Systematic screening and analysis of disease-relevant mutations using prime editors

Jinman Park

Department of Medical Science

The Graduate School, Yonsei University

Systematic screening and analysis of disease-relevant mutations using prime editors

Directed by Professor Hyongbum Henry Kim

The Doctoral Dissertation
submitted to the Department of Medical Science,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Jinman Park

June 2024

This certifies that the Doctoral Dissertation of Jinman Park is approved.

Thesis Supervisor: Hyongbum Henry Kim

Thesis Committee Member#1: Sung-Rae Cho

Thesis Committee Member#2: Hae Jeong Park

Thesis Committee Member#3: Taeyoung Park

Thesis Committee Member#4: Dong Woo Chae

The Graduate School
Yonsei University

December 2023

ACKNOWLEDGEMENTS

I would like to thank first and foremost, my parents, who started and sustained my existence and motivated me to always strive to be better. I thank them for their quiet support and unspoken prayers. I dedicate this work to their sacrifices.

I would like to thank my advisor and mentor, Hyongbum Henry Kim. He has supported me since the beginning with advice on life and science. I will always be grateful for his patience and mentorship.

I would like to express feelings of warm gratitude to all my close lab members. I thank you for your discussions, your support, your encouragement, and all the laughs. You have shown me what it means to be resolute, hard-working, independent scientists, but most of all, kind humans.

I dedicate this work to my grandparents. They welcomed me back to Korea in 2011 and supported me throughout my service and graduate school day. I wish they were still around to ask when I will be graduating one more time. Thank you for your stories, your cooking, and your love.

Lastly, I dedicate this work to my wife, my best friend, and my love, Caitlin. I am humbled by your strength, intelligence, and kindness. Your laughter gives my life meaning. I thank you for standing by me throughout this endeavor. You have sacrificed more than you will ever admit. This work bears the weight of all your love and patience. I hope that it can serve as a foundation for everything we want to achieve in the future.

<TABLE OF CONTENTS>

LIST OF FIGURES

ABSTRACT

Systematic screening and analysis of disease-relevant mutations using prime editors

Jinman Park

Department of Medical Science

The Graduate School, Yonsei University

(Directed by Professor Hyongbum Henry Kim)

Manipulating DNA materials through recombinant DNA technologies has been the foundation of modern biological and medical research. Each innovative breakthrough through re-engineered techniques from nature has opened new frontiers for understanding complex genetic mechanisms and created new opportunities for tackling genetic diseases. First discovered in bacteria as a form of adaptive immunity against viral genetic materials, elucidation of the CRISPR-Cas9 system has allowed for its wide application in nearly all fields of biology and medicine, such as the analysis of large-scale screenings for therapeutic targets of various drugs in the treatment of cancers and hereditary diseases[1]. Since then, concerted efforts by laboratories across the world have aided in the advancement of its application potential, specificity, and programmability for genome editing, leading to the development various Cas9 variants and base editors[2-4]. Most recently, Dr. Liu's group introduced prime editing, a bio-engineered form of the CRISPR-Cas9 system that intrinsically alleviated many of the limitations of the canonical system by combining a reverse transcriptase to the Cas9 protein. Notably, prime editing has significantly improved genome editing by

allowing for the introduction of potentially any combination of specific genetic alterations without requiring donor DNAs or double-strand breaks[5]. However, determining the optimal conditions for improving prime editing efficiencies in various experimental factors required extensive time and resources. Our previous effort evaluated the efficacies of around 50K pairs of prime editing guide RNAs (pegRNAs) and their target sequences in human cells. In doing so, we determined features that affect prime editing efficiency and constructed three computational models that can predict pegRNA efficiencies. Although our efforts provided valuable insights for practical applications of prime editing in future studies, our approach was limited to a set of specific alteration types and positions. In this study, we aim to expand our data a degree of magnitude to 600K pairs of pegRNAs and their efficiencies in inducing any combination of alterations up to 3 nucleotides in size. In doing so, we aim to identify and evaluate the impact of factors that contribute to prime editing efficiency. In addition, we will carefully curate our pegRNA and target pairs using the extensive repertoire of disease relevant mutations available on the ClinVar database so that their prime editing efficiencies could be better evaluated within the context of disease therapy and modeling. We also aim to evaluate the optimal prime editing conditions in various cell lines and compare other variants of prime editors that have been recently reported[5]. Taken together, we found that our vastly expanded pegRNA design can cover up to 87% of reported disease-relevant mutations. Using our large-scale profiling data, we will develop a significantly improved prediction model based on the latest deep learning-based algorithms. Our work is expected to provide a more comprehensive tool to aid future works in expanding the application of prime editing in basic and clinical research efforts.

Key words: prime editing, high-throughput profiling, deep-learning

Systematic screening and analysis of disease-relevant mutations using prime editors

Jinman Park

Department of Medical Science

The Graduate School, Yonsei University

(Directed by Professor Hyongbum Henry Kim)

I. INTRODUCTION

First reported by the Liu group in 2019, prime editing (PE) has demonstrated immense potential in gene editing by allowing for any combination of specific alteration to the genome including insertions, deletions, and all 12 single point mutations[1]. Biochemically engineered, prime editors are composed of a Cas9 nickase–reverse transcriptase fusion protein and a prime editing guide RNA (pegRNA). The pegRNA consists of a guide sequence that recognizes a target sequence, a tracrRNA scaffold sequence, a primer binding site (PBS) necessary for reverse transcription (RT) initiation and an RT template that is designed to induce the desired genetic alteration. Multiple improved variants of prime editors have been since reported, including PE1, PE2, PE3, PE4, and PE5[1,2].They are distinguished by their biochemical properties and most notably their improved efficiencies in introducing genetic alterations at target sequences.

PE1 is the least efficient and less likely to be widely adapted. PE3 systems implements an additional single guide RNA (sgRNA) that can improve efficiencies according to specific targets in the genome. However, due to the additional sgRNA with the pegRNA, PE3 systems have been

shown to be more susceptible to the introduction of unintended off-target alterations. As the only difference between PE2 and PE3 systems is the additional sgRNA, the systematic evaluation of PE2 efficiencies in various genetic context is expected to apply consistently to PE3 systems. PE4 and PE5 are the most recent variants of the PE systems in which specific DNA repair mechanism within the cell can be transiently suppressed with the co-expression of a dominant negative MMR (DNA mismatch repair) protein, MLH1dn. PE4 and PE5 are PE2 with MLH1dn and PE3 with MLH1dn, respectively [2].

Previous works that have successfully implemented large-scale high-throughput evaluation of various Cas proteins and base editors have shown that the efficiency in inducing genetic alteration at target locations can depend on various genetic and epigenetic factors[3-6]. Such large-scale data have allowed the development of computational models that can accurately predict the activity levels of guide RNAs in various genetic contexts. Accordingly, these prediction tools have aided research efforts involving CRISPR nucleases. Previous efforts from our group have constructed a large-scale pegRNA library to systematically evaluate the efficiencies of around 50K pegRNAs in a high-throughput manner. Using this data, we were able to develop a computational tool, DeepPE, to predict pegRNA activities in specific, yet limited, context. The pegRNA library constructed for the evaluation of the 50K pegRNAs were specifically designed for a single point mutation induced at a single location.

In this study, we expanded our pegRNA library design to include more than 500K pegRNAs, a substantial improvement in scope, which will examine all possible 3-nt long alterations with single nucleotide resolution, including insertions, deletions, and substitutions. Importantly, we curated our pegRNAs using disease-relevant mutations that have been validated and reported on the ClinVar database[7]. In doing so, we conducted a systematic

evaluation of prime editing efficiencies in disease-relevant context and further demonstrated its potential in clinical applications and disease therapy. We developed DeepPrime, a high-performance prediction model, through our well-established high-throughput library profiling techniques combined with the latest convolutional and recurrent neural network algorithms. Taken together, we conducted a more comprehensive evaluation of pegRNA activities in diverse disease-based context to improve the performance and prediction power of our deep learning-based model that is expected to be a valuable tool in aiding future research and application of prime editing.

II. MATERIALS AND METHODS

1. Designing large-scale libraries for evaluation of prime editing

A. General oligonucleotide library preparation

The oligonucleotide pools containing pegRNA-target sequence pairs used in this study were synthesized by Twist Bioscience (San Francisco, CA). Each oligonucleotide contained the following elements: a 19-nt guide sequence, BsmBI restriction site #1, a 10~15-nt barcode sequence (barcode 1), BsmBI restriction site #2, the RT template sequence, the PBS sequence, a poly-T sequence, a 14~18-nt barcode sequence (barcode 2), and a corresponding 74-nt wide target sequence that included the PAM and RT template binding region. Barcode 1 was included to minimize template switching during PCR amplification, while barcode 2 (located upstream of the target sequence) allowed the identification of individual pegRNA and target sequence pairs after deep sequencing. Oligonucleotides that included unintended BsmBI restriction sites in their sequences were excluded.

B. Design of Library-Profiling

To evaluate the prime editing efficiencies under various pegRNA conditions, we designed a library of 47,839 oligonucleotides. These pegRNAs were selected from 40 seed target sequences from our previous study[6] that exhibited high editing efficiencies. The top half of the target sequences exhibited 70-75% editing efficiencies while the bottom half demonstrated 50-55% efficiencies. For each of these 40 high efficiency targets, we generated a 74-nt wide target sequence and designed 1,196 pegRNAs-target sequences pairs with various range of PBS and RT template lengths and various edit positions, lengths, and types. A total of 81

oligonucleotides containing the BsmBI cut site were excluded. We categorized the pegRNA-target sequence pairs into 8 groups as follows.

Group 1: Effect of PBS length

The intended edit was set to +5 G to C conversion. The pegRNAs were designed with a combination of RT template lengths fixed to 5, 12, 20, 33 and 50-nt and PBS length ranged from 1 to 17-nt. Of the total 3,400 oligonucleotides (40 seed targets x 17 PBS lengths X 5 RT template lengths), 4 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 3,396 pegRNA-target pairs.

Group 2: Effect of RT template length

The intended edit was set to +5 G to C conversion. The pegRNAs were designed with a combination of PBS lengths fixed at 7, 12, 17-nt and RT template length ranged from 5 to 40, 42, 44, 46, 48, and 50-nt. Of the total 4,920 oligonucleotides (40 seed targets x 3 PBS lengths X 41 RT template lengths), 4 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 4,917 pegRNA-target pairs.

Group 3: Effect of edit position

For intended edits of substitutions, pegRNAs were designed with a combination of PBS length was fixed to 12-nt and RT template lengths at 5, 12, 20, 30, 50-nt with all possible 1bp substitution of A•C•G•T-to-T•G•C•A set for all edit positions. Of the 4,800 oligonucleotides (40 seed targets x (5+12+20+33+50 positions for RT template lengths 5, 12, 20, 33, and 50-nt, respectively)), 12 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 4,788 pegRNA-target pairs. For intended edits of insertion or deletion, the pegRNAs were designed with fixed lengths of PBS and RT template at 12-nt and 22-nt, respectively, and intended insertion of AGG or CCT and a 3bp deletion at +1 to +12 edit positions from the nicking site. Of the 1,440 oligonucleotides (40 seed targets x 12 positions x 3

edit type (AGG insertion, CCT insertion, and 3-nt deletion)), 2 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 1,438 pegRNA-target pairs.

Group 4: Effect of edit type

The pegRNAs were designed with a fixed PBS length of 12-nt and RT template length up to 40-nt with intended edits of 1bp substitution, 3bp insertion of AGG or CCT, or a 3bp deletion at +1, +5, +12, +20 positions from the nicking site. For substitution or insertions, the minimal RHA length was 0 while for deletions, the minimal RHA was 1-nt long. Of the 19,640 oligonucleotides (40 seed targets x 491 of 3′ extensions (126+118+118+129 for A•C•G•T-to-T•G•C•A 1bp substitution, AGG insertion, CCT insertion, and 3-nt deletion, respectively)), 27 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 19,613 pegRNA-target pairs.

