Ph.D. Dissertation of Science

Systematic Analysis of Digenic Interactions to Discover Genetic Interactions Affecting Cancer Survival

- Synthetic Dosage Cancer Survival Analysis -

암 생존에 영향을 미치는 유전적 상호작용을 발견하기 위한 유전적 상호작용의 체계적인 분석: 정량 합성 암 생존 분석

February 2023

THE DEPARTMENT OF BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Jeong Hoon Lee

Systematic Analysis of Digenic Interactions to Discover Genetic Interactions Affecting Cancer Survival

- Synthetic Dosage Cancer Survival Analysis -

Advisor

Professor Ju Han Kim, M.D., Ph. D.

Submitting a Ph.D. Dissertation of Science

February 2023

THE DEPARTMENT OF BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Jeong Hoon Lee

Confirming the Ph.D. Dissertation written by
Jeong Hoon Lee

February 2023

Chair	황 대 희 (Seal)
Vice Chair	김 주 한 (Seal)
Examiner	한 경 숙 (Seal)
Examiner	김 준 태 (Seal)
Examiner	이 상 우 (Seal)

ABSTRACT

Genetic interactions occur when two or more gene mutations combine to generate a phenotype such as cell lethality. Several anticancer therapies have exploited genetic interactions by targeting somatic mutations and the overexpression of oncogenes; these therapies target tumor pathways for survival without affecting normal cell. Because this concept of genetic interactions utilizes cell lethality as a phenotype, numerous bottlenecks exist in the discovery of new genetic interactions using computational methods. To overcome this limitation, I defined the phenotype of genetic interactions at the patient level and not at the cell level. In this study, I propose synthetic dosage cancer survival (SDCS), a modified concept of synthetic dosage lethality, in which a combination of a mutation and an overexpressed gene causes cell lethality. SDCS involves a pair of genes, in which a combination of a mutation and overexpression of a gene leads to significant differences in patient survival. A gene combination that improved patient survival was defined as a positive SDCS pair, whereas one that worsened prognosis was defined as a negative SDCS pair. SDCS pairs were identified and validated using two databases: The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). The genotype-tissue expression (GTEx) database was used as a control. To confirm the possibility of the over-expressed genes of SDCS as druggable targets, the Genomics of Drug Sensitivity in Cancer (GDSC) database was used to validate the SDCS pairs. Twenty-two positive and 35 negative SDCS pairs were identified and validated. These SDCS pairs comprised 18 gene disruptions and 52 overexpressed genes. The combination of *PIK3CA* disruption and *MTOR* overexpression, which is a negative SDCS pair, is a potential drug target that has

recently attracted interest as a dual inhibitor for breast cancer and was validated through drug sensitivity analysis. Breast cancer cell line samples with *PIK3CA* mutations and *MTOR* overexpression were significantly sensitive to omipalisib and OSI-027, which are MTOR inhibitors. This observation suggests that the genes included in the SDCS pairs could be potential candidates for developing new drugs for cancer therapy. Thus, SDCS analysis can help to identify novel therapeutic and prognostic targets.

Keywords: Negative genetic interaction, Synthetic dosage lethality, Gene expression analysis, Drug sensitivity, Target discovery, Prognosis marker

Student Number: 2015-20509

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER 1. Introduction

1.1. Genetic interaction

Genetic interactions occur when two or more genetic events combine to generate an unexpected phenotype such as cell lethality (Brough et al., 2011; Kaelin, 2005; Kuzmin et al., 2018; O'Neil et al., 2017; Typas et al., 2008). This is a type of targeted therapy enables individualized treatment based on the characteristics of cancer cells in each patient. With recent advances in genome sequencing, changes in a patient's genomics can be identified rapidly to characterize the biological functions of cancer cells and to identify vulnerabilities that can be exploited to selectively kill cancer cells using therapeutics(Katti et al., 2022; Molina et al., 2018). Therefore, if we uncover genetic interactions, more personalized therapeutics for cancer patients could be developed, thereby improving patient prognosis.

Genetic interactions can be categorized into two types: negative and positive (Dixon et al., 2009; Kuzmin et al., 2018; O'Neil et al., 2017; Tong et al., 2004). In a negative genetic interaction, two or more genetic events contribute to a more serious fitness defect in a cell than expected for each genetic event. Representative examples include synthetic lethality (SL) and synthetic dosage lethality (SDL) (Figure 1)(Paul et al., 2014). SL occurs when two combinatory genetic alterations generate a lethal phenotype in a cell, but individual genetic alterations do not. The most representative drug developed using SL is the PARP inhibitor (Hodgson et al., 2018; Lord & Ashworth, 2017; Tutt et al., 2009). Two genes, BRCA and PARP, were representative

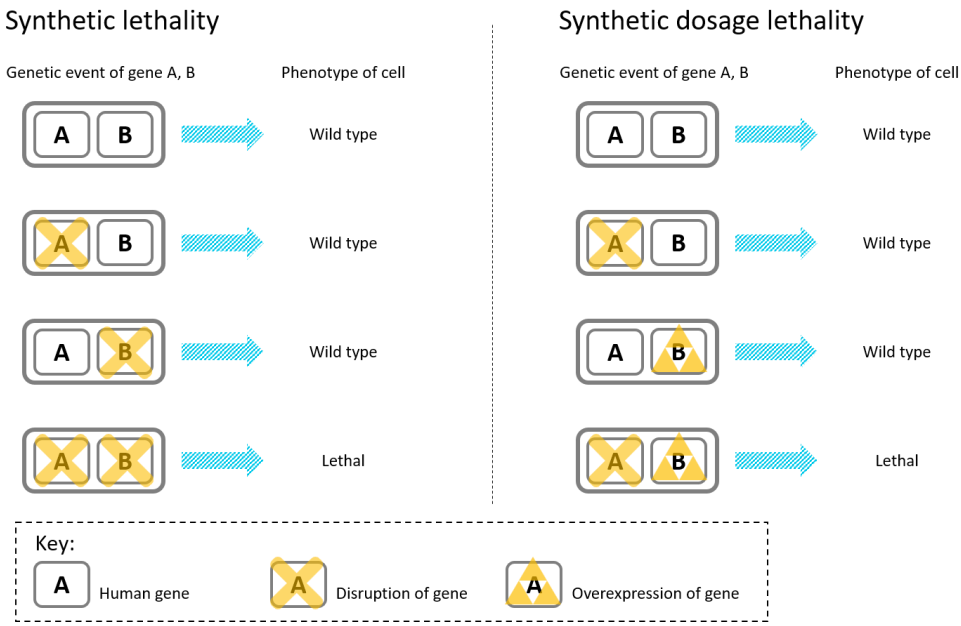

Figure 1. Negative genetic interactions, SL and SDL. Two examples of digenic interactions that negatively affect cell fitness: SL in which cell lethality occurs because of a simultaneous event involving two genes; SDL in which the combination of a genetic event with overexpression of another gene causes lethality.

SL pairs, and the *PARP* gene can be exploited in patients with breast cancer, where genetic alterations in *BRCA* are common. SDL refers to cases in which gene mutations and overexpression cause cell death but not individually (Kroll et al., 1996; Measday & Hieter, 2002; Megchelenbrink et al., 2015). Unlike normal cells, cancer cells have somatic mutations and exhibit gene overexpression; thus, by exploiting the concepts of SL and SDL, cancer cells can be specifically treated. Alternatively, positive genetic interactions occur when a combination of genetic alterations results in greater fitness than expected.

Negative genetic interactions change the cell phenotype to lethality(Baryshnikova et al., 2013; Vizeacoumar et al., 2013). This can be the underlying concept for developing cancer therapies to selectively kill cancer cells without significantly affecting the normal cells of cancer patients by utilizing the characteristics of cancer cells that have somatic mutations(Han et al., 2019). Similarly, cancer-cell-specific death can be induced by utilizing SDL because numerous oncogenes are overexpressed in cancer cells (O'Neil et al., 2017). Figure 2 shows a method for developing cancer-cell–specific anticancer drugs using negative genetic interactions. Cancer cells with a disruption in gene A can induce SL using an inhibitor of gene B, as if the cells had a combination of genetic alterations in both genes A and B. The inhibitor of gene B did not have a significant effect on normal cells without genetic alterations to gene A. This is also applicable to SDL. An inhibitor of gene B, a partner of SDL, can be applied for cancer therapy by exploiting the overexpression of gene A, which is specific to cancer cells.

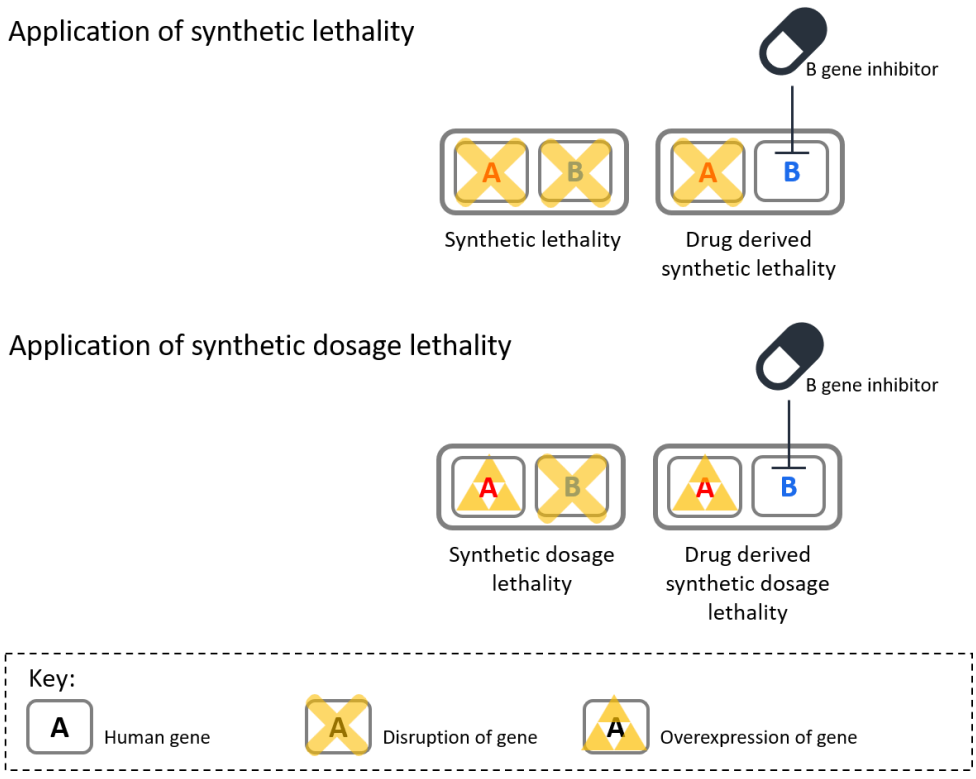

Figure 2. Application of negative genetic interactions. Drug-derived SL can be induced using an inhibitor of the partner gene. SDL can also exploit cancer cells using inhibitors.

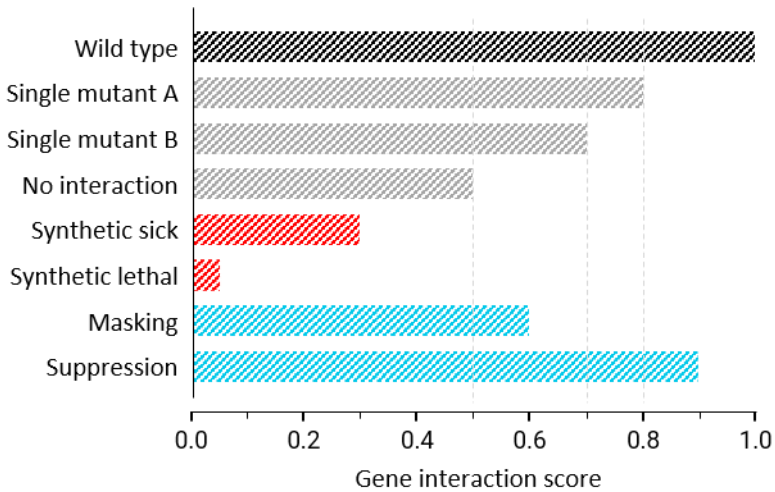

Figure 3. Digenic interaction types classified by genetic interaction score with multiplicative models suggested by Mani et al. (2008).

The degree of genetic interaction is calculated mainly based on the growth rate or colony size of cells as a phenotype (Tong et al., 2004). This was performed to evaluate the effects of combining different genetic interactions. Multiplicative models have previously been used to measure genetic interactions (Mani et al., 2008; van Leeuwen et al., 2017). By measuring the change in the phenotype of cells according to the genetic events of genes A/B, the genetic interactions of the A/B gene pair can be categorized (Figure 3). If the effects of the event associated with gene A and those associated with gene B are cumulative, the two genes are defined as having no genetic interactions. However, if the events associated with the two genes negatively affect the phenotype of the cell, it is called synthetic sick, and the case in which most cells die, it is called synthetic lethal, which is a negative genetic interaction. In contrast, when the events of two genes occur simultaneously, and if they affect the phenotype to a lesser extent than the cumulative sum of each effect, the process is called masking. If this leads to a better phenotype than the effect of each event, this is a suppression. Digenetic interactions, masking, and suppression are defined as positive genetic interactions.

With the advent of screening methods, including yeast, drug, RNA interference (RNAi), and clustered regularly interspaced short palindromic repeat (CRISPR) screening, a profound increase in the number of candidate genetic interactions has been observed, and drugs targeting many of these interactions are currently in the developmental stages (Castells-Roca et al., 2021; Katti et al., 2022; Luo et al., 2009; McDonald III et al., 2017; C. Wang et al., 2019). Synthetic genetic array (SGA) analysis was used to successfully combine double mutants in the yeast genome to identify candidate synthetic lethal pairs comprising ~550,000 negative genetic

interactions and ~350,000 positive interactions among ~18 million double mutant pairs (Ooi et al., 2006; Yan Tong & Boone, 2006). Despite numerous experiments based on the attractive concept of SL-based target therapy, only PARPi has entered the clinic to date(Mateo et al., 2019). Because of the large number of combinatorial explosions, it is very expensive and time-consuming screening for all possible genetic interactions is expensive and time consuming.

1.2. Computational approach for genetic interactions

Computational approaches can reduce the candidate SL gene pairs that need to be screened. Computational methods can be divided into four categories: statistical-, network-, classical machine-learning–based methods, and deep-learning-based methods (J. Wang et al., 2022). Data mining synthetic lethality identification pipeline (DAISY) is a representative statistical-based method that is based on the assumption that SL pairs tend to be co-expressed but not inactivated simultaneously (Jerby-Arnon et al., 2014; Lee et al., 2021; Srihari et al., 2015). Network-based approaches are based on protein–protein interactions, co-expression networks, signaling networks, or metabolism networks that describe biological interactions. Machine learning and deep-learning–based methods have also recently attracted attention for identifying genetic interactions (Li et al., 2019; Madhukar et al., 2015; Wan et al., 2020). In particular, graph neural networks (GNNs) can efficiently utilize graph-structured data (Cai et al., 2020; S. Wang et al., 2021). SL and SDL can be structured in a graph format, numerous methods to identify new genetic interactions using GNNs have been proposed.

Although computational approaches can complement experimental methods, there are limitations to predicting genetic interactions based on data (J. Wang et al., 2022). Negative genetic interactions causing strong lethality cannot be confirmed in patients with cancer because the cells are dead and cannot be observed in the patient, which is a characteristic property of SL and SDL. In other words, analysis is possible only at the cellular level, which limits its clinical use. Therefore, a computational approach can be used to suggest candidates only at the molecular or cellular level, which makes verification at the clinical level challenging.

1.3. Overview of thesis

This study aimed to identify extended genetic interactions using the concept of synthetic dosage cancer survival (SDCS). The proposed SDCS analysis involves identifying genetic interactions using the patient's prognosis and not cell death as a phenotype. Similar to genetic interactions, SDCS can be divided into two categories: positive SDCS pairs that improve a patient's survival and negative SDCS pairs that worsen a patient's survival. Both cases can be used as prognostic markers for patients and as important information for the development of anticancer drugs.

