European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.13 – Document describing the analytical tools implemented and additional requirements that have been requested for future implementation

Due date of deliverable: 30th September 2017

Actual submission data: 20th September 2017

Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including Commission Services)	
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

DELIVERABLE INFORMATION

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	115446
Document	
Deliverable number:	D4.13
Deliverable title:	Document describing the analytical tools implemented and additional requirements that have been requested for future implementation
Deliverable version:	
Due date of deliverable:	September 30th 2017
Actual submission date:	September 20th 2017
Leader:	Reinhard Schneider, Manfred Hendlich
Editors:	Jay Bergereon, Neil Fitch
Authors:	Mansoor Saqi; Irina Balaur; Bertrand De Meulder; Diane Lefaudeux; Alexander Mazein; Charles Auffray; Wei Gu; Venkata Satagopam; Sascha Herzinger; Denny Verbeeck; Adriano Barbosa; Manfred Hendlich; Reinhard Schneider
Reviewers:	Jay Bergereon
Participating beneficiaries:	EISBM, UL
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Manfred Hendlich and Reinhard Schneider
Work Package participants:	Adriano Barbosa; Wei Gu; Venkata Satagopam; Emmanuel Van der Stuyft; Francisco Bonachela-Capdevila; Bertrand De Meulder; Kavita Rege
Estimated person-months for deliverable:	15
Nature:	Report
Version:	1.14
Draft/Final:	Final
No of pages (including cover):	11
Keywords:	Analytical tools, data curation

CONTENTS

1 Abstract

eTRIKS aims to provide an infrastructure for storage and retrieval of clinical and omics data collected from translational medicine studies, together with associated functionality for analysis and visualisation. The data analysis requirements for translational medicine are wide, and the focus for tools created by eTRIKS has been to widen access to a few commonly used but non-trivial types of analyses. Specifically, core tools have been included for exploration of clinical and omics data, cohort selection, and approaches have been developed for the contextualisation of disease related genes. In addition, a computing infrastructure (the eTRIKS Analytics Environment or eAE) has been developed for managing and analysing large scale data to enable compute intensive analyses to be more accessible to translational medicine researchers.

2 Introduction

In this document, we describe the eTRIKS analytics tools that have been developed during the course of the project together with a discussion of additional requirements that might be implemented in the future.

As described previously (D4.6), provision of analytic requirements has two components, namely, provision of a set of core tools to address common analysis requirements, and provision contextual information to results emerging from analysis workflows. As many analyses in systems medicine are CPU and/or memory intensive, a computational infrastructure to enable these types of analyses to be carried out routinely, the eTRIKS Analytics Environment (eAE) was also developed.

The core tools developed focus on common data types generated in the major eTRIKS supported projects, which currently includes largely clinical and transcriptomics data. These tools need to be robust and have good visualisation capabilities. The tools developed have included: smartR for exploratory data analysis and interactive visualisation, and HiDome, for cohort selection and statistics on high-dimensional data (omics data), as well as accessible interfaces for two commonly used programs, one for correlating networks of co-expressed genes to clinical features (Langfelder & Horvath, 2008) and the other, a fast network based approach to subtype discovery using combined multiple omics data types (Wang et al, 2014).

Exploratory work has been carried out on developing approaches to contextualisation. The problem faced by the researcher is that identification of biomarkers (molecular patterns that are discriminative between disease and control samples for example) usually provides no mechanistic insight into the disease. However, identification of the pathways and processes in which the genes in the molecular signatures are involved, together with other information, may lead to better mechanistic understanding. The approaches developed have been to (i) use a graph database approach to integrate disease specific information with background information (ii) develop detailed, curated pathway based maps of the hallmark

pathways that characterise the disease. Both representations facilitate data interpretation by allowing experimental information to be overlaid onto the associated networks or pathway maps.

Introductory video webinars describing some of these tools are available on the eTRIKS website.

