Discovery of novel EGFR VUS related to drug sensitivity using high-throughput prime editing screening

Hyeong-Cheol Oh

The Graduate School
Yonsei University
Department of Medicine

Discovery of novel EGFR VUS related to drug sensitivity using high-throughput prime editing screening

A Dissertation Submitted
to the Department of Medicine
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medicine

Hyeong-Cheol Oh

June 2024

This certifies that the Dissertation of Hyeong-Cheol Oh is approved.

[signature]

Thesis Supervisor Hyongbum Kim

[signature]

Thesis Committee Member Sun Young Rha

[signature]

Thesis Committee Member Jae-Ho Cheong

[signature]

Thesis Committee Member Sung-Rae Cho

[signature]

Thesis Committee Member Tae-Min Kim

The Graduate School
Yonsei University
June 2024

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to the individuals who played a pivotal role in the successful completion of this dissertation. Foremost, I am truly grateful to Professor Hyongbum Kim, my dissertation advisor, for his invaluable guidance, encouragement, and steadfast support. I consider myself extremely fortunate to have been under his mentorship.

I also extend my sincere thanks to the esteemed members of the dissertation committee. Professors Sun Young Rha and Jae-Ho Cheong provided insightful advice on the significance of a Ph.D. degree and drug-resistant cancer variants. Professors Sung-Rae Cho and Tae-Min Kim offered critical comments that greatly enhanced my understanding of cancer genomics.

Special appreciation goes to Seungho Lee and Younggwang Kim, whose assistance was crucial in my grasp of CRISPR screening and introduction to the realms of bioinformatics. I am grateful to all my lab members.

Lastly, I extend my deepest thanks to my beloved wife, Kilju Hong, for her unwavering support and belief in me. I sincerely hope that someday my dear son, Jihan Oh, and our eagerly awaited second child will have the opportunity to read my paper.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Discovery of novel EGFR VUS related to drug sensitivity using high-throughput prime editing screening

Variants of uncertain significance (VUS) pose challenges in the clinical interpretation of genetic information across various diseases. Despite numerous efforts by researchers to decipher the implications of these mutations, the majority of single nucleotide substitution mutations remain classified as VUSs. Existing methodologies have attempted to elucidate variant effects using approaches such as cDNA expression, Cas9-based homology-directed repair, base editing, and prime editing. However, these methods encounter limitations stemming from non-physiological approaches, low editing efficiency, inadequate coverage, and imprecise assessments, respectively, thereby failing to conclusively specify variant effects. Here, we developed SynPrime, a novel method that incorporate an additional synonymous edit near the targeted SNV, demonstrating remarkably high accuracy in evaluating variant effects. Through the application of SynPrime, we functionally assessed 2,476 SNVs within the EGFR gene, encompassing 99% of all potential variants within the canonical tyrosine kinase domain (exons 18 to 21). Moreover, we determined the resistance profiles of 95% of all conceivable EGFR protein variants encoded in the entire tyrosine kinase domain (exons 18 to 24) against afatinib, osimertinib, and osimertinib in the presence of the co-occurring T790M mutation. Our study demonstrates the potential to significantly improve the precision of variant functional assessment and contributes to addressing the issue of VUS by being applied to other genes and diseases.

Key words : VUS, prime editing, functional screening, saturation genome editing

I. INTRODUCTION

Recent advances in high-throughput sequencing technologies have precipitated a rapid identification of numerous genetic variants. However, a considerable fraction of these variants remains categorized as variants of uncertain significance (VUS). The lack of detailed functional data for these variants significantly impedes the clinical management of diseases associated with them and limits the ability to predict treatment outcomes accurately.

For instance, targeted therapies using tyrosine kinase inhibitors (TKIs) such as afatinib and osimertinib have shown promising results in the treatment of lung cancers harboring specific mutations in the epidermal growth factor receptor (EGFR) gene[1-6]. These TKIs function by competitively binding to the intracellular ATP-binding site of EGFR, effectively inhibiting ATP attachment and subsequent receptor activation[7]. They are particularly effective against certain EGFR mutations, such as in-frame deletions in exon 19 and the L858R mutation[8,9]. However, the emergence of TKI-resistant variants poses a significant challenge, as these often result from additional substitution mutations[2,10,11]. Although a subset of drug-resistant variants has been identified, the majority of EGFR variants are still classified as VUS with regard to their resistance to TKIs. These variants occupy a clinical gray area, complicating treatment decisions and potentially leading to less effective therapeutic outcomes[12]. Therefore, establishing a comprehensive resistance profile for all EGFR variants against representative TKIs like afatinib and osimertinib could significantly enhance clinical outcomes by facilitating the selection of the most effective TKI for individual patients and devising strategies to counteract drug resistance.

Despite various efforts to elucidate the effects of VUS, there is a notable gap in methodologies that can precisely determine their functional impacts. Traditional approaches, such as the use of cDNA-based transgene libraries, while useful, often fail to replicate the natural biological effects of variants because of potential overexpression artifacts[13-16]. An emerging solution involves generating variants directly at their endogenous loci, which is expected to yield more accurate functional insights. Techniques such as homology-directed repair (HDR), base editing[17-19], and prime editing[20] have been employed for this purpose. saturation genome editing (SGE) was previously conducted to map all possible single nucleotide variants (SNVs) of the BRCA1 gene using HDR, but the limited efficiency of HDR in mammalian cells—often necessitating the use of haploid cells for precise

evaluations—has restricted its broader application[21]. Base editing, known for its higher overall efficiency, can unfortunately address only a subset of possible SNVs due to its specific mutation capabilities and the frequent occurrence of bystander effects[22-26]. Prime editing offers a promising alternative, capable of introducing precise small-scale genetic modifications without inducing double-strand breaks[20]. It has been applied to evaluate variants within selected regions of NPC1 and BRCA2 in a modified HEK293 cell line; however, this method was able to generate and assess only 12% to 34% of all possible SNVs in each targeted exon, leaving a significant proportion as VUS[27]. Thus, there remains an urgent need for a robust and efficient method for SGE that can comprehensively generate and functionally test nearly all possible SNVs at endogenous sites in disease-relevant non-haploid cells.

In this study, we introduce SynPrime, a groundbreaking method that incorporates a synonymous edit adjacent to the target single nucleotide variant (SNV). This addition significantly enhances the precision and efficiency of variant generation and evaluation. Using SynPrime, we have successfully generated and cataloged 2,476 EGFR SNVs, which translate into 1,726 distinct protein variants. This dataset includes 99% (1,137 out of 1,146) of all conceivable protein variants derived from SNVs within the canonical tyrosine kinase domain sequence (exons 18 to 21) and 95% (1,726 out of 1,817) of all protein variants across the entire tyrosine kinase domain (exons 18 to 24)[28]. Further, we conducted an exhaustive resistance profiling of these variants against afatinib and osimertinib, including their interaction with the T790M mutation, using PC-9 cells. These cells feature a prevalent EGFR exon 19 in-frame deletion that typically confers increased sensitivity to TKIs[8]. By generating a comprehensive drug resistance profile related to EGFR, our study significantly contributes to enhancing the precision of therapeutic decision-making in clinical settings. Moreover, the successful application of SynPrime to evaluate nearly all possible variants within a clinically relevant region of EGFR paves the way for a deeper understanding of the role of VUS in other genes and pathological conditions.

II. MATERIALS AND METHODS

2.1. Design of the pegRNA libraries

Utilizing the reference human genome (hg38), we designed pegRNA libraries to introduce all possible nucleotide substitutions across the coding DNA sequence (CDS) of the EGFR gene, spanning exons 18 to 24 (denoted as Syn-exon18 to Syn-exon24). The PC-9 cell line, characterized by a 15-bp in-frame deletion within exon 19 of EGFR, was specifically targeted by designing pegRNAs based on this deletion variant. Initial selection of pegRNAs was performed by calculating their DeepPrime-FT scores, prioritizing those designed to induce single-nucleotide variants (SNVs). PegRNAs with left homology arms (LHAs) under 5 bp were excluded, and the top three pegRNAs per SNV were selected based on their scores. Additionally, to ensure the integrity of the editing process, synonymous substitutions were incorporated into different codons within the right homology arm (RHA) of the reverse transcriptase (RT) template. This strategy was intended to (i) differentiate true edited sequences from erroneous or partially incorporated sequences during sequencing, and (ii) disrupt the native DNA repair pathways potentially improving editing outcomes. Out of 2,610 evaluated SNVs, 95 were excluded due to constraints on the maximum allowable RT template length, leading to a library consisting of 2,515 SNVs. For 36 SNVs, no synonymous substitutions were incorporated due to the absence of viable silent substitution options. The library also included 198 sham-editing and 198 nontargeting pegRNAs as negative controls.

For the ribosomal protein L15 (RPL15), located in exon 2, pegRNAs were similarly designed to target all possible 1-bp substitutions within its CDS. Following the exclusion of pegRNAs with insufficient LHAs, the top three pegRNAs for each SNV were selected based on the highest DeepPrime-FT scores. Each selected pegRNA also incorporated an additional synonymous substitution within the LHA of the RT template, resulting in a total of 1,492 pegRNAs targeting 500 SNVs.

In designing pegRNAs for exon 20 of EGFR, we employed a pilot version of the DeepPrime-FT model, tailored for NRCH-PE2max in HEK293T cells, to calculate efficiency scores for all possible nucleotide changes within the CDS. The selection criteria mirrored those used for the Syn-exon18 to Syn-exon24 library, with the top three pegRNAs for each SNV being prioritized based on their DeepPrime-FT score. To establish robust controls, 86 sham-editing and 85 nontargeting pegRNAs were included, culminating in a library comprising 1,845 pegRNAs.

2.2. Construction and cloning of pegRNA libraries

The design of the EGFR-targeting pegRNA libraries was subdivided by exon to create specific

libraries for exons 18 through 24, each enriched with control pegRNAs including both sham-editing and nontargeting pegRNAs.

For the assembly of these libraries, oligonucleotides synthesized by Twist Bioscience were structured into 228-bp segments containing critical components essential for effective prime editing. These components included: (1) a 19-bp homology arm situated downstream of a U6 promoter; (2) a 20-bp sequence starting with a "G", followed by a 19-bp spacer; (3) a random 18-bp sequence bordered by BsmBI restriction sites; (4) an RT template combined with a primer binding site; (5) an 8-bp linker derived from pegLIT tools, followed by a 37-bp tevopreQ1 sequence and a 7-bp poly-T tail; (6) a unique 18-bp barcode for individual pegRNA identification, prefixed by a constant "GTCAG" sequence for analytic consistency; (7) variable-length "Buffer" sequences to standardize oligonucleotide lengths across different pegRNAs; and (8) a 20-bp unique homology arm for targeted library amplification.

These oligonucleotides, each approximating 4 ng in mass, were subjected to PCR amplification using a high-fidelity enzyme system under stringent conditions designed to minimize errors. The amplicons were then size selected via agarose gel electrophoresis and assembled into a pre-linearized vector using a seamless cloning method. The assembled vectors were then transformed into electrocompetent cells to generate a comprehensive plasmid library without scaffold sequences.

Further refinements involved synthesizing an enhanced version of the SpCas9 sgRNA scaffold[31], which included strategic BsmBI sites for seamless cloning. These elements were assembled with previously generated scaffoldless plasmids in a ligation reaction, pooled, precipitated, and then transformed to construct the finalized plasmid libraries.

2.3. Cell lines and culture

PC-9 cells, instrumental for testing EGFR-targeted pegRNA efficiency due to their known genetic deletion, were cultured under standard conditions to ensure vitality and responsiveness. Cells were passaged every 3 days and maintained in 37°C incubators with 5% CO2. Optimal culture conditions, including specific nutrient and antibiotic concentrations, were meticulously upheld to support ongoing cellular health and experimental reliability.

2.4. Construction of plasmid vectors

To generate a lentiviral vector for expressing PEmax, three linear DNA fragments were assembled: (1) a linearized lentiviral backbone from Lenti-Cas9 Blast (Addgene #52962)[32] (XbaI and EcoRI (NEB); (2) a sequence encoding Cas9 containing R221K, N394K and H840A mutations amplified from pCMV-PEmax-p2A-BSD[34] (Addgene #174821) using primer pair 4/5; and (3) optimized MMLV-RT and NLS sequences and a blasticidin resistance gene amplified from pCMV-PEmax-p2A-BSD using primer pair 6/7. The assembled vector is referred to as pLenti-PEmax-p2A-BSD.

To generate a lentiviral vector for expressing plenti-gRNA-puromycin-p2A-luciferase, four linear DNA fragments were assembled: (1) a linearized lentiviral backbone from Lenti-gRNA-puro (Addgene #84752)[30] digested with FseI-NcoI; (2) the sequence encoding EF-1a and the puromycin resistance gene amplified from Lenti-gRNA-Puro using primer pair 16/17; (3) the sequence encoding luciferase amplified from pLenti HRE-Luc PGK Hygro (Addgene #118706)[33].

2.5. Lentivirus production

HEK293T cells were seeded in culture dishes and later transfected using Lipofectamine 3000. The plasmids used were specifically aimed at lentiviral production with particular molar ratios provided for precision.

The medium containing the virus was collected at specific time points post-transfection, processed to remove cell debris, filtered, and stored at -80°C.

2.6. Generation of cell lines

Prime editor expressing PC-9 Cell: PC-9 cells were transfected with CRISPR-Cas9 and sgRNA plasmids to knockout the MLH1 gene. Post-selection, these cells were further transduced with lentiviral vectors to express prime editors and selected again using blasticidin.

T790M-expressing PC-9 cells: MLH1-knockout PC-9 cells were transfected with a mixture of a plasmid expressing BE4max (pCMV-BE4max; Addgene #112093)[35] and a plasmid expressing sgRNA (pLKO5.sgRNA.EFS.tRFP; Addgene #57823)[36]. Cells expressing the transgene were sorted using FACS at day 7 post-transfection. these cells were further transduced with lentiviral vectors to express prime editors and selected again using blasticidin.

2.7. SynPrime pooled assay

For the experiments, each exon was assayed in duplicate. Cells were seeded at approximately 10,000 cells per pegRNA, 24 hours prior to the transduction with the lentiviral library. The multiplicity of infection (MOI) was set to 0.5 to ensure a representation of each pegRNA in about 5,000 cells. After 24 hours of infection, the medium was replaced with one containing puromycin, and the cells were cultured for an additional nine days. Following the puromycin removal on day 0, cells were harvested to achieve 5,000-fold coverage of the pegRNA libraries. Subsequently, cells were segregated into a treatment group receiving drugs—afatinib at 3 nM and Osimertinib at 8 nM—and a control group, both maintained for 10,000-fold coverage of the pegRNA libraries. Cells were passaged every three days over a ten-day period and then collected for genomic DNA extraction.

In the case of RPL15 evaluations, the process mirrored the protocol up to the puromycin removal, post which cells were cultured for four additional days before harvesting. The cells were then maintained for 10,000-fold pegRNA library coverage, passaged every three days, and collected after ten days for genomic DNA extraction.

2.8. Genomic DNA preparation and deep sequencing

Genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega) following the manufacturer's instructions. For deep sequencing, pegRNA-specific barcodes and unique molecular identifiers (UMIs) were amplified from the genomic DNA using a two-step PCR process with the 2X Pfu PCR Smart Mix (Solgent). The first PCR involved multiple 50-μL reactions, each containing 3 μg of genomic DNA, 20 pmol each of forward and reverse primers (targeting both endogenous genomic regions and RT template-specific sequences), and 25 μL of PCR pre-mix. The thermal cycling conditions were set as follows: an initial denaturation at 95 °C for 2 minutes, followed by 20 cycles of 95 °C for 30 seconds, 60 °C for 30 seconds, and 72 °C for 40 seconds, concluding with a final extension at 72 °C for 5 minutes.

The resulting amplicons, representing over 10,000-fold coverage of the libraries, were pooled, concentrated using the MEGAquick-Spin Total Fragment DNA Purification Kit (iNtRON Biotechnology), and size-selected via agarose gel electrophoresis. In the second PCR, 20 ng of the purified PCR products were used in two separate reactions containing 20 pmol of Illumina indexing primers. This step consisted of 5 cycles using the same thermal profile as the first PCR. The final PCR products were then size-selected and sequenced on a NovaSeq 6000 (Illumina).

2.9. Data analysis and variant filtering

To identify single nucleotide variants (SNVs) introduced by the Syn-RPL15 and Syn-exon18 through Syn-exon24 libraries, we constructed a SNV reference sequence library from the coding sequence (CDS) and adjacent intronic regions (NM_005228.5). This reference included only the intended SNVs and synonymous substitutions, devoid of other mismatches or indels. A similar approach was taken for the NRCH-exon20 library.

Sequencing reads were aligned to the SNV reference sequence library, and only those reads that perfectly matched the intended SNV sequences were considered. Unedited wild-type reads were identified by their perfect alignment to the reference sequence (NM_005228.5). To discern true prime-edited variants from sequencing or library preparation artifacts, we calculated the odds ratio (OR) and P-values using Fisher's exact test, comparing the reads from day 0 to those from unedited cells. True edited SNVs were defined as those with a P-value less than 0.05 and an OR greater than 3, with an additional filtering criterion of at least 0.5 reads per million (RPM) at day 0. These stringent criteria ensure the reliability and significance of detected SNVs.

$$Odds\ ratio = \frac{(Reads\ of\ SNV\ at\ day\ 0 + 1)\ /\ (Reads\ of\ wild\ type\ sequence\ at\ day\ 0 + 1)}{(Reads\ of\ \ SNV\ in\ unedited\ cell + 1\ )/(\ Reads\ of\ wild\ type\ sequence\ in\ \ unedited\ cell + 1)}$$

2.10. Calculation of adjusted LFCs

For syn-RPL15 data, Log2 fold change (LFC) values for each single nucleotide variant (SNV) were computed by contrasting allele frequencies between cells at day 10 and day 0. Regarding Syn-exon20 data, LFCs were determined by comparing allele frequencies in untreated cells at day 10 with those at day 0 for EGFR dependency analysis. Additionally, LFCs were assessed by contrasting allele frequencies in drug-treated cells at day 10 with those in untreated cells to establish drug resistance profiles. Subsequently, these LFC values underwent normalization utilizing LOWESS (Locally Weighted Scatterplot Smoothing) regression.