SDCS analysis deviates from genetic interactions at the cellular level, and investigates genetic interactions at the patient level. Through the genetic interactions identified by SDCS analysis, novel therapies utilizing the relationships between genetic interactions can be developed. For example, inhibiting the over-expressed gene of SDCS or investigating the role of the overexpressed gene could provide candidates for druggable targets in cancer cells.

CHAPTER 2. MATERIALS AND METHODS

2.1. Overview of the TCGA and ICGC database

The Cancer Genome Atlas (TCGA) is a multi-omics project to create a comprehensive “atlas” of the multi-omics cancer genome profiles to catalog and identify causes of cancer(Tomczak et al., 2015; Weinstein et al., 2013). The International Cancer Genome Consortium (ICGC) is a large-scale cancer genome study based on multiomics cancer genome profiles (Zhang et al., 2011, 2019). Both databases provide DNA and RNA sequencing (RNA-Seq) data acquired from the same patient with clinical information. The UCSC Xena platform (https://https://xenabrowser.net/) provides public multi-omics data and clinical datasets from large-scale genome studies, including TCGA, ICGC, and the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012; Goldman et al., 2017, 2019). These multi-omics databases include whole-exome sequencing (WXS), RNA-Seq mRNA expression, and copy number variation (CNV). All the data provided by TCGA, ICGC, and the genotype-tissue expression (GTEx) were downloaded from the UCSC Xena platform(Consortium et al., 2015). The Catalog of Somatic Mutations in Cancer (COSMIC) and Genomics of Drug Sensitivity in Cancer (GDSC) data were downloaded from official websites (Bamford et al., 2004; Tate et al., 2019; Yang et al., 2012).

2.2. Clinical information

Clinical information on 19 types of cancer was collected from TCGA and ICGC databases. Important clinical variables were collected for each cancer type, and progression-free survival (PFS) was used as the dependent variable. The clinical variables included age, sex, stage, microsatellite instability (MSI), smoking status, alcohol consumption, grade, necrosis percentage, residual cancer, and focality. Because the important clinical variables differed for each cancer type, they were assigned differently according to cancer type and used as input variables for analysis. Table 1 shows a list of the 19 cancer types included in this analysis and a list of clinical variables used. Survival was used as the dependent variable and PFS was calculated. The variable most commonly used for all cancers was age. For urogenital cancer types that occur predominantly in a specific gender, such as breast, cervical, prostate, and uterine cancers, the gender variable was not used as an input. In the ICGC database used for validation, only the dependent variables PFS and age were used. The types of clinical variables were different, and there were numerous missing values; therefore, it was not possible to compose data identical to TCGA data.

In this study, normal samples were used for comparison with cancer samples. Normal RNA-seq samples included paired-normal samples obtained from TCGA database and data from the Genotype-Tissue Expression (GTEx) database. When using normal RNA-seq data, arbitrarily generated clinical information was used because the clinical variables of the samples were normal. Because cells in the normal sample did not die due to cancer, all survival events were assigned as censored, and longer follow-up times than those of the cancer patients with the longest follow-up time for each cancer type were arbitrarily assigned to all samples.

Table 1. List of cancer types and clinical variables used in the TCGA database.

Cancer Type	Clinical variables
Bladder urothelial carcinoma (BLCA)	Survival, age, gender and stage
Breast invasive carcinoma (BRCA)	Survival, age and stage
Cervical and endocervical cancers (CESC)	Survival, age and stage
Colon adenocarcinoma (COAD)	Survival, age, gender, stage and microsatellite instability
Glioblastoma multiforme (GBM)	Survival, age, gender
Head and Neck squamous cell carcinoma (HNSC)	Survival, age, gender, smoking status and alcohol
Kidney renal clear cell carcinoma (KIRC)	Survival, age, gender, stage, tumour grade and tumor necrosis percent
Kidney renal papillary cell carcinoma (KIRP)	Survival, age, gender, stage and tumor necrosis percent
Brain Lower Grade Glioma (LGG)	Survival, age, gender and tumour grade
Liver hepatocellular carcinoma (LIHC)	Survival, age, gender, residual and tumour grade
Lung adenocarcinoma (LUAD)	Survival, age, gender, stage and smoking status
Lung squamous cell carcinoma (LUSC)	Survival, age, gender, stage and smoking status
Ovarian serous cystadenocarcinoma (OV)	Survival, age, gender, stage and tumour grade
Pancreatic adenocarcinoma (PAAD)	Survival, age, gender and stage
Prostate adenocarcinoma (PRAD)	Survival, age and stage and residual
Skin Cutaneous Melanoma (SKCM)	Survival, age, gender and stage
Stomach adenocarcinoma (STAD)	Survival, age, gender and stage
Thyroid carcinoma (THCA)	Survival, age, gender, stage and focality
Uterine Corpus Endometrial Carcinoma (UCEC)	Survival, age and stage

Table 2. GTEx tissue matched based on the organ for the TCGA/ICGC cancer type.

TCGA	TCGA Full Name	GTEx Database
BLCA	Bladder urothelial carcinoma	Bladder
BRCA	Breast invasive carcinoma	Breast
CESC	Cervical and endocervical cancers	Cervix Uteri
COAD	Colon adenocarcinoma	Colon
GBM	Glioblastoma multiforme	Brain
HNSC	Head and Neck squamous cell carcinoma	Salivary Gland
KIRC	Kidney renal clear cell carcinoma	Kidney
KIRP	Kidney renal papillary cell carcinoma	Kidney
LGG	Brain Lower Grade Glioma	Brain
LIHC	Liver hepatocellular carcinoma	Liver
LUAD	Lung adenocarcinoma	Lung
LUSC	Lung squamous cell carcinoma	Lung
OV	Ovarian serous cystadenocarcinoma	Ovary
PAAD	Pancreatic adenocarcinoma	Pancreas
PRAD	Prostate adenocarcinoma	Prostate
SKCM	Skin Cutaneous Melanoma	Skin
STAD	Stomach adenocarcinoma	Stomach
THCA	Thyroid carcinoma	Thyroid
UCEC	Uterine Corpus Endometrial Carcinoma	Uterus

To match the RNA-Seq samples from the GTEx database to the same cancer tissue samples, data derived from the same organ for each cancer type in the TCGA database were matched (Table 2).

2.3. Whole exome sequencing

2.3.1. Next generation sequencing

For DNA sequencing data, multicenter mutation calling multiple cancer (MC3) project data were used (Ellrott et al., 2018). This project enables robust cross-tumor-type analysis of variance and batch effects introduced by DNA extraction, hybridization capture, and sequencing. This project provides refined DNA sequencing data, including pipelines: Alignment, The Genome Analysis Toolkit, MuTect, and Indelocator, as well as Pindel, MuSE, Radia, Varscan, and Somatic Sniper (Benjamin et al., 2019; Koboldt et al., 2012; Larson et al., 2012; McKenna et al., 2010; Ye et al., 2009).

2.3.2. Variant annotation

For all TCGA/ICGC DNA sequence samples, variants, insertions, and deletions (INDELs) were annotated using the ensemble variant effect predictor (VEP)(McLaren et al., 2016). Annotation includes the Sorting Intolerant from Tolerant (SIFT) algorithm, which predicts the effect of coding variants on protein function, and PolyPhen v2 (Polymorphism Phenotyping v2), which predicts the impact of an amino acid substitution on the structure and function of a human protein

(Adzhubei et al., 2013; Sim et al., 2012). The combined annotation-dependent depletion (CADD), which predicts the deleteriousness of variants and INDELs, was included(Kircher et al., 2014). LoFtool, which measures gene intolerance and susceptibility using the frequency of loss-of-function (LoF) mutations, was also included (Fadista et al., 2017). After using VEP, the variant with the most severe consequence was used for multiple annotated variants and INDELs. LoF consequences included transcript ablation, splice acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, start lost, and transcript amplification (Table 3). These eight consequences are cases that have high impacts, as predicted by SNPEff and SNPSift (Cingolani, 2022; Cingolani, Patel, et al., 2012; Cingolani, Platts, et al., 2012). If copy number variation (CNV) leads to a complete loss, the gene is disrupted. Finally, if a variant or INDEL was classified as a pathogenic variant in the ClinVar database, this gene was considered a disrupted gene (Landrum et al., 2016). If none of the above conditions were satisfied, the gene was not considered to be disrupted.

2.3.3. Genetic disruption

In this SDCS analysis, genetic interactions were investigated at the gene level, but not at the variant and INDEL levels. Therefore, variants and INDELs were aggregated into units of genes to construct a binary matrix according to the presence or absence of gene disruption. When any one of the variants, INDELs, or CNV requirements were satisfied, it was determined that a gene disruption was determined. If one condition was unsatisfactory, it was considered that there was no gene disruption. First, if at least one loss-of-function variant exists in a gene, the gene is

considered a disruption gene. The consequences of the loss-of-function variants were as follows: transcript ablation, splice acceptor variant, splice donor variant, stop gained, frame-shift variant, stop lost, and start lost. Second, a disruption gene is a gene containing variants with a deleterious score for pathogenicity prediction. For the pathogenicity prediction algorithm, the SIFT, PolyPhen-v2, and CADD scores were used. The threshold of each deleterious score was 0.05 or less for SIFT, 0.85 or more for PolyPhen-v2, and 30 or more for CADD score. Third, using the CNV change at the gene level, GISTIC2 created a threshold of the data provided by TCGA, and genes with completely lost copy numbers with a value of-2 were defined as disruption genes. Finally, genes with mutations classified as pathogenic variants in the ClinVar database were defined as disruption genes. Genes that did not satisfy all the above four conditions were considered wild-type genes. We excluded 2086 genes with a LoFtool score of 0.85 or higher from this analysis because they were disrupted genes that occur frequently in normal samples.

Table 3. The consequences of the variants and INDELs in order of severity estimated by Ensembl.

Sequence ontology	Description
Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature
Splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an intron
Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron
Stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript
Frame-shift variant	A sequence variant that disrupts the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three
Stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
Start lost	A codon variant that changes at least one base of the canonical start codon
Transcript amplification	A feature amplification of a region containing a transcript
Inframe insertion	An inframe nonsynonymous variant that inserts bases into the coding sequence
Inframe deletion	An inframe nonsynonymous variant that deletes bases from the coding sequence
Missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved

2.4. RNA-Seq

2.4.1. Pre-processing of RNA-Seq data

TCGA provides RNA-Seq data for cancer cells and some normal cells, and the GTEx database provides RNA-Seq data for normal tissues. The RNA-Seq samples from these two databases differ in terms of the type of sample, cancer, and normal but are also provided by different sources. Therefore, combining the two datasets not only considers the difference between cancer and normal tissues but also considers that they come from different sources. The goal of the Toil recomputing project is to unify the RNA-Seq pipeline between databases(Vivian et al., 2017). However, the Toil RNA-Seq Recompute Compendium using a uniform pipeline cannot solve the batch effect caused by two different databases. Therefore, RNA-seq data processed considering the batch effect of TCGA and GTEx were used in this study (Figure 4) (Leek, 2014; Q. Wang et al., 2018).

Using the example of bladder cancer, the distribution of the two databases from the unified RNA-Seq pipeline samples provided by the Toil recompute project and batch-effect-corrected samples were visualized by dimensional reduction using the uniform manifold approximation and projection for dimension reduction (UMAP) algorithm (Figure 5) (McInnes et al., 2018). When only ICGC data were used, Data provided by the Toil recompute project were used when only ICGC data were used. Table 4 describes the number of cancer and normal samples according to the database used and cancer type.

Table 4. The consequences of the variants and INDELs in order of severity estimated by Ensembl.

Cancer type	Number of cancer samples		Number of normal samples	
	TCGA	ICGC	TCGA	GTEx
BLCA	398	294	19	9
BRCA	736	970	83	179
CESC	291	241	3	10
COAD	209	390	18	308
GBM	138	155	0	1141
HNSC	489	461	43	55
KIRC	327	345	67	28
KIRP	276	216	31	28
LGG	489	431	0	1141
LIHC	352	281	50	110
LUAD	490	475	58	288
LUSC	478	411	46	288
OV	204	186	0	88
PAAD	175	130	4	167
PRAD	479	370	52	100
SKCM	462	426	1	556
STAD	409	414	33	174
THCA	487	480	58	279
UCEC	150	494	6	78

2.4.2. Normalization

To remove genes with low expression from the RNA-Seq data, we transformed the read counts to counts per million (CPM) on a log scale. The function ‘cpm’ is provided in the R package edgeR to compare the relative mRNA expression levels among the different samples by adjusting the library size of whole read counts (Robinson et al., 2010). Genes with a CPM value > 1 in less than half of all samples were considered as lowly expressed genes and filtered out. This process was performed independently for each cancer type.

Then, trimmed mean of M-value (TMM) normalization was used to simultaneously adjust the library size and composition of the RNA population to estimate the appropriate relative regularization factor unaffected by outliers(Robinson & Oshlack, 2010). Using the voom function provided in R package 'limma', RNA-Seq mRNA expression data were transformed into logCPM data (Law et al., 2014; Ritchie et al., 2015). TCGA and GTEx data were normalized, and ICGC data were processed independently. Of the normalized mRNA expressions, only genes belonging to the cancer gene consensus (CGC) were used for analysis (Bamford et al., 2004; Tate et al., 2019).

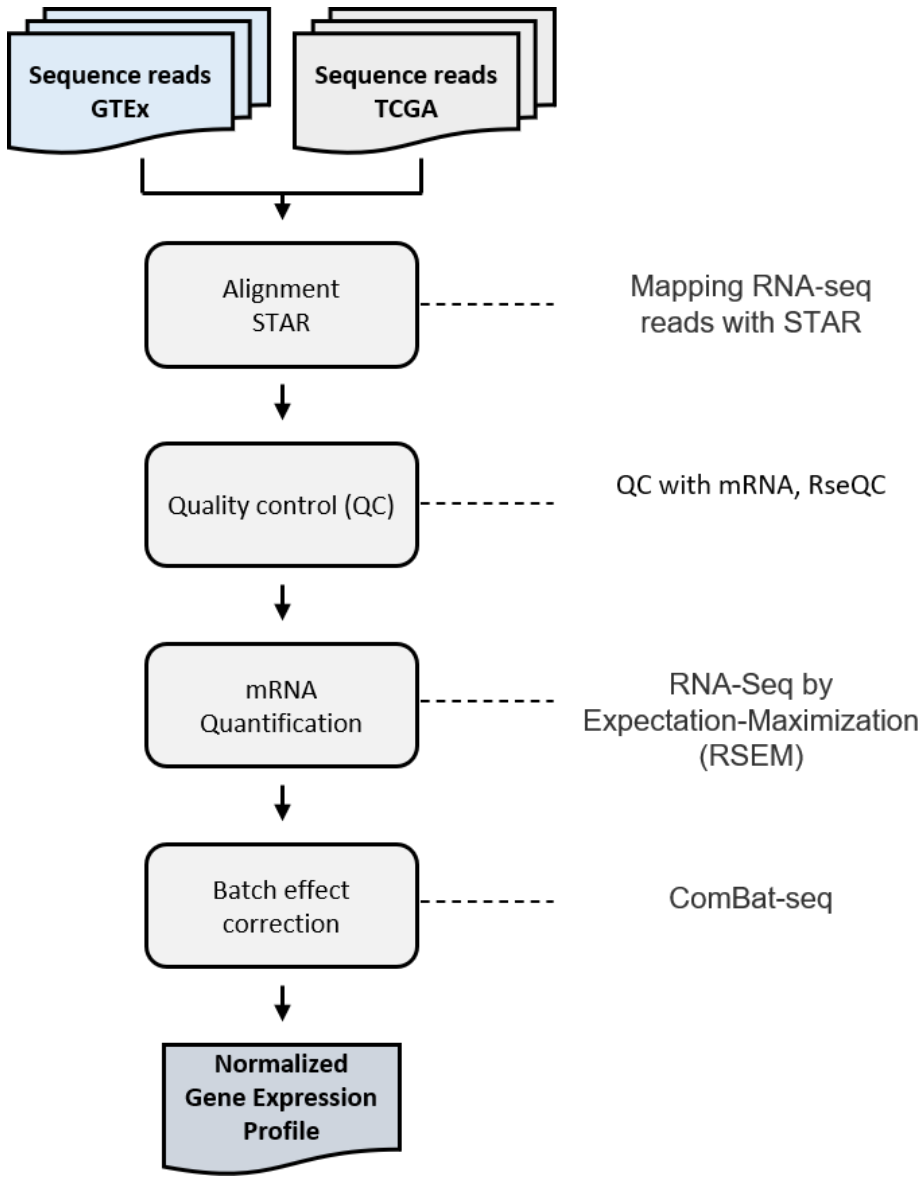

Figure 4. Integration of cancer and normal RNA sequencing data from the TCGA/GTEx database.

