Correcting signal biases and detecting regulatory elements in STARR-seq data

Young-Sook Kim, Graham D. Johnson, Jungkyun Seo, et al.

Genome Res. published online March 15, 2021
Access the most recent version at doi:10.1101/gr.269209.120

Correcting signal biases and detecting regulatory elements in STARR-seq data.

Young-Sook Kim[1,2,3,4,5] , Graham D. Johnson[1,2,3,4], Jungkyun Seo[1,2,3,4,5], Alejandro Barrera[1,2,3,4], Thomas N. Cowart[1,4], William H. Majoros[1,2,3,4,5], Alejandro Ochoa[1,2,3,4,5], Andrew S. Allen[1,2,4,5], Timothy E. Reddy[1,2,3,4,5*].

[1]Department of Biostatistics & Bioinformatics, Division of Integrative Genomics, Duke University Medical School, Durham, NC 27710, USA

[2]Center for Genomic and Computational Biology, Duke University Medical School, Durham, NC 27710, USA

[3]Center for Advanced Genomic Technologies, Duke University, Durham, NC 27710, USA

[4]Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC 27710, USA

[5]Program in Computational Biology & Bioinformatics, Duke University, Durham, NC 27710, USA

Authors' email addresses

Young-Sook Kim: yk162@duke.edu

Graham D. Johnson: grahams.mailbox@gmail.com

Jungkyun Seo: jungkyun.seo@duke.edu

Alejandro Barrera: alejandro.barrera@duke.edu

Thomas N. Cowart: thomas.cowart@duke.edu

William H. Majoros: william.majoros@duke.edu

Alejandro Ochoa: alejandro.ochoa@duke.edu

Andrew S. Allen: andrew.s.allen@duke.edu

Timothy E. Reddy (corresponding author): tim.reddy@duke.edu

Running title

Detecting regulatory elements in STARR-seq data

Keywords

Self-transcribing active regulatory region sequencing (STARR-seq), High-throughput reporter assay, Technical bias, Generalized linear model (GLM), Peak caller, Regulatory elements

Abstract

High-throughput reporter assays such as self-transcribing active regulatory region sequencing (STARR-seq) have made it possible to measure regulatory element activity across the entire human genome at once. The resulting data, however, present substantial analytical challenges. Here, we identify technical biases that explain most of the variance in STARR-seq data. We then develop a statistical model to correct those biases and to improve detection of regulatory elements. This approach substantially improves precision and recall over current methods, improves detection of both activating and repressive regulatory elements, and controls for false discoveries despite strong local correlations in signal.

Introduction

Gene regulation is of foundational importance to nearly all biological processes, and variation in gene regulatory activity plays a major role in human disease risk (Lee and Young 2013; Parker et al. 2013; Finucane et al. 2015). A major step toward measuring regulatory activity across the human genome has been the development of high-throughput reporter assays such as STARR-seq (Arnold et al. 2013) that allow regulatory element activity to be quantified with high-throughput sequencing rather than with optical detection of a fluorescent or luminescent signal.

High-throughput reporter assays create substantial analytical challenges that are distinct from other sequencing-based genomic assays. There is significant local variation in high-throughput reporter assay signal. We show here that, across data from several labs, most of that variation can be explained by features of the underlying genomic sequence and experimental procedures, rather than due to regulatory element activity. For example, nucleotide composition

can alter PCR efficiency leading to under- and over-representation of some sequences. Meanwhile, highly repetitive sequences often do not align uniquely to the human reference genome, also biasing signal estimates. Additional analytical challenge include that STARR-seq signals can be both positive and negative, reflecting activation and repression; and that the boundaries of regulatory elements are typically unknown, and must therefore be estimated from the data. Those challenges together impact signal representations, hinder estimation of regulatory element activity, and cause false positives and false negatives when left unaddressed.

Taken together, key requirements of statistical methods to analyze STARR-seq data are the ability to identify and estimate the effect of both activating and repressing regulatory elements, while also correcting for underlying sequence biases in high-throughput reporter assays. A statistical model was recently introduced that corrects technical biases and detects regulatory elements in STARR-seq but the model is limited to detecting only activating regulatory elements (Lee et al. 2020). Considering repression is crucial gene regulation mechanism (Courey and Jia 2001), overlooking repressive elements may limit understanding of gene regulation with STARR-seq. To overcome that challenge, our correcting reads and analysis of differentially active elements (CRADLE) model takes a two-step approach. First, CRADLE uses a generalized linear regression model to estimate and correct major biases that we have identified in STARR-seq data. Next, CRADLE detects regions with statistically significant regulatory activity from the bias-corrected signals while rigorously controlling FDR. In doing so, CRADLE substantially improves the use of STARR-seq by providing a robust estimation of regulatory activity and improved visualization of raw signals.

Results

DNA sequence biases STARR-seq signals

To identify sources of signal variance in STARR-seq, we analyzed data from two whole genome STARR-seq studies completed in different labs and in different human cell models: A549

(Johnson et al. 2018) and HeLa-S3 cells (Muerdter et al. 2018). Each study followed a similar protocol in which an input STARR-seq library was generated by cloning randomly fragmented genomic DNA into the 3' untranslated region (UTR) of a reporter gene. The input library was then assayed by transfecting it into cultured human cells where the cloned DNA fragments regulate their own transcription into mRNA. The expression of each random fragment as mRNA was then measured with high-throughput sequencing. Finally, regulatory activity was estimated by comparing the expression of each fragment in the output library relative to its abundance in the input library.

The STARR-seq input libraries exhibited substantially more signal variance than is observed in the controls for other genomic assays such as for ChIP-seq (Fig. 1A-B; (The ENCODE Project Consortium 2012)). That variance in STARR-seq input signal was consistent across replicates and between studies (Fig. 1C), more so than for ChIP-seq input signal (Supplemental Fig. S1). Here, we analyzed four potential sources of variance in STARR-seq signal: (i) DNA structure influencing DNA fragmentation, cloning or other enzymatic reactions, thus affecting which DNA fragments are available in the assay library (Poptsova et al. 2014); (ii) differences in the Gibbs free energy of DNA fragments influencing multiplex PCR efficiency, leading to preferential amplification of some fragments (Cheung et al. 2011; Benjamini and Speed 2012; Hansen et al. 2012; Jiang et al. 2015; Love et al. 2016; Teng and Irizarry 2017); (iii) G-quadraplexes in the genome impairing amplification by DNA polymerase (Chambers et al. 2015; Rhodes and Lipps 2015); and (iv) biases resulting from differences in the mappability of short read sequences to the reference human genome, for example due to repetitive sequences (Derrien et al. 2012).

We found evidence that each source of bias influences the signal observed when sequencing STARR-seq libraries. To model biases due to DNA secondary structure, we computationally estimated the minor groove width (MGW) and propeller twist (ProT) at the 5' ends of DNA fragments (Zhou et al. 2013). We analyzed 5' ends of DNA fragments because they are

not modified in the end-repairing process of generating STARR-seq libraries (Poptsova et al. 2014). We observed distinct biases in 5' MGW and ProT (Fig. 1D). Consistent with signal biases due to preferential fragmentation, ApG and GpG dinucleotides are most underrepresented at the 5' ends of STARR-seq fragments, which were previously reported to be less prone to shearing (Supplemental Fig. S2; (Poptsova et al. 2014)). To estimate biases due to differences in the thermodynamic stability of complementary DNA strands, G-quadruplex structure, and mappability, we binned the genome into 500bp windows. We then used data from previous studies to estimate the Gibbs free energy of the duplexed DNA strands (Protozanova et al. 2004), stability of G-quadruplex structure (Chambers et al. 2015), and the fraction of redundant mappable positions in the reference genome for each window (Derrien et al. 2012) (Fig. 1D). Fragments with the highest Gibbs free energy, highly stable G-quadruplex structure, and low mappability all had substantially depleted STARR-seq signals. Those trends were consistent across both whole genome STARR-seq studies (Johnson et al. 2018; Muerdter et al. 2018).

To evaluate whether biases in estimated Gibbs free energy are due to differences in PCR amplification efficiency, we generated DNA fragment libraries from a bacterial artificial chromosome (BAC) using between three and 18 cycles of PCR. The BAC contained 211 kb surrounding the *PER1* gene on human Chromosome 17. DNA fragments with extreme Gibbs free energy were depleted from the final library, and particularly so after 12 or more PCR cycles (Fig. 1E). That observation also indicates that signal from output STARR-seq libraries will have more severe PCR-related biases than that from input libraries due to the additional 15-16 PCR cycles used (Johnson et al. 2018; Muerdter et al. 2018); and that minimizing PCR can substantially reduce this source of bias.

Modeling biases in STARR-seq signal guides improved experimental designs.

To model the above biases in STARR-seq signal, we developed a generalized linear regression model (GLM) with covariates to model DNA structure (Zhou et al. 2013) in fragment-

ends, annealing and denaturing efficiency of DNA fragments related to their Gibbs free energy (Protozanova et al. 2004), stability of G-quadruplex structure (Chambers et al. 2015), and mappability (Derrien et al. 2012) as a reduced set of independent variables (Fig. 2A). We then fit that model to predict biases in STARR-seq signals across the genome (Fig. 2A). To improve model fit, particularly at the extremes of STARR-seq signal, we separately modeled regions with high STARR-seq signal that we observed to have significantly different coefficients for biases related to the fragment Gibbs free energy (Fig. 2B). We used a biased structured sampling approach to better fit the tails of the signal distribution (Fig. 2C; see Methods). The model fit was robust to the specific thresholds used in the biased structured sampling approach (Supplemental Fig. S3). Together, these adjustments improve model fit at the extremes of STARR-seq signal where biases are most strong and thus most likely to impact analysis.