Group 5: Effect of PAM co-editing

The pegRNAs were designed with fixed PBS and RT template lengths of 12-nt and 22-nt, respectively. For intended edits of substitutions, the pegRNAs were designed to install A•C•G•T-to-T•G•C•A 1bp substitutions at the +1, +2, +3, +4, and +8 positions from the nicking site while simultaneously installing all possible 16 PAM co-edits at the +5 and +6 positions (N**GG** of the PAM). Of the 3,200 oligonucleotides (40 seed targets x 5 edit positions x 16 types of PAM editing), 4 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 3,196 pegRNA-target pairs. For intended edits of insertions, pegRNAs were designed to insert AGG or CTT at the +1, +4, or +8 positions from the nicking site and simultaneously install a +5 G to C conversion. Of the 480 oligonucleotides (40 seed targets x 3 edit positions x 2 insertion types (AGG or CCT) x 2 types of PAM co-edits (with or without +5 G to C conversion)), 2 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 478

pegRNA-target pairs. For intended edits of deletions, pegRNAs were designed to install a 3bp deletion at the +1 or +8 positions from the nicking site and simultaneously install a +5 G to C conversion. Of the 160 oligonucleotides (40 seed targets x 2 edit positions x 2 types of PAM co-edits (with or without +5 G to C conversion)), a single oligonucleotide was excluded for BsmBI recognized sequences to a final count of 159 pegRNA-target pairs.

Group 6: Effect of edit length – substitution

The pegRNAs were designed with fixed PBS and RT template lengths of 12-nt and 22-nt, respectively. From the positions of +1, +2, +4, +7, +8, +9, +10, +11, +12, +13 and +14 from the nicking site, up to 10 edit positions were randomly chosen to install a substitution. For cases that included an intended edit size of 3bp substitution, a simultaneous co-edit of +5 G to C conversion was also installed. Random selection of 1 to 10 edit positions was conducted 5 times to yield 55 pegRNAs per see target. Of the 2,200 oligonucleotides (40 seed targets x 11 RT-PBS selections x 5 iterations), 7 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 2,193 pegRNA-target pairs.

Group 7: Effect of edit length – insertion and deletion

The pegRNAs were designed with fixed PBS and RT template lengths of 12-nt and 22-nt, respectively, with intended edits of insertion at 1 to 10, 12, 15, and 20-nt in size or deletions at 1 to 10, 12, 15, 20, and 30-nt in size. The edits were installed at positions +2, +5, +10, and +15 from the nicking site. The intended edits of insertions were designed from two template sequences, Type I: AGGATCGATCCTGTACTTGC, Type II: CCTGACAACGCTTAGACAGA, where according to the edit size, the insert was spliced from the template sequence starting at their 5′ end. For example, an intended edit of 4bp insertion would yield two pegRNAs for inserting

AGGA or CCTG, Type I and Type II, respectively. The pegRNA designs where the edit position and insertion length combined to exceed the RT template length of 22-nt were excluded. Of the 5,760 oligonucleotides (40 seed targets x 144 edit lengths (44+44+56 for Type I insertions, Type II insertions, and deletions, respectively)), 11 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 5,749 pegRNA-target pairs.

Group 8: Effect of inserted sequence

The pegRNAs were designed with fixed PBS and RT template lengths of 12-nt and 22-nt, respectively, with intended edits of all possible 2bp insertions at the +2, +5, and +10 positions from the nicking site. Of the 1,920 oligonucleotides (40 seed targets x 3 edit positions x 16 2bp insertions), 5 oligonucleotides were excluded for BsmBI recognized sequences to a final count of 1,915 pegRNA-target pairs.

C. Design of Library-ClinVar

To evaluate prime editing efficiencies for installation and correction of disease relevant mutations, we designed a library of 549,168 pairs of pegRNA and target sequences. We selected variants that exhibited continuous 1 to 3bp substitutions, insertions, and deletions that were classified as pathogenic or likely pathogenic in the ClinVar database (version dated 2020-04-20). We extracted a 60-nt size flanking window around the variant position and determined all possible PAM sequences in the top and bottom DNA strands that could be used to designed guide RNAs for installing (disease modeling) or correcting (disease therapeutics) the target variant. A seed target sequence of 74-nt in the surrounding context containing each spacer sequence was extracted, and pegRNAs were generated from a combination of all possible PBS sequence (1 to 17-nt) and RT template sequences (from

minimum length to edit variant up to 50-nt). The final pegRNA library consisted of eight randomly selected pegRNAs from each seed target. The library included pegRNAs for installing (disease modeling) and correcting (disease therapeutics) variant edits. As the ClinVar database heavily favors single nucleotide variants, the proportion of pegRNAs targeting 2-3bp variants was adjusted accordingly to reflect the actual ClinVar distribution to ensure a dataset with proportionally accurate representation of the ClinVar variants. Lastly, oligonucleotides with internal BsmBI cut site were removed from the selection process.

D. Design of Library-Small

Library-Small was derived from the pegRNAs selected as the test dataset used from Library-ClinVar where high efficiency pegRNAs were selected for further modification to evaluate prime editing efficiency under various cell line and alternate PE system condition. Of the total 6,000 pegRNA-target pairs in the library, there were 1,495 pegRNAs-target pairs for each disease modeling and therapeutics, respectively, in which half were randomly selected and the other half was proportionally selected from 0%, 0 to 1%, 1 to 5% and over 5% editing efficiency ranges. 2,990 additional pegRNAs included randomly altered NGG PAM sequence to a NNN PAM sequence to examine the effect of PAM variants on prime editing. Lastly, 20 pegRNAs were included in 5-folds redundancy (5 x 4 pegRNAs) that exhibited the highest editing efficiencies from our previous study as positive controls.

2. Construction of large-scale libraries

A. Plasmid library preparation

The plasmid library containing pairs of pegRNA-encoding and corresponding target sequences was prepared using a two-step cloning

process: (Step I) Gibson assembly and (Step II) restriction enzyme-induced cutting and ligation. Uncoupling between paired guide RNA and target sequences during oligonucleotide amplification via PCR is effectively prevented by this two-step process[8]. The multistep procedure was adapted and modified from a previously reported method[9].

Step I: Construction of the initial plasmid library containing the pegRNA-encoding and target sequence pairs.

The oligonucleotide pool was amplified via PCR for 15 cycles using Phusion Polymerase (NEB) and gel purified. The Lenti_gRNA-Puro vector (Addgene #84752) was digested with BsmBI enzyme (NEB) at 55℃ at least 3 hours. The linearized vector was then treated with 1 μl of Quick CIP at 37℃ for 10 minutes, followed by gel-purification. Gibson assembly was used to assemble the amplified pool of oligonucleotides with the linearized Lenti_gRNA-Puro vector. After isopropanol precipitation, the assembled products were transformed into electrocompetent cells (Lucigen) using a MicroPulser (Bio-Rad). SOC media was then added to the transformation mixture, which was incubated at 37℃ for 1 hour. The cells were then spread and incubated on Luria-Bertani (LB) agar plates containing 50 μg/ml carbenicillin. Small fractions of the culture (0.1, 0.01, and 0.001 μl) were separately spread to allow determination of the library coverage. Plasmids were extracted from the total harvested colonies using QIAGEN Plasmid Maxi kit (QIAGEN). The calculated coverages of this initial plasmid library Profiling & ClinVar, Library-Off, Library-Small-PE2 and Library-Small-PE4 were 986X, 2,486X, 2,210X and 500X the number of oligonucleotides for each library, respectively.

Step II: sgRNA scaffold insertion.

The initial plasmid library produced in Step I was digested with BsmBI for at least 6 hours, followed by treatment with 1 μl of Quick CIP at

37℃ for 10 minutes. The digested product was gel-purified after size-selection on a 0.6% agarose gel.

Independently, an insert fragment containing either the conventional sgRNA scaffold sequence in the pRG2 plasmid (Addgene #104174) or the optimized sgRNA scaffold sequence from the previous study[10] was PCR-amplified using Phusion Polymerase and a primer pair with a BsmBI restriction site in each member of the pair followed by TOPO vector cloning (T-blunt vector; Solgent). The TOPO vector containing the insert fragment was digested with BsmBI for at least 12 hours and gel-purified on a 2% agarose gel to isolate the scaffold sequence. The purified insert was ligated with the digested initial plasmid library vector using T4 ligase (Enzynomics) at 16℃ for 3 hours (vector and insert; 1:10 weight ratio). The ligation products were purified by isopropanol precipitation and electroporated into Endura electrocompetent cells (Lucigen). Colonies were harvested and the final plasmid library was extracted using QIAGEN Plasmid Maxi kit. The calculated coverages of this initial plasmid library Profiling and ClinVar, Library-Off, Library-Small-Conv, Library-Small-Opti and Library-Small-PE4-Opti were 353X, 6,371X, 6,015X, 8,630X and 1,183X the number of oligonucleotides for each library, respectively.

B. Prime editor-encoding lentiviral plasmids

pLenti-PE2-BSD (Addgene #161514) and pLenti-NG-PE2-BSD (Addgene #176933) previously generated were used in this study for evaluating PE efficiencies of the original PE2 and NG-PE2, PEmax, NRCH-PE, and NRCH-PEmax -encoding fragment were amplified by PCR with Phusion High-Fidelity DNA Polymerase (NEB, M0530L). To prepare the lentiviral backbone vector, pLenti-PE2-BSD was digested with XcmI and BamHI (for cloning with hyPE fragment) or XbaI and EcoRI (for cloning

with PEmax, and NRCH-PEmax) restriction enzymes (NEB). After digestion, the vector was treated with 1uL of Quick CIP (NEB, M0525L) for 10 minutes at 37℃. The prime editor fragment amplicons and digested pLenti-PE2-BSD backbone vector were separated via 1% or 2% agarose gel electrophoresis, purified with a MEGAquick-spin™ Plus Total Fragment DNA Purification Kit (iNtRON Biotechnology, 17290), and assembled using NEBuilder HiFi DNA Assembly Master Mix (NEB, E2621L) according to the manufacturer's protocol. All plasmids were verified by sanger sequencing, and the primers used for cloning are summarized in **Supplementary Table 6 (Make Table 1).**

C. Culture and selection conditions for cell lines

HEK293T, HCT116, HeLa, DLD1, A549, and NIH3T3 cells were cultured in DMEM (Dulbecco's Modified Eagle Medium; Thermo Fisher Scientific) supplemented with 10% fetal bovine serum. MDA-MB-231 was cultured in Roswell Park Memorial Institute (RPMI) 1640 medium containing HEPES (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum. All cell types were maintained below 80% confluency at 37°C and 5% CO_2 and passaged every 3-4 days. For the high-throughput experiments, each cell line was seeding to 150-mm dishes according to the following cell density: HEK293THCT116: $1.2x10^7$ or $6x10^6$ cells for PE2 or PE4 system respectively, HCT116: $1.2x10^7$ cells, MDA-MB-231: $1.2x10^7$ cells, HeLa and NIH3T3: $6x10^6$ cells, DLD1and A549: $8x10^6$ cells. Unless otherwise noted, blasticidin S (BSD) selection concentration was 5ng/ml for NIH3T3 and 10ng/ml for all other cell lines while the puromycin selection concentration was 1μg/ml.

D. Production of lentivirus

HEK293T cells were seeded on 100-mm or 150-mm cell culture dishes (55,000 cells / cm^2) containing Dulbecco's Modified Eagle Medium (DMEM). 15 hours later, the DMEM was exchanged with fresh medium containing 25 μM chloroquine diphosphate, after which the cells were incubated for up to 5 hours. The transfer plasmid, psPAX2 (Addgene #12260), was mixed with pMD2.G (Addgene #12259) at a molar ratio of 1.3:0.72:1.64 and co-transfected into HEK293T cells using polyethyleneimine. 15 hours after transfection, cells were refreshed with maintaining medium. At 48 hours after transfection, the lentivirus-containing supernatant was collected, filtered through a Millex-HV 0.45-μm low protein-binding membrane (Millipore), aliquoted, and stored at -80℃. To determine the virus titer, serial dilutions of a viral aliquot were transduced into cells in the presence of polybrene (8 μg/ml). Both untransduced cells and cells treated with the serially diluted virus were cultured in the presence of puromycin (Invitrogen).

