Optimization of Epigenome-wide CRISPR-Cas9 Knockout Screen Analysis to Prioritize Cancer Therapeutics

by

Ezgi Kurt

A Dissertation Submitted to the
Graduate School of Health Sciences
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Cellular and Molecular Medicine

KOÇ
ÜNİVERSİTESİ

Ocak 31, 2023

Optimization of Epigenome-wide CRISPR-Cas9 Knockout Screen Analysis to Prioritize Cancer Therapeutics

Koç University

Graduate School of Health Sciences

This is to certify that I have examined this copy of a master's thesis by

Ezgi KURT

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Prof. Tuğba BAĞCI ÖNDER (Advisor)

Prof. Tamer ÖNDER

Assoc. Prof. Şerif ŞENTÜRK

Assoc. Prof. Gözde KORKMAZ

Assist. Prof. Hamzah SYED (Co-Advisor)

Date: 31/01/2023

To my mom and my sister

ABSTRACT

Optimization of Epigenome-wide CRISPR-Cas9 Knockout Screen Analysis to Prioritize Cancer Therapeutics

Ezgi Kurt

Master of Science in Cellular and Molecular Medicine

January 31, 2023

Cancer target identification has been expanded by genome-wide, high-throughput CRISPR knockout screens. Using CRISPR-Cas9 knockdown screening, cancer cell survival genes are identified, and new targeted treatments are developed. Genome-scale knockout screens help with the discovery of essential genes that are needed for cancer cell growth. The examination of the screening data produced by this newly developed technology offers several difficulties.

A variety of epigenetic modifiers are the focus of the epigenetic knockout library EPIKOL. Five EPIKOL screenings in two separate cell lines were employed in this study (Prostate and triple-negative breast cancer).

Variations in sample size, sgRNA knockout efficiency, and the distribution of read counts make it difficult to interpret findings from CRISPR-Cas9 knockout screen data. Since off-target effects can cause drug development to progress in the wrong direction, it is particularly crucial to comprehend what the results of a genetic screen indicate. Therefore, a promising computational algorithm capable of handling various screening library types and read counts is required. There are multiple methods for genome-wide CRISPR screen analysis; however, not all of them are suitable for small-scale screening libraries. False positives appear to be widespread in small-scale library screen analysis. Our objective is to create the most efficient method for screening small-scale CRISPR libraries.

To achieve this objective using EPIKOL data, we compared the outcomes of three distinct methods. The implemented algorithms are MAGeCK, CRISPRcleanR, and BAGEL2. This study illustrates that altering the method of normalization or differential expression analysis can enhance the number of hit genes in small-scale libraries. According to the results of the analysis, BAGEL2 discovers more hit genes than other approaches.

ÖZETÇE

Kanser Tedavilerine Öncelik Vermek İçin Epigenom Genelinde CRISPR-Cas9 Tekniğinin Analizinin Optimizasyonu

Ezgi Kurt

Hücresel ve Moleküler Tıp, Yüksek Lisans

Ocak 31, 2023

CRISPR sistemi ile yapılan genomik çalışmalar kanser tedavilerinde kullanılmak üzere ilaçların geliştirilmesinin hızlanmasına yardım eden bir tekniktir. CRISPR-Cas9 tekniği kanser hücrenin hayatta kalmasında önemli olan genlerin bulunmasında ve bunlara yönelik tedavilerin uygulanmasında kullanılır. Nispeten yeni bir teknoloji olduğu için, bu teknoloji tarafından elde edilen verilerinin analiz edilmesi birkaç sorunu beraberinde getirir. EPIKOL, çok çeşitli epigenetik değiştiricileri hedefleyen, epigenom çapında bir CRISPR kütüphanesidir. Bu çalışmada iki farklı hücre hattında (Prostat ve üçlü negatif meme kanseri) beş EPIKOL verisi kullanıldı.

CRISPR-Cas9 deneylerinin analizindeki zorluklar; kullanılan örneklerin kopya sayısının az olması, sgRNA'lerin etkinliği ve okuma sayılarının dağılımı deneyler arasında farklılık gösterebilir. Hedef dışı etkilerin ilaç geliştirme çabalarını yanıltabileceği düşünüldüğünde, genetik tarama verilerinin doğru biçimde yorumlanması özellikle önemlidir. Bu nedenle, farklı tarama kütüphane türlerini ve okuma sayılarını işleyebilen güçlü bir hesaplama algoritması gereklidir. Genom çapında CRISPR tarama analizi için çeşitli yaklaşımlar mevcuttur, ancak bunların tümü küçük ölçekli tarama kütüphanelerine uygun değildir. Küçük ölçekli kütüphanelerle yapılan deneylerin analizinde yanlış pozitiflerin yaygın olduğu görülmektedir. Amacımız, küçük ölçekli CRISPR kütüphanelerine uygun en iyi sonucu veren algoritmayı tanımlamaktır.

EPIKOL verilerini kullanarak bu amaca ulaşmak için üç farklı algoritma kullandık ve sonuçlarını karşılaştırdık. Kullanılan algoritmalar: MAGeCK, CRISPRcleanR ve BAGEL2. Bu çalışma, normalizasyon yönteminin veya gen önem sıralanması için kullanılan yöntemin değiştirilmesinin, küçük ölçekli kütüphanelerde daha güvenilir gen listesinin bulunabileceğini göstermektedir. Analiz sonuçları, BAGEL2'nin diğer yöntemlere göre daha fazla sağ kalım için önemli genleri tanımladığını ve analizde kullanılan algoritmanın normalizasyon yönteminin değiştirilmesinin, daha az yanlış pozitif gen bulunan sonuç verebileceğini göstermektedir.

AKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude and appreciation to my advisors, Prof. Tuba Bağcı Önder and Asst. Prof. Hamzah Syed. I'd would like to express my thankfulness to them for their guidance, support, understanding, and encouragement throughout my master's degree. I also would like to thank my thesis defence committee members Prof. Tamer Önder, Assoc. Prof. Şerif Şentürk and Assoc. Prof. Gözde Korkmaz. My sincere thanks to my friend and KUBioCore member Begüm Durdar for being with me throughout this journey.

I owe the greatest debt of gratitude to my parents, Hatice and Mehmet Ali, who have stood alongside me through everything, who have never wavered in their belief in me, and who have always pushed me to do my best. I can't express how grateful I am to my sister, Özge, for always being there for me and supporting me in any way she can. She was always there to pick me up when I was feeling down. I couldn't have done it without her support. I am extremely happy to have her as my big sister.

I would like to thank my psychologist, Seval Yarış Bakır, for years of encouragement and support as I worked to become a better version of myself. Her assistance has been crucial to my success. I probably wouldn't be here without her help.

I'm also grateful for my dear friends, İsra Atan, Döndü Selcan Akkuş and Elhan Taka who have always been there for me when things were hard.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BAGEL	Bayesian Analysis of Gene Essentiality
BF	Bayes Factor
CCR	CRISPRcleanR
CN	Copy Number
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DSB	Double Strand Break
EPIKOL	Epigenetic Knockout Library
FDR	False Discovery Rate
LFC	Log Fold Change
FN	False Negative
FP	False Positive
MAGeCK	Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout
NGS	Next Generation Sequencing
PCa	Prostate Cancer
PDT	Population Doubling Time
TNBC	Triple Negative Breast Cancer
sgRNA	Single Guide RNA
PDT	Population Doubling Time
RRA	Robust Ranking Aggregation
TP	True Positive
TPM	Transcript Per Million

Chapter 1:
INTRODUCTION

The second most prevalent cause of death in the world is cancer. It kills around 8 million individuals worldwide each year (Bray et al., 2013; Tarver, 2012). Cancer is a hereditary disease produced by alterations in the genome that can arise in any part of the human body. Causes for these changes can be errors that occur during cell division, the genes which are inherited from the parents, lifestyle, or environmental substances. All cancer types have different causes behind it. Therefore, it is difficult to understand the underlying mechanism and develop new therapeutic drugs.

Changes in the epigenetic control of genome activity have a role in cancer that is almost as important as changes to the genomic coding information. Genes' chromatin states, which serve as on/off switches for gene expression, are established by epigenetic alterations to DNA and histones as well as by changes in chromatin-remodeling pathways. It changes how a locus or chromosome works but doesn't change the DNA sequence. Several studies have found that cancer genomes often have epigenetic changes (Esteller, 2002; Handy et al., 2011; Plass et al., 2013; Zhou et al., 2018).

Changes at the epigenome level allow pre-malignant cells to acquire cancer-like features such as aggressiveness, environmental sensitivity, and resistance to treatment (Hanahan & Weinberg, 2011). To illustrate, in many carcinoma, DNA hypermethylation blocks tumor-suppressor expression, while DNA hypomethylation stimulates protooncogene expression (Esteller, 2008; Kuang et al., 2020; Kulis & Esteller, 2010; Saghafinia et al., 2018; Xu et al., 2019). Many epigenetic modifiers associated with cancer have been found in recent cancer genome sequencing studies (Plass et al., 2013). During the formation of malignant tumors, it is believed that driver mutations increase cellular plasticity. Since epigenetic regulators play such a vital role in cancer, small-molecule inhibitors that target them are promising anticancer agents. These therapies have shown effectiveness against a range of cancer types (Ganesan et al., 2019). First-generation drugs have had minimal therapeutic advantages due to their high toxicity (Linnekamp et al., 2017; Stahl et al., 2016). Clinical studies are being carried out to evaluate novel compounds in order to address these limitations (Berdasco & Esteller, 2019; Cheng et al., 2019; Mohammad et al., 2019; Pfister & Ashworth, 2017). Using

epigenetic inhibitors, it is possible to target cancer in a particular way by taking advantage of the synthetic lethal interaction between epigenetic modifiers (Morel et al., 2017, 2020; Pfister & Ashworth, 2017; Wilson et al., 2010; Wu et al., 2018; Yang et al., 2019).

CRISPR/Cas9 genome editing technology is fast, easy, and effective and it has drastically accelerated the study of functional genomics (Doudna & Charpentier, 2014). Multiplexed sgRNA libraries are created easily to examine gene functions in pooled genome-wide knockout screens (Sanjana et al., 2014; Wang et al., 2014). Under certain selective pressure, positive selection screens can determine essential genes (Shalem et al., 2014; Wang et al., 2014) while negative selection screens identify essential genes in various contexts (Evers et al., 2016b; Hart et al., 2015; Wang et al., 2015). Even though several phenotypes can be examined together at the same using genome wide CRISPR knockout libraries, designing and executing out such experiments are time consuming and expensive. Secondary screens focusing on the pathways found in the initial screen are typically carried out in order to eliminate false-negative outcomes and obtain high confidence leads. In contrast to genome-wide libraries that contain a fixed number of sgRNAs per gene, sgRNA counts per gene may be increased in targeted libraries to improve the accuracy of the acquired phenotype. As a consequence, targeted sgRNA libraries have been created as a method to solve these difficulties by decreasing costs and time while simultaneously enhancing yield and signal-to-noise ratio (Doench, 2018; Miles et al., 2016).

EPIKOL, or the Epigenetic Knockout Library, is a CRISPR-Cas9 library that targets a vast array of epigenetic regulators and cofactors (Yedier-Bayram et al., 2022). sgRNAs were used to target chosen epigenetic modifiers from the EpiFactors database to develop the EPIKOL (Medvedeva et al., 2015). Using previously created genome wide CRISPR knockout libraries, EPIKOL employs a set of 10 sgRNAs to target a single gene at a time (Doench et al., 2016; Sanjana et al., 2014). Additional sgRNAs were produced using the CCTop and E-CRISP tools in cases where the total number of sgRNAs per gene was fewer than 10 as a result of overlapping sequences in previous libraries (Heigwer et al., 2014; Stemmer et al., 2015).

The goal of EPIKOL is to target a wider variety of epigenetic modifiers. Due to its comprehensive epigenome-wide targets and comparatively high number of sgRNAs per gene, this library will make it simpler to undertake studies examining the functional

effects of epigenetic modifiers in a range of situations. 719 epigenetic modifiers, 25 context-specific controls, 35 pan-essential genes, and 80 non-targeting controls are the targets of the 7870 sgRNAs that make up EPIKOL (Yedier-Bayram et al., 2022).

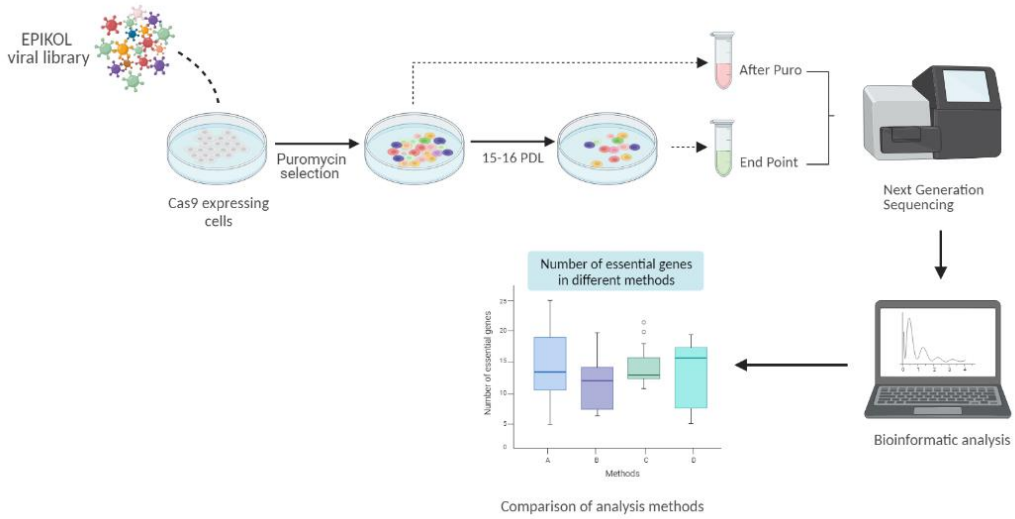


Figure 1 Overview of the screening procedure for EPIKOL.

In each screen, EPIKOL was used to infect Cas9-expressing stable cell lines at a low multiplicity of infection. Cells were collected after 3 days of library infection with puromycin selection, which was used to provide a baseline for the sgRNA distribution. These samples are known as After Puro (AP) samples. The rest of the cells were kept in the culture until they reached 15–16 population doubling times (PDTs). When they reached the required PDT, another sample was collected. This served as the experiment's end point samples (EP) (**Figure *1***). The cells collected in each time point compared to see the sgRNA depletion. Next generation sequencing (NGS) result is further analyzed to identify significantly depleted genes.

The current chapter introduces the concept of CRISPR-Cas9 screens in cancer studies using EPIKOL library, and the methods used in the analysis of CRISPR screen data. The limitations with current CRISPR data analysis tools are outlined. The introduction concludes with defining the aims and explaining the motivation behind the research in this thesis.

1.1 Background to CRISPR-KO Screens

Archaea and bacteria have developed adaptive immune mechanisms, called clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated

(Cas), that are used to degrade foreign nucleic acids with directing of short RNAs (Mali et al., 2013). Genome editing techniques based on CRISPR-Cas9 are changing the field of genetics (Cong et al., 2013; Mali et al., 2013). Because of its great efficiency and specificity, the CRISPR-Cas9 system is used to mutagenize genes by inducing DNA double-strand breaks (DSB), either at the gene level or on a genome-wide scale. This allows the systematic study of loss-of-function phenotypes (Iorio et al., 2018).

Researchers have created genome-wide screening tools for CRISPR knockouts (CRISPR-KO) (Koike-Yusa et al., 2014; Shalem et al., 2014; Wang et al., 2014). With this method, essential cancer-related genes are being screened on a large scale (Iorio et al., 2018). CRISPR-KO screens are often used to detect essential genes for cancer cell viability and to create novel targeted therapies. Cas9 endonuclease is delivered into cells, followed by or in combination with the introduction of a targeted sgRNA pool library. The genes of interest are targeted by a library containing multiple single guide RNA (sgRNA) for each gene to enable a robust identification of gene essentiality. Researchers can determine which specific sgRNAs are differentially expressed in control and test samples with the help of analysis strategies, thereby targeting a gene that could be critical to the survival of cancer cells.

The CRISPR-KO screening system has a mode of action drawback which is DSB induction. The DNA damage response triggered by DSBs may lead to cell cycle arrest and, in some cases, cell death (Aguirre et al., 2016; Munoz et al., 2016; Wang et al., 2014). This becomes a problem for doing whole-genome CRISPR-KO screening in cancer cells since the CN of the cancer cells' genomes changes regularly, leading to extensive Cas9 DNA damage. Therefore, depletion of these genes occurs by DSBs at the genes amplified region in a pooled CRISPR-KO screen. This happens regardless of their essentiality, so they are mistakenly referred to as fitness genes. The CN-associated effect can lead to a high false-positive rate, so correcting these effects is crucial for interpreting the results of CRISPR-KO screening. Several approaches, such as scanning the dataset for biased regions and eliminating them from subsequent analysis, have been presented. (Wang et al., 2015), which results in the exclusion of CN-amplified regions containing potentially biologically relevant genes or using a piecewise linear model to infer true gene dependences based on CN profiles across large cell lines panels (Meyers et al., 2017). Compared to the shRNA knock-down technology, CRISPR-Cas9 system provides greater specificity and effectiveness in gene knockout (Evers et al., 2016b;

Hart et al., 2015). Using genome-scale knockout screens, it is possible to unbiasedly identify genes whose disruption inhibits proliferation relative to wild-type cells ("essential genes"), and to curate pan- and context-dependent essential genes used to identify potential drug targets for specific tumor genotypes (Aguirre et al., 2016; Behan et al., 2019; Hart et al., 2015; Meyers et al., 2017; Steinhart et al., 2016; Wang et al., 2015, 2017). In targeted drug development, off-target effects can mislead the whole experiment, which makes the analysis of genetic screening data important (Lin et al., 2019).

