CRISPR-Decryptr reveals cis-regulatory elements from noncoding perturbation screens

Anders Rasmussen[1,*], Tarmo Äijö[1,*], Mariano Ignacio Gabitto[1], Nicholas Carriero[3], Neville Sanjana[4], Jane Skok[5], Richard Bonneau[1,2,5,**]

[1] Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY 10010, USA.
[2] New York University, Center for Data Science, New York, NY 10010, USA.
[3] Scientific Computing Core, Flatiron Institute, Simons Foundation, New York, NY, 10010, USA.
[4] New York Genome Center, New York, NY, 10013, USA.
[5] New York University, Department of Biology, New York, NY 10012, USA
* These authors contributed equally to this work
** To whom correspondence should be addressed

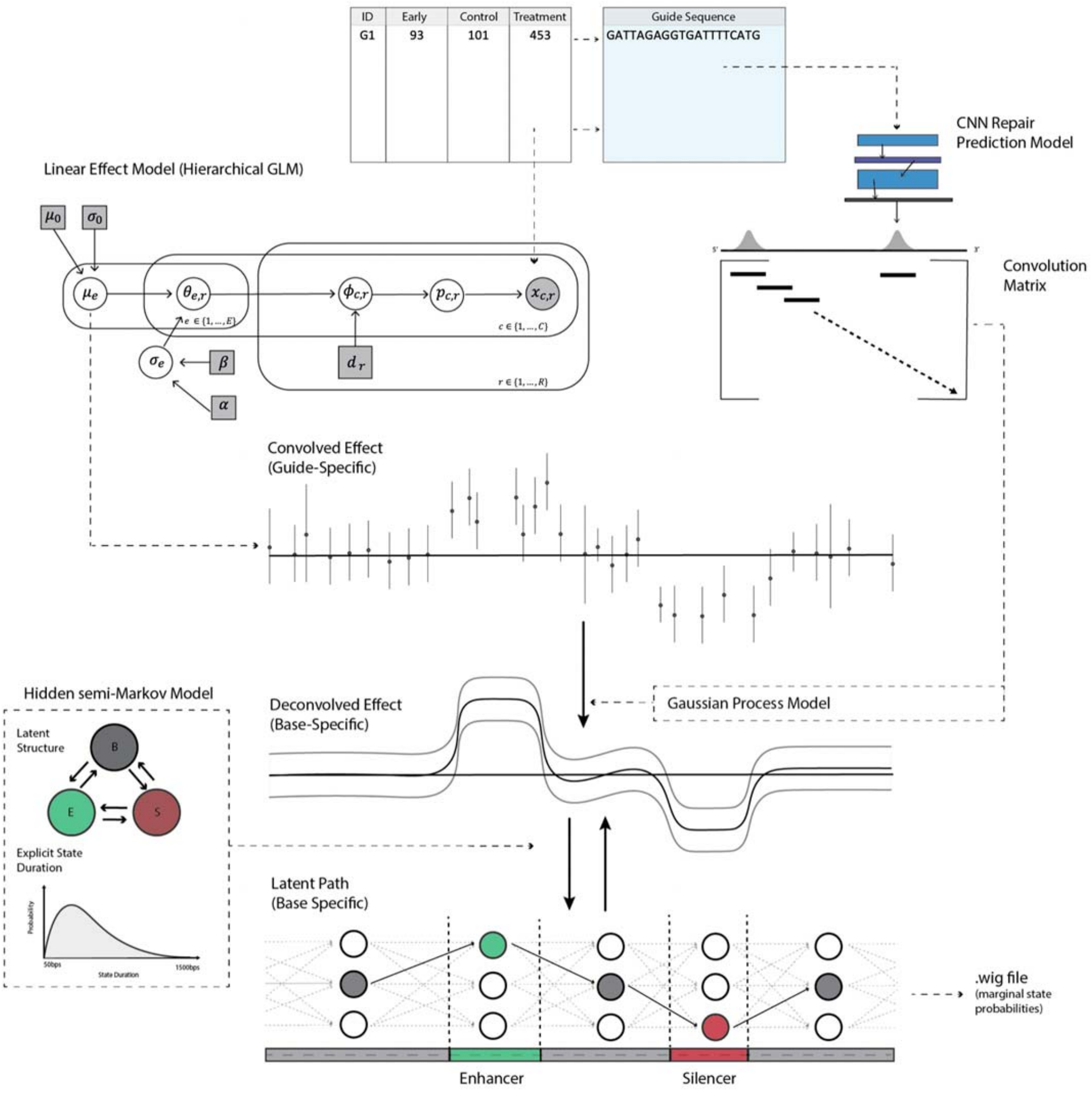
Abstract:

Clustered Regularly Interspace Short Palindromic Repeats (CRISPR)-Cas9 genome editing methods provide the tools necessary to examine phenotypic impacts of targeted perturbations in high-throughput screens. While these technologies have the potential to reveal functional elements with direct therapeutic applications, statistical techniques to analyze noncoding screen data remain limited. We present CRISPR-Decryptr, a computational tool for the analysis of CRISPR noncoding screens. Our method leverages experimental design: accounting for multiple conditions, controls, and replicates to infer the regulatory landscape of noncoding genomic regions. We validate our method on a variety of mutagenesis, CRISPR activation, and CRISPR interference screens, extracting new insights from previously published data.

