Block Forests: random forests for blocks of clinical and omics covariate data

Roman Hornung[1], Marvin N. Wright[2,3]

March, 19th, 2019

[1] **I**nstitute for Medical Information Processing, **B**iometry and **E**pidemiology, University of Munich, Germany

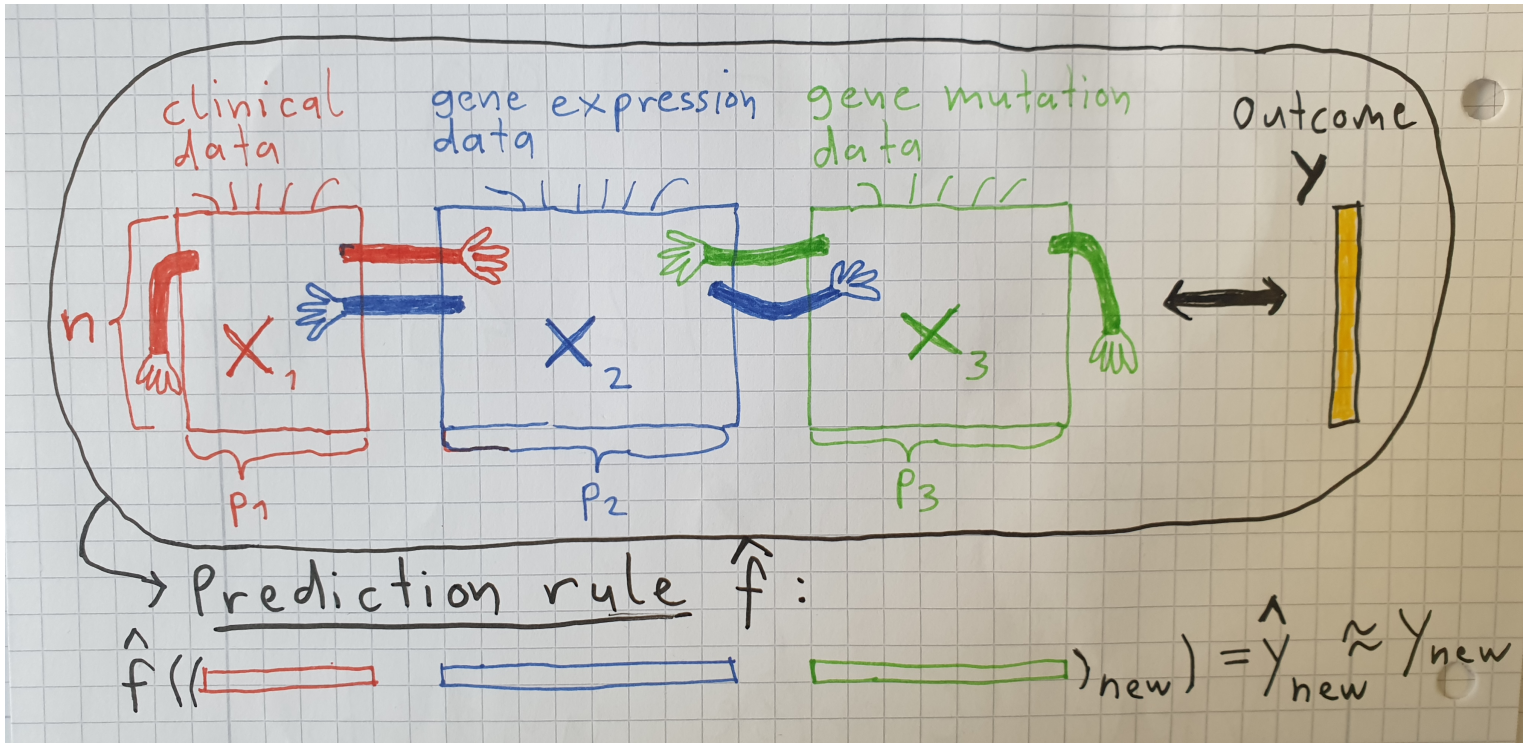
[2] Leibniz Institute for Prevention Research and Epidemiology - **BIPS**, Bremen, Germany

[3] Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark

Introduction

Last few years: more and more **data** available that feature **omics** measurements of **several types for the same patients** (**multi-omics data**)

⇒ **New** possibility: **combine several types** of omics data **for prediction** modeling



Introduction

Multi-omics data have **complex** structures...

