Universiteit Leiden

Master Computer Science

Streamlining Computational Workflow of Dual CRISPR Screen Library Design and Screen Analysis

Name:	Minkang
Student ID:	Tan]s2160293
Date:	28/06/2020
Specialisation:	Data Science: CS
1st supervisor:	Dr. Erwin Bakker (LIACS)
2nd supervisor:	Dr. Baoxu Pang (LUMC)
	Dr. M.S.K. Lew (LIACS)

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

The gene coding region only represents less than 2% of the human genome, and the non-coding genome also plays a vital role in normal cellular activities by regulating the expression of genes. However, the current research on the function of regulatory elements in non-coding regions is still under-studied. To determine the function of the non-coding regulatory regions, a dual-CRISPR screen system was developed to produce genome-wide deletions of non-coding DNA and study the biological consequences. This unique CRISPR screen application still calls for a novel, standardized, end-to-end computational pipeline.

This work focused on the design and analysis pipeline for the genome-wide dual CRISPR screen system targeting the non-coding regulatory regions in the human genome. The dual CRISPR screen library design pipeline is based on an empirical scoring model, which aims to generate multiple guide RNA pairs with high target efficiency and low off-target rate, for each potential regulatory region in the genome. The CRISPR screen analysis pipeline deconvolutes pooled screen results and integrates heterogeneous data to identify non-coding elements that are significantly enriched in specific biological processes. The benchmark performance showed that the proposed pipeline algorithm for the screen analysis outperformed the other two mainstream screen analysis tools for its sensitivity to identify essential genes. Also, the selected pipeline could also perform bias correction and normalization of the CRISPR screens and provide solutions for visualization and clustering analysis of screen results, which enhance the power of mining biology information from dual CRISPR screen experiments.

This computational pipeline has been tested using CRISPR screen datasets produced in the group for the library design and data analysis. Applications include the CRISPR screen analyses of identification of membrane importers of anthracycline drugs, identification of non-coding regulatory regions for drug resistance and potential immunotherapy targets discovery using dual CRISPR screen.

KEY WORDS: Non-coding genome; Dual-CRISPR screen; Computational pipeline; Guide RNA design

Table of contents

Chapter 1

Introduction

1.1 Noncoding DNA

Deoxyribonucleic acid (DNA) is the basic building block that carries the genetic information necessary for the synthesis of ribonucleic acid (RNA) and protein in biological cells. It is an essential biological macromolecule for the development and normal operation of organisms. The combinations of 4 nucleotides in DNA constitutes genetic information. The genetic information can be transcribed into RNA, which serves as the intermediate template to be translated into polypeptides to form proteins.

Non-coding DNA sequences are components of the genome that do not encode protein sequences. Recent studies have shown that less than 2% of the human genome contains the genetic information for protein sequences. Small part of the remaining non-coding region DNA is transcribed into functional non-coding RNA molecules, such as transfer RNA, ribosomal RNA and regulatory RNA. In addition, many non-coding DNA plays an important role in the regulation of transcription and translation of protein-coding sequences, scaffold attachment regions, DNA replication origin, centromere, and telomere.

The International Encyclopedia of DNA elements (ENCODE) project found that a large percentage of human genomic DNA may have certain biological functions, breaking the long-standing controversy of calling non-coding regions "junk DNA", for example, enhancers, promoters, and silencers. Functional genomics studies of non-coding regions have also found that non-coding DNA is an important participant in epigenetic activities and complex genetic interaction networks.

1.2 Dual CRISPR Screen System

1.2.1 CRIPSR/Cas9 Gene Editing Technology

The CRISPR/Cas9 (Clustered regularly interspaced short palindromic repeat sequences/CRISPR-associated protein 9) technology emerged in recent years has become a powerful tool for systematic analysis of genomes due to its convenient and precise gene editing feature.

The CRISPR/Cas9 system is composed of single guide RNA (sgRNA) and Cas9 endonuclease. Among them, the sgRNA is composed of a constant part (multiple stem loops form a fixed scaffold sequence that binds to Cas9) and a variable part (20 bp sequence complementary to the target DNA sequence at the 5'end). The unidirectional RNA can guide the Cas9 protein to the target DNA site for the specific cleavage once recognizing the protospacer adjacent motif (PAM) site. The Cas9 protein from S. pyogenes (SpyCas9) is the most commonly used one which recognizes PAM motifs with a 5'-NGG-3' pattern. If the PAM motif is not present in the target sequence, the Cas9 protein will not cleave the sequence. Only when the target protospacer is complementarity to the 20-bp crRNA sequence and the specific Cas proteins bind the PAM sequence, the CRISPR/Cas9 system will generate a DNA double-strand break (DSB).

After the double-strand break formed, the nonhomologous end-joining (NHEJ) and homologous recombination (HDR) mechanisms within the cell will be activated to repair the broken DNA. For the NHEJ mechanism, the DNA sequence may not be restored as it was and the induction of insertions or deletions (indels) will result in the loss of gene function.

1.2.2 Dual CRISPR Screen system

Most DNA in the genome does not encode proteins, but many of these non-coding regions play a role in the regulation of gene expression. Regulatory elements of non-coding regions, such as promoters, enhancers, silencers, and insulators, play a vital role in the gene regulatory network. Many studies have also shown that more than 90% of disease-related genetic variations occur in non-coding regions. In order to understand how these non-coding regions participate in the regulation of gene expression, CRISPR/Cas screening technology has been used by researchers to identify non-coding regions with specific biological functions.

The widely used CRISPRko screening technology based on single-guide RNA introduces indel to cause the loss of function of the protein-coding genes. However, this technology still faces challenges when it is applied to screen the non-coding DNAs. The length of the non-coding regions is generally more than 200bp, such as transcription factor (TF) binding sites. The insertion or deletion of short fragments (indel) produced by a single-guide-mediated

DNA double-strand break, is not sufficient to completely abolish the binding of the TF of the non-coding region.

To achieve the loss-of-function screening of non-coding DNA, a lentiviral dual CRISPR screening technology based on paired gRNA (pair guide RNA, pgRNA) was proposed. As shown in Figure 1.1, this system uses two different gRNAs targeting both the upstream and downstream of the non-coding regulatory regions. Each gRNA can bind to the catalytically active Cas9 protein and simultaneously cause cleavage at the upstream and downstream of the target site, therefore completely removing the non-coding regions from the genome. Then the effects of this loss of the non-coding regions could be studied in the context of many biological processes, such as cell growth and cell differentiation. This cleavage activity will repair DNA breaks through the NHEJ mechanism of non-homologous ends of the cells. This dual CRISPR library screening method has been proved to achieve high efficiency and high throughput screening.

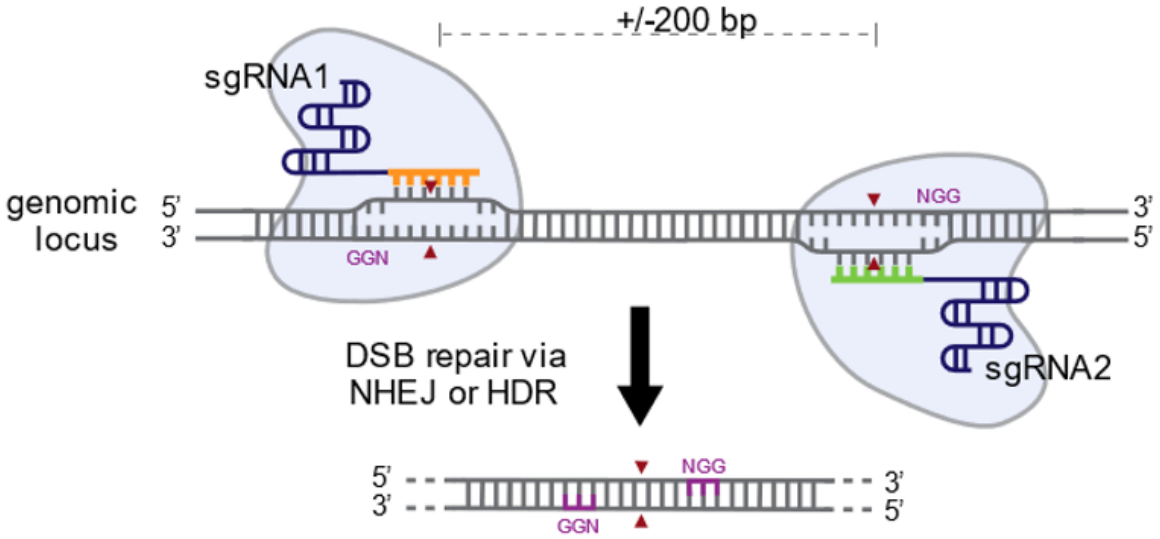


Fig. 1.1 A schematic of the dual CRISPR lentiviral system.

1.3 Computational pipeline for dual CRISPR screening

A typical workflow of CRISPR screen experiment consists of library preparation, viral transduction, selection for infected cells, drug treatment, genomic DNA isolation, PCR amplification, high-throughput sequencing and bioinformatics analysis. The simplicity of the CRISPR-Cas system enables large-scale experimentation, which brings new challenges for the design and analysis of CRISPR screen experiment. To solve these challenges, numerous

software tools and analytical methods have been developed, including resources to design optimal guide RNAs for various modes of manipulation and to analyze the data from pooled genetic screens.

As a novel method to study the function of non-coding regions, dual CRISPR screening calls for customized software tools for the unique bioinformatic design and analysis requirements of CRISPR deletion. First, as the dual CRISPR knockout approach employs two guide RNAs flanking the target non-coding region, the optimal sgRNA pairs are required before performing the screen studies. To select the optimal individual protospacer targeting sequences (key component of sgRNA), two key criteria need to be considered: (1) maximizing the efficiency of the designed sequence at generating mutations, and (2) minimizing "off-targeting" effect, or the propensity for recognizing similar but undesired sites in the genome. Second, next-generation sequencing for a dual CRISPR pooled screen primarily returns demultiplexed FASTQ files of all samples. To leverage the results for pooled screening analysis, screen deconvolution was performed. During deconvolution, the guide sequence is extracted from the read and aligned to the guide library, resulting in a read count table of the total number of reads per guide in each sample. Third, screen performance is assessed at the level of individual paired guides before mapping to guide pairs to non-coding regions. After that, screen analysis tools could identify statistically significant hits (essential non-coding regions) from guide-level performance data. Optional data improvement and custom analysis could be added to interrogate the screen results.

As this field continues to progress and numerous tools emerge, a clear ongoing challenge is not only to innovate, but to actively maintain and improve existing tools to obtain high-quality library design and reliable screen analysis for dual CRISPR experiments. To address that, we present a comprehensive computational pipeline integrating custom analysis scripts and existing set of excellent computational resources for CRISPR-Cas experiments. This user-friendly and end-to-end solution enables convenient deployment and efficient execution on the diverse computation platforms. Our pipeline is capable of designing optimised pairs of guide RNAs to perform genome-wide knockout screens of non-coding DNAs. Furthermore, our pipeline is compatible with CRISPR screens directed by both single guide RNA and pair guide RNAs in screen analysis steps like screen deconvolution and hit identification. Our method balances sensitivity and specificity by generating consensus predictions from multiple mature algorithms. Benchmarking on ground-truth sets demonstrates the improved sensitivity we achieved with respect to existing methods.

Chapter 2

Related Work

This chapter described the recent research of computational tools and pipelines for CRISPR screen experiments, mainly focusing on the guide design and CRISPR screen analysis tools. The guide design tools could be divided into three classes: alignment-based, hypothesis-driven and learning-based. Two mainstream screen analysis algorithms: parametric and non-parametric methods are also compared to identify the suitable ones for our research. Finally, based on the testing and comparison, comprehensive and integrated computational workflow solutions for CRISPR screen experiments are made.

2.1 Guide RNA Design Tools

CRISPR technology has transformed many fields of biology research. As it applications broadened, many software tools have been developed to design the guide RNA that paired with the protospacer targeting sequence. For most of these tools, the key factors for guide RNA evaluation are:

1. The efficiency of a given guide RNA sequence (the proper recognition of the target sequences);

2. The off-target possibility (the tendency to identify similar but undesired genomic sites that are unrelated to the target sequences).

Some scoring models that are based on experimental data have been developed to predict the efficiency of gRNAs. Most gRNA design tools can be used to identify an unique targeting site throughout the genome, with reduced possible off-target effects. The current sgRNA design tools can be roughly divided into three types:

Alignment-based Methods

This method is based on sequence alignment, that is, the choice of sgRNA is determined only by the position of PAM;

Hypothesis-driven Methods

This method is based on hypothetical prior knowledge, which comprehensively considers the impact of multiple specific factors (such as GC content, exon position, etc.) for the sgRNA target efficiency;

Learning-based Methods

The learning-based method is using a comprehensively trained model to consider different characteristics, and then to predict the sgRNA cleavage efficiency.

The latter two types of tools perform better than the alignment-based method, because they consider different sequences and chromatin features, and integrate machine learning models for prediction. Different calculation tools should be selected according to different sgRNA design scenarios: when users only need to design sgRNA based on PAM, it is recommended to use CRISPRseek [1] and Cas-Offinder [2]. These two tools can also be used for other non-traditional Cas proteins. For high-efficiency gRNA design, optimal prediction performance is preferred. Learning-based tools such as sgRNA-designer [3], CRISPRscan [4], and SSC [5] that use update design rules perform much better than others.

2.2 CRISPR Screen Analysis Tools

After the screen experiment, next-generation sequencing is used to determine the abundance of gRNA sequences in the samples under different perturbation conditions. However, due to the large-scale and aggregated screening results obtained by CRISPR screening and the heterogeneity of different sample sources, a statistical algorithm is needed to properly analyze the screen results.

Many screening tools have been developed to identify the significantly selected genes from the CRISPR genome-wide screening results. The most critical step is to statistically determine the positive/negative selectivity of the selected genes. At present, such tools can be roughly divided into two categories, based on the computational models used:

2.2.1 Parametric-based Methods

Parametric-based methods use statistical model to fit the sgRNA screen data which is usually very overdispersed. Choices of models include Poisson, negative binomial, or Gaussian distribution which are commonly used to model read counts in RNA-seq analysis. Then, statistical hypothesis test (e.g. Student's t-test) is applied to measure differences in sgRNA levels based on the aforementioned distribution assumption. Finally, to identify the candidate genes, combined probability test (e.g. Fisher's method, RRA and MLE) is used to combine the sgRNA-level significance scores (P-value) of each gene and determine the gene-level significance.

Parametric methods such as MAGeCK [6], CB2 [7], HitSelect [8] and PinAPL-Py [9] have different performances because they use slightly different distribution assumptions. And the significant scores generated by each tool are also different. MAGeCK usesa custom RRA significant score, and CB2 uses through the FDR error detection rate performance.

2.2.2 Nonparametric-based Methods

Nonparametric-based methods abandons the scheme of parametric modeling of gRNAs in the screen results. Because of the limitations of parametric modeling itself, it is not optimal to characterize the true distribution of data. Using substitution or other non-parametric testing methods can more flexibly process heterogeneous data and directly define the gene-level saliency selection indicators.

Non-parametric methods such as PBNPA [10] and RIGER [11] are more suitable for saliency analysis to identify genes in different scenarios, but are slightly worse in sensitivity than parameter methods.

