European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.7 – 2nd progress report on Data Curation

Due date of deliverable: October 2014

Actual submission date: 08/01/2015

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:	D4.7
Deliverable title:	2nd progress report on Data Curation
Deliverable version:	
Due date of deliverable:	October 1st 2014
Actual submission date:	
Leader:	Reinhard Schneider, Manfred Hendlich, and Fabien Richard
Editors:	
Authors:	Serge Eifes; Adriano Barbosa; Wei Gu; David Henderson; Nathalie Jullian; Ioannis Pandis; Venkata Satagopam; Emmanuel Van der Stuyft
Reviewers:	Robin Munro; Philippe Rocca-Serra
Participating beneficiaries:	
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Fabien Richard, Manfred Hendlich and Reinhard Schneider
Work Package participants:	
Estimated person-months for deliverable:	
Nature:	
Version:	
Draft/Final:	Final
No of pages (including cover):	26
Keywords:	

2nd progress report on Data Curation

1 Abstract

Data curation has been provided for several different Innovative Medicines Initiative (IMI) projects as well as for the European Translational Information and Knowledge Management services (eTRIKS) Public server.

Four IMI projects have so far collaborated with eTRIKS in the area of data curation. These are UBIOPRED, OncoTrack, PredictTB and ABIRISK. Additionally RA-MAP, an MRC funded public-private collaborative project, has also engaged with eTRIKS for help with curation activities.

In this document, we present an overview of the different projects where eTRIKS has provided curation support. A short project description is provided about each project, along with information on which data types have been curated, the methods and algorithms applied to the data, and which problems had to be resolved during the curation process.

2 Introduction

The IMI is Europe's largest public-private initiative. It is focused on accelerating the development of better and safer medicines for patients. Data intensive translational research, as is used by many IMI projects, requires a knowledge management (KM) environment that allows for the provision of sustainable access to research data in an integrated manner.

The eTRIKS project specifically focuses on building a sustainable KM platform and providing support at the level of data management throughout the life cycle of a given translational research project. In this context, the value of data curation guarantees the sustainability of the data and facilitates the analysis and integration of highly complex clinical and multi-omics data.

In this report, we describe the curation efforts provided by the eTRIKS consortium during the second year of the project. A global overview on the support we have provided to the different collaborating IMI projects will be given. We focus on the following information for each of the projects:

- generic project description,
- information on the data types that have been curated,
- methods and algorithms applied to the data and
- specific problems that have been faced during curation.

Besides the curation support for IMI projects, we also give an overview about the data, which has been curated, for the eTRIKS Public server and details on the curation training that has been provided to the different projects.

3 Overview on eTRIKS supported projects

3.1 ABIRISK

Contributors: Wei Gu; Nathalie Jullian; Fabien Richard; Serge Eifes

3.1.1 Project description

ABIRISK will have access to large cohorts of patients treated with biopharmaceutical products (BPs). Analyses of the mechanisms and consequences of immunization against BPs require extensive post-marketing follow-up of patients, with comparisons of several BPs and various clinical conditions treated with the same BP. Sufficient numbers of patients must be included in each subgroup for the reliable evaluation of independent parameters. There is also a need for high-quality data generated by centres familiar with clinical research. The ABIRISK consortium has been designed to meet all of these requirements in order to target three types of disorders:

- Hemophilia A
- Multiple sclerosis
- Inflammatory diseases: inflammatory rheumatisms (including rheumatoid arthritis) and inflammatory bowel diseases

The ABIRISK Project collects data both retrospectively from patients suffering from various types of diseases and treated with various BPs at European centres with a high level of experience in clinical research and will prospectively recruit additional patients in dedicated studies during the 5 years of its program. Guidelines and Standard Operating Protocols (SOPs) for the study of anti-drug (AD) immunization will be established and used to standardize the collection of prospective data from these patients.

3.1.2 What data types have been curated?

Clinical data with multiple visits that has been provided by the Karolinska Institute (KI), Sweden and University of Innsbruck (UI), Austria have been curated so far. A more detailed overview on this data is given in **Table 1**.

Data provider	Wave	Number of variables	Number of patients	Number of visit records
UI	1	34	4500	12000
UI	2	93	4500	12000
KI	1	42	6300	16000

Table 1: Overview on clinical data curated for the ABIRISK project

3.1.3 What methods/algorithms and/or pipelines have been developed/used?

The University of Luxembourg (UL) team has developed a data quality-checking pipeline to check data integrity defined in the data control definition file (see **Figure 1**). In the data control definition file, features of each column in the data file have

been defined (e.g. data type, whether the column is compulsory or not). Then the data quality-checking pipeline will go through all the data columns in the data file and check if the data fulfil the definitions.

Filename	DataType	ColumnNumber	DataLabel	CDISCLabel	TerminologyLabel	IsCompulsory
IMU_TEST_FILE_wave2.txt	Int	1	OMIT			N
IMU_TEST_FILE_wave2.txt	String	2	SUBJ_ID			Y
IMU_TEST_FILE_wave2.txt	String	3	TEST_ID_UNIQUE		C66786	Y
IMU_TEST_FILE_wave2.txt	String	4	VISIT_NAME	CDISC:VISIT		Y
IMU_TEST_FILE_wave2.txt	Date	5	SAMPLE_DATE	CDISC:LBDT	C66731	N
IMU_TEST_FILE_wave2.txt	String	6	TREAT	CDISC:EXTRT		N
IMU_TEST_FILE_wave2.txt	Date	7	DOSE_LATEST_DATE	CDISC:EXENDT		N

Figure 1: example of the control definition file

3.1.4 What problems have been encountered?

In the ABIRISK tranSMART 1.1 server (hosted by eTRIKS), the data export function could not generate correct output for visit of categorical data. The visit information of the categorical data was lost during the exporting. This was found to be a general problem of the tranSMART 1.1 version. A fix was provided by the UL team, in the form of a new exporting function developed to add the missing visit information. This solution overcame this issue in TM 1.1. The new TM version (1.2) has solved this problem.

