Transcriptome based classifier to distinguish mycosis and eczema

Richa

Institute of Computational Biology
Helmholtz Center Munich

27 July 2017

Objective

to build a molecular classifier distinguishing Mycosis fungoides (MF) & Atopic eczema (AE)

Motivation

in early stages mycosis is indistinguishable from eczema, using current methods of diagnosis based on clinical and histological attributes

MF or AE

MF or AE

(a) AE

(b) MF

Big picture

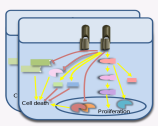
Approach

Cohort	47 AE & 9 MF patients
Sample Collection	6mm punch biopsy (lesional and nonlesional skin)
RNA Sequencing	HiSeq4000 (paired end: 2*150bp)
Sequence alignment	STAR aligner (hg19)
Data preprocessing	Bioconductor

Differential expression analysis

Molecular Classification

Functional characterization

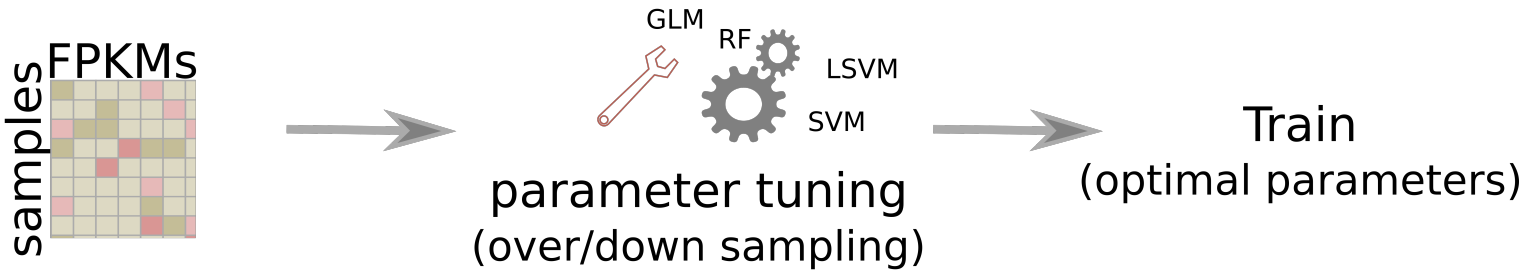


Feature ranking

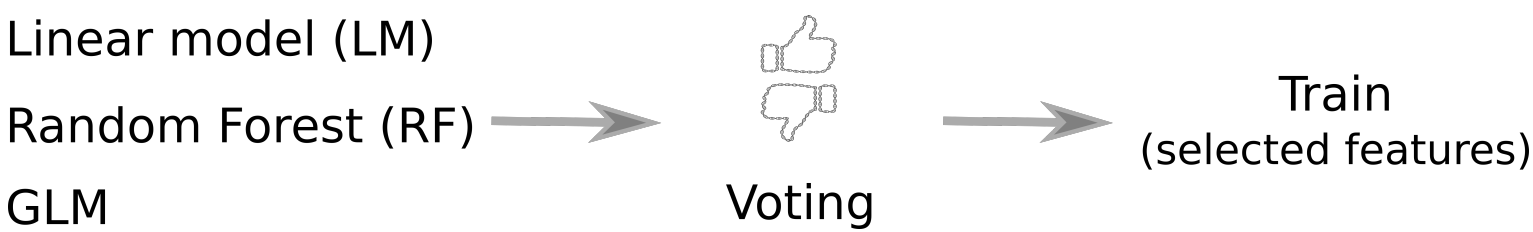
Classifier: protocol

Feature selection

Parameter optimization



Model training



Classifier: features

$$FPKM_i = \frac{ER_i x 10^9}{EL_i x MR} \tag{1}$$

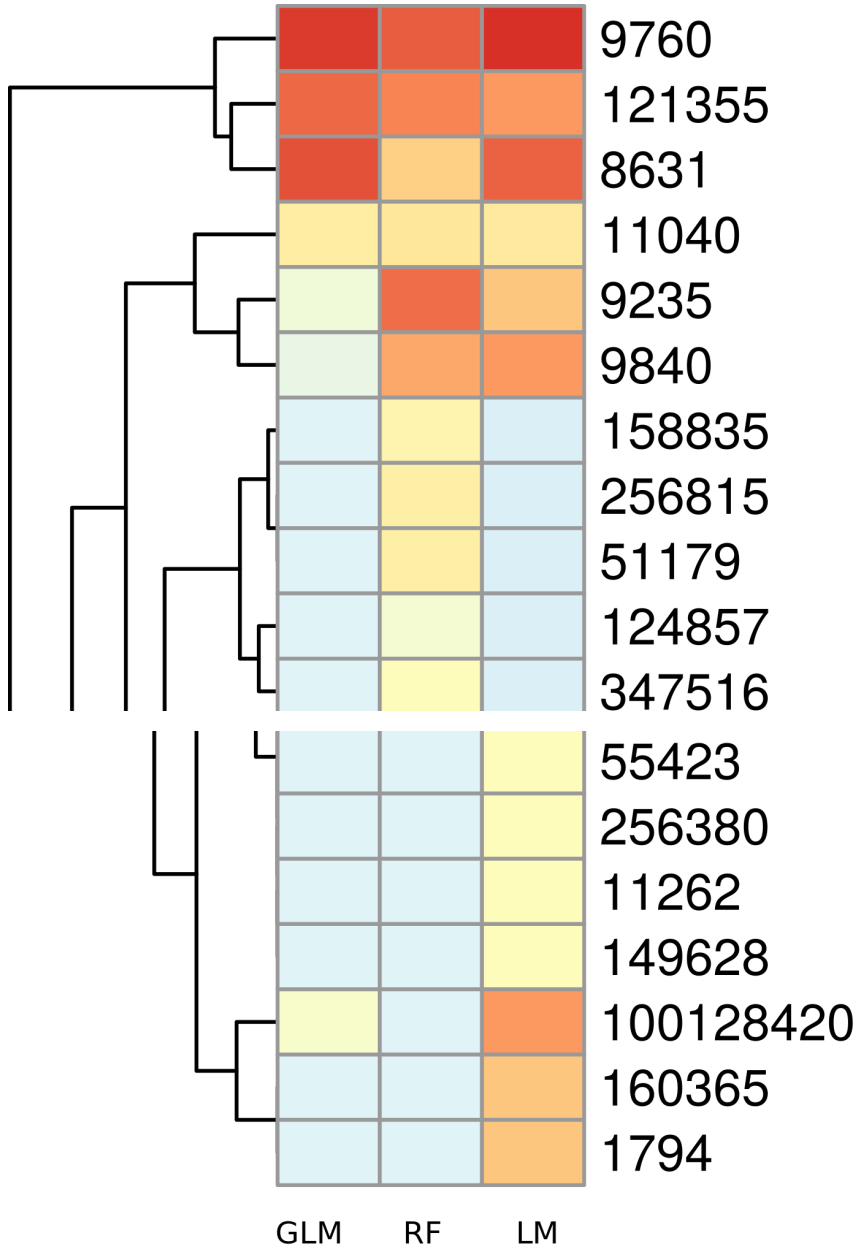
ER is number of reads, EL is exon length, MR stands of total mapped reads