Overall, the GLM fit the observed signals with R^2 up to 0.75 for input STARR-seq libraries (Fig. 2D). The model fit output libraries less well than the input libraries, due in part to regulatory activity also contributing to differences in STARR-seq signal. Still, the GLM showed comparable performance to using input library in predicting biases of output library, despite of the high degrees of freedom (Supplemental Fig. S4). The GLM had significantly better fit than the model which simply binned genome and used GC content in each bin as a covariate (Supplemental Fig. S5). We think the improved fit of the GLM over the simple GC-content model is partially due to the non-linear relationship of GC content and signal (Fig. 1D). Residuals from the model approximately follow a normal distribution (Fig. 2E), supporting model fit. We also estimated the extent to which each covariate independently explained variation in STARR-seq signal (Fig. 2F). Overall, the median of the explained variation across the two studies showed fragment-end sequences and Gibbs free energy explained the greatest amount of signal variation. In the data from Johnson et al. (Johnson et al. 2018), G-quadruplex bias in the input STARR-seq library and mappability bias in the output STARR-seq library had a negative marginal contribution to total predictive power but the effects were minor. Meanwhile, in the Muerdter et al. dataset (Muerdter

et al. 2018), Gibbs free energy was the major contributor to signal biases, showing relatively large variance between replicates. This shows distinctive bias effects in the two input libraries, which aligns with relatively small correlation between two signals compared to Johnson et al. (Fig. 1C). These findings indicate that most of the variance in STARR-seq signal can be attributed to technical biases; and that it is important to model distinct relative contributions of those biases in different STARR-seq library preparations.

Most of the parameters we modeled are not readily mitigated by modifying experimental procedures. As examples, reducing PCR cycles may not be feasible when template is limited, and DNA fragmentation is required for STARR-seq. Therefore, we investigated whether the GLM can instead statistically correct biases in STARR-seq signal. First, we fit the above GLM to fragment sequencing libraries generated with different numbers of PCR cycles and calculated the amount of variance explained by the GLM. Consistent with our earlier observation that additional PCR cycles increased Gibbs free energy bias (Fig. 1E), the model explained more signal variance when more PCR cycles were used (Fig. 2G). There was also a monotonic increase in the coefficients for fragment annealing and denaturing efficiency based on the Gibbs free energy (Fig. 2H). Those results demonstrate that the GLM can correct different amounts of bias resulting from different experimental designs.

Removing technical biases in STARR-seq improves visualization

Visualizing signals from functional genomic assays is often a critical step in quality control, experiment interpretation, integrative analysis, and hypothesis generation. Because substantial signal variation in STARR-seq is due to the underlying DNA sequence, however, it is challenging to gain useful information from visual inspection of uncorrected STARR-seq signals. That visualization can be substantially improved by instead using the residuals from the GLM. For example, across the two genome-wide studies analyzed here (Johnson et al. 2018; Muerdter et al. 2018), the GLM reduced signal variance by between 40 and 80%, resulting in approximately

zero-centered corrected signals (Fig. 3A-B). Further demonstrating generality across specific experimental procedures, the GLM also effectively corrected biases due to different amounts of PCR (Fig. 3C-D). An example of the resulting correction for a 19 kb region is demonstrated in Fig. 3E, where the GLM residuals allows for clearer visual identification of elements with activating or repressive regulatory effects. For example, a region near the *PER1* gene that is well-known to have activating regulatory activity in response activation of the glucocorticoid receptor (NR3C1) (Reddy et al. 2012; Johnson et al. 2018) showed much clearer indication of activity after correction (Fig. 3F). Similarly, an example of a repressive element that is bound by the REST repressor (The ENCODE Project Consortium 2012) is also better represented in corrected signals compared to uncorrected observed signals (Fig. 3G). To generalize the argument that correcting biases better reflects regulatory activity, we compared observed and corrected signal for NR3C1-binding regions and REST-binding regions that had corresponding motifs (Supplemental Fig. S6; (The ENCODE Project Consortium 2012)). Overall, corrected signal provides more stable background signal and shows clearer regulatory activity (Supplemental Fig. S6). Together, these results demonstrate that our model accounts for a substantial variation of signals in STARR-seq data and improves visualization of signals.

Correcting biases improves detection of regulatory signals embedded in STARR-seq data

To next detect genomic regions with significant STARR-seq activity, we developed a new method that rigorously models two key features of the assay. First, STARR-seq measures both activation and repression of reporter gene expression (Johnson et al. 2018), and thus being able to detect both activation and repression is important. Second, local STARR-seq signals are highly correlated (e.g. Fig. 3E). That correlation, if not appropriately considered, can lead to nonconservative control of type I errors if not modeled (Lun and Smyth 2014).

To overcome those challenges, we developed a two-step statistical approach to merge locally correlated signals while maintaining well-calibrated control of the false-discovery rate

(FDR) (Fig. 4A). Briefly, our approach first detects signals in broad genomic regions, and then identifies more specific sources of signal variation within those regions. The approach is based on previous work from Benjamini (Benjamini and Hochberg 1995; Benjamini and Bogomolov 2014). To increase power of detecting regulatory elements, we also used independent filtering to remove regions without enough signal variation to reject the null (Fig. 4A-B) (Bourgon et al. 2010).

To demonstrate the benefit of correcting technical biases when detecting regulatory elements in STARR-seq data, we simulated whole genome STARR-seq signals with embedded activating and repressive regulatory elements across a range of effect sizes (Fig. 4C-D). We then used the method described above to detect regulatory elements in corrected or uncorrected signals. When detecting regulatory elements with uncorrected signals, we used statistical tests based on a Poisson distribution to avoid unfairly reducing the performance due to violating key assumptions of a *t*-test. Specifically, we used two approaches: 1) fitting uncorrected signals to a Poisson GLM and Wald tests to reject the null ('Uncorrected 1'), and 2) using a Poisson distribution with the mean of uncorrected input signals as a null distribution and testing for a significant difference in the means of uncorrected output signals ('Uncorrected 2').

Overall, correcting biases with the GLM substantially improved the precision of detecting regulatory signals, especially at more stringent detection thresholds (Fig. 4C-D). In contrast, the majority of regulatory elements with uncorrected signals were false positives (Supplemental Fig. S7). Performance improvement was particularly pronounced when detecting repression (Fig. 4D), where the area under the precision recall curve (AUPRC) increased by 0.64 when correcting signals.

Overall, repressive signals are more difficult to detect. In the repression simulation with corrected signals, recall and precision were worse than in the activation simulation by as much as 0.43 in AUPRC. The decreased AUPRC was mainly due to small simulated output signals of repressive regulatory elements that were largely filtered out by the overall variance filter.

However, this simulation result still shows correcting technical biases helps to decrease false positives in detecting both activation and repression.

Improved detection of regulatory elements in STARR-seq data

We used the CRADLE method described above to call regulatory elements in data from two published whole genome STARR-seq studies (Johnson et al. 2018; Muerdter et al. 2018). Muerdter et al. measured differential regulatory activity in response to inhibitors that blocked interferon response (Muerdter et al. 2018). The study reported 12,010 inhibitor-responsive regulatory elements with 2,892 repressive elements, with their analysis pipeline that used binomial distribution and hypergeometric tests (Arnold et al. 2013; Muerdter et al. 2018). CRADLE detected a similar number of regulatory elements at 20% FDR (N = 11,997), 815 of which were repressive (Supplemental Table S1). While the activating elements detected by each method had overlap up to 46%, repressive elements were largely different between the methods (Fig. 4E).

To investigate the biological properties of regulatory elements detected exclusively by each method, we used motif enrichment analysis to detect potential biologically important sequence signals in the non-overlapping sets of regulatory elements. Motifs for interferon-responsive transcription factors (TFs) were most strongly enriched in the CRADLE-exclusive repressive elements (Fig. 4F, Supplemental Table S2). In contrast, activator-protein 1 (AP-1) TF motifs were most significantly enriched in the repressive elements unique to the Muerdter et al. analysis (Muerdter et al. 2018), with interferon response motifs ranked lower by enrichment (Fig. 4F, Supplemental Table S3). The motifs enriched in shared repressive regulatory elements of CRADLE and Muerdter et al. overall corresponded with the motifs enriched in CRADLE-exclusive repressive regulatory elements (Supplemental Table S2, S4). In addition, the CRADLE-exclusive repressive elements showed higher enrichment for IRF3 ChIP-seq signal (Fig. 4G). We noted that CRADLE estimated positive effects for 1,704 repressive elements uniquely detected by Muerdter et al. (Supplemental Fig. S8) (Muerdter et al. 2018), suggesting they are false positives due to

the biases in STARR-seq signal. Indeed, subsequent motif analysis of those repressive elements with positive effects revealed enriched NF-κB motifs, not corresponding to the experimental design.

The Johnson et al. study used STARR-seq to measure changes in regulatory activity in response to the dexamethasone (dex) across time (Johnson et al. 2018). The study used MACS2 (Zhang et al. 2008) and edgeR (Robinson et al. 2010) together to identify 4,835 dex-responsive regulatory elements at 0.05 FDR with 3,311 activating elements. With the data from Johnson et al. (Johnson et al. 2018), we used CRADLE to detect regulatory elements both in untreated A549 cells and in response to dex at the same 0.05 FDR (Supplemental Table S5-S6). That analysis identified 10% more dex-responsive regulatory elements (N = 5,368) than the methods used by Johnson et al. (Johnson et al. 2018), with 4,683 activating and 685 repressive dex-responsive regulatory elements (Supplemental Table S6). As with the comparison to the Muerdter et al. analysis, we observed little overlap in repressive elements while activating elements showed up to 70% overlap (Fig. 4H). Overall, those repressive regulatory elements identified by CRADLE in each study had higher control library signals than activating regulatory elements, demonstrating CRADLE requires a region to have enough coverage to be reliably detected as repressive (Supplemental Fig. S9).