Three steps make up the analysis pipeline for a loss-of-function CRISPR screen: First, reads are mapped to the CRISPR library's guide sequences to create a table of read counts. Next, counts across samples are normalized and fold changes at the guide level are calculated. Finally, guide level data is combined with gene level fitness scores.

1.2 Challenges which can occur in the analysis of CRISPR screens

CRISPR/Cas9 knockout technology is a powerful tool for eliminating genes at the DNA level. Since it’s a rather newly discovered technology, analysis of the screening data generated by this technology demonstrates several challenges.

Firstly, these studies are conducted with a small number of replicates. Replicates help to boost the power to detect associations using statistical models that calculate the read count variance and evaluate the statistical significance of the difference between the treatment and control samples. In both negative and positive selection experiments, there is a substantial degree of variability in the observed sgRNA abundance (Li et al., 2014). Usually, count data is analyzed using a negative binomial model.

Second, various sgRNAs targeting the same gene may have distinct properties and knockdown efficiency. A comprehensive approach that takes these parameters into account is required for the collecting of information from several sgRNAs (Fu et al., 2013; Hsu et al., 2013).

Thirdly, depending on the experiment design and its screening library, the distribution of the read counts of the CRISPR/Cas9 knockout screening experiment will be different. Therefore, it requires a powerful normalization method, which can handle the different types of screening libraries and its read counts.

There are some competing algorithms which have been developed for CRISPR screens such as edgeR's "processAmplicons" function, BAGEL2 and CRISPRcleanR. In below there is a brief overview of the existing algorithms used for CRISPR knockout screen analysis. Each algorithm is accompanied by a short explanation of its intended usage and a summary of the mathematics used in its study.

MAGeCK stands for Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK) screens. It is a commonly used method for the analysis of CRISPR/Cas9 screening data (Li et al., 2014). It identifies essential sgRNA, genes and pathways from the pooled CRISPR/Cas9 knockout screens. Essentiality determines the genes that are essential for a cell to survive. If that gene is eliminated from the cell, the cell will not survive. MAGeCK has a high degree of sensitivity and can distinguish between biologically significant pathways and cell type-specific essential genes. It can simultaneously perform positive and negative selection screens. Compared to previous techniques based on high-throughput sequencing counts, MAGeCK detects substantially chosen sgRNAs and genes with high sensitivity. In addition to being robust to varying sequencing depths and different numbers of sgRNAs targeted to each gene, this will also allow for more cost-effective CRISPR/Cas9 knockout screening studies.

CRISPRcleanR is a computational framework created in the open-source computer languages Python and R. It detects biased genomic areas from CRISPR-KO screening and provides corrected read counts and log fold change (LFC) values of individual sgRNAs within biased genomic regions. It decreases false positives while preserving the true positive rate of known essential genes constant, and it enables the identification of essential genes even within focally amplified regions (Iorio et al., 2018).

BAGEL2 is the improved version of the Bayesian Analysis of Gene Essentiality (BAGEL). BAGEL is a method that used to identify differentially expressed genes in CRISPR knockout screens (Hart et al., 2014; Hart & Moffat, 2016). It will first calculate the log likelihood of a gene which belongs to essential or non-essential class. It produces a log Bayes Factor (BF) based on statistical confidence and biological effect size. The classifier is trained using a gold standard reference collection of predicted core-essential and non-essential genes. These sets come from gene expression and genetic screen research (Hart et al., 2014, 2017). When compared to other algorithms, these classifier provide an unbiased setting.

The edgeR package provides a comprehensive solution for screen data analysis. It begins with unprocessed sequence reads and produces a list of candidate genes for further biological evaluation. The raw data from a fastq file is first transformed into a matrix of counts, with the option to allow for modest changes in sequence position and mismatches. Diagnostic charts, normalizations, and differential representations may be used to prioritize data statically using either exact testing or generalized linear modeling, which can handle complicated experimental designs (Ritchie et al., 2014).

Copy number variations tend to be large in cancer cells. When Cas proteins are used to cut copies of DNA with high copy numbers, there can be gene-independent effects due to cell death caused by DNA damage. Meyers et al. proposed CERES to perform this correction computationally. By comparing each gene's guides across all cancer cell lines evaluated, the CERES algorithm normalizes for cell line-specific alterations. After that, the CERES model is fitted to the experimental data using an alternating least squares regression for each gene across all cell lines. However, a CERES experiment can only be undertaken when several screens are completed on diverse cell types with established copy number profiles. Otherwise, copy number effects cannot be detected.

1.3 Objective and Thesis Outline

The main objective of this study is to design and execute an optimal analysis protocol and pipeline with application to the EPIKOL data. To achieve this, our secondary objective is to benchmark the algorithms comparing computational tools for normalization, read counting, quality control, differential representation analysis and gene level analysis. The other secondary objective of this thesis is to identify and prioritize targets in tumors while expanding the range of targets, improving the success rate, and accelerating the development of new cancer therapies. The results of this study will allow for the generation of a framework to prioritize cancer drug targets and suggest specific new epigenetic targets.

Chapter 2:
METHODOLOGY AND SOFTWARE FOR CRISPR-SCREEN ANALYSIS

CRISPR screens are a great way to find new druggable targets and find out which genes are being targeted by existing drugs, but there are still some problems with how they are analyzed. Analysis may become more challenging due to variations in guide RNA efficiency. Variable gene impact sizes can lead to a bias toward identifying those genes with large effects. Excessive cutting in high copy number areas might cause cell death and which leads to false positives in CRISPR-KO screening. These issues make it more difficult to examine the data and often result in false positives and false negatives. Although numerous algorithms have been developed with the aim of eliminating these issues, it is still remains to be explored how these algorithms affect the analysis and inference of biological results and how well these algorithms compare to one another.

Table 1 Properties of software tools used for analysis of CRISPR screens.

	MAGeCK	CRISPRcleanR	BAGEL2	CERES	edgeR
User interface	Command line	R	Python	R	R
Input format	Demultiplex FASTQ files; read counts	Read counts	Log fold changes	sgRNA-depletion and copy number data	Demultiplex FASTQ files
Single or multiple screens	Both	Single screen	Single screen	Multiple screens	Single screen
Guide RNA ranking method	Negative binomial	Median-ratio method	Not applied to guide RNA ranking	Hyperparameter optimization	Negative binomial
Gene ranking method	α-RRA P value	Not applied to gene ranking	Bayes factors	Average guide score	Not applied to gene ranking

The step they used	Normalization and esentiality measurements	Normalization	Esentiality measurements	Estimate gene-dependency levels while adjusting for copy number	Normalization and differential representation

In this section, some of the programs developed for the CRISPR screens are going to be explained. The programs are CRISPRcleanR, BAGEL2, MAGeCK, edgeR, CERES. Their application, prerequisites, and role in the analysis will be discussed.

2.1 CRISPRcleanR

Functional genetics gene inactivation investigations employ the CRISPR-Cas9 system as a tool. It has become feasible to study which genes, when inactivated, are capable of specifically lowering the viability of cancer cells (Evers et al., 2016a; Hart et al., 2015; Wang et al., 2014). To enhance inactivation efficiency and minimize topological biases, CRISPR-Cas9 libraries use multiple sgRNA targeting the same gene (Haeussler et al., 2016).

Comparing sgRNA abundances in control and test samples may help researchers determine which sgRNAs are produced differentially in the experiment, therefore enabling them to target a gene that may be crucial for cancer cell viability. Novel drug targets can be identified through these types of screens and their analysis (Steinhart et al., 2016; Tzelepis et al., 2016). Double-strand breaks are the primary downside of the CRISPR-KO screening technique. When there are double strand breaks in DNA, it may induce cell cycle arrest and possibly death (Aguirre et al., 2016; Munoz et al., 2016; Wang et al., 2017). During genome-wide CRISPR-KO screening, cancer cells have frequent CN mutations, leading in significant Cas9-induced DNA damage. Double strand breaks in amplified areas of the genome result in gene depletion regardless of essentiality; these genes are usually referred to as fitness genes. False-positive rates are high because to CN variations. This effect must be considered in order to fully comprehend the results of CRISPR-KO screening (Iorio et al., 2018). Scanning biased regions of the data and removal of them from downstream analysis has been proposed (Wang et al., 2015). Because of that, biologically relevant genes in the CN-amplified regions may be excluded. Piecewise linear model may be used to infer true gene

dependencies from large panels of cell lines with CN profiles to fix the CN alterations (Meyers et al., 2017).

CRISPRcleanR package address this copy number associated effect and corrects it. To identify biased genomic areas in CRISPR-KO screens, the software employs an unsupervised approach. Within these areas, the program offers both corrected read counts and log fold changes (LFC) of each sgRNA. False positives are minimized, but true positives for known essential genes remain substantially intact. With this method, essential genes can even be detected within focally amplified regions (Iorio et al., 2018). This program is not a replacement for MAGeCK but it can be used as an additional step.

CRISPRcleanR uses a circular binary segmentation algorithm, which is developed to analyze array-based comparative genomic hybridization assays (AB et al., 2004; ES & AB, 2007). In a cell line, this binary segmentation is applied directly to the genome-wide patterns of sgRNA LFCs across individual chromosomes. Then, it detects multiple sgRNAs with adequately equal LFC in the genomic segments. Using mean-centering, LFCs values are corrected if these segments contain at least a couple of sgRNAs targeting different genes, indicating that CRISPR-Cas9 targeting a gene occur independently (in segments that contain sgRNAs targeting a minimum number of genes, CRISPR-Cas9 targeting is likely to occur in a gene-independent manner, and LFCs values are corrected using mean-centered analysis.).

Functions from the DNAcopy R package are embedded inside the CRISPRcleanR algorithm (*Bioconductor - DNAcopy*, n.d.). In addition, it possesses several features that make it versatile, statistically robust, and also practical for downstream applications. These features are listed as follows:

- It works in an unsupervised manner. There is no priori defined sets of essential genes or chromosomal CN information needed for it to work.
- By applying a LFC correction, all gene depletion scores can be used in subsequent analyses.
- LFC is analyzed at the sgRNA level to gain resolution and enable for various levels of on-target efficiency, enabling algorithms such as BAGEL that require input at the sgRNA level to determine gene depletion score (Hart & Moffat, 2016).
- Lastly, CRISPRcleanR uses data from individual cell lines to correct logFC values. Comparatively, other computational correction methods depend on

the number of cells analyzed to determine their performance (Meyers et al., 2017). These implies that CRISPRcleanR is ideal for the analysis of data not only from large scale CRISPR-KO studies but also from small scale CRISPR-KO studies.

Without any assumptions regarding gene essentiality, CRISPRcleanR can be used in an unsupervised setting. It has been showed that CRISRPcleanR does not have adverse effect on the overall essentiality profile of a cell line. The 3 functions from CRISPRcleanR package that are used in this study for normalizations are explained in the following sections.

Before CRISPRcleanR package could be used, EPIKOL library needs to have an index file which shows the sgRNA information. This file needs to be created.

2.1.1 ccr.NormfoldChanges Function

It is one of the many functions in CRISPRcleanR package. Normalizing the raw sgRNA count median-ratio and calculating the log fold-changes for the sgRNAs in the sample, ccr.NormfoldChanges function would be used. sgRNA counts in a data frame or in a tsv file are normalized by using the gene wise median of ratios method or scaling by total number of reads per sample. It calculates the log fold change between replicates of the transfected library and controls. A list of normalized counts and log fold-changes is returned as output. Two tsv file is produced at the end; one containing the normalized count, and the other containing log fold-changes information.

To run this function two files, need as an input. These files are library annotation and read count file.

Library annotation is a data frame containing the annotation of sgRNAs with a row for each sgRNA and each column has information on targeted genes, genomic coordinates which exon, chromosome, strand and start and end position of the sgRNA.

Read count file. This file is read into R as a data frame which contains raw sgRNA counts. It must have one row for each sgRNA, and the following columns headers are; sgRNA, which contains alphanumeric identifiers of the sgRNA under consideration; gene, containing HGNC symbols of the genes targeted by the sgRNA under

consideration; columns for controls and library-transfected samples sgRNA counts are listed after that.

Other parameters using in this function such as deciding the normalization method (ScalingByTotalReads or MedRatios), number of control replicates, minimum reads which defines a filter threshold value for sgRNA compared to the counts in the control sample, could differ according to purpose of the experiment.

2.1.2 ccr.logFCs2chromPos Function

This function maps genome-wide sgRNAs' log fold changes (averaged across replicates) on the genome and returns them sorted according to the position of their targeted region on the chromosomes.

It needs a foldchange file. This should be stored in a data frame. First column has information sgRNAs identifiers, usually the sequence of the sgRNAs. Second column has the HGNC symbols of the targeted genes for each sgRNA. The other columns contain sgRNAs log fold change, one column per library transfected replicate. ccr.NormfoldChanges function produces the log foldchange data from the raw count files. Secondly, the library annotation file needs to be inputted.

2.1.3 ccr.GWclean Function

The ccr.GWclean function, find and fix biased sgRNA log fold-changes that are thought to be caused by responses to CRISPR-Cas9 targeting that are not gene-dependent.

The corrected sgRNAs fold-changes are returned in a list (as a data frame), together with the annotation of the identified segments (in another data frame) and a vector of strings containing all the genome-sorted sgRNAs' identifiers. Additionally, it generates a plot for each chromosome that shows segments with the equal log fold-change in sgRNAs before and after the correction.

A pooled library of single guide RNAs (sgRNAs) targeting protein-coding genes that have been transfected into an in vitro model that is stably expressing Cas9 is used to create gene essentiality profiles. The essentiality profile measures the loss/gain-of-fitness brought on by each sgRNA-targeting and is expressed as the difference in LFCs. It is calculated by comparing the abundance of the sgRNAs in the plasmid pool used for

viral production, at the beginning of the experiment, or in any other control condition, and the abundance of the sgRNAs at the end point after cell purification. This function applies a circular binary segmentation method to the input genome-wide pattern of LFCs to discover genomic areas containing sgRNAs with sufficiently equal LFC and targeting a minimum number of distinct genes (AB et al., 2004; ES & AB, 2007). When targeting several contiguous genes, it is rare to see the same loss/gain-of-fitness impact; thus, the LFCs of these sections are adjusted if specific requirements are satisfied since they are thought to be biased by some property of the relevant genomic segment.

A data frame containing the log fold changes of all sgRNAs across the genome must be provided as input to this function. The "ccr.logFCs2chromPos" function can be used to create this from a data frame containing the log fold changes for sgRNAs that were obtained from raw sgRNA counts using the "ccr.NormfoldChanges" function.

2.2 BAGEL2

BAGEL2 calculates the log probability that a gene belongs to the "essential" or "nonessential" class in the context of a typical genome-scale knockout screen in a cell line and provides a log Bayes Factor (BF) that combines statistical confidence with biological effect magnitude. The classifier is trained using gold-standard reference sets of genes that are considered essential and non-essential. These data sets are based on genetic screening and gene expression investigations (Hart et al., 2014, 2017). This will serve as an unbiased benchmark for comparing the performance of other algorithms. The common reference non-essential and gene essential lists are available inside the algorithm.

BAGEL was modified in many ways to make it faster and more precise: Off-target effects of sgRNA were controlled, and the dynamic range of BFs was enlarged, allowing the discovery of tumor suppressor genes that induce improved cellular fitness when knocked out. While these changes were presented, the core algorithm remains intact. But copy number artifacts are not addressed by BAGEL2. Correct these effects CRISPRcleanR package is recommended by the developers (Iorio et al., 2018; Kim & Hart, 2021).

Depending on the user's selections, the BAGEL2 method resamples the dataset and splits it into a training set and a test set using either 10-fold cross-validation or bootstrapping.

A fold change distribution for essential and non-essential models is calculated using kernel density estimation for each iteration of sampling, using all the guides targeting control essential or non-essential genes in the sample. Log ratio of Pr(Ess)/Pr(Non) is used to estimate the log BF at the guide level.

The thresholds are calculated by BAGEL2 using a log decay function, which makes BAGEL2 more suitable for small CRISPR library screens.

Log Bayes Factors are calculated at sgRNA level using linear regression model, and replicate-level BF is summed at sgRNA level. Using the sum of sgRNA-level BF gene-level Bayes Factor is calculated.

BAGEL2 “bf” function used for calculating the essentiality from the fold change file. Later, benchmarking was carried out using the reference genes and the "pr" function in BAGEL2 (Hart et al., 2014).

2.2.1 Correcting multi-target effect

“bf” function of the BAGEL2 is used for correcting the sgRNA multi-target effects. The dropout phenotype of a guide RNA is caused by a combination of events, including the deletion of the target gene, locus independent DNA breakage, and off-target effects from another locus, perhaps involving non-additive genetic interactions. To remove guide-level effects unrelated to the targeted gene, a multi-targeting correction method has been built. The algorithm calculates and removes the "incremental BF" caused by off-target DNA cleavage within a single mismatch while eliminating the confusing effects of off-target gene deletion (Kim & Hart, 2021).

2.2.2 F-measure and false discovery rate (FDR) calculation

The cumulative mean of accuracy and recall at the Bayes Factor 5 threshold is known as the F-measure or F1 score (BF = 5), and it may be used to evaluate how well an essentiality prediction performed. F-measure calculated using a precision-recall table generated by BAGEL pr function.

$$Precision = \frac{TP}{TP + EP}$$

$$Recall = \frac{TP}{TP + FN}$$

TP = positives in reference + core-essential set

FP = positives in reference \kern-1ptnon-essential set

FN = negatives in reference core-essential set

Precision is a metric that indicates how many positive forecasts were right (true positives). Recall is an indicator of the number of positive instances in the data were properly predicted by the classifier. It is also referred to as sensitivity.

