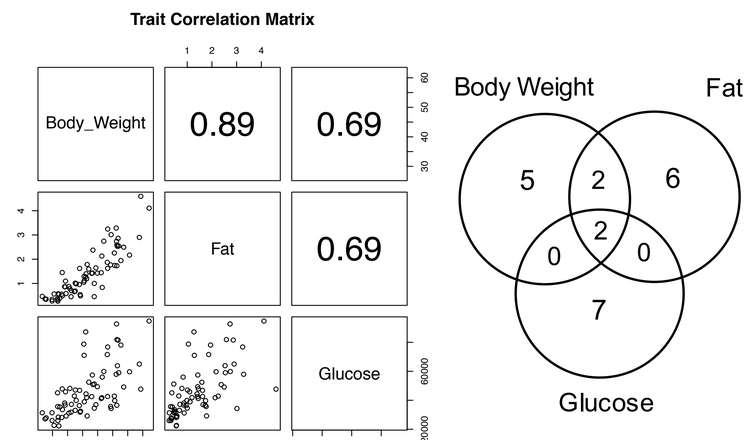
Fig !. Correlation heatmap of variable importance rank for top 100 variables. Each ntree & mtry combination was repeated 10 times, the resulting variable ranks were compared for each setup by correlating ranks. Default forest size (500 trees) produces different importance ranking for each realisation: the resulting ranks don't correlate neither within 500-tree forests, nor with the ranks of bigger forests. However, forests of 5000-40000 of trees yield ranks that are more similar to each other. Thus, to get a reliable importance-ordered list of variables, bigger forest size is necessary. Mtry influence is lower. However, for mtry value, close to default (1/3 of variables, here 1030), a middle-size forest of 5000 trees yields a variable list, very similar to list of 20000-40000 trees.

Algorithm Performance





For highly correlated features, also predictors selected are shared (see Venn diagram)



References

1. Breiman L. Random Forests. Machine Learning.45(1):5-32.
2. Williams EG, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann CA, et al. Systems proteomics of liver mitochondria function. Science. 2016;352(6291):aad0189.

ETH zürich FNSNF NNPrECISE life science zurich IBM Research