2.11. Modelling positional biases and normalization across exons

To account for sequence context influences on editing efficiency, LFCs for each SNV were normalized using synonymous SNVs, which are known for their neutral LFC as they do not induce

amino acid changes. We employed LOWESS regression to compute the regressed synonymous SNV LFC at each position in every exon. Standardized LFC values were then obtained by subtracting the synonymous SNV LFC at each position (LOWESS regressed) and dividing by the standard deviation of synonymous SNVs in each exon for inter-exon comparison. Finally, the standardized LFC of each replicate was averaged to obtain the adjusted LFC[21].

$$Standardized\ LFC = \frac{LFC\ of\ each\ SNV - Synonymous\ LFC\ (Lowess\ regressed)}{Standard\ deviation\ of\ synonymous\ SNVs}$$

To calculate the adjusted LFC based on protein variants, we averaged the adjusted LFC values of SNVs inducing the same protein variant.

2.12. SNV drug resistance classifications

Drug resistance induced by the Syn-exon18 through Syn-exon24 libraries was analyzed using cutoff values determined by the distribution of resistance scores of synonymous SNVs within each exon. SNVs were classified as follows:

- Resistance: SNV resistance scores above the 99.7th percentile relative to scores of synonymous SNVs in both replicates.
- Sensitive: SNV resistance scores below the 95th percentile relative to scores of synonymous SNVs in both replicates.
- Intermediate: SNVs falling into neither the ‘Resistance’ nor ‘Sensitive’ categories.

Classification based on protein variants followed similar criteria employed in SNV-based classification.

2.13. pegRNA barcode count and library-based screening analysis

PegRNA barcode counting and fold change calculation of SNV frequencies between treatment and control groups were performed using an in-house Python code[24]. We identified a barcode motif allowing a single base mismatch and a 20-bp unique homology sequence, permitting one base mismatch. UMIs were collapsed based on fold change in read count, and subsequent classification

of pegRNAs was based on normalized LFCs in both replicates using cutoff values determined by the distribution of the LFCs of non-targeting and sham-editing pegRNAs in each exon.

2.14. Resistance profiles of EGFR variants in previously published literature

Sixteen SNVs within the EGFR gene from the ClinVar database were identified[38]. SNVs conferring resistance to afatinib, osimertinib in the absence of T790M, or osimertinib in the presence of T790M in at least one dataset were manually verified against the PubMed database[39-41].

2.15. Evaluation of resistance in conventional manner

PegRNA sequences designed to generate specific SNVs were cloned into BsmBI-linearized Lenti gRNA Puro. MLH1-knockout PC-9 cells were transduced with lentivirus harboring pegRNA sequences.

In total, 3 million PEmax-expressing MLH1-knockout PC-9 cells per pegRNA were seeded in 150-mm culture dishes 24 hours before transduction in triplicate. The cells were infected with lentivirus harboring pegRNA sequences at a high MOI (~1). For a negative control, a separate population of 3 million cells were infected with lentivirus harboring empty vectors. The day after transduction, the medium was replaced with fresh medium containing puromycin, after which cells were incubated for an additional nine days.

Ten days after infection, cells transduced with pegRNA sequences and cells transduced with an empty vector were mixed at a 25:75 ratio. These mixed cell populations were then divided into untreated and drug-treated conditions. Deep sequencing of amplified pegRNA-targeted genomic sites was performed for analysis.

2.16. In vivo xenograft studies

All animal study protocols involving mice were ethically approved by the Institutional Animal Care and Use Committee (IACUC) of Yonsei University Health System (Seoul, Korea). We designed an experimental setup aimed at evaluating the efficacy of specific single nucleotide variants (SNVs) using pegRNA-mediated genome editing. Initially, we identified three pegRNAs targeting SNVs with the highest resistance scores, namely p.C797S, p.S811F, and p.S811Y. Additionally, we selected 5 pegRNAs encoding SNVs classified as intermediate and 15 pegRNAs encoding SNVs

classified as sensitive. To provide a comprehensive assessment, 2 pegRNAs encoding intermediate and 3 pegRNAs encoding sensitive synonymous SNVs were also included. This compilation of 28 pegRNAs constituted the Syn-vivo-exon20 library.

Subsequent to lentiviral transduction of target cells at a Multiplicity of Infection (MOI) of 0.3, infected cells were cultured for 24 hours. The culture medium was then replaced with fresh medium supplemented with puromycin, and the cells were cultured for an additional 9 days. Upon puromycin withdrawal (designated as day 0), $1x10^7$ cells were subcutaneously implanted into NOD/Shi-scid, IL-2Rγ KOJic (NOG) mice along with Matrigel. The mice were stratified into three groups, each comprising paired mice: one receiving the drug challenge and the other remaining untreated. At the time of cell implantation, all mice were 6 weeks old to ensure consistency across the experimental cohort. Ten days post-implantation, tumor formation was confirmed in all mice through in vivo imaging using an IVIS system. Following tumor confirmation, the mice underwent an oral drug regimen for a period of two weeks. The untreated group received a vehicle consisting of 5% DMSO, 40% PEG300, 5% TWEEN, and 50% distilled water administered orally once daily. Conversely, the drug-challenged group received the vehicle along with osimertinib at a dose of 2.5 mg/kg, also administered orally once daily.

Throughout the experimental period, mice were closely monitored on a daily basis to assess general mobility, morbidity, and mortality. Body weight was recorded twice a week to evaluate any potential adverse effects of the drug regimen. Following the completion of the treatment period, all mice from both experimental groups were humanely euthanized, and tumors were excised for genomic DNA extraction and subsequent deep sequencing analysis.

2.17. Statistical analysis

All statistical tests described were performed as two-tailed tests using Python software packages.

2.18. Data visualization

Figures were created with Python. Schematics were created with BioRender.com. PyMOL (version 2.5.5) was used to map the variants onto the following crystal structures from the Protein Data Bank: PDB: 6JXT, (EGFR complex with osimertinib).

2.19. Data and Code availability

We have submitted the deep sequencing data from this study to the National Center of Biotechnology Information's Sequence Read Archive under accession number PRJNA1018283.

Screening data were analyzed with in-house custom Python scripts and MAGeCK (version 0.5.9.3). Custom Python scripts (version 3.8.16) were used to generate input files based on grouped UMIs for MAGeCK. They are available at https://github.com/oreolic/SGE_EGFR and https://github.com/po99044/SynPrime

III. RESULTS

3.1. Sequencing errors can hinder accurate identification of SNVs induced by prime editing

Sequencing errors pose a significant challenge to the precise identification of single nucleotide variants (SNVs) induced by prime editing. Previous endeavors in saturation genome editing (SGE), aimed at exploring all possible SNVs within a target region, have predominantly relied on haploid or partially haploid cells due to editing efficiency limitations. However, since most human cells are diploid, analyzing variants in non-haploid cells offers a more accurate representation of disease pathophysiology. Additionally, the conversion of haploid cells to diploid cells is facile, and only a limited range of haploid cell types are available, restricting the widespread application of SGE. In this study, we employed PC-9, a non-haploid lung adenocarcinoma cell line harboring an in-frame deletion in EGFR exon 19 (E746-A750 deletion), a prevalent mutation conferring sensitivity to tyrosine kinase inhibitors (TKIs) in non-small cell lung cancer (NSCLC)[9].

Efficient SGE is crucial for the effective and accurate functional analysis of potential variants. To enhance SGE efficiency for EGFR mutations, we initially utilized NRCH-PE2max, employing Cas9-NRCH as the nickase domain due to its broad PAM compatibility, surpassing other SpCas9 variants with different PAM compatibilities[44-46]. PC-9 cells expressing NRCH-PEmax were generated via lentiviral transduction (**Figure 1**). We designed 1,674 prime editing guide RNAs (pegRNAs) using DeepPrime-FT, a deep learning model predicting pegRNA efficiencies, to induce SGE in the 186 bp-long exon 20 of EGFR[44]. Subsequently, we constructed a lentiviral pegRNA library, NRCH-exon20, containing 1,845 pegRNAs, including the designed pegRNAs and control

pegRNAs (**Figure 1b**). To further enhance prime editing efficiency, we introduced dominant negative MLH1 (MLH1dn) via lentiviral delivery, inhibiting the DNA mismatch repair (MMR) system[47].

Ten days after delivering the NRCH-exon20 library, deep sequencing of endogenous EGFR exon 20 was conducted. Analysis of the deep sequencing data revealed that the percentage of reads with a single substitution in the prime-edited cell population was only marginally higher (10%) than that observed in the unedited control cell population (9.2%) (**Figure 1c**). This observation suggests that a considerable proportion of reads with a single substitution in the prime-edited cell population may be attributed to sequencing or PCR errors. To determine the level of prime editing for each substitution, we compared the read counts of each SNV between the prime-edited and unedited cells, calculating the odds ratio and P-value using Fisher's exact test. Considering an intended SNV as significantly generated if it met specific criteria (odds ratio > 3 and $P < 0.05$, with no mutations at the remaining 185 positions), only 6% (34 out of 558 possible SNVs) were identified as significantly generated (**Figure 1d-f.**). The median frequency of these 34 SNVs among all exon 20 sequencing reads was merely 0.043% (range: 0.0028% to 0.12%). While these frequencies were generally higher than the median read frequency of 0.011% (range: 0.00050% to 0.13%) for all SNVs, they were still too low to reliably identify SNVs by sequencing, considering sequencing errors are approximately 0.1% per base pair[48,49].

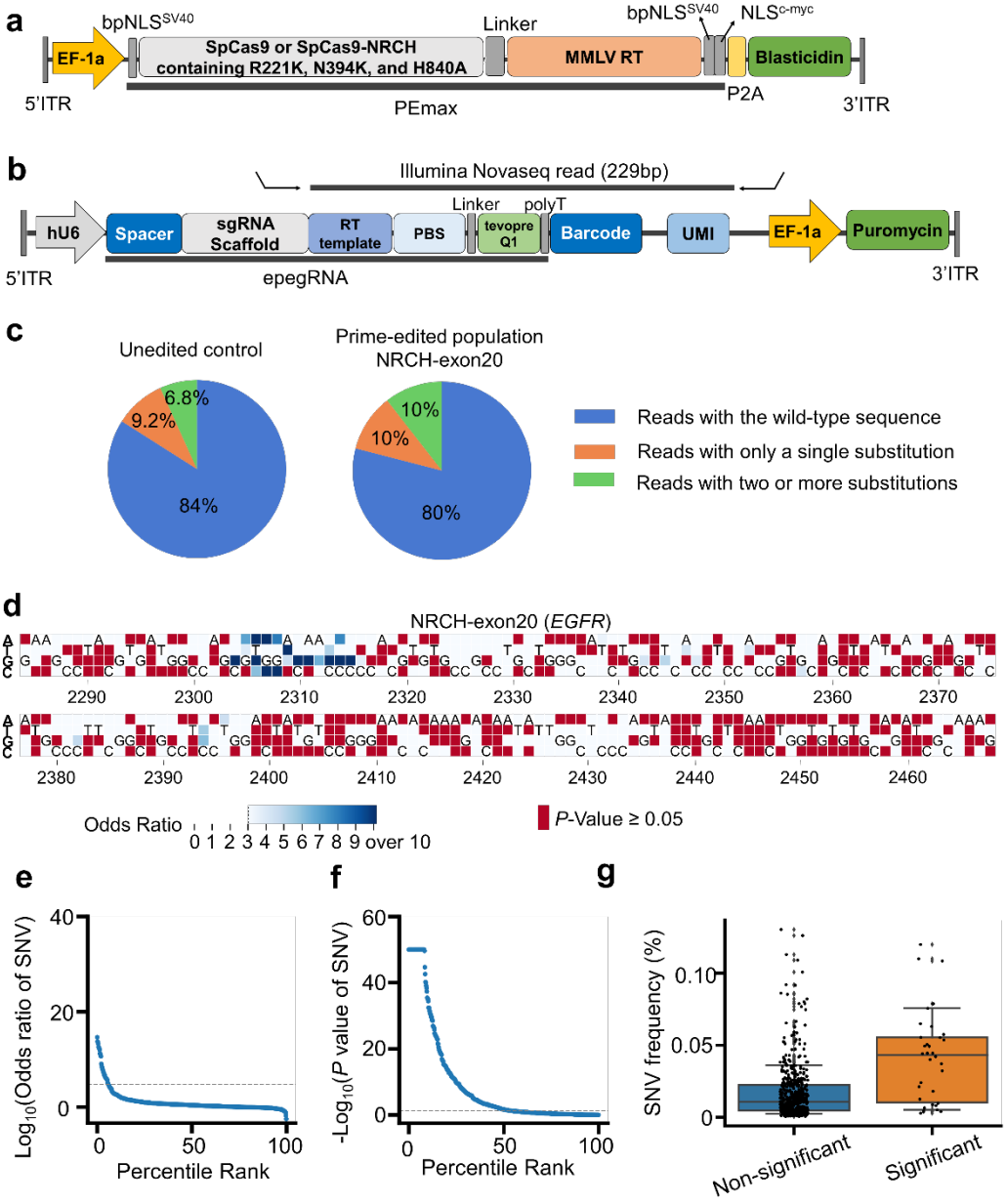


Figure 1. Sequencing errors can hinder the accurate identification of SNVs induced by prime editing. **(a),** Lentiviral vector maps illustrating the expression cassette for the prime editor (PEmax). Components include bipartite nuclear localization signal (bpNLS), human codon-optimized Moloney murine leukemia virus reverse transcriptase (MMLV-RT), and inverted terminal repeat

(ITR) sequences. **(b)**, Representation of lentiviral vector maps showing the locations of PCR primers used for deep sequencing of a 229-bp region encompassing the RT-template, primer binding site (PBS), pegRNA barcode, and unique molecular identifier (UMI). **(c)**, Proportion of sequencing reads containing substitution(s) in unedited PC-9 cells expressing NRCH-PEmax (left, unedited control) compared to those ten days post-transduction with the NRCH-exon20 library (right) for editing exon 20 of EGFR. **(d),** Heatmap displaying odds ratios and/or P-values of 558 SNVs generated by prime editing in exon 20 of EGFR ten days post-transduction with the NRCH-exon20 library. SNVs with P-values > 0.05 by the two-sided Fisher's exact test are indicated in red, where odds ratios are not displayed. SNVs with odds ratios < 3 are represented with a white background. The bottom numbers on the heatmap correspond to the EGFR coding sequence position, with the nucleotide in the reference sequence indicated. **(e-f)**, Distribution of odds ratios (e) and P-values (f) of 558 SNVs in cells transduced with the NRCH-exon20 library. Dashed horizontal lines denote the thresholds where odds ratio = 3 (e) and P-value = 0.05 (f). **(g)**, Distribution of observed SNV frequencies in PC-9 cells expressing PEmax ten days post-transduction with NRCH-exon20. The number of SNVs, n = 524 (Nonsignificant), n = 34 (Significant). Box plots depict the 25th, 50th, and 75th percentiles, with whiskers indicating the 10th and 90th percentiles.

3.2. SynPrime enables precise identification of SNVs induced by prime editing

To enhance the precision of SNV detection, we engineered a synonymous mutation adjacent to the targeted mutation, serving as a robust indicator of successful prime editing (**Figure 2a**). This strategy is expected to provide a reliable marker and potentially increase the efficacy of the desired edit by reducing the activity of the mismatch repair (MMR) system[47]. Furthermore, we utilized the final published version of DeepPrime-FT for selecting highly efficient pegRNAs. Predictive analytics from DeepPrime-FT indicated superior performance of PEmax over NRCH-PEmax (**Figure 2b**), prompting us to utilize PEmax in subsequent experiments. A lentiviral library, designated as Syn-exon20, was constructed, comprising sequences for 1,762 pegRNAs. This includes 1,674 pegRNAs for the 186 bp segment of EGFR exon 20, aiming to introduce synonymous substitutions alongside the intended edits, and 88 control pegRNAs (**Methods**).

We hypothesized that a complete knockout of MLH1 would further enhance prime editing efficiencies compared to merely expressing a dominant-negative MLH1 (MLH1dn). A 99% deletion/indel frequency at the MLH1 locus was achieved in PC-9 cells using SpCas9 and two sgRNAs targeting MLH1, indicating effective knockout (data not shown). Upon assessing prime editing outcomes, we observed significantly higher efficiencies in MLH1-knockout cells than in both MLH1dn-expressing and wild-type PC-9 cells (**Figure 2c**). Additionally, drug affinity analysis revealed no significant difference in IC50 values between MLH1 wild-type and knockout conditions (**Figure 3**).

Following the transduction of PEmax-expressing MLH1-knockout cells with the Syn-exon20 library, deep sequencing of EGFR exon 20 was performed ten days post-transduction. The median frequency of detected SNVs, both with and without the synonymous mutations, was 0.038% (range: 0.0014% to 0.36%) (**Figure 2d**), reflecting a 3.4-fold increase from the initial approach. Moreover, a 3.8-fold increase in the median odds ratio and enhanced -log10(P-values) were observed. These improved SNV detection metrics were attributed to four primary factors: (i) selection of the optimal PE system (PEmax instead of NRCH-PEmax), (ii) refined pegRNA design by leveraging the completed version of DeepPrime-FT, (iii) introduction of synonymous mutations, and (iv) strategic knockout of MLH1 rather than using MLH1dn. As a result, 60% (334 out of 558) of the potential SNVs were successfully identified as significantly generated (**Figure 2d and Figure 2e top**). Despite this substantial improvement, achieving near-complete SGE still requires further

enhancements.

