REPOTRIAL

An in silico-based approach to improve the efficacy and precision of drug REPurpOsing TRIALs for a mechanism-based patient cohort with predominant cerebro-cardiovascular phenotypes

D1.6 Software for detection of pathway co-enrichment for extracting of patient(group)-specific pathways also enriched in target sites of effective drugs

Project acronym:	REPO-TRIAL
Grant Agreement:	777111
Project Duration:	01 February 2018 – 31 January 2023 (60 months)
Version:	V1
Date:	30/08/2019
WP Leader:	Jan Baumbach (11 TUM)
Authors:	Elisa Anastasi (03 UNEW), Jan Baumbach (11 TUM), Charlotte Brown (03 UNEW), Simon Cockell (03 UNEW), Spencer Du (03 UNEW), Keith Flanagan (03 UNEW), Tim Kacprowski (11 TUM), Sepideh Sadegh (11 TUM), Anil Wipat (03 UNEW), Gihanna Galindez (11 TUM), James Skelton (03 UNEW)
Due date of deliverable	31/08/2019 (Month 19)
Actual submission date	22/09/2019

Abbreviations

UM	Universiteit Maastricht
UNEW	University of Newcastle upon Tyne
UKE	Universitaetsklinikum Essen
MHH	Medizinische Hochschule Hannover
UMCU	Universitair Medisch Centrum Utrecht
BIOCRATES	Biocrates Life Sciences AG
SomaLogic	Somalogic Limited
HMPC	Mucke Hermann
concentris	concentris Research Management GmbH
TUM	Technische Universität München

Table of Contents

1. Executive Summary

Systems medicine requires the accurate identification of genes and pathways that mechanistically define a disease phenotype. While biomarker signatures derived from single-omics analyses have proven useful for disease diagnosis and prognosis, they rarely explain the underlying mechanism. We developed Grand Forest, an ensemble learning method that extends Random Forests and integrates experimental data with molecular interaction networks to discover relevant endophenotypes and their defining molecular subnetworks. These endophenotypes represent patient-specific subnetworks or pathways. Through it's web interface, Grand Forest also provides functionality for follow-up analyses and identification of drug targets enriched in identified endohpenotypes. The Grand Forest algorithm will be integrated into our Cytoscape App CypoTrial for straight-forward access to the wealth of data and networks we host in the REPO-TRIAL DB. A manuscript describing GrandForest is currently under review in Systems Medicine and this report is largely based on this manuscript.

2. Deliverable report

We developed Grand Forest (Graph-guided Random Forest), a novel kind of module discovery method. A key aspect is the integration of omics data with networks. When building the forest of decision trees, each tree is enforced to represent a randomly sampled subnetwork from the whole network and consecutive splits within the tree have to be direct neighbors in said subnetwork. The estimated feature importance then allows to extract a highly connected gene set that explains the phenotype under scrutiny. The subnetwork induced by the most important genes is then extracted and returned as result. Figure 1 provides an overview of the unsupervised algorithm for *de novo* endophenotyping.

Grand Forest is freely available at https://grandforest.compbio.sdu.dk where we provide the source code, a package for the R programming language and an easy-to-use online analysis platform. The web server allows users to upload a gene expression data set and analyze their data using two different workflows: an supervised workflow and an unsupervised workflow. Users can either use one of the provided gene interaction networks or upload their own network. Enrichment analysis tools are provided for both workflows, to enable searching for overrepresented Gene Ontology terms, pathways and disease associations among the extract genes. Even the extraction and visualization of networks of drugs and miRNAs targeting the genes in a module, to search for potentially druggable targets, is possible. See Figure 2 for a graphical overview. We will also integrate the developed algorithm into our Cytoscape App CypoTrial for easy access to the wealth of data and network we host in the REPO-TRIAL DB.