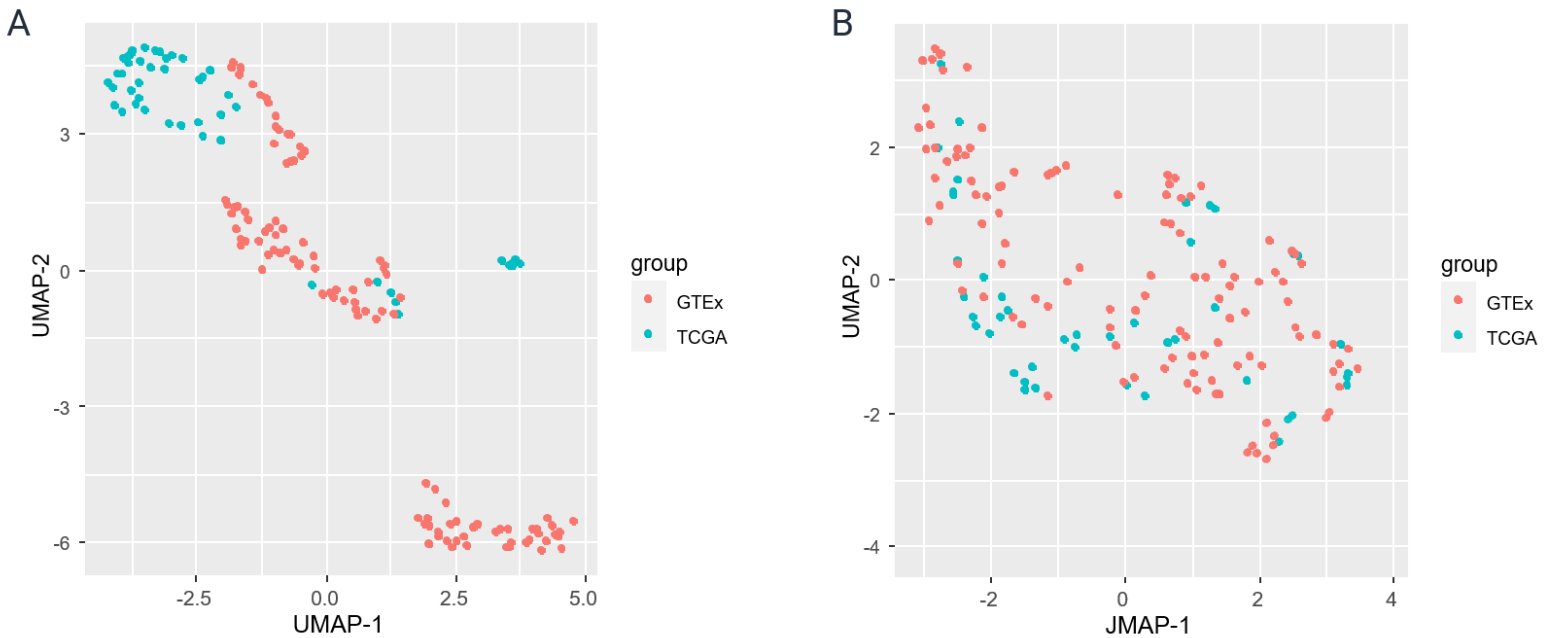

Figure 5. The distribution of bladder mRNA expression levels visualized by UMAP. (A) Unified RNA-Seq pipeline samples provided by the Toil recompute project. (B) Batch-effect corrected samples.

2.5. Cell line data

Cell line samples from the COSMIC database included DNA sequences from WXS and RNA-Seq mRNA expression and copy number analysis data quantified by the Affymetrix SNP6.0 array. Cancer variants through expectation maximization (CaVEMan) and Pindel were used on the tumor and normal paired samples, and copy number variations were quantified using the predicting integral copy numbers in cancer (PICNIC) algorithm (Greenman et al., 2010; Stephens et al., 2012; Ye et al., 2009). The frequency of variants > 0.0014 in the 1000 Genomes database (http://browser.1000genomes.org/) and frequency of variants > 0.00025 in ESP6500 (https://evs.gs.washington.edu/) were removed (Siva, 2008). Variants with minor allele frequencies were removed from the database. Variants found in in-house COSMIC normal samples were excluded. Variants and INDELs not present in the cDNA region were excluded from the study.

2.6. Drug sensitivity

The GDSC database provides drug sensitivity data in the form of half-maximal inhibitory concentration (IC_{50}) by screening drugs on cancer cell line samples (Yang et al., 2012). DNA sequence and mRNA expression of the cell line samples were obtained from the COSMIC database, which allowed the analysis of the effects of genomic alterations on drug sensitivity (Bamford et al., 2004; Tate et al., 2019). The GDSC database contains DNA sequences and mRNA expression data for 982 cell lines, with IC_{50} values of 449 drugs for these cell lines.

2.7. Synthetic Dosage Cancer Survival (SDCS)

2.7.1. SDCS analysis

An overview of the concept of SDCS analysis is presented in Figure 6. The SDCS pair is a combination of two genetic modifications, one gene disruption and one overexpression, significantly affecting cancer patient's survival. Although disruption of each gene or overexpression of one gene alone does not affect patient survival alone, if the simultaneous occurrence of disruption of one gene and overexpression of another significantly affects the patient's prognosis, it is the SDCS gene pair. To find the SDCS pair, two different genes were selected individually from the gene disruption matrix and the expression matrix. Then, an analysis was performed to identify candidate SDCS pairs based on the TCGA database for the 19 cancer types. The workflow scheme for discovering SDCS pairs from the TCGA database is illustrated in Figure 6.

A gene combination that improved patient survival was defined as a positive SDCS, and a gene combination that worsened prognosis was defined as a negative SDCS. SDCS was identified and validated using TCGA and ICGC databases. The SDCS pairs found in TCGA database were verified using the ICGC database. Only gene pairs that were reconfirmed to be SDCS pairs during verification were defined as SDCS pairs.

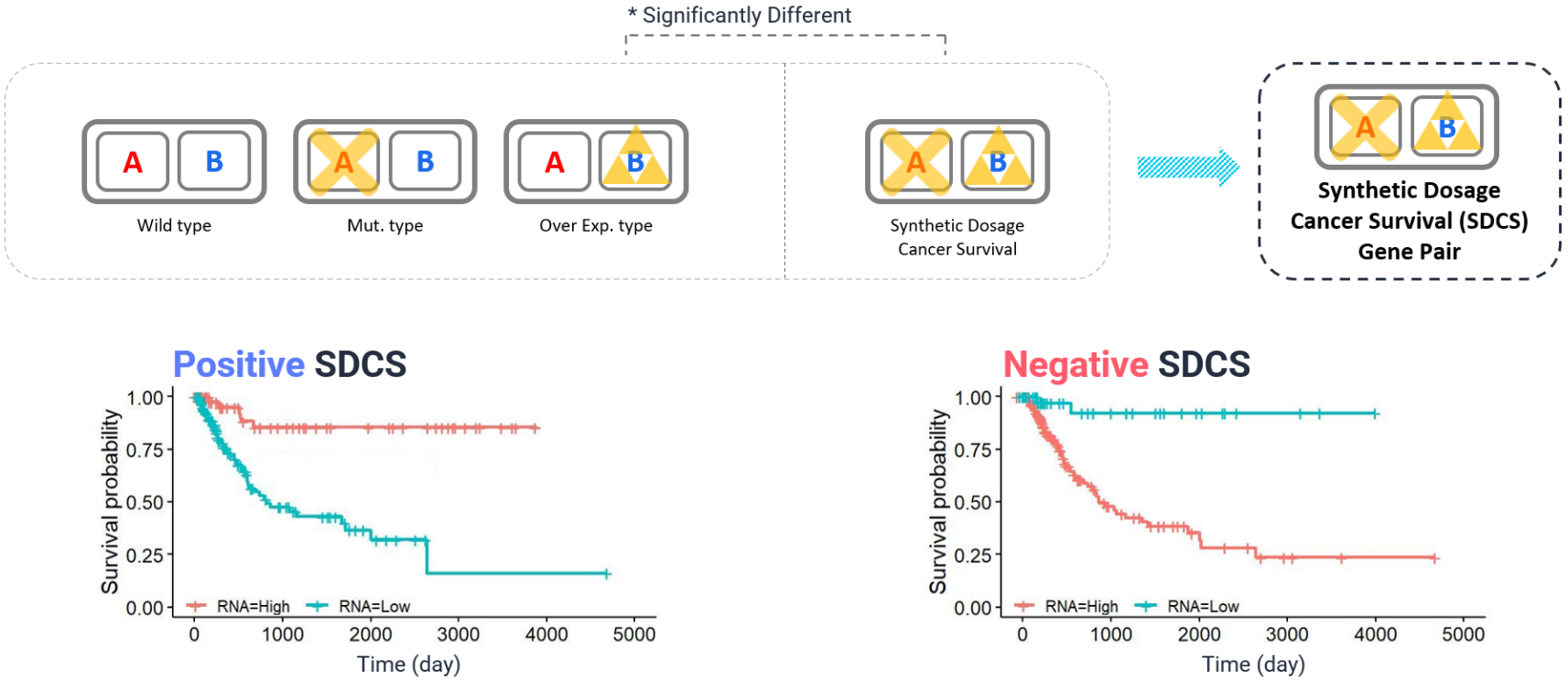

Figure 6. The concept of SDCS analysis and the example of two type of SDCS pairs. A combination of two genetic modifications, one gene disruption and one overexpression, significantly affects prognosis in a patient. When survival improves, this is called a positive SDCS pair, and when survival significantly deteriorates, it is called a negative SDCS.

2.7.2. Statistical analysis

To identify SDCS pairs using the gene disruption matrix, gene expression matrix, and clinical variables, three survival analyses were performed for the two selected genes. First, in the patient group without gene disruption, the expression level of one gene should not be significant for survival when fitting a Cox proportional hazards model with clinical covariates. Expression data from TCGA normal samples and GTEx normal tissues were also added during survival analysis for expression. As there were no clinical variables in the normal sample, random data were generated and used. Among the dependent variables of the normal sample, all events were censored, and for follow-up time, the highest value was arbitrarily assigned to all patients. As for the clinical variables of normal samples used as covariates in survival analysis, clinical variables (shown in Table 1) were randomly generated for each cancer type. Among the clinical variables, age and sex were randomly selected from the available samples. In addition, pathological staging, tumor grade, and focality were assigned to stage 0, grade 0, and no lesions, respectively, which do not exist in cancer patients. MSI variables for the normal sample were all assigned as microsatellite stability, no smoking, and no alcohol intake. Because there were no lesions in the normal sample, the necrosis percentage was assigned as 0% and residual was assigned as non-existent. In the group without gene disruption, the expression levels of these normal samples were not significant for survival when fitted using the Cox proportional hazards model ($p > 0.05$). Second, in the group with gene disruption, the expression level, including that of the normal sample, should be significant when fitted with the Cox proportional hazards model for survival ($p < 0.05$). Thirdly, among samples with gene disruption, the group with

a high expression level through two-means clustering should show a significant difference in survival compared to all samples without gene disruption. At this time, the direction, positive or negative, of the coefficient in the survival analysis of the entire first gene disruption group and the third analysis should be same as that of the SDCS pair. In the Cox proportional-hazards model, if the direction of coefficient of expression was negative, it was defined as a positive SDCS, and if the coefficient was positive, it was defined as a negative SDCS.

The workflow scheme for deriving the gene disruption matrix and generating the gene expression profile from the raw data to infer the SDCS pair is shown in Figure 7.

2.7.3. Positive SDCS and negative SDCS

Positive SDCS refers to patients with a significantly improved prognosis as the expression level increases in the gene disruption group. In contrast, negative SDCS refers to patients with a significantly worse prognosis, as the expression level increases in the gene disruption group. Figure 8 shows the positive and negative SDCS values.

In the case of positive SDCS, gene expression complements gene disruption to improve patient survival; therefore, analysis of the mechanism of gene B may be an important mechanism for drug development. For example, if the disruption gene of a positive SDCS is a tumor suppressor gene, the gene expression can be considered to complement the loss of function of the tumor suppressor. Therefore, if the regulation mechanism of the over-expressed gene is studied, hints on whether it suppresses cancer progression can be obtained. In the case of negative SDCS, gene

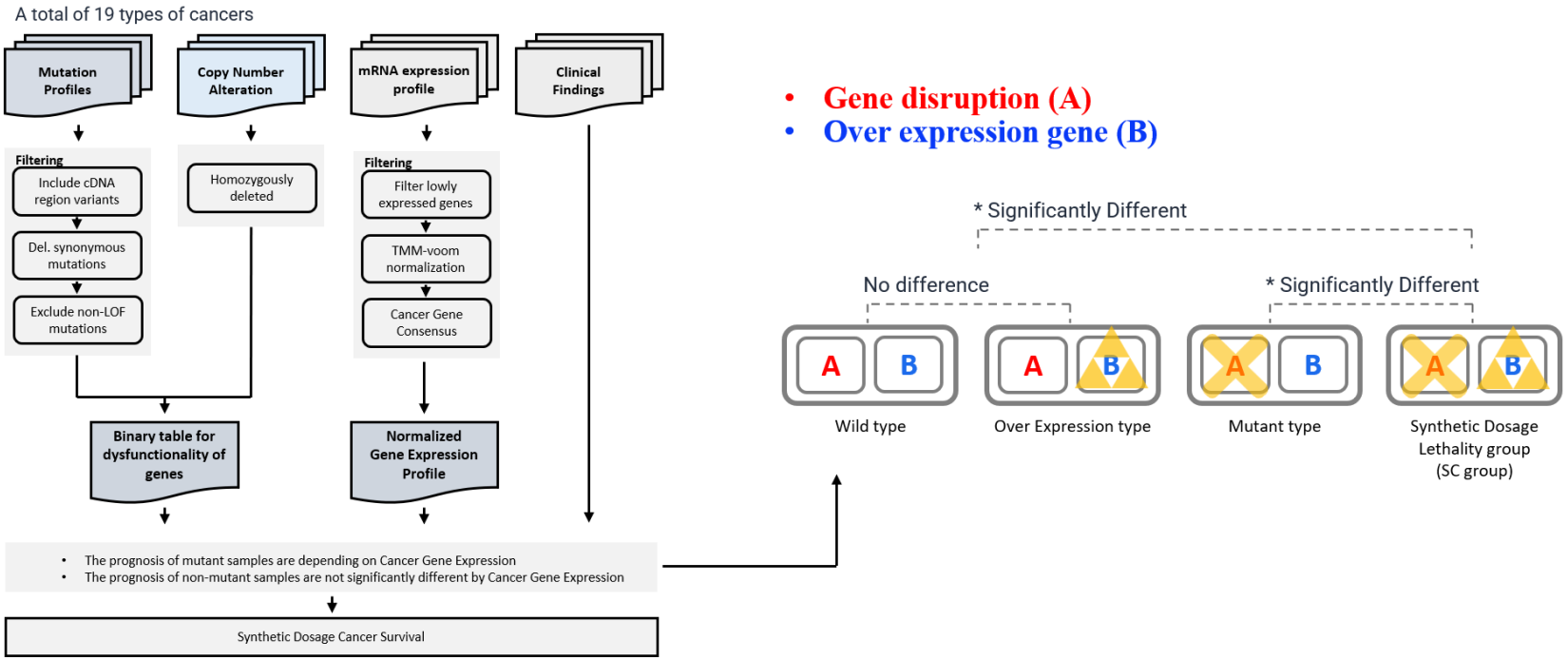

Figure 7. The workflow scheme from the derivation of the gene disruption matrix and the gene expression profile to the inference of the SDCS pair.