For refined SNV identification accuracy, we introduced an additional synonymous mutation as a marker. By exclusively analyzing 'double hits'—sequences featuring both the intended substitution and the synonymous mutation—the median frequency of intended SNVs decreased to 0.015% (range: 0.00006% to 0.29%) (**Figure 2d**). However, this 'double hit' counting approach, termed SynPrime (Synonymous Prime Editing), led to a dramatic 252-fold increase in the median odds ratio and an overall rise in -log10(P-values) compared to the method of counting both 'double hits' and 'single hits' (sequences containing only the intended substitution). Remarkably, the SynPrime method enabled the generation and identification of 99% (555 out of 558) of all possible SNVs (**Figure 2e**), demonstrating near-complete saturation genome editing.

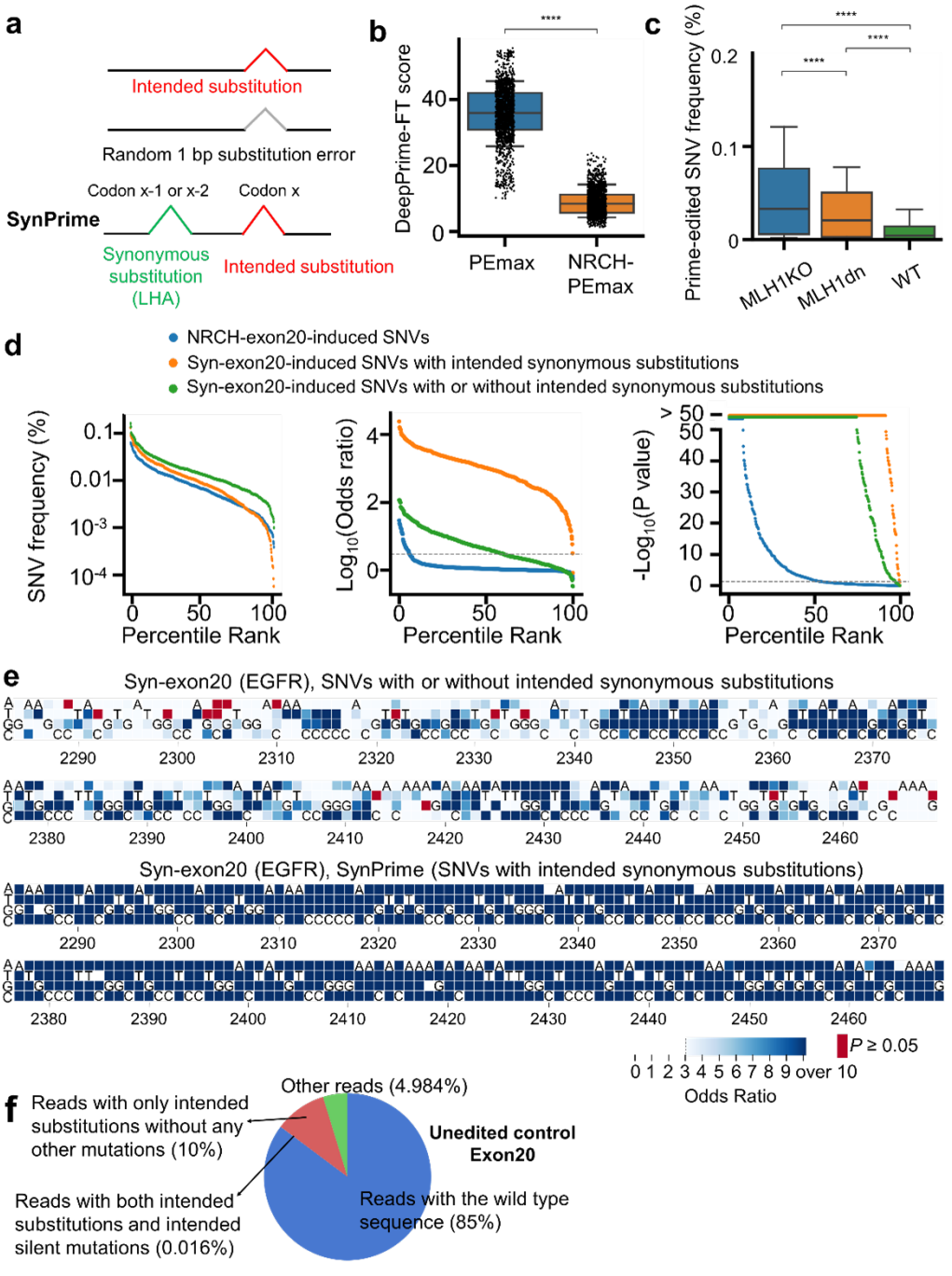


Figure 2. SynPrime enables the accurate identification of prime editing-induced SNVs. (a), Schematic representation of the SynPrime methodology which includes an additional synonymous mutation proximal to the target site. This approach is designed to enhance the detectability and verification of prime editing events. **(b)**, Distribution of efficiency scores from DeepPrime-FT for 1,674 pegRNAs designed for use with PEmax versus those designed for NRCH-PEmax. Box plots illustrate the 25th, 50th (median), and 75th percentiles, while whiskers extend from the 10th to the 90th percentiles. ****$P < 10^{-4}$ (Two-sided Student's t-test). **(c)**, Box plots showing the frequencies of 558 distinct SNVs generated by prime editing in PC-9 cells, categorized by MLH1 status. The central line represents the median, with boxes marking the interquartile range (25th to 75th percentiles), and whiskers extending from the 10th to the 90th percentiles. ****$P < 10^{-4}$ (two-sided paired t-test). **(d)**, Comparative analysis of SNV frequencies, odds ratios, and P-values in unedited control versus prime-edited cells (NRCH-exon20). The dashed lines indicate critical thresholds where the odds ratio reaches 3 (middle) and P-value meets 0.05 (right), serving as benchmarks for significant editing events. **(e),** Heatmap visualization of the odds ratios and P-values for 558 SNVs in EGFR exon 20, ten days post-transduction with the Syn-exon20 library. SNVs are coded: red indicates P-values greater than 0.05 as per a two-sided Fisher's exact test (suggesting non-significance and thus absence of displayed odds ratios); odds ratios less than 3 are displayed on a white background. The heatmap's bottom axis labels the nucleotide positions within the EGFR coding sequence, providing a detailed spatial context for observed edits. Editing identification is categorized by the detection of solely the intended edit (top) and the simultaneous presence of both the intended and synonymous edits (bottom). **(f)**, Proportion of sequencing reads from unedited control PC-9 cells across EGFR exon 20.

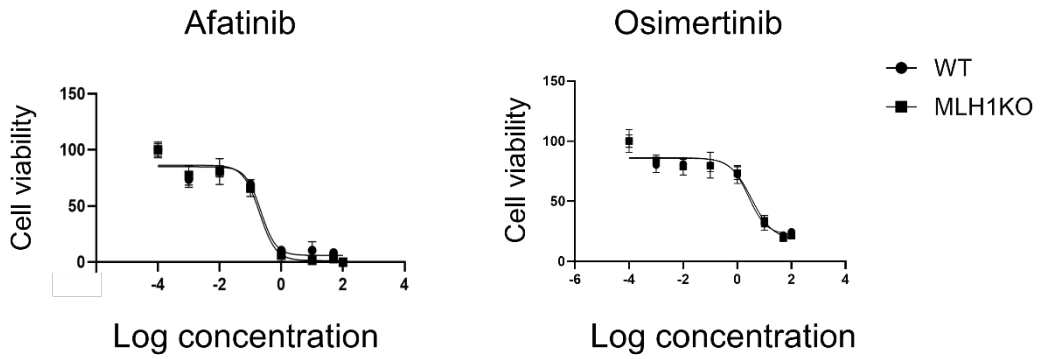


Figure 3. IC50 measurement according to the MLH1 status against afatinib and osimertinib

We posited that the marked improvement in SNV identification accuracy provided by the SynPrime method was primarily due to its efficacy in eliminating reads containing errors from wild-type alleles (i.e., false positives). To test this, we sequenced the 186-bp region of EGFR exon 20 in unedited PC-9 cells, discovering that reads with single base-pair substitution errors—indicative of false positives—accounted for 10% of the total reads (**Figure 2f**). n sharp contrast, the application of the SynPrime strategy, which involves detecting sequences with both the intended and a synonymous substitution, dramatically reduced the proportion of these erroneous reads to merely 0.016%, representing a substantial 625-fold decrease in false positives.

Furthermore, we determined the point mutation error rate within this sequencing experiment to be approximately 0.06% per base pair. This rate is comparable to, or slightly better than, the widely recognized standard sequencing error rate of 0.1%[48,49]. Extending our analysis to other genomic regions in the unedited control cells, we observed an increase in the frequency of false-positive SNVs with the extension of the sequenced region length (**Figure 4**). Nevertheless, the point mutation rates per base pair remained consistent across these regions, ranging from 0.038% to 0.095%, with a median of 0.055%.

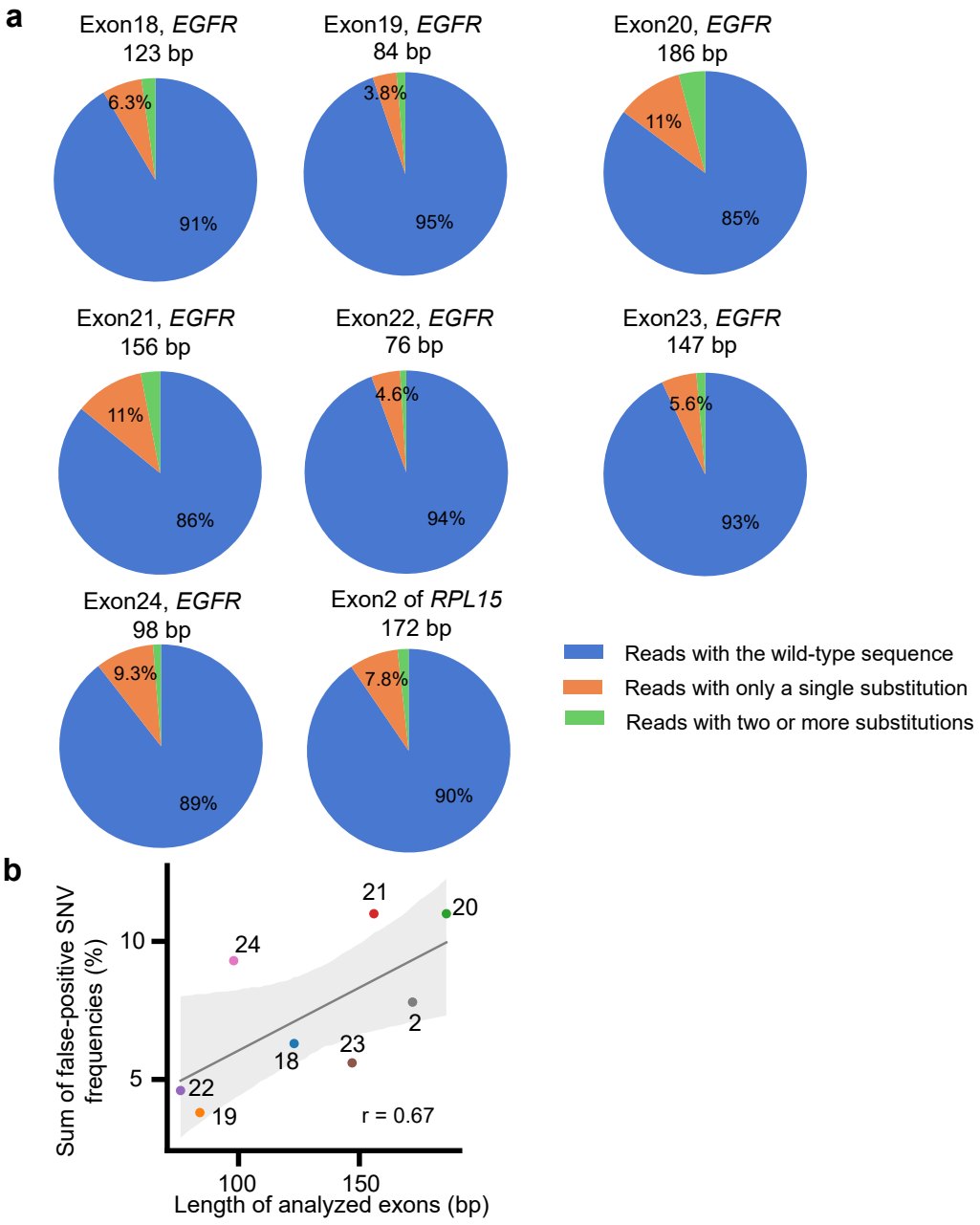


Figure 4. Frequencies of false-positive SNV reads. (a), Distribution of false-positive single nucleotide variant (SNV) reads in sequencing data from unedited PC-9 cells. The chart displays the proportion of sequencing reads that include substitutions, aligned with the respective exon lengths.

(b), Analysis of the relationship between exon length and the accumulation of false-positive SNV reads. Each data point represents an exon from the EGFR or RPL15 genes, annotated with its respective exon number. The dataset includes eight exons, and the Pearson correlation coefficient (r) is calculated to quantify the strength of the correlation.

3.3. Benchmarking SynPrime-based functional evaluation of SNVs in an essential gene

To evaluate the effectiveness of our SynPrime approach in conducting near-saturation functional analyses of SNVs, we developed a pegRNA library designated as Syn-RPL15. This library includes 1,492 pegRNA sequences engineered to generate 500 SNVs, which account for 97% of the 516 possible SNVs in the 172-bp long exon 2 of the essential gene RPL15, each accompanied by a synonymous substitution (Methods Section). MLH1-knockout PC-9 cells expressing PEmax were then transduced with the Syn-RPL15 library. Following transduction, puromycin selection was applied to remove untransduced cells, completing this process within five days (**Figure 5a**).

Deep sequencing was performed on these cells five days post-transduction (Day 0), revealing that the median frequency of double-hit SNVs (sequences with both the intended substitution and the synonymous marker) was 0.0083%, with a variation ranging from 0.00% to 0.28%. This frequency represented a significant 464-fold increase compared to the median frequency of 1.8 x 10^-5% observed in unedited cells (**Figure 5b**). Analysis of the total read counts revealed that double-hit SNVs accounted for 10% of the reads in prime-edited cells versus only 0.02% in unedited control cells (**Figure 5c**). This indicates that the SynPrime method enhanced the detection of intended SNVs by a factor of 500 (10%/0.02%). This dual-marker approach enabled the identification of 440 out of the 500 designed SNVs, equivalent to 88% coverage of the intended targets and 85% coverage of all potential SNVs in exon 2 of RPL15 (**Figure 5 d-f**).