3 Computational Infrastructure for large datasets – the eAE

In order to make compute intensive analyses more accessible to the translational medicine researcher, the eTRIKS Analytical Environment (eAE) has been developed. The eAE provides a modular framework for managing and analysing large scale data, in particular the massive amount of data produced by high-throughput technologies. This framework mandates the development of a scalable, on demand, performant and resilient to failure, solution. The modularity of this framework also enable us to add new components (public or private) or replace ageing ones by more performant or proprietary ones. Since flexibility and modularity are the core properties of cloud computing, this makes cloud computing an ideal candidate for supporting this need for an almost organic growth. The eAE relies on mature open source technologies with strong supporting communities, such as tranSMART, OpenStack, Jupyter and the Apache Spark stack. Some of the problems for which the eAE is being used include: GWAS analysis on large scale data (>1TB); Sleep scoring and staging using Deep Learning methods (GPU based) - in this use case the exploration of the parameters and cross validation mandated the batching of 31 GPU jobs concurrently; Unbiased modelling using high frequency model generation; Unified environment enabling reproducible research and open science; Major speed up on statistical analysis (CF general statistics workflow as an illustration); Large scale omics analysis

4 Core Tools

4.1 HiDome

HiDome is a tool to allow selection of patient groupings (creation of cohorts) by filtering on particular values of the associated high dimensional datasets, for example high and low values of the expression of a given gene. Since these values depend highly on the technology, normalization method and type of high dimensional data, researchers might not generally have an intuition towards the range of these values. Therefore HiDome shows a histogram of the chosen dimension to aid the creation of such a filter, as shown in Figure 2. Multiple filters can be employed to include additional high dimensional datasets as well as clinical measurements (e.g. age of the patient). HiDome can also be used in the Summary Statistics and Grid View in tranSMART to show and compare values for a specific biomarker upon dragging a high dimensional node into the window, thus facilitating hypothesis generation (see Figure 1a and 1b).

4.2 smartR

smartR uses web technologies such as AngularJS and D3.js to provide interactive visual analytics to allow data exploration in tranSMART, of both clinical and omics data types. The type of workflows include: heatmap analyses for gene expression datasets onto which information such as disease status can be overlaid thereby allowing the clusters to be explored in the context of phenotype; correlation analyses; selecting subsets of genes for which pathway enrichment analyses can be performed; survival analysis; logistic regression, mapping patients across studies, connecting to Ingenuity Pathway Analysis Tools, mapping sequence variants across patient populations. Importantly, smartR also provides a framework that allows easier implementation of new workflows. smartR has been published recently in Bioinformatics (Herzinger et al., 2017). Some examples of smartR output is shown in Figure 2(a-d).

4.3 Shiny Apps

In order to make some commonly used tools easier to use by non bioinformaticians the Shiny web application framework for the R programming language was employed. The shiny web applications are designed to allow complex bioinformatics analyses to be performed without any code input from the users, allowing researchers with no coding experience at all to use them. User friendly interfaces were built with Shiny for three tools (i) SNF (Similarity Network Fusion) a tool for disease subtype discovery through integration of multiple omics data types (Wang et al, 2014) (ii) WGCNA (Weighted Gene Coexpression Network Analysis (Langfelder & Horvath, 2008) a widely used tool for finding clusters of correlated genes from gene expression data and correlating these clusters to clinical traits. The Shiny application for WGCNA allows users to upload their own molecular measurements and related clinical measurements, perform the WGCNA analysis semi-automatically and extract meaningful results easily. (iii) GSVA Gene Set Variation Analysis (Hanzelmann et al, 2013) a tool that calculates the variation in the expression of gene sets (signatures) across a population. It is an unsupervised approach that allows us to observe the variation in the activity of a gene set of interest, over an entire sample population. It produces an enrichment score per gene set and sample, which varies between +1 and -1. A value of +1 indicates that all genes in the gene set of interest are expressed at the top end, while -1 that all genes are expressed at the bottom end of the spectrum of gene expression in the entire expression dataset. For example, Figure 3 below, visualises that result of applying GSVA to U-BIOPRED sputum gene expression data, using a gene expression signature of Th2 cells activation. The GSVA application allows the user to submit their own signatures of interest and apply GSVA analysis, without a need for programming background.

