

# **Biomarker signatures discovery to support cancer diagnosis**

Towards an accurate and robust machine  
learning strategy

**Ahmed Debit, PhD Student (\*)**

[adebit@uliege.be](mailto:adebit@uliege.be)

(\*) GIGA-R, Medical Genomics, University of Liège, Belgium

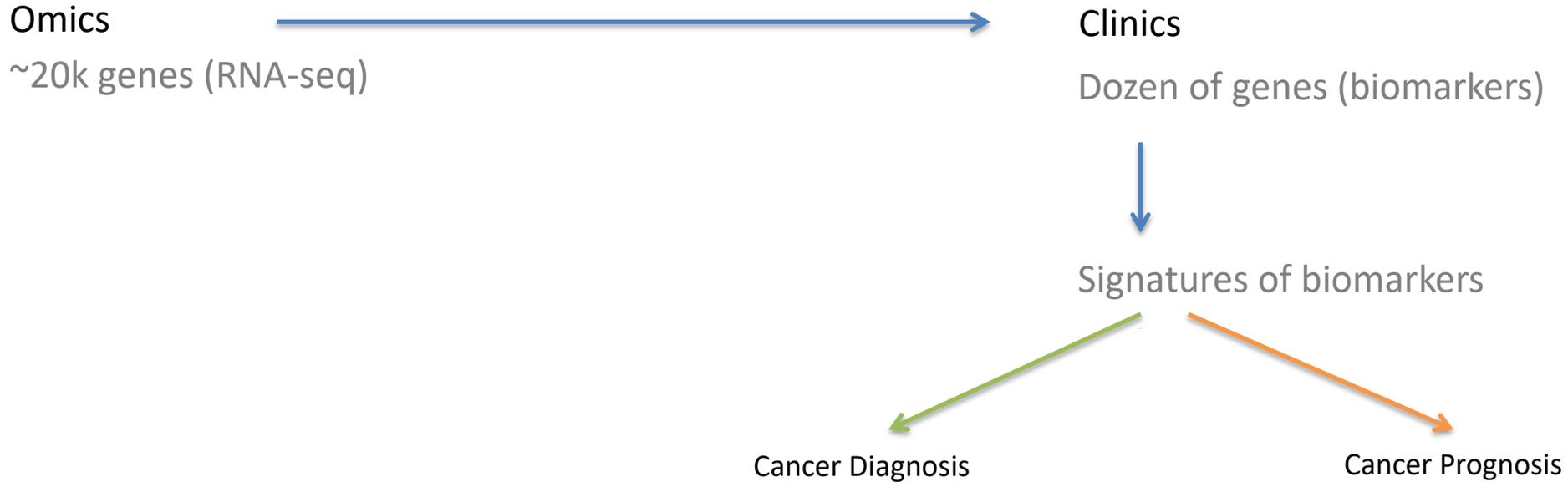
# Context

- Diagnosis of Breast Cancer
- Breast Cancer treatment response
- Design of short biomarker signatures

