공학 박사 학위 논문

Identification of Novel Biomarkers and Drug Candidates via Web-based Bioinformatic Analysis in Fusion-Positive Cancer

융합 유전자 양성 암에서의 생물정보학 분석 웹 어플리케이션 기반 신규 바이오마커 및 약물후보물질 발굴

2023 년 2 월

서울대학교 대학원

바이오엔지니어링 협동과정 전공

정 재 헌

Identification of Novel Biomarkers and Drug Candidates via Web-based Bioinformatic Analysis in Fusion-Positive Cancer

융합 유전자 양성 암에서의 생물정보학 분석 웹 어플리케이션 기반
신규 바이오마커 및 약물후보물질 발굴

지도교수 허 찬 영
이 논문을 공학박사학위논문으로 제출함

2023 년 2 월

서울대학교 대학원
바이오엔지니어링 협동과정 전공
정 재 헌

정 재 헌의 박사 학위논문을 인준함
2023 년 2 월

위 원 장 강 성 범 (인)
부 위 원 장 허 찬 영 (인)
위 원 이 세 준 (인)
위 원 윤 재 원 (인)
위 원 김 홍 숙 (인)

Abstract

Identification of Novel Biomarkers and Drug Candidates via Web-based Bioinformatic Analysis in Fusion-Positive Cancer

Jae Heon Jeong

Department of Engineering

The Graduate School

Seoul National University

At present, cancer continues to be a major health problem around the world. The breakthroughs in cancer biology via development of high-throughput sequencing technologies have enabled the development of novel diagnostic and therapeutic approaches. Recently, due to the specificity of fusion genes in cancer, a large number of them have been identified as cancer biomarkers. Since, about 10,000 fusion genes have been discovered via development of sequencing technologies, over 90% of them lack functional mechanisms and therapeutic agents.

Insight into the molecular mechanisms and the identification of potential therapeutic and diagnostic biomarkers in cancer are both greatly aided by the molecular profiling data organized by The Cancer Genome Atlas (TCGA) database. It is significant to offer online platform that make it easy for cancer researchers and clinicians (regardless of their level of computational expertise) to access, analyze, visualize, and

interpret cancer transcriptomic data. cBioPortal, miRGator v 3.0, TANRIC, and ISOexpresso are well known public resources that aid researchers to analyze TCGA data. Here, we report DRPORTAL, an easy to use, interactive web-portal to investigate potential therapies targeting 9,950 fusion genes based on new bioinformatical strategy. DRPORTAL uses TCGA level 3 RNA-seq and clinical data from 33 cancer types. DRPORTAL systematically infer potential drug candidates in fusion positive cancer in four steps: 1) we first extract fusion expression correlated genes as well as age, sex, alive status, TNM cancer stages or other clinical features across fusion positive and negative tumor samples, and 2) select oncogenic signaling pathways containing those genes; and 3) construct a drug-target network using the CIViC and OncoKB database, and 4) finally, prioritize suitable cancer drugs.

I have exploited two different fusion positive cancer (ESR1-CCDC170 fusion positive breast cancer, PTRPRK-RSPO3 fusion positive colorectal cancer) to validate the reliability of novel therapeutic strategy.

In ESR1-CCDC170 fusion positive breast cancer, six major cancer-related signaling pathways (p53, ATR/ATM, FOXM1, hedgehog, cell cycle, and Aurora B) were significantly altered. Further investigation revealed that nine genes (AURKB, HDAC2, PLK1, CENPA, CHEK1, CHEK2, RB1, CCNA2, and MDM2) in coordination with E:C fusion were found to be common denominators in three or more of these pathways, thereby making them promising gene biomarkers for target therapy. Among the 21 putative actionable drugs inferred by drug-target network analysis, palbociclib, alpelisib, ribociclib, dexamethasone, checkpoint kinase inhibitor AXD 7762, irinotecan, milademetan tosylate, R05045337, cisplatin, prexasertib, and olaparib were considered promising drug candidates targeting genes involved in at least two E:C fusion-related pathways.

In PTRPRK-RSPO3 fusion positive colorectal cancer, 2505 genes were altered in RNA expression specific. By pathway analysis based on the altered genes, ten major cancer-related signaling pathways (Apoptosis, Direct p53, EGFR, ErbB, JAK-STAT, tyrosine kinases, Pathways in Cancer, SCF-KIT, VEGFR, and WNT-related Pathway)

were significantly altered in P:R fusion-positive CRC. Among these pathways, the most altered cancer genes (ALK, ACSL3, AXIN, MYC, TP53, GNAQ, ACVR2A, and FAS) specific for P:R fusion and involved in multiple cancer pathways were considered to have a key role in P:R fusion-positive CRC. Based on the drug-target network analysis, crizotinib, alectinib, lorlatinib, brigatinib, ceritinib, erdafitinib, infigratinib and pemigatinib were selected as putative therapeutic candidates.

Based on the two experiments, we confirmed that DRPORTAL can greatly help cancer biologists and clinicians to identify trustable therapeutic targets and applicable drug candidates for fusion positive cancer.

keywords: Bioinformatics, Drug repurposing, TCGA, Gene fusion, Cancer, Web Resource, Bio Platform, Computational Genomics, Transcriptomics, Differentially Expressed Genes Analysis

Student Number: 2019-20994

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Research background

Currently, cancer continues to be a major health problem around the world. According to worldwide cancer statistics for 2022, there will be an expected 25 million new cases of cancer and 13.3 million deaths from cancer until 2030. Finding a way to reduce the mortality caused by cancer is the fundamental goal of society, governments, the medical, and the scientific community. The breakthroughs in cancer biology via development of high-throughput sequencing technologies have enabled the identification of novel diagnostic and therapeutic approaches.

Cancer is caused by a variety of abnormalities to the genome, involving single nucleotide polymorphisms (SNPs), copy number variations, and chromosomal rearrangement. Researchers and clinicians can get a comprehensive understanding of cancer by having access to genomic complexity in tumorigenesis.

However, fusion genes are a form of fusion product with two originally separated genes resulting from the DNA structural rearrangement. A proportion of fusion genes are transcribed into fusion transcripts. These fusion transcripts could synthesize fusion proteins by maintaining the reading frame of their parental genes, by regulating parental expression, or by functioning as long non-coding fusion RNAs. Many fusion genes have been explored and utilized as cancer biomarkers due to their specificity in cancer.

Furthermore, many of these fusion genes have been identified as oncogenes with the ability to induce tumorigenesis and have been utilized as therapeutic targets. These fusion transcripts have also been reported as promising biomarkers since their detection could be used to discern the presence of certain types of cancer cells.

Imatinib, a tyrosine kinase inhibitor, was approved in 2001 for the treatment of Philadelphia chromosome (BCR-ABL)-positive CML as the first fusion-targeted drug. Crizotinib is a renowned tyrosine kinase inhibitor that targets the ALK kinase domain of the EML4-ALK fusion gene in non-small cell lung cancer (NSCLC). In non-small cell lung cancer, the EML4-ALK fusion gene was found to be an oncogenic driver. Patients with NSCLC with EML4-ALK fusion gene were given approval to treat with crizotinib in 2011.

Despite the fact that around 10,000 fusion genes have been discovered in last five years, more than 90% of them lack functional mechanisms and applicable therapeutic agents.

1.2 Research aims

Recent advancements in high throughput technologies, such as next-generation sequencing (NGS) and microarrays, have allowed clinical cancer researchers to examine molecular alterations in DNA, RNA, and proteins on a large scale. Using various data platforms, the Cancer Genome Atlas (TCGA) consortium has generated molecular profiles of over ten thousand samples associated with 33 cancer types, leading to numerous studies involving the genomic and molecular characterization of specific cancer types. Computational methods for predictive repurposing provide a relatively rapid and mechanistically agnostic strategy for identifying therapeutic targets that can be utilized into clinic fields; this may also be the only option for drug development in some fusion-positive cancers for which pathophysiological mechanisms are lacking.

Prior to the first, in Chapter 2, I will introduce the concept of molecular-targeted therapy, drug repurposing and computational approaches which successfully identified therapeutic targets using gene expression profiles. Many of the recently discerned targets and repurposed drug candidates are a consequence of the extensive usage of computational techniques. For this reason, I will first discuss how these techniques have been successfully developed from molecular-targeted therapy to computational methods and introduce good examples of these approaches in previous studies.

In Chapter 3, DRPORTAL, interactive web-portal, based on novel bioinformatic techniques to unearth potential drug candidates of fusion positive cancer will be introduced. Through massive development of

sequencing technology, about 10,000 fusion genes have been identified within five years. While a large number of fusion genes have specificity in cancer and have been identified as oncogenic makers, only limited number of them have applicable drugs. Herein, I will introduce DRPORTAL's main functions and discuss the reason why this could be extremely helpful in accelerating fusion-positive cancer research.

In Chapter 4, I will discuss about the molecular-pathological profiles of ESR1-CCDC170 fusion positive cancer using aforementioned novel computational techniques. The result of this study are divided in to clinicopathological characteristics, differentially expressed genes, target pathways and therapeutic agents, and based on this, predictive ability of DRPORTAL's therapeutic mechanism has been successfully validated.

Finally, in Chapter 5, I would also like to discuss another application example of former therapeutic strategy which elucidated cell signaling and therapeutic targets in PTPRK-RSPO3 fusion-positive colorectal cancer. As in Chapter 4, following in silico studies were performed and confirmed.

Chapter 2 Computational approaches in cancer therapy

2.1 Molecular targeted therapy

Molecular targeted therapy is the application of chemical compounds or other substances that target specific molecules to inhibit the tumorigenesis and proliferation of cancer. In the late 1800s, Paul Rich originally outlined the concept of the "magic bullet," from which targeted therapy was formed. It was first used to describe the capacity of a drug to target microorganisms selectively, but the approach has now been broadened to cancer therapies [1].

Since the Federal Drug Administration (FDA) approved rituximab in 1997, there have been 71 molecular-targeted drugs approved, and 18 of them could be used for multiple indication. While the number of these drugs increased, the number of “non-molecular-targeted” drugs such as cytotoxic pharmaceuticals and antihormonal agents reduced, even when pegylated and novel formulations of existing cytotoxic drugs are taken into consideration. No more than 28 drugs of these “non-molecular-targeted” drugs have been approved or reapproved for new indications since 1997- 2017 [2, 3].

Selecting the appropriate targets is crucial for an effective application of molecular targeted therapies in cancer. Differentiation of the genetic profiles which leads to mutations or variations in proteins and receptors is one of the primary drivers of tumorigenesis by regulating cell survival and proliferation. This specific genetic differentiation, which is appeared to be

differed from cancer and normal cells, can be utilized as molecular targets in the development of molecular targeted therapies [4]. Researchers have been able to interrogate molecular therapies to suppress tumor proliferation and progression by investigating physiology and features of certain molecular targets in cancer.

Interestingly, cancer biomarkers can be identified using genome sequencing, which allows researcher to distinguish genes expression profiles between normal and malignant cells and identify alterations in those expressions [5]. Various cancer genomes have been analyzed using sequencing technology to disclose the genetic heterogeneity between malignant and normal cells within an individual. Among the various targets selected for molecular targeted therapy are growth factors, signaling molecules, cell-cycle proteins, apoptosis modulators, and molecules that promote angiogenesis [6]. Understanding and discerning a specific target enables the development of successful and effective drugs.

2.2 Drugs used in molecular targeted therapy

The activities and properties of molecular targeted drugs exploited in cancer therapy may differ. Depending on the targets, they function on cell surface antigens, growth factors, receptors, or signal transduction pathways which are known to regulate cell cycle progression, cell death, metastasis, and angiogenesis [7]. Small molecules, monoclonal anti-bodies, immunotherapeutic cancer vaccines, and gene therapy are the four major types of molecular-targeted therapeutic methods [1].

To eliminate cancer cells, drugs applied as molecular targeted therapy inhibit signaling that promote cancer cell growth, disrupt with cell cycle regulation, or induce cell death. These drugs may activate the immune system to attack not just cancer cells, but also components of tumor microenvironment [8]. When combined with chemotherapy, these drugs can also prevent the growth and spread of tumors and make resistant tumor more sensitive to other therapies [9].

2.3 Necessity of drug repurposing

Conventional drug development process encompass target finding and validation, lead identification via high-throughput screening, and lead optimization through medicinal chemistry. In pre-clinical stage, substance efficacy and pharmacology (Administration, Metabolism, Distribution, Elimination "ADME") are assessed in animal models, along with toxicity, specificity, and drug interaction investigations.

Despite technological advancements and increased understanding of molecular biology, translation of these insights into therapeutic improvements has been much slower than anticipated [10, 11]. High attrition rates, longer times to provide new pharmaceuticals to market, and fluctuating regulatory standards are compelling the global pharmaceuticals to raise costs of new drugs [12, 13]. It has been calculated that for every dollar spent on research and development (R&D), less than a dollar is returned on average, which may make the pharmaceutical industry a less attractive investment option [14].

Therefore, a large number of cancer drugs on the market are too expensive in present for the vast majority of patients around the world, and there are some studies that new drugs may not have meaningful therapeutic benefits. In addition, no correlation has been identified between drug costs and the advantages to patients [15, 16]. Because of this severe issue, researchers at universities and other non-profit organizations have been coming up a with new idea, drug repurposing [17, 18].

2.3 Drug repurposing

Drug repurposing (also known as drug repositioning) is a method for mapping new indications for approved or experimental drugs that have initial medical indication [10]. Compared to developing a new drug for a given indication, this approach has numerous benefits. First, the risk of failure is much lower; It is because repurposed drug has proven to be adequately safe in preclinical models and humans from previous trials, it is less likely to fail in new indications from a safety standpoint. Second, the drug development timeline may be shortened since much of the preclinical trials, safety evaluation, and, in some cases, formulation development will have been finished already. Third, less money is required, although this may differ considerably depending on the stage and process of candidate's development [19]. Although repurposed drug may have similar regulatory and phase III costs, significant amount of costs may be saved in preclinical and phase I and II costs.

When taken together, these benefits suggest that developing repurposed drugs could yield a faster and safer return on investment, as well

as lower expense once failures have been considering (In fact, bringing a repurposed drug to market is calculated to cost approximately $300 million, whereas bringing a new drug to market is calculated to cost $2–3 billion [20]).

Conclusively, repurposed drugs can elucidate novel targets and pathways that could be used in further studies. Drug repurposing has conventionally been a result of chance and serendipity; once a drug was discovered to have an off-target effect or a newly identified on-target effect, it was applied for commercial use. The effective repurposing of sildenafil citrate for erectile dysfunction depended on retrospective clinical experience, while the successful repurposing of thalidomide for erythema nodosum leprosum (ENL) and multiple myeloma was based on serendipity, rather than a systematic method [10]. As of 2012, global sales of Viagra made from the sildenafil, which had been created as an antihypertensive but repurposed by Pfizer for the treatment of erectile dysfunction reached $2.05 billion, giving Viagra a 47% share of the erectile dysfunction market. Thalidomide, when it was discovered that pregnant women who took the sedative thalidomide during the first trimester of their pregnancies were at risk for having babies with severe skeletal birth abnormalities, the drug was pulled off the market worldwide within 4 years [10].

As a consequence, more systematic and computational approaches to the identification of repurposable drugs have been encouraged. PubMed data reveal that a large number of studies about drug repurposing has increased massively since 2004 [21]. These methods have led to the discovery of many potential drugs candidate, some of which are now undergoing advanced-stage clinical trials in multiple cancer types [14].

2.4 Repurposed drugs in cancer

In cancer, thalidomide, repurposed drug, is now considered as standard therapy. The FDA authorized this drug in 1998 as combination with dexamethasone for the treatment of newly diagnosed multiple myeloma. Nonetheless, the National Comprehensive Cancer Network (NCCN) guidelines now recommend this drug as a main therapeutic option in combination with bortezomib and dexamethasone. It is interesting to note that thalidomide is not often utilized in the United States but is more accessible and economical in other regions of the countries with less resources [22].

Treatment for acute promyelocytic leukemia now includes arsenic trioxide, which was previously used in traditional Chinese medicine, and all-trans retinoic acid (ATRA), which had been used since 1962 for skin disorders but was licensed by the FDA in 2000. Only these three drugs have been successfully repurposed for use in cancer treatment [23, 24].

Additionally, repurposing drug candidates have been unearthed novel mechanisms into molecular pathophysiology of cancer. The finding in 2001 that AMP-activated protein kinase (AMPK), was the target of metformin, followed by the evidence that AMPK is also a cancer target, is an example that exemplifies this case [25]. Investigating the anticancer effects of previously approved drug is another method of off-target toxicity. For example, hydralazine and procainamide, originally used to treat autoimmune disorders, have been repurposed to DNA methyltransferase inhibitors for cancer drugs [26].

2.5 Computational approaches using gene expression profile in cancer

Identification of target candidates for a given indication (hypothesis generation); systematic assessment of the pharmacological effects in animal models; and evaluation of efficacy in phase II clinical trials (assuming there is adequate safety data from phase I, which have already evaluated for original indication) are the three big phases of a drug repurposing. Among these three phases, step 1 — Identifying the appropriate drug candidates for new indication with a high level of confidence is the most important, and there are many new tools for hypothesis generation. These systematic processes may be classified into computational and experimental approaches, both of which can synergistically exploited. These two main categories include clinical data.

Researchers are able to obtain vast amounts of experimental data as a result of development in technologies such as next-generation sequencing and rapidly decreasing sequencing costs. These data include high-throughput DNA and RNA sequencing, mass spectrometry, metabolomics and transcriptomic data, phenotyping data, and many more. In addition, large amounts of clinical data are increasingly viable through electronic health records (EHRs), clinical trials, and biobanks. Big data refers to data collections that are so big and complicated that conventional data processing techniques are inadequate [27].

Consequently, computational approaches, often referred to as bioinformatic analysis, have emerged. Bioinformatics employs various computational techniques, including sequence and structural alignment, designing databases, data mining, macromolecular geometry, building

phylogenetic trees, predicting protein structure and function, identifying new genes, and clustering expression data. Bioinformatics is becoming more commonly used and many of the candidates for repurposing that have been recently discovered come from this computational approach.

Computational analysis can be conducted using a various type of data, including transcriptomics, genomes, proteomics, epigenomics, and metabolomic profiles, adverse effects, phenotypes, or a combination of these. Since the publicly accessible databases for transcriptomic data are well-known and contain normalized data, they are also used in other well-established resources such as cBioPortal, miRGator v3.0, TANRIC, and ISOexpresso. This is reason why transcriptomic profile from TCGA database was interrogated into DRPORTAL, also. As so, I will introduce good example of computational approaches based on transcriptomics, genomics, and pathway analysis that are applicated in DRPORTAL's bioinformatical analyses.

Transcriptomic profiles can offer a list of under- and over- expressed genes in an experimental environment, such as disease versus normal or drug-treatment group versus control. These gene lists may then be exploited to identify dysregulated pathways or networks. The Connectivity Map (CMap) is a good example of this concept; it is a huge collection of transcriptomes from cell lines treated with 1300 drug-like compounds that uses a pattern-matching approach to unearth differences and similarities in complicated diseases [28]. Using publicly accessible transcriptomic profiles, Iorio and colleagues built a drug network based on the "guilt by association" concept to identify drugs that had a similar transcriptional signature and, hence, a perceived similar mechanism of action [29, 30]. Based only on these transcriptomic profiles, a drug network was constructed with 1,302 drugs as

nodes and 41,047 edges (showing similarity between pairs of drugs). Remarkably, nine known anticancer drugs' mechanisms were successfully validated using this network, demonstrating its predictive ability.

As a result of the significant progress in human genetics, "druggable" targets may be identified by identifying the associations between genes and diseases. The use of genetic or genomic methods, particularly large-scale genetic investigations such as genome-wide association studies, are reported to be twice as likely to be approved compared to drugs with no such links [31, 32]. With this method, it was possible to identify the genes that encode for the drug targets of tamoxifen (ESR1) and aromatase inhibitors (CYP19A1), which are correlated to genetic differences that increase the risk of breast and endometrial cancer [33, 34]. These two drugs are FDA-approved for breast and endometrial cancer.

Approaches based on pathways or networks have been used extensively to discover medications or pharmacological targets with potential for repurposing [35]. As mentioned before, despite the fact that some of the putative targets identified by genomics or transcriptomic data may be directly amenable as therapeutic targets, these genes are often not ideal therapeutic targets. In this case, a pathway-based strategy could give information about genes that are either upstream or downstream of the target and could be used to find new uses for them [36]. Network analysis is the construction of drug or illness networks based on gene expression patterns, disease pathophysiology, and protein interactions to help in the identification of repurposing candidates. Several signature matching studies based on the transcriptome profile also use the network analysis method [29, 37].

2.6 Computational approaches of drug repurposing in fusion positive cancer

Since some fusion genes have been found to be dominant in various cancer, researchers discovered that these genes have significant functions in tumorigenesis [38, 39]. Therefore, diverse efforts have been progressed to identify the molecular pathological profiles of the fusion gene and confirmed that successful targeted therapies have been accomplished when these genes were inhibited [40, 41].

Imatinib, a tyrosine kinase inhibitor, was approved in 2001 for the treatment of Philadelphia chromosome (BCR-ABL)-positive CML as the first fusion-targeted drug. Crizotinib is a renowned tyrosine kinase inhibitor that targets the ALK kinase domain of the EML4-ALK fusion gene in non-small cell lung cancer (NSCLC). Patients with NSCLC with EML4-ALK fusion gene were given approval to treat with crizotinib in 2011.

However, there are many other fusion genes that cannot be directly targeted with this conventional approach. In order to overcome this limitation, it was necessary to newly explore the targetable genes and signaling pathways that function in the downstream level of fusion gene. Thus, current studies that elucidated the novel targets and applicable drugs with computational approaches have been emerged [38, 42-44].

First, in the case of using tyrosine kinase inhibitors (TKIs) as standard molecular targeted therapy for BCR-ABL fusion positive chronic myeloid leukemia (CML), researchers identified novel targetable markers to prevent drug resistance caused by TKIs. In this study, DEG analysis was processed using gene expression data (GEO database) between pre-treatment and post-

treatment of TKI, and target pathways were identified through two pathway analysis (gene ontology; GO terms and KEGG pathway). Therapeutic compounds which have high correlation with the target genes were selected through drug response gene signature database, Connectivity Map. However, in the case of Cmap database, it only comprises gene signature data for four cancer cell lines (breast cancer, prostate cancer, leukemia, and skin cancer). Therefore, it is difficult to search for therapeutic agents in other types of cancer [43].

In other study, RUNX1-RUNX1T1 fusion gene plays a crucial role in the tumorigenesis of Acute Myeloid Leukemias (AML), but since direct targeting is difficult, researchers tried to investigate signaling pathways that can be targeted. Using TCGA mRNA expression data, they found 293 genes highly related to the RUNX1-RUNX1T1 fusion, and based on these genes, two pathway analysis (ORA and GSEA) were performed to identify fusion-related signaling pathways. As a result, it was possible to discern cyclooxygenase (COX), vascular endothelial growth factor receptor (VEGFR), platelet-derived growth factor receptor (PDGFR), and fibroblast growth factor receptor (FGFR) pathways. In vitro experiments were also conducted to validate whether the identified signaling pathways were actually targets of R:R fusion positive AML. As a consequence, when the 4 pathways were pharmacologically inhibited, it was confirmed that the proliferation of AML cells significantly decreased in fusion positive compared to fusion negative [42].

In last study, molecular signaling was investigated and potential therapeutic targets were curated for TMPRSS2-ERG fusion positive cancer, which is likely to account for nearly 50% of prostate cancer patients. Then,

3,870 differentially expressed genes between fusion-positive and negative groups were identified, and pathway analysis was conducted based on these genes. Finally, based on the drug-target database (CIViC), 55 drug candidates were repurposed to these targets, and in vitro experiments discerned the six drugs from previous candidates which were effective to fusion positive cancer [44].

Taken together, all these three studies identified target genes and pathways via DEG analysis and pathway analysis. Based on the public transcriptomic data (GEO, TCGA), significantly altered genes were selected and, pathway analysis was conducted based on these DEGs to investigate cellular processes.

In RUNX1-RUNX1T1 fusion and TMPRSS2-ERG fusion studies, in vitro experiments were also performed to verify the reliability of target pathways and genes found in silico, and it was successfully validated because a fraction of drugs successfully reduced the proliferation of fusion positive cancer cell by inhibiting the targets [42, 44]. From this point of view, bioinformatic analysis commonly used in previous drug repurposing studies is effective in identifying putative target genes and pathways and can prioritize some effective drug candidates which can be utilized in further experiments.

However, these fusion positive cancers have been analyzed individually only in a very limited number. With the development of deep sequencing and detection algorithms, more than 90% of 10,000 fusion genes have been identified within the last 5 years, and the cellular mechanism has not been fully identified since most of the fusion genes have been recently discovered [38, 45].

If the aforementioned computational analysis is performed for each fusion positive cancer, it will take a long time and not be easy to reproduce for some clinicians and cancer researchers who lack bioinformatic or programming skills. Furthermore, it will require more effort and time to change some experimental conditions or update numerous, complex databases. Therefore, interactive web-resource can help these researchers to access various types of complex databases and investigate the molecular pathological characteristics easily.

Chapter 3 DRPORTAL: A Web-Portal for repurposing potential drug candidates in fusion positive cancer

3.1 Introduction

Next-generation sequencing (NGS) and microarrays are examples of recently developed high throughput technologies that have allowed cancer researchers to investigate molecular alterations in DNA, RNA, and proteins on a massive scale [46-48]. Using numerous data platforms (including DNA methylation and copy number, as well as RNA and protein expression), the Cancer Genome Atlas (TCGA) collaboration has constructed molecular profiles of over ten thousand samples linked to 33 types of cancer, generating a large number of studies including the genomic and molecular characterization of individual cancer types [49-68].

Clinicians and researchers who lack bioinformatics abilities have difficulty to undertake in-depth analysis of TCGA cancer genomics data due to its massive complexity and accessibility in diverse data formats. Numerous analytical platforms have been developed in order to facilitate fundamental data queries. cBioPortal is one such application that allows users to input gene sets for a specific type of cancer. cBioPortal provides RNA level expression data, mutation events, copy number variations, protein expression by Reverse Phase Protein Array (RPPA), a survival plot, and a list of co-expressed and mutually expressed genes for each gene that is queried. Other

tools like as miRGator v3.0 [69], TANRIC [70], and ISOexpresso [71] can be used to examine differential expression of specific biomolecules including miRNA, lincRNA, and transcript isoforms. Using TCGA data, the Gene-Drug Interaction for Survival in Cancer (GDISC) [72] web platform evaluates the effect of gene-drug interactions on various cancer types. The Stanford Cancer Genome Atlas Clinical Explorer (Stanford-TCGA-CE) [73] aids in the detection of associations between genomic/proteomic characteristics and clinical parameters, hence easing the identification of clinically relevant genes. PROGgeneV2 enables extensive survival analysis of publicly accessible gene expression data, such as TCGA [74]. Using TCGA and other public cDNA, Affymetrix, and Illumina microarray data, Oncomine [75, 76] offers an interactive platform for gene expression profiling.

