Genome Biology

METHOD Open Access

MAUDE: inferring expression changes in sorting-based CRISPR screens

Carl G. de Boer[1,2*†], John P. Ray[3†], Nir Hacohen[1,3,4] and Aviv Regev[1,5*]

* Correspondence: carl.deboer@ubc.ca; aregev@broadinstitute.org
†Carl G. de Boer and John P. Ray contributed equally to this work.
[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA
Full list of author information is available at the end of the article

Abstract

Improved methods are needed to model CRISPR screen data for interrogation of genetic elements that alter reporter gene expression readout. We create MAUDE (Mean Alterations Using Discrete Expression) for quantifying the impact of guide RNAs on a target gene's expression in a pooled, sorting-based expression screen. MAUDE quantifies guide-level effects by modeling the distribution of cells across sorting expression bins. It then combines guides to estimate the statistical significance and effect size of targeted genetic elements. We demonstrate that MAUDE outperforms previous approaches and provide experimental design guidelines to best leverage MAUDE, which is available on https://github.com/Carldeboer/MAUDE.

Keywords: CRISPR/Cas9, Enhancers, Gene regulation, Transcriptional regulation, Gene expression, Pooled screen, R

Background

Pooled CRISPR/Cas9 screens with a readout of a reporter gene's expression have improved our understanding of how both *cis* and *trans* regulators control gene expression. CRISPR/Cas9 and related systems use complementarity between an RNA guide and the genomic DNA to direct Cas9 to specific loci. Engineered Cas9 proteins with different enzymatic and regulatory activities now allow researchers to mutate DNA [1], modify the chromatin state [2] of the targeted locus, inertly bind loci to inhibit transcription [3], or recruit activators [4]. Using these tools, pooled screens with libraries of guides targeting genes or the non-coding genome and measuring the impact on a reporter gene's expression can identify *trans* regulators or *cis*-regulatory elements that affect the reporter's expression [1, 2, 5–7]. For instance, CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa) have been used to respectively repress or activate elements in a targeted *cis*-regulatory region and identify enhancers of a gene's expression [2, 4, 8, 9]. In each case, we estimate how each guide affects the targeted (or reporter) gene's expression by sorting cells into discrete expression bins and measuring the guide's abundance in the bins. Such screens help dissect enhancer regulatory logic [5, 8] and identify the genetic variation most likely to contribute to common human disease [9].

Despite the growing number of gene expression CRISPR screening strategies, computational methods to determine which guides and elements alter gene expression remain relatively ad hoc. In general, detecting guides that impact expression relies on finding those with a biased distribution across the expression bins. While tools have been created to model and analyze the data for selection-based screens [10], the analysis requirements for sorting-based screens differ significantly. Most methods used to analyze sorting-based screens rely on one of two strategies: (1) using the log fold change between the guide abundances in the high and low expression bins directly [5, 6, 9], or (2) repurposing RNA-seq differential expression tools [11, 12] to find changes in guide abundance between high and low bins, taking advantage of their ability to account for differences in sequencing depth [6, 13]. However, both strategies are difficult to apply if more than two sorting bins are used and do not leverage the unique character of these data.

Here, we describe MAUDE (Mean Alterations Using Discrete Expression), an analytical framework for estimating the effect of CRISPR guides (or other perturbagens) on expression, as measured by sorting into discrete expression bins and sequencing. MAUDE maximizes the likelihood of the observed sequencing data over the cell-sorting bins to estimate the mean expression for cells containing each guide. It then combines the resulting Z-scores from multiple guides targeting the same element into element-level effect sizes and estimates statistical significance. Elements can be defined a priori, by annotating guides with the element they target, or identified de novo by using a sliding window to combine guides targeting neighboring chromosomal locations. MAUDE is highly sensitive at finding expression-altering guides and elements, is adept at estimating effect sizes, and outperforms existing approaches. Finally, we provide guidance on the design of expression-based CRISPR screens to maximize the information gleaned.

Results

MAUDE identifies expression-altering guides by their distribution across bins

A sorting-based screen takes the following form: (1) A set of possible perturbations, each of which is encoded in DNA and may or may not modify the mean gene expression level by some amount (for sorting-based CRISPR screen, these perturbations are guide RNAs), is introduced into (2) a population of cells, such that each receives one or more of these perturbations, and is allowed to grow. (3) Afterwards, the cells are sorted by Fluorescence Activated Cell Sorting (FACS) into discrete bins based on their expression level of a gene of interest (a continuous value). FACS expression measurements are often log-normal, but are not required to be so. (4) The sorted cells, as well as unsorted control cells, are then lysed and their DNA isolated. (5) Guide DNA barcodes are usually amplified with PCR and sequenced. This sequencing-mediated sampling process can be represented as a sampling from a negative binomial distribution [11, 14, 15]. Consequently, (6) at the end of the experiment, we have guide RNA read counts for each of several bins, and an unsorted control bin.

MAUDE aims to learn the mean expression level that results from each guide-directed perturbation. If expression is bimodal (e.g., perturbations alter the fraction of expressing cells), MAUDE guide-level effect sizes correspond to the *average* effect of the perturbation (see below). MAUDE requires as input that the sorting bins be sequenced and the sizes of each bin (% of total cells in the distribution) are known, that the screen includes negative control guides that should not affect expression, and that the unsorted input library be sequenced to quantify overall library composition. The unsorted input library should be harvested from the same population of cells used to sort at the time of sorting to minimize differences in composition that might result from guides altering cellular fitness. Furthermore, the negative control guides included should either be non-targeting (as in a CRISPRi or CRISPRa screen), or target an irrelevant locus (if using enzymatically active Cas9, since DNA damage can induce global expression changes).

MAUDE estimates the most likely mean expression level effect of a given guide g, $\hat{\mu}_g$, by maximizing the likelihood of the observed reads per bin under these assumptions (Fig. 1). To do so, it estimates the number of reads of guide g expected to be in each expression bin b, for a given mean expression level μ_g. It estimates the optimal $\hat{\mu}_g$ as the one that maximizes the log likelihood of the observed number of reads given the expected number of reads in each bin (Fig. 1a—left).