3.2 OncoTrack

Contributors: Adriano Barbosa; Serge Eifes; Wei Gu; David Henderson; Gino Marchetti; Nathalie Jullian; Ioannis Pandis; Anthony Rowe; Venkata Satagopam; Emmanuel Van der Stuyft

3.2.1 Project description

The goal of OncoTrack is to identify and characterize biological markers that will help the understanding of the variable make-up of tumours and how this affects the way colorectal cancer patients respond to treatment. This question is being tackled with cutting edge laboratory-based genome sequencing techniques coupled to novel computer modelling approaches to study both the biological heterogeneity of colon cancers (i.e. patient to patient variability) as well as tumour variation within the patient – for example, by comparing primary tumours with metastases. For more details see section "Objectives" on the OncoTrack website[1].

The results of this research is expected to allow the identification and qualification of a set of biomarkers that may be used to guide patient therapy, provide immediate feedback upon the effects of treatment and ultimately, indicate likely outcome of disease management – i.e. Oncology Tracking.

3.2.2 OncoTrack data landscape

The OncoTrack data landscape is schematically represented in **Figure 2**.

[1] http://www.oncotrack.eu

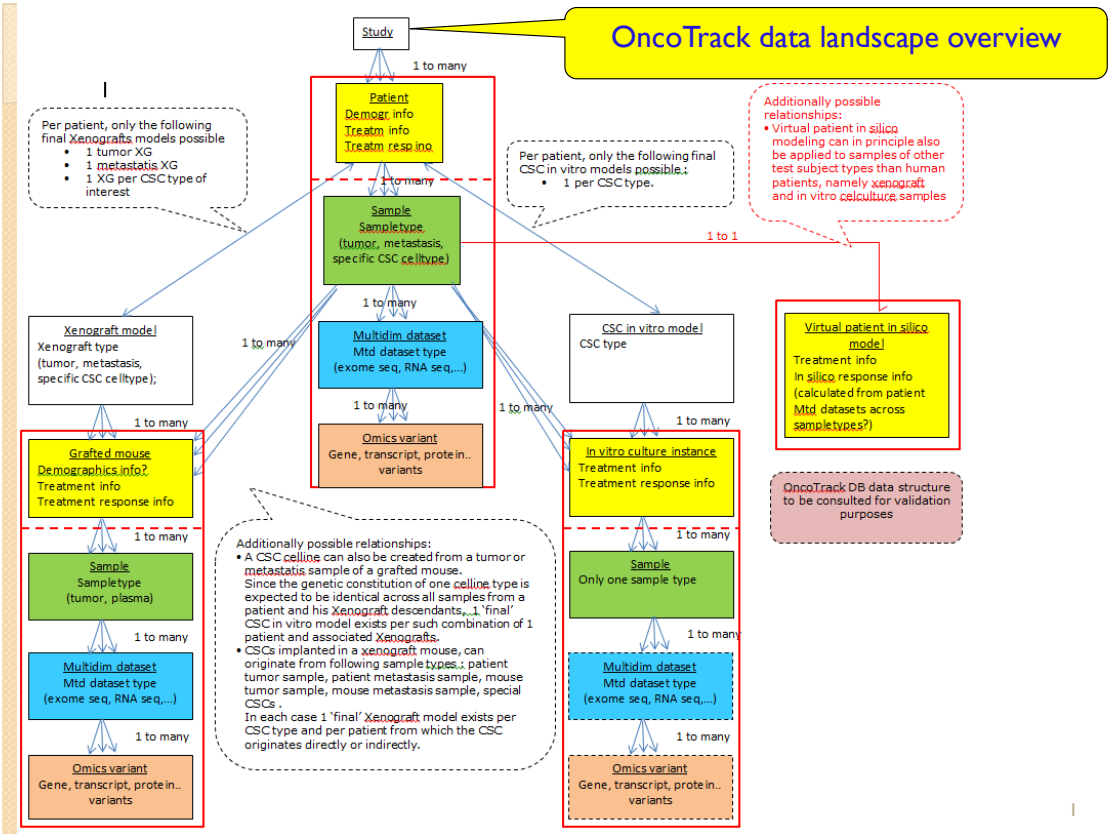


Figure 2: Schematic representation of the OncoTrack data landscape

3.2.3 What data types have been curated?

The low dimensional data currently available in OncoTrack for the different subject types involved in the research (patients, xenografts, cell-lines and in silico models) have been curated into separate studies for each subject type, together with reference data to capture the interrelationships between the different subjects (e.g. what is the parent patient from which a xenograft was created?)

> So far, > 100 variables of > 300 Patients, > 400 Cell line treatment groups, > 750 Xenograft treatment groups and > 3000 *In silico* model treatment groups have been curated.

Given the importance of genomic variation in OncoTrack's research, work is currently ongoing to enable scientists to view - for a cohort selection of samples in tranSMART 1.2 - genomic variation information that can be captured in the VCF format. This is based on an integration with the Dalliance genome browser. A proof of concept is now available based on public VCF data. Integration of OncoTrack data is planned shortly.

Once the above functionality is available, we will reduce to a sub-selection only the overall mutation summary we experimented within our earlier implementation in the low dimensional concept tree space. This allowed cohort selection based on whether any combination of genes of interest was mutated or not. However, should a full-sized experiment be considered, such query would likely over-stress the system by the sheer amount of concepts needed to cover that information.

Other multidimensional data being worked on includes methylation and RNA-Seq data.

- M-scaled differential Methylation values between an individual patient's tumour vs. healthy sample on a probe-by-probe basis have already been successfully uploaded into tranSMART.
- Based on interaction with Lee Butcher (OncoTrack methylation expert) we have agreed to focus on Beta-scaled methylation values at a higher aggregation level (gene regions as well as 'differentially methylated regions' which may span across genes). We are awaiting sample data from Lee Butcher (University College London) to experiment with their curation and upload.

- The curation of RNA-Seq data into the OncoTrack tranSMART work environment has been successfully tested based on public RNA-Seq data. As soon as such data are available from OncoTrack DB, these will be made available in tranSMART.