E. Preparation of PE2-expressing cell lines

We adopted the PE2-expressing HCT116 and MDA-MB-231 cell lines produced from our previous study. To generate each PE-expressing cell line used in this experiment, we created and maintained a lentiviral vector of each prime editor as described above at -80℃ until the cell line was ready to be prepared. For transduction, HEK293T, HeLa, DLD1, A549, and NIH3T3 cells were seeded on a 6 well plate with $2x10^5$ cells per well. After 12-24 hours, lentivirus was transduced in various amounts (0.8 μl-1mL per well) with polybrene. Following 24-48 hours after transduction, BSD selection was initiated by replacing the media. To produce a cell line containing only a single copy of the prime editor encoding gene, the cell lines with cell survival rates under 30% or less were selected for further experiments. All cell lines

were verified using prime editor targeted PCR and sanger sequencing.

F. Delivery of pegRNA-target library into the prime editor expressing cell for PE2 system

For the high-throughput experiment, we transferred the pegRNA-target paired library to the cell as described in our previous study. Briefly, the lentivirus containing the pegRNA and target-encoding plasmid library in the lentiviral backbone plasmid was prepared and stored at -80℃ until the PE-expressing cell line was ready. For lentiviral library transduction, PE-expressing cell lines were seeded in 150-mm dish and then incubated for 12 hours. Next, lentiviral library was transduced into the cells at 0.5 MOI to achieve greater than 500 coverage above the initial number of oligonucleotides. At 12 hours after transduction, the culture medium was replaced with DMEM containing 10% FBS and puromycin (2 μg ml). The cells were harvested 8 days (for Library- Profiling / ClinVar), 7 days (for Library-Small) after library transduction.

3. Bioinformatic platforms for library design and evaluation

A. Input processing for pegRNA design.

To facilitate a simply yet robust design system, the major databases for obtaining reference gene information on the human genome and genetic variants were curated, cleaned, and indexed for fast and resource-efficient manipulation. Databases included the NCBI's RefSeq gene information (REF), Ensembl database for vertebrate genomes (REF), the HUGO Gene Nomenclature Committee, HGNC, database, where the gene name and symbol, with corresponding identification and accession numbers for representative gene forms were cross-referenced and verified manually. In doing so, a comprehensive reference list of the major 18K human genes and their collective IDs from each database was established. Furthermore, the

transcription start and end positions as well as the distribution of each exon start and end positions were procured from the Matched Annotation from NCBI and EMBL-EBI (MANE) database to established a uniform representative transcript information through all future studies. Our reference source of human genes will continuously be updated and curated using the automated preprocessing scripts that require the most up-to-date or desired data file version from each database portal.

In addition to the inputs that utilize gene identifiers, our platform also can process variants from the ClinVar and COSMIC databases. The mutation identifiers from the archives can be used as inputs in which the preprocessed and index variant information can be fetched efficiently and all feasible pegRNAs targeting the variant for installing (disease model) or correcting (disease therapy) the mutation will be designed.

Lastly, basic inputs of sequence or genic positions of interest can be used to obtain a basic output of pegRNA and features. However, certainly annotation information regarding genic location or variant pathogenicity are omitted.

B. Analysis of prime editing efficiencies

For analysis of deep-sequencing data, we used in-house Python scripts that were adopted and expanded from our previous study. Each pegRNA and target sequence pair was identified via a 36-nt sequence (12-nt sequence that contained PBS domain of pegRNAs + 18-nt barcode + 6-nt sequences that include 4-nt 5′ neighboring sequence of target sequence and 2-nt ‘5 target sequence). The reads containing the specified edits without unintended mutations within the wide target sequence were considered to represent PE2-induced mutations. To exclude the background prime editing frequency originating from array synthesis and PCR amplification procedures, we normalized the observed prime editing frequency using the background

prime editing frequency determined in the absence of PE2 as shown below.

$$= \frac{\begin{array}{c}Read\ counts\ with\ intended\ edit\ and\ specified\ barcode - \\ (Total\ read\ counts\ with\ specified\ barcode\ \times\ background\ prime\ editing\ frequency) \div 100\end{array}}{\begin{array}{c}Total\ read\ counts\ with\ specified\ barcode - \\ (Total\ read\ counts\ with\ specified\ barcode\ \times\ background\ prime\ editing\ frequency) \div 100\end{array}} \times 100$$

Deep sequencing data were filtered to improve the accuracy of our analysis. pegRNA and target sequence pairs for which the deep sequencing read counts were below 200 or the background prime editing frequencies were above 5% were excluded as reported previously.

C. Analysis of prime editing byproducts

We determined the possible byproduct edits from prime editing by examining the top 20K pegRNAs by predicted activity and their indel and substitution frequencies. Only the reads containing imperfect or random edits were selected for analysis. The insertion, deletion, and substitution frequencies within a 5-nt window regions around the nicking position and RT template end position were compared to that of the non-window regions. Accordingly, the frequencies were normalized based on the size of the region and the total read count and compared along with the frequencies from the background samples.

D. Data preprocessing for machine learning

We combined the NGS read counts of replicates sorted by barcodes and obtained prime editing efficiency data as described above. We filtered pegRNAs with less than 200 reads without any unintended mutation, or greater than 5% background PE efficiencies. For extraction of features from the pegRNA and corresponding target sequences, we used the biopython (1.79), ViennaRNA package (**2.5.0**), and DeepSpCas9 to calculate the melting

temperature, GC counts, GC contents, minimum free energy, and DeepSpCas9 score. These features were combined with other sequence-based features named “Biofeatures” including length of RT-PBS, RHA length, edit type, edit position, and edit length using in-house Python scripts.

E. Conventional machine learning-based model generation

To compare various machine learning models that predict PE efficiency for pegRNA, we used the Pycaret package (2.3.10)[11]. To prepare datasets for model training, we add wide target sequence, PBS, and RT template sequence 1- and 2-nt motif features to HT-B by one-hot encoding. Tm, GC count, GC contents, MFE, and DeepSpCas9 score were normalized by z-score to produce a dataset with a total of 2,956 features. For evaluating estimator performance, we performed five-fold cross-validation. We used the default parameters as the other data preprocessing options. We generate linear regression, Lasso, Ridge, ElasticNet, Bayesian ridge, random forest, gradient boosting, extra tree, XGboost, CatBoost, and LightGBM regression model using the default parameters of pycaret. LightGBM, CatBoost, and XGboost, which had the best performance, were tuned using random grid search to select better performance models as final models and compare performance.

4. Development of deep learning-based model, DeepPrime

DeepPrime is a deep learning-based computational model that predicts the prime editing efficiency of any target gene and corresponding PBS and RT template lengths to introduce a 1 to 3-nt substitution, insertion, or deletion at position from +1 to +30. DeepPrime is implemented with PyTorch and takes a sequence of unedited and edited, and type of prime editing as an input. The input sequence processing module of DeepPrime consists of four convolutional layers and a recurrent neural network. Each convolutional layer

uses a kernel with a width of 3 and a stride of 1 and preserves the length by performing zero padding at both ends. The number of channels is 128, 108, 108, and 128, respectively, and average pooling is performed after the 2nd, 3rd, and 4th convolution operations. Input sequences are one-hot encoded into four channels and fed to the convolutional module. The output of the convolutional module is input to a bidirectional gated recurrent unit (GRU) to analyze long-distance interactions and positional characteristics of a gene. The GRU hidden state is 128-dimensional, and the output is linearly projected as a 12-dimensional vector. In addition, DeepPrime has a separate three-layer perceptron module for analyzing the physicochemical properties of pegRNA and target sequences (previously mentioned Biofeatures) Through this, 128-dimensional features are extracted and then concatenated with the 12-dimensional RNN output to create a 140-dimensional vector. This vector finally goes through a linear projection and outputs one regression floating point value as an output. For hyperparameter optimization, we used AdamW, cosine annealing learning rate scheduler, and Bayesian optimization.

5. SynDesign pipeline and web portal access

We have established a simple yet powerful bioinformatics pipeline that combines systematic targeting of gene or variant of interests, design and evaluation of saturation pegRNA library using DeepPrime and DeepPrime-FT and the automated incorporation of synonymous mutation markers in oligonucleotide constructs. This pipeline has been optimized for online access through our web portal called SynDesign. This webtool provides an end-to-end interface for targeting genes or variants of interest using prime editing in a highly programmable and automated manner. SynDesign facilitates precise gene targeting through extensive curation of major public databases. Users can effortlessly target their gene of interest using common

identifiers such as the gene symbol, GI from NCBI, NMID from RefSeq, Ensembl ID, and the HGNC ID. The integration of variant archives from ClinVar and COSMIC databases enhances the tool's utility, enabling users to focus on specific variants associated with human diseases.

In addition, with a single gene or variant input, SynDesign streamlines saturation genome editing (SGE). Our platform automates the design of all feasible pegRNAs and assesses their efficiency using advanced prediction models, DeepPrime and DeepPrime-FT. The extensive coverage of the prime editors and cell types of DeepPrime and DeepPrime-FT establishes SynDesign as the most up-to-date automation of pegRNA design and evaluation for precision SGE research.

Lastly, SynDesign's results section ensures a smooth transition to the saturation design of oligonucleotides. This includes incorporating synonymous mutation markers and other critical features from the top pegRNA designs. The platform facilitates comparative analysis, allowing users to assess the efficiency of top pegRNAs against others specific to the corresponding prime editor and cell type. This contextual information provides a more complete understanding of their efficiency.

The web portal includes an easy-to-follow tutorial with extensive examples for input and output. The help section provides visual examples of all the parameters and their impact on the data processing. We believe that with our expanded inputs, automation of key steps, and an in-depth guide on proper usage with optimal parameters will significantly facilitate the application of prime editing to future studies.

III. RESULTS

1. Factors that impact prime editing efficiency

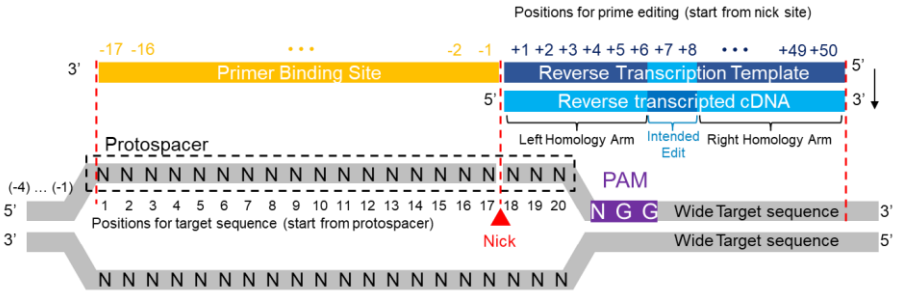

Figure 1. Schematic of prime editing guide RNA.

Prime editing efficiency can be affected by the parameters of its design at a given target sequence such as the lengths of prime binding site (PBS) and the reverse transcriptase template (RTT) regions in the pegRNA[1,6] (Figure 1). Our previous efforts explored a pre-determined set of PBS lengths and RTT lengths ((7, 9, 11, 13, 15, and 17-nt), (10, 12, 15, and 20-nt), respectively), which reveal certain insights into the importance of these parameters when designing effective pegRNAs[6]. Here, we expanded our scope to examine the effects of PBS and RTT lengths more systematically. We explored all experimentally feasible ranges of PBS and RTT to understand their impact on PE efficiency at a single-nucleotide resolution. We deployed all possible PBS (1~17-nt) and RTTs (5~35-nt) at various target sequences to induce a specific G to C substitution at the +5 edit position **(**Figure 2**)**.

In doing so, we found that for pegRNAs with 12-nt and 20-nt RTTs, the highest average efficiencies were observed when PBS lengths were 11-nt (average efficiency 13%) and 12-nt (8.5%), respectively. (Figure 3A), which

is consistent with our previous findings[6].