- Strongly **overlapping information** between different omics **blocks**
- **Differing levels of predictive information** between **blocks** that depend on the outcome considered
- **Interactions between** variables from different **blocks**

The prediction method **Random Forest** (**RF**) **captures complex dependency structures** between outcome and covariates.

$\Rightarrow$ **Goal of the project:** Develop **RF variant for multi-omics data** that **exploits** the **information** contained in such data by considering their specific structure.

Quick reminder: Splitting in RF

- Each **tree** decision rule in a RF performs a **series of binary decisions**, where **each decision** is obtained using a threshold (**split point**) in the values of **one of the covariates**.

- In the **construction of a RF**, a **split** point is obtained by first **randomly drawing** a number 'mtry' (default: $\sqrt{p}$) of all **covariates** and second **determining** that **split** in the drawn covariates that is **best according to a split criterion**.

Five potential RF variants for multi-omics data

Potential RF variants differ with respect to **split selection**:

- "**VarProb**":
  1. **Sample** $\sum_{m=1}^{M} \sqrt{p_m}$ variables, where a variable from block m is sampled **with probability prob$_\mathbf{m}$**.
  2. **Split** according to the highest split-criterion value.
- "**SplitWeights**":
  1. **Sample** $\sum_{m=1}^{M} \sqrt{p_m}$ variables with equal sampling probabilities.
  2. **Split** according to the highest weighted split-criterion value **using block-specific weights w$_\mathbf{m}$** ($m = 1, \ldots, M$, $\max\{w_1, \ldots, w_M\} = 1$).
- "**BlockVarSel**":
  1. **Sample** $\sqrt{p_m}$ variables **from block m**, $m = 1, \ldots, M$.
  2. Perform **step 2 of SplitWeights**.

Potential RF variants for multi-omics data

- "**RandomBlock**":
  1. **Sample one block m*** from the M blocks, where block m has selection probability $\widetilde{prob}_m$ ($\sum_{m=1}^{M} \widetilde{prob}_m = 1$)
  2. **Sample** $\sqrt{p_{m^*}}$ variables **from block m***.
  3. **Split** according to the highest split-criterion value.

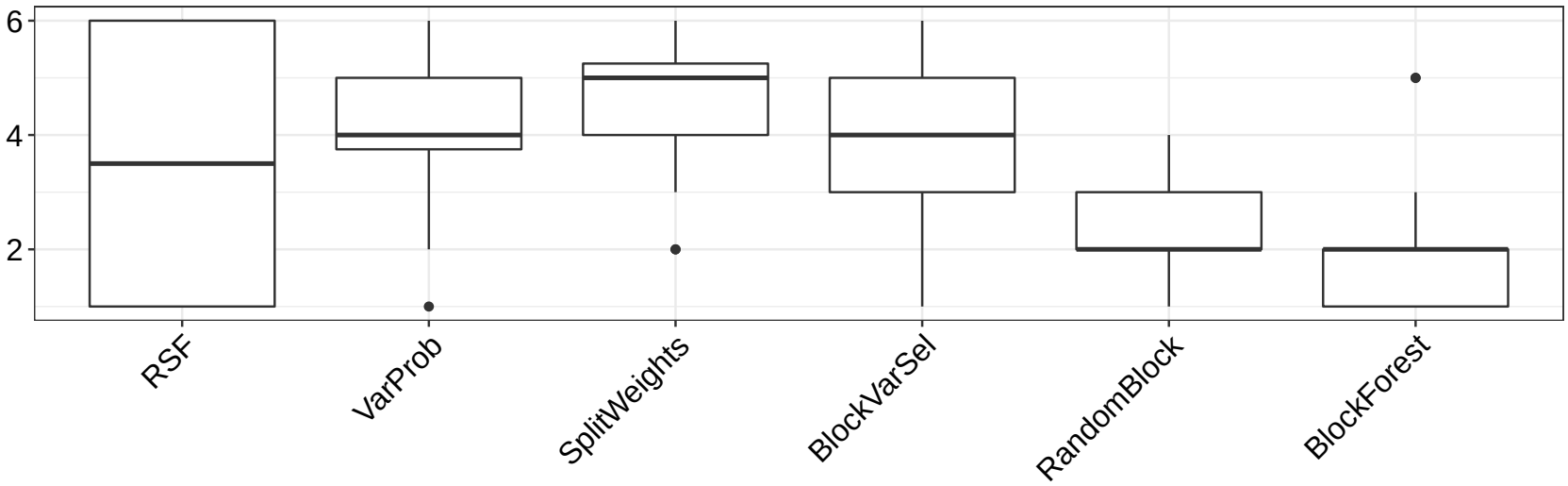
- "**BlockForest**":
  1. **Sample** each **block with probability 0.5**.
  2. **Perform steps 1 and 2 of BlockVarSel** considering only the sampled blocks.

