# Beta-binomial modeling of CRISPR pooled screen data identifies target genes with greater sensitivity and fewer false negatives

Hyun-Hwan Jeong, Seon Young Kim, Maxime W.C. Rousseaux, et al.

*Genome Res.* published online April 23, 2019
Access the most recent version at doi:10.1101/gr.245571.118

# Beta-binomial modeling of CRISPR pooled screen data identifies target genes with greater sensitivity and fewer false negatives

Hyun-Hwan Jeong[1,2], Seon Young Kim[1,2], Maxime W. C. Rousseaux[1,2,+], Huda Y. Zoghbi[1,2,3,4], Zhandong Liu[2,3*]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA

[2]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, Texas, USA

[3]Department of Pediatrics, Baylor College of Medicine, Houston, Texas, USA

[4]Howard Hughes Medical Institute, Houston, Texas, USA

[+]Present Address: University of Ottawa Brain and Mind Research Institute and Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada

[*]Correspondence should be addressed to: zhandong.liu@bcm.edu

## Abstract

The simplicity and cost-effectiveness of CRISPR technology have made high-throughput pooled screening approaches accessible to virtually any lab. Analyzing the large sequencing data derived from these studies, however, still demands considerable bioinformatics expertise. Various methods have been developed to lessen this requirement, but there are still three tasks for accurate CRISPR screen analysis that involve bioinformatic know-how if not prowess: designing a proper statistical hypothesis test for robust target identification, developing an accurate mapping algorithm to quantify sgRNA levels, and minimizing the parameters necessary that need to be fine-tuned. To make CRISPR screen analysis more reliable as well as more readily accessible, we have developed a new algorithm, called CRISPRBetaBinomial or $CB^2$ (https://CRAN.R-project.org/package=CB2). Based on the beta-binomial distribution, which is better suited to sgRNA data, $CB^2$ outperforms the eight most commonly used methods (HiTSelect, MAGeCK, PBNPA, PinAPL-Py, RIGER, RSA, ScreenBEAM, and sgRSEA) in both accurately quantifying sgRNAs and identifying target genes, with greater sensitivity and a much lower false discovery rate. It also accommodates staggered sgRNA sequences. In conjunction with CRISPRcloud, $CB^2$ will bring CRISPR screen analysis within reach for a wider community of researchers.

# Introduction

Genetic screens have become a favored tool for gathering information about disease pathogenesis and cellular biology. Initially, these screens were performed using chemical mutagenesis or RNA inference (RNAi), which are effective but laborious processes (Mohr et al. 2014; Park et al. 2013; Schlabach et al. 2008; Silva et al. 2008; Simon et al. 2015). Larger-scale, pooled approaches were finally made feasible by the advent of microarray (DeJesus et al. 2016; Gilbert et al. 2014; Luo et al. 2009; Shalem et al. 2014; Wang et al. 2014). Pooled shRNA (short-hairpin RNA) libraries can be barcoded and packaged into viruses, which are used to infect a population of cells that are then selected for a desired phenotype (e.g., growth or fluorescence). Hybridizing microarray soon followed for hit identification (Paddison et al. 2004). In later iterations of this approach, next-generation sequencing (NGS) was used to identify hits (Hu and Luo 2012).

The development and optimization of CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein 9) systems have propelled pooled screen approaches into even wider use. Besides their relatively simplicity and low cost, the robustness of hit identification has reduced the requirement for redundancy in the number of targeting single-guide RNAs (sgRNAs), which allows the same size library to be more diverse (Sanjana et al. 2014; Xu et al. 2015). Moreover, these pooled libraries have been made accessible through repositories such as Addgene (https://www.addgene.org/). CRISPR/Cas9 pooled screens are thus within the technical reach of most biomedical researchers (Doench 2017). For all this accessibility on the experimental side, however, the resulting bioinformatics datasets are enormous and complex. The analysis of thousands of genetic perturbations demands considerable bioinformatics expertise.

In our experience, most users are unaware of whether their tool of choice models the data according to Poisson, negative binomial, or even Gaussian distribution, nor what the relative strengths and weaknesses of these models are. The current roster of tools bear the imprint of the history of RNA-seq data analysis: RNA-seq data were initially modeled using Poisson distributions (Marioni et al. 2008), which is a natural choice for simple read counts. Poisson distribution assumes that the mean and variance are equal, however, and with biological data, the variance is often greater than the mean. Analytic methods therefore turned to negative binomial distribution, which can handle overdispersed data (Anders and Huber 2010; Love et al. 2014). A number of popular tools still use negative binomial distribution to analyze sgRNA screen data (Li et al. 2014; Spahn et al. 2017), even though the structure of the sgRNA screen data is very different from that of RNA-seq. In the latter, there is huge variation in transcript lengths, from 60bp to 2.4Mbp, but all sgRNAs for any given gene are designed to have the same length. This often leads to the variance being less than the mean (Supplemental Figure 1). We hypothesized that a beta-binomial model,

in which the variance can be either larger or smaller than the mean, would better fit the data and more accurately identify changes in sgRNA.

We therefore developed CB$^2$, a new web-based algorithm that uses beta-binomial distribution, and compared its performance with that of the eight most commonly used algorithms (HiTSelect (Diaz et al. 2015), MAGeCK (Li et al. 2014), PBNPA (Jia et al. 2017), PinAPL-Py (Spahn et al. 2017), RIGER (Luo et al. 2009), RSA (König et al. 2007), ScreenBEAM (Yu et al. 2016), and sgRSEA (Noh 2015)), which encompass both parametric and nonparametric approaches (Table 1). We applied all these methods to ten different biological datasets, taken from fields ranging from cancer to basic cell biology (Koike-Yusa et al. 2014; Parnas et al. 2015; Evers et al. 2016; Golden et al. 2017; Li et al. 2018; Sanson et al. 2018).

# Results

## CB$^2$ is more sensitive in target gene identification than existing methods

Identifying candidates by a statistical hypothesis test is a key component in any screen analysis. In CB$^2$, we adapted a beta-binomial model (Baggerly et al. 2003) with a modified Student's *t*-test to measure differences in single-guide RNA (sgRNA) levels, followed by Fisher's combined probability test (Fisher 1925) to estimate the gene-level significance. We chose Fisher's method for two reasons: first, to keep the entire pipeline parametric, and second, to keep CB$^2$ as fast as possible (Robust Rank Aggregation [RRA] requires permuting the data—a non-parametric feature—so it runs slower than Fisher's method). Furthermore, when we compared Fisher's method against RRA, we found that RRA is not effective in combining the p-values estimated by beta-binomial distribution (Supplemental Figure 2).

We compared CB$^2$ with eight state-of-the-art methods on three benchmark datasets evaluating gene essentiality (Evers et al. 2016) using different technologies: CRISPRn (CRISPR nuclease gene knockout via Cas9) and CRISPRi (CRISPRinterference, a CRISPR/Cas9 system with a catalytically inactive Cas9 fused to the transcriptional repressor KRAB which results in gene repression). These benchmark datasets (CRISPRn-RT112, CRISPRn-UMUC3, and CRISPRi-RT112) were constructed based on 46 genes that are essential for cell survival and 47 genes that are non-essential. We first tested whether these methods could easily distinguish between essential and non-essential genes. We found that each method clearly discriminates essentiality by their gene rankings (Supplemental Figure 3A). In addition, gene rankings obtained from each method, except PBNPA for the CRISPRi-RT112 dataset, are highly correlated (Supplemental Figure 3B , $R^2$ is $[0.86,0.98]$ for CRISPRn-RT112, $[0.85,0.98]$ for CRISPRn-UMUC3, and $[0.72,0.96]$ for CRISPRi-RT112). CB$^2$, ScreenBEAM, and MAGeCK produced very similar gene rankings across all the benchmark datasets. We also compared the Precision-Recall (PR) and Receiver

operating characteristic (ROC) curves across the methods and calculated the Area under the curve (AUC) of each. $CB^2$ recorded the best PR-AUC and ROC-AUC scores for both CRISPRn screen datasets (Supplemental Figure 4 & 5). HiTSelect had the best PR-AUC and ROC-AUC scores for the CRISPRi-RT112 dataset for which $CB^2$ achieved comparable scores (Supplemental Figure 6). Although the gene ranking is similar among these methods (Supplemental Figure 3), the estimated p-values and FDRs are very different. These results highlight the importance of using FDR to guide the gene selection process.

While several CRISPR screens (Aguirre et al. 2016; Parnas et al. 2015; Zhou et al. 2014) have prioritized candidate hits by ranking, they do not provide statistical estimates of error rates. These methods, therefore, rely on an arbitrary rule to select the top candidate genes and are prone to biased selections and high hit attrition rates. One solution is candidate selection by a quantitative statistical measure such as a p-value or a false positive rate (FDR) cut-off. To assess the detection powers of FDR of established CRISPR screen analysis methods and $CB^2$, we measured the F1-score ($2 \cdot \frac{precision \cdot recall}{precision + recall}$) of each method, i.e., the harmonic average of the precision and the recall, for FDR thresholds ranging from 10% to 0.01%. $CB^2$ outperformed all other methods at every FDR cut-off level, and all other methods lost their detection powers at more rigorous FDRs (Figure 1 A&B and Supplemental Figure 7). In other words, all methods demonstrated a small type-I error due to the strong lethality phenotype of the CRISPR assay, but $CB^2$ demonstrated a significantly lower type-II error than the other methods (Supplemental Figure 8). Across all paradigms tested with different FDR cut-offs, $CB^2$ performed the best, with a much larger F1-score and recall. Thus, $CB^2$ is both sensitive and specific in selecting candidate genes.

