Combining Massive Oligo Pools and Artificial Intelligence (AI) to Predict Guide RNA Efficiency for Prime Editing Screens

Discover how Twist's synthesis capabilities were leveraged to develop prediction models for cutting prime editing technology.

ABSTRACT

A key challenge exists whenever a new CRISPR application is developed: How do you predict which guide RNA sequence will maximize the chance of an edit? Prime editing is one of the newest CRISPR technologies. It is highly promising as it can install small and precise mutations with a unique reverse transcriptase-dependent mechanism that eliminates the need for donor DNA and double-strand breaks. Due to its precise editing capabilities, it has potential applications in CRISPR therapeutics for treating genetic disease. However, relatively little is known about the rules that govern the performance of prime editor 2 (PE2), the chimeric enzyme on which second-generation prime editors are based.

In this application note, massive 300mer oligo libraries encoded prime editing guide RNA (pegRNA) molecules leveraged in a high-throughput evaluation pipeline to create data-sets large enough to robustly train high-performance computational models for predicting PE2 efficiency. Deep learning prediction models outperformed conventional machine learning when large, but not small, data-sets were used for model training. The computational models described here provide the first comprehensive resource for designing highly efficient pegRNAs that can be used with PE2.

INTRODUCTION

Correcting the 75,000 genetic variants known to cause disease in humans remains a substantial hurdle for wild-type Cas9 (Anzalone et al, 2019). This fact has spurred the refinement of the CRISPR/Cas9 system, leading to the advent of new CRISPR technologies allowing precise base editing and, more recently, prime editing (Anzalone et al, 2020). Based on catalytically impaired Cas9 variants, these new technologies avoid the consequences of double-strand breaks, including indel mutations (Mali et al, 2013), translocations (Kosicki et al, 2018), and cell death (Haapaneiemi et al, 2018).

Prime editing was developed to provide a more flexible alternative to base editing (Anzalone et al, 2020). Base editing relies on deaminases fused to a catalytically impaired Cas9 to insert point mutations at targeted loci. However, the available deaminases can only generate four transition mutations (C➔T, G➔A, A➔G, and T➔C). Prime editors use an attached reverse transcriptase (RT) to "rewrite" the sequence at a genomic site nicked by Cas9 nickase.

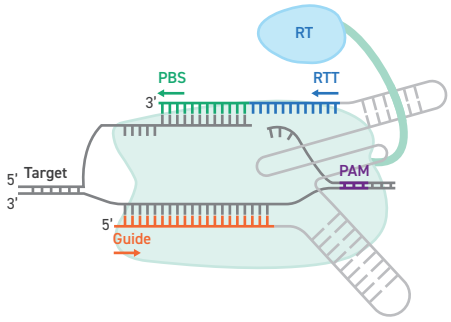

Figure 1: Schematic of a prime editor. pegRNAs encode sequences that determine the genomic target site (guide) and the desired edit (RT template). In addition, pegRNAs include sequences with which the Cas9 nickase (tracrRNA) and reverse transcriptase domains (the primer binding site, PBS) of PE2 interact. When PE2 associates with its genomic target, the Cas9 nickase domain nicks the target DNA strand containing the protospacer adjacent motif (PAM). The PBS of the pegRNA then hybridizes with the 3' end of the target DNA, priming the reverse transcriptase domain to incorporate the desired edit by synthesis. Cellular repair mechanisms then modify the complementary strand accordingly.

The pegRNA dictates the genomic target site and provides the template for reverse transcription, incorporating the desired edit (Figure 1). Encoding the edit in the pegRNA makes all 12 transition mutations possible, as well as small insertions and deletions (Anzalone et al, 2019).

A key challenge exists whenever a new CRISPR application is developed: How do you predict which guide RNA sequence will maximize the chance of an edit? Answering this challenge requires an understanding of the underlying guide, target DNA, and enzyme properties driving editing efficiency, thus necessitating access to a significant body of research. However, AI-based computational models are emerging as a powerful tool for predicting highly efficient guide RNAs within various gene editing approaches, enabling researchers to avoid the intensive upstream work.

Two types of these models exist: conventional machine learning models and deep learning models. Whereas the former still requires system features (e.g., CG content, melting temperature, etc) to be defined manually, the latter extracts information on these features automatically from training data. Automatic feature extraction makes deep learning models more effective than conventional machine learning at defining the targeting rules of

newer gene editing technologies. Even for well-understood gene editors like wild-type Cas9, deep learning models outperform their machine learning counterparts (Kim et al, 2019). But these performance benefits are only realized when deep learning models are trained on massive, robust data-sets.