$$g_i = log_2 \frac{(FPKM^l)_i}{(FPKM^{nl})_i} \tag{2}$$

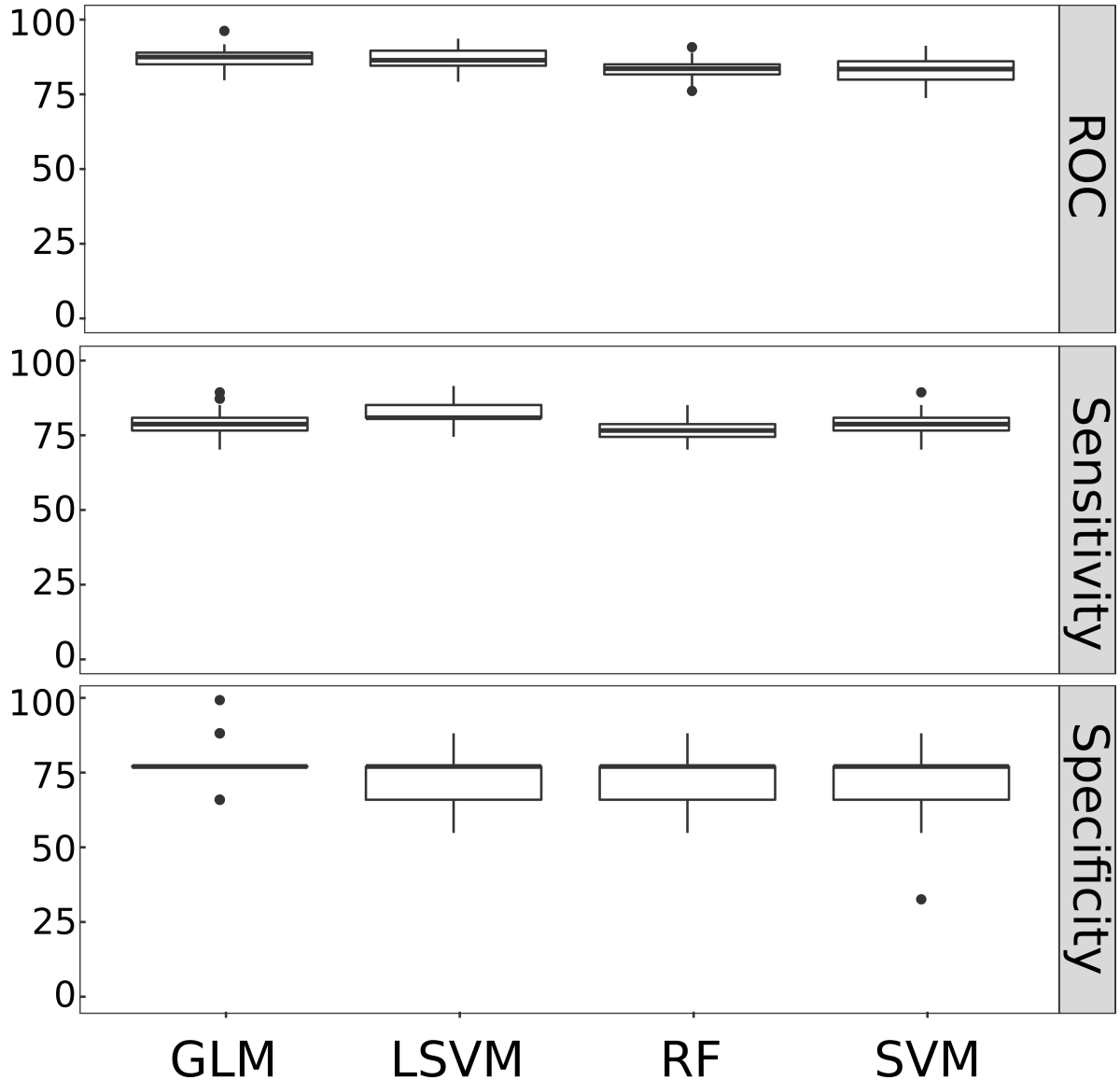
Classifier: models

1. regularization methods
2. kernel based methods: linear SVM and radial SVM
3. tree based methods: random forest

Classifier: feature selection (under sampling)



Classifier: performance (under sampling)



Classifier: performance measures

Sensitivity (true positive rate)

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

Specificity (true negative rate)

$$TNR = \frac{TN}{TN + FP} \tag{4}$$

Classifier: performance measures

100 people are tested for disease. 15 people have the disease; 85 people are not diseased. So, prevalence is 15%:

