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)	

D6.1 Open-source software supporting the molecular tumour characterization pipeline and predicting perturbed pathways

Due date of deliverable: Month 18 - 28.02.2017
Actual submission date: 23.02.2017

Organization name of lead contractor for this deliverable: University Hospital Zurich (USZ)
Organization name of other involved partners: ETH
Personnel involved: Mitch Levesque, Phil Cheng, Thomas Thurnherr and Niko Beerenwinkel

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

In collaboration with the group of Prof. Niko Beerenwinkel and BeDataDriven, we developed statistical models and computational tools for data integration with the goal of predicting treatment outcome, based on subclonal diversity, mutation load, immune signatures and specific prognostic genetic lesions. These models will support clinical decision-making based on the molecular profiles of individual, patient derived heterogeneous tumours. The integrated data include the results of our molecular tumour characterization pipeline, TCGA signatures, clinical outcomes, and clinical covariates collected by the University Hospital Zurich (USZ). These data will be interpreted in terms of functional cellular pathways resulting in individual pathway perturbation profiles, which can point to possible treatment options, including approved and potential targeted therapies, or drug combinations targeting several subclones simultaneously.

Research progress

In this action, we developed software to integrate multiple omic datasets and clinical features, starting with the TCGA data as a proof of concept: TCGAbrowser: http://www.tcgabrowser.com, source code http://www.github.com/pcheng84/TCGAbrowser). TCGAbrowser is an open-source web portal for mining and analysing all TCGA tumour data. It is specifically designed for on-the-fly single gene analysis across RNAseq, exome, copy number, DNA methylation, protein and clinical datasets. The tool takes a single gene and splits the samples according to high and low expression and performs survival analysis, differential mutation, copy number, gene expression and DNA methylation analyses. Furthermore, pathway analysis is performed on the differential gene expression analyses will several pathway analysis tools like GSVA and Reactome pathway analysis. The purpose of this tool is (i) to allow biologists to investigate the effect of gene expression phenotypes on survival in human data and (ii) to support biologists with the design of experiments by suggesting clinically relevant candidate genes for knock-down/overexpression experiments in cellular or in vivo disease models.

TCGAbrowser will be finalized as a package in R/Bioconductor[1] with an interactive user interface using shiny. TCGAbrowser has convenience wrappers to a number of third-party R/Bioconductor packages which include limma[1], GSVA[2], survival[3], and Complex Heatmap[4]. The shiny app of TCGAbrowser is a real-time self-service exploration, visualization and analysis platform of gene expression-based patient phenotypes in all cancers and multiple platform data. It holds promise to

[1] http://www.bioconductor.org

become even more versatile by contributions from the authors and the scientific community. Possible extensions are the integration of data from all platforms and all cancer types to facilitate pan-cancer integrative multi-omics data mining for everyone. Specifically, the TCGAbrowser is testing the usability of the MultiAssayexperiment data containers developed by Dr. Martin Morgan in the past, integrating drug gene interaction analysis developed by the group of Prof. Niko Beerenwinkel from WP4, and working to optimize analysis runtimes with BeDataDriven from WP9. This deliverable will be one part of the open-source software for supporting molecular tumour characterization.

References

1 Smyth, G. K. Limma: linear models for microarray data. 397-420 (Springer, 2005).

2 Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14, 7, doi:10.1186/1471-2105-14-7 (2013).

3 A Package for Survival Analysis in S . v. 2.38 (https://CRAN.R-project.org/package=survival . (2015).

4 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847-2849, doi:10.1093/bioinformatics/btw313 (2016).

5 Cheng, P, Dummer, R, Levesque, MP. Data mining The Cancer Genome Atlas for novel insights into melanoma biology and patient survival. Swiss Medical Weekly 28:6, 763 (2015).