Researchers at Yonsei University leveraged their experience in developing deep learning models for Cas9 (Kim et al, 2019), Cas9 variants (Kim et al, 2020a), and Cas12a (Kim et al, 2018) to generate high-performance prediction models for prime editor 2 (PE2), an engineered prime editor with improved editing capabilities (Anzalone et al, 2019). Kim et al (2020b) took advantage of the long length limits of Twist oligos (300mer) to implement a high-throughput evaluation pipeline that co-delivers pegRNAs and synthetic prime editing targets in the same oligo. By directing PE2 at genomically integrated, synthetic targets rather than endogenous loci, Kim et al (2020b) were able to more easily generate the massive mutational data-set necessary to create a robust deep learning model for predicting PE2 efficiency.

Using two paired oligo libraries—one large (48,000 paired sequences) and one small (6,800 paired sequences)—this application note demonstrates that data-set size is a key parameter that impacts the performance of deep learning models. In doing so, this application note also highlights the importance of various oligo synthesis parameters—including fidelity, uniformity, volume, and length—to the development of high-performance prime editing workflows.

WORKFLOW

The workflow, from initial oligo design to computational model selection, follows four general steps **(Figure 2).**

1. **Oligo design:** Paired oligos are designed as 250mer oligos. They include the pegRNA sequence, a barcode sequence, and a synthetic target sequence.
2. **High-throughput evaluation:** Paired oligos are cloned into lentiviral vectors for transduction in HEK293T cells at low multiplicity of infection to ensure a single integrant per cell. After selection, cells are transfected with a PE2-encoding plasmid to initiate prime editing. Genomic DNA is purified from the HEK293 cell library several days after transfection for NGS analysis of genomically integrated, synthetic target sites.
3. **Computational model training:** Data-sets generated by the high-throughput evaluation pipeline are used to train and evaluate conventional machine learning and deep learning models.
4. **Computational model selection:** Prediction models that result in the highest concordance between predicted and measured PE2 editing efficiencies are selected. Selected models provide a key resource for researchers implementing prime editing in their research.

RESULTS

Pool Uniformity Before and After PE2 Transfection

Highly uniform, evenly represented oligo libraries ensure the algorithm is trained on unbiased data, maximizing accuracy of resulting predictions. **Figure 3** shows the distribution of paired oligo sequences in the HEK293 cell library before and after transfection with the PE2-encoding plasmid. Sequence uniformity was maintained after PE2 transfection, with ~90% of paired oligo sequences represented within 1 log. Totals of 95.3% and 91.8% of paired oligo sequences were identified with >200 NGS reads before and after PE2 transfection, respectively. These data indicate that Twist oligos can generate highly uniform, evenly represented oligo libraries with few, if any, errors.

Computational Models Predict PE2 Efficiency

Two libraries were used to generate the mutational data-sets for training and evaluating PE2 prediction algorithms: a massive library containing 48,000 paired oligos (**Library 1**) and a substantially smaller library of 6,800 paired oligos (**Library 2**). These libraries were initially designed to test the effect of PBS and RT length (**Library 1; see Figure 1**) and editing type and position (**Library 2**) on prime editing efficiency with PE2. The mutational data-sets generated by these libraries were further split into training and test data-sets to train and evaluate computational models, respectively.

Figure 4 shows the relative performance of deep learning and conventional learning models for three prime editing parameters: PBS and RT template length (**Figure 4A**), editing type (**Figure 4B**), and editing position (**Figure 4C**). Among the computational models developed to predict the effect of PBS and RT template length on PE2 efficiency, the deep learning model (DeepPE), which was trained on the larger data-set (**Library 1**), outperformed conventional machine learning models, as demonstrated by a higher correlation between predicted and measured PE2 activity levels (**Figure 4A**). In contrast, conventional machine learning models outperformed deep learning models when a much smaller data-set was used for training, as was the case for models developed to predict the effect of editing type (PE_type, **Figure 4B**) and position (PE_position, **Figure 4C**) on PE2 efficiency. These comparisons emphasize that the potential predictive power of deep learning models in this application can only be realized when they are trained on massive data-sets.

DISCUSSION

Prime editing has emerged as a powerful strategy for introducing small and precise mutations, but little is known about the factors that affect PE2's editing efficiency. Instead of undertaking the large volume of work required to identify all properties affecting efficiency, deep learning models can make accurate inferences about a system without a priori knowledge of its properties when trained on enough high quality data. The computational models developed by Kim et al (2020b) provide the first comprehensive resource for designing highly efficient pegRNAs that can be used with PE2. Importantly, this resource (available at http://deepcrispr.