Researchers have created various computational tools to assist them in conducting specific analyses of TCGA data.; however, there is need for new tool to analyze the molecular pathological features across a large number of fusion positive cancer (n=9,950). Due to the specificity of fusion genes in cancer, some of them have been developed as cancer biomarkers. Since over 9,000 fusion genes have been discovered in last 5 years, 90% of them lack functional insights and applicable drugs.

Here, we report DRPORTAL (Figure 1-2), an easy to use, interactive web-portal to investigate potential therapies targeting 9,950 fusion genes based on new bioinformatical strategy. DRPORTAL utilizes TCGA level 3 RNA-seq and clinical profiles from 33 cancer types. The web resource systematically infer potential drug candidates in fusion positive cancer in four steps: 1) we first extract fusion expression correlated genes as well as age, sex, survival status, TNM cancer stages or other clinical features across

fusion positive and negative tumor samples, and 2) select oncogenic signaling pathways containing those genes; and 3) construct a drug-target network using the CIViC and OncoKB database, and 4) finally, prioritize suitable cancer drugs.

This resource serves as a foundation for validating fusion genes through computer simulation and discovering potential drug candidates. Thus, DRPORTAL could be extremely helpful in accelerating fusion-positive cancer research.

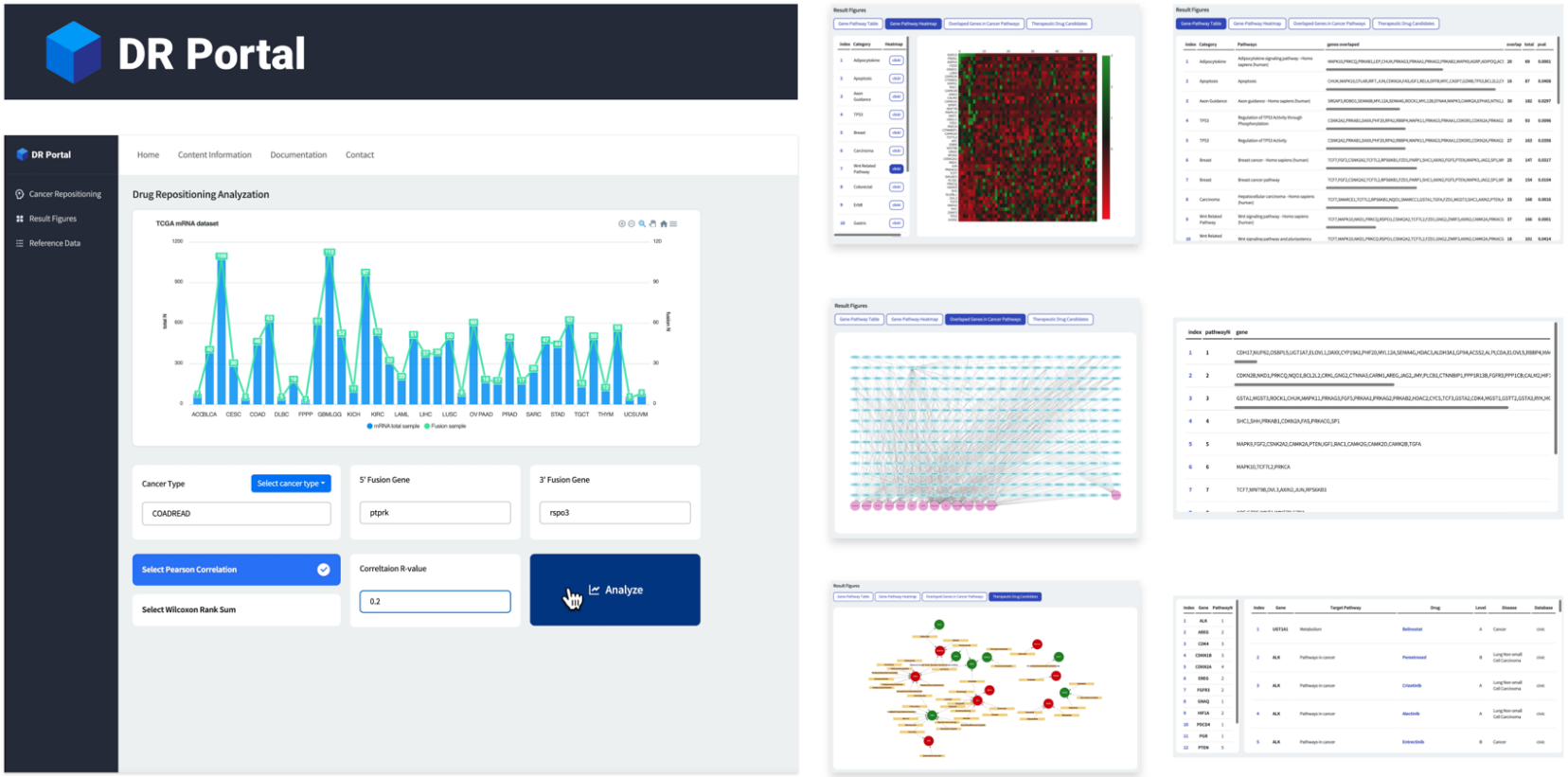


Figure 1. Overall schema of DRPORTAL.

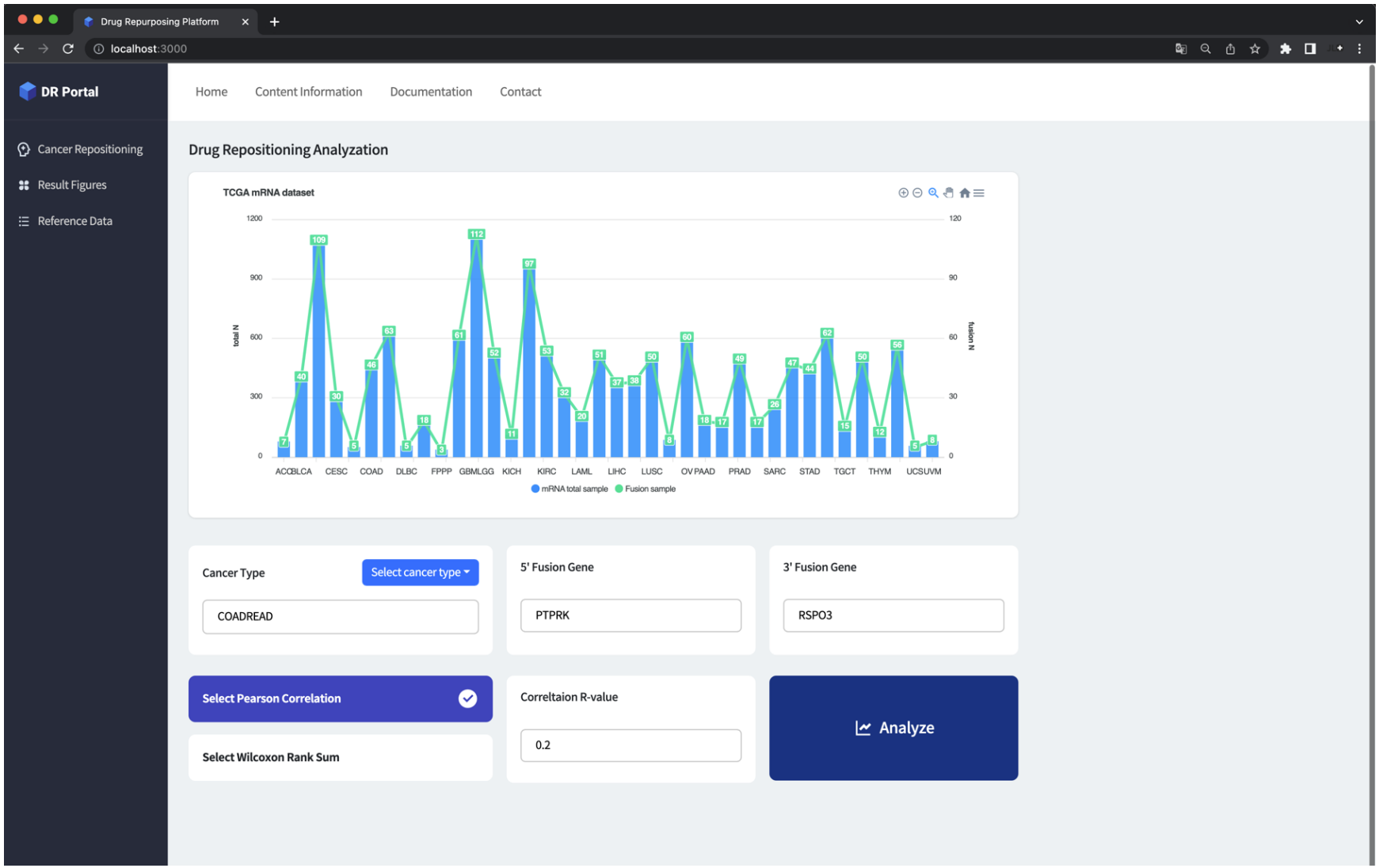


Figure 2. DRPORTAL input page. Researcher can select cancer type, 5’ fusion gene, 3’ fusion gene, method of DEG analysis, cut-off value of DEG analysis .

3.2 Materials and Methods

3.2.1 mRNA expression data

Gene level 3 (RNA-seq by expectation-maximization, RSEM) mRNA expression with normalized read count values of 33 TCGA cancer types was obtained from the Broad GDAC Firehose website (https://gdac.broadinstitute.org). "illuminahiseq_rnaseqv2 RSEM_genes_normalized (MD5)" files were obtained for primary solid tumor for each cancer. This file includes gene expression values estimated by RSEM algorithm for 20,531 genes; Each column represents the patient ID, while each row represents the entrez gene ID. Related clinical feature data, including information about age, sex, mutation annotation format (MAF) files, molecular subtypes, and tumor-node-metastasis (TNM) stages, were obtained from the website mentioned above.

3.2.2 Case-control selection and Differential Expressed Genes analysis

RNA expression data from TCGA were made into a two-dimensional matrix composed of the selected fusion tumor samples and control tumor samples. I downloaded 9,950 fusion gene samples using the TCGA fusion gene data portal (The Jackson Laboratory, https://www.tumorfusions.org). For control tumor sample selection, 50 samples were randomly selected among the samples with low fusion gene expression (less than the median value of reference gene RNA expression). I used two different statistic methods to select differentially expressed genes between two tumor groups.

First, Wilcoxon rank sum tests were performed between fusion-positive and fusion-negative patients to select genes in coordination with fusion gene in the RNA expression. Then, above tests were repeated 100 times. Based on the median of *p*-values from 100 tests, 20,531 genes were sorted in decreasing order and selected only when it satisfies "cut-off" over 80 times among 100 tests. Using this condition, mostly affected genes were selected by "cut-off" *p*-value (default adjusted *p*-value < 0.01).

Second, Pearson correlation tests were performed in fusion-positive and fusion-negative cases to obtain R-values of 20,531 genes in correlation with fusion gene **in RNA expression. Then, above tests were repeated 100 times also. Based on the median of absolute R values from 100 tests, 20,531 genes were sorted in decreasing order. Using the median of absolute R values, mostly affected 2,505 genes were selected by correlation cut-off (default R > 0.2).

3.2.3 Identification of key altered genes via ConsensusPathDB (CPDB) and Over-representation analysis (ORA)

Pathway enrichment analysis is useful for researchers to gain an understanding of the underlying mechanisms of gene lists obtained from large-scale (omics) experiments, particularly DEGs (Differentially Expressed Genes). This approach is used to find biological pathways that have more genes from a given gene list than what would be expected by random chance. ConsensusPathDB (http://consensuspathdb.org/) is a meta-database combining interactions of diverse types from 31 public resources for humans. According to BioCarta (http://www.biocarta.com/], 177 biological pathways

were combined from the following sources: INOH, KEGG, NetPath, PID, Reactome and Wikipathways. Using ConsensusPathDB, researchers commonly evaluate lists of genes, proteins, and metabolites against sets of molecular interactions defined by pathways, Gene Ontology and network neighborhoods and retrieve complex molecular neighborhoods formed by heterogeneous interaction types.

The aforementioned differentially expressed genes were used to perform over-representation analysis (ORA), pathway analysis, using ConsensusPathDB. In over-representation analysis, the p-value is determined using the hypergeometric distribution which reflects the significance of the observed relationship between the input gene list and the members of the pathway in comparison to random expectations. Analyzing the ontological features and the proportion of duplicated genes, the pathways enriched with chosen differentially expressed genes were collapsed into cancer-related pathways, having key altered genes as components.

3.2.4 Inferring and prioritizing actionable drugs via CIViC and OncoKB

The “Clinical Evidence Summaries” data, released on 1 December 2021, were downloaded from the Clinical Interpretations of Variants in Cancer (CIViC) website (https://civic.genome), and the “Actionable Variants” data were accessed and downloaded on 1 December 2021 from the OncoKB website (http://oncokb.org/). Each drug database comprise 3,374 actionable variations (470 genes) and 670 variants (161 genes). CIViC is a knowledgebase that is created by a combination of experts and the public to provide information about the significance of inherited and somatic variants in

cancer treatment, diagnosis, prognosis, and predisposition. OncoKB also annotates the biologic and oncogenic effects and prognostic and predictive significance of somatic molecular alterations. The information about the potential treatment impact of a specific molecular alteration is sorted according to the level of evidence that it can predict drug response. The evidence is based on the labeling and guidelines of the US Food and Drug Administration, National Comprehensive Cancer Network, recommendations of disease-focused expert groups and scientific literature.

OncoKB database is from MSKCC (Memorial Sloan Kettering Cancer Center), and CIViC database is from Washington University. As such, these are all databases curated by experts, and in our experience, when comparing the above two databases with experimental studies and literature reviews, the results were more than 95% accurate.

A clinical evidence statement is a piece of information that has been carefully selected from reliable medical literature, it refers to the variant or genomic event that has an impact on cancer predisposition, diagnosis, prognosis, or the prediction of response to therapy, that has been manually curated. Evidence level in each database describes the robustness of the study supporting the evidence item.

In CIViC database, five different evidence levels are supported: “A - Validated association”, “B - Clinical evidence”, “C - Case study”, “D - Preclinical evidence”, and “E - Inferential association”. Clinical evidence A drugs have a proven or clinical consensus on the variant association in human medicine. Typically, these evidence items describe Phase III clinical trials or have associated companion diagnostics. Clinical evidence B drugs have

typically large clinical trials or other primary patient data supporting the clinical association. These evidence items usually include more than 5 patients supporting the claim made in the evidence statement.

In OncoKB database, four different evidence levels are supported: “1 - FDA-approved drug as FDA-recognized”, “2 - FDA-approved drug as standard care”, “3 - Compelling clinical evidence”, and “4 - Compelling biological evidence”. Level 1 drugs are FDA-recognized biomarker predictive of response to an FDA-approved drug in this indication. Level 2 drugs are standard care biomarker recommended by the NCCN or other professional guidelines predictive of response to an FDA-approved drug in this indication.

Thus, exploiting the drugs that have high clinical evidence level in both drug-target database, a total of 740 CIViC (level A -Validated association, level B -Clinical evidence) and 182 OncoKB (level 1 - FDA approved drug as FDA-recognized, level 2 - FDA approved drug as standard care) drugs were matched to key altered genes and fusion-related pathways.

Then drug-target relationships were prioritized based on the scenario that properly working cancer drugs are generally inhibitors for activated oncogenes or activators for down-regulated tumor suppressor genes.

3.2.5 Statistical analysis and data visualization

All statistical analyses, including the Pearson correlation tests, Wilcoxon rank sum test, and Over-representation analysis were performed using the open software Python 3.10.9. SciPy (https://docs.scipy.org/), a

Python package, a collection of mathematical algorithms and convenience functions built on the NumPy extension of Python was used for statistical analyses. Plotly.js, Open-Source Graphing Libraries was used to visualize an RNA expression heatmap and all the other graphical figures. The comprehensive network between targetable drugs and therapeutic agents was analyzed and illustrated using Cytoscape 3.5.3. Also signaling pathways and overlapped genes network is also visualized by Cytoscape. In this study, statistical significance was determined as a p-value of 0.05 and false detection rate (FDR) as a q-value of 0.01.

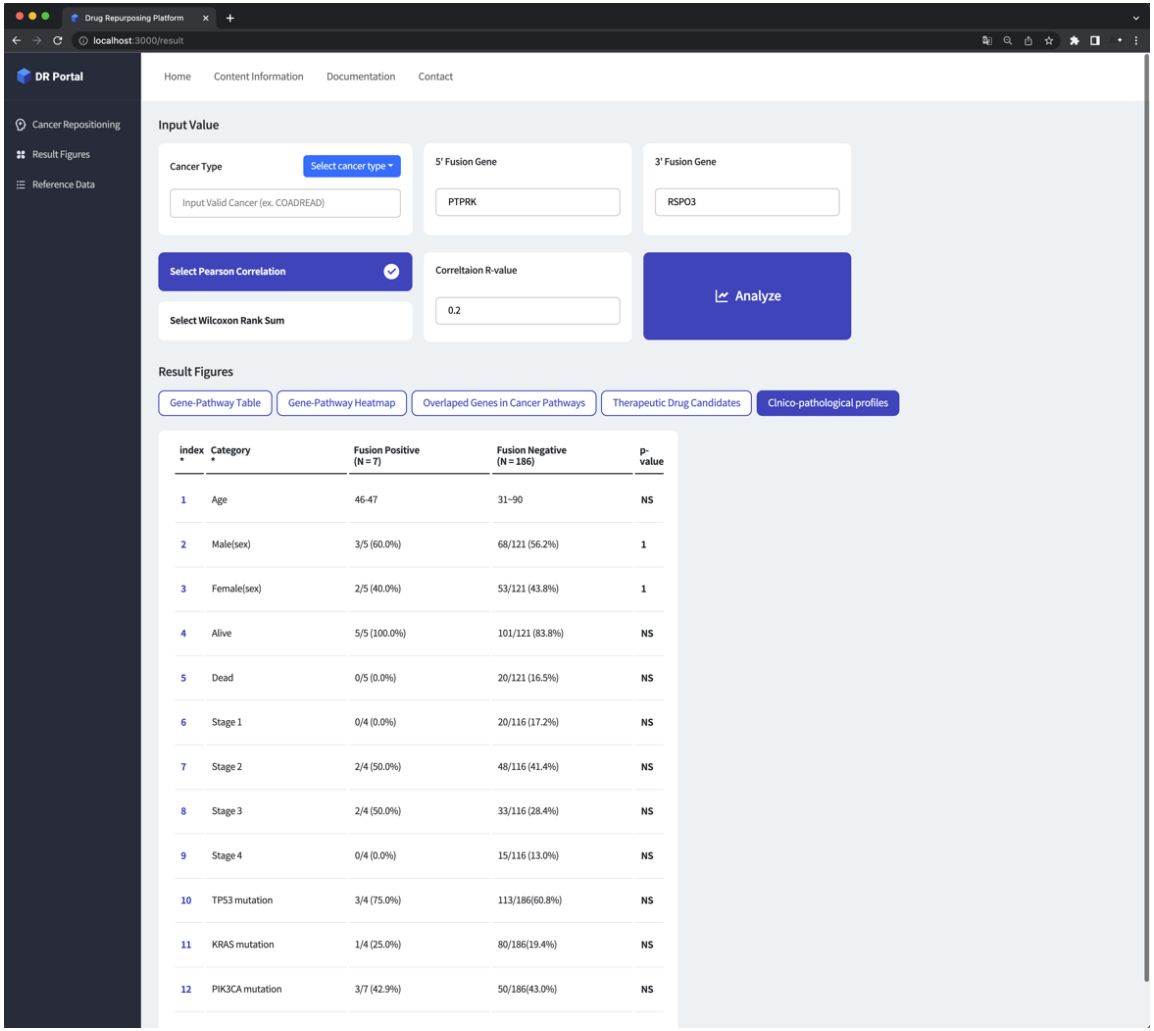
3.2.6 Web-application development via software programming languages

Front-end software engineering was developed by open-source library, React.js, more commonly known as React, is a free, open-source JavaScript library. It works best to build user interfaces by combining sections of code (components) into full websites. Building rest APIs of back-end software engineering was developed by open-source library, Fast API, a new, high-performance web framework for creating APIs in Python 3.7 or later. It uses Python's standard type hints to build the API. The server that is used to host the service was created by AWS (Amazon Web Services), which is a wide-ranging and constantly developing cloud computing platform offered by Amazon. It includes a combination of Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) options.

3.3 Result and Discussion

3.3.1 Clinicopathological characteristics

First, the clinicopathological characteristics of fusion-positive and fusion-negative patients among total cancer patients in Broad GDAC Firehose were analyzed. By comparing cancer-related parameter such as age, sex, alive status, Tumor Node Metastasis stage, and Mutation Annotation Files, researchers were able to analyze specific clinical features of fusion positive cancer patients among other control patients (Figure 3).



index	Category	Fusion Positive (N = 7)	Fusion Negative (N = 186)	p-value
1	Age	46-47	31~90	NS
2	Male(sex)	3/5 (60.0%)	68/121 (56.2%)	1
3	Female(sex)	2/5 (40.0%)	53/121 (43.8%)	1
4	Alive	5/5 (100.0%)	101/121 (83.8%)	NS
5	Dead	0/5 (0.0%)	20/121 (16.5%)	NS
6	Stage 1	0/4 (0.0%)	20/116 (17.2%)	NS
7	Stage 2	2/4 (50.0%)	48/116 (41.4%)	NS
8	Stage 3	2/4 (50.0%)	33/116 (28.4%)	NS
9	Stage 4	0/4 (0.0%)	15/116 (13.0%)	NS
10	TP53 mutation	3/4 (75.0%)	113/186(60.8%)	NS
11	KRAS mutation	1/4 (25.0%)	80/186(19.4%)	NS
12	PIK3CA mutation	3/7 (42.9%)	50/186(43.0%)	NS

Figure 3. DRPORTAL output page listing clinicopathological characteristic of fusion positive and negative patients. It comprises age, sex, alive status, TNM stages, and mutation information.

3.3.2 Key genes and pathways altered in fusion-positive cancer

Differentially expressed genes obtained by Wilcoxon rank sum test either Pearson correlation test were inputted for performing the over-representation analysis (ORA) by using ConsensusPathDB to select fusion-related cancer pathways.

I utilized 37 cancer signaling pathways with frequent genetic variations, with key cancer genes established in previous research (TCGA [77], KEGG pathway [78]), and focused on pathway members likely to be cancer drivers or therapeutic targets.

The pathways used are: (1) Adipocytokine (2) Apoptosis, (3) Axon Guidance, (4) Bladder cancer, (5) Brest cancer, (6) Carcinoma, (7) Cell cycle, (8) Colorectal cancer, (9) EGFR, (10) Endometrial, (11) ErbB, (12) FGFR, (13) Gastric cancer, (14) Hippo, (15) Irinotecan, (16) JAC-STAT, (17) Leukemia, (18) Lung cancer, (19) Melanoma, (20) Metabolism, (21) MicroRNAs, (22) Migrations, (23) Migration, (24) Notch, (25) P53, (26) Pathways in cancer, (27) PI3K, (28) Prostaglandin, (29) Prostate cancer, (30) SCF-KIT, (31) Stem cell, (32) Thyroid cancer, (33) TNF, (34) TP53, (35) Tyrosine kinase, (36) VEGF, and (37) WNT.

Fusion-related cancer pathways were discerned among these 37 cancer-signaling pathways and hypergeometric distribution p-value was exploited for ORA (q-value < 0.01). As a result, Figure 4 comprises key altered genes which were enriched in fusion-related cancer pathways.

Gene expression heatmaps of obtained cancer-related pathways correlated with fusion mRNA expression were visualized in Figure 5. The x-axis is indicative of the sample, while the y-axis is indicative of its respective

RNA expression. The RNA expression was converted into z-score prior to representation on the heatmap.

Putative target genes involved in multiple cancer-related pathways of fusion-positive cancer were visualized via Cytoscape (Figure 6). These genes are important for tumor proliferation and maintenance specific to fusion-positive patients. Putative target genes involved in multiple pathways were organized in Figure 7.

