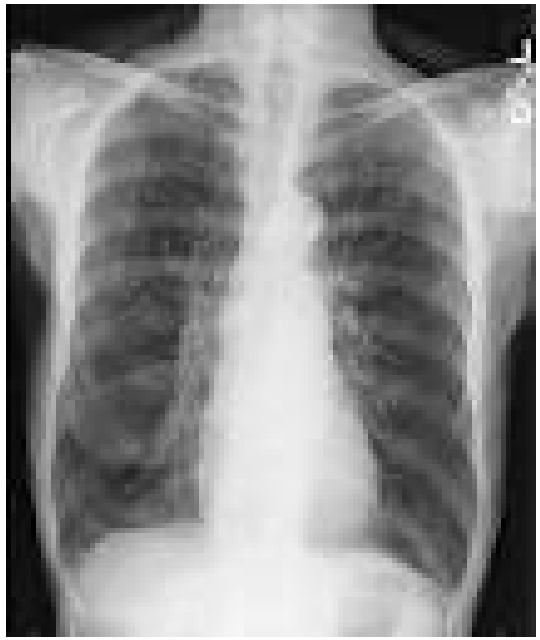
Random generalized linear model predictor for COPD diagnostics

Lin Song

Team leader: Steve Horvath

Human Genetics and Biostatistics
University of California, Los Angeles



Outline

- Construction of RGLM predictor
  - Definition.
  - Step-by-step construction.
  - Pros and cons.
- Major issues to be solved in COPD sub-challenge
- COPD data pre-processing
- COPD classification
- Discussion

Motivation

- RGLM: Random Generalized Linear Model
  - While GLMs often perform well with few covariates, RGLM also performs well for high dimensional data e.g. gene expression data.
  - Excellent prediction performance in many applications including cancer gene expression data and machine learning benchmark data (Song L et al. submitted).

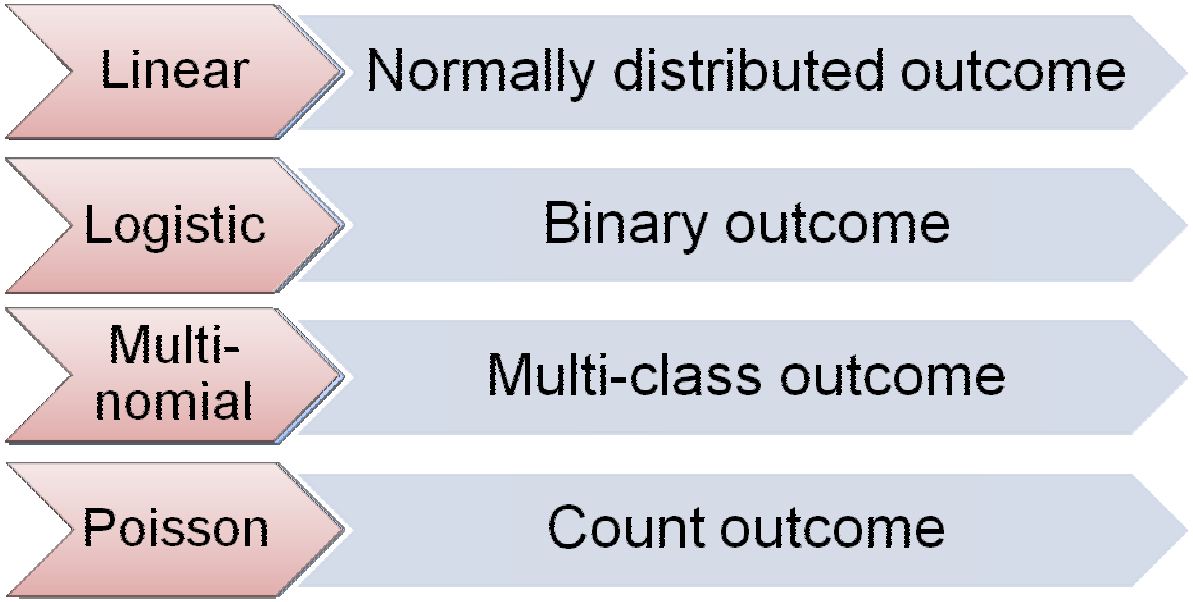
How does RGLM perform in real life diagnostic challenges?