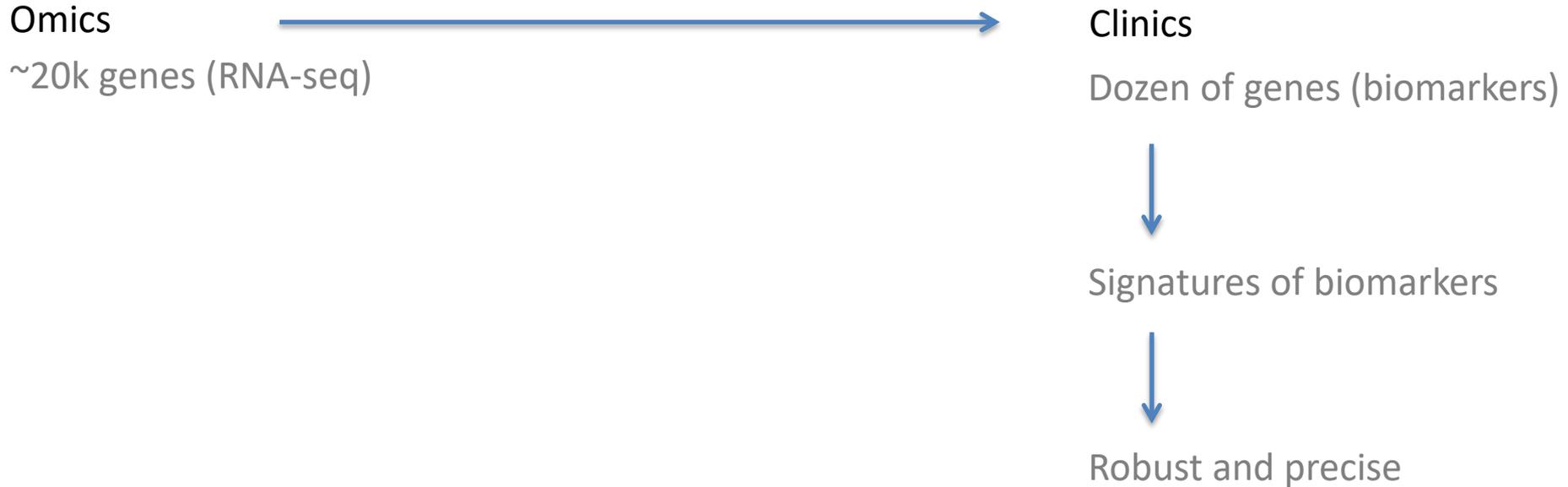
# From Omics to Clinics



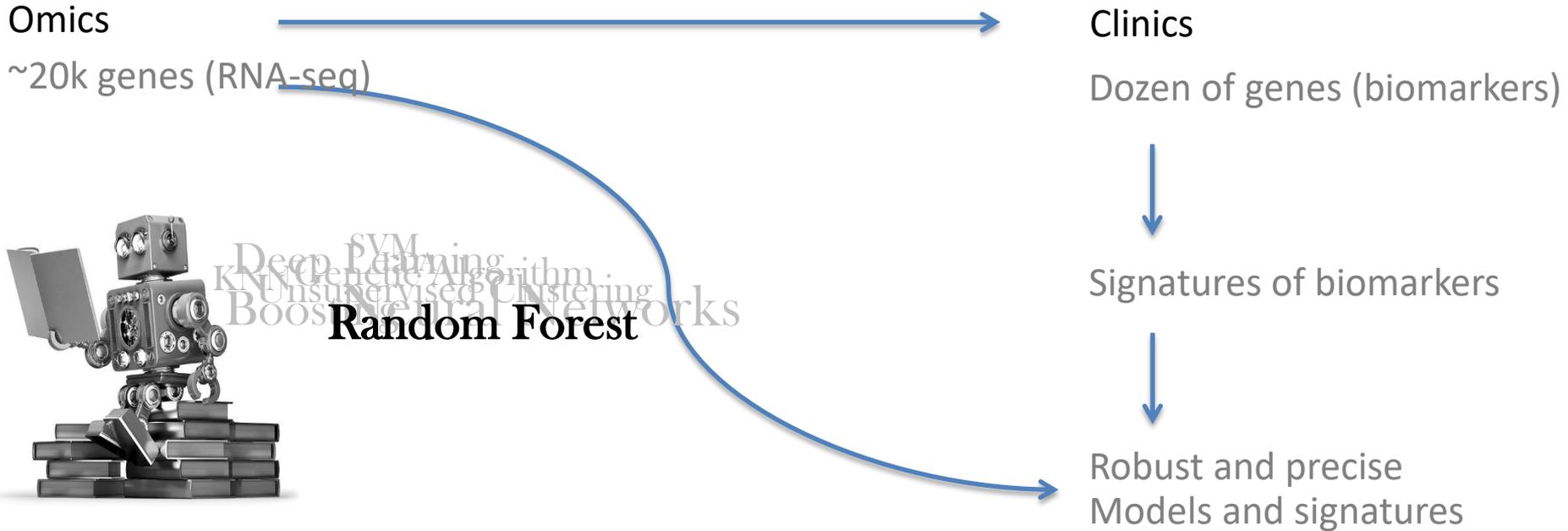
# From Omics to Clinics



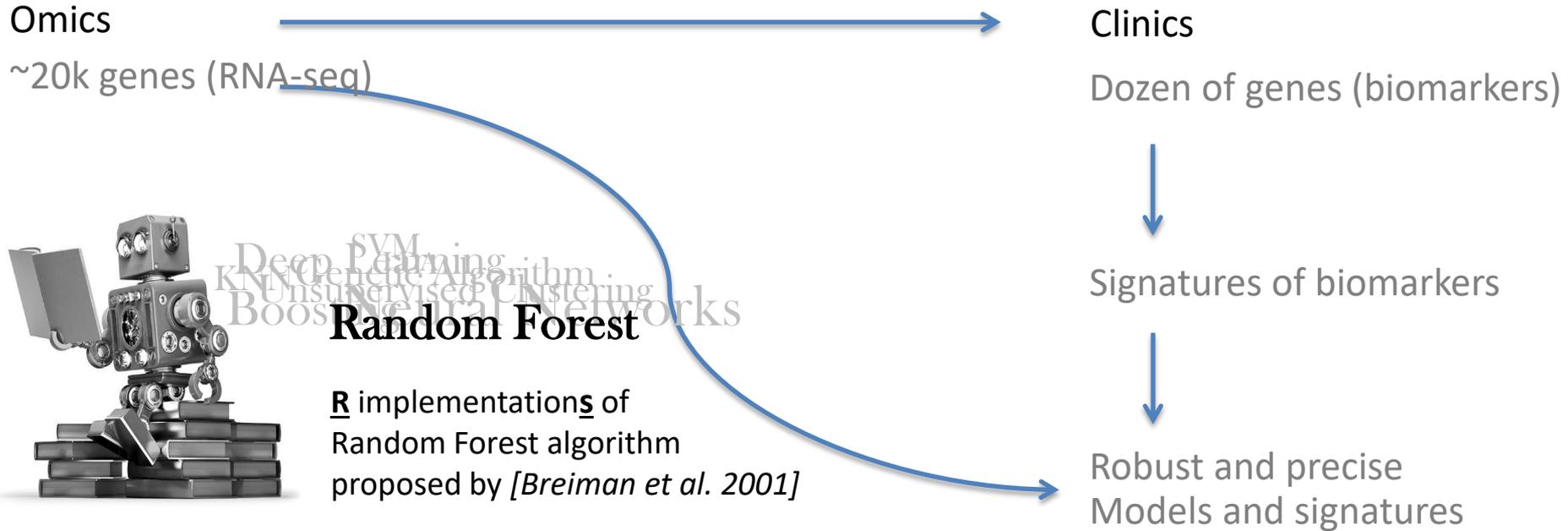
# From Omics to Clinics



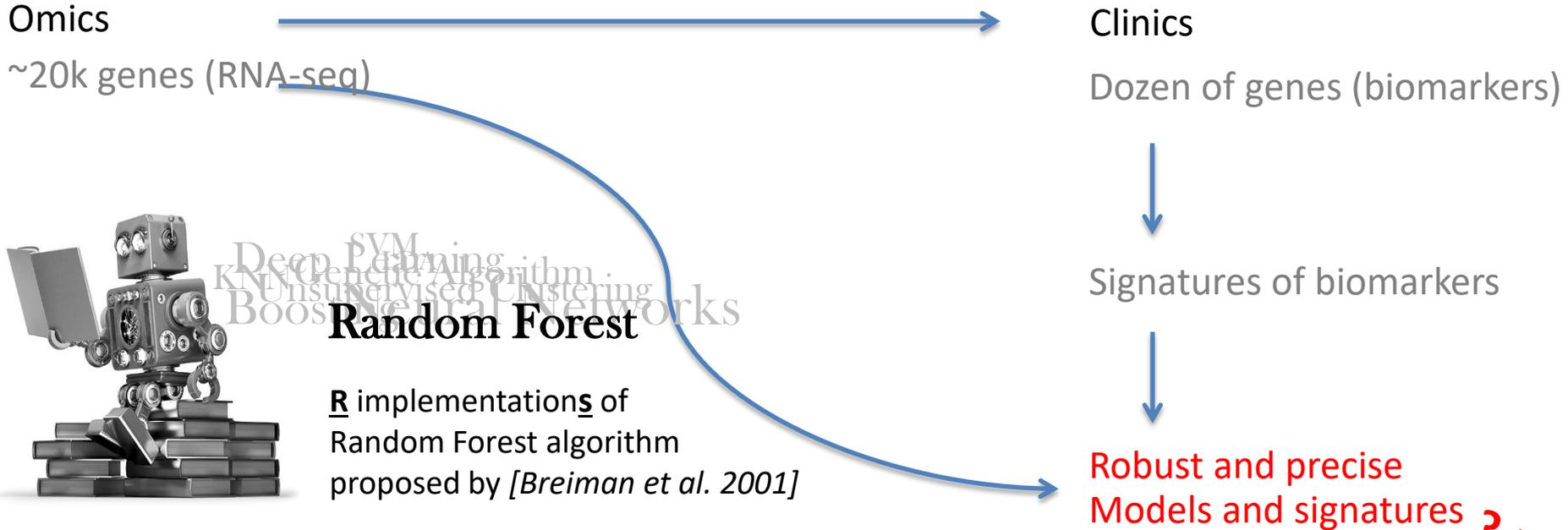
# From Omics to Clinics



# From Omics to Clinics



# From Omics to Clinics



# Objectives

**Toward a robust RF method for the Biological question asked**

Which method is suitable for which dataset (platform/technology) ?

# Objectives

- Empirical comparison of random forest based methods
- Differences/Similarities of RF methods → groups of methods
- Designing a high stability score to rank RF methods

**Toward a robust RF method for the Biological question asked**

Which method is suitable for which dataset (platform/technology) ?