Patient clinical data are not yet available for reasons related to data privacy. As these become available, these too will be curated into tranSMART.

For each of the data types currently curated, an export script has been written to draw the relevant data from the OncoTrack data repository environment (OncoTrack DB) and to automatically transform them into the column-mapping format required for data upload into tranSMART. Since OncoTrack data are collected on an ongoing basis, a next logical step could consist in the automation of the data curation and upload process at regular time intervals.

3.2.4 What methods/algorithms and pipelines have been developed/used?

By adopting tranSMART 1.2 we get transparent access to:

- flexible data export functionality based on specifically selected cohorts/subsets. This cohort-based data can then be directly accessed by analytical tools outside of the tranSMART environment
- data import functionality into the Galaxy workflow environment which supports flexible data processing/analysis pipelines based on the exported cohort data from tranSMART.

Since it became very clear that the degree to which tranSMART will eventually add value to OncoTrack critically depends on the flexibility with which the interrelationships between different subject types in the overall OncoTrack data landscape can be integrated into querying and analysis, we tried to address the current tranSMART limitations in bridging across subject types in two complementary ways:

- Via the tranSMART API, and leveraging the above-referred referential data that provide the links between subject types.

  To illustrate the API capabilities, an R script is being developed to create a report of how different subject types align as to their response to similar treatment.

  This should illustrate the great potential of an interaction platform, where

scientists can iteratively formulate their investigational questions and have these addressed by bio-informaticians, with API-based access to the tranSMART data warehouse.

- Via the development of a script to allow for the cascading from within the tranSMART web interface of different query filters across the different subject-types in the overall OncoTrack data landscape. Such scripting enables users to work through the following example query cascade:
    - In the xenograft study, create two cohorts based on treatment response:
        - Cohort 1: responders; Cohort 2: non-responders
    - In the corresponding patient study, create two cohorts to investigate patient-based genomic variations which could explain the response vs non-response in their corresponding child-xenografts.
        - Cohort 1: all parent patients of xenograft in the responder cohort
        - Cohort 2: all parent patients of xenografts in the non-responder cohort

3.2.5 What problems have been encountered?

The section here below gives an overview of the problems we have encountered with tranSMART version 1.1.

A core limitation we struggled with (as explained above) boils down to this:

- tranSMART can only capture 1 value per low dimensional tree concept. This was the cause why we had to cut up the study data into separate studies per subject type, since concepts for patient-derived subject types typically have multiple values for the same patient.
- Given its dependence on the i2b2 layer, the extension of tranSMART query filtering capabilities to include sub-lists of categorical variables seemed like a big problem. A workaround to this problem was implemented in a script as explained above.

  Given the importance to OncoTrack of the above two bullet points, a set of slides were included in the appendix to articulate the problem for which the team eventually found a workaround.

Performance and stability issues were met quite often, both with the upload and when using tranSMART.

Given the fact that API and VCF-genomic data support are crucial to OncoTrack and missing in tranSMART 1.1, OncoTrack decided to take the risk of becoming early

adopters of tranSMART 1.2, as this and many other desirable features were available.

On the critical path of making the above-described functionality operationally accessible to the OncoTrack scientists is the deployment of the Luxembourg-hosted test environment on the operational eTRIKS tranSMART hosting environment in Lyon.

Prerequisites for that are:

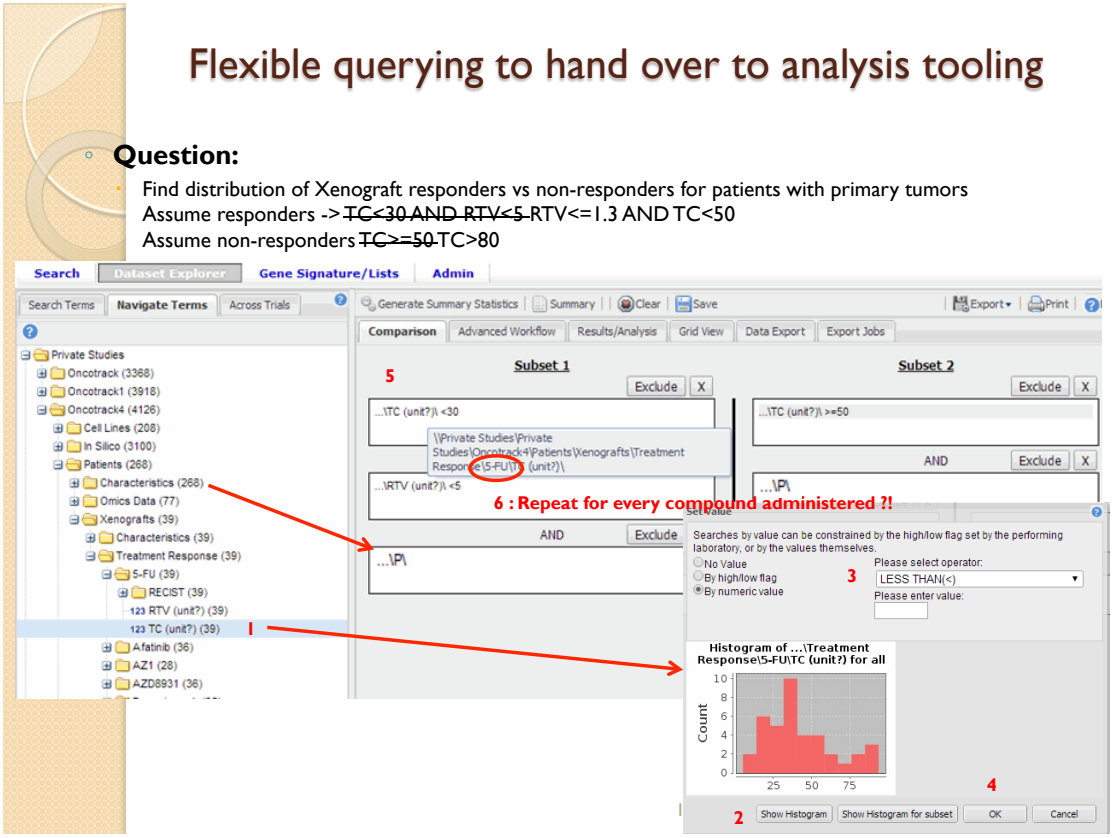
- Getting CDAs/MTAs in place between OncoTrack and CC-Lyon/eTRIKS
- The deployment of an operational eTRIKS tranSMART 1.2 version.

