Identification of longevity biomarkers in *Nothobranchius furzeri* by Transcriptome Measurement and Random Forest Analysis

U. Menzel[1], S. Priebe[1], M. Baumgart[2], M. Groth[2], A. Cellerino[2], R. Guthke[1]
[1]Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute
[2]Leibniz Institute for Age Research - Fritz-Lipmann Institute

Fin biospies were obtained from 152 individuals of the short-lived teleost fish *Nothobranchius furzeri* (maximum lifespan ~ 60 weeks), at the age of 10 weeks and 20 weeks. The biopsies were taken without sacrificing the fish, so that lifespan data were available for each individual. Transcriptome data have been generated for these samples using RNA-Seq on the Illumina platform. In order to identify genes which are predictive for lifespan, a Random Forest analysis has been made, considering the expression at both time points as well as the change of the expression between the two time points. Based on this analysis, we can conclude that differences in lifespan manifest in gene expression already at young age. On the other hand, the analysis reveals that batch effects hamper the identification of generally applicable biomarkers for the prediction of lifespan.

Introduction

Nothobranchius furzeri is a favorable model organism for age research because it inhabits ephemeral habitats that last on average 75 days, thereby limiting it's natural lifespan to a few months[1]. However, despite of ephemerality, differences in lifespan between individuals can be substantial. In order to identify genes determining lifespan, the transcriptomes of the fishes have been measured at the age of 10 weeks and 20 weeks, respectively, and the lifespan data have been recorded simultaneously. Fishes were subdivided into 3 lifespan groups: short-lived, medium-lived and long-lived, each including 10 individuals, as illustrated in Fig. 1. These group labels were used as response variable in the Random Forest analysis described below.

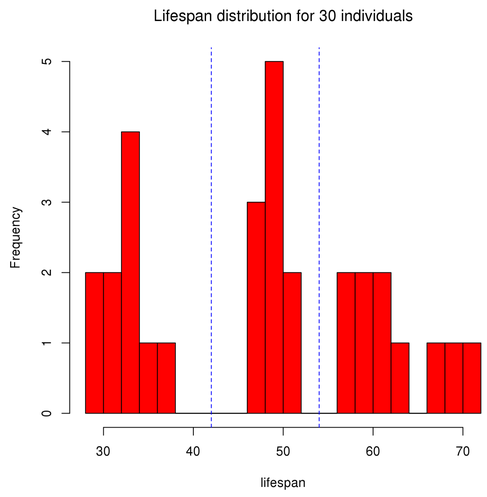


Fig. 1: Histogram of the life-span of 30 individuals of *N. furzeri*. The blue vertical lines separate the short-lived, medium -lived, and long-lived individuals.

Decision Tree

A Decision Tree is a classifier that is favourable when the number of variables (genes) is higher than the number of observations (samples). The tree pinpoints genes that are informative with regard to some attribute, e.g. with regard to age. In the current analysis, we pre-selected 1000 genes most strongly correlated with lifespan. In Fig. 2, a tree is shown which is based on the most informative genes. The tree was constructed by using the RPKM values of these genes as predictors, and the respective age group (short-lived, medium-lived, long-lived) as response.

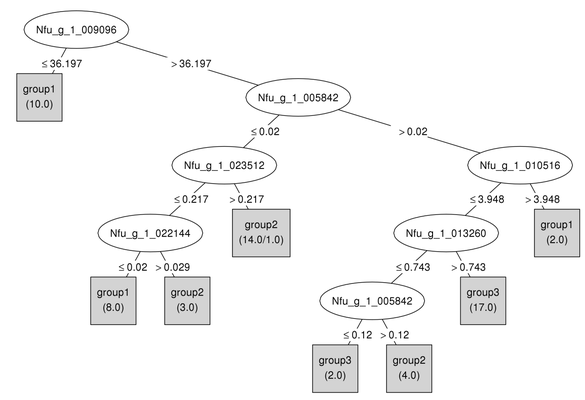
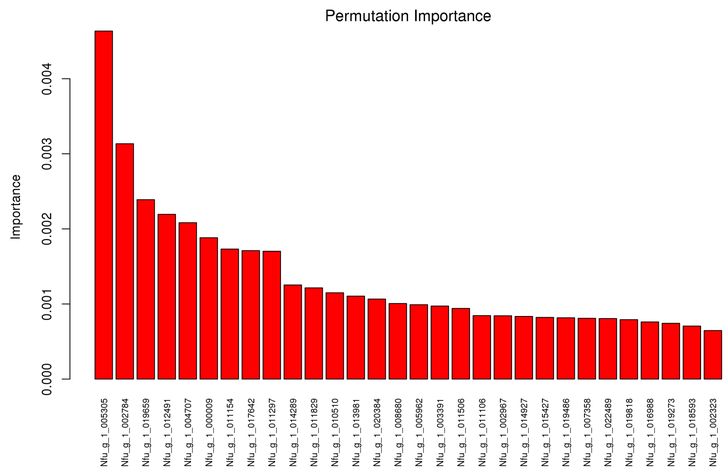