The first step in the CRISPR screen analysis is mapping raw sequencing reads to their corresponding CRISPR libraries. Read count file would be produced at the end of the mapping step. After that, CRISPRcleanR is used for normalization method. With reference to guide location and fold change, it has been used to remediate amplicon-induced artifacts. With the normalized counts, BAGEL2 functions could be proceeded.

```
BAGEL.py bf -i [fold change file] -o [output file] -e [reference essentials]
-n [reference nonessentials] -c [columns to test]
```

BAGEL “bf” function needs some input files to run. These files are:

- Foldchange file, which has the information of the screen data. The format of this file should be first column is the code of the sgRNA which is the same format as the CRISPRcleanR. Second column has the gene name of sgRNA target, the rest of the columns should contain the LFC of the sample.
- Output file name. This information needs to be mentioned.
- A text file containing the training set of essential genes must be supplied.
- Training set of nonessential genes need to be provided in a text file.
- The function must accept a set of input file columns to analyze, separated by commas.

At the end of this command, it generates a log2 Bayes factor for every gene; positive BFs suggest a high degree of certainty that the gene is essential.

The output file has; gene name, mean Bayes Factor across all iterations, std deviation of BFs, and how many times the gene was included in the test set for each iteration.

```
BAGEL.py pr -i [Bayes factors] -o [output file] -e [reference essentials] -n
[reference nonessentials]
```

BAGEL “pr” function needs files below to run. These files are:

- Output file, which should specify the name of the output file.
- Reference essentials file. This should have the information of known essential genes.
- Reference non-essential file. This file should have the information of known non-essential genes.

2.3 MAGECK

MAGeCK stands for Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK) screens. It is a commonly used method for the analysis of CRISPR/Cas9 screening data (Li et al., 2014).

MAGeCK is intended for use in genome wide CRISPR/Cas9 knockout investigations to discover both positively and negatively chosen short guide RNAs and genes. There are four steps in the MAGeCK procedure: i) read count normalization, ii) mean-variance modeling, iii) sgRNA ranking and iv) gene ranking.

To account for the impacts of library sizes (or sequencing depth) on read count distributions, the read counts from distinct samples are first normalized using the median ratio approach (Anders & Huber, 2010; Wang et al., 2014).

To establish the statistical significance of variations in sgRNA abundance between conditions, it is required to quantify the variance of the read counts within one condition (usually the control samples). As a result, sgRNA read counts estimated from an NB distribution are utilized to determine if sgRNA abundance changes substantially between treatments and controls.

P-values generated from the NB model are used to rank sgRNAs, and separately positive or negative genes are identified using a revised robust ranking aggregation (RRA) algorithm called α-RRA. If a gene has no influence on selection, the sgRNAs that target it should be uniformly spread among the top-ranked list of all sgRNAs, according to the α-RRA approach. The α-RRA prioritizes genes with higher sgRNA rankings than predicted based on the variance in sgRNA rankings observed as compared to the uniform null model. Permutation-based statistics are used by the α-RRA to determine the skew's statistical significance (Li et al., 2014).

2.4 edgeR

edgeR offers a complete analysis solution for screen data. It uses the raw sequence reads as input and returns a list of candidate genes that can be tested in the next step as the output. edgeR analysis involves a number of procedures that must be taken to ensure that high-quality raw sequence reads are used and that statistically significant changes in screen representation are discovered.

2.5 CERES

Cell proliferation assays in genome-wide CRISPR-Cas9 knockout screens have been discovered to be influenced by the genomic CN of the areas targeted by the sgRNA-Cas9 complex (Aguirre et al., 2016; Munoz et al., 2016; Wang et al., 2015). Cas9 may be used to target DNA sequences in high CN gain areas and produce numerous DNA DSBs, which block the G2 cell cycle and trigger a gene-independent DNA damage response (Aguirre et al., 2016). Even at modest levels of CN amplification and deletion, DNA cleavage has a systematic, sequence-independent influence on cell survival. The presence of many genomic amplifications in cancer cell lines prevents accurate interpretation of experiments performed on these cells. False positives are primarily caused by genes located in these regions (Aguirre et al., 2016; Munoz et al., 2016). The CERES approach, which estimates gene dependence from essentiality screens by computing copy-number effects, allows for precise interpretation of gene dependency at all CN levels.

The gene-knockout effect and the copy-number effect are combined together by CERES to calculate the level of sgRNA depletion. The CERES infers all the parameters through the fitting of the model to the observed data by using the alternating least squares regression method. Scores of 0 and 1 are produced from these predicted gene-knockout effects to indicate the median impact of common core essential genes and nonessential genes, respectively. The CERES model's piecewise linear spline functions allow for flexible simulation of the copy-number effect's saturation at large cut counts (Meyers et al., 2017).

All cell lines screened for sgRNA depletion and CN data are input to CERES. To determine depletion values as part of the inference process, CERES combines gene knockout and copy number effects with a guiding activity score parameter. Then,

CERES generates the parameter values that have the highest chance of producing data that is compatible with the model.

2.6 Study Pipeline

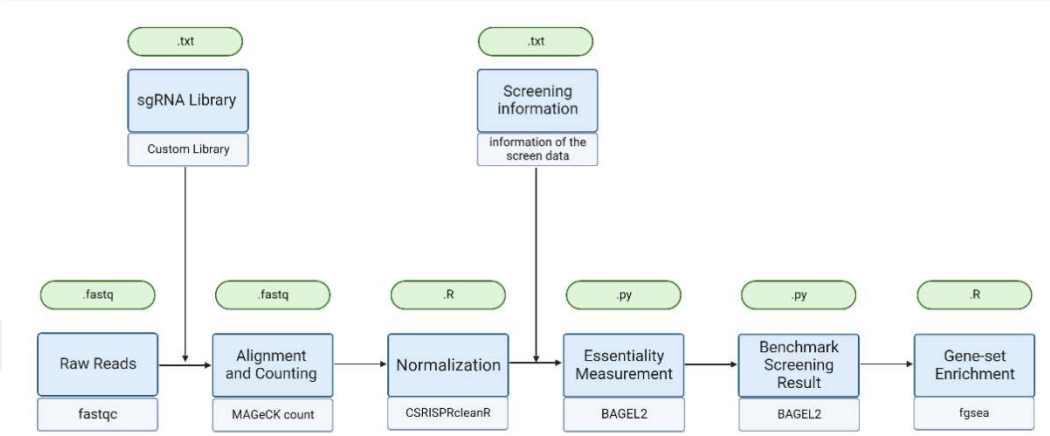


Figure 2 Optimal pipeline for CRISPR screen data analysis. It also shows the required format for each algorithms input.

The SRA database was used to download all of the data used in this study. The data are available under the accession number SRP318541. SRA toolkit used for downloading these FASTQ files. It is a command-line tool available in an Anaconda environment. The fasterq-dump function was used to download the files.

```
fasterq-dump --threads 3 --progress SRR14428521
```

After the download of the files, quality control with FastQC was done. FastQC is used for performing quality control tests on raw sequence data.

2.6.1 Annotation

The major output of next-generation sequencing for a CRISPR-Cas pooled screen is FASTQ files, that contain both sequence and quality data for each run. In the first step of pooled screening analysis, called deconvolution, the guide sequence is extracted from the read and aligned to the guide library. This generates a table listing the total number of readings per guide for each sample. MAGeCK program used for generating the count file. The MAGeCK count function observes and normalizes reads per sgRNA and normalize. Using the median ratio method, the read counts were modified to account for library size. It generates normalized counts from FASTQ files. To utilize this feature:

- The annotation file for the library must be entered. The library annotation file is a text file containing the unique identifiers and sequences of the targeted genes in the screens.
- Sequence files must be uploaded in the FASTQ format.
- Normalization method must be stated.
- Both the experiment and sample names must be supplied in the command.

Each sample replicate's reads per sgRNA and normalization is done separately in the count function.

```
mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
median --pdf-report -n lncap_AP1 --sample-label lncap_AP_R1 --fastq
/scratch/users/ekurt20/SRR14428514_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
median --pdf-report -n lncap_AP2 --sample-label lncap_AP_R2 --fastq
/scratch/users/ekurt20/SRR14428515_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
median --pdf-report -n lncap_AP3 --sample-label lncap_AP_R3 --fastq
/scratch/users/ekurt20/SRR14428516_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
median --pdf-report -n lncap_EP1 --sample-label lncap_EP_R1 --fastq
/scratch/users/ekurt20/epikol_data/SRR14428517_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
median --pdf-report -n lncap_EP2 --sample-label lncap_EP_R2 --fastq
/scratch/users/ekurt20/epikol_data/SRR14428518_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
median --pdf-report -n lncap_EP3 --sample-label lncap_EP_R3 --fastq
/scratch/users/ekurt20/epikol_data/SRR14428519_1.fastq
```

Produced count files should merge into one for each sample. MAGeCK test command works with one count file and performs comparisons based on sample names inside that one file. Prior to merging, sort the count files by library sgRNA ID to prevent any errors.

‘awk’ command used for sorting the files based on the sgRNA ID.

```
awk 'NR == 1; NR > 1 {print $0 | "sort -n"}' lncap_EP3.count_normalized.txt >
lncap_EP3.count_normalized_sorted.txt

awk 'NR == 1; NR > 1 {print $0 | "sort -n"}' lncap_EP2.count_normalized.txt >
lncap_EP2.count_normalized_sorted.txt

awk 'NR == 1; NR > 1 {print $0 | "sort -n"}' lncap_EP1.count_normalized.txt >
lncap_EP1.count_normalized_sorted.txt

awk 'NR == 1; NR > 1 {print $0 | "sort -n"}' lncap_AP1.count_normalized.txt >
lncap_AP1.count_normalized_sorted.txt

awk 'NR == 1; NR > 1 {print $0 | "sort -n"}' lncap_AP2.count_normalized.txt >
lncap_AP2.count_normalized_sorted.txt

awk 'NR == 1; NR > 1 {print $0 | "sort -n"}' lncap_AP3.count_normalized.txt >
lncap_AP3.count_normalized_sorted.txt
```

Then the sorted files are combined into one file. This merge is performed using the paste function.

```
paste lncap_AP1.count_normalized_sorted.txt
lncap_AP2.count_normalized_sorted.txt lncap_AP3.count_normalized_sorted.txt >
lncap_AP.normalized_megred_count.txt

paste lncap_EP1.count_normalized_sorted.txt
lncap_EP2.count_normalized_sorted.txt lncap_EP3.count_normalized_sorted.txt >
lncap_EP.normalized_megred_count.txt

paste lncap_AP.normalized_megred_count.txt
lncap_EP.normalized_megred_count.txt >
lncap_AP_EP.normalized_megred_count.txt
```

Keep the gene name and library sgRNA ID columns from the first file, as well as the count columns from each of the other files, due to the format that is needed. This is done by ‘awk’ command.

```
awk 'BEGIN {OFS="\t"};{print $1, $2, $3, $6, $9, $12, $15, $18}'
lncap_AP_EP.normalized_megred_count.txt >
lncap_AP_EP_grouped.normalized_countfile.txt
```

After that, the variation in guide abundance needs to be calculated, which is often expressed as a log-fold difference between treatment and control conditions. The MAGeCK Test command could now be executed. The normalized counts file, test and control groups to be compared, a name for the comparison, and the normalizing technique must be supplied into the Mageck test function. A PDF report might be requested in addition.

```
mageck test -k lncap_AP_EP_grouped.normalized_countfile.txt -t
lncap_EP_R1,lncap_EP_R2,lncap_EP_R3 -c lncap_AP_R1,lncap_AP_R2,lncap_AP_R3 -n
LNCaP_EPvsAP --norm-method median --pdf-report
```

Screen-level quality statistics could be investigated using guide-level performance data by running the mageck test function.

2.6.2 Normalization / Preprocessing of sgRNA counts and CRISPR bias correction

CRISPRcleanR is a R package, created to help with read count normalization on CRISPR screens. So, raw read counts are needed for it. Read counts for this study produced with MAGeCK count algorithm. Count function used like it was mentioned above, only the normalization method is chosen as none.

```
mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
none --pdf-report -n lncap_AP1 --sample-label lncap_AP_R1 --fastq
/scratch/users/ekurt20/SRR14428514_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
none --pdf-report -n lncap_AP2 --sample-label lncap_AP_R2 --fastq
/scratch/users/ekurt20/SRR14428515_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
none --pdf-report -n lncap_AP3 --sample-label lncap_AP_R3 --fastq
/scratch/users/ekurt20/SRR14428516_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
none --pdf-report -n lncap_EP1 --sample-label lncap_EP_R1 --fastq
/scratch/users/ekurt20/epikol_data/SRR14428517_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
none --pdf-report -n lncap_EP2 --sample-label lncap_EP_R2 --fastq
/scratch/users/ekurt20/epikol_data/SRR14428518_1.fastq

mageck count -l /scratch/users/ekurt20/epikol_sgrna_library.txt --norm-method
none --pdf-report -n lncap_EP3 --sample-label lncap_EP_R3 --fastq
/scratch/users/ekurt20/epikol_data/SRR14428519_1.fastq
```

All raw read counts produced with the MAGeCK count function were merged into one file. It was the final read count file that was to be used as an input for the CRISPRcleanR package.

The CRISPR screen library file is also necessary for the CRISPRcleanR package to work properly. The file must be in the proper format for analysis to proceed. It must be a data frame containing the targeted genes and their respective sgRNAs, along with some additional information. It should be formatted, with one row for each sgRNA and an alphanumeric code identifier for the sgRNAs. The library file should have information for the targeted genes, the exon of the targeted genomic region, the chromosome on which the targeted region is located, the starting genomic coordinate of the targeted genomic region, and the ending genomic coordinate of the targeted genomic region. sgRNA information obtained from a BLAT search (Human BLAT Search, n.d.) is used to generate an EPIKOL library file in the proper format. Each sgRNA is run in the BLAT database, and the results are collected in an Excel file. This

was eventually converted to R, formatted as a data frame, and incorporated into the library's annotation file.

The ccr.NormfoldChanges function is used for normalizing the sgRNA counts. Counts produced with MAGeCK (no normalization) needs to be a data frame. This will go into the ccr.NormfoldChanges function's DFrame argument. The EPIKOL library file will used for the library annotation argument. The number of controls for the sample needs to be specified inside the command.

```
normfoldchanges <- ccr.NormfoldChanges(fn, Dframe = lncapcount, display=TRUE,
                                       saveToFig= TRUE,
                                       outdir = './',
                                       min_reads = 30, EXPname = 'LNCaP',
                                       libraryAnnotation=epikol,ncontrols=3,
                                       method='MedRatios')
```

Firstly, with this command, sgRNAs with fewer than 30 reads will be eliminated from the control samples. The replicates were then adjusted using a median-ratio approach to account for the influence of library sizes and read count distributions. Individual sgRNA depletion/enrichment fold changes were assessed using read counts at the individual replicate level after puromycin selection and after 15 population doubling times. CRISPRcleanR's ccr.NormfoldChanges function was utilized for this purpose. This function gives back a list of two data frames with log fold-changes and normalized counts.

The ccr.logFCs2chromPos function is used to sort sgRNA log fold changes based on their target gene's position on the chromosome. The function takes normalized log foldchange output from ccr.NormfoldChanges and the library file is used. Because the library file contains genomic location information, it will be employed. After executing the code, a data frame of logFCs classified by chromosome position will be produced.

```
gwSortedFCs <- ccr.logFCs2chromPos(normfoldchanges$logFCs,epikolol2)
```

The ccr.GWclean function detects and corrects biased sgRNA log fold changes caused by gene-independent reactions due to CRISPR-Cas9 targeting. This takes the sorted sgRNA logFCs information and produces corrected logFCs. The corrected sgRNA fold-changes, segment annotation, and genome-sorted sgRNA identifers are returned in a list containing a separate data frame.

```
correctedFCs <-ccr.GWclean(gwSortedFCs,display=FALSE,label='LNCaP_gwclean')
```

2.6.3 BAGEL analysis

The fold change values of the CRISPRcleanR corrected sgRNAs were used as input into BAGEL2. It is a Python-based tool that calculates gene-level BF by sgRNAs on a per-gene basis by averaging as opposed to summing. BAGEL2's Bf function computes the log2 bayes factor for each gene. For this purpose, a list of common non-essential and common essential genes is necessary. They must be supplied as text files. CRISPRcleanR-normalized sgRNA foldchanges must be provided in a tab-delimited text file. The number of columns to examine must be specified within the function.

```
python BAGEL.py bf -i /scratch/users/ekurt20/epikol_wo/lncap_norm_lfc.txt -o
lncap.bf  -e /scratch/users/ekurt20/crisprcleanr/commonessential.txt -n
/scratch/users/ekurt20/crisprcleanr/nonessential.txt -c 1,2,3
```

The BAGEL2's pr function is used to calculate the precision-recall plot. The input file is the output from the BAGEL bf function, and the essential and common non-essential gene lists are specified in the command. The output is a ranked list of genes with precision/recall determined at each location for simple visualization.

```
python BAGEL.py pr -i /scratch/users/ekurt20/bagel/lncap.bf -o lncap.pr -e
/scratch/users/ekurt20/crisprcleanr/commonessential.txt -n
/scratch/users/ekurt20/crisprcleanr/nonessential.txt
```

Chapter 3:
RESULTS

Three different analysis tools are used to identify negatively selected sgRNA in EPIKOL screens. The results from these three methods were compared to see which method would better on EPIKOL screens. Which algorithm performed better on EPIKOL screen data choose based on the FDR values, number of false positive hit genes. The study included two cancer types and five cancer cell lines in total. Triple-negative breast cancer (TNBC) cell lines are MDA-MB-231, SUM149PT, and SUM159PT. Prostate cancer (PCa) cell lines are LNCaP and DU145.