# Materials and Methods

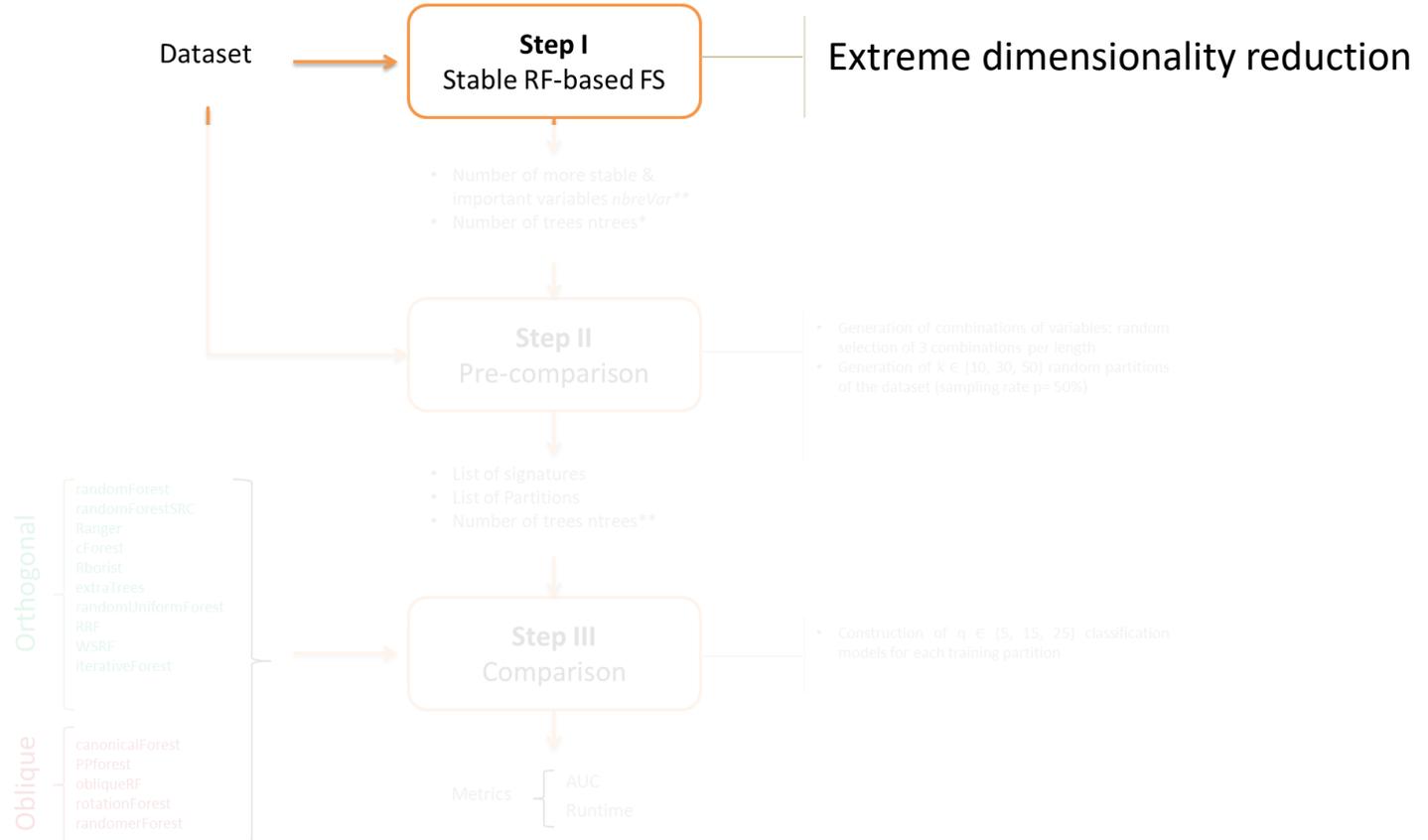
- **Datasets (Perfectly balanced)**

- TCGA-BRCA (RPKM): 182 samples x 9560 genes
- TCGA-LUSC (RPKM): 96 samples x 9262 genes

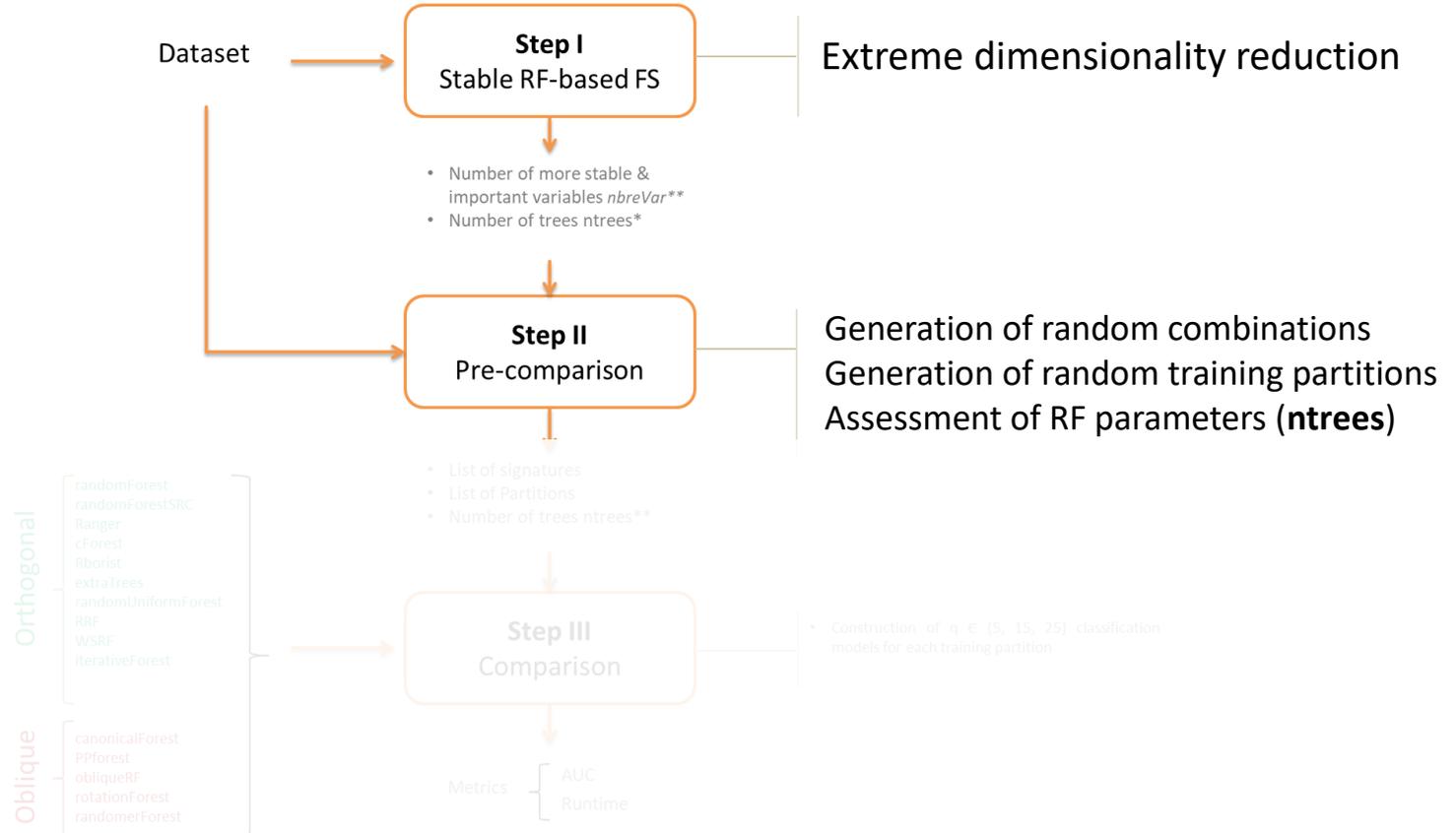
- **Main classification question**

The difference between paired **Tumor / Normal** samples will be used as a **strong classification** parameter, allowing for **strong** modeling only