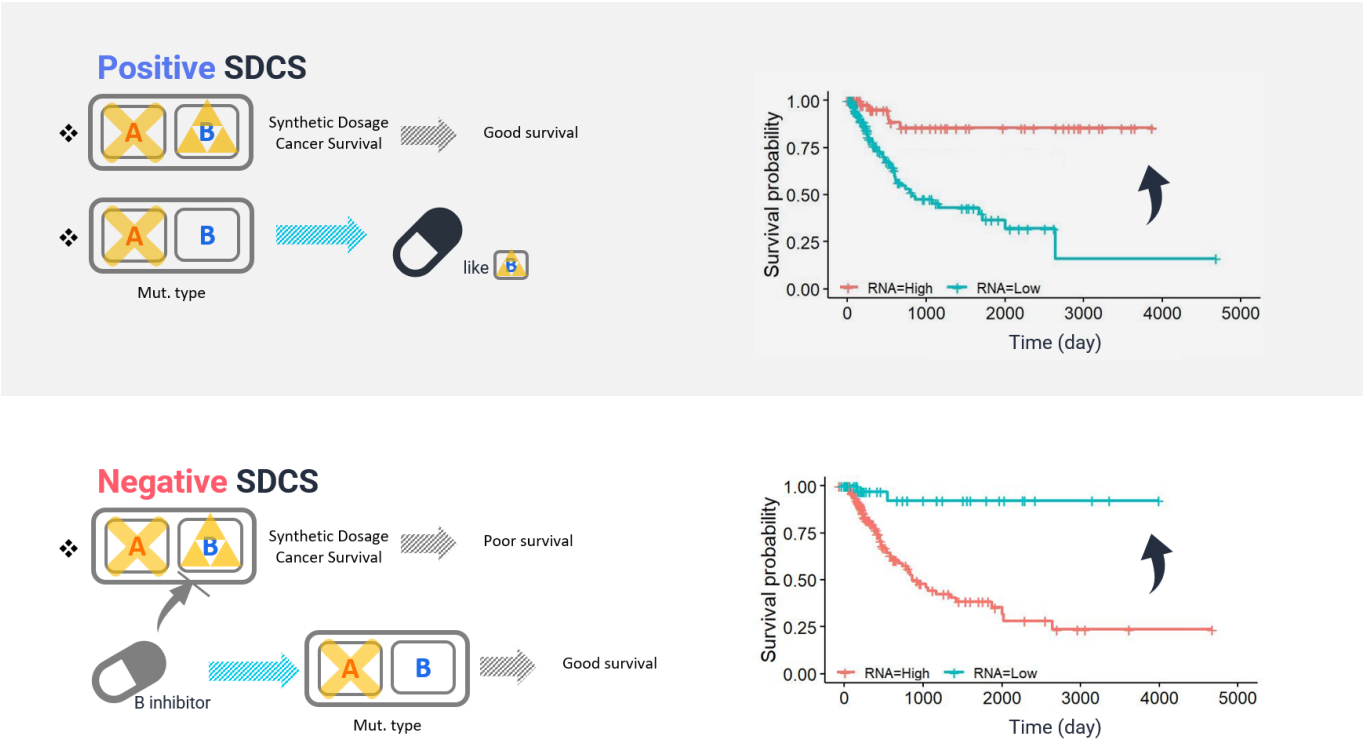

Figure 8. Illustration of application of drug therapy according to positive SDCS and negative SDCS. In the case of positive SDCS, since expression complements gene disruption, analysis of the mechanism of the B gene may be an important mechanism for drug development. In the case of negative SDCS, the expression deteriorates the survival of the patient, so studies on inhibition of this gene are warranted.

expression deteriorates the survival of patients with gene disruption; thus, the over-expressed gene may be related to cancer progression, and it may be meaningful to analyze the mechanism and study drugs to suppress it.

CHAPTER 3. RESULTS

3.1. Inferred SDCS pairs

Table 5 lists the numbers of SDCS pairs derived in this study. The p-value cut-off was used to compare the highest p-value between TCGA and ICGC databases. If one of the p-values from the two databases did not satisfy the p-value cut-off, the gene pair was not included in the SDCS pair. Based on $p < 0.05$, 138 SDCS pairs were derived, of which 71 were positive and 59 were negative SDCS pairs.

To focus on more significant SDCS pairs, 57 SDCS pairs were included in the subsequent analysis with a cut-off of $p < 0.01$. In all SDCS pairs, the number of disrupted and overexpressed genes was 18 and 52, respectively. The disrupted gene most frequently included in the SDCS pair was *TP53*, which was included in a total of 12 SDCS pairs. Among them, two were included in the positive SDCS and 10 were included in the negative SDCS pairs. The *PIK3CA* gene was included nine times in SDCS pairs, *KRAS* gene six times, and *IDH1* gene six times. In contrast, genes that are overexpressed in SDCS pairs are less likely to be included repeatedly than disrupted genes. Only *ERBB3*, *GMPS*, *KRAS*, *MET*, and *ZRSR2* were included in the SDCS pairs twice, and the remaining genes formed only one SDCS pair. Therefore, centered on disrupted genes, a network is formed in which overexpressed

genes form leaf nodes.

Table 6 shows the composition of genes belonging to positive SDCS pairs, the results of survival analysis according to the expression of disrupted genes from TCGA and ICGC databases, and the results of survival analysis according to the expression of genes of patients not included in the disrupted gene group. As indicated by the SDCS analysis, gene expression had a significant prognostic effect on survival only in the group of disrupted genes. In the absence of a disrupted gene, even if the expression level of the expressed gene was changed, survival was not affected at all. In contrast, negative SDCS pairs have a significant effect on survival, such as positive SDCS, but because the coefficient for expression in the Cox proportional-hazards model is positive, higher expression worsens the patients' prognoses (Table 7).

Table 5. The number of SDCS pairs according to the threshold for the maximum value among the p-values from both TCGA and ICGC databases.

Threshold of p.value Minimum TCGA and ICGC	nSDCS (**P**/**R**)
$p < 0.05$	130 (71/59)
$p < 0.01$	57 (35/22)
$p < 0.001$	19 (12/7)
$p < 0.0001$	6 (5/1)
$p < 0.00001$	2 (1/1)

Table 6. The positive SDCS pairs and analysis results in TCGA/ICGC database.

CancerType	Disrupted gene	Expressed gene	TCGA database				ICGC database			
			Disrupted group		Wildtype group		Disrupted group		Wildtype group	
			Coef.	P.value	Coef.	P.value	Coef.	P.value	Coef.	P.value
BLCA	TTN	ZRSR2	-1.213	<0.001	0.012	0.960	-2.215	<0.001	-0.184	0.424
BLCA	TP53	ZRSR2	-1.359	<0.001	0.197	0.349	-1.214	<0.001	-0.311	0.240
BLCA	PIK3CA	CDH1	-0.232	<0.001	-0.063	0.130	-0.331	<0.001	-0.065	0.485
BLCA	PIK3CA	ERBB3	-0.285	<0.001	-0.021	0.677	-0.408	<0.001	-0.069	0.417
LGG	IDH1	PTPRD	-0.401	<0.001	-0.099	0.480	-0.660	<0.001	-0.254	0.247
LUAD	KEAP1	AXIN2	-0.740	<0.001	0.083	0.332	-0.647	<0.001	0.063	0.534
BRCA	PIK3CA	CBLC	-0.777	<0.001	-0.193	0.065	-0.329	<0.001	-0.005	0.955
PAAD	KRAS	RNF43	-0.659	<0.001	-0.225	0.468	-0.805	0.001	0.118	0.738
LUSC	TTN	EIF1AX	-0.555	<0.001	0.085	0.597	-0.846	0.001	-0.333	0.122
STAD	SYNE1	MSI2	-1.284	<0.001	-0.110	0.503	-1.188	0.002	-0.258	0.216
BLCA	KMT2D	IKBKB	-0.549	<0.001	-0.226	0.158	-1.331	0.002	-0.349	0.110

BLCA	TP53	COX6C	-0.435	<0.001	0.187	0.159	-0.644	0.002	0.063	0.810
BLCA	KMT2D	MYD88	-0.533	<0.001	-0.161	0.273	-1.064	0.003	-0.153	0.490
LUAD	KEAP1	WIF1	-0.435	<0.001	0.000	0.991	-0.287	0.003	0.018	0.665
LUSC	SI	LEF1	-1.075	<0.001	-0.059	0.327	-0.716	0.003	-0.071	0.429
LUAD	USH2A	KDR	-0.684	<0.001	0.029	0.692	-0.500	0.003	0.003	0.970
LUSC	ZFHX4	VTI1A	-1.131	<0.001	0.272	0.196	-1.407	0.006	0.001	0.998
LUAD	USH2A	NFIB	-0.441	<0.001	0.018	0.850	-0.503	0.006	-0.132	0.296
COAD	PIK3CA	SETDB1	-9.191	<0.001	-0.043	0.903	-2.471	0.007	0.135	0.786
BLCA	PIK3CA	GATA3	-0.447	<0.001	-0.044	0.228	-0.294	0.008	-0.089	0.076
LUAD	FAM135B	ELN	-0.320	<0.001	0.003	0.953	-0.437	0.008	0.007	0.913
BLCA	PIK3CA	TRIM24	-0.691	<0.001	0.024	0.747	-0.771	0.010	-0.037	0.761

Table 7. The positive SDCS pairs and analysis results in TCGA/ICGC database.

			TCGA database				ICGC database			
			Disrupted group		Wildtype group		Disrupted group		Wildtype group	
CancerType	Disrupted Gene	Expressed gene	Coef.	P.value	Coef.	P.value	Coef.	P.value	Coef.	P.value
LGG	TP53	CDK4	0.443	<0.001	-0.114	0.424	0.954	<0.001	0.189	0.209
LGG	IDH1	MAP3K1	0.721	<0.001	-0.066	0.633	0.876	<0.001	0.350	0.083
LGG	IDH1	CBFB	0.723	<0.001	0.084	0.773	1.203	<0.001	0.368	0.321
LGG	TP53	DDIT3	0.408	<0.001	-0.083	0.609	0.853	<0.001	-0.061	0.784
LUSC	CNTNAP5	N4BP2	1.008	<0.001	-0.025	0.843	2.047	<0.001	0.099	0.547
PAAD	KRAS	SND1	1.122	<0.001	-1.158	0.062	2.515	<0.001	-0.956	0.352
BLCA	PIK3CA	ABL2	1.158	<0.001	0.127	0.234	1.601	<0.001	0.326	0.066
PAAD	TP53	GMPS	1.600	<0.001	0.539	0.287	1.904	<0.001	1.137	0.086
PAAD	KRAS	CNBP	1.981	<0.001	-0.157	0.871	2.409	<0.001	1.344	0.348

LGG	IDH1	ERBB3	0.259	<0.001	0.003	0.972	0.437	<0.001	-0.004	0.976
BRCA	PIK3CA	MTOR	1.499	<0.001	0.013	0.942	2.279	<0.001	-0.248	0.441
BLCA	TP53	SMAD4	0.602	<0.001	0.358	0.057	0.902	<0.001	0.564	0.079
LGG	IDH1	MEN1	0.924	<0.001	0.083	0.865	1.413	0.001	1.015	0.112
LUAD	KEAP1	BIRC3	0.442	<0.001	0.052	0.464	0.414	0.001	0.155	0.072
PAAD	TP53	NPM1	1.365	<0.001	0.241	0.652	2.105	0.001	0.804	0.211
BLCA	TTN	SETBP1	0.422	<0.001	0.043	0.617	0.588	0.001	0.144	0.306
BLCA	TP53	NCOR1	0.512	<0.001	0.047	0.780	0.893	0.002	-0.025	0.941
HNSC	TP53	MAP2K1	0.463	<0.001	0.134	0.603	0.644	0.002	0.174	0.614
PAAD	KRAS	GMPS	1.388	<0.001	0.171	0.767	1.469	0.002	1.321	0.144
BLCA	TP53	SNX29	0.572	<0.001	0.005	0.974	0.518	0.003	0.146	0.553
LUAD	NRXN1	KRAS	1.171	<0.001	0.129	0.310	0.746	0.003	0.242	0.160
LGG	IDH1	BCORL1	1.164	<0.001	0.203	0.289	1.143	0.004	0.143	0.589
PAAD	TP53	MET	0.814	<0.001	0.382	0.054	0.751	0.004	0.635	0.052

LUAD	PTPRD	HNRNPA2B1	1.530	<0.001	0.034	0.858	1.547	0.004	0.106	0.667
BLCA	PIK3CA	WWTR1	0.555	<0.001	0.099	0.143	0.711	0.004	0.119	0.319
PAAD	TP53	HSP90AB1	0.862	<0.001	-0.579	0.360	1.251	0.005	-1.007	0.181
PAAD	KRAS	POT1	1.120	<0.001	-2.003	0.103	1.799	0.005	-1.547	0.319
STAD	SYNE1	TNC	0.459	<0.001	-0.023	0.705	0.394	0.005	0.035	0.640
LUAD	ANK2	ID3	0.594	<0.001	0.006	0.935	0.860	0.005	0.075	0.397
PAAD	KRAS	MET	0.825	<0.001	0.619	0.060	0.709	0.005	0.701	0.093
BLCA	MACF1	PBRM1	1.146	<0.001	-0.013	0.916	1.472	0.006	-0.057	0.799
UCEC	KIF1B	SMC1A	355.544	<0.001	-0.290	0.592	2.276	0.007	0.345	0.231
KIRC	TTN	TPM3	0.993	<0.001	0.406	0.196	2.325	0.007	0.555	0.094
LUAD	ZFHX4	KRAS	0.442	<0.001	0.236	0.131	0.504	0.008	0.203	0.285
LUAD	NRXN1	ETNK1	1.080	<0.001	0.130	0.261	0.936	0.010	-0.052	0.757

3.2. Validation of SDCS pairs

3.2.1. Positive SDCS

Among the positive SDCS pairs, the most significant SDCS pairs were two pairs formed by *TTN* and *TP53* gene disruption with *ZRSR2* expression in bladder cancer. Among patients with bladder cancer, 208 out of TP53 mutant patients were identified in TCGA and 127 out of 294 in the ICGC database. Tenascin-N (TNN) is predicted to be involved in several processes such as the generation of neurons and regulation of osteoblast differentiation. *TP53* encodes a tumor suppressor protein that regulates cell cycle arrest, apoptosis, and DNA repair. *ZRSR2* encodes an essential splicing factor that is predicted to be involved in network interactions during spliceosome assembly. The pair of *TNN* gene disruption and *ZRSR2* expression is an SDCS pair, and the survival of bladder cancer patients from TCGA database based on the status of both genes is shown in Figure 9. Among patients with *TNN* disruption, the expression level of *ZRSR2* was significant when normal samples were included ($p < 0.001$), and patients divided into two means clustering also showed a significant survival difference in cancer samples ($p < 0.001$). However, in patients without TNN disruption, *ZRSR2* expression was not significantly different in terms of survival ($p > 0.05$). The PFS according to the status of this SDCS pair showed the same pattern in patients with bladder cancer in the ICGC database (Figure 10). Similarly, in the two-mean clustering of the *ZRSR2* gene in the TNN-disrupted group, the two groups showed a significant difference in survival ($p < 0.001$), but not in the TNN wild-type group.

In all positive SDCS, 22 SDCS pairs comprising 12 disrupted genes and 21

overexpressed genes were visualized (Figure 11). Blue nodes are disrupted genes, and yellow nodes are overexpressed genes. The colors of the edges indicate the type of cancer. Bladder cancer had the most SDCS pairs with nine SDCS pairs, followed by lung adenocarcinoma with five, and lung squamous cell carcinoma with three. Edge width is the negative log of the *p*-value, which indicates the significance of the survival analysis.

