Ammene ton CARTable!

Luke Hayden

8 Novembre 2018

Classic modelling

Simple linear regression:

Age = X(marker1) + c

We try to find values for x & c that come as close as possible to solving the equation for each set of values for *Age* and *marker1* we have.

Two predictors:

Age = X(marker1) + Y(marker2) + c

Many predictors

Age = X(marker1) + Y(marker2) + Z(marker3) + W(marker4) + ... + c

Where we have many different markers, we can find values of x,y,z,w, etc that solve this equation very well but don’t provide predictive power: we call this overfitting

How do we avoid overfitting?

We want:

Modelling approach that can capture the signal without simply reproducing all the noise present in our dataset
To maximise predictive power

Data partitioning:

train-test split
cross-validation)

Model type

Ensemble methods!

Model parameters

Exploring parameter space

Machine Learning terminology

Supervised vs unsupervised learning

Unsupervised learning: find the shape of the data (
(eg: PCA, kmeans clustering)
Supervised learning: train an algorithm to recapitulate the examples it sees in a dataset
(eg: linear regression)

Classification vs Regression

Classification: categorise examples into one of a number of discrete categories
Regression: determine value along range

Classification and Regression Trees

Decision tree

Classify or perform regression by asking binary questions of data: whether value of marker X is above or below key value Y, whther marker Z is above or below. . . ..

Random Forest

Ensemble of decision trees, each using a random subset of the predictors to classify/perform regression on a random subset of the data
Resists overfitting

Gradient Boosting Machine

Start with simple model (eg: mean of values in training dataset)
Stepwise improvement (boosting) of this model by adding decision trees to progressively build a better model

Random Forest parameters

ntree: number of trees

mtry: Number of variables randomly sampled as candidates at each split

min.node.size: sets depth of trees

cross-validation folds: number of repartitions of data for testing

splitting model: variance or "extratrees"

GBM parameters

number of iterations, i.e. trees, (called n.trees in the gbm function)

complexity of the tree, called interaction.depth

learning rate: how quickly the algorithm adapts, called shrinkage

the minimum number of training set samples in a node to commence splitting (n.minobsinnode)

Model tuning

Trying to manually tune every parameter by building huge numbers of real models is extremely tiresome

Caret

R package to allow optimisation of tuning parameters for model building

Can provide a tuning grid with a range of parameters to be tested

Small models are built with all possible combinations of these parameters, then final model built under best-performing parameter set

My project as example

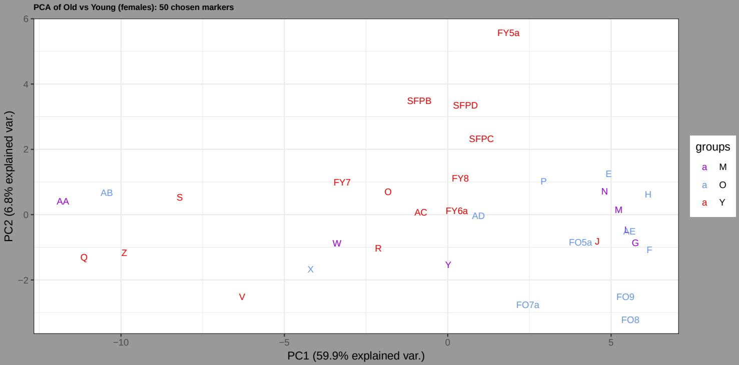
Project

Examine the effect of regeneration on the molecular age profile of *Parhyale* limbs

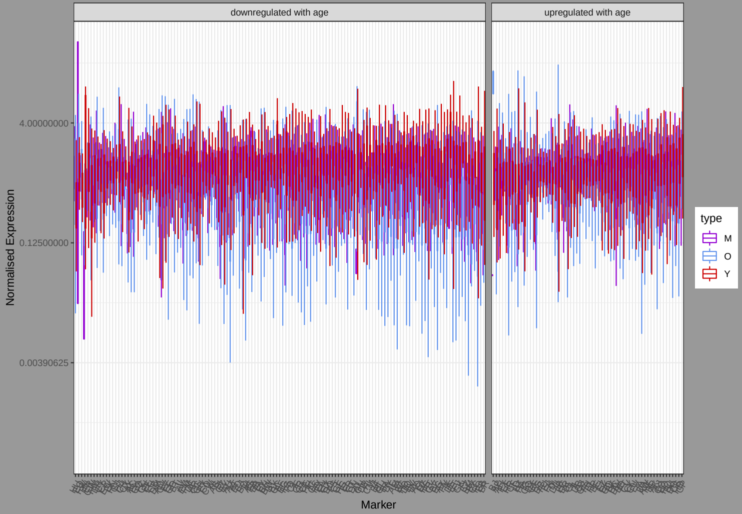
Designing codeset

*Nanostring as method to quantify gene expression
*200 genes in codeset
-195 genes chosen on the basis of differential expression analysis
-5 control genes: do not vary in expression between conditions