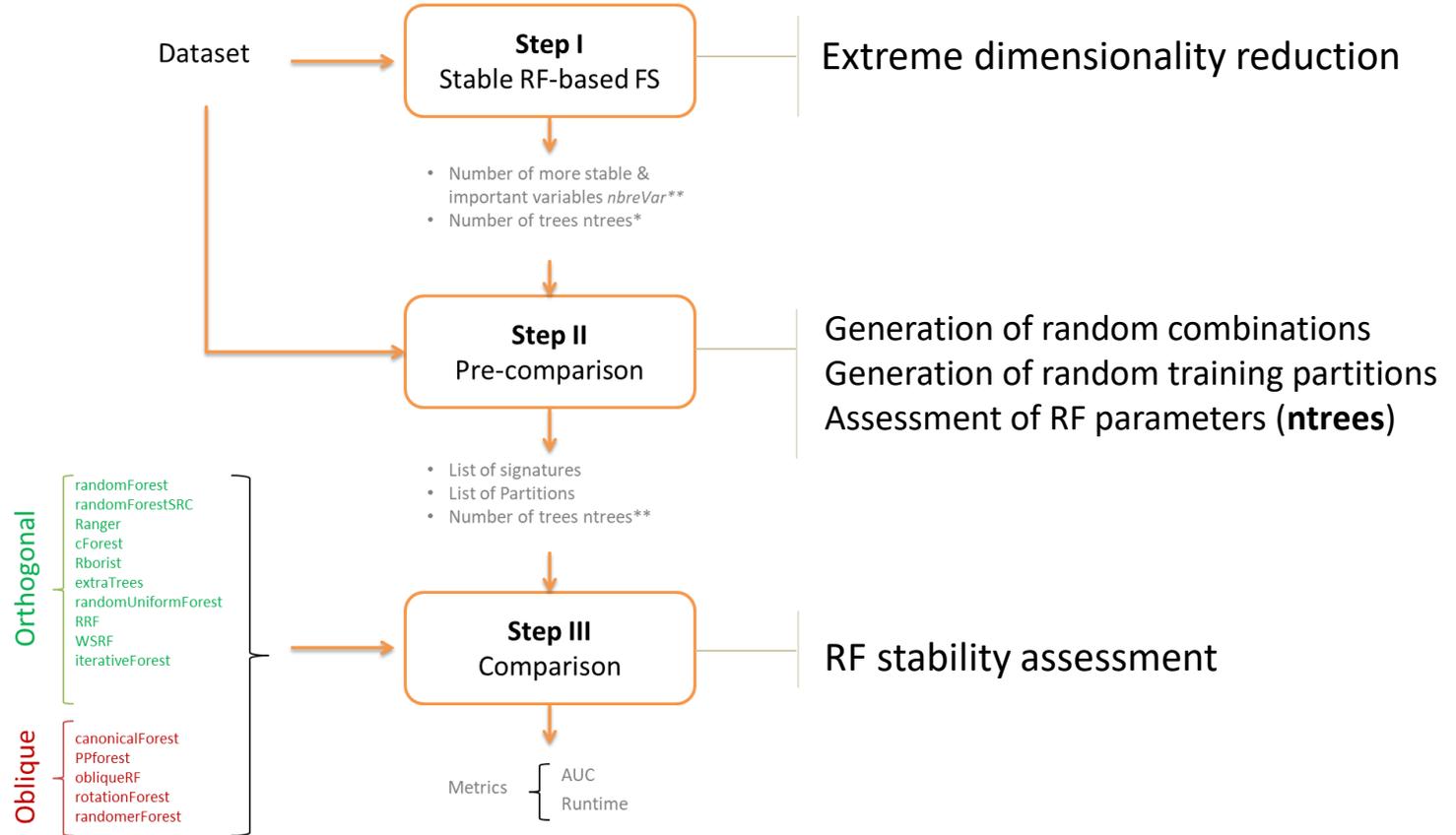
# Overview of the method



# Overview of the method



# Overview of the method

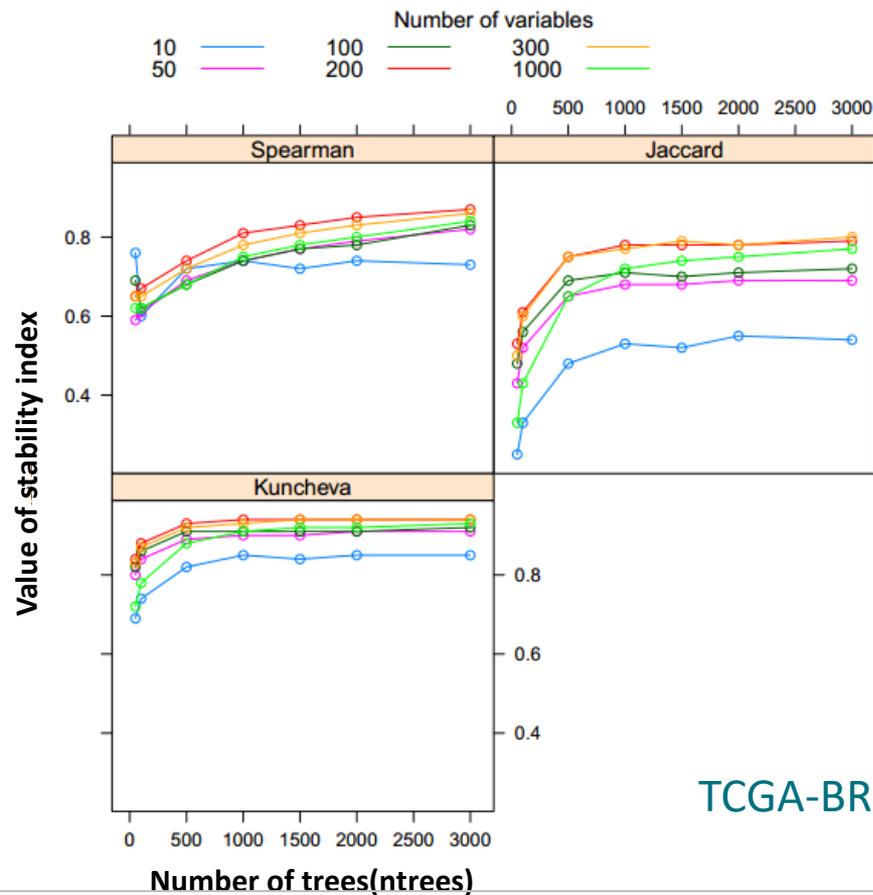


---

# Stable Feature Selection

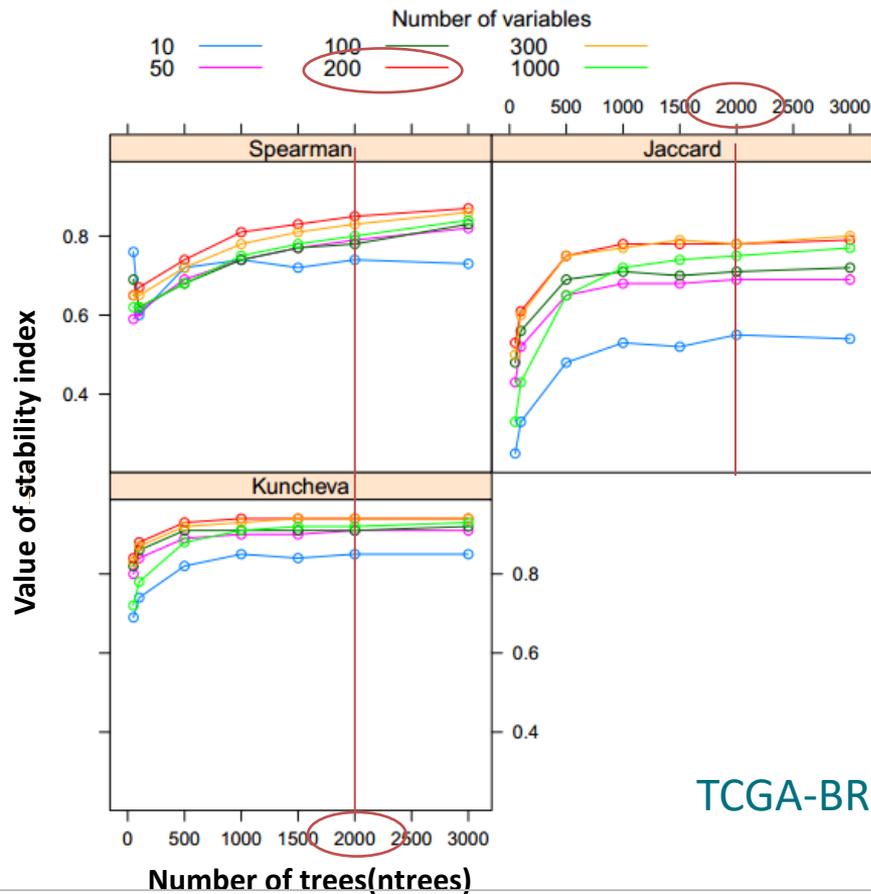
## Extreme dimensionality reduction

# First pass Feature Selection results



TCGA-BRCA dataset

# First pass Feature Selection results

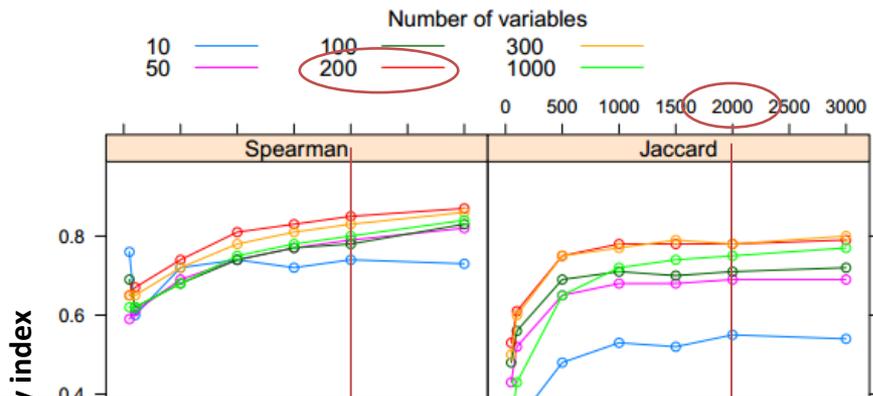


$ntrees^* = 2000$

$nVar^* = 200$

TCGA-BRCA dataset

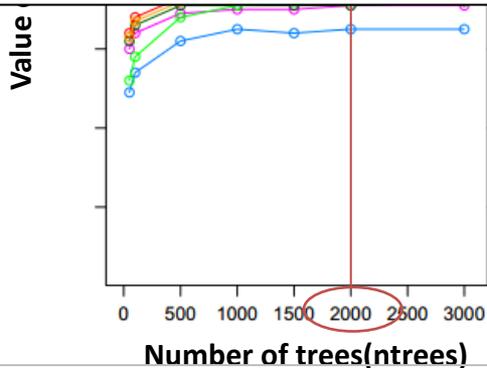
# First pass Feature Selection results



$ntrees^* = 2000$

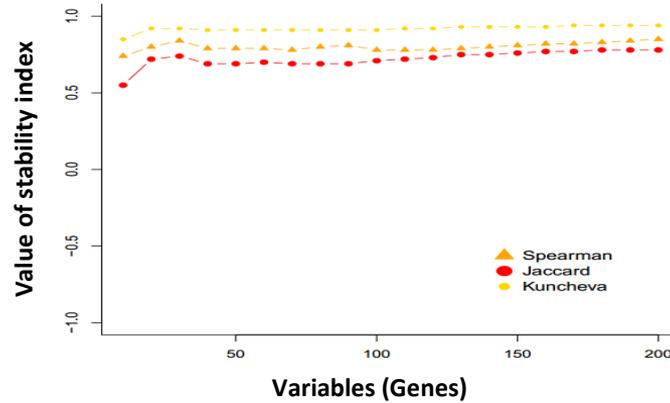
$nVar^* = 200$

~9000 to **200** variables (Genes)