In general, the consensus of significant genes identified by various tools during a positive selection is high. However the results of negative selections are often mixed with a certain percentage of false-positive components, which is due to many confounding factors in the experiment (including sgRNA cleavage efficiency and targeted copy number gain). Such limitations can be alleviated by more appropriate model assumptions or a subsequent deviation correction. In practice, it is also possible to combine the results from multiple screening analysis tools to identify the consensus list of genes.

2.3 CRISPR Screen Computational Pipeline

Although many individual tools have been developed for CRISPR screening experiments, there is currently no standardized and complete calculation process tool from library design to screening result processing and analysis.

For the design of dual CRISPR screening libraries, there are also some tools such as CRISPETa [12] that can be used to design screening libraries (targeting lncRNA and other non-coding regions), with a website design capabilities. However, its calculation process cannot implement flexible parameter configuration (such as defining the length of the regions for gRNA search), and its calculation process does not support the batch design of the target screening libraries deployed in the calculation cluster. The input format of CRISPETa must provide the search area sequence information, and the output format also does not support personalized configuration, for example, the oligonucleotide sequence connecting two gRNAs cannot be customized.

In the analysis of the results of dual CRISPR screening experiments, MAGeCK-VISPR [13] and MAGeCKFlute [14] can be used for the screening experiment analysis from sequencing raw data comparison to screening gene function annotation end-to-end analysis process, but the reliability of its data analysis results completely depends on the quality of its internally integrated gRNA comparison to the screening library. The MAGeCK RRA/MLE algorithm cannot achieve the quality control and comparison of the original data of the dual CRISPR screening and sequencing, nor can it achieve the personalized processing of the dual CRISPR screening results (such as correcting the false positive results caused by the targeted copy number variation regions). There is an urgent need for a more flexible and effective design and analysis pipeline for dual CRISPR screening experiments.

Chapter 3

Method

In this chapter, the pipeline programming framework Snakemake will be introduced, as well as the design schematics of three pipelines for dual CRISPR screen experiment: dual CRISPR screen library design pipeline, paired gRNA quantification pipeline and CRISPR screen analysis pipeline.

3.1 Snakemake Workflow Management System

Snakemake [15] is a tool for bioinformatics analysis workflow management. It is based on the powerful execution environment of Python, where the workflows are described with human-readable grammar rules, thereby creating a reproducible and extensible bio-data analysis pipeline.

Snakemake defines a workflow in a 'Snakefile' through a domain-specific language similar to the standard Python syntax. The workflow is consists of rules that denote how to create output files from input files. The workflow is implied by dependencies between the rules that arise from one rule needing an output file of another as an input file. The input and output files of each rule can contain multiple named wildcards, which can be used to determine the actual file name and further infer the rule dependencies.

Snakemake pipeline can be easily deployed on different computational platforms such as servers, clusters, grids and cloud environments. There is no need to modify the pre-defined rules and the job execution will be limited by the available computational resource such as CPU cores, memory or the number of GPUs. Snakemake can conveniently manage workflow-related software versions using Conda or Singularity. Because Snakemake rule can be defined in the form of Shell commands, pure Python code, external Python, or R scripts, it is easy to integrate software tools written by different languages into the workflow.

Upon invocation of the defined workflow, Snakemake will create a directed acyclic graph (DAG), which represents the plan for rule execution. An example of Snakemake DAG is shown in Figure 3.1. The node of DAG denotes one specific job and the directed edge between job A and job B means that the rule below job B requires the output of job A as an input file. The path in the DAG represents a series of operations that must be performed sequentially. The two disjoint paths in DAG can be executed in parallel, and since a single job can use multiple threads, it is possible to optimize the CPU usage by specifying the number of thread usages of a job given the available kernel threshold.

Fig. 3.1 Example of Snakemake workflow DAG.

In order to efficiently design different types of non-coding functional regions at the same time, such as enhancers and ultra-conserved regions, the computational process of dual CRISPR screening library design was extended in the Snakemake execution environment. In the meanwhile, Snakemake is also used to quantify the paired gRNA abundance in multiple samples and identify the essential regions in different perturbations to ensure the scalability and reproducibility of the screen analysis results.

3.2 Dual CRISPR Screen Library Design Pipeline

In order to design a good screen library for the dual CRISPR screen experiments, a Snakemake-based computational pipeline was made to design paired guide RNA with high targeting efficiency and low off-target rate to accurately target the genomics regions of our interest.

The sgRNA sequence is composed of protospacer sequence and a constant scaffold region. The main goal of sgRNA design is to select the optimal protospacers, that is, the 20bp sequence before the PAM sequence, to achieve efficient and specific targeting of the target area. The design of pgRNA for dual CRISPR screening systems mainly includes the following three steps: defining the search regions, selecting protospacer sequences, and prioritizing gRNA pairs. The schematic of dual CRISPR screen library design is shown in Figure 3.2.

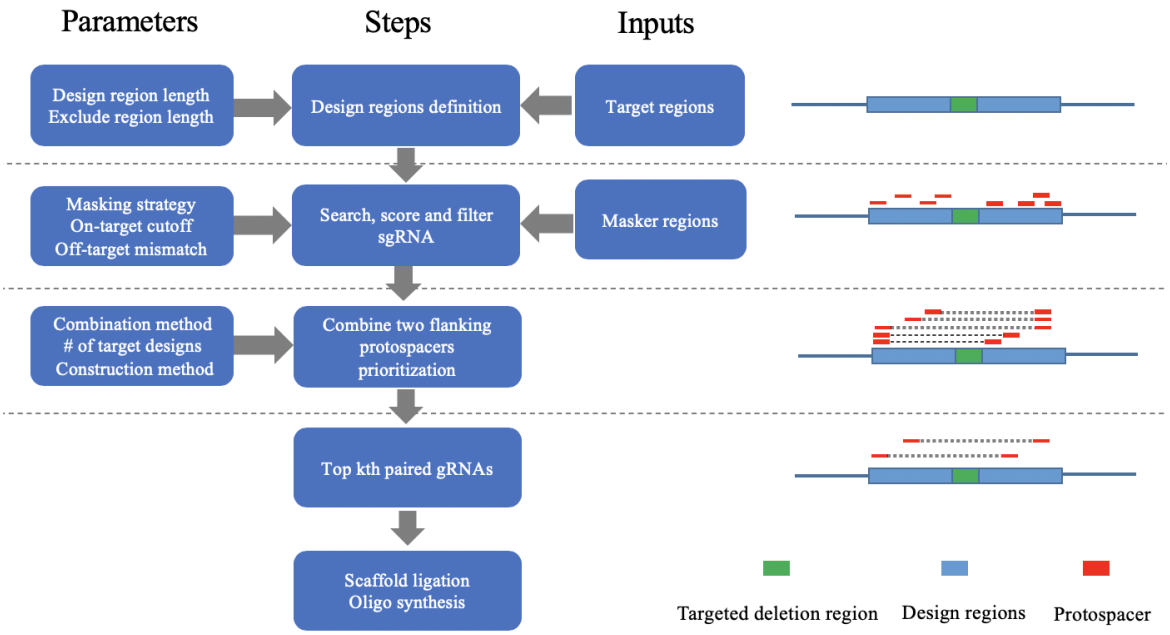


Fig. 3.2 Schematic of dual CRIPSR screen library design pipeline.

The main principle to make the tool is to have flexibility and scalability during the process of design. Users can adjust the design scheme and algorithm parameter configurations according to their requirements, otherwise it will be set as default. The algorithm can perform parallel pgRNA design: targeting multiple regions according to the size of the screen library. The user-definable program variables and their default parameter values are set as shown in Table3.1.

Table 3.1 Configuration of dual CRISPR screen library design pipeline.

Parameter	Default	Description
Target regions		Input file path in BED format
Positive masker region		Favorable design region in BED format
Negative masker reegion		Unfavorable design region in BED format
Output suffix		Suffix for output design screen library file path name
# of design pairs	15	Maximum # of gRNA pairs return
Design region length	200bp	Length of flanking regions for candidate protospacers search
Exclusion region length	50bp	Length of flanking regions for separating target and design regions
PAM	NGG	PAM motif pattern to search candidate protospacer part of gRNA
# of off-target mismatch	1,0,0,x,x	Maximum # of 1,2,3,4,5 base(s) mismatch for gRNA protospacer
sgRNA on-target cutoff	0.2	Minimum prediction on-target efficiency of sgRNA
pgRNA on-target cutoff	0.4	Minimum prediction on-target efficiency of pgRNA
Diversity	0.5	Maximum fraction of guide pair design containing unique sgRNA
Combination method	sum	sum or product operation to combine two flanking sgRNA scores

3.2.1 Design Region Definition

The target regions on the genome are usually provided in BED format, which mainly contains the name of the chromosome, the start, and the end coordinate.

Firstly, the gRNA design regions can be defined within a fixed distance upstream and downstream of the target deletion regions. Users can also choose to separate design regions from the target deletion regions at an interval of user-defined length (exclusion regions). Secondly, users can also specify two types of masker regions, positive masker region and negative masker region. The gRNA designs falling into the positive masker regions will be prioritized and the designs in the negative masker region will have lower priority. The positive masker region mainly refers to the open chromatin regions sensitive to DNase I enzyme while the negative masker region includes repeated regions and heterochromatin regions. Thirdly, the deletion of non-coding regions can also apply different library construction methods, such as defining the gRNA search area within a given length within the 5' and 3' sides of the target region or specifying the fixed distance between the upstream and downstream of the target area as the gRNA design regions. Candidate protospacer sequences which complementarily

bind to the 20bp of sgRNA 5' end are distributed within these defined design regions. Most of the above operations can be implemented by the intersection method of BEDtools [16].

3.2.2 Protospacer Selection

To find candidates for gRNA designs, CRISPRseek [1] is used to scan all feasible protospacer sequence which is defined as 20mer sites adjacent to the PAM motif pattern (5'-NGG-3') in the design region.

When predicting the cutting efficiency of DNA double-strand breaks induced by the Cas9 protein directed by the gRNA, logistic regression model proposed by sgRNA Designer [3] is used to predict the targeting efficiency of candidate protospacers. The scoring algorithm uses a supervised learning method and the prediction model was trained on the experimental data of 6085 and 1151 sgRNAs distributed on 6 mice and 3 human genes respectively. To acquire the nucleic acid preference information, sequence features are extracted from the internal 20bp sequence of the sgRNA design as well as its adjacent 5bp sequence. The prediction model incorporates 72 sequence features including position-specific motifs of single nucleotide and dinucleotide and also GC counts of sgRNA to predict the target efficiency of sgRNA.

When analyzing the potential off-target effects of the sgRNA design, the CRISPR Analyser tool [17] was used. By specifying the maximum number of sites with a given number of mismatches inside 20bp sequence, CRISPR Analyser can search for all possible genomic off-target sites the design sgRNA might bind to. The default off-target truncation threshold is set as "0: 1, 1: 0, 2: 0, 3: x, 4: x", indicating that there are no other genomic sites with 2 mismatches in the sequence, and the symbol "x" indicates gigantic.

Combined with the threshold of the predicted target efficiency and the number of off-target bindings, most candidate protospacer sequences located within non-coding regions can be filtered, which could ensure the high efficiency and specificity of the final gRNA design.

3.2.3 Prioritizing gRNA pairs

To select the optimal pairs of gRNAs which can efficiently and accurately delete the regions of interest, the pipeline first enumerates all feasible paired gRNAs combinations from both sides of target regions and prioritizes them by an empirical ranking method. Here we mainly refer to the gRNA target efficiency score predicted by the logistic regression scoring model, and combine the predicted scores of each pair of gRNAs (addition or multiplication), and also consider the length of the deleted sequence of pgRNA. Experimental data shows that when the length of the sequence (including the target interval) actually targeted by pgRNA is

short, it can be deleted more effectively. However, only considering actual deletion length may lead to the sgRNA design on both sides tends to be concentrated in a limited range on the design regions, which does not cover the complete design area well. The empirical ranking method takes the predictive scores for each pair of pgRNA into consideration and makes the sgRNA designs evenly distributed within the design region. In the meanwhile, the algorithm may only consider a few sgRNA designs with higher prediction efficiency and these sgRNAs tend to appear in almost all pgRNA designs, which is not optimal to the diversity of design libraries. To alleviate this problem, the algorithm introduces a diversity indicator to control the proportion of returned pgRNA containing the same sgRNA. As a default, the maximum fraction of unique sgRNA in the final design is 0.5. Considering the difference in plasmid construction methods, promoters that start expressing sgRNA have a certain sequence preference for sgRNA. In some designs, U6 promoter requires the 5' end sgRNA linked to it to be guanine nucleotide "G". According to this limitation, the returned pgRNA pairs can be further filtered, and the user-specified maximum number of higher priority pgRNA pairs can be output.

Finally, for the convenience of screen library synthesis, the algorithm will assemble the output pgRNA and partial scaffold sequence of sgRNA on both sides to generate oligonucleotides sequences to be synthesized. The scaffold sequence contains sites for restriction enzyme cleavage, to insert the full length scaffold sequence (200bp) by Gibson assembly. The synthesized oligonucleotides sequences in the final screen library share the same scaffold and only the paired sgRNA parts are unique sequences . This may result in unwanted homologous recombinations among guide RNA pairs. To alleviate the homologous recombination between the synthesized oligonucleotides, the program introduces a fixed-length (10bp) random sequence in the middle of the designed sequences.

3.3 Paired gRNA Quantification Pipeline

To identify the function of specific genes or non-coding regions, CRISPR screen experiment compares the effects of knocking out target genes or non-coding regions. The common practice is to determine and compare the gRNA abundance of the target genes or non-coding regions among the samples from different perturbation conditions. The changes of gRNA abundance between treatment and control samples are used as the signature to determine whether the candidate genes or non-coding regions play important roles in phenotypes like cell proliferation and survival.

The Illumina sequencers can simultaneously sequence libraries from different samples during a single run through via barcoding the samples, which greatly saves the time and

cost of sequencing. The method of multiplexing sequencing is to add a unique barcode to each DNA fragment during library preparation, so that reads from different samples can be identified and separated. CRISPR screen experiments normally require multiple perturbations and each case may have multiple biological repeats. Another computational pipeline was developed to handle this scalable and reproducible computation task. The schematic of paired gRNA quantification pipeline is shown in Figure 3.3:

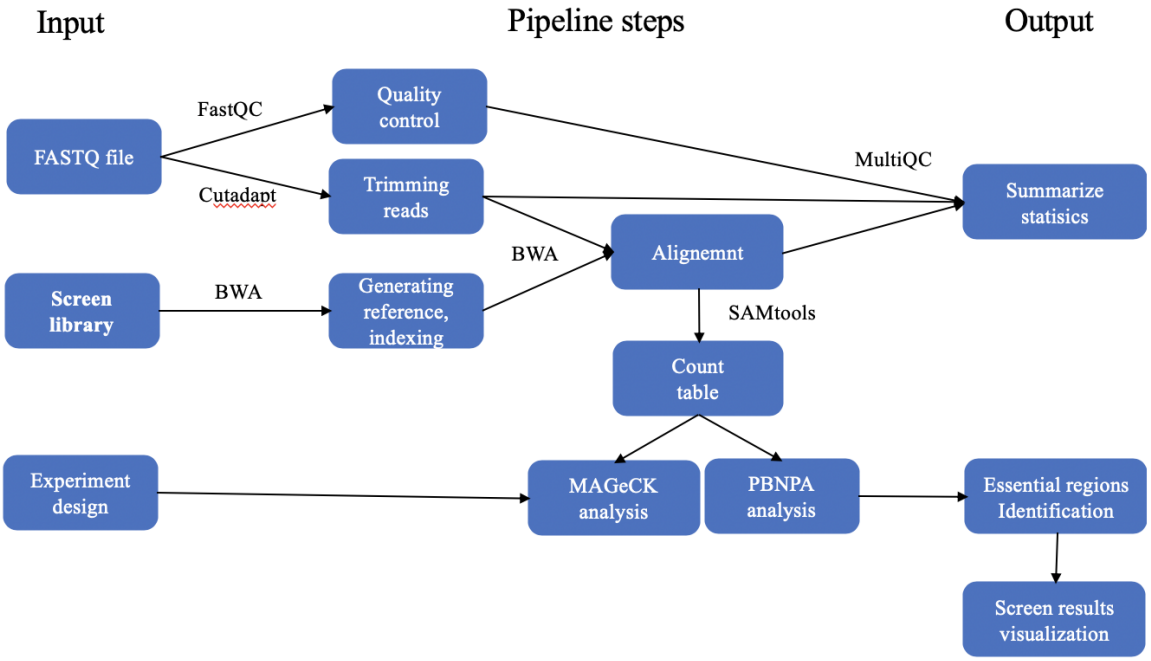


Fig. 3.3 Schematic of paired gRNA quantification pipeline.