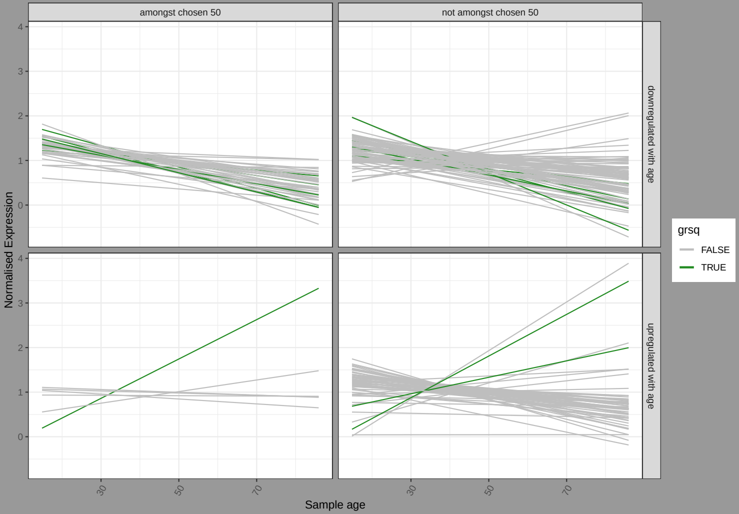
Young vs old separation: PCA old vs young



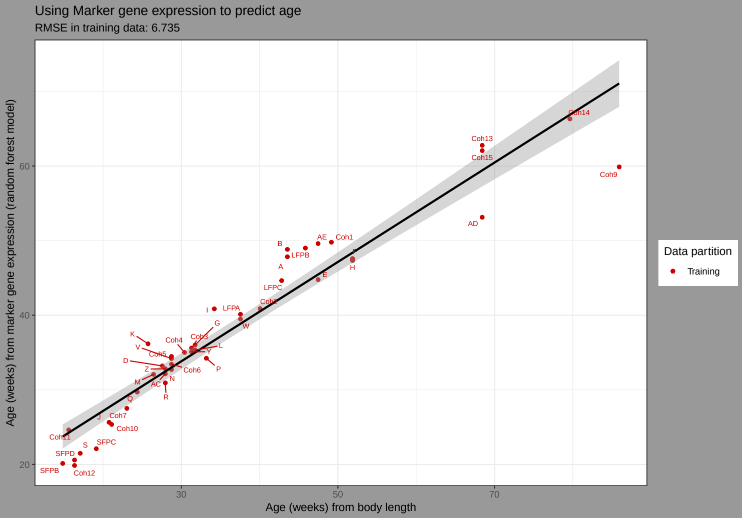
Young vs old separation: Old vs young by marker