Figure 3: NGS analysis showing uniform distributions of oligo sequences in the HEK293T cell library before and after transfection with PE2-encoding plasmid. ~90% of gel-purified sequences were represented within 1 log in both cases.

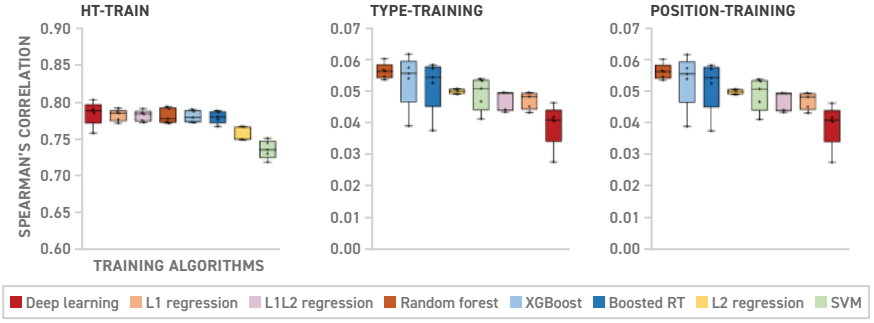

Figure 4: Performance of deep learning and conventional machine learning (L1 regression, L1L2 regression, L2 regression, Random forest, XGBoost, Boosted RT, SVM) models for PE2 efficiency. Datapoints represent the Spearman's correlation coefficient between measured and predicted PE2 activity (n = 5). NS, not significant; two-sided Steiger's test.

info/DeepPE) eliminates the need to test multiple pegRNAs before starting a prime editing experiment. Instead, users can identify highly effective pegRNAs by simply entering their desired genomic target sequence.

Guiding pegRNA selection in this web tool are three computational models: a deep-learning model that predicts the optimal PBS and RT lengths for a given target sequence (DeepPE) and two conventional machine learning models that predict effective pegRNAs of various editing types and positions (PE_type and PE_position). Although deep learning models generally outperform their conventional machine learning counterparts, this was not the case for PE_type and PE_position. This is because substantially smaller data-sets were used to train these models. The predictive power of deep learning was only realized with DeepPE—a model that was trained with a comparably massive data-set.

The data-sets used to train the algorithms described here were generated by a high-throughput evaluation pipeline that couples pegRNA expression cassettes to corresponding artificial target sequences in paired sequences (Kim et al, 2018). This approach eliminates the laborious, expensive, and error-prone process of amplifying tens of thousands of editing events from endogenous loci. With artificial target sequences, editing events can be amplified en masse for sequencing with a single primer pair. The synthetic nature of the target sequence also ensures that mutations reflect true editing events rather than endogenous mutations. Twist's massively parallel oligo synthesis platform accommodates both the volume (50,000+ oligos) and oligo length requirements (250–300mer) of this paired pegRNA-target screening pipeline.

High sequence fidelity reduces the background mutational frequencies in CRISPR screening applications. This is particularly important for prime editing applications because prime editing events generally amount to small mutations—similar to those that might otherwise be introduced during oligo synthesis or PCR amplification. Sequence errors were mitigated in these experiments through the use of Twist's low-error oligo synthesis platform and by minimizing the amount of PCR amplification during the preparation of the cell screening library.

Overall, Twist's oligo synthesis platform provides the fidelity, uniformity, volume, and length required to generate massive CRISPR data-sets, which are valuable for a diverse array of applications, including the development of high-performance, prime editing prediction algorithms.

METHODS

These methods are abbreviated from Kim et al. (2020b).

Plasmids

The LentiCas9-Blast plasmid (Addgene no. 52962) was linearized by double digestion with AgeI and BamHI (NEB). The sequence encoding PE2 was PCR amplified from pCMV-PE2 (Addgene no. 132775) using Sol 2× pfu PCR Smart Mix (SolGent) and inserted into the linearized LentiCas9-Blast plasmid using NEBuilder HiFi DNA Assembly Kit (NEB). The final assembled plasmid was called pLenti-PE2-BSD.

