SOUND
Statistical Multi-Omics Understanding

Project Deliverable

Project acronym: **SOUND**	GA number: **633974**
Project title: **Statistical Multi-Omics Understanding of Patient Data**	
Funding Scheme: Collaborative Project (H2020-PHC-2014-2015/H2020-PHC-2014-two-stage) Health, novel medical developments	
Project start date: **01 September 2015**	Duration: **36 months**
Project's coordinator: Dr Wolfgang Huber (European Molecular Biology Laboratory, Heidelberg)	

D1.2 Open-source software implementing end-to-end analysis and integrating the results from T1.1-1.5

Due date of deliverable: Month 36 – 30.08.2018
Actual submission date: 30.08.2018

Organization name of lead contractor for this deliverable: European Molecular Biology Laboratory (EMBL)
Organization name of other involved partners: DKFZ
Personnel involved: Junyan Lu, Jennifer Hüllein, Britta Velten and Wolfgang Huber (EMBL), Michelle da Silva Liberio, Sebastian Scheinost and Thorsten Zenz (DKFZ)

Project co-funded by the European Commission within the H2020 Program (2015-2018)		
Dissemination Level		
PU	Public	x
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Deliverable description and summary

One of the aims of Work Package 1 was to implement open-source software and a chain of tools, on the basis of the infrastructures and methods developed in the Work Package, for common and challenging tasks such as data pre-processing, quality assessment and control, integration, statistical inference, visualization and publication-quality reporting. Those software and chain of tools were aimed to be easily accessible by bioinformaticians and clinicians with or without coding experience, in order to lower the barrier to entry into this transdisciplinary field of functional profiling of patient cells by providing simple, robust and easy-to-use solutions.

The most important outcome of Work Package 1 is to package, publish and document the statistical tools and platform initially designed for the clinical partner DKFZ (in collaboration with the Heidelberg University Hospital and the National Center for Tumor Diseases Heidelberg) and make them available as open-source software and a chain of tools, which make them easily accessible and deployable beyond the consortium.