3.2.2. Negative SDCS

In negative SDCS, the combination of disruption and overexpression of the two genes worsened the patient's prognosis, and 35 negative SDCS pairs were found in this study. Among these, the most significant negative SDCS pair was *TP53* disruption and *CDK4* overexpression in patients with low-grade glioma. Among patients with low-grade glioma in the TCGA database, 253 of 499 patients had *TP53* mutations, while 186 of 431 patients in the ICGC database had mutations in the gene. In *TP53* mutants, the higher the expression level of the *CDK4* gene, the worse the prognosis in both TCGA and ICGC databases ($p < 0.001$), and there was no significant difference in survival in patients with the *TP53* wild-type allele.

In all negative SDCS, 35 SDCS pairs comprising 14 disrupted genes and 32 overexpressed genes were visualized (Figure 12). In the network with positive SDCS pairs, blue nodes are disrupted genes, and yellow nodes are overexpressed genes. The colors of the edges indicate the type of cancer. Prostate cancer had the most SDCS pairs at 9, followed by low-grade glioma at 7, bladder cancer at 7, and lung adenocarcinoma at 6. Edge width is a negative log of the *p*-value, which indicates

the significance of the survival analysis.

3.2.3. Biological interactions of SDCS

Among the SDCS pairs, physically interacting proteins in humans, as reported in the BioGRID database, were selected to analyze the relationship between disrupted and overexpressed genes(Stark et al., 2006). There were 33 experimental reports that the disrupted gene and overexpressed gene physically interacted with each other, and there were 12 unique SDCS pairs (Table 8). Among them, two pairs were positive SDCS pairs and 10 were negative SDCS pairs. Among the types of experimental systems are SL and negative genetics, there are three SDCS pairs, all of which are negative SDCS pairs: *TP53* disruption and *CDK4* overexpression in low-grade glioma, and *KRAS* disruption, *GMPS* overexpression, *KRAS* disruption, and *POT1* overexpression in prostate cancer.

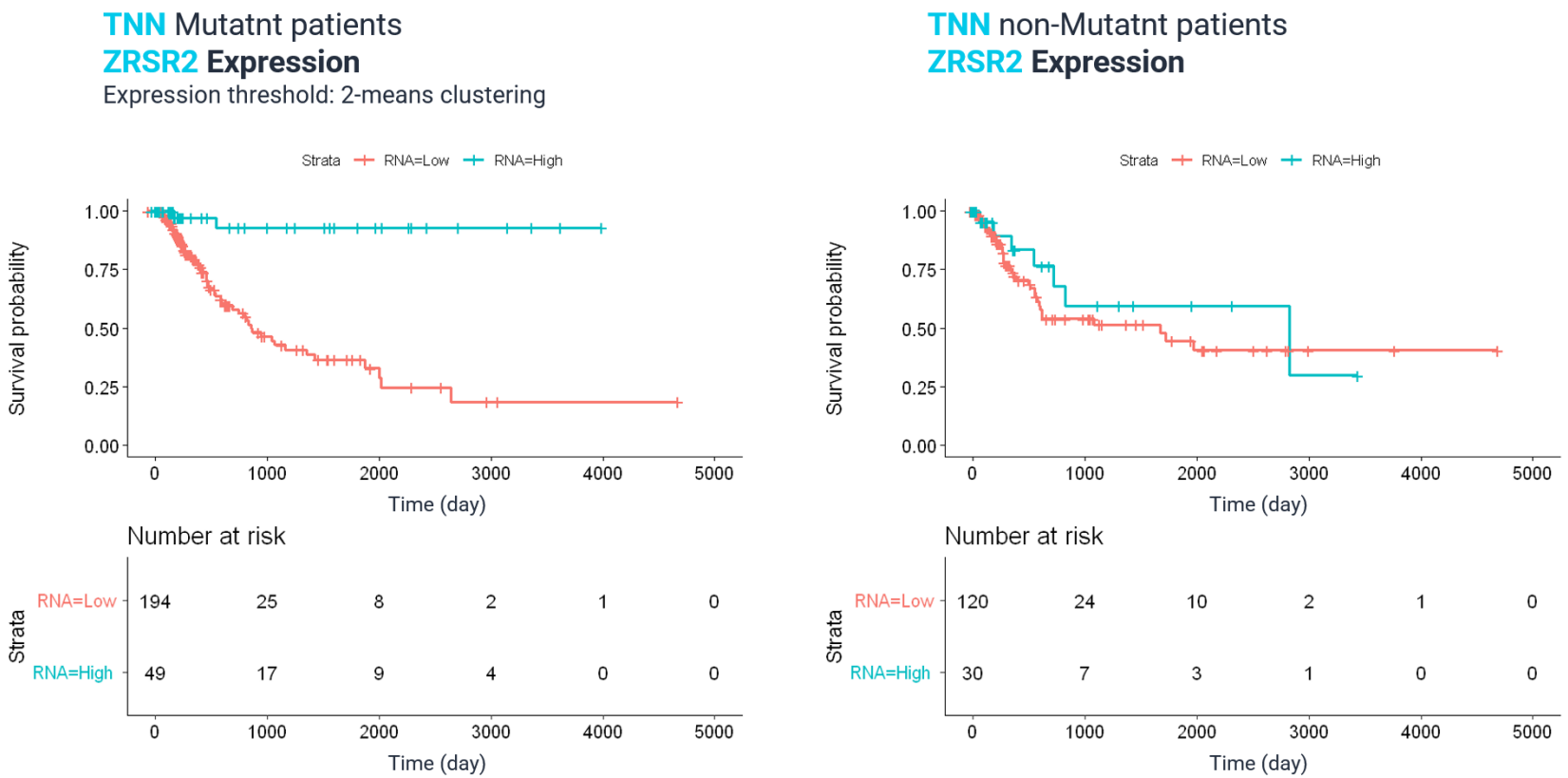

Figure 9. Survival analysis of overexpression of *ZRSR2* in two groups according to *TNN* gene disruption in TCGA bladder cancer patients.

ZRSR2 overexpression in patients with *TNN* disruption is a positive SDCS pair that improves patient prognosis.

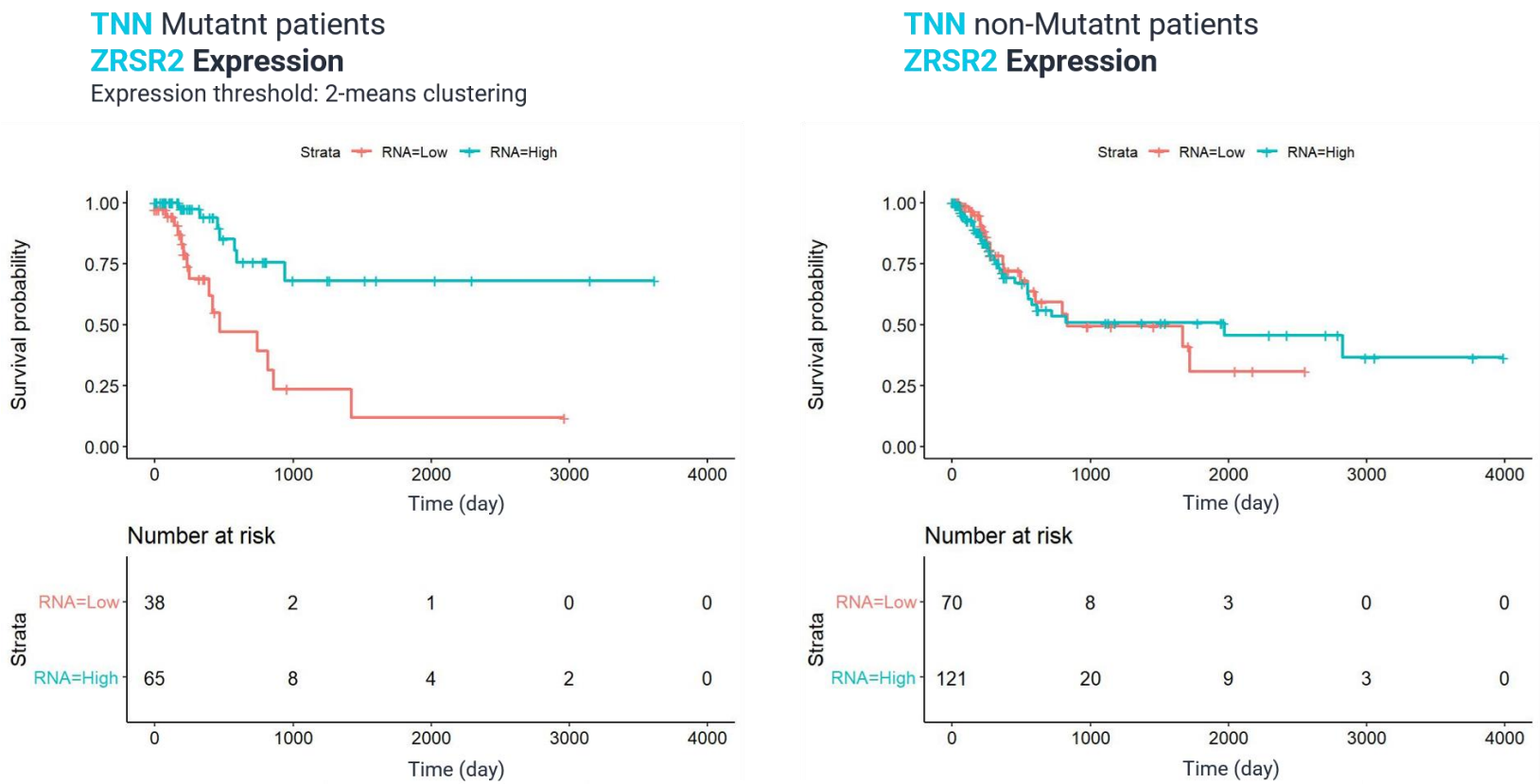

Figure 10. Survival analysis of overexpression of *ZRSR2* gene in two groups according to *TNN* gene disruption in ICGC bladder cancer patients. *ZRSR2* overexpression in patients with *TNN* disruption is a positive SDCS pair that improves patient prognosis.

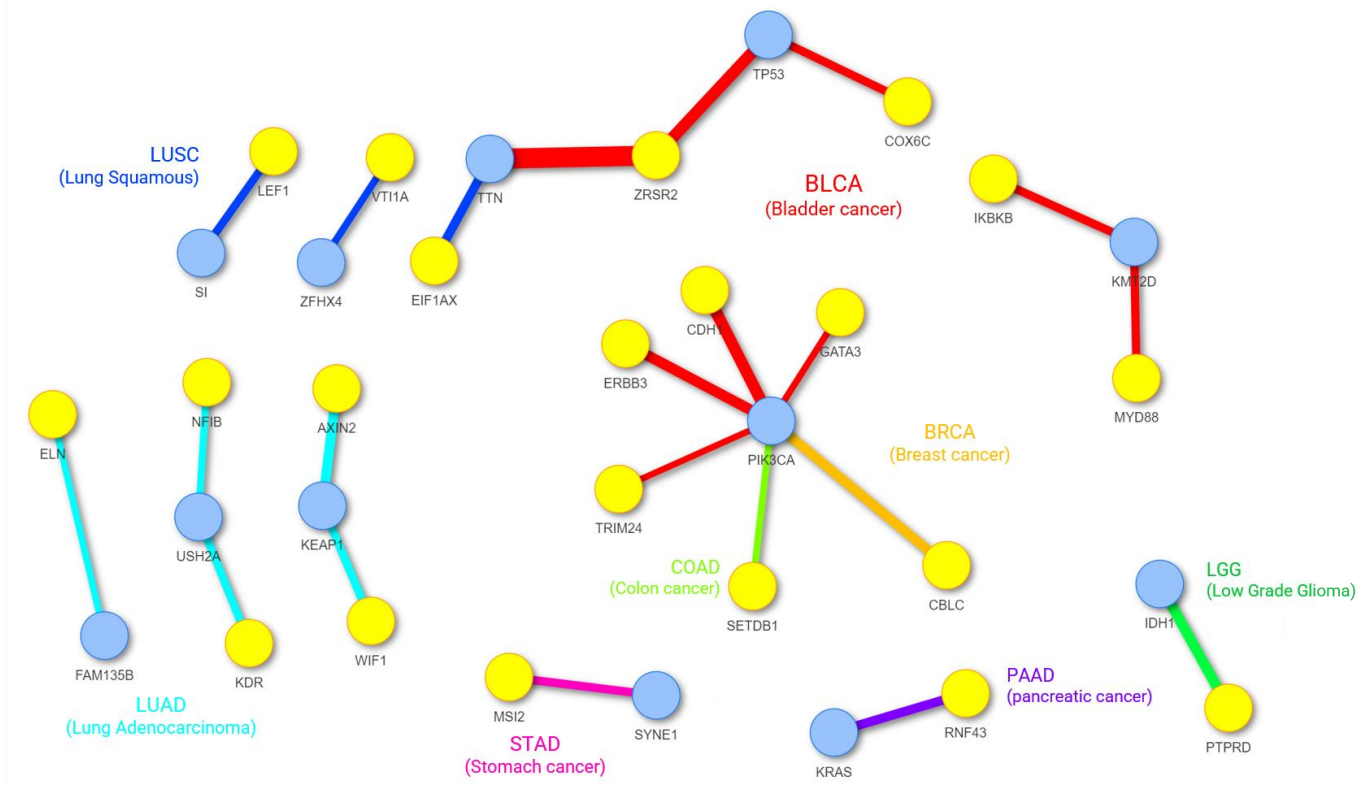

Figure 11. Visualization of the positive SDCS pair as a network. Blue nodes are disrupted genes, and yellow nodes are over-expressed genes. The color of the edge indicates the cancer type. The edge width is the negative log of p. value, which indicates the significance of survival analysis.

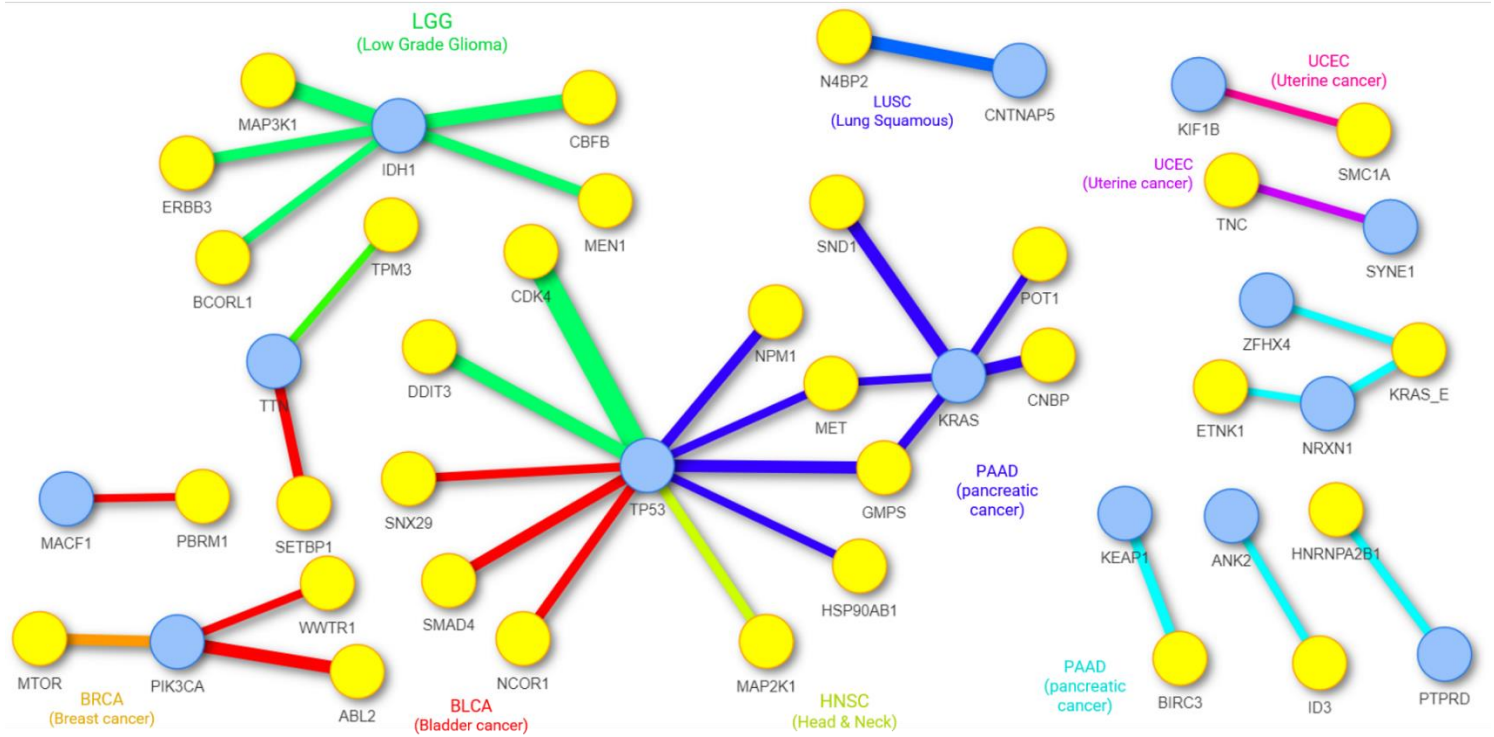

Figure 12. Visualization of the Negative SDCS pair as a network. Blue nodes are disrupted genes, and yellow nodes are over-expression genes. The color of the edge indicates the cancer type. The edge width is the negative log of the p value, which indicates the significance of survival analysis.