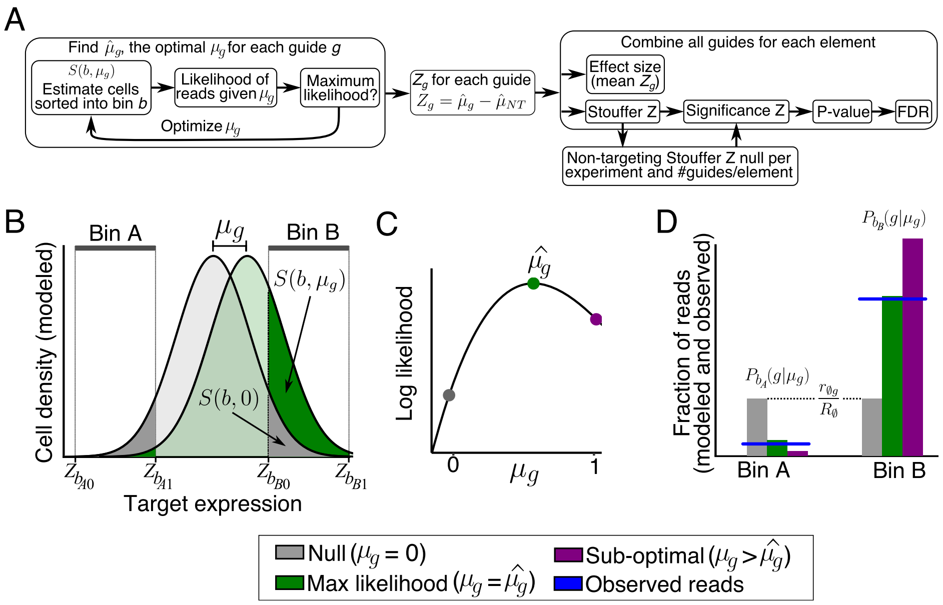
A
Find $\hat{\mu}_g$, the optimal μ_g for each guide g
$S(b, \mu_g)$
Estimate cells sorted into bin b
Likelihood of reads given μ_g
Maximum likelihood?
Optimize μ_g
Z_g for each guide
$Z_g = \hat{\mu}_g - \hat{\mu}_{NT}$
Combine all guides for each element
Effect size (mean Z_g)
Stouffer Z
Significance Z
P-value
FDR
Non-targeting Stouffer Z null per experiment and #guides/element
B
Bin A
μ_g
Bin B
$S(b, \mu_g)$
$S(b, 0)$
Cell density (modeled)
$Z_{b_{A0}}$
$Z_{b_{A1}}$
$Z_{b_{B0}}$
$Z_{b_{B1}}$
Target expression
C
$\hat{\mu_g}$
Log likelihood
0
μ_g
1
D
$P_{b_B}(g|\mu_g)$
$P_{b_A}(g|\mu_g)$
$\frac{r_{\emptyset g}}{R_\emptyset}$
Fraction of reads (modeled and observed)
Bin A
Bin B
Null ($\mu_g = 0$)
Max likelihood ($\mu_g = \hat{\mu_g}$)
Sub-optimal ($\mu_g > \hat{\mu_g}$)
Observed reads

Fig. 1 MAUDE approach to scoring expression-based screens. **a** Method overview. **b**–**d** MAUDE approach to estimating the optimal mean expression level per guide. **b** Estimation of cell density distribution for each guide. Cell density distribution (*y* axis) of target gene expression (*x* axis) for each guide is modeled as a normal distribution, with mean μ_g. μ_g is optimized by calculating, for each bin (bins A and B, top), the fraction of cells expected to be in the bin under the null model (no overall change in expression; $S(b, 0)$) and the fraction of cells expected in the bin given the current value of μ_g ($S(b, \mu_g)$). **c** Estimation of optimal μ_g. We find the optimal $\hat{\mu}_g$ by calculating the log likelihood (*y* axis) of the observed bin read abundances given each value of μ_g (*x* axis). **d** Expected number of guide reads per bin. The fraction of reads (*y* axis) of guide g ($P_b(g|\,\mu_g)$) expected in each bin b (*x* axis), for different values of μ_g (colors)

To achieve this, MAUDE proceeds in several steps. First, it uses the unsorted (input) library to estimate the library composition, defined as $r_{\varnothing g}/R_{\varnothing}$, the fraction of reads from guide g ($r_{\varnothing g}$) out of the total reads ($R_{\varnothing}$). Next, using the overall fraction of *cells* observed in each expression bin (recorded during cell sorting), it quantile normalizes the sorting expression space to a standard normal reference ($\mu = 0$ and $\sigma = 1$). For each bin, it calculates the Z-scores in standard normal space (Z_{b0} and Z_{b1} for lower and upper limit, respectively) that capture the same quantile as the bin captured in the original (FACS) data. Since most guides in such screens do not alter expression, the vast majority of the overall distribution represents the expression σ associated with no effect (e.g., wild type). Thus, for modeling simplicity, we can assume that each guide has the same σ—that of the overall distribution. It then defines $S(b, \mu_g)$ as the fraction of cells containing guide g expected by the model to have been sorted into bin b (with the bin defined by Z_{b0} and Z_{b1}, respectively). $S(b, \mu_g)$ is calculated using the normal cumulative distribution function, $CDF_{norm}()$, yielding the probability that a cell sampled from a normal distribution with mean μ_g will have an expression (x) less than the sorting upper bin bound Z_{b1}, but not smaller than the lower bin bound Z_{b0}:

$$S\left(b, \mu_g\right) = CDF_{norm}\left(x + \mu_g \leq Z_{b1}\right) - CDF_{norm}\left(x + \mu_g \leq Z_{b0}\right) \quad (1)$$

For a given μ_g, MAUDE calculates the expected expression distribution (Fig. 1b). The expected fraction of cells in bin b that contain guide g is:

$$P_b\left(g|\mu_g\right) = \frac{r_{\varnothing g}}{R_{\varnothing}} \times \frac{S\left(b, \mu_g\right)}{S(b, 0)} \quad (2)$$

Intuitively, if guide g does not affect expression, μ_g is 0, and so $P_b(g| \mu_g)$ is just the fraction of the overall library occupied by the guide ($r_{\varnothing g}/R_{\varnothing}$); if the value of μ_g results in a higher fraction of cells containing guide g in bin b, the guide is enriched in that bin ($P_b(g| \mu_g) > r_{\varnothing g}/R_{\varnothing}$; Fig. 1b—right); depletion results when fewer cells end up in the bin ($P_b(g| \mu_g) < r_{\varnothing g}/R_{\varnothing}$; Fig. 1b—left).