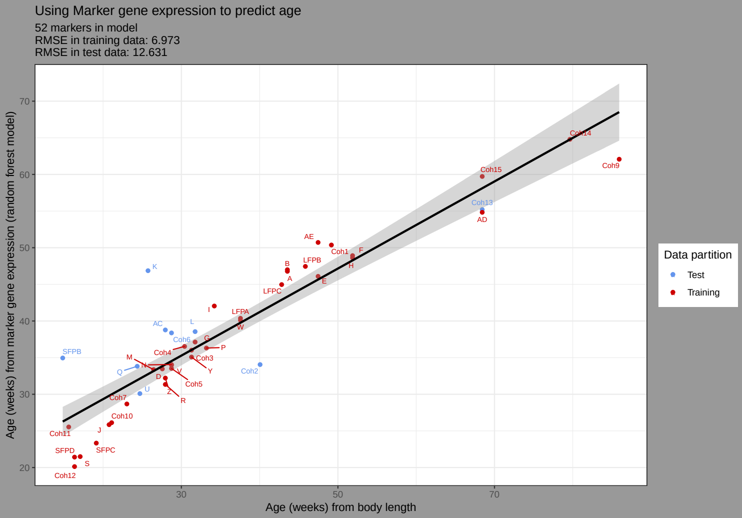
Expression/length relationship



Initial attempts



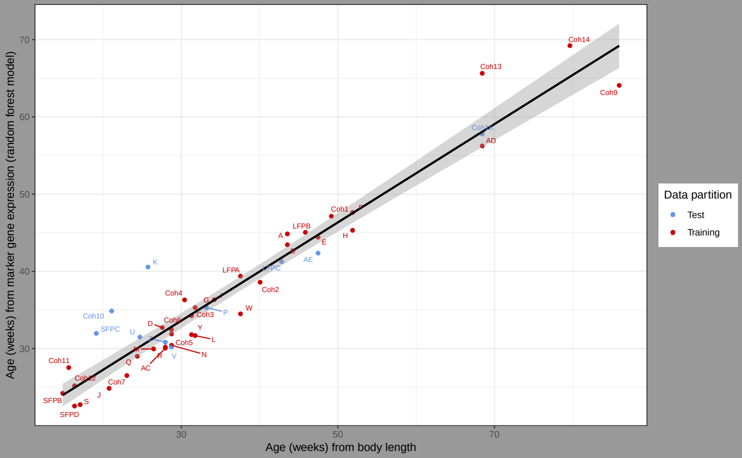
But. . . . . . .


Using Marker gene expression to predict age
52 markers in model
RMSE in training data: 6.973
RMSE in test data: 12.631
Age (weeks) from marker gene expression (random forest model)
Age (weeks) from body length
70
60
50
40
30
20
30
50
70
Data partition
● Test
● Training
Coh14
Coh9
Coh15
Coh13
AD
AE
Coh1
F
H
B
LFPB
A
E
LFPC
K
I
LFPA
W
AC
L
Coh6
G
P
Coh4
Coh3
SFPB
M
N
Q
V
Y
Coh2
D
Coh5
U
Z
R
Coh7
Coh10
J
Coh11
SFPD
SFPC
S
Coh12