Table 8. Physically interacting protein SDCS pairs in humans from the BioGRID database.

Gene_A	Gene_B	Experimental system	Pubmed ID	SDCS
PIK3CA	ERBB3	Affinity Capture-MS	24189400	Prognostic
		Affinity Capture-Western	11546794	Prognostic
TP53	ZRSR2	Affinity Capture-MS	32807901	Prognostic
TP53	CDK4	Negative Genetic	30762338	Negative
GMPS	TP53	Affinity Capture-Western	24462112	Negative
		Affinity Capture-Western	33742136	Negative
KRAS	GMPS	Synthetic Lethality	28700943	Negative
		Negative Genetic	34373451	Negative
KRAS	MET	Proximity Label-MS	34079125	Negative
KRAS	POT1	Synthetic Lethality	28700943	Negative
TP53	NPM1	Affinity Capture-Western	15144954	Negative
		Reconstituted Complex	16376884	Negative
		Reconstituted Complex	12080348	Negative
		Affinity Capture-Western	12080348	Negative
		Affinity Capture-Western	15964625	Negative
		Affinity Capture-Western	15310764	Negative
TP53	CDK4	Affinity Capture-Western	28218424	Negative
		FRET	28205554	Negative
TP53	GMPS	Affinity Capture-Western	24462112	Negative
TP53	HSP90AB1	Affinity Capture-MS	23443559	Negative
		Affinity Capture-MS	32807901	Negative
TP53	MET	FRET	28205554	Negative
TP53	NCOR1	Affinity Capture-Western	19011633	Negative
		Co-localization	24157709	Negative
		Proximity Label-MS	34795231	Negative
TP53	NPM1	Affinity Capture-Western	16740634	Negative
		Affinity Capture-Western	15144954	Negative
		Affinity Capture-Western	12080348	Negative
		Affinity Capture-MS	23443559	Negative
		Affinity Capture-MS	31152661	Negative
		Affinity Capture-MS	32807901	Negative
		Reconstituted Complex	15082766	Negative

Table 9. A list of drugs that act as inhibitor of overexpression gene of negative SDCS pairs.

Drug	Gene	Interaction
OMIPALISIB	MTOR	inhibitor (inhibitory)
OSI-027	MTOR	inhibitor (inhibitory)
PI-103	MTOR	inhibitor (inhibitory)
DACTOLISIB	MTOR	inhibitor (inhibitory)
TEMSIROLIMUS	MTOR	inhibitor (inhibitory)
CABOZANTINIB	MET	inhibitor (inhibitory), antagonist (inhibitory)
AMUVATINIB	MET	inhibitor (inhibitory)
PHA-665752	MET	inhibitor (inhibitory)
CRIZOTINIB	MET	inhibitor (inhibitory)
SELUMETINIB	MAP2K1	allosteric modulator, inhibitor (inhibitory)
TRAMETINIB	MAP2K1	inhibitor (inhibitory), antagonist (inhibitory)
CI-1040	MAP2K1	inhibitor (inhibitory), allosteric modulator
SELUMETINIB	KRAS	inhibitor (inhibitory)
TANESPIMYCIN	HSP90AB1	inhibitor (inhibitory)
PHA-793887	CDK4	inhibitor (inhibitory)
AT-7519	CDK4	inhibitor (inhibitory)
PALBOCICLIB	CDK4	inhibitor (inhibitory)
TOZASERTIB	ABL2	inhibitor (inhibitory)

3.3. Drug sensitivity analysis

To validate the positive and negative effects of the SDCS pairs, genomic data of the COSMIC cell lines and response data of GDSC drugs were utilized. Among the obtained SDCS pairs, in the case of genes inhibiting gene expression, it was assumed that cell-line drug sensitivity would differ according to the status of the disrupted gene and expressed gene. Among the available drugs with IC_{50} values in the GDSC database, drugs that inhibit over-expressed genes in the SDCS pairs were investigated using the DGIdb (Figure 13)(Griffith et al., 2013). Among them, 11 drugs inhibited the expressed genes of positive SDCS pairs and 20 drugs inhibited the expression of negative SDCS pairs (Table 9). The target gene of the drugs was one *KDR* gene, which formed a positive SDCS pair. The *KDR* gene forms a pair with the *USH2A* disruption, which is found in patients with lung adenocarcinoma. In negative SDCS, there were 18 drug-gene-inhibiting relationships with *MTOR*, *MET*, *MAP2K1*, *KRAS*, *HSP90AB1*, *CDK4*, and *ABL2* genes.

Among the negative SDCS pairs, it was assumed that cell lines with high overexpression levels and gene disruption of the SDCS pairs were sensitive to drug reactivity. By matching the tissue of the cell line for each SDCS cancer type, I attempted to identify a drug that showed a significant difference in terms of the expression level of the drug among the cell lines with gene disruption, but there was no significant difference among the cell lines without gene disruption. Among them, the most significant drugs were omipalisib and OSI-027, which inhibit *MTOR*. Because *MTOR* inhibitors have recently attracted attention as a potential treatment for breast cancer, this analysis was performed on breast cancer cell lines(Bhagwat et

al., 2011; Lukey et al., 2019). I compared the sensitivity of omipalisib and OSI-027 to the expression level of *MTOR* in breast cancer cell lines (Figure 14). For omipalisib, the p-value was 0.94 by the Wilcoxon signed-rank results comparing both groups according to *MTOR* expression. As for the sensitivity of OSI-027, the *p*-value obtained by the Wilcoxon signed-rank test comparing the two groups according to the expression of *MTOR* was 0.23. However, mutation of the *PIK3CA* gene constituting the SDCS pair with *MTOR* expression significantly increased the sensitivity of the two drugs according to *MTOR* expression. Omipalisib was only highly sensitive to the *PIK3CA* mutant group in breast cancer cell with *MTOR* overexpression ($p = 0.005$) (Figure 15). OSI-027 was also highly sensitive only to *PIK3CA* mutant samples with *MTOR* overexpression in breast cancer cell lines ($p = 0.015$) (Figure 16). MTOR inhibiting drugs could be used to create the SDCS pair by inhibiting MTOR within PICK3CA disrupted breast cancer cell lines.

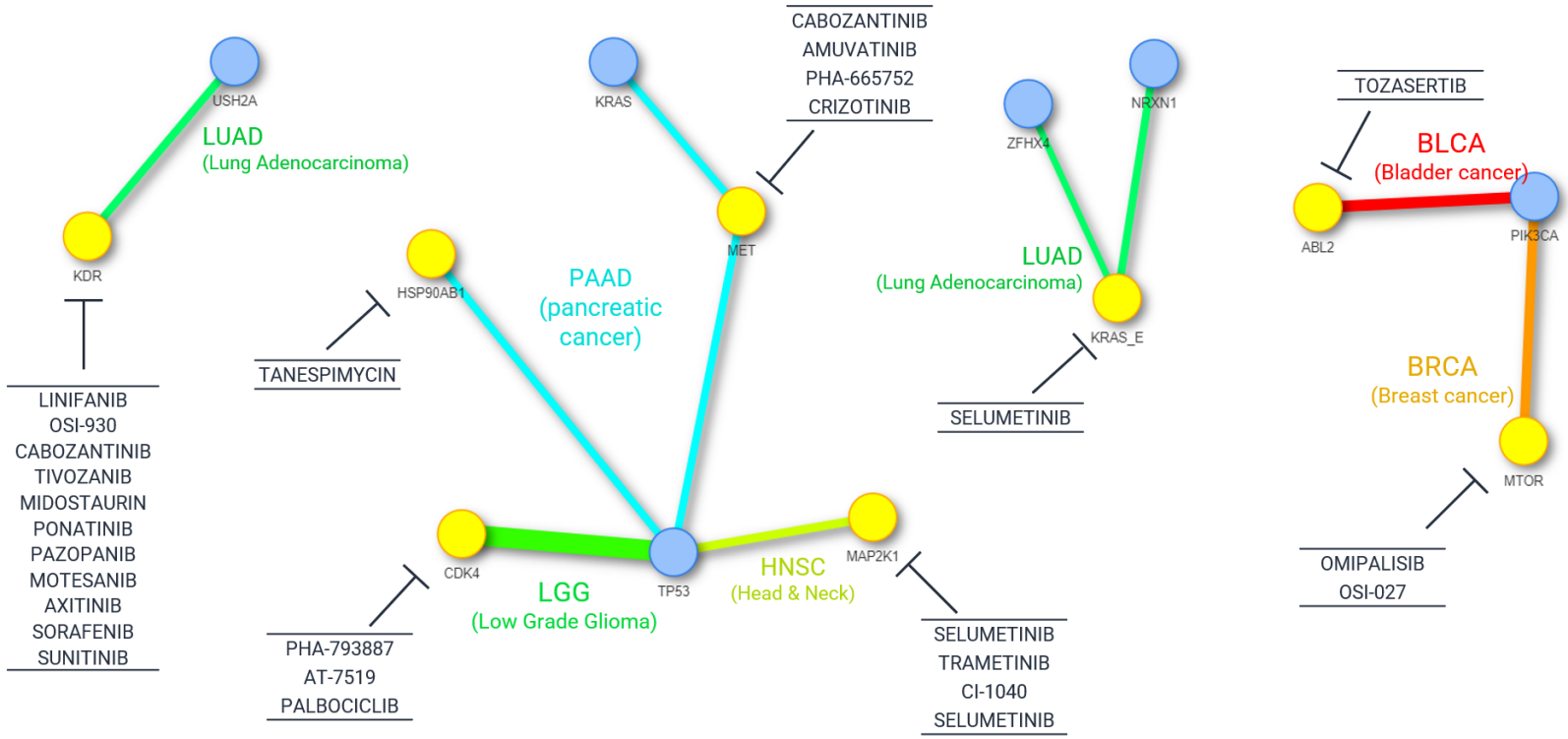

Figure 13. List of drugs capable of inhibiting over-expressed genes among SDCS pairs. The color of the edge indicates the cancer type. The edge width is the negative log of the *p* value, which indicates the significance of survival analysis.

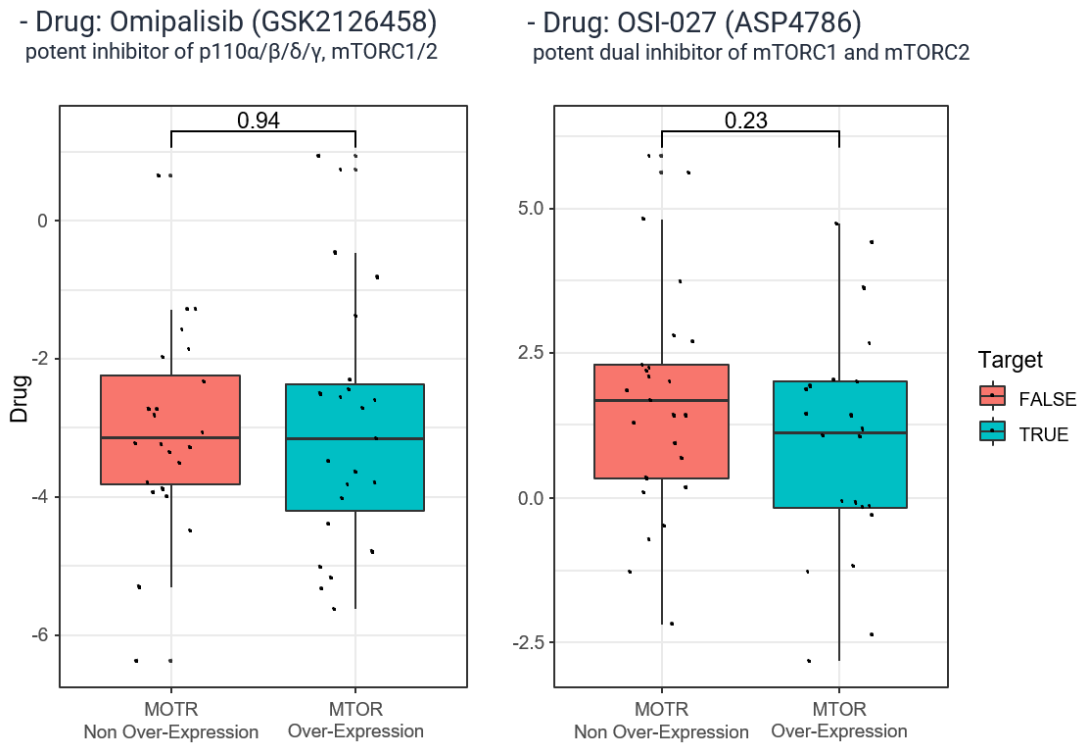

Figure 14. In breast cancer cell-lines, the mRNA expression level of *MTOR* gene was not significantly associated to the reactivity of two *MTOR* inhibitors, Omipalisib and OSI-027.

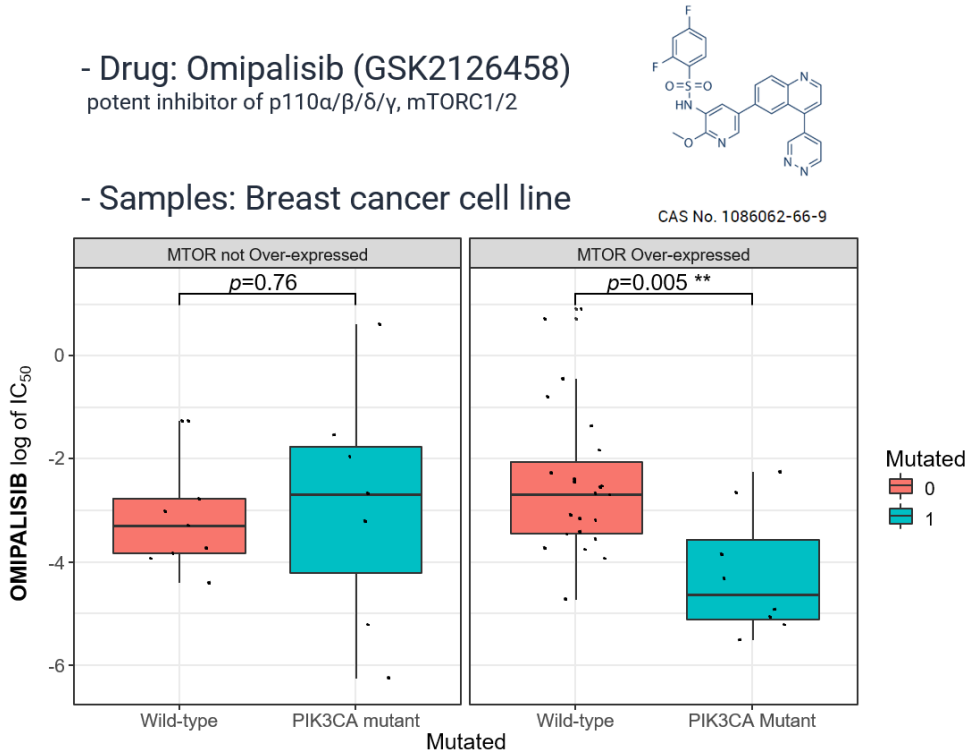

Figure 15. Among samples with over-expression of *MTOR* gene in breast cancer cell-line, *PIK3CA* mutant cell-line is significantly sensitive to the *MTOR* inhibitor Omipalisib.