4.4 Galaxy workflows

Most bioinformatics analyses involve multiple steps such as data pre-processing, statistical analysis, and visualisation, with each step involving many separate operations. The steps can be captured in a pipeline or workflow. Reusable workflows describing the sequence of component computational steps necessary for an analysis are important in bioinformatics as they facilitate reproducibility of analyses. They also insulate a non-bioinformatics end user from the details of the component methodological processes involved in getting the final output. The use of Galaxy

workflows (Afgan et al, 2016) to analyse data stored in the tranSMART platform has been investigated. A marker selection workflow, for example, can capture all the steps required in identifying genes differentially expressed between disease samples and healthy controls, from transcriptomics data sets. The workflow of a pipeline can specify the data re-processing steps such as quality control, normalisation as well as the statistical tests to be used and various steps in the workflow can be configured.

5 Tools for contextualisation

Two approaches have been taken to contextualisation of experimental results emerging from systems medicine studies, one using graph based data integration and the other using a pathway-centric community driven approach that integrates expertly curated disease hallmark paths.

5.1 Graph-based data integration

This approach involves mapping disease implicated genes on a network of background knowledge. The popular graph database Neo4j, has been used to integrate multiple heterogeneous types of data such as protein interactions, sequence similarity, pathway membership, disease-gene association, and drug target associations. Neo4j provides a robust persistence mechanism and has powerful functionality (the Cypher query language) for finding connections between entities (such as genes, pathways, diseases) using graph traversal methods. Network neighbourhoods of disease associated genes can be explored visually, and the framework allows new data types to be added with relative ease. We have made the Java code used to parse the data and populate the graph database available on GitHub.

5.2 Construction of hallmark disease pathways

These are curated by domain experts and represented as maps in standard computer readable formats (generally SBGN, the Systems Biology Graphical Notation). MINERVA is a platform that was developed (outside of eTRIKS) at LCSB which uses the Google maps API for envisioning of pathway diagrams that allows experimental data to be explored in the contact of known molecular pathways and processes). Work has been carried out in eTRIKS to link tranSMART to MINERVA to provide a more seamless framework for hypothesis generation. Community building efforts are underway to establish a framework for integrated, expertly curated signalling, metabolic and gene regulatory pathways relevant to particular diseases (www.disease-maps.org). A pipeline for automatic merging of pathway modules has been developed. The pipeline was successfully applied for developing the AsthmaMap, where known asthma implicated pathways were translated into the SBGN Activity Flow (AF) standard, and automatically merged, additionally curated and then presented in yEd Graph Editor. This approach can be applied for designing various disease maps, combining pathway modules from different pathway databases. As part of the Disease Maps Project (http://disease-maps.org/), there is a plan to link all the maps to MINERVA and tranSMART as was done previously for the Parkinson's Disease Map (Fujita et al., 2014).

6 Additional requirements

The provisioning of tools in a complex translational medicine informatics platform involves a trade-off between offering a wide range of functionality including advanced analytical methods, and the need for robustness and ease of use. A key feature is the requirement to easily capture subsets of the data and of exporting the subsets in a variety of formats for use with other tools and for detailed analysis by expert bioinformaticians. It is recognised that many users have their own favourite analysis tools, both open source and proprietary. There are however a number of commonly used bioinformatics analyses that, although methodologically straightforward, are complex for the non-expert data analyst to perform, and inclusion of these would empower experimental researchers, and help them to 'suggest the next question'.

The tools developed by eTRIKS need to be extended to support other data types. Data types that have been supported to date have been largely those associated with clinical attributes and gene expression reflecting the types of data currently available in the supported projects and in the public domain.