In each cell line, end point and after puro samples collected. The after-puro samples act as a control group in the analysis. Comparing EP samples to the control group revealed depleted sgRNA and their targeted genes.

The first analyzing method with MAGeCK, raw sgRNA read counts taken from FASTQ files and normalized first by log transformation and then comparing to a control condition. After that sgRNA ranking, gene ranking performed to identify hit genes (Li et al., 2014).

Second method was similar to the first method, only the normalization of the sgRNA reads were different. CRISPRcleanR (CCR) employed for the normalization of the sgRNA read counts. The median-ratio method is used by CCR to take into consideration the effects of library sizes and read count distributions. A circular binary segmentation approach is used directly on the genome-wide patterns of sgRNA LFCs across individual chromosomes in a cell line in CRISPRcleanR, as opposed to MAGeCK normalization (Iorio et al., 2018). Obtaining the raw read counts and essentiality measurements were done with MAGeCK commands.

BAGEL2 application employed in the third method. When computing gene-level Bayesian factors via sgRNAs on a targeted-gene basis, the BAGEL implementation averages the results rather than adding them together. Its implementation computes gene-level BFs using reference sets of predefined essential and non-essential genes (Kim & Hart, 2021). A screens LFC file is supplied into the BAGEL2. This file was obtained with the assistance of MAGeCK and CCR. MAGeCK was used to obtain raw read counts, and the CRR was used to normalize sgRNA and obtain the LFC file.

Various stringent criteria are used to determine the essential genes. In MAGeCK, genes with neg|fdr<0.15 and neg|logFC<0.5 were deemed as essential. MAGeCK gives both negative and the positive selection values in a table. The negative section of MAGeCK's results is also of particular importance given that this is a knock-out screen. In BAGEL2, genes with BF>7 were considered as an essential gene.

At the end of each method, essential gene lists were produced. These lists compared to each other based on the number of essential genes found in each method. A variety of epigenetic modifiers were discovered to be considerably reduced in the TNBC cell lines in MAGeCK analysis; MDA-MB-231 (81), SUM149PT (87), and SUM159PT (28) (**Table *2***). **Table *2*** demonstrates that BAGEL2 will provide a greater number of essential genes than the other approaches.

These programs are designed for genome-wide CRISPR screens. One data set from a genome-wide study was retrieved. To show that these programs have the same impact on analysis, the Brunello library data set on A375 cells were acquired from SRA (Sanson et al., 2018). On this set of data, the same analysis techniques are used, and the outcomes are displayed.

Table 2 Number of essential genes found in each method.

Cell line	MAGeCK	CCR-MAGECK	BAGEL2
LNCaP	91	86	128
DU145	122	98	163
MDA-MB-231	81	92	121
SUM149PT	87	73	116
SUM159PT	28	31	61

3.1 TNBC Knockout screening results

Negative selection (drop-out) screens were carried out to identify the epigenetic modifiers that contribute to triple negative breast cancer (TNBC) cell fitness. MDA-MB-231, SUM149PT, and SUM159PT are three distinct TNBC cell lines that were analyzed with three different methods.

The number of essential gene found in each analysis method were visualized. The genes who fitted to the conditions pointed in above classified as essential to the cell survival.

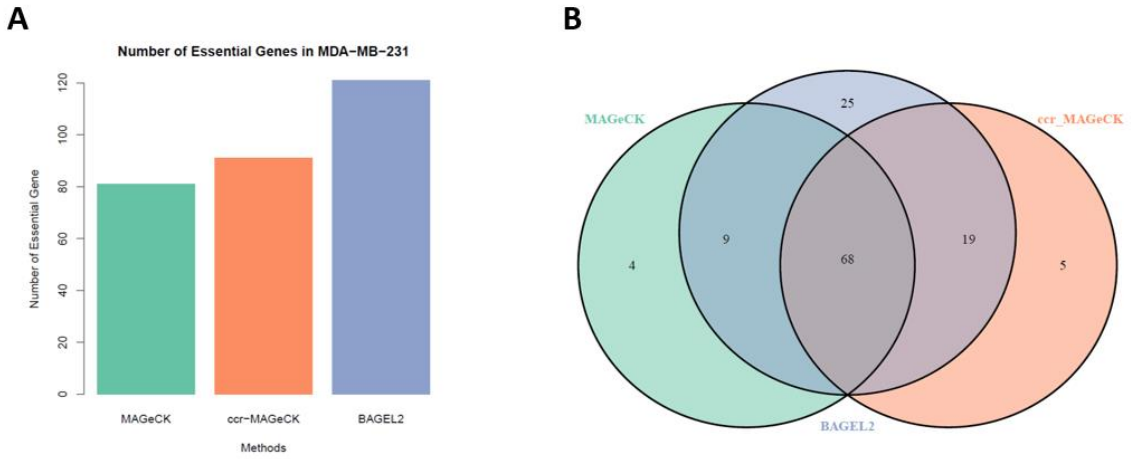


Figure 3 Number of essential genes in MDA-MB-231 cell line. A Essential genes found in each three analysis method were shown in bar chart. **B** Venn diagram shows the number of common essential genes in the methods.

BAGEL2 returns a list including a greater number of essential genes than the other two methods (**Figure 3**). In all of the three methods, 68 common essential genes were observed. The number of genes which haven't been able by other method is higher in BAGEL2. BAGEL2 detects 25 essential genes in MDA-MB-231 cells that were not detectable by other methods. While ccr-normalized MAGeCK reveals 5 distinct essential genes, MAGeCK uncovers only 4 (**Figure 3B**).

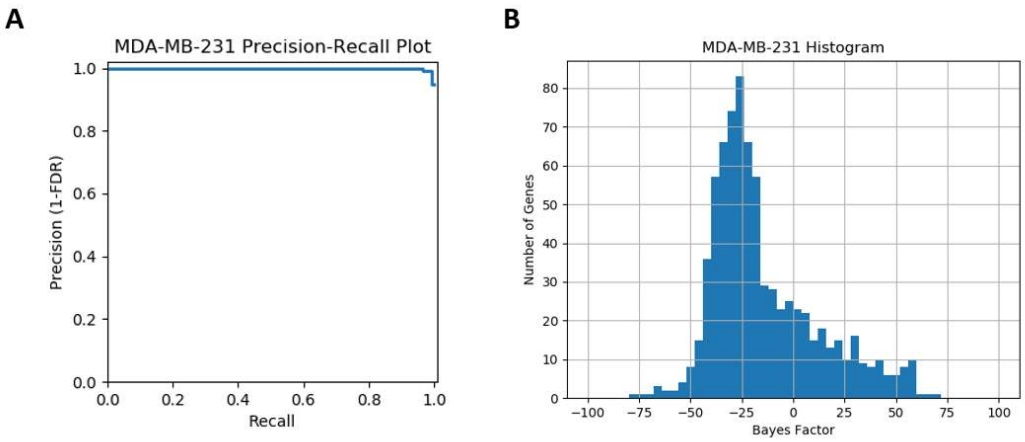


Figure 4 BAGEL2 results of MDA-MB-231 cell line shown in the graphs. A The precision and recall plot displays the recall and precision for MDA-MB-231 ranked genes at each point. **B** The histogram graph shows the distribution of the Bayes factor.

The output of BAGEL2 is shown as a precision recall plot and histogram (**Figure *4***). The differentially expressed genes are displayed using a volcano plot to visualize the MAGeCK and CCR-MAGeCK output (**Figure *5***).

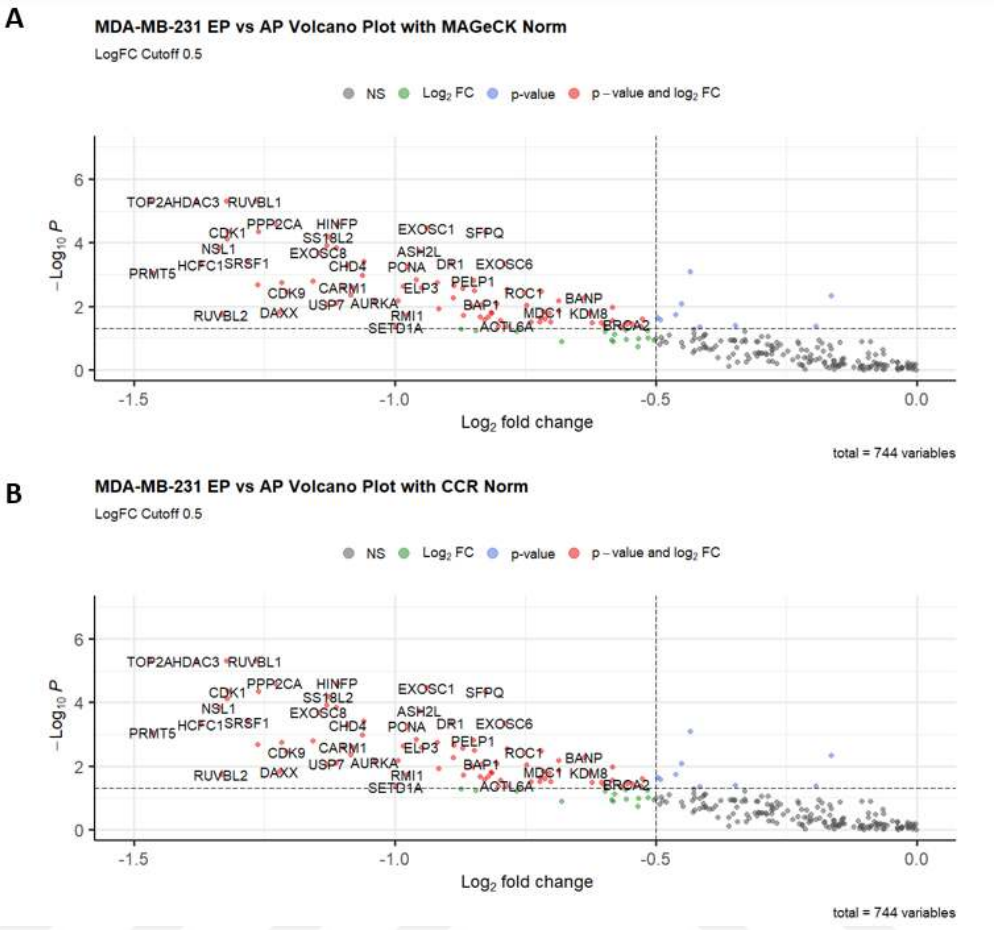


Figure 5 Volcano plots showing the differentially expressed genes in MDA-MB-231 cells. A Differentially expressed gene with MAGeCK method was shown. **B** Differentially genes with CCR-MAGeCK combination were showed.

In SUM149PT cells, BAGEL2 also provides a higher number of essential genes compared to other algorithms (**Figure *6*A**). In all three techniques, 62 of the genes were identified as common. BAGEL2 revealed additionally 32 differentially expressed genes in the SUM159PT cells. CCR- MAGeCK gives 2 distinct genes, whereas MAGeCK distinguishes 8 (**Figure *6*B**).

SUM149PT screening results used for creating volcano plot (**Figure 7**). Using the same criteria as in MDA-MB-231 cells, the data from MAGeCK and CCR-MAGeCK were used respectively to create the volcano plots. BAGEL2 output is provided as a precision recall plot and histogram in SUM149PT cells (**Figure 8**).

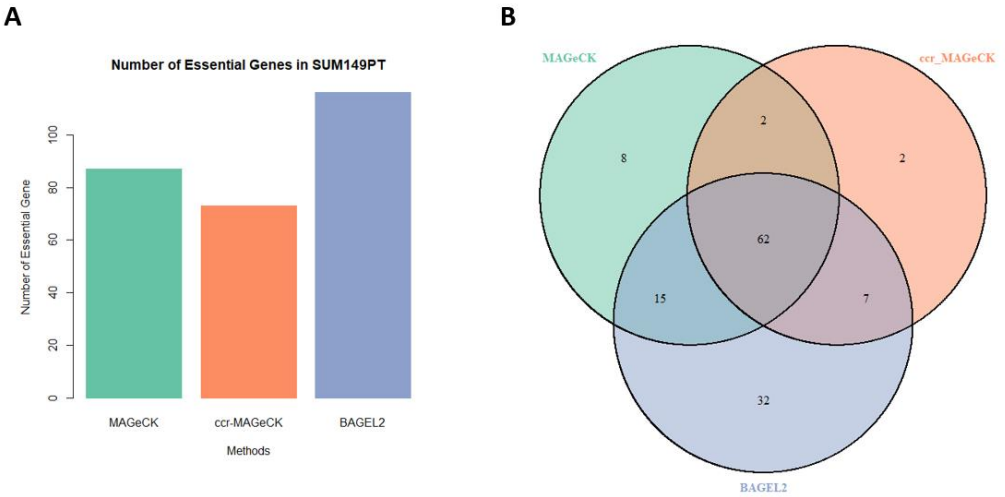


Figure 6 Number of essential genes in SUM149PT knockout screen. A Essential genes found in each method. **B** Venn diagram shows the number of common essential genes found in the methods.

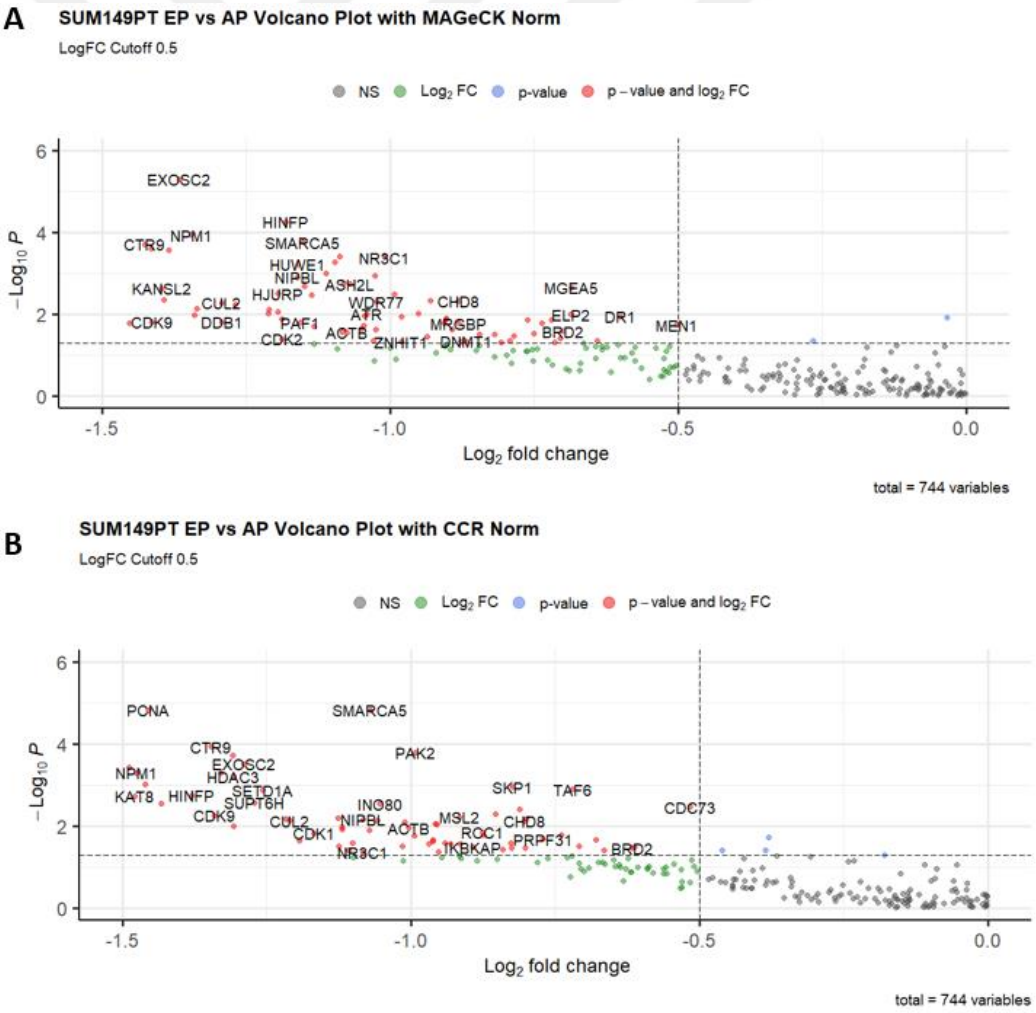


Figure 7 Volcano plots showing the differentially expressed genes in SUM149PT cells. A Differentially expressed gene with MAGeCK method was shown. **B** Differentially genes with CCR-MAGeCK combination were showed.