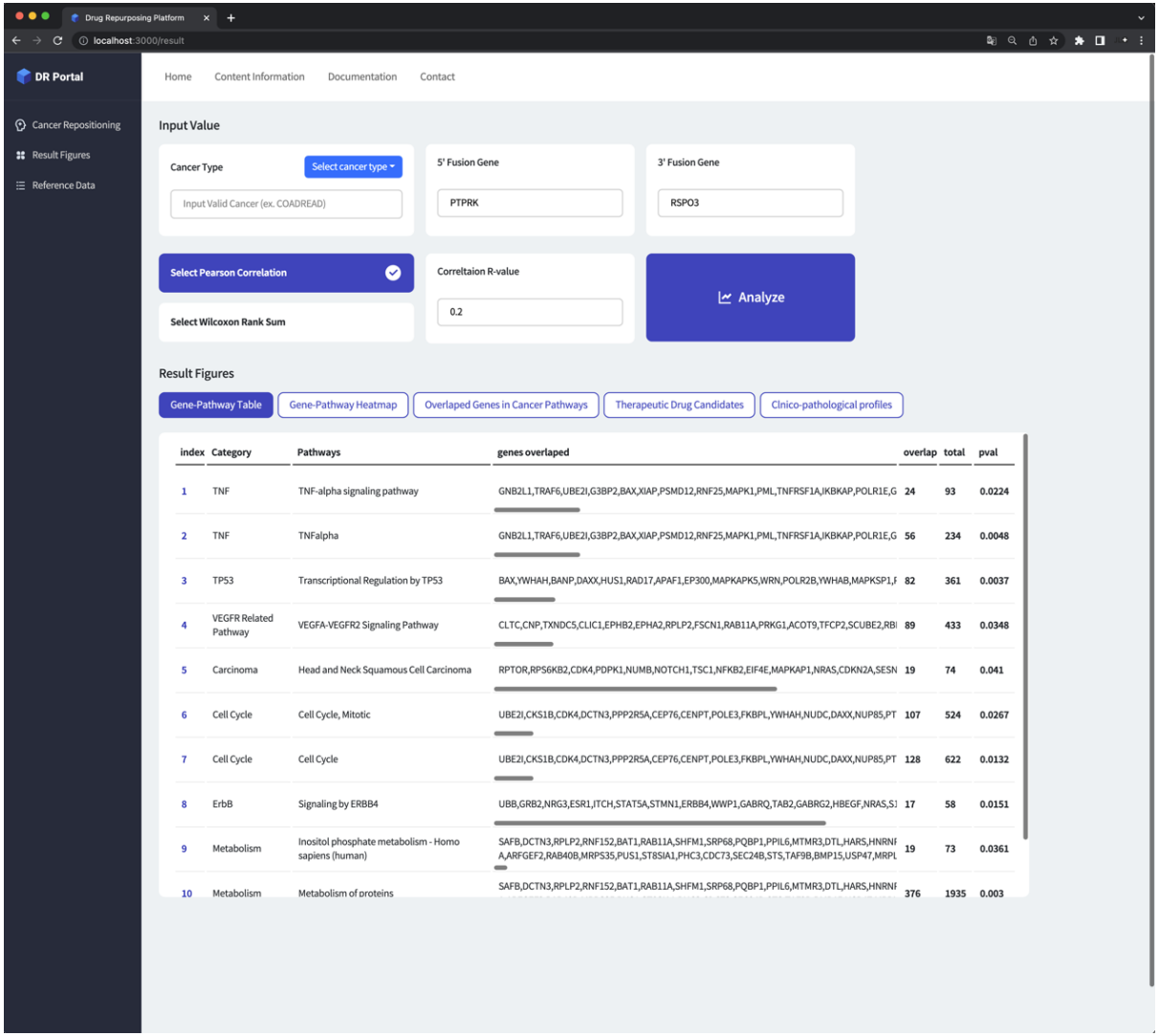


Figure 4. DRPORTAL output page listing key genes and pathways altered in fusion-positive cancer

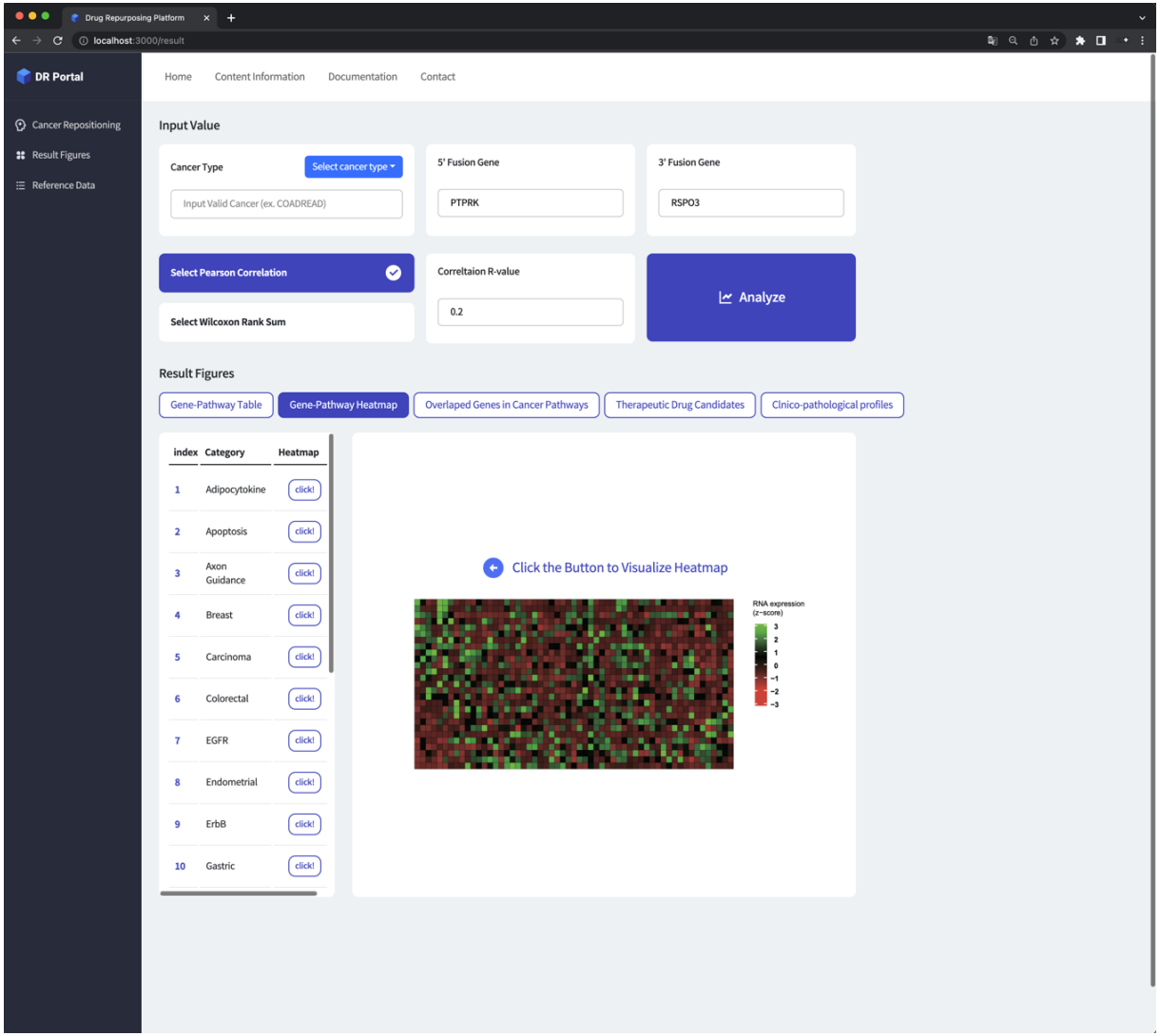
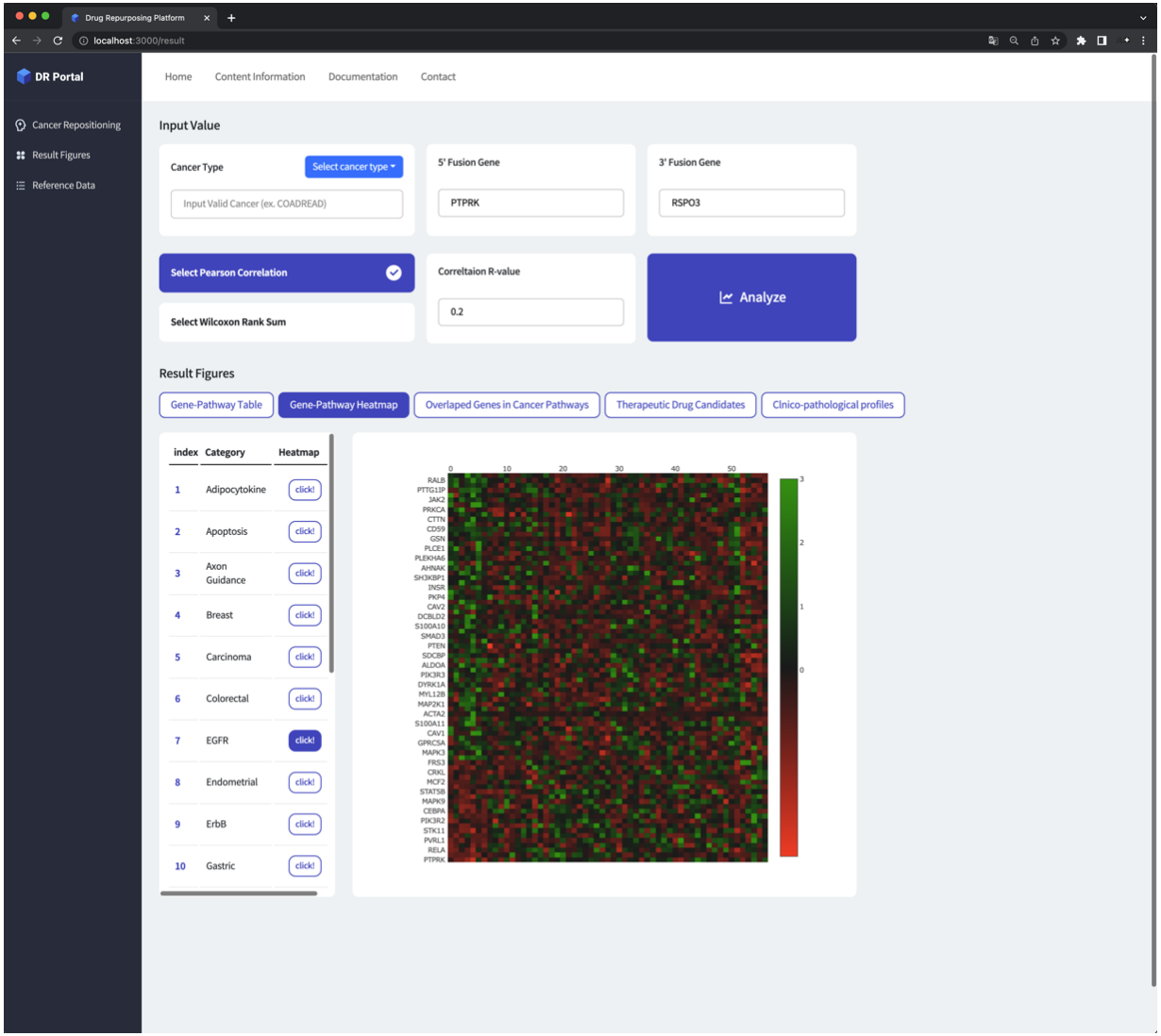


Figure 5. DRPORTAL output page visualizing gene expression heatmap of cancer-related pathways enriched with genes that were correlated to fusion gene's RNA expression.


Drug Repurposing Platform
localhost:3000/result
DR Portal
Home
Content Information
Documentation
Contact
Cancer Repositioning
Result Figures
Reference Data
Input Value
Cancer Type
Select cancer type
Input Valid Cancer (ex. COADREAD)
5' Fusion Gene
PTPRK
3' Fusion Gene
RSPO3
Select Pearson Correlation
Select Wilcoxon Rank Sum
Correltaion R-value
0.2
Analyze
Result Figures
Gene-Pathway Table
Gene-Pathway Heatmap
Overlaped Genes in Cancer Pathways
Therapeutic Drug Candidates
Clnico-pathological profiles
index
Category
Heatmap
1 Adipocytokine click!
2 Apoptosis click!
3 Axon Guidance click!
4 Breast click!
5 Carcinoma click!
6 Colorectal click!
7 EGFR click!
8 Endometrial click!
9 ErbB click!
10 Gastric click!
0
10
20
30
40
50
RALB
PTTG1IP
JAK2
PRKCA
CTTN
CD59
GSN
PLCE1
PLEKHA6
AHNAK
SH3KBP1
INSR
PKP4
CAV2
DCBLD2
S100A10
SMAD3
PTEN
SDCBP
ALDOA
PIK3R3
DYRK1A
MYL12B
MAP2K1
ACTA2
S100A11
CAV1
GPRC5A
MAPK3
FRS3
CRKL
MCF2
STAT5B
MAPK9
CEBPA
PIK3R2
STK11
PVRL1
RELA
PTPRK
3
2
1
0

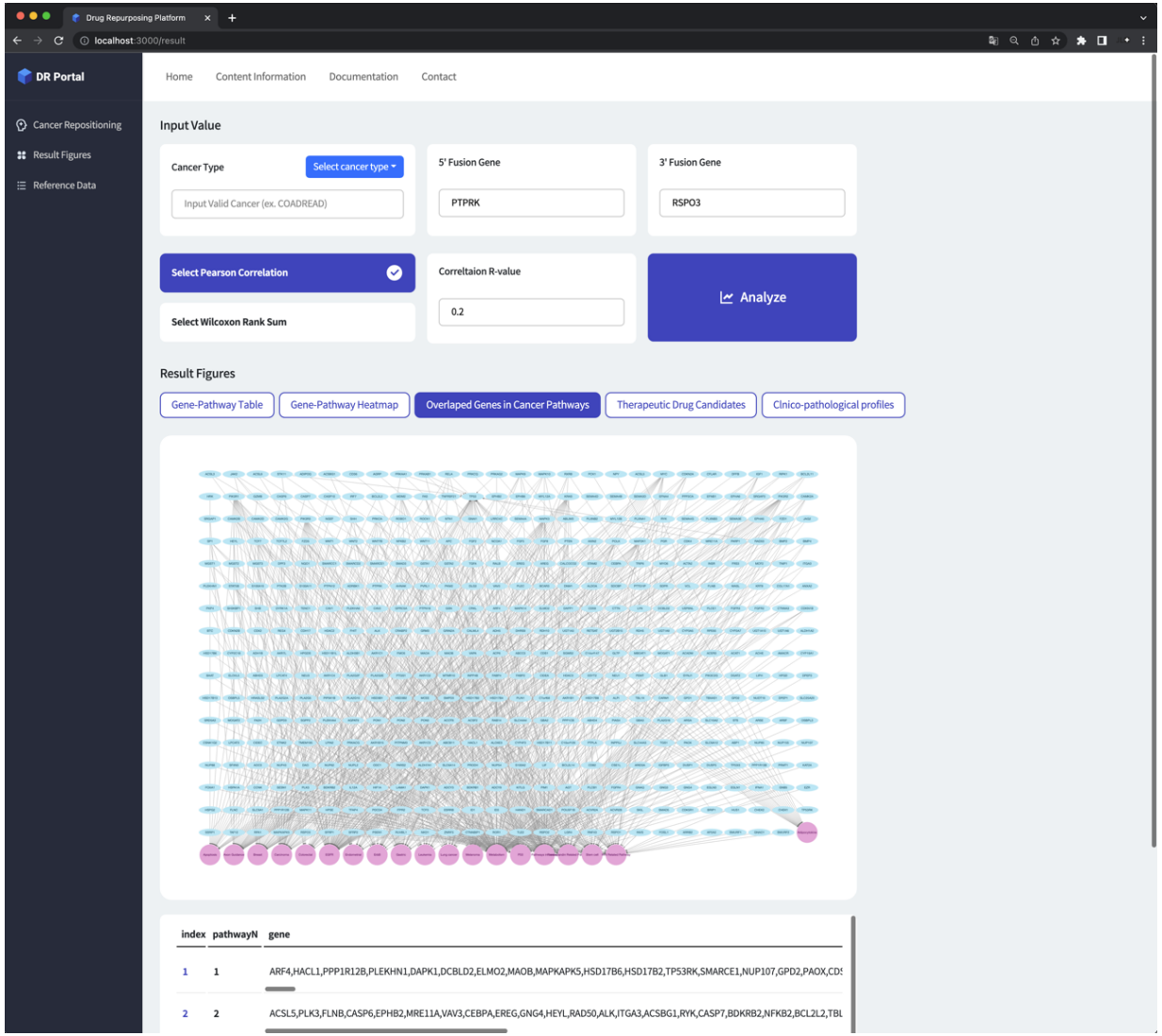


Figure 6. DRPORTAL output page visualizing putative target genes involved in multiple pathways of fusion-positive cancer.

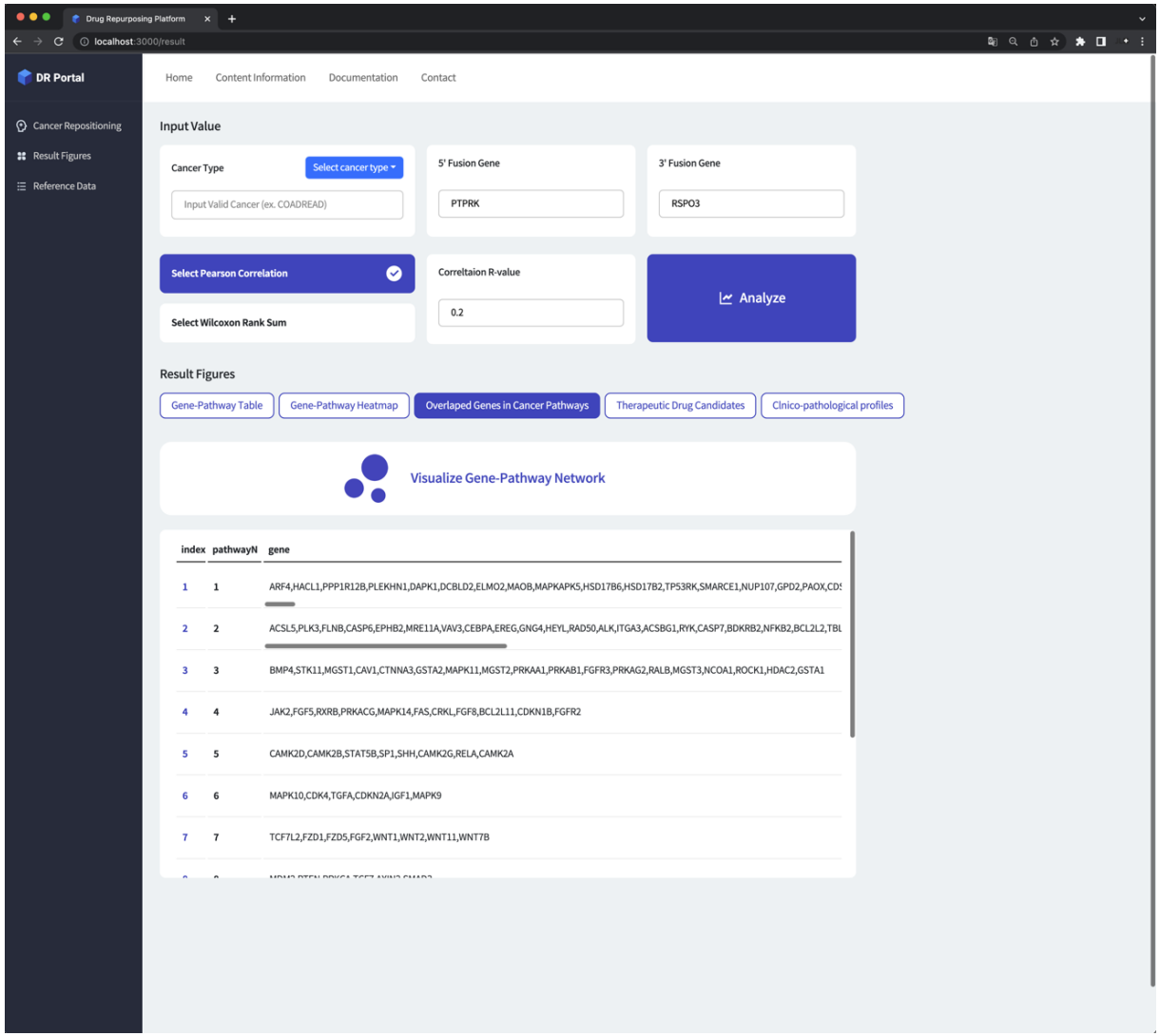


Figure 7. DRPORTAL output page listing putative target genes involved in multiple pathways of fusion-positive cancer.

3.3.3 Identification of actionable targets and potential therapeutic choice using network analysis

Actionable target genes and drugs that have high clinical evidence were extracted by mapping key altered genes from pathway analysis in the following drug databases. Drug-target network was visualized based on CIViC (n = 673) and OncoKB (n = 262) databases (Figure 8-9): Yellow boxes, drugs; red circles, genes that are over-expressed in fusion-positive cancer; green circles, genes that are under-expressed in fusion-positive cancer. Drug-target table comprises putative target genes, target pathways, actionable drugs, clinical evidence of drugs and original indication of cancer drugs.

However, there are three classes for activation signaling including hotspot mutation, amplification, and overexpression. For this study, RNA sequence-based overexpression was considered an activating signal. CIViC and OncoKB drug database provides information on the relationship between the activating signaling (three classes mentioned above) and available drugs. For example, MET activating mutations are including amplification, over-expression, and activating point mutations and the three class of mutations are mostly sharing target-drug sensitivity (capmatinib, tepotinib). So, the three types of mutations were considered as showing similar target-drug sensitivity in silico level. Although the sensitivity of the drugs may differ according to the various types of signal activation, the purpose of this study is to enroll as many drugs with high potential as possible. It would be ideal for these hypotheses to be validated with further additional experimentations. However, the scope of this study does not encompass validation experiments and will take into consideration for future studies.

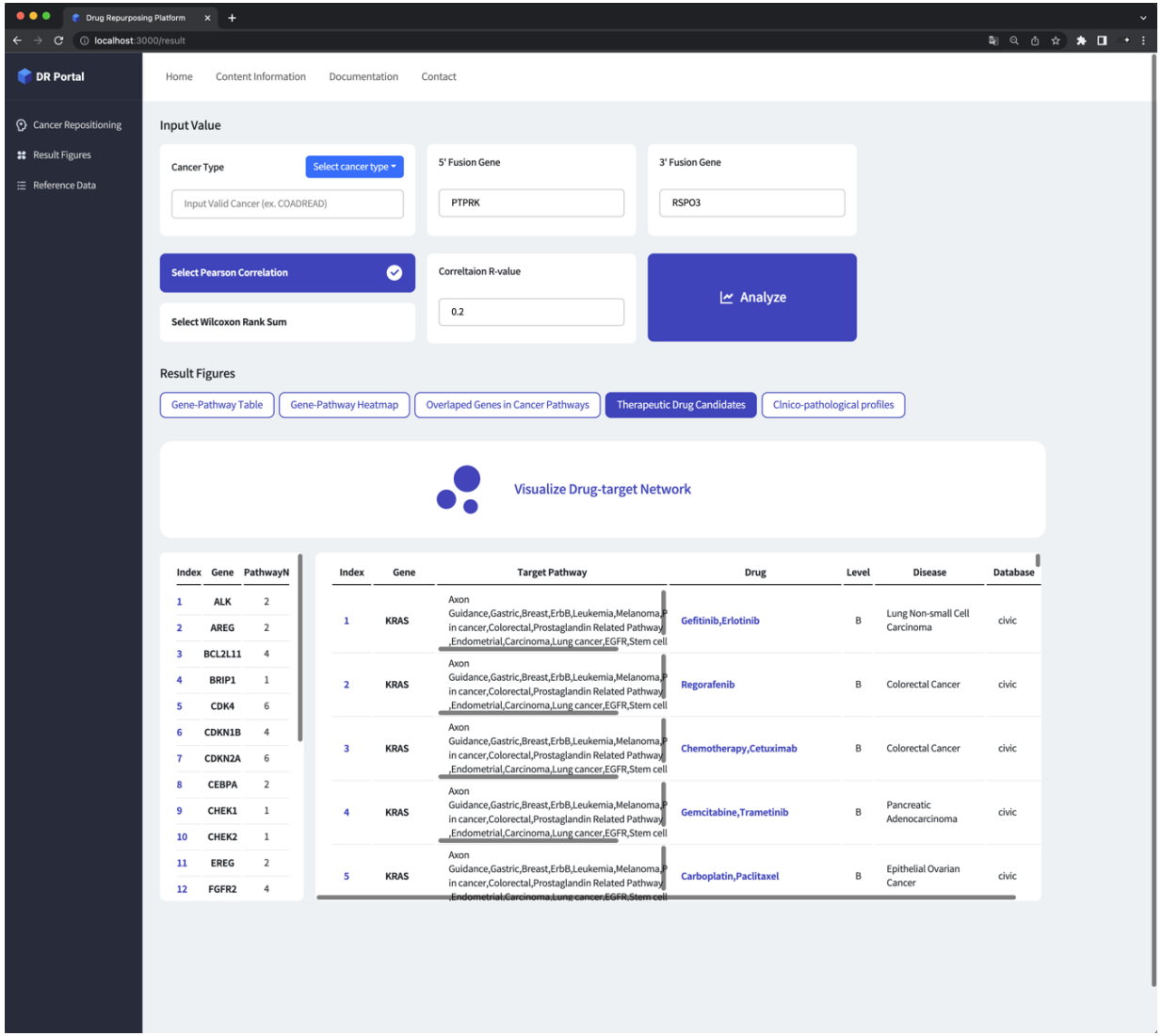


Figure 8. DRPORTAL output page listing drug-gene-pathway table of fusion-positive cancer

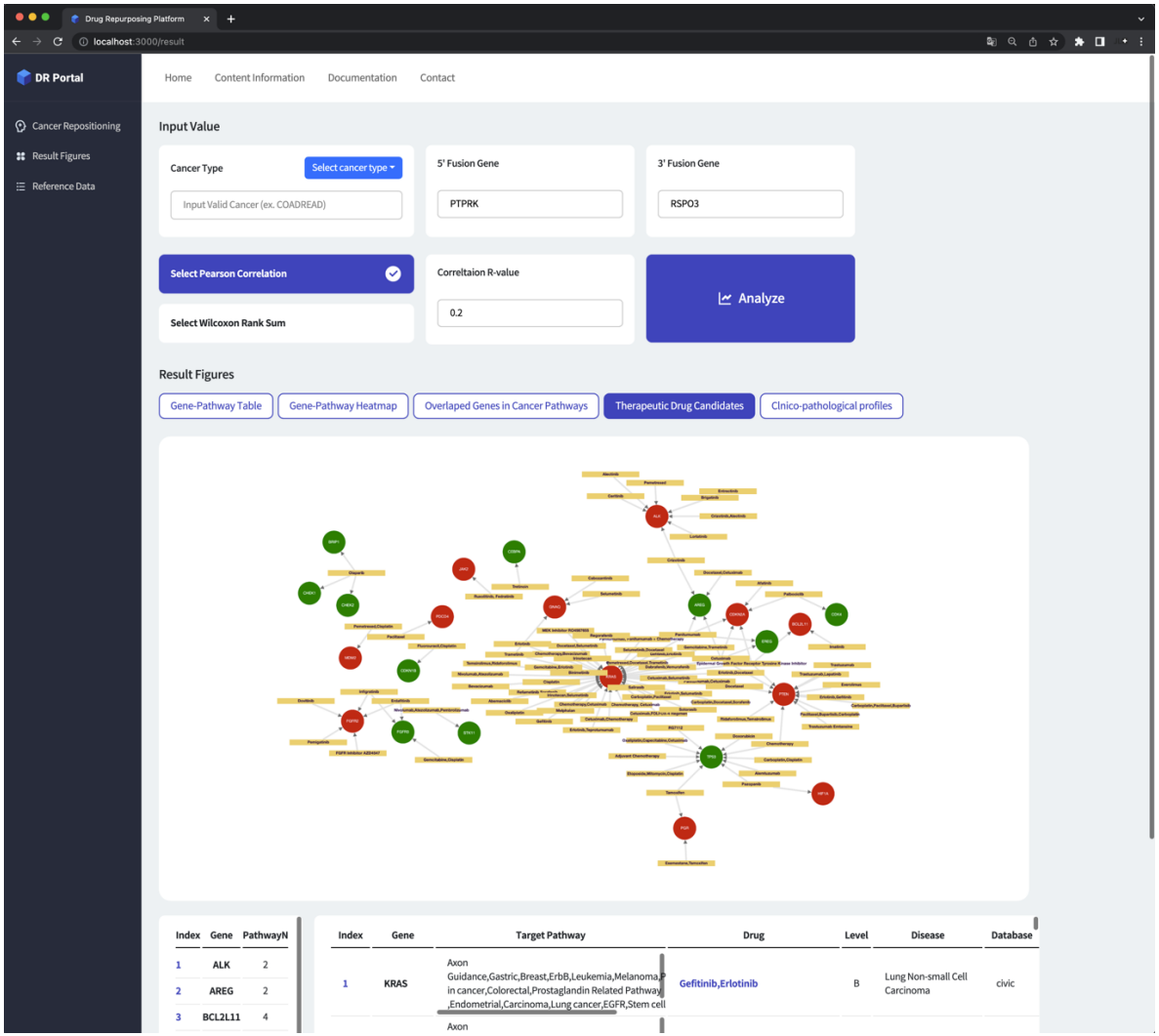


Figure 9. DRPORTAL output page visualizing drug-target network of fusion-positive cancer

3.3.4 Discussion

There are numerous web portals for drug repurposing in cancer previously. Drug Repurposing Hub (http://www.broadinstitute.org/repurposing) has manually curated collection of 4,707 experimentally validated drugs. The collection includes 3,422 drugs that are marketed worldwide or tested in human clinical trials. DrugSig (https://biotechlab.fudan.edu.cn/database/drugsig) is a drug response gene signatures database containing more than 1,300 drugs, 7,000 microarray and 800 targets. RepurposeDB (http://repurpsedb.dudelylab.org) is a collection of 253 repurposed drugs, drug target and diseases, which was assembled, indexed, and annotated from public data. DrugCentral (http://drugcentral.org) integrates structure, bioactivity, regulatory, pharmacologic actions, and indications for around 4,444 active drugs approved by regulatory agencies. KsRepo (http://github.com/adam-sam-brown/ksRepo) interrogates any case/control disease study expression profile. Researchers can use any pair of disease expression dataset and compound exposure database with the constraint that they are mappable to a single, common identifier system. DeSigN (http://design.cancerresearch.my) can be used to identify drugs with unknown efficacy against cancer cell lines. It consists of a set of differentially expressed genes (DEG) signatures, a pattern-matching algorithm and reference database. RE:fine drugs (http://drug repurposing.nationwidechildrens.org/search) is an interactive website for search and discovery of drug repurposing candidates from GWAS and PheWAS repurposing datasets.