Main:

Information garnered from pooled CRISPR perturbation screens impacts decisions that have therapeutic implications. Genome-wide knockout and noncoding screens have been used to identify new therapeutic targets, to reveal genes responsible for anti-cancer drug resistance, and to map functional elements in leukemia cell lines.[1,2,3,4] As researchers in academia and industry make greater use of improving gene editing technologies, computational approaches that tackle the unique challenges posed by their experimental design are of increasing importance. Methods employed for knockout screens are designed to assess the impact of perturbing a genome-wide set of pre-delineated coding regions. [5,6] However, analysis of CRISPR noncoding screens, which employ saturated guide libraries to reveal *cis*-regulatory elements, necessitate distinct experimental considerations. Most importantly, classification of functional elements without *a priori* knowledge of their location or size requires integrating information across perturbations within genomic proximity, an aspect that renders existing knockout methods inapplicable to these experimental designs. Literature on methods for analyzing noncoding screens is scarce, with only a single method published that addresses one of the many aspects of noncoding screen analysis.[7]

CRISPR-Decryptr utilizes techniques from Bayesian inference, signal processing, and latent variable models to integrate data and experimental design, allowing the end-user to make precise conclusions about their noncoding screen results (**Figure 1**; *Methods*). A Bayesian hierarchical generalized linear model (GLM) serves as the mathematical formulation from which perturbation-specific effect on phenotype are inferred[8, 9]. The model leverages experimental conditions, controls, and replicates in a single numerical procedure implemented with Markov Chain Monte Carlo, allowing for rigorous statistical treatment of parameter uncertainty (*Methods 2.2*). Effects are mapped to a base-by-base level of granularity through a Gaussian process-based model (*Methods 2.4*)[10]. This deconvolution fully accounts for guide-specificity, off-target effects and, if applicable, double-strand break (DSB) repair uncertainty (*Methods 2.3*)[11, 12]. A hidden semi-Markov model (HsMM) incorporates spatial information to decode the latent regulatory landscape of interest, revealing enhancers and silencers in the noncoding genome (*Methods 2.5*)[13]. Regulatory element calls and guide-specific effects are exported in bioinformatics file formats such as Browser Extendable Data (.bed) and Wiggle (.wig) that can easily be explored in genomic visualization software such as the Integrative Genomics Viewer (IGV)[14].



ID	Early	Control	Treatment
G1	93	101	453

Guide Sequence
GATTAGAGGTGATTTTCATG

Figure 1: Overview of the CRISPR-Decryptr method for the analysis of noncoding screens. The hierarchical GLM infers pertubation-specific regulatory effect on phenotype from raw guide RNA (gRNA) counts (*top left*). Guide RNA sequences are used to construct a convolution matrix accounting for specificity, off-target effects, and repair uncertainty in the case of mutagenesis screens (*top right*). Finally, iterating between Gaussian Process deconvolution and HsMM training and prediction reveals base-specific effects and ultimately the latent state path of interest (*bottom half*).

We validated CRISPR-Decryptr on noncoding screens of distinct experimental designs, including CRISPR mutagenesis, CRISPR activation (CRISPRa), and CRISPR interference (CRISPRi) screens (**Figure 2**) [4, 15, 16]. In the CRISPR mutagenesis screen (Canver *et al.*), three intronic DNAse hypersensitivity sites (DHS) within *BCL11A* were perturbed in human umbilical cord blood-derived erythroid progenitor (HUDEP) cells[15]. These sites, termed DHS +62, +58, and +55, are known to impact fetal hemoglobin (HgF) levels from prior published research, with the enhancer identified in DHS +58 having proven a successful therapeutic target in two patients with hemoglobinopathies.[17] When applied to this dataset, CRISPR-Decryptr produced regulatory state calls in agreement with the original analysis (**Figure 2A** and **Supplementary Figure 3.1.1**). The CRISPR activation screen we re-analyzed (Simionov *et al.*) targeted the *IL2RA* and *CD69* gene loci in Jurkat T-cells.[16] To measure phenotypic change, the FACS sort cells into a "negative", "low", "medium", and "high" bins of IL2RA and CD69 based on expression levels. Analysis of the two gene loci with CRISPR-Decryptr recalls the enhancers from the original analysis, as well as novel putative enhancers that are correlated with DNAse-seq and H3K27ac from the Jurkat-T Cell line (**Supplementary Figures 3.2.2** and **3.2.3**). Finally, the re-analysis of the Fulco et al. CRISPRi screen of the GATA1 gene loci revealed similar regulatory element calls to the original analysis.[4] (**Figure 2C** and **Supplementary Figure 4.3.1**).