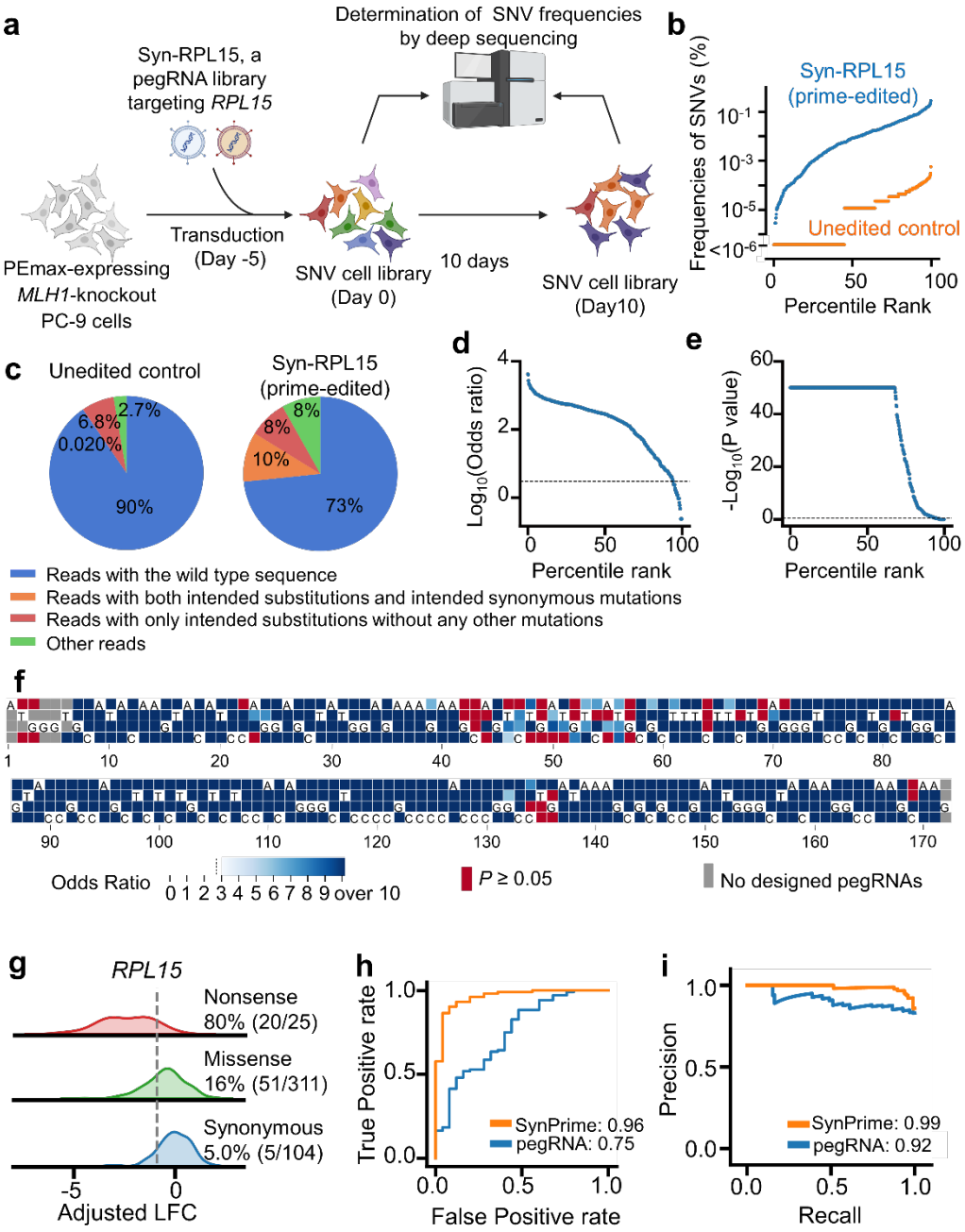


Figure 5. SynPrime evaluation of SNVs in *RPL15*. (a), Schematic depiction of the high-throughput SynPrime method for evaluating variants in RPL15. **(b)**, Observation of the frequencies

of 500 SNVs in unedited control cells compared to prime-edited cells post-treatment with syn-RPL15. **(c)**, Comparison of sequencing reads containing both intended substitutions and synonymous mutations versus reads containing only intended substitutions without additional mutations in unedited PEmax-expressing PC-9 cells (left, unedited control) and in cells ten days post-transduction with Syn-RPL15 (right). **(d-e)**, Display of odds ratios (d) and P-values (e) for 500 SNVs in PEmax-expressing cells ten days post-transduction with Syn-RPL15. Dashed horizontal lines denote thresholds where odds ratio equals 3 (d) and P-value equals 0.05 (e). **(f)**, Heatmap illustrating odds ratios and/or P-values of 516 SNVs generated by prime editing in exon 2 of RPL15 ten days post-transduction with Syn-RPL15. SNVs with P-values greater than 0.05 by a two-sided Fisher's exact test are depicted in red, with odds ratios not shown. SNVs with odds ratios lower than 3 are represented in a white background. SNVs without designed pegRNAs are displayed as gray boxes. The bottom axis indicates the nucleotide positions in the RPL15 coding sequence, with the reference sequence nucleotide provided at each position. **(g)**, Kernel density estimation plots of adjusted Log Fold Changes (LFCs) of SNVs in RPL15 categorized by SNV type. For each category, the number and percentage of SNVs with adjusted LFC values below a specified cutoff (represented by the gray dashed vertical line), equivalent to the 5th percentile of adjusted LFC values of synonymous mutations, are shown. **(h-i),** Receiver operating characteristic (ROC) (h) and precision-recall (i) curves for adjusted LFCs of SNVs determined by SynPrime (orange) and analysis based on pegRNA abundance (blue) for sets of nonsense (n = 25) versus synonymous SNVs (n = 104) in exon 2 of RPL15. Area under curve values are indicated.

The prime-edited cells underwent a further 10-day culture period, during which log2-fold changes (LFCs) in SNV frequencies at day 10 were calculated relative to those at day 0 (i.e., 5 days post-transduction of Syn-RPL15), providing insights into the degree of RPL15 function impairment associated with each SNV (**Figure 5a**). To address positional biases in LFCs of synonymous SNVs, LOWESS (Locally Weighted Scatterplot Smoothing)[21] regression was employed for normalization, resulting in an increase in the area under the ROC (AUROC) from 0.92 to 0.96 (**Figure 6a, b, Methods**). Correlations between replicate adjusted LFC values were robust (**Figure 6c**). As anticipated, SNVs leading to nonsense mutations exhibited significant depletion compared to those inducing synonymous mutations (P = 1.5 X10-23, Student's t-test, **Figure 5g**). Receiver-Operator Characteristics (ROC) and Precision-Recall curve (PRC) analyses were conducted, assuming that nonsense SNVs abolish RPL15 function (resulting in cell depletion), while synonymous mutations preserve RPL15 function. The prime editing-based high-throughput evaluations effectively discriminated between loss-of-function and intact RPL15 function, yielding an AUROC of 0.96 and an area under the PRC curve (AUPRC) of 0.99 (**Figure 5h-i**), indicative of the high accuracy of our SynPrime approach.

Seventy-one SNVs (16% of 440), corresponding to 62 protein variants (20% of 307), were classified as depleting (i.e., causing RPL15 function loss, with LFC values < the 5th percentile of synonymous mutation LFC values), while 189 SNVs (43%), encoding 178 protein variants (58%), were classified as non-depleting (i.e., associated with intact RPL15 function, with LFCs > the 20th percentile of synonymous mutation LFCs). The remaining 180 SNVs (41%), encoding 67 protein variants (22%), were categorized as intermediate (with LFCs falling between the 5th and 20th percentiles of synonymous mutation LFCs) (**Figure 6 d-e**). In summary, utilizing the SynPrime approach, we functionally categorized 440 SNVs (85% of all possible SNVs), encompassing 307 protein variants, constituting 85% of all potential protein variants encoded in exon 2 of RPL15.

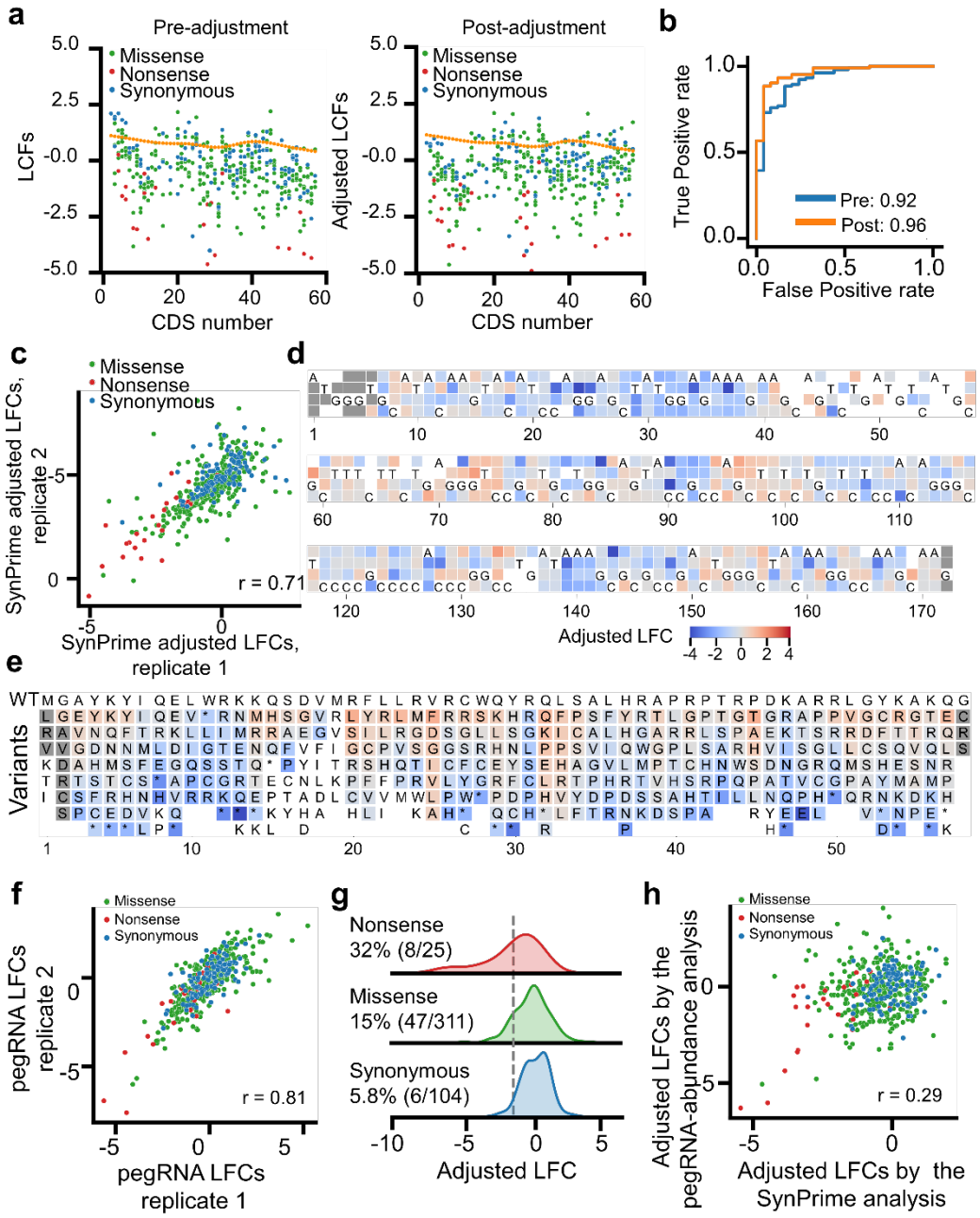


Figure 6. Optimization of SynPrime and comparison with pegRNA-based analysis.

(a), To mitigate the effects of positional biases in Log2 Fold Changes (LFCs), we employed Locally

Weighted Scatterplot Smoothing (LOWESS) regression, leveraging synonymous Single Nucleotide Variants (SNVs) presumed to be functionally inert. The LOWESS regression curves, depicted in orange, reveal a pronounced alleviation of positional biases, particularly evident in the overall depletion of nonsense SNVs post-adjustment. **(b)**, Utilizing Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) analysis, we evaluated the impact of positional bias correction on sets of nonsense (n = 25) versus synonymous SNVs (n = 104) within exon 2 of RPL15. Pre-adjustment and post-adjustment ROC-AUC values are compared, highlighting the efficacy of positional bias correction. **(c)**, We assessed the consistency of SynPrime Log2 Fold Change (LFC) values across two biological replicates, calculating Pearson correlation coefficients (r). This analysis, encompassing 440 SNVs, underscores the robust reproducibility of SynPrime measurements. **(d-e)**, Heatmaps portray the adjusted LFCs of 516 SNVs (d) and 336 protein variants (e) induced by prime editing within exon 2 of RPL15. SNVs and protein variants failing to meet significance criteria ($P > 0.05$ or odds ratios < 3) are excluded, denoted by white backgrounds. Additionally, SNVs lacking corresponding pegRNAs are represented by gray boxes. Annotation at the bottom of each heatmap denotes the position within the RPL15 coding sequence (d) and amino acid sequence (e), with reference sequence nucleotides (d) and amino acids (e; WT, wild-type) provided. **(f)**, The correlation between adjusted LFC values of SNVs determined via pegRNA abundance-based analysis across two biological replicates is assessed, yielding Pearson correlation coefficients (r) for 440 SNVs. **(g)**, Kernel density estimation plots illustrate the distribution of adjusted LFCs for SNVs in RPL15, categorized based on pegRNA abundance-based analysis. For each category, the proportion of SNVs below a defined cutoff value, corresponding to the 5th percentile of synonymous mutation LFCs, is presented. **(h)**, We explore the correlation between adjusted LFC values derived from SynPrime evaluations and those obtained via pegRNA abundance-based analysis, evaluating 440 SNVs and calculating Pearson correlation coefficients (r) to delineate the concordance between methodologies.

3.4. Accuracy comparison of SynPrime and pegRNA abundance analyses in RPL15

In lieu of evaluating endogenous sites, an alternative approach involves computing Log2 Fold Changes (LFCs) of lentivirally integrated pegRNA sequences, a method previously employed for high-throughput functional assessments[24,25,50-55]. The LFCs for each Single Nucleotide Variant (SNV) were computed using data obtained from three corresponding pegRNAs, employing the MAGeCK algorithm[24,43]. A strong correlation was observed between LFCs of SNV frequencies calculated from pegRNA abundance data across two replicates (Pearson r = 0.81, **Figure 6f**). As expected, LFC values of pegRNAs for nonsense SNVs were significantly lower than those for synonymous SNVs (P = 2.2 X 10^{-7}, Student's t-test, **Figure 6g**). However, the reduction in LFC, as evaluated by pegRNA abundance-based analysis, was less pronounced compared to the SynPrime approach based on direct sequencing of endogenous sites (**Figure 5f and Figure 6g**); Only 32% of nonsense SNVs exhibited a slight depletion in the pegRNA abundance-based analysis, while 80% were strongly depleted by the SynPrime analysis. This disparity suggests that the SynPrime approach may offer higher accuracy and lower susceptibility to errors compared to pegRNA abundance-based analysis.

Additionally, we conducted Receiver Operating Characteristic (ROC) and Precision-Recall Curve (PRC) analyses, assuming that pegRNAs inducing nonsense substitutions would lead to the loss of RPL15 function, while negative control pegRNAs inducing synonymous prime editing would not affect RPL15 function. AUROC and AUPRC values were found to be 0.75 and 0.92, (**Figure 5 h-i**) respectively, which were lower than the corresponding values obtained from the SynPrime approach. We observed only a modest correlation between LFCs determined by endogenous site sequencing (SynPrime) and those determined by pegRNA abundance sequencing (r = 0.29, **Figure 6h**). These findings collectively suggest that the SynPrime approach offers superior accuracy compared to pegRNA abundance-based analyses, indicating that direct sequencing of endogenous sites may yield more precise functional evaluations of variants.

3.5. Saturating SNV generation in exons encoding the EGFR tyrosine kinase domain

To comprehensively investigate the ramifications of saturating point mutations within the EGFR tyrosine kinase domain on drug resistance, we employed the SynPrime methodology. This domain, spanning 290 amino acids, necessitating 870 coding sequences, is encoded by seven exons (exons 18-24) (**Figure 7a**). Using DeepPrime-FT[31], we meticulously designed 7,488 pegRNAs, allowing for the induction of 96% of all possible SNVs (corresponding to 97% of all possible protein variants) within the tyrosine kinase domain. Each substitution was targeted by 2-3 pegRNAs, resulting in a vast library poised to introduce comprehensive mutational coverage. Specifically, we generated one pegRNA library per exon, denoted as Syn-exon18 through to Syn-exon24. Additionally, to serve as controls, we included negative control pegRNAs (sham-editing or nontargeting), constituting approximately 5% of the pegRNAs in each library (**Methods**).

Subsequent to library generation, SGE was initiated by independently transducing each library into PEmax-expressing MLH1 knockout PC-9 cells, followed by a 10-day cultivation period to facilitate sufficient prime editing (**Figure 7b**). Deep sequencing analysis revealed that the sums of the two-hit SNV reads (reads with the intended substitution and an additional intended synonymous mutation) in both the unedited control and prime-edited cell populations were a median of 0.009% and 19%, respectively, of the total reads (**Figure 8**), indicating that the frequency of identified SNVs was a median of 2,111-fold (= 19%/0.0009%) higher than the frequency of false positive PCR and sequencing errors. Upon exclusion of 39 non-significant SNVs (odds ratio $\leq$ 3 or P-value $\geq$ 0.05 in any replicate), we identified 2,476 significant SNVs (95% of all 2,610 possible SNVs), corresponding to 1,726 protein variants (95% of all 1,817 possible protein variants) (**Figure 7c-e and Figure 10**).

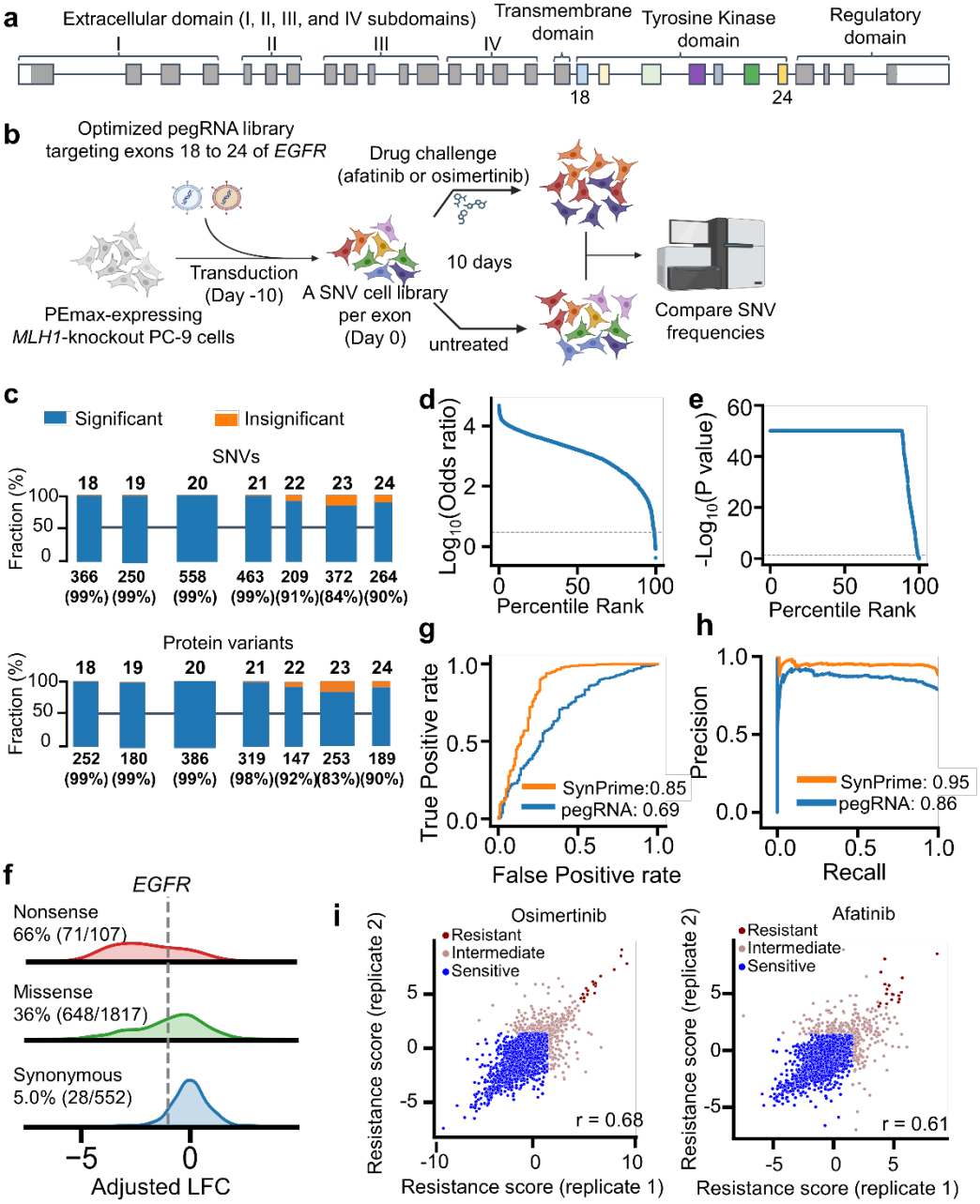


Figure 7. Saturating SNV generation in the region encoding the EGFR tyrosine kinase domain. (a), Schematic representation of *EGFR*. The tyrosine kinase domain is encoded by exons 18 – 24 within the EGFR gene. The extracellular domain comprises subdomains I, II, III, and IV.