Song L, Langfelder P, Horvath S. Random generalized linear model: a highly accurate and interpretable ensemble predictor. Submitted.

Construction of RGLM

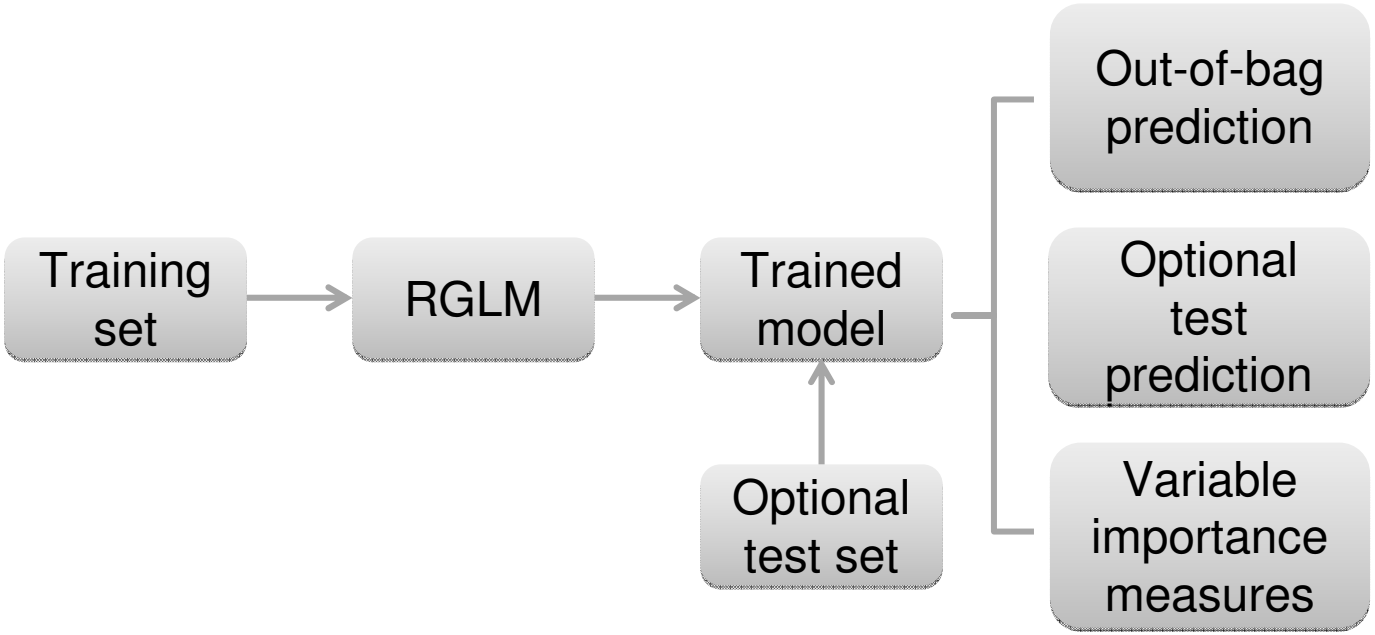
Generalized linear model

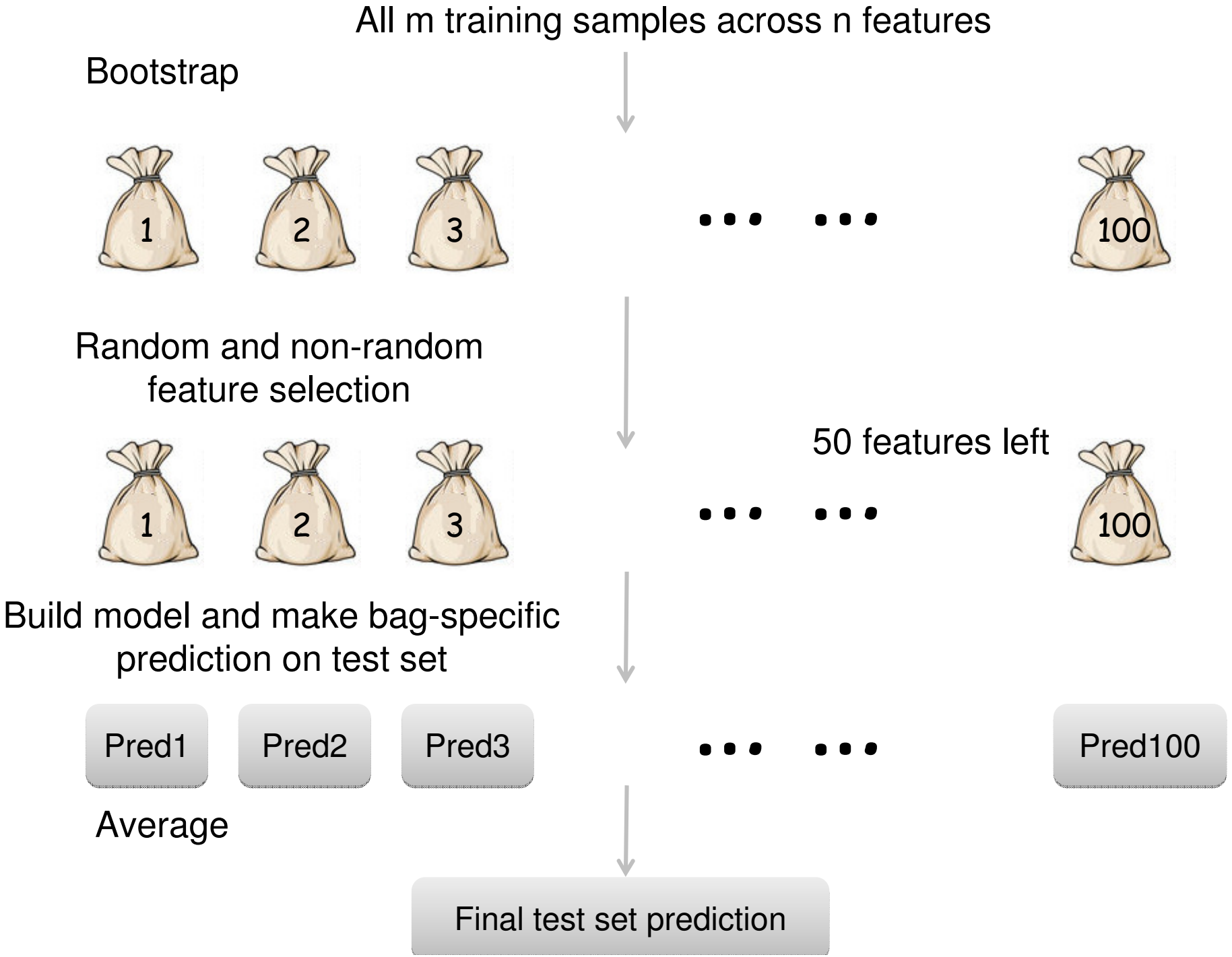
- Generalized linear model
  - Flexible generalization of ordinary linear regression.
  - Allows for outcomes that have other than a normal distribution.
- Examples



What is RGLM

- RGLM: an ensemble predictor based on bootstrap aggregation (bagging) of generalized linear models whose covariates are selected using forward regression according to AIC criteria.