However, previous drug repurposing portals generally provide database related to pharmacologic indications and drug response gene signatures. And some of them have tried to discover prognosis-related biomarkers using differentially expressed genes (DEGs) analysis. To our best knowledge, our study differs from previous strategies in two respects. First, the purpose of this study is to discover novel targets and therapeutics related to original mutations by analyzing downstream pathways and genes affected by target mutations that cannot be directly targeted. Second, our study is based on a structural variation (fusion by DNA structural variation) that is a driver mutation in fusion positive cancer. As consequence, almost all genes correlated with gene fusion are downstream-level genes affected by fusion. In this aspect, our study is different from other studies, and, for example, it is not clear whether prognosis-related biomarkers found by previous portals are primary driver or is affected by other drivers in those approaches.

In addition, even though a large number of fusion genes have been identified to have oncogenic function and successfully developed as therapeutic targets, most of them confronts obstacles due to the absence of high throughput analyzation tool. Although only a limited number of them have been studied in previous research, over 9,000 fusion genes lack functional insights and druggable targets. Thus, there remains a need for fast and reliable bioinformatic web-resource allowing cancer researcher to examine biological backgrounds of fusion positive cancer.

By providing an interactive web-resource, all researchers can investigate the landscape of molecular signaling and curate potential therapeutic agents for any fusion positive cancer regardless of their bioinformatics abilities. DRPORTAL can suggest three main utilities to cancer

researchers, clinicians and pharmaceutical company researchers

First, for fusion genes that cannot be directly targeted, DRPORTAL can search for new druggable targets and applicable therapeutic agents. In addition, it can also help to expand molecular pathological insights into fusion gene by identifying genes and cell signaling affected by the fusion gene. Besides, for fusion genes that targeted drugs already exist, DRPORTAL can suggest other drug candidates that can act synergistically in combination.

DRPORTAL also allows researchers to easily conduct analysis and customize various parameters, such as the statistical techniques used and the cut-off values. Also, the ability to visualize the results interactively through graphical figures is also useful, as it can help researchers better understand and interpret the results.

Finally, since all processes and database linkage are automated, even if there is an update to the gene expression profile, fusion gene, signaling pathways, and drug-target database, researchers can continue the analysis without any extra effort. It is certainly true that automating processes and linking them to databases can make research more efficient, as it can eliminate the need for manual data entry and reduce the potential for errors. Automation can also help researchers save time, as they can focus on other aspects of their work rather than having to spend time on data management tasks.

In summary, DRPORTAL will be able to maximize the speed of drug repurposing research on fusion positive cancer for clinicians, biologists, and pharmaceutical company researchers.

Since DRPORTAL identifies targets and drug candidates within in silico

level, further experimental validation is needed. However, in vitro experiments were previously performed on computationally identified targets, and confirmed the proliferation of fusion positive cancer cell was successfully reduced by inhibiting these targets.

Rationally designed targeted therapies are likely to have solid scientific basis [79]. This can be supported by the large number of molecular targeted drugs that have already been approved and successfully used. (A drug targeting BCR-ABL in chronic myeloid leukemia [80], c-kit in gastrointestinal stromal tumors [81], EGFR in non- small-cell lung cancer [82], monoclonal antibody in HER2 positive breast cancer [83]). However, molecular-targeted drugs have little effect on symptom control and survival due to genetic heterogeneous of tumors in clinical areas. In particular, it is reported that the effect on progression-free survival of common solid tumors usually lasts only for a few weeks to a few months [84, 85].

In summary, we believe that DRPORTAL, which has novel therapeutic strategy specific to fusion-positive cancer, can greatly aid cancer biologists and clinicians not only with identifying novel diagnostic and therapeutic targets but also investigating the mechanisms of fusion gene by analyzing various molecular-pathological characteristics. We hope that our findings will be the steppingstone for future investigations, leading to the promotion of a targeted cancer therapy.

Chapter 4 Identification of new therapeutic targets and applicable drug candidates in ESR1-CCDC170 fusion positive breast cancer

4.1 Introduction

Breast cancer, aside from skin cancer, is the most commonly diagnosed cancer in women worldwide [86]. Recent statistics report the emergence of 250,000 new cases of breast cancer solely in 2017 contributing to the 12% of women diagnosed with breast cancer in the United States [87]. Molecular classification divides breast cancer into four major classes: luminal A, luminal B, and human epidermal growth factor receptor 2 (HER2)-enriched (HER2-E), and basal-like subtype [87]. Among them, luminal B remains to be the most common subtype in young women, accounting for 15-20% of total breast cancer cases, and within luminal B, ESR1-CCDC170 fusion positive subtype, constituting 6 to 8% of the luminal B class, persists to be the most dominant subtype [88-95].

ESR1-CCDC170 fusion causing chimeric mRNA is known to be formed by a tandem duplication at 6q25.1 location on coiled-coil domain containing 170 (CCDC170) adjacent to ESR1 gene [89, 96]. It has been reported that polymorphism of CCDC170 gene correlates with breast cancer susceptibility [97, 98]. ESR1-CCDC170 fusion-positive patients undergoing ER-positive (ER+) breast cancer endocrine therapy have demonstrated reduced treatment efficiency and growth of aggressive ER+ breast cancer [99]. Although its effect has been studied in relation to ovarian cancer, the molecular signaling

involved in the induction of ESR1-CCDC170 fusion-positive breast cancer has yet to be elucidated [100].

Herein, we systematically analyze the molecular pathological features of ESR1-CCDC170 fusion-positive breast cancer through the data analysis of TCGA and identified the activated oncogenic pathways. In addition, putative target genes and actionable drugs were inferred and prioritized by performing network analysis using both transcriptomic signatures and the drug-target databases such as OncoKB and CIViC.

4.2 Materials and Methods

4.2.1 Sample acquisition and quality control

Gene level 3 (RSEM) mRNA expression with normalized read count values of the Cancer Genome Atlas (TCGA) breast cancer carcinoma (BRCA) was obtained from the Broad GDAC Firehouse website (https://gdac.broadinstitute.org). Related clinical features data including information about the samples' MAF files, molecular subtypes, and TNM stages were obtained from the website mentioned above.

4.2.2 Case-Control selection

Previous study confirmed 319 fusion genes in TCGA clinical breast cancer tumors [101]. Unlike other in-frame fusion genes, ESR1-CCDC170 is known as breast cancer-specific oncogenic fusion gene. Using the TCGA fusion gene data portal (The Jackson Laboratory,

https://www.tumorfusions.org), we identified eleven samples of CCDC170 fusion, which were cross-checked with increased CCDC170 expression level. Furthermore, only tumor samples that have the barcode 01A were selectively chosen by disregarding other types of tumor samples, 11A (Normal) or 06A (Metastasized). Among the remaining samples, 50 samples with the highest expression of CCDC170 were confirmed as up-regulated controls for analyzing the network within the non-coding region of the fusion gene. From the upregulated control samples, 2 outlier samples were filtered out using the IQR (Inter Quartile Range) method. The same number of controls (n=48) with the lowest expression of CCDC170 were then selected from the remaining samples.

4.2.3 Selection of genes affected by E:C Fusion

RNA expression data from TCGA were made into two-dimensional matrix comprised of the selected 11 fusion samples and 48 control samples. Each column represents the patient ID while each row represents the gene name. Based on the RNA expression matrix, variance tests were conducted using independent two-sample t-tests. To select genes in coordination with CCDC170 in RNA expression, t-tests were performed between E:C fusion-positive and fusion-negative cases. Mostly affected 1,000 genes were selected (adjust p-value < 2.0 E-08).

4.2.4 Pathway analysis via ConsensusPathDB (CPDB) and Over-Representation

The selected 1,000 genes that correlate to the reference gene (CCDC170) from the aforementioned RNA expression data were used to perform over-representation analysis (ORA) via ConsensusPathDB (CPDB, http://cpdb.molgen.mpg.de/CPDB) using recent protocols. 113 biological pathways were merged from the following sources, according to data from BioCarta (http://www.biocarta.com), INOH [102], KEGG [103], NetPath [104], PID [105], Reactome [106] and Wikipathways [107]. In consideration of the ontological characteristics and the proportion of duplicated genes, the pathways, which were enriched with selected 1,000 genes (q-value < 0.05), were condensed into 15 cancer-related pathways, and their component were 184 genes.

4.2.5 Druggable pathway analysis via CIViC and OncoKB

The “Clinical Evidence Summaries” data, released on October 1, 2017, was downloaded from the Clinical Interpretations of Variants in Cancer (CIViC) website (https://civic.genome.wustl.edu/releases), and the “Actionable Variants” data was accessed and downloaded on October 17, 2017 from the OncoKB website (http://oncokb.org/). 673 CIViC variants (181 genes) with expected therapy efficacy 148 OncoKB actionable variants (53 genes) were integrated. 113 CCDC170-correlated genes were matched to the CIViC and OncoKB variants.

4.2.6 Statistical analysis and data visualization

Open software R version 3.4.3 was used to process all statistical analysis for selecting genes correlated to CCDC170 including the variance test and independent two-sample t-test. RNA expression heatmap was also visualized using Complexheatmap, a package for R. KEGG mapper (https://www.genome.jp/kegg/mapper.html) was used to visualize target pathways related to DNA damage response. Cytoscape version 3.5.3 was used to analyze and express the complex network between targetable drugs and therapeutic agents. Our study defined statistical significance with p-value of < 0.05 and false detection rate (FDR) with q-value of <0.001.

4.3 Results

4.3.1 Clinicopathological characteristics

We checked the clinicopathological characteristics of 11 ESR1-CCDC170 fusion-positive and 48 fusion negative patients amongst 1,093 breast cancer patients in Broad GDAC Firehose (Figure 10, Table1). Two significant differences were identified between fusion-positive and negative patients. First, CCDC170-fusion-positive patients had a high rate of ER-positive (90.0%) and PR positive (60.0%), whereas fusion-negative patients displayed significantly lower rates of 15.9% and 4.7%, respectively ($p < 0.05$). Additionally, HER2 immunohistochemistry (IHC) results showed significantly higher rate of 3+ for fusion-positive patients than for patients with fusion-negative (44.4% vs 6.5%, $p < 0.05$, Table 1). According to the findings above,

CCDC170-fusion-positive BRCA appears to closely resemble characteristics typical of triple-positive breast cancer in this cohort.

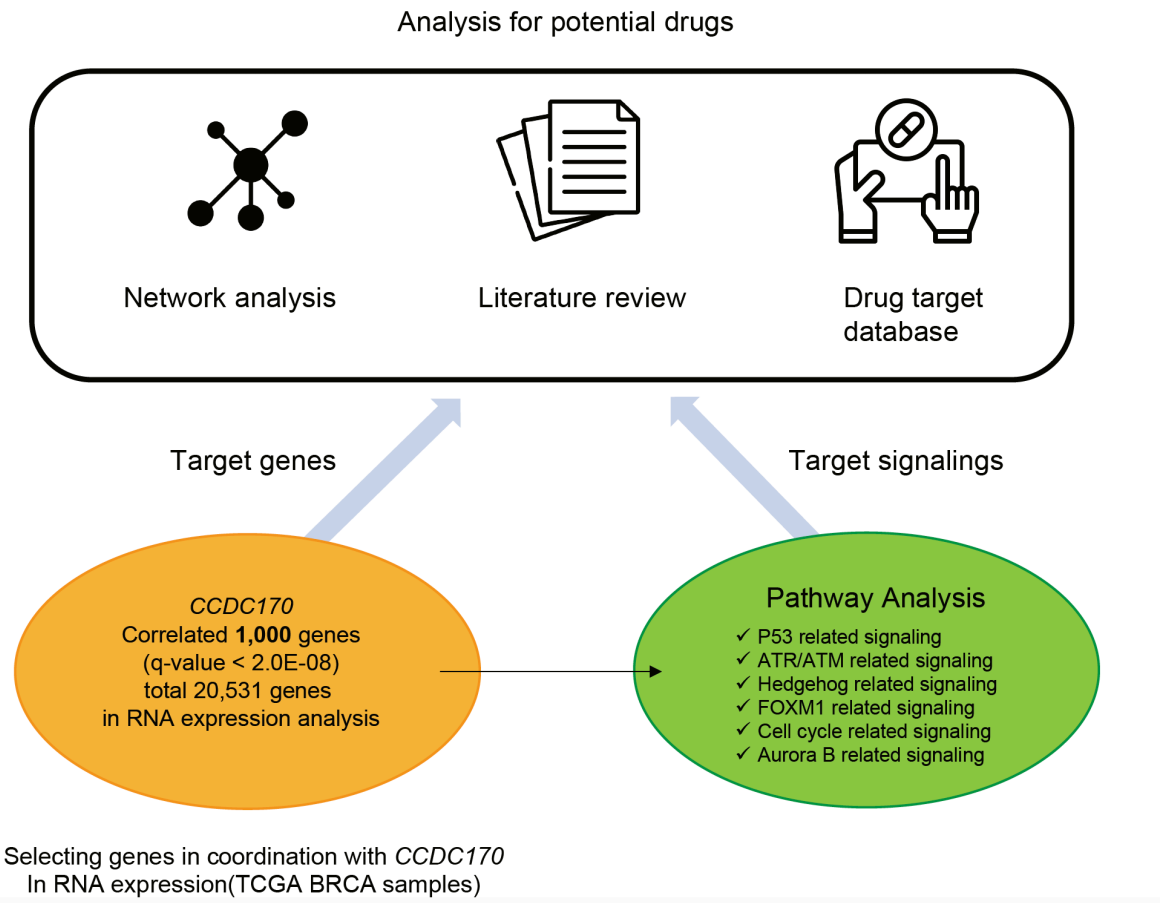


Figure 10. Overall schematics. Transcriptome data for breast cancer (BRCA) was obtained from the Broad GDAC Firehose database. Following the RNA measurement analysis of a total of 20,531 genes, 1,000 genes correlated with CCDC170 were selected. (q< 2.0 E-08). Over-representation analysis of the 1,000 genes demonstrated significant relation to six major cancer-related pathways (p53, ATR/ARM, hedgehog, FOXM1, cell cycle, Aurora B). Potential gene targets and drug candidates were isolated via drug-network analysis using a drug-target database on genes correlated to CCDC170 and literature review.

Table 1. Comparisons in clinical and pathological characteristics of ESR1-CCDC170 fusion positive, negative BRCA patients and control cohorts. The clinical and pathological characteristics between ESR1-CCDC170 fusion positive, negative BRCA, and control cohorts were compared.

	Total (N=929)	**control (N=44)**	**fusion (N=10)**	**p values**
age	48~70	44~68	49~72	NS
sex				
- female	929 (100.0%)	44 (100.0%)	10 (100.0%)	
Vital status				1
- alive	837 (90.1%)	38 (86.4%)	9 (90.0%)	
- dead	92 (9.9%)	6 (13.6%)	1 (10.0%)	
Stage				NS
- stage I	84 (9.1%)	1 (2.3%)	1 (10.0%)	
- stage Ia	68 (7.4%)	3 (6.8%)	1 (10.0%)	
- stage Ib	5 (0.5%)	0 (0.0%)	0 (0.0%)	
- stage II	4 (0.4%)	0 (0.0%)	0 (0.0%)	
- stage IIa	311 (33.7%)	23 (52.3%)	2 (20.0%)	
- stage IIb	215 (23.3%)	8 (18.2%)	3 (30.0%)	
- stage IIIa	132 (14.3%)	4 (9.1%)	2 (20.0%)	
- stage IIIb	24 (2.6%)	0 (0.0%)	0 (0.0%)	
- stage IIIc	52 (5.6%)	3 (6.8%)	1 (10.0%)	
- stage Iv	14 (1.5%)	1 (2.3%)	0 (0.0%)	
- stage x	12 (1.3%)	1 (2.3%)	0 (0.0%)	
ER status				0
- positive	685 (77.2%)	7 (15.9%)	9 (90.0%)	
- negative	200 (22.5%)	37 (84.1%)	0 (0.0%)	
- indeterminate	2 (0.2%)	0 (0.0%)	1 (10.0%)	
PR status				0
- positive	594 (67.0%)	2 (4.7%)	6 (60.0%)	

- negative	288 (32.5%)	41 (95.3%)	4 (40.0%)	
- indeterminate	4 (0.5%)	0 (0.0%)	0 (0.0%)	
HER2 IHC				0.013
0	57 (10.6%)	8 (25.8%)	0 (0.0%)	
- 1+	234 (43.4%)	11 (35.5%)	3 (30.0%)	
- 2+	171 (31.7%)	10 (31.2%)	1 (10.0%)	
- 3+	77 (14.3%)	2 (6.5%)	4 (40.0%)	
NA	0 (0.0%)	0 (0.0%)	2 (20.0%)	
subtype				0
- Basal	168 (18.1%)	40 (90.9%)	0 (0.0%)	
- Her2	74 (8.0%)	4 (9.1%)	1 (9.1%)	
- LumA	493 (53.1%)	0 (0.0%)	4 (40.0%)	
- LumB	194 (20.9%)	0 (0.0%)	5 (45.5%)	
PIK3CA mutation	293 (31.5.5%)	3 (6.8%)	1 (10.0%)	NS
CDH1 mutation	91 (9.8%)	0 (0.0%)	1 (10.0%)	NS
TP53 mutation	266 (28.6%)	29 (65.9%)	3 (30.0%)	NS
BRCA1 mutation	11 (1.2%)	0 (0.0%)	0 (0.0%)	NA
BRCA2 mutation	12 (1.3%)	1 (2.7%)	1 (10.0%)	NS

Secondly, the pathological subtype of fusion-positive patients was found to have a high proportion of Luminal A (40.0%) and Luminal B subtype (45.5%), while 90.9% of the 48 cases with the lowest CCDC170 expression were found to be basal-type ($p < 0.05$, Table 1). Taken together, the CCDC170 fusion-positive BRCA showed a mutually exclusive relationship with the basal-type breast cancer cells.

On the other hand, no significant differences in the age, sex, vital status, and TNM stage was observed between the two groups. In addition, the five gene variants (PIK3CA, CDH1, TP53, and BRCA1/2) frequently found in BRCA showed no significant difference between the two groups. Based on these results, CCDC170 fusion positive BRCA patients have distinct pathological characteristics in terms of tumor subtype and triple positive tendency.

4.3.2 Key pathways and genes altered in E:C fusion-positive breast cancer

One thousand genes were obtained by an independent t-test (Q < 2.0E-08) and inputted for performing the Over Representation Analysis (ORA) of the ConsensusPathDB website to select cancer-related pathways. As a result, a total of six cancer-related pathways (p53, ATR/ATM, FOXM1, Hedgehog, Cell cycle, Aurora B related signaling pathways) were discerned.

In the six major cancer-related pathways, 137 genes were significantly over- or under-expressed in the CCDC170 fusion-positive cases compared to the CCDC170 fusion-negative controls. (Figure 11, Supplement 1).

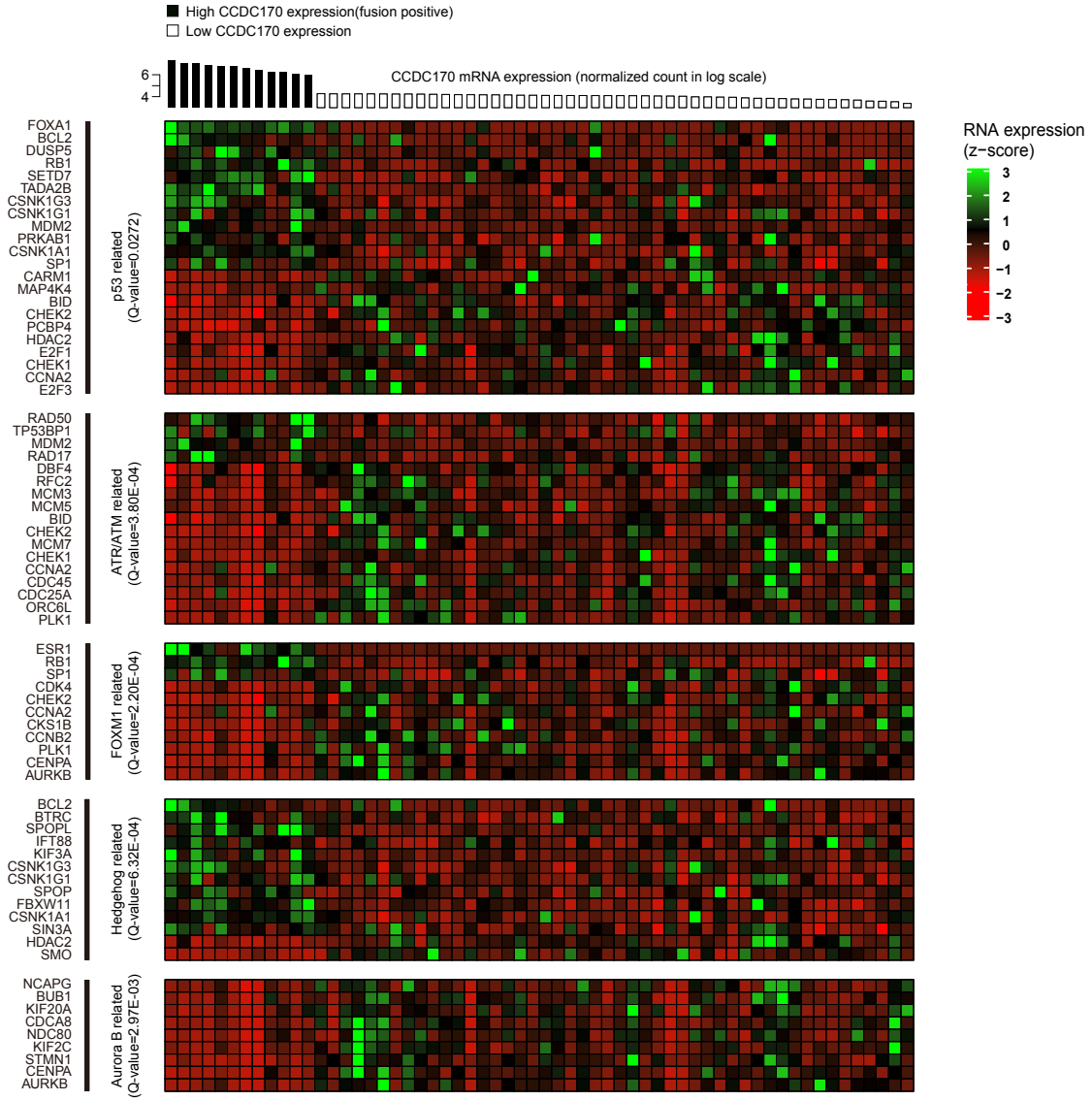


Figure 11. Gene expression heatmap of cancer-related pathways correlated with CCDC170 RNA expression. Of the analyzed genes, 72 of genes associated with P53, ATR/ATM, FOXM1, hedgehog, and aurora demonstrated significant differences in expression in CCDC170 fusion-positive BRCA samples when compared to the control group. Over-representation analysis using CPDB yielded statistically significant pathways related to cancer ($q < 0.05$). The x-axis is indicative of the sample, while the y-axis is indicative of its respective RNA expression. The RNA expression was converted into z-score prior to representation on the heatmap.

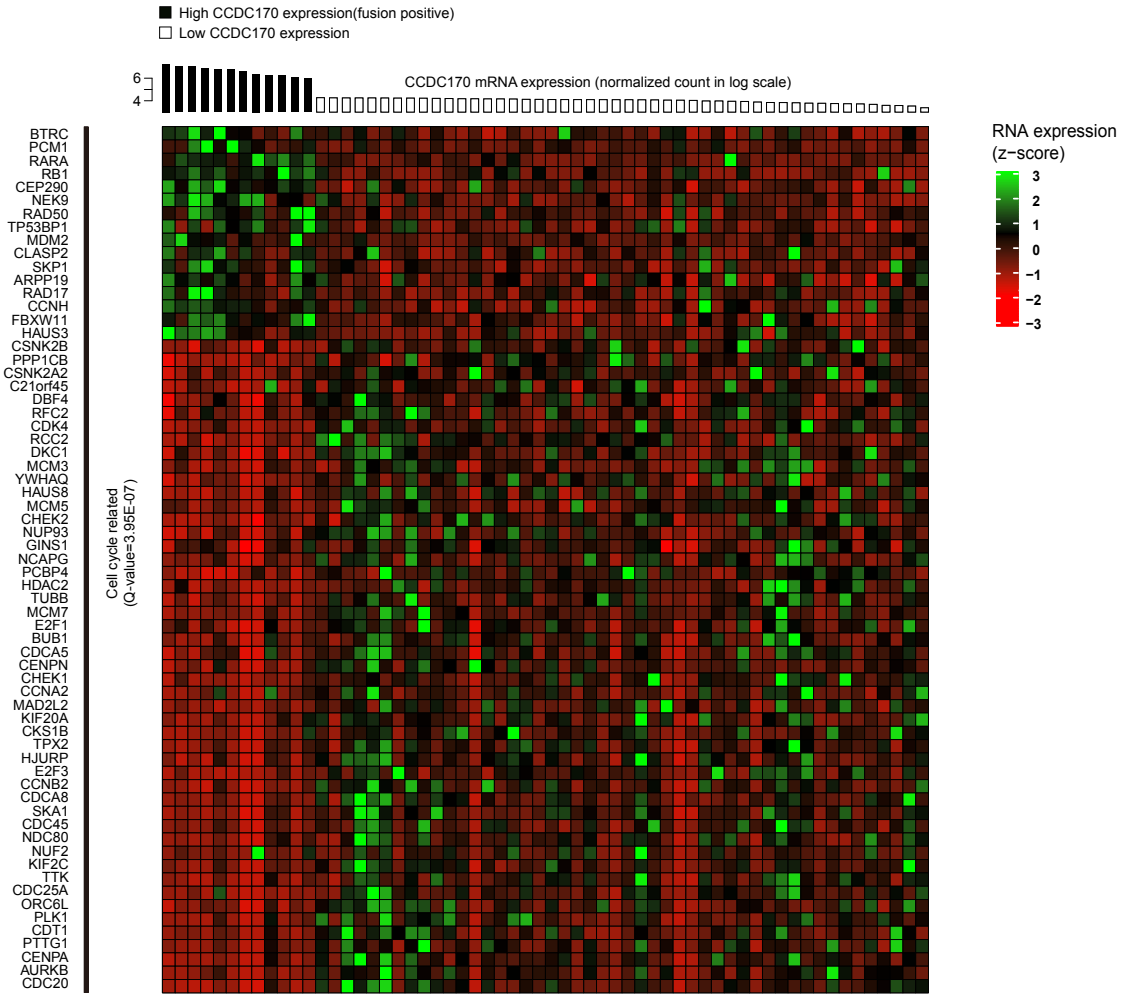


Figure S1: Heatmap of cell cycle-related and other miscellaneous genes with altered expression correlating to CCDC170 expression.