Finally, MAUDE learns the optimal μ_g given the observed guide reads, by maximizing the (log) likelihood of observing the reads from the guide g in each bin b (r_{gb}), given the overall number of reads observed in that bin (R_b; Fig. 1c, d). To reduce the impact of cases where a guide cannot be quantified reliably due to low coverage, MAUDE also includes a prior favoring no change in expression via a pseudocount added uniformly to all expression bins (see the "Methods" section). We assume reads are sampled using a negative binomial distribution (with probability density function $PDF_{NB}()$) [11, 14, 15]. Thus, MAUDE maximizes the following equation separately for each guide g to learn its optimal $\hat{\mu}_g$:

$$\hat{\mu}_g = \underset{\mu_g}{\operatorname{argmax}} \sum^{\text{bins } b} \log\left(PDF_{NB}\left(r_{gb}, R_b, P_b\left(g|\mu_g\right)\right)\right) \quad (3)$$

Once MAUDE learned $\hat{\mu}_g$ for each guide, it subtracts from each the average $\hat{\mu}_g$ of the negative control guides ($\hat{\mu}_{NT}$) to convert each guide's effect on expression to a Z-score (Z_g; Fig. 1a—middle):

$$Z_g = \hat{\mu}_g - \hat{\mu}_{NT} \tag{4}$$

This step is necessary because targeting guides, whose true mean is sometimes not 0, are also included in the overall distribution and may have shifted the overall mean. Thus, Z_g represents the number of standard deviations (SDs) by which guide g has altered gene expression compared to the negative control guides. Importantly, these guide Z-scores can easily be inspected to detect possible off-target effects or other outliers, and nominate specific guides for further validation.

MAUDE calculates element-level effect size and significance by aggregating signal across guides

With a Z-score representing the expression change for every guide, MAUDE calculates element-level effect size and statistical significance (Fig. 1a—right), using either (1) known element annotations or (2) a sliding window across the locus of interest to define elements. Examples of potential annotated elements are putative enhancers or protein coding genes. Sliding windows require that the region be tiled with guides, with the final resolution and sensitivity proportional to the tiling density, but require no prior knowledge about the region. We combine all guides within a pre-specified window size (e.g., 500 bp), requiring a minimum number of guides per window (e.g., 5), and testing all possible windows with distinct guide sets. The element's effect size, expressed as a Z-score, is the average Z-score of the guides targeting that element.

To estimate statistical significance, we combine guide-level Z-scores for all guides targeting that element (by Stouffer's method) into a single Z-score, representing a signed significance of the regulatory change ($Z > 0$ representing upregulation and $Z < 0$ representing downregulation). To minimize the effects of outliers, the experimental noise, and the number of guides per element (which can greatly affect the robustness of these Stouffer Z-scores), we repeatedly sample the negative control guides for each experiment and every possible number of guides/element, to create null distributions. We then scale the Stouffer Z-scores by the standard deviation of the corresponding negative control null (i.e., the null for that experiment with the same number of guides/element). Finally, the resulting significance Z-scores are converted into P values using the normal cumulative distribution function ($\mu = 0$, $\sigma = 1$), for the upper tail (upregulation), the lower tail (downregulation), or the minimum of the two (either), with corresponding Benjamini-Hochberg FDR correction. Although MAUDE does not require biological replicates, replicates allow the detection of replicate-specific experimental noise, and so we recommend having at least two replicates and considering only those elements that are active in multiple replicates.

MAUDE correctly identifies differential elements in simulated data

To test MAUDE's performance and underlying assumptions, we next simulated two experiments, each with 200 targeted elements and 5 guides per element. We compared two ways in which expression might be altered by targeting of an element: (1) a shift from one unimodal distribution of expression levels to another, reflecting an impact on expression affecting each perturbed cell (μ_g ranging from 0.01 to 1 SDs, in 0.01 increments; "mean-altering"; Fig. 2a), or (2) having two fixed distributions of expression levels, as would be the case in two sub-populations, with the perturbation causing a

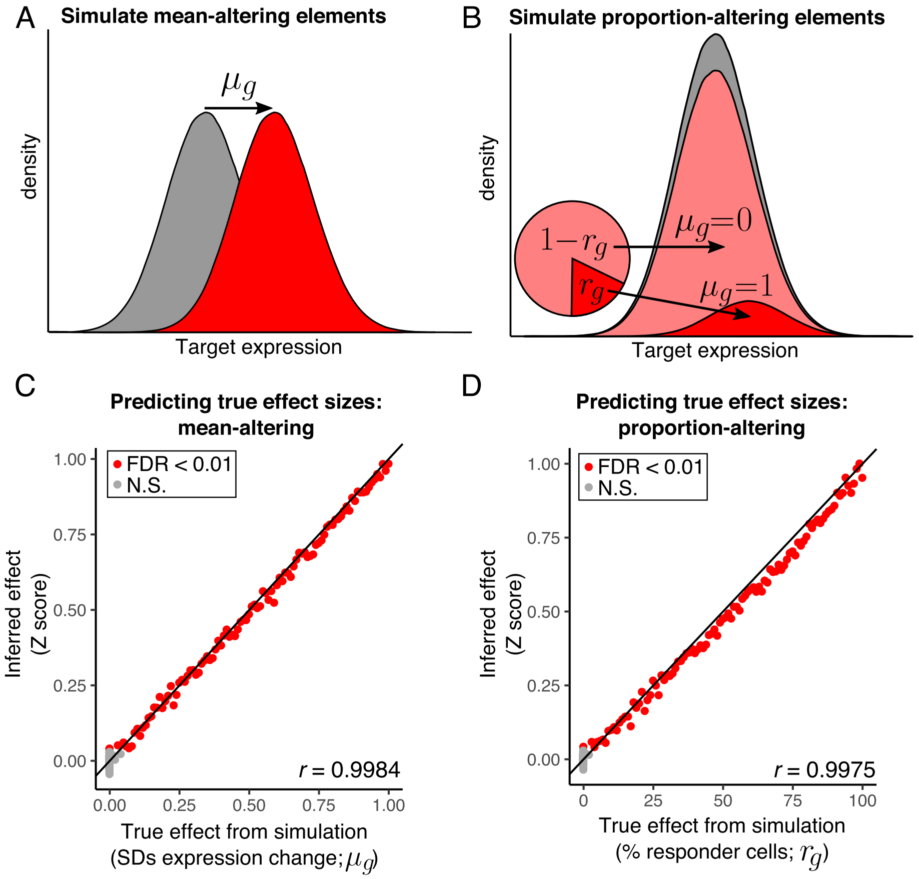

Fig. 2 MAUDE correctly identifies functional elements and effect sizes in simulated data. **a** Simulation for mean-altering elements. A mean-altering element is modeled as changing the mean expression μ_g (between 0.01 and 1) of cells with the effective element (red curve) compared to cells with ineffective elements or negative control guides (gray curve). **b** Simulation of proportion-altering elements. A proportion-altering element is modeled as resulting in a different proportion of cells with altered expression (r_g), from 1 to 100%. On average, r_g% of cells would have expression increased by $\mu_g = 1$ (dark red curve), and $(1 - r_g)$% of cells would have no change in expression ($\mu_g = 0$; light red curve) relative to ineffective elements and negative control guides (gray curve). **c**, **d** Effect sizes are correctly estimated. The true effect used to generate the simulated data (*x* axis) vs. the inferred effect size (*y* axis) for **c** mean-altering elements (where the true effect is μ_g) and **d** proportion-altering elements (where the true effect is r_g). Elements meeting a 1% FDR are shown in red and others in gray. Pearson's *r* is shown in lower right (considering effective elements only). Black line: $y = x$

shift in the relative proportion of cells in each expression mode ($\mu_g = 1$; the fraction of affected cells r_g ranges from 1 to 100%, in 1% increments; "proportion-altering"; Fig. 2b). In either case, 100 "effective" elements have an impact, and the other 100 "ineffective" elements have no effect on expression. We generated these synthetic data, simulating sorting cells by expression into six different 10% bins on the extremes of expression (see the "Methods" section).