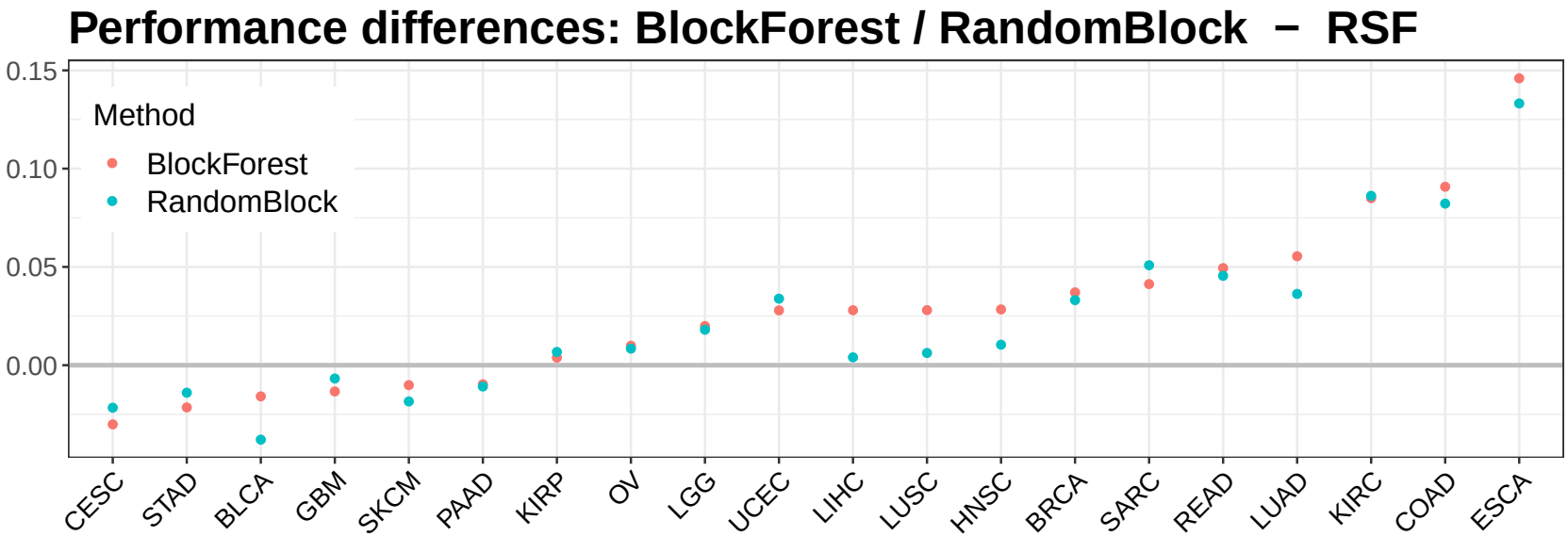
The **tuning parameter values** are **optimized** on the training data **by** repeatedly considering different candidate values and using the candidate values associated with the **smallest out-of-bag error**.