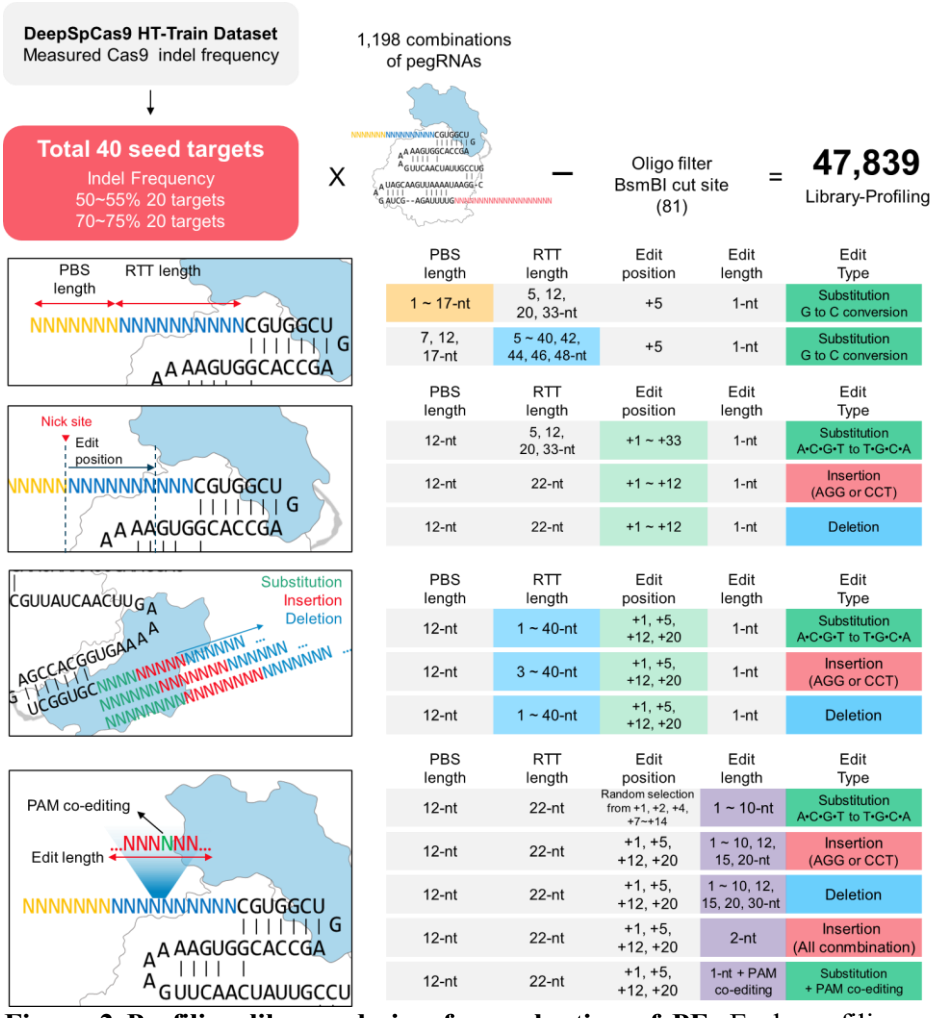

Figure 2 Profiling library design for evaluation of PE. Each profiling group focused on evaluating a specific pegRNA feature on PE efficiency. A pre-determined set of 40 seed target sequence was designated for each library with varying lengths of PBS and RTT lengths, editing positions within the pegRNA with respect to the PAM sequence, or editing types of substitutions, insertions, or deletions. Library-Profiling was comprised of a total of 47,839 pegRNAs designs.

In evaluating pegRNAs with 5-nt or 33-nt long RTTs, our findings indicate

the editing efficiencies were relatively low in identifying the optimal PBS lengths despite pegRNAs with 5-nt RTTs showing relatively high efficiency levels when paired with PBS lengths of 6 to 11-nt. We found that pegRNAs with 33-nt RTTs also exhibited relatively high activity levels with PBS at 12 to 17-nts. Collectively, our results indicate that using 11-nt PBS for RTT lengths 12-nt or lower and 12-nt PBS for RTT lengths at 12-nt or longer will yield generally high PE efficiencies.

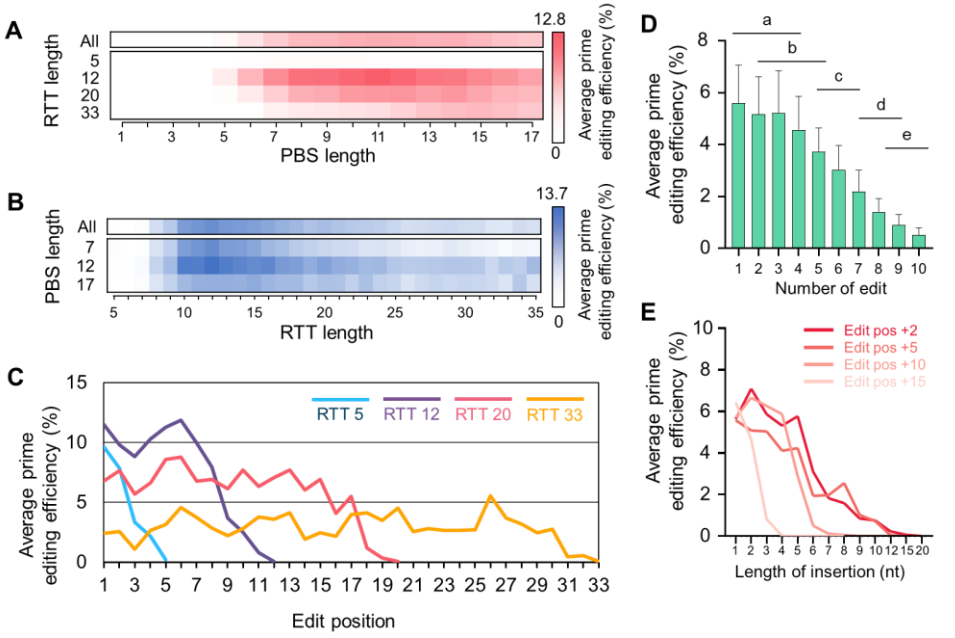

Figure 3. **Evaluation of factors that impact PE efficiency.** (A) The heat maps demonstrate the average PE efficiencies induced by pegRNAs designed with all possible PBS sequence lengths (1~17-nt) paired with fixed RTT lengths of 5, 12, 20, or 33-nt or (B) various RTT lengths (5~35-nt) with fixed PBS lengths 7, 12, or 17-nts. The number of pegRNA and target sequence pairs were n = 2,201 and 2,929, respectively. (C) Effect of the editing position on PE efficiencies for inducing 1bp substitutions. Each line indicates average efficiency at positions +1 to +33 with RTT at 5, 12, 20, and 33-nts. The length of PBS was fixed to 12-nts. The number of pegRNA and target sequence pairs are n = 195, 460, 683, and 474, for pegRNAs 5-, 12-, 20-, and 33-nts RT template, respectively. (D) Effect of edited nucleotide count on prime editing efficiency when inducing 1~10bp sized substitutions. Edit positions were randomly chosen among the positions +1, +2, +4, and +7 to+14 on the target sequence. The lengths of PBS and RT template were fixed to 13-nts and 22-

nts, respectively. The number of pegRNA and target sequence pairs was n = 1,920. Error bars represent 95% confidence interval. (E) Line plots indicating the average prime editing efficiency when introducing different lengths (1 to 10, 12, 25, and 20-nts) of insertion positions +2, +5, +10, and +15. The lengths of PBS and RT template were fixed to 12-nts and 22-nts, respectively. The number of pegRNA and target sequence pairs was n = 3,052.

Next, we assessed the impact of RTT lengths on editing efficiencies using pegRNAs with fixed PBS lengths of 7, 12, and 17-nt for RTT lengths ranging from 5 to 48-nt in length. Such a scope was previously unexplored, and our analysis showed that independent of PBS lengths, RTTs at 12-nt and within a 2-nt window exhibited the most efficient editing levels (Figure 3B). Examination of Library-ClinVar suggested consistent finding in line with that of our previous analyses[6]. Unique to the RTT, the intended edit position directly impacts its size such that a 1bp substitution intended for the +20 edit position requires a minimum RTT length of 20-nt. When we investigated the optimal RTT lengths for intended edit positions at +1, +5, +12, and +20, the optimal editing was observed when the RTT lengths were large enough to encompass the edit position and allow for a minimum right homology arm (RHA) of 4 to 7-nt (Figure 1). Larger insertions and deletions indicated consistent requirements of at least 7-nt and 9-nt RHA, respectively (Figure 4). Importantly, all editing types demonstrated that once a minimum RHA is achieved, there was an immediate diminishing return in RHA size as its length past 6-nt yielded reductions in efficiencies. Additionally, we were able to determine consistent findings for RHA requirement from analysis using Library-ClinVar. Altogether, we found that using pegRNAs designed with 7 to 9-nt RHA to achieve general high performance in prime editing.

RHA length	Edit length 1nt			Edit length 2nt			Edit length 3nt			Total
	Sub	Ins	Del	Sub	Ins	Del	Sub	Ins	Del	
0	0.13			0.01			0.11			0.12
1	0.39	0.08	0.09	0.04	0.05	0.04	0.01	0.12	0.10	0.23
2	1.16	0.24	0.20	0.23	0.06	0.05	0.06	0.10	0.27	0.65
3	2.58	0.65	0.39	0.74	0.16	0.12	0.22	0.07	0.09	1.48
4	3.49	1.59	0.83	1.46	0.59	0.20	0.51	0.24	0.16	2.21
5	3.72	2.53	1.42	2.39	1.23	0.42	1.15	0.87	0.15	2.68
6	3.71	3.02	2.27	2.51	2.12	0.72	1.28	0.93	0.34	2.97
7	3.64	3.32	2.69	2.31	2.60	1.21	1.41	1.44	0.83	3.12
8	3.49	3.24	2.97	2.72	2.93	1.91	1.45	1.90	0.91	3.14
9	3.41	3.18	3.35	2.82	3.02	2.23	1.23	1.69	1.44	3.19
10	3.29	3.14	3.14	2.25	2.86	2.86	1.13	1.80	1.66	3.12
11	3.00	2.93	3.36	1.88	2.65	2.50	1.13	1.88	2.04	2.90
12	2.90	2.68	3.02	1.94	2.70	2.44	1.09	1.78	1.23	2.73
13	2.67	2.81	2.77	2.40	2.60	2.43	1.00	2.11	1.45	2.62
14	2.65	2.62	2.81	2.80	2.53	2.38	1.38	1.71	1.29	2.57
15	2.40	2.73	2.70	2.14	2.37	1.94	0.65	1.48	1.28	2.40
16	2.33	2.31	2.34	1.63	2.32	2.26	1.04	1.65	1.35	2.25
17	2.30	2.04	2.39	2.52	1.83	2.07	1.05	1.52	1.37	2.16
18	2.35	2.12	2.42	1.64	2.32	2.20	1.02	1.78	1.50	2.25
19	2.13	2.02	2.03	1.30	2.10	2.10	0.77	1.36	1.03	2.02
20	2.12	1.74	2.09	1.82	2.11	2.34	1.00	1.00	2.15	2.00
21	1.90	1.84	1.98	1.48	1.96	1.87	0.51	1.04	1.48	1.84
22	1.65	1.81	2.11	1.60	1.94	1.56	1.02	1.57	1.58	1.75
23	1.88	1.45	1.67	2.37	1.70	2.09	0.65	0.88	1.28	1.72
24	1.81	1.53	1.62	1.44	1.47	1.49	0.85	1.04	1.34	1.64
25	1.55	1.42	1.44	1.64	1.74	1.88	0.46	0.78	1.18	1.49
26	1.51	1.00	1.45	1.47	1.19	1.51	0.81	1.55	1.37	1.37
27	1.38	1.04	1.67	1.50	0.93	1.56	0.86	1.12	1.45	1.32
28	1.22	1.32	0.99	0.98	1.06	1.20	0.99	1.17	1.49	1.19
29	1.47	1.22	1.10	0.80	1.05	1.37	0.73	0.00	1.09	1.31
30	1.27	1.14	1.04			1.57	0.07	0.32	0.39	1.19
Total	2.57	2.08	1.94	1.70	1.67	1.42	0.86	1.02	1.01	2.18

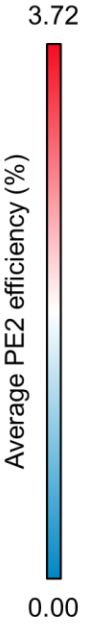

Figure 4. RHA length and edit type effects prime editing. Heatmap shows the average prime editing efficiency of pegRNAs with all evaluated edit types and sizes (1 to 3bp substitution, insertion, or deletion) with RHA lengths (0 to 30-nts).

We also investigated the effect of the editing position by examining pegRNAs designed to install 1bp substitutions at positions +1 to +33 with RTT lengths fixed at 5, 12, 20, and 33-nt and found consistent PE2 efficiencies up to edit positions +2, +8, +17, and +30, respectively, suggesting that with a minimum RHA length, editing position in general exhibit similar editing efficiencies, except at positions +5 and +6 which affect the PAM sequence (Figure 4**)**.