MAUDE analysis was adept at identifying effective elements and accurately estimated effect sizes and fractions of responding cells. In both simulations, it detected as significant (1% Benjamini-Hochberg FDR) 97 of the 100 effective elements, with effect sizes as low as 0.03 SDs (mean-altering; Fig. 2c) or 3% of cells responding (proportion-altering; Fig. 2d), indicating that MAUDE is sensitive to even small changes. As expected, each simulation included 1–2 false positives (2/99 and 1/98 for mean-altering and proportion-altering simulations, respectively; estimated effect size of ~ 0.04 SDs for both), consistent with our 1% FDR, indicating that our P values are well-calibrated. The correlation between inferred and actual effect size was very high (Pearson's $r = \sim 0.998$ for both simulations; only considering the 100 simulation-defined effective elements). In the proportion-altering simulation, the estimated effect size corresponds to the average affect size (e.g., 10% of cells shifting by 1 SD results in an average effect size of 0.1). Overall, MAUDE performs extremely well on synthetic data and can identify perturbations that shift the mean expression or the fraction of expressing cells with high sensitivity and specificity.

MAUDE outperforms other approaches on a variety of experimental datasets

We next evaluated MAUDE's performance on two previously published experimental datasets: a CRISPRa tiling screen for CD69 expression using cell surface CD69 protein as a readout in Jurkat cells [9], which was conducted in duplicate, and both CRISPRi and CRISPRa screens performed at the *TNFAIP3* locus using *TNFAIP3* mRNA-based expression readout in a variety of cell types, each performed in at least duplicate [16]. We compared MAUDE to the log fold change calculated relative to unsorted cells ("log(high/unsorted)"), the log fold change comparing the two most extreme sorting bins ("log(high/low)"), edgeR [11], and DESeq2 [12], looking for differential guides between the high and low expression bins, following previous studies [6, 13].

We assessed the sensitivity and accuracy of each approach by three criteria, where possible: (1) higher similarity between effect sizes of adjacent guides, which are expected to more often target the same regulatory element, compared to pairs of randomly selected guides; (2) ability to distinguish promoter-targeting guides from other targeting guides because promoter-targeting guides are a positive control expected to greatly alter expression; and (3) similarity between the effect sizes estimated for each replicate, which should be high if a method is successful.

MAUDE performed well on all three measures, outperforming the other approaches on most datasets and criteria (Fig. 3a–c). MAUDE most easily distinguished adjacent from randomly paired guides in 3 of 5 datasets (by AUROC, considering only datasets with at least one method having significant (rank sum $P < 0.01$) performance; Fig. 3a). As expected, adjacent guides had more similar effect sizes than randomly paired guides. MAUDE most easily distinguished promoter-targeting guides from other genome-targeting guides in 6 of 8 datasets (by AUROC; Fig. 3b). Finally, the estimated guide-level effects were much more highly correlated between replicates for MAUDE than for the two potential log fold change approaches for all (10 of 10) datasets (edgeR and DESeq2 do not return replicate-level statistics; Fig. 3c).

MAUDE highlights new elements regulating CD69 expression

We next asked if MAUDE yielded any new insights into CD69 regulation, which were not highlighted in the original analysis. We generated element-level statistics using a 200-bp sliding window across the locus, combining the two experimental replicates by requiring elements to have FDR < 0.01 in both replicates and consistent expression changes. The minimum element effect size that was reproducibly found by MAUDE was a change of 0.12 SDs.