To validate the newly identified repressive elements, we again used motif enrichment analysis to identify potential sequence signals consistent with repressive elements. The motif for the RE1-silencing transcription factor (REST), a well-characterized repressive factor (Chong et al. 1995), was most enriched in repressive regulatory elements in untreated A549 cells (Fig. 4I, Supplemental Table S7). Meanwhile, the motif enrichment in dex-responsive regulatory elements exclusively detected from CRADLE corresponded to previous findings about NR3C1 biology. Namely, for the dex-responsive activating regulatory elements, the NR3C1 DNA binding motif was most enriched followed by co-factor AP-1 transcription family (Supplemental Table S8), corresponding to the motifs enriched in the shared dex-responsive activating regulatory elements

(Supplemental Table S9). For the dex-responsive repressive regulatory elements, the AP-1 motif was most enriched, consistent with the role of AP-1 in NR3C1-mediated activation and repression (Supplemental Table S10; (Gupte et al. 2013; Johnson et al. 2018; McDowell et al. 2018)).

We also validated some of the 240 A549 steady-state REST-binding repressive regulatory elements using two independent studies (Supplemental Fig. S10; (van Arensbergen et al. 2019; Doni Jayavelu et al. 2020)). Though neither of these studies used A549 cells, we assumed the repressive activity of the REST-binding repressive regulatory elements could be validated in other cell models because REST is a common repressor in diverse cell lines. We intersected the REST-binding repressive regulatory elements with the regions tested by Jayavelu et al. (Doni Jayavelu et al. 2020) that used massively parallel reporter assay (MPRA) test repressive activity, and observed 30 elements in common (Supplemental Fig. S10). Of those, 27 elements (90%) had repressive activity in Jayavelu et al. while two elements did not have coverage and one element did not exhibit repressive activity. The one non-repressive element is likely due to the small overlap with their tested region (27bp) that did not cover the REST motif. We also compared our repressive element calls with the data from a genome-wide survey of regulatory elements (SuRE) signal (van Arensbergen et al. 2019). The SuRE signal in the REST-binding repressive regulatory elements exhibited repression in that study that was significantly different from both random and activator-binding elements (Supplemental Fig. S10).

Discussion

We demonstrated a substantial fraction of the variation in STARR-seq signal can be explained by DNA sequence features that are related to experimental artifacts rather than regulatory element activity. Overall, biases in PCR amplification had some of the strongest impacts on sequence biases in STARR-seq, and we show here that minimizing the amount of PCR can reduce variation in signals. DNA structure bias at the ends of fragments is possibly caused by preferential fragmentation, cloning, or efficiency as an enzymatic substrate. The

efficiency of adding adaptors in cloning or in reverse transcription could be also affected by DNA sequences or structures at fragment-ends (Zheng et al. 2011). Potential opportunities to mitigate those biases could include using multiple enzymes from different species that have different sequence biases, or further refinement of reaction conditions. Similarly, increasing read length could mitigate mappability-induced biases by decreasing the mappable space in the genome; and G-quadruplex structure bias might be alleviated by optimizing experimental conditions to de-stabilize those structures. However, mitigating technical biases using a statistical model is much faster and easier. We demonstrated the GLM had significant predictive power that led to substantially stabilized STARR-seq signals. Indeed, corrected signals showed noticeably reduced variance and improved visualization of regulatory activity.

With corrected signals from the GLM, we detected regulatory elements with substantially improved accuracy compared to previous models. CRADLE especially improved the identification of repressive regulatory elements that were challenging to detect previously, as we demonstrated via simulations, comparisons to other studies, and through investigation of DNA binding motifs for repressive factors. That improvement will allow for a more complete understanding of the diversity of regulatory element activity across the human genome.

Lee et al. also recently addressed the need to model biases in STARR-seq to improve detection of regulatory elements (Lee et al. 2020). Conceptually, both approaches model physical characteristics of genomic sequence that substantially influence STARR-seq signal, and develop novel peak calling approaches. In terms of implementation, there are differences in model parameters (e.g. how CRADLE models PCR biases), model fitting (e.g. weighted sampling of the tails of the coverage distribution), and peak calling methods. In terms of performance evaluation, we evaluated CRADLE over a broader range of simulations, and in doing so we demonstrated that the reported False Discovery Rates are well-calibrated. We also showed that CRADLE is especially able to detect repressed regulatory elements.

Our work on CRADLE also opens up the possibility of developing analogous statistical models for other high-throughput sequencing technologies. Many high-throughput sequencing technologies share common experimental steps that cause technical biases in STARR-seq. In that regard, CRADLE exemplifies how those biases can be statistically modeled and corrected, thus allowing effect estimation and peak calling from data with a mean of zero. Of course, each sequencing technology may have assumptions that are distinct from STARR-seq and have other major bias effects that were not modeled in CRADLE. For example, antibody specificity might be one of major bias sources in ChIP-seq. More studies need to be done to determine the scope of the applicability of CRADLE.

Methods

Downloaded data

For STARR-seq data, we downloaded FASTQ files of whole genome STARR-seq that used A549 and HeLa-S3 cells from Johnson et al. (Johnson et al. 2018) and Muerdter et al. (Muerdter et al. 2018), respectively. Those files were downloaded from NCBI Gene Expression Omnibus (GEO) repository (Barrett et al. 2013) with accession codes available in those studies (GSE114063, GSE100432).

For ChIP-seq data (The ENCODE Project Consortium 2012; Davis et al. 2018), we downloaded ChIP-seq FASTQ files with following GEO accession codes: GSE91296 for A549 control ChIP-seq; GSE91275 for A549 0hr-dex-treated NR3C1 ChIP-seq; GSE91235 for A549 12 hr-dex-treated NR3C1 ChIP-seq; GSE92032 for HeLa-S3 control ChIP-seq; GSE101280 for A549 REST control ChIP-seq; GSE101362 for A549 REST ChIP-seq; GSM935570 for HeLa-S3 IRF3 ChIP-seq; GSM935339 for HeLa-S3 IRF3 control ChIP-seq.

Processing of high-throughput sequence data

FASTQs files were aligned to the human genome reference assembly hg38 with Bowtie 2 (version 2.3.4.3; (Langmead and Salzberg 2012)), using the --sensitive option and requiring a MAPQ of at least 30. Fragments were discarded if they are aligned to gap, centromere, and telomere that are available in UCSC Gap and Centromere table browser (Hinrichs et al. 2006) and ENCODE blacklist regions (Amemiya et al. 2019). Alignment of paired-end datasets were further restricted to require properly paired alignments. Unnormalized and RPKM-normalized (--binSize 1) bigWig files were generated by bamCoverage subcommand in deepTools (version 3.0.1; (Ramirez et al. 2016)) using --extendReads. The reported average fragment length was used to extend reads when generating single-end bigWigs. Unnormalized and normalized bigWig files were used for CRADLE inputs files and for visualizing signals in genome browser tracks, respectively. A549 ATAC-seq FASTQs were processed as above but were aligned to hg19 and required a less stringent MAPQ score (≥5). Peaks were called for ChIP-seq datasets using MACS2 (Zhang et al. 2008) with a FDR threshold of 0.05. For NR3C1-binding sites, we first called peaks using MACS2 (Zhang et al. 2008), independently for 0 hr-dex-treated NR3C1 ChIP-seq samples and 12 hr-dex-treated NR3C1 ChIP-seq samples with respective control ChIP-seq samples. Then we merged those peaks and used edgeR (Robinson et al. 2010) to perform differential testing at FDR 0.05 and selected peaks with positive effect size to detect NR3C1-binding regions. The coordinates of autosomal inhibitor-responsive regulatory elements previously reported in hg19 (Muerdter et al. 2018) were converted to hg38 with liftOver (Hinrichs et al. 2006).

PER1 BAC library preparation and sequencing

Purified *PER1* bacterial artificial chromosomes (BACs) (CH17-212C17; Chr17:7,981,103-8,192,310) were harvested from *E. coli* using standard protocols. Following DNA shearing using the Covaris S2 instrument, the BAC DNAs were size-selected using solid phase reversible immobilization (SPRI) beads. STARR-seq insert libraries were prepared using the NEBNext DNA Library Prep Master Mix kit and 50 ng of template DNA. Adapted DNAs were enriched in triplicate

reactions via 3, 6, 12, or 18 cycles of PCR using the NEB Q5 PCR kit. The resulting libraries were characterized on the Agilent Tape Station prior to 50-cycles of paired-end sequencing on the Illumina MiSeq platform. FASTQs were aligned as above. We checked duplicated rate for each cycle using Picard (MarkDuplicates, version 2.14.0). The mean duplicate rate for each cycle is as follows: 0.3% in Cycle 3; 0.5% in Cycle 6; 1.3% in Cycle 12; 1.9% in Cycle 18.

Data processing for bias covariates

To obtain DNA structure parameters for fragment-end bias, we estimated minor groove width (MGW) and propeller twist (ProT) for all 5-mers (total 1,024 sequences) using DNAshape (Zhou et al. 2013). For Gibbs free energy parameters, we used the estimated Gibbs free energy for all dimers (Protozanova et al. 2004). For G-quadruplex structure parameters, we used bigWig files that reported stability of G-quadruplex structure in whole genome with accession code GSE63874 in GEO (Chambers et al. 2015). For mappability scores, we downloaded the human mappability score bigWig files for 36-mer and 50-mer (Derrien et al. 2012), using accession codes (ENCSR821KQV, ENCSR093EEM) in ENCODE (The ENCODE Project Consortium 2012; Davis et al. 2018). For those G-quadruplex structure bigWig and mappability bigWig files, the genomic coordinates were in hg19 assembly so we used liftOver tool (Hinrichs et al. 2006) to convert them to hg38.