- Prevalence of Disease:
  $T_{disease}$/ Total × 100,
  15/100 × 100 = 15%

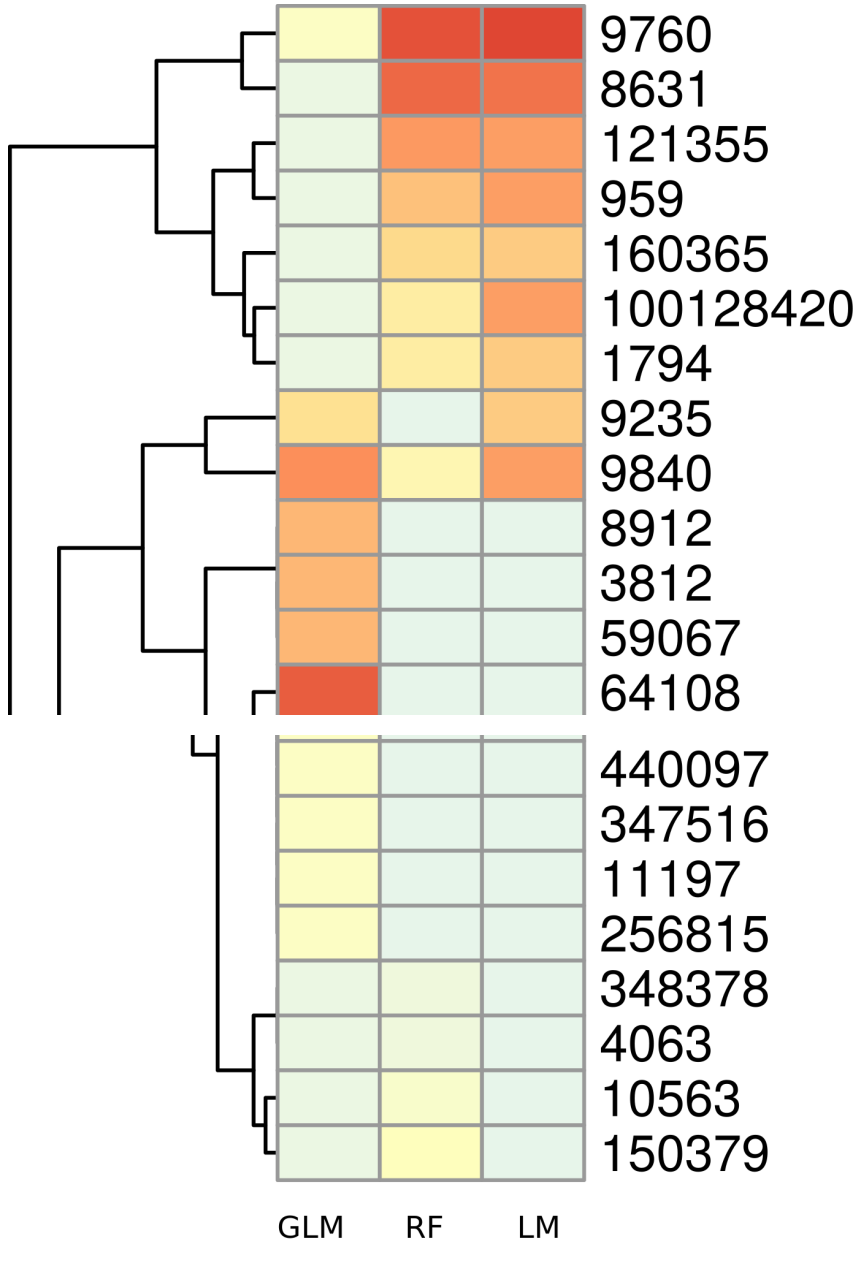
Sensitivity is two-thirds, so the test is able to detect two-thirds of the people with disease. The test misses one-third of the people who have disease.