To understand the differences produced by these methods, we next tested a prototypical essential gene, *RPL5*, to compare the gene-level enrichment across datasets and analytical tools. In the first CRISPR screen on an RT112 cell line, we expected to see the depletion of sgRNAs targeting *RPL5* in group $T_1$. Out of the 10 sgRNAs that target this gene in the CRISPRn-RT112 dataset, six showed a strong decrease in the group $T_1$. $CB^2$ estimated an FDR of $2.07 \times 10^{-19}$, while only three other methods (HiTSelect, ScreenBEAM, and sgRSEA) estimated FDR to be $< 0.01$ (Figure 2A). Next, we looked at the same gene in the UMUC3 cell line. Five of the 10 sgRNAs targeting *RPL5* decreased in group $T_1$, and all five sgRNAs were among those identified in RT112 cell line. $CB^2$ estimated an FDR of $3.81 \times 10^{-10}$ for *RPL5*, while none of the other methods identified it to be significantly depleted with an FDR cutoff below 0.01 (Figure 2B). Lastly, in the CRISPRi-RT112 dataset, three of seven sgRNAs indicated depletions, but only $CB^2$ was able to estimate the FDR of $2.78 \times 10^{-8}$, and the other methods did not count *RPL5* as a hit in the dataset (Figure 2C). Overall, $CB^2$ produced more reliable hit identification than other methods based on statistical cut-offs for the gene tested (*COPS8* and *RPL27*, Supplemental Figure 12 & 13).

We performed the same analysis on two distinct datasets (Sanson et al. (2018)) to determine how $CB^2$ performs compared to other methods for genome-wide screening analysis. Sanson et al. used novel optimized libraries for genome-wide CRISPRn (Brunello), CRISPRi (Dolcetto), and CRISPRa (Calabrese) screening and showed these libraries outperform other previously established libraries, such as GeCKO (Sanjana et al. 2014) and hCRISPRi-v2 (Horlbeck et al. 2016). For our analysis, we chose two screens for benchmarking, both datasets from a dropout screen in A375 melanoma cells: one used the Brunello CRISPRn library with tracr-v2 tracrRNA (CRISPRn-A375), the other used the Dolcetto CRISPRi Set A library (CRISPRi-A375). Each dataset contains a control sample (plasmid DNA) and three biological replicates. We used the gold-standard gene sets of 1,580 essential and 927 non-essential genes reported in (Hart et al. 2014, 2015) to assess the performance of the methods. (We excluded PinAPL-Py from benchmarking since it does not report statistics when the input contains only one control sample.) $CB^2$ outperformed other methods at the stringent FDR cutoff level (Figure 2 and Supplemental Figure 9). $CB^2$ outperformed all other methods in F-1, precision and recall measures at the stringent FDR cut-offs on (Sanson et al. 2018)'s A375 genome-wide screen datasets. F1-score (top), precision (middle), and recall (bottom) for each method on two benchmark datasets are presented as a function of FDR cut-off values (Figure 1 C & D and Supplemental Figure 9), and provided higher AUC values of PR and ROC curves than the other methods (Supplemental Figure 10 & 11).

These results indicate that $CB^2$ more accurately estimates the gene-level FDR. The use of FDR in selecting hits is critical in real data analysis since the arbitrary selection of top genes is purely heuristic. $CB^2$ is better at identifying true hits based on multiple concordant sgRNAs targeting the same gene.

## $CB^2$ is more sensitive in target gene detection than existing methods

To test the idea that a beta-binomial model would better fit the data and more accurately identify changes in sgRNA, we compared the sgRNA level statistics on several CRISPR pooled libraries containing non-targeting sgRNAs as negative controls. Non-targeting sgRNAs are not supposed to show any differential enrichment and can be used to assess the quality of the method. (Parnas et al. 2015) used Mouse CRISPR Knockout Pooled Library for their genome-wide screen (GeCKO v2, Addgene #1000000052, #1000000053), which contains 1,000 non-targeting sgRNAs. We therefore used this dataset to measure the specificity with which $CB^2$ and MAGeCK detect true negatives. We compared the unadjusted p-values for sgRNAs since the FDR is controlled at the gene level.

$CB^2$ showed greater specificity (the proportion of actual negatives that are correctly identified) than MAGeCK across a wide range of p-value thresholds. At a p-value threshold of 0.01, $CB^2$ had a specificity of 86% vs. MAGeCK's 68% (Figure 3A). Next, we plotted the log-fold change against the p-value levels

in a volcano plot. The majority of the negative control sgRNAs were correctly detected by $CB^2$, whereas MAGeCK demonstrated a one-side long tail for positively changed sgRNAs producing inflated p-values for a group of negative controls (Figure 3B). Many of these false positives showed extremely low p-values (ranging from $10^{-5}$ to $10^{-40}$), indicating a strong selection bias. To understand the impact of this selection bias, we analyzed the rest of the sgRNA library with both methods. At the same threshold ($p < 0.01$), $CB^2$ selected 12,648 sgRNAs while MAGeCK selected 31,381 sgRNAs; 2,971 sgRNAs were identified by both methods (Figure 4B). We applied the same analysis to the CRISPRn-A375 dataset (Sanson et al. 2018) which contains 1,000 non-targeting sgRNAs. $CB^2$ shows higher specificity than MAGeCK, except when setting a p-value cut-off at 0.2. Similar p-value distributions shown in Figure 3 were also found for this dataset (Supplemental Figure 14).

We plotted sgRNAs unique to each method on a heatmap, which showed a high concordance within each experimental group for sgRNAs unique to $CB^2$ (Figure 4A). In contrast, sgRNAs identified by MAGeCK showed a much noisier pattern, and samples from the same experimental group could not be clustered together based on these differentially enriched sgRNAs (Figure 4A).

Next, we performed the same sgRNA-level comparisons on two additional datasets. In the first study, a differentiation screen was conducted to identify target genes that maintain naive pluripotency (Li et al. 2018) using the Mouse Improved Genome-wide Knockout CRISPR Library v2 (Addgene #67988). The library contains 91,319 sgRNAs targeting 18,542 mouse genes. To identify differentially enriched sgRNAs, we kept the same threshold ($p < 0.01$) for both $CB^2$ and MAGeCK. $CB^2$ identified 732 sgRNAs while MAGeCK identified 5,105 sgRNAs (395 sgRNAs shared between the two) (Figure 4C and 4D), and we observed the same trend as in Parnas et al.'s screening dataset (Parnas et al. 2015) (Figure 4A and B). We found the same trend on the Evers et al.'s datasets (Evers et al. 2016) (Supplemental Figure 15). Thus, $CB^2$'s accurate sgRNA-level statistics are attributable to its use of the beta-binomial model.

## $CB^2$ provides more accurate alignment without parameter tuning

Many CRISPR pooled screens use in-house scripts to quantify sgRNA abundance (Gilbert et al. 2014; Golden et al. 2017; Iorio et al. 2018; Li et al. 2018) or other alignment algorithms for RNA-seq (DeJesus et al. 2016; Hart et al. 2015; Parnas et al. 2015; Sanjana et al. 2014). These codes are often not shared publicly and are not easily reusable. Both MAGeCK and PinAPL-Py provide an integrated mapping function, but PinAPL-Py requires complex parameter tuning and MAGeCK samples only the first million reads to estimate the location of sgRNAs in FASTQ files. Furthermore, there is no systematic comparison of mapping accuracy in the literature, and users lack reliable guidelines for selecting mapping tools. We,

therefore, introduced an adaptive hash-mapping algorithm into $CB^2$ and tested all three methods on six published datasets (Supplemental Table 1).

$CB^2$ showed consistently greater mapping accuracy than MAGeCK and PinAPL-Py (Figure 5A). To understand why, we studied the reads that are mapped by $CB^2$ but not MAGeCK or PinAPL-Py in the CRISPRn-RT112 dataset (Evers et al. 2016). MAGeCK mapped 64% of the reads, compared to 75% by $CB^2$ (Figure 5A). This is primarily due to the fact that MAGeCK estimates the trimming windows using the first $N$ reads from the input ($N$ is 100,000 by default). There is no guarantee that these windows are optimal for the rest of the input files. If a sgRNA locates outside of the precomputed windows, MAGeCK will fail to detect it (Figure 5B). PinAPL-Py does not precompute sgRNA locations based on a subset of reads but uses Cutadapt (Martin 2011) for flexible trimming followed by the Bowtie 2-based alignment (Langmead and Salzberg 2012). We found that PinAPL-Py failed to identify some of the sgRNAs because reads fail to align due to the incorrect trimming from Cutadapt even under several different tuning parameters (Figure 5B). This is likely because of the frequent indels that occurred in the 5' adapter sequence region of the reads (see Figure 5B)—usually, the quad-nucleotide sequence 'CACC', which is part of the U6 promoter, precedes the sgRNA sequence. In contrast, the 'GTTT' sequence, which is the first four nucleotides of the sgRNA scaffold sequence, was present in all the reads. Given the fidelity of the GTTT sequence, the sequences mapped by CB2 but missed by other algorithms are likely accurate and not false positives . $CB^2$ is currently the only CRISPR/Cas9 online screen analysis tool with parameter-free mapping and high accuracy in sgRNA quantification.

