NAR Genomics and Bioinformatics, 2024, **6**, lqae059
https://doi.org/10.1093/nargab/lqae059
Advance access publication date: 25 May 2024
Pipelines and Workflows for Biological Data Analysis

PitViper: a software for comparative meta-analysis and annotation of functional screening data

Paul-Arthur Meslin[1,*], Lois M. Kelly[1], Salima Benbarche[1], Séverine Lecourt[1,2], Kevin H. Lin[3], Justine C. Rutter[3], Christopher F. Bassil[3], Raphael Itzykson[1,4], Kris C. Wood[3], Alexandre Puissant[1,†] and Camille Lobry[1,*,†]

[1]Université de Paris Cité, Inserm U944 and CNRS UMR 7212, Institut de Recherche Saint Louis, Hôpital Saint Louis, APHP, 75010 Paris, France
[2]Inserm U1279, Gustave Roussy Institute, Université Paris-Saclay, Villejuif, France
[3]Department of Pharmacology and Cancer Biology, Duke University, Durham, NC, USA
[4]Department of Hematology, Saint Louis Hospital, Assistance Publique-Hôpitaux de Paris (APHP), Paris, France

[*]To whom correspondence should be addressed. Tel: +33153724041; Email: paul-arthur.meslin@inserm.fr
Correspondence may also be address to Camille Lobry. Tel: +33153724041; Email: camille.lobry@inserm.fr
[†]The last two authors should be regarded as Joint Last Authors.

Abstract

Recent advancements in shRNA and Cas protein technologies have enabled functional screening methods targeting genes or non-coding regions using single or pooled shRNA and sgRNA. CRISPR-based systems have also been developed for modulating DNA accessibility, resulting in CRISPR-mediated interference (CRISPRi) or activation (CRISPRa) of targeted genes or genomic DNA elements. However, there is still a lack of software tools for integrating diverse array of functional genomics screening outputs that could offer a cohesive framework for comprehensive data integration. Here, we developed PitViper, a flexible and interactive open-source software designed to fill this gap, providing reliable results for the type of elements being screened. It is an end-to-end automated and reproducible bioinformatics pipeline integrating gold-standard methods for functional screening analysis. Our sensitivity analyses demonstrate that PitViper is a useful tool for identifying potential super-enhancer liabilities in a leukemia cell line through genome-wide CRISPRi-based screening. It offers a robust, flexible, and interactive solution for integrating data analysis and reanalysis from functional screening methods, making it a valuable resource for researchers in the field.

Introduction

The discovery of genetic elements such as short hairpin RNA (shRNA) which can mediate selected RNA degradation, or the type II Cas protein which can precisely cleave genomic regions under the guidance of a single-guide RNA (sgRNA) molecule (1), has in recent years fostered an unprecedented profusion of functional screening methodologies. These screening approaches rely on the use of multiple single or pooled shRNA and sgRNA which offer cost-effective systems for simultaneously targeting genes or non-coding genomic regions of interest (2,3). Such systems enable researchers to query the role of thousands of genes in the maintenance of a given cell phenotype. High-throughput screening approaches using pooled shRNA and sgRNA libraries have been deployed, for instance, to identify putative tumor-promoting and -repressing genes that influence disease progression or to pinpoint drug sensitizer or resister genes whose modulation affects specific drug resistance phenotypes. Other CRISPR-based systems involving a non-functional Cas9 nuclease, called CRISPR/dCas9, have more recently been generated to modulate the accessibility of DNA to the transcriptional machinery at the transcription initiation site or recruit transcriptional activating or repressive effector domains, resulting in CRISPR-mediated interference (CRISPRi) or activation (CRISPRa) of targeted genes or genomic DNA elements (4,5).

The overall representation of every single sgRNA or shRNA in a given library—which is introduced by lentiviral or retroviral infection of the bulk cell population—is generally evaluated by high-throughput sequencing of the PCR-amplified DNA region encoding this given sgRNA or shRNA. An initial sequencing (T_0) of the input library is performed to estimate the abundance of each CRISPR guide or shRNA at the beginning of the experiment. The growth advantage or disadvantage which is imposed on cells by the targeting of specific candidate genes by selected sgRNAs or shRNAs may affect over time their abundance within the bulk cell population. Further downstream sequencing is then performed to obtain the final representation (T_f) of these candidate hits. For instance, if a CRISPR guide silences a cancer cell dependency, this guide will decrease in relative abundance from the first to the second sequencing of the cell population. Conversely, if a CRISPR guide targets a critical tumor-repressing gene, its abundance relative to the other sgRNAs in the library will increase.

The noticeable increase of many functional screening protocols was ultimately followed by the design of multiple computational algorithms based on statistical models which were inherent to the type of protocol being developed for the identification of essential elements (6–9). Despite the extraordinary statistical potential of these functional screening analysis methods to achieve robust identification of hit candidates, there is still a lack of computational tools which can integrate

Received: December 8, 2023. **Revised:** March 19, 2024. **Editorial Decision:** May 6, 2024. **Accepted:** May 10, 2024

multiple data analysis pipelines into one comprehensive and straightforward resource interface which would exploit multiple capabilities: i) this resource needs to be accessible to all research scientists with long-range expertise in computational biology who should be able to easily adjust and customize their threshold parameters, and tailor their analysis; ii) it needs to be flexible to handle a variety of functional screening data types (i.e. shRNA-, CRISPR/Cas9-based systems) and provide the most significant and reliable output results according to the type of elements being screened (i.e. coding or non-coding genomic regions) in any particular experimental settings; iii) output results need to be interactive to simultaneously provide, compare, and visualize results from multiple screening analysis pipelines; and, iv) these results have to be easily exploited to query additional databases and provide a contextualizing framework for follow-up validation studies.

Here, we present PitViper, for Processing, InTerpretation and VIsualization of PoolEd screening Results (https://github.com/lobrylab/PitViper), a publicly available, robust, and comprehensive platform. PitViper was designed with the intent of assisting computational biologists and researchers who may not benefit from extensive computer expertise, in performing cutting-edge integrative screening data analysis. The extended capabilities of PitViper allow the user to carry out *de novo* analysis and reanalysis of screening experiments whose sequencing data are provided under the most common format types. PitViper performs read quantification, normalization, and essentiality analysis using gold-standard as well as additional customized and advanced methods which were specially designed for its deployment. Simultaneous integration of all generated data can be comprehensively visualized as heatmaps and network diagrams depicting depletion or enrichment of CRISPR guides and shRNAs, or targeted coding and non-coding genomic regions of interest. These visualizations are interactive, programmable, customizable, and integrated into easily shareable reports which are amenable to publication. Importantly, PitViper includes a module that enables the interpretation and the contextualization of results by querying external gene ontology, gene and protein expression, dependency, and pathway enrichment databases, thereby facilitating the prioritization of candidate hits for follow-up biological studies.

Materials and methods

Biological resources

Two published CRISPR/Cas9-based screening studies were used as training datasets for the design and deployment of PitViper: (i) a first study, which used a sgRNA library targeting metabolism-related genes to identify metabolic vulnerabilities in acute myeloid leukemia (AML) cells treated with the BCL-2 inhibitor venetoclax (10); and, (ii) a second study which used a custom-made sgRNA library to map super-enhancer (SE) genomic regions critical for the proliferation of ETO2-GLIS2$^+$ acute megakaryoblastic leukemia (AMKL) cells (11).

Cell culture

The human M-07e cell line was derived from a 6-month-old patient with acute megakaryoblastic leukemia (AMKL) (12). M-07e cells were cultured in MEM-α (Sigma-Aldrich, M6199) supplemented with 20% FBS, Penicillin (100U/ml)-Streptomycin (100μg/mL) (Sigma-Aldrich, P4333) and 5 ng/ml of human GM-CSF (PeproTech, 300-03). HEK293T cells were cultured with DMEM (Sigma-Aldrich, D6429) 10% FBS (Sigma-Aldrich) and 100 U/mL penicillin–streptomycin.

CRISPR/Cas9-mediated silencing of super-enhancer regions

M-07e cells expressing sgRNA and dCas9-KRAB targeting the SE region 66 were produced using the following oligonucleotide sequences: SE_66_peak1_7FOR 5′ CACCGCTCTCGAGTGAGAAGTTGC 3′ SE_66_peak1_7REV 5′ AAACGCAACTTCTCACTCGAGAGC 3′. SgRNAs targeting Renilla (not present in the human genome) were used as a control. These oligonucleotides were synthesized (Sigma Aldrich), annealed, phosphorylated, and ligated into the linearized lentiviral vector pLV-hU6-sgRNA hUbC-dCas9-KRAB-T2a-GFP (Addgene #71237) upon BsmBI digestion.