All m training samples across n features
Bootstrap
1
2
3
100
Random and non-random feature selection
50 features left
1
2
3
100
Build model and make bag-specific prediction on test set
Pred1
Pred2
Pred3
Pred100
Average
Final test set prediction

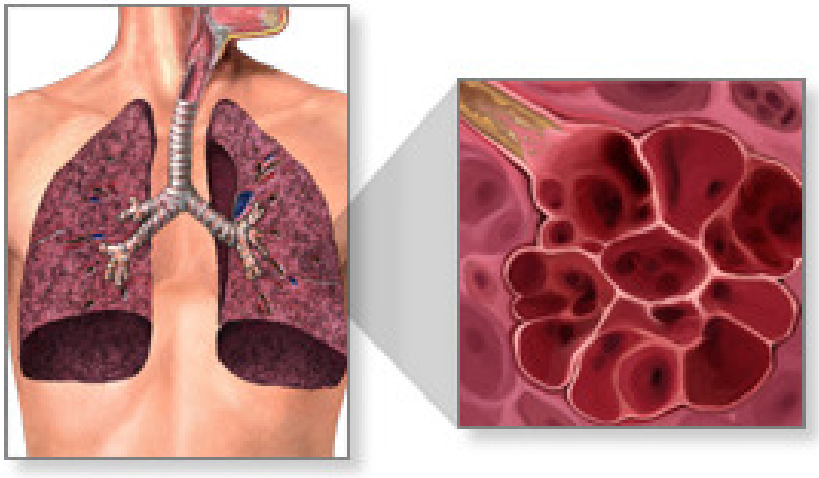
Construction of RGLM

1. Bootstrap samples of observations (bags) are generated from the training data.
2. A subset of features is randomly selected for each bag (default 20%).
3. Feature selection is carried out in each bag using a univariate GLM. Only the features with the most significant univariate significance will become candidate covariates for multivariate regression.
4. A forward selected generalized linear regression model is fitted to the training data in each bag. The fitted model is then employed to make bag-specific predictions of the test data.
5. The predictions of each multivariate model are aggregated across bags to get final predictions.

RGLM characteristics

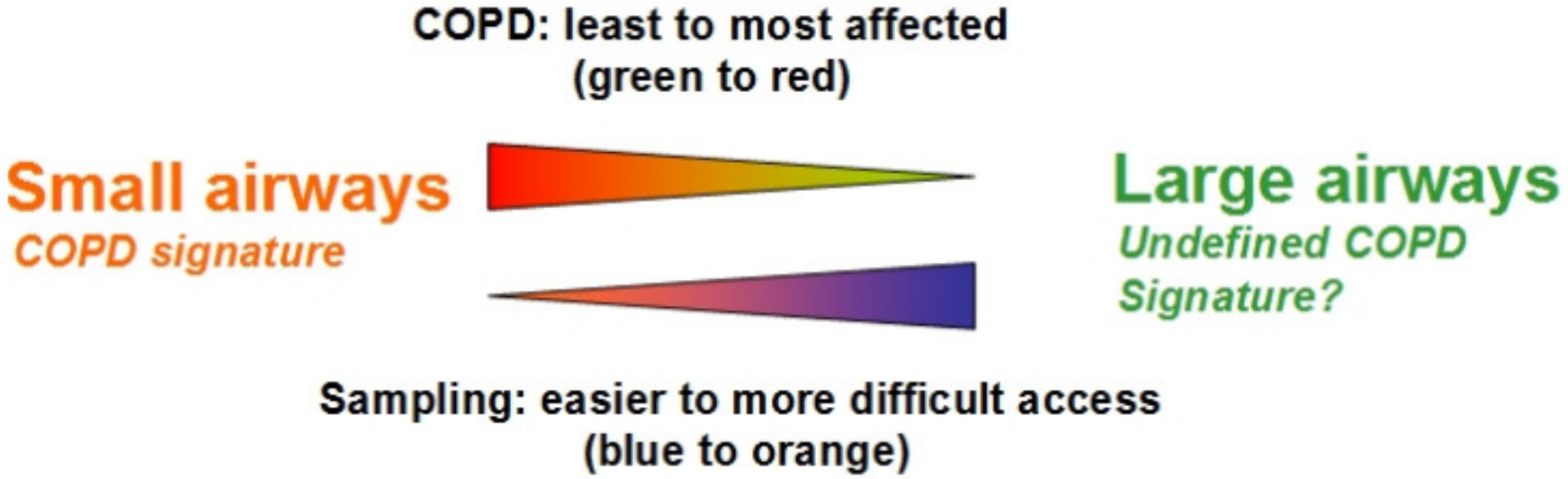
- Pros
  - Excellent accuracy.
  - Easy to interpret compared to other ensemble predictors.
  - Out-of-bag prediction.
  - Variable importance measures.
  - Users can specify “mandatory covariates” that will contribute to prediction across all bags, e.g. demographic and clinical data.
- Cons
  - Computational intensive, slower than common predictors.
- Why “random” GLM?
  - Bootstrapping.
  - For each bootstrap sample, a random subset of feature is selected.