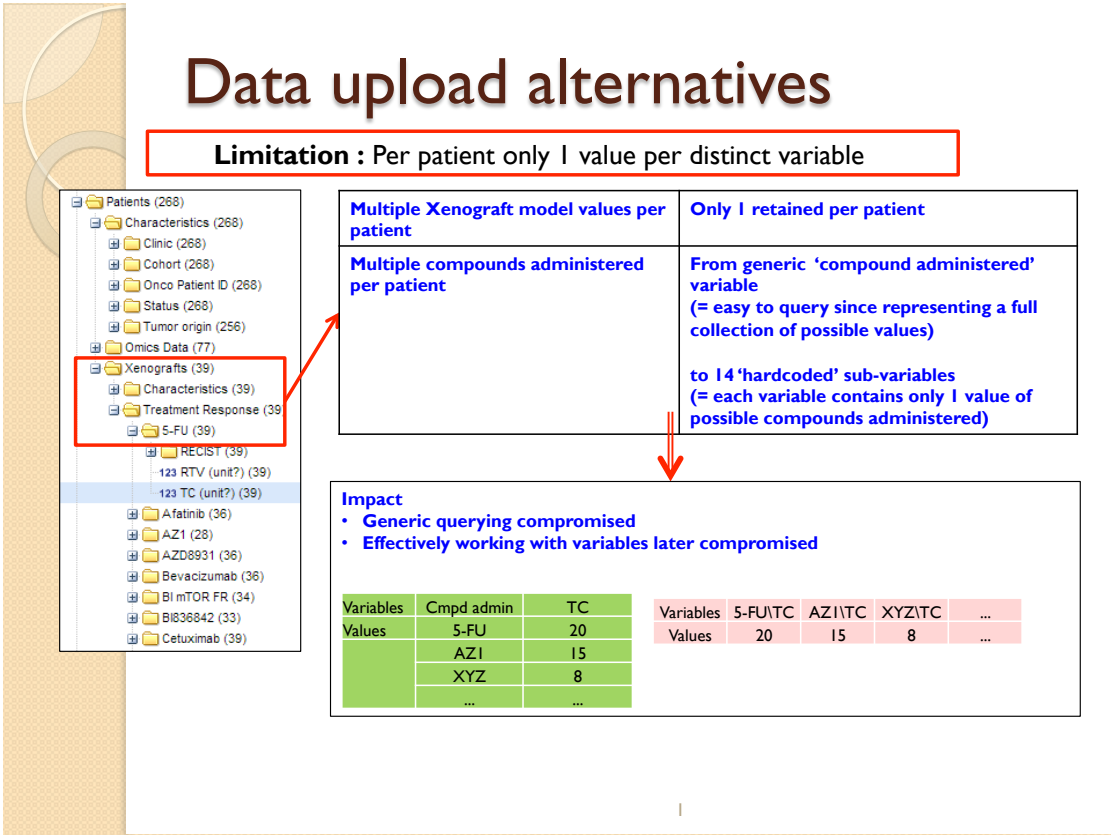
Upfront work, independent of these prerequisites, has been done by the Lyon team (including deployment of test environment and LDAP setup).

The data curation and upload work, which has been difficult and slow in its initial stages, picked up significant speed and effectiveness latterly. Two key factors in that have been:

- The dedication of significant bandwidth by the Luxembourg team with great data curation and bio-informatics expertise in the OncoTrack data curation effort, as agreed to in the latest eTRIKS resource team meeting.
- The availability of tranSMART 1.2 functionality

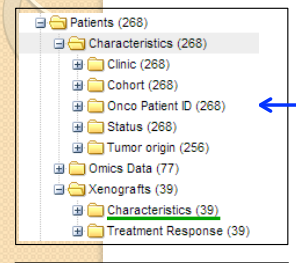
3.2.6 Appendix: Articulation of a key-issue (addressed via scripting by Wei Gu, University of Luxembourg)

