Co-funded by the Horizon 2020
Framework Programme of the
European Union

Grant agreement No. 671555

ExCAPE
Exascale Compound Activity Prediction Engines

Future and Emerging Technologies (FET)

Call: H2020-FETHPC-2014
Topic: FETHPC-1-2014
Type of action: RIA

Deliverable D3.9
PublicCancer
Public cancer cell line datasets

Due date of deliverable: 30.05.2017
Actual submission date: 30.09.2017

Start date of Project: 1.9.2015 Duration: 36 months

Responsible Consortium Partner: AZ
Contributing Consortium Partners: JPNV, IDEA
Name of author(s): Jiangming Sun, Hongming Chen (AZ), Vladimir Chupakhin (JPNV), Nina Jeliazkova (IDEA)
Internal reviewer(s): Felipe Golib (JP)
Revision: V2.0

Project co-funded by the European Union within the H2020 Framework Programme (2014-2020)		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme 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)	

Document revision tracking

This page is used to follow the deliverable production from its first version until it has been reviewed by the assessment team. Please give details in the table below about successive releases.

Release number	Date	Reason of this release and/or validation	Dissemination of this release (task level, WP/ST level, Project Office Manager, Industrial Steering Committee, etc)
V1.0	16.06.2017	First draft for discussion	WP3 project partner
V2.0	29.08.2017	Second draft for release	PU

Glossary

GDSC	The Genomics of Drug Sensitivity in Cancer Project
CCLE	The Cancer Cell Line Encyclopedia
CTRP	The Cancer Therapeutics Response Portal
GE	Gene expression
Mut	Mutation
CNV	Copy number variation
Methy	Methylation
ECFP	Extended-Connectivity Fingerprint
IC50	Half maximal inhibitory concentration
pIC50	Negative log10 of the IC50 value in molar concentration

Link to Tasks

Task number	Work from task carried out	Deviations from task technical content, with motivation and summary of impact
T3.1	Collection and curation of public and proprietary ADMET and Chemogenomics datasets	The task was carried out as expected

Table of Contents

1 Executive summary

D3.9 is focused on the curation of public cancer cell line data for ExCAPE project. The data set includes NCI60 drug response data set, the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC) drug cancer cell line sensitivity dataset. These datasets include compound chemical information and cell line genomic features. For genomic information, features like gene expression, copy number alteration and mutation are collected. For chemical related features, ECFP6 fingerprint is included. These cancer cell line data will be used as benchmark dataset for multi-task learning algorithms developed in WP1 and WP2.

The reason for late submission was due to the complexity of cell line data and extensive consultation among WP3 partners for data curation.

2 Introduction – Aim

For decades, human immortal cancer cell lines have constituted an accessible, easily usable set of biological models for cancer biology investigations and exploring of the potential efficacy of anticancer drugs. The drug/cell line sensitivity data has been accumulated in public domain and pharmaceutical industry. The first attempt to establish such resources is the NCI-60 Human Tumor Cell Lines Screen that has served the global cancer research community for more than 20 years [1]. There are also two other publicly available large-scale pharmacogenomics resources: The Genomics of Drug Sensitivity in Cancer (GDSC) [2] and the Cancer Cell Line Encyclopedia (CCLE) [3] databases. Large number of genomic features like gene expression, copy number alteration and mutation are curated in these databases, as well as the compound information. These databases can be useful for building predictive models of cell line-drug sensitivity.

In ExCAPE project, curating public cancer cell line data is one of our goals and these cancer cell line data will be used as a benchmark dataset for the multi-task learning algorithms developed in WP1 and WP2. By building predictive models for cell line drug sensitivity, our aim is to identify chemical, genomic and their interaction patterns of cancers that predict response to anti-cancer drugs and ultimately, to discover biomarkers which can predict drug-sensitivity/resistance for precise medicines.