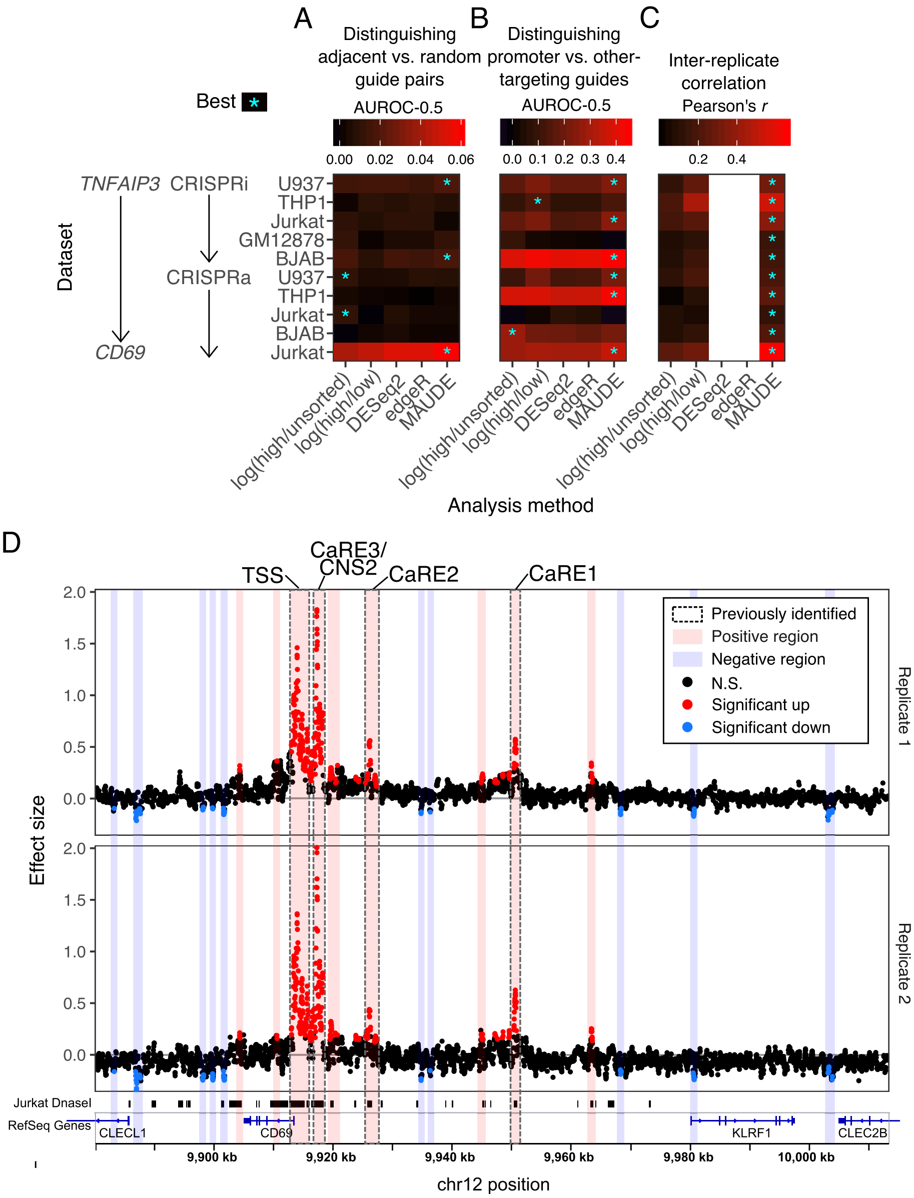

Fig. 3 MAUDE outperforms other methods and yields insight into CD69 regulation. **a**–**c** MAUDE outperforms other methods by three criteria. Comparison between methods (*x* axes) in the analysis of CRISPRi and CRISPRa screens (*y* axis), including *TNFAIP3* RNA expression-based screens [16] and a CD69 CRISPRa protein expression-based tiling screen [9]. We evaluated the performance of each method on each dataset (color; higher values correspond to better performance) and indicated the best performing approach for each dataset and criterion with an asterisk (only indicated in cases where one or more methods had a statistically significant performance ($P < 0.01$ for rank sum test (**a**, **b**) or Pearson's correlation (**c**))). Three evaluation criteria were used: **a** relative similarity in effect sizes of adjacent guides (compared to randomly selected pairs of guides, by AUROC), (**b**) ability to distinguish promoter-targeting guides from other targeting guides, and (**c**) Pearson's correlation between the effect size estimates of both replicates. Correlation between replicates was not possible for edgeR or DESeq2 as they combine replicates to estimate effect sizes and statistics. **d** Reanalysis of CD69 CRISPRa screen data with MAUDE's sliding window approach highlights additional plausible regulatory elements. Effect size (*y* axes) vs. genomic position (*x* axis) for two replicates (top and bottom), with elements (points) colored by those that significantly increased (red) or decreased (blue) expression of CD69 and those that did not (black). Bottom: gene annotations (blue) and Jurkat DnaseI hypersensitivity (black). Vertical bars: clusters of significant elements identified by MAUDE—regions that increase CD69 expression (red) and regions that decrease CD69 expression (blue), with previously identified regions indicated (gray border)

MAUDE re-discovered the four regions previously called as being CRISPRa-sensitive (Fig. 3d, gray bars), and called 15 additional regions as responsive to CRISPRa. Of these, 10 appeared to cause a downregulation of CD69 when activated by CRISPRa (Fig. 3d, red bars). Although none of these are in open chromatin regions in Jurkat cells, two are adjacent to the promoters of other nearby genes (CLECL1 and CLEC2B), suggesting these may act by competition with the CD69 promoter. Similar opposing effects on expression have been previously reported when targeting the promoters of neighboring genes with CRISPRi [8, 13]. The remaining five CRISPRa-sensitive regions we identified caused upregulation of CD69; all five of these regions overlapped with Jurkat open chromatin regions (Fig. 3d, red bars), as did the CRISPRa-sensitive regions originally identified [9]. Overall, 85% of CD69-upregulating elements identified by MAUDE were within Jurkat open chromatin and were closer to open chromatin than expected by chance (rank sum $P < 10^{-15}$; distances to open chromatin vs. distances to randomly placed open chromatin; see the "Methods" section).

MAUDE helps design bin number and width to enhance screen sensitivity

While an experimenter can set the sizes of the expression bins, finding an optimal bin configuration experimentally is laborious and costly, and typically not pursued in practice. We reasoned that our simulation framework can readily test different expression bin configurations to help choose the best one under our assumptions.

To this end, we accounted for several experimental considerations. First, cell sorters can have between two to six bins available, and the optimal bin size may differ when different numbers of bins are used. Second, the tails of the distribution are the most informative because they show the greatest ratios of effective vs. ineffective guides, and this ratio increases as bins get more extreme. However, the more extreme the bins, the fewer cells will be captured, resulting in higher sampling noise. Finally, it is desirable to have each bin capture approximately the same number of cells to facilitate uniform treatment of samples (e.g., genomic DNA isolation and library preparation).

With these considerations in mind, we tested by simulation how bin size and number altered the ability to identify differentially active elements (Fig. 4a). We altered our simulation framework to include more elements (1000 ineffective and 1000 effective) and, to make the recovery of effective elements more challenging, focused on smaller effect sizes (100 each with effect sizes from 0.01 to 0.1, at 0.01 increments). Here, the smallest effect sizes are the most difficult to recover and only the most sensitive approaches will recover them. Our bin configurations included both uniform percentile bins and more complex bin configurations, always placing bins symmetrically at both tails of the distribution, and within each tail, bins are contiguous (Fig. 4a, right). Uniform binning schedules were also tested with two and four bins, always retaining the most extreme bins. Otherwise, six bins were used. We evaluated bin configurations for their accuracy in predicting effect sizes (Pearson's r) and their sensitivity in recovering functional elements (true positives).

The best bin configurations include more bins and cover of most of the distribution. When using non-uniform bins, greater resolution (i.e., smaller bin sizes) at the tails of the distribution (i.e., A/F scheme, Fig. 4a) resulted in both greater sensitivity and accuracy than greater resolution at the inner bins (Fig. 4b, c). Having more sorting bins was