7 References

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. Nucleic Acids Res. 2016 Jul 8;44(W1):W3-W10

Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, Crespo I, Perumal TM, Jurkowski W, Antony PM, Diederich N, Buttini M, Kodama A, Satagopam VP, Eifes S, Del Sol A, Schneider R, Kitano H, Balling R. Mol Neurobiol. 2014 Feb;49(1):88-102

Gawron P, Ostaszewski M, Satagopam V, Gebel S, Mazein A, Kuzma M, Zorzan S, McGee F, Otjacques B, Balling R, Schneider R. NPJ Syst Biol Appl. 2016 Sep 22;2:16020

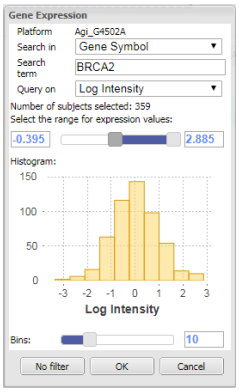
Herzinger S, Gu W, Satagopam V, Eifes S, Rege K, Barbosa-Silva A, Schneider R; eTRIKS Consortium., Bioinformatics. 2017 Jul 15;33(14):2229-2231

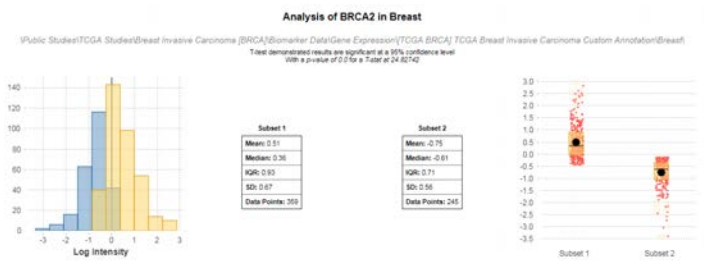
Langfelder P, Horvath S. BMC Bioinformatics. 2008 Dec 29;9:559

Sonja Hanzelmann, Robert Castelo and Justin Guinney *BMC Bioinformatics* 2013, 14:7 Satagopam V, Gu W, Eifes S, Gawron P, Ostaszewski M, Gebel S, Barbosa-Silva A, Balling R, Schneider R. Big Data. 2016 Jun;4(2):97-10

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A Nat Methods. 2014 Mar;11(3):333-7

8 Appendix





Figures(1a) HiDome filter selection dialog (1b) Visualization of a specific biomarker in the high dimensional dataset