- Sensitivity:
  A/(A + C) × 100
  10/15 × 100 = 67%

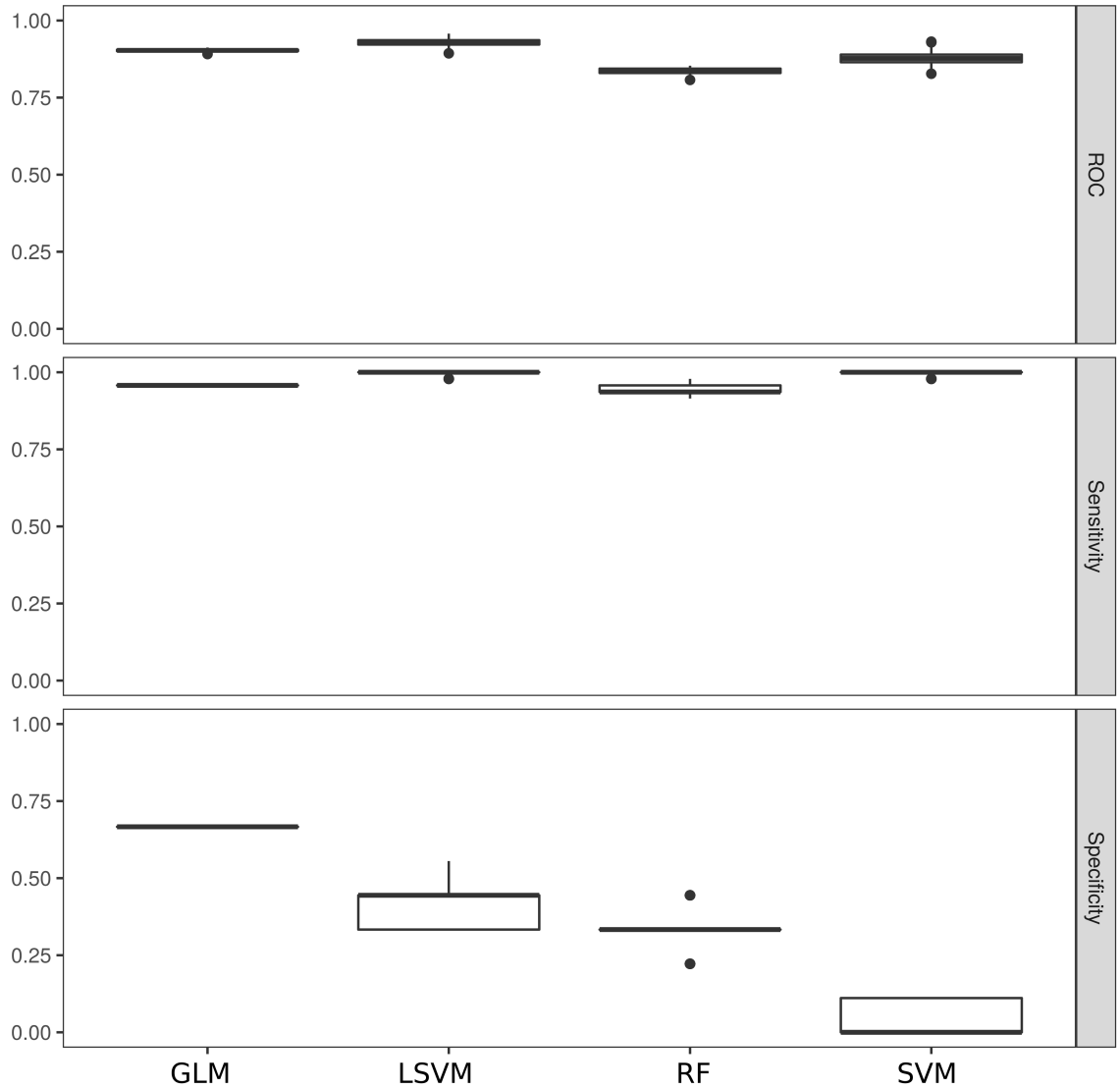
The test has 53% specificity. In other words, 45 persons out of 85 persons with negative results are truly negative and 40 individuals test positive for a disease which they do not have.

- Specificity:
  D/(D + B) × 100
  45/85 × 100 = 53%

Classifier: feature selection (over sampling)



Classifier: performance (over sampling)