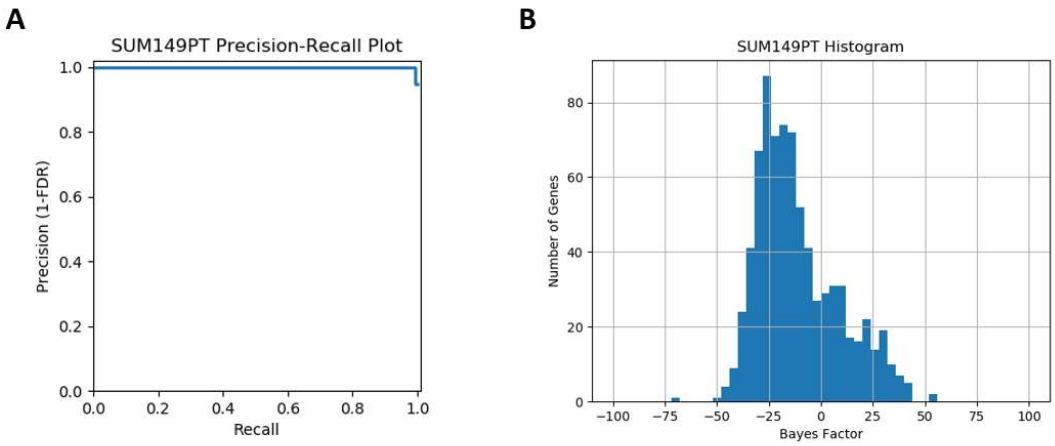


Figure 8 BAGEL2 results of SUM149PT cell line shown in the graphs. A The precision and recall plot displays the recall and precision for SUM149PT ranked genes at each point. **B** The histogram graph shows the distribution of the Bayes factor.

BAGEL2 delivers a greater number of essential genes than other algorithms in SUM159PT cells (**Figure *9***). In each of the three approaches, 17 genes were found in common. BAGEL2 found an additional 23 genes with differential expression in SUM159PT cells. CCR- MAGeCK identifies only one unique gene, whereas MAGeCK identifies 3 (**Figure *9*B**). The precision recall and histogram plot were created using the BAGEL2 output of SUM159PT cells (**Figure 10**). The result from MAGeCK and CCR-MAGeCK used for creating volcano plots in R (**Figure 11**).

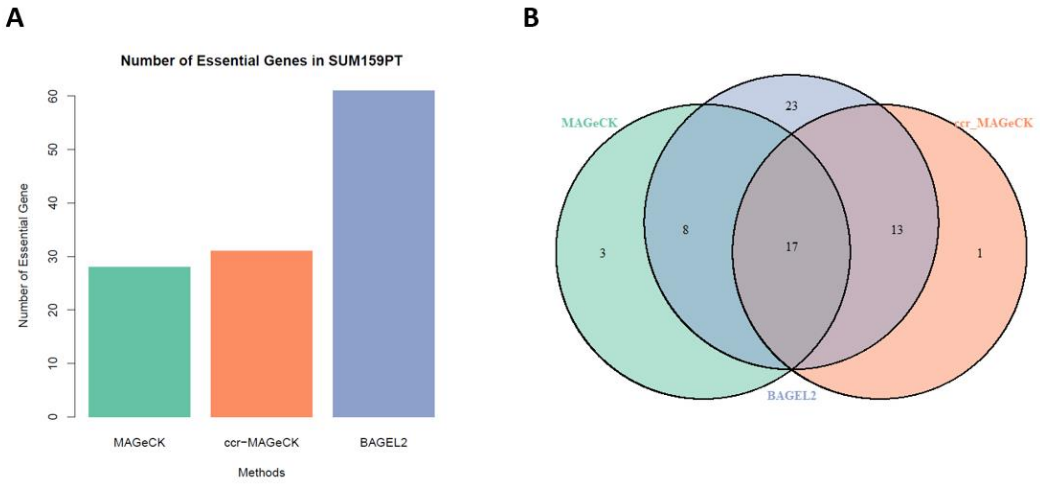


Figure 9 Number of essential genes in SUM159PT knockout screen. A Essential genes found in each method. **B** Venn diagram shows the number of common essential genes found in the methods.

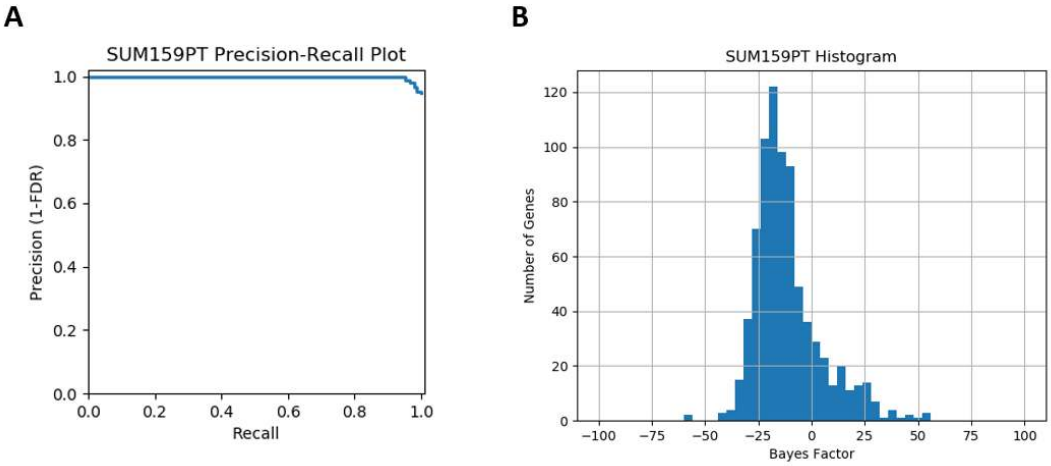


Figure 10 BAGEL2 results of SUM159PT cell line shown in the graphs. A The precision and recall plot displays the recall and precision for SUM159PT ranked genes at each point. **B** The histogram graph shows the distribution of the Bayes factor.

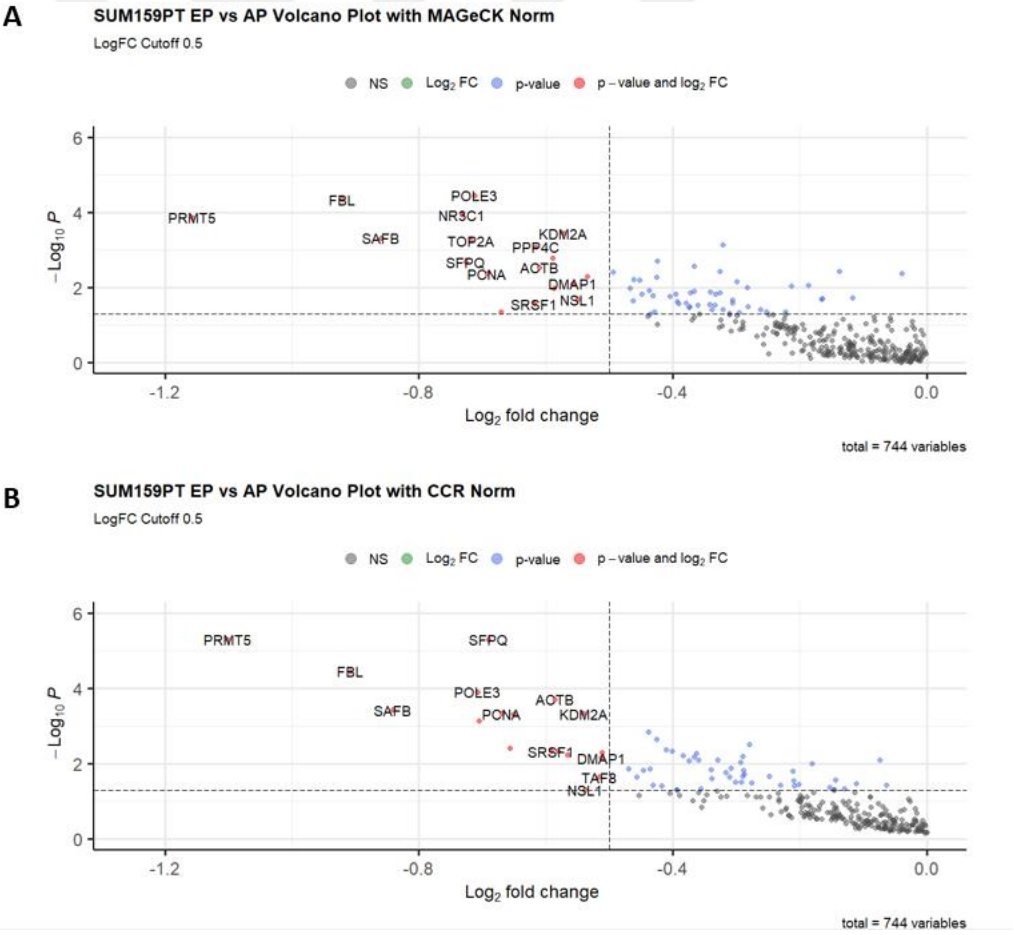


Figure 11 Volcano plots showing the differentially expressed geness in SUM159PT cells. A Differentially expressed gene with MAGeCK method was shown. **B** Differentially genes with CCR-MAGeCK combination were showed.

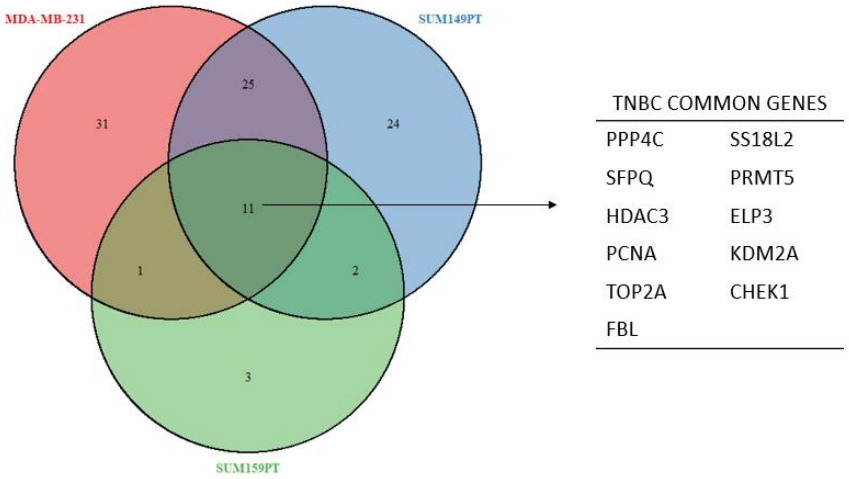


Figure 12 Venn diagram showing the common essential genes in TNBC cells.

The number of essential genes identified by each technique in TNBC cell lines was then combined to determine the common genes (**Figure 12**). 11 genes were found significantly depleted in all of them. PRMT5 and HDAC3, epigenetic modifiers that previously associated with breast cancer cell viability, were served as a positive control for the analysis methods (Cui et al., 2018; Vinet et al., 2019). Rest of TNBC common hit gene list used for gene set enrichment analysis. SS18L2 is a homolog of the SS18 gene, which has been linked to synovial sarcoma chromosomal translocation. The precise role of SS18L2 in synovial sarcoma or any other cancer is unknown. SS18L2 showed a high depletion score on EPIKOL screens. It is found to be common in each analysis method that was used.

3.2 Prostate Cancer Knock-out Screening Results

Two prostate cancer CRISPR screen used which are LNCaP and DU145. Both screening data analyzed with three different methods. At the end of analysis results first compared within the same cell line and then combined to see the common essential genes in the prostate cancer.

BAGEL2 provides a list that contains more essential genes than the other two techniques in the LNCaP cells (**Figure 13**). 73 essential genes were identified as common using each of the three approaches. CCR-MAGeCK, MAGeCK, and BALEL2 detect 1, 6, and 35 essential genes in LNCaP cells, respectively, that were undetectable

by other techniques (**Figure 13B**). The precision recall and histogram plot were created using the BAGEL2 output of LNCaP cells (**Figure 14**). The distribution of essential genes was presented using volcano plots. MAGeCK and CCR-MAGeCK results used in the volcano plots (**Figure 15**).

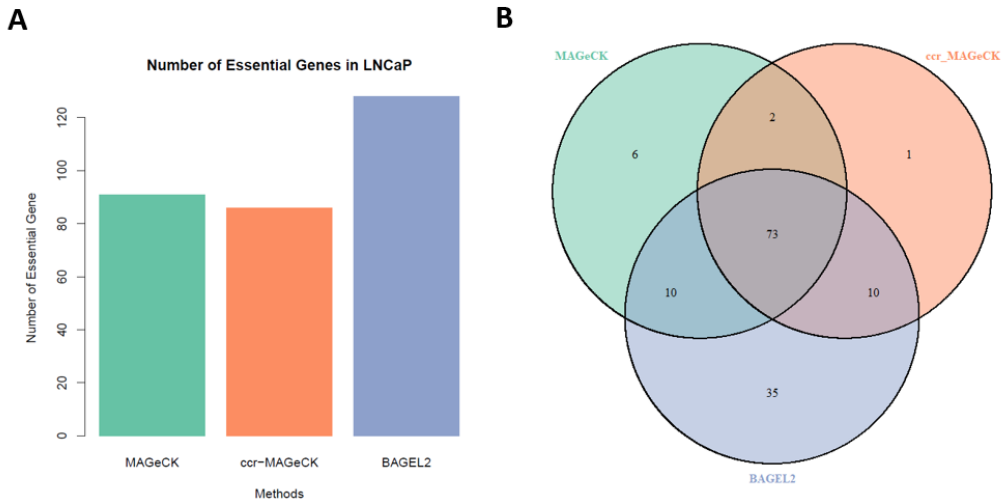


Figure 13 Number of essential genes in LNCaP knockout screen. A Essential genes found in each method. **B** Venn diagram shows the number of common essential genes found in the methods.

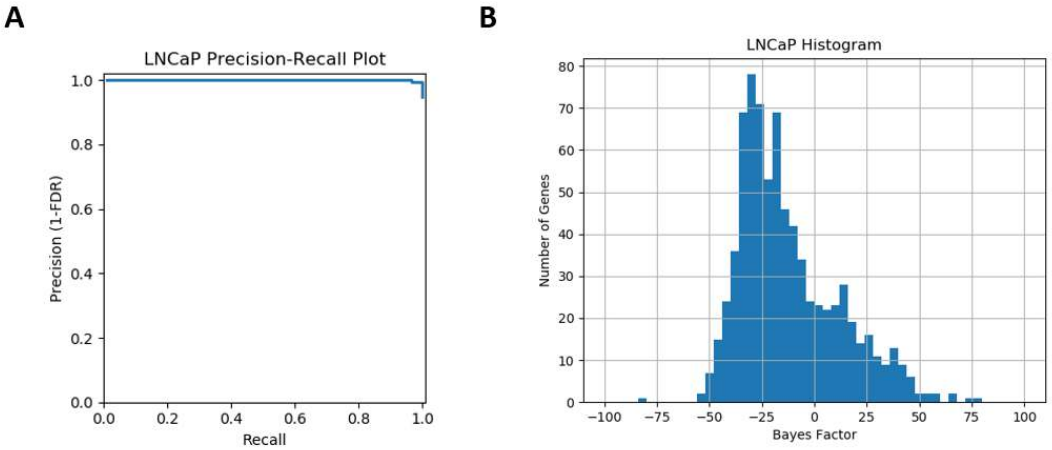


Figure 14 BAGEL2 results of LNCaP cell line shown in the graphs. A The precision and recall plot displays the recall and precision for LNCaP ranked genes at each point. **B** The histogram graph shows the distribution of the Bayes factor.

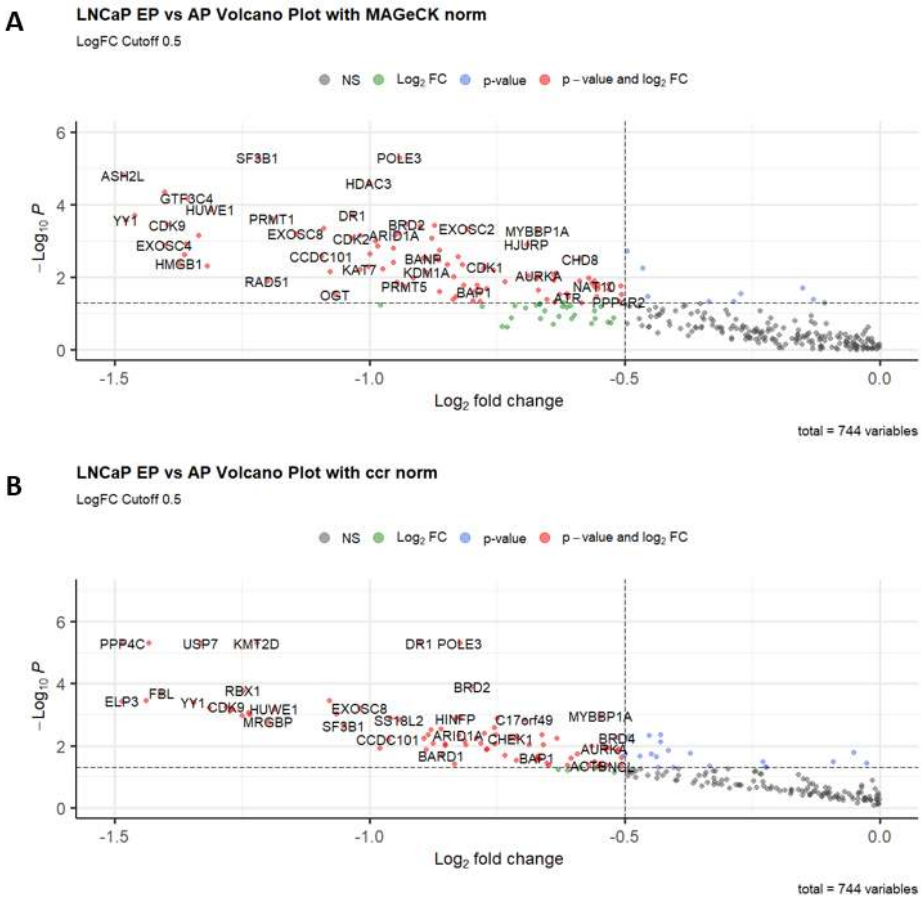


Figure 15 Volcano plots showing the differentially expressed genes in LNCaP cells. A Differentially expressed gene with MAGeCK method was shown. **B** Differentially genes with CCR-MAGeCK combination were showed.