(b), A schematic depiction of the functional assay utilizing pooled pegRNA libraries to induce SNVs in EGFR. Each exon of EGFR was targeted by a distinct pegRNA library, resulting in the creation of seven libraries corresponding to the seven exons of EGFR. **(c),** The number and percentage of significantly generated SNVs and protein variants in each EGFR exon are depicted. Significantly generated SNVs were defined as those with frequencies > 0.0005%, odds ratios > 3, and P-values < 0.05 at day 0. (**d-e)**, Illustration of odds ratios (d) and P-values (e) of 2,515 SNVs in PEmax-expressing cells ten days post-transduction with Syn-exon18, Syn-exon19, …, and Syn-exon24 pegRNA libraries. Dashed horizontal lines indicate thresholds for odds ratio = 3 (d) and P-value = 0.05 (e). **(f)**, Kernel density estimation plots of adjusted Log2 Fold Changes (LFCs) of SNVs in the region encoding the EGFR tyrosine kinase domain, categorized by SNV type. The percentage of SNVs with adjusted LFC values lower than the 5th percentile of adjusted LFC values of synonymous mutations is displayed for each category. **(g-h),** Receiver Operating Characteristic (ROC) (g) and Precision-Recall (h) curves for adjusted LFCs of SNVs determined by SynPrime (orange) and analysis based on pegRNA abundance (blue) for sets of nonsense and synonymous SNVs in exons 18-24 of EGFR. Area under curve values are provided. **(i)**, Correlation between SynPrime LFC values following treatment with afatinib (left) and osimertinib (right) in two biological replicates. SNV classification is indicated by dot color. Pearson correlation coefficients are shown.

3.6. Accuracy of SynPrime vs. pegRNA abundance analyses in *EGFR*

Upon evaluating the SNV libraries at day 10 (20 days post-transduction of pegRNA libraries), we observed a significant depletion of SNVs inducing nonsense mutations compared to those eliciting synonymous mutations ($P = 1.8 \text{ X } 10^{-67}$, Student's t-test, **Figure 7f**), indicating the reliance of PC-9 cells on EGFR signaling. However, in pegRNA abundance-based analysis, the depletion of nonsense SNVs was less pronounced than that observed with SynPrime (**Figure 9a-b**). SynPrime exhibited superior performance, as evidenced by AUROC and AUPRC values of 0.85 and 0.95, respectively, while pegRNA abundance-based analyses yielded lower values of 0.69 and 0.86, respectively. Moreover, there was only a modest correlation between Log2 Fold Changes (LFCs) determined by endogenous site sequencing (SynPrime) and pegRNA abundance sequencing (r = 0.40), consistent with observations in the RPL15 analysis (**Figure 9c**). These findings collectively affirm the greater accuracy of the SynPrime approach compared to pegRNA abundance-based analyses in elucidating EGFR dynamics.

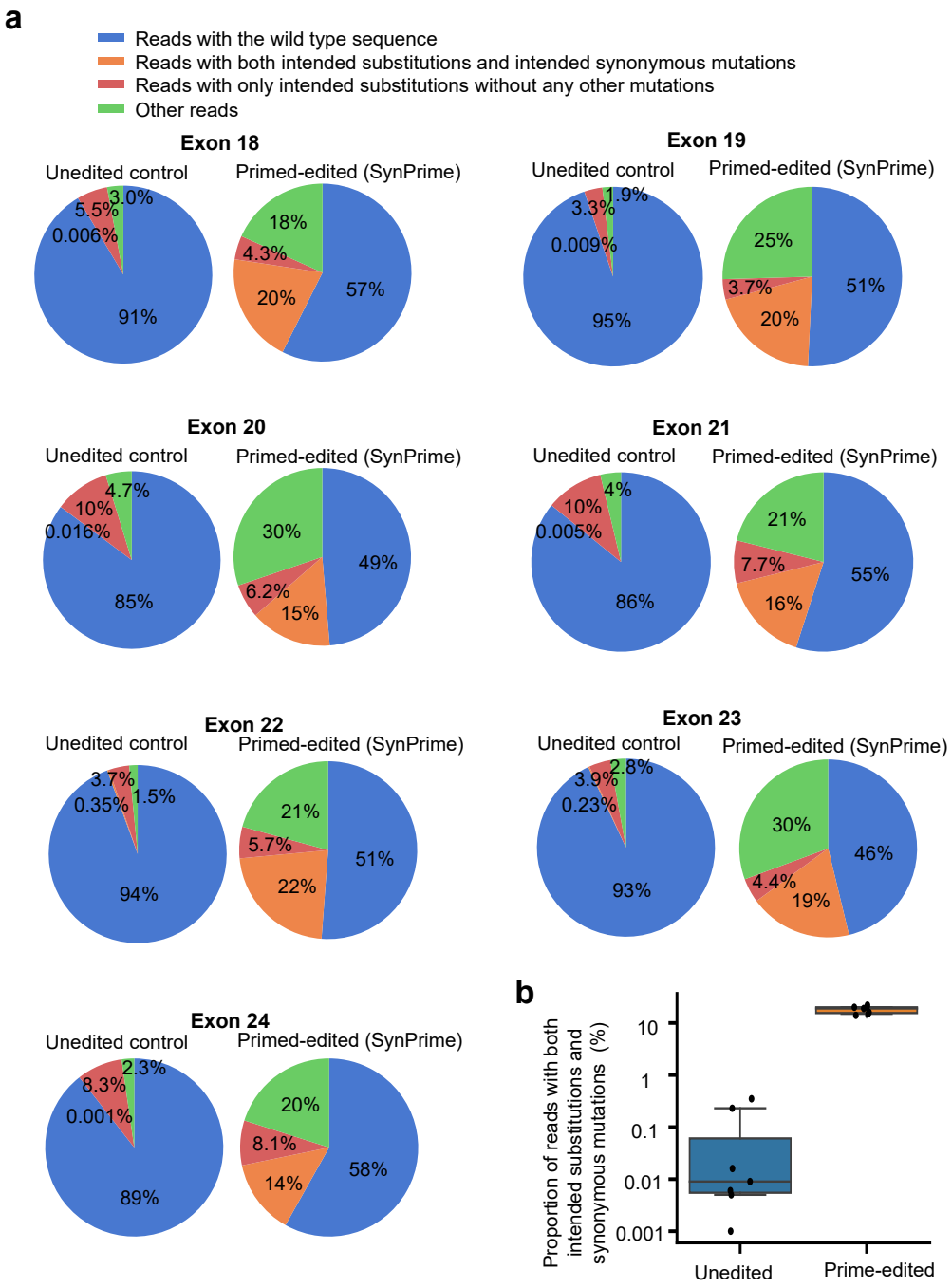


Figure 8. Proportions of sequencing reads.

(a), This figure illustrates the proportions of sequencing reads containing both intended substitutions and intended synonymous mutations, as well as those containing only intended substitutions without any other mutations. The data are presented for unedited PEmax-expressing PC-9 cells (left, unedited control) and for cells ten days after transduction with the indicated pegRNA library (right, Syn-exon18, Syn-exon19, …, Syn-exon24) corresponding to the respective EGFR exon. **(b)**, boxes represent the 25th, 50th, and 75th percentiles of the data distribution, while whiskers depict the 10th and 90th percentiles. The analysis encompasses seven exons (n = 7), providing a comprehensive view of the distribution of sequencing reads across the EGFR exonic regions before and after pegRNA library transduction.

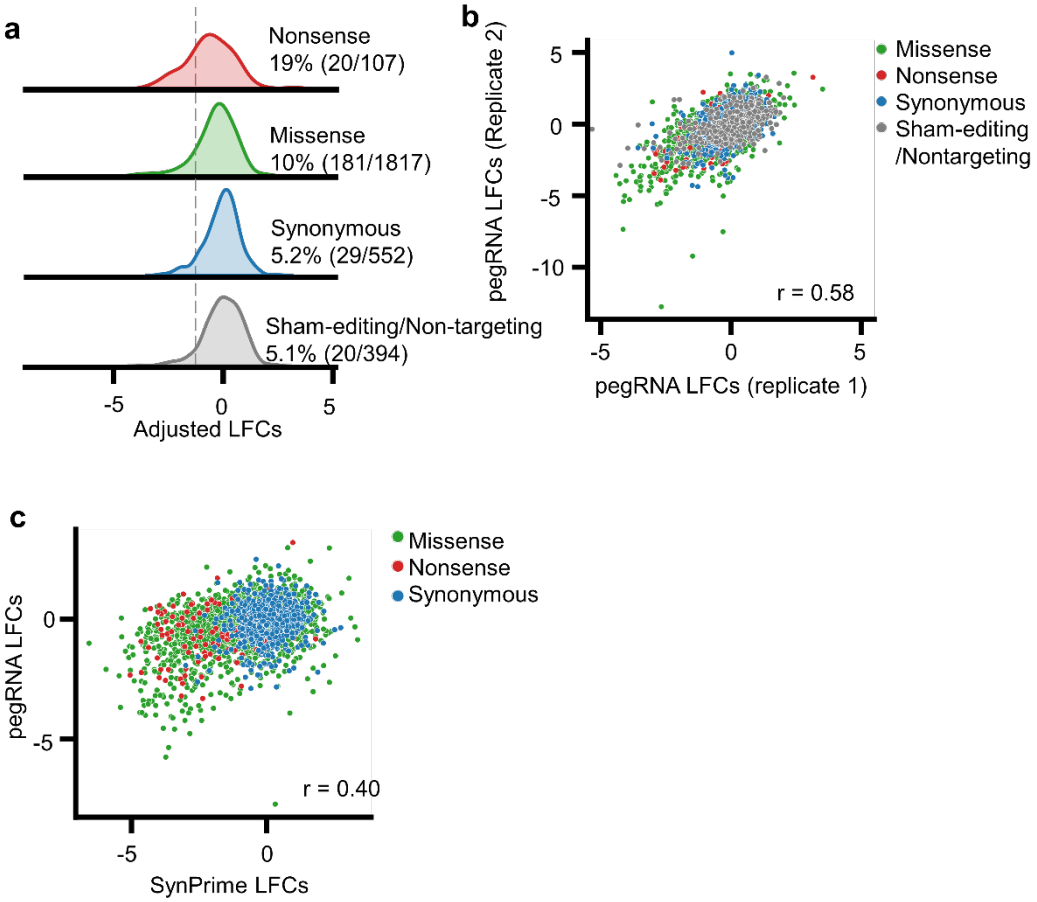


Figure 9. Depletion of nonsense *EGFR* SNVs assessed by PegRNA abundance-based analyses. **(a)**, These plots depict the distribution of adjusted Log2 Fold Changes (LFCs) of Single Nucleotide Variants (SNVs) in EGFR, as determined by pegRNA abundance-based analysis, categorized by SNV type. For each category, the number and percentage of SNVs with adjusted LFC values lower than a defined cutoff value (the gray dashed vertical line), representing the 5th percentile of adjusted LFC values of sham-editing/non-targeting pegRNAs, are presented. **(b)**, The correlation between adjusted LFC values of SNVs determined by pegRNA abundance-based analysis in two biological replicates is depicted. The Pearson correlation coefficient (r) is provided. The analysis encompasses missense, nonsense, and synonymous SNVs, with corresponding counts of 1,817, 107, and 552, respectively. Additionally, the number of sham-editing and nontargeting pegRNAs is specified as 394. **(c)**, This panel illustrates the correlation between adjusted LFC values of SNVs calculated from

SynPrime evaluations and those derived from pegRNA abundance-based analysis. The Pearson correlation coefficient (r) is presented. The analysis includes a total of 2,476 SNVs, with counts of missense, nonsense, and synonymous SNVs as 1,817, 107, and 552, respectively.

Figure 10. Identified SNVs in EGFR TK domain. This heatmap displays odds ratios and/or P-values of 2,610 SNVs generated by prime editing in exons 18-24 of EGFR, observed ten days post-transduction of the Syn-exon18, Syn-exon19, …, Syn-exon24 libraries. SNVs with P-values greater

than 0.05 by the two-sided Fisher's exact test are highlighted in red; in such cases, odds ratios are not depicted. SNVs with odds ratios lower than 3 are represented with a white background. The numbers at the bottom of each heatmap denote the position in the EGFR coding sequence, with the nucleotide in the reference sequence indicated. Edited reads were identified based on the presence of both the intended edit and an additional synonymous edit. This heatmap provides a comprehensive overview of the distribution of identified SNVs within the EGFR tyrosine kinase domain, highlighting significant alterations induced by prime editing across the targeted exonic regions.

3.7. Complete resistance profiles of 2,476 *EGFR* SNVs against afatinib and osimertinib

We extended the culture of the SNV libraries for an additional 10 days under three distinct experimental conditions: a control condition without any treatment, treatment with the 2nd generation TKI afatinib, and treatment with the 3rd generation TKI osimertinib (**Figure 7b**). LFCs of SNV frequencies in the TKI-treated arms were compared with those in the untreated arm to ascertain their impact on resistance to these two TKIs (**Methods**). These normalized LFCs were referred to as resistance scores. Notably, the resistance scores of SNVs in both the osimertinib and afatinib arms from two different replicates exhibited a strong correlation (**Figure 7i**). Based on these resistance scores, SNVs were categorized into "resistant" (exceeding the resistance score of synonymous SNVs in the 99.7th percentile in both replicates), "sensitive" (below the resistance score of synonymous SNVs in the 95th percentile in both replicates), and "intermediate" (remaining SNVs) classifications.

Out of the 1,726 protein variants, 158 were encoded by two SNVs and 20 by three SNVs. Correlations between the resistance scores of the two members of each pair were notably high (**Figure 11a, left and middle**). By averaging SNV resistance scores, the resistance score of each protein variant was determined, showing strong correlations across replicates (**Figure 11b, left and middle**). In the case of afatinib, 2,138 SNVs (encoding 1,476 protein variants) were classified as sensitive, 320 as intermediate, and 18 as resistant (**Figure 12, 14**). Notably, 88% of the 16 resistant protein variants were previously unreported (**Figure 11c**). Similarly, for osimertinib, 2,102 SNVs (encoding 1,448 protein variants) were sensitive, 357 were intermediate, and 17 were resistant (**Figure 13, 15**). Of these, 80% of the 15 resistant protein variants were previously unreported (**Figure 11c**).