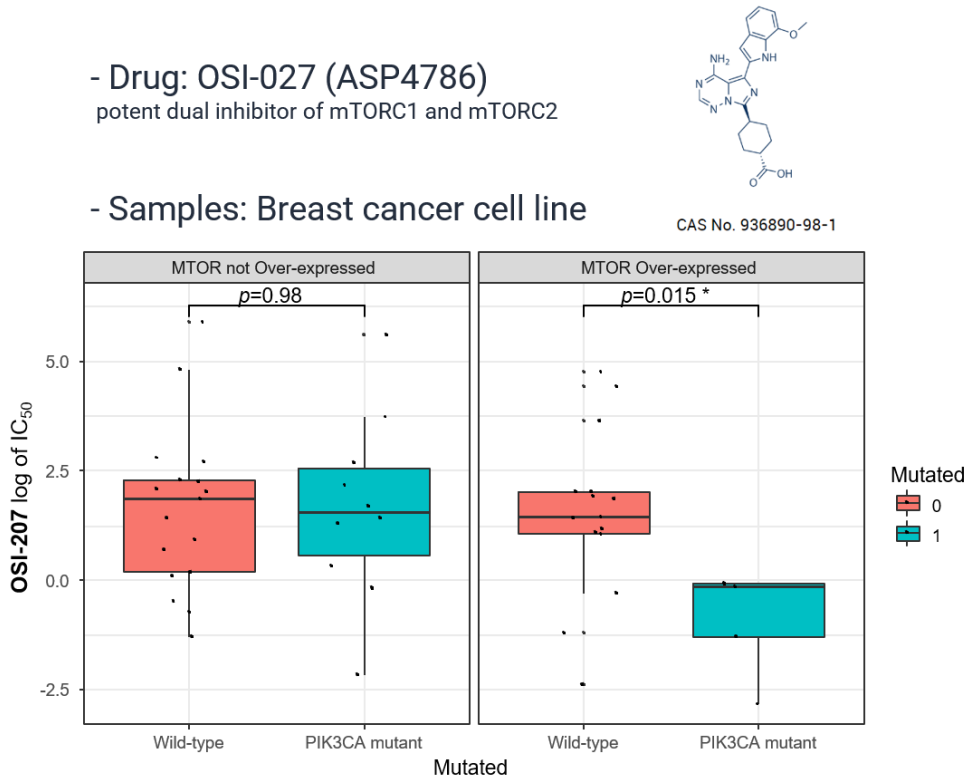

Figure 16. Among samples with over-expression of *MTOR* gene in breast cancer cell-line, *PIK3CA* mutant cell-line is significantly sensitive to the *MTOR* inhibitor OSI-027 drug.

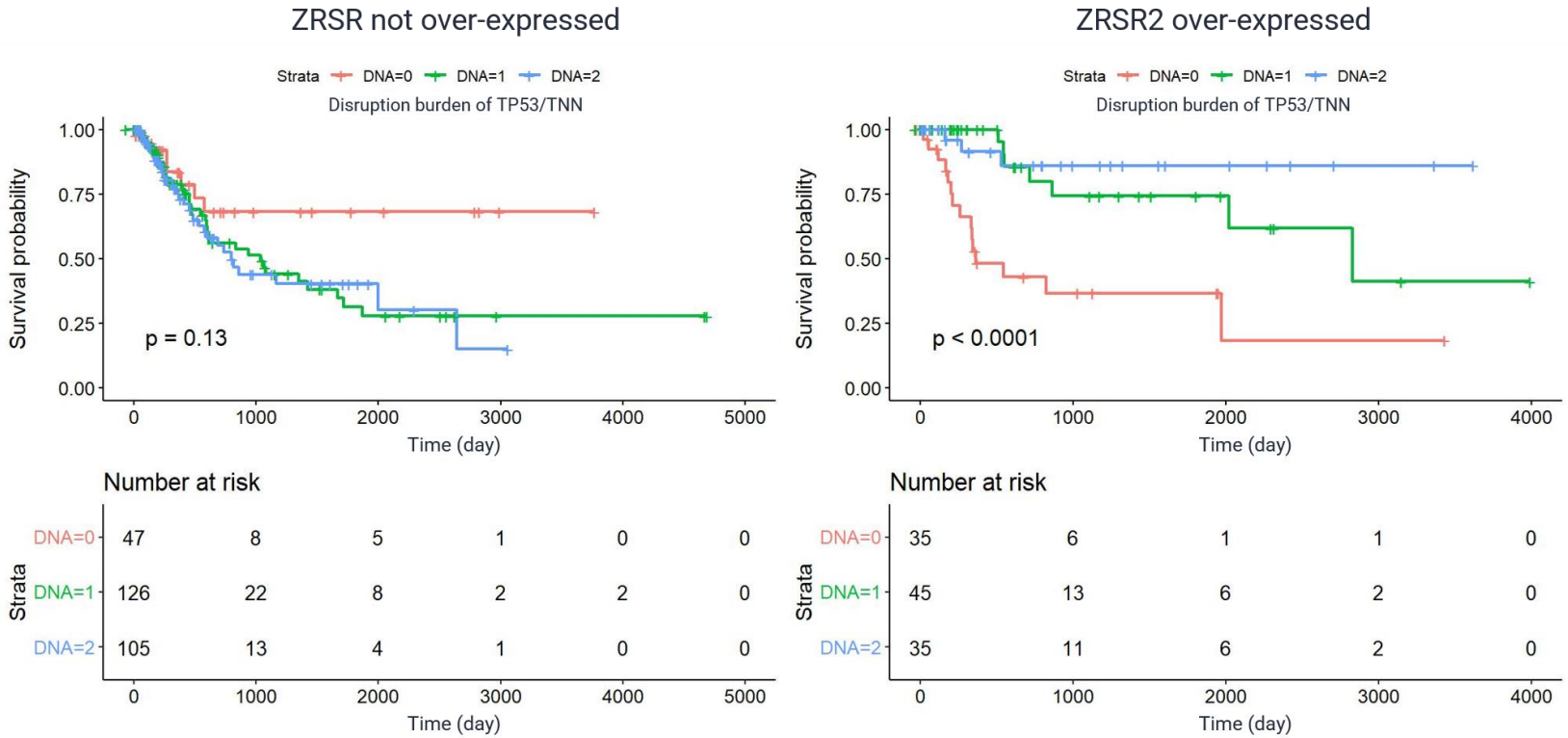

Figure 17. The cumulative effect of the two positive SDCS gene disruption of two gene, *TP53* and *TNN*according to expression of ZRSR2 in TCGA database.

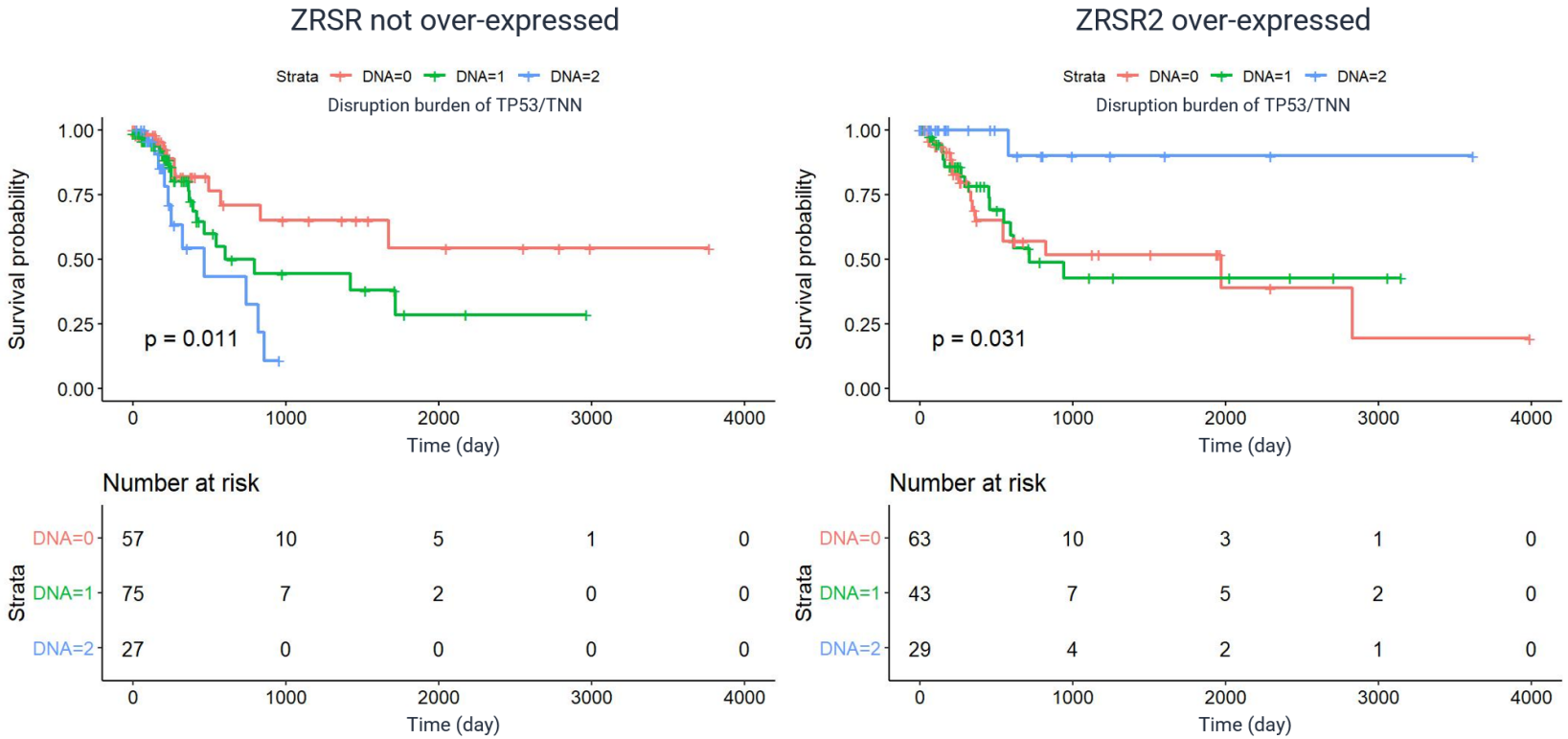

Figure 18. The cumulative effect of the two positive SDCS gene disruption of two gene, *TP53* and *TNN*according to expression of ZRSR2 in TCGA database.

CHAPTER 4. DISCUSSION AND CONCLUSION

1. DISCUSSION

In this study for SDCS analysis, to identify the genetic interaction with patient's survival as the phenotype, 57 SDCS pairs that satisfied data from both TCGA and ICGC multi-omics databases were identified. SDCS disruption genes are all frequently disrupted genes in cancer patients, and expression genes are composed of cancer gene consensus; therefore, they will be useful and important candidates for the development of cancer therapy. Among these genes, interestingly, 14 genes out of 57 pairs of two genes that make up a pair showed physical interactions (Stark et al., 2006). With respect to the number of combinations of all genes used in the analysis, the odds ratio for physically interacting SDCS pairs was 4.295. Therefore, it can be seen that the genes in the SDCS pairs physically influence each other in a large proportion of genome.

An interesting pair in this study was the *PIC3CA* disruption and CDK4 overexpression. Recently, dual inhibitors of *MTOR* and *PIK3CA* have been in the spotlight as therapies for breast cancer. Numerous studies have demonstrated that the dual inhibition of both genes can kill cancer cells. Among the 35 negative SDCS pairs found in this study, one was a combination of *PIK3CA* disruption and *MTOR* overexpression. We found that among the *PIK3CA* disrupted cell lines, *MTOR* overexpressing cells were very sensitive to two MTOR inhibitors, omipalisib and OSI-027 (Bhagwat et al., 2011; Lukey et al., 2019). Furthermore, we found that the *MTOR* over-expressing breast cancer cell lines overexpressing MTOR with *PIK3CA* disruption were highly sensitive to two *MTOR* inhibitors, omipalisib and OSI-027.

Therefore, dual inhibitors of *PIK3CA* and *MTOR* re-verified that there was a negative genetic interaction in the SDCS pairs found in our study. In addition, among patients with *TP53* disruption in low-grade glioma, CDK4 overexpression was found to lead to very poor survival. The CDK4/6 inhibitor PD0332991 is a therapeutic agent that has attracted attention for the treatment of glioblastoma, suggesting that the inhibition of *CDK4* overexpression, which was found to be an overexpressed gene in an SDCS pair in our study, could be a new therapeutic technique (Barton et al., 2013; Cen et al., 2012; Liu et al., 2018). This means that the remaining SDCS pairs are worth testing as new candidate drug targets. Based on the gene-drug relationship provided by DGIdb, 211 drugs inhibited the 52 overexpressed genes. Further studies targeting the overexpression of genes would be a suitable line of research.

Two disruption genes, *TP53* and *TNN*, which form pairs with *ZRSR2* overexpression, forming positive SDCS pairs, are notable. In patients overexpressing *ZRSR2*, the higher the disruption burden of *TP53* and *TNN*, the better the survival using TCGA database (Figure 17). However, in patients that did not overexpress *ZRSR2*, disruption of *TNN* and *TP53* showed no relationship with survival. The SDCS pairs that we found were considered to have a cumulative effect on survival, which was found in the same cancer type. Furthermore, the survival effect could be demonstrated by the validity in the ICGC database (Figure 18). This suggests that, in complex biological networks, targeting multiple genes may play a positive role in patient survival.

A representative contribution of this study is the discovery of new biomarkers that cannot be identified as single biomarkers by combining survival and genetic interactions. In particular, 17 drugs inhibited the overexpression of genes in negative

SDCS pairs, which worsened the patient's prognosis. This can help in the development of personalized anticancer drugs based on the genomic status of patients. Currently, among drugs developed based on genetic interactions, *PARP* inhibitors targeting *BRCA* mutant patients are the only ones used clinically. Therefore, the development of drugs based on gene disruptions that occur frequently in cancer patients is very important. Therefore, future research on the overexpression of genes in SDCS pairs will aid in the development of anticancer drugs based on genetic interactions.

However, this study had some limitations. First, although the results were validated using TCGA and ICGC databases, drug sensitivity could not be investigated for all SDCS pairs because of the lack of samples for each cancer type using cell-line data. Nevertheless, it was confirmed that the PIK3CA/MTOR SDCS pair acted as a dual inhibitor in breast cancer cell-lines. Among SDCS pairs, drug experiments based on interesting target SDCS over-expressed genes must be performed. In addition, since the mortality and censored rates are different for each cancer type, there is a problem in terms of selecting the same statistical threshold for all cancer types. Even a single SDCS pair cannot be found in some cancer types. However, it is expected that many meaningful pairs can be identified by adjusting this threshold.

2. CONCLUSION

Using SDCS analysis, which identified genetic interactions using survival as a phenotype, we identified novel genetic interactions that could not be derived from previous computational methods. Centering on frequently mutated cancer-related genes, overexpression of major cancer-related genes included in the cancer gene consensus formed a bipartite network. The SDCS pairs in this network were independent prognostic markers and were verified using both TCGA and ICGC databases. These gene pairs were also significantly enriched in physical interaction databases. Moreover, the PIK3CA/MTOR gene pair is a negative SDCS pair that has recently attracted attention as a potential target for dual inhibitors. Therefore, the SDCS pairs derived in our study will help in the development of anticancer drugs and personalized medicine.

BIBLIOGRAPHY

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *76*(1), 7–20.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., & others. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, *91*(2), 355–358.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. v, Sonkin, D., & others. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607.

Barton, K. L., Misuraca, K., Cordero, F., Dobrikova, E., Min, H. D., Gromeier, M., Kirsch, D. G., & Becher, O. J. (2013). PD-0332991, a CDK4/6 inhibitor, significantly prolongs survival in a genetically engineered mouse model of brainstem glioma. *PloS One*, *8*(10), e77639.

Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B., & Boone, C. (2013). Genetic interaction networks: toward an understanding of heritability. *Annual Review of Genomics and Human Genetics*, *14*, 111–133.