Then, we assessed the effect of the number of edited nucleotides on prime editing and found that PE2 efficiencies were similar for 1 to 4bp substitutions, decreased for 5 to 7bp substitutions and drastically decreased for 8 to 10bp substitutions (Figure 1D). We next examined the effect of inserted nucleotide length on prime editing efficiency. We observed the insertion efficiencies were similar for 1 to 5bp insertions and decreased for 6bp or greater insertion sizes when the editing positions were +2 or +5 (Figure 1E).

2. Role of last templated nucleotide on PE

It has been previously reported that the last templated nucleotide of the RTT should not be a guanine (G) to prevent RTTs locating a cytosine (C) close to the 3′ hairpin of the sgRNA scaffold[1]. However, findings from our previous study showed contradictory results as pegRNAs with RTTs 20-nt exhibited consistent findings the previous work but pegRNAs with RTTs at 15-nt, indicated no importance of the last templated nucleotide to PE efficiency. Accordingly, when pegRNAs with RTT lengths at 10 or 12-nt were used, G as the last templated nucleotide yielded highest average PE efficiencies. To elucidate the underlying mechanism behind the contradicting observations, we explored the role of the last templated nucleotide using the vast Library-ClinVar data. As a result, we found that the average editing

efficiencies followed the optimal pattern of C > T > A > G all edit types (Figure 5).

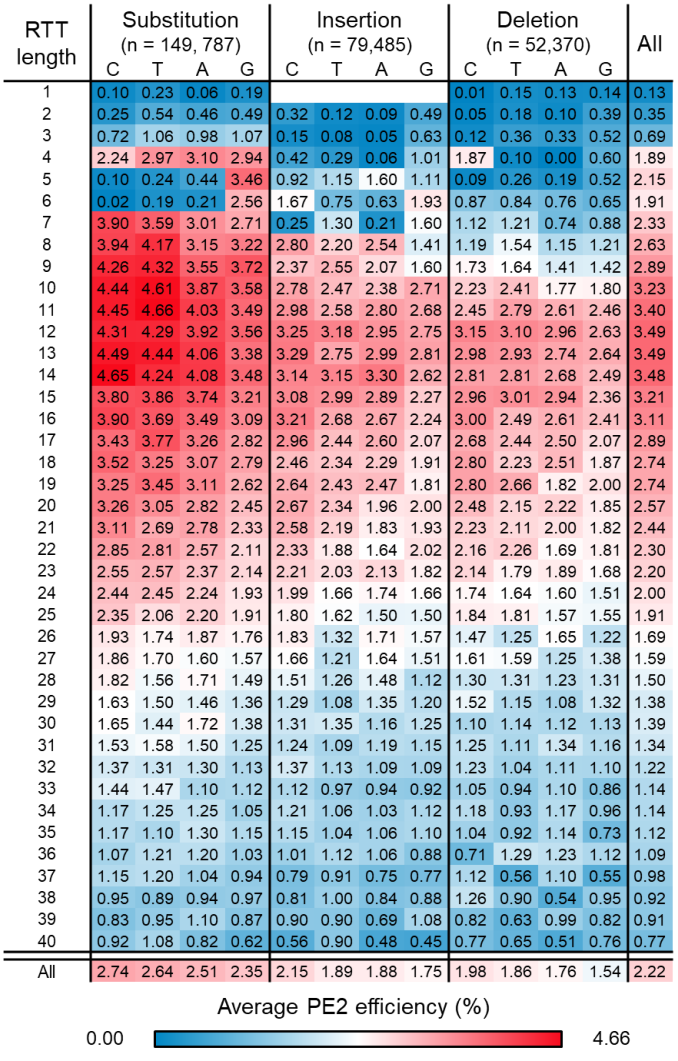

Figure 5. Effect of last templated nucleotide on prime editing efficiency. Heatmap showing the effect of last templated nucleotide on prime editing efficiency according to RTT lengths and edit types. Each value represents the average PE2 efficiency. Those with an average PE2 efficiency of 4 or more were indicated in white letters.

3. Role of PAM co-editing on PE

Previous reports showed that simultaneous perturbation of the PAM sequence, traditionally NGG, can result in improved editing efficiencies by blocking the re-binding of the Cas9 protein to the target sequence, which can lead to multiple nicking of the reverse-transcribed DNA strand before the repair of the complementary strand[1,6]. Accordingly, the two G's (NGG) corresponding to the +5 and +6 positions on the wide target sequence (Figure 1) can be altered to any of the 15 dinucleotide sequence combinations. We investigated the impact of all 15 co-edits on PE efficiency by designing pegRNAs to induce 1bp substitutions at positions +1, +2, +3, +4, and +8 simultaneously with all possible PAM co-edit dinucleotides. When compared to the same pegRNAs without the PAM co-edits, we found that PE efficiencies were improved by on average 1.7-folds, in which the NAT co-edit exhibiting the highest efficiency improvements (Figure 6). On the other hand, NCT, NCC, and NCA yielded the lowest improvement and should be avoided. Interestingly, we found that for edit positions +1, +2, and +3, PAM editing induced higher editing efficiencies than those of positions +4 and +8, which may be potentially due to the synergistic effects between the close proximity of the intended edit and the PAM co-edit that blocks the nickase activity of the Cas9 protein.

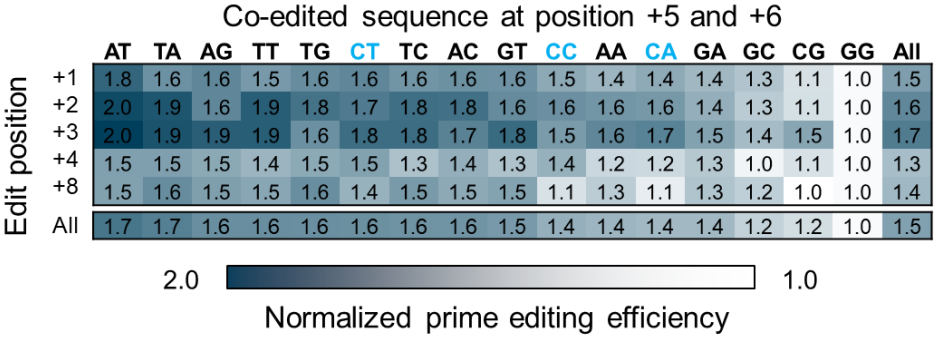

Figure 6. Effect of PAM co-editing on PE2 efficiency. Evaluation of PE

when introducing 1bp substitutions at positions +1, +2, +3, +4, and +8 with co-edited +5 and +6 positions on the PAM (NGG). The average prime editing efficiencies with PAM co-editing were normalized to that of without PAM co-editing. The number of pegRNA and target sequence pairs was n = 2,686.

4. DeepPrime is a powerful tool for predicting prime editing efficiency

Our previous study focused on the development of accurate models for the prediction of editing efficiencies of Cas12a[4], Cas9[3], base editors[12]. In addition, our most recent study evaluated PE2 activity levels and developed three prediction models, DeepPE, PE_type and PE_position. However, the models were limited by the scope of experimental conditions of the high-throughput library design[6]. DeepPE was limited to a single intended edit of G to C substitution at the +5 edit position with only 24 different combinations of PBS and RTT lengths. PE_type can predict prime editing efficiencies of pegRNAs with only 13-nt long PBS and predominantly RHA of 14-nt for 24 edit types at limited edit positions and PE_position can predict editing efficiencies of pegRNAs with only 13-nt PBS and 20-nt RTT for intended editing of 1bp substitution. Notably, the two models were developed using conventional machine learning algorithms as opposed to the more robust deep learning based techniques. In addition, the size of the dataset used for training the models were limited to sets of 3,775 and 1,774 pegRNAs, respectively, yielding modest performances of R = 0.47 and 0.56 for PE_type, PE_position, respectively. Development of a more comprehensive model for evaluating pegRNAs with potentially all experimentally feasible combination of parameters can reveal a clearer landscape of the factors impacting prime editing efficiency. Our Library-ClinVar contains 549,618 pairs of pegRNA and target sequences with 850 (= 17 x 50) combinations of PBS and RTT lengths for all edit types including up to 3bp substitutions, insertions, and

deletions across an extensive edit position range of +1 to +30 positions. We exhaustively examined the prime editing efficiencies of the 549,618 pairs of pegRNA and target sequences. However, human capacity for determining the feature map of such an enormous amount of data is impossible. After best-practice methods for reducing data noise and errors, we split the data into two data sets, ClinVar_Train (n = 259,910) and ClinVar_Test (n = 28,883), by random sampling with unique representation by each pegRNA. As the scale of ClinVar_Train significantly overshadows those of DeepPE, PE_type, and PE_train, we conducted comparative analysis of its performance across various conventional machine learning-based and deep-learning based models. Subsequently, we tested each model using 5-fold cross-validation, and found that a convolutional neural network (CNN) based algorithm, not unlike those of previous studies, combined with gated recurrent units (GRU), a key component of recurrent neural networks (Figure 7), exhibited the highest performance and significantly greater that of the next best algorithm (CNN with attention module) (P = 2.3 x 10^{-2}; Steiger's test) (Figure 8A). Accordingly, we adopted this algorithm developed using CNN with GRU-based approach, as our main prime editing prediction model, DeepPrime.

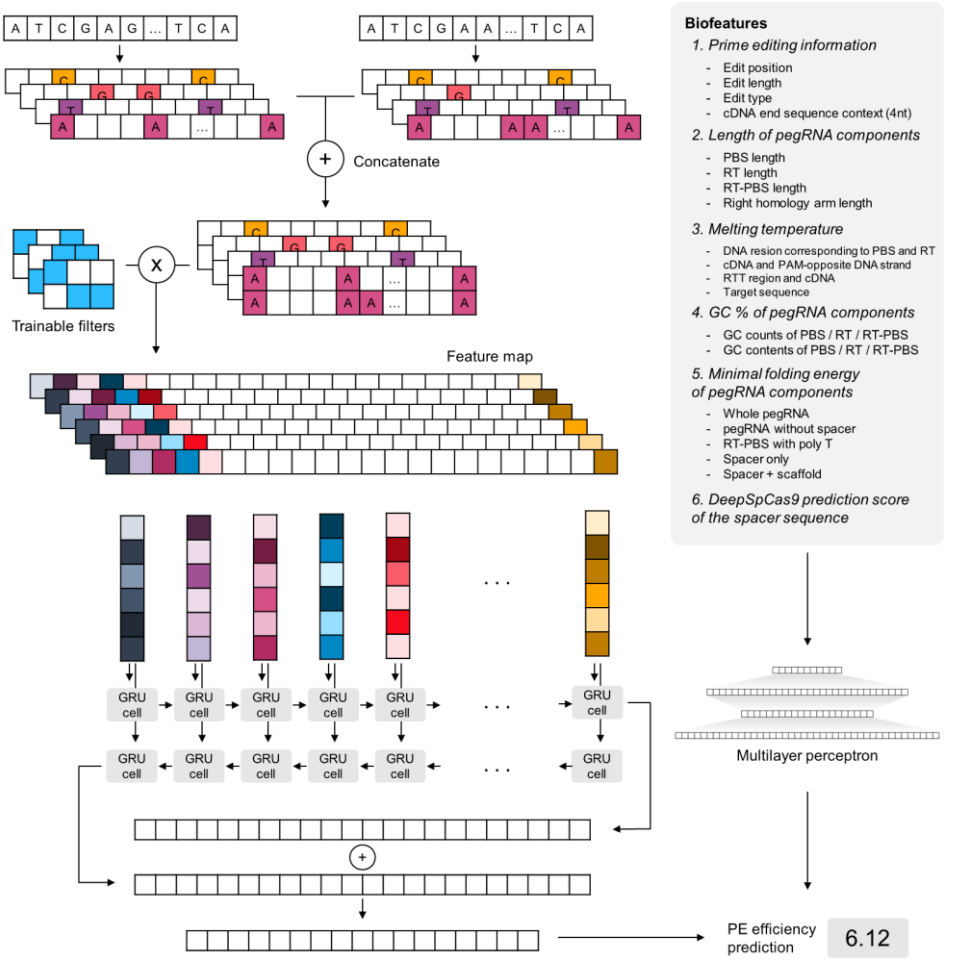

Figure 7. DeepPrime model development schematic. A diagram of DeepPrime model architecture. Inputs for the prediction model require non-edited target and prime-edited target sequences, pegRNA information (guide/PBS/RT), and an additional "Biofeatures." The non- / prime edited target sequences are converted into one-hot encoding, fixed-length array data, and merged into a vector. Then, the target sequence vector is trained with a convolutional filter to extract a feature map, and then a flattened layer is made through a bidirectional GRU. In addition, feature information extracted from target sequence and pegRNAs creates another layer through multilayer perceptron and combines these two layers to create a final output layer.