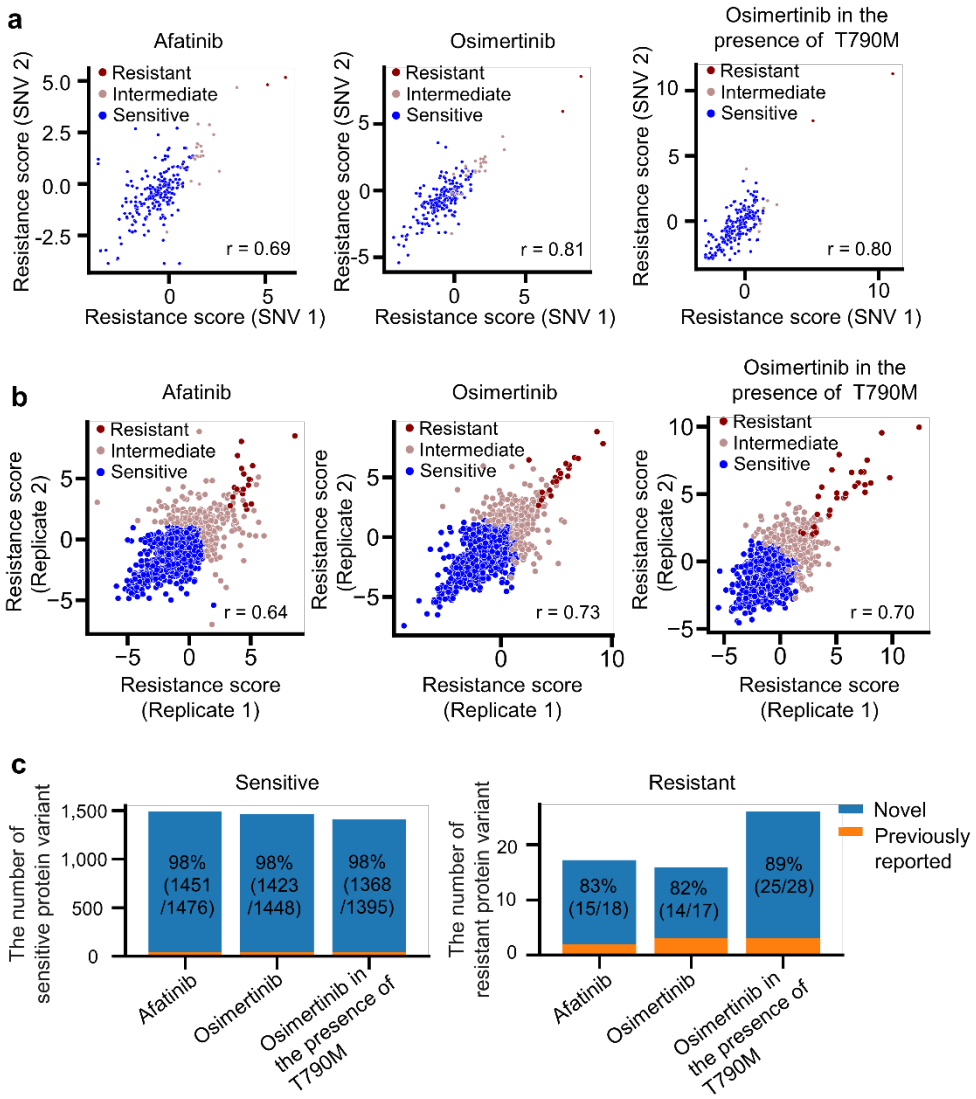


Figure 11. Assessment of resistance profiles of EGFR protein variants. (a), Correlation between resistance scores post-treatment with afatinib (left), osimertinib without T790M mutation (middle), and osimertinib with T790M mutation (right) in pairs of SNVs encoding identical protein variants. Each protein variant's classification is indicated by dot color. Pearson correlation coefficients (r) are displayed. Number of SNV pairs: n = 218 (left), 218 (middle), and 210 (right). **(b)**, Correlation between resistance scores of protein variants post-treatment with afatinib (left), osimertinib without

T790M mutation (middle), and osimertinib with T790M mutation (right) across two biological replicates. Each protein variant's classification is indicated by dot color. Pearson correlation coefficients are presented. Number of protein variants: n = 1,726 (left), 1,726 (middle), and 1,671 (right). **(c)**, Enumeration of sensitive and resistant protein variants functionally categorized in the present study. The proportion of protein variants lacking previously reported data on their drug resistance effect, among all sensitive or resistant variants, is depicted on the blue bars.

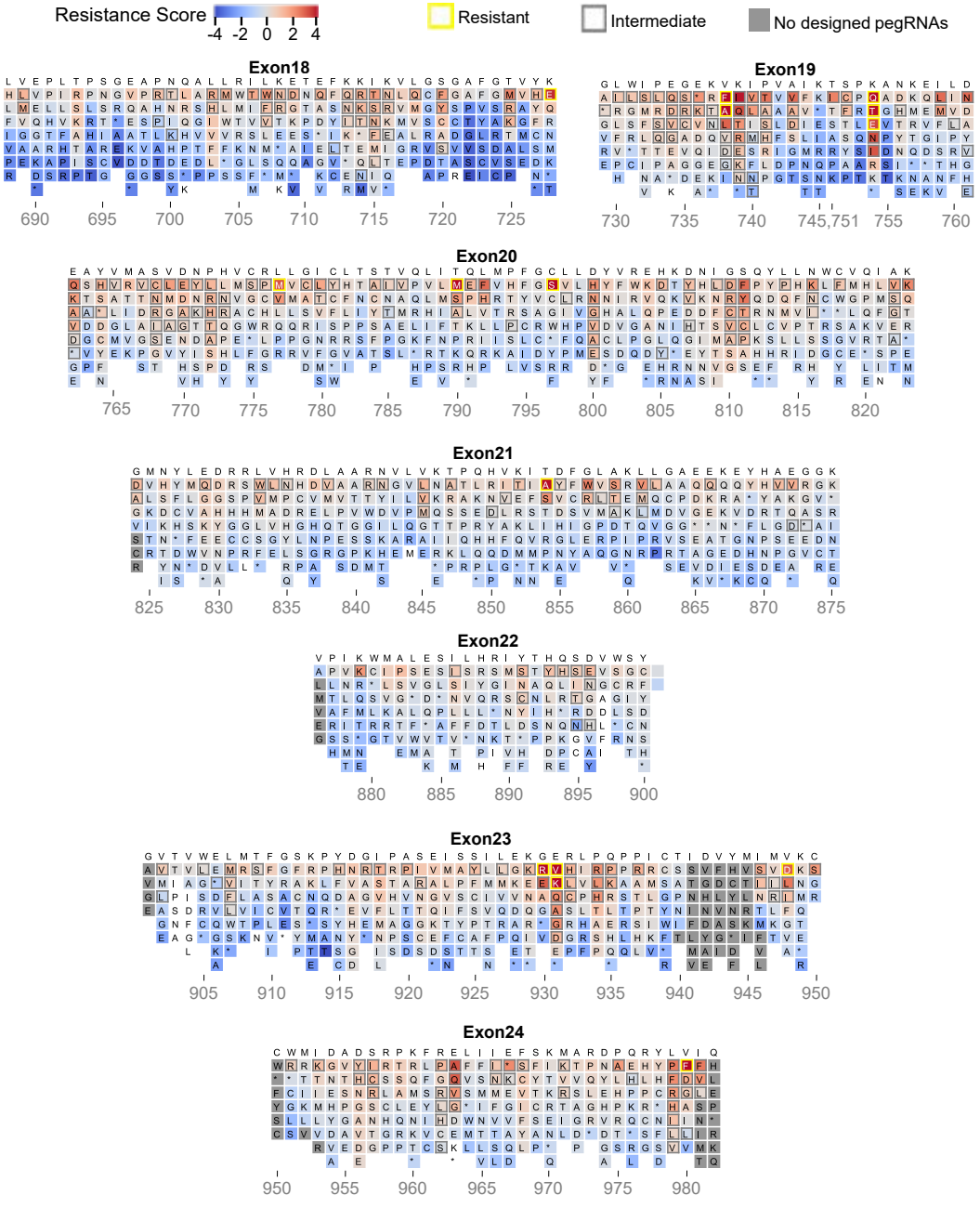


Figure 12. Heatmap illustrating afatinib resistance scores of 1,817 protein variants.

These variants were generated via prime editing in exons 18-24 of EGFR in PC-9 cells. Yellow-

outlined and gray-outlined boxes denote protein variants inducing resistant and intermediate phenotypes, respectively. Numbers at the bottom represent positions in the EGFR amino acid sequence, with the reference sequence amino acids displayed at the top. Excluded from the analysis are 30 protein variants with P-values > 0.05 or odds ratios < 3 (depicted with white background) and protein variants without designed pegRNAs (gray boxes).

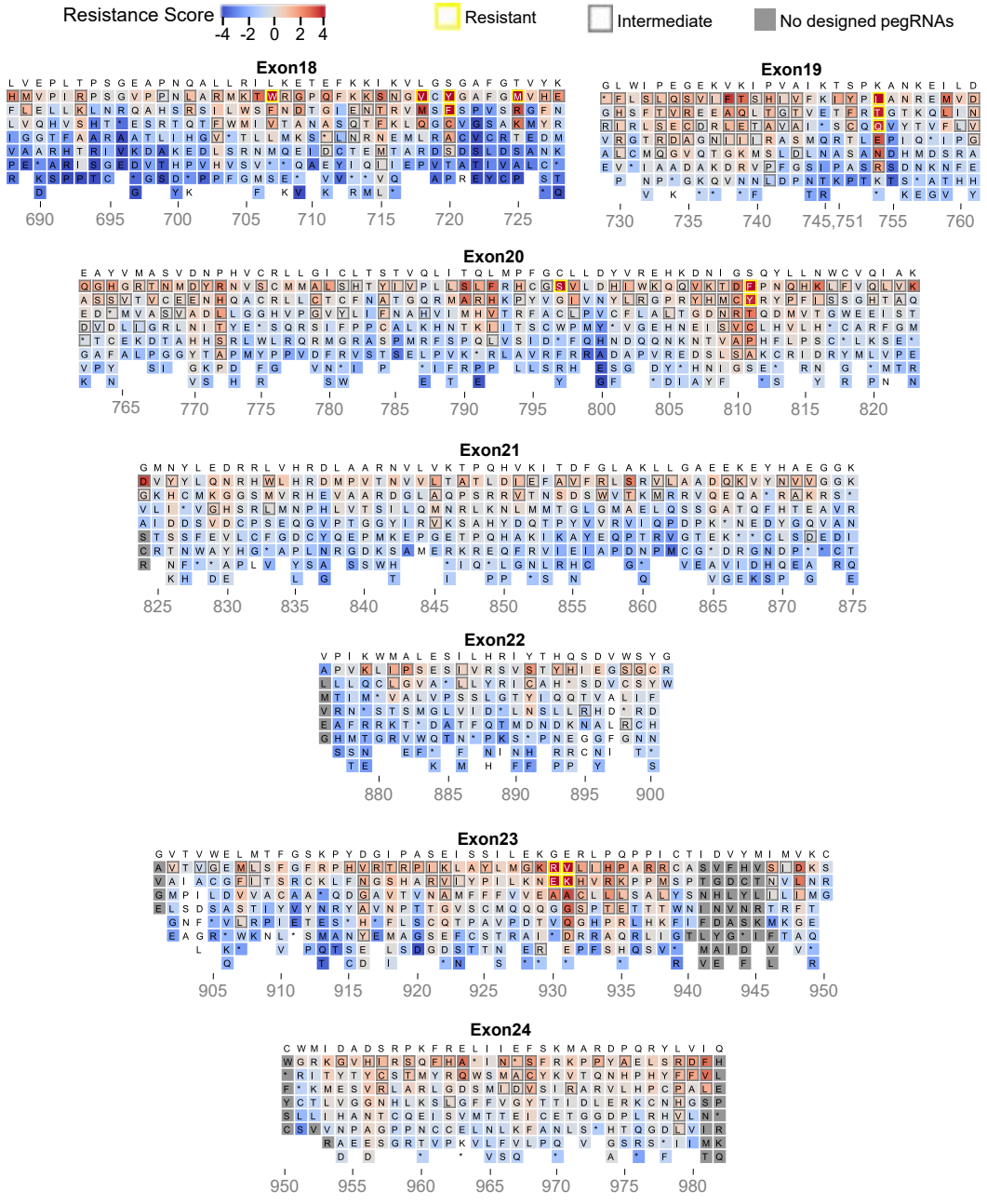


Figure 13. Heatmap illustrating osimertinib resistance scores of 1,817 protein variants.

These variants were generated via prime editing in exons 18-24 of EGFR in PC-9 cells. Yellow-

outlined and gray-outlined boxes represent protein variants inducing resistant and intermediate phenotypes, respectively. Numbers at the bottom represent positions in the EGFR amino acid sequence, with the reference sequence amino acids displayed at the top. Excluded from the analysis are 30 protein variants with P-values > 0.05 or odds ratios < 3 (depicted with white background) and protein variants without designed pegRNAs (gray boxes).

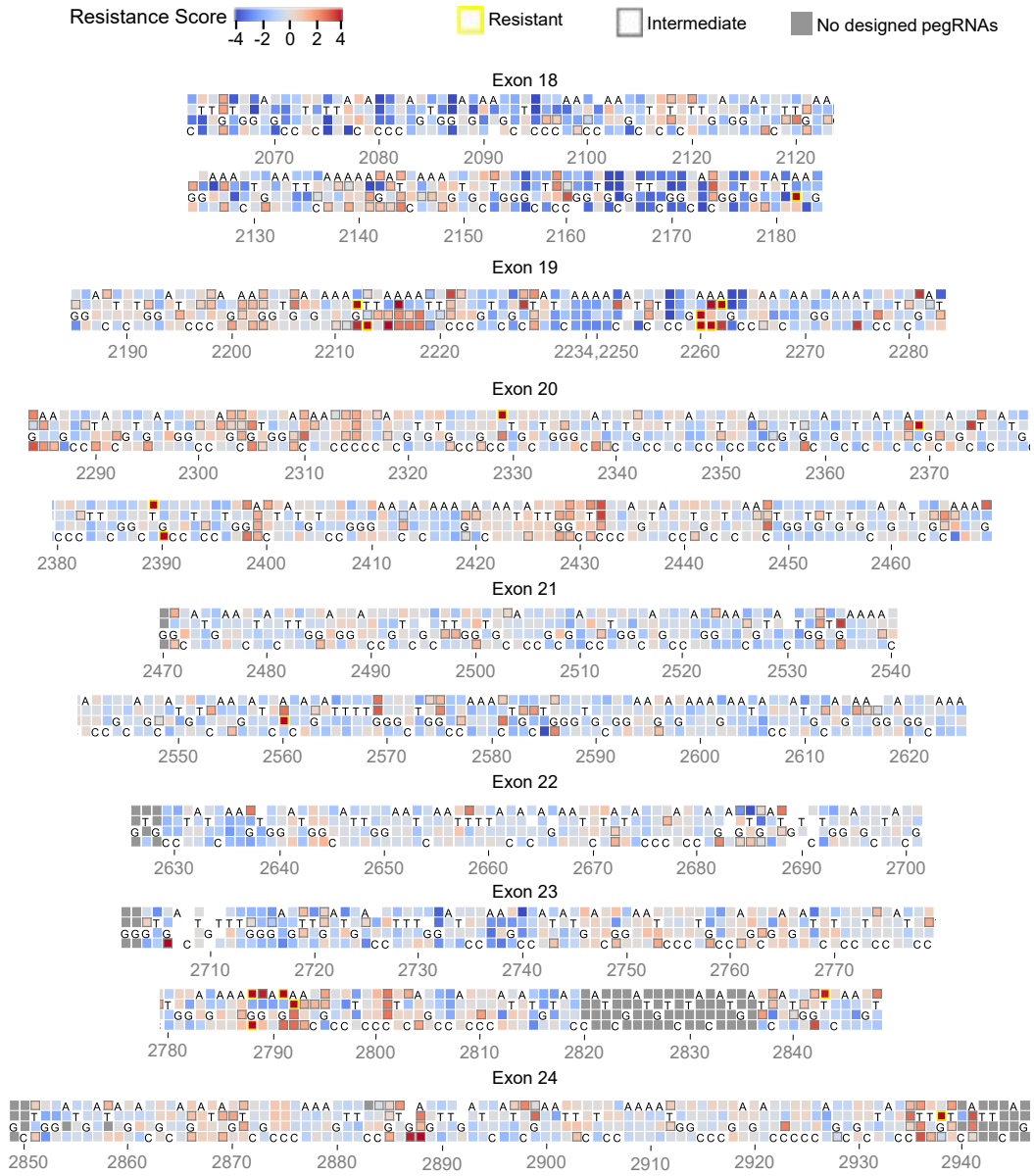


Figure 14. Heatmap illustrating afatinib resistance scores of 2,610 SNVs. These variants were generated via prime editing in exons 18-24 of EGFR in PC-9 cells. Yellow-outlined and gray-outlined boxes represent SNVs inducing resistant and intermediate phenotypes, respectively. Numbers at the bottom represent positions in the EGFR coding sequence, with nucleotides from

the reference sequence displayed. Excluded from the analysis are 40 SNVs with P-values > 0.05 or odds ratios < 3 (depicted with white background) and SNVs without designed pegRNAs (gray boxes).

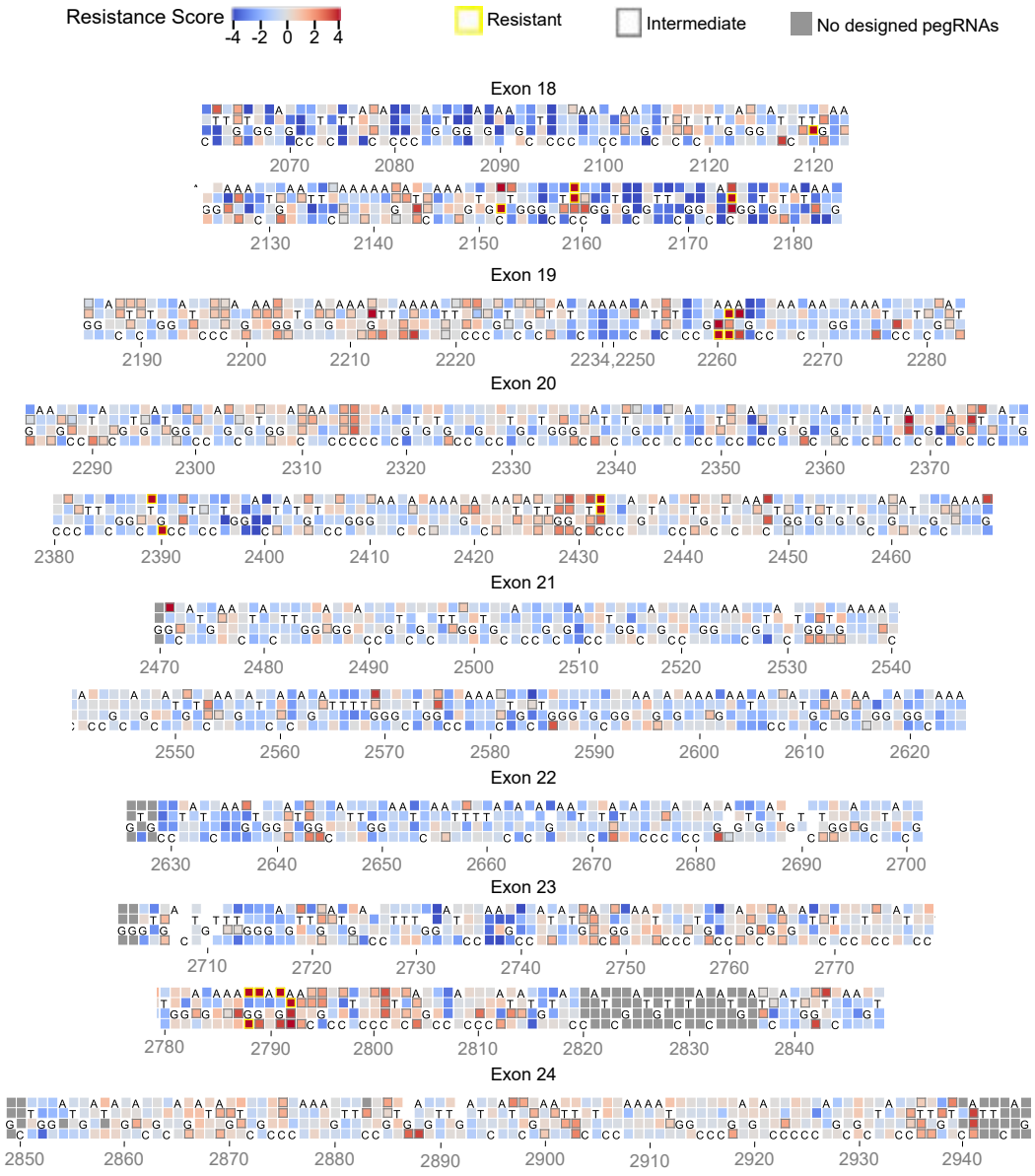


Figure 15. Heatmap illustrating osimertinib resistance scores of 2,610 SNVs.