In DU145 cells, BAGEL2 generates a list with more hit genes than the other two approaches (**Figure 16**). Using each of the three techniques, 89 differentially expressed genes were found to be shared. Only MAGeCK and BAGEL2 detect different essential genes within the dataset. 5 and 37 genes were identified, respectively (**Figure 16B**). The precision recall and histogram plot were created using the BAGEL2 output of DU145 cells (**Figure 17**). Volcano charts used for showing the gene distribution with MAGeCK and CCR-MAGeCK algorithm (**Figure 18**).

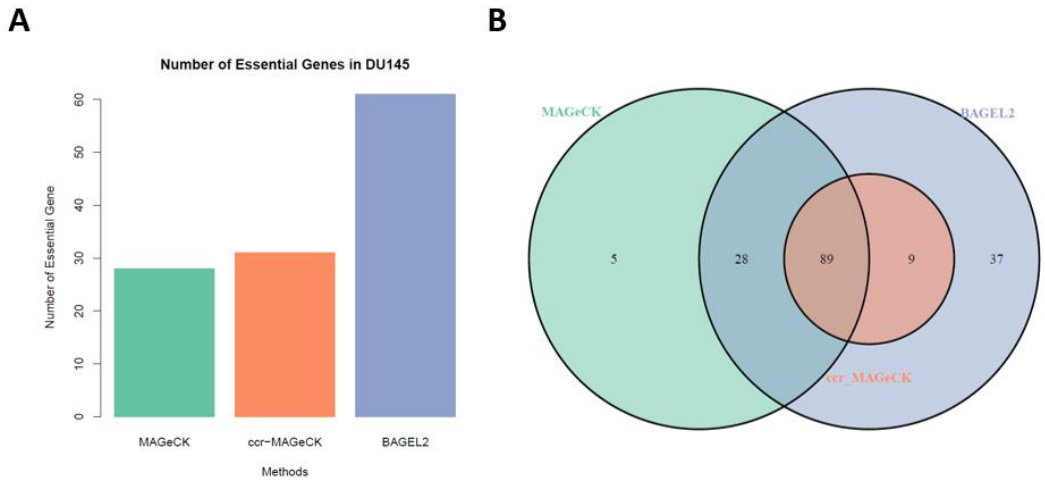


Figure 16 Number of essential genes in DU145 knockout screen. A Essential genes found in each method. **B** Venn diagram shows the number of common essential genes found in the methods.

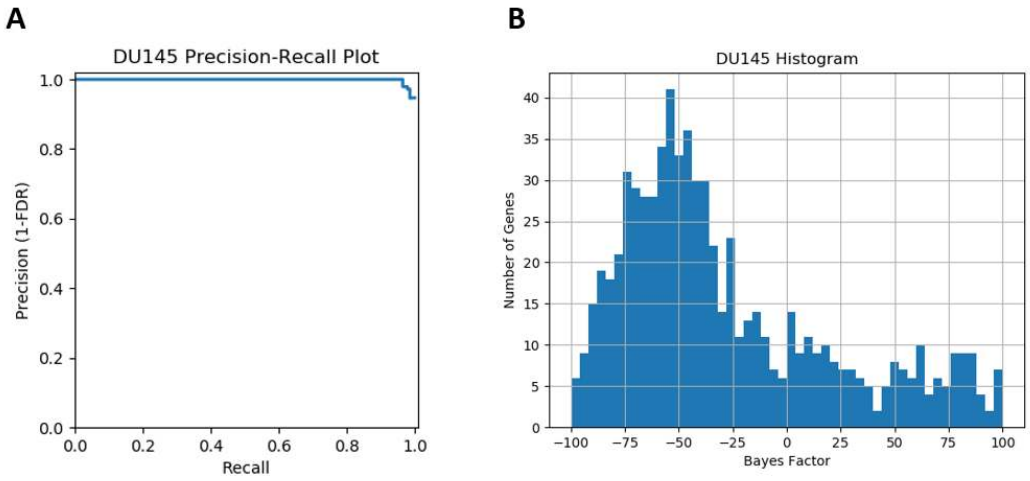


Figure 17 BAGEL2 results of DU145 cell line shown in the graphs. A The precision and recall plot displays the recall and precision for DU145 ranked genes at each point. **B** The histogram graph shows the distribution of the Bayes factor.

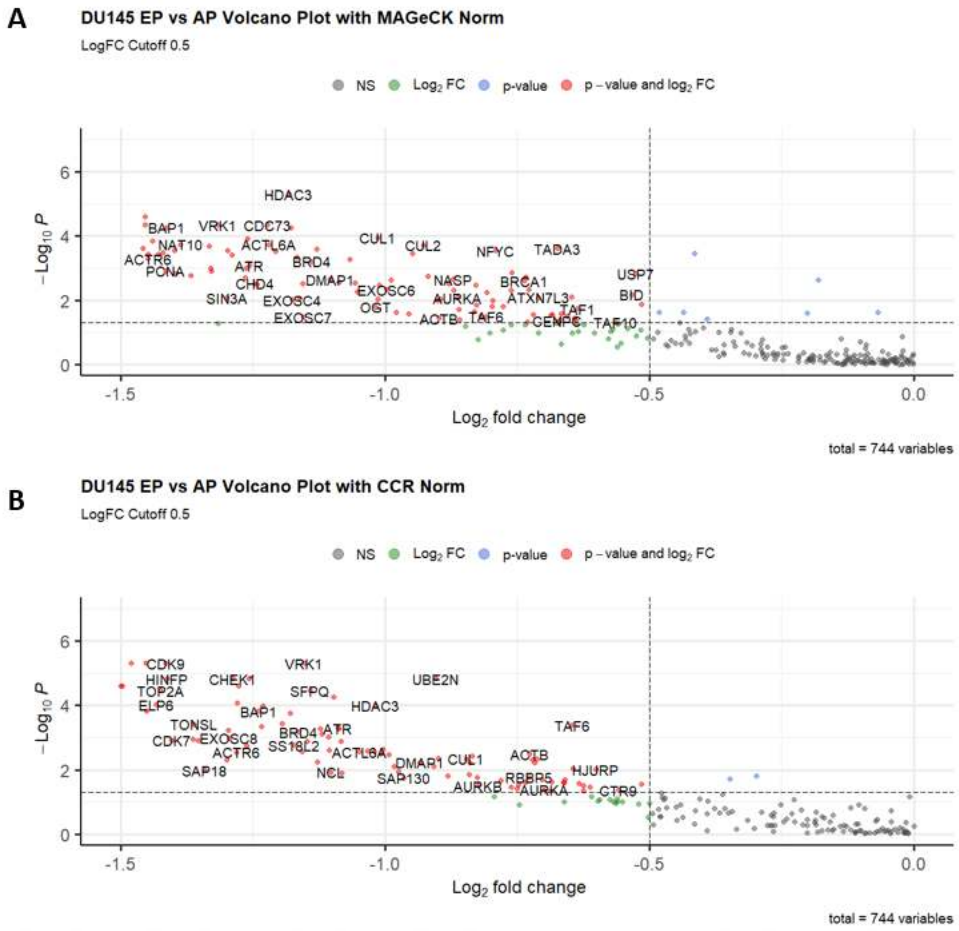
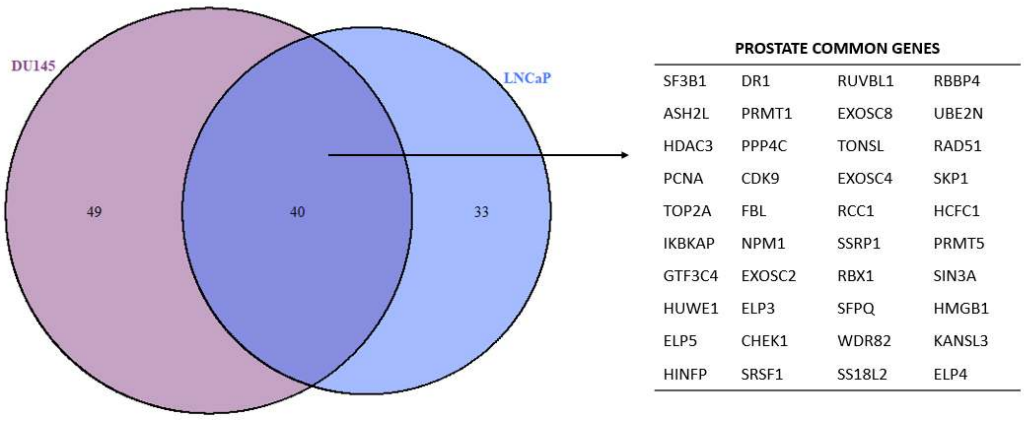


Figure 18 Volcano plots showing the differentially expressed gene in DU145 cells. A Differentially expressed gene with MAGeCK method was shown. **B** Differentially genes with CCR-MAGeCK combination were showed.



PROSTATE COMMON GENES			
SF3B1	DR1	RUVBL1	RBBP4
ASH2L	PRMT1	EXOSC8	UBE2N
HDAC3	PPP4C	TONSL	RAD51
PCNA	CDK9	EXOSC4	SKP1
TOP2A	FBL	RCC1	HCFC1
IKBKAP	NPM1	SSRP1	PRMT5
GTF3C4	EXOSC2	RBX1	SIN3A
HUWE1	ELP3	SFPQ	HMGB1
ELP5	CHEK1	WDR82	KANSL3
HINFP	SRSF1	SS18L2	ELP4

Figure 19 Venn diagram showing the common essential genes in Prostate cancer cells.

The number of essential genes identified by each method in PCa cell lines was then combined to identify the common genes (**Figure *19***). Forty genes were discovered to be severely decreased in each cell line. PRMT1 was previously connected with the viability of prostate cancer cells, and therefore it served as a positive control for the methodological approaches (Tang et al., 2022). The remainder of the PCa common hit gene list is used for gene set enrichment analysis. Again, SS18L2 gene found significantly depleted in prostate cancer cell lines with EPIKOL.

SS18L2 highly downregulated in each cell line. Both TNBC and PCa results shows that the SS18L2 could play important role in the regulation of cell cycle and cell fitness.

CCLE RNA-seq log TPM data collected from DepMap is used to assess the number of false-positive genes detected by each approach (*DepMap: The Cancer Dependency Map Project at Broad Institute*, n.d.). A log-expression below 1.0 was used to indicate false positives for each cell line. To continue with FP estimation only MDA-MB-231, LNCaP, and DU145 data were found in DepMap for our screening results. The number of false positives in each cell line was assessed, compared to the list of hit genes, and counted for each approach. In each method's results, there were no false positives.

After FP examination, we can conclude that BAGEL2 identified a greater number of hit genes than other methods. It found 121, 128, and 163 genes, respectively, in MDA-MB-231, LNCaP, and DU145 cells. These values are significantly greater than MAGeCK and the CRR normalized MAGeCK analysis approaches.

A375 cells downloaded from SRA were analyzed with each of the three approaches, and plots were constructed to validate the methodologies. There are two replicates of the A375 cell line. These replicates were examined separately, and a Brunello library plasmid sample was used as a control to determine which genes were differentially expressed. A375 A represents the cell line's first replicate, while A375 B indicates the dropout screen's second replicate.

BAGEL2 returns a list including a greater number of essential genes than the other two methods in genome-wide library screen of A375 cells (**Figure 20).** In all of the three methods, 1368 common essential genes were observed. The number of genes which haven’t been able by other method is higher in BAGEL2. BAGEL2 detects 232 essential genes in A375_A cells that were not detectable by other methods. While ccr-normalized MAGeCK reveals 27 distinct essential genes, MAGeCK uncovers 57 (**Figure 20B**).

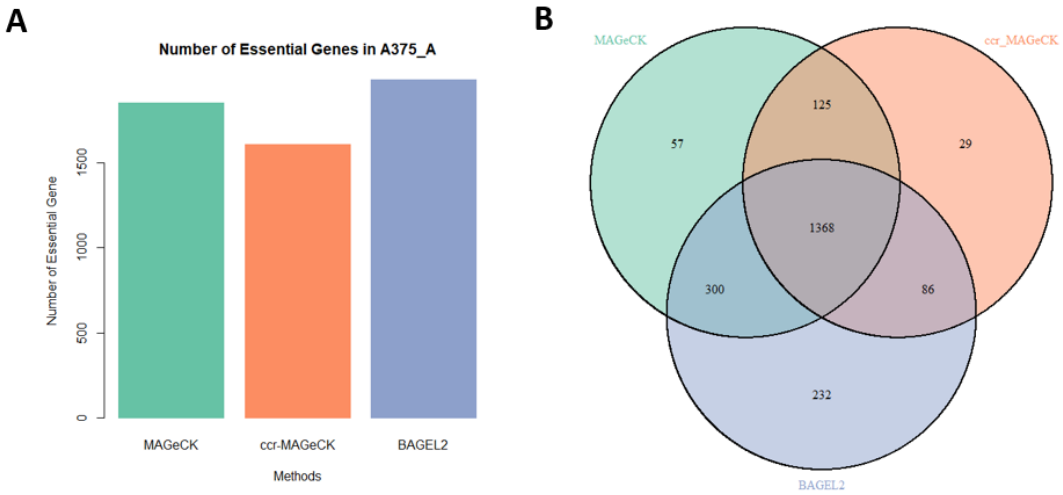


Figure 20 Number of essential genes in A375_A cell line. A Essential genes found in each three analysis method were shown in bar chart. **B** Venn diagram shows the number of common essential genes in the methods.

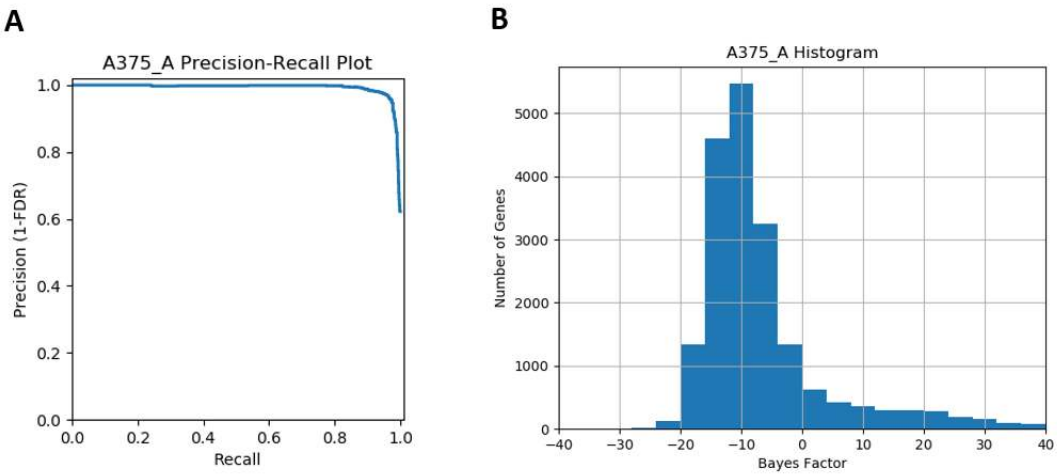


Figure 21 BAGEL2 results of A375_A cell line shown in the graphs. A The precision and recall plot displays the recall and precision for A375_A ranked genes at each point. **B** The histogram graph shows the distribution of the Bayes factor.

The precision recall and histogram plot were created using the BAGEL2 output of A375_A cells (**Figure 21**). The distribution of essential genes was presented using volcano plots. MAGeCK and CCR-MAGeCK results used in the volcano plots (**Figure 22**).

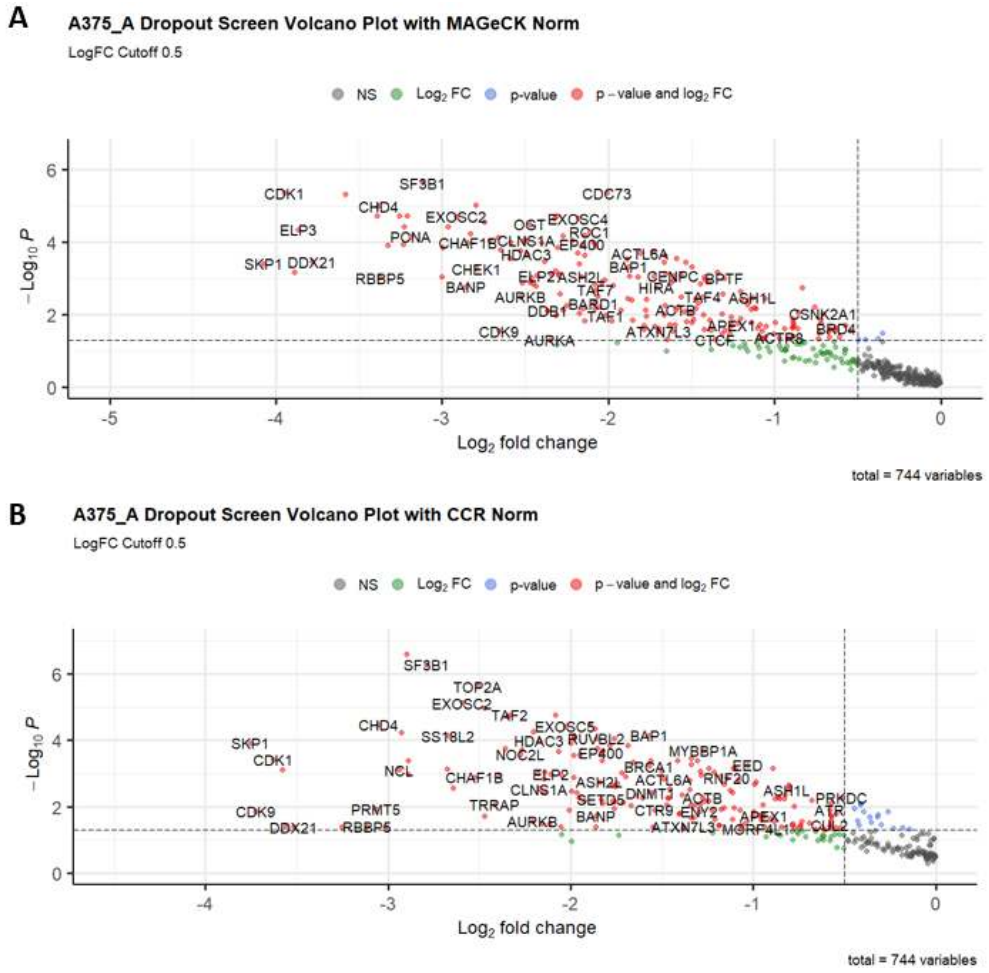


Figure 22 Volcano plots showing the differentially expressed gene in A375_A cells. A Differentially expressed gene with MAGeCK method was shown. **B** Differentially genes with CCR-MAGeCK combination were showed.