COPD sub-challenge



Major challenges

- Small vs large airways.



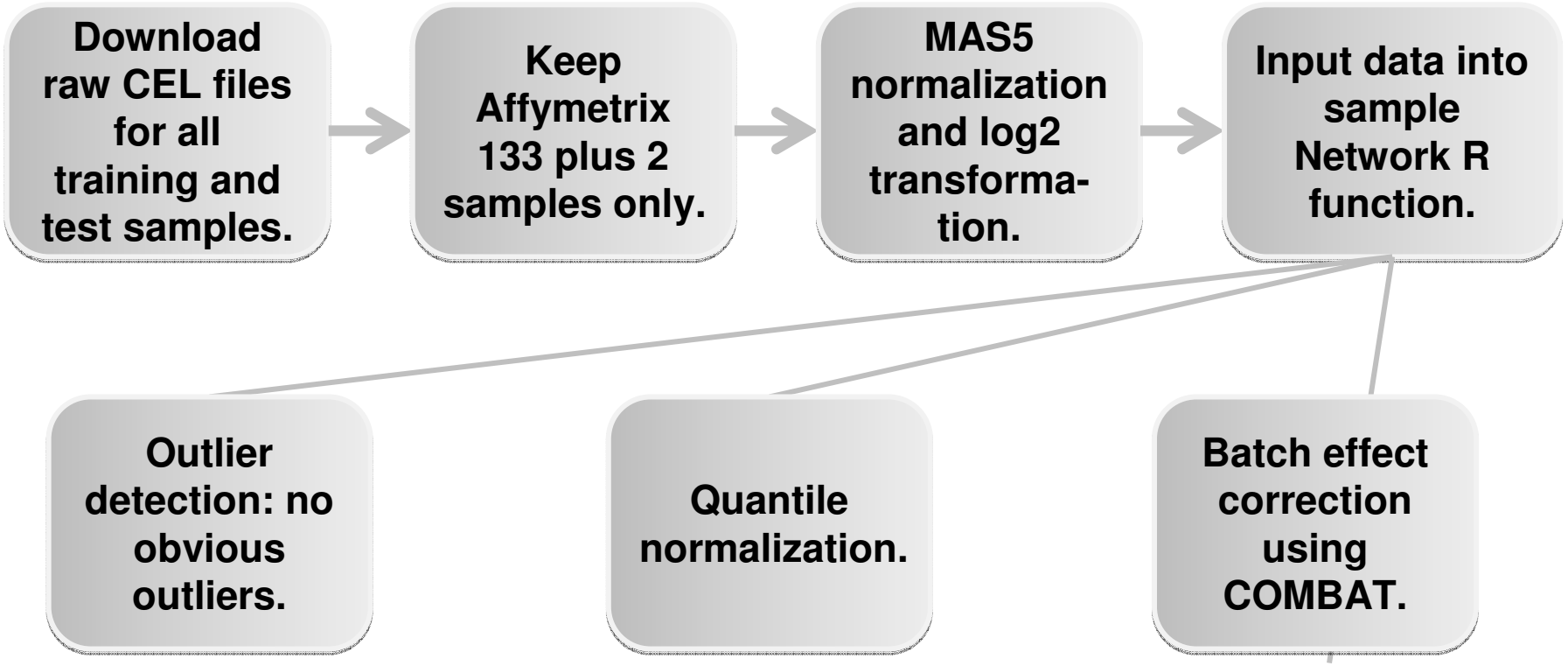
  - Training set cases: small airways only.
  - Test set cases and controls: large airways only.