We have described a statistical technique for analyzing CRISPR noncoding screen data and illustrated the accuracy of CRISPR-Decryptr on three distinct perturbation technologies, demonstrating the method's ability to reveal novel insights from a diverse set of experimental designs. CRISPR-Decryptr will be a valuable component in future attempts to identify functional genomic elements and their link to phenotypic traits, enabling target identification and synthetic biology in biomedical and biotechnological settings.

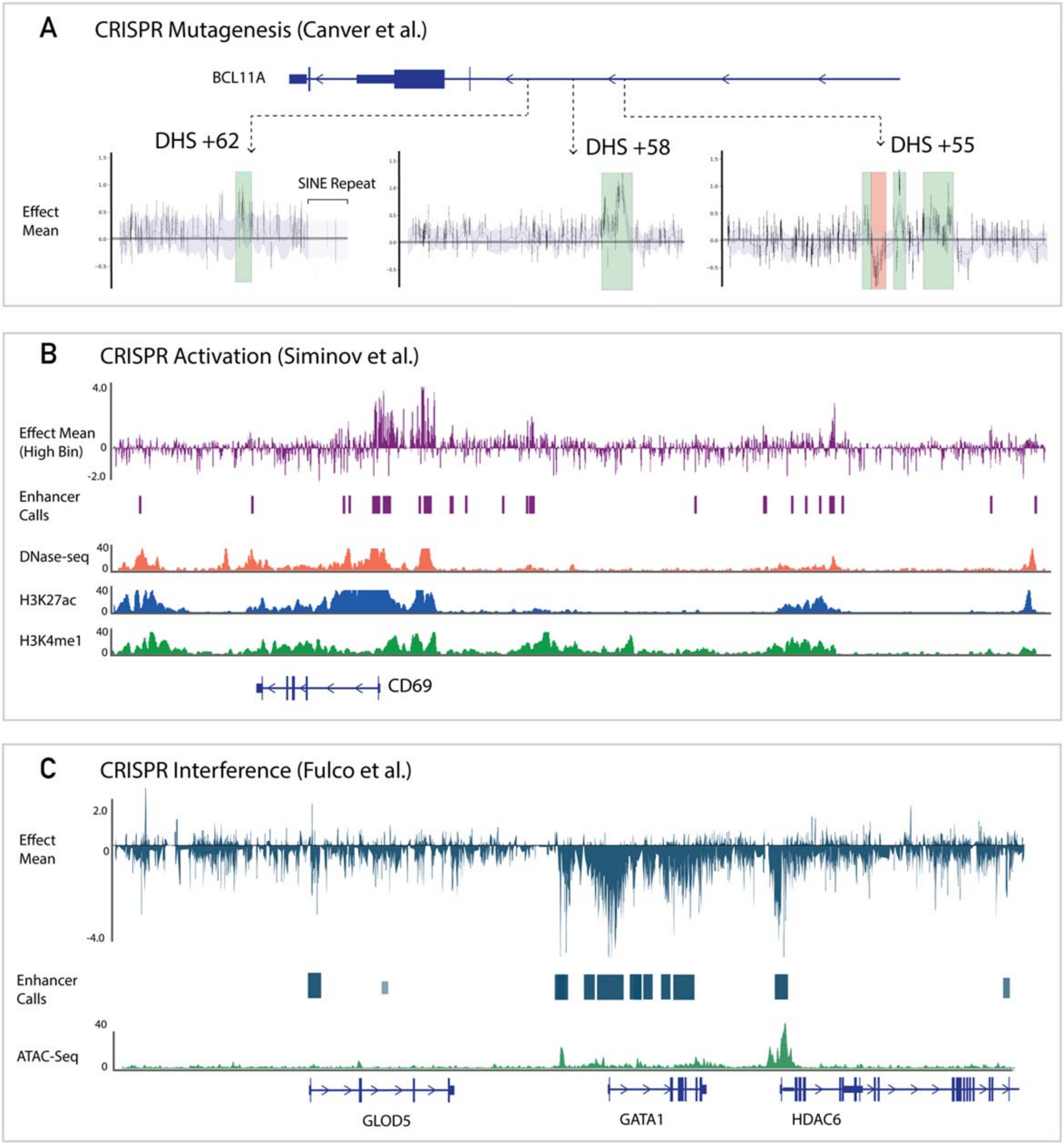


Figure 2: Regulatory elements classified by CRISPR-Decryptr for three published noncoding screens. **A**: Analysis mutagenesis screen targeting BCL11A DHS sites reveals similar enhancer and silencer locations as in the original publications. **B***:* Analysis CRISPRa screen targeting CD69 promoter region reveals novel enhancer calls. **C**: Analysis CRISPRi screen targeting GATA1 gene loci reveals the same enhancer calls as in the original analysis.

Code Availability:

CRISPR-Decryptr code and readme are located at: https://github.com/anders-w-rasmussen/crispr_decryptr

Data Availability:

All data is available at https://github.com/anders-w-rasmussen/crispr_decryptr

No restrictions on data are applicable here. All data used were publicly available.

Acknowledgements:

This research was made possible by the Simons Foundation. RB and MG acknowledge support from the following sources: NIH R01DK103358, Simons Foundation, NSF-IOS-1546218, R35GM122515, NSF CBET-1728858, and NIH R01AI130945

Contributions:

A.R. conceived of the model with guidance and oversight from R.B. and T.Ä. The inference step of the method and Gaussian Process deconvolution were formulated and coded by T.Ä. Off-target, repair outcome prediction, and the Gaussian Process / HsMM iterative procedure were formulated by A.R. The majority of HsMM code is adapted from previous work by done by M.I.G and A.R. in developing the ChromA algorithm. Rules for HsMM parameter updates and variational methods were developed by M.I.G. N.C. was instrumental in advising on high-performance computing considerations. N.S. and J.S. provided important insight into noncoding screens from the viewpoint of experimentalists and were central in bringing the need for this statistical method to A.R. and R.B.'s attention. A.R. did analyses of published data, wrote the paper supplement, and wrote the CRISPR-Decryptr software. All authors contributed to the writing of the manuscript.