RNA extraction, quantitative real-time PCR (qRT-PCR)

RNA was extracted using the RNeasy RNA Micro Kit (Qiagen) per the manufacturer's instructions and quantified using Qubit (ThermoFisher Scientific). One microgram of RNA was then reverse transcribed using SuperScript™ IV Reverse Transcriptase (ThermoFisher, 18090200), dNTP (ThermoFisher, 18427013), RNaseOUT™ Recombinant Ribonuclease Inhibitor (ThermoFisher, 10777019), Oligo(dT)20 Primer (ThermoFisher, 18418020). Quantitative PCR was performed using Kapa SYBR Fast Master mix (Sigma-Aldrich, KK4622) and the primers used were hCSF2RB fwd: 5′ AACGGGATCTGGAGCGAGTG 3′, hCSF2RB rev: 5′ AGATCACGATGAGGGCCAGC 3′, hGAPDH fwd: CTTTTGCGTCGCCAGCCGAG, hGAPDH rev: CCAGGCGCCCAATACGACCA. Each mRNA level was measured as a fluorescent signal normalized based on the signal for β-actin. Relative quantification was determined by the ΔΔCt method and normalized according to β-actin.

Lentivirus production

HEK293T cells were grown to 70–80% confluence in a 15-cm plate. Twenty four hours after seeding, cells were transfected using 4 μg VSVG, 8 μg PAX2, 10 μg plasmid of interest and PEI (1 μg/μl). The medium was changed after 6 h of incubation at 37°C and 5% CO_2. Viral supernatant was collected 48 h after transfection and filtered through a 0.45 μm filter (Millipore, SLHP033RB). The resulting virus was then ultracentrifuged with M-07e cells at 2300 rpm for 90 min with polybrene at 1 μg/ml (Sigma-Aldrich, TR-1003-G).

FACs analysis

Cells were analyzed on FACS CantoII using FACS Diva software (BD Biosciences) every 2 days for 24 days after infection of M-07e cells with CRISPR-Cas9 constructs.

Directional scoring method

To complement already published methods for essentiality analysis, we implemented an additional solution that we named as *Directional Scoring Method* based on sg/shRNA filtering. DESeq2 (13) was used on a count matrix generated by MAGeCK 'counts' to identify sg/shRNA with significant differential abundances between conditions. Replicate conditions are used by DESeq2 for the calculation of the $\log_2$ fold change and the adjusted *P*-value (FDR). An efficiency score was then computed for each sg/shRNA based on the following formula:

$$Efficiency = \begin{cases} -log_{10}\left(FDR\right) \times \ log_2\left(fold\ change\right), \\ if abs\left(log_2\left(fold\ change\right)\right) \geq 2\ and\ FDR \leq 0.25 \\ 0\ otherwise \end{cases}$$

Fold Change threshold and FDR threshold being customizable in the software graphical user interface, and displayed values are giving as a default example.

Scores were summarized at the gene level by counting both the number of strictly positive (n_p) and strictly negative (n_n) scores for each gene. Then, each gene was categorized as positive essential, negative essential, ambiguous, or unchanged according to the following criteria:

$$Category\ = \begin{cases} Positive\ essential\ if\ n_p \geq t\ and\ n_n \\ lt; t \\ Negative\ essential\ if\ n_n \geq t\ and\ n_p \\ lt; t \\ Ambiguous\ if\ n_n \geq t\ and\ n_p \geq t \\ Unchanged\ otherwise \end{cases}$$

where t is a threshold defined by the user and set at 2 by default. Gene-level scores were defined as the mean score of positive guides for positive essential elements or the mean score of negative guides for negative essential elements. Otherwise, gene-level scores were set to zero.

Single-guide or ShRNA enrichment analysis (SSREA)

A second method was developed in PitViper based on the global ranking of guides using either the *signal-to-noise* metrics, as previously described (14). This is calculated using the difference of means of replicates values scaled by the standard deviation between normalized read counts of replicates (referred as 'signal to noise ratio', $SNR\ = \frac{\mu_B - \mu_A}{\sigma_A + \sigma_B}$), or the *log_2 fold-change* ($log_2 FC\ = \frac{\mu_B}{\mu_A}$) (where μ is the mean of values in a condition and σ is the standard deviation of the condition), of each guide between two distinct conditions. Computed metrics for each guide and their rank were used to perform a single-guide or shRNA enrichment analysis (SSREA) using the fGSEA R's package (15). The different shRNA or sgRNA targeting the same gene or genomic element are considered as a set, in which enrichment in one condition versus the other is computed using the GSEA algorithm on the global pre-ranked list of sh/sgRNAs. Eventually, targeted elements are ranked by their computed Normalized Enrichment Score and defined as significant for a threshold False Discovery Rate inferior or equal to 0.25 by default.

Computational resources and software

The analyses were performed on a desktop computer with 8Gb of RAM in <15 min for four conditions with three replicates and a total of 7395 guides targeting 452 elements. PitViper was successfully used on several operating systems using Conda and Docker: Ubuntu 18.04.6 and up, Windows 10 (with Windows Subsystem for Linux and Conda), and macOS. The list of software used for the design of PitViper is shown in Supplementary Table S1.

Various functions were developed to generate publication-ready figures in high-quality vectorial format (SVG) using python3 package Altair (16) or Plotly (PNG). In addition to interactive figures, function parametrization was facilitated by using the ipywidgets python3 package to generate graphical forms inside Jupyter Notebooks (17).

MAGeCK MLE results were ranked by beta scores, which represent a metric similar to the log fold-change transformation. MAGeCK RRA scoring hits were ranked by $\log_2$ fold-changes. CRISPhieRmix (9) results were ranked according to the mean of the top 3 guides targeting each gene. The mean $\log_2$ fold change value for the three guides targeting each gene, based on the absolute $\log_2$ fold change, was calculated using the DESeq2 package in R. Normalized Enrichment Scores from the fGSEA package were used for ranking (defined by the package's authors as the 'enrichment score normalized to mean enrichment of random samples of the same size'). BAGEL2 results are ranked by the computed Bayesian Factor, where positive values indicate a $\log_2$ fold-change distribution likelihood greater under the essentiality hypothesis than non-essentiality. Finally, results from our directional scoring method are ranked by the average of products of $\log_2$ fold-change of guides by their $-\log_{10}$ adjusted *P*-value, as described above.

Reanalysis of shRNA screen data

The shRNA screen data from Lin *et al.* was analyzed through PitViper starting from a raw count matrix. The first step involved normalizing the count matrix with MAGeCK count command using the 'total' normalization method. DESeq2 was used to compute $\log_2$(Fold Change) and adjusted *P*-values of guides between conditions. Then, all six methods implemented in PitViper were applied on the data. MAGeCK MLE was used with the following parameters: '–adjust-method fdr –genes-varmodeling 0 –norm-method total –permutation-round 2'. MAGeCK RRA was used with the following parameters: '–gene-lfc-method median –adjust-method fdr –remove-zero-threshold 0 –sort-criteria neg –gene-test-fdr-threshold 0.25 –norm-method total –remove-zero both'. CRISPhieRmix was used with the following parameters: 'BIMODAL = TRUE, mu = -4, screenType = 'LOF'. Furthermore, for each gene, the average $\log_2$(Fold Change) value was computed using the top 3 of shRNAs ranked by $\log_2$(Fold Change) computed by DESeq2. The 50 control guides designed in the experiment were used as negative control guides for CRISPhieRmix. BAGEL2 was used with essential and non-essential genes list from the original publication of BAGEL. SSREA was used with the 'signal-to-noise' ranking method. The Directional Scoring Method was used with the following parameters: shRNAs FDR threshold = 0.25, minimal number of shRNAs threshold = 2 and shRNA $\log_2$(Fold Change) threshold = 1.

Results of essentiality estimation from each method were filtered at gene-level by applying the following thresholds. MAGeCK MLE: FDR < 0.05 and beta < 0, MAGeCK RRA: FDR < 0.05 and LFC < 0, CRISPhieRmix: local FDR < 0.05 and top-3 shRNAs $\log_2$(Fold Change) average < 0, BAGEL2: Bayesian factor > 0, Directional Scoring Method: score < 0, SSREA: FDR < 0.25 and NES < 0. Genes at the intersection of all methods were used for DepMap (CRISPR dependency, protein expression and deleterious mutations heatmaps) and GeneMania network visualizations. Genes at the union of all methods were used for RRA ranking aggregation and EnrichR analysis.

Reanalysis of CRISPRi screen data

The CRISPRi screen data from Benbarche et al. was analyzed through PitViper starting from raw FASTQ files. The first step involved the quantification of each sgRNA in each condition and replicate with MAGeCK count algorithm with default parameters. MAGeCK MLE, MAGeCK RRA, CRISPhieRmix, SSREA, and the Directional Scoring Method, were executed with the same default parameters than the shRNA data. To create a set of negative control sgRNAs, we randomly sampled 300 sgRNAs from non-significant guides between day 0 and day 21 in DESeq2 analysis. We used this set of negative control guides with CRISPhieRmix.

Results of essentiality estimation from each method were filtered at super-enhancer-level by applying the following thresholds. MAGeCK MLE: FDR < 0.25 and beta < 0, MAGeCK RRA: FDR < 0.25 and LFC < 0, CRISPhieRmix: local FDR < 0.25 and top-3 shRNAs log2(Fold Change) average < 0, Directional Scoring Method: score < 0, SSREA: FDR < 0.25 and NES < 0. Genes at the union of all methods were used for RRA ranking aggregation.