3 The Genomics of Drug Sensitivity in Cancer Project (GDSC)

The GDSC is one of the large-scale efforts to characterize genomes, mRNA expression, and anti-cancer drug dose-responses across cell lines [2]. Those genomic features and drug response data like IC50 were well established in a recent study [4]. The GDSC cancer cell line dataset comprise cell inhibition (IC50, GI50) data for 265 anti-cancer drugs cross 990 different cell lines. The GDSC cell line genome data were downloaded from GDSC web site, the gene expression data of a set of 1496 cancer related genes (GE), 241 driver mutations (Mut), 417 recurrent copy number altered chromosomal segments (CNV) and 279 informative CpG sites (Methylation) data were compiled to represent cell line genome features.

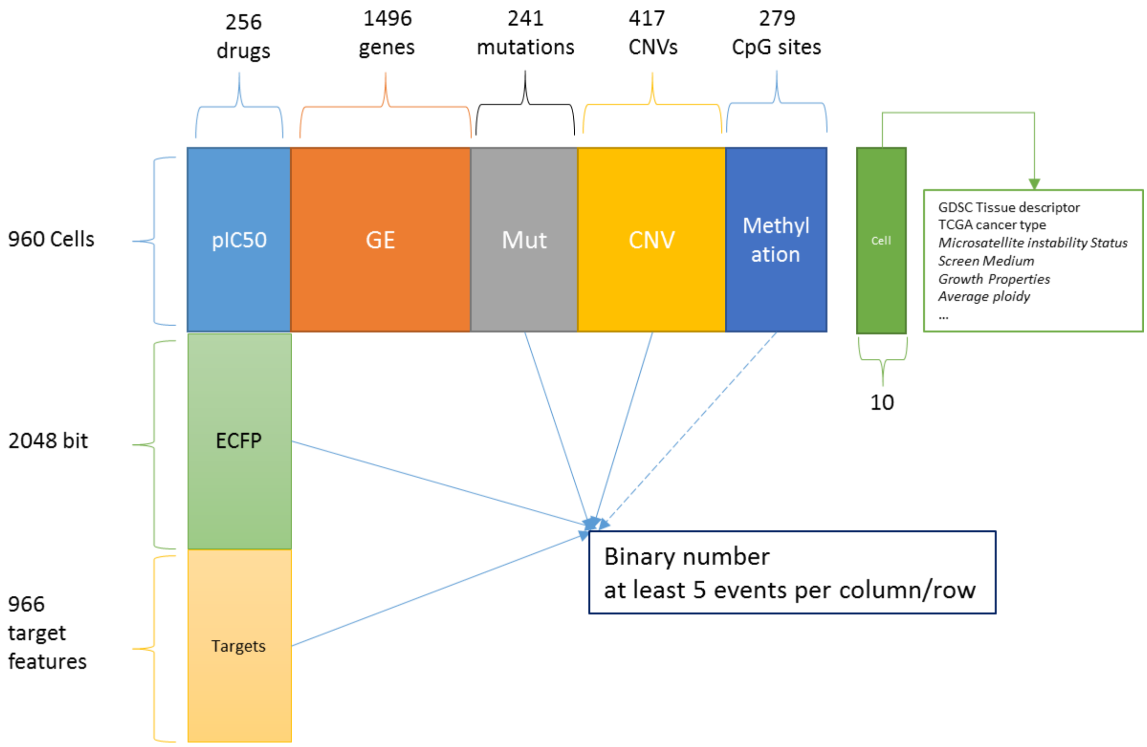

Figure s3-1. The composition of GDSC data matrix

3.1 *Cell line*

The GDSC collection comprises 1001 human tumour cell lines. Cell lines have been categorized based on disease relevant tissue descriptions (GDSC descriptions 1 and 2, site and Histology), as well as using the TCGA tumour type descriptions. Analysis of microsatellite instability (MSI) was carried out to check if genetic hypermutability that resulted from impaired DNA mismatch repair. Cells were grown in RPMI or DMEM/F12 medium supplemented with 5% or 10% FBS and penicillin/streptomycin in order to minimize the potential effect of varying the media on sensitivity to therapeutic compounds in the assays. Cells were cultured using three different growth properties: adherent, semi-adherent or suspension.