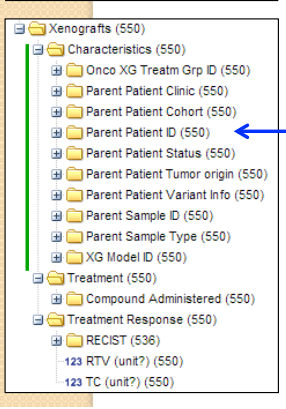




Data upload alternatives

Limitation : Per patient only 1 value per distinct variable





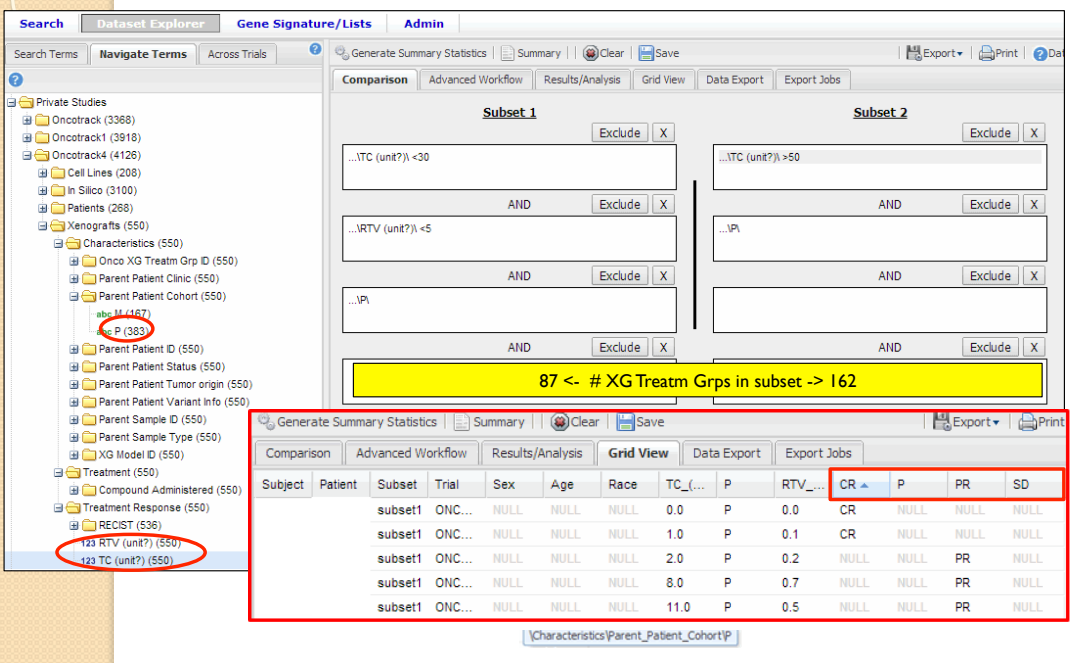
Patient ID	XG Treatm Grp
597-CM-M	597-CM-M-MF-01-04-XEN-01-BI_mTOR_FR
597-CM-M	597-CM-M-MF-01-04-XEN-01-BI836842
597-CM-M	597-CM-M-MF-01-04-XEN-01-Regorafenib
597-CM-M	597-CM-M-MF-01-04-XEN-01-Selumetinib
633-CB-P	633-CB-P-TF-01-04-XEN-01-BI_mTOR_FR
...	...

Cascading queries between distinct studies with 1xN relationship between them
- **Via joint variable – patient id**

Requirements
- **Export list with distinct Patient IDs related to a selection in either study**
- **Import that list as a selection criterium in the other study**

Interim solution
- **Lists of interest as characteristics in complementary study**

Flexible querying to hand over to analysis tooling



3.3 U-BIOPRED

Contributors: Ioannis Pandis; Kai Sun; Florian Guitton

3.3.1 Project description

U-BIOPRED is a multiple stakeholder, EU-IMI funded, severe asthma research project, comprising of longitudinal clinical (adult and pediatric cohorts) studies and associated animal, *in vitro* and *in silico* model translational studies, using multi-omics technologies, with an aim to create novel classifiers for better describing asthma disease heterogeneity.

3.3.2 What data types have been curated?

Samples collected from both human subjects and models are profiled using the following omics technologies:

- Transcriptomics - Affymetrix Gene Arrays
- Proteomics - MSE-based Label-free technology
- Lipidomics - ESI, HPLC-MS, GS-MS, LC-QTOF and MRM, depending on the lipid subset profiled
- Breathomics - eNOSE, GC/MS, NMR
- Genetics - Affymetrix SNP Arrays

N.B: Somalogic refers to the SOMAscan platform[2].

[2] http://www.somalogic.com/Products-Services/SOMAscan.aspx

The data types and volume of data curated which have been loaded into the U-BIOPRED eTRIKS/tranSMART instance to date, are shown in **Table 2**.

Technology	Dataset	Number of Subjects/Samples	Number of features	Total Features
Transcriptomics	Blood	202	54,675	11,044,350
	Bronchial Brushings	145	54,675	7,927,875
	Bronchial Biopsies	136	54,675	7,435,800
	Nasal Brushings	78	54,675	4,264,650
	Sputum	139	54,675	7,599,825
				38,272,500
Proteomics	Serum Somalogic	598	1,129	675,142
	Serum MS using MS^e technology[3]	131	130	17,030
	Sputum MS using MS^e technology[3]	99	3,231	319,869
				1,012,041
Lipidomics	Sputum MS using MS^e technology[3]	128	15,622	1,999,616
	Urine Eicosanoids	584	13	7,592
	Sputum Eicosanoids	319	13	4,147
				2,011,355
Beathomics	Adult eNOSE	106	190	20,140
	Adult GC/MS	64	7,036	450,304
	Paediatric eNOSE	106	190	20,140
	Paediatric GC/MS	64	7,036	450,304
				940,888
Clinical Data	Adult	617	2,388	1,473,396
	Paediatric	260	1,685	438,100
				1,911,496
Total				**88,296,560**

Table 2: Overview of the data types and volume of curated data in the U-BIOPRED eTRIKS/tranSMART instance

3.3.3 What methods/algorithms and/or pipelines have been developed/used?

3.3.3.1 Scripts

Custom scripts have been developed for pre-processing the clinical and lab results datasets, using R, Perl and Python languages and are publicly available on Github[4].

[3] http://www.medscape.com/viewarticle/814689_3
[4] https://github.com/pandis83/UBIO_Scripts

3.3.3.2 Annotation files

For all omics datasets, custom annotation files were created in order for them to be "loadable" into tranSMART. This was achieved by leveraging the standard gene expression (GEX) platform annotation file structure. Furthermore, for proteomics datasets, all UniProt IDs were converted to EntrezGene IDs, thus enabling the filtering by pathway functionality available in several Advanced Workflows.

3.3.3.3 Loading scripts

Standard Kettle Scripts were used for all loading purposes.

3.3.3.4 Data Quality Control (QC)

Multiple data sources and diverse data types make the entire data management process highly error prone. Thus, to ensure provision of high quality data to the consortium, apart from automated rule-based check included in the pre-processing scripts, a data QC process was implemented.

3.3.3.5 Data QC process

- All data labelled "PROVISIONAL DATASET" was prepared and loaded into tranSMART as a secure study, which no consortium member except for platform admin and curators can access.
- Approximately 10 "testers" who are domain experts and omics platform experts were identified.
- A group comprised of the above testers was created and provided full access to the PROVISIONAL DATASET.
- An issue-tracking sheet was set up using Smartsheet (smartsheet.com) by BIOSCI consulting, that only the testers and data management team could have access.
- Testers were given two weeks, in which they were assigned to investigate the dataset and perform analytical queries.
- All identified issues were reported on the issue tracker.
- Once the issues had been resolved the amended dataset was loaded into tranSMART and all consortium members were granted access.