Simulation of CRISPR screening counts data

CRISPR screening counts data were simulated to benchmark tools implemented in the PitViper pipeline. Raw counts data were simulated using the simulation framework developed in Bodapati *et al.* (18). Three parameters were investigated: the number of guides per gene n, the guide targeting efficiency b (probability that a given guide will efficiently target a given genomic or RNA region), and the multiplicity of efficient guides e (the number of guides which efficiently target a gene of interest). Parameter $n \in \{5, 10, 25, 50, 75\ 100\}$, parameter $b \in \{0.2, 0.4, 0.6, 0.8\}$, and parameter $e \in \{0.2, 0.8\}$. For each combination of n, e and b, three simulations were performed with different random seeds to produce simulation replicates. Default parameters for all simulations include 10 000 nonessential genes and 150 negative essential genes and a depth factor of 1000.

Benchmarking of essentiality detection methods

The PitViper pipeline was automatically run over each simulated dataset with default parameters. BAGEL2 was not included in the benchmark analysis because it requires prior knowledge of essential and nonessential genes, which are not available in the simulated dataset. For MAGeCK MLE, MAGeCK RRA, CRISPhieRmix and SSREA, all genes with a computed FDR value <0.25 were considered negative essential predictions. CRISPhieRmix was run with 300 negative control guides simulated with parameters: $b = 0.8$ and $e = 0.8$. The same set of 300 $\log_2$(fold-change) values was used as negative control guides for all simulation combinations.

For each tool and for each simulation, numbers of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions were computed by comparing them to the list of known true essential genes simulated previously. Then, precision and recall were computed as follows:

$$precision = \frac{TP}{TP + FP}$$

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

Finally, the $F1$ score was computed using the precision and recall by considering the harmonic mean of the two values to get one single metric to compare all methods:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Data and code availability

M07e H3K27Ac and H3K4me3 ChIP-seq are available at ArrayExpress database (E-MTAB-4367). M07e ATAC-seq data is available in the Gene Expression Omnibus database under accession code GSE131462.

Pitviper code is available at https://github.com/lobrylab/PitViper.

Results

PitViper is a comprehensive computational tool for functional screen integration and visualization

PitViper was organized as follows: first, a pipeline was produced using Snakemake (19), a workflow management system for creating reproducible and scalable data analyses. Then, a command-line interface (CLI) was developed to facilitate the use of PitViper in an automated and reproducible manner. This CLI allows easy reanalysis of previously generated results. Finally, a graphical user interface (GUI) running locally on a server based on the Flask python package allows users to set up and run PitViper.

All dependencies can be installed with the Conda package manager using a YAML file containing the specification of all dependencies or using a Dockerfile. Actionable Jupyter Notebook reports can be automatically generated for dynamic viewing, and results can be exported as HTML reports. Interactivity and module parameterization were achieved using Altair (16), a declarative statistical visualization library for Python, and the python library Ipywidgets, respectively. Reports are customizable and contain quality control, single tool results, or multiple method integration and data annotation with external tools, such as EnrichR (20–22), GeneMANIA (23), or The Cancer Dependency Map Project (DepMap) (24,25) (Figure 1A and B).

PitViper was designed to integrate most common file formats, including unaligned raw sequences in FASTQ format, aligned sequences in BAM format, or count matrices provided as text files from pooled shRNA-, CRISPR/Cas9- and CRISPR/dCas9-based screens. In the first case, the preliminary alignment of sequencing results can be performed with Bowtie2 (26), allowing the user to define parameters adapted to various experimental designs. MAGeCK count is available as an alternative method for alignment and is subsequently used to generate read count tables and normalize reads from raw FASTQ files or already aligned BAM files (Figure 1A). Alternatively, an existing count matrix can be used to start the analysis after this step. Therefore, PitViper accepts as input all state-of-the-art file types usually generated for pooled screening such as FASTQ, BAM and count matrix.

Once a normalized read count table is generated, subsequent analyses (all of which are incorporated into PitViper) are performed using a selection of well-reported algorithms relying on distinct, yet complementary, statistical assumptions: (i) MAGeCK RRA (6) and MAGeCK MLE (7), which were published and developed by the same group for the analysis of CRISPR/Cas9 knockout screens; (ii) BAGEL2 (8), which

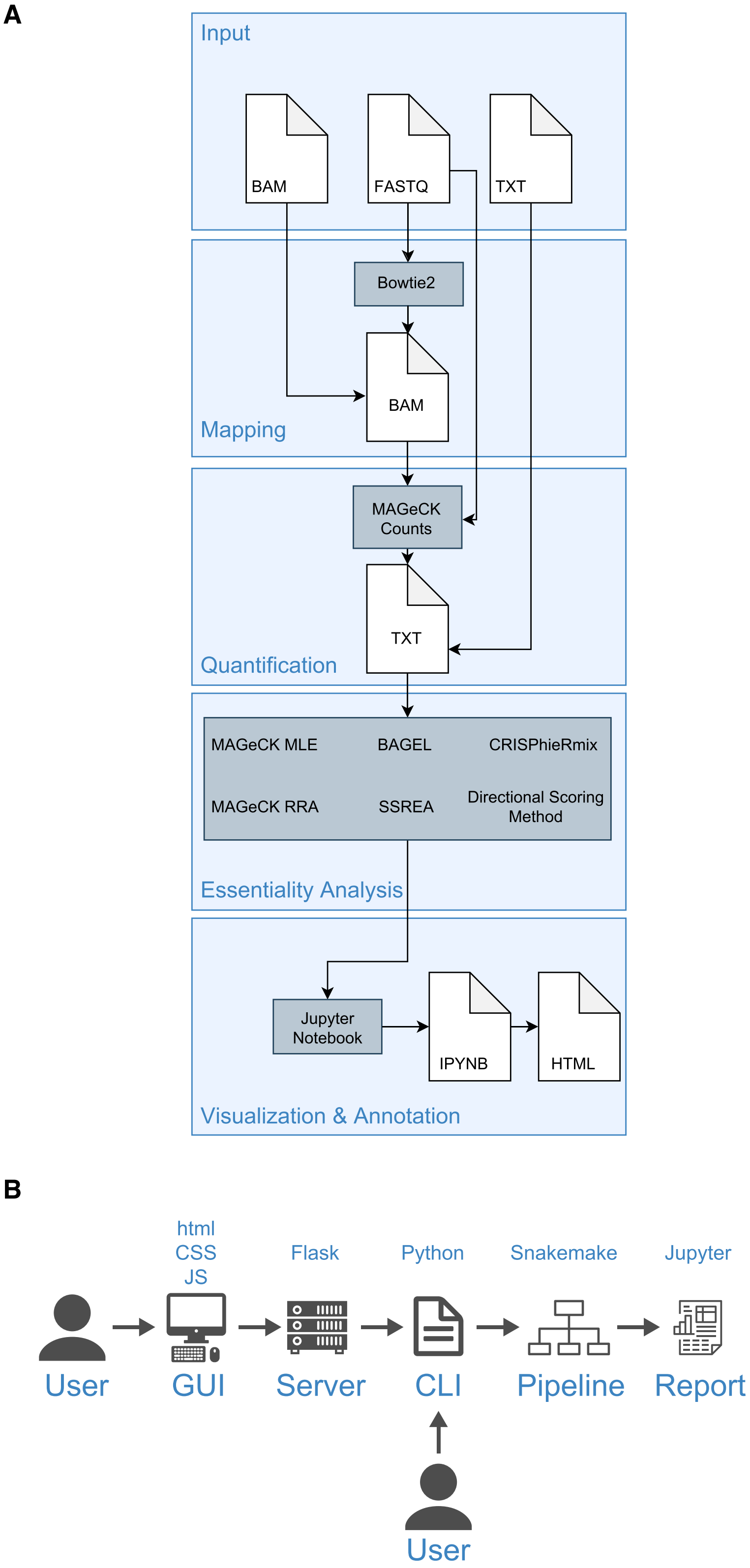


Figure 1. PitViper, Snakemake Workflow for Functional Screening Data Analysis. (**A**). PitViper Snakemake workflow, describing the main modules involved in screening data analysis. (**B**). PitViper module organization showing how users can upload their data either through Graphical User Interface (GUI) or Command Line Interface (CLI).

was designed as a supervised learning method to identify essential genes in CRISPR/Cas9 knockout screens using reference sets of essential and non-essential genes; and, (iii) CRISPhieRmix (9), which was developed to consider the variability of guide efficiency to inhibit or activate genes. Moreover, we have also included two additional analysis modules that we have designed with the ultimate goal of increasing the power of detection of essential hits. The first module which is referred to as *SSREA* (*Single-guide or ShRNA Enrichment Analysis*), consists of the re-purposing of the GSEA algorithm (14). Following the observation that a substantial proportion of CRISPRi/a guides are inefficient to affect the expression of their cognate target genes robustly and reliably, we sought to develop a method that would score the enrichment of features with low or barely significant representation changes. As sgRNA or shRNA overall representation follows a normal distribution, we decided to apply the GSEA scoring method. We used the features (sgRNA or shRNA) targeting a particular gene or genomic region as a set that should be scored among overall ranked features. A second module, referred to as the *Directional Scoring Method*, was designed based on the DESeq2 algorithm to filter out ineffective features and to rank candidate hits according to their aggregated signal across all effective features. After completion of all selected methods, an interactive Jupyter Notebook report is generated and can be exported into an interactive and shareable HTML report (Figure 1A).