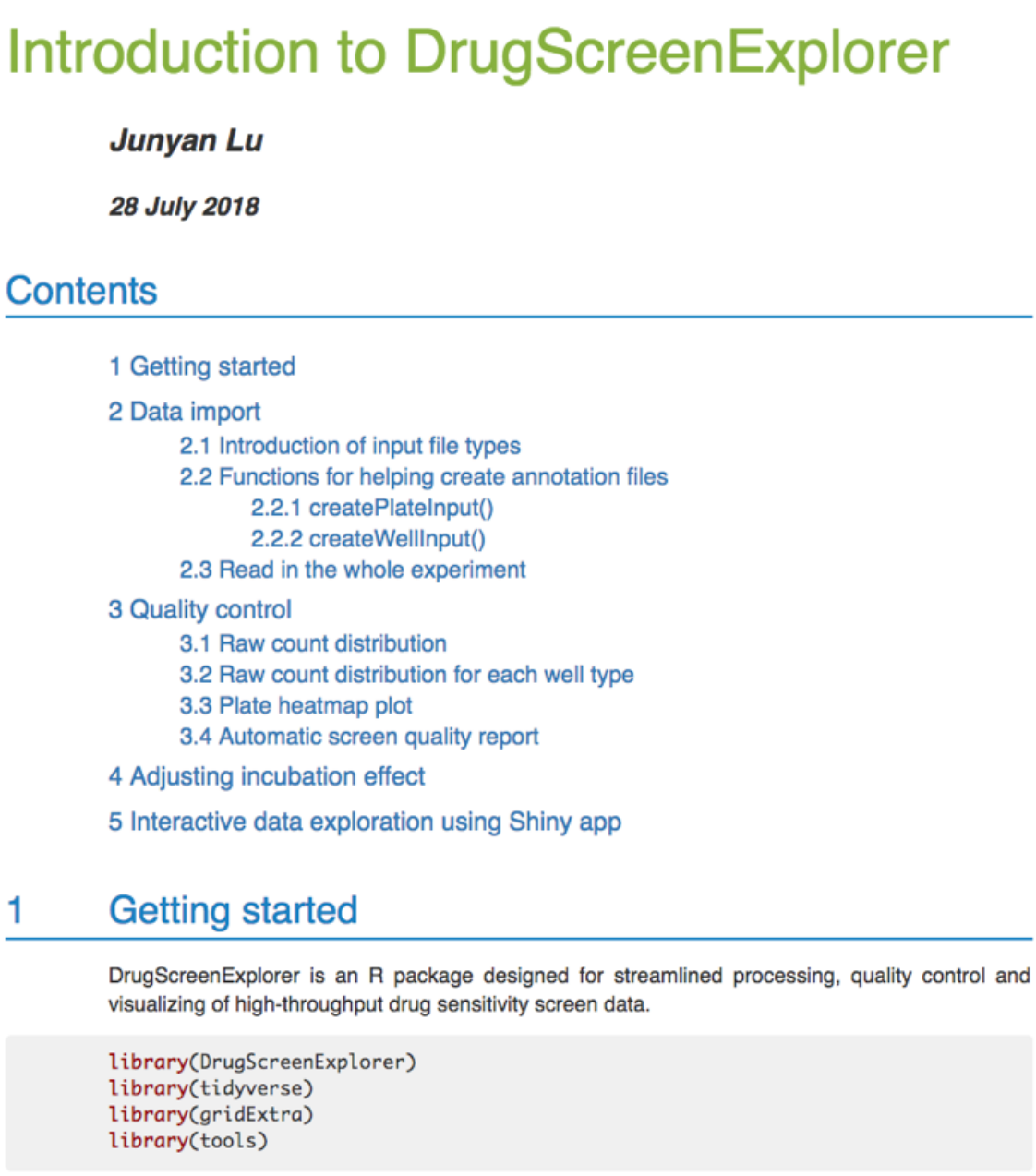
Starting from the R and Shiny based online platform *DrugScreenExplorer* described in Deliverable 1.1, which were mainly used for hosting and presenting data from the clinical partner DKFZ, we extended it into an open-source R package, which features automatic screening of imported data, quality checking, reporting and interactive visualization. In order to tackle the challenge of efficiently managing patient-derived samples and other multi-omics data, we developed the *PatientExplorer*, an R and Shiny based platform, which works as the frontend for the SQL based *TumorBank*. The *PatientExplorer* can seamlessly interact with *DrugScreenExplorer* and facilitate fast clinical translation. In addition, in order to further understand the complex interactions between patient derived sample drug sensitivity phenotype and their genomic, transcriptomic, epigenetic and clinical features, we developed a novel multi-variate statistical model for integrative multi-omics analysis, called Multi-Omics Factor Analysis (MOFA), which identifies common underlying factors (latent variables) shared between the drug response data and the 'omics data types. MOFA (see below) has been packaged as a R/Python package and can therefore be readily used by researchers.

By lowering the entry barrier, the outcome of WP1 expands the number of researchers conducting high-throughput functional assays on patient-derived samples and multi-omics understanding of disease mechanisms. In the following, we describe the software and how it integrates the results of tasks 1.1 to 1.5.

Research progress

1. DrugScreenExplorer package

DrugScreenExplorer is an open-source R package designed for streamlined processing, quality control and interactive visualising of high-throughput drug sensitivity screen data. It can be downloaded at: https://github.com/lujunyan1118/DrugScreenExplorer. The figure below shows the title page of the vignette of DrugScreenExplorer package. The full document is provided at the above location as part of the package. This package is licensed under GNU General Public License (GPL) v3: https://github.com/lujunyan1118/DrugScreenExplorer/blob/master/LICENSE.md.