This computational pipeline allows automatically quantifying pgRNA abundance in different samples. It also allows users to customize program running configurations. All the parameter settings and default values are shown in Table 3.2:

3.3.1 Reference Preparation and Indexing

Since each pair of pgRNA uniquely targets a certain region with fixed deletion length, thus the abundance of pgRNA in each sample can be determined by sequence alignment. The dual CRISPR screen library design contains the information of design ID, coordinate information, protospacer and scaffold sequence for each pgRNA, which could be used as the alignment reference. To implement the aligner tools such as BWA-MEM [18] and Bowtie2 [19], the pipeline first constructs the FASTA file and generates the index file based on the designed screen library automatically. In theory, each pair of reads generated by Illumina pair-end

Table 3.2 Configuration of paired gRNA quantification pipeline.

Parameter	Default	Description
Input		Input file path in FASTQ format
Reference		Design library file path in FASTA format
Output suffix		Suffix for output design screen library file path name
Extension length	20bp	Extend the protospacer into scaffold region by 20bp
Mismatch in scaffold	3	Maximum # of mismatch base(s) in scaffold regions
Trimming length	20bp	Maximum # of gRNA pairs return
Mapping quality cutoff	15	Minimum mapping quality score
Mapping mismatch	0	Maximum # of mismatch base(s) in protospacer

sequencing can be specifically aligned to a reference sequence, where the reference sequence is composed of protospacer and partial scaffold sequences on both sides.

3.3.2 Quality Control

By using FastQC [20] quality control tool, the characteristics of raw data like sequencing quality score, GC content, sequence length distribution, sequence repeat level, k-mer overexpression and whether the primers and adaptors in the sequencing data are contaminated, etc. will be further checked. Based on the quality control report, first we could filter out part of the poor quality data. Then, the information of paired gRNA sequences can be extracted by the pattern of U6 and H1 promoter sequences at both sides by tools such as Cutadapt [21]. Cutadapt will remove the i5 and i7 primers in the sequencing fragments first and also reads containing U6 and H1 promoters will be filtered out by setting the criteria of discarding untrimmed reads.

For the dual CRISPR system to delete the target regions, there is an additional clone step to add a full length scaffold. To ensure the extended scaffold ligation in place, each pair-end read will be trimmed to 40bp which should theoretically include 5bp overhang and 15bp cloned scaffold sequence as well as 20bp protospacer. The scaffold region might suffer from a failure of proper cloning into a full-length guide RNA and bad quality of sequencing, which both lead to these reads useless. To prevent these confounding factors being considered into the exact pgRNA quantification and affecting the downstream screen analysis, invalid reads will be filtered out too. By aligning the partial scaffold regions to the reference, only reads with an acceptable number of base mismatch can be retained for the final quantification step.

3.3.3 Alignment

After quality control and read trimming, the sequencing data containing the proper protospacer sequence of pgRNA are kept. The aligner like BWA-MEM or Bowtie2 will be employed to align the paired protospacer sequences to the reference screen library. Bowtie2 is more efficient in alignment step while BWA-MEM can support customized algorithm settings and the accuracy is higher. The pipeline has integrated both aligner package and also has the flexibility to adjust the alignment configurations such as the maximum number of mismatched nucleotides and the minimum number of matched nucleotides. Reads mapped to multiple reference pgRNA designs will be discarded. Finally, the alignment results will be stored in BAM format and the poor mapping quality reads will be filtered out from the BAM file. SAMtools [22] is used to filter out reads with poor mapping quality or mapped multiple times to different regions. The frequency of reads falling into the unique pgRNA sequences will be counted and the BAM files will be converted into a count table which can be used for subsequent screen analysis. MultiQC [23] tool is employed to summarize the quality statistics information across the complete workflow.

3.4 CRISPR Screen Analysis Pipeline

3.4.1 Essential Regions Identification

After obtaining the count table of pgRNA abundance in all samples, the next step is to identify gRNAs that are significantly enriched or depleted using rigorous statistical tests, in different biological settings.The pipeline integrates two such methods, MAGeCK [6] and PBNPA [10], to achieve sensitive and consensus identification performance. MAGeCK uses negative binomial distribution to model the abundance of each pgRNA and estimate the variance of read count under different conditions:

$$R_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2) \tag{3.1}$$

Here R_{ij} denotes the read counts of jth pgRNA design in sample i, μ_{ij} and σ_{ij}^2 are the estimated mean and variance of this pgRNA abundance respectively. Then the estimated variance of different perturbations are compared with that of the control experiment to determine whether this pgRNA varies significantly by statistical hypothesis testing. The essentiality of pgRNAs are ranked by P-Value in statistical hypothesis testing. Here the P-Value is the probability of finding the observed or more extreme results when the null hypothesis (H_0) of a study question is true. The null hypothesis is usually an hypothesis of

"no difference". Most researchers refer to statistically significant as $P < 0.05$. In the last step, MAGeCK uses a robust ranking aggregation algorithm called α-RRA to decide whether the non-coding region is selected to be positive or negative under the corresponding filtering conditions.

Another non-parametric method, PBNPA, directly estimates P-value at the gene level by permuting sgRNA labels, and thus it avoids restrictive distributional assumptions. To get more reliable identification results, the pipeline algorithm uses the Fisher method to combine the P-values at the gene level calculated by these two methods. The calculation of comprehensive P-Value by Fisher method is as follows:

$$X_{2k}^2 \sim -2\sum_{i=1}^{k} \ln(p_i) \tag{3.2}$$

Here p_i denotes the P-Value at the gene level calculated by various screening tools and is combined into a chi-square statistic X^2 with a degree of freedom of $2k$ by the Fisher method. A comprehensive P-Value at gene level can be obtained by the chi-square test.

Finally, the comprehensive P-Value will be adjusted into a quantitative statistical measure called false-positive rate (FDR) through the Benjamini-Hochberg process. Because if we use the P-Value less than 5% as cut-off, it is possible to reject the true null hypothesis by mistake. But the Benjamini-Hochberg process can avoid this Type I error and reduce the error discovery rate. The calculation of Benjamini-Hochberg process is defined as follows:

1. Sort the P-Values in ascending order;
2. Assign the priority level i to the P-Values. For example, the smallest level is 1, and the second smallest level is 2.
3. Use the formula $(i/m) * Q$ to calculate the Benjamini-Hochberg critical value for each p value.

Here i is defined as the ranking of each p value, m is the total number of tests, and Q is the level of false discovery rate.

In many CRISPR screen experiments, it is necessary to associate the significance of gene-level changes under different perturbations to obtain deeper biological insights of these genes or non-coding regions, for example, during a drug resistance study. For this purpose, the pipeline also incorporates MAGeCK-MLE [13] algorithm to calculate a significance measure beta score by maximum likelihood. The beta score indicates not only the directionality of the selection through the "+/-" symbol, but also the magnitude of the value that reflects the significance of the gene level change. Screen experiment design matrix will be generated

automatically to implement the MAGeCK-MLE algorithm and the significance score will be output in a standard format, which is convenient to link the changes in the gene/region level under different perturbations for subsequent downstream analysis.

3.4.2 Bias Correction

Since dual CRISPR screen system works based on the strategy of loss-of-function by deleting regions of interest. However, the CRISPR screen readout involved many confounding factors such as targeting regions with copy number variation and variance in CRISPR cutting efficiency. Therefore, the pipeline also incorporates bias correction methods to eliminate these biases in dual CRISPR screen readout.

Correction of deviations targeting CNV amplified regions

Copy number variation (CNV) is a structural mutation in which the copy number of a particular gene is changed from normal levels, and is commonly found in cancer cells to cause cell dysfunction. When a Cas9/gRNA targets at a single copy site in the copy number variation region, it will in fact make multiple cuts at the same time. This will trigger a strong DNA damage response that may cause cell cycle arrest and reduce cell proliferation. Therefore, even if the targeted non-coding region has nothing to do with the phenotype alteration, targeting CNV regions will lead to pgRNA depletion during the screen and thus constitute as a false positives hit in the negative screen selection. The CERES [24] tool jointly models the deviation caused by copy number variation and the potential effect of knocking out the targeted region, and uses the constrained least squares optimization model to decouple the two effects. This method is integrated in the pipeline to correct the effect of copy number variation within different cell lines during the dual CRISPR screening.

$$D_{ij} = q_i\Big(\sum_{k_i \in G_i} (h_k + g_{kj}) + f_j\Big(\sum_{l \in L_i} C_{lj}\Big)\Big) + o_i + \varepsilon \tag{3.3}$$

Here ε is independent Gaussian noise term, o_i explains the noise in the i-th sgRNA abundance measurement from pooled screen results. The actual gene knockout effect is the sum of cell line-specific effect g_{kj} and shared effect h_k, which is the aggregation effect of all i-th sgRNA (G_i) targeting this gene. The copy number effect is modeled by spline linear regression coefficient f_i and determined by the target locus L_i and the copy number C_{lj} of each locus. Finally, the cumulative knockout effect is scaled by the fraction score q_i limited between 0 and 1 to mitigate the effects of low-quality reagents.

CERES needs the known copy number variation profile in the corresponding cell line as an input. This can be retrieved from the Cancer Cell Line Encyclopedia (CCLE) [25] that contains copy number variation profiles of most genes in commonly used cell lines. The ENCODE database also gives a genome-wide copy number variation profile of K562 cell line including three status (normal copy number, amplification and heterozygous deletion).

Correction of the deviation of different sgRNA target efficiency

Different gRNA may show varied cleavage efficiency Thus sgRNA with low cutting efficiency may confound the interpretation of the screen results. To correct this bias, MAGeCK-MLE provides an option to include gRNA cleavage efficiency as an optional input. The cleavage efficiency of a gRNA can be predicted by the SSC [5] algorithm based on the gRNA sequence features. Through the regression modeling, the real knockout effect and bias induced by different sgRNA with varied cutting efficiency will be jointly modeled and decoupled to obtain the real knockout effect.

3.4.3 Preprocessing of screen results

Essential gene filtering

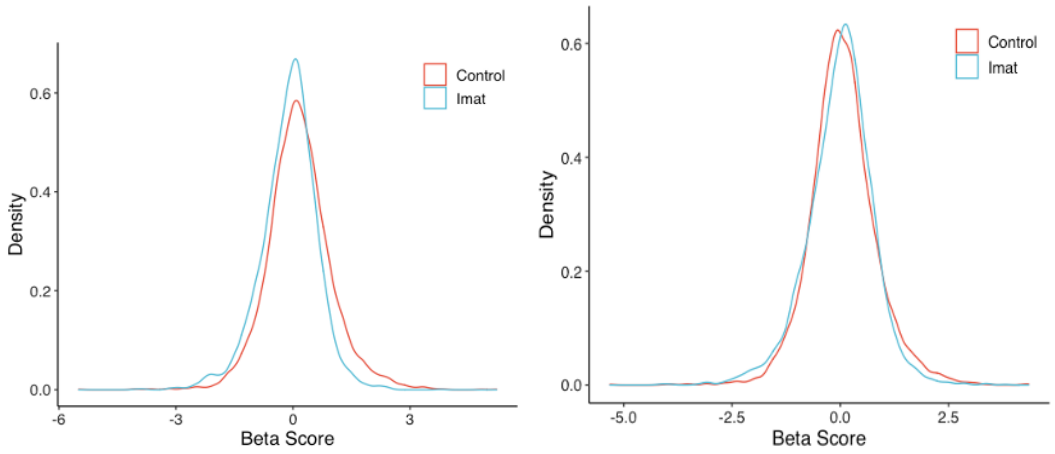
In order to discover new therapeutic targets or detect new candidate gene functions, it is necessary to filter out genes that are known to be vital for cell survival when preprocessing the screen result. Depmap is a pre-clinical reference map that provides a way to define and predict genes that are critical for cell viability. After omitting common essential genes from the data, the similarity between the CRISPR screen with Depmap screens will be computed to ensure no essential gene left.

Data normalization

In CRISPR screen experiments, it's difficult to ensure all samples are in the same cell cycle and the significance scores of different samples are not directly comparable. Therefore, the pipeline developed a cell cycle normalization method to shorten the cell cycle gap under different conditions before conducting downstream analysis.

Experimental batch effect removal

In biology experiments, a batch effect occurs when non-biological factors in an experiment cause changes in the data produced by the experiment. Such effects can lead to inaccurate



(a) Before normalization (b) After normalization

Fig. 3.4 Beta score distribution before and after cell cycle normalization.

conclusions when they affect the readouts.Before further analyzing the screen data, the pipeline tests whether the batch effect exists in the experiments and then remove it.

As Figure 3.5 reveals, the green dots are interfered by the same noise signal and thus these data show a high similarity but heterogeneous indeed. The batch effects can be removed by a regression test.

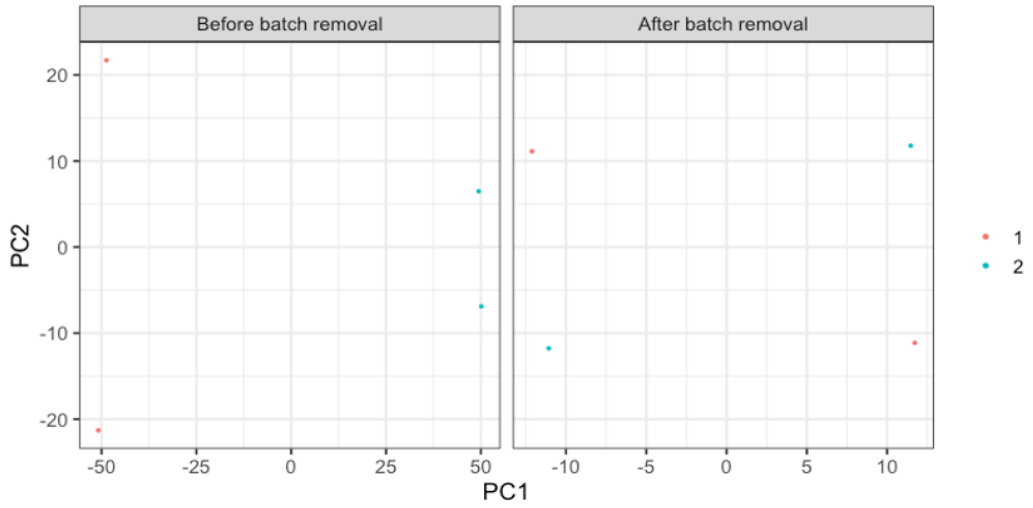


Fig. 3.5 Batch effect removal on simulated data.