3.2 Driver mutation

Variants were calling from whole exome sequencing (Agilent SureSelectXT Human All Exon 50Mb bait set) data and were screened against approx. 8,000 normal samples to remove sequencing artefacts and germline variants as well as variants in the dbSNP database. The remaining putatively somatic variants were classed as validated if present in other large scale cell-line sequencing datasets like CCLE, NCI-60 and previous findings. The remaining variants, together with other high quality variants were named as cell line variants. Variants from both the cell lines and tumors were screened against a derived large-scale clinical datasets to identify the recurrent variants most likely to contribute to carcinogenesis (‘driver mutations’). The diver mutations were encoded as a binary event matrix (1: present, 0: absent).

3.3 Copy number

The chromosomal segments copy numbers were measured by Affymetrix SNP6.0 Array in cells. A set of Recurrent Copy number altered chromosomal segments (RACSs) would be obtained if such segment copy numbers in cells was also presented in human tumors (TCGA, COAD and READ). Finally, copy number alteration were encoded as a binary event matrix (1: present, 0: absent).

3.4 DNA Methylation

The Cell line methylation profile was measured by Illumina Human Methylation 450 Array. After pre-processing, the beta signal of each CpG island was analyzed in the context of each cancer type. Due to the high tissue specificity of the methylation profiles, a systematic Hartigan’s dip test for unimodality was executed with the aim of identifying a set of CpG islands for which this signal did not distribute unimodally. Such CpG islands were deemed unlikely to be tissue specific, hence consistently hyper methylated or hypo methylated across all the samples and were considered informative. Finally, a binary event matrix with hyper-methylation status of such informative CpG islands (iCpG) were constructed if such iCpGs were also present in human tumors (TCGA).

3.5 *Gene expression*

Gene expression were quantified by Affymetrix Human Genome U219 Array. Raw data intensity values were normalized by Robust Multi-array Average (RMA) [5] approach. That is background corrected, log2 transformed and then quantile normalized. Probe was annotated using human genome version 19. Further data processing was performed to remove batch effects caused by growth properties. An empirical Bayes method Comba [6] following was applied to remove effects of batch.

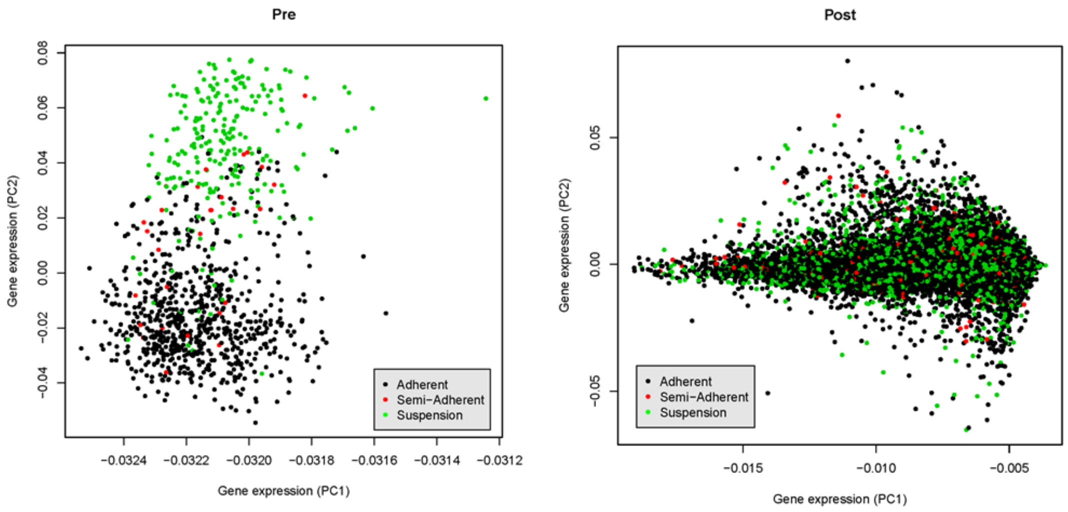

Figure s3-2. Overview of gene expression data before (left) and after (right) batch removal.

3.6 *Drug response*

Fluorescence intensity data from screening plates for each dose response curve is fitted using a multi-level fixed effect model [7] to estimate IC50. Natural log half maximal inhibitory concentration (IC50) and Area under the dose-response curve (AUCs) values for all screened cell lines were download from http://www.cancerrxgene.org/downloads and IC50 data was converted to pIC50. A pIC50 matrix with 960 cell as rows and 256 drugs as column was constructed. 18.2% of values were missing in the matrix since they were not yet measured.