Introduction to DrugScreenExplorer

Junyan Lu

28 July 2018

Contents

1 Getting started

DrugScreenExplorer is an R package designed for streamlined processing, quality control and visualizing of high-throughput drug sensitivity screen data.

```
library(DrugScreenExplorer)
library(tidyverse)
library(gridExtra)
library(tools)
```

Figure 1: Title page of the *DrugScreenExplorer* package vignette

2. TumorBank and PatientExplorer

Biobanks provide the infrastructure to manage biological material at high quality standards and to correlate corresponding data with clinical outcomes, and are therefore powerful resources for biomedical research. We developed an access-controlled SQL-based database to manage incoming biological specimens and analysis workflows for patient-derived blood cancer samples from a biobank at the National Center for Tumour Diseases (NCT) (A). Published software packages (e.g. dplyr, DBI, pool package, https://db.rstudio.com/) facilitate the interaction with databases in R and allows the integration of sample information into bioinformatics analysis pipelines. In addition, we developed the R/Shiny-based user interface CRUD to Create, Read, Update and Delete (based on https://www.r-bloggers.com/shiny-crud-app/) records that support the researchers in planning and conduction of experiments. This application is available on GitHub under GPL v3 and can be downloaded from https://github.com/Jen-NY/RLIMS.git.

To further facilitate the use of omics data sets, that are large in scale and come in diverse data structures, in precision medicine, we organized data in centralized repositories for raw and processed data and analysis objects. To explore and visualize the results from genetic and molecular analyses, we developed a R/Shiny based web interface called *patientExplorer*. Molecular and genetic annotations coming from different sources were combined into a patient metadata object to increase data coverage and quality by maintenance of a well-documented and curated data set (B). Common identifiers allow the integration of biological features with experimental results (C).

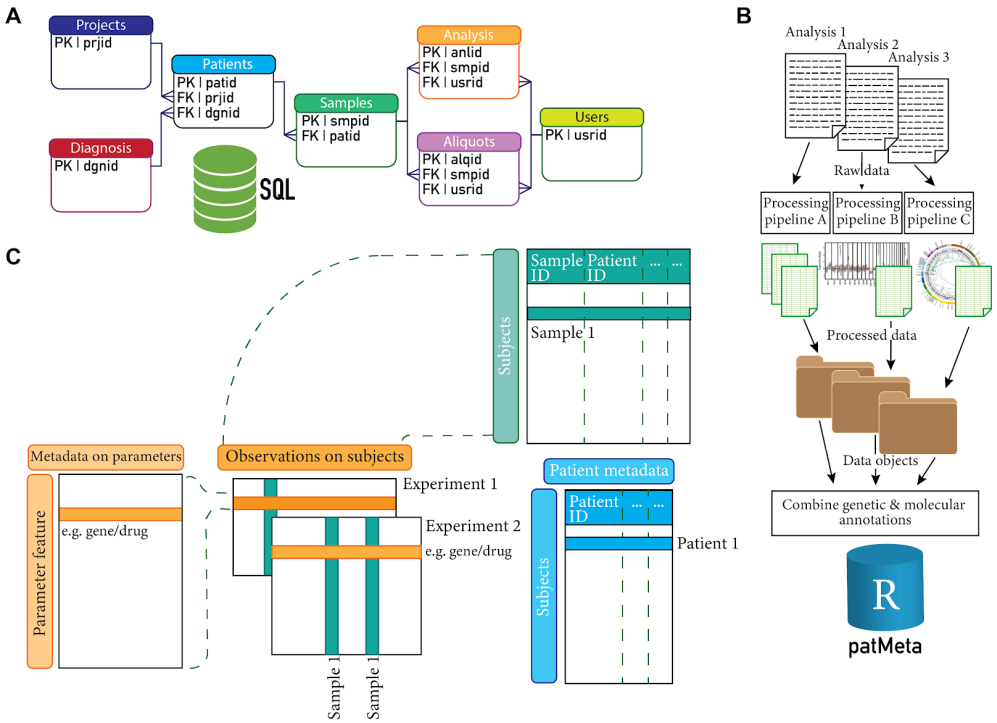


Figure 2: TumorBank and PatientExplorer

3. Multi-Omics Factor Analysis

Multi-Omics Factor Analysis (MOFA) is a computational method that provides a general framework for the unsupervised integration of data from different omic types. Intuitively, MOFA can be viewed as a versatile and statistically rigorous generalization of principal component analysis to multi-omics data. Given several data matrices with measurements of multiple omics data types on the same or on overlapping sets of samples, MOFA infers an interpretable low-dimensional data representation in terms of (hidden) factors. These factors represent the major sources of variation across the samples and can be shared by multiple omic types or omic-type specific. The method is able to cope with missing values, missing assays and can model different data types.