3.4.4 Visualization of screening results

Using the ggplot2 package, the screen analysis pipeline provides a number of visualization options for displaying positively or negatively selected genes, including volcano plot, rank plot and scatter plot. For the screen results of MAGeCK-MLE, the beta scores under different perturbations are usually used to get biology insights, for example, in a drug resistance screen study to identify important genes. A nine-square scatter plot is often used to visualize screen results between different treatments and control conditions to identify genes or non-coding elements relevant to a particular treatment.

3.4.5 Clustering analysis of the screening results

Regulatory elements such as enhancers which are in close proximity on the genome sometimes co-operate or play redundant roles in the regulatory network to regulate the same transcription activity. These enhancers form enhancer clusters [26] or super-enhancers [27] which are both associated to cell identity and are bound by transcription regulators with higher intensity than individual *cis*-regulatory element [28]. As it is expected that some enhancers may act redundantly, it is possible that removing one enhancer from a enhancer cluster will only have a marginal effect on the phenotype of interest. It is therefore possible that integrating the clustering information of the target regions into the dual CRISPR knockout screen would lead to the identification of cluster regions.

An unsupervised clustering algorithm CREAM [29] was integrated in the pipeline to identify clusters of *cis*-regulatory elements (COREs), such as enhancers, promoters. CREAM uses genome-wide maps of genomic regions such as DNaseI, ATAC or ChIP-Seq data as input and considers proximity of the elements within chromosomes of a given sample to identify COREs in the following steps:

- It identifies window size or the maximum allowed distance between the elements within each CORE,
- It identifies number of elements which should be clustered as a CORE,
- It calls COREs,
- It filters the COREs with lowest order which does not pass the threshold considered in the pipeline.

Then, BEDTools *intersect* method was applied to detect the overlaps between the identified CORES in specific cell type and the hits from dual CRISPR knockout screen. Candidate clusters of hits or CORE regions were output for function validation in the next step.

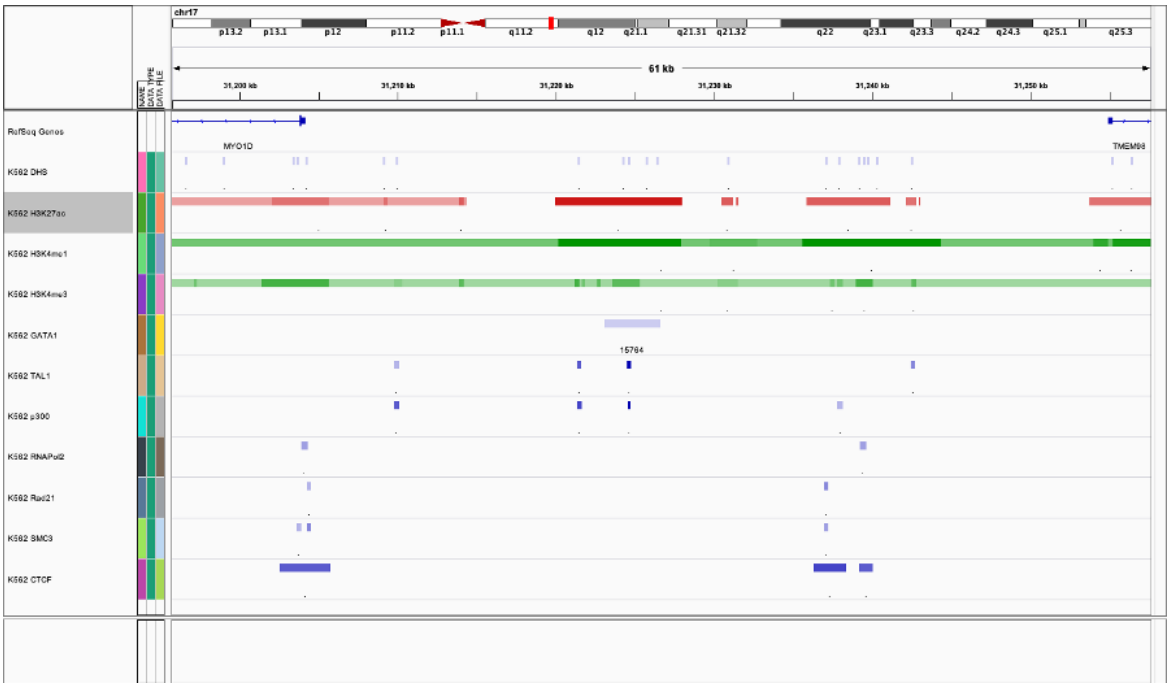


Fig. 3.6 Identify super enhancer from screen results by incorporating ChIP-seq data.

Taking the super enhancer in Figure 3.6 as an example, when the distance between enhancers in a certain genomic region is about 10-12.5 kb, these regions have the same selectivity or show similar functions from the screen results. For these enhancer clustering regions, ChIP-seq signals [30] like transcription factor Med1 or histone modification signal such as H3K27ac could be combined to further infer the presence of super enhancers.

3.5 Dynamic graphical interface for Snakemake workflows

To expose our pipelines to a wider audience, a graphical user interface (GUI) was designed by Sequanix [31] to offer the ability to edit the configuration file interactively. Sequanix interface is written in PyQt, which is a Python binding of the cross-platform GUI toolkit Qt.

The user interface is consisted of a main dialog, a Snakemake dialog and a Preferences dialog. To run the pipeline in the main dialog as Figure 3.7 shows, first, users need to select the generic pipeline of interest by following steps: 1) select the *Generic pipelines* tab, 2) select the *Snakefile* tab, 3) click on the *Browse* button to select the pipeline file.

Because Snakemake pipelines are made of two parts: a pipeline and an optional configuration file. Usually, the pipeline is called Snakefile which contains the code of the pipeline itself. In the Snakefile, the pipeline is linked to an external configuration file which is encoded in YAML or JSON format. So the second step is to click the *Config file* tab, navigate to the

configuration file path and select it. Once done, the configuration of the pipeline will be loaded in the Config parameters tab. The directory of input data specified in the configuration file will be interpreted and a widget is available to select the directory where to find the data.

Third, users need to select the working directory where the pipeline and its configuration file will be stored as well as the results of the execution. To do so, just click on the working directory tab and select or create a directory.

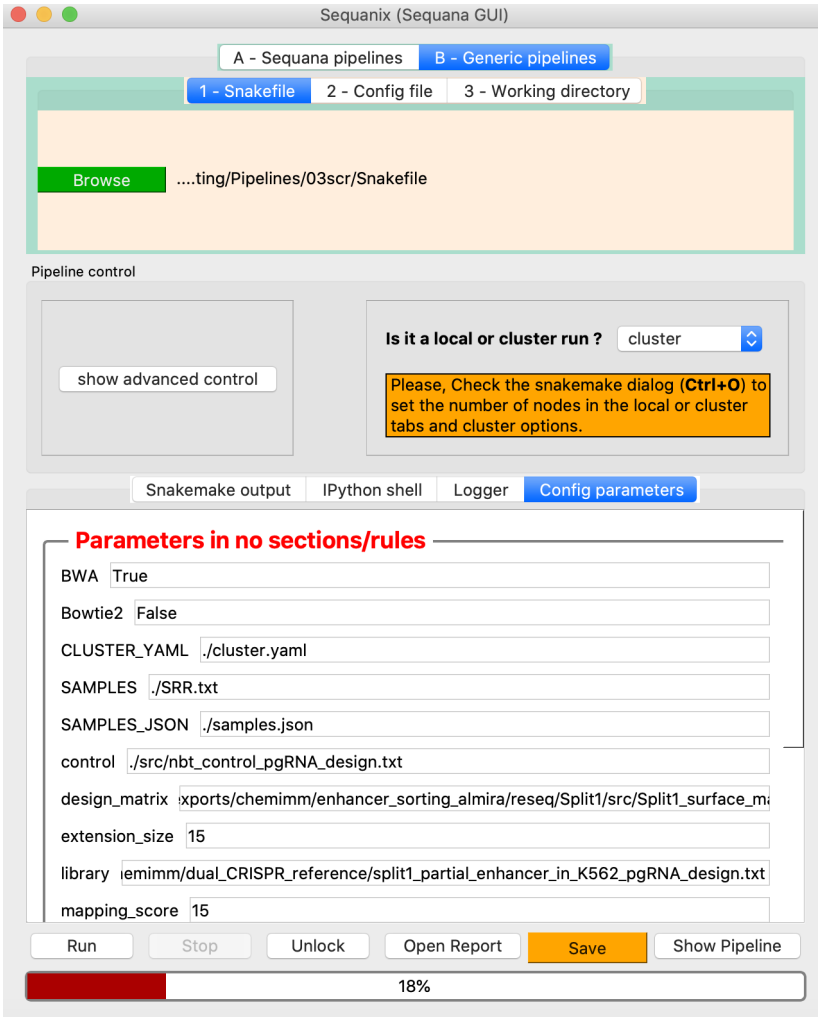


Fig. 3.7 Snapshot of the Sequanix graphical user interface (GUI) for pgRNA quantification pipeline.

Furthermore, to run the pipelines on clusters using various scheduler frameworks without changing the pipeline code, users just need to switch the *local* mode to *cluster* mode in the main window as shown in the Figure 3.7. Then, users could tune the configurations related to execution of the Snakemake pipeline on clusters such as specific job scheduler information or number of CPUs to be used.

Finally, users could perform following steps to debug or run the pipeline:

- Save the project: click the *Save* button in the bottom bar. This saves the configuration and pipeline files in the working directory defined above.
- Check the project: click the *Show pipeline* button. This shows an image of the pipeline.
- Run the project: click the *Run* button that should now be clickable. The output of the Snakemake execution is shown and the progress bar will move showing the stage of the analysis.
- Stop the analysis: click the *Stop* button if you made a mistake in the configuration or simply want to stop the current analysis.

Wait until the Run button is released. The Stop button should now be disabled. At the end of the analysis, if everything executes correctly, the progress bar's color switches to green, otherwise to red. The results are summarised into a file named "summary.txt" in the working directory.

Chapter 4

Experiments

The design principles for dual CRISPR screen library design, paired gRNA quantification and CRISPR screen analysis pipelines have been stated in *Methods* chapter. To test the computational pipeline design for dual CRISPR screening, results from several CRISPR screen experiments generated in the lab were used to evaluate the performance of these three modules. The details of experiment setup including tested software version, program running configuration and benchmark data set were described in this chapter.

4.1 Experiment Setup

4.1.1 Designing paired gRNA targeting diverse non-coding elements

To test the performance of dual CRISPR screen library design pipeline, pairs of guide RNAs targeting a variety of non-coding genomic elements were designed. These non-coding genomic elements don't rely on transcription to work and the indel induced by canonical single CRISPR knockout strategy can't ensure to abolish functions of non-coding regions completely. The precise deletion produced by dual CRISPR system could accomplish the loss-of-function studies for the non-coding genomic elements more efficiently. For example, the effect of a enhancer to regulate the expression of its target gene will be eliminated because there is no binding site for its associated activator after dual CRISPR deletion.

Overview of the target regions

The non-coding genomic elements used for testing were sampled from 1,747 validated enhancers from the VISTA enhancer database [32], 4,351 ultraconserved elements from the UCNEbase database [33] (2,139 are located in the intergenic region, 1,713 in intron and 499

in UTR regions). The random intergenic regions were also tested as negative controls, which are composed of the regions except for coding and non-coding genes (including introns and exons) removed from hg38, 10kb upstream/downstream of the gene region.

Pipeline configuration and parameter setting

The positive mask region selects the DNaseI hypersensitivity sites where Cas9 has a higher cutting efficiency. The negative mask region selects the repeat regions that have relatively high off-target rates. Regarding the data source of the positive/negative mask region, DNaseI hypersensitive sites and repetitive regions are derived from the experimental data of ENCODE project and downloaded from UCSC browser.

For the parameter setting, as Table 3.1 shows, the design region of the gRNA is defined as 200bp upstream/downstream of the target region and the PAM motif is set as "NGG". The off-target parameter is set as default 1,0,0,x,x, meaning that the designed sequences can't align with other genomic sites that have more than 2 mismatches. The threshold value for one-sided predicted targeting efficiency is set to 0.2 and the predicted target efficiency threshold for each pair of designed pgRNAs is 0.4. The maximum number of output pgRNAs for each target region is 10. To control the diversity of gRNA design list, one unique gRNA design on either side is not allowed to appear more than 5 times in the final design. The design list is directly sorted according to the target efficiency and gRNAs with a lower score are excluded.

Off-target analysis

To control the performance of off-target prediction, the parameter of mismatch cutoff was adjusted and the pipeline was performed on targeting the 86 validated enhancers on chromosome 22. By relaxing the potential off-target cutoff, it will probably improve the fraction of successfully targeted regions in the designed sgRNA pairs library. The cutoff parameter was adjusted from 1,0,0,x,x as default to 1,1,5,x,x. To measure the performance, percentages of full depth and predicted on-target efficiency were used. Here full depth refers to the percent of targets receiving n=10 sgRNA pair designs. On-target efficiency is predicted by sgRNA designer, which is an experimentally trained logistic regression model employing 72 sequence features.

4.1.2 Deconvolution of pooled dual CRISPR screen results

The paired gRNA quantification pipeline was tested using dual CRISPR screen data generated within the group. The primary outputs of the next-generation sequencing of a CRISPR-Cas

pooled screen are FASTQ files, which contain both the sequence and quality information for each read. In deconvolution step, the guide sequence is extracted from the read and aligned to the guide library, resulting in a read count table of the total number of reads per guide in each sample. This also applies to screen analysis of the single gRNAs targeting genes.

Overview of the dual CRISPR screen data

We performed dual CRISPR screen on K562 cell line in which upon removing the non-coding regions would affect the expression of four different antigens (MHC class I, PD-L1, CD61 and CD14) on the cell membrane individually. To determine the essential non-coding regions affecting the expression of these four surface markers respectively, the paired gRNA quantification pipeline was used to demultiplex FASTQ files from different samples. The amount of sequencing reads generated from MHC class I+, PD-L1+, CD61+ and CD14+ cells ranges from to 300,000 to 22 million. Three biology replicates were performed to gain statically power. This sums up to 30 screen samples which are need to be deconvoluted in total.

Parameter setting

As the Table 3.2 depicts, the reads will be trimmed and left with maximum 45 bp in the adaptor removal step. The sequence will contain 5bp overhang from U6/H1 promoter, 20bp guide sequence and 20bp extension into scaffold region. The maximum mismatches is set as 3 bases and the clean reads for alignment will only keep 20bp guide part on each pair-end. The minimum mapping quality score is defined as 15 and no mismatch base is tolerant in final alignment results. The pipeline was running on the workstation CentOS8, 86Gb of memory and 12 CPUs (Intel E5-2697 CPU), with maximum 6 threads for each alignment job.