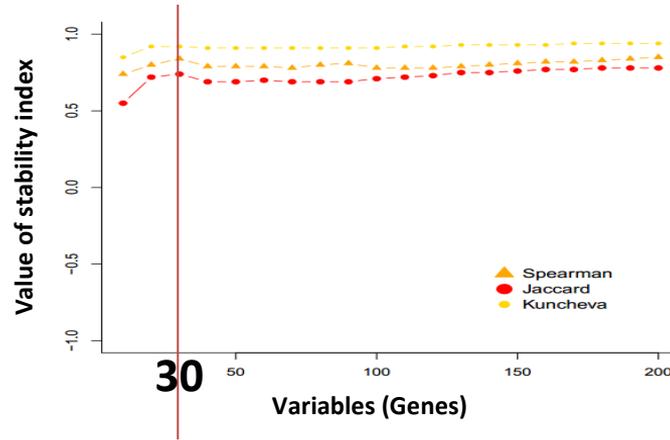
TCGA-BRCA dataset

# Second pass Feature Selection results



TCGA-BRCA dataset

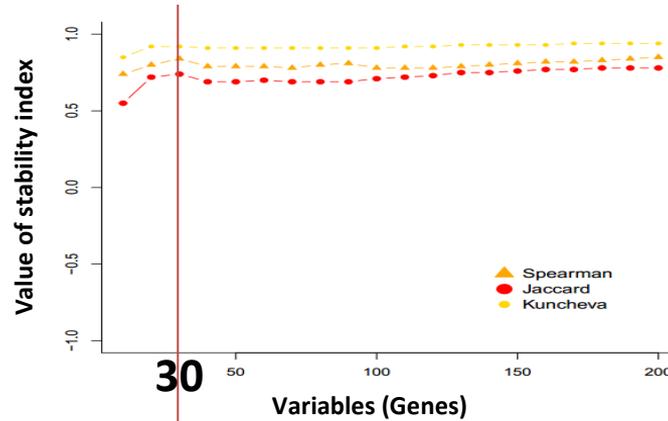
# Second pass Feature Selection results



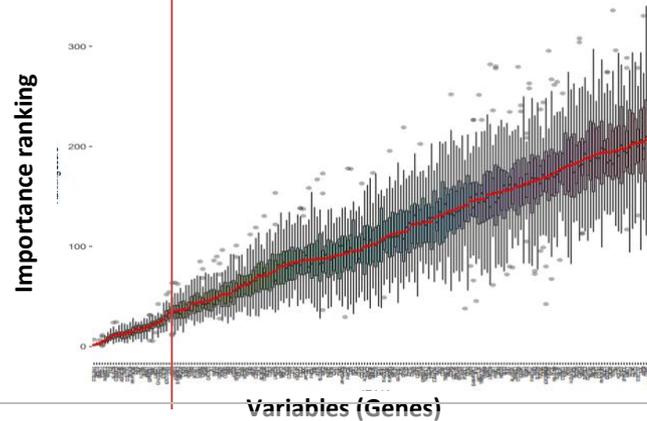
$$nVar^{**} = 30$$

TCGA-BRCA dataset

# Second pass Feature Selection results

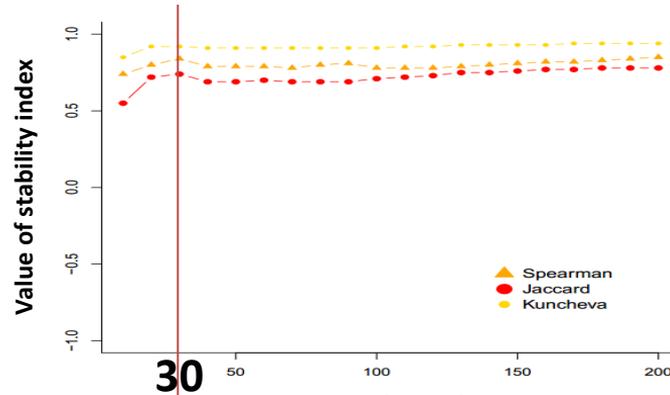


$$nVar^{**} = 30$$



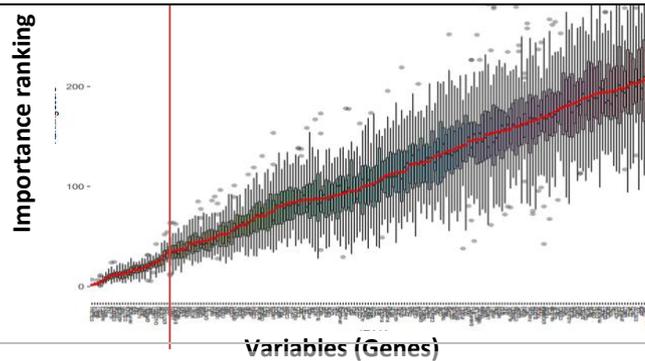
TCGA-BRCA dataset

# Second pass Feature Selection results



$$nVar^{**} = 30$$

~200 to **30** variables (Genes)



TCGA-BRCA dataset

# Pre-Comparison

Assessment of RF parameters & Generation of random combinations

# Pre-comparison

## I- Generation of random combinations (Cancer signatures)

- Multiple predictive models using combinations of different lengths



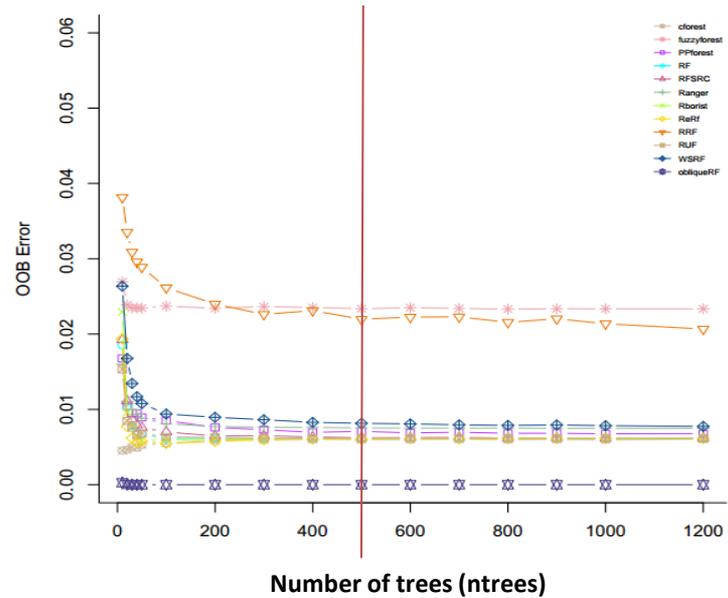
## II- Generation of random training partitions

- 50 random training partitions

training partition = a set of samples used to construct a model

# Pre-comparison

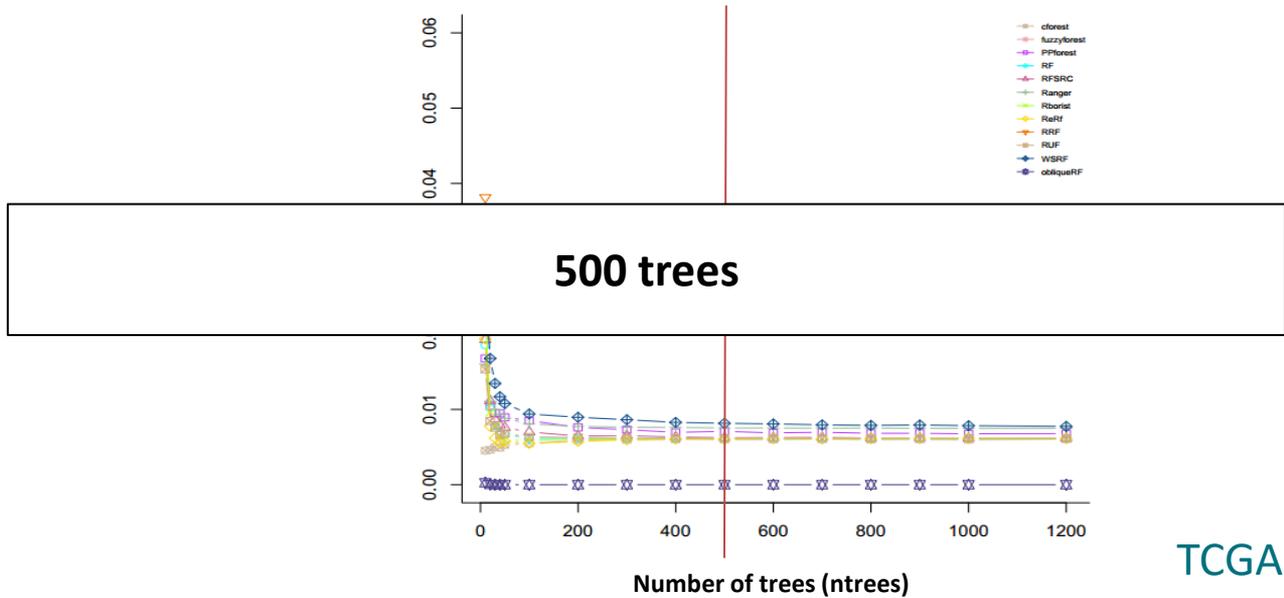
## III- Tuning the parameter *ntrees* for each RF method



TCGA-BRCA dataset

# Pre-comparison

## III- Tuning the parameter *ntrees* for each RF method



TCGA-BRCA dataset

# Summary of step I + step II

Dataset	<i>nVar</i>	<i>nVar</i> *	<i>nVar</i> **	<i>ntrees</i> *	<i>ntrees</i> **	Nbre combinations
TCGA-BRCA	9560	200	30	2000	500	78
TCGA-LUSC	9262	200	10	2000	500	21

# Comparison

Random Forest stability assessment

# Random Forest Method Comparison

- Comparison of RF methods under **same** conditions
- Using **same** random training partitions
- Assessing the **same** signatures
- On computational cores of **same** characteristics

# Random Forest Method Comparison

- For each signature, we'll focus on:
  - **50 resampling** to build the Training and the Validation set.
  - **25 modeling and validations.**