## $CB^2$ is accessible, secure, and does not require a steep learning curve with CRISPRcloud

We had previously developed a web-based application called CRISPRcloud that could run any statistical testing and mapping algorithm through the cloud-based infrastructure provided by Amazon Web Service (Jeong et al. 2017). We implemented $CB^2$ in the platform and added new features to increase speed and data security (Supplemental Figure 16). CRISPRcloud is compatible with most modern web browsers (Google Chrome version 69 and later, Apple Safari version 11 and later, and Mozilla Firefox 48.0 and later) and operating systems (iOS, Windows, and Linux). Our fast client-side sgRNA mapping program reduces input files of several gigabytes into a single megabyte-sized file. By transferring a much smaller file through the Internet, CRISPRcloud decreases transfer time and prevents the sharing of raw input files, thereby eliminating downloading errors and data privacy issues in one step. Our adaptive mapping algorithm, via Angular (https://angular.io/) and TypeScript (https://www.typescriptlang.org), provides an open-source front-end web application platform. The enormous computing power needed to perform

these operations mean that platforms built with a centralized server solution will have load-balancing problems when many users submit their requests simultaneously, leading to the longer user waiting times and raising the risk of system-wide failure. $CB^2$ therefore provides a decentralized, cloud-computing-based, scalable service through a combination of AWS infrastructure that includes Amazon Elastic Compute Cloud (EC2) (https://aws.amazon.com/ec2/), Amazon Simple Storage Service (S3) (https://aws.amazon.com/s3/) and Amazon Simple Queue Service (SQS) (https://aws.amazon.com/sqs/). With this infrastructure, we launched a web service of CRISPRcloud. CRISPRcloud enables researchers with no programming background to pre-process, check the quality, statistically analyze, query, and visualize their CRISPR/Cas9 pooled screening data (Supplemental Table 2).

# Discussion

The number of datasets for CRISPR/Cas9 screens in Gene Expression Omnibus have more than tripled each of the past three years (39 datasets in 2015, 121 datasets in 2016, and 408 datasets in 2017). This expansion has outpaced the development of methods for analyzing the data, most of which employ statistical models that are better suited to RNA-seq than to sgRNA data. Here we took into account the difference between the two types of data to develop a new algorithm, $CB^2$, and show that it is more sensitive, specific, and selective than eight other leading tools.

We focused first on the central task for any analytic tool being applied to sgRNA data: statistical hypothesis testing to identify target genes accurately from the screening data. Several methods that facilitate analysis of RNAi pooled screening data (Dutta et al. 2016; König et al. 2007; Luo et al. 2008; Shao et al. 2013) are not compatible with CRISPR/Cas9 pooled screening data because of differences in effect size, sequence determinants, and on- vs. off-target effects (Li et al. 2014). MAGeCK was the first tool specifically developed to analyze CRISPR/Cas9 pooled screening data, and it combines a negative-binomial distribution model with a modified robust ranking aggregation (RRA) algorithm (Li et al. 2014). Subsequent methods (Table 1) employed different strategies to improve the accuracy of data analysis, but to our knowledge there has never been a thoroughgoing attempt to benchmark these methods and determine which performs best with sgRNA data. Our choice of the beta-binomial distribution, which is not the approach used by any of these analytic tools, was justified by both the theoretical and empirical considerations and proved able to provide far fewer false positives than these other methods at comparable FDR thresholds.

$CB^2$ also addressed another difficult task: quantifying sgRNA from next-generation sequencing data. Except for the quantification algorithm provided by MAGeCK, most studies use in-house algorithms or

extend established methods that were optimized for RNA-seq. CB$^2$ proved capable of fast and accurate alignment, with the ability to handle indels.

Last but not least is the challenge of making powerful tools readily accessible to the research community. Of the existing tools, PinAPL-Py (Spahn et al. 2017) and CRISPRcloud (Jeong et al. 2017) are the only two that support a graphical web interface and require no additional program installation. These programs are an important first step toward enabling the scientists who are actually generating the CRISPR/Cas9 screen data to analyze their large datasets. They still have limitations, however: for PinAPL-Py, users still need to provide the adaptor sequence to be trimmed, the error tolerance rate, and the quality threshold for trimmed reads. CRISPRcloud is the only framework into which the user can plug in any statistical tools or mapping algorithms but transferring a large amount of sequence data over the internet has inherent disadvantages such as long transfer times, vulnerability to file copying errors, and possible data security breaches. Taking advantage of the CRISPRcloud framework, CB$^2$ is fully web-based and designed to require only the minimal number of parameters for data analysis, since fewer parameters means shorter learning curves for the majority of users. CB$^2$'s power and accessibility will enable more labs to extract biologically relevant discoveries from CRISPR pooled screens.

# Methods and Materials

## Statistical hypothesis testing using beta-binomial distribution for sgRNA-level differential analysis

We adapted a beta-binomial model proposed for Serial Analysis of Gene Expression (SAGE) by (Baggerly et al. 2003). Specifically, let $p_i$ be the true proportion of an sgRNA in sample $i$. We assume the value of $p_i$ can vary from sample to sample and follows a beta distribution, $p_i \sim Beta(\alpha,\ \beta)$. Let $X_i$ denote the number of read counts for a sgRNA in the $i^{\text{th}}$ sample. We assume $X_i$ follows a binomial distribution, $X_i|p_i \sim Binomial(n_i, p_i)$, where $n_i$ is the total number of mapped reads in sample $i$. To combine the estimated $\hat{p_i}$ across multiple samples of the same treatment group, we proposed a linear model $p^A = \sum w_i p_i$, where *i* is the index for samples and *w* is the weight vector for samples in group A. Baggerly et al. (2003) proved that as long as $w^T 1 = 1$, the expectation of $E(p^A)$ is unbiased. The value of *w* is estimated through gradient descent methods by minimizing the variance on $p^A$. Baggerly et al. (2003) showed that $w_i \propto \left[\frac{1}{\alpha+\beta} + \frac{1}{n_i}\right]^{-1}$.

CB$^2$ performs the sgRNA-level differential analysis between two groups using a Student *t*-test like statistic (Baggerly et al. 2003):

$$t = \frac{p_B - p_A}{\sqrt{V_B + V_A}},$$

where $p_A$ and $p_B$ are the proportions of sgRNA, and $V_A$ and $V_B$ are the group variances of sgRNA, for groups A and B, respectively. Test statistic $t$ represents the strength of the difference of sgRNA abundance between groups $A$ and $B$. In other words, a large positive $t$-value indicates that the quantity of sgRNA in group $B$ is greater than in group $A$, and a large negative $t$-value indicates that the quantity of sgRNA in group $B$ is less than that in group $A$.

The variance is estimated by

$$\hat{V} = \max\left[\frac{\sum w_i^2 \hat{p}_i^2 - (\sum w_i^2)\,\hat{p}^2}{1 - (\sum w_i^2)}, \frac{\frac{\sum X_i}{\sum n_i}\left(1 - \frac{\sum X_i}{\sum n_i}\right)}{\sum n_i}\right].$$

To measure the statistical significance of the difference, we approximate the $p$-value of a given $t$ in a Student's $t$-distribution with a degree of freedom (df) defined by

$$df = \frac{(V_A + V_B)^2}{\frac{V_A^2}{n_A - 1} + \frac{V_B^2}{n_B - 1}},$$

where $n_A$ and $n_B$ are the numbers of replicates in groups $A$ and $B$.

## sgRNA p-value aggregation for gene-level statistics

Because multiple significant sgRNAs targeting the same gene hold greater biological significance than a single significant sgRNA, we must aggregate $p$-values to increase confidence in target identification. To do so, we combine p-values of sgRNAs for a target gene using Fisher's method (Fisher 1925) to assess overall differences at the gene level. The combined chi-square statistical test is used:

$$\chi_{2k}^2 \sim -2\sum_{j=1}^{k} ln(p_j),$$

where $k$ is the number of sgRNAs targeting a gene in the screen and $p_j$ is the $p$-value of $j$-th sgRNA for the gene. $\chi^2$ follows a chi-squared distribution with $2k$ degrees of freedom. To correct for multiple hypothesis testing, we adapted the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) to estimate the False Discovery Rate (FDR).

## Gene-level statistics benchmarking on existing methods

We used three different CRISPRn/CRISPRi pooled screen datasets (Evers et al. 2016) (RT112 and UMUC3 cell line screens with CRISPR; RT112 cell line screen with CRISPRi) and two different genome-wide CRISPRn/CRISPRi pooled screen datasets (Sanson et al. 2018) (A375 cell line screens), which provide ground-truth labels of essentiality for each gene. With those screening datasets and labels, we benchmarked the accuracy of essential gene detection by $CB^2$ with eight other published methods (HiTSelect, MAGeCK, PBNPA, PinAPL-Py, RIGER, RSA, ScreenBEAM, and sgRSEA). We computed the False Discovery Rate (FDR) for each gene from each method in the benchmark and set five different levels of FDR cut-off (0.1, 0.05, 0.01, 0.005, 0.001) for essential gene classification. For example, if we set FDR cut-off to 0.1, then a gene is predicted to be essential in the cell line if the FDR of the gene falls below the cut-off value. We calculated recall (a recall value close to 1 indicates a prediction with a low false negative rate), precision (a precision value close to 1 indicates a prediction with a low false positive rate), and F-measure (the harmonic mean of precision and recall) of all the methods at each FDR level. To allow each method to archive its best performance, we tuned parameters for the five parameter-tunable methods – MAGeCK (permutation parameter for RRA test), PBNPA (no.sim parameter), RIGER (alpha parameter), sgRSEA (multiplier parameter) and ScreenBEAM (burnin parameter) on CRISPRn-RT112 dataset. The F-measure was used as a measure for the parameter tuning. Most of the methods showed robust performance regardless of the varied parameters, except sgRSEA and RIGER (Supplemental Figure 17). Therefore, we used the default parameter for MAGeCK, PBNPA, and ScreenBEAM for other datasets and used an optimized parameter for sgRSEA and RIGER. We also calculated the AUC (Area Under the Curve) of the Precision and Recall curves and Receiver Operating Characteristic curves of all the methods with FDR values. All of the data and scripts for the benchmarking are available at https://github.com/hyunhwaj/CB2-Experiments. In addition, the data and the scripts can be found in the Supplemental Material. Parameters used in these experiments are described below.