Of the six pathways, two concerning p53 and ATR/ATM-related signaling pathway were associated with DNA damage response. Mapping with the KEGG pathway revealed 22 genes that are involved in the p53-related pathway and 17 genes, the ATR/ATM-related pathway. Both pathways are highly relevant to the promotion and maintenance of the cell cycle (Figure 12). Genes with multiple hits of more than 2 that coincide for both, p53 and ATR/ATM-related signaling pathways, are CCNA2(CycA), MDM2, CHEK1(Chk1), CHEK2(Chk2).

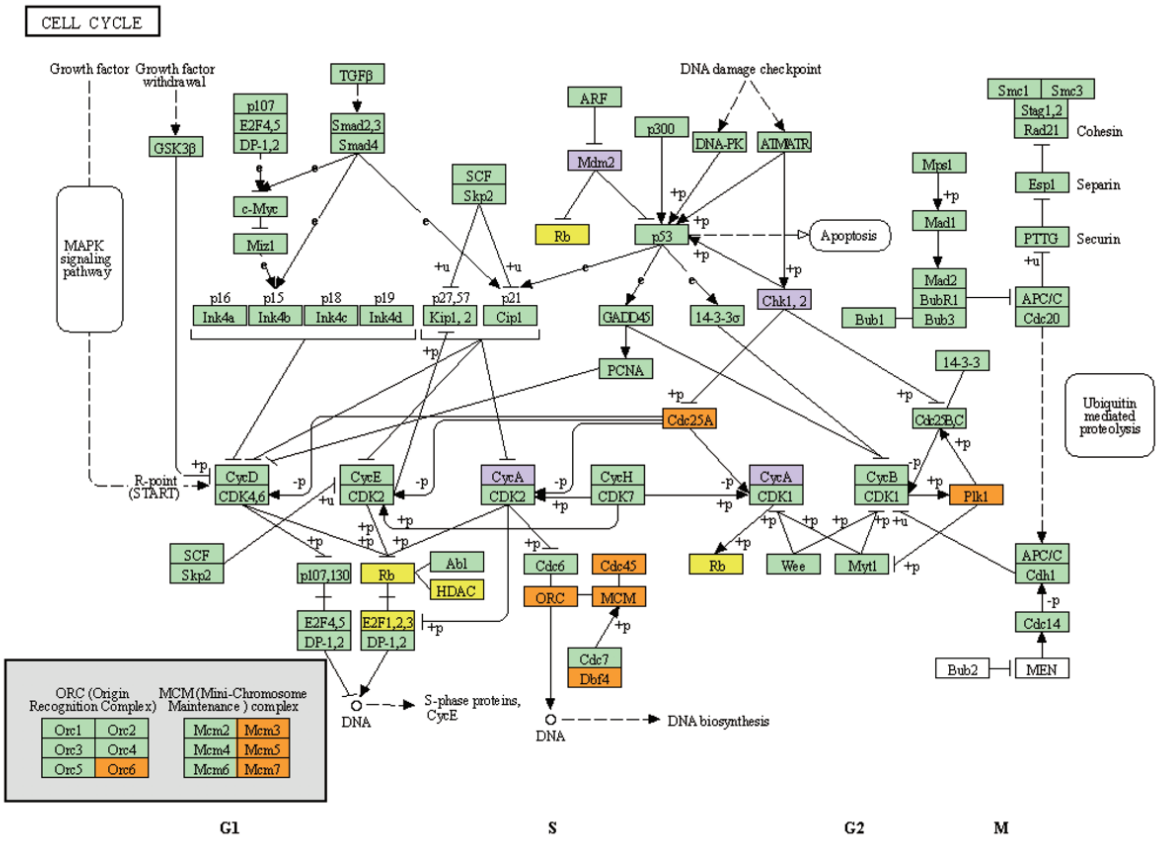


Figure 12. Over- and under-expressed genes are enriched in human cell cycle pathway. The KEGG pathway map for the human cell cycle signaling pathway, has04110, was visualized using the KEGG Mapper. Among the pathways, p53 and ATR/ATM shared significant correlation with the identified genes; genes associated with p53 signaling pathway are boxed in yellow, ATR/ATM, in orange, and common denominators for both pathways, in purple.

For genes with hits of more than 3 with implications of its role in multiple pathways, 39 genes were identified (Figure 13). Of which, AURKB, HDAC2, PLK1, CENPA, CHEK1, CHEK2, RB1, and MDM2 were included in at least three pathways that are important for tumor proliferation and maintenance specific to ESR1-CCDC170 fusion positive BRCA patients.

Further investigation of the 48 samples of highest mRNA levels of CCDC170 with the Differentially Expressed Gene (DEG) analysis showed similar patterns as the TCGA data obtained above in fusion-positive samples when compared to the control samples (Supplementary Figure 2). This suggests that similar cell signaling is activated not just with fusion but other possibilities in CCDC170 overexpression.

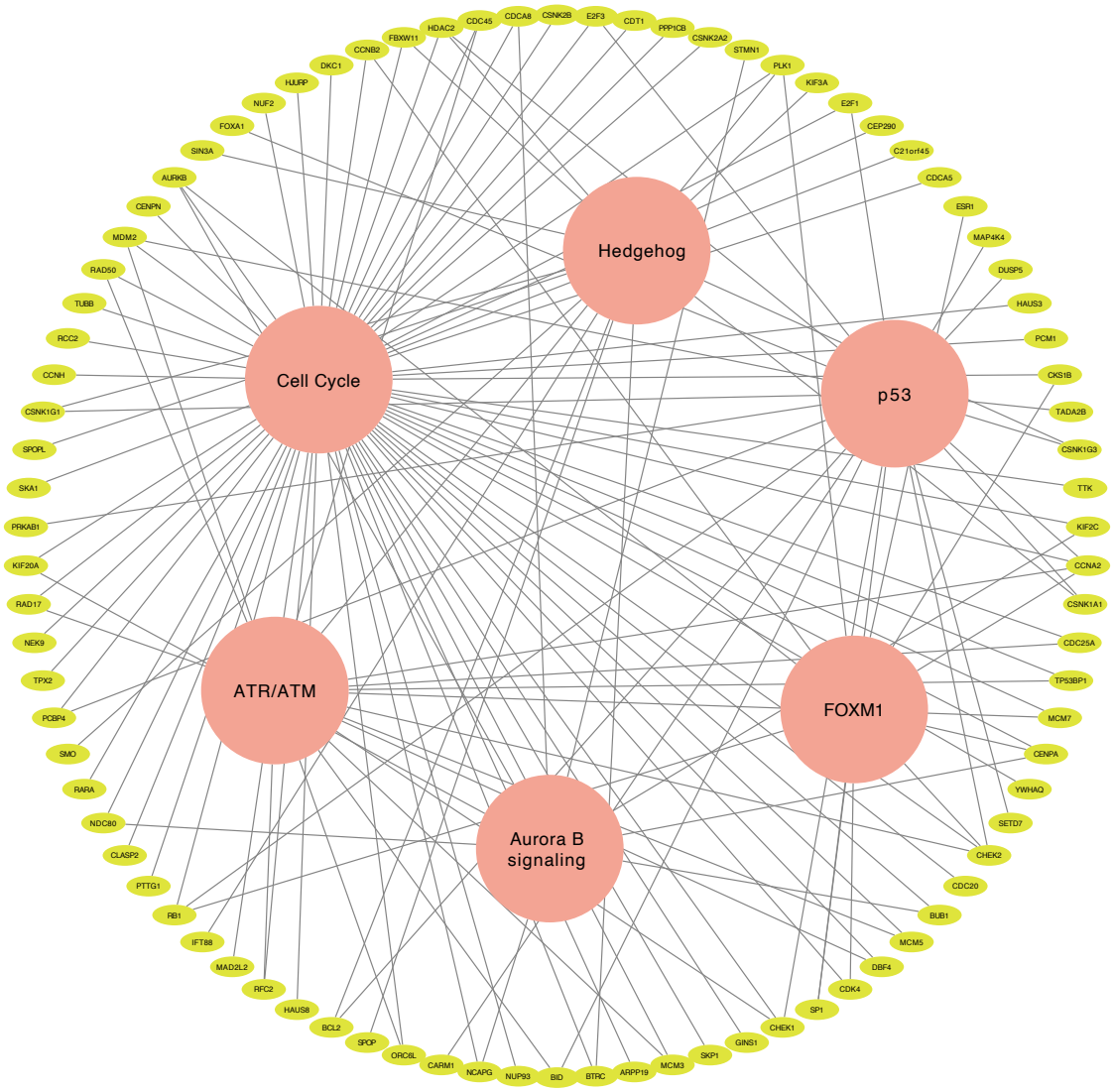


Figure 13. Putative target genes involved in multiple pathways of ESR1-CCDC170 fusion-positive cancer. 6 major cancer signaling pathways associated with p53, ATR/ATM, FOXM1, hedgehog, cell cycle and aurora B in accordance with its respective genes were visualized. Potential gene candidates involved in these pathways were discerned.

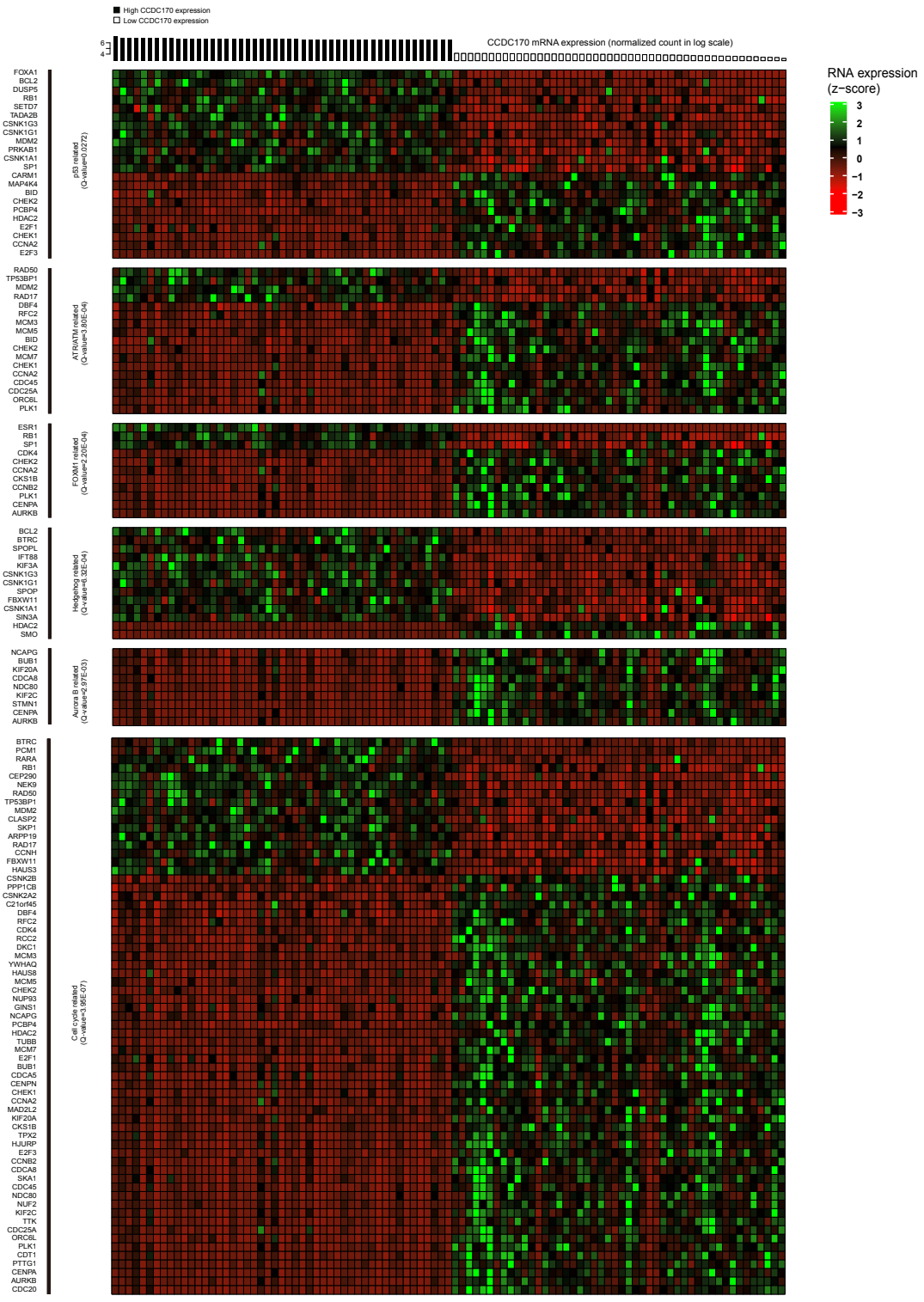


Figure S2: Heatmap of cancer-related and other miscellaneous genes with altered expression correlating to CCDC170 expression in DEG analysis

4.3.3 Identification of actionable targets and potential therapeutic choice using network analysis

Actionable target genes and potentially available drugs were extracted by inputting the 137 genes in the following drug databases, CIViC (n=673) and OncoKB (n=262). ESR1, CDK4, RAD50, CHEK1, MDM2, and SMO were mapped as targetable genes. Results indicated the following drug-target relationship: letrozole, palbociclib, fulvestrant, AZD9496, tamoxifen for ESR1; palbociclib, alpelisib, ribociclib, dexamethasone for CDK4; checkpoint kinase inhibitor AZD7762, irinotecan for RAD50; cisplatin, prexasertiv, olaparib for CHEK1; milademetan tosylate, RO5045337 for MDM2; PSI, vismodegib, patidegib, arsenic trioxide for SMO (Figure 14).

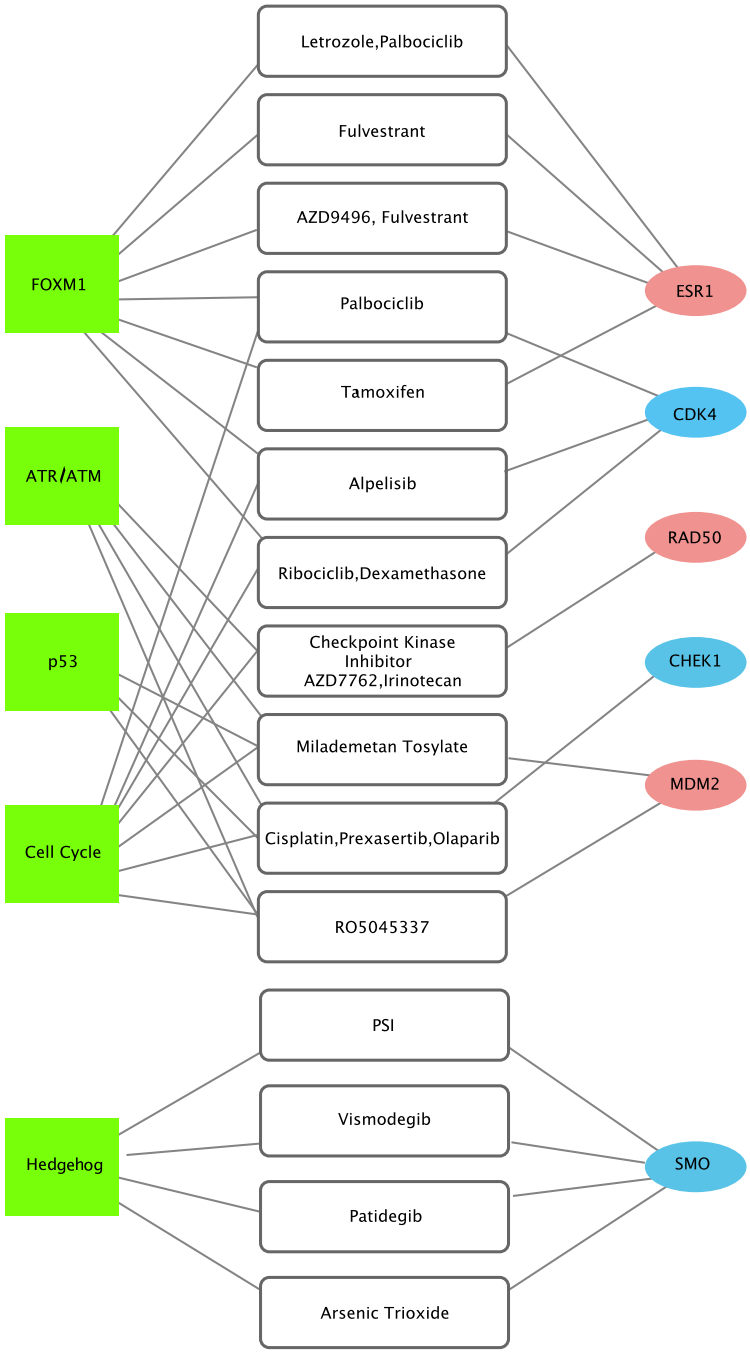


Figure 14. Drug-target network of ESR1-CCDC170 fusion-positive BRCA cancer. Network visualization was demonstrated with Cytoscape, and drug-target relation was identified with Civic and OncoKB. Green boxes are representative of pathways, white boxes, of drugs, and oval boxes, of genes. Red, oval boxes are genes that are over-expressed in fusion-positive cancer whereas blue, oval boxes are genes that are under-expressed in fusion-positive cancer.

Observing the druggable target genes associated with the main pathways of E:C fusion-positive BRCA, ESR1, CDK4 genes were included in the FOXM1-related signaling pathway, RAD50 in the ATR/ATM-related signaling pathway, CHEK1, MDM2 genes in the P53-related signaling pathway, CDK4, RAD50, and MDM2 genes in the cell cycle-related signaling pathway, and SMO genes in the hedgehog-related signaling pathway. Interestingly, four of the six targetable genes, CDK4, RAD50, CHEK1, and MDM2 were involved in two or more major cancer-related pathways. In the case of MDM2, three of the six pathways associated with E:C fusion-positive were identified to be involved.

4.3.4 Discussion

In this study, characteristics concerning ER-positive molecular subtype in *CCDC170*-subtype breast cancer was identified, and genes specifically regulated in E:C BRCA were identified and screened for BRCA-cancer related signaling pathways. Also, information regarding optimal treatment targets and drugs for targeted therapy were provided. E:C fusion-positive BRCA requires a new therapeutic approach to overcome its relatively low response to hormone therapy [91, 99, 108-110].

A recent study on the potential targeted therapy for E:C BRCA performed by Li et al. was met with limitations with regards to a restricted number of cell line samples and proteins [99]. Our study has addressed this issue by performing analysis on a sufficient number of case-control using

TCGA human cancer samples and systematically testing the DEGs using more than 20,000 genes and cancer specific pathways. Finally, we were able to propose a number of potential drugs with promising therapeutic effects.

The common early treatment options for breast cancer are generally divided into conventional chemotherapy (i.e., Adriamycin, cyclophosphamide, paclitaxel, docetaxel), endocrine therapy (tamoxifen, letrozole, anastrozole, exmenestane), ERBB-targeted therapy (trastuzumab, pertuzumab), and combination treatment methods according to the pathological and molecular classification of breast cancer [87]. In case of metastatic breast cancer, CDK4/6 inhibitor and PARP inhibitor are considered to be additional options [87].

Among repositioned drugs inferred in our study, CDK4/6 inhibitor (Palbociclib), cisplatin, and PARP inhibitor are the drugs used as standard treatments for breast cancer patients with or without metastasis. On the other hand, AZD9496, alpelisib, dexamethasone, checkpoint kinase inhibitor AZD7762, irinotecan, cisplatin, prexasertiv, milademtan tosylate, R05045337, PSI, vismodegib, patidegib, and arsenic trioxide are seen as putative actionable drugs that can be used for E:C fusion positive BRCA proceeding in-vitro and in-vivo validation.

AURKB, HDAC2, PLK1, CENPA, CHEK1, CHEK2, RB1, and *MDM2* genes, which were included in at least three pathways, are expected to play an important role in the promotion and maintenance of *CCDC170* subtype breast cancer [111]. For instance, *PLK1* may act as a tumor suppressor gene that regulates estrogen receptor (ER)-regulated gene transcription in breast cancer; *RB1* gene, also a tumor suppressor gene, however, frequently

lost in triple-negative breast cancer [112]; *CENPA* is a significant prognostic marker for ER-positive patient [113]; *HDAC2* and *CHEK2* genes have been significantly correlated to *CCDC170* fusion-subtype and have been reported to be associated with DDR functioning [114, 115], which is also suggestive of *CCDC170* fusion subtype's relation to DDR.

In summary, this study presents core biomarkers and potentially actionable drugs specific to E:C fusion-positive breast cancer. Via in-vitro experimentation, these candidates were confirmed to be strongly associated with this type of cancer and their roles were verified by discerning its associated signaling pathways. We hope that our findings will be the steppingstone for future investigations leading to the promotion of targeted cancer therapy.

4.4 Conclusion

Amongst the various types of breast cancer, luminal B subtype is the most common in young women, and *ESR1-CCDC170* (E:C) fusion is the most frequent oncogenic fusion driver of the luminal B subtype. Nevertheless, treatments targeting E:C fusion has not been well established yet. Hence, the aim of this study is to investigate for potential therapies targeting E:C fusion based on systematic bioinformatical analysis of The Cancer Genome Atlas (TCGA) data. 1,000 genes related were extracted using transcriptome analysis, and major signaling pathways associated with breast cancer were identified with over-representation analysis. Then, we conducted drug-target network analysis based on OncoKB and CIViC database, and finally selected potentially applicable drug candidates. Six major cancer-related signaling pathways (p53, ATR/ATM, FOXM1, hedgehog, cell cycle, and Aurora B) were significantly altered in E:C fusion positive cases of breast cancer. Further investigation revealed that eight genes (*AURKB, HDAC2, PLK1, CENPA, CHEK1, CHEK2, RB1,* and *MDM2*) in coordination with E:C fusion were found to be common denominators for three or more of these pathways, thereby making them promising gene biomarkers for target therapy. Among the 21 putative actionable drugs inferred by drug-target network analysis, palbociclib, alpelisib, ribociclib, dexamethasone, checkpoint kinase inhibitor AXD 7762, irinotecan, milademtan tosylate, R05045337, cisplatin, prexasertib, and olaparib were considered as promising drug candidates targeting genes involved in at least two E:C fusion-related pathways.

Chapter 5 Investigation of cell signalings and therapeutic targets in PTPRK-RSPO3 fusion-positive colorectal cancer.

5.1 Introduction

Colorectal cancer (CRC) is the third most fatal and fourth most diagnosed cancer worldwide, according to 2018 global cancer data released by the IARC. Approximately 2 million new cases were recorded in year 2018 alone, resulting in approximately 1 million fatalities [86, 116, 117]. With the development of NGS technology, simultaneous detection of various mutations in colorectal cancer has become possible, including SNV, INDEL, CNV, fusion, and MSI [118-121]. The reason that the detection of these mutations is important is that it can be used as a target therapy for gene mutations, for example, EGFR-inhibitor for *KRAS* wildtype CRC and immune checkpoint inhibitors for MSI-high solid tumor [122].

The efforts to develop targeted drugs for treating colorectal cancer are increasing, however, the candidates of target drugs other signaling pathways besides EGFR and mismatch Repair are limited. WNT signaling is known as a major pathway in colorectal cancer and mostly is activated by mutation of the *APC* gene, which plays an important role in the pathogenesis of colorectal cancer [123-127]. Recently, *PTPRK-RSPO3* (P:R) fusion also contributes to the activation of WNT signaling and causes colorectal cancer, and this mutation is mutually exclusive with the *APC* mutation and is recognized as another important mutation contributing to the development of

colorectal cancer [128-131]. Recent studies have reported that LGK974 and *RSPO3* antibodies may be beneficial at in vitro and in vivo levels, however, the development of targeted therapeutics for colorectal cancer patients with P:R fusions is still in its infancy [132, 133].

Herein, we systematically inferred drug candidates in P:R fusion colorectal cancer. First, we extracted *RSPO3* expression correlated genes and selected oncogenic cell signal pathways containing those genes. Then, we constructed a drug-target network in P:R fusion colorectal cancer using the drug-target database, and finally, we prioritized a suitable therapeutic agent.

5.2 Materials and Methods

5.2.1 Sample collection and quality control

The Broad GDAC Firehouse website (https://gdac.broadinstitute.org) provided gene level 3 (RSEM) mRNA expression with normalized read count values of the Cancer Genome Atlas (TCGA) colorectal cancer (CRC). The above-mentioned website provided information on the samples' MAF files, TNM stages, and molecular subtypes among other clinical characteristics.

5.2.2 Case-Control selection and selection of genes affected by P:R Fusion

We found seven samples with *PTPRK-RSPO3* fusion using the TCGA fusion gene data portal (The Jackson Laboratory, https://www.tumorfusions.org), which were cross-checked with elevated *RSPO3* expression levels. Additionally, 372 tumor samples with the barcode 01A were chosen, with other types of tumor samples, 11A (Normal) or 06A(Metastasized), being excluded. For control sample selection, 50 samples were randomly selected among the samples with low *RSPO3* expression (less than the median value of *RSPO3* RNA expression, N=186). To obtain R-values of 20,531 genes in correlation with *RSPO3* in RNA expression, Pearson correlation-tests were performed in seven *PTPRK-RSPO3* fusion-positive cases and 50 controls. Then, above tests were repeated 100 times. Based on the median of absolute R values from 100 tests, 20,531 genes were sorted in decreasing order. Using the median of absolute R values, mostly affected 2,505 genes were selected by correlation cut-off ($R > 0.2$). The R cut-off

value, 0.2, were selected based on the 100,000 permutation tests. For each permutation test, randomly ordered expression values of a randomly selected gene were tested using Pearson correlation test with the expression values of reference gene (*RSPO3*) in 7 cases and 50 controls. After 100,000 tests, falsely selected genes correlated with *RSPO3* are assumed to be 0.37% when the cut-off value for R was 0.2.

5.2.3 Pathway analysis via ConsensusPathDB (CPDB)

The aforementioned R-value data were used to perform over-representation analysis (ORA) using ConsensusPathDB (CPDB, http://cpdb.molgen.mpg.de/CPDB) using recent protocols. 113 biological pathways were merged from the following sources, according to data from BioCarta (http://www.biocarta.com), INOH [102], KEGG [103], NetPath [104], PID [105], Reactome [106] and Wikipathways [107]. Analyzing the ontological features and the proportion of duplicated genes, the pathways enriched with chosen 2,505 genes (q-value < 0.05) were collapsed into 10 cancer-related pathways, having 848 genes as components.