3.3.4 What problems have been encountered?

The only issue encountered was the amount of time needed to upload the data. Currently, the entire clinical and omics datasets require ca. 60 hours for complete upload. This was addressed by only performing the data upload on weekends and by disabling access to tranSMART by rerouting the URL (transmart.doc.ic.ac.uk) to a "maintenance page" and informing all consortium members via e-mail, 2-3 days in advance.

4 Curation training and documentation

To provide better support to the curators of the different IMI projects, eTRIKS has created an IMI curator training. This training by members of the University of Luxembourg, which happened on March 20th and 21st 2014 in Luxembourg had a special focus on familiarizing the participants with the eTRIKS/tranSMART curation environment and the corresponding software tools, namely the Extract, Transform, Load (ETL) scripts[5] and the FCL4tranSMART/ICE tool[6]. The ETL procedures allow the experienced user to upload the curated data to the tranSMART data warehouse via the command line. FCL4tranSMART is a software tool that guides the curator via graphical user interface through the curation process from the uncurated data files to the final upload of the curated data to the tranSMART database. Training material and documentation has been made available.

A number of training videos focusing specifically on the use of the FCL4tranSMART/ICE tool have been created by collaborators from Sanofi and made accessible to the IMI data curator community.

A detailed overview on the curator training and documentation can be found in **Table 3**.

Title	Presenter	Organization	Type of training	Data Type	Link
eTRIKS-Data curation and upload training	S. Eifes, W. Gu & V. Satagopam	University of Luxembourg	Onsite & documentation	Clinical & gene expression	[7]
ICE: introduction	C. Raillère & D. Peyruc	Sanofi	Video	Raw data files	[8]
ICE: loading of clinical data	C. Raillère & D. Peyruc	Sanofi	Video	Raw data files	[9]
ICE: loading of gene expression data	C. Raillère & D. Peyruc	Sanofi	Video	Gene Expression	[10]

Table 3: Overview on curator training/documentation

[5] https://git.etriks.org/transmart-dse-etl/tree/master/DSE/Kettle/Kettle-ETL
[6] https://github.com/transmart/tranSMART-ETL/tree/master/FCL4tranSMART
[7] https://app.smartsheet.com/b/home
[8] http://www.youtube.com/watch?v=lTdRbtaXb24&feature=youtu.be
[9] http://www.youtube.com/watch?v=53BttxoIZXE
[10] http://www.youtube.com/watch?v=MrXW21sogmc

5 eTRIKS Public server

Contributors: Serge Eifes; Wei Gu; Adriano Barbosa; Venkata Satagopam; Nathalie Jullian

5.1 Introduction

As described in deliverable D4.5 (1st Progress report on Data Curation, section 1.2 "Aim of the Public server delivery package"), the main objective of the eTRIKS Public server is to provide a public eTRIKS/tranSMART server[11] giving access to highly curated and standardized public studies. This server should make public studies that are of interest for the different IMI projects accessible to the public. An added value for this public data is the application of eTRIKS data curation and quality standards facilitating the integrated analysis of these studies in the eTRIKS tranSMART software.

In this section, we give an overview on the different data resources and studies that have been curated and uploaded to the Public server. We provide details on which data types have been curated, the methodologies and tools that we have used during the data curation and upload, as well as the problems and limitations that have been encountered.

5.2 Advertisement for Public study curation requests

To reach out to the other IMI projects and make them aware of the capabilities of the eTRIKS project in general and the Public server more specifically, we decided to provide an advertisement[12] for Public study curation requests to the IMI consortium, which was announced in their newsletter.

This advertisement gave detailed information on the eTRIKS data curation and loading service for public studies that we offer to other IMI projects. Following curation by our curation team, the studies are made available at the Public server. In conjunction with the advertisement and to allow for the proper collection of the incoming curation requests, we developed an online curation request form. The corresponding URL was made available in the advertisement[13].

5.3 Gene expression Omnibus

5.3.1 Description

The Gene Expression Omnibus (GEO)[14] is a public functional genomic repository for data submitted by the research community. GEO contains high-throughput microarray and next generation sequence data, and currently encompasses more than 32000 public studies (Barrett et al., 2013).

[11] https://public.transmart.etriks.org/transmart/
[12] https://portal.etriks.org/Portal/pdf/IMI_Public_server_advertisement_v1.pdf
[13] http://tinyurl.com/qfpxtgd
[14] http://www.ncbi.nlm.nih.gov/geo/

For gene expression data it is important to be accompanied by the contextual biological and processing details under which experiments were performed. A lack of this "metadata" can render the data itself meaningless. GEO is a MIAME-compliant infrastructure and it supports fully annotated records encompassing biological as well as descriptive metadata (Barrett et al., 2007).

GEO is a database that allows a unified access to thousands of valuable public gene expression studies (Barrett et al., 2011). This makes it particularly interesting as data resource for study curation in the context of tranSMART Dataset Explorer.

Over the last few months a considerable amount of new GEO studies have been curated by the eTRIKS Public server team (on request by the projects RA-MAP and OncoTrack). Curation has been performed according to the standards defined in deliverable "D4.5 1st Progress report on Data Curation". The curated studies are available on the Public server. A full overview of these studies can be found in **Table 4**.