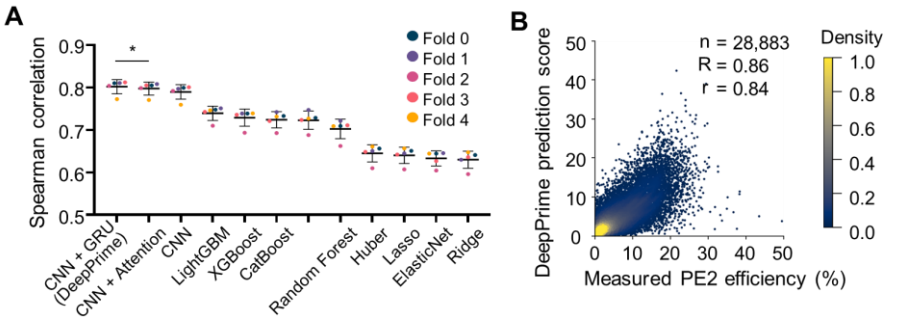

Figure 8. Comparison of DeepPrime model performance. (A) Comparative analysis of various model algorithms for predicting PE efficiency. Each dot represents the Spearman's correlation coefficient between the measured and predicted prime editing efficiency levels using five-fold cross-validation (total, n = 5 correlation coefficients). Statistical analysis of the top two algorithms is as shown (***P = 9.4 x 10-4; two-sided Steiger's test). The bar and error bar indicate the average and standard deviation of coefficient values, respectively. (B) Validation of DeepPrime using ClinVar_Test. Dot color gradient was generated using Kernel Density Estimation (KDE) with a Gaussian kernel. Spearman's (R) and Pearson's (r) correlation coefficients are as indicated.

Testing of DeepPrime was conducted first using the non-redundant test dataset, ClinVar_Test. As a result, we were able to confirm the high performance of DeepPrime with a Spearman's and Pearson's correlation coefficient (R = 0.86 and r = 0.84, respectively). As expected, these findings were notably higher than that of DeepPE, PE_type, and PE_position[6] (Figure 8B). We segmented our testing further to the different factors that impact PE efficiency to ascertain whether DeepPrime exhibited consistent prediction power. Nine subsets of ClinVar_Test were created to focus on the nine different combination of intended edit types (1 to 3bp substitutions, insertions, and deletions). Our findings indicated high Spearman's (R) and Pearson's (r) correlation coefficients for each of the nine edit type combinations, indicating

that DeepPrime is a robust and versatile prediction model that can serve as a valuable asset in a variety of pegRNA configuration (Figure 9A).

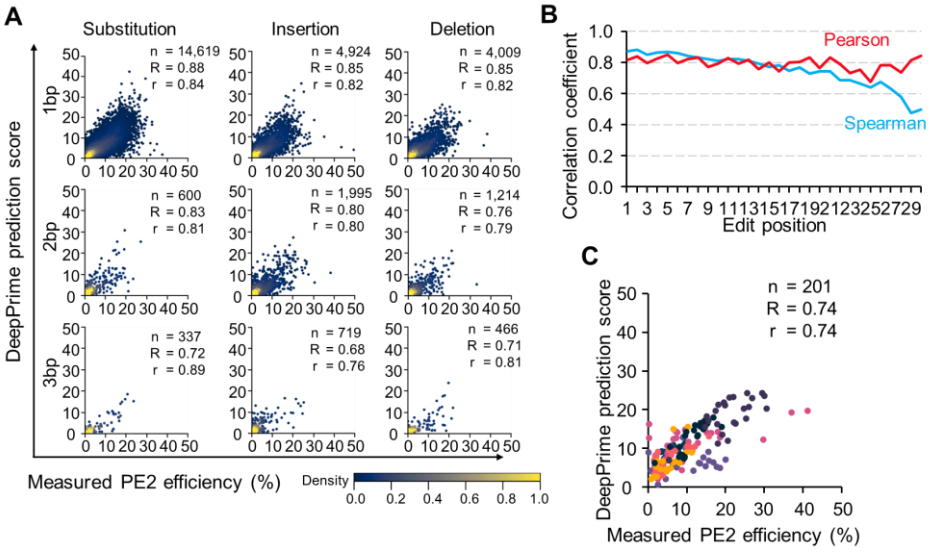

Figure 9. **DeepPrime model profiling and validation.** (A) Evaluation of DeepPrime robust prediction power across all edit configurations. Otherwise as Figure 7. (B) DeepPrime performance across various edit positions according to Pearson's and Spearman's correlation coefficients. (C) Evaluation of DeepPrime using an independent test dataset[1], which include prime editing efficiencies obtained from endogenous sites in HEK293T cells. The number of pegRNAs, Spearman's (R) and Pearson's (r) correlation coefficients are shown.

Furthermore, our testing of DeepPrime also indicated high performance across editing positions, Pearson's coefficients ranged from 0.68 to 0.85 at intended positions +1 to +30 while Spearman's ranged from 0.63 to 0.88 at positions +1 to +27 but dropped off at positions +28 to +30 to 0.47 to 0.58 (Figure 9B). Most importantly, DeepPrime was assessed using an independent dataset published previously that measured prime editing at endogenous sites, DeepPrime demonstrated excellent Spearman's and

Pearson's correlation coefficient of R = 0.74 and r = 0.74, respectively, indicating its robust performance in predicting PE2 efficiencies at endogenous sites (Figure 9C).

5. DeepPrime identifies important context features impacting PE

To systematically ascertain the degree of contribution from each feature associated with prime editing efficiency, we performed SHAP (SHapley Additive exPlanations) analysis using 2,956 features that include the melting temperature, GC counts, GC contents, the minimum self-folding free energy of various regions in the pegRNAs, PBS and the RTT lengths, edit types, edit positions, edit lengths, RHA lengths, and the DeepSpCas9 scores[3]. In addition, we included direct sequence information, such as all mononucleotides and dinucleotides that are position independent and dependent. We distinguished features as favored or disfavored when its values were associated with high and low prime editing efficiencies (**Figure 10**). We found that the features distinguished for their association with high PE efficiency were generally in line with the previously reported features[6].

The SHAP analysis showed that the GC count of the PBS sequence was the most favored feature associated with high PE efficiency. Accordingly, Tm of PBS and number of 'C' in PBS were also favored and were within the top 10 important features. In our previous study, we identified the GC counts in PBS and Tm of PBS as one of the top three features[6]. The difference in findings is potentially due to the increase in the diversity of PBS compositions in the ClinVar_train dataset including very short PBS of 1 to 6-nt in size.

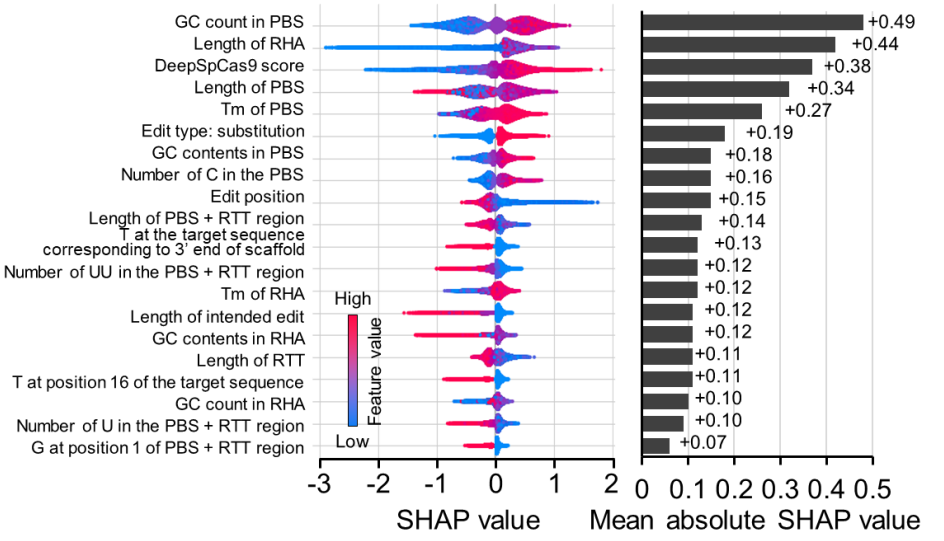

Figure 10. **DeepPrime features analysis.** The 20 most important features associated with prime editing efficiencies. Tree SHAP determined the feature importance. The summary violin plot (the left graph) represents each target sequence as a dot where its position on the x-axis reflects the SHAP value. High SHAP value indicates that feature is associated with high prime editing efficiency. The color of dot shown in red, or blue represents high or low value of the relevant feature for that particular target sequence, respectively.

We expect that higher GC count and Tm of the PBS can facilitate an increase in affinity between the pegRNA and the nicked strand of the target DNA, which would mediate an increase in reverse transcription rate. To investigate this thoroughly, we compared the prime editing efficiencies across different Tm, GC count, and length of PBS. We found that all three factors affected prime editing efficiency with GC counts consistently being the most important feature associated with efficient prime editing. With fixed PBS lengths, our findings remained similar with the average prime editing efficiencies varied at a large scale depending on the GC count (favored) and Tm (favored) (Figure 10). Importantly, the GC count and Tm of PBS that yielded the highest average efficiencies in general, were higher than 6 and 25

degrees, respectively (Figure 10). The optimal range of Tm also varied depending on PBS length and GC count. Once GC count was higher than or equal to 6, pegRNAs with longer PBS than 13-nt led to lower prime editing efficiencies. Consistently, GC contents in PBS was 7th important feature (favored) and pegRNAs with GC contents higher than 60% were associated with high prime editing efficiency as long as the PBS length is longer than 6-nt (Figure 10).