The PitViper graphical user interface consists of: (i) a general settings panel; (ii) an input file upload panel; and, (iii) tool selection and parametrization panels. First, users are asked to enter a short and informative text string used by PitViper to create a directory to store all results (Supplementary Figure S1A). In the 'screen type' tab, users can choose between two options depending on the type of screening results that they intend to analyze. They can either select the 'gene' option if screening results are linked to known gene symbols, or 'not genes' if the screen targets are, for instance, enhancer regions, promoters, or other elements of the genome (Supplementary Figure S1B). For this latter case, a BED file containing targeted region localization can be uploaded to eventually allow annotation with proximal genes. In the last section, users are finally asked to select the file format selected for the analysis, which will be achieved in subsequent sections (Supplementary Figure S1C). Lastly, the number of jobs to run in parallel as well as the consideration of applying a filter to sg/shRNAs with raw counts values below a specified threshold before starting the analysis are to be addressed. (Supplementary Figure S1D, E).

In the input file upload panel, a sample sheet describing the relationship between conditions, replicates, and raw sequencing files is required. In addition, users can upload optional files including (i) a library file for read alignment if raw data were provided in the aforementioned section; (ii) a list of negative control features targeting known neutral genes; or, (iii) a count matrix file, if already available (Supplementary Figure S1F). Count matrices can also be used as input data to facilitate the reanalysis of previously generated or reported data. Several algorithms optionally use negative controls for essentiality identification (MAGeCK MLE, MAGeCK RRA, and CRISPhieRmix).

In the parameter panel, the PitViper GUI integrates options and parameters with expendable help from the tools' documentation. Specific parameters are already pre-selected by default; however, the user has the option of modifying any pre-applied settings to customize the analysis (Supplementary Figures S1G–J). Once all inputs and parameters are defined, the analysis is initiated, and a configuration file is automatically generated in YAML format in a configuration sub-folder. All chosen settings are summarized in this file to allow the PitViper Snakemake pipeline to run. This configuration file can be used as input for the PitViper CLI to reproduce an analysis with the same parameters, or to restart an interrupted analysis. Coupled with the Conda environment or the Dockerfile, this procedure facilitates results reproducibility. Alternatively, an empty configuration file can be edited and used with the CLI directly. YAML format is convenient for both users and automated modifications.

The sample sheet contains information on the relationship between conditions and replicates. For analysis starting from FASTQ/BAM files, paths to files are needed in an additional column. Raw data are not uploaded through the GUI, but rather defined in the sample sheet to limit unnecessary data transfer. All files uploaded by users are ultimately stored in the PitViper resources folder specific to the ongoing analysis.

PitViper outputs vary depending on starting data format and selected methods. Results include raw and normalized read count tables generated by the MAGeCK count software, gene-level results and a Jupyter Notebook report. This notebook document is ultimately generated to allow users to browse figures and results in an interactive and customizable environment. These notebooks can then be exported into a shareable file format, such as HTML. Raw and preprocessed data are directly embedded within the report and downloadable. Information from all steps of the PitViper Snakemake pipeline is stored in a log directory to archive the analysis process.

PitViper allows functional screen quality control assessment and identification, visualization of scoring candidate hits

To depict the analytic features and main functions of PitViper, we selected a CRISPR/Cas9 screening dataset in which a library of sgRNAs targeting metabolism-related genes was queried in an AML cell line, THP1. THP1 cells were treated for 11 and 21 days with either vehicle or a BCL2 inhibitor, venetoclax, at two drug concentrations (100 and 500 nM). Two replicates per condition were used, except for the input condition, in which only one replicate was sufficient to ensure the overall sgRNA library representation.

Several layers of quality control were performed throughout the pipeline. Normalized read count distribution was shown after $\log_2$ transformation of all replicates (Figure 2A). Principal Component Analysis (PCA) was then performed using the 'sklearn' python3 package[20] to control for consistency of read counts across replicates (Figure 2B). Taking the example of a top-scoring depleted gene in this screen, *TPI1*, we could visualize normalized read counts corresponding to all five *TPI1*-targeting sgRNAs as a z-score-transformed heatmap or a line chart across the various conditions (Figure 2C and D).

As exemplified by the automatic implementation of the MAGeCK MLE analysis pipeline, results from each functional screen analysis method could be visualized as a waterfall plot (Figure 3A). This depicts all non-significant (FDR ≥ 0.05) and significant hits (FDR < 0.05) from the top-enriched to the

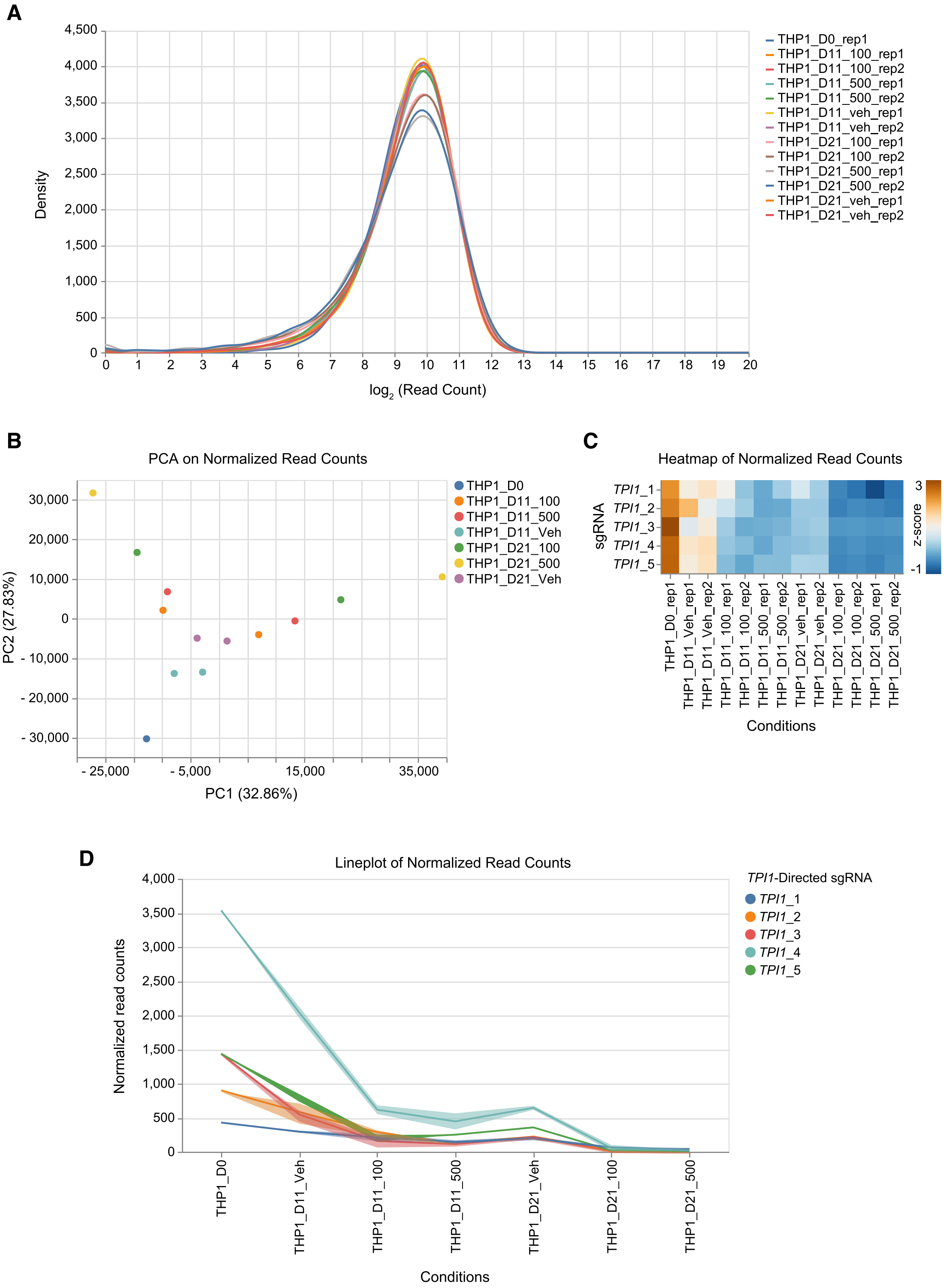
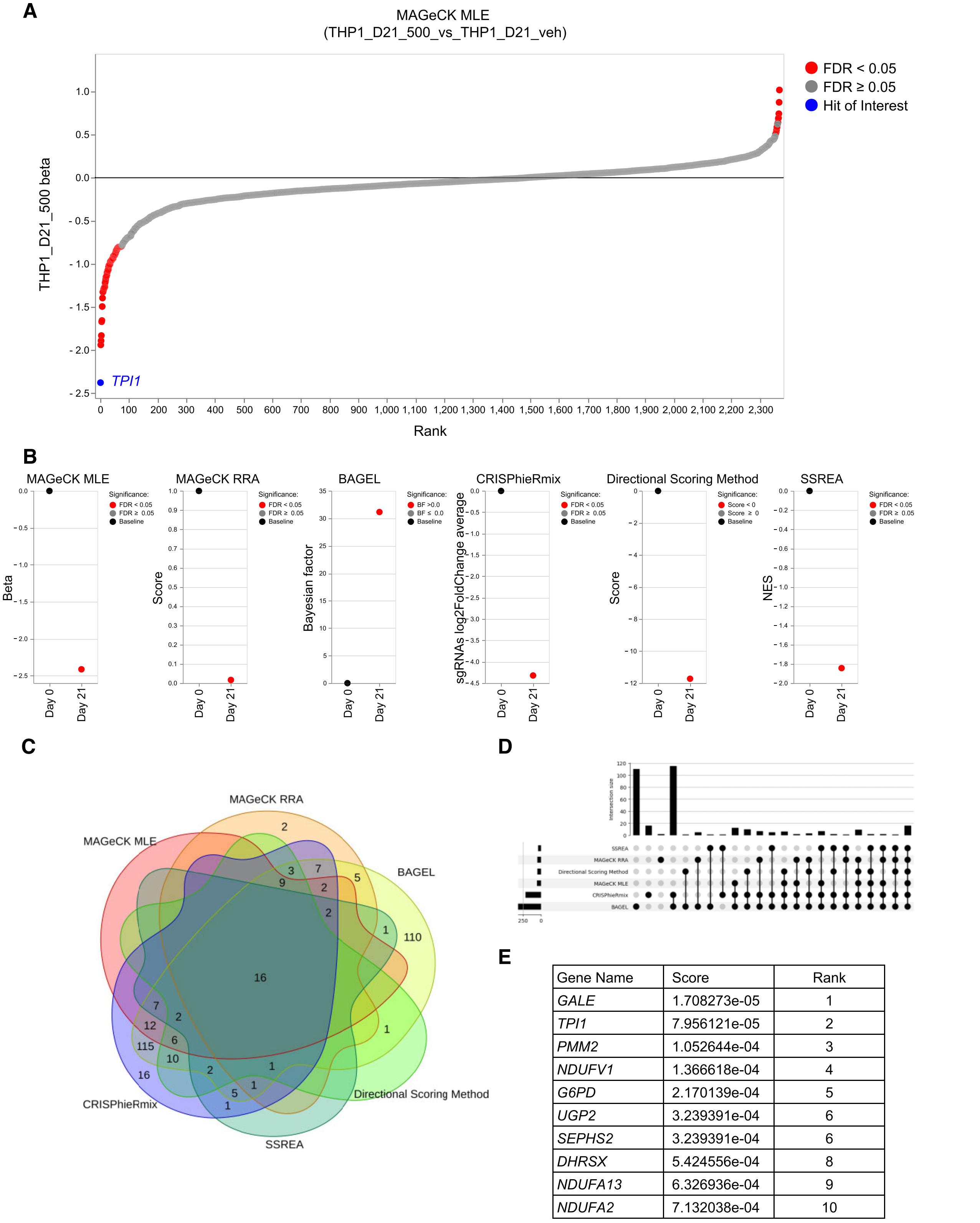