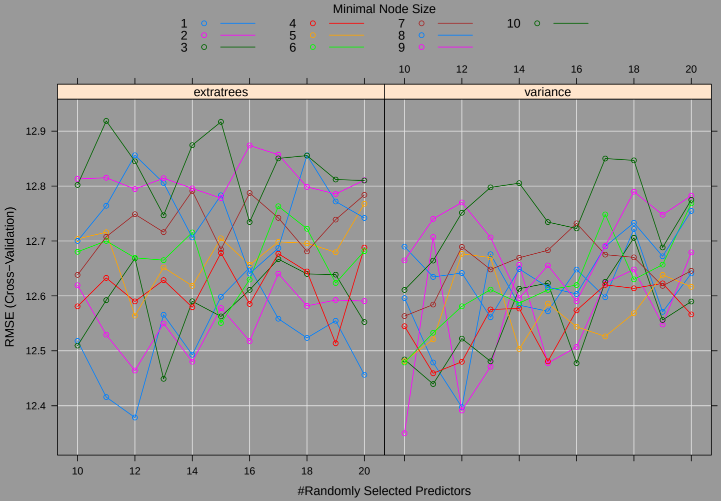
40-fold cross-validation

Using Marker gene expression to predict age

55 markers in model
RMSE in training data: 6.321
RMSE in test data: 8.812



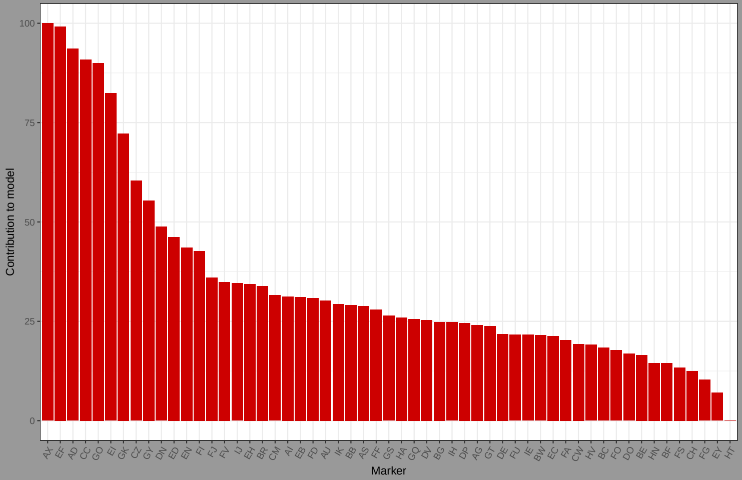
Tuning model parameters



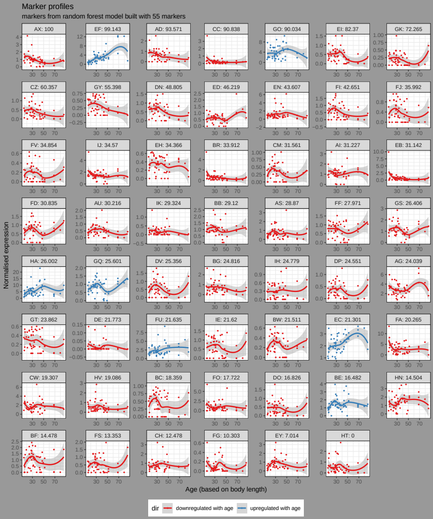
Contributions to model

Marker contributions

random forest model built with 55 markers

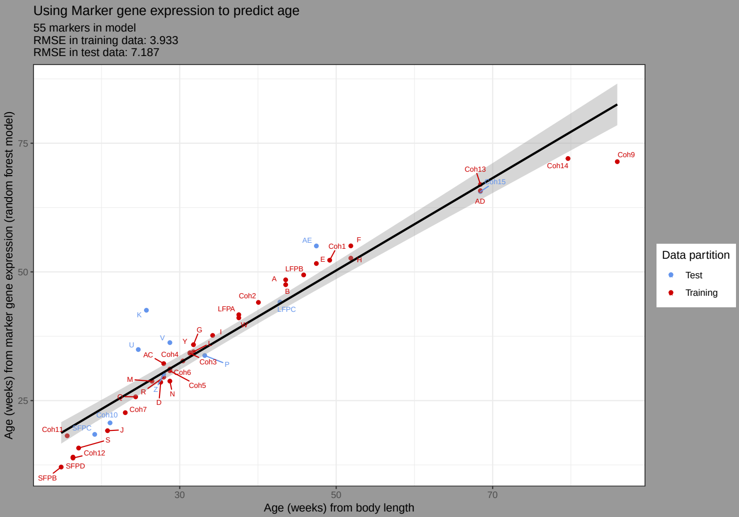


Marker profiles



It looks like some optimisation on the basis of these profiles may

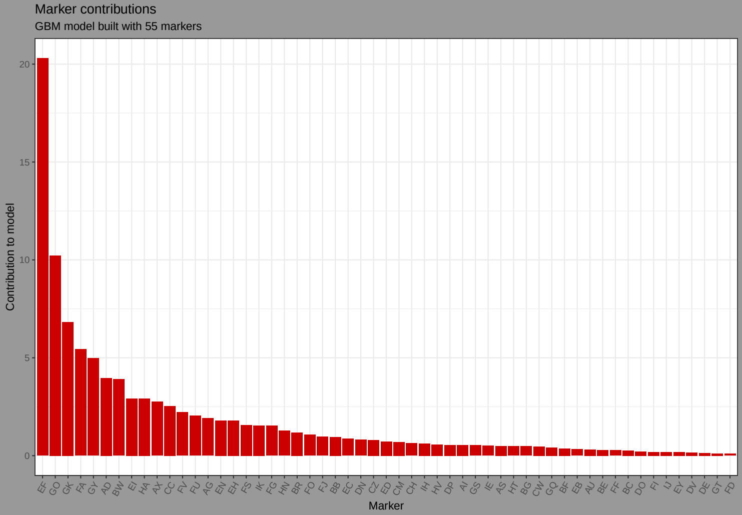
GBM



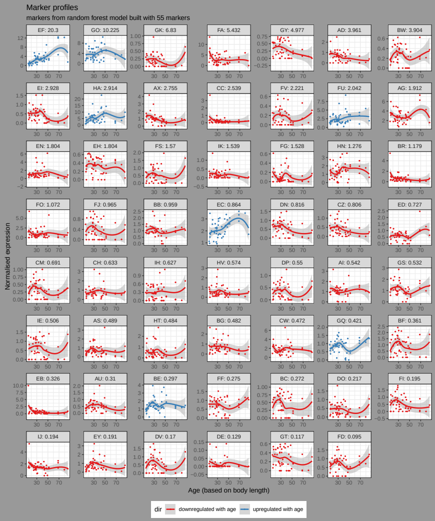
Using Marker gene expression to predict age