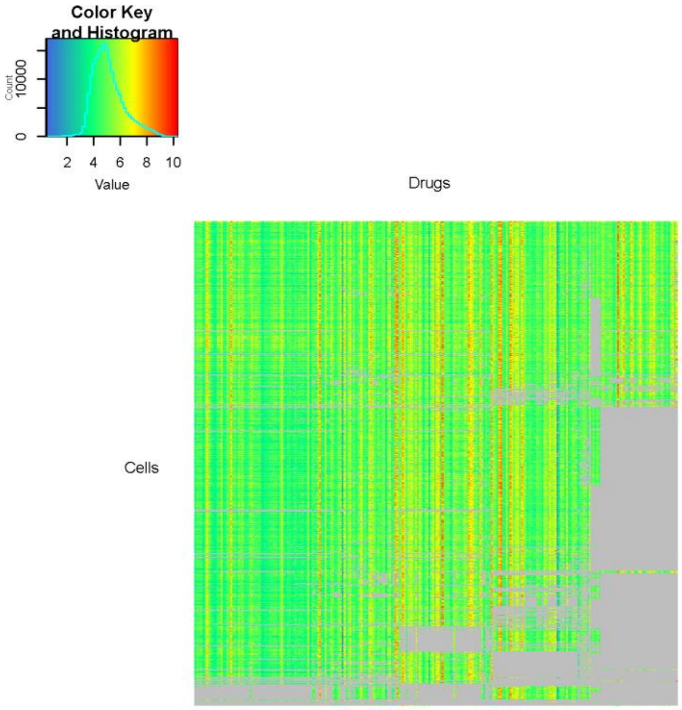

Figure s3-3. Distribution of drug response data in the GDSC. The pIC50 were showed in the heat map. Grey color donated missing values.

3.7 Compound

Structures of 256 compounds can be obtained searching by drug name. Using SMILES as input, compound chemical features were represented by a 2048-length bit string Extended-Connectivity Fingerprints (ECFPs). The jCompoundMapper software [8] was used to generate ECFP fingerprints by setting search depth to 6.

3.8 Network feature

Moreover, we identified the known annotated targets of those drugs as primary targets and then using these primary targets as input to further extend the target list by retrieving the first layer neighbors of the primary target from protein-protein interaction (PPI) networks [9-10]. These first layer neighbor proteins in the PPI are also regarded as targets for a specific drug and are combined with the primary targets to form the target fingerprints for each drug. In total a binary 966 bit-vector (i.e. target fingerprint) were composed for each drug.

4 The Cancer Cell Line Encyclopedia (CCLE)

The CCLE [3] is an ongoing project 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.

4.1 Cell line

The CCLE collection comprises 1 046 human tumor cell lines. Cell lines have been categorized based on disease relevant descriptions like site Histology as well as using the TCGA tumor type descriptions.

4.2 Copy number

The chromosomal segments copy numbers were measured by Affymetrix SNP6.0 Array in cells. Raw Affymetrix CEL files were converted to a single value for each probe set representing a SNP allele or a copy number probe. Copy numbers were then inferred based upon estimating probe set specific linear calibration curves, followed by normalization by the most similar HapMap normal samples. Segmentation of normalized log2 ratios (specifically, log2(CN/2)) was performed using the circular binary segmentation (CBS) algorithm. Finally, copy number alteration were encoded as a binary event matrix (1: present, 0: absent).

4.3 Mutations

Mutation information was obtained both by using massively parallel sequencing of >1,600 genes and by mass spectrometric genotyping (OncoMap), which interrogated 492 mutations in 33 known oncogenes and tumour suppressors. Finally, mutations were encoded as a binary event matrix (1: present, 0: absent).

4.4 Gene expression

Gene expression were quantified by Affymetrix U133 Plus 2.0 arrays. Raw Affymetrix CEL files were converted to a single value for each probe set using Robust Multi-array Average (RMA) and normalized using quantile normalization.