These variants were generated via prime editing in exons 18-24 of EGFR in PC-9 cells. Yellow-outlined and gray-outlined boxes represent SNVs inducing resistant and intermediate phenotypes, respectively. Numbers at the bottom represent positions in the EGFR coding sequence, with

nucleotides from the reference sequence displayed. Excluded from the analysis are 40 SNVs with P-values > 0.05 or odds ratios < 3 (depicted with white background) and SNVs without designed pegRNAs (gray boxes).

3.8. Resistance profiles of 2,391 *EGFR* SNVs against osimertinib in PC-9-T790M cells

T790M, a prevalent gatekeeper mutation, confers resistance to 1st and 2nd generation TKIs[10,56]. Considering osimertinib's efficacy as a frontline treatment for T790M-positive advanced NSCLC patients[57], we investigated the resistance profiles of SNVs against osimertinib in PC-9 cells harboring T790M. Through cytosine base editing, we introduced the T790M (c.2369C>T) mutation into the PEmax-expressing PC-9 cell line, generating PC-9-T790M (**Figure 16, Methods**). Sanger sequencing and deep sequencing confirmed the homozygous nature of the T790M mutation, with a frequency of alleles containing T790M at 99.8%. SynPrime evaluation was conducted in these PC-9-T790M cells using osimertinib, considering the resistance conferred by the T790M mutation to afatinib. Resistance scores of SNVs and protein variants in the PC-9-T790M cells exhibited high correlation between two replicates (**Figure 16e**). Strong correlations were also observed between resistance scores of SNVs encoding identical protein variants (**Figure 11a, right**), in both afatinib and osimertinib arms. The resistance scores of protein variants from two different replicates displayed good correlation as well (**Figure 11b, right**). We classified 2,028, 335, and 28 SNVs, encoding 1,395, 281, and 26 protein variants, as sensitive, intermediate, and resistant, respectively (**Figure 17-18**). Notably, 18 (69%) of the 26 protein variants causing resistance have not been previously reported (**Figure 11c**).

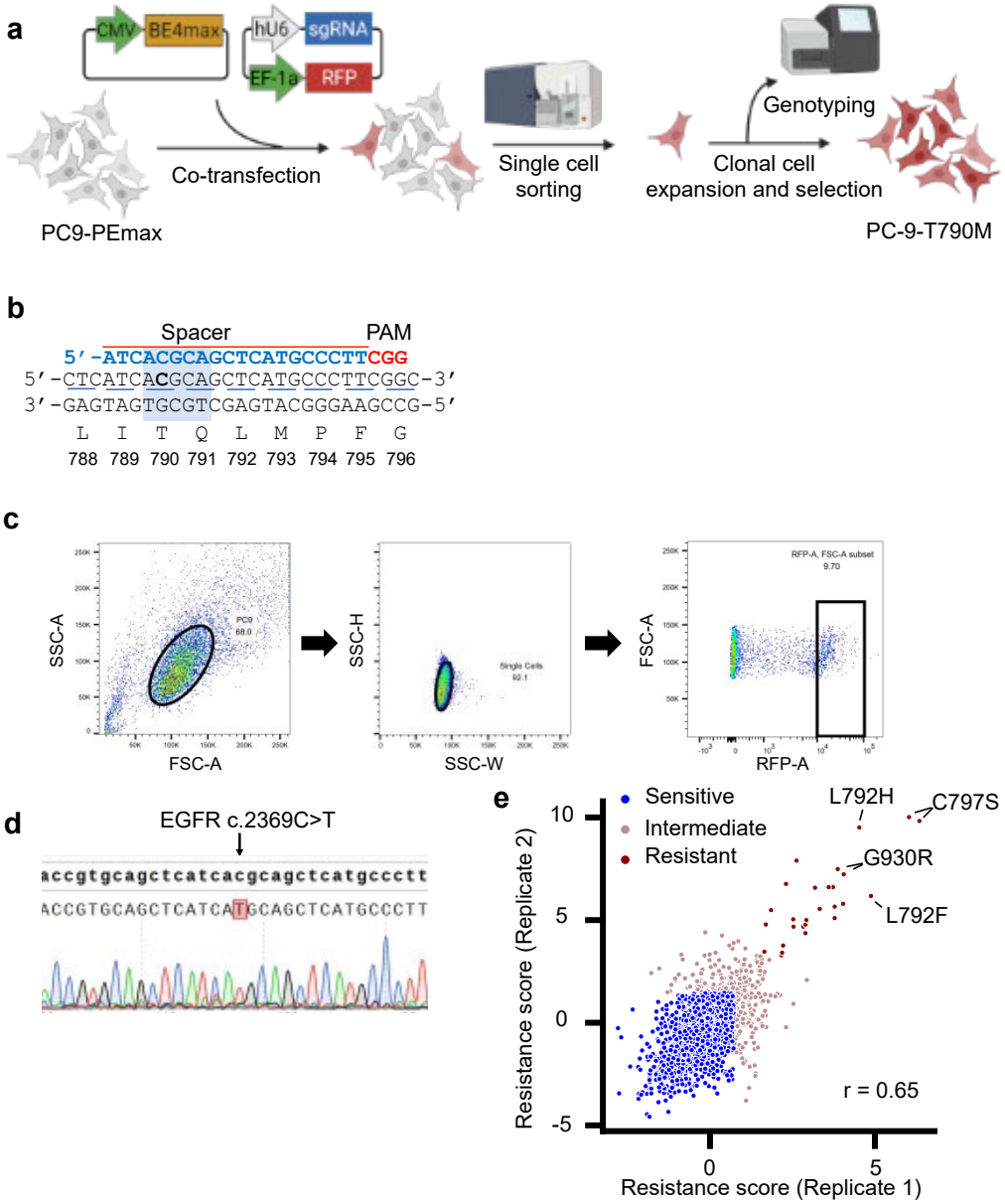


Figure 16. Generation of PC-9 cells harboring the T790M mutation. (a), Illustration depicting the methodology employed to generate PC-9 cells with the T790M mutation. PEmax-expressing MLH1-knockout PC-9 cells underwent co-transfection with plasmid vectors expressing BE4max, the requisite sgRNA, and red fluorescent protein (RFP). Subsequently, single cells were isolated via

flow cytometry and underwent expansion. Subsets of single cell-derived clones were lysed and underwent deep sequencing to ascertain the outcomes of base editing. **(b)**, Schematic representation of the target sequence for base editing in the EGFR gene, particularly highlighting the c.2369C>T alteration (p.T790M, depicted in bold). The shaded region delineates the expected cytosine base editing window. **(c)**, Illustration demonstrating the flow cytometry gating strategy utilized for sorting the transfected cells. **(d)**, Display of Sanger sequencing outcomes from the chosen base-edited clone, confirming the occurrence of c.2369C>T editing. **(e)**, Correlation analysis exhibiting the relationship between SynPrime LFC (Log2 Fold Change) values subsequent to osimertinib treatment in PC-9 cells bearing the T790M mutation across two independent biological replicates. Each single nucleotide variant (SNV) is classified by dot color. The Pearson correlation coefficient is indicated. The total number of SNVs examined is n = 2,391.

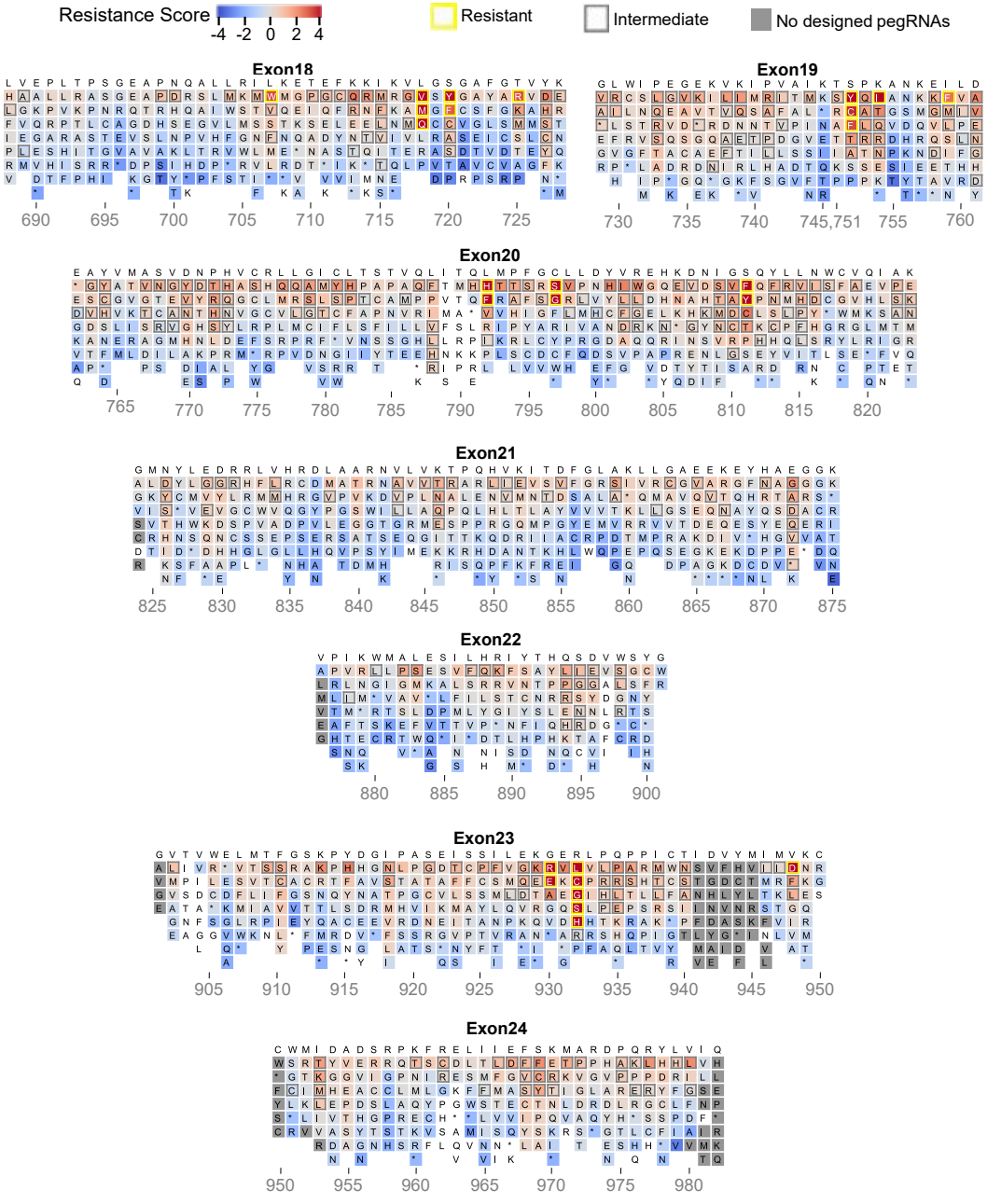


Figure 17. Heatmap illustrating osimertinib resistance scores of 1,817 protein variants in the presence of a co-occuring T790M mutation. These variants were generated through prime editing

within exons 18-24 of the EGFR gene in PC-9 cells containing the T790M mutation. Protein variants inducing resistant and intermediate phenotypes are demarcated by boxes outlined in yellow and gray, respectively. The numerical annotations at the bottom of each heatmap denote the positions within the EGFR amino acid sequence. For each position, the amino acid in the reference sequence is displayed at the top. Ninety-four protein variants with P-values exceeding 0.05 or odds ratios lower than 3 were excluded from the analysis and are depicted in a white background. Protein variants for which no pegRNAs were designed are represented by gray boxes.

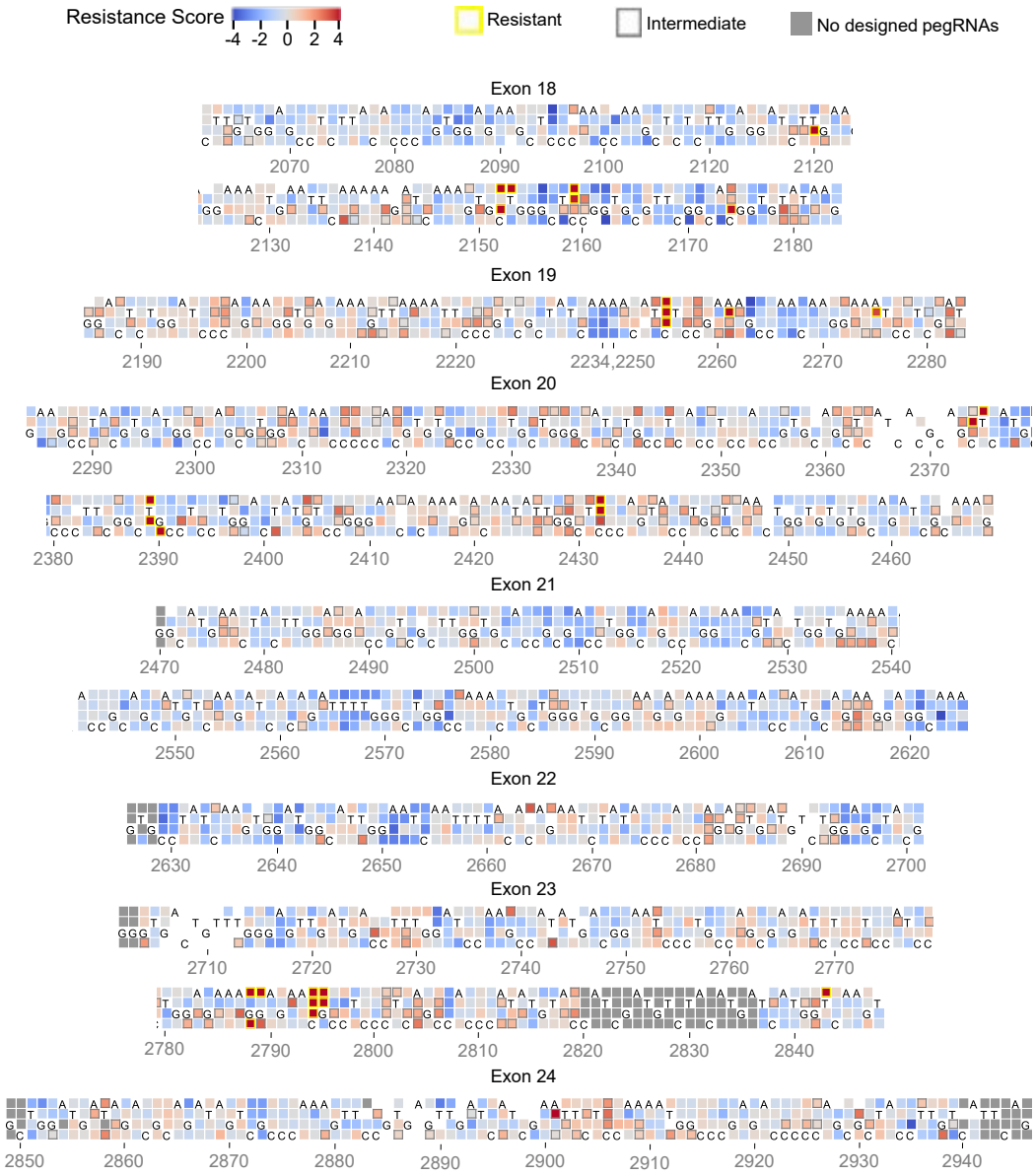


Figure 18. Heatmap exhibiting osimertinib resistance scores of 2,610 SNVs in the presence of the co-occuring T790M mutation. These SNVs were generated via prime editing within exons 18-24 of the EGFR gene in PC-9 cells containing the T790M mutation. SNVs associated with resistant and intermediate phenotypes are delineated by boxes outlined in yellow and gray, respectively. The

numerical annotations at the bottom of each heatmap denote the positions within the EGFR coding sequence. For each position, the nucleotide in the reference sequence is displayed. One hundred twenty-five SNVs with P-values exceeding 0.05 or odds ratios lower than 3 were excluded from the analysis and are depicted in a white background. SNVs for which no pegRNAs were designed are represented by gray boxes.