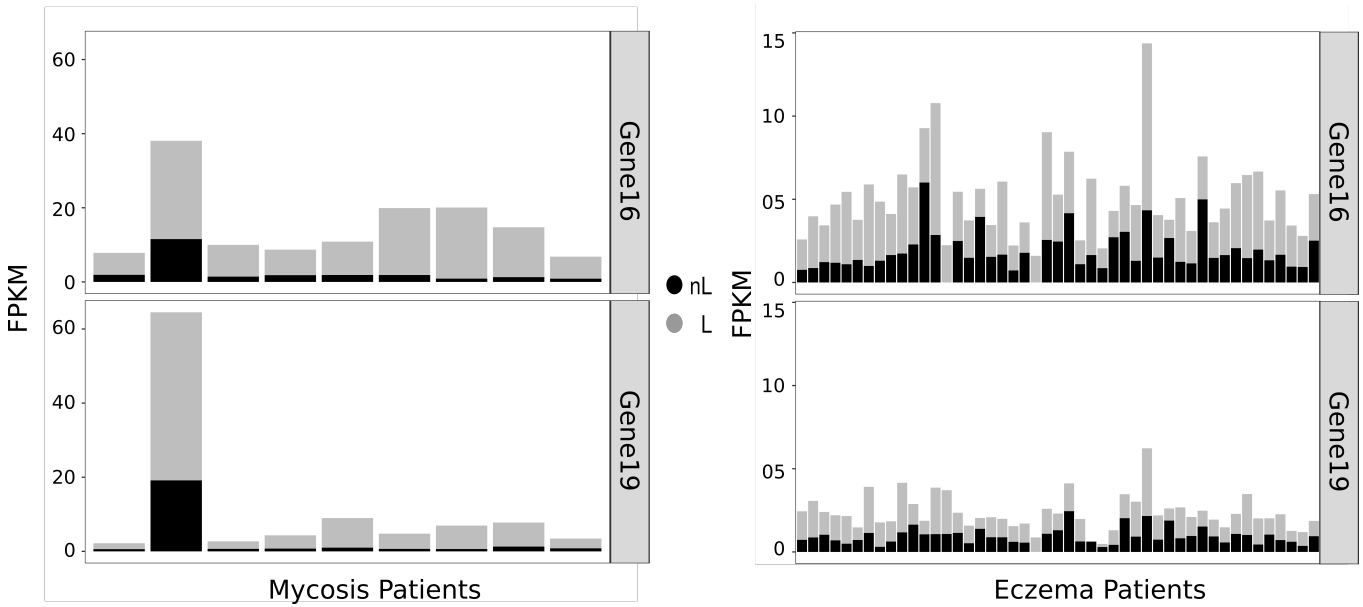


Classifier: selected features

TOX - thymocyte selection associated high mobility group box. This protein may function to regulate T-cell development. Huang Y, et al.(Blood, 2015) showed that aberrant TOX plays a role in cutaneous t-cell lymphoma i.e MF.

SKAP1 - src kinase associated phosphoprotein 1. It encodes a T cell adaptor protein

Classifier: selected features



Outlook

1. optimize feature ranking
2. validation of selected features

3. transcript level classifier
4. classifier based on lesional samples

Collaborations & Acknowledgments

Department of Dermatology and Allergy @ TUM
Center for Allergy and Environment (ZAUM)
Kilian Eyerich & AG- Eyerich

Computational Cell Maps @ ICB
Nikola Mueller & CCM
Institute of Computational Biology @ HMGU
Fabian Theis & ICB

thank you!

Questions? Suggestions?