### $CB^2$

$CB^2$ R package (https://CRAN.R-project.org/package=CB2) were used in the benchmarking (R Core Team 2019). Benchmarking of $CB^2$ was performed without parameter tuning since $CB^2$ is parameter-free. FDR values for negative changes between two different groups from $CB^2$ statistical analysis were used for benchmarking.

### HiTSelect

We ran HiTSelect MATLAB package (https://github.com/diazlab/HiTSelect). Normalization by Sequencing depth option was selected for benchmarking.

**MAGeCK**

MAGeCK version 0.5.8 was used for benchmarking. We ran MAGeCK with the 'mageck test' command with the following parameters: '--norm-method' 'median' and '--adjust-method' 'fdr'. We performed 100 permutations for the modified robust ranking aggregation (RRA) algorithm to estimate the gene-level statistics on the benchmark datasets.

**ScreenBEAM**

ScreenBEAM R package (version 1.0.0, https://github.com/jyyu/ScreenBEAM) was used for benchmarking 'data.type' parameter was set as 'NGS,' and 'do.normalization' was set as TRUE, 'nitt' and 'burnin' parameters for Bayesian computing were set at 15000 and 5000. ScreenBEAM does not provide the one-sided $p$-value for negative selection, so for the FDR comparison with other methods, we changed the FDR of a gene to 1 if the $\beta$ of the gene is greater than $0$.

**PinAPL-Py**

We used the PinAPL-Py website (http://pinapl-py.ucsd.edu) to perform the benchmarking. For the sgRNA read counting, we used 'GGCTTTATATATCTTGTGGAAAGGACGAAACACCG, GCTTTATATATCTTGTGGAAAGGACGAAACACCG,' and 'CTTTATATATCTTGTGGAAAGGACGAAACACCG,' were used for 'seq_5_end' parameters of 'CRISPRn-RT112', 'CRISPRn-UMUC3', and 'CRISPRi-RT112' datasets. We used CPM normalization and set the GeneMetric parameter as 'aRRA' to perform a modified robust ranking aggregation (RRA). We used the combined FDR values for each gene in the benchmarking.

**RIGER**

We used the Java implementation of RIGER (version 2.0, https://github.com/broadinstitute/rigerj) to perform the benchmarking. We set the 'alpha' parameter at 0.1 on (Evers et al. 2016) datasets and at 1.0 on (Sanson et al. 2018) datasets. $log_2$ fold-change values calculated by CB[2] were used as an input of RIGER.

**RSA**

We used the Python implementation of RSA (version 1.9, https://admin-ext.gnf.org/publications/RSA/). $log_2$ fold-change values calculated by CB[2] were used as an input of RSA.

**sgRSEA**

sgRSEA R package (version 0.1, https://cran.r-project.org/web/packages/sgRSEA/) was used for benchmarking. we set the multiplier at 30.

### PBNPA

PBNPA R PACKAGE (version 0.0.2, https://cran.r-project.org/web/packages/PBNPA/) was used for benchmarking. We set the sim.no parameter at 10.

## Specificity measure at sgRNA level

The specificity of detecting true negative from the negative control sgRNAs is measured using $\frac{\sum_{i=1}^{N} 1_A(p_i<\theta)}{N}$, where $N$ is the number of non-targeting sgRNA, $p_i$ is the estimated p-value of the i-th sgRNA, and $\theta$ is the p-value threshold, and $1_A$ is the indicator function.

## Algorithm for quantifying sgRNA abundance

### Previous sgRNA abundance quantification methods

Recently published tools for CRISPR pooled screen analysis, including CRISPRcloud (Jeong et al. 2017), MAGeCK (Li et al. 2014), CRISPRAnalyzeR (Winter et al. 2017), and PinAPL-Py (Spahn et al. 2017) provide different methods for estimating the abundance of sgRNAs in each sample from pooled libraries. In most cases, input data consist of raw FASTQ-format sequencing result files.

CRISPRcloud was the first tool to offer an online user-defined, light-weight quantification method which proceeds on the user-client side. In contrast, CRISPRAnalyzeR and PinAPL-Py run their quantification methods on the server-side. As a result, CRISPRcloud minimized information passed through the Internet by transferring only the processed count matrix to the cloud storage.

However, as pointed out by Spahn et al. (2017), CRISPRcloud quantification algorithm can produce erroneous mapping results if the sgRNA sequences are staggered, because CRISPRcloud extracts sgRNA sequence for each read at a fixed location. Another limitation of CRISPRcloud is the fact that the user must decide where the extraction site is. Nevertheless, CRISPRcloud did not require tuning and was thus arguably more user-friendly than other tools. For instance, in PinAPL-Py, users need to set many tuning parameters for sgRNA quantification: adapter error rate parameter for trimming, matching and ambiguity thresholds and parameters for alignment seed for the Bowtie alignment (Spahn et al. 2017).

We engineered $CB^2$ to address precisely these issues. As a result, users no longer need to perform complicated parameter tuning for the sgRNA abundance quantification; one must simply provide the input files to $CB^2$.

## The binary representation of sgRNA sequence lowers the cost of computation

We used a binary representation for sgRNA sequence. This approach is memory-efficient and improves the user experience at the client-side (Melsted and Pritchard 2011). It only needs $max(K, 2M)$ bits to store an sgRNA-sequence, where *M* is the length of the sequence and *K* is the bit size to store a primitive integer in the machine (usually 64 bits) because we only need two bits to save a nucleotide (i.e., 'A' is '00', 'C' is '01', 'G' is '11', and 'T' is '10'). The memory size is about half of that required for storing a character string of the sequence, i.e., 160 bits are needed to store a 20nt sgRNA sequence. Another benefit of binary representation is that it lowers the time complexity for the shift operator when comparing all k-mers of an sgRNA read using a sliding window. This is an essential function for the quantification algorithm in $CB^2$. Compared to the string shift operator functions, such as string copy, substring extraction, and concatenation, the binary representation produces significantly shorter running times. Algorithm 2 in Supplemental Methods illustrates how a sgRNA library converted to a bit sequence and store the converted sequence into a hash table.

## Sliding window-based algorithm gives a high-resolution quantification with comparable running time

With a binary representation, we run the quantification algorithm as follows: First, we build a hash table for the reference library, with each key of the library in the hash table converted to the binary representation. Second, for each read, we scan the sequence of the read from $5'$ to $3'$ with the sliding window. In the $i$-th iteration, the sliding window contains a substring of the read sequence from $i$ to $i + k - 1$, where $k$ is the length of the sgRNAs. The substring is also converted to a binary sequence, and the hash table is quickly checked to see if the sequence in the sliding window exists in the reference library. If the sequence is found in the hash table, then the count of the sequence is increased by one, and then the algorithm proceeds to the next read. Otherwise, it moves to $i + 1$-th iteration and the bit-shift method will be applied to take the next sliding window. Algorithm 1 in Supplemental Methods represents a procedure of the Sliding window-based algorithm.

For the case of a reverse complement sequenced sample, the entire procedure is repeated on the reverse complement reference sgRNA library and scanning the read from 3' to 5'. After both assays are performed (5' to 3' and 3' to 5' with the reverse complement reference sequences), mapping results between both sequences are compared. The one with a larger count corresponds to the correct sequence mapping. We compared the mappability of $CB^2$ to those of MAGeCK (Li et al. 2015) and PinAPL-Py (Spahn et al. 2017) across multiple datasets from previous studies (Figure 5).

## Acknowledgments

This work has been supported by National Institute of General Medical Sciences R01-GM120033, National Science Foundation - Division of Mathematical Sciences DMS-1263932, Cancer Prevention Research Institute of Texas RP170387, Houston Endowment, and Chao Family Foundation (Z.L.), Huffington Foundation, Howard Hughes Medical Institute (H.Y.Z.), and the Parkinson's Foundation Stanley Fahn Junior Faculty Award PF-JFA-1762 (M.W.C.R.). We thank V. L. Brandt for editing the manuscript.


## Author contributions

H-H.J., M.W.C.R., H.Y.Z., and Z.L. designed the study. H-H.J. and S.Y.K. implemented the $CB^2$ software. H-H.J. and Z.L. performed analysis. Z.L. supervised the project. H-H.J., M.W.C.R., and Z.L. wrote the manuscript with input from all the authors.

## Competing interests

The authors declare no competing interests.

## References


Aguirre AJ, Meyers RM, Weir BA, Vazquez F, Zhang C-Z, Ben-David U, Cook A, Ha G, Harrington WF, Doshi MB, et al. 2016. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discovery*.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome biology* **11**: R106.

Baggerly KA, Deng L, Morris JS, Aldaz CM. 2003. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19**: 1477–1483.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**: 289–300.

DeJesus R, Moretti F, McAllister G, Wang Z, Bergman P, Liu S, Frias E, Alford J, Reece-Hoyes JS, Lindeman A, et al. 2016. Functional CRISPR screening identifies the ufmylation pathway as a regulator of SQSTM1/p62. *eLife* **5**.

Diaz AA, Qin H, Ramalho-Santos M, Song JS. 2015. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic acids research* **43**: e16.

Doench JG. 2017. Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews Genetics* **19**: 67–80.

Dutta B, Azhir A, Merino L-H, Guo Y, Revanur S, Madhamshettiwar PB, Germain RN, Smith JA, Simpson KJ, Martin SE, et al. 2016. An interactive web-based application for Comprehensive Analysis of RNAi-screen Data. *Nature communications* **7**: 10578.

Evers B, Jastrzebski K, Heijmans JPM, Grernrum W, Beijersbergen RL, Bernards R. 2016. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nature biotechnology* **34**: 11–14.

Fisher RA. 1925. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**: 647–61.