Figure 2. Quality control assessment. (**A**) Density distribution of $\log_2$ (read counts) across replicates. (**B**) Visualization of two first principal components of the normalized count matrix. Colored by condition. (**C**) Heatmap of z-score-transformed and normalized read counts of sgRNAs targeting a representative scoring gene hit, *TPI1*. (**D**) Line chart of normalized read counts of sgRNAs targeting *TPI1* gene. Lines represent the mean of the normalized read counts across replicates per condition, along the standard deviation.



Gene Name	Score	Rank
GALE	1.708273e-05	1
TPI1	7.956121e-05	2
PMM2	1.052644e-04	3
NDUFV1	1.366618e-04	4
G6PD	2.170139e-04	5
UGP2	3.239391e-04	6
SEPHS2	3.239391e-04	6
DHRSX	5.424556e-04	8
NDUFA13	6.326936e-04	9
NDUFA2	7.132038e-04	10

Figure 3. Visualization of scoring hits through various functional screening analysis methods. (**A**) Waterfall plot from the top-enriched to the top-depleted candidate hit following comparison of the THP1_D21_500 to the THP1_D21_Vehicle condition. Top-scoring gene, *TPI1*, exemplified in blue. Significant depleted or enriched hits (FDR < 0.05) based on the MAGeCK MLE method are shown in red. Y-axis representing the MAGeCK MLE-calculated beta score for each gene. Genes are ranked according to their beta scores. (**B**) Representative scoring candidate hit, *TPI1*, across all available screening analysis tools. Significant depletion of *TPI1* in each condition compared to vehicle is shown in red. Significance threshold established from each method's algorithm. (**C**) Venn diagram of the overlap of genes which scored according to each indicated tool. (**D**) Upset plot of the overlap of genes which scored according to each indicated tool. (**E**) Top-10 robust rank aggregate of the union of scoring hits across all methods from (C). Each hit is ranked according to its RRA score.

top-depleted candidates such as *TPI1*. Results for *TPI1* were illustrated across all conditions and tools (Figure 3B). Additionally, the depiction of other genes of interest could be facilitated using a search tool, allowing users to query any genes ranked in the waterfall plot based on their respective scoring criteria. Scoring hit candidates were finally compared and exemplified in a Venn diagram, which could be automatically designed by selecting any or all of the six screening analysis tools available in PitViper (Figure 3C). Default parameters were preselected, but the user is free to modify the scoring parameters of any tool to relax or restrict the number of included hits. The ranking of hits at the convergence of the selected methods was then used to define a hit prioritization ranking score across all selected screening analysis tools using the Robust Rank Aggregate (RRA) R's package (27) (Figure 3D).

PitViper allows multi-modal interrogation of scoring candidate hits with external multi-omic datasets and pathway enrichment tools

The DepMap R package (28) (v1.12.0.) automatically downloads the latest available version of DepMap datasets which include (i) whole-exome sequencing data; (ii) transcriptomic and proteomic data; and, (iii) a whole-genome dependency map of a wide panel of cancer cell lines which are representative of numerous tissues. Using these various DepMap datasets, we, for instance, queried cell lines from hematopoietic tissue by the list of top-scoring RRA-ranked hits and performed hierarchical clustering on both proteins and CRISPR DepMap dependencies, and cell lines (Figure 4A and B). In addition, if any mutation annotated as deleterious was identified in any of these scoring genes across any of the cell lines referenced in the DepMap dataset, we could visualize it in an interactive heatmap (Figure 4C). Hovering over each figure provides additional information about the type of events by measures such as protein expression changes or the type of genetic alteration. If scoring hits are associated with gene symbols, it is possible to investigate and identify additional scoring hit-related genes with external annotation tools such as GeneMANIA. Once a set of genes has been obtained, a hyperlink can be generated to open a GeneMANIA website page with selected genes as a queried set (Figure 4D). Genesets over-representation analysis (ORA) in the essential genes was carried out using a wrapper for EnrichR's python API. Results were represented for each selected gene set as dot or bar plots (Figure 4E). Users are allowed to select genesets libraries from the list available on the original EnrichR website and analyze them directly in the report. Lastly, when multiple conditions are compared to the same reference, such as the library representation at day 0, it's possible to visualize and compare gene-level scores between these conditions. This helps in identifying genes with unique patterns. For instance, we utilized this module to find genes that showed a more significant negative essentiality score when the treated condition at day 21 is compared to the condition at day 0, than day 21 treated with DMSO (Figure 4F) and were able to reproduce enrichment of Heme Biosynthesis pathway as previously published (10) using an associated Enrichr databases interrogation module (Figure 4G).

Comparative benchmarking analysis was employed to assess the performance of the different solutions implemented in PitViper, using simulated functional screening count data (23) (Supplementary Figure S2). We noticed significant performance fluctuations across the range of simulation parameters employed, with F1 scores varying between 0 and 0.89 across all combinations. As the number of sgRNAs n and their efficiency e increased, the performance improved for all methods. However, it was observed that when a relatively high number of sgRNAs were utilized per gene, with $n > 10$ and $n > 75$, for instance, the Differential Scoring Method and MAGeCK MLE were, respectively, outperformed by the other solutions. Importantly, each method was identified as top performer based on the F1 score for a minimum of four combinations (Supplementary Figures S2A and S2B). In addition, the parameter space highlighting the optimal performance of each method relative to the others exhibited significant variability: CRISPhieRmix showed superior performance with low targeting efficiency and a reduced multiplicity of efficient guides. In contrast, MAGeCK MLE displayed its peak performance under conditions of high sgRNA targeting efficiency and high multiplicity of efficient guides, with the important caveat that the number of sgRNAs used does not exceed 75; otherwise, its performance drastically declines, irrespective of targeting efficiency or the multiplicity of efficient guides. Both MAGeCK RRA and SSREA displayed a similar behavioral pattern, outperforming other methods when employing a high number of sgRNAs per target gene. Conversely, the Directional Scoring method showcased slight advantages in scenarios where the number of sgRNAs designed per gene and their efficiency were low (Supplementary Figure S3). Collectively, these findings underscore the attractive capability of PitViper in integrating these diverse analysis tools, thereby constituting a qualitative resource to comprehensively analyze a diversified array of screening systems.