4.1.3 Detection performance comparisons for genome-wide genetic screen

The CRISPR screen analysis pipeline was designed by integrating parametric and non-parametric hit identification algorithms using the Fisher's method, which is supposed to be able to explore outside the consensus field and achieve sensitive detection of essential hits. To show that, the selected screen analysis algorithm was compared with other mainstream parametric and non-parametric algorithms on the genome-wide CRISPR screen benchmark sets.

Benchmark datasets

To test the performance of the screening analysis module, two published whole genome screening data sets are used and reanalyzed here: CRISPRn-A375 and CRISPRi-A375 [34]. These two data sets are from the genome-wide CRISPR knockout and CRISPR interference experiments. Both screening experiments were performed on the A375 melanoma cell line. These experiments used optimized the design of gRNA, which is more efficient than the original genome-wide screening library GeCKO [35].

Each data set contains a control experiment sample that is the empty plasmid DNA and three biological replication experiments. To test the performance of the algorithm, the gold standard gene set was selected which includes 1580 key genes and 927 non-key genes reported by Hart et al. [36, 37]

Evaluation metrics

To evaluate the reliability of the proposed comprehensive screen analysis algorithm compared with other canonical methods, the main performance metrics used include specificity, sensitivity, precision and recall. In addition to the precision and recall rate, a comprehensive index F1 score is also introduced to measure the performance of each method on two benchmark sets. The definitions of these measures are described in formulas 4.1, 4.2, 4.3 and 4.4:

$$\text{sensitivity} = \text{recall} = \frac{\text{Detected essential regions}}{\text{Total essential regions}} \tag{4.1}$$

$$\text{precision} = \frac{\text{Detected essential regions}}{\text{Total detected regions}} \tag{4.2}$$

$$\text{specificity} = \frac{\text{Detected nonessential regions}}{\text{Total nonessential regions}} \tag{4.3}$$

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4.4}$$

PR curve (precision-recall curve) and ROC curve (receiver operating characteristic curve) were used to illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The PR-AUC value and ROC-AUC value, the area of under the PR curve or ROC curve respectively, were calculated to show how good it could measure the separability.

4.2 Workstation and software dependencies

The performance test of all three pipelines was performed on the workstation running CentOS8, 86Gb of memory and 12 CPUs (Intel E5-2697 CPU). The workstation uses Slurm 20.02 as workload manager system to schedule the user-submit job execution queue.

The three modules: dual CRISPR screen library design, paired gRNA quantification and CRISPR screen analysis pipelines are managed in a stream through Snakemake language. The main scripting languages used in Snakemake file are Shell, Python and R.

The software and versions required for testing are shown in Table 4.1:

Table 4.1 Softwares and their versions during pipeline development.

Pipeline module	Software version	Author
Dual CRISPR Screen Library Design	BEDtools v2.28.0	Quinlan et al. 2010
	CRISPRseek v3.11	Zhu et al. 2014
Paired gRNA Quantification	FastQC v0.11.8	Andrews et al. 2010
	Cutadapt 1.18	Martin et al. 2011
	Bowtie2 v2.4.0	Langmead et al. 2012
	BWA 0.7.17	Li et al. 2013
	SAMtools 1.4.1	Li et al. 2009
	MultiQC v1.8	Ewels et al. 2016
CRISPR Screen Analysis	MAGeCK 0.5.8	Li et al. 2014
	PBNPA 0.0.3	Jia et al. 2017
	VISPR 0.4.14	Li et al. 2015
	MAGeCKFlute 1.8.0	Wang et al. 2019

Chapter 5

Results

This chapter mainly describes the observations from the experiments which were designed to evaluate the three pipelines' performance. Two case studies on anthracycline single gRNA screens were also tested using the pipelines to illustrate the research application value.

5.1 Pre-generation of dual screen library targeting diverse non-coding genomic elements

By completing the test run of guide design pipeline on the sampled input regions, results of pre-generated guide RNA pairs targeting the diverse non-coding regions are shown in Table 5.1:

Table 5.1 Summary of demo run results by targeting non-coding regions.

Data Set	# Input targets / total	# Designed targets	# Designed sgRNA pairs
Enhancer	100 / 1747	87	757
Ultraconserved elements	200 / 4351	185	1619
Random intergenic	100 / 3170	65	503

When running the CRISPRseek tool to search for all candidate target protospacer adjacent to PAM, most of the pgRNA designs are filtered out due to the overlap between the design region and the repeat region, or low targeting efficiency predicted by logistic regression model. The depth of designed library, which means the percent of targets which return required number of unique sgRNA pairs (10 designs) regarding the total number of input regions, reached 70% on average. This case can be improved regarding the partial depth of designed library, which doesn't require enough unique sgRNA pairs for targets (<10 designs).

The number of targets with valid designed sgRNA pairs were returned (≥ 1) as Table 5.1 shows. The partial depth of library targeting enhancers and ultraconserved elements achieves 87% and 92.5%. The low depth of guide design targeting random intergenic regions is caused by overlap with repeats and low targeting efficiency.

5.2 Off-target analysis of sgRNA pairs targeting enhancers

To verify whether the off-target effects existing in the pre-generated screen library, 86 enhancers on chromosome 22 were selected as targeting regions to design sgRNA pairs. The experiments were repeated several times, where in each case the off-target cutoff parameter was changed and compared. The experimental results are shown in Table 5.2:

Table 5.2 Off-target analysis of pgRNA targeting 86 validated enhancers on chromosome 22.

Parameter	Mean on-target score	# Designed targets	# Designed pgRNAs
1,0,0,x,x	0.52	74	644
1,0,1,x,x	0.57	78	730
1,0,2,x,x	0.6	80	751
1,0,3,x,x	0.6	80	760
1,0,4,x,x	0.61	81	766
1,1,5,x,x	0.6	82	779

With the number of allowed off-target mismatches increasing, both the full and partial coverage depth of screen library design increased. The setting "1,1,5,x,x" allows similar sites with maximum 1 base and 2 bases mismatches in the genome actually designs guide sequences targeting 82 of 86 input regions. In the meanwhile, it also show the tendency of off-target existing in the designed screen library.

Moreover, it can also be observed the predicted sgRNA targeting efficiency mildly improves after relaxing the restriction of the perfect match. As recent research [38, 39, 3] pointed out that the off-target effects might affect library efficiency greatly, the pipeline takes the cutoff parameter 1,0,0,x,x as default to ensure the reliable CRISPR screen result.

5.3 End-to-end solution for CRISPR screen deconvolution

During the deconvolution of the dual CRISPR screen data, the FastQC reports were generated to check the sequencing-level metrics like GC content. Adapters were removed from the FASTQ data and then guide sequence information was extracted from the raw reads.

After checking the intermediate trimming results, the experimental errors like primer contamination, absence of guide sequence and poor sequencing quality of scaffold were found in the raw data. It brings problem to quantify the accurate abundance of pgRNA in each sample. The pipeline applies strict quality control and filtered out the invalid reads from the FASTQ files. Both guide sequences (20bp) from pair-end read were extracted and aligned to the guide library to obtain the raw count table. The quality statistics of each sample were checked after alignment and results of the percentage of mapped reads were shown in Figure 5.1:

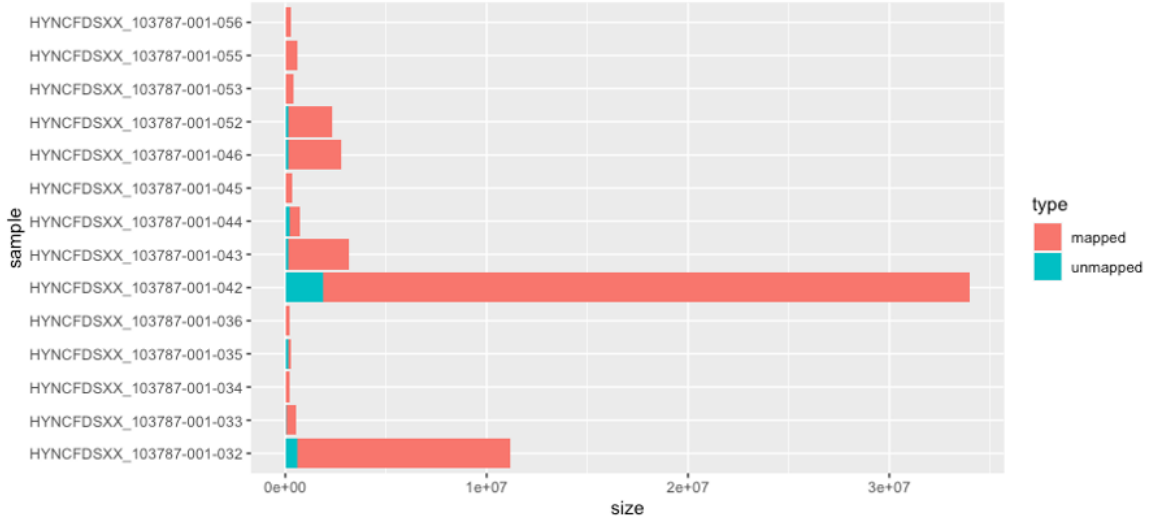


Fig. 5.1 Reads mapping quality after quality control.

The mapping rate of each sample (including unsorted and sorted) reaches around 90% while the contaminated one is only 40%, which indicates the alignment result is quite convincing. Because different samples vary greatly in size, contamination induced by experimental errors was strictly removed and relatively fair percentage of valid reads in each sample was achieved. This step makes sure that the reliable results will be used in the downstream analysis.

5.4 Sensitive detection of essential hits from genome-wide screen

To evaluate the performance of pipeline screen analysis algorithm, four CRISPR screen analysis algorithms (parametric methods: MAGeCK and CB2, non-parametric methods:

PBNPA and RIGER) were tested on the two benchmark sets CRISPRn-A375 and CRISPRi-A375. The performance of the four tools is shown in Figure 5.2:

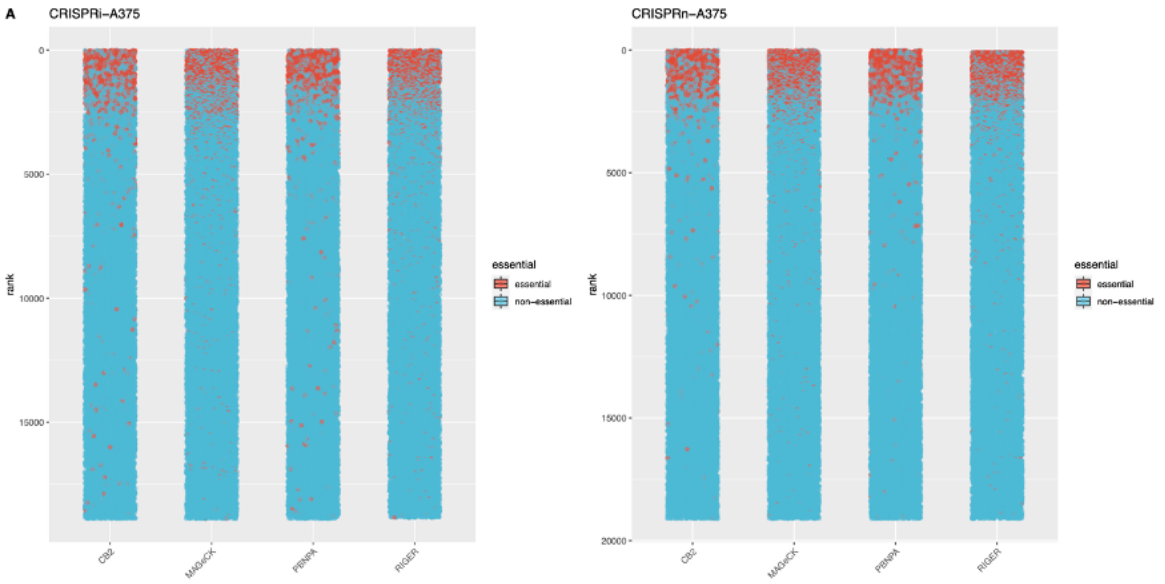


Fig. 5.2 Comparison of gene priority rankings obtained by the four screening methods on the CRISPRn-A375 and CRISPRi-A375 benchmark sets, where red dots indicate essential genes and blue dots indicate non-essential genes.

All the four methods can recognize most of the significant genes from the benchmark sets and prioritize them in the top rank. And a small percent of the essential hits diffuse in the middle and bottom. According to the gene rankings output by four screen analysis algorithms, the correlations between each two methods can be calculated, as shown in Figure 5.3:

According to the correlation matrix, CB2 and RIGER assign the similar essential genes with other screen analysis algorithms on both CRISPRn-A375 and CRISPRi-A375. In contrast, MAGeCK and PBNPA rank essential genes less consistent with the other two methods. This also reveals MAGeCK and PBNPA have more potential to explore essential genes outside the consensus field.

The correlation analysis also shows both parametric and non-parametric methods show similar performance within the group. Even though MAGeCK and CB2 use different distribution assumptions, as MAGeCK assumes the screen readout follows the negative binomial distribution while CB2 uses beta binomial distribution. Also these two methods use different aggregation algorithm to perform gene-level analysis based on multiple targeting sgRNA significance. MAGeCK uses the permutation-based RRA algorithm while CB2 uses the Fisher's method to integrate the significance score P-value. The same is true for PBNPA and RIGER, which are also non-parametric methods. PBNPA uses a permutation-based

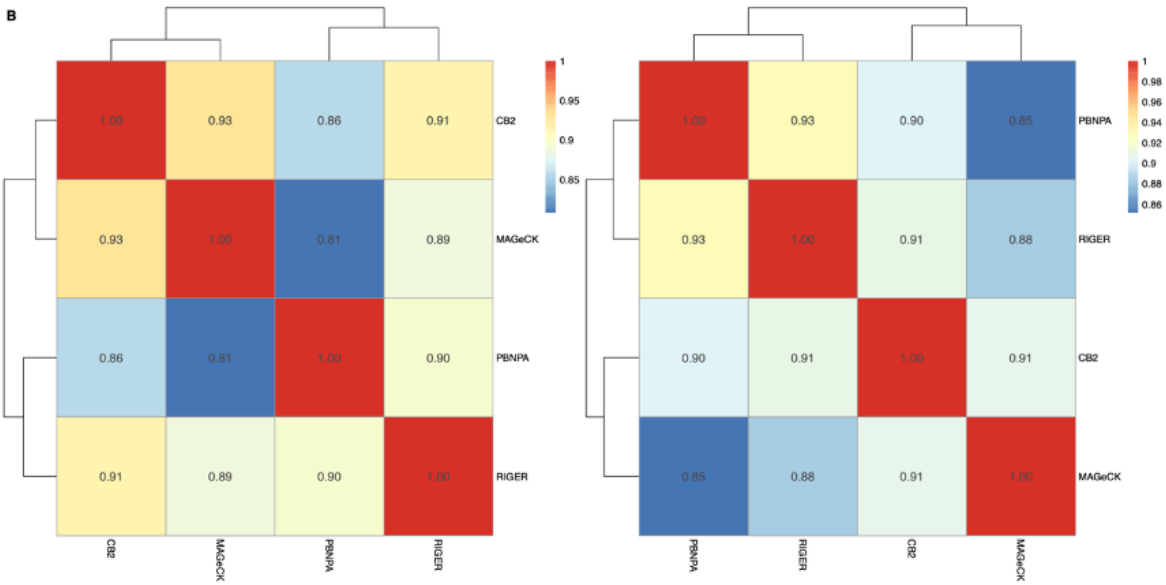
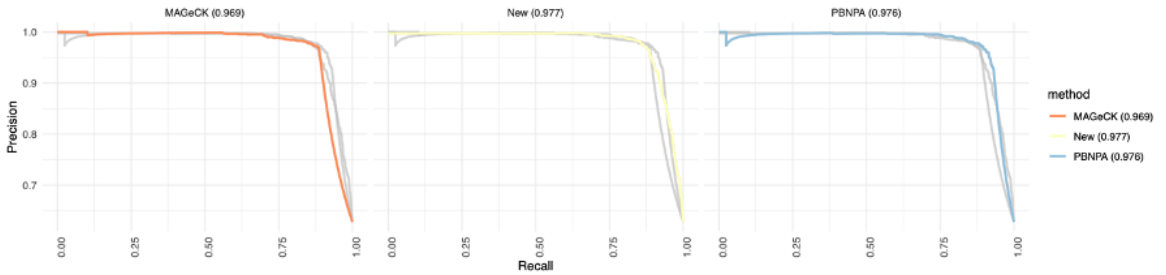
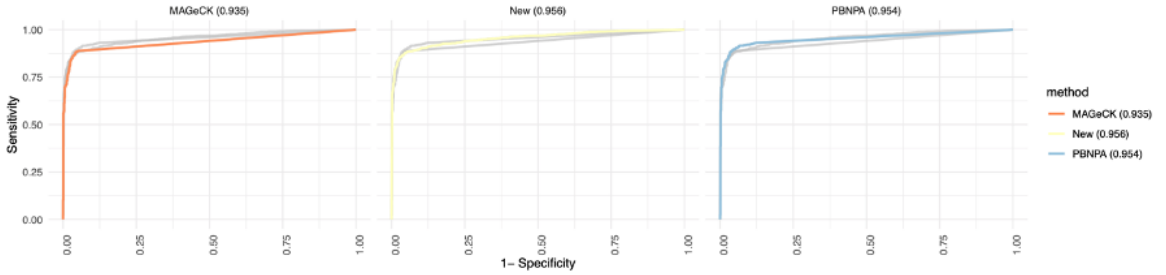