4.5 Drug response

Eight-point dose–response curves were generated for 24 anticancer drugs across 504 cell lines using an automated compound-screening platform. These curves were represented by a logistical sigmoidal function with a maximal effect level, the concentration at half-maximal activity of the compound (EC50), a Hill coefficient representing the sigmoidal transition, and the concentration at which the drug response reached an absolute inhibition of 50% (IC50).

4.6 Compound

Compound structures were retrieved from GDSC if the compound/drug name is identical. Otherwise, structures were searched in the PubChem using compound name as input. Structures of all 24 anticancer drugs were showed as SMILES format.

5 The Cancer Therapeutics Response Portal (CTRP)

The CTRP [11] is a living resource that links genetic, lineage, and other cellular features of cancer cell lines to small-molecule sensitivity with the goal of accelerating discovery of patient-matched cancer therapeutics. The CTRP v1 provides open access to the results obtained through quantitatively measuring the sensitivity of 242 genetically characterized cancer-cell lines (CCLs) to a 185 small-molecule probes and drugs. In CTRP v2, 481 compounds and drugs sensitivity across 860 CCLs.

5.1 Cell line and genomic data

The genomic data, analysis and visualization for cell lines used in CTRP can be accessed from the Cancer Cell Line Encyclopedia.

5.2 Drug response

The response of CCLs to each member of the Informer Set over a 16-point concentration range using an automated, high-throughput workflow, fit concentration-response curves, and calculated the area under the curve (AUC) as a measure of sensitivity.

5.3 Compound

The structures of 355 and 545 compounds were provided by CTRP V1 and V2 in the SMILES format, respectively.

6 NCI-60 data

The NCI-60, a panel of 60 diverse human cancer cell lines used by the Developmental Therapeutics Program of the U.S. National Cancer Institute to screen over 100,000 chemical compounds and natural products since 1990 [12]. A web application CellMiner [13] generated by the Genomics & Bioinformatics Group, LMP, CCR, NCI that facilitates systems biology through the retrieval and integration of the molecular and pharmacological data sets for the

NCI-60 cell lines. In this delivery, genomic features of 60 cancer cell lines and drug response over 20 000 compounds were included.

6.1 Cell line

The NCI-60 collection comprises 60 human tumor cell lines. Cell lines have been categorized based on tissue relevant descriptions like site Histology and tissue of origin. The ploidy and doubling time of cells were also provided.

6.2 Copy number

DNA copy numbers for all genes were determined by the integration of probes from 1) the Human Genome CGH Microarray 44A (Agilent Technologies) with 44 k probes, 2) the H19 CGH 385K WG Tiling v2.0 array (Roche NimbleGen Systems) with 385 k probes, 3) the GeneChip Human Mapping 500 k Array Set (Affymetrix Technologies) with 500 k probes, and 4) the Human Human1 Mv1_C Beadchip array (Illumina) with 1,100 k probes. Data for these microarrays can be accessed at CellMiner.

6.3 Mutations

Mutation information was measured by Exome-seq using paired-end 80-mer reads on an Illumina Genome Analyzer IIx instrument (Illumina). Fastq files were aligned against the reference human genome build 19 (hg19) using the Burrows-Wheeler Aligner. Alignment files were base quality score recalibrated and locally realigned around indels with GATK and marked for duplicates using PICARD tools (picard.sourceforge.net). Alignment files and variant calls can be accessed from the links provided. Consensus genotype calls were generated using samtools mpileup and annotated using the Annovar package. Variants were further filtered for the SureSelect bait region, a minimum read depth of 6 and a minimum quality score of 30 for single nucleotide variant (SNV) and 60 for indels, producing the final variant calls. Mutations can be retrieved from CellMiner.

6.4 DNA methylation

The Cell line methylation profile was measured by Illumina Human Methylation 450 Array. Approximately 450,000 probes querying the methylation status of CpG sites within and outside of genes. Beta values was normalized to a value between 0 (unmethylated) and 1 (methylated) using the R methylumi package.