In A375_B cells, BAGEL2 generates a list with more essential genes than the other two methods (**Figure 23A**). 1205 common essential genes were discovered using all three techniques. In BAGEL2, there are more genes that haven't been discovered using previous techniques. In A375_B cells, 230 essential genes that could not be found using other techniques are discovered by BAGEL2. Whereas ccr-normalized MAGeCK exposes 35 unique essential genes, MAGeCK reveals 63 (**Figure 23B**).

The precision recall and histogram plot were created using the BAGEL2 output of A375_B cells (**Figure 24**). The distribution of essential genes was presented using volcano plots. MAGeCK and CCR-MAGeCK results used in the volcano plots (**Figure 22**).

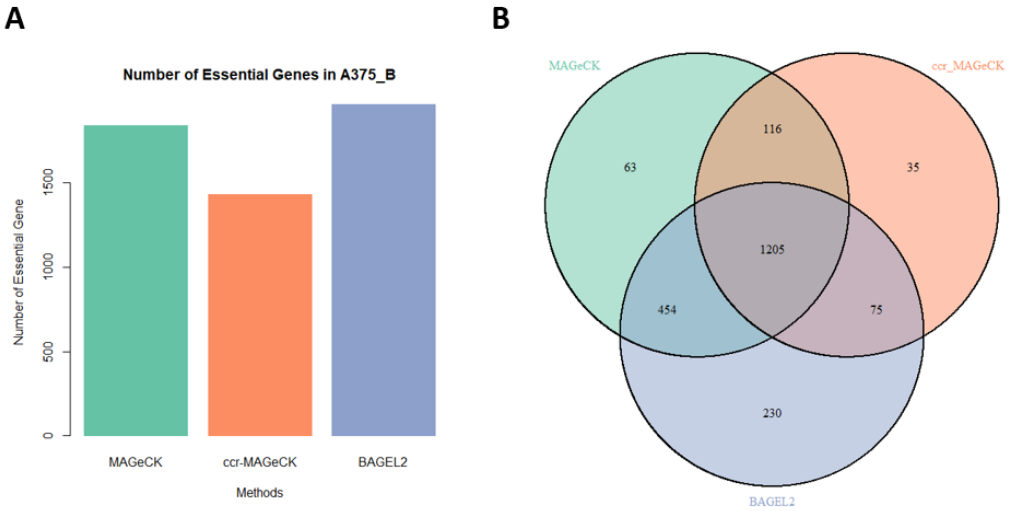


Figure 23 Number of essential genes in A375_B cell line. A Essential genes found in each three analysis method were shown in bar chart. **B** Venn diagram shows the number of common essential genes in the methods.

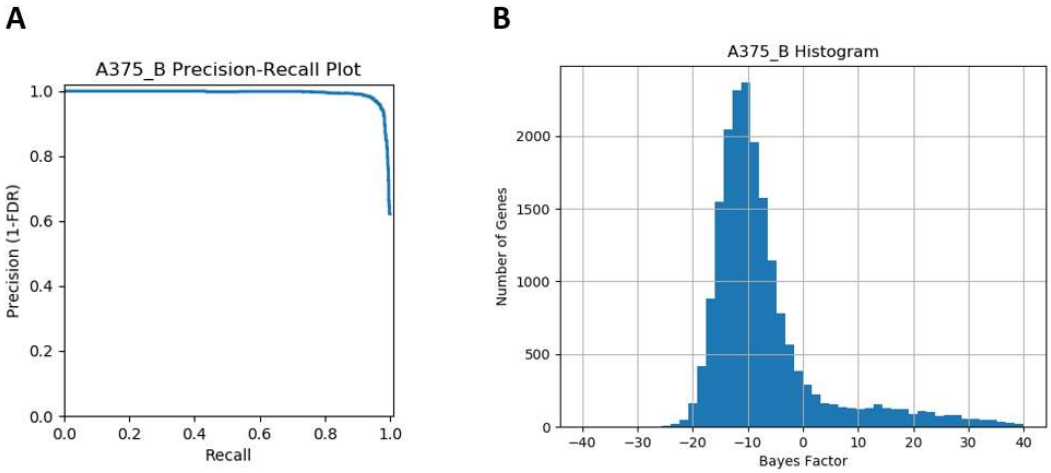


Figure 24 BAGEL2 results of A375_B cell line shown in the graphs. A The precision and recall plot displays the recall and precision for A375_B ranked genes at each point. **B** The histogram graph shows the distribution of the Bayes factor.

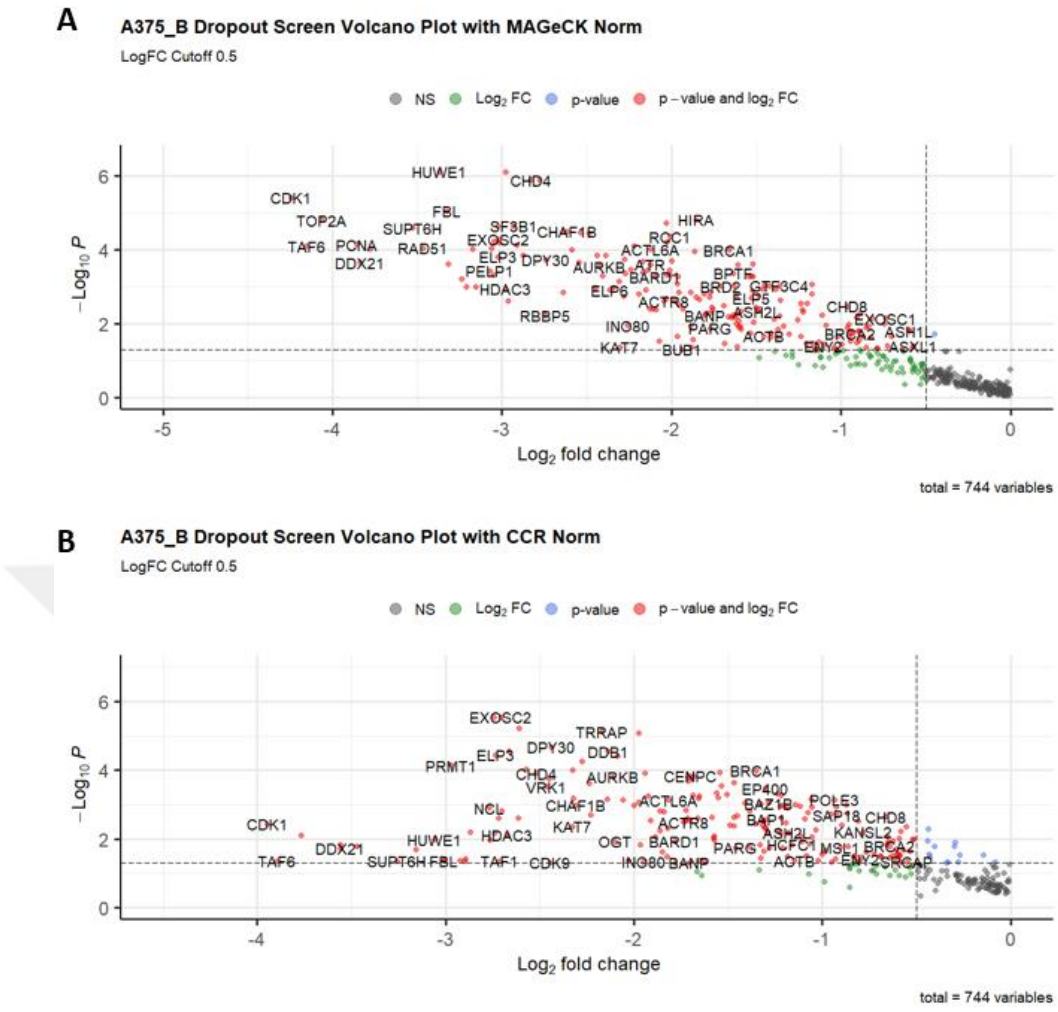


Figure 25 Volcano plots showing the differentially expressed gene in A375_B cells. A Differentially expressed gene with MAGeCK method was shown. **B** Differentially genes with CCR-MAGeCK combination were showed.

False positive evaluation has been done on A375 cells. There were no false positive detected in each replicate with all the methods.

Chapter 4:
DISCUSSION

Finding essential genes in mammalian cells using CRISPR-Cas9 knockdown screening is a promising strategy. Quantifying the efficiency of target modification is a crucial step in the application of CRISPR technology. From CRISPR-Cas9 knockout screens, various computational methods are available for identifying essential sgRNAs, genes, and pathways. The outcome of the experiment may also be affected by other factors, such as the sequence context and how well each sgRNA knocks out a gene. These will result in high false positive gene rate. This issue needs to be fixed in the analysis process.

Usually, the tools have been developed for genome-wide library screens and they perform well. Unfortunately, there is not optimized analysis tool for analyzing focused library screens. The analysis of small-scale library screens done with MAGeCK in most of the publication. But in MAGeCK sometimes FDR value can lead to identify false positive genes. Alternative analysis method needs to be explored.

In this work, the EPIKOL screen data of five cancer cell lines were examined using three approaches to establish the correlation between their hit gene lists. Two types of prostate cancer (LNCaP and DU145) and three types of triple-negative breast cancer (MDA-MB-231, SUM149PT, and SUM159PT) cell lines are selected. One Brunello library screen (A375) is included in the analysis to show the reliability of the methods.

GSEA searches for common hit genes to obtain additional information on specific hit genes. A literature search was conducted on these genes. We were unable to determine which method would be advantageous for small CRISPR screens. To evaluate it, GSEA can be applied to each method and its results can be compared to those of the other methods. This can give an idea of which method can uncover previously unexplored epigenetic targets that are more specific.

We used three distinct techniques to demonstrate in EPIKOL library screens that the SS18L2 gene is significantly depleted in all 5-cell lines. SS18L2 is a homolog of SS18, a transcriptional co-activator that interacts with numerous proteins involved in (epigenetic) gene regulation. SS18 was discovered in the ATP-dependent chromatin remodeling complex BAF (SWI/SNF) and an unidentified gene (de Bruijn & van Kessel, 2006). SS18 controls typical expression patterns from enhancers and promoters

along with BAF47; however, malignant gene translocation (SS18-SSX) in synovial sarcoma evicts BAF47 from the complex and novel fusion oncoprotein activates bivalent genes (McBride et al., 2018). The function of SS18L2 in cancer is unknown, and no studies have been conducted to investigate its impact on triple negative breast cancer.

Another gene which is found significantly depleted in EPIKOL screens was ASH2L. ASH2L is a key component of a multimeric histone methyltransferase complex that is important in the maintenance of active gene transcription by inducing histone H3 lysine 4 (H3K4) trimethylation. The absence of ASH2L protein results in a widespread decrease in H3K4 trimethylation (Mittal et al., 2018; Steward et al., 2006). The presence of an amino-terminal plant homeodomain (PHD) finger and winged-helix (WH) motifs in ASH2L suggests that it has the ability to detect DNA and histones (Callebaut & Mornon, 2012; Chen et al., 2011). Therefore, ASH2L levels in cancer might be utilized as a biomarker to predict the response to genotoxins (Constantin & Widmann, 2020).

All sgRNA location information (exon, chromosome, start and end positions) should be saved in a table during library production. These details must be included in the CRISPRcleanR annotation file. The CCR sorts LFCs of sgRNAs according to the gene on the chromosome that they target. It corrects sgRNA LFC on an individual chromosome basis. All EPIKOL sgRNA chromosome locations could not be found using BLAT. Consequently, CCR did not perform the correction of sgRNA LFC correctly. Due to improper LFC correction, BAGEL2 calculations cannot be relied upon. Because BAGEL2 uses LFCs that have been corrected using CCR. This issue needs to be fixed in further research.

CCR could correct the effects of copy number on screen data. CN is related to the number of false-positive results (FP). Downloading CCLE RNA-seq log TPM data from DepMap enables the estimation of false positives. Log-expression below 1.0 was defined as a false-positive for each cell line. After removing FP genes, we can count the number of false positives present in each algorithm's essential gene calls. This will allow us to compare the number of positive calls made by each algorithm and select the method that produces the fewest FP genes for our screening analysis. Eliminating the effect of copy number from data analysis facilitate the selection of hit genes more accurate and reliable. In the end, we discovered that the combination of CCR and

BAGEL2 provides more essential genes than the most used algorithm, MAGeCK. To see the applicability of the method, the same pipeline needs to be applied to different small scale library screen data. For example, Kinome or Transcription Factor CRISPR-Cas9 library screens data can be analyzed with these methods and see if same outcome appear. It is extremely important to study this analysis pipeline in different small scale libraries so that we can say for sure that BAGEL2 gives the most reliable hit gene list.

The amount of data needed to determine which method works best on small library screens is inadequate. It must be studied further by incorporating more variables and diverse small-scale library screen data. The CERES algorithm must be executed for a reliable comparison.

In this thesis, new analysis pipeline for focused CRISPR-Cas9 libraries were proposed and applied to EPIKOL library screen data. The result of the analysis show that there are better algorithms for analyzing small scale CRISPR screens. BAGEL2 could be more applicable to focused library screens because of the number of hit genes present in the output file. To test reliability of the BAGEL2 same pipeline should applied to other focused libraries and the results should be compared.

BIBLIOGRAPHY

AB, O., ES, V., R, L., & M, W. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)*, *5*(4). https://doi.org/10.1093/BIOSTATISTICS/KXH008

Aguirre, A. J., Meyers, R. M., Weir, B. A., Vazquez, F., Zhang, C. Z., Ben-David, U., Cook, A., Ha, G., Harrington, W. F., Doshi, M. B., Kost-Alimova, M., Gill, S., Xu, H., Ali, L. D., Jiang, G., Pantel, S., Lee, Y., Goodale, A., Cherniack, A. D., … Hahn, W. C. (2016). Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discovery*, *6*(8), 914–929. https://doi.org/10.1158/2159-8290.CD-16-0154/333303/AM/GENOMIC-COPY-NUMBER-DICTATES-A-GENE-INDEPENDENT

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), 1–12. https://doi.org/10.1186/GB-2010-11-10-R106/COMMENTS

Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., Ansari, R., Harper, S., Jackson, D. A., McRae, R., Pooley, R., Wilkinson, P., van der Meer, D., Dow, D., Buser-Doepner, C., … Garnett, M. J. (2019). Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature 2019 568:7753*, *568*(7753), 511–516. https://doi.org/10.1038/s41586-019-1103-9

Berdasco, M., & Esteller, M. (2019). Clinical epigenetics: seizing opportunities for translation. *Nature Reviews. Genetics*, *20*(2), 109–127. https://doi.org/10.1038/S41576-018-0074-2

Bioconductor - DNAcopy. (n.d.). Retrieved October 30, 2022, from https://bioconductor.org/packages/release/bioc/html/DNAcopy.html

Bray, F., Ren, J. S., Masuyer, E., & Ferlay, J. (2013). Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer*, *132*(5), 1133–1145. https://doi.org/10.1002/IJC.27711

Callebaut, I., & Mornon, J. P. (2012). The PWAPA cassette: Intimate association of a PHD-like finger and a winged-helix domain in proteins included in histone-modifying complexes. *Biochimie*, *94*(9), 2006–2012. https://doi.org/10.1016/J.BIOCHI.2012.05.025

Chen, Y., Wan, B., Wang, K. C., Cao, F., Yang, Y., Protacio, A., Dou, Y., Chang, H. Y., & Lei, M. (2011). Crystal structure of the N-terminal region of human Ash2L shows a winged-helix motif involved in DNA binding. *EMBO Reports*, *12*(8), 797. https://doi.org/10.1038/EMBOR.2011.101

Cheng, Y., He, C., Wang, M., Ma, X., Mo, F., Yang, S., Han, J., & Wei, X. (2019). Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduction and Targeted Therapy 2019 4:1*, *4*(1), 1–39. https://doi.org/10.1038/s41392-019-0095-0

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, *339*(6121), 819–823. https://doi.org/10.1126/SCIENCE.1231143

Constantin, D., & Widmann, C. (2020). ASH2L drives proliferation and sensitivity to bleomycin and other genotoxins in Hodgkin's lymphoma and testicular cancer cells. *Cell Death & Disease 2021 11:11*, *11*(11), 1–16. https://doi.org/10.1038/s41419-020-03231-0