4-nt PAM sequence	SpCas9	SpCas9 -NG	NG-PE2	NRCH-PE2	PE2	PEmax	PE4max
NAAA	0.0	0.2	0.8	0.5	0.0	0.1	0.2
NAAC	0.0	0.2	2.1	5.9	0.1	0.2	0.3
NAAG	0.0	0.6	2.1	1.1	0.1	0.1	0.1
NAAT	0.0	0.2	0.4	1.1	0.0	0.1	0.1
NACA	0.0	0.2	1.0	3.5	0.0	0.1	0.2
NACC	0.0	0.1	0.4	3.1	0.1	0.1	0.0
NACG	0.0	0.5	1.8	0.8	0.1	0.1	0.2
NACT	0.0	0.2	1.7	10.4	0.1	0.1	0.3
NAGA	0.4	0.5	1.8	1.0	1.6	4.0	4.1
NAGC	0.5	0.4	2.1	3.1	3.4	10.5	12.0
NAGG	0.5	0.7	1.1	0.7	2.4	5.7	7.0
NAGT	0.4	0.6	0.8	0.6	0.7	2.2	2.2
NATA	0.0	0.3	0.2	0.2	0.0	0.0	0.0
NATC	0.0	0.2	1.1	3.5	0.1	0.1	0.1
NATG	0.0	0.7	1.7	2.0	0.1	0.1	0.0
NATT	0.0	0.2	0.3	1.2	0.0	0.1	0.0
NCAA	0.0	0.0	0.1	0.1	0.1	0.0	0.1
NCAC	0.0	0.0	0.1	0.4	0.0	0.1	-0.1
NCAG	0.0	0.1	0.0	0.0	0.1	0.1	0.0
NCAT	0.0	0.0	0.4	0.8	0.0	0.1	0.2
NCCA	0.0	0.0	0.1	0.1	0.0	0.0	0.1
NCCC	0.0	0.0	0.3	3.7	0.0	0.0	0.0
NCCG	0.0	0.1	0.0	0.1	0.0	0.1	0.2
NCCT	0.0	0.0	0.2	1.4	0.0	0.1	0.1
NCGA	0.0	0.1	0.2	0.2	0.1	0.1	0.2
NCGC	0.1	0.1	0.1	0.1	0.0	0.1	0.2
NCGG	0.4	0.2	1.7	0.9	3.2	8.6	9.4
NCGT	0.0	0.1	0.3	0.2	0.0	0.1	0.2
NCTA	0.0	0.0	0.2	0.2	0.0	0.0	0.2
NCTC	0.0	0.0	0.0	0.1	0.0	0.0	0.0
NCTG	0.0	0.2	0.4	0.2	0.0	0.0	0.0
NCTT	0.0	0.0	0.2	0.8	0.0	0.0	0.0
NGAA	0.3	0.7	4.5	9.7	2.4	6.8	8.2
NGAC	0.3	0.5	1.9	6.0	1.4	4.7	5.2
NGAG	0.3	0.9	5.8	6.6	2.7	7.0	7.7
NGAT	0.3	0.7	1.5	3.7	0.4	0.9	0.4
NGCA	0.1	0.6	2.7	6.5	0.3	0.7	0.6
NGCC	0.1	0.4	2.3	8.5	0.6	2.2	2.8
NGCG	0.1	0.9	0.9	0.9	0.0	0.1	0.0
NGCT	0.1	0.6	2.5	6.6	0.6	1.5	1.6
NGGA	1.0	0.9	4.2	5.8	7.4	13.7	20.3
NGGC	1.0	0.7	3.5	6.8	7.1	13.4	19.1
NGGG	0.9	1.0	4.4	6.0	7.4	13.8	19.1
NGGT	1.0	0.9	4.7	8.2	8.4	15.8	23.0
NGTA	0.0	0.8	5.0	9.2	0.1	0.1	0.1
NGTC	0.1	0.7	0.6	3.7	0.1	0.2	0.3
NGTG	0.0	1.0	4.0	6.3	0.1	0.1	0.1
NGTT	0.0	0.8	7.7	15.8	1.0	1.8	2.2
NTAA	0.0	0.1	0.1	0.0	0.0	0.0	0.1
NTAC	0.0	0.1	0.6	1.3	0.0	0.1	0.1
NTAG	0.0	0.3	0.2	0.0	0.0	0.1	0.1
NTAT	0.0	0.1	0.1	0.1	0.0	0.0	0.0
NTCA	0.0	0.0	0.0	0.2	0.0	0.1	0.1
NTCC	0.0	0.0	0.0	0.2	0.1	0.1	0.1
NTCG	0.0	0.2	0.2	0.0	0.1	0.2	0.3
NTCT	0.0	0.1	0.1	0.4	0.0	0.0	0.0
NTGA	0.1	0.2	1.8	0.6	0.6	1.8	1.8
NTGC	0.1	0.1	0.8	1.7	0.8	2.6	2.8
NTGG	0.4	0.5	0.9	0.2	2.5	7.2	7.7
NTGT	0.1	0.3	0.4	0.1	0.2	0.2	0.3
NTTA	0.0	0.1	0.1	0.0	0.0	0.0	0.0
NTTC	0.0	0.0	0.1	0.3	0.1	0.1	0.5
NTTG	0.0	0.3	0.6	0.1	0.0	0.1	0.3
NTTT	0.0	0.0	0.2	0.7	0.0	0.0	0.0

Figure 11. **DeepPrime PAM compatibility analysis.** Alternative Cas9 nuclease and prime editing systems were evaluated. The relative Cas9 activity are shown in blue and average PE efficiencies for each PE system are shown in red (n = 2,588 for each).

6. Further methods for improving PE using optimized molecular components

It has been reported that prime editors and Cas9 activities are interdependent due to prime editors being based on bioengineered Cas9 proteins[6], it is feasible to hypothesize an improvement in PE efficiency using experimental techniques that improve Cas9 activity. One key aspect for improvement is the use of an optimized sgRNA scaffold that has a 5nt longer loop and TTTC instead of TTTT that has been reported to improve Cas9 activity[13]. Accordingly, we implemented and compared the potential improvement in PE efficiency between conventional and optimized scaffolds using Library-Small. Notably, our findings indicate that using an optimized scaffold infrastructure for pegRNA design can improve PE efficiencies by on average 1.25-folds over conventional pegRNAs. Our analysis showed that optimized scaffold design exhibited improved PE efficiencies in 79% (1,674/2,132) of the pegRNA and target sequence pairs that were examined. As PEs contain the Cas9 nickase domain, we expect that Cas9 nucleases and PE variants would exhibit similar PAM. To uncover this phenomenon, we compared the average efficiencies of prime editing and nuclease-induced indel generation at target sequences with varying 3-nt PAM sequences. As a result, we found that high correlation between them, suggesting that PEs and Cas9 share an overlapping repertoire of PAM compatibilities (Figure 11).

Traditional NGG PAM sequences present certain challenges when targeting regions of the genome with a targetable PAM sequence. To expand the targetable PAM repertoire for future studies, we applied our high-

throughput screening for all possible non-NGG PAM sequence when we expanded the definition of the PAM sequence as a sequence that yields higher than 1% editing at the associated target sequences 7 days post-transduction of the pegRNA-target pairwise library using Library-Small with optimal pegRNAs. We found 35 of the 64 3-nt PAM sequences can be used as PAMs by at least one of the PE2 variants (i.e., PE2, NG-PE2, NRCH-PE2) (Figure 11). Especially, at target sequences with NGTN, NGAN, NGCH, and NACH PAM sequence, NRCH-PE2 showed high average prime editing efficiencies, whereas PE2 showed high average efficiency of 7.3% at target sequences exhibiting the NGGN PAM. At target sequences with NWGA and NAMG PAM sequences, NG-PE2 showed the highest efficiency although the average efficiencies are 1.9% and 1.8%, respectively.

Prime editing efficiency can also be enhanced by using recently reported prime editor improvements including PE2max and PE4max[2]. We applied our high-through put screening and analysis as described and examined the editing efficiencies by PE2, PE2max, and PE4max using Library-Small. When compared to the canonical PE2, we found that when combined with optimal pegRNA conditions, PE2max and PE4max exhibited 1.9 and 2.7-folds improvement in general prime editing efficiency, respectively (Figure 11).

Lastly, we explored the variation in PE efficiency according to different cell types as they have been reported to express varying levels of mismatch repair (MMR)-related components such as *MSH2*, *MSH6*, *MLH*, and *PMS2*[2]. These intrinsic MMR-related proteins have been shown to negatively affect PE efficiencies[2]. Our experiments have predominantly utilized HEK293T cells which is partially MMR-deficient due to hypermethylation of the MLH1 promoter[14]. We conducted a more in-depth evaluation of the MMR effect on prime editing efficiency by conducting high-throughput analyses on two additional cell types, HCT116, an MMR-

deficient cell line, and MDA-MB-231, an MMR-proficient cell line using Library-Small. Our finding from comparing the PE2 efficiencies measured in HEK293T cells to those of HCT116 and MDA-MB-231 cells, we found strong correlations (R = 0.92, r = 0.89) and modest correlation (R = 0.81, r = 0.63), respectively (Figure 11). We further investigated a previous report that edit types are associated with prime editing efficiencies in HEK293T cells due to varying MMR mechanisms on different DNA alterations. Accordingly, we compared prime editing efficiencies in HEK293T cells with those of MDA-MB-231 and HCT116 cells under all nine different editing configurations using Library-Small. Interestingly, we found differential preference of MMR mechanism for certain edit configurations where MDA-MB-231 cells demonstrated higher PE efficiencies for 1bp G to C and A to G substitutions, similar or slight decrease for 1 to 3bp insertions, and a drastic reduction for 1 to 3bp substitutions. These effects were not observed in MMR-deficient HCT116 cells (Figure 11).

7. Role of synonymous mutation markers for improving on-target efficiency and sequence analysis.

It has been reported that in addition to multiple biochemical improvements to the Cas9 protein and the reverse transcriptase, the optimal structural improvements to the pegRNA scaffold complex, a secondary edit adjacent to the intended edit that is silent through a synonymous amino acid change can improve on-target efficiency by diluting the intrinsic mismatch repair mechanism (Figure 12). Furthermore, the secondary edit functions as a sequence marker for properly edited reads that can clearly distinguish the intended and secondary edit combination among other byproducts caused by sequencing or PCR-based random errors and noise. This dual function system has been implemented in the SynDesign platform to automatically

incorporate the secondary synonymous marker at a user-specified region of the pegRNA. Based on previous observation, we have established a default priority parameter of the LHA, PAM co-edit, and RHA.

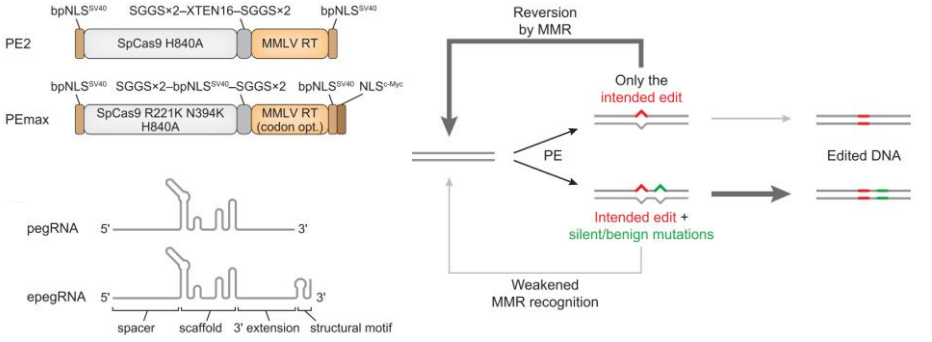

Figure 11. **Improved prime editing efficiency using optimized components.** The +max system utilizes human codon-optimized RT, a 34-aa linker containing a bipartite SV40 NLS, an additional C-terminal c-Myc NLS and R221K N394K. The +epegRNA takes advantage of structural variants that perform better than the canonical pegRNA. The silent mutation adjacent to the intended edit dilutes the cellular mismatch repair machinery and leads to higher on-target efficiency[16].

8. Overall performance optimization of SynDesign.

Our analysis pipeline has been optimized extensively for online access and simultaneous use by multiple users. First, the input tables for all databases were preprocessed and indexed according to gene identifier using random access locating through the built-in function, seek(). Furthermore, a reference point for each gene identifier was made to address “Not Found” exceptions extremely quickly. The preprocessing step of the input databases allows for fast and resource-efficient location of the correct gene information, exon positions, and genome sequence retrieval.

For processing, the webserver currently hosts multiple web portals that are independently run on two Nvidia RTX 3090 GPUs that allocate

resources according to the size of input. For example, larger input loads involving full gene runs or large exons (3k bps or larger) are distributed among 4 jobs across both GPUS. Smaller input loads such as 1k bps or smaller exon targets or single variant targets are distributed across 2 jobs on a single GPU. This job scheduling allows the GPU-optimized feature engineering and predicting scoring steps to run as efficiently as possible.

Lastly, the pipeline allows for the construction of a precision prime editing library that incorporates the silent synonymous mutation markers in the priority of the left homology arm, PAM sequence, and right homology arm.