Golden RJ, Chen B, Li T, Braun J, Manjunath H, Chen X, Wu J, Schmid V, Chang T-C, Kopp F, et al. 2017. An Argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature* **542**: 197–202.

Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. 2014. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular systems biology* **10**: 733.

Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, et al. 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**: 1515–1526.

Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, et al. 2016. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* **5**: e19760.

Hu G, Luo J. 2012. A primer on using pooled shRNA libraries for functional genomic screens. *Acta Biochimica et Biophysica Sinica* **44**: 103–112.

Iorio F, Behan FM, Gonçalves E, Bhosle SG, Chen E, Shepherd R, Beaver C, Ansari R, Pooley R, Wilkinson P, et al. 2018. Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics* **19**: 604.

Jeong HH, Kim SY, Rousseaux MW, Zoghbi HY, Liu Z. 2017. CRISPRcloud: A secure cloud-based pipeline for CRISPR pooled screen deconvolution. *Bioinformatics* **33**: 2963–2965.

Jia G, Wang X, Xiao G. 2017. A permutation-based non-parametric analysis of CRISPR screen data. *BMC genomics* **18**: 545.

Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. 2014. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature biotechnology* **32**: 267–73.

König R, Chiang C, Tu BP, Yan SF, DeJesus PD, Romero A, Bergauer T, Orth A, Krueger U, Zhou Y, et al. 2007. A probability-based approach for the analysis of large-scale RNAi screens. *Nature Methods* **4**: 847–849.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357.

Li M, Yu JSL, Tilgner K, Ong SH, Koike-Yusa H, Yusa K. 2018. Genome-wide CRISPR-KO Screen Uncovers mTORC1-Mediated Gsk3 Regulation in Naive Pluripotency Maintenance and Dissolution. *Cell Reports*.

Li W, Köster J, Xu H, Chen C-H, Xiao T, Liu JS, Brown M, Liu XS. 2015. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome biology* **16**: 281.

Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. 2014. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome biology* **15**: 554.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**: 550.

Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, Hinkle G, Boehm JS, Beroukhim R, Weir BA, et al. 2008. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences* **105**: 20380–20385.

Luo J, Emanuele MJ, Li D, Creighton CJ, Schlabach MR, Westbrook TF, Wong K-K, Elledge SJ. 2009. A Genome-wide RNAi Screen Identifies Multiple Synthetic Lethal Interactions with the Ras Oncogene. *Cell* **137**: 835–848.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18**: 1509–1517.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10.

Melsted P, Pritchard JK. 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* **12**: 333.

Mohr SE, Smith JA, Shamu CE, Neumüller RA, Perrimon N. 2014. RNAi screening comes of age: improved techniques and complementary approaches. *Nature reviews Molecular cell biology* **15**: 591–600.

Noh J. 2015. *sgRSEA: Enrichment Analysis of CRISPR/Cas9 Knockout Screen Data*. https://cran.r-project.org/web/packages/sgRSEA/.

Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K, et al. 2004. A resource for large-scale RNA-interference-based screens in mammals. *Nature* **428**: 427–431.

Park J, Al-Ramahi I, Tan Q, Mollema N, Diaz-Garcia JR, Gallego-Flores T, Lu H-C, Lagalwar S, Duvick L, Kang H, et al. 2013. RAS–MAPK–MSK1 pathway modulates ataxin 1 protein levels and toxicity in SCA1. *Nature* **498**: 325–331.

Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, Przybylski D, Platt RJ, Tirosh I, Sanjana NE, et al. 2015. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**: 675–86.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org.

Sanjana NE, Shalem O, Zhang F. 2014. Improved vectors and genome-wide libraries for CRISPR screening. *Nature methods* **11**: 783.

Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, Vaimberg EW, Goodale A, Root DE, Piccioni F, et al. 2018. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature communications* **9**: 5416.

Schlabach MR, Luo J, Solimini NL, Hu G, Xu Q, Li MZ, Zhao Z, Smogorzewska A, Sowa ME, Ang XL, et al. 2008. Cancer proliferation gene discovery through functional genomics. *Science (New York, NY)* **319**: 620–4.

Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, et al. 2014. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343**: 84–87.

Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, Schumacher SE, Zack TI, Beroukhim R, Garraway LA, et al. 2013. ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome research* **23**: 665–678.

Silva JM, Marran K, Parker JS, Silva J, Golding M, Schlabach MR, Elledge SJ, Hannon GJ, Chang K. 2008. Profiling Essential Genes in Human Mammary Cells by Multiplex RNAi Screening. *Science* **319**: 617–620.

Simon MM, Moresco EMY, Bull KR, Kumar S, Mallon A-M, Beutler B, Potter PK. 2015. Current strategies for mutation detection in phenotype-driven screens utilising next generation sequencing. *Mammalian genome : official journal of the International Mammalian Genome Society* **26**: 486–500.

Spahn PN, Bath T, Weiss RJ, Kim J, Esko JD, Lewis NE, Harismendy O. 2017. PinAPL-Py: A comprehensive web-application for the analysis of CRISPR/Cas9 screens. *Scientific reports* **7**: 15854.

Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science (New York, NY)* **343**: 80–4.

Winter J, Schwering M, Pelz O, Rauscher B, Zhan T, Heigwer F, Boutros M. 2017. CRISPRAnalyzeR: Interactive analysis, annotation and documentation of pooled CRISPR screens. *bioRxiv*.

Yu J, Silva J, Califano A. 2016. ScreenBEAM: a novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling. *Bioinformatics (Oxford, England)* **32**: 260–7.

Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W. 2014. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**: 487–491.

# Figure legends

Figure 1. $CB^2$ offers robust target identification with high precision and recall. (A-B) Benchmark results using data from Evers et al. (2016). (A) Heatmaps illustrate FDRs of gene statistics from each of nine leading high-complexity pooled screen analysis tools. (B) F1-score measurements at different FDR cut-offs across all methods. At commonly used FDR cut-offs, $CB^2$ can identify most of the essential genes with high rates of precision and recall. (C-D) Same representation as in (A-B), using data from Sanson et al. (2018).

Figure 2. $CB^2$ detects essential genes missed by other leading methods: the case of *RPL5*. sgRNA quantification for *RPL5* in cell line (A) RT112, (B) UMUC3 using CRISPRn and (C) RT112 using the CRISPRi library. The top panels show CPM (count per million) of sgRNAs that target *RPL5* for each group ($T_0$ and $T_1$), and the bottom panels indicate the reported the FDR for *RPL5* in each screen across all the methods. A horizontal line at $\mathrm{FDR} = 0.01$ is used as a threshold for statistical cutoff. $CB^2$ outperforms all other methods of identifying *RPL5* as an essential gene across all benchmark datasets.

Figure 3. Comparison of False Positive Rate for non-targeting sgRNAs on (Parnas et al. 2015)'s screen data. (A) Specificity comparison between $CB^2$ and MAGeCK for the six different p-value thresholds. The y-axis indicates specificity, and the x-axis indicates the level of the p-value threshold for the specificity calculation. (B) Volcano plots of the p-value of non-targeting sgRNAs. The y-axis indicates the negative logarithm value of p-value, and the x-axis indicates the $\log_2$ value of fold-change. All of the data points are from negative control sgRNAs. False Positive were plotted in red. Horizontal blue lines at $p = 0.01$ indicates the threshold for statistical cutoff.

Figure 4. Discrepancy of sgRNA-level statistics between $CB^2$ and MAGeCK in two public CRISPR pooled screens. (A) Heatmaps of normalized read counts of the detected sgRNAs from the screen data in Parnas et al. (2015). Left: A heatmap of sgRNAs detected by $CB^2$ only. Right: A heatmap of sgRNAs detected by MAGeCK only. (B) Venn diagrams of sgRNAs detected by $CB^2$ and MAGeCK from the screen data in Parnas et al. (2015). (C-D) Same representations of (A-B) using data from Li et al. (2018).

Figure 5. $CB^2$ outperforms MAGeCK and PinAPL-Py in the percentage of mapped reads over six benchmark datasets. (A) Read mappability of $CB^2$, MAGeCK, and PinAPL-Py across six different datasets. (B) Representative examples of reads that were not mapped by MAGeCK or PinAPL-Py. Adapters are highlighted with green, sgRNAs with a red. Yellow boxes show the predicted locations of sgRNAs by each method. Several parameters were used to optimize performances of PinAPL-Py.

