Multi-omics data analysis in tranSMART using the Cell Line Use Case dataset

Ward Weistra* [1], Janneke Schoots-van der Ploeg[1], Jochem Bijlard[2], Kees van Bochove[1], Mariska Bierkens[3,6], Sanne Abeln[2], Jan-Willem Boiten[4], Wim van der Linden[5], Jeroen A.M. Belien[3], Gerrit A. Meijer[6], Remond J.A. Fijneman[3,6]

[1] The Hyve, Weg der Verenigde Naties 1, 3527 KT Utrecht, The Netherlands
    The Hyve, Cambridge Innovation Center, 1 Broadway, 14th Floor, Cambridge, MA 02142, United States
[2] Centre for Integrative Bioinformatics (IBIVU), VU University, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands
[3] VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands
[4] CTMM-TraIT, Centre for Translational Molecular Medicine, High Tech Campus 84, 5656 AG, Eindhoven, The Netherlands
[5] Philips Research, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
[6] Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

*E-mail: ward@thehyve.nl. Tel: +31 (0)30 7009713.

We empower scientists by building on open source software

New data types and multi-omics analytics

With the establishment of next generation sequencing and the advent of proteomics and metabolomics in the translational research domain, there is an **increasing need for integrating different multi-omics data types**. In the course of different recent projects, **The Hyve has integrated new multi-omics capabilities in the translational research platform tranSMART** [1] (http://transmartfoundation.org).

In addition to low dimensional data (demographics, clinical, imaging measurements) and mRNA expression data, tranSMART now supports analysis and visualisation of:

- ArrayCGH and miRNA microarray data
- mRNA expression and genomic variants from RNA-Seq and DNA-Seq
- Proteomics and metabolomics data

TranSMART already has multiple ways to **analyze multi-omics data** (figure 1a, b). On top of those the GenePrint analysis (from the cBioportal OncoPrint [2]) has been added to allow integrated analysis of mRNA expression, proteomics, genomic variants and aCGH in a single visualization (figure 1c).
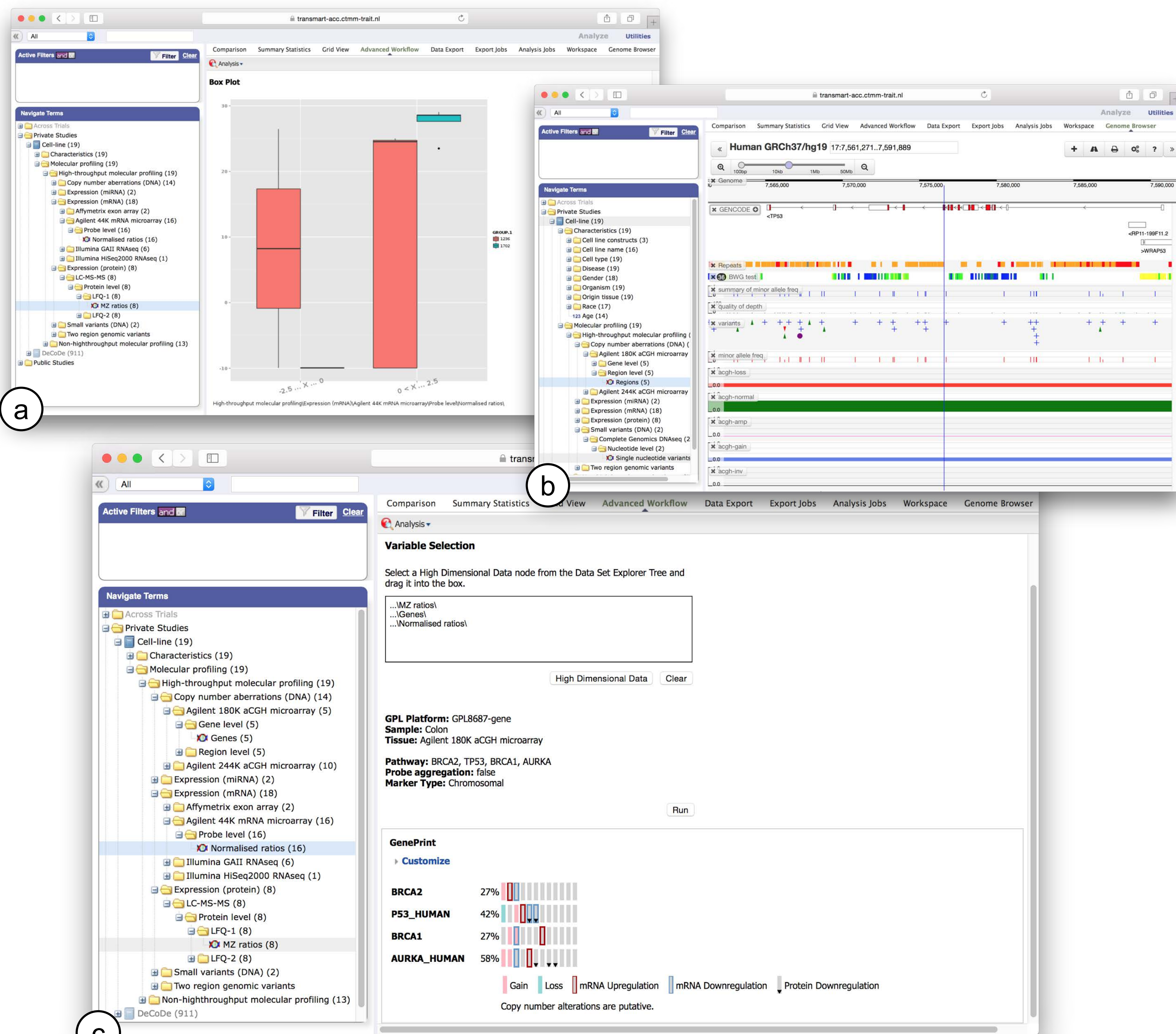
▲ **Figure 1: Multi-omics analysis in TranSMART.** *In every screenshot the data tree on the left shows the organization of the CLUC study (see right alinea), the right panel the analyses.*
(a) Boxplot *of TP53 gene (over or under) expression vs the AURKA and TP53 protein expression.*
(b) Genome browser *showing genomic variants alongside copy number alteration.*
(c) GenePrint *showing the co-occurrence of copy number aberrations and gene and protein expression alterations on four cancer genes.*