Measuring technical biases in STARR-seq libraries

To investigate fragment-end bias, we counted the frequency of 5-mers starting 2 bp upstream of the 5’ end of positive strands of fragments in STARR-seq input libraries (Johnson et al. 2018; Muerdter et al. 2018). To identify enriched fragmentation sites, we compared that observed 5-mer frequency distribution to that observed in the reference genome (hg38) excluding gaps, centromeres, telomeres that are available in UCSC Gap and Centromere table browser (Hinrichs et al. 2006), and ENCODE blacklist regions (Amemiya et al. 2019). To examine Gibbs free energy,

G-quadruplex structure, and mappability bias, we binned human Chromosome 1 into 500 bp windows using a 250 bp stride. We estimated the amount of potential technical bias in a window by calculating the mean of per-base measure of those biases using previously reported values: Gibbs free energy value (Protozanova et al. 2004); the percent of mismatch for G-quadruplex structure bias (Chambers et al. 2015); mappability score (Derrien et al. 2012). This analysis was limited to the *PER1* BAC when estimating Gibbs free energy bias in the *PER1* BAC library.

Correcting technical biases in STARR-seq

We used the technical bias covariates in a general linearized model (GLM) with a Poisson distribution and log link to correct STARR-seq signals. An estimate of the 90th percentile of observed coverage in input libraries ($Isig_{P90}$) was calculated using 1 kb bookended regions. To ensure the GLM models effects across the range of observed signals, we trained the model using a structured sampling strategy to select bookended regions without replacement such that the final training set is approximately 10^6 bases in length. We evenly partitioned the training set to fit regions with input signal above and below $Isig_{P90}$. The set of regions below $Isig_{P90}$, were further evenly partitioned into the following percentile bins of observed coverage: [0, 20); [20, 40); [40, 60); [60, 80); [80, 90). To ensure representation across the upper tail of the STARR-seq signal distribution, regions above $Isig_{P90}$ were asymmetrically partitioned as follows: 62.5% of regions were evenly divided into the following percentile bins of observed coverage, [90, 92); [92, 94); [94, 96); [96, 98); [98, 99), while the remaining 37.5 % of regions were binned into the 99th percentile of coverage. With Muerdter et al. et al. data, we empirically found preferentially sampling Chromosome X in the last two bins improved performance.

Single base positions with observed input signal ($Isig_{pos}$) above and below $Isig_{P90}$ were independently fit to the GLM. To predict the total bias effects at each single base position we used windows twice the length of the median fragment length (*L*) centered on the position of interest

(Fig. 2A). We assumed each position was covered by *L* number of hypothetical fragments of *L* bp length with each overlapping by a single base. We then multiplied the same bias covariates for all fragments in that window with each covariate to the power of a unique beta as below.

$$\begin{aligned} Observed\ signal_{pos} \sim \prod_{i=pos-L+1}^{L} & MGW_i^{\beta_1 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_1' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)} \\ & \cdot ProT_i^{\beta_2 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_2' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)} \\ & \cdot Anneal_i^{\beta_3 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_3' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)} \\ & \cdot Denature_i^{\beta_4 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_4' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)} \\ & \cdot Gquad_i^{\beta_5 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_5' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)} \\ & \cdot Map_i^{\beta_6 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_6' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)} \end{aligned} \quad (1)$$

Each beta coefficient represents the relative effect of each bias predictor. Here, we assumed the set of betas is the same for all overlapping fragments. Then in log space, observed signal can be estimated with using the sums of bias covariates in the GLM as follows.

$$\begin{aligned} \log\left(E\left(Observed\ signal_{pos}\right)\right) = & \left(\beta_1 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_1' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)\right) \sum_{i=pos-L+1}^{L} \log(MGW_i) \\ & + \left(\beta_2 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_2' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)\right) \sum_{i=pos-L+1}^{L} \log(ProT_i) \\ & + \left(\beta_3 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_3' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)\right) \sum_{i=pos-L+1}^{L} \log(Anneal_i) \\ & + \left(\beta_4 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_4' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)\right) \sum_{i=pos-L+1}^{L} \log(Denature_i) \\ & + \left(\beta_5 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_5' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)\right) \sum_{i=pos-L+1}^{L} \log(Gquad_i) \\ & + \left(\beta_6 \cdot I\left(Isig_{pos} < Isig_{P90}\right) + \beta_6' \cdot \left(1 - I\left(Isig_{pos} < Isig_{P90}\right)\right)\right) \sum_{i=pos-L+1}^{L} \log(Map_i) \end{aligned} \quad (2)$$

MGW and ProT values were calculated using DNAshape (Zhou et al. 2013) for the two 5-mers starting from two bp external to both 5' ends of each hypothetical fragment. MGW_i and $ProT_i$. was obtained by multiplying those two MGW and ProT values, respectively.

We used the the nearest-neighbor model (SantaLucia 1998; Protozanova et al. 2004) to estimate the Gibbs free energy of each hypothetical fragment. To estimate the relative melting temperature

(T_m) of each fragment, we divided Gibbs free energy of a hypothetical fragment by the number of dimers in that hypothetical fragment and by the fixed entropy value (Protozanova et al. 2004). T_m values were normalized to range of [0, 1]. To model the non-linear dependency of annealing and denaturing efficiencies to T_m, normalized T_m values in the i th fragment ($T_{m,i}$) were mapped to two exponential functions as below.

$$Anneal_i = \left(e^{T_{m,i}} - \frac{10^6 - e}{10^6 - 1}\right) / \left(\frac{10^{-6} - 1}{1 - e}\right)$$
$$Denature_i = \left(e^{-T_{m,i}} - \frac{10^6 - e}{10^6 - 1}\right) / \left(\frac{10^{-6} - 1}{1 - e}\right) \quad (3)$$

Those mapped values were used for $Anneal_i$ and $Denature_i$ in the GLM Model.

To obtain $Gquad_i$ for each hypothetical fragment, we used the maximum of G-quadruplex structure stability value in that sequence (Chambers et al. 2015). To obtain Map_i for each hypothetical fragment, we used a k-mer mappability score file (Derrien et al. 2012) where k-mer is a sequencing length, and multiplied the mappability scores of both ends of a fragment.

After fitting the GLM, the bias predicted by the model at each base position was removed by subtracting the estimated bias effect from the observed signal. To avoid false positives, positions with fewer than ten observed overlapping fragments and with no signal in output libraries are not reported in the corrected signal file. The minimum number of observed overlapping fragments required for a position to be reported is parameterized (-mi) in the correctBias subcommand in CRADLE.

Modeling technical biases in STARR-seq libraries with only GC content

To show our sophisticated approach of modeling bias has better fit than simply modeling GC content, we took a following approach. We binned genome with non-overlapping sliding window with six different window sizes ranging from 10bp to 1000bp. We randomly selected approximately

10^6 bases in length for a train set. Then with the train set, we calculated GC content in each bin of which size corresponded to the chosen window size and used the GC content as a covariate in fitting GLM with Poisson distribution and log link. We independently fitted each replicate in the GLM. Then we used the resulting coefficients (intercept and the coefficient of GC content) to estimate bias impact for the regions that were not in the train set.

Normalizing signals in STARR-seq

To make the corrected signals from the GLM comparable between replicates, for example by correcting for overall differences in sequencing, we normalized STARR-seq signals between replicates using linear regression. We used the training set sampled as mentioned above and regressed per-nucleotide signal from each library against the a common replicate of the input library. We estimated the slope in the linear regression and divided observed signals in each library by that slope estimate.

Evaluating model fit

To determine how well the CRADLE GLM explained variance in observed signal, we calculated R^2 with observed and predicted signals across Chromosome 1 for each STARR-seq library. To calculate R^2, we fitted the GLM in CRADLE, and calculated the sum of squares (SSQ) by adding up squared residuals. Then, we calculated total SSQ, the sum of squared difference of observed signals and the mean signal. With using the SSQ and total SSQ, we calculated R^2 with the following equation: $R^2 = 1 - (SSQ/totalSSQ)$.

Evaluating the contribution of each covariate to model fit

To estimate the contribution of each bias type, we calculated semi-partial correlations for Chromosome 1 using the GLM. To assess each technical bias type, we excluded bias covariates that model corresponding bias type in fitting the GLM. For example, we excluded ‘anneal’ and

'denature' covariates when assessing Gibbs free energy bias impact. The R^2 of these models were calculated as above and subtracted from the R^2 of the full model.

Calling regulatory elements with CRADLE

Genomic regions possessing regulatory activity were identified using a modified Benjamini method (Benjamini and Bogomolov 2014). We first binned the genome into windows (1.5x *L*) and determined the effect size of each window by subtracting the mean corrected signal in input libraries from the mean corrected signal in output libraries. Each window was classified to one of the three types, using the following standard:

$Type(window_x) = 1$ if (effect size > 0 and |effect size| > 99th percentile of absolute effect sizes)

$Type(window_x) = -1$ if (effect size < 0 and |effect size| > 99th percentile of absolute effect sizes)

$Type(window_x) = 0$ else

The threshold of the 99th percentile of absolute effect sizes was chosen to classify windows because the majority of windows are not expected to encode regulatory activity. Contiguous windows of the same type, including Type 0, were merged to form regions for statistical testing. These regions were then binned with non-overlapping bins of which length is 1/6x *L*. In each bin, the input and output STARR-seq signals were compared using Welch's *t*-test (Welch 1947) to account for potential differences in variance. Individual bin-level *P*-values from the same region were merged to a region-level *P*-value via the Simes' method (RJ 1986). To increase our power to detect potential regulatory regions for final testing, regions with small overall variance were removed from further analysis, independently of the statistical test used (Bourgon et al. 2010). Specifically, we ranked regions according to their overall variance and then applied the overall variance filter that removed 0-90% of regions with low variance using 10% intervals. *P*-values for regions passing each threshold were subjected to the first BH procedure (Benjamini and

Hochberg 1995) using a parameterized FDR value (-fdr). Then, we chose the threshold of the overall variance filter that returned the greatest number of selected regions from the first BH procedure. To identify bins that have regulatory activity, bin-level *P*-values from the Welch's *t*-test in the selected regions were then subjected to the second BH procedure with new FDR adjusted by following equation:

new FDR = (pre-determined FDR) x (the number of selected regions / total number of regions)

Contiguous bins that encode regulatory activity with the same sign of effect sizes were merged in the final output and the minimum *P*-value was reported.