A

CRISPRn–RT112

CRISPRn–UMUC3

CRISPRi–RT112

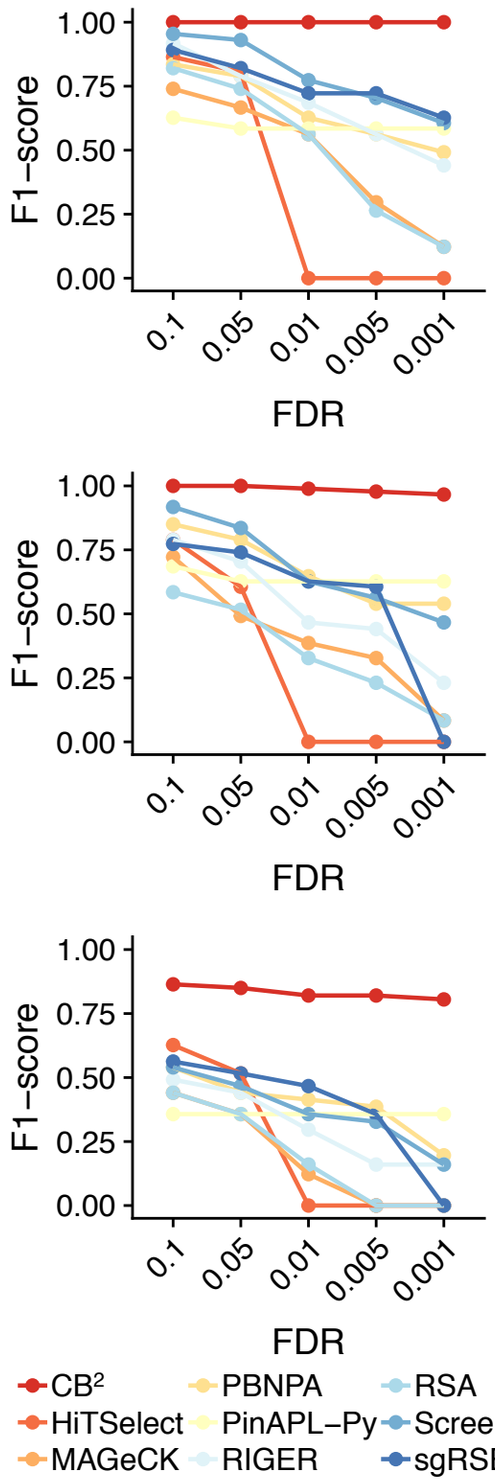
CRISPRn–A375

CRISPRi–A375

CB²
ScreenBEAM
PBNPA
sgRSEA
HitSelect
MAGeCK
RIGER
RSA
PinAPL–Py

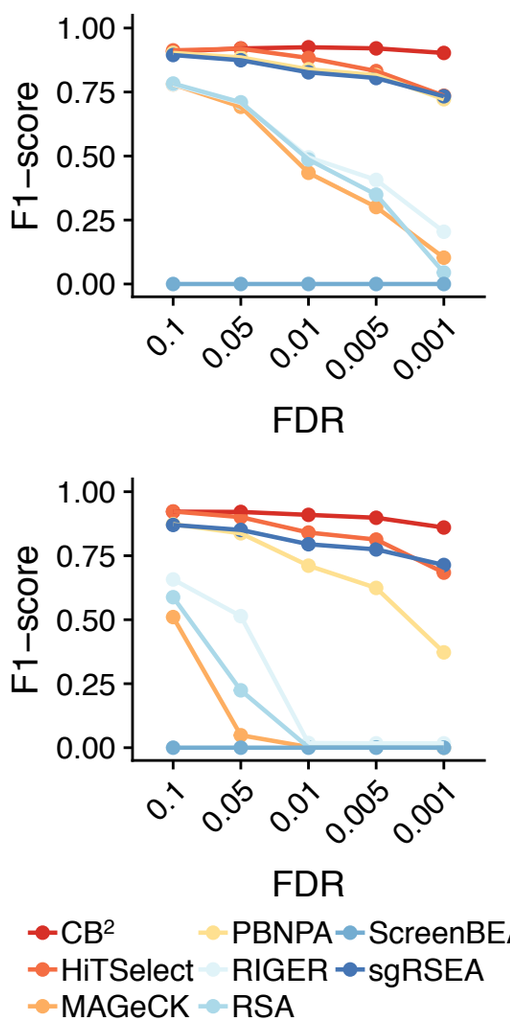
–log₁₀(FDR) ≥8 6 4 2 0

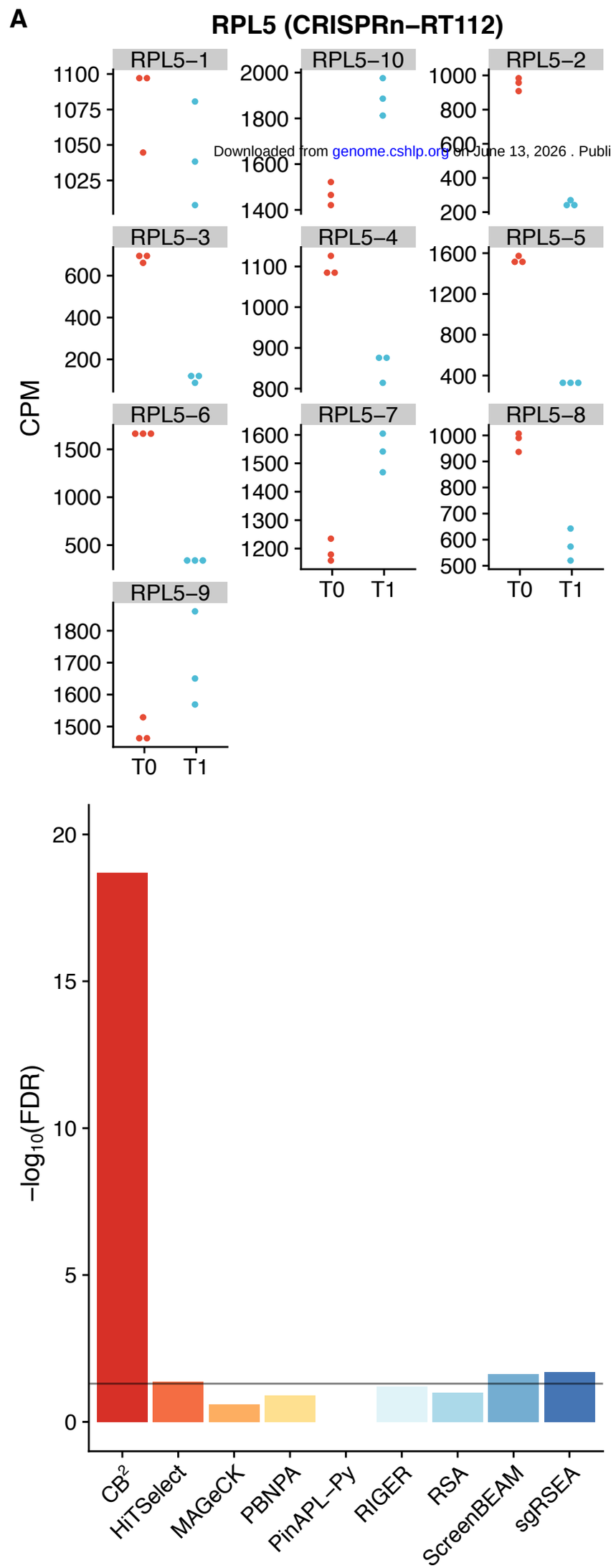
B

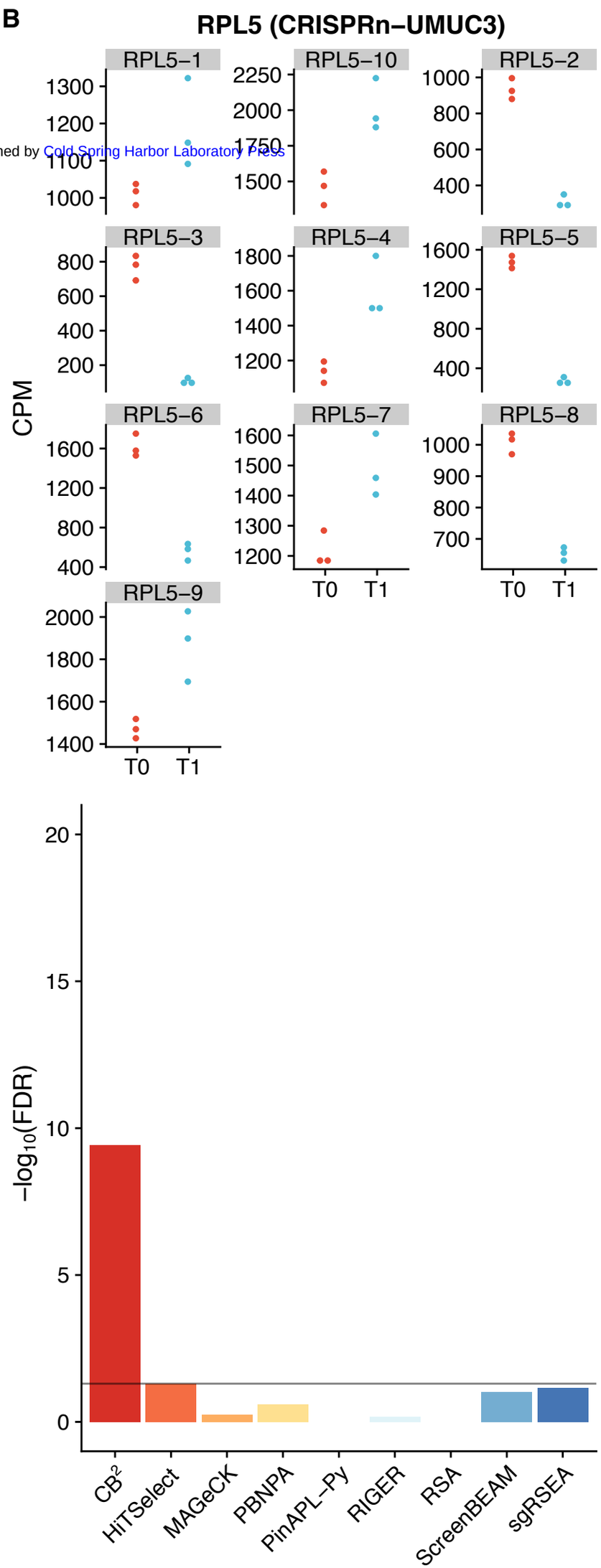


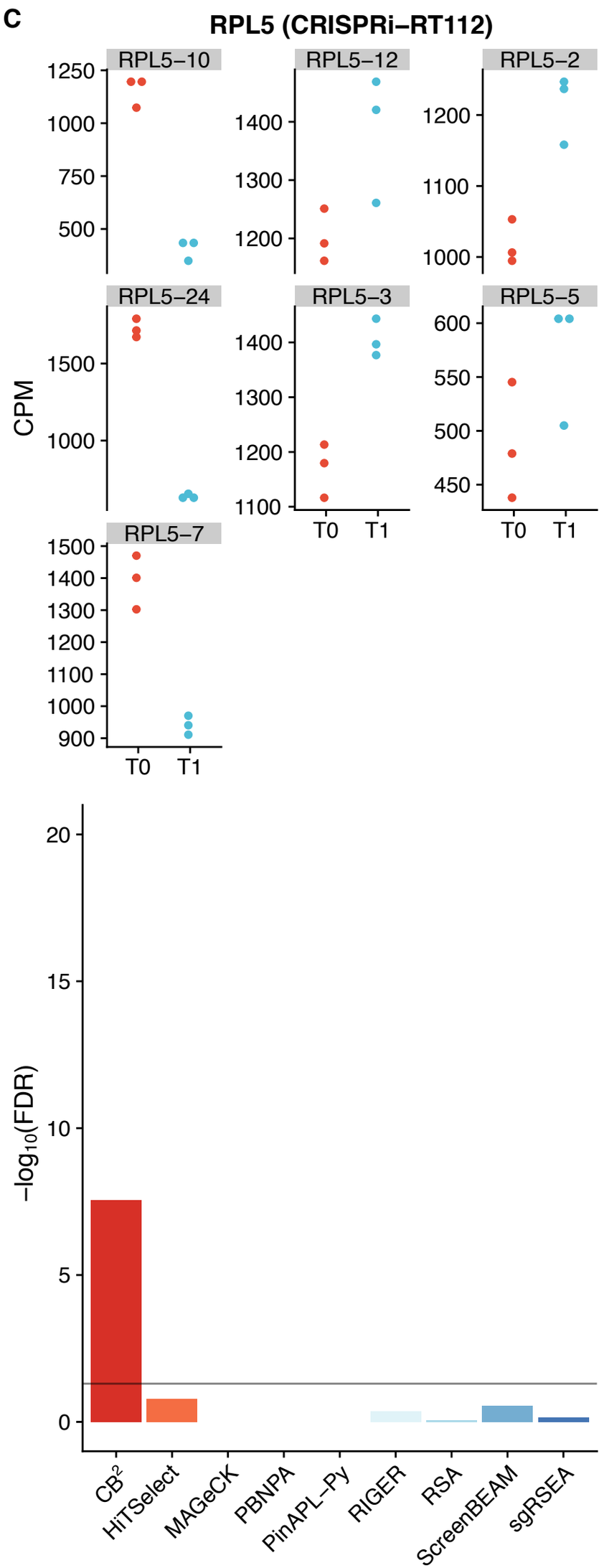

C

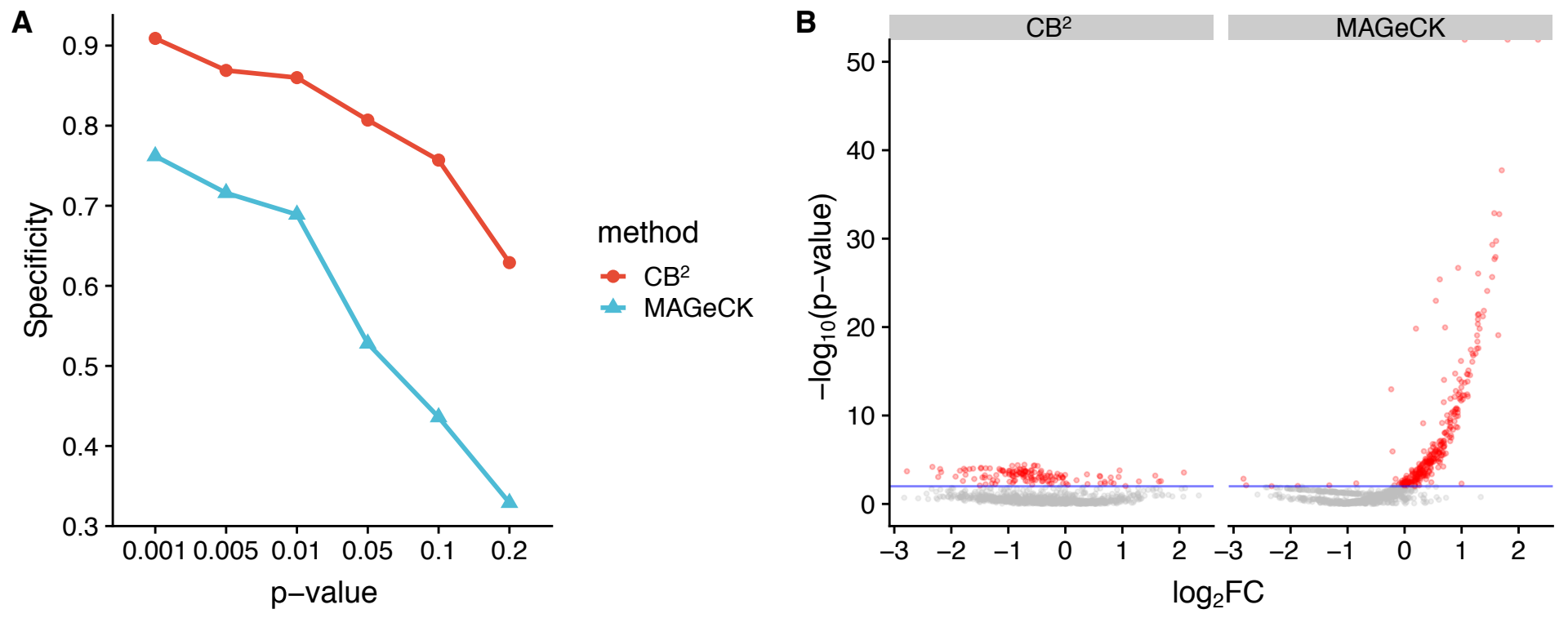
D

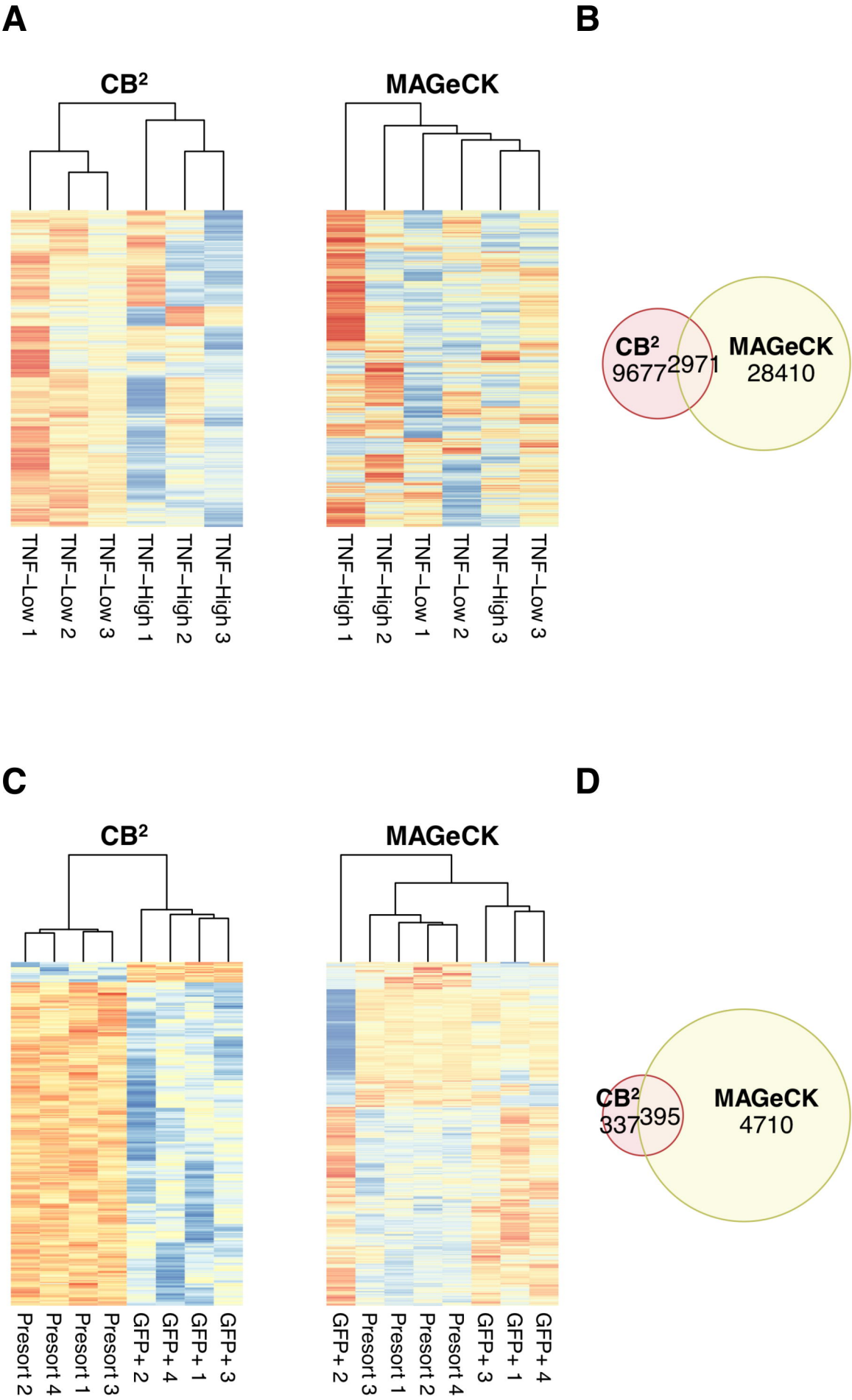


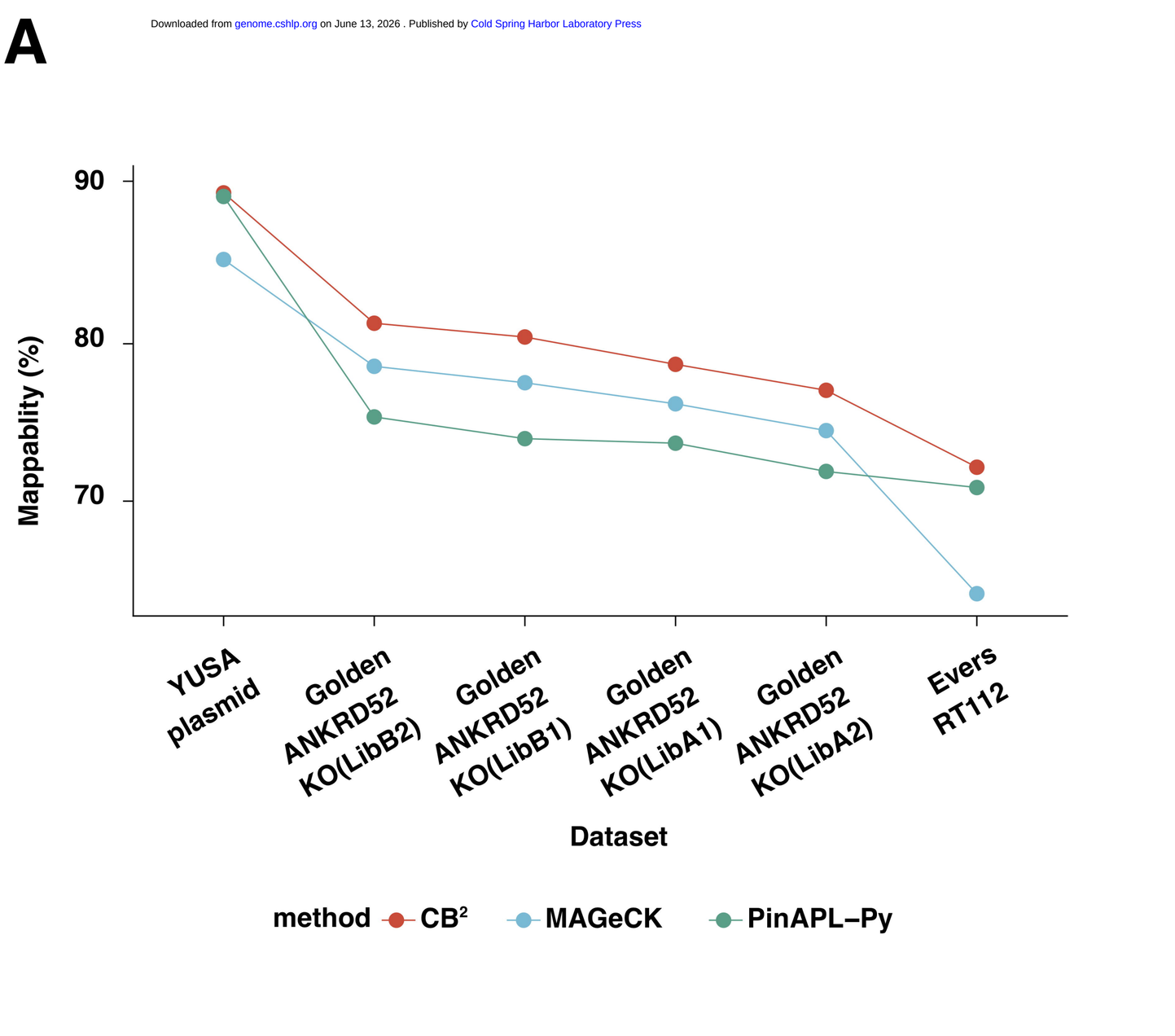



A
Specificity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.001
0.005
0.01
0.05
0.1
0.2
p–value
method
CB²
MAGeCK
B
CB²
MAGeCK
−log₁₀(p–value)
50
40
30
20
10
0
−3
−2
−1
0
1
2
log₂FC


A
CB²
MAGeCK
TNF–Low 1
TNF–Low 2