Major challenges

- Platform inconsistency.
  - HuGeneFL GeneChips, HG-U133plus2, HG-U133A.
- Batch effects.
  - Training samples come from 13 GEO data sets.
  - Experimental equipments and time are not the same.
- Clinical information.
  - Smoking status and dose, gender, age, race available in both training and test sets.
- Classification strategies.

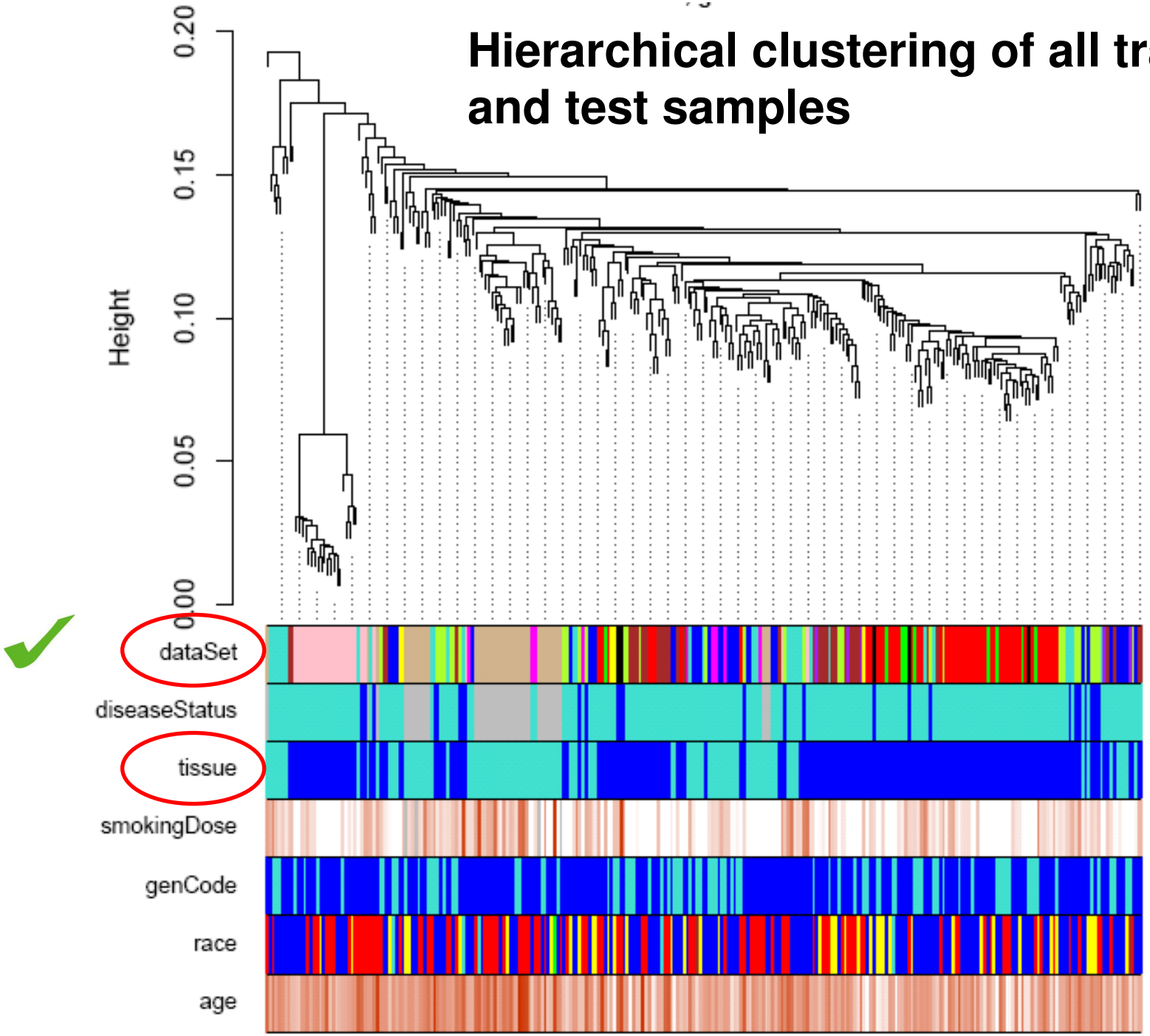
COPD data preprocessing

COPD data preprocessing



Oldham MC, Langfelder P, Horvath S (2012) Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. BMC Syst Biol. 2012 Jun 12;6(1):63.
Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics* **8**(1):118-127.

Hierarchical clustering of all training and test samples



Data after preprocessing

- For processing the data we used all probes, i.e. no feature selection was carried out.
- All training samples are pooled together into one large training set in order to increase power.
- Training set: 237 samples X 54675 probes.
  - 26 COPD patients, 211 controls. Highly unbalanced.
- Test set: 40 samples X 54675 probes.