Simulation of STARR-seq signals

To evaluate the performance of CRADLE, we simulated STARR-seq signals that maintained the observed sequence biases and expected variance across replicates. STARR-seq signals were simulated using a negative binomial distribution and mean-variance relationships estimated independently for input and output libraries from previously published STARR-seq data (Johnson et al. 2018). Simulated input and output signal matrices, generated using 300 bp bookended windows along Chromosome 1, were used to estimate mean-dispersion relationship in DESeq2 (Love et al. 2014) prior to interpolation with the Scipy.interpolate.interp1d command in Python. To generate a set of pre-defined regulatory elements (N = 50,504), we randomly sampled ~0.5% of total windows requiring that the selected windows to be in at least the 70th percentile of coverage in the published input libraries. For each pre-defined regulatory element, we randomly assigned an absolute fold change [2, 3, or 4] and regulatory activity type [activating or repressing]. Pre-defined sets of regulatory elements with a specific fold change and regulatory activity type were generated as above using the specified fold change and regulatory activity types as described in text.

Five simulated STARR-seq input and output signals were generated using a negative binomial distribution. The mean parameters used to generate the simulated input and output signals were

determined by calculating the mean window counts using the published input libraries (Johnson et al. 2018). The variance parameters were determined using either the input or output interpolation analyses described above. The mean parameters used to generate the simulated output signals were adjusted for pre-defined regulatory elements windows by multiplying or dividing the mean signal by the pre-determined fold change and determining the corresponding variance parameter.

Detecting regulatory elements in simulated data

To evaluate the effect of correcting STARR-seq signals on identifying regulatory elements, we used CRADLE to call regulatory elements before and after correcting biases in the simulated datasets. Due to the normality assumption in Welch's *t*-test, we modified the CRADLE approach described above to call regulatory activity in uncorrected simulated signals. In place of the Welch's *t*-test, we used two alternative statistical approaches to compare uncorrected simulated input and output signals. First, we used a Poisson GLM as follows:

$$\log(E(signal_i)) = \beta_0 + \beta_1 \times (data\ type_i) \qquad data\ type_i = \begin{cases} 0 & if\ signal_i\ is\ from\ input\ library \\ 1 & if\ signal_i\ is\ from\ output\ library \end{cases} \quad (4)$$

We then performed the Wald test for β_1 with f distribution. Second, we followed a similar approach as used by MACS2 (Zhang et al. 2008). We used the mean input bin signal as the mean parameter in a Poisson distribution to calculate a *P*-value for the mean output bin signal. We called regulatory activity in corrected simulated signals as described above.

Motif enrichment analysis

Motif enrichment analysis was performed using the findMotifsGenome subcommand in the HOMER (Heinz et al. 2010) 4.10.1 software suite using the following parameters: -size given -mis 3 -mset vertebrates.

Plotting Heatmaps

Heatmaps were plotted using deepTools command “computeMatrix” using a reference-point and “plotHeatmap” (Ramirez et al. 2016). In both cases, we specified single nucleotide resolution sing the option --binSize 1.

TF occupancy in CRADLE regulatory elements

To determine whether regulatory elements called by CRADLE are bound by TFs, we used the CRADLE pipeline to detect A549 steady-state activating and repressive regulatory elements from a previously published study (Johnson et al. 2018). We used the findMotifsGenome subcommand in the HOMER suite (version 4.10.1; (Heinz et al. 2010)) with the following parameters, -size given -mis 0 -mset vertebrates -find, to detect REST motifs in each repressive element and FOSL2, JUNB, and GABPA motifs in activating elements. For each element that encoded a specified motif, we intersected those elements with ENCODE ChIP-seq peaks for the corresponding TF in A549 cells (The ENCODE Project Consortium 2012; Davis et al. 2018).

Validation of REST occupied repressive regulatory elements

Regions tested by Jayavelu et al. in K562 cells (N = 7,440) (Doni Jayavelu et al. 2020) were intersected with repressive regulatory elements identified by CRADLE in A549 cells that also contained a REST motif and were bound by REST in the same cell line (N = 240) (The ENCODE Project Consortium 2012). Reported fold change values in K562 cells were compared for the intersection set except the two elements without coverage (N = 28), regions predicted by Jayavelu et al. to be repressive elements (N = 3,001), and control regions (N = 40).

Signals from genome-wide Survey of Regulatory Elements (SuRE) in HepG2 and K562 cells (van Arensbergen et al. 2019) were compared in specific sets of regulatory elements identified by CRADLE in A549 cells. These regulatory elements included activating regulatory elements that contained either a FOSL2, GABPA, or JUNB motif and were bound by the corresponding TF in

A549 (The ENCODE Project Consortium 2012) or repressive elements that likewise contained a REST motif and were bound by REST (N = 240; (The ENCODE Project Consortium 2012)). A549 regulatory elements that contained a SNP in the genomes assayed in the SuRE study were excluded on a per genome basis. The minimum and maximum number of SNP-filtered elements compared for each TF are as follows: FOSL2 N=650-651; GABPA N=401-402; JUNB N=723; REST N=102.

We randomly generated a set of regions (N = 240) of fixed length (430 bp) controlling for accessibility (The ENCODE Project Consortium 2012) and dinucleotide composition. The fixed length was set to the median length of the compared repressive elements. In generating random regions, we excluded regions that overlapped gaps, centromeres, and telomeres that are available in UCSC Gap and Centromere table browser (Hinrichs et al. 2006) and ENCODE blacklist regions (Amemiya et al. 2019), or the following features defined by ChromHMM (Ernst and Kellis 2017) in K562 and HepG2 cells: promoters, promoter flanking regions, enhancers, CTCF enriched sites, and repressed regions. After applying the SNP-filter described above, we obtained 94 random regions.

Data access

The *PER1* BAC datasets generated in this study have been submitted to NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE149914.

CRADLE is implemented in Python and the source code is available in the Supplemental code. CRADLE can be freely downloadable either from GitHub (https://github.com/ReddyLab/CRADLE) or pip (pip install cradle). Instructions for installing and running CRADLE are available on the CRADLE GitHub page.

Acknowledgments

We thank Greg Crawford and David MacAlpine for their helpful comments and advice in developing this work.

We also thank funding from NIH grants R01HD085227 (T.E.R), UM1HG009428 (Y..K, J.S., A.B., W.H.M., A.S.A., T.E.R), UM1HG009428 (A.B., T.E.R), and R01DK104927 (A.B., T.E.R.); G.D.J. was supported by NIH fellowship F32DK115188.

Author Contributions

Conceptualization and project administration was by Y.K. and T.E.R.; data curation, formal analysis, investigation, validation, and visualization was by Y.K.; supervision and funding acquisition was by T.E.R.; methodology was by Y.K., T.E.R, G.D.J., J.S., W.H.M, A.O. and A.S.A.; software development was by Y.K., T.N.C. and A.B.; writing of the original draft was by Y.K. and T.E.R; and review and editing was by all of the authors.

Disclosure declaration

The authors declare that they have no competing interests.

References

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**: 9354.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074-1077.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M et al. 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**: D991-995.

Benjamini Y, Bogomolov M. 2014. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society* **76**: 297-318.

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.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**: e72.

Bourgon R, Gentleman R, Huber W. 2010. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A* **107**: 9546-9551.

Chambers VS, Marsico G, Boutell JM, Di Antonio M, Smith GP, Balasubramanian S. 2015. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**: 877-881.

Cheung MS, Down TA, Latorre I, Ahringer J. 2011. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res* **39**: e103.

Chong JA, Tapia-Ramirez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G. 1995. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**: 949-957.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Courey AJ, Jia S. 2001. Transcriptional repression: the long and the short of it. *Genes Dev* **15**: 2786-2796.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**: D794-d801.

Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* **7**: e30377.

Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. 2020. Candidate silencer elements for the human and mouse genomes. *Nat Commun* **11**: 1061.

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**: 2478-2492.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**: 1228-1235.

Gupte R, Muse GW, Chinenov Y, Adelman K, Rogatsky I. 2013. Glucocorticoid receptor represses proinflammatory genes at distinct steps of the transcription cycle. *Proc Natl Acad Sci U S A* **110**: 14616-14621.

Hansen KD, Irizarry RA, Wu Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**: 204-216.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576-589.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**: D590-598.

Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. 2015. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res* **43**: e39.

Johnson GD, Barrera A, McDowell IC, D'Ippolito AM, Majoros WH, Vockley CM, Wang X, Allen AS, Reddy TE. 2018. Human genome-wide measurement of drug-responsive regulatory activity. *Nat Commun* **9**: 5317.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Lee D, Shi M, Moran J, Wall M, Zhang J, Liu J, Fitzgerald D, Kyono Y, Ma L, White KP et al. 2020. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol* **21**: 298.

Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell* **152**: 1237-1251.

Love MI, Hogenesch JB, Irizarry RA. 2016. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat Biotechnol* **34**: 1287-1291.

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

Lun AT, Smyth GK. 2014. De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res* **42**: e95.

McDowell IC, Barrera A, D'Ippolito AM, Vockley CM, Hong LK, Leichter SM, Bartelt LC, Majoros WH, Song L, Safi A et al. 2018. Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Res* **28**: 1272-1284.

Muerdter F, Boryn LM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR et al. 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**: 141-149.

Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, van Bueren KL, Chines PS, Narisu N, Black BL et al. 2013. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* **110**: 17921-17926.

Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, Polozov RV, Nechipurenko YD, Grokhovsky SL. 2014. Non-random DNA fragmentation in next-generation sequencing. *Sci Rep* **4**: 4532.

Protozanova E, Yakovchuk P, Frank-Kamenetskii MD. 2004. Stacked-unstacked equilibrium at the nick site of DNA. *J Mol Biol* **342**: 775-785.

Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160-165.

Reddy TE, Gertz J, Crawford GE, Garabedian MJ, Myers RM. 2012. The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. *Mol Cell Biol* **32**: 3756-3767.