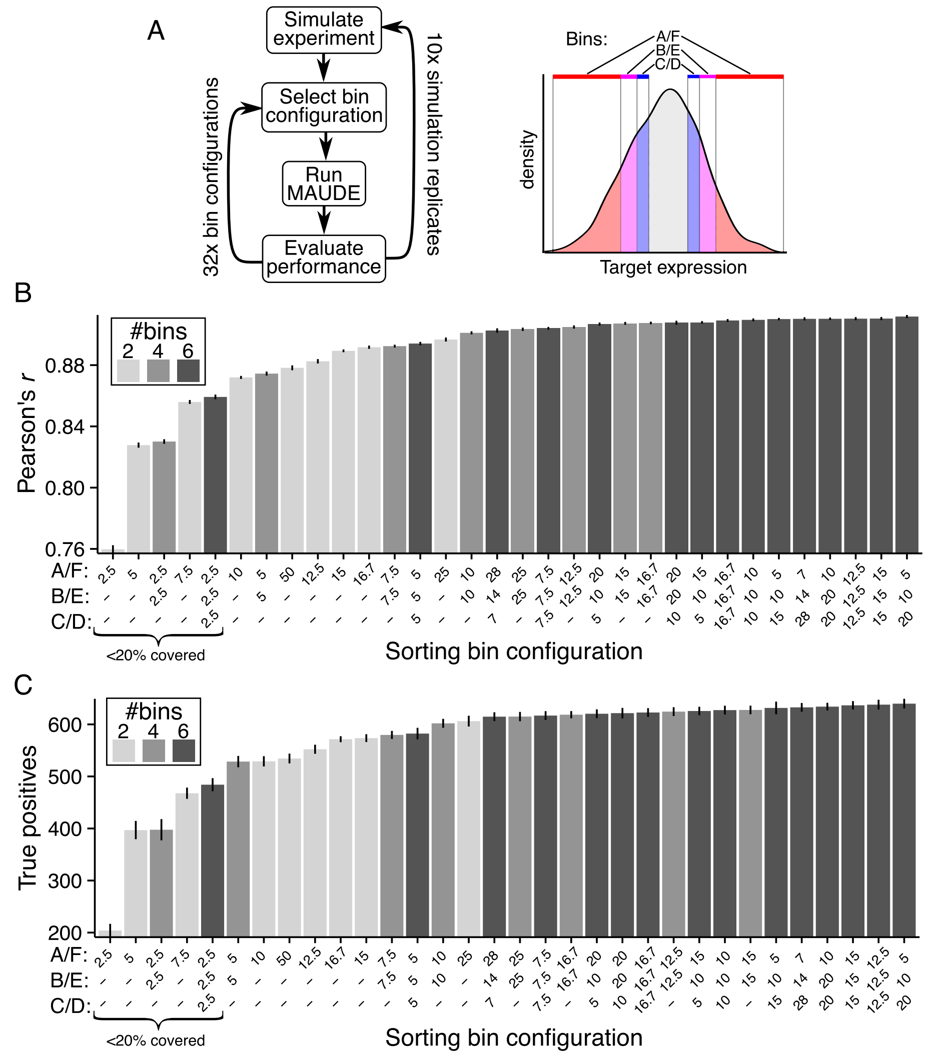

Fig. 4 MAUDE simulations propose optimal cell-sorting bin configurations. **a** Evaluation of expression bin configurations by simulation. Left: an experiment is simulated (10 replicates, top), followed by selection of each bin configuration, MAUDE analysis, and evaluation of performance for true positive recovery and effect size prediction (by Pearson's *r*). Right: each bin configuration had 2–6 bins that are contiguous on each end and symmetric about the center, progressing gradually with inclusion from outer to innermost: A and F—outermost and always included, B and E (when present)—contiguous with A and F (respectively) and more central, and C and D (when present)—contiguous with B and E (respectively) and innermost. **b**, **c** Identification of optimal bin configurations. Performance (*y* axis) of each bin configuration (*x* axis) for **b** Pearson's *r* between the true effect sizes (as defined in the simulation) and the predicted effect sizes (by MAUDE), and **c** the number of true positives (elements defined as having an effect in the simulation that are predicted to have an effect by the analysis). *x* axis labels: bin sizes expressed as the percent of the distribution covered by each of the paired corresponding bins, sorted in order of increasing performance. Error bars: standard error of the mean (SEM) for the 10 simulated experiments. Gray shading: number of bins in the configuration

always better, but particularly when bins are small, and going from two to four bins was a much more significant increase than from four to six bins (Fig. 4b, c). Overall, performance was relatively poor when bins covered less than 20% of the overall distribution (Fig. 4b, c—indicated at left). However, performance is also poor if bins are too large, presumably because resolution is reduced. For example, having two 50% bins was worse than two 25% bins, and four 25% bins was worse than four 12.5% bins (Fig. 4b, c). The best configuration tested by both measures had A and F cover 5%, B and E cover 10%, and C and D cover 20%, but uniform bins covering 12.5–15% had similar performance (Fig. 4b, c) and yield uniform cell numbers per sample, and so are more practical. Consistent with a previous simulation [17], we found that the top and bottom quartiles are optimal when only two sorting bins are available. Overall, this simulation indicates that the greatest sensitivity and accuracy is achieved by having more bins, good performance can be achieved with uniform bins, and each of the bins should cover about ~ 15% of the distribution (~ 25% if using only two bins).

Discussion

MAUDE is a principled framework for the analysis and design of pooled CRISPR screens with an expression readout. By estimating the effects of guides and elements from the distribution of guides across bins, we show that MAUDE is both sensitive and specific on simulated data and more accurate on real data, allowing the recovery of additional regulatory elements. Notably, although our sliding window approach can be used for identifying regulatory regions in a tiling experiment, MAUDE can also be used in combination with other approaches [18]. MAUDE's simulation framework can help experimental design to maximize a screen's success.

Although we focus our testing on CRISPR screens aiming to identify regulatory elements with a guide RNA perturbagen, which are the current common examples, we expect any experiment with a binned expression readout and sequencing to quantify perturbations could use MAUDE analysis. This could include perturbations of genes [6], other types of perturbagens (e.g., RNAi), direct readouts of mutation (e.g., via base editors), or even reporter assays [19].

Conclusions

MAUDE is a highly sensitive and accurate approach for identifying functional elements in a binned expression screen. MAUDE estimates the effect size on expression of individual perturbagens (e.g., CRISPR guide RNAs) and combines perturbagens to estimate the effects of genetic elements. MAUDE performs well in simulations and outperforms other methods on experimental data by all three independent evaluation criteria. Finally, by simulation, we identify which cell-sorting bin configurations are optimal.

As research focused on the function on regulatory variants and regions, which are disproportionately contributing to the genetic basis of common disease [20], we expect that pooled genetic screens with expression readouts will become much more common. Moreover, many such variants and elements are expected to have smaller effect sizes, emphasizing the need for sensitive detection and accurate effect size estimates. We anticipate that MAUDE—which we implemented as an R package available on GitHub (https://github.com/Carldeboer/MAUDE)—will become an important tool in the effort to map from variants to function.

Methods

Implementation and usage

MAUDE is implemented in R. Tutorials are provided on the MAUDE website (https://github.com/Carldeboer/MAUDE). Users provide a data.frame, with columns containing the bin counts (one column per bin, plus one for the unsorted cells), as well as columns annotating the data included in each row, including guide ID, experimental identifiers (e.g., replicate, condition, etc.), whether or not the guide is a negative control guide, and any other guide-associated data (e.g., genomic locus). Users also provide a data.frame containing the bin sizes, with one row per bin per experiment, and columns corresponding to the Z-score bounds of each bin ("binStartZ" and "binEndZ") and the corresponding bin cumulative distribution function percentiles ("binStartQ" and "binEndQ"). Using MAUDE's "findGuideHitsAllScreens" function, and providing the experimental design data.frame, read count data.frame, and bin bound data.frame, MAUDE will calculate the optimal mean expression for each guide, separately for each experiment, returning the mean guide expression ($\hat{\mu}_g$), Z-score (Z_g), and log likelihood ratio for each guide/experiment pair. By default, for each guide g and bin b, 10 pseudocounts are added to the read count ($r_{g,b}$) for every million reads of coverage in that bin (R_b) to reduce the noise resulting from poor coverage. R's "optimize" function is used for finding the guide mean expression with the highest log likelihood. The fraction of cells expected to be in a bin given the current μ_g is calculated with the "pnorm" function, scaled by the overall abundance in the library of that guide and the fraction of cells sorted into the bin, as in Eq. (2). The log likelihood for each guide/bin is calculated as using the "dnbinom" function, with x = number of reads for this guide in this bin, size = number of reads total for this bin, and prob = 1 – the fraction of this bin expected to be occupied by this guide, as in Eq. (3). The "findGuideHitsAllScreens" will return the guide-level statistics as a data.frame.