Case study from a CRISPRi screening analysis of super-enhancer-regulated gene network essentialities

PitViper was used to reanalyze a genome-wide CRISPRi-based screen of super-enhancer (SE) genomic regions which was developed to (i) identify potential super-enhancer liabilities in an *ETO2-GLIS2*-positive AMKL cell line, M-07e; and, (ii) functionally map the rich clusters of genes whose expression is controlled by these critical SEs, and which might thereby cooperate to promote the growth of AMKL cells[24]. PitViper analysis started from raw FASTQ files. MAGeCK MLE, MAGeCK RRA, CRISPhieRmix, SSREA and the directional scoring method, were executed with default parameters. MAGeCK 'counts' module was used to directly quantify the number of reads associated with each guide in each condition. BAGEL2 was excluded from this analysis because this method was not robust enough due to the lack of prior knowledge on essential versus nonessential SEs. The intersection of top-depleted hits from the five different analysis methods pinpointed previously identified hits (Figure 5A). In addition, one SE (SE_66) was pinpointed as essential by two methods, excluding MAGeCK MLE, MAGeCK RRA and CRISPhieRmix, and was also not identified in the original screening report. The lack of power of MAGeCK MLE and MAGeCK RRA to detect this particular SE region as a scoring hit came from the fact that out of the 29 different sgRNAs directed against SE_66, only a subset of them exhibited significant depletion between day 0 and day 21 (Figure 5B). Accordingly, SE_66

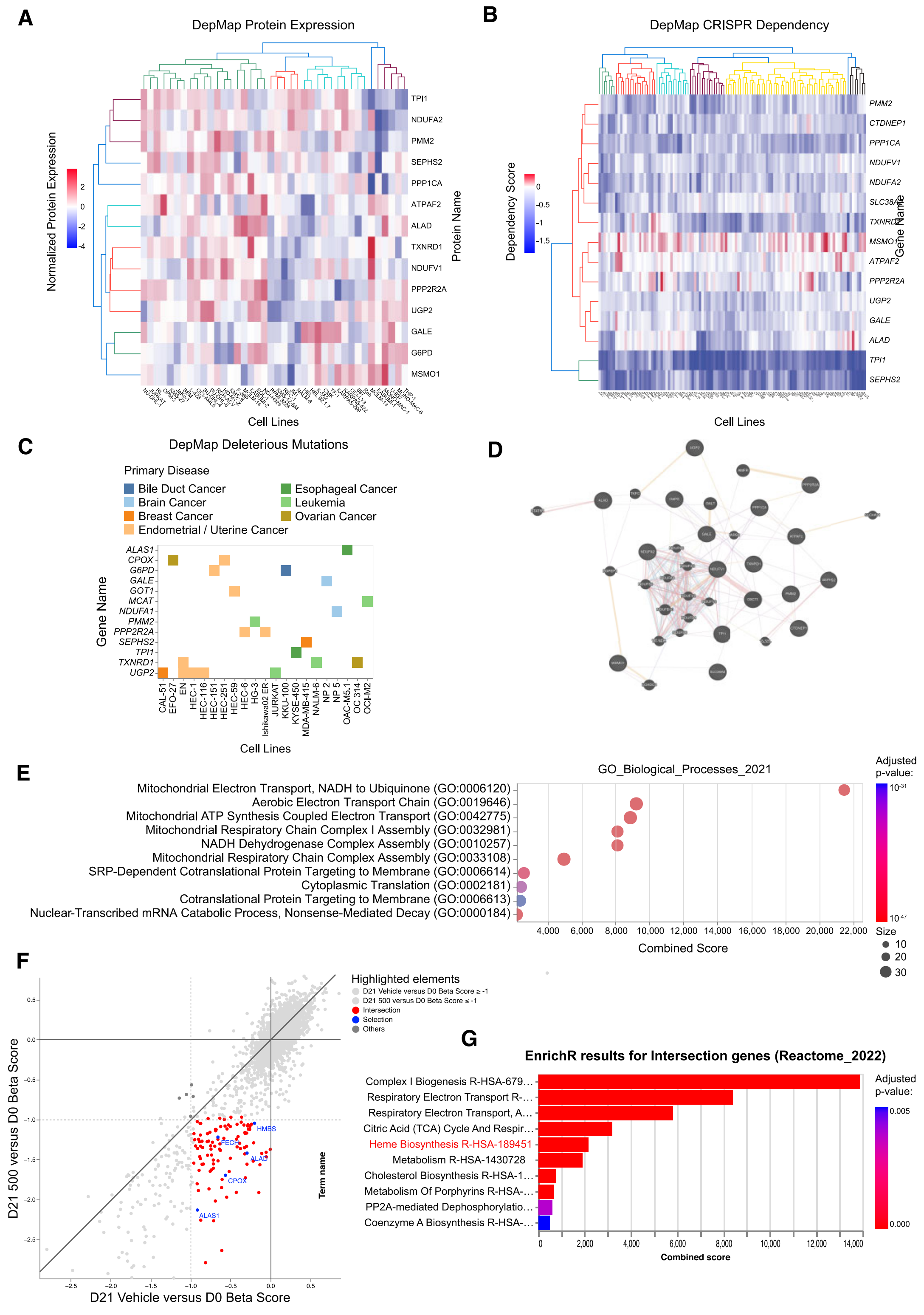


Figure 4. Interactive visualization of additional expression, mutation, interaction, dependency and pathway annotation tools. (A–C) Interactive heatmap of the protein levels (**A**), CRISPR dependency scores (**B**) and deleterious mutations (**C**) of each scoring hit at the intersection of all analysis methods from Figure 3C. Out of 16 hits, the 14, 15 and 8 for which expression (A), dependency (B) and mutation (C) information were available across all cell lines, respectively, are shown on the heatmap. Results were obtained from The Cancer Dependency Map Project (DepMap). Negative essential hits with FDR < 0.05 for all methods but SSREA (FDR < 0.25). (**D**) Network visualization generated with GeneMANIA website using the intersection of essential genes. (**E**) Overrepresentation analysis results obtained using EnrichR python's API with 'GeneOntology: Biological Processes' on the union of all hits are represented as a dot plot. Negative essential hits with FDR < 0.05 for all methods but SSREA (FDR < 0.25). (**F**) Scatter plot depicting the MLE beta score of each gene between 'Day 21 Vehicle against Day 0' normalized counts and 'Day 21 500 against Day 0' normalized counts. Genes highlighted in red are at the intersection of genes with a beta score ≥ -1 in D21 Vehicle vs D0 and a beta score ≤−1 in D21 500 versus D0. Genes enriched in the Reactome 'Heme Biosynthesis' pathway are depicted in blue. (**G**) Bar plot depicting combined score of Enrichr analyses of intersection genes using Reactome database. Heme Biosynthesis pathway is highlighted in red.