TNF–Low 3
TNF–High 1
TNF–High 2
TNF–High 3
TNF–High 1
TNF–High 2
TNF–Low 1
TNF–Low 2
TNF–High 3
TNF–Low 3
B
CB²
9677
2971
MAGeCK
28410
C
CB²
MAGeCK
Presort 2
Presort 4
Presort 1
Presort 3
GFP+ 2
GFP+ 4
GFP+ 1
GFP+ 3
GFP+ 2
Presort 3
Presort 1
Presort 2
Presort 4
GFP+ 3
GFP+ 1
GFP+ 4
D
CB²
337
395
MAGeCK
4710



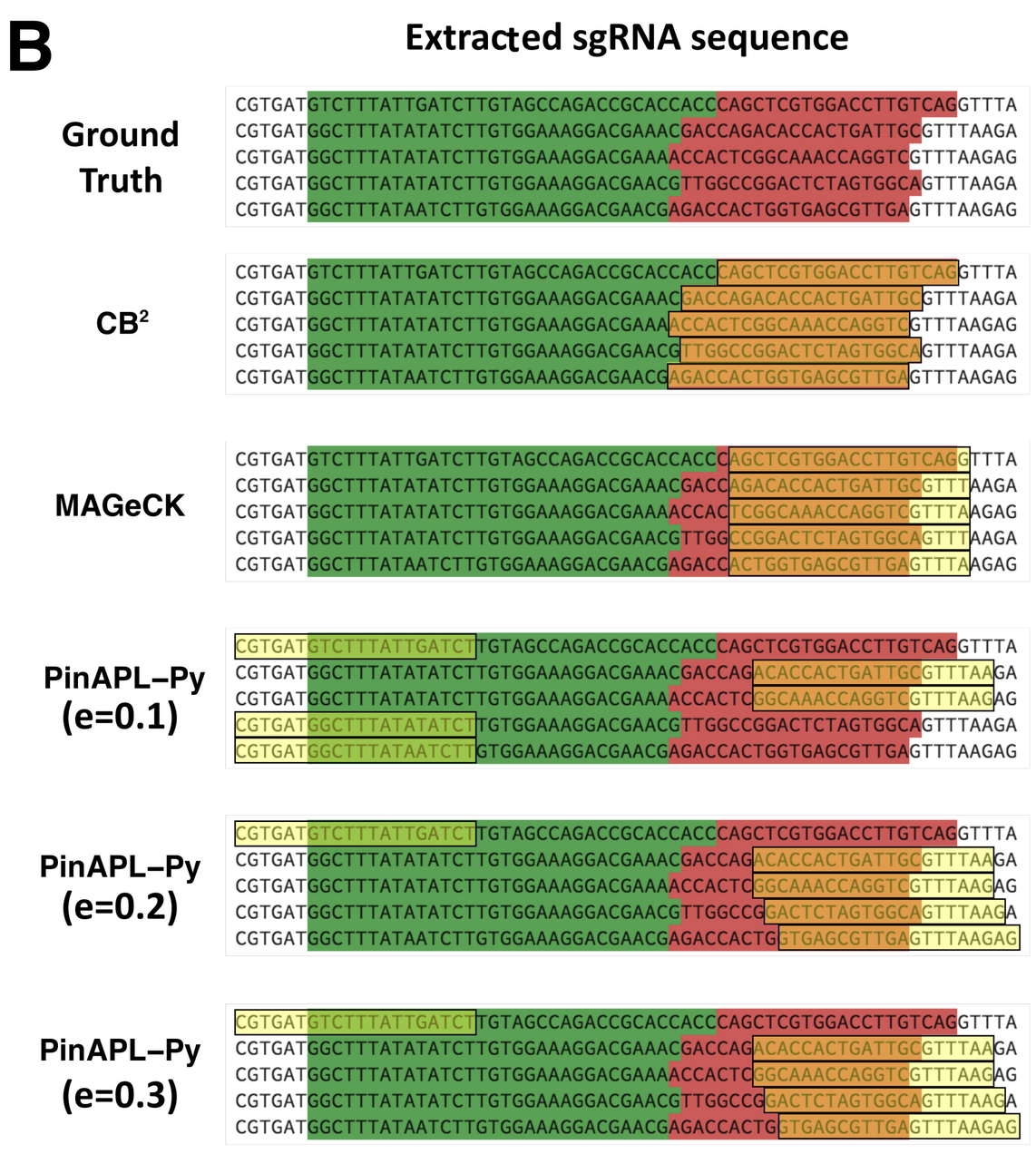



**Table 1.** Statistical models used by CB2 and existing methods. All of the methods were used in the target identification benchmarking.

| Name | sgRNA-level statistics | gene-level statistics |
|---|---|---|
| $CB^2$ | Beta-binomial distribution | Fisher's method |
| HitSelect (Diaz et al. 2015) | Poisson distribution (active number of sgRNAs) | Stochastic multi-objective ranking method for gene-level statistics |
| MAGeCK (Li et al. 2015) | Negative-binomial distribution | $\alpha$RRA and MLE for the gene-level statistics |
| PBNPA (Jia et al. 2017) | | Non-parametric permutation-test for each replicate |
| ScreenBeam (Yu et al. 2016) | Normal distribution | Bayesian hierarchical modeling |
| sgRSEA (Noh 2015) | | Non-parametric permutation-test |
| PinAPL-Py (Spahn et al. 2017) | Negative-binomial model (control samples) | $\alpha$RRA and STARS |
| RIGER (Luo et al. 2008) | | Kolmogorov–Smirnov-based non-parametric statistics for the gene-level statistics |
| RSA (König et al. 2007) | hypergeometric distribution (sgRNA ranking) | Ranking-based statistics for the gene-level statistics |