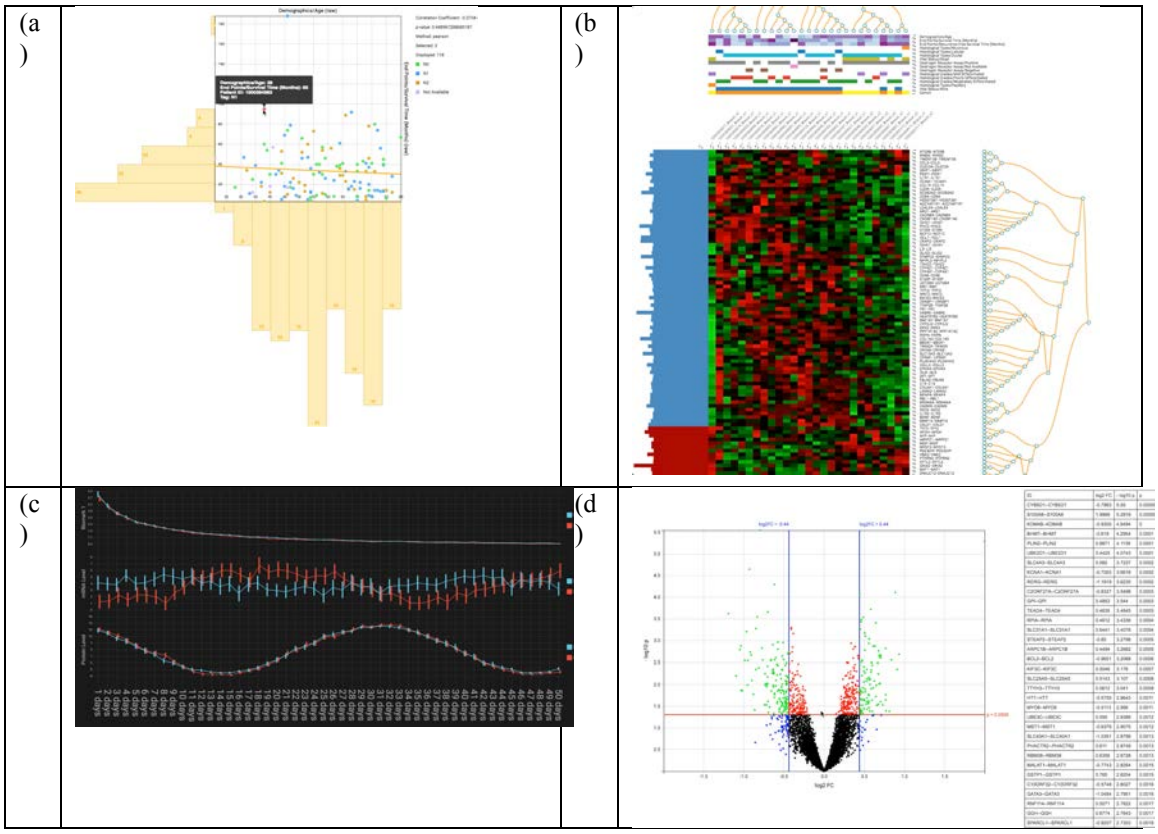


Figure 2(a-d): Examples of smartR output:

(a) **Correlation Analysis.** Shown is a scatterplot with histograms for the respective axes. Statistics and other plot elements adapt dynamically to certain user-triggered events.

(b) **The SmartR Heat Map.** Seen is the interactive heat map in tranSMART based on the breast cancer mRNA data of the GEO study GSE4382. (Sorlie et al., 2003). It is fully sortable and contains many interactive elements.

(c) **Line graph.** Shown is a plot that visualizes machine generated time series data. Besides many visual helpers that are triggered by mouse-over events, this visualization has a manual sortable x-axis and different methods for defining the shown timelines.

(d) **Volcano plot.** Shown is the widely known volcano plot that has been enhanced by some dynamic elements. The limiters are drag-enabled and trigger an update for the right-hand table displaying the most significant genes.

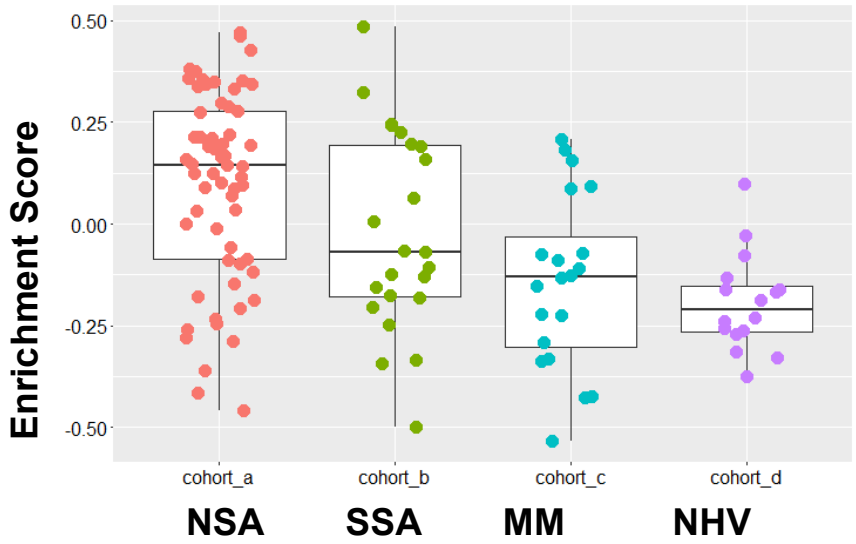


Figure 3: GSVA analysis of sputum gene expression, using a signature of Th2 cell activation. Each dot corresponds to the enrichment score of a study participant, for the aforementioned signature. NSA: Non-smoking severe asthma, SSA: smoking severe asthma, Mild/moderate asthmatic, NHV: Normal healthy volunteers. There is evident increase in the enrichment of the Th2 gene signature in subpopulations of the severe asthma sufferers (NSA and SSA).