[1] Athey, Brian D., et al. "tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research."*AMIA Summits on Translational Science Proceedings 2013 (2013): 6.*
[2] Gao, Jianjiong, et al. "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal." Science signaling 6.269 (2013): pl1.
[3] Bierkens, Mariska & Bijlard, Jochem "The TraIT cell line use case." Manuscript in preparation

TranSMART in CTMM TraIT

One example of these multi-omics data analysis developments is the work done within the Dutch **CTMM Translational Research IT** (TraIT) project (http://www.ctmm-trait.nl), where tranSMART functions as the central data integration platform where exploratory analysis and hypothesis generation can be performed.

tranSMART v1.2

Regarding molecular data, **four domains** are accommodated within TraIT:

1. DNA and RNA arrays
2. DNA and RNA next generation sequencing
3. Mass spectrometry proteomics
4. Non-highthroughput molecular profiling

	CACO2	CACO2_AURKA	COLO205	COLO320	DLD1	HCT116	HCT116_MLH1	HCT15	HT29	LIM1863	LS174T	LS513	PC346c	RKO	SW1398	SW48	SW480	SW480_AURKA	VCaP	Biological information
DNA and RNA arrays																				
ArrayCGH, Agilent, 180K		✓				✓		✓	✓					✓		✓				CNA
ArrayCGH, Agilent, 244K	✓		✓	✓	✓		✓		✓		✓		✓	✓		✓	✓			CNA
mRNA Arrays, Agilent, 44K	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓		✓	✓	✓	✓			mRNA expression
Affymetrix Exon Array							✓												✓	mRNA expression
miRNA, Agilent							✓												✓	miRNA expression
DNA and RNA next generation sequencing																				
DNA-Seq, Complete Genomics										✓		✓							✓	SNV
RNA-Seq, Illumina GAII	✓				✓		✓				✓				✓		✓			SNV & mRNA expression
mRNA-Seq, HiSeq 2000							✓													SNV & mRNA expression
Mass spectrometry proteomics																				
LC-MS/MS	✓		✓	✓	✓		✓									✓	✓		✓	Protein quantification
Non-highthroughput molecular profiling																				
Microsatellite instability status	✓		✓	✓	✓		✓	✓	✓		✓	✓		✓	✓	✓	✓			Microsatellite (in)stability
DNA mutation	✓		✓	✓	✓		✓	✓	✓		✓	✓		✓	✓	✓	✓		✓	DNA mutation status
MLPA-gain-P146	✓				✓				✓										✓	CNA
MLPA-loss-X006	✓						✓													CNA
Promoter methylation	✓				✓		✓				✓					✓				Promotor methylation status
mRNA expression							✓													mRNA expression

▲ **Table 1: The experimental molecular data for the CLUC cell lines.**

The TraIT project recently added the **Cell Line Use Case** (CLUC) [3] to tranSMART. The CLUC is a collection of data on colorectal and prostate cell lines from an exceptionally broad set of platforms, as shown in Table 1.

This diverse set is used to:

- Standardize data formats and data processing pipelines from the four domains
- Test the integration pipelines within the TraIT translational toolset

By incorporating the same platforms as used for ongoing research projects, this cell line set gives a representative test set comparable to real patient data, without the legal burden of handling personal data. The TraIT Cell Line Use Case **will soon be made available freely** under an open data license.