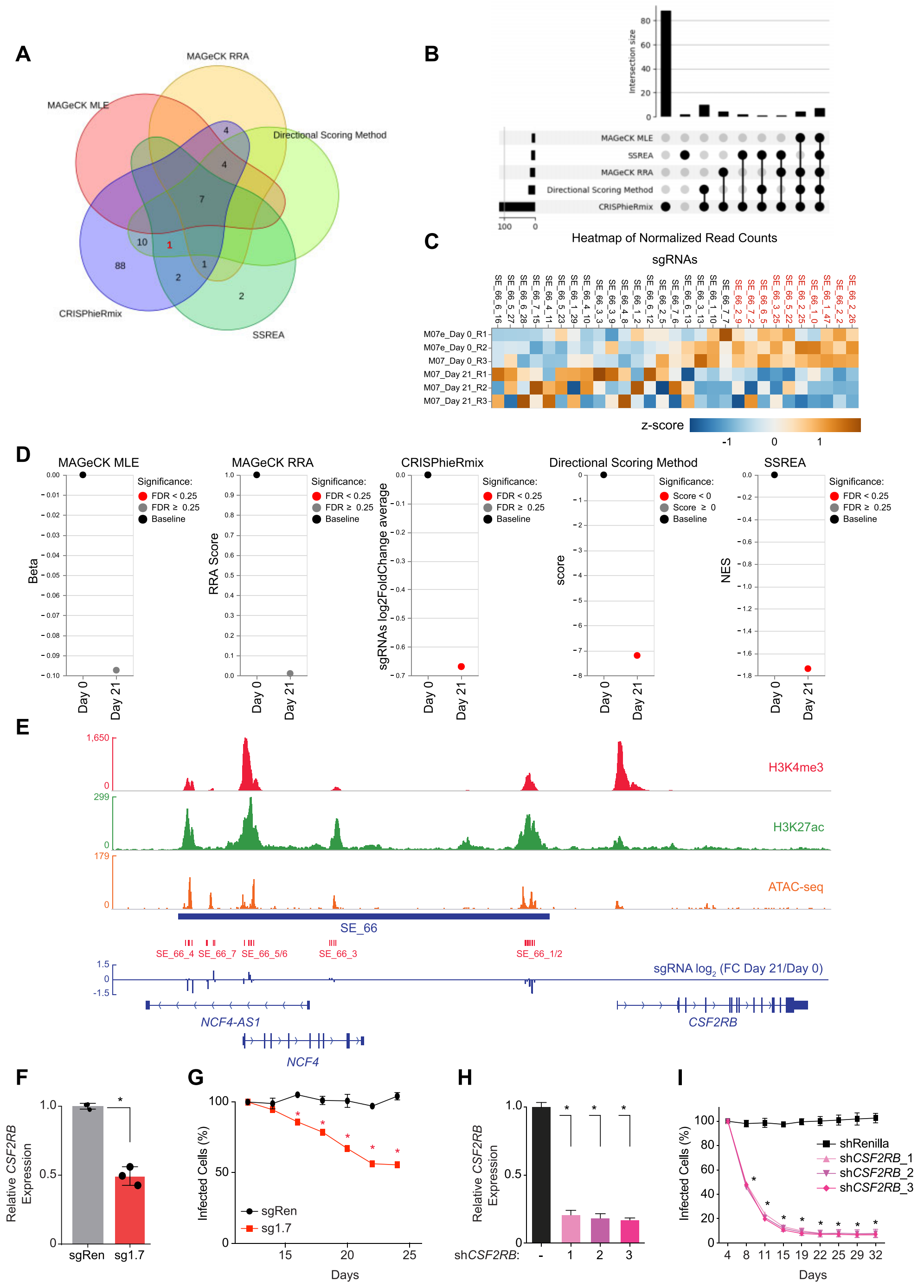


Figure 5. Case study: super-enhancer 66 (SE_66) is identified by three independent functional analysis methods as critical for M-07e cell growth. (**A**) Venn diagram of the number of super enhancers (SEs) identified at the intersection of all methods (FDR < 0.25 and negative selection) as critical dependencies. SE_66 (shown in red) is scored by 2 out of the five implemented methods. (**B**) Upset plot of the number of super enhancers (SEs) identified at the intersection of all methods (FDR < 0.25 and negative selection) as critical dependencies. (**C**) Heatmap of *z*-score-transformed normalized expression of sgRNAs targeting SE_66 between day 0 and day 21. sgRNA are sorted by FDR as estimated by MAGeCK RRA, from the less significant (left) to the more significant (right). sgRNA with FDR value < 0.25 are depicted in red. (**D**) Analysis of the essentially for SE_66 by MAGeCK MLE, MAGeCK RRA, CRISPhieRmix, SSREA and the directional scoring method. Gray and red dots indicate non-significant and significant depletion, respectively, according to the significance calculation of each method. (**E**) Gene track showing SE_66 localization in the human genome (hg19), annotation of the SE_66 proximal genes, *NCF4* and *CSF2RB*, the position of the sgRNA and their $\log_2$ fold-change of representation between day 0 and day 21 as well as normalized read density histograms for H3K27ac- and H3K4me3- ChIP-seq and chromatin accessibility assessed by ATAC-seq. (**F**) qRT-PCR analysis of *CSF2RB* mRNA levels in M-07e cells transduced with a SE_66-targeting sgRNA (sg1.7) and a control Renilla-targeting sgRNA (sgRen). Error bars represent mean ± SD of three replicate per condition. (**G**) Proportion of GFP-positive M-07e cells following CRISPRi inhibition of SE_66 with a SE_66-targeting sgRNA (sg1.7) compared to a control non-targeting sgRNA (sgRen) and normalized to day 8 after infection. Error bars represent mean ± SD of three replicate per condition. (**H**) qRT-PCR analysis of *CSF2RB* mRNA levels in M-07e cells transduced with three different *CSF2RB*-targeting shRNA vs a control Renilla-targeting shRNA. Error bars represent mean ± SD of three technical replicates. (**I**) Proportion of GFP-positive M07e cells following *CSF2RB* silencing using three *CSF2RB*-directed shRNAs compared to a control Renilla-directed shRNA and normalized to day 4 after infection. Error bars represent mean ± SD of three technical replicates. (F–I) *P*-value calculated using Welch's *t*-test. * $P < 0.05$.

was identified as significantly depleted over time only by the SSREA, and Directional Scoring methods (Figure 5C).

SE_66 is located on chromosome 22 proximal to the *CSF2RB* gene (Figure 5D). The protein encoded by this gene is a common beta chain subunit to type I cytokine receptors which could be activated by various cytokines, including interleukin-3 (IL-3) or granulocyte-macrophage colony-stimulating factor (GM-CSF). These two cytokines are required for M-07e growth[25], suggesting that the inactivation of SE_66 could lead to the repression of *CSF2RB* which, in turn, could affect the IL-3- or GM-CSF-mediated M-07e cell growth. To test this hypothesis, we performed a qRT-PCR analysis of *CSF2RB* mRNA levels in M-07e cells infected with either a Renilla non-targeting control or a SE_66-directed sgRNA, sg1.7. We confirmed that the inactivation of SE_66 reduced the expression of CSF2RB (Figure 5E). Decreased *CSF2RB* expression was associated with a sustained impairment of M-07e cell growth over time in comparison with the control sgRenilla-transduced cells (Figure 5F). To confirm the effect of *CSF2RB* repression on M-07e cell growth, we performed shRNA-mediated knockdown of this gene using three independent shRNAs (Figure 5G). Upon knockdown of *CSF2RB*, we confirmed indeed that M-07e cell growth was significantly impaired, thereby corroborating our previous results observed upon SE_66 inhibition (Figure 5H).

Together, our results show that PitViper enables users to exploit simultaneously and at once multiple screening analysis methods. These include two newly developed methods within the frame of the PitViper pipeline, to nominate scoring hits that could not be otherwise considered as noteworthy by specific and individual methods. We ultimately bench-proofed the robustness of our PitViper analysis pipeline by showing the relevance of one of these questionable hit candidates as a true essentiality in our model system.

Discussion

High-throughput functional screens constitute a set of powerful techniques for the identification, at a global scale, of coding and non-coding genomic elements which play a critical role in maintaining a phenotype of interest. In this study, we developed PitViper, an automated pipeline for rapid, easily reproducible analysis of newly generated data and the reanalysis of existing functional screening data (shRNA, CRISPR-Cas9 and CRISPR-dCas9) with multiple complementary methods for essentiality analysis. This solution was built on a Snakemake pipeline associated with an HTML-based graphical user interface and Jupyter Notebook reports.

One of the most important criteria to consider when using a screening analysis solution is the statistical accuracy of the given tool to pinpoint all possible true-positive candidate hits reliably and robustly, without retaining too many false positives. This accuracy depends on a trade-off between the sensitivity (likelihood of the identification of a positive hit, conditioned on truly being positive) and the specificity (likelihood of the identification of a negative hit, conditioned on truly being negative) of the prediction. The experimental design of the screen may affect the statistical potency of the platform being queried. For instance, BAGEL2 represents an invaluable tool to identify hits in a system where a list of curated and validated essential and non-essential elements is available. MAGECK MLE is more comprehensive than BAGEL2 to pinpoint essential elements in all possible conditions, especially in a context in which the efficiency of targeting guides is high. Its robustness tends to decrease, however, when guide efficiency is suboptimal (18). CRISPhieRmix was developed specifically for CRISPR/dCas9 screens and models the phenotypic effect of each gene. It is based on the mixture of a null distribution from sgRNAs targeting non-essential elements or ineffective sgRNAs targeting essential elements and an alternative distribution of effective sgRNAs targeting essential elements. We indeed observed that CRISPhieRmix identified the highest number of potential hits among all methods for a CRISPRi data but also tend to identify higher number of false positive in our benchmark. In addition, we have implemented in PitViper two additional methods: a repurposing of the GSEA algorithm and a method based on the directional filtering, scoring, and aggregation of sh/sgRNAs at the gene level. To validate these results, and to confirm that these two methods provide an accurate prediction of true-positive hit candidates, we conducted a case study in which we used PitViper to reanalyze a dataset of a genome-wide CRISPRi-based screen of super-enhancers in the M-07e cell line. In doing so, we were able to identify new SEs, which did not score as critical hits using standard one-tool analysis strategies. We validated the role of one of them, the SE_66, as a growth dependency in M0-7e cells via the modulation of *CSF2RB* expression. Simulations of functional screening data demonstrated that the outcomes of each method varied based on factors such as the count of sh/sgRNAs designed for each gene, their targeting efficiency, and the multiplicity of efficiency of the targeting elements. Of note, the directional scoring method displays the overall worst performances as expected by its the rather heuristic nature and we recommend usage of the other analysis methods that are more statistically sound. By offering the possibility of combining these multiple screening analysis tools into one platform, PitViper represents a beneficial resource for minimizing the trade-off between sensitivity and specificity and maximizing the chances of identifying all possible truly positive screened hits.