Competing Interests:

A.R. owns stock in Editas medicine and 10x Genomics. T.Ä. owns stock in 10x Genomics.

R.B. has ongoing or recent consulting or advisory relationships with Eli Lily, Merus, Merck and Epistemic AI.

References:

[1] Wei, L., Lee, D., Law, C. *et al.* Genome-wide CRISPR/Cas9 library screening identified PHGDH as a critical driver for Sorafenib resistance in HCC. *Nat Commun* **10,** 4681 (2019).

[2] Lau, M., Ghazanfar, S., Parkin, A. *et al.* Systematic functional identification of cancer multi-drug resistance genes. *Genome Biol* **21,** 27 (2020).

[3] Fellmann, C., Gowen, B., Lin, P. *et al.* Cornerstones of CRISPR–Cas in drug discovery and therapy. *Nat Rev Drug Discov* **16,** 89–100 (2017).

[4] Fulco, C.P, Munschauer, M. Anyoha, R. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).

[5] Li, W., Xu, H., Xiao, T. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15,** 554 (2014).

[6] Allen, F., Khodak, A. Behan, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Research* **29**, 464-471 (2019)

[7] Hsu, J.Y., Fulco, C.P., Cole, M.A. *et al.* CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. *Nat Methods* **15,** 992–993 (2018).

[8] Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models* (Cambridge university press, Cambridge, 2006).

[9] Gelman, Andrew, et al. *Bayesian data analysis* (CRC press, Boca Raton, FL, 2013).

[10] Rasmussen, C.E., Williams, C.K. *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA, 2006)

[11] Hsu, P., Scott, D., Weinstein, J. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31,** 827–832 (2013).

[12] Allen, F., Crepaldi, L., Alsinet, C. *et al.* Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat Biotechnol* **37,** 64–72 (2019).

[13] Gabitto, M.I., Rasmussen, A., Wapinski, O. *et al.* Characterizing chromatin landscape from aggregate and single-cell genomic assays using flexible duration modeling. *Nat Commun* **11,** 747 (2020).

[14] Robinson, J., Thorvaldsdóttir, H., Winckler, W. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29,** 24–26 (2011).

[15] Canver, M., Smith, E., Sher, F. *et al. BCL11A* enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis. *Nature* **527,** 192–197 (2015).

[16] Simeonov, D., Gowen, B., Boontanrart, M. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549,** 111–115 (2017).

[17] Bauer DE, et al. An Erythroid Enhancer of BCL11A Subject to Genetic Variation Determines Fetal Hemoglobin Level. *Science* **342**, 253–257 (2013).