Oligonucleotide Library Design

A pool of 250mer oligonucleotides containing 54,836 paired pegRNA and target sequences (called "paired oligos" hereafter) was synthesized by Twist Bioscience. The paired oligo design is illustrated in Figure 2. This pool contained two libraries:

1. **Library 1 (48,000 paired oligos):** 24 combinations of PBS and RT template lengths (i.e., six PBS lengths [7, 9, 11, 13, 15, 17 nts] × four RT template lengths [10, 12, 15, 20 nts] = 24 combinations) for 2,000 unique paired oligos (24 × 2,000 = 48,000 total). The 2,000 target sequences were randomly selected from human protein-coding genes.
2. **Library 2 (6,800 paired oligos):** 34 templates were generated for 200 target sequences randomly selected from the 2,000 used in Library 1.

An additional 36 paired oligos were designed to target sequences used in Anzalone et al (2019). These paired oligos were used to correlate PE2 efficiencies at integrated sequences versus endogenous sites.

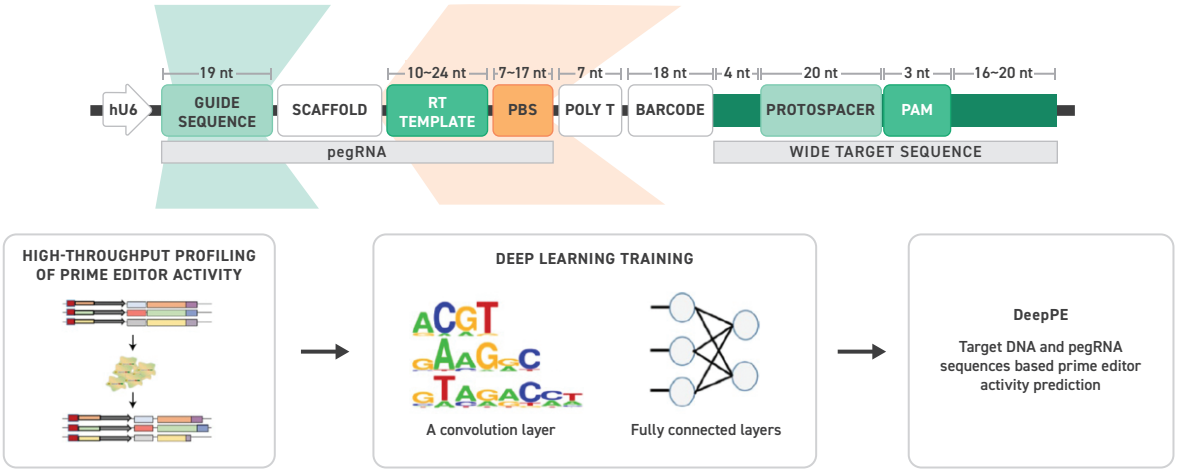

Figure 2: Workflow from oligo design to computation model selection. Paired oligos were designed to encode the pegRNA sequence, a corresponding target sequence, and a barcode for identification. Prime editing was measured in a HEK293T cell library transfected with a PE2-encoding plasmid. The resulting NGS data was used to train and evaluate machine learning models. Deep learning models were developed using a convolutional neural network (CNN) that uses a convolution layer to extract simpler, local features and fully connected layers to aggregate these local features into more complex, abstract features during model training.

Plasmid Library Preparation

Paired oligos were cloned into lentiviral transfer vectors using a two-step process involving Gibson assembly followed by restriction enzyme digestion and ligation. First, the paired oligo pool was PCR amplified using Phusion polymerase (15 cycles; NEB) and gel-purified (MEGAquick-spin, iNtRON Biotech). Lenti-gRNA-Puro (Addgene no. 84752) was linearized using BsmBI (NEB), dephosphorylated with CIP, and gel-purified. Gibson assembly was used to insert paired oligos into the linearized Lenti-gRNA-Puro backbone. The resulting plasmid library was transformed in electrocompetent cells (Lucigen), incubated, and extracted (calculated coverage: 113×). In the second step, the extracted, Gibson assembled plasmid library was linearized with BsmBI to insert the sgRNA scaffold sequence from pGR2 (Addgene no 104174). The final plasmid library was transformed in electrocompetent cells (Lucigen), incubated, and extracted (final calculated coverage: 785×).

Generation of HEK293 Cell Library

Lentiviral pools encoding the Libraries 1 and 2 were generated by transfecting HEK293T cells with the lentiviral plasmid library, psPAX2, and pMD2.G (1.2:0.72:1.64 ratio) using polyethylenimine. Lentiviral supernatants were collected 48 hrs later, filtered, and stored at –80°C until use.

HEK293T cells were transduced with the paired oligo library at a MOI of 0.3. After overnight incubation, puromycin (2 µg ml-1, 5 days) was to select for successfully transduced cells. The library was maintained at 500× coverage throughout this process.