Rhodes D, Lipps HJ. 2015. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**: 8627-8637.

RJ S. 1986. An improved Bonferroni procedure for multiple tests of significance. Vol 73, pp. 751-754. Biometrika.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

SantaLucia J, Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* **95**: 1460-1465.

Teng M, Irizarry RA. 2017. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. *Genome Res* **27**: 1930-1938.

van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, van der Weide RH, Teunissen H, Vosa U, Franke L et al. 2019. High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet* **51**: 1160-1169.

Welch BL. 1947. The generalisation of student's problems when several different population variances are involved. *Biometrika* **34**: 28-35.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zheng W, Chung LM, Zhao H. 2011. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* **12**: 290.

Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* **41**: W56-62.

Figure legends

Figure 1. Technical biases affect STARR-seq signal. (A) STARR-seq input libraries have higher signal variance than ChIP-seq input control libraries. Variance in per base signal in individual RPKM-normalized libraries are plotted for Chromosome 1. The error bars indicate variance between replicates. The number of replicates plotted is as follows; six replicates for STARR-seq A549 data; two replicates for STARR-seq HeLa-S3data; three replicates for ChIP-seq A549 data; two replicates for ChIP-seq HeLa-S3 data; three replicates for ChIP-seq LNCaP data. (B) Representative browser signal tracks are shown for STARR-seq and ChIP-seq input libraries (Chr1:11197048-11236707). Signals are RPKM-normalized. (C) Pearson's correlations of STARR-seq input library signals in 1 bp windows along Chromosome 1. (D) DNA sequence biases impact STARR-seq signals. STARR-seq signals are plotted for 500 bp windows with varying degrees of bias for the following physical properties of DNA: fragment-end DNA structures, Gibbs free energy, G-quadruplex structure, and mappability. Whiskers extend 1.5 times the interquartile range. Center lines in the boxes show the medians. In plots of fragment-end bias, minor groove width (MGW) and propellar twist (ProT) are plotted and the ideal is y=0. In plots of other biases, the ideal line is the median signal. (E) PCR amplification introduces bias into STARR-seq libraries. Impact of Gibbs free energy bias is shown for *PER1* BAC libraries amplified with different number of PCR cycles (3, 6, 12, and 18 cycles). Each point represents the

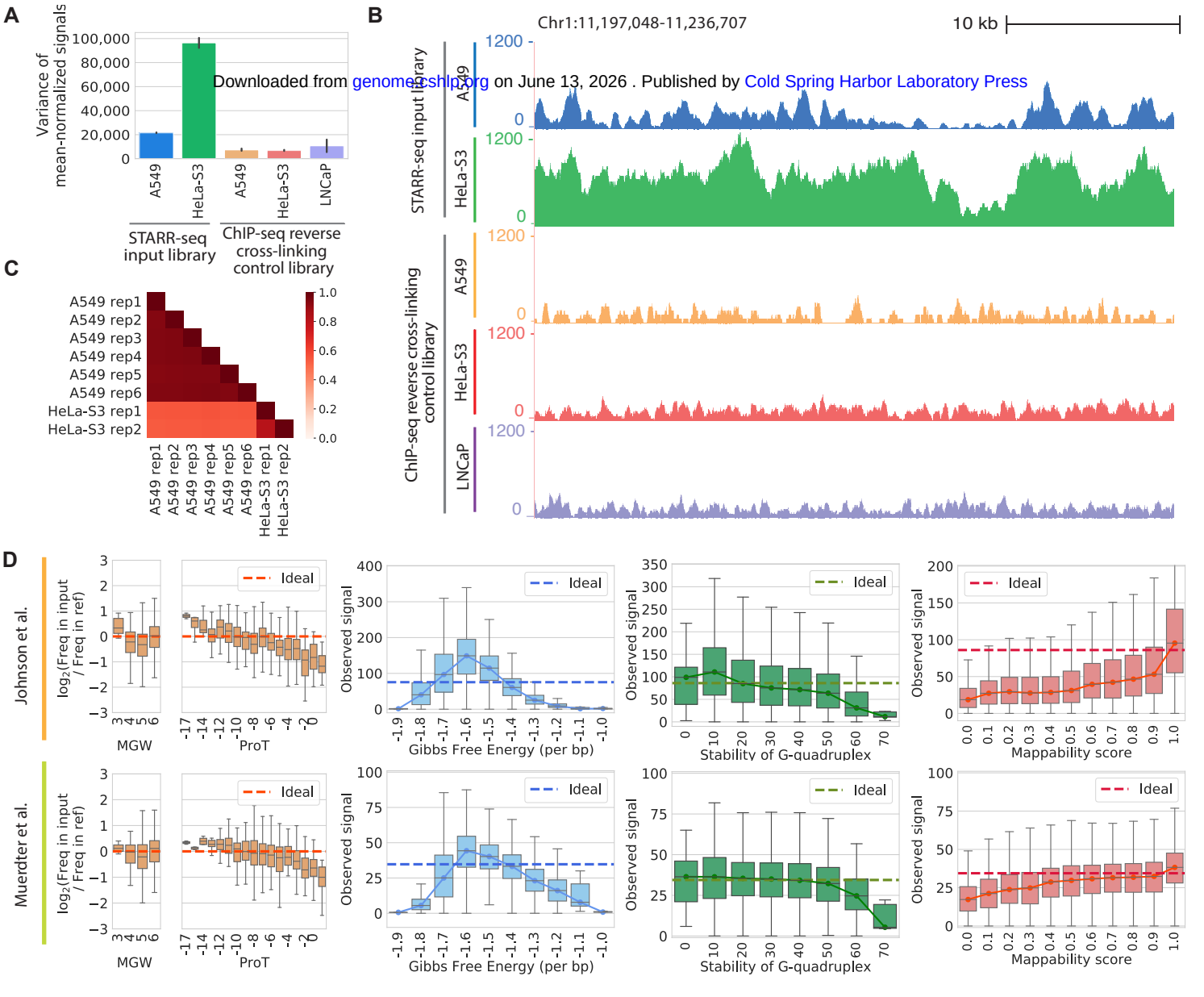
sum of signals in a 500 bp window from three technical replicates. The solid line is a lowess fit line. The dashed ideal line is the median signal across all windows.

Figure 2. The CRADLE GLM approach accurately predicts signal bias. (A) Equation of the GLM to predict the impact of technical biases and approach used by CRADLE to calculate bias covariates. To estimate bias effects for each position (blue), we used a window centered on that position that was twice the median fragment length, L. We assume L number of fragments (green) in a window and that each fragment is L-bp in length. To calculate each bias covariate for the position, we combined quantitative measures from L fragments. Notation: pos, single-bp position; MGW_{pos}, minor grove width; $ProT_{pos}$, propeller twist; $Anneal_{pos}$, annealing efficiency; $Denature_{pos}$, denaturation efficiency; $Gquad_{pos}$, G-quadruplex structure; Map_{pos}, mappability. (B)-(F) The results from the GLM fitted with Johnson et al. STARR-seq data (six input libraries and five 0hr-dex-treated output libraries) and Muerdter et al. STARR-seq data (two input libraries and two no-inhibitor-treated output libraries). For (C)-(F), the results were visualized for Chromosome 1. (B) Coefficients in input libraries for regions with signals above and below 90th percentile ('Regions with high input signal' and 'The rest of regions', respectively). (C) Ratio of the sum of squared errors with structured sampling to the sum of squared errors with random sampling are plotted for regions with extremely high signals (above 99th percentile). (D) Variance explained by CRADLE are plotted. The R^2 values are from GLMs fitted with input and output STARR-seq libraries. The error bars indicate variance between replicates. (E) Distribution of GLM residuals and the STARR-seq effect sizes are shown after correction. (F) Squared semi-partial correlations are shown for fragment-end, Gibbs free energy, G-quadruplex, and mappability covariates. The error bars indicate variance between replicates. (G) The R^2 values of the GLMs are shown for *PER1* BAC libraries amplified with different number of PCR cycles. (H) Coefficients

of anneal and denature covariates are shown for the GLM fitted with *PER1* BAC libraries. The error bars show 95% confidence interval.