55 markers in model
RMSE in training data: 3.933
RMSE in test data: 7.187

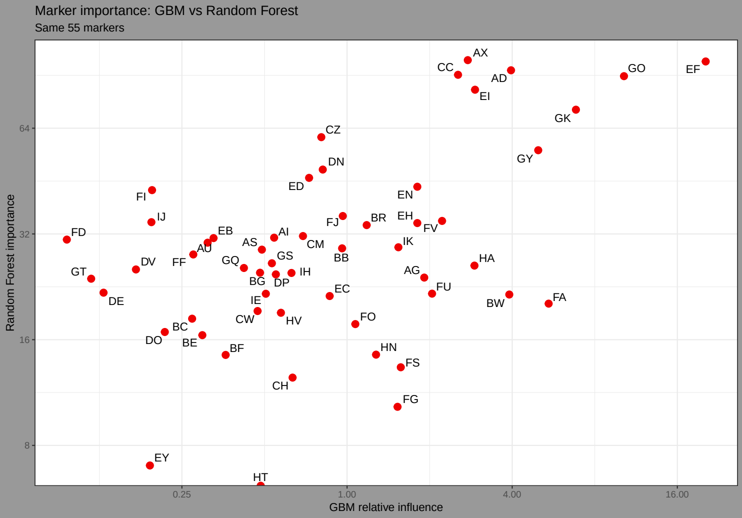
Marker contributions



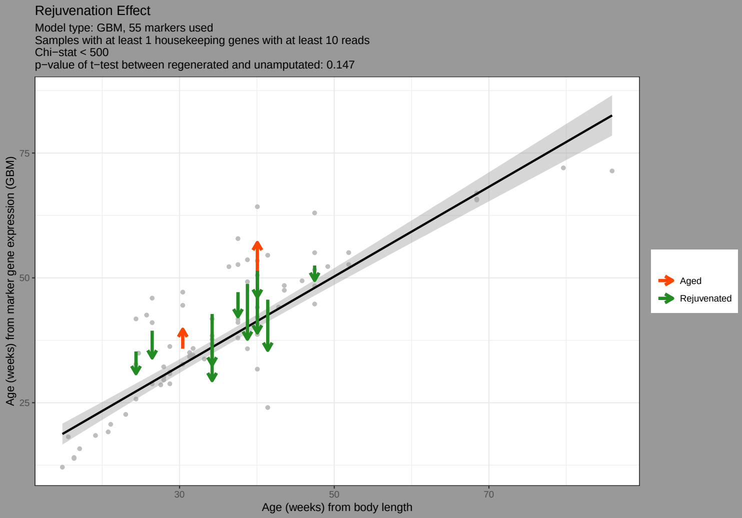
Marker contributions and profiles



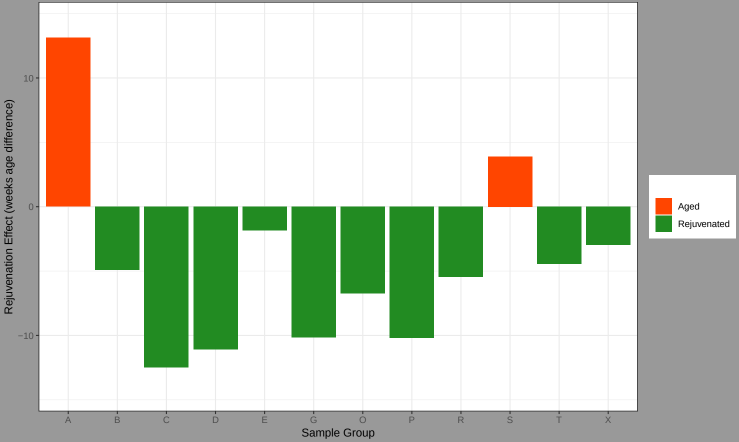
GBM vs Random Forest



Does regeneration rejuvenate?



Does regeneration rejuvenate?



Rejuvenation Effect

Model type: RANGER, 61 markers used
Samples with at least 1 housekeeping genes with at least 10 reads
Chi−stat < 500
p−value of t−test between regenerated and unamputated: 0.147