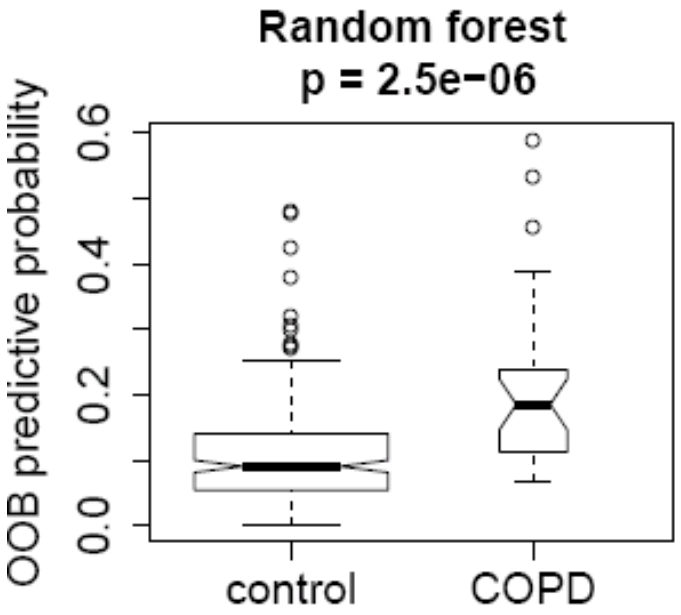
COPD classification

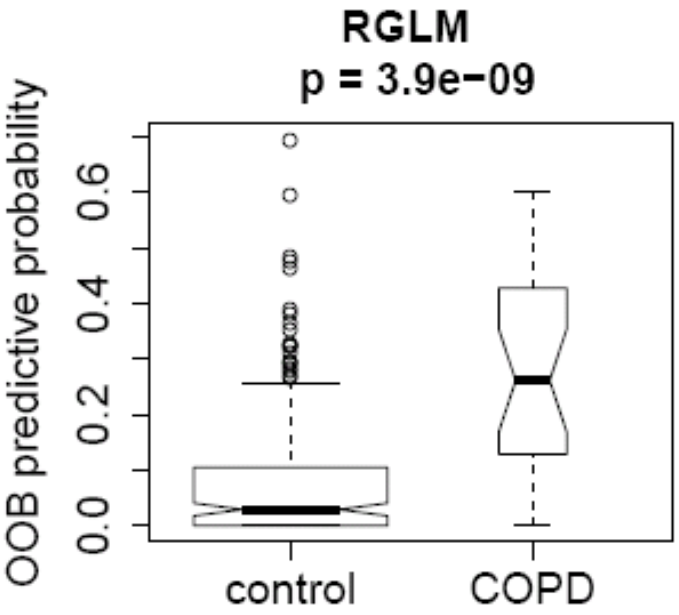
Classification strategy

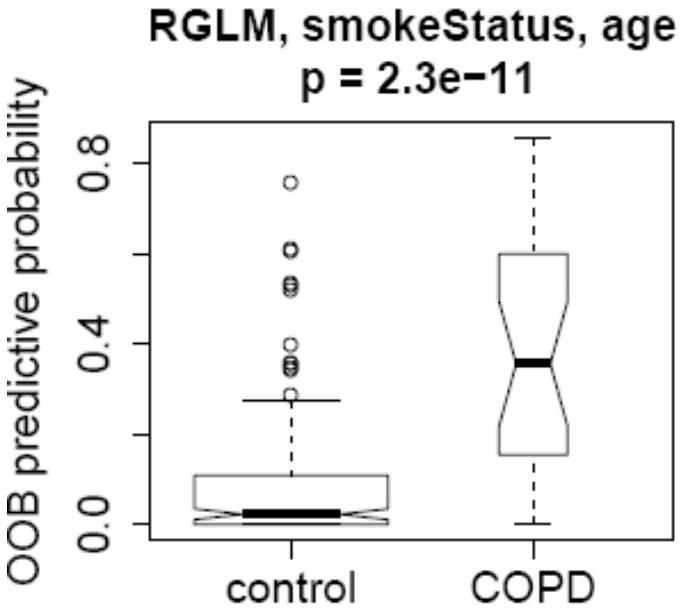
- Classification strategy selection criteria:

  Training set out-of-bag (OOB) prediction accuracy.
- Strategies considered:
  - RGLM vs commonly used predictors such as random forest.
  - Whether and how to use clinical information.
  - Smoking and aging are known risk factors for COPD.
  - RGLM mandatory covariates: force variables into each individual weak learner.

Young RP, Hopkins RJ, Christmas T, Black PN, Metcalf P, Gamble GD (2009). COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. Eur. Respir. J. 34 (2): 380–6.
Kazuhiro Ito, Peter J. Barnes (2009). COPD as a Disease of Accelerated Lung Aging. CHEST.2009;135(1):173-180.







- RGLM is superior to the random forest on these data.
- Adding smoke status and age as mandatory covariates are beneficial.

Using prior biological knowledge

- Smoking is a major risk factor for COPD.
  - Almost all lifelong smokers will eventually develop COPD.
- We know the smoking status of test samples.
- Therefore, we assume half smokers would be COPD patients.
- Re-calibrate test set predictive probabilities so that half smokers have predictive probability >0.5.
  - We did a linear transformation of the predictive probabilities on the log-scale.
- This insight allows us to counter the bias resulting from severely unbalanced training data.
  - 26 COPD patients and 211 controls.

Discussion

- 423 probes finally contribute to RGLM prediction.
- Top probes selected by RGLM from genes such as REPIN1, KDM3A, ZNF565 and C2orf70, are highly informative to predict COPD. It is not clear, however, whether these genes are biologically relevant to COPD.
- Our goal was prediction and not learning biology.
- When it comes to learning biology, we would recommend a systems biologic approach: weighted gene co-expression network analysis (WGCNA) but this is a different topic.

* RGLM is available as an R function randomGLMpredictor in the WGCNA R package.

Acknowledgement

- Steve Horvath, PhD, ScD

  Professor of Biostatistics & Human Genetics
  UCLA
  Team leader

- Peter Langfelder, PhD

  Biostatistician, UCLA