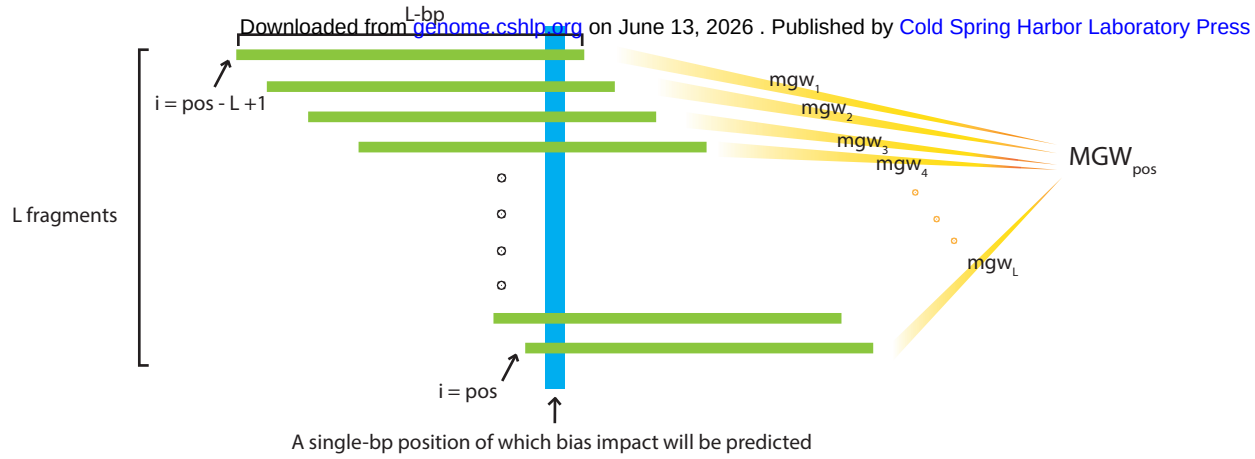
Figure 3. The CRADLE GLM approach corrects technical bias. (A) STARR-seq signals are plotted for 500 bp windows along Chromosome 1 after removing technical bias with CRADLE. Signal is balanced despite varying degrees of technical biases. The ideal line is the median corrected signal. Whiskers extend 1.5 times the interquartile range. Center lines in the boxes show the medians. (B) Variance in observed signals and CRADLE corrected signals in 1 bp windows are shown along Chromosome 1. The error bars indicate variance between replicates; six input libraries and five 0hr-dex-treated output libraries in Johnson et al. are plotted.; two input libraries and two no-inhibitor-treated output libraries in Muerdter et al. are plotted. (C) STARR-seq signals are shown for *PER1* BAC libraries amplified with different number of PCR after removing technical bias with CRADLE. Each point represents the sum of corrected signals in a 500 bp window from three technical replicates. The solid line is a lowess fit line. The dashed ideal line is the median signal across all windows. (D) Variance in observed signals and CRADLE corrected signals in 1 bp windows are shown after correcting the *PER1* BAC libraries. (E) Representative signal tracks are shown for STARR-seq input libraries before and after CRADLE correction (Chr2:29772197-29791543). (F) STARR-seq and ChIP-seq signal tracks are shown in the dex-responsive *PER1* locus. Observed and corrected signal of Johnson et al. are presented for 0hr-dex-treated (untreated) and 12hr-dex-treated output libraries. ChIP-seq signal tracks are not corrected. The highlighted region (Chr17:8151204-8152809) is a known dex-responsive activating regulatory element. (G) STARR-seq signal tracks are shown for the TMEM63C locus. Observed and corrected signal of Johnson et al. are presented for input and 0hr-dex-treated output libraries. The highlighted region (Chr14:77207895-77210261) contains a REST motif and is occupied by REST in multiple cell types.

Figure 4. Detection of regulatory elements with CRADLE. (A) CRADLE regulatory element pipeline is shown in diagram. Effect sizes are calculated in windows of uniform length. Contiguous windows with similar effect sizes are merged into regions prior to filtering regions with small variance. Regions are binned and a statistical test is performed on each bin to compare corrected input and output signals. Bin-level *P*-values are merged to generate a region-level *P*-value prior to performing a region-level Benjamini-Hochberg (BH) procedure. Regions selected by the first BH procedure were used to perform a bin-level BH procedure to identify regulatory elements. (B) The number of detected regulatory elements is dependent on the variance filter. (C)-(D) Precision-recall curves, using corrected and uncorrected signals in the simulation study. To detect regulatory elements with uncorrected signals, two statistical approaches were used: 1) fitting uncorrected signals to Poisson GLM and performing Wald test ('Uncorrected 1') and 2) Using a Poisson distribution with the mean of uncorrected input signals as a null distribution and testing the significance of the mean of uncorrected output signals ('Uncorrected 2'). (C) Precision-recall curve when signals are simulated with mixed fold change (2, 3, 4) and a mix of activating and repressive elements. (D) Precision-recall curve when signals are simulated with a fixed fold change (FC) and with a fixed regulatory activity (either activating or repressive). (E)-(G) Comparison of inhibitor-responsive regulatory elements detected by CRADLE and Muerdter et al. (E) The venn diagram shows the overlap of regulatory elements detected by both studies. (F) Transcription factor motif enrichment is shown for inhibitor-responsive repressive regulatory elements exclusively detected by each study. Rank* is the rank of motif in the other study. (G) The mean of IRE3 ChIP-seq effect size is plotted for inhibitor-responsive repressive regulatory elements exclusively detected by each study. (H) The venn diagram shows the overlap of dex-responsive activating and repressive regulatory elements detected by CRADLE and Johnson et al. (I) Transcription factor motif enrichment in A549 steady-state repressive regulatory elements detected by CRADLE.

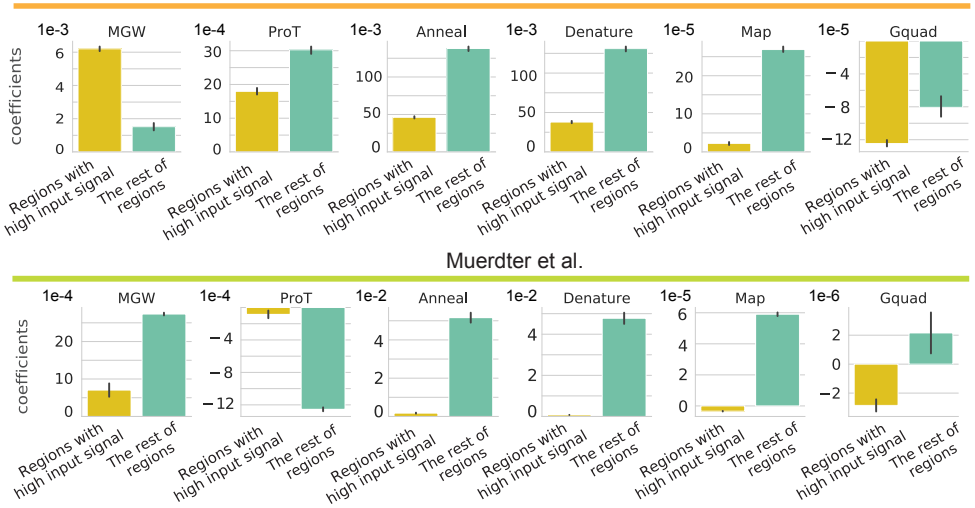


A

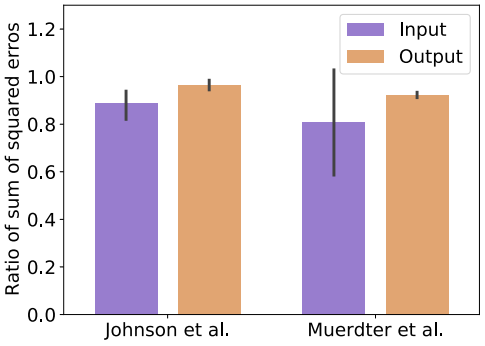
$$\log\left(\mathrm{E}(Observed\ signal_{pos})\right) = \beta_0 + \beta_1 \cdot MGW_{pos} + \beta_2 \cdot ProT_{pos} + \beta_3 \cdot Anneal_{pos} + \beta_4 \cdot Denature_{pos} + \beta_5 \cdot Gquad_{pos} + \beta_6 \cdot Map_{pos}$$