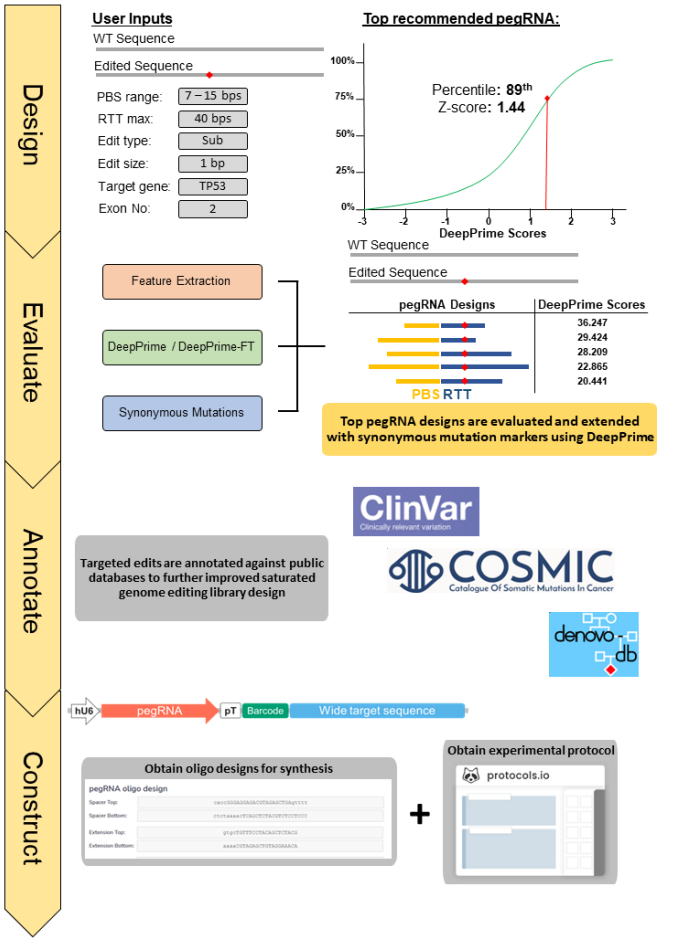

Figure 12. **Flowchart of SynDesign pipeline.** The schematic shows each major analysis step of SynDesign from the initial input to pegRNA design, feature engineering, and evaluation using DeepPrime/DeepPrime-FT. Annotation and library construction are greatly aided through the automated inclusion of synonymous mutation markers as well as annotation against public variant databases.

IV. DISCUSSION

Here we designed, synthesized, and evaluated hundreds of thousands of prime editing guide RNAs (pegRNAs) in a systematic, high-throughput manner. In doing so, we profiled these pegRNAs to distill valuable factors that impact prime editing in various molecular contexts. Our well-establish library screening methodologies combined with highly optimized analysis pipelines for determining PE efficiencies revealed key parameters in pegRNA design and function that will aid future studies aimed at utilizing prime editing in important biomedical applications. Notably, our library was based on the ClinVar database, a vast resource of human disease-relevant mutations, where the pegRNAs were designed to install or correct these mutations to investigate disease modeling or disease therapeutic conditions. Our focus on disease-relevant mutations provides useful context and platform for addressing variants that were previously difficult or impossible to profile in-depth. Accordingly, we found that with our pegRNAs designed to target ClinVar variants inducing substitutions, insertions, and deletion up to 3bp, we can cover up to 88% of the variants reported as pathogenic or likely pathogenic.

Key factors that impact PE efficiency in human cells are the sizes and sequence context of the main pegRNA components[1,6]. In general, we found that RTT length of 12-nt or less should be paired with a PBS length of 11-nt and RTT length of 13-nt and longer yields the best performance when

designed with a PBS length of 12-nt. The optimal nucleotide for the last templated positions has been a focus of debate in previous studies[1,6]. However, our extensive analysis revealed that cytosine, at the last template positions provides the most efficient prime editing performance. In general, PAM co-editing has been shown to yield improved PE efficiencies across all 15 non-GG dinucleotide variations. However, our findings show that there is a distinct hierarchy in efficiency where NAT leads to the best performance in general with CT, CC, and CA being the least effective forms of PAM co-editing. Taken together, our in-depth profiling of pegRNA performance across various design conditions and target context has demonstrated key components and their optimal range in size and nucleotide composition for maximizing general primed editing performance.

As the massive profiling experiments cannot be replicated by those seeking optimal pegRNA designs specific to their research requirements, we sought to train a deep-learning based model using our data to evaluate pegRNAs according to the user-specific parameters and predict their efficiencies in real-time without the need for time or resources of experimental procedures. DeepPrime, our convolutional neural network with recurrent neural network components is in part a hybrid model that combines CNN with bidirectional gated recurrent units (GRUs). When finely tuned using our pegRNA "Biofeatures," we found that DeepPrime exhibits high performance in predicting PE efficiencies in various contexts. DeepPrime also exhibited excellent reproducibility of endogenous targets and independent datasets.

With recent improvements to PE and its components, designing the most appropriate genome editing system for a specific experiment can be challenging. To remedy this issue and expand the application of prime editing and saturation genome editing, we generated an easy-to-use web portal for accessing our powerful analysis tool for designing and optimizing pegRNA

designs under various cellular and sequence contexts. Our web portal allows researchers to evaluate all possible pegRNA designs for a specific target gene with parameters adjusted for their experimental conditions without resource-heavy experimental procedures. SynDesign also empowers saturation prime editing library designs by automatically installing synonymous mutation markers for improved precision and on-target efficiency. Additional annotations can be conducted against public variant databases to improve designs that target disease models and improve prime editing applications in therapeutics. We expect that our analysis platform will serve as a valuable tool for future studies in exploring prime editing for understanding human disease and expanding its therapeutic potential.

V. CONCLUSION

Genome editing has established a vast field of research in manipulating human DNA for understanding human disease and developing effective therapeutic methods. As the most recent innovation in genome editing, prime editing (PE) allows for the installation or correction of virtually any desired nucleotide composition in the human genome[1,6]. Understanding the factors that impact PE efficiency is of paramount importance for its safe and effective application in biomedical research. We profiled PEs in various cellular and sequence conditions using our high-throughput library screening techniques in concert with a highly optimized PE analysis pipeline that computationally determines PE levels with additional "Biofeatures" such as thermal dynamic properties, sequence context, and secondary structures. We developed DeepPrime[15] using our large-scale profiling and feature data as the train and test set to develop a deep-learning model that utilizes convolutional neural network with bidirectional gated recurrent units. We found that DeepPrime consistently outperforms other conventional machine-learning methods by providing more accurate prediction of PE levels. We established a user-friendly web portal for accessing our PE data and analysis tools in addition to DeepPrime for future works in prime editing. We expect that DeepPrime, and our web portal will serve to better facilitate the understanding and application of prime editing in biomedical research.

REFERENCES

1. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 2019;576:149-57.
2. Chen PJ, Hussmann JA, Yan J, Knipping F, Ravisankar P, Chen PF, et al. Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. Cell 2021;184:5635-52 e29.
3. Kim HK, Kim Y, Lee S, Min S, Bae JY, Choi JW, et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. Sci Adv 2019;5:eaax9249.
4. Kim HK, Min S, Song M, Jung S, Choi JW, Kim Y, et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat Biotechnol 2018;36:239-41.
5. Kim HK, Song M, Lee J, Menon AV, Jung S, Kang YM, et al. In vivo high-throughput profiling of CRISPR-Cpf1 activity. Nat Methods 2017;14:153-9.
6. Kim HK, Yu G, Park J, Min S, Lee S, Yoon S, et al. Predicting the efficiency of prime editing guide RNAs in human cells. Nat Biotechnol 2021;39:198-206.
7. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016;44:D862-8.
8. Du D, Roguev A, Gordon DE, Chen M, Chen SH, Shales M, et al. Genetic interaction mapping in mammalian cells using CRISPR interference. Nature Methods 2017;14:577-588.
9. Shen JP, Zhao D, Sasik R, Luebeck J, Birmingham A, Bojorquez-Gomez A, et al. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nat Methods 2017;14:573-6.
10. Kim N, Kim HK, Lee S, Seo JH, Choi JW, Park J, et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. Nat Biotechnol 2020;38:1328-36.
11. Ali M. PyCaret: An open source, low-code machine learning library in Python. PyCaret version 2020;2.
12. Song M, Kim HK, Lee S, Kim Y, Seo SY, Park J, et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. Nat Biotechnol 2020;38:1037-43.
13. Dang Y, Jia G, Choi J, Ma H, Anaya E, Ye C, et al. Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome

Biol 2015;16:280.

14. Trojan J, Zeuzem S, Randolph A, Hemmerle C, Brieger A, Raedle J, et al. Functional analysis of hMLH1 variants and HNPCC-related mutations using a human expression system. Gastroenterology 2002;122:211-9.

15. Yu G, Kim HK, Park J, Kwak H, Cheong Y, Kim D, et al. Prediction of efficiencies for diverse prime editing systems in multiple cell types. Cell 2023;186:2256-72 e23.

16. Chen, P.J., Hussmann, J.A., Yan, J., Knipping, F., Ravisankar, P., Chen, P.F., Chen, C., Nelson, J.W., Newby, G.A., Sahin, M., et al. (2021). Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. Cell *184*, 5635-5652.e5629.

ABSTRACT (IN KOREAN)

프라임에디터를 이용한 질병관련유전자 돌연변이 기능 분석

<지도교수 김형범>

연세대학교 대학원 의과학과

박 진 만

재조합 DNA 기술을 통한 DNA 물질 조작은 현재 생물학 및 의학 연구의 기초가 되었다. 자연에서 재설계되거나 재구성된 기술을 통한 각각의 혁신적인 이노베이션은 복잡한 유전 메커니즘을 이해하기 위한 새로운 지평을 열었고 유전 질환을 해결할 수 있는 새로운 기회를 창출했다. 바이러스 유전 물질에 대한 적응 면역의 한 형태로 박테리아에서 처음 발견된 CRISPR-Cas9 시스템의 연구는 다양한 치료 표적에 대한 대규모 스크리닝 분석과 대다수의 생물학 및 의학 분야에 적용할 수 있게 되었다. 그 이후로 전 세계 실험실의 여러 노력으로 게놈 편집에 대한 응용 가능성, 특이성 및 프로그래밍 가능성이 향상되어 다양한 Cas9 변이체 및 기본 편집기가 개발되었다. 가장 최근에 Liu 의 그룹은 역전사효소를 Cas9 단백질에 결합하여 표준 시스템의 많은 한계를 본질적으로 개선하는 CRISPR-Cas9 시스템의 생체 공학 형태인 프라임에디팅을 도입했다. 특히, 프라임에디팅은 기증자 DNA 또는 이중 가닥 절단 없이 특정 유전자 변형의 잠재적인 조합을 도입할 수 있게 함으로써 게놈 편집을 크게 개선했다. 그러나 다양한 실험적 요인에서 프라임에디팅 효율을 향상시키기 위한 최적의 조건을 결정하는 데는 많은 시간과 자원이 필요하다. 이전 노력은 인간 세포에서 약 50K 쌍의 프라임에디팅 가이드 RNA(pegRNA)와 표적 서열의 효능을 평가했다. 이를 통해 프라임에디팅 효율성에 영향을 미치는 기능을 결정하고 pegRNA 효율성을 예측할 수 있는 세 가지 계산 모델을 구성했다. 우리의 노력이 미래 연구에서 프라임에디팅의 실제 적용을 위한 귀중한

통찰력을 제공했지만, 우리의 접근 방식은 일련의 특정 변경 유형 및 위치로 제한되었다.
본 연구에서는 데이터를 600K 쌍의 pegRNA 로 크게 확장하고 최대 3 개의 염기서열 크기까지 변경 조합을 유도하는 효율성을 목표로 합니다. 이를 통해 주요 편집 효율성에 기여하는 요소의 영향을 식별하고 평가하는 것을 목표로 한다. 또한, 우리는 ClinVar 데이터베이스에서 사용할 수 있는 광범위한 질병 관련 돌연변이 레퍼토리를 사용하여 pegRNA 및 표적 쌍을 신중하게 선별하여 질병 치료 및 모델링의 맥락에서 주요 편집 효율성을 더 잘 평가할 수 있게 도입예정이다. 우리는 또한 다양한 세포줄기에서 최적의 프라임에디팅 조건을 평가하고 최근에 보고된 프라임에디팅의 다른 변이체를 비교하는 것을 목표로 한다. 종합하면, 우리는 광범위하게 확장된 pegRNA 디자인이 보고된 질병 관련 돌연변이의 최대 88%를 포함할 수 있음을 확인했다. 우리의 대규모 프로파일링 데이터를 사용하여 최신 딥러닝 기반 알고리즘을 기반으로 크게 개선된 예측 모델을 개발했다. 우리의 노력은 기본 및 임상 연구 노력에서 프라임에디팅의 적용을 확장하는 미래 작업에 크게 도움이 되고 포괄적인 도구를 제공할 것으로 기대한다.

핵심되는 말 : 프라임에디팅, 고처리량 실험, 딥러닝 모델링

PUBLICATION LIST

Park J, Yu G, Seo SY, Yang J, Kim HH. SynDesign: web-based prime editing guide RNA design and evaluation tool for saturation genome editing. Nucleic Acid Research 2023: gkae304, https://doi.org/10.1093/nar/gkae304.