Fig. 5.3 Pairwise correlations between gene priority rankings obtained by the four screening methods on the CRISPRn-A375 and CRISPRi-A375 benchmark sets.

non-parametric test method while RIGER uses Kolmogorov–Smirnov non-parametric test to obtain gene-level significance score.

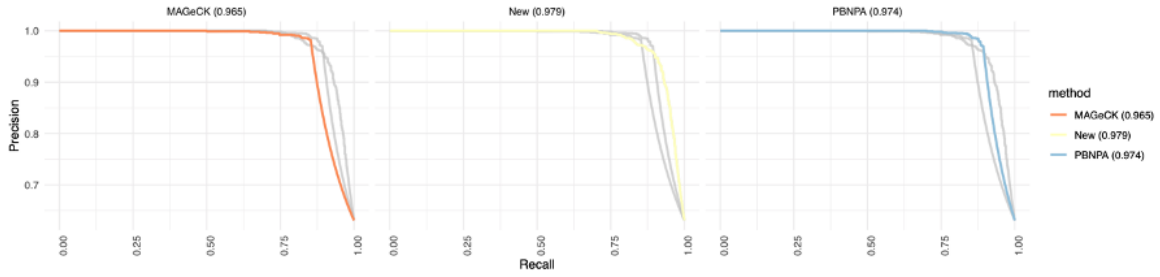


(a) PR curve.

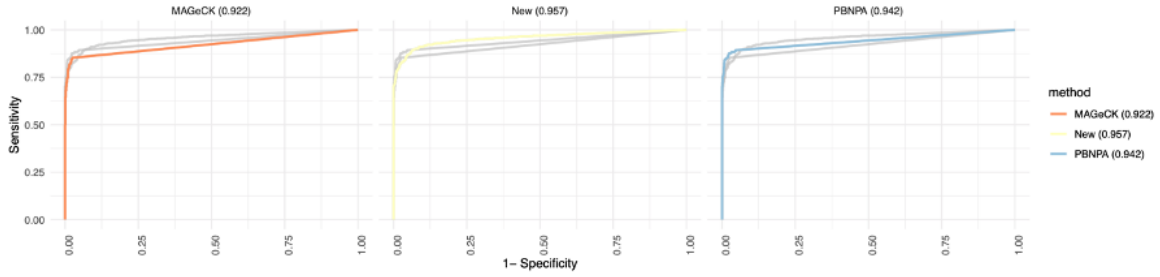


(b) ROC curve.

Fig. 5.4 Performance comparison of three screening methods on the benchmark set CRISPRn-A375.



(a) PR curve.



(b) ROC curve.

Fig. 5.5 Performance comparison of three screening methods on the benchmark set CRISPRi-A375.

Our pipeline can identify more reliable hits by Fisher method incorporating parametric and non-parametric methods as describe in the *Method* chapter. To compare our pipeline screen analysis algorithm with other algorithms, the PR-AUC value and ROC-AUC value can be obtained by calculating the area of under the PR curve, shown in Figure 5.4a and 5.5a, and ROC curve, shown in Figure 5.4b and 5.5b. Our pipeline achieves good separation performances on CRISPRn-A375 (PR-AUC = 0.977, ROC-AUC = 0.956) and CRISPRi-A375 (PR-AUC = 0.979, ROC-AUC = 0.957), which are better than the other two methods.

The original output of MAGeCK and PBNPA was converted into FDR, by adjusting the FDR threshold range from 0.1 to 0.001, to compare with our pipeline algorithm's output. The FDRs we obtained at each cutoff point are better than the other methods, indicating that our method can detect even more essential genes under more strict conditions.

Existing CRISPR screen analysis algorithms are prone to suffer from the Type I error (False positive) due to many factors as mentioned in *Method* chapter, for example, targeting the CNV regions will lead to false positives in the screen results. The comprehensive evaluation of hits by our pipeline, which combines the strengths of both parametric modeling and non-parametric test, can give more robust results thereby effectively reducing the Type II error (False negative).

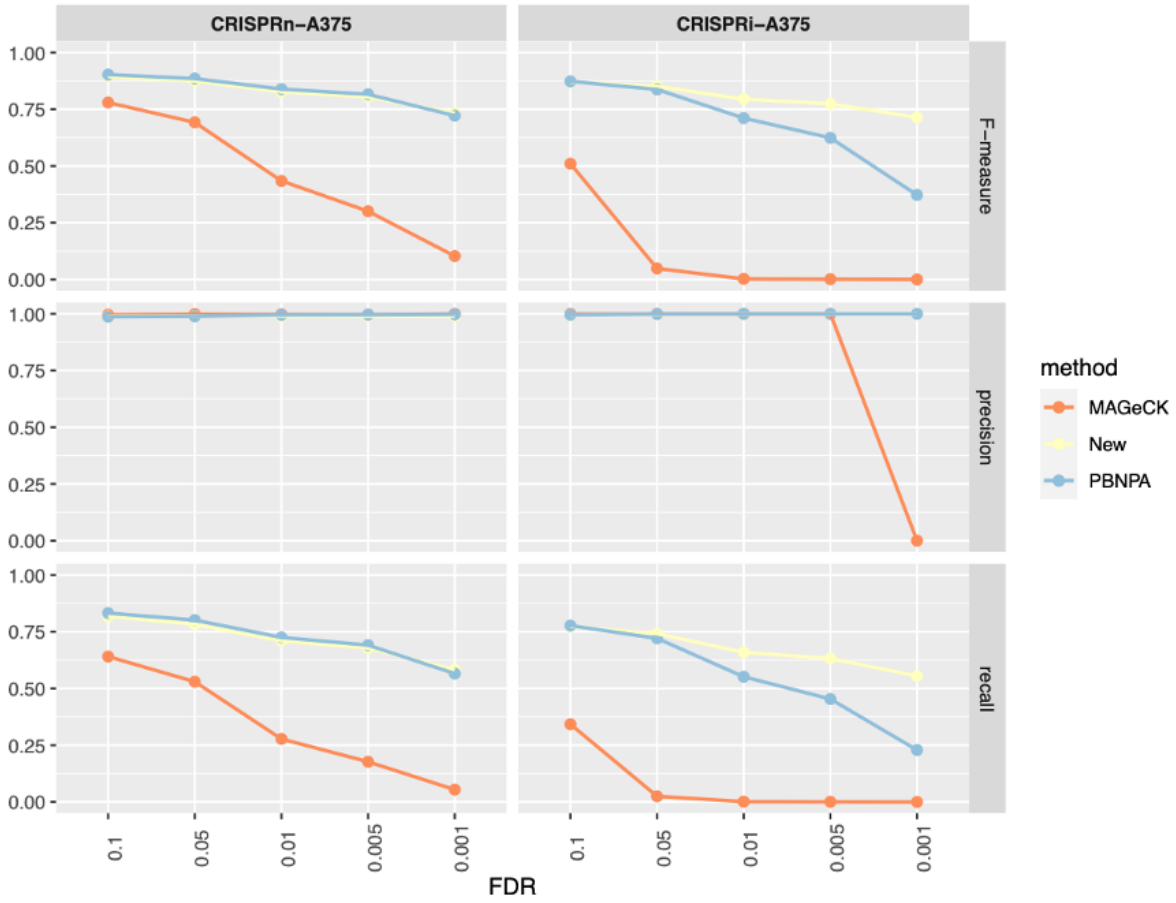


Fig. 5.6 Three screening methods to detect essential genes on the benchmark set CRISPRi-A375 and CRISPRn-A375, comparison of F1 score, precision and recall at each FDR cutoff.

5.5 Application Case Study: Anthracycline Chemotherapy

Anthracycline antibiotics are a class of chemotherapeutic drugs derived from the strain of Streptomyces bovis, including Doxorubicin and Aclarubicin. They are widely used in the treatment of many tumors, such as acute leukemia, breast cancer and ovarian cancer.They have the advantages of broad anti-tumor spectrum and strong anti-tumor effect. However, anthracycline drugs can also cause a series of side effects, such as cardiotoxicity, so the cumulative dose is limited in treatment.

To study the mechanism of action like drug resistance and side effects of anthracycline, systematic loss-of-function strategies like CRISPR activation and dual CRISPR deletion are applied. CRISPR activation screening method is used to study genes related to the transmembrane transport of anthracyclines. And dual CRISPR deletion screening method is used to identify non-coding regions that are essential to render cells resistant to anthracyclines.

The cell line used in the screening test was K562, which was the first established human immortalized myeloid leukemia cell line. The K562 cell line is also the cell line used by the ENCODE project to analyze all heredity, epigenetics and trans-transcription factor binding sites. Therefore, more genetica and epigenetic data are available from the K562 cell line for downstream analysis.

The CRISPR activation screening library contains the sgRNAs targeting 1,313 membrane transporter encoding genes from TransportDB 2.0 [40]. This sgRNA library was contributed by Feng Zhang's lab in MIT. The dual CRISPR deletion screening library is composed of dual CRISPRs targeting non-coding regions, which was designed by Xiao Li from Michael Snyder's lab at Stanford University. Two schemes were applied to design the paired gRNAs for dual CRISPR screen, with guide pairs targeting inside and outside both ends of the region of interest. Negative controls are added in all screen library designs.

Both Library 1 and Split 1 target partial enhancer region in K562 cells, Library 1 gRNA design region is 200bp upstream and downstream of the enhancer region while Split 1 gRNA design region is the 5'and 3'ends in the enhancer region. Library 2 targets ultraconserved elements and validated enhancer regions. In addition, both Library 1 and Library 2 introduced 15,821 pgRNAs to target 6,000 random intergenic regions of pgRNA. Finally, 1,070 pairs of pgRNA were introduced as positive and negative controls in each screening library. The design of the library used in all screening experiments is shown in Table 5.3:

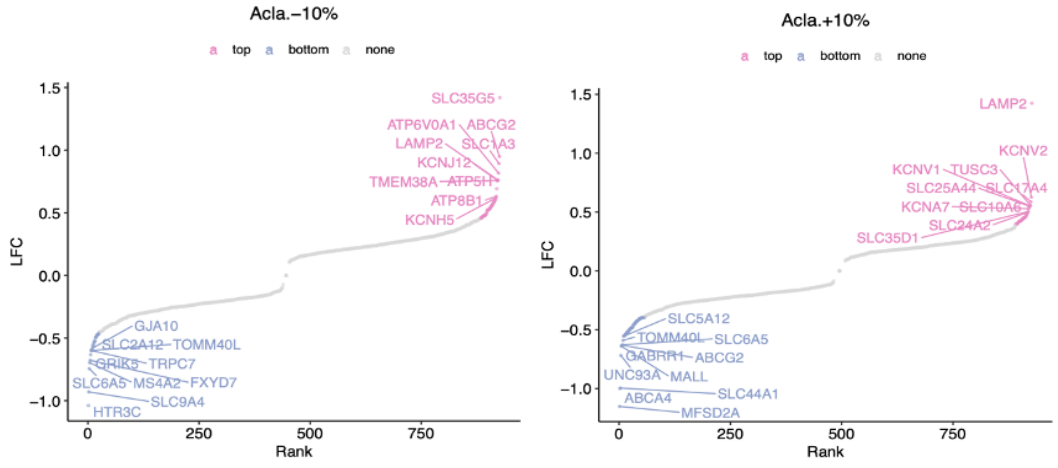
Table 5.3 Screen libraries and major target regions in case study.

Main target area	Guide type	Design scheme	# Designed guides
Partial enhancer regions in K562 cell	pair	outside	116,696
Validated enhancer, ultraconserved elements	pair	outside	51,678
Partial enhancer regions in K562 cell	pair	inside	87,921
Partial enhancer regions in K562 cell	pair	inside	87,347
Membrane transport protein	single	-	46,091

5.5.1 Case 1: Anthracycline Transmembrane Transport

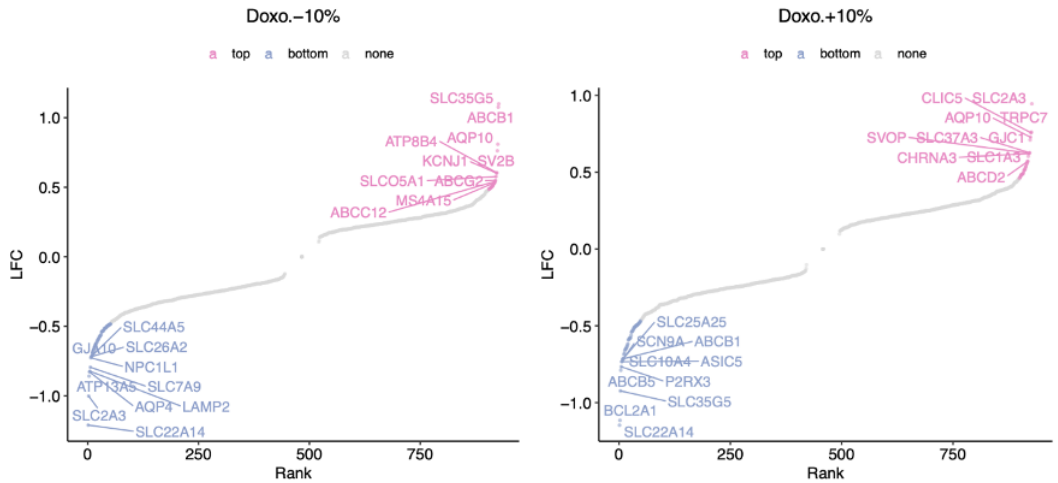
Membrane transport protein is a carrier that participates in molecular diffusion on the cell membrane, and can selectively allow non-freely diffusing small molecules to penetrate the plasma membrane. Membrane transporters can be either membrane integrins or large transmembrane molecular complexes. Although the membrane transport protein involved in promoting diffusion has no enzyme activity, it has the characteristics of enzyme catalysis, which can achieve the highest rate, specificity and competitive inhibition.