Once guide-level statistics have been calculated for each experiment, they can then be combined to obtain element-level statistics. Using MAUDE's "getTilingElementwiseStats" function, one can identify elements de novo by tiling across the region, providing the experimental design data.frame and the guide-level statistics data.frame, including columns denoting the genomic coordinates of the guide (chromosome, position). By default, all guides within each 500 bp window are combined in this sliding window approach, requiring a minimum of five guides per window (parameters which can and should be customized, dependent on the density of tiled guides), and testing all possible windows with unique guide sets. Alternatively, MAUDE provides the "getElementwiseStats" function to calculate element-level statistics given element annotations as an additional column in the guide-level statistics data.frame.

In either case, element-level statistics are calculated by combining all Z-scores for guides in the element. Three Z-scores are calculated for each element: (1) an effect size—the mean of the guides' Z-scores; (2) the Stouffer Z-score—the guides' Z-scores combined with Stouffer's method ("stoufferZ"); and (3) a significance Z-score—the Stouffer Z-score scaled using the appropriate negative control null model ("significanceZ"). To create robust null models used to scale the statistical Z-scores, we sample each negative control guide up to 10 times each. For each experiment and number of guides per element, we calculate the standard deviation of the null Stouffer Z-scores. By dividing by this standard deviation, we ensure that the null now has a standard deviation of 1, and so we can treat them as true Z-scores, calculating P values with the "pnorm" function. P values can be calculated with one tail (up-/downregulation) or two tails (either). FDRs are then calculated per-experiment using the Benjamini-Hochberg procedure.

MAUDE *Z*-scores as a percentage of normal expression

Effect sizes as estimated by MAUDE represent a deviation (Z-score) from the overall distribution. If one desires to estimate expression as a percentage of normal expression, two additional measurements are required: the mean expression of wild type (EL_{WT}) and of a null mutant (EL_{NULL}), as measured by FACS (here expressed in log space). In the case of Flow-FISH data, EL_{NULL} could be more easily obtained by performing the FISH component without the RNA-recognizing probe. Here, a log expression value x's percentage of normal expression is defined as:

$$\%EL(x) = 100\% \times \frac{\left(10^{x} - 10^{EL_{NULL}}\right)}{\left(10^{EL_{WT}} - 10^{EL_{NULL}}\right)}$$

Z-scores as output by MAUDE (Z) are linearly related to log FACS expression values (x) as follows, with the mean μ and standard deviation σ calculated from the sorting experiment FACS data:

$$x = \sigma Z + \mu$$

In practice, this μ should be very close to EL_{WT} when most guides in the library do not affect expression of the target gene.

Synthetic data generation

For each simulation, we included 1000 targeting guides (5 per element for 200 elements) and 1000 negative control guides designed to have no effect. Using a larger number of negative controls produces more robust statistics, and we chose this number to be on the low end of what other studies have used. The underlying abundance of each guide (A_g) followed a Poisson distribution with mean 1000 (representing library construction noise; $A_g \sim Pois(1000)$). The number of cells sorted for each guide (S_g) followed another Poisson distribution with a mean corresponding to the abundance of that guide within the library ($S_g \sim Pois(A_g)$). For the mean-altering simulation (Fig. 2a, c), expression levels of cells (E_{gc}) were simulated by sampling from a normal distribution with a mean corresponding to the assigned mean expression of that guide ($E_{gc} \sim Norm(\mu_g)$) for each sorted cell c. For the proportion-altering simulation (Fig. 2b, d), cells were partitioned into responders and non-responders with a probability proportional to the assigned fraction of cells whose expression is changed by the perturbation (r_g), and the expression followed a normal distribution with a mean 0 for non-responder cells and 1 for responder cells. The number of cells "sorted" into each bin (S_{bg}) using the bounds of each bin is:

$$S_{bg} = \sum^{i} Z_{b0} \leq E_{gi} \leq Z_{b1}$$

Reads were simulated with a negative binomial distribution with the number of reads for each bin equal to ten times the number of cells sorted into that bin and the probability of selecting a guide equal to its fraction of cells within that bin (Fig. 2c):

$$r_{bg} \sim NB\left(10 \times \sum^{i} S_{bi}, \frac{S_{bg}}{\sum^{i} S_{bi}}\right)$$

MAUDE, log fold change, edgeR, and DESeq2 application to CD69 data

Raw count data for the CD69 CRISPRa screen was downloaded from PubMed Central (Simeonov et al. Supplementary Table 1) [9]. *TNFAIP3* CRISPRi and CRISPRa data were downloaded from NCBI's GEO database (accession GSE136693) [16]. The log fold change values were calculated using the counts per million (CPM) values for each bin, with "log(high/low)" as $\log2((CPM_{high}+1)/(CPM_{low}+1))$. The case of "log(high/unsorted)" was calculated as $\log2((CPM_{high}+1)/(CPM_{unsorted}+1))$. For CRISPRi data, CPM_{low} and CPM_{high} were swapped since the perturbation aimed to reduce expression. EdgeR (v 3.16.1) analysis was done with default parameters by comparing high to baseline bins, for each of the two replicates, together, normalizing the data (calcNormFactors) and performing a likelihood ratio test (estimateDisp, glmFit, glmLRT), and the significance *P* values and estimated log fold change retrieved (topTags). DESeq2 (v 1.14.1) analysis was done with default parameters by creating a DESeq2 experiment using the count data, comparing the high and low bins, using both replicates (DESeqDataSetFromMatrix), and then comparing the expression between these bins (using the "DESeq" function). MAUDE analysis was performed with default parameters on the count data for each bin, including the unsorted bin. We estimated the fraction of cells in each bin for use in MAUDE analysis by reconstructing the expression distribution in Simeonov et al. Extended Data Fig. 1a [9] using the "digitize" R package [21], using the same data for both replicates. Using the actual bin proportions for each replicate would likely increase MAUDE's performance. Bin fractions for the Ray et al. *TNFAIP3* screens [16] were as provided on GEO (GSE136693).