Study name	Project	Data source	Data type(s)
Asano(2012) GSE36757	RA-MAP	GEO	Clin/LD+GEX
Badot(2009) GSE15602	RA-MAP	GEO	Clin/LD +GEX
Bansard(2011) GSE11827	RA-MAP	GEO	Clin/LD +GEX
Del Rey(2010) GSE21959	RA-MAP	GEO	Clin+GEX
Julia(2009) GSE15316	RA-MAP	GEO	Clin/LD +GEX
Kabuyama(2008) GSE9329	RA-MAP	GEO	Clin/LD +GEX
Kusaoi(2011) GSE30662	RA-MAP	GEO	Clin/LD +GEX
Lee(2011) GSE27390	RA-MAP	GEO	Clin/LD +GEX
Lequerre(2006) GSE3592	RA-MAP	GEO	Clin/LD +GEX
Lequerre(2009) GSE13026	RA-MAP	GEO	Clin/LD +GEX
Lindberg(2010) GSE21537	RA-MAP	GEO	Clin/LD +GEX
Mathieu(2008) GSE11575	RA-MAP	GEO	Clin/LD +GEX
Mesko(2012) GSE25160	RA-MAP	GEO	Clin/LD +GEX
Nishimoto(2008) GSE12653	RA-MAP	GEO	Clin/LD +GEX
Teixeira(2009) GSE15573	RA-MAP	GEO	Clin/LD +GEX
Toonen(2012) GSE33377	RA-MAP	GEO	Clin/LD +GEX
van Baarsen(2010) GSE19821	RA-MAP	GEO	Clin/LD +GEX
GSK Cell Lines	OncoTrack	caBIG	Clin/LD +GEX
CCLE	OncoTrack	?	Clin+DrugProf+GEX

Table 4: "Study name" indicates the name of the study on the eTRIKS Public server. "Projects" shows the name of the project that has suggested the study to be uploaded to the server. "Data source" corresponds to the data repository where the original data has been retrieved (GEO: NCBI Gene Expression Omnibus; caBIG: National Cancer Institute cancer Bioinformatics Grid). "Data type(s)" indicates the different types of data that have been curated and loaded to the server (Clin: Clinical/low dimensional; GEX: Gene Expression; DrugProf: Drug Profiling

5.3.2 What data types have been curated

We have curated the available clinical data in conjunction with the gene expression data.

5.3.3 What methods/algorithms/pipelines have been used or developed?

The GEO-Dataset Explorer curation pipeline[15] (for more details see deliverable "D4.5 1st Progress report on Data Curation", section "1.4.2.1 GEO Dataset Explorer pipeline") was used to generate the corresponding standard format files. The curated files were then uploaded to tranSMART using the Kettle Pentaho[16] ETL scripts[17].

5.3.4 Problems encountered

The metadata uploaded by the study data owners to GEO are currently not following the same standard. For some studies, metadata are not provided in the relevant or required fields. These data are frequently found back in other fields. This lack of consistency at source requires a lot of tuning of the pipeline to rectify such problems

5.4 GSK Cancer Cell Line Genomic Profiling Data

5.4.1 Description

GlaxoSmithKline (GSK) Cancer Cell Line Genomic Profiling Data (Greshock et al., 2010) provides access to the genomic profiling data for more than 300 human cancer cell lines. This data initially generated by GSK has been made publicly available and encompasses a large variety of cancer cell types.

5.4.2 What data types have been curated?

We have curated experimental metadata and cancer cell line data to be uploaded with the ETL procedures for study metadata and clinical data[17]. Furthermore we have curated the gene expression data. **Table 5** provides a summary of the data that has been loaded into the eTRIKS Public server instance.

	Total count
Samples	950
Cell lines	319
Diseases	58
Variables loaded	28
Variables curated	44

Table 5: Overview on GSK Cancer Cell Line Genomic Profiling Data available on the eTRIKS Public server.

5.4.3 What methods/algorithms and/or pipelines have been developed/used?

A dedicated R script has been developed that allows the generation of standard format files (SFF). These SFF have been used as input for the corresponding Kettle Pentaho ETL procedures.

[15] https://git.etriks.org/serge.eifes/geo_deapp_pipeline/tree/master
[16] http://community.pentaho.com/projects/data-integration/
[17] https://git.etriks.org/transmart-dse-etl/tree/master/DSE/Kettle/Kettle-ETL

5.4.4 What problems have been encountered?

/

5.5 Broad-Novartis Cancer Cell Line Encyclopedia

5.5.1 Description

The Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) project is a collaboration between the Broad Institute, and the Novartis Institutes for Biomedical Research and its Genomics Institute of the Novartis Research Foundation to conduct a detailed genetic and pharmacologic characterization of a large panel of human cancer models, to develop integrated computational analyses that link distinct pharmacologic vulnerabilities to genomic patterns and to translate cell line integrative genomics into cancer patient stratification. The CCLE provides public access to genomic data, analysis and visualization for about 1000 cell lines. [18]

5.5.2 What data types have been curated?

Data Types curated:

The following data has been curated and made available on the Public server:

- 1046 cell lines
- 24 anti cancer drugs, tested against 504 cell lines
- RNA Gene expression
- Clinical characterization

Curation process:

Drug profiling: for each drug tested, 3 values are reported: Activity area, Amax and IC50. A Pentaho script has been used for transforming the drug profiling data into a matrix format file that can be uploaded into tranSMART v1.1 The final matrix contains 1 line per sample and 1 column per variable, as shown below in **Figure 3** for 11 examples: the first column contains the cell line name, the columns starting with the "ActArea_", "Amax_" and "IC50_" prefix contain respectively the values for the activity area, the Amax and the IC50 of a given drug. A more in depth description for the corresponding clinical standard format file structure can be found under [19]

CCLE Cell Line Name	ActArea_17-AAG_HSP90	ActArea_Erlotinib_EGFR	...	Amax_17-AAG_HSP9	Amax_Erlotinib_EGFR	...	IC50_17-AAG_HSP90	IC50_AEW541_IGF1R
1321N1_CENTRAL_NERVOUS_SYSTEM	3	0.2		-72.1	-24.3		0.2	8
22RV1_PROSTATE	3.1	0.2		-76.3	0.7		0.3	2.3
42MGBA_CENTRAL_NERVOUS_SYSTEM	5.1	0.9		-80.4	-46.8		0.1	2.7
5637_URINARY_TRACT	3.5	1.9		-91.7	-76.7		0.1	5
639V_URINARY_TRACT	3.8	0.3		-89.6	-3.5		0.2	1.7
697_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	3.7	1.6		-91.8	-82.6		0.4	4.3
769P_KIDNEY	3	1.9		-72.5	-68		0.2	8
786O_KIDNEY	3.3	0.1		-95.4	-16		0.3	7.6
8305C_THYROID	3.5	0.5		-71	-24.9		0.4	5
8505C_THYROID	3.7	0.1		-92.8	-1.4		0.3	8
8MGBA_CENTRAL_NERVOUS_SYSTEM	3.6	0.5		-80.6	-51		0.2	3.9

Figure 3: Example for the clinical standard format file showing the data for 11 examples

[18] https://www.broadinstitute.org/ccle/home

[19] https://wiki.transmartfoundation.org/display/TSMTGPL/Clinical+Data

Probe annotation files have been generated automatically for the GPL570 platform. The file that is required for the probes to be identified must contain the 4 following columns: GPL_ID (platform name), PROBESET (probe name), GENE_SYMBOL, GENE_ID, and ORGANISM (Human here). The PROBESET corresponds to the name of the probe as it is listed in the experimental file, while the GENE_SYMBOL and GENE_NAME relate to the label of the genes.