In order to study which membrane transporters can specifically select anthracycline molecules to enter the cell, CRISPR activation (CRISPRa) experiments were designed to screen for genes encoding membrane transporters. A screening library targeting membrane transporter genes was designed by sgRNA design pipeline. FACS sort was used to enrich cells that uptake more drugs and guide RNAs enriched from this population was sequenced. Cell population imported less drugs was also studied as the negative control. By applying gRNA qualification pipeline, the gRNA abundance of different samples are efficiently deconvoluted. Finally, essential genes under different drug treatments (Doxorubicin and Aclarubicin drug are enriched and depleted in K562 cells) are identified through the screening analysis pipeline.



(a) -10% fluorescence intensity distribution

(b) +10% fluorescence intensity distribution



(c) -10% fluorescence intensity distribution

(d) +10% fluorescence intensity distribution

Fig. 5.7 Screening results of membrane transporter after Aclarubicin and Doxorubicin treatment, K562 cells are sorted by fluorescence flow sorting.

After treatment with Aclarubicin and Doxorubicin, K562 cells were enriched (Acla +10% as fig. 5.7a and Doxo +10% as fig. 5.7c) and depleted (Acla -10% as fig. 5.7b and Doxo -10% as fig. 5.7d). According to the screen results, ABCB1 and ABCG2 are known to encode membrane transporter genes that excrete drugs, and are negatively selected under the condition of anthracycline molecule enrichment. Anthracycline molecules are positively selected under depleted conditions. At the same time, SLC family membrane transporters SLC2A3 and SLC16A2 were identified that may promote doxorubicin uptaking into cells.

After further analysis of the volcano plot of membrane transporter enrichment (Doxo +10%) after doxorubicin treatment, as shown in Figure 5.8, the intracellular content of sgRNA targeting ABCG2, MFSD2A and SLC35G5 can be seen significantly decreased, indicating that after the transporter is activated, the drug is excreted faster outside the cell; and after activating SLC2A3 and SLC18B1, the drug diffuses into the cell faster.

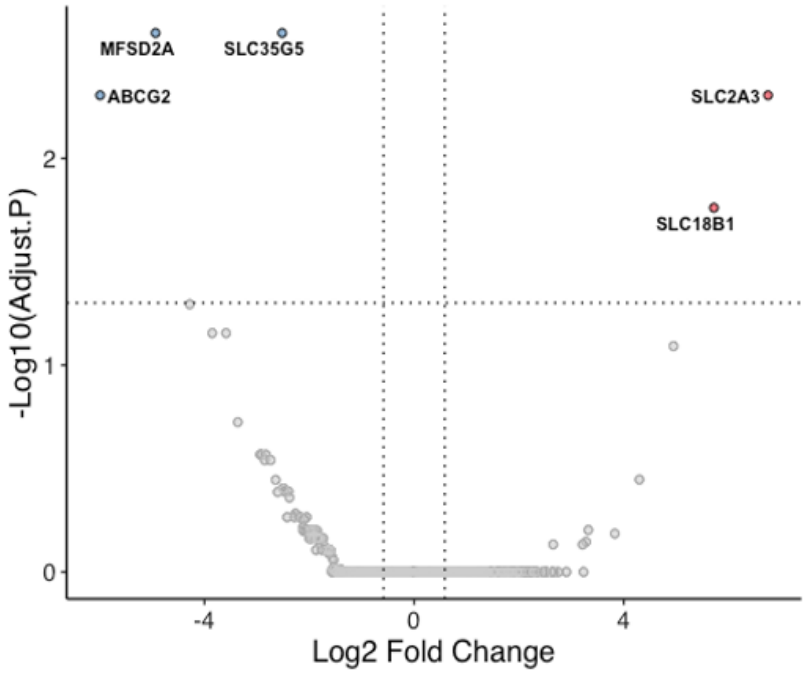


Fig. 5.8 Volcano plot of drug-enriched membrane transporter after doxorubicin treatment.

Pathway enrichment analysis of the top-ranked genes also proves that, as shown in Figure 5.9, anthracyclines diffuse into cells are similar to biological process such as minerals absorption, synaptic vesicle circulation and nicotine addiction.

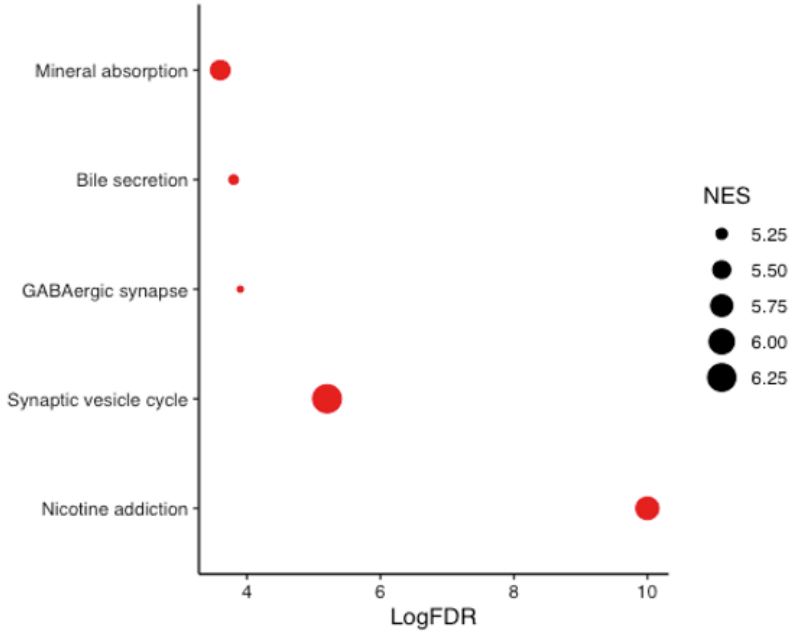


Fig. 5.9 Analysis of membrane transporter pathway enrichment of drug enrichment after doxorubicin treatment.

5.5.2 Case 2: Drug Resistance Analysis

Anthracycline drugs are known to inhibit topoisomerase II, causing DNA double-strand breaks and accelerating the process of apoptosis. However, for the target of drug action, which gene functional regions play a key role in drug resistance is not known from the current research results. By using dual CRISPR screen, the role of functional elements on drug resistance will be studied.

K562 cells infected with the dual crispr library treated with two drugs (doxorubicin and imatinib) were screened for survival pressure. After 15 days treatment, samples are send to next-generation sequencing. The calculation pipeline of the guide RNA content determination is also used to determine the pgRNA content in each sample. Finally, the calculation pipeline of the screening analysis is used to determine the functional non-coding region to be verified later.

Before analysis, the consistency of biological experiments under the same screening conditions and the correlation of samples under different screening conditions were compared. To normalize the original library count, the logarithmic transformation of reads per million (RPM) standardization was used for each sample. Here Spearman Coeffient was used to calculate the correlations as shown in Figure 5.10:

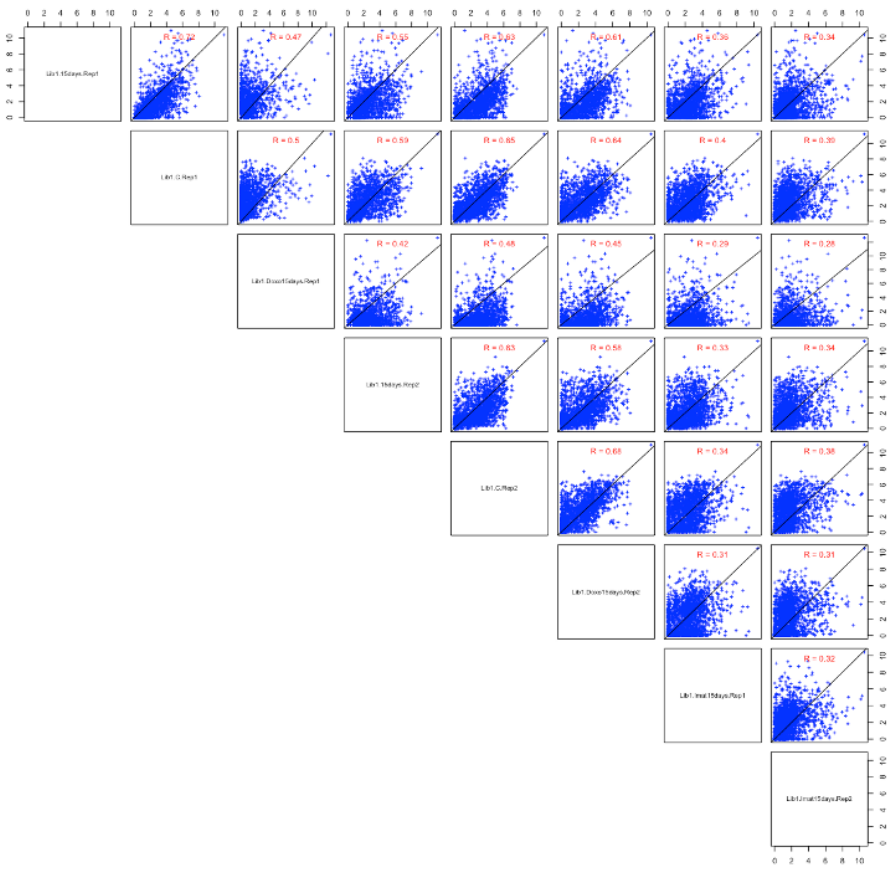


Fig. 5.10 Correlation of sample RPM after anthracycline treatment.

It can be seen that samples after the same drug treatment has a high correlation, indicating the biological replicates are of good quality. But the correlation is not high compared across different perturbation conditions, indicating some non-coding genomic elements have significant selectivity advantage towards different treatments.

Figures 5.11 and 5.12 show the results of dual CRISPR screen analysis. Figures 5.11 shows the positively selected essential non-coding regions with higher priority (top 10) after doxorubicin treatment, which are mainly enrich in the putative enhancer regions. To observe each pgRNA abundance changes and validate the screen performance of each guide design, we zoom into the chr9:133738411-13373861 region as the screen analysis pop out. All the designed guide pairs targeting this region consistently show the tendency of increasing, which verifies the guide targeting efficiency. The biological function of the hit, putative enhancer region chr9: 133738411-13373861, will be validated experimentally.

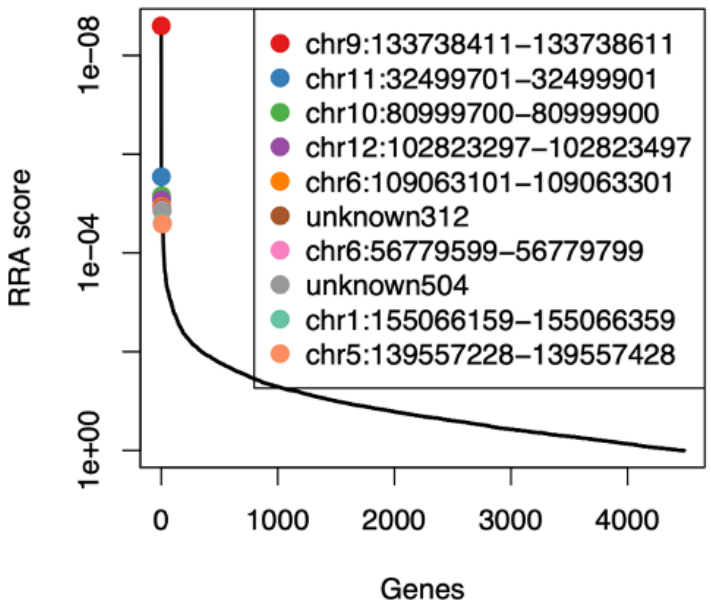


Fig. 5.11 Results of MAGeCK screen analysis.

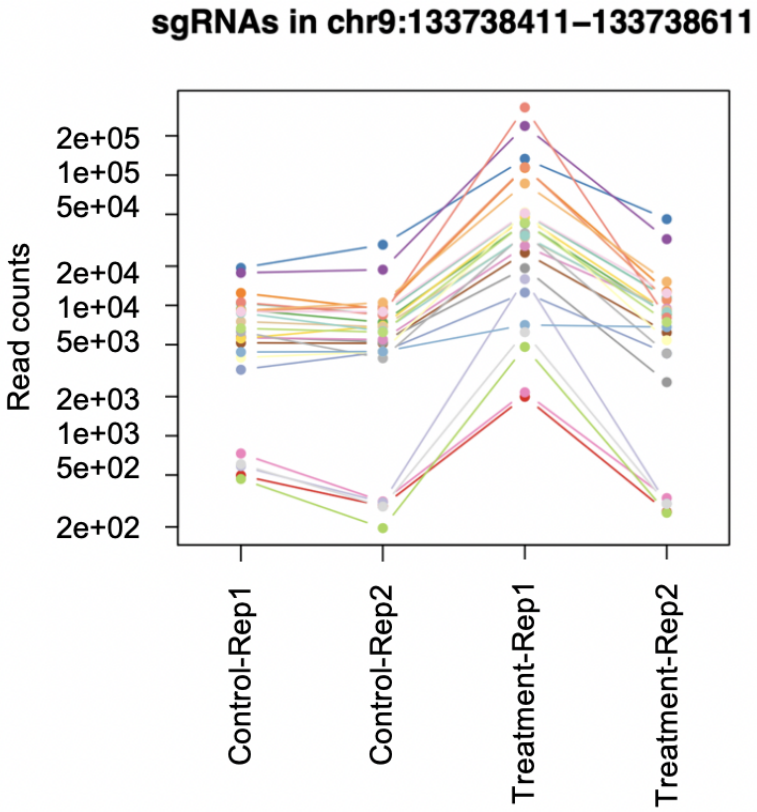


Fig. 5.12 Abundance change of pgRNA targeting chr9: 133738411-13373861.

In the last step, the significance scores under different perturbations (usually compared with control), like beta score output by MAGeCK-MLE algorithm, need to be associated to infer the biology meaning behind the screen results.

As shown in Figure 5.13, the non-coding regions in the upper-middle and the bottom-middle are screened out as positive selection and negative selection regions after comparing the doxorubicin treatment with the control. The hits in the upper-middle are associated with drug resistance, and the hits in the bottom-middle functions as drug sensitive non-coding regions, which both play important roles in cell survival.