- Analysis of:
  - **Coefficient of Variation of 1,250 models & AUCs**

} 1,250 models & AUCs

# Random Forest Method Comparison

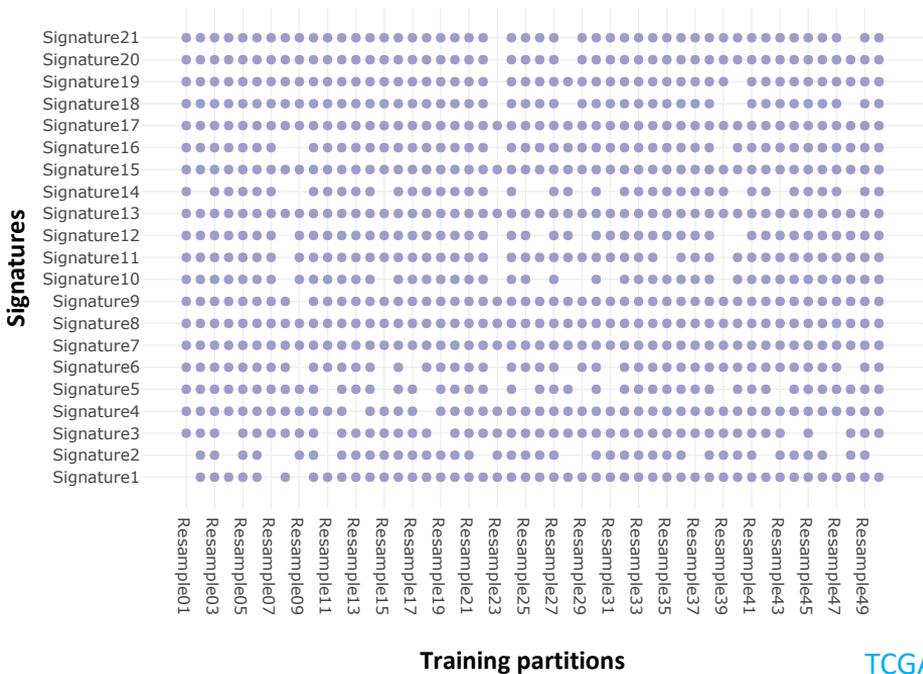
- For each signature, we'll focus on:
  - **50 resampling** to build the Training and the Validation set.
  - **25 modeling and validations.**