PitViper is easy to parametrize and run, supporting most common input formats (FASTQ files, BAM files, and counts matrix), making it well suited for analysis of *de novo* or already-published data. Too often, the re-analysis of bioinformatics results requires reconstruction and deduction on the part of those trying to reproduce the results. However, it becomes fully reproducible when researchers generating the results provide a complete and continuous framework of their analysis pipeline. To promote reproducibility, we have ensured that these considerations are applied in PitViper without requiring additional work for data and screening analysis pipeline sharing. Genomic and epigenomic data such as shRNA screens targeting genes or CRISPR screens targeting custom regions are supported. Actionable and shareable reports with interactive publication-ready visualizations made to improve results annotations are generated. Embedding of raw and normalized counts in the report facilitates further reproducibility, access to data and encourage transparency. External programs such as EnrichR for pathway enrichment, GeneMania for network reconstruction or DepMap for dependency identification are available in the report and allow quick access to relevant biological information in the data. From a computational standpoint, the Jupyter Notebook report template allows developers to easily add new modules or modify reports to suit their specific needs. Installation of computational dependencies is possible through Conda or Docker to facilitate reproducibility and is complementary to the ability to rerun a previous PitViper analysis with the same exact parameters using a configuration file. We expect PitViper to offer a rapid, easy to use, scalable and reproducible solution for screen essentiality analysis and to improve the detection of essential elements by offering a broader range of computational methods adapted to multiple screening workflows.

Data availability

M07e H3K27Ac and H3K4me3 ChIP-seq are available at ArrayExpress database (E-MTAB-4367). M07e ATAC-seq data is available in the Gene Expression Omnibus database under accession code GSE131462.

Pitviper code is available at https://github.com/lobrylab/PitViper and https://doi.org/10.5281/zenodo.11128707.

All correspondence, material and code request should be addressed to either Dr Camille Lobry (camille.lobry@inserm.fr) or Paul-Arthur Meslin (paul-arthur.meslin@inserm.fr).

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We would like to thank B. Jost, C. Thibault-Charpentier, M. Cerciat and all the member of the GenomeEast sequencing facility for their advice and expert sequencing. We would like to thanks N. Setterblad, C Doliger, S. Duchez, A. Alberdi, V. Parietti and all members of the 'Plateforme Saint Louis' for expert sorting and sequencing.

Author contributions: P.A.: conceptualization, formal analysis, investigation, visualization, methodology, writing-original draft. L.M.K.: formal analysis, investigation, visualization. S.B.: formal analysis, investigation, visualization. S.L.: formal analysis, investigation, visualization. K.H.L.: resources. J.C.R.: resources. C.F.B.: resources. R.I.: conceptualization. K.C.W.: resources, conceptualization. A.P.: conceptualization, formal analysis, investigation, writing-original draft, funding acquisition, supervision. C.L.: conceptualization, formal analysis, investigation, methodology, writing-original draft, funding acquisition, supervision.

Funding

Institut National du Cancer [PLBIO-2020-074 to C.L., PLBIO-2020-246 to A.P. and C.L.]; ERC Starting program [758848 to A.P.] and ERC Consolidator Program [101088563 to A.P.]; C.L. is supported by The ARC Foundation. A.P. is supported by the St Louis Association and is an FSER laureate.

Conflict of interest statement

K.C.W. is a co-founder, consultant, and equity holder at Tavros Therapeutics and Celldom, is a consultant and equity holder at Simple Therapeutics and Decrypt Biomedicine, and has performed consulting work for Guidepoint Global, Bantam Pharmaceuticals, and Apple Tree Partners. C.L. is a scientific board member of Adlin Science. The remaining authors declare no competing interests.

References

1. Garneau,J.E., Dupuis,M.È., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadán,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.
2. Paddison,P.J., Caudy,A.A., Bernstein,E., Hannon,G.J. and Conklin,D.S. (2002) Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev.*, **16**, 948–958.
3. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816.
4. Larson,M.H., Gilbert,L.A., Wang,X., Lim,W.A., Weissman,J.S. and Qi,L.S. (2013) CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.*, **8**, 2180–2196.
5. Gilbert,L.A., Horlbeck,M.A., Adamson,B., Villalta,J.E., Chen,Y., Whitehead,E.H., Guimaraes,C., Panning,B., Ploegh,H.L., Bassik,M.C., *et al.* (2014) Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, **159**, 647–661.
6. Li,W., Xu,H., Xiao,T., Cong,L., Love,M.I., Zhang,F., Irizarry,R.A., Liu,J.S., Brown,M. and Liu,X.S. (2014) MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.*, **15**, 554.
7. Li,W., Köster,J., Xu,H., Chen,C.H., Xiao,T., Liu,J.S., Brown,M. and Liu,X.S. (2015) Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol.*, **16**, 281.
8. Kim,E. and Hart,T. (2021) Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier. *Genome Med.*, **13**, 2.
9. Daley,T.P., Lin,Z., Lin,X., Liu,Y., Wong,W.H. and Qi,L.S. (2018) CRISPhieRmix: a hierarchical mixture model for CRISPR pooled screens. *Genome Biol.*, **19**, 159.
10. Lin,K.H., Xie,A., Rutter,J.C., Ahn,Y.-r., Lloyd-Cowden,J.M., Nichols,A.G., Soderquist,R.S., Koves,T.R., Muoio,D.M., MacIver,N.J., *et al.* (2019) Systematic dissection of the metabolic-apoptotic interface in AML reveals heme biosynthesis to Be a regulator of drug sensitivity. *Cell Metab.*, **29**, 1217–1231.
11. Benbarche,S., Lopez,C.K., Salataj,E., Aid,Z., Thirant,C., Laiguillon,M.-C., Lecourt,S., Belloucif,Y., Vaganay,C., Antonini,M., *et al.* (2022) Screening of ETO2-GLIS2–induced Super Enhancers identifies targetable cooperative dependencies in acute megakaryoblastic leukemia. *Sci. Adv.*, **8**, eabg9455.
12. Avanzi,G.C., Lista,P., Giovinazzo,B., Miniero,R., Saglio,G., Benetton,G., Coda,R., Cattoretti,G. and Pegoraro,L. (1988) Selective growth response to IL-3 of a human leukaemic cell line with megakaryoblastic features. *Br. J. Haematol.*, **69**, 359–366.
13. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
14. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
15. Korotkevich,G., Sukhov,V., Budin,N., Shpak,B., Artyomov,M.N. and Sergushichev,A. (2021) Fast gene set enrichment analysis. bioRxiv doi: https://doi.org/10.1101/060012, 01 February 2021, preprint: not peer reviewed.
16. VanderPlas,J., Granger,B., Heer,J., Moritz,D., Wongsuphasawat,K., Satyanarayan,A., Lees,E., Timofeev,I., Welsh,B. and Sievert,S. (2018) Altair: interactive statistical visualizations for Python. *J. Open Source Softw.*, **3**, 1057.
17. Kluyver,T., Ragan-Kelley,B., Pérez,F., Granger,B., Bussonnier,M., Frederic,J., Kelley,K., Hamrick,J., Grout,J., Corlay,S., *et al.* (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas* . IOS Press, pp. 87–90.

18. Bodapati,S., Daley,T.P., Lin,X., Zou,J. and Qi,L.S. (2020) A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome Biol.*, **21**, 62.
19. Köster,J., Mölder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., *et al.* (2021) Sustainable data analysis with Snakemake. *F1000Research*, **10**, 33.
20. Chen,E.Y., Tan,C.M., Kou,Y., Duan,Q., Wang,Z., Meirelles,G.V., Clark,N.R. and Ma'ayan,A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.*, **14**, 128.
21. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A., *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
22. Xie,Z., Bailey,A., Kuleshov,M.V., Clarke,D.J.B., Evangelista,J.E., Jenkins,S.L., Lachmann,A., Wojciechowicz,M.L., Kropiwnicki,E., Jagodnik,K.M., *et al.* (2021) Gene set knowledge discovery with Enrichr. *Curr. Protoc.*, **1**, e90.
23. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T., *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
24. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G.V., Sonkin,D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
25. Stransky,N., Ghandi,M., Kryukov,G.V., Garraway,L.A., Lehár,J., Liu,M., Sonkin,D., Kauffmann,A., Venkatesan,K., Edelman,E.J., *et al.* (2015) Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, **528**, 84–87.
26. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
27. Kolde,R., Laur,S., Adler,P. and Vilo,J. (2012) Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, **28**, 573–580.
28. Killian,T. and Gatto,L. (2021) Exploiting the DepMap cancer dependency data using the depmap R package. *F1000Research*, **10**, 416.

Received: December 8, 2023. **Revised:** March 19, 2024. **Editorial Decision:** May 6, 2024. **Accepted:** May 10, 2024