PE2 Delivery

pLenti-PE2-BSD was delivered to the HEK293T cell library by transient transfection using Lipofectamine 2000 (Thermo Fisher Scientific) per the manufacturer's instructions. Cells were harvested 4.8 days post-transfection.

Measurement of PE2 Efficiencies at Endogenous Sites

A set of 33 pegRNA-encoding plasmids were selected from the plasmid library to validate the results of the high-throughput experiment. HEK293T cells were transiently transfected with pLenti-PE2-PSD and pegRNA-encoding plasmid using Lipofectamine 2000 or TransIT-2020 (Mirus) per the manufacturer's instructions. Puromycin was used to select for pegRNA-expressing cells. Cells were harvested 4.5 (Endo-BR1, Endo-BR2) or 7 days post-transfection (Endo-BR3).

Deep Sequencing and Analysis

Genomic DNA was extracted using Wizard Genomic DNA Purification Kit (Promega). Barcodes and target sequences were PCR amplified using 2× Taq PCR Smart Mix (SolGent). These PCRs were performed such that 700× coverage was achieved for each library. After gel purification, the resulting DNA was appended with Illumina adapter and barcode sequences for deep sequencing.

Custom Python scripts were used to analyze deep sequencing data. Paired oligo sequences were identified by unique barcode sequences inserted during oligo design. PE2-induced mutations were defined as edits not accompanied by unintended mutations within the wide target sequence. The background mutation frequency (resulting from the library preparation) was computed and subtracted from the observed prime editing frequencies as described (Kim et al, 2020b). Paired oligos with fewer than 200 reads or background mutation frequencies higher than 5% were excluded from analysis.

Conventional Machine Learning-based Model Training

Seven models were trained using conventional machine learning algorithms: XGBoost, gradient-boosted regression tree (Boosted RT), random forest, L1-regularized linear regression, L2-regularized linear regression, L1L2-regularized linear regression, and support vector machine (SVM). Five-fold cross-validation was performed to select models. Each model was selected from 144 models.

Development of Deep Learning-based Algorithms

Three deep learning-based computational models were developed: DeepPE (predicts optimal PBS/RT template lengths), PE_type (predicts optimal editing type), and PE_position (predicts optimal editing position). The code for each model is included in the original publication (Kim et al, 2020b).

DeepPE was trained using a dataset consisting of PE2 editing efficiencies for 38,692 pegRNAs, including nucleotide sequences (target, RT, and PBS sequences) and 20 additional features (e.g., melting temperature, GC content, etc.). Nucleotide sequences were converted into four-dimensional binary matrices by one-hot encoding.

DeepPE was developed using a convolutional neural network containing a convolutional layer and a fully connected layer (**Figure 1**). Nine models were tested in total. The model that produced the highest Spearman's correlation coefficients between empirical and predicted activity levels during five-fold cross-validation was selected. This model was called DeepPE.

PE_type and PE_position were developed using a multilayer perceptron (MLP) because the convolutional neural network performed poorly. Each model was selected from 18 MLP models containing similar architectures and parameters as DeepPE, but without convolutions.

Statistical Analysis

Spearman's correlation coefficients were analyzed by a two-sided Steiger's test.

REFERENCES

Anzalone, A. V. et al (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. Nature, 576(7785), 149–157.

Anzalone, A. V., Koblan, L. W., & Liu, D. R. (2020) Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nature Biotechnology, 38(7), 824–844.

Haapaniemi, E. et al (2018) CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. Nature Medicine 24, 927–930.

Kim, H. K. et al (2018) Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nature Biotechnology, 36(3), 239–241.

Kim, H. K. et al (2019) SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. Science Advances, 5(11), eaax9249.

Kim, N. et al (2020a) Prediction of the sequence-specific cleavage activity of Cas9 variants. Nature Biotechnology, 38(11), 1328–1336.

Kim, H. K. et al. (2020b) Predicting the efficiency of prime editing guide RNAs in human cells. Nature Biotechnology, 10.1038/s41587-020-0677-y.

Kosicki, M., Tomberg, K. & Bradley, A. (2018) Repair of double-strand breaks induced by CRISPR– Cas9 leads to large deletions and complex rearrangements. Nature Biotechnology, 36, 765–771.

Mali, P. et al (2013) RNA-guided human genome engineering via Cas9. Science, 339, 823–826.

Song, M. et al (2020) Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. Nature Biotechnology, 38(9), 1037–1043.