Comparison study using real multi-omics data sets - Design

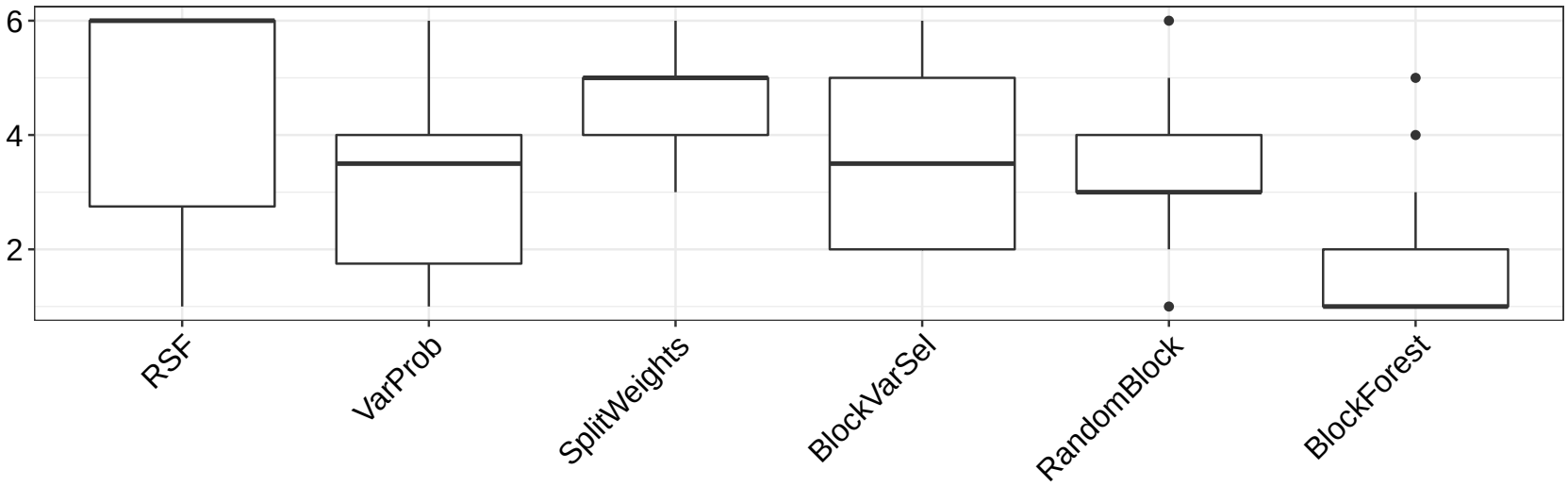
- Data:
  - **20 data sets** with **survival outcome** downloaded from the The Cancer Genome Atlas (**TCGA**) database
  - Each data set features patients of a different cancer type.
  - **Five blocks**: clinical covariates ($p < 10$), copy number variation, methylation, miRNA, mRNA
- Study design:
  - **Six compared methods**: five RF variants, Random Survival Forest (**RSF**, reference method)
  - **Two settings**: 1) "**multi-omics case**": use all available blocks for each data set; 2) "**clinical+RNA case**": use only the clinical block and the RNA block.
  - **Performance assessment**: Harrel's **C index** values estimated using five times repeated 5-fold **cross-validation**