6.5 Gene expression

Transcript expression for each gene was determined through the integration of all pertinent probes from five platforms: the Human Genome U95 Set (HG-U95); the Human Genome U133 (HG-U133); the Human Genome U133 Plus 2.0 Arrays (HG-U133 Plus 2.0); the GeneChip Human Exon 1.0 ST array (GH Exon 1.0 ST) and the Whole Human Genome Oligo Microarray (WHG). Probe intensity values that pass quality controls are transformed to z-scores. Average z-scores were determined for each gene for each cell line. The CellMiner tools provide a portal for rapid data retrieval of transcripts for 22,379 genes along with activity reports for 20,503 chemical compounds including 102 drugs approved by the U.S. Food and Drug Administration.

6.6 Drug response

Drug activity levels expressed as 50% growth-inhibitory levels (GI50) are determined by the DTP (http://dtp.nci.nih.gov/) at 48 hours using the sulforhodamine B assay. Negative log10 (GI50) values of sulforhodamine B assay for about 50K compounds, including more than 20,000 that passed quality control, 158 Food and Drug Administration approved and 79 clinical trial drugs. Higher values equate to higher sensitivity of cell lines.

6.7 Compound

The structures of 21,012 compounds were provided by DTP-NCI60 in the SMILES format.

7 Conclusion

Four large scale public cancer cell line data were collected in this delivery. It comprise genomic features like gene expression, mutation, copy number and chemical feature like ECFP. The data was cleaned and ready to be used as a starting point to predict cancer drug response. In particular, the GDSC, a well-established data, shared genomic features with AstraZeneca proprietary data. This unique features made GDSC quite suitable for testing machine learning algorithms within ExCAPE project for WP1 and WP2.

8 Annexes

Data storage at IT4I Salomon server with the path

/scratch/work/project/excape-public/data/az_cancer_cell_line/

9 References

1. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006, 6, 813-23.

2. GDSC database, Http://Www.Cancerrxgene.Org.

3. CCLE database, Https://Portals.Broadinstitute.Org/Ccle/Home.

4. Iorio, F.; Knijnenburg, T. A.; Vis, D. J.; Bignell, G. R.; Menden, M. P.; Schubert, M.; Aben, N.; Goncalves, E.; Barthorpe, S.; Lightfoot, H.; Cokelaer, T.; Greninger, P.; van Dyk, E.; Chang, H.; de Silva, H.; Heyn, H.; Deng, X.; Egan, R. K.; Liu, Q.; Mironenko, T.; Mitropoulos, X.; Richardson, L.; Wang, J.; Zhang, T.; Moran, S.; Sayols, S.; Soleimani, M.; Tamborero, D.; Lopez-Bigas, N.; Ross-Macdonald, P.; Esteller, M.; Gray, N. S.; Haber, D. A.; Stratton, M. R.; Benes, C. H.; Wessels, L. F.; Saez-Rodriguez, J.; McDermott, U.; Garnett, M. J., A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **2016,** *166*, 740-54.

5. Irizarry, R. A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y. D.; Antonellis, K. J.; Scherf, U.; Speed, T. P., Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* **2003,** *4*, 249-64.

6. Johnson, W. E.; Li, C.; Rabinovic, A., Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **2007,** *8*, 118-127.

7. Vis, D. J.; Bombardelli, L.; Lightfoot, H.; Iorio, F.; Garnett, M. J.; Wessels, L. F., Multilevel Models Improve Precision and Speed of Ic50 Estimates. *Pharmacogenomics* **2016,** *17*, 691-700.

8. Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Zell, A., Jcompoundmapper: An Open Source Java Library and Command-Line Tool for Chemical Fingerprints. *Journal of cheminformatics* **2011,** *3*, 3.

9. Fazekas, D.; Koltai, M.; Turei, D.; Modos, D.; Palfy, M.; Dul, Z.; Zsakai, L.; Szalay-Beko, M.; Lenti, K.; Farkas, I. J.; Vellai, T.; Csermely, P.; Korcsmaros, T., Signalink 2 - a Signaling Pathway Resource with Multi-Layered Regulatory Networks. *BMC Syst Biol* **2013,** *7*, 7.

10. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C., String V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic acids research* **2015,** *43*, D447-52.

11. CTRP database, Http://Portals.Broadinstitute.Org/Ctrp/.

12. NCI-60 database, Https://Dtp.Cancer.Gov/Discovery_Development/Nci-60/.

13. CellMiner. Https://Discover.Nci.Nih.Gov/Cellminer/.