Extra slides

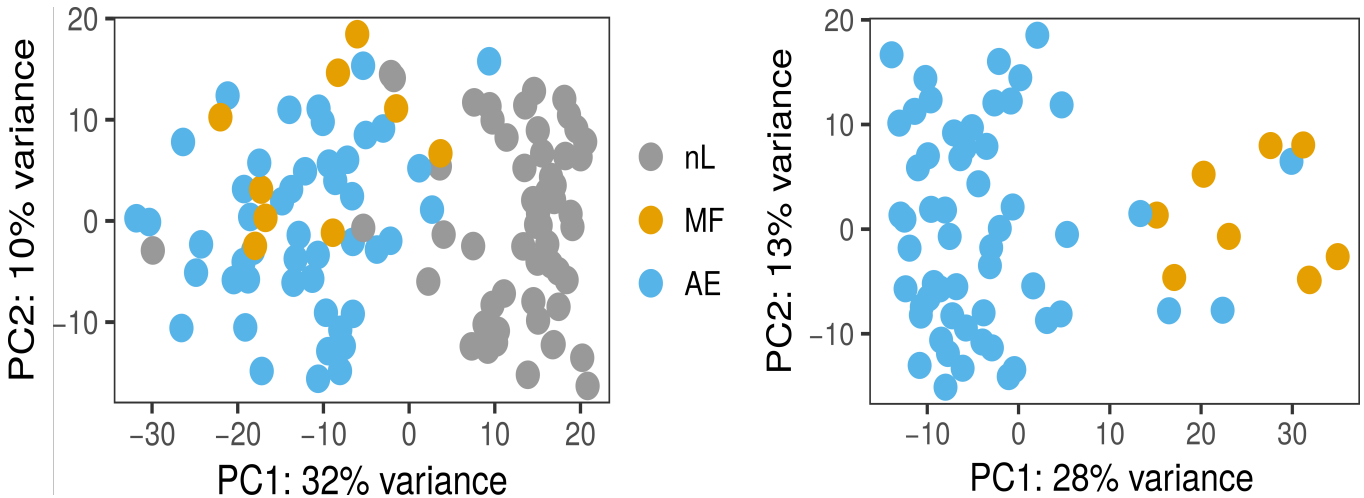
Classifier

Classifier: parameter optimization

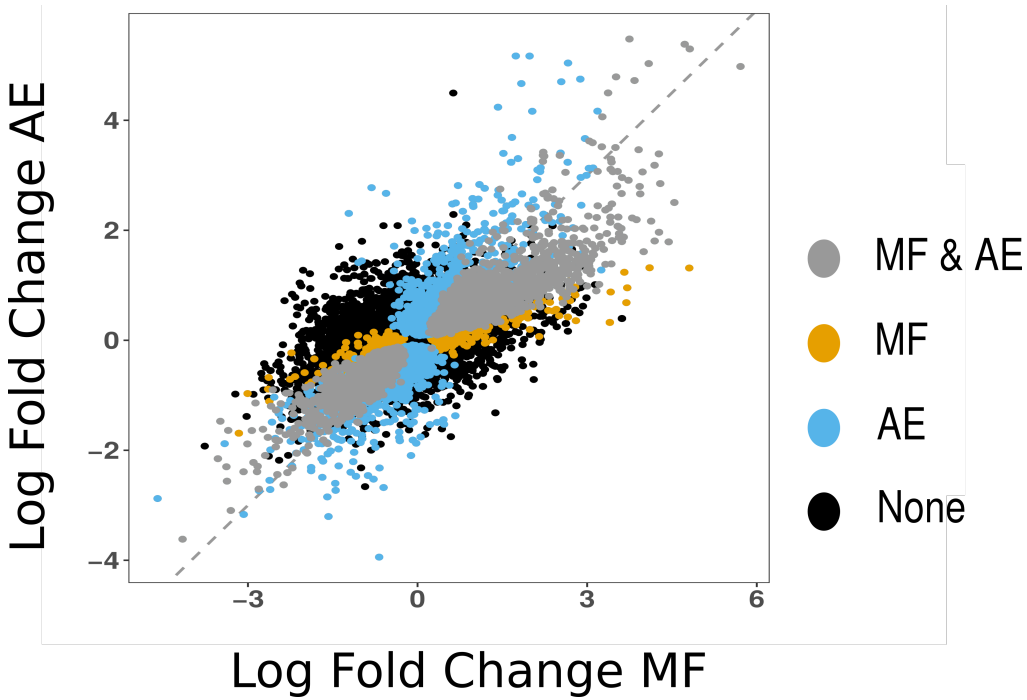
```
Define sets of model parameter values to evaluate
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples
        [Optional] Pre–process the data
        Fit the model on the remainder
        Predict the hold–out samples
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

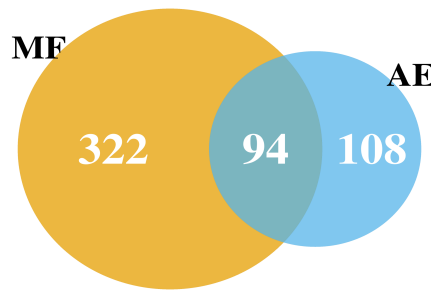
DEGS and Funtional analysis

Distribution of patients



Differentially expression genes





Pathway analysis