APPRAISR: A model for predicting guide RNA efficacy for a hypercompact CRISPR-Cas epigenetic editor using machine learning

Tyler Borrman, Melanie Silvis, Zaki Jawaid, Tengyu Ko, Xiao Yang, Jackson Powell, Daniel Hart
Epicrispr Biotechnologies, South San Francisco 94080 CA, USA

Abstract

Introduction

To treat diseases driven by gene dysregulation, we developed GEMS (Gene Expression Modulation System), a CRISPR-based epigenetic editing platform to modulate therapeutic target genes. GEMS consists of three highly optimized components: Cas molecule, guide RNA, and modulator. For compatibility across all delivery systems, including AAV, we have developed a hypercompact dead Cas molecule, dCasONYX, derived from the Cas12 family (for which the first clinical trial has been initiated: NCT06907875). Successful CRISPR editing of therapeutic targets depends on an effective and locus-specific guide RNA. Consequently, guide RNA design is critical to the development of viable and safe CRISPR-based therapies. Machine learning has proven to be a successful method to predict guide RNA efficacy, improve hit rates, and uncover guide design principles. However, most machine learning studies focus on *Streptococcus pyogenes* Cas9 (SpCas9) and its nuclease activity, leaving guide design principles for dCasONYX and its epigenetic modality unknown.

Methods

To discover the guide design principles of dCasONYX we assayed the suppression effects of thousands of guides targeting over 160 genes in a CRISPR interference (CRISPRi) essential gene screen. Using this screen we developed APPRAISR (AI-Powered Prediction of RNA-guIded Suppression Responses), a GEMS scoring model for guide RNA efficacy. Guide RNA sequence, guide RNA energy scores, and target chromatin features were used as input for modeling. Prediction performance was evaluated for linear, tree-based, and deep learning architectures.

Results

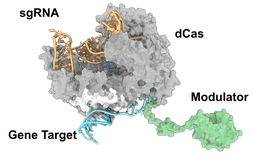
We discovered guide RNA sequence features to be predictive of guide RNA efficacy and found conserved guide RNA nucleotide preferences between dCasONYX CRISPRi and Cas12a nuclease activity. Testing APPRAISR in an independent CRISPRi screen, we observed a significant correlation between predicted and experimental guide suppression supporting the model's utility across different experiments and cell types.

Conclusion

APPRAISR's guide efficacy predictions are expected to accelerate CRISPR-based therapy development by focusing on high-confidence candidates early in the drug development process. Additionally, its feature analysis expands the set of guide design rules for Cas molecules and editing modalities providing a resource for the broader CRISPR/Cas community.

Gene Expression Modulation System (GEMS)

GEMS components



Delivery via any single vector

GEMS advantages

Precise – Precisely activates or suppresses genes without cutting DNA.

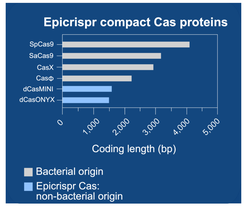
Versatile – Regulates any gene, anywhere, unlocking opportunities across a wide range of diseases.

Durable – Induces long-lasting changes in gene expression.

Safe – Non-immunotoxic and avoids permanent DNA modification, reducing risk of off-target effects.

Compact - smaller than competitors, enabling delivery through any single delivery vector.

Hypercompact dCasONYX for epigenetic editing

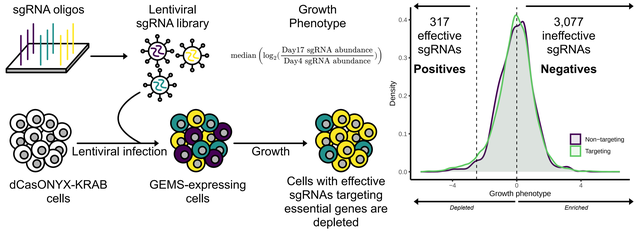


Low Human Pre-existing Immunogenicity

No cellular or humoral response to dCasONYX when tested with >60 donors

- dCasONYX is 1/3 size of Cas9, enabling efficient in vivo therapy using AAV, also with lower risk of pre-existing immune response.
- Clinical asset EPI-321, a potential cure for FSHD (Facioscapulohumeral Muscular Dystrophy), uses dCasONYX (Clinical trial: NCT06907875).