For each evaluation criterion, we used effect sizes or significance values, as appropriate. Effect sizes were the mean of the two replicates' guide *Z*-scores (MAUDE), the estimated log fold change (edgeR and DESeq2), or the mean log fold changes for the two replicates (for log(high/low) and log(high/unsorted)). Significance values were combined replicates' *Z*-scores by Stouffer's method (MAUDE), estimated *P* values (edgeR and DESeq2), and the mean of the two replicates' log fold changes (others).

For testing whether adjacent guides have similar effect sizes, we sorted the guides by the genomic coordinates of the target site and, for each pair of adjacent guides, calculated the absolute difference in effect sizes. To create the random-pairing distribution, this was repeated, including each guide 10 times (to improve our "random" estimate), and randomizing the order of the guides. The distributions in absolute effect size differences between adjacent and randomly paired guides were compared with the area under the ROC curve statistic (AUROC), and how much it differed from that expected by chance (0.5; i.e., AUROC-0.5). Here, the AUROC is equivalent to the fraction of randomly paired guides whose absolute effect size differences are greater than those of adjacent guides.

The AUROC statistic was also used to compare the guide significance distributions between promoter-targeting guides and other genomic targeting guides, where the AUROC is proportional to the fraction of promoter-targeting guides that are more significant than non-promoter-targeting guides. Promoter-targeting guides were defined as those surrounding the *CD69* and *TNFAIP3* TSSs (between 9912997 and 9913996 on chromosome 12, and 138187040 and 138189439 on chromosome 6, respectively; hg19).

For testing the correlation between replicates, the log fold changes and MAUDE guide *Z*-scores were used directly. When there were more than two replicates, the largest correlation coefficient was used as the performance measure.

Element-level statistics for CD69 screen were calculated with the sliding window approach, using a 200-bp window and requiring a minimum of five guides per element. To estimate the significance of the overlap between CD69-activating elements and Jurkat open chromatin (DNaseI hypersensitivity), the distance distributions between open chromatin sites and MAUDE-identified CD69-activating elements were compared for actual vs. randomized open chromatin sites. The randomization placed open chromatin elements in between the starts of the first and last open chromatin sites in the locus (chr12: 9885788–9973087), accepting only random placements with no overlap between open chromatin sites, and preserving their widths. One hundred such randomizations were created, each time computing the distances between the MAUDE-identified CD69-upregulating elements and the nearest randomly placed open chromatin element. The distance distributions for the actual open chromatin data vs. the randomized open chromatin data were then compared using a two-tailed Wilcoxon rank sum test.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02046-8.

Additional file 1. Review history.

Acknowledgements

We thank Benjamin Doughty and Charles Fulco for helpful discussions.

Peer review information

Yixin Yao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 1.

Authors' contributions

CGD conceived of and implemented the algorithms with help from JPR. CGD drafted the manuscript, with input from JPR, AR, and NH. All authors read and approved the final manuscript.

Funding

CGD was supported by a Fellowship from the Canadian Institutes for Health Research and the NIH (K99-HG009920-01). JPR is funded by NIH 5 F32 AI129249. Work was supported by the Klarman Cell Observatory and HHMI (AR), the Center for Cell Circuits, a Center of Excellence in Genome Science (RM1HG006193), and the NHGRI (R01HG008131-01; NH). NH is the David P. Ryan endowed chair, and thanks Sandra, Sarah, and Arthur Irving for support.

Availability of data and materials

All data analyzed during this study are included in the supplementary information of Simeonov et al. (Supplementary Table 1) [9] and on GEO (GSE136693 [22]). MAUDE is available as an R package on GitHub [23] (https://github.com/Carldeboer/MAUDE) under the MIT License. The MAUDE release at the time of manuscript submission is available through Zenodo [24] (https://doi.org/10.5281/zenodo.3697319).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

AR is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and an SAB member of ThermoFisher Scientific, Neogene Therapeutics, Asimov, and Syros Pharmaceuticals. NH is a co-founder of Neon Therapeutics, and equity holder in BioNTech, and consultant for Related Sciences. All other authors declare no competing interests.

Author details

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA. [2]School of Biomedical Engineering, University of British Columbia, Vancouver, BC V6T 1Z3, Canada. [3]Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA. [4]Center for Cancer Research, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

Received: 25 October 2019 Accepted: 14 May 2020
Published online: 03 June 2020

References

1. Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, Cheng C, Regev A, Zhang F. High-resolution interrogation of functional elements in the noncoding genome. Science. 2016;353:1545–9.
2. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science. 2016;354:769–73.
3. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013;154:442–51.
4. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature. 2015;517:583–8.
5. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. 2015;527:192–7.
6. Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, Przybylski D, Platt RJ, Tirosh I, Sanjana NE, et al. A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. Cell. 2015;162:675–86.
7. Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, Zwart W, Elkon R, Agami R. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat Biotechnol. 2016;34:192–8.
8. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA, et al: Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51(12):1664–69.
9. Simeonov DR, Gowen BG, Boontanrart M, Roth TL, Gagnon JD, Mumbach MR, Satpathy AT, Lee Y, Bray NL, Chan AY, et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature. 2017;549:111–5.
10. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15:554.
11. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
13. Klann TS, Black JB, Chellappan M, Safi A, Song L, Hilton IB, Crawford GE, Reddy TE, Gersbach CA. CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat Biotechnol. 2017;35:561–8.
14. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
15. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
16. Ray JP, de Boer CG, Fulco CP, Lareau CA, Kanai M, Ulirsch JC, Tewhey R, Ludwig LS, Reilly SK, Bergman DT, et al. Prioritizing disease and trait causal variants at the TNFAIP3 locus using functional and genomic features. Nat Commun. 2020;11:1237.
17. Nagy T, Kampmann M. CRISPulator: a discrete simulation tool for pooled genetic screens. BMC Bioinformatics. 2017;18:347.
18. Hsu JY, Fulco CP, Cole MA, Canver MC, Pellin D, Sher F, Farouni R, Clement K, Guo JA, Biasco L, et al. CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. Nat Methods. 2018;15:992–3.
19. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 2012;30:521–30.
20. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
21. Poisot T. The digitize package: extracting numerical data from scatterplots. The R Journal. 2011;3:25–6.
22. Ray JP, de Boer C: Assessing the ability of various genomic features to prioritize causal non-coding variants associated with diseases and traits [CRISPR guide-seq]. GSE136693. Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136693 (2020).
23. de Boer C: MAUDE: Mean Alterations Using Discrete Expression. GitHub. 2019. github.com/Carldeboer/MAUDE. Accessed 25 May 2020.
24. de Boer C: MAUDE: first official release Zenodo. 2020. https://doi.org/10.5281/zenodo.3697319. Accessed 25 May 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.