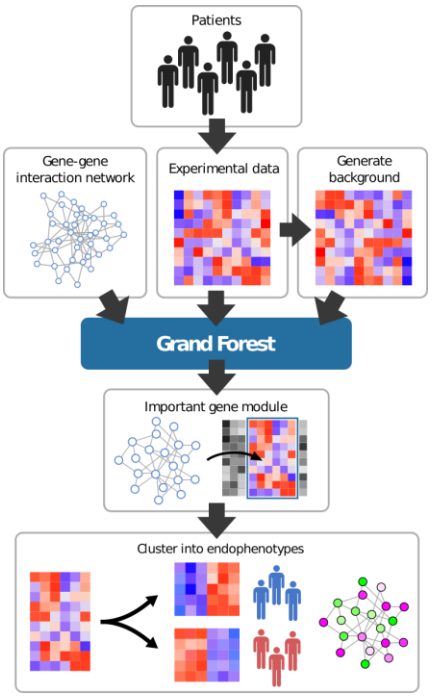
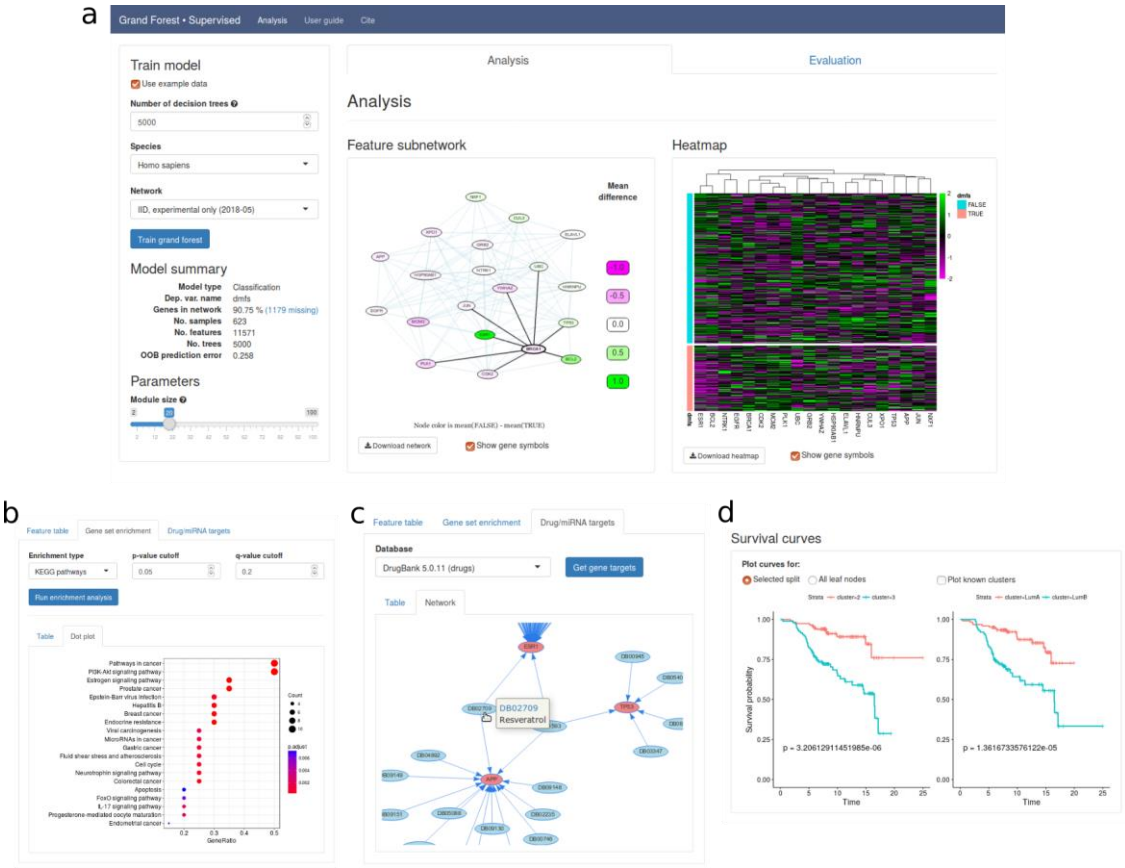


Figure 1: Overview of the unsupervised Grand Forest workflow. A model is trained to recognize unlabeled patients from a generated background distribution. From the trained model highly informative genes are then selected and used to stratify the patients into groups with different endophenotype.

Figure 2: Overview of key features of Grand Forest web server. (a) Main panel of the supervised analysis. (b) Dot plot visualizing enrichment results. (c) A drug (blue nodes) -target (red nodes) network. (d) Survival curves comparing de novo endophenotypes to known disease subtypes.



3. Tables and other supporting documents where applicable and necessary

n/a

4. Conclusion

We developed a new method for disease-gene module discovery by integrating genomic profiling data with molecular interactions networks. We also introduce the first network-based *de novo* endophenotyping methodology, allowing analysis on unlabelled data. We provide a comprehensive web server in order to make our methodology easily available to researchers and will integrate it into the REPO-TRIAL DB framework.

5. Acknowledgement and Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

This report reflects only the author's views and the European Union is not liable for any use that may be made of the information contained therein.