High throughput identification of effective and ineffective guide RNAs for dCasONYX mediated gene suppression

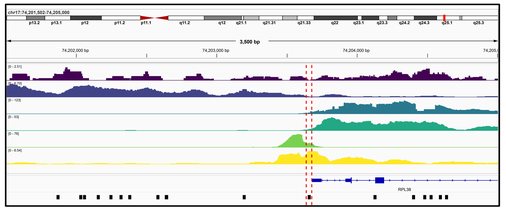


Essential gene screen identifies effective and ineffective dCasONYX guides

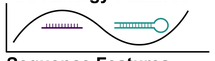
- 229 essential genes were selected for suppression targeting by dCasONYX based on public screens.
- dCasONYX was fused to KRAB suppression domain for effective gene silencing.
- Library of >8,000 sgRNAs targeting TTTR PAMs at essential gene promoters (-2kb to +1kb of transcription start site) were designed along with non-targeting controls.
- sgRNA libraries were transduced into a K562 cell line stably expressing dCasONYX-KRAB and growth phenotypes were evaluated by comparing day 4 and day 17 sgRNA abundance.
- Non-targeting and targeting sgRNA growth phenotype distributions were used to assign thresholds for positive and negative classifications. Guides with growth phenotypes < -2.5 were classified as positive effective guides while gRNAs with growth phenotypes > 0 were classified as negative ineffective guides.
- Evaluating genes with at least one positive guide, the total essential gene screen dataset consists of **3,394** gRNAs targeting **167** human genes.

Input features for modeling guide RNA efficacy

Chromatin Features



RNA Energy Features

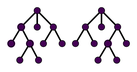


Sequence Features

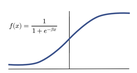
- sgRNAs were annotated for chromatin features based on the mean signal at their target location for 6 selected ENCODE K562 chromatin tracks. Distance to the gene TSS was also included as a chromatin feature for modeling.
- RNA energy features: Minimum free energy (MFE) = predicted energy of the most stable RNA conformation; Spacer Interaction = average bp interaction probabilities between spacer and scaffold.
- sgRNA 20 bp spacer sequence was one-hot encoded to generate 80 sequence features.

Model architecture and performance

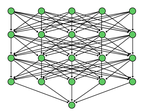
Random Forest



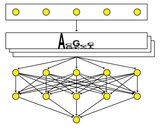
(L1/L2) Logistic Regression



Feed Forward Neural Network (FNN)

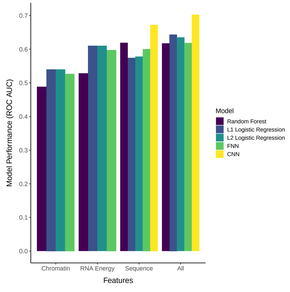


Convolutional Neural Network (CNN)



Architecture

- 5 different machine learning architectures were evaluated for guide efficacy prediction performance.
- Models were trained on 4 separate datasets of input features: Chromatin, RNA Energy, Sequence, and All (the full set of input features).
- Performance was measured by pooled ROC AUC from a leave-one-gene-out cross-validation scheme on the essential gene screen dataset.

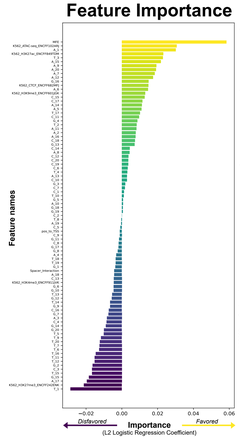


Performance

- Training on all input features led to the greatest performance for linear and deep learning models.
- Models trained on RNA energy or sequence features alone outperformed counterpart models trained solely on target chromatin features.
- The CNN model trained on all features outperformed all other models with a ROC AUC = **0.70**.

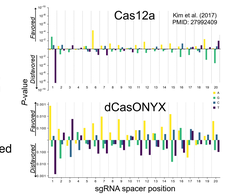
Features driving guide RNA efficacy

Feature Importance

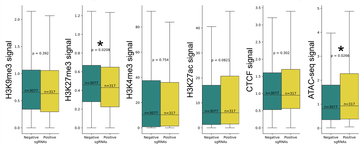


Conserved nucleotide favorability

- High MFE, ATAC-seq signal and adenine at position 1 are the most favorable features.
- Thymine at position 1, high H3K27me3 signal, and adenine at position 17 are the most disfavorable features.
- Conserved nucleotide preferences are observed between Cas12a and dCasONYX-KRAB systems.



Target chromatin favorability

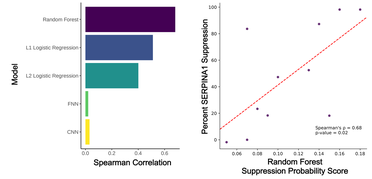


- Positive guides have significantly higher ATAC-signal.
- Positive guides have significantly lower H3K27me3 signal.

Model validation and predicted screening improvement

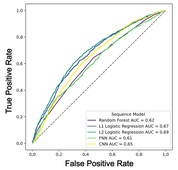
SERPINA1 Dataset

scRNA-seq guide screen to suppress the SERPINA1 gene via dCasONYX-KRAB in Huh7 cancer cell line



Kim et al. Dataset

Indel frequency of Cas12a in HEK293T cells (PMID: 29431740)



- The Sequence models trained on the essential gene screen dataset were tested against two independent datasets to validate performance and test generalizability.
- The Random Forest model was selected for **APPRAISR v1** given its performance across all datasets and generalizability across cell types.
- **Given the hit rate estimated by the essential gene screen we expect APPRAISR to drop the required number of guides to screen by 44% to identify a functional hit.**

Predicted Improvement