Once trained, the inferred factors and their loadings can be used for a range of downstream analyses, including the visualisation of samples in factor space, the automatic annotation of factors using (gene set) enrichment analysis, the identification of outliers (e.g. due to sample swaps) and the imputation of missing values.

We have made an open-source implementation of MOFA in R and Python available from **https://github.com/bioFAM/MOFA.**

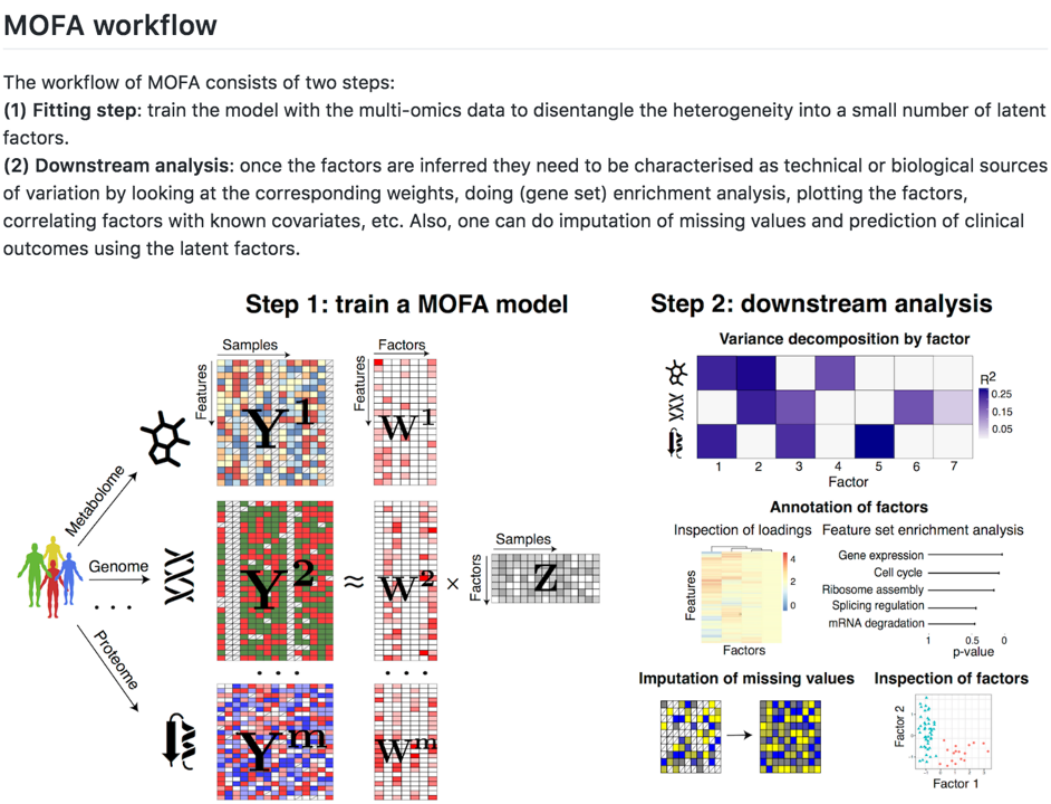


Figure 3: MOFA workflow

Publications

1. Argelaguet,R., Velten,B., Arnol,D., Dietrich,S., Zenz,T., Marioni,J.C., Buettner,F., Huber,W. and Stegle,O. (2018) Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol., 14, e8124.
2. Lu,J., Walther,T., Mougiakakos,D. and Zenz,T. Energy metabolism influences drug sensitivity and is co- determined by genetic variants in CLL. *(Under journal review)*
3. Dietrich,S., Oleś,M., Lu,J., Sellner,L., Anders,S., Velten,B., Wu,B., Hüllein,J., da Silva Liberio,M., Walther,T., et al.(2017) Drug-perturbation-based stratification of blood cancer. J. Clin. Invest., 128.