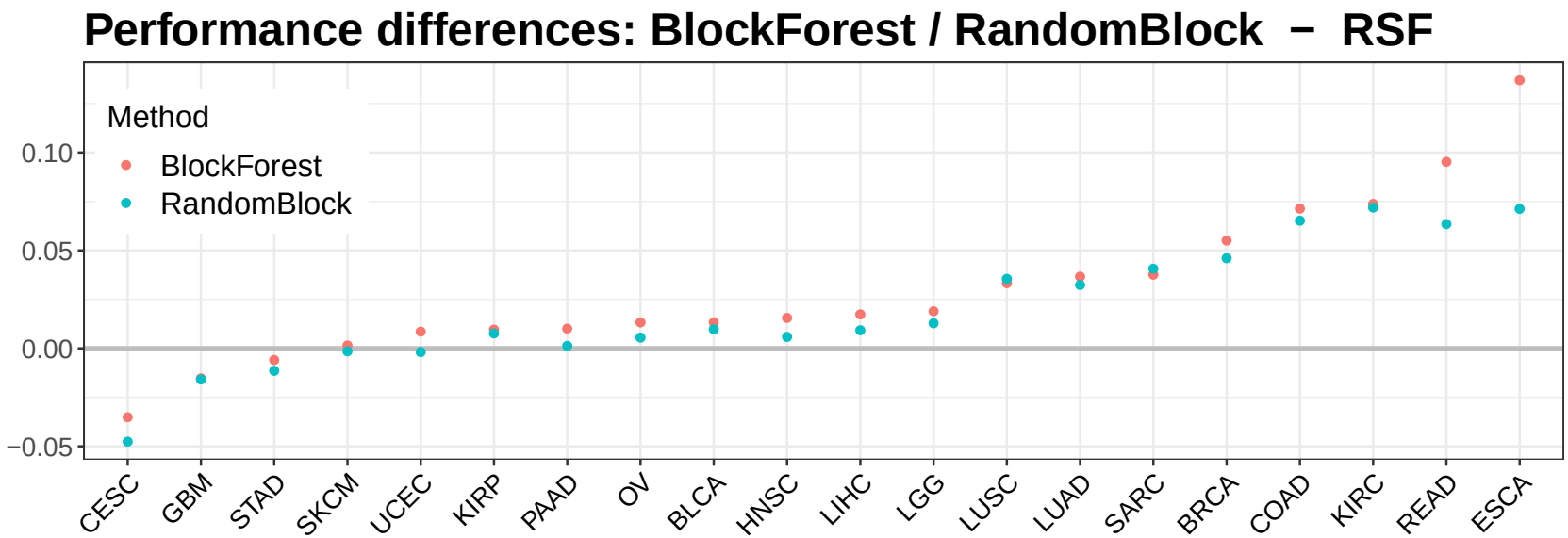
Comparison study using real multi-omics data sets - Results: multi-omics case





Comparison study using real multi-omics data sets - Results: clinical+RNA case





Further results & Discussion

- **Variant BlockForest significantly better than RF** in both settings (paired t test; adjusted P values 0.027 (multi-omics) and 0.010 (clinical+RNA))
- Best methods, BlockForest and RandomBlock, both **randomize the block choice - tackle information overlap** between the blocks.
- Block-specific **weighting** in particular **important** with respect to the **clinical block** - small number of variables, but high prognostic relevance

Outlook

- Performances in the clinical+RNA case, in general, slightly better than in the multi-omics case - Using **clinical plus RNA information may often be sufficient**.
- **CRAN/github R package** `blockForest` (**fork** of RF package `ranger`): **all 5** considered **variants** for binary, survival, and metric outcome (default variant: "BlockForest").
- **Technical Report**: Hornung, R., Wright, M. N. (2018). Block Forests: random forests for blocks of clinical and omics covariate data. Technical Report No. 219, Department of Statistics, LMU.

References and **thank you for your attention!**


Breiman, L. (2001).
Random forests.
Machine Learning **45,** 5–32.


Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008).
Random survival forests.
The Annals of Applied Statistics **2,** 841–860.


Wright, M. N. and Ziegler, A. (2017).
ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.
Journal of Statistical Software **77,** 1–17.


Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B., and Ma, S. (2015).
Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA.
Briefings in Bioinformatics **16,** 291–303.


Huang, S., Chaudhary, K., and Garmire, L. X. (2017).
More is better: Recent progress in multi-omics data integration methods.
Frontiers in Genetics **8,** 84.

TCGA: `https://cancergenome.nih.gov`

R package: `https://github.com/bips-hb/blockForest`