- Analysis of:
  - **Coefficient of Variation of 1,250 models & AUCs**

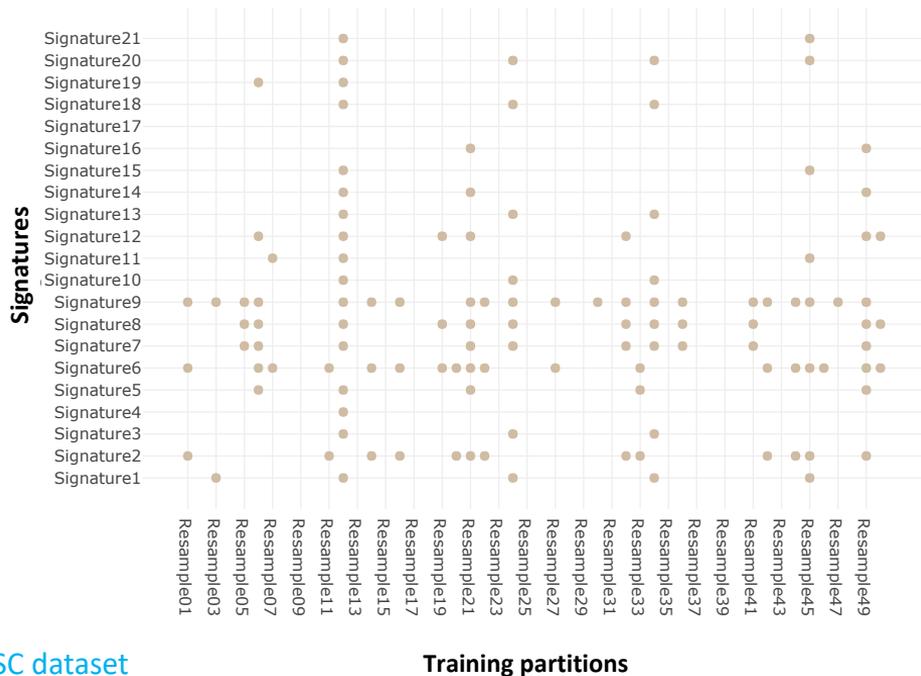
} 1,250 models & AUCs

**Clinics** → **Hyper Stability : CV == 0**

# Hyper Stability discriminates RF methods

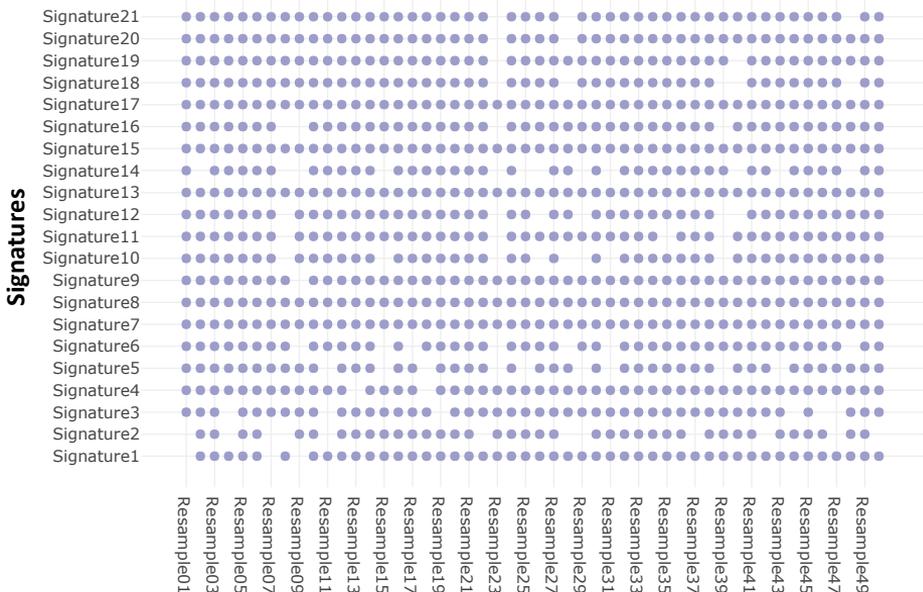


TCGA-LUSC dataset



# Hyper Stability discriminates RF methods

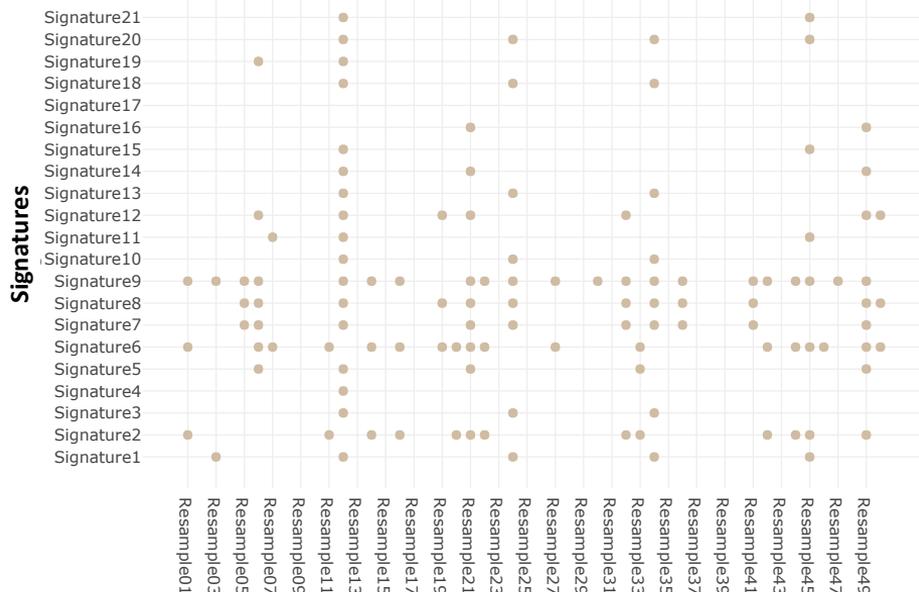
Best Methods



Training partitions

TCGA-LUSC dataset

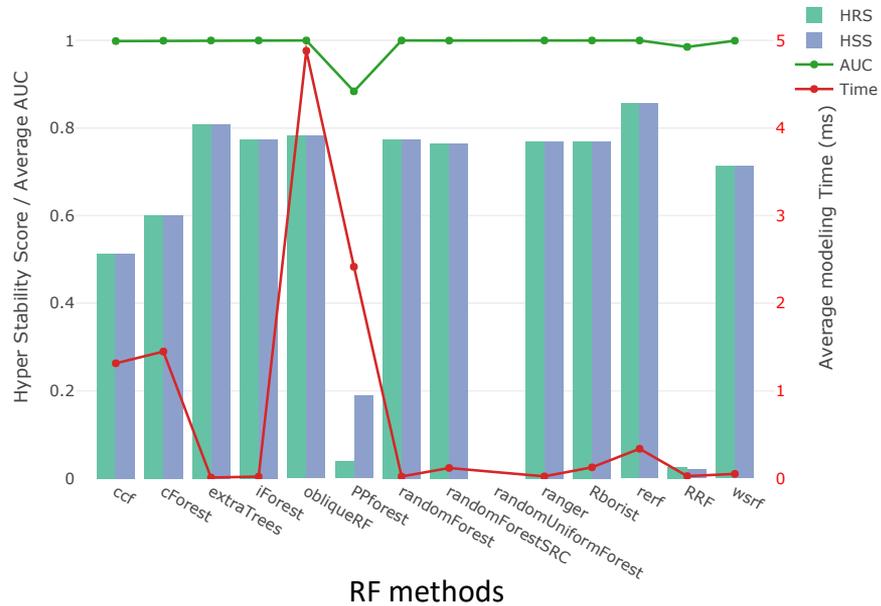
Worst Methods



Training partitions

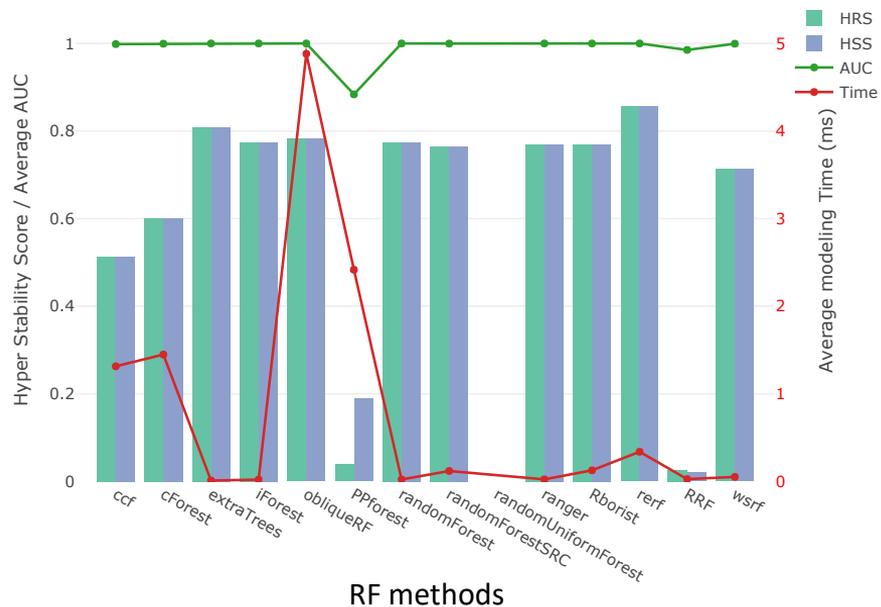
# Hyper Stability Score helps finding the best method(s)

TCGA-BRCA

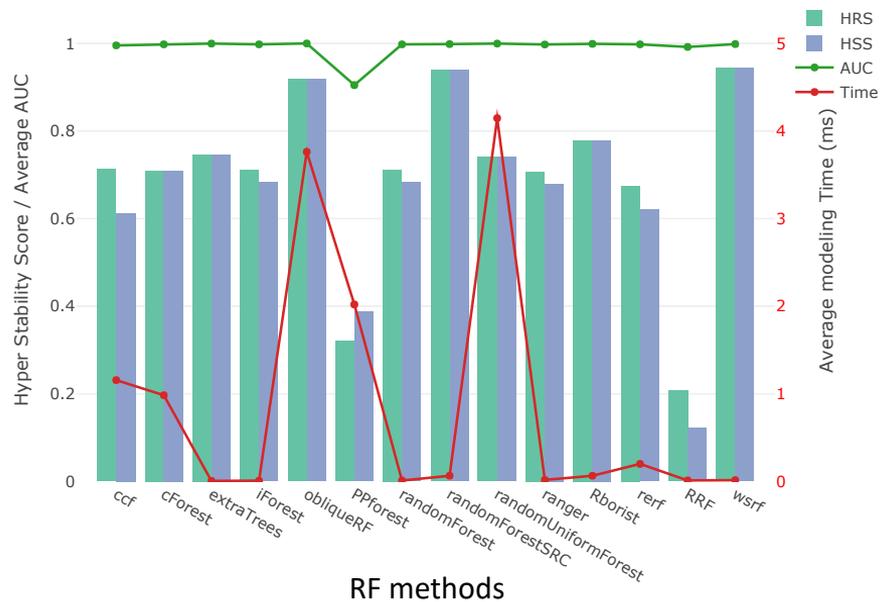


# Hyper Stability Score is dataset dependent

## TCGA-BRCA



## TCGA-LUSC



# Conclusions

- The AUC precision is dataset dependent
  - The Methods are dataset dependent.

- Trade-off:
  - AUC precision (hyper-stability)
  - Average AUC value
  - Modeling Time



Classification of classification methods

Towards robust signatures and predictions

# Perspectives

- Datasets of same dimensions
- Same datasets coming from other platforms: micro-array, ...
- Other machine learning methods

# Acknowledgment

## Human Genetics (GIGA)

- Prof. Vincent Bours, MD, PhD
- Christophe Poulet, PhD
- Corinne Fasquelle, Ir

## Oncology (CHU-Liege)

- Prof. Guy Jerusalem, MD
- Dr. Claire Josse
- Jerome Thiry

## BIO3 Unit (GIGA)

- Prof. Kristel Van Steen

## CBIO, Mines ParisTech, Institut Curie

- Chloe-Agathe Azencott, PhD