5.2.4 Inferring and prioritizing actionable drugs

The “Clinical Evidence Summaries” data was downloaded from the Clinical Interpretations of Variants in Cancer (CIViC) website (https://civic.genome.wustl.edu/releases) on July 1, 2021, and the “Actionable Variants” data was accessed and downloaded from the Precision Oncology

Knowledge Base (OncoKB) website (http://oncokb.org/) on July 1, 2021. RSPO3-crrelated genes were annotated using 673 CIViC variations (181 genes) with predicted treatment effectiveness and 148 OncoKB actionable variants (53 genes). Then drug-target relationships were prioritized based on the scenario that properly working cancer drugs are generally inhibitors for activated oncogenes or activators for down-regulated tumor suppressor genes.

5.2.5 Statistical analysis and data visualization

All statistical analyses, including the Pearson correlation-tests, were performed using the open software R version 3.4.4. Complexheatmap, a R package, was used to visualize an RNA expression heatmap. KEGG mapper (https://www.genome.jp/kegg/mapper.html) was used to display target genes associated to WNT signaling pathway. The comprehensive network between targetable drugs and therapeutic agents was analyzed and illustrated using Cytoscape 3.5.3. In this study, statistical significance was determined as a p-value of 0.05 and false detection rate (FDR) as a q-value of 0.2 in over-representation analysis.

5.3 Results and Discussion

5.3.1 Clinicopathological characteristics

Using the Broad GDAC Firehose (Figure 15), 443 colorectal cancers were screened for P:R fusion positive cases. 7 patients demonstrated presence of the fusion mutation whereas the remaining 416 patients were negative for the fusion based on the clinic-pathological characteristics (Table 2).

There showed no definite statistical significance of histological type, age, sex, vital status and TNM stage between fusion-positive cases and controls. Notably, no other mutation driver was identified in P:R fusion-positive patients, showing mutual exclusiveness. However, one case of microsatellite instability-high (MSI-H) was identified in these P:R fusion positive patients, implying the possibility of the co-occurrence of two oncogenic aberrations.

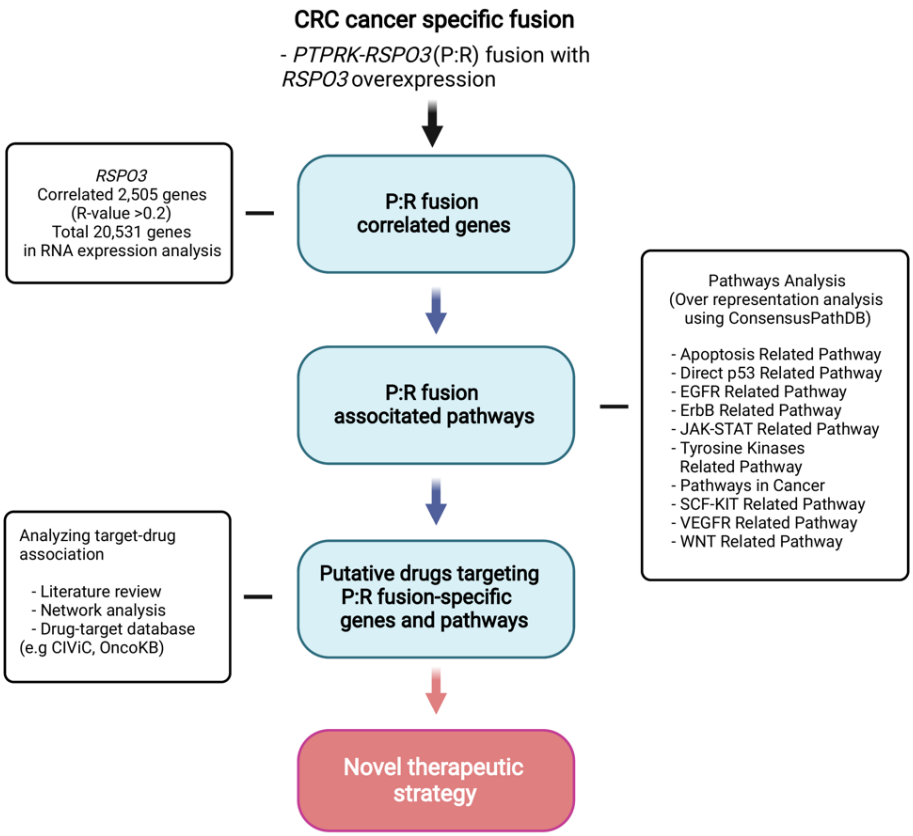


Fig 15. Overall design of this study. Transcriptome data for colorectal cancer (CRC) was attained from the Broad GDAC Firehose database. Following the RNA expression analysis of a total of 20,531 genes, 2,505 genes correlated with RSPO3 expression were selected. (R-value > 0.2, see Methods). Over-representation analysis of the 2,505 genes showed significant relation to 10 major cancer-related pathways (Apoptosis Related Pathway, Direct p53 Related Pathway, EGFR Related Pathway, ErbB Related Pathway, JAK-STAT Related Pathway, JAK-STAT Related Pathway, Tyrosine Kinases Related Pathway, Pathways in Cancer, SCF-KIT Related Pathway, VEGFR Related Pathway, WNT Related Pathway). Potential targets and repurposed drugs were inferred by analyzing target-drug associations via literature reviews and network analysis using the differentially expressed gene list and target-drug databases.

5.3.2 Key genes and pathways altered in P:R fusion-positive colorectal cancer

By the correlations test and permutation tests (See Methods), 2,505 genes were passed the Pearson correlation-test with an R-value greater than the cut-off value. Eighteen genes including PPP1R12B, NPY, VIP, C2orf72, IQGAP2, SYT2, ADCYAP1, ZNF385D, SCIN, MAGEE2, SDCBP2, AHCYL2, C6orf105, ZNF229, BTNL8, SLC7A14, GPR88, and ASTN1 showed good correlation ($R > 0.5$) with RSPO3 in RNA expression (Table S1).

Table 2. Clinicopathological characteristics of PTPRK-RSPO3 fusion-positive and fusion-negative cases in TCGA colorectal cancer.

	*Fusion (N=7)	**Control (N=186)	***p values
Age	46~76	31~90	NS
Sex			1
- Male	3/5 (60.0%)	68/121 (56.2%)	
- Female	2/5 (40.0%)	53/121 (43.8%)	
Vital status			NS
- Alive	5/5 (100.0%)	101/121 (83.8%)	
- Dead	0/5 (0.0%)	20/121 (16.5%)	
Stage			NS
- Stage I	0/4 (0.0%)	20/116 (17.2%)	
- Stage II	2/4 (50.0%)	48/116 (41.4%)	
- Stage III	2/4 (50.0%)	33/116 (28.4%)	
- Stage Iv	0/4 (0.0%)	15/116 (13.0%)	
Microsatellite instability			NS
- MSI-high	1/5 (20.0%)	18/121 (14.5%)	
- MSI-low	0/5 (0.0%)	19/121 (15.9%)	
- MSS	4/5 (80.0%)	84/121 (69.4%)	
Histological type			NS
- Adenocarcinoma	3/4 (75.0%)	113/120 (88.0%)	

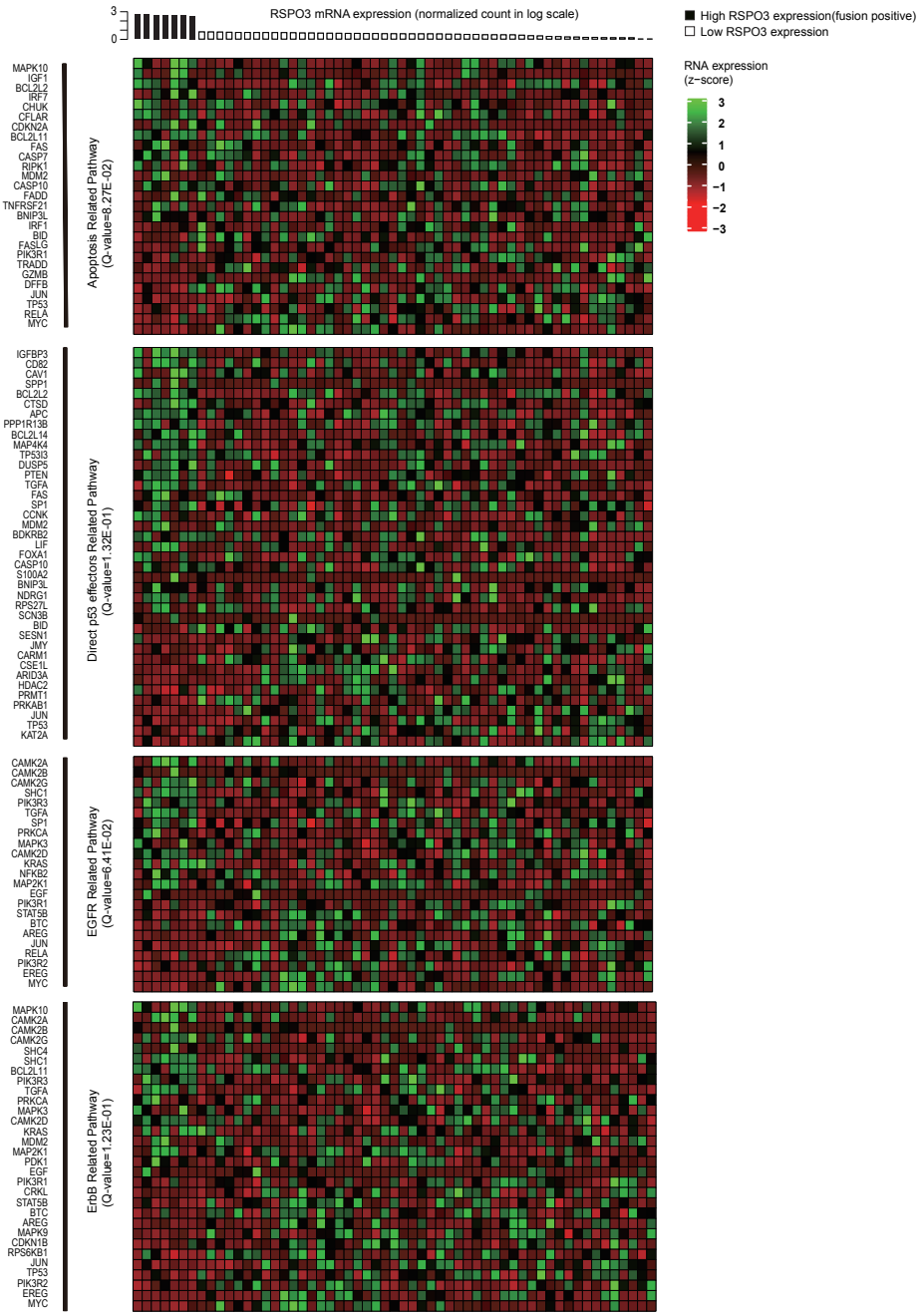
-Mucinous Adenocarcinoma	1/4 (25.0%)	7/120 (12.0%)	
Mutation profile			
- TP53 mutation	3/7 (42.9%)	113/186 (60.8%)	NS
- KRAS mutation	2/7 (28.6%)	80/186 (19.4%)	NS
- PIK3CA mutation	2/7 (28.6%)	50/186 (43.0%)	NS
- PTEN mutation	1/7 (14.3%)	15/186 (8.0%)	NS
- BRAF mutation	2/7 (28.6%)	23/186 (12.7%)	NS

* Samples harboring PTPRK-RSPO3 fusion

** Control group was extracted from samples demonstrating the lower median of RSPO3 mRNA expressions.

***p-value was calculated between fusion positive samples and controls using moonBook R package.

In pathway analysis, ten different pathways were shown to be statistically significant: apoptosis-related pathway, direct p53-related pathway, EGFR-related pathway, ErbB-related pathway, JAK-STAT-related pathway, JAK-STAT-related pathway, tyrosine kinases-related pathway, pathways in cancer, SCF-KIT-related pathway, VEGFR-related pathway, and WNT-related pathway. Of these pathways, the P:R fusion-positive cases in comparison to the P:R fusion-negative control demonstrated 848 significantly over- or under-expressed RNA expressions of 848 genes (Figure 16-17 and Figure S3-S4).


RSPO3 mRNA expression (normalized count in log scale)
3
0
High RSPO3 expression(fusion positive)
Low RSPO3 expression
RNA expression
(z−score)
3
2
1
0
−1
−2
−3
MAPK10
IGF1
BCL2L2
IRF7
CHUK
CFLAR
CDKN2A
BCL2L11
FAS
CASP7
RIPK1
MDM2
CASP10
FADD
TNFRSF21
BNIP3L
IRF1
BID
FASLG
PIK3R1
TRADD
GZMB
DFFB
JUN
TP53
RELA
MYC
Apoptosis Related Pathway
(Q-value=8.27E-02)
IGFBP3
CD82
CAV1
SPP1
BCL2L2
CTSD
APC
PPP1R13B
BCL2L14
MAP4K4
TP53I3
DUSP5
PTEN
TGFA
FAS
SP1
CCNK
MDM2
BDKRB2
LIF
FOXA1
CASP10
S100A2
BNIP3L
NDRG1
RPS27L
SCN3B
BID
SESN1
JMY
CARM1
CSE1L
ARID3A
HDAC2
PRMT1
PRKAB1
JUN
TP53
KAT2A
Direct p53 effectors Related Pathway
(Q-value=1.32E-01)
CAMK2A
CAMK2B
CAMK2G
SHC1
PIK3R3
TGFA
SP1
PRKCA
MAPK3
CAMK2D
KRAS
NFKB2
MAP2K1
EGF
PIK3R1
STAT5B
BTC
AREG
JUN
RELA
PIK3R2
EREG
MYC
EGFR Related Pathway
(Q-value=6.41E-02)
MAPK10
CAMK2A
CAMK2B
CAMK2G
SHC4
SHC1
BCL2L11
PIK3R3
TGFA
PRKCA
MAPK3
CAMK2D
KRAS
MDM2
MAP2K1
PDK1
EGF
PIK3R1
CRKL
STAT5B
BTC
AREG
MAPK9
CDKN1B
RPS6KB1
JUN
TP53
PIK3R2
EREG
MYC
ErbB Related Pathway
(Q-value=1.23E-01)

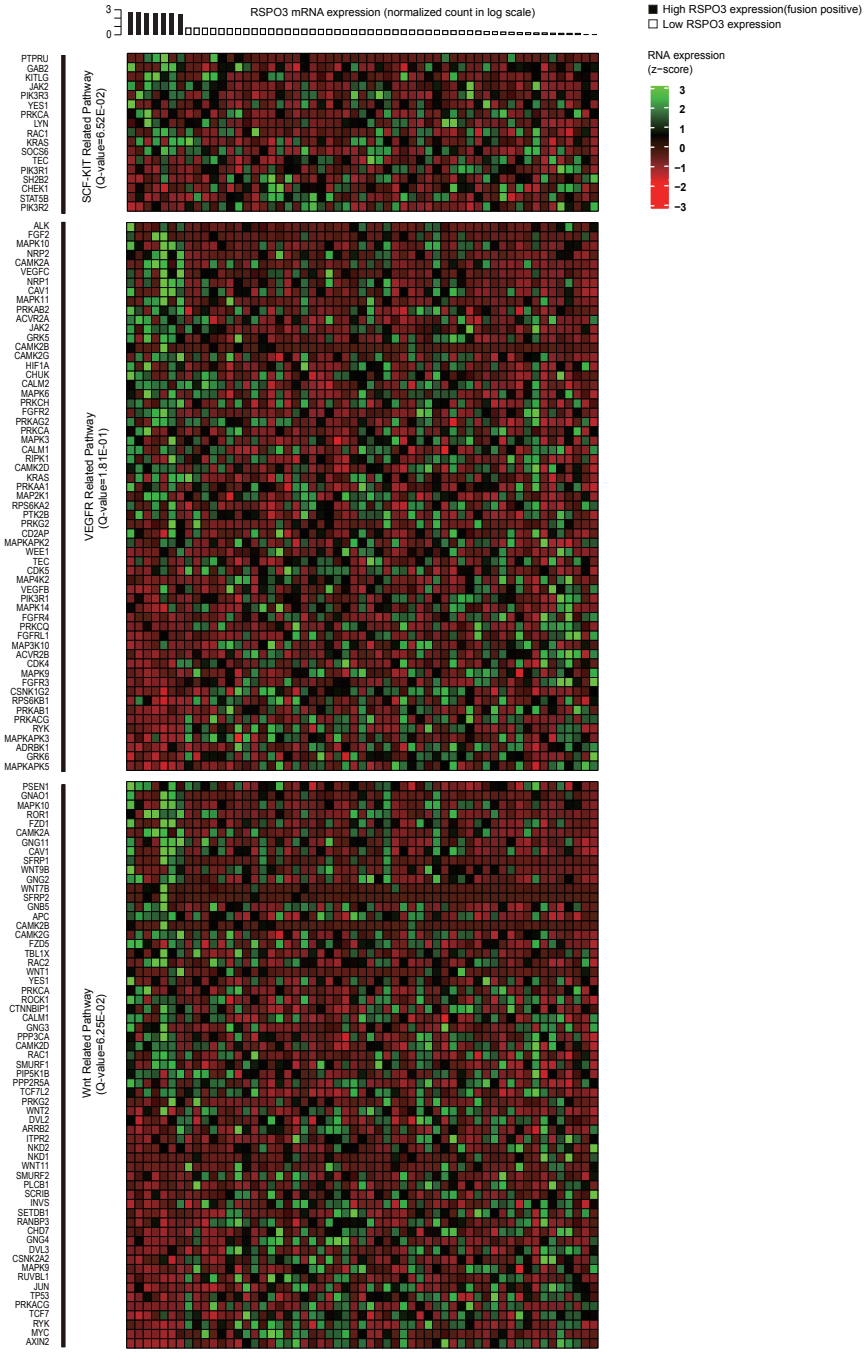


Fig 16. Gene expression heatmap of 7 cancer-related pathways enriched with genes that were correlated to RSPO3 in RNA expression. A total of 256 genes associated with Apoptosis, Direct p53, EGFR, ErbB, SCF-KIT, VEGFR, WNT signaling showed significant differences in expression between RSPO3 fusion-positive colorectal samples and the control samples (see details in methods). The RNA expression was transformed to z-score. The x-axis represents the sample, and the y-axis represents the RNA expression.

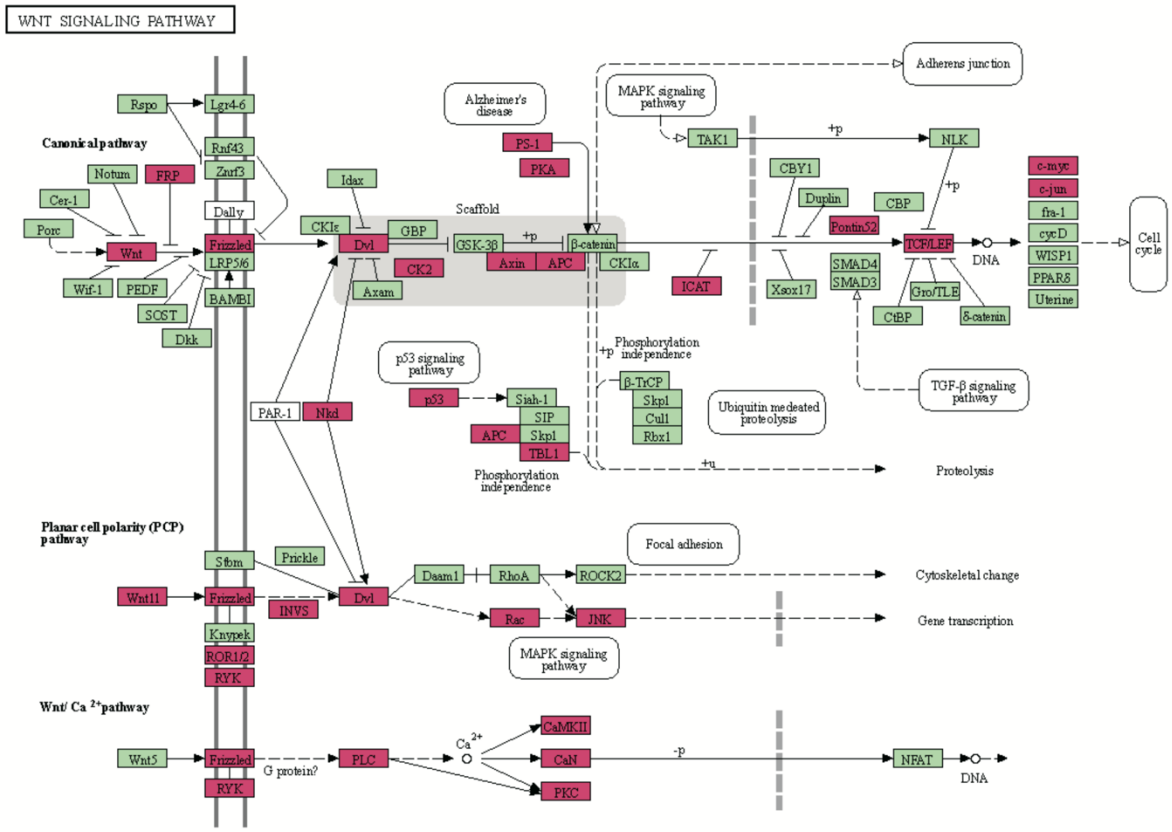
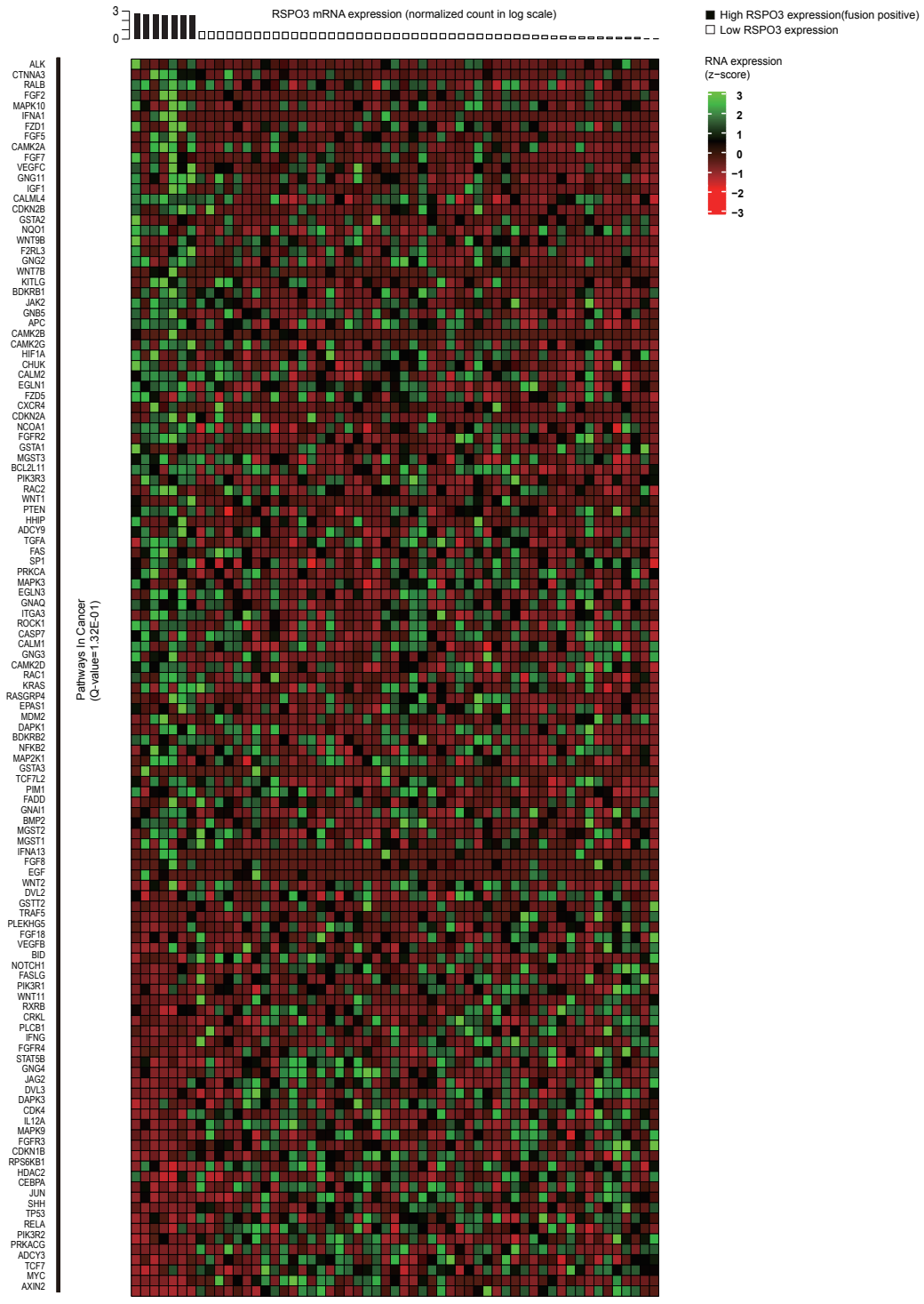


Fig 17. Over- and under-expressed genes are highlighted in WNT signaling pathway. The KEGG pathway map for the human WNT signaling pathway (hsa 04310) was illustrated using the KEGG Mapper; genes correlated with RSPO3 expression are colored in pink.