Data tree organization :
The CCLE study contains 4 blocks of data listed under "Biomarker Data", "Demographics", "Drug Profiling" and "Samples". The "Biomarker Data" node contains the Affymetrix mRNA gene expression data. The "Demographics" node contains information related to the gender of the original subject. The "Drug Profiling" node contains the drug data measures for the 24 anti-cancer compounds (IC50, Activity Area and Amax as defined in the original paper). The compounds are classified into 3 sub-nodes depending on their mode of action: "cytotoxic compounds", "kinase inhibitors" and "other targeted compounds". The "Sample" node contains details about the cell lines: Cell line name, histology and histology subtypes, name of primary site and origin of the cell line (source). The data tree as it is shown in tranSMART is shown in **Figure 4**.

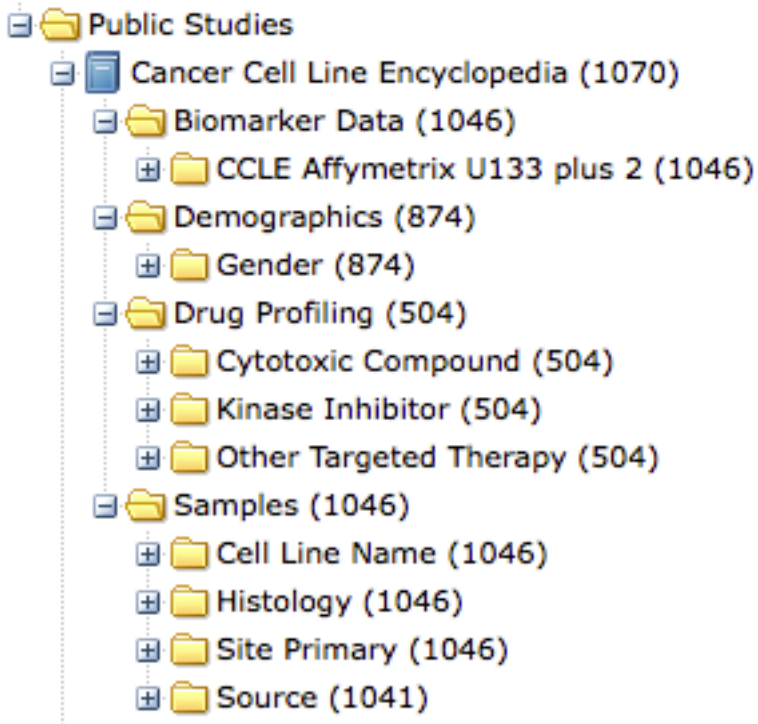


Figure 4: Data tree in tranSMART for the CCLE data.

5.5.3 What methods/algorithms/pipelines have been used or developed?

Kettle Pentaho ETL scripts have been used for the upload of clinical and high dimensional data.

5.5.4 Problems encountered?

A patch had to be incorporated in order to handle the cell line names, as they are made of up to 80 characters while the current tranSMART instance could handle a maximum of 20 characters.

6 Discussion on encountered problems

One of the current challenges in biomedical research is the existence of a multitude of different omics data types. There is no data warehouse that allows for the handling of all the existing omics data types and provides access to their corresponding data analytic algorithms. A number of such limitations in tranSMART version 1.1 have been encountered and reported. To overcome these limitations we have developed a number of fixes and strategies to deal with this type of data. Following the requests and needs of different supported IMI projects, a number of new data types have been added in tranSMART 1.2 including DNA/RNA sequencing, (i.e. VCF, mRNA/miRNA), QPCR, Proteomic, aCGH and time series for high dimensional data.

The runtime for the ETL procedures, especially for large data sets is problematic. This problem may also prevent accessibility to the tranSMART front-end, or will at least mean that it might be very limiting for the users/data analysts. Various discussions focusing on how to best solve this problem have taken place. A possible solution might be to combine the relational database management system that tranSMART relies on, with a NoSQL-based approach for storing omics/high dimensional data.

The eTRIKS project is in part focused on enabling cross-study analyses. Achieving cross-study data comparability requires homogenous data curation across studies. To allow this, homogenous ontology/terminology usage across studies is needed. Furthermore, a comparable data tree representation in tranSMART is required. On-going efforts of eTRIKS work package 3 are focusing on solving these issues.

The access for curators to project specific data is currently limited due to the requirement to have material transfer agreements (MTA) and confidential disclosure agreements (CDA) in place between the implicated institutions; this means the institutions of the data providers and the data curators. In some cases the IMI projects might encompass a large number of different partners, so the eTRIKS curating institutions have to sign CDAs and MTAs with all implicated data providers from the project before getting access to the data. Such legal constraints can cause considerable delays before eTRIKS is able to curate the data.

7 Bibliography:

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–7. doi:10.1038/nature11003

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., … Soboleva, A. (2011). NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Research*, *39*(Database issue), D1005–10. doi:10.1093/nar/gkq1184

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., … Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Research*, *35*(Database issue), D760–5. doi:10.1093/nar/gkl887

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991–5. doi:10.1093/nar/gks1193

Greshock, J., Bachman, K. E., Degenhardt, Y. Y., Jing, J., Wen, Y. H., Eastman, S., … Wooster, R. (2010). Molecular target class is predictive of in vitro response profile. *Cancer Research*, *70*(9), 3677–86. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20406975