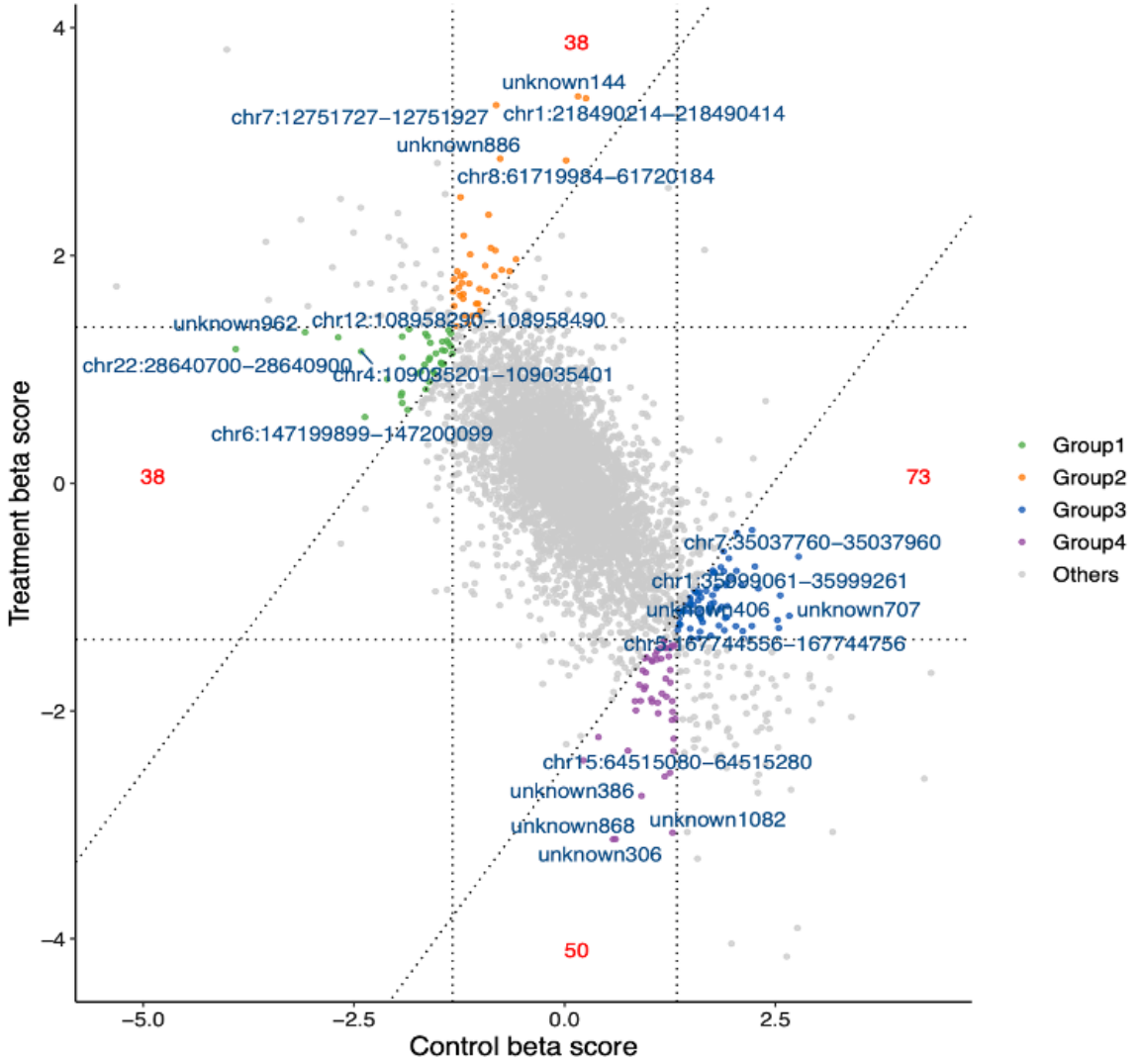


Fig. 5.13 Nine square scatter plot of beta scores after treatment with doxorubicin and the control group.

Chapter 6

Discussion and Conclusion

6.1 Discussion

Non-coding regions in the genome have an important regulatory effect on cell activities. As a technique for identifying the function of non-coding regions, the dual CRISPR screening system is gaining more and more attention because of its advantages such as it could remove the target regions completely. Therefore, there is also an urgent need for a computer-aided design and analysis process for the dual CRISPR screening system, from designing efficient screening libraries, to deconvolution of high-throughput sequencing screening experiment results, to accurately identification of candidate regions and final visualization and annotation of screening results. This report is mainly devoted to solving these practical application problems and developed a set of computational pipelines based on Snakemake workflow management tool, covering the computation tasks from dual CRISPR system screening library design, pgRNA abundance quantification and screening result analysis. These pipelines can be easily deployed on computing platforms such as clusters and cloud servers. Users can flexibly adjust program operating parameters according to their needs to achieve personalized CRISPR screening experiment design and data analysis. The calculation process realizes the end-to-end operation, and the user input the file in the format required by the program to return to the standardized output.

The highlights of this article are:

- A detailed review of the components of the dual CRISPR screening system and the experimental design process of the dual CRISPR screening to identify functional non-coding regions. A supplementary explanation for the precautions in the design of double CRISPR screening experiments, such as off-target effects.

- Before elaborating the tool development process, the characteristics of the current mainstream CRISPR design tools and analysis software were carefully investigated, which provided convenience for subsequent research to screen experiment-related software and build calculation processes according to experiment needs.

- According to the characteristics of the dual-CRISPR screening system, the differences between the dual-CRISPR screening library design and the CRISPRko or CRISPRa guided by sgRNA are specifically explained. These differences mainly include the definition of the target design area, the scoring mechanism of the paired gRNA and the role of the scaffold area connected to the gRNA. In the test session of the screening library design module, the optimized pgRNA was designed for different types of target regions, and the off-target effect in the library design session was analyzed.

- The sequencing results of high-throughput screening experiments often introduce many confounding factors. The pipeline filters these confounding factors from the original data to obtain the effect of knocking out real genes or non-coding regions. Quality control operations on the original sequencing data can ensure high-quality alignment and unbiased determination of pgRNA abundance.

- Based on the changes in pgRNA abundance under different perturbations, a combination of parametric modeling and non-parametric testing is used to sensitively identify candidate key genes or non-coding regions from heterogeneous screening data.

- The results of screening experiments are carefully handled by the pipeline. Because they might include false positive hits caused by variations in copy number and differences in sgRNA targeting efficiency. The screen data may also suffer from problems like cell cycle and batch effects, which is resolved by incorporating useful processing steps in the pipeline.

- To further interrogate the data from dual CRISPR knockout screen, plotting scripts was integrated in the pipelines to facilitates visualizing and explaining the screening results. Clustering analysis was done to explore the candidate clusters of functional non-coding regions.

- A GUI was designed for new users to easily and visually edit configuration files of expert-validated pipelines and can interactively execute these production-ready workflows.

6.2 Perspectives

Reasonable screening library design is a key step for the successful implementation of CRISPR screening experiments. At present, most of the commonly used library design for screening experiments are experimentally verified public data sets, and the auxiliary library design tools are mostly for scanning PAM motif patterns to search for potential targets, protospacer sequence. With the accumulation of screening library data for different application scenarios, library design methods based on sequence pattern mining can be further improved to design more efficient and accurate guide RNA for CRISPR screening experiments. Furthermore, the synthesis process of screening libraries with dual CRISPR systems is relatively complicated. Efficient targeting activity of pgRNA can save researchers a lot of experimental effort and economic costs.

The dual CRISPR screening technology used to identify the function of non-coding regions is not yet mature, and the analysis of experimental results after screening also lacks a standardized processing procedure. Especially for identifying key non-coding regions that affect phenotypes, most of the currently available analysis tools are developed for experiments such as RNAi or CRISPRko. There is no analysis tool for the more complex dual CRISPR screening of targeted regions. In future, data from transcriptional factor binding, DNA modification, chromatin accessibility states, sequence variation and other complex factors can be integrated into the analysis pipeline. It is still a problem to identify the real genetic effects of the knock-out target area from these factors. For example, false positive hits caused by copy number variation. Because the CNV gain state is cell-line specific, and the current main research results are only revealing the CNV state near the protein-coding gene, it is difficult to obtain the CNV state targeting the non-coding region. CNV states in the non-coding regions are still needed.

The statistical modeling of screening results is often based on certain assumptions. For instance, MAGeCK assumes that the gRNA content follows a negative binomial distribution, and CB2 assumes that the gRNA content follows the beta binomial distribution. Reasonable probability assumptions can better describe the characteristics of the data itself, especially for screening experiments where multiple pairs of pgRNA are often designed to target the same region. Different pgRNAs have different statistical distribution characteristics, such as the abnormality of individual pgRNA in a targeted region. Enrichment or depletion often leads to misjudgment of the targeted region's selectivity. A more reasonable model assumption is needed to describe this phenomenon, such as hierarchical mixed model modeling.

In all, the functional identification of non-coding regions can help people better understand the network regulation mechanism on the genome. Dual CRISPR screening technology as a potential research method will be further improved in the future.

References

[1] Lihua J Zhu, Benjamin R Holmes, Neil Aronin, and Michael H Brodsky. Crisprseek: a bioconductor package to identify target-specific guide rnas for crispr-cas9 genome-editing systems. *PloS one*, 9(9), 2014.

[2] Sangsu Bae, Jeongbin Park, and Jin-Soo Kim. Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 rna-guided endonucleases. *Bioinformatics*, 30(10):1473–1475, 2014.

[3] John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2):184, 2016.

[4] Miguel A Moreno-Mateos, Charles E Vejnar, Jean-Denis Beaudoin, Juan P Fernandez, Emily K Mis, Mustafa K Khokha, and Antonio J Giraldez. Crisprscan: designing highly efficient sgrnas for crispr-cas9 targeting in vivo. *Nature methods*, 12(10):982–988, 2015.

[5] Han Xu, Tengfei Xiao, Chen-Hao Chen, Wei Li, Clifford A Meyer, Qiu Wu, Di Wu, Le Cong, Feng Zhang, Jun S Liu, et al. Sequence determinants of improved crispr sgrna design. *Genome research*, 25(8):1147–1157, 2015.

[6] Wei Li, Han Xu, Tengfei Xiao, Le Cong, Michael I Love, Feng Zhang, Rafael A Irizarry, Jun S Liu, Myles Brown, and X Shirley Liu. Mageck enables robust identification of essential genes from genome-scale crispr/cas9 knockout screens. *Genome biology*, 15(12):554, 2014.

[7] Hyun-Hwan Jeong, Seon Young Kim, Maxime WC Rousseaux, Huda Y Zoghbi, and Zhandong Liu. Beta-binomial modeling of crispr pooled screen data identifies target genes with greater sensitivity and fewer false negatives. *Genome research*, 29(6):999–1008, 2019.

[8] Aaron A Diaz, Han Qin, Miguel Ramalho-Santos, and Jun S Song. Hitselect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic acids research*, 43(3):e16–e16, 2015.

[9] Philipp N Spahn, Tyler Bath, Ryan J Weiss, Jihoon Kim, Jeffrey D Esko, Nathan E Lewis, and Olivier Harismendy. Pinapl-py: A comprehensive web-application for the analysis of crispr/cas9 screens. *Scientific reports*, 7(1):1–8, 2017.

[10] Gaoxiang Jia, Xinlei Wang, and Guanghua Xiao. A permutation-based non-parametric analysis of crispr screen data. *BMC genomics*, 18(1):545, 2017.

[11] Biao Luo, Hiu Wing Cheung, Aravind Subramanian, Tanaz Sharifnia, Michael Okamoto, Xiaoping Yang, Greg Hinkle, Jesse S Boehm, Rameen Beroukhim, Barbara A Weir, et al. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences*, 105(51):20380–20385, 2008.

[12] Carlos Pulido-Quetglas, Estel Aparicio-Prat, Carme Arnan, Taisia Polidori, Toni Hermoso, Emilio Palumbo, Julia Ponomarenko, Roderic Guigo, and Rory Johnson. Scalable design of paired crispr guide rnas for genomic deletion. *PLoS computational biology*, 13(3), 2017.

[13] Wei Li, Johannes Köster, Han Xu, Chen-Hao Chen, Tengfei Xiao, Jun S Liu, Myles Brown, and X Shirley Liu. Quality control, modeling, and visualization of crispr screens with mageck-vispr. *Genome biology*, 16(1):281, 2015.

[14] Binbin Wang, Mei Wang, Wubing Zhang, Tengfei Xiao, Chen-Hao Chen, Alexander Wu, Feizhen Wu, Nicole Traugh, Xiaoqing Wang, Ziyi Li, et al. Integrative analysis of pooled crispr genetic screens using mageckflute. *Nature protocols*, 14(3):756–780, 2019.

[15] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[16] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[17] Bin Shen, Wensheng Zhang, Jun Zhang, Jiankui Zhou, Jianying Wang, Li Chen, Lu Wang, Alex Hodgkins, Vivek Iyer, Xingxu Huang, et al. Efficient genome modification by crispr-cas9 nickase with minimal off-target effects. *Nature methods*, 11(4):399, 2014.

[18] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[19] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

[20] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[21] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[22] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[23] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[24] Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, et al. Computational correction of copy number effect improves specificity of crispr–cas9 essentiality screens in cancer cells. *Nature genetics*, 49(12):1779–1784, 2017.

[25] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[26] Justin Malin, Mohamed Radhouane Aniba, and Sridhar Hannenhalli. Enhancer networks revealed by correlated dnase hypersensitivity states of enhancers. *Nucleic acids research*, 41(14):6828–6838, 2013.

[27] Xi Wang, Murray J Cairns, and Jian Yan. Super-enhancers in transcriptional regulation and genome organization. *Nucleic acids research*, 47(22):11481–11496, 2019.

[28] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947, 2013.

[29] Seyed Ali Madani Tonekaboni, Parisa Mazrooei, Victor Kofia, Benjamin Haibe-Kains, and Mathieu Lupien. Identifying clusters of cis-regulatory elements underpinning tad structures and lineage-specific regulatory networks. *Genome research*, 29(10):1733–1743, 2019.

[30] Sebastian Pott and Jason D Lieb. What are super-enhancers? *Nature genetics*, 47(1):8–12, 2015.

[31] Dimitri Desvillechabrol, Rachel Legendre, Claire Rioualen, Christiane Bouchier, Jacques Van Helden, Sean Kennedy, and Thomas Cokelaer. Sequanix: a dynamic graphical interface for snakemake workflows. *Bioinformatics*, 34(11):1934–1936, 2018.

[32] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A Pennacchio. Vista enhancer browser—a database of tissue-specific human enhancers. *Nucleic acids research*, 35(suppl_1):D88–D92, 2007.

[33] Slavica Dimitrieva and Philipp Bucher. Ucnebase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic acids research*, 41(D1):D101–D109, 2013.

[34] Kendall R Sanson, Ruth E Hanna, Mudra Hegde, Katherine F Donovan, Christine Strand, Meagan E Sullender, Emma W Vaimberg, Amy Goodale, David E Root, Federica Piccioni, et al. Optimized libraries for crispr-cas9 genetic screens with multiple modalities. *Nature communications*, 9(1):1–15, 2018.

[35] Neville E Sanjana, Ophir Shalem, and Feng Zhang. Improved vectors and genome-wide libraries for crispr screening. *Nature methods*, 11(8):783, 2014.

[36] Traver Hart, Kevin R Brown, Fabrice Sircoulomb, Robert Rottapel, and Jason Moffat. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular systems biology*, 10(7):733, 2014.

[37] Traver Hart, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R Brown, Graham MacLeod, Monika Mis, Michal Zimmermann, Amelie Fradet-Turcotte, Song Sun, et al. High-resolution crispr screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, 163(6):1515–1526, 2015.

[38] Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, 31(9):827–832, 2013.

[39] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, et al. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.

[40] Liam DH Elbourne, Sasha G Tetu, Karl A Hassan, and Ian T Paulsen. Transportdb 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Research*, 45(D1):D320–D324, 2017.