Fig. 2: Decision tree constructed on the basis of the most informative genes.

Nfu_g_1_009096 spna2 spectrin alpha 2
Nfu_g_1_005842 slc13a1 solute carrier family 13
Nfu_g_1_023512 CCDC96 coiled-coil domain

Fig. 3: Biomarkers found using Random Forest

#1: "spectrin alpha 2": scaffold proteins that stabi-lize the plasma membrane
#4 "clathrin" major protein component of the cytoplasmic face of intracellular organelles
#7 "mitochondrial ribosomal protein" ….



Random Forest

A Random Forest (RF) is an ensemble classifier based on many Decision Trees (Breiman, Cutler[2]). The classifier calculates the variable importance (Fig. 3), a measure of the explanatory power of a gene with regard to lifespan. Validation shows that 95% of the samples were correctly classified with the help of the expression values of the genes identified to be most important. These genes can be considered as biomarkers. Moreover, the RF calculates a proximity matrix which estimates how similar the expression profiles of the samples are. This matrix can be used to construct Multi-Dimensional Scaling (MDS-) plots illustrating the similarity of the samples in a comprehensible way.

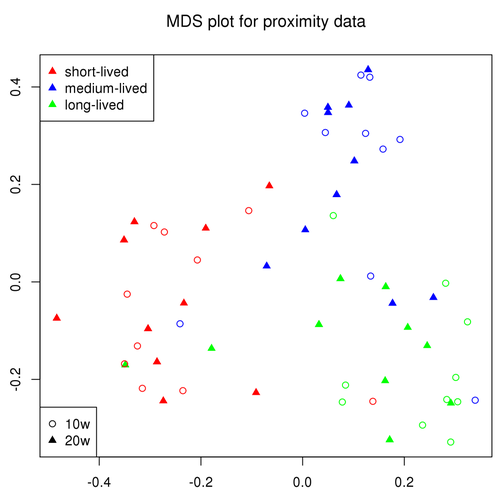


Fig. 4: MDS plot for 30 samples of *N. furzeri*.

Multi-Dimensional Scaling plots

The MDS-plot (Fig. 4) assesses the similarity of gene expression of the most informative genes. Each point in the plot represents the measured expression profile belonging to an individual, while color indicates it's subsequently recorded lifespan. It turns out that the samples fairly well cluster according to the recorded lifespan, confirming the lifespan-predictive power of the identified biomarkers, and allowing the conclusion that lifetime can, with some accuracy, already be predicted at a younger age (10/20 weeks). Furthermore, it was found that also the change of gene expression between 10 and 20 weeks has good predictive power with regard to lifespan. The accuracy of a prediction can be measured by subdividing the samples into test- and training set. Table 1 shows the observed and the predicted lifetime groups, reporting only 1 false prediction out of 11. However, because of overfitting, the prediction error is much higher if training- and test set originate from different batches, indicating the need for even higher sample numbers when reliable statistical models are to be established .

observed / predicted	short-lived	medium-lived	long-lived
short-lived	3	0	1
medium-lived	0	3	0
long-lived	0	0	4

Table 1: Observed and predicted lifetime groups by the Random Forest model for a test set with 11 samples.

[1] Baumgart et al. (2014) Aging Cell. July 25
[2] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1).