RSPO3 mRNA expression (normalized count in log scale)
3
0
High RSPO3 expression(fusion positive)
Low RSPO3 expression
RNA expression
(z−score)
3
2
1
0
−1
−2
−3
Pathways In Cancer
(Q-value=1.32E-01)
ALK
CTNNA3
RALB
FGF2
MAPK10
IFNA1
FZD1
FGF5
CAMK2A
FGF7
VEGFC
GNG11
IGF1
CALML4
CDKN2B
GSTA2
NQO1
WNT9B
F2RL3
GNG2
WNT7B
KITLG
BDKRB1
JAK2
GNB5
APC
CAMK2B
CAMK2G
HIF1A
CHUK
CALM2
EGLN1
FZD5
CXCR4
CDKN2A
NCOA1
FGFR2
GSTA1
MGST3
BCL2L11
PIK3R3
RAC2
WNT1
PTEN
HHIP
ADCY9
TGFA
FAS
SP1
PRKCA
MAPK3
EGLN3
GNAQ
ITGA3
ROCK1
CASP7
CALM1
GNG3
CAMK2D
RAC1
KRAS
RASGRP4
EPAS1
MDM2
DAPK1
BDKRB2
NFKB2
MAP2K1
GSTA3
TCF7L2
PIM1
FADD
GNAI1
BMP2
MGST2
MGST1
IFNA13
FGF8
EGF
WNT2
DVL2
GSTT2
TRAF5
PLEKHG5
FGF18
VEGFB
BID
NOTCH1
FASLG
PIK3R1
WNT11
RXRB
CRKL
PLCB1
IFNG
FGFR4
STAT5B
GNG4
JAG2
DVL3
DAPK3
CDK4
IL12A
MAPK9
FGFR3
CDKN1B
RPS6KB1
HDAC2
CEBPA
JUN
SHH
TP53
RELA
PIK3R2
PRKACG
ADCY3
TCF7
MYC
AXIN2

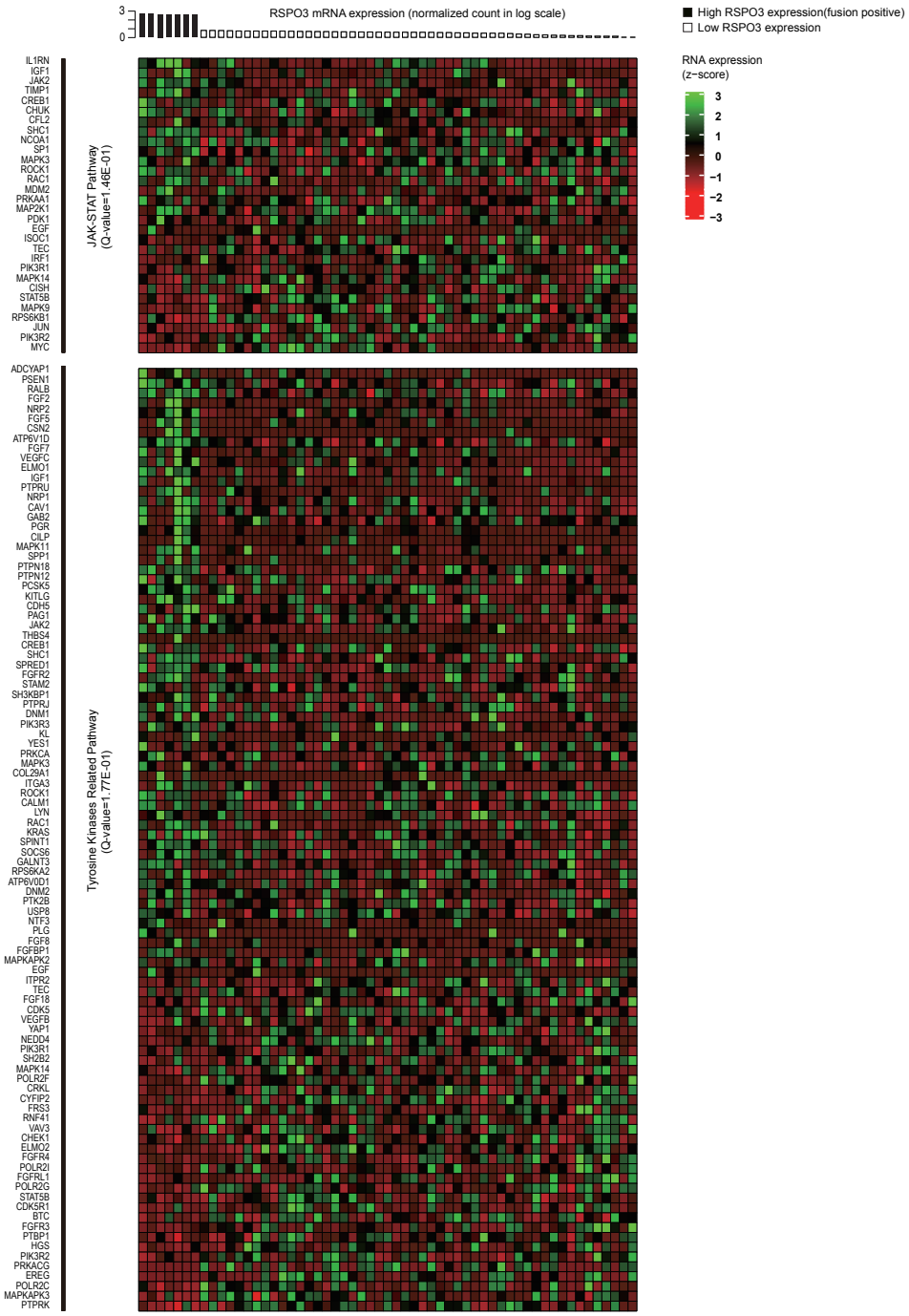


Figure S3. Gene expression heatmap of cancer-related pathways enriched with genes correlated to RSPO3 in RNA expression.

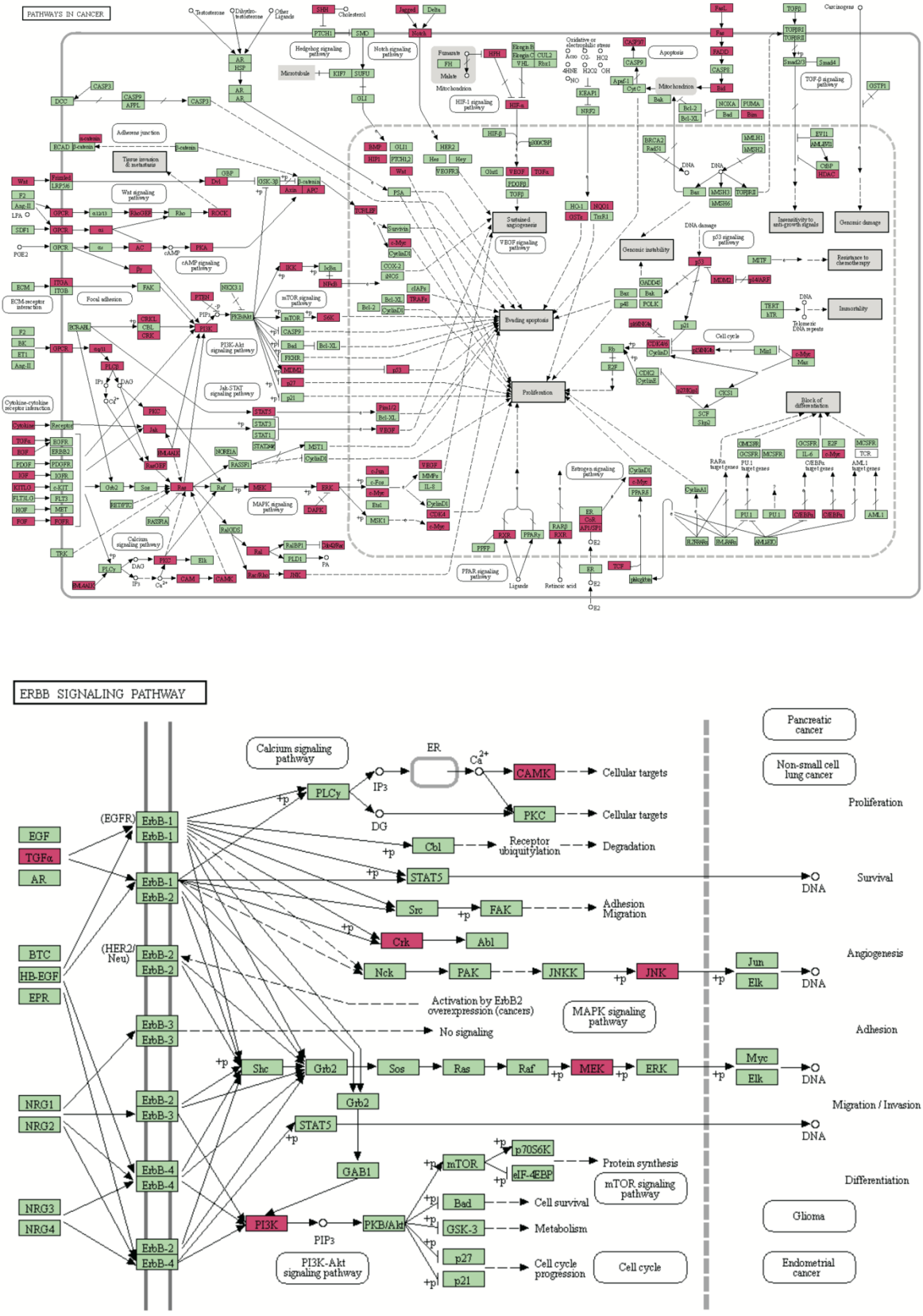


Figure S4. The KEGG pathway maps for the human ERBB signaling pathway and pathways in cancer using the KEGG Mapper; genes correlated with RSPO3 expression are colored in pink.

Among the 848 genes, 36 genes were annotated as cancer genes using the cancer gene census database offered by the CatalogueOfSomaticMutationsInCancer DB (COSMIC, https://cancer.sanger.ac.uk). Of these, ten genes from highest R-values were as follows: *ALK, ACSL3, AXIN2, PTPRK, CDX2, MYC, TP53, GNAQ, ACVR2A*, and *FAS*. *RSPO3*-correlated cancer genes involved in more than four pathways were as follows; *JUN* was involved the most in 9 of 10 pathways, *APC*, 6 pathways, *AXIN2*, 5 pathways, *FGFR2*, 5 pathways, *JAK2*, 7 pathways, *MDM2*, 5 pathways, *MYC*, 9 pathways, *RAC1*, 8 pathways, and lastly, *TP53*, 7 pathways. In addition, 4 genes were associated with 4 pathways, 4 genes in 3 pathways and 7 genes in 2 pathways (Figure S5). Amongst these genes, those with a correlation R-value greater than 0.3 were *ALK*(R=0.44), *ACSL3*(R=0.43), *AXIN*(R=-0.38), *MYC* (R=-0.34), *TP53* (R=-0.33), *GNAQ* (R=0.31), *ACVR2A* (R=0.31), and *FAS* (R=0.31).

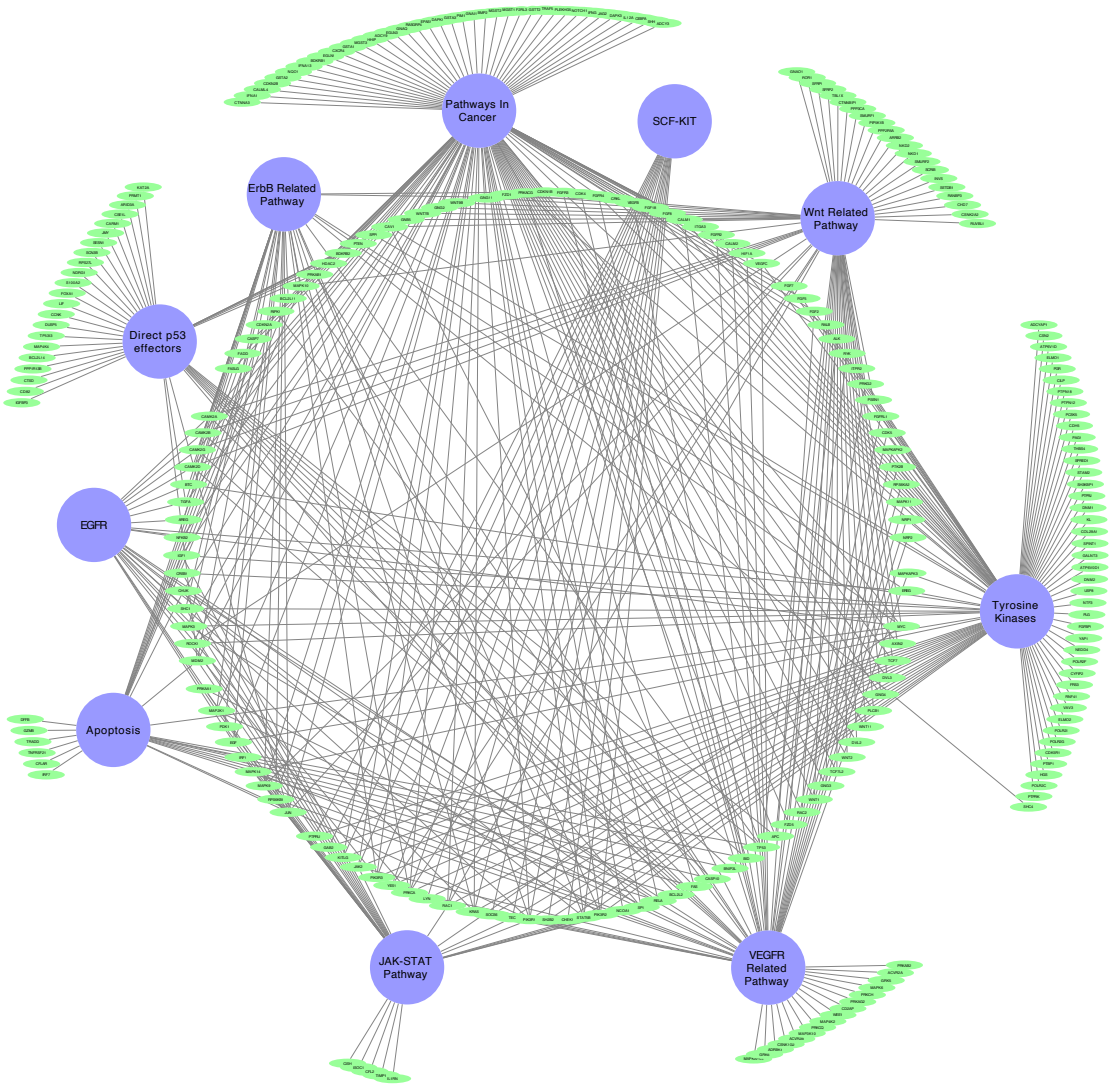


Figure S5. Putative target genes involved in multiple pathways of PTPRK-RSPO3 fusion-positive cancer.

5.3.3 Identification of actionable targets and potential therapeutic choice using network analysis

By matching the 848 genes included in the 10 pathways associated with P:R fusion-positive colorectal cancer using the CiVIC database and OncoKB, we were able to infer 673 and 262 drugs to have actionable target potential.

In the CIViC database, following 19 genes among 848 genes were related with actionable drugs: ALK, FGFR2, TP53, HIF1A, EPAS1, KRAS, CEBPA, NOTCH1, STK11, JAK2, PGR, RAD50, PIK3R1, CDKN1B, NQO1, NT5E, MAP2K1, GNAQ, and PTEN. ALK was identified to be a class A-type drug (Proven/consensus association in human medicine) which can be targeted using crizotinib, alectinib and ceritinib. In other class-type (B, C, D, and E) gene-drug association, additional thirteen drugs were found, considering the scenario for inhibitors for activated oncogenes or activators for down-regulated tumor suppressor genes (Figure 18A).

When using the OncoKB database, 4 genes (KRAS, FGFR2, ALK, and JAK2) were identified and they were all included in the inferred results using CIViC database. Level 1 drugs (FDA-recognized biomarker predictive of response to an FDA-approved drug) for target genes are as follows: lorlatinib, brigatinib, crizotinib, ceritinib, alectinib for ALK; erdafitinib, infigratinib, pemigatinib for FGFR2; sotorasib for KRAS. These cancer drugs appeared to be inhibitors for activated oncogenes (Figure 18B).

Of 19 druggable genes, five were involved in the multiple pathway: *FGFR2* for 5 cancer-related pathways (gastric cancer pathway, VEGFR

related pathway, tyrosine kinases, stem cell, and pathways In cancer); *JAK2* for 7 cancer-related pathways (JAK-STAT pathway, adipocytokine, VEGFR related pathway, SCF-KIT, pathways in cancer, tyrosine kinases pathway, and stem cell pathway); *ALK* for 2 cancer-related pathways (VEGFR related pathway and pathways in cancer); *HIF1A* for 2 cancer-related pathways (VEGFR related pathway and pathways in cancer); *STK112* for 2 cancer-related pathways (adipocytokine-related pathway, energy metabolism-related pathway).

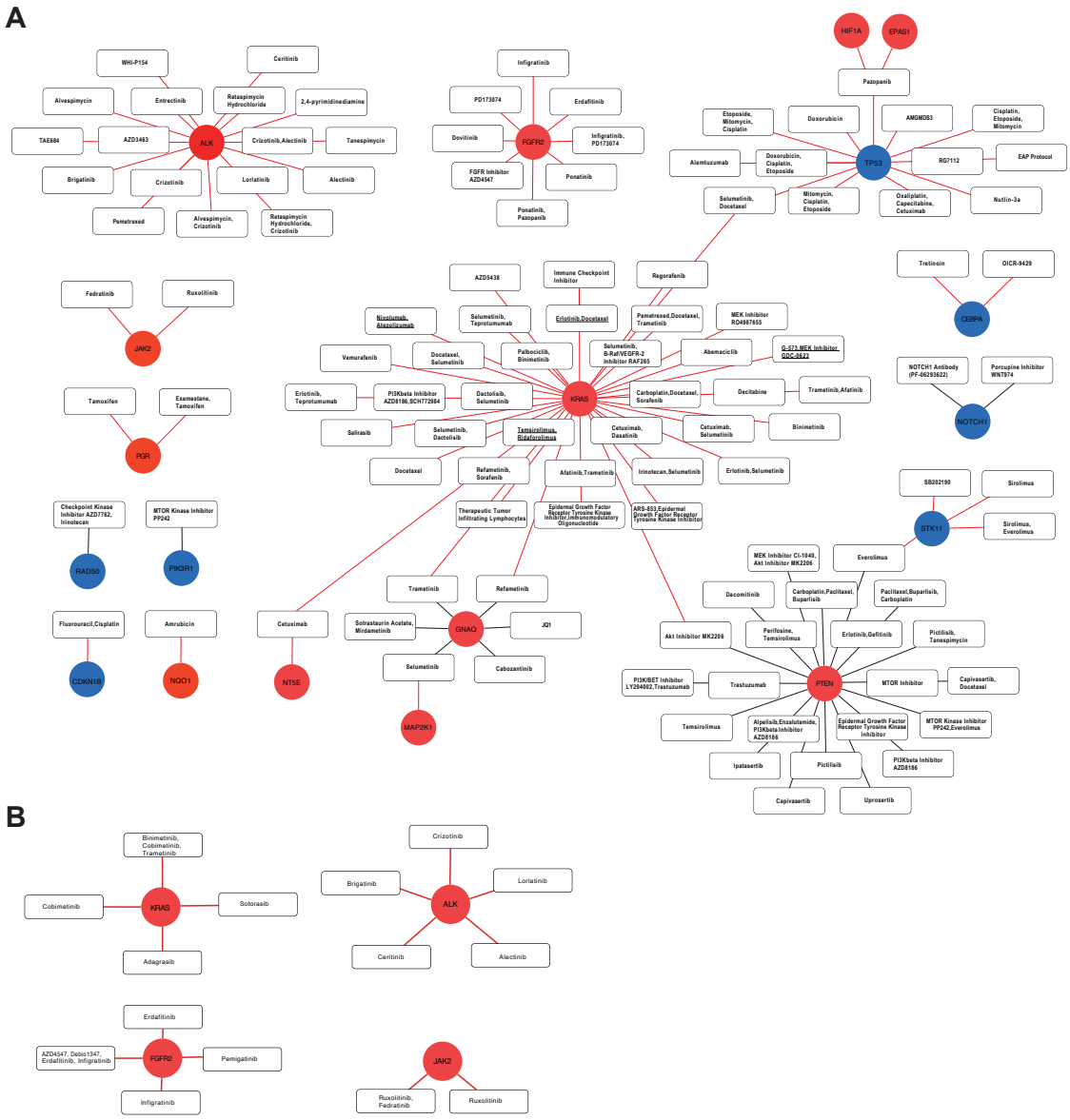


Fig 18. Inferred drug-target network in PTPRK-RSPO3 fusion-positive colorectal cancer. Drug-target relation was obtained based on CIViC and OncoKB databases: white boxes, drugs; circles, underlined white boxes, substitute drugs; genes; red circles, genes that are over-expressed in fusion-positive cancer; blue circles, genes that are under-expressed in fusion-positive cancer. The red lines are prioritized drug-target relationships based on the scenario that properly working cancer drugs are generally inhibitors for activated oncogenes or activators for down-regulated tumor suppressor genes.

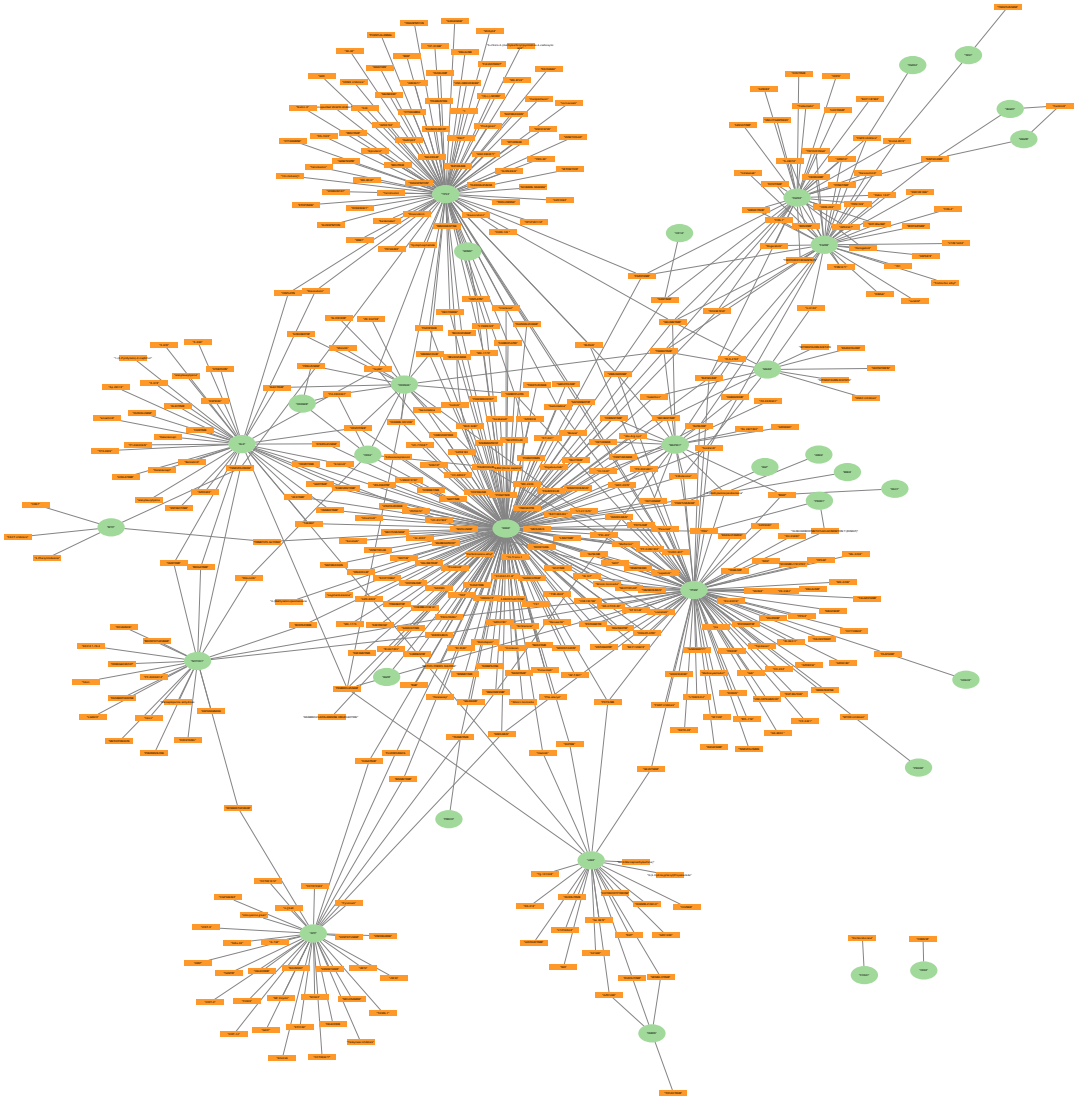


Figure S6. Inferred drug-target network in PTPRK-RSPO3 fusion-positive colorectal cancer based on VICC database.

5.3.4 Discussion

In this study, drug candidates were identified in P:R fusion colorectal cancer as follows. First, genes correlated with *RSPO3* RNA expression were extracted, and oncogenic cell signaling pathways including these genes were selected. We then used the drug target database to build a drug target network in P:R fusion colorectal cancer and prioritize suitable therapeutics (Fig 4). As a result, this study is expected to provide an opportunity to try a wider range of therapeutics in colorectal cancer, where EGFR inhibitors and ICI are limitedly used as targeted therapeutics [107].

Previous studies that systematically explore gene biomarkers with bioinformatics analysis in colorectal cancer have focused on discovering prognosis-related biomarkers using differentially expressed genes (DEGs) analysis and machine learning techniques [134, 135]. To our best knowledge, our study differs from previous studies in two respects. First, the purpose of this study is to discover novel targets and therapeutics related to original mutations by analyzing downstream pathways and genes affected by target mutations that cannot be directly targeted. Second, our study is based on a structural variation (P:R fusion by DNA structural variation) that is a driver mutation in colorectal cancer. As consequence, almost all genes correlated with P:R fusion are downstream-level genes affected by fusion. In this aspect, our study is different from other studies, and, for example, it is not clear whether COL11A1 is a primary driver or is affected by other drivers in the study by Ritwik et al [135].

The WNT signaling pathway is an important mediator in tissue homeostasis and recovery while it acts an important role in tumor-

development of colorectal cancer [132]. Both in vitro experiments in human-colon cancer cell line HT-29 and in vivo experiments in CRISPR-based xenograft mice provided the evidence that *RSPO3* fusion gene was involved in the initiation and development of CRC via activating WNT signaling [128, 133]. This means that human CRC is a sensitive tumor for WNT-targeted treatment, suggesting that *RSPO3* fusion gene can be an effective therapeutic target.

As a result of our analysis, it is interesting that *ALK* up-regulated in P:R fusion-positive CRC has the following three characteristics. First, the correlation between *ALK* RNA expression and *RSPO3* RNA expression in P:R fusion-positive CRC was the highest among COSMIC common oncogenes (R=0.44). Second, *ALK* was a gene involved in multiple cancer pathways. Finally, *ALK* inhibitors are FDA-approved therapeutics that perform well in other carcinomas (e.g. lung cancer) [136, 137]. Taken together, in-silico analysis showed that *ALK* inhibitors were highly likely to act in P:R fusion positivity [138].

Despite the limited number of samples, the clinical characteristics of P:R-positive and P:R-negative patients were found to be similar. This indicates that even if the clinical properties are similar, the molecular properties may be different, which may require treatment to target the molecular properties. One interesting point is that P:R fusions can also be found in MSI-H. In this case, further clinical evaluation is needed to determine if there is a synergistic effect between ICI and the targeted therapy we propose.

In summary, we were able to present key indicators and clinically

viable therapeutics for P:R fusion-positive CRC. Our findings will serve as a steppingstone for future research in the development of precision medicine targeting colorectal cancer.

5.4 Conclusion

Colorectal cancer (CRC) is one of the most deadly and common diseases in the world, accounting for over 881,000 casualties in 2018. The *PTPRK-RSPO3* (P:R) fusion is a structural variation in CRC and well known for its ability to activate WNT signaling and tumorigenesis. However, till now, therapeutic targets and actionable drugs are limited in this subtype of cancer. The purpose of this study is to identify key genes and cancer-related pathways specific for P:R fusion-positive CRC. In addition, we also inferred the actionable drugs in bioinformatics analysis using the Cancer Genome Atlas (TCGA) data. 2,505 genes were altered in RNA expression specific for P:R fusion-positive CRC. By pathway analysis based on the altered genes, ten major cancer-related signaling pathways (Apoptosis, Direct p53, EGFR, ErbB, JAK-STAT, tyrosine kinases, Pathways in Cancer, SCF-KIT, VEGFR, and WNT-related Pathway) were significantly altered in P:R fusion-positive CRC. Among these pathways, the most altered cancer genes (*ALK, ACSL3, AXIN, MYC, TP53, GNAQ, ACVR2A*, and *FAS*) specific for P:R fusion and involved in multiple cancer pathways were considered to have a key role in P:R fusion-positive CRC. Based on the drug-target network analysis, crizotinib, alectinib, lorlatinib, brigatinib, ceritinib, erdafitinib, infigratinib and pemigatinib were selected as putative therapeutic candidates, since they were already used in routine clinical practice in other cancer types and target genes of the drugs were involved in multiple cancer-pathways.