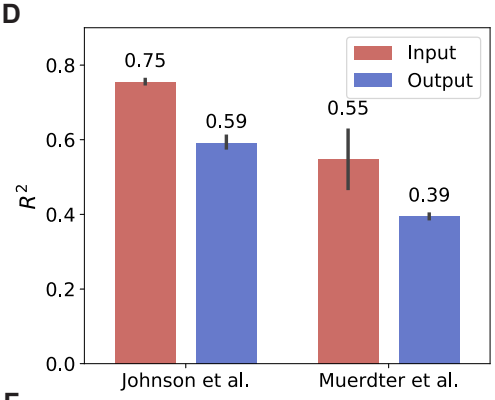
B



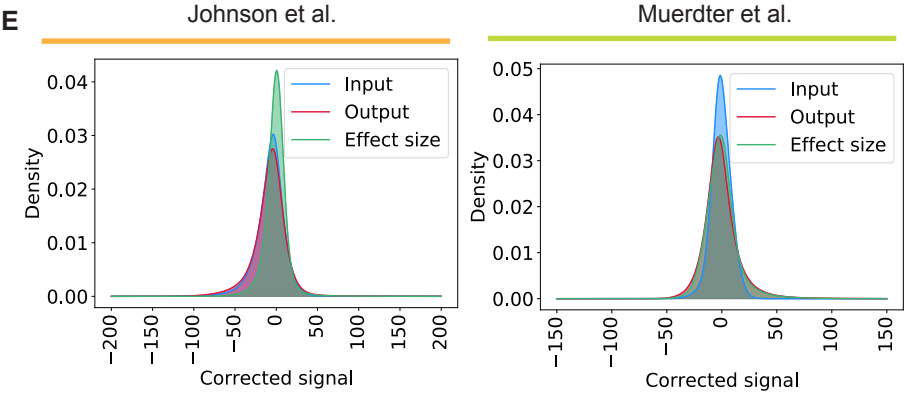
C



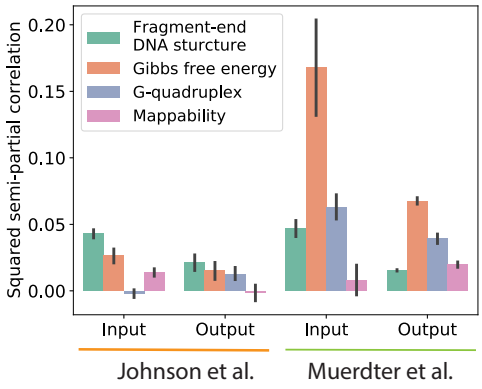
D



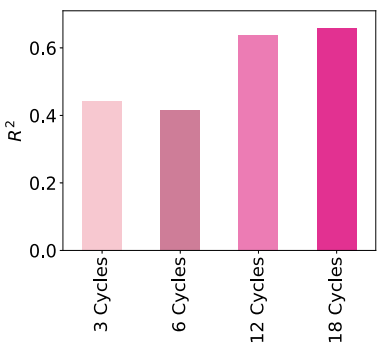
E



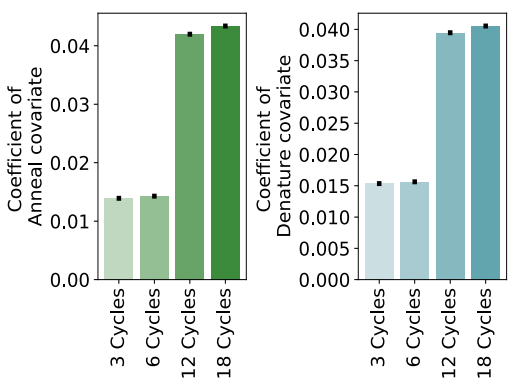
F

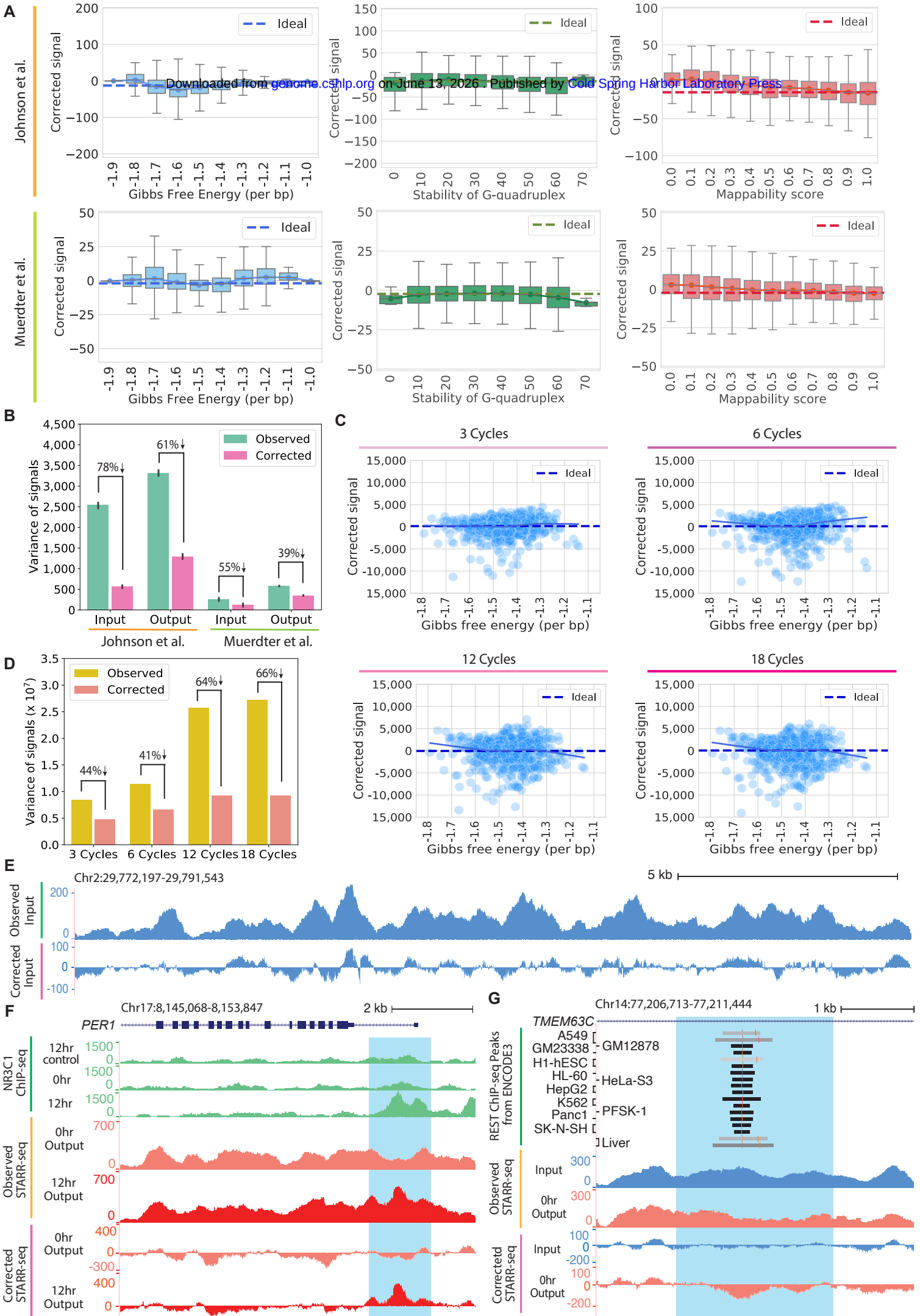


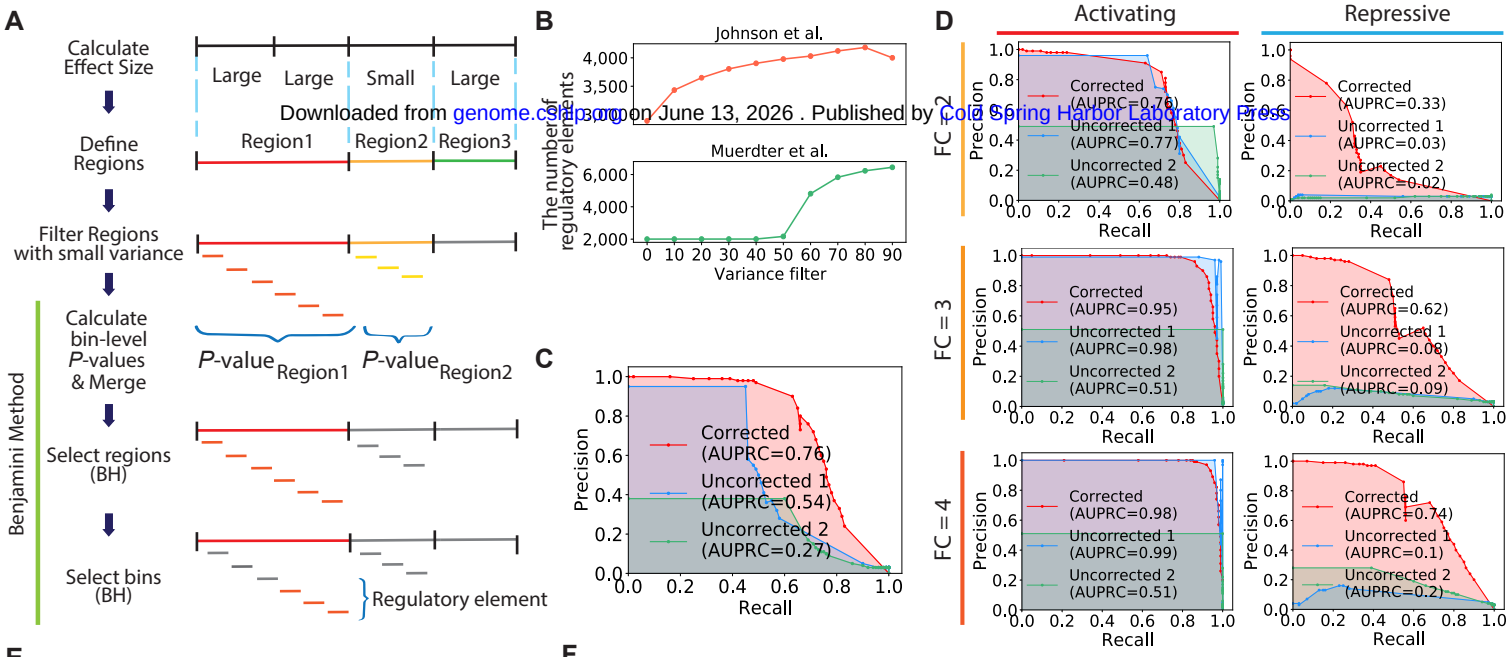
G

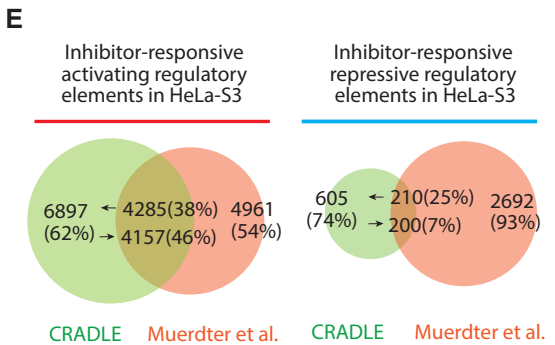


H









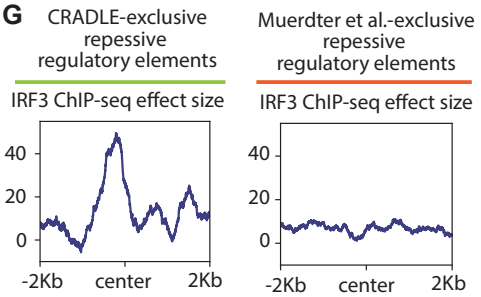
F

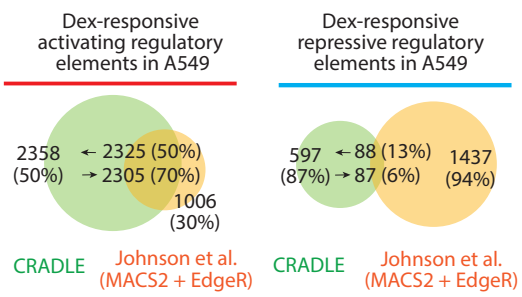
CRADLE-exclusive repressive regulatory elements

Rank	Motif	Symbol	P-value	Rank*
1		ISRE	10^{-46}	6
2		IRF2	10^{-41}	9
3		IRF3	10^{-36}	11
4		IRF1	10^{-33}	12
5		IRF8	10^{-31}	13
6		SPI1:IRF8	10^{-8}	33

Muerdter et al.-exclusive repressive regulatory elements

Rank	Motif	Symbol	P-value	Rank*
1		JUN	10^{-214}	48
2		FOSL2	10^{-210}	22
3		FRA2	10^{-192}	47
4		JUNB	10^{-177}	67
5		FRA1	10^{-175}	57
6		ISRE	10^{-165}	1





I

Repressive elements in the untreated A549 cells

Rank	Motif	Symbol	P-value
1		REST	10^{-102}
2		MYC	10^{-39}
3		EPAS1	10^{-29}
4		MYOD1	10^{-28}