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). Calling somatic SNVs and indels with Mutect2. *BioRxiv*, 861054.

Bhagwat, S. v, Gokhale, P. C., Crew, A. P., Cooke, A., Yao, Y., Mantis, C., Kahler, J., Workman, J., Bittner, M., Dudkin, L., & others. (2011). Preclinical Characterization of

OSI-027, a Potent and Selective Inhibitor of mTORC1 and mTORC2: Distinct from RapamycinPreclinical Profile of Dual mTORC1/2 Inhibitor OSI-027. *Molecular Cancer Therapeutics*, *10*(8), 1394–1406.

Brough, R., Frankum, J. R., Costa-Cabral, S., Lord, C. J., & Ashworth, A. (2011). Searching for synthetic lethality in cancer. *Current Opinion in Genetics & Development*, *21*(1), 34–41.

Cai, R., Chen, X., Fang, Y., Wu, M., & Hao, Y. (2020). Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. *Bioinformatics*, *36*(16), 4458–4465.

Castells-Roca, L., Tejero, E., Rodr\'\iguez-Santiago, B., & Surrallés, J. (2021). CRISPR screens in synthetic lethality and combinatorial therapies for cancer. *Cancers*, *13*(7), 1591.

Cen, L., Carlson, B. L., Schroeder, M. A., Ostrem, J. L., Kitange, G. J., Mladek, A. C., Fink, S. R., Decker, P. A., Wu, W., Kim, J.-S., & others. (2012). p16-Cdk4-Rb axis controls sensitivity to a cyclin-dependent kinase inhibitor PD0332991 in glioblastoma xenograft cells. *Neuro-Oncology*, *14*(7), 870–881.

Cingolani, P. (2022). Variant Annotation and Functional Prediction: SnpEff. In *Variant Calling* (pp. 289–314). Springer.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, *3*, 35.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster

strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.

Consortium, Gte., Ardlie, K. G., Deluca, D. S., Segrè, A. v, Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., & others. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, *348*(6235), 648–660.

Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B., Boone, C., & others. (2009). Systematic mapping of genetic interaction networks. *Annual Review of Genetics*, *43*(1), 601–625.

Ellrott, K., Bailey, M., Saksena, G., Covington, K., Kandoth, C., Stewart, C., McLellan, M., Sofia, H., Hutter, C., Getz, G., & others. (2018). Multi-center mutation calling in multiple cancers: the MC3 project. *Cancer Research*, *78*(13_Supplement), 926.

Fadista, J., Oskolkov, N., Hansson, O., & Groop, L. (2017). LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*, *33*(4), 471–474.

Goldman, M., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., & others. (2019). The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *Biorxiv*, 326470.

Goldman, M., Craft, B., Zhu, J., & Haussler, D. (2017). The UCSC Xena system for cancer genomics data visualization and interpretation. *Cancer Research*, *77*(13_Supplement), 2584.

Greenman, C. D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., & others. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, *11*(1), 164–175.

Griffith, M., Griffith, O. L., Coffman, A. C., Weible, J. v, McMichael, J. F., Spies, N. C., Koval, J., Das, I., Callaway, M. B., Eldred, J. M., & others. (2013). DGIdb: mining the druggable genome. *Nature Methods*, *10*(12), 1209–1210.

Han, Y., Wang, C., Dong, Q., Chen, T., Yang, F., Liu, Y., Chen, B., Zhao, Z., Qi, L., Zhao, W., & others. (2019). Genetic interaction-based biomarkers identification for drug resistance and sensitivity in cancer cells. *Molecular Therapy-Nucleic Acids*, *17*, 688–700.

Hodgson, D. R., Dougherty, B. A., Lai, Z., Fielding, A., Grinsted, L., Spencer, S., O'connor, M. J., Ho, T. W., Robertson, J. D., Lanchbury, J. S., & others. (2018). Candidate biomarkers of PARP inhibitor sensitivity in ovarian cancer beyond the BRCA genes. *British Journal of Cancer*, *119*(11), 1401–1409.

Jerby-Arnon, L., Pfetzer, N., Waldman, Y. Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P. A., & others. (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, *158*(5), 1199–1209.

Kaelin, W. G. (2005). The concept of synthetic lethality in the context of anticancer therapy. *Nature Reviews Cancer*, *5*(9), 689–698.

Katti, A., Diaz, B. J., Caragine, C. M., Sanjana, N. E., & Dow, L. E. (2022). CRISPR in cancer biology and therapy. *Nature Reviews Cancer*, *22*(5), 259–279.

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: somatic mutation and copy

number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576.

Kroll, E. S., Hyland, K. M., Hieter, P., & Li, J. J. (1996). Establishing genetic interactions by a synthetic dosage lethality phenotype. *Genetics*, *143*(1), 95–102.

Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A., Mattiazzi Usaj, M., van Leeuwen, J., & others. (2018). Systematic analysis of complex genetic interactions. *Science*, *360*(6386), eaao1729.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., & others. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., & Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, *28*(3), 311–317.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), 1–17.

Lee, J. H., Lee, K. H., & Kim, J. H. (2021). In Silico Inference of Synthetic Cytotoxic Interactions from Paclitaxel Responses. *International Journal of Molecular Sciences*, *22*(3), 1097.

Leek, J. T. (2014). Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, *42*(21), e161–e161.

Li, J., Lu, L., Zhang, Y.-H., Liu, M., Chen, L., Huang, T., & Cai, Y.-D. (2019). Identification of synthetic lethality based on a functional network by using machine learning

algorithms. *Journal of Cellular Biochemistry*, *120*(1), 405–416.

Liu, M., Liu, H., & Chen, J. (2018). Mechanisms of the CDK4/6 inhibitor palbociclib (PD 0332991) and its future application in cancer treatment. *Oncology Reports*, *39*(3), 901–911.

Lord, C. J., & Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. *Science*, *355*(6330), 1152–1158.

Lukey, P. T., Harrison, S. A., Yang, S., Man, Y., Holman, B. F., Rashidnasab, A., Azzopardi, G., Grayer, M., Simpson, J. K., Bareille, P., & others. (2019). A randomised, placebo-controlled study of omipalisib (PI3K/mTOR) in idiopathic pulmonary fibrosis. *European Respiratory Journal*, *53*(3).

Luo, J., Emanuele, M. J., Li, D., Creighton, C. J., Schlabach, M. R., Westbrook, T. F., Wong, K.-K., & Elledge, S. J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell*, *137*(5), 835–848.

Madhukar, N. S., Elemento, O., & Pandey, G. (2015). Prediction of genetic interactions using machine learning and network properties. *Frontiers in Bioengineering and Biotechnology*, *3*, 172.

Mani, R., st. Onge, R. P., Hartman IV, J. L., Giaever, G., & Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences*, *105*(9), 3461–3466.

Mateo, J., Lord, C. J., Serra, V., Tutt, A., Balmaña, J., Castroviejo-Bermejo, M., Cruz, C., Oaknin, A., Kaye, S. B., & de Bono, J. S. (2019). A decade of clinical development of PARP inhibitors in perspective. *Annals of Oncology*, *30*(9), 1437–1447.

McDonald III, E. R., de Weck, A., Schlabach, M. R., Billy, E., Mavrakis, K. J., Hoffman, G.

R., Belur, D., Castelletti, D., Frias, E., Gampa, K., & others. (2017). Project DRIVE: a compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. *Cell*, *170*(3), 577–592.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint ArXiv:1802.03426*.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & others. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology*, *17*(1), 1–14.

Measday, V., & Hieter, P. (2002). Synthetic dosage lethality. In *Methods in enzymology* (Vol. 350, pp. 316–326). Elsevier.

Megchelenbrink, W., Katzir, R., Lu, X., Ruppin, E., & Notebaart, R. A. (2015). Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival. *Proceedings of the National Academy of Sciences*, *112*(39), 12217–12222.

Molina, J. R., Sun, Y., Protopopova, M., Gera, S., Bandi, M., Bristow, C., McAfoos, T., Morlacchi, P., Ackroyd, J., Agip, A.-N. A., & others. (2018). An inhibitor of oxidative phosphorylation exploits cancer vulnerability. *Nature Medicine*, *24*(7), 1036–1046.

O'Neil, N. J., Bailey, M. L., & Hieter, P. (2017). Synthetic lethality and cancer. *Nature Reviews Genetics*, *18*(10), 613–623.

Ooi, S. L., Pan, X., Peyser, B. D., Ye, P., Meluh, P. B., Yuan, D. S., Irizarry, R. A., Bader, J. S., Spencer, F. A., & Boeke, J. D. (2006). Global synthetic-lethality analysis and yeast functional profiling. *TRENDS in Genetics*, *22*(1), 56–63.

Paul, J. M., Templeton, S. D., Baharani, A., Freywald, A., & Vizeacoumar, F. J. (2014). Building high-resolution synthetic lethal networks: a 'Google map'of the cancer cell. *Trends in Molecular Medicine*, *20*(12), 704–715.

Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), 1–9.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*(W1), W452–W457.

Siva, N. (2008). 1000 Genomes project. *Nature Biotechnology*, *26*(3), 256–257.

Srihari, S., Singla, J., Wong, L., & Ragan, M. A. (2015). Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biology Direct*, *10*(1), 1–18.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, *34*(suppl_1), D535–D539.

Stephens, P. J., Tarpey, P. S., Davies, H., van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., & others. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature*, *486*(7403), 400–404.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., & others. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, *47*(D1), D941–D947.

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, *2015*(1), 68–77.

Tong, A. H. Y., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., & others. (2004). Global mapping of the yeast genetic interaction network. *Science*, *303*(5659), 808–813.

Tutt, A., Robson, M., Garber, J. E., Domchek, S., Audeh, M. W., Weitzel, J. N., Friedlander, M., & Carmichael, J. (2009). Phase II trial of the oral PARP inhibitor olaparib in BRCA-deficient advanced breast cancer. *Journal of Clinical Oncology*, *27*(18_suppl), CRA501–CRA501.

Typas, A., Nichols, R. J., Siegele, D. A., Shales, M., Collins, S. R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B. L., & others. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. *Nature Methods*, *5*(9), 781–787.

van Leeuwen, J., Boone, C., & Andrews, B. J. (2017). Mapping a diversity of genetic interactions in yeast. *Current Opinion in Systems Biology*, *6*, 14–21.

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., & others. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, *35*(4),

314–316.

Vizeacoumar, F. J., Arnold, R., Vizeacoumar, F. S., Chandrashekhar, M., Buzina, A., Young, J. T. F., Kwan, J. H. M., Sayad, A., Mero, P., Lawo, S., & others. (2013). A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. *Molecular Systems Biology*, *9*(1), 696.

Wan, F., Li, S., Tian, T., Lei, Y., Zhao, D., & Zeng, J. (2020). EXP2SL: a machine learning framework for cell-line-specific synthetic lethality prediction. *Frontiers in Pharmacology*, *11*, 112.

Wang, C., Wang, G., Feng, X., Shepherd, P., Zhang, J., Tang, M., Chen, Z., Srivastava, M., McLaughlin, M. E., Navone, N. M., & others. (2019). Genome-wide CRISPR screens reveal synthetic lethality of RNASEH2 deficiency and ATR inhibition. *Oncogene*, *38*(14), 2451–2463.

Wang, J., Zhang, Q., Han, J., Zhao, Y., Zhao, C., Yan, B., Dai, C., Wu, L., Wen, Y., Zhang, Y., & others. (2022). Computational methods, databases and tools for synthetic lethality prediction. *Briefings in Bioinformatics*, *23*(3), bbac106.

Wang, Q., Armenia, J., Zhang, C., Penson, A. v, Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B. E., Iacobuzio-Donahue, C. A., & others. (2018). Unifying cancer and normal RNA sequencing data from different sources. *Scientific Data*, *5*(1), 1–8.

Wang, S., Xu, F., Li, Y., Wang, J., Zhang, K., Liu, Y., Wu, M., & Zheng, J. (2021). KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers. *Bioinformatics*, *37*(Supplement_1), i418–i425.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120.

Yan Tong, A. H., & Boone, C. (2006). Synthetic genetic array analysis in Saccharomyces cerevisiae. In *Yeast Protocol* (pp. 171–191). Springer.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., & others. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, *41*(D1), D955–D961.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, *25*(21), 2865–2871.

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L. D., & Ferretti, V. (2019). The international cancer genome consortium data portal. *Nature Biotechnology*, *37*(4), 367–369.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., & others. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database*, *2011*.

ABSTRACT IN KOREAN

유전자의 상호작용은 두개 이상의 유전자에 동시적으로 문제가 발생했을 때, 세포의 사멸, 성장 등 세포에 표현형을 나타내는 것을 말한다. 많은 항암치료제들이 암세포의 특이적인 DNA 돌연변이, 혹은 RNA의 과발현을 타겟하여, 유전자 상호작용을 인위적으로 발생시켜 정상세포에는 영향을 주지 않고 암세포를 효율적으로 사멸시키는 것을 목표로 한다. 하지만 그럼에도 불구하고 지금까지의 유전자 상호작용을 밝혀내는 연구들은 사멸이라는 세포의 표현형에 초점을 맞춰왔기 때문에, 새로운 유전자 상호작용을 밝혀내기 쉽지 않았다. 왜냐하면, 세포의 상호작용이 존재하는 순간 세포가 사멸하기 때문에 환자에서 관찰을 할 수 없었기 때문이다. 이러한 한계를 극복하기 위해서, 이 연구에서는 유전자 상호작용을 세포단위가 아닌 환자단위에서 분석하였다. 이 연구에서, 나는 정량 합성 암 생존 (Synthetic Dosage Cancer Survival; SDCS) 분석이라는 방법을 제안한다. 이는 한 유전자의 돌연변이와 한 유전자의 과발현의 조합이 세포를 사멸시키는 정량 합성 치사 (Synthetic dosage lethality) 개념을 기반으로 한다. SDCS는 한 유전자의 돌연변이와 한 유전자의 과발현이 환자의 생존에 유의한 변화를 이끌어내는 유전자 상호작용을 말한다. SDCS 조합이 환자의 예후를 증진시킨다면 positive SDCS로 정의하였으며, 만약 환자의 예후를 악화시킨다면 negative SDCS로 정의하였다. SDCS 조합은 The Cancer Genome Atlas (TCGA) 데이터베이스와 International Cancer Genome Consortium (ICGC) 두가지 데이터베이스를 기반으로 검증하였다. Genotype-Tissue Expression (GTEx) 데이터베이스를 활용하여 대조군의 보충적 데이터로 활용하였다. SDCS 조합에 포함되는 과발현 유전자를 타겟하는 약물들의 민감성을 통해 SDCS의 중요성을 검증하기 위하여, Genomics of Drug Sensitivity in Cancer (GDSC) 데이터베이스를

활용하였다. 22개의 positive SDCS 조합과 35개의 negative SDCS 조합을 TCGA와 ICGC 데이터베이스에서 동시적으로 검증하였다. SDCS 조합을 이루는 유전자들 중, 돌연변이에 해당되는 유전자는 18개가 포함되었으며 과발현 유전자는 52개가 포함되었다. *PIK3CA* 유전자의 돌연변이와 *MTOR* 유전자 과발현은 negative SDCS 조합 중 하나로 이 연구에서 검증되었으며, 이는 현재 이중 억제제로 각광받는 유전자 조합으로 유방암의 치료제로 개발되고 있는 유전자 조합이다. GDSC 데이터베이스에서 MTOR의 과발현과 PIK3CA 돌연변이가 동시적으로 있는 세포주는, MTOR를 억제하는 Omipalsib 약물과 OSI-027약물에 유의하게 민감한 것이 확인되었다. 이러한 관찰은 SDCS 조합이 새로운 항암 치료제를 개발하는데 있어서 중요한 후보 타겟 유전자라는 것을 시사한다. 따라서 SDCS 분석은 새로운 항암제를 개발하는 것에 중요한 방법론으로서 도움을 줄 것이다.

주요어: 유전자 상호작용, 정량 합성 치사, 유전자 발현분석, 약물 민감도 분석, 유전자 타겟 발굴, 예후 마커 발굴

학번: 2015-20509