References

1. Lee YT, Tan YJ, Oon CE. Molecular targeted therapy: Treating cancer with specificity. Eur J Pharmacol. 2018;834:188-96.

2. DeVita VT, Jr., Chu E. A history of cancer chemotherapy. Cancer Res. 2008;68(21):8643-53.

3. Duenas-Gonzalez A, Gonzalez-Fierro A. Barriers for Pharmaceutical Innovation With Focus in Cancer Drugs, the Case of Mexico. Ther Innov Regul Sci. 2020;54(2):342-52.

4. Rosland GV, Engelsen AS. Novel points of attack for targeted cancer therapy. Basic Clin Pharmacol Toxicol. 2015;116(1):9-18.

5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-58.

6. Chabner BA, Roberts TG, Jr. Timeline: Chemotherapy and the war on cancer. Nat Rev Cancer. 2005;5(1):65-72.

7. Saijo N. Progress in cancer chemotherapy with special stress on molecular-targeted therapy. Jpn J Clin Oncol. 2010;40(9):855-62.

8. Amer MH. Gene therapy for cancer: present status and future perspective. Mol Cell Ther. 2014;2:27.

9. Tsai MJ, Chang WA, Huang MS, Kuo PL. Tumor microenvironment: a new treatment target for cancer. ISRN Biochem. 2014;2014:351959.

10. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3(8):673-83.

11. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11(3):191-200.

12. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical

R&D. Nat Rev Drug Discov. 2011;10(6):428-38.

13. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov. 2015;14(7):475-86.

14. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41-58.

15. Danzon PM, Towse A. Differential pricing for pharmaceuticals: reconciling access, R&D and patents. Int J Health Care Finance Econ. 2003;3(3):183-205.

16. Serajuddin HK, Serajuddin AT. Value of pharmaceuticals: ensuring the future of research and development. J Am Pharm Assoc (2003). 2006;46(4):511-6.

17. Verkman AS. Drug discovery in academia. Am J Physiol Cell Physiol. 2004;286(3):C465-74.

18. Barton JH, Emanuel EJ. The patents-based pharmaceutical development process: rationale, problems, and potential reforms. JAMA. 2005;294(16):2075-82.

19. Breckenridge A, Jacob R. Overcoming the legal and regulatory barriers to drug repurposing. Nat Rev Drug Discov. 2019;18(1):1-2.

20. Nosengo N. Can you teach old drugs new tricks? Nature. 2016;534(7607):314-6.

21. Shah RR, Stonier PD. Repurposing old drugs in oncology: Opportunities with clinical and regulatory challenges ahead. J Clin Pharm Ther. 2019;44(1):6-22.

22. Mohty M, Terpos E, Mateos MV, Cavo M, Lejniece S, Beksac M, et al. Multiple Myeloma Treatment in Real-world Clinical Practice: Results of a Prospective, Multinational, Noninterventional Study. Clin Lymphoma Myeloma Leuk. 2018;18(10):e401-e19.

23. Rao Y, Li R, Zhang D. A drug from poison: how the therapeutic effect of arsenic trioxide on acute promyelocytic leukemia was discovered. Sci China Life Sci.

2013;56(6):495-502.

24. Huang ME, Ye YC, Chen SR, Chai JR, Lu JX, Zhoa L, et al. Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia. Blood. 1988;72(2):567-72.

25. Zhou G, Myers R, Li Y, Chen Y, Shen X, Fenyk-Melody J, et al. Role of AMP-activated protein kinase in mechanism of metformin action. J Clin Invest. 2001;108(8):1167-74.

26. Segura-Pacheco B, Trejo-Becerril C, Perez-Cardenas E, Taja-Chayeb L, Mariscal I, Chavez A, et al. Reactivation of tumor suppressor genes by the cardiovascular drugs hydralazine and procainamide and their potential use in cancer therapy. Clin Cancer Res. 2003;9(5):1596-603.

27. Gligorijevic V, Malod-Dognin N, Przulj N. Integrative methods for analyzing big data in precision medicine. Proteomics. 2016;16(5):741-58.

28. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929-35.

29. Iorio F, Isacchi A, di Bernardo D, Brunetti-Pierri N. Identification of small molecules enhancing autophagic function from drug network analysis. Autophagy. 2010;6(8):1204-5.

30. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010;107(33):14621-6.

31. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47(8):856-60.

32. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5-22.

33. Dunning AM, Michailidou K, Kuchenbaecker KB, Thompson D, French JD, Beesley J, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nat Genet. 2016;48(4):374-86.

34. Thompson DJ, O'Mara TA, Glubb DM, Painter JN, Cheng T, Folkerd E, et al. CYP19A1 fine-mapping and Mendelian randomization: estradiol is causal for endometrial cancer. Endocr Relat Cancer. 2016;23(2):77-91.

35. Smith SB, Dampier W, Tozeren A, Brown JR, Magid-Slav M. Identification of common biological pathways and drug targets across multiple respiratory viruses based on human host gene expression analysis. PLoS One. 2012;7(3):e33174.

36. Greene CS, Voight BF. Pathway and network-based strategies to translate genetic discoveries into effective therapies. Hum Mol Genet. 2016;25(R2):R94-R8.

37. Iorio F, Saez-Rodriguez J, di Bernardo D. Network based elucidation of drug response: from modulators to targets. BMC Syst Biol. 2013;7:139.

38. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007;7(4):233-45.

39. Edwards PA. Fusion genes and chromosome translocations in the common epithelial cancers. J Pathol. 2010;220(2):244-54.

40. An X, Tiwari AK, Sun Y, Ding PR, Ashby CR, Jr., Chen ZS. BCR-ABL tyrosine kinase inhibitors in the treatment of Philadelphia chromosome positive chronic myeloid leukemia: a review. Leuk Res. 2010;34(10):1255-68.

41. Awad MM, Shaw AT. ALK inhibitors in non-small cell lung cancer: crizotinib and beyond. Clin Adv Hematol Oncol. 2014;12(7):429-39.

42. Yun JW, Bae YK, Cho SY, Koo H, Kim HJ, Nam DH, et al. Elucidation of Novel Therapeutic Targets for Acute Myeloid Leukemias with RUNX1-RUNX1T1 Fusion. Int J Mol Sci. 2019;20(7).

43. Natarajan A, Thangarajan R, Kesavan S. Repurposing Drugs by In Silico Methods to Target BCR Kinase Domain in Chronic Myeloid Leukemia. Asian Pac J Cancer Prev. 2019;20(11):3399-406.

44. Yun JW, Lee S, Chun S, Lee KW, Kim J, Kim HS. Comprehensive analysis of oncogenic signatures and consequent repurposed drugs in TMPRSS2:ERG fusion-positive prostate cancer. Clin Transl Med. 2021;11(5):e420.

45. Latysheva NS, Babu MM. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res. 2016;44(10):4487-503.

46. Sheehan KM, Calvert VS, Kay EW, Lu Y, Fishman D, Espina V, et al. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. Mol Cell Proteomics. 2005;4(4):346-55.

47. Trevino V, Falciani F, Barrera-Saldana HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol Med. 2007;13(9-10):527-41.

48. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586-97.

49. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8.

50. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-7.

51. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.

52. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519-25.

53. Cancer Genome Atlas Research N, Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N

Engl J Med. 2013;368(22):2059-74.

54. Cancer Genome Atlas Research N, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67-73.

55. Cancer Genome Atlas Research N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315-22.

56. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543-50.

57. Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513(7517):202-9.

58. Cancer Genome Atlas Research N. Integrated genomic characterization of papillary thyroid carcinoma. Cell. 2014;159(3):676-90.

59. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576-82.

60. Cancer Genome Atlas N. Genomic Classification of Cutaneous Melanoma. Cell. 2015;161(7):1681-96.

61. Cancer Genome Atlas Research N. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163(4):1011-25.

62. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015;163(2):506-19.

63. Cancer Genome Atlas Research N, Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N Engl J Med. 2016;374(2):135-45.

64. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al.

Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 2016;164(3):550-63.

65. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell. 2016;29(5):723-36.

66. Cancer Genome Atlas Research N, Albert Einstein College of M, Analytical Biological S, Barretos Cancer H, Baylor College of M, Beckman Research Institute of City of H, et al. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017;543(7645):378-84.

67. Cancer Genome Atlas Research N, Analysis Working Group: Asan U, Agency BCC, Brigham, Women's H, Broad I, et al. Integrated genomic characterization of oesophageal carcinoma. Nature. 2017;541(7636):169-75.

68. Fishbein L, Leshchiner I, Walter V, Danilova L, Robertson AG, Johnson AR, et al. Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma. Cancer Cell. 2017;31(2):181-93.

69. Cho S, Jang I, Jun Y, Yoon S, Ko M, Kwon Y, et al. MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. Nucleic Acids Res. 2013;41(Database issue):D252-7.

70. Li J, Han L, Roebuck P, Diao L, Liu L, Yuan Y, et al. TANRIC: An Interactive Open Platform to Explore the Function of lncRNAs in Cancer. Cancer Res. 2015;75(18):3728-37.

71. Yang IS, Son H, Kim S, Kim S. ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. BMC Genomics. 2016;17(1):631.

72. Spainhour JCG, Lim J, Qiu P. GDISC: a web portal for integrative analysis of gene-drug interaction for survival in cancer. Bioinformatics. 2017;33(9):1426-8.

73. Lee H, Palm J, Grimes SM, Ji HP. The Cancer Genome Atlas Clinical Explorer: a

web and mobile interface for identifying clinical-genomic driver associations. Genome Med. 2015;7:112.

74. Goswami CP, Nakshatri H. PROGgeneV2: enhancements on the existing database. BMC Cancer. 2014;14:970.

75. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004;6(1):1-6.

76. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia. 2007;9(2):166-80.

77. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321-37 e10.

78. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.

79. Mahon FX. Discontinuation of TKI therapy and 'functional' cure for CML. Best Pract Res Clin Haematol. 2016;29(3):308-13.

80. Braun TP, Eide CA, Druker BJ. Response and Resistance to BCR-ABL1-Targeted Therapies. Cancer Cell. 2020;37(4):530-42.

81. Al-Share B, Alloghbi A, Al Hallak MN, Uddin H, Azmi A, Mohammad RM, et al. Gastrointestinal stromal tumor: a review of current and emerging therapies. Cancer Metastasis Rev. 2021;40(2):625-41.

82. Yoneda K, Imanishi N, Ichiki Y, Tanaka F. Treatment of Non-small Cell Lung Cancer with EGFR-mutations. J UOEH. 2019;41(2):153-63.

83. Li L, Zhang D, Liu B, Lv D, Zhai J, Guan X, et al. Antibody-drug conjugates in HER2-positive breast cancer. Chin Med J (Engl). 2021;135(3):261-7.

84. Ocana A, Tannock IF. When are "positive" clinical trials in oncology truly positive? J Natl Cancer Inst. 2011;103(1):16-20.

85. Arnedos M, Soria JC, Andre F, Tursz T. Personalized treatments of cancer patients: a reality in daily practice, a costly dream or a shared vision of the future from the oncology community? Cancer Treat Rev. 2014;40(10):1192-8.

86. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.

87. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321(3):288-300.

88. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol. 2014;5(3):412-24.

89. Veeraraghavan J, Tan Y, Cao XX, Kim JA, Wang X, Chamness GC, et al. Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. Nat Commun. 2014;5:4577.

90. Goksu SS, Tastekin D, Arslan D, Gunduz S, Tatli AM, Unal D, et al. Clinicopathologic features and molecular subtypes of breast cancer in young women (age </=35). Asian Pac J Cancer Prev. 2014;15(16):6665-8.

91. Hartmaier RJ, Trabucco SE, Priedigkeit N, Chung JH, Parachoniak CA, Vanden Borre P, et al. Recurrent hyperactive ESR1 fusion proteins in endocrine therapy-resistant breast cancer. Ann Oncol. 2018;29(4):872-80.

92. Matissek KJ, Onozato ML, Sun S, Zheng Z, Schultz A, Lee J, et al. Expressed Gene Fusions as Frequent Drivers of Poor Outcomes in Hormone Receptor-Positive Breast Cancer. Cancer Discov. 2018;8(3):336-53.

93. Giltnane JM, Hutchinson KE, Stricker TP, Formisano L, Young CD, Estrada MV, et

al. Genomic profiling of ER(+) breast cancers after short-term estrogen suppression reveals alterations associated with endocrine resistance. Sci Transl Med. 2017;9(402).

94. Fimereli D, Fumagalli D, Brown D, Gacquer D, Rothe F, Salgado R, et al. Genomic hotspots but few recurrent fusion genes in breast cancer. Genes Chromosomes Cancer. 2018;57(7):331-8.

95. Lei JT, Shao J, Zhang J, Iglesia M, Chan DW, Cao J, et al. Functional Annotation of ESR1 Gene Fusions in Estrogen Receptor-Positive Breast Cancer. Cell Rep. 2018;24(6):1434-44 e7.

96. Veeraraghavan J, Ma J, Hu Y, Wang XS. Recurrent and pathological gene fusions in breast cancer: current advances in genomic discovery and clinical implications. Breast Cancer Res Treat. 2016;158(2):219-32.

97. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet. 2010;42(6):504-7.

98. Wang Y, He Y, Qin Z, Jiang Y, Jin G, Ma H, et al. Evaluation of functional genetic variants at 6q25.1 and risk of breast cancer in a Chinese population. Breast Cancer Res. 2014;16(4):422.

99. Li L, Lin L, Veeraraghavan J, Hu Y, Wang X, Lee S, et al. Therapeutic role of recurrent ESR1-CCDC170 gene fusions in breast cancer endocrine resistance. Breast Cancer Res. 2020;22(1):84.

100. Yang SYC, Lheureux S, Karakasis K, Burnier JV, Bruce JP, Clouthier DL, et al. Landscape of genomic alterations in high-grade serous ovarian cancer from exceptional long- and short-term survivors. Genome Med. 2018;10(1):81.

101. Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene.

2015;34(37):4845-54.

102. Wilson RG, Smith AN, Bird CC. Immunohistochemical detection of abnormal cell proliferation in colonic mucosa of subjects with polyps. J Clin Pathol. 1990;43(9):744-7.

103. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D61.

104. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, et al. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 2010;11(1):R3.

105. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37(Database issue):D674-9.

106. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res. 2016;44(D1):D481-7.

107. Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, et al. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016;44(D1):D488-94.

108. Ma CX, Reinert T, Chmielewska I, Ellis MJ. Mechanisms of aromatase inhibitor resistance. Nat Rev Cancer. 2015;15(5):261-75.

109. Jeselsohn R, Buchwalter G, De Angelis C, Brown M, Schiff R. ESR1 mutations-a mechanism for acquired endocrine resistance in breast cancer. Nat Rev Clin Oncol. 2015;12(10):573-83.

110. Lei JT, Gou X, Seker S, Ellis MJ. ESR1 alterations and metastasis in estrogen receptor positive breast cancer. J Cancer Metastasis Treat. 2019;5.

111. Wierer M, Verde G, Pisano P, Molina H, Font-Mateu J, Di Croce L, et al. PLK1 signaling in breast cancer cells cooperates with estrogen receptor-dependent gene

transcription. Cell Rep. 2013;3(6):2021-32.

112. Robinson TJ, Liu JC, Vizeacoumar F, Sun T, Maclean N, Egan SE, et al. RB1 status in triple negative breast cancer cells dictates response to radiation treatment and selective therapeutic drugs. PLoS One. 2013;8(11):e78641.

113. McGovern SL, Qi Y, Pusztai L, Symmans WF, Buchholz TA. Centromere protein-A, an essential centromere protein, is a prognostic marker for relapse in estrogen receptor-positive breast cancer. Breast Cancer Res. 2012;14(3):R72.

114. Shan W, Jiang Y, Yu H, Huang Q, Liu L, Guo X, et al. HDAC2 overexpression correlates with aggressive clinicopathological features and DNA-damage response pathway of breast cancer. Am J Cancer Res. 2017;7(5):1213-26.

115. Nevanlinna H, Bartek J. The CHEK2 gene and inherited breast cancer susceptibility. Oncogene. 2006;25(43):5912-9.

116. Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Prz Gastroenterol. 2019;14(2):89-103.

117. Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. CA Cancer J Clin. 2014;64(2):104-17.

118. Fearon ER. Molecular genetics of colorectal cancer. Annu Rev Pathol. 2011;6:479-507.

119. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318(5853):1108-13.

120. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, et al. Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PLoS One. 2010;5(12):e15661.

121. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat

Genet. 2011;43(10):964-8.

122. Xie YH, Chen YX, Fang JY. Comprehensive review of targeted therapy for colorectal cancer. Signal Transduct Target Ther. 2020;5(1):22.

123. Sansom OJ, Reed KR, Hayes AJ, Ireland H, Brinkmann H, Newton IP, et al. Loss of Apc in vivo immediately perturbs Wnt signaling, differentiation, and migration. Genes Dev. 2004;18(12):1385-90.

124. Barker N, Ridgway RA, van Es JH, van de Wetering M, Begthel H, van den Born M, et al. Crypt stem cells as the cells-of-origin of intestinal cancer. Nature. 2009;457(7229):608-11.

125. Moser AR, Pitot HC, Dove WF. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. Science. 1990;247(4940):322-4.

126. Hinoi T, Akyol A, Theisen BK, Ferguson DO, Greenson JK, Williams BO, et al. Mouse model of colonic adenoma-carcinoma progression based on somatic Apc inactivation. Cancer Res. 2007;67(20):9721-30.

127. Korinek V, Barker N, Morin PJ, van Wichen D, de Weger R, Kinzler KW, et al. Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. Science. 1997;275(5307):1784-7.

128. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature. 2012;488(7413):660-4.

129. Giannakis M, Hodis E, Jasmine Mu X, Yamauchi M, Rosenbluh J, Cibulskis K, et al. RNF43 is frequently mutated in colorectal and endometrial cancers. Nat Genet. 2014;46(12):1264-6.

130. Madan B, Ke Z, Harmston N, Ho SY, Frois AO, Alam J, et al. Wnt addiction of genetically defined cancers reversed by PORCN inhibition. Oncogene. 2016;35(17):2197-207.

131. Shinmura K, Kahyo T, Kato H, Igarashi H, Matsuura S, Nakamura S, et al. RSPO

fusion transcripts in colorectal cancer in Japanese population. Mol Biol Rep. 2014;41(8):5375-84.

132. Schatoff EM, Leach BI, Dow LE. Wnt Signaling and Colorectal Cancer. Curr Colorectal Cancer Rep. 2017;13(2):101-10.

133. Han T, Schatoff EM, Murphy C, Zafra MP, Wilkinson JE, Elemento O, et al. R-Spondin chromosome rearrangements drive Wnt-dependent tumour initiation and maintenance in the intestine. Nat Commun. 2017;8:15945.

134. Chen L, Lu D, Sun K, Xu Y, Hu P, Li X, et al. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. Gene. 2019;692:119-25.

135. Patra R, Das NC, Mukherjee S. Exploring the Differential Expression and Prognostic Significance of the COL11A1 Gene in Human Colorectal Carcinoma: An Integrated Bioinformatics Approach. Front Genet. 2021;12:608313.

136. Sahu A, Prabhash K, Noronha V, Joshi A, Desai S. Crizotinib: A comprehensive review. South Asian J Cancer. 2013;2(2):91-7.

137. Qian H, Gao F, Wang H, Ma F. The efficacy and safety of crizotinib in the treatment of anaplastic lymphoma kinase-positive non-small cell lung cancer: a meta-analysis of clinical trials. BMC Cancer. 2014;14:683.

138. Lev A, Deihimi S, Shagisultanova E, Xiu J, Lulla AR, Dicker DT, et al. Preclinical rationale for combination of crizotinib with mitomycin C for the treatment of advanced colorectal cancer. Cancer Biol Ther. 2017;18(9):694-704.

Abstract in Korean

1. 국문요약

현재까지도 암은 전세계적으로 높은 발병률과 치사율을 보이고 있다. 차세대 염기서열 시퀀싱 기술의 개발을 비롯한 암 생물학 분야의 획기적인 발전은 새로운 진단 및 치료 방법들의 개발을 가능하도록 하였다. 최근에는, 융합 유전자 (gene fusion)들이 각종 암종에서 빈번하게 발견되면서 주요 치료 타겟 및 암의 예후 마커로써 활용되고 있다. 딥 시퀀싱 기술의 발전을 통해 약 1만개의 융합 유전자가 최근 5년 동안 발견되었지만, 이들 중 90% 이상은 분자병리학적인 메커니즘 연구 및 치료제 개발이 제대로 진행되지 못한 상황이다.

분자 메커니즘 연구 및 잠재적인 치료 마커의 탐색은 모두 The Cancer Genome Atlas(TCGA) 데이터베이스에 의해 구성된 암 유전체 빅데이터에 의해 빠른 속도로 발전할 수 있었다. 이에, 임상의 및 암 연구자들에게 소프트웨어 능력과 관계없이 복잡한 암 유전체 데이터에 쉽게 접근, 분석, 시각화 그리고 해석까지 도와주는 다양한 웹 리소스들이 개발되었다. cBioPortal, miRGator v 3.0, TANRIC 및 ISO express는 연구자들이 TCGA 데이터베이스를 분석하는데 다양한 기능을 제공하는 대표적인 웹 포탈들이다. 따라서, 이번 연구에서는 새로운 생물정보학 분석기법을 기반으로 약 1만개의 융합 유전자들의 분자병리학적인 특징들을 분석하고, 잠재적인 치료 타겟까지 탐색하는 웹 포탈인, DRPORTAL을 개발하였다. DRPORTAL은 TCGA로부터 33개 암종에 대한 유전체 및 임상 빅데이터, Jackson laboratory의 유전자 융합 빅데이터, CIViC과 OncoKB의 항암제 빅데이터, 그리고 ConsensuPathDataBase로부터의 세포신호경로 빅데이터를 활용하였다. DRPORTAL은 4단계에 걸쳐 융합 유전자 양성암의 잠재적 약물 후보들을 분석한다. 1) 먼저 융합 유전자 양성 환자군과 음성 환자군에서 연령, 성별, 생존 상태, TNM 암 단계, 유전자 변이 정보 등의 임상병리학적인 특징들을 분석하고, DEG (Differentially Expressed Genes) 분석을 통해 융합 유전자와 상관관계가 높은 유전자군들을 추출한다. 이후, 2) 높은 관계성을 가지는 유전자들을 기반으로 Pathway 분석을 진행하여 주요 암

관련 신호 경로들을 식별한다. 그리고 3) CIViC 및 OncoKB 약물 데이터베이스를 활용하여 약물 표적 네트워크를 구축하고, 4) 마지막으로 선별된 항암제 후보물질들의 우선 순위를 지정해준다.

DRPORTAL에 사용된 생물정보학 분석기법의 신뢰성을 검증하기 위해 *ESR1-CCDC170* (E:C) 융합 유전자 양성 유방암 연구와 *PTRPRK-RSPO3* (P:R) 융합 유전자 양성 대장암 연구를 진행하였다.

E:C 융합 유전자 양성 유방암 연구를 진행하기 위해 유방암 환자의 mRNA 발현량 데이터에서 11명의 E:C 융합 유전자 양성 환자 샘플과 48명의 음성 환자 샘플을 분석하였다. 그 결과, E:C 융합 유전자 양성 유방암의 임상병리학적인 특징이 triple-positive 유방암과 비슷했으며, basal-type 유방암과는 상호 배타적인 관계인 것으로 확인되었다. 또한 6개 주요 암세포 신호 경로들이 타겟 경로들로 식별되었다 (p53, ATR/ATM, FOXM1, Hedgehog, Cell cycle, Aurora B 관련 세포 신호 경로). 타겟 유전자들중 8개 유전자 (*AURKB, HDAC2, PLK1, CENPA, CHEK1, CHEK2, RB1, MDM2*)들이 E:C 융합 유전자의 RNA 발현량과 상관관계가 높았고, 3개 이상의 암 세포 신호 경로들에서 공통적으로 발견되었다. 마지막으로, 약물 데이터베이스로부터 21개의 후보 물질들을 타겟 유전자들과 재배치시켰으며, 그 중 palbociclib, alpelisib, ribociclib, dexamethasone, checkpoint kinase inhibitor AZD7762, irinotecan, milademtan tosylate, R05045337, cisplatin, prexasertib, olaparib 약물들은 2개 이상의 암세포 신호 경로에 포함된 유전자들을 타겟하고 있었다.

P:R 융합 유전자 양성 대장암 연구를 진행하기 위해 mRNA 발현량 데이터에서 7명의 P:R 융합 유전자 양성 환자 샘플과 50개의 음성 환자 샘플을 분석하였다. 그 결과, 앞선 유방암 연구와 같은 DEG 분석을 통해, 2,505개의 유전자가 P:R 융합 유전자와 상관관계가 높은 유전자군으로 식별되었다. 이 유전자군을 기반으로 Pathway 분석을 진행하였고, 10가지 주요 암 세포 신호 경로들(Apoptosis, Direct p53, EGFR, ErbB, JAK-STAT, tyrosine kinases, Pathways in Cancer, SCF-KIT, VEGFR 및 WNT 관련 세포 신호 경로)이 확인되었다. 선별된 세포 신호 경로들 중에서 P:R 융합 유전자에 특이적이고 여러 암 경로에 동시에 관여하는 주요 타겟 유전자 (*ALK, ACSL3, AXIN, MYC, TP53, GNAQ, ACVR2A, FAS*)들은 P:R 융합 유전자 양성 대장암에서 핵심적인 역할을 하는

것으로 예상한다. 마지막으로, 약물 데이터베이스로부터 crizotinib, alectinib, lorlatinib, brigatinib, ceritinib, erdafitinib, infigratinib 그리고 pemigatinib 약물들이 타겟 유전자들과 연결되었으며, 해당 약물들은 이미 다른 암종에서 상용화되고 있기 때문에 이번 연구에서 최종 약물 후보 물질들로 선정되었다.

결과적으로, 이번 연구를 통해 임상의와 여러 연구자들이 생물정보학 분석능력에 관계없이 약 1만개 융합 유전자들에 대한 유전체 데이터를 쉽게 분석, 해석 그리고 시각화까지 할 수 있을 것으로 예상하며, 더 나아가 아직 분자병리학적인 메커니즘이 밝혀지지 않은 융합 유전자 양성 암종들에 대해 잠재적인 약물후보물질들을 제공해줄 수 있다는 점에서 정밀 의료의 발전과 암 치료 개선에 기여했다고 예상한다.

주요어 : 생물정보학, 약물재배치, 융합 유전자, 전산 유전체학, 전사체학, DEG 분석

학 번 : 2019-20994