Cui, Z., Xie, M., Wu, Z., & Shi, Y. (2018). Relationship Between Histone Deacetylase 3 (HDAC3) and Breast Cancer. *Medical Science Monitor*, *24*, 2456–2464. https://doi.org/10.12659/MSM.906576

de Bruijn, D. R. H., & van Kessel, A. G. (2006). Common origin of the human synovial sarcoma associated SS18 and SS18L1 gene loci. *Cytogenetic and Genome Research*, *112*(3–4), 222–226. https://doi.org/10.1159/000089874

DepMap: The Cancer Dependency Map Project at Broad Institute. (n.d.). Retrieved January 24, 2023, from https://depmap.org/portal/

Doench, J. G. (2018). Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews. Genetics*, *19*(2), 67–80. https://doi.org/10.1038/NRG.2017.97

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology 2015 34:2*, *34*(2), 184–191. https://doi.org/10.1038/nbt.3437

Doudna, J. A., & Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science (New York, N.Y.)*, *346*(6213). https://doi.org/10.1126/SCIENCE.1258096

ES, V., & AB, O. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics (Oxford, England)*, *23*(6). https://doi.org/10.1093/BIOINFORMATICS/BTL646

Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, *21*(35), 5427–5440. https://doi.org/10.1038/SJ.ONC.1205600

Esteller, M. (2008). Epigenetics in cancer. *The New England Journal of Medicine*, *358*(11), 1148–1159. https://doi.org/10.1056/NEJMRA072067

Evers, B., Jastrzebski, K., Heijmans, J. P. M., Grernrum, W., Beijersbergen, R. L., & Bernards, R. (2016a). CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nature Biotechnology 2016 34:6*, *34*(6), 631–633. https://doi.org/10.1038/nbt.3536

Evers, B., Jastrzebski, K., Heijmans, J. P. M., Grernrum, W., Beijersbergen, R. L., & Bernards, R. (2016b). CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nature Biotechnology*, *34*(6), 631–633. https://doi.org/10.1038/NBT.3536

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., & Sander, J. D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, *31*(9), 822–826. https://doi.org/10.1038/NBT.2623

Ganesan, A., Arimondo, P. B., Rots, M. G., Jeronimo, C., & Berdasco, M. (2019). The timeline of epigenetic drug discovery: from reality to dreams. *Clinical Epigenetics*, *11*(1). https://doi.org/10.1186/S13148-019-0776-0

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J. B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., Joly, J. S., & Concordet, J. P. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology*, *17*(1), 1–12. https://doi.org/10.1186/S13059-016-1012-2/FIGURES/5

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/J.CELL.2011.02.013

Handy, D. E., Castro, R., & Loscalzo, J. (2011). Epigenetic Modifications: Basic Mechanisms and Role in Cardiovascular Disease. *Circulation*, *123*(19), 2145. https://doi.org/10.1161/CIRCULATIONAHA.110.956839

Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R., & Moffat, J. (2014). Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular Systems Biology*, *10*(7), 733. https://doi.org/10.15252/MSB.20145216

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., & Moffat, J. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, *163*(6), 1515–1526. https://doi.org/10.1016/J.CELL.2015.11.015

Hart, T., & Moffat, J. (2016). BAGEL: A computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics*, *17*(1), 1–7. https://doi.org/10.1186/S12859-016-1015-8/FIGURES/4

Hart, T., Tong, A. H. Y., Chan, K., van Leeuwen, J., Seetharaman, A., Aregger, M., Chandrashekhar, M., Hustedt, N., Seth, S., Noonan, A., Habsid, A., Sizova, O., Nedyalkova, L., Climie, R., Tworzyanski, L., Lawson, K., Sartori, M. A., Alibeh, S., Tieu, D., … Moffat, J. (2017). Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3: Genes, Genomes, Genetics*, *7*(8), 2719–2727. https://doi.org/10.1534/G3.117.041277/-/DC1

Heigwer, F., Kerr, G., & Boutros, M. (2014). E-CRISP: fast CRISPR target site identification. *Nature Methods 2014 11:2*, *11*(2), 122–123. https://doi.org/10.1038/nmeth.2812

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., & Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, *31*(9), 827–832. https://doi.org/10.1038/NBT.2647

Iorio, F., Behan, F. M., Gonçalves, E., Bhosle, S. G., Chen, E., Shepherd, R., Beaver, C., Ansari, R., Pooley, R., Wilkinson, P., Harper, S., Butler, A. P., Stronach, E. A., Saez-Rodriguez, J., Yusa, K., & Garnett, M. J. (2018). Unsupervised correction of

gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics*, *19*(1), 1–16. https://doi.org/10.1186/S12864-018-4989-Y/FIGURES/6

Kim, E., & Hart, T. (2021). Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier. *Genome Medicine*, *13*(1), 1–11. https://doi.org/10.1186/S13073-020-00809-3/FIGURES/3

Koike-Yusa, H., Li, Y., Tan, E. P., Velasco-Herrera, M. D. C., & Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, *32*(3), 267–273. https://doi.org/10.1038/NBT.2800

Kuang, Y., Wang, Y., Zhai, W., Wang, X., Zhang, B., Xu, M., Guo, S., Ke, M., Jia, B., & Liu, H. (2020). Genome-Wide Analysis of Methylation-Driven Genes and Identification of an Eight-Gene Panel for Prognosis Prediction in Breast Cancer. *Frontiers in Genetics*, *11*, 301. https://doi.org/10.3389/FGENE.2020.00301/BIBTEX

Kulis, M., & Esteller, M. (2010). DNA Methylation and Cancer. *Advances in Genetics*, *70*(C), 27–56. https://doi.org/10.1016/B978-0-12-380866-0.60002-2

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., & Liu, X. S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology*, *15*(12), 554. https://doi.org/10.1186/s13059-014-0554-4

Lin, A., Giuliano, C. J., Palladino, A., John, K. M., Abramowicz, C., Yuan, M. lou, Sausville, E. L., Lukow, D. A., Liu, L., Chait, A. R., Galluzzo, Z. C., Tucker, C., & Sheltzer, J. M. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine*, *11*(509). https://doi.org/10.1126/SCITRANSLMED.AAW8412

Linnekamp, J. F., Butter, R., Spijker, R., Medema, J. P., & van Laarhoven, H. W. M. (2017). Clinical and biological effects of demethylating agents on solid tumours - A systematic review. *Cancer Treatment Reviews*, *54*, 10–23. https://doi.org/10.1016/J.CTRV.2017.01.004

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9 NIH Public Access. *Science*, *339*(6121), 823–826. https://doi.org/10.1126/science.1232033

McBride, M. J., Pulice, J. L., Beird, H. C., Ingram, D. R., D'Avino, A. R., Shern, J. F., Charville, G. W., Hornick, J. L., Nakayama, R. T., Garcia-Rivera, E. M., Araujo, D. M., Wang, W. L., Tsai, J. W., Yeagley, M., Wagner, A. J., Futreal, P. A., Khan, J., Lazar, A. J., & Kadoch, C. (2018). The SS18-SSX Fusion Oncoprotein Hijacks BAF Complex Targeting and Function to Drive Synovial Sarcoma. *Cancer Cell*, *33*(6), 1128-1141.e7. https://doi.org/10.1016/j.ccell.2018.05.002

Medvedeva, Y. A., Lennartsson, A., Ehsani, R., Kulakovskiy, I. v., Vorontsov, I. E., Panahandeh, P., Khimulya, G., Kasukawa, T., & Drabløs, F. (2015). EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database : The Journal of Biological Databases and Curation*, *2015*. https://doi.org/10.1093/DATABASE/BAV067

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. v., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang, G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., … Tsherniak, A. (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics 2017 49:12*, *49*(12), 1779–1784. https://doi.org/10.1038/ng.3984

Miles, L. A., Garippa, R. J., & Poirier, J. T. (2016). Design, execution, and analysis of pooled in vitro CRISPR/Cas9 screens. *The FEBS Journal*, *283*(17), 3170–3180. https://doi.org/10.1111/FEBS.13770

Mittal, A., Hobor, F., Zhang, Y., Martin, S. R., Gamblin, S. J., Ramos, A., & Wilson, J. R. (2018). The structure of the RbBP5 β-propeller domain reveals a surface with potential nucleic acid binding sites. *Nucleic Acids Research*, *46*(7), 3802. https://doi.org/10.1093/NAR/GKY199

Mohammad, H. P., Barbash, O., & Creasy, C. L. (2019). Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. *Nature Medicine 2019 25:3*, *25*(3), 403–418. https://doi.org/10.1038/s41591-019-0376-8

Morel, D., Almouzni, G., Soria, J. C., & Postel-Vinay, S. (2017). Targeting chromatin defects in selected solid tumors based on oncogene addiction, synthetic lethality and epigenetic antagonism. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *28*(2), 254–269. https://doi.org/10.1093/ANNONC/MDW552

Morel, D., Jeffery, D., Aspeslagh, S., Almouzni, G., & Postel-Vinay, S. (2020). Combining epigenetic drugs with other therapies for solid tumours — past lessons and future promise. *Nature Reviews Clinical Oncology*, *17*(2), 91–107. https://doi.org/10.1038/S41571-019-0267-4

Munoz, D. M., Cassiani, P. J., Li, L., Billy, E., Korn, J. M., Jones, M. D., Golji, J., Ruddy, D. A., Yu, K., McAllister, G., Deweck, A., Abramowski, D., Wan, J., Shirley, M. D., Neshat, S. Y., Rakiec, D., de Beaumont, R., Weber, O., Kauffmann, A., … Schlabach, M. R. (2016). CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discovery*, *6*(8), 900–913. https://doi.org/10.1158/2159-8290.CD-16-0178

Pfister, S. X., & Ashworth, A. (2017). Marked for death: targeting epigenetic changes in cancer. *Nature Reviews Drug Discovery 2017 16:4*, *16*(4), 241–263. https://doi.org/10.1038/nrd.2016.256

Plass, C., Pfister, S. M., Lindroth, A. M., Bogatyrova, O., Claus, R., & Lichter, P. (2013). Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nature Reviews Genetics 2013 14:11*, *14*(11), 765–780. https://doi.org/10.1038/nrg3554

Ritchie, M. E., Dai, Z., Sheridan, J. M., Gearing, L. J., Moore, D. L., Su, S., Wormald, S., Wilcox, S., O'Connor, L., Dickins, R. A., & Blewitt, M. E. (2014). edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. *F1000Research*, *3*. https://doi.org/10.12688/F1000RESEARCH.3928.2

Saghafinia, S., Mina, M., Riggi, N., Hanahan, D., & Ciriello, G. (2018). Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Reports*, *25*(4), 1066-1080.e8. https://doi.org/10.1016/J.CELREP.2018.09.082

Sanjana, N. E., Shalem, O., & Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. *Nature Methods 2014 11:8*, *11*(8), 783–784. https://doi.org/10.1038/nmeth.3047

Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., & Doench, J. G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications 2018 9:1*, *9*(1), 1–15. https://doi.org/10.1038/s41467-018-07901-8

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., & Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science (New York, N.Y.)*, *343*(6166), 84–87. https://doi.org/10.1126/SCIENCE.1247005

Stahl, M., Gore, S. D., Vey, N., & Prebet, T. (2016). Lost in translation? Ten years of development of histone deacetylase inhibitors in acute myeloid leukemia and myelodysplastic syndromes. *Http://Dx.Doi.Org/10.1517/13543784.2016.1146251*, *25*(3), 307–317. https://doi.org/10.1517/13543784.2016.1146251

Steinhart, Z., Pavlovic, Z., Chandrashekhar, M., Hart, T., Wang, X., Zhang, X., Robitaille, M., Brown, K. R., Jaksani, S., Overmeer, R., Boj, S. F., Adams, J., Pan, J., Clevers, H., Sidhu, S., Moffat, J., & Angers, S. (2016). Genome-wide CRISPR screens reveal a Wnt–FZD5 signaling circuit as a druggable vulnerability of RNF43-mutant pancreatic tumors. *Nature Medicine 2016 23:1*, *23*(1), 60–68. https://doi.org/10.1038/nm.4219

Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J., & Mateo, J. L. (2015). CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLOS ONE*, *10*(4), e0124633. https://doi.org/10.1371/JOURNAL.PONE.0124633

Steward, M. M., Lee, J. S., O'Donovan, A., Wyatt, M., Bernstein, B. E., & Shilatifard, A. (2006). Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nature Structural & Molecular Biology*, *13*(9), 852–854. https://doi.org/10.1038/NSMB1131

Tang, S., Sethunath, V., Metaferia, N. Y., Nogueira, M. F., Gallant, D. S., Garner, E. R., Lairson, L. A., Penney, C. M., Li, J., Gelbard, M. K., Alaiwi, S. A., Seo, J. H., Hwang, J. H., Strathdee, C. A., Baca, S. C., AbuHammad, S., Zhang, X., Doench, J. G., Hahn, W. C., … Viswanathan, S. R. (2022). A genome-scale CRISPR screen reveals PRMT1 as a critical regulator of androgen receptor signaling in prostate cancer. *Cell Reports*, *38*(8). https://doi.org/10.1016/J.CELREP.2022.110417

Tarver, T. (2012). Cancer Facts & Figures 2012. American Cancer Society (ACS). *Http://Dx.Doi.Org/10.1080/15398285.2012.701177*, *16*(3), 366–367. https://doi.org/10.1080/15398285.2012.701177

Tzelepis, K., Koike-Yusa, H., de Braekeleer, E., Li, Y., Metzakopian, E., Dovey, O. M., Mupo, A., Grinkevich, V., Li, M., Mazan, M., Gozdecka, M., Ohnishi, S., Cooper, J., Patel, M., McKerrell, T., Chen, B., Domingues, A. F., Gallipoli, P., Teichmann,

S., … Yusa, K. (2016). A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Reports*, *17*(4), 1193–1205. https://doi.org/10.1016/J.CELREP.2016.09.079

Vinet, M., Suresh, S., Maire, V., Monchecourt, C., Némati, F., Lesage, L., Pierre, F., Ye, M., Lescure, A., Brisson, A., Meseure, D., Nicolas, A., Rigaill, G., Marangoni, E., del Nery, E., Roman-Roman, S., & Dubois, T. (2019). Protein arginine methyltransferase 5: A novel therapeutic target for triple-negative breast cancers. *Cancer Medicine*, *8*(5), 2414–2428. https://doi.org/10.1002/CAM4.2114

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., & Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science (New York, N.Y.)*, *350*(6264), 1096–1101. https://doi.org/10.1126/SCIENCE.AAC7041

Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science (New York, N.Y.)*, *343*(6166), 80–84. https://doi.org/10.1126/SCIENCE.1246981

Wang, T., Yu, H., Hughes, N. W., Liu, B., Kendirli, A., Klein, K., Chen, W. W., Lander, E. S., & Sabatini, D. M. (2017). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell*, *168*(5), 890-903.e15. https://doi.org/10.1016/J.CELL.2017.01.013/ATTACHMENT/F8541A23-1E22-4719-A9B5-EEB75C35099E/MMC7.XLSX

Wilson, B. G., Wang, X., Shen, X., McKenna, E. S., Lemieux, M. E., Cho, Y. J., Koellhoffer, E. C., Pomeroy, S. L., Orkin, S. H., & Roberts, C. W. M. (2010). Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell*, *18*(4), 316–328. https://doi.org/10.1016/J.CCR.2010.09.006

Wu, C., Lyu, J., Yang, E. J., Liu, Y., Zhang, B., & Shim, J. S. (2018). Targeting AURKA-CDC25C axis to induce synthetic lethality in ARID1A-deficient colorectal cancer cells. *Nature Communications 2018 9:1*, *9*(1), 1–14. https://doi.org/10.1038/s41467-018-05694-4

Xu, N., Wu, Y. P., Ke, Z. bin, Liang, Y. C., Cai, H., Su, W. T., Tao, X., Chen, S. H., Zheng, Q. S., Wei, Y., & Xue, X. Y. (2019). Identification of key DNA methylation-driven genes in prostate adenocarcinoma: An integrative analysis of

TCGA methylation data. *Journal of Translational Medicine*, *17*(1), 1–15. https://doi.org/10.1186/S12967-019-2065-2/FIGURES/14

Yang, H., Cui, W., & Wang, L. (2019). Epigenetic synthetic lethality approaches in cancer therapy. *Clinical Epigenetics*, *11*(1), 1–7. https://doi.org/10.1186/S13148-019-0734-X/FIGURES/1

Yedier-Bayram, O., Gokbayrak, B., Kayabolen, A., Aksu, A. C., Cavga, A. D., Cingöz, A., Kala, E. Y., Karabiyik, G., Günsay, R., Esin, B., Morova, T., Uyulur, F., Syed, H., Philpott, M., Cribbs, A. P., Kung, S. H. Y., Lack, N. A., Onder, T. T., & Bagci-Onder, T. (2022). EPIKOL, a chromatin-focused CRISPR/Cas9-based screening platform, to identify cancer-specific epigenetic vulnerabilities. *Cell Death & Disease 2022 13:8*, *13*(8), 1–11. https://doi.org/10.1038/s41419-022-05146-4

Zhou, Z., Rajasingh, S., Barani, B., Samanta, S., Dawn, B., Wang, R., & Rajasingh, J. (2018). Therapy of Infectious Diseases Using Epigenetic Approaches. *Epigenetics in Human Disease*, 689–715. https://doi.org/10.1016/B978-0-12-812215-0.00022-4

Appendix A: Code Availability

The codes used in the analysis are available through GitHub at https://github.com/ezgikrt/EPIKOL