3.9. Resistance profiles of *EGFR* SNVs against afatinib and osimertinib in the absence of T790M and against osimertinib in the presence of T790M

We conducted a comprehensive analysis to delineate the resistance profiles of SNVs within the EGFR tyrosine kinase domain against afatinib and osimertinib in PC-9 cells. This investigation encompassed three conditions: afatinib and osimertinib treatments in the absence of T790M mutation, and osimertinib treatment in the presence of T790M mutation. Through SynPrime evaluation, we provide an extensive catalog of resistance profiles against these therapeutics for nearly all possible SNVs (**Figure 19, Table 1**). Our analysis revealed several noteworthy findings. Firstly, consistent with prior knowledge, the T790M mutation exhibited resistance to afatinib but not to osimertinib (**Figure 19a**), corroborating its role as a common gatekeeper mutation conferring resistance to 1st and 2nd generation TKIs[10,56]. Furthermore, the C797S mutation, a well-characterized resistance mechanism to both osimertinib and afatinib, demonstrated resistance across all three conditions[11,58,59]. The results of our SynPrime evaluation are compatible with representative previous reports about TKI resistance profiles.

In our investigation, we observed that nine distinct protein variants, notably including L718V, exhibited resistance specifically to osimertinib, while remaining susceptible to afatinib. This finding suggests a potential therapeutic avenue utilizing afatinib for patients bearing such mutations. Our results align with existing evidence, including a case report highlighting the association between the L718V mutation and osimertinib resistance, while maintaining sensitivity to afatinib[60]. Moreover, our study identified L718Q as another variant conferring resistance to osimertinib, particularly evident in PC-9-T790M cells, with an intermediary effect compared to afatinib and osimertinib in PC-9 cells. This observation correlates with prior research indicating that L718Q-mediated resistance to osimertinib in NSCLC is independent of T790M status, likely attributable to spatial constraints and altered hydrophobic interactions with osimertinib [15,61,62].

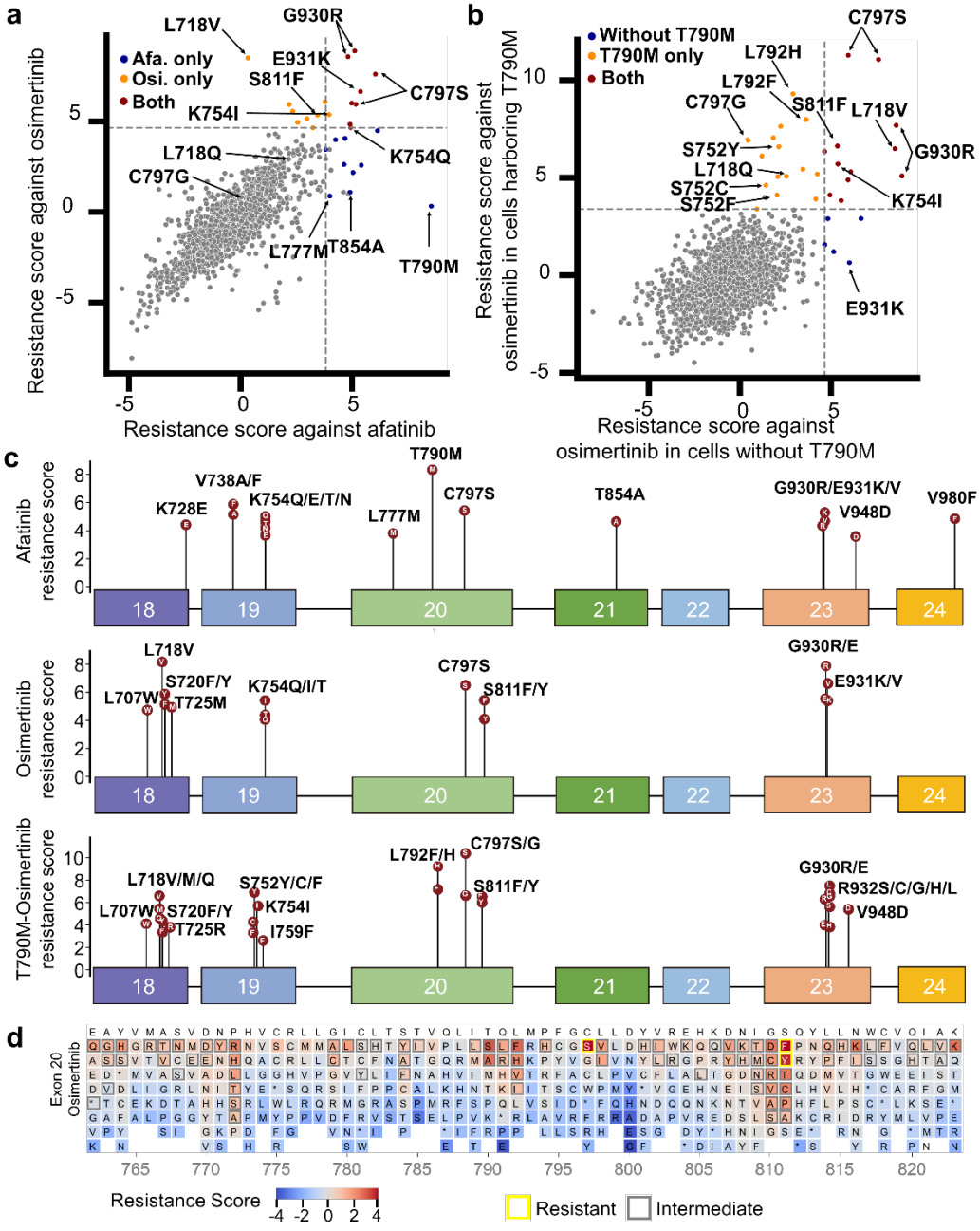


Figure 19. Landscape of TKI resistance mediated by SNVs within the EGFR tyrosine kinase domain. (a,b), Resistance profiles of EGFR SNVs in PC-9 cells devoid of the T790M gatekeeper

mutation, assessing resistance against osimertinib with and without the T790M mutation in PC-9 cells. **(c)**, Lollipop plot illustrating SNVs in exons 18 to 24 of EGFR and their corresponding resistance scores against Afatinib, Osimertinib, and Osimertinib in the presence of a co-occurring T790M mutation. **(d)**, Heatmap depicting Osimertinib resistance scores of 1,817 protein variants induced by prime editing in exon 20 of EGFR in PC-9 cells. variants leading to resistant and intermediate phenotypes are highlighted in yellow and gray outlines, respectively. excluded variants with p-values greater than 0.05 or odds ratios lower than 3 are shown with a white background, ensuring statistical rigor in the analysis.

Table 1. Protein variants classified as resistant in this study

Protein variants	SynPrime Afatinib	SynPrime Osimertinib	SynPrime T790M-osimertinib	Evidence of resistance in previously published literature
L707W	Intermediate	Resistant	Resistant	Not reported
L718V	Sensitive	Resistant	Resistant	Resistant to osimertinib, sensitive to afatinib[48]
L718M	Sensitive	Intermediate	Resistant	Not reported
L718Q	Intermediate	Intermediate	Resistant	Resistant to osimertinib[49], afatinib sensitive[50]
S720Y	Intermediate	Resistant	Resistant	Not reported
S720F	Intermediate	Resistant	Resistant	Not reported
T725M	Intermediate	Resistant	Intermediate	Sensitive to afatinib[3]
T725R	Intermediate	Intermediate	Resistant	Resistant to gefitinib[72]
K728E	Resistant	Intermediate	Intermediate	Not reported
V738F	Resistant	Intermediate	Intermediate	Not reported
V738A	Resistant	Intermediate	Intermediate	Not reported
S752Y	Sensitive	Intermediate	Resistant	Not reported
S752F	Intermediate	Intermediate	Resistant	Not reported
S752C	Intermediate	Intermediate	Resistant	Acquired after osimertinib treatment[51]
K754E	Resistant	Intermediate	Sensitive	Not reported
K754Q	Resistant	Resistant	Intermediate	Not reported
K754T	Resistant	Resistant	Intermediate	Not reported
K754N	Resistant	Intermediate	Intermediate	Not reported
K754I	Intermediate	Resistant	Resistant	Acquired after osimertinib treatment[57]
I759F	Sensitive	Sensitive	Resistant	Not reported

L777M	Resistant	Sensitive	Sensitive	Resistant to erlotinib[55]
T790M	Resistant	Sensitive	Sensitive	Resistant to afatinib[47], sensitive to osimertinib[73]
L792F	Intermediate	Intermediate	Resistant	Resistant to osimertinib[47] and afatinib[47]
L792H	Intermediate	Intermediate	Resistant	Resistant to osimertinib[49,74]
C797S	Resistant	Resistant	Resistant	Resistant to osimertinib[11] and afatinib[47]
C797G	Sensitive	Sensitive	Resistant	Resistant to osimertinib[52,53] in the presence with T790M
S811Y	Intermediate	Resistant	Resistant	Not reported
S811F	Intermediate	Resistant	Resistant	Not reported
T854A	Resistant	Intermediate	Sensitive	Resistant to gefitinib[54], Sensitive to osimertinib[56]
G930R	Resistant	Resistant	Resistant	Not reported
G930E	Intermediate	Resistant	Resistant	Not reported
E931K	Resistant	Resistant	Sensitive	Not reported
E931V	Resistant	Resistant	Intermediate	Not reported
R932S	Sensitive	Intermediate	Resistant	Not reported
R932C	Intermediate	Intermediate	Resistant	Not reported
R932G	Sensitive	Sensitive	Resistant	Not reported
R932H	Intermediate	Intermediate	Resistant	Not reported
R932L	Intermediate	Intermediate	Resistant	Not reported
V948D	Resistant	Intermediate	Resistant	Not reported
V980F	Resistant	Intermediate	Sensitive	Not reported

Additionally, our investigation uncovered 16 protein variants, such as S752Y/C/F and C797G, that conferred osimertinib resistance only in the presence of T790M. Conversely, in the absence of T790M, these variants displayed an intermediate or sensitive response to osimertinib, suggesting its potential efficacy in such scenarios. These findings are consistent with previous reports associating S752C[63] with osimertinib resistance post-T790M treatment and linking C797G to osimertinib resistance in the presence of T790M[64,65]. Remarkably, cells carrying the C797G mutation exhibited sensitivity to both afatinib and osimertinib in the absence of T790M, indicating the viability of both drugs for treating patients with this mutation in such contexts.

Moreover, L777M and T854A mutations were identified as resistant to afatinib, corroborating prior findings implicating them as acquired resistance mutations to first-generation TKIs[66,67], while demonstrating sensitivity to osimertinib treatment[68]. Although the resistance profiles of these mutations in relation to afatinib, a second-generation TKI, remain unexplored, our assessment suggests osimertinib as a potential therapeutic option for tumors harboring L777M or T854A mutations.

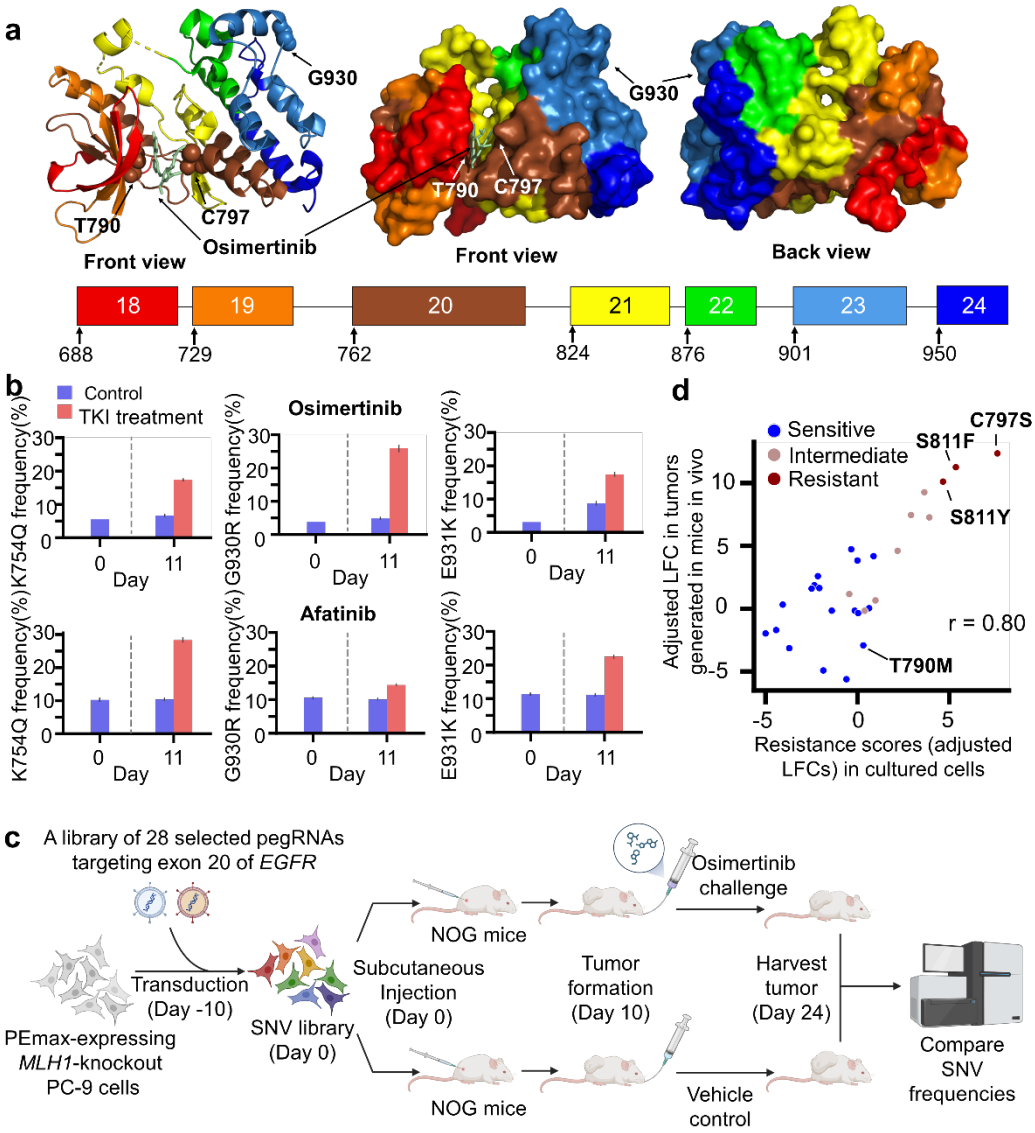
We identified clustered, resistance conferring SNVs that affected K754 (encoded in exon 19), S811 (exon 20), and G930 and E931 (exon 23). K754I conferred resistance against osimertinib irrespective of T790M status, aligning with clinical observations of its enrichment following osimertinib treatment in T790M-bearing patients[69].

Our investigation underscores the importance of exploring the less-investigated exons 22-24 of the EGFR gene, which lie outside the canonical ATP binding pocket delineated by exons 18-21[70,71] (**Figure 20a**). Notably, G930R, identified in exon 23, demonstrates uniform resistance across all treatment arms, indicating its pivotal role in conferring resistance to both afatinib and osimertinib. Additionally, E931K, also situated in exon 23, exhibits resistance against both afatinib and osimertinib exclusively in the absence of T790M. These nuanced responses underscore the intricate interplay between mutation status and drug efficacy, highlighting the need for comprehensive molecular profiling in guiding treatment decisions. Crystallographic studies elucidating the structural implications of residues E931 and G930 further support our findings, providing mechanistic insights into their role in modulating EGFR activation and response to TKIs[71,72]. Overall, our study expands the understanding of EGFR mutation profiles, suggesting potential drug choices tailored to specific mutation contexts.

3.10. Evaluation of TKI resistance in a conventional cell-based manner and in murine models

To assess the resistance of cells bearing newly identified SNVs in our study conventionally, lentivirus encoding pegRNAs targeting EGFR and generating SNVs (K754Q, G930R, or E931K) were individually delivered to PEmax-expressing MLH1-knockout PC-9 cells (**Figure 21**). These prime-edited cells were then mixed with control cells at a 25:75 ratio. Subsequently, the mixed cell populations underwent untreated conditions or were treated with afatinib or osimertinib, cultured for 10 days, and subjected to deep sequencing analysis. Notably, the mean frequencies of SNVs encoding K754Q, G930R, and E931K increased 2.6-fold (from 17% to 6.7%), 5.4-fold (from 26% to 4.8%), and 2.0-fold (from 17% to 8.7%), respectively, in the presence of osimertinib, and 2.7-fold (from 28% to 10%), 1.4-fold (from 14% to 10%), and 2-fold (from 22% to 11%) in the presence of afatinib, compared to the untreated control (**Figure 20b**). These findings substantiate the role of these SNVs in conferring resistance to both osimertinib and afatinib.

Moreover, we determined the half-maximal inhibitory concentrations (IC50s) of the tyrosine kinase inhibitors afatinib and osimertinib on PC-9 cells expressing newly identified resistant (K754Q, E931K, G930R) and sensitive (G796R, G796G) EGFR mutations, alongside well-characterized variants such as T790M and C797S. Our analysis revealed that cells harboring sensitive mutations exhibited lower IC50s compared to those carrying resistant variants (**Figure 20e**). Additionally, strong positive correlations were observed between resistance scores derived from the SynPrime algorithm and IC50 values (Spearman's R and Pearson's r values ranged from 0.80 to 0.91 across both drugs), thereby confirming that higher resistance scores correlate with elevated IC50s. These results validate the efficacy of the SynPrime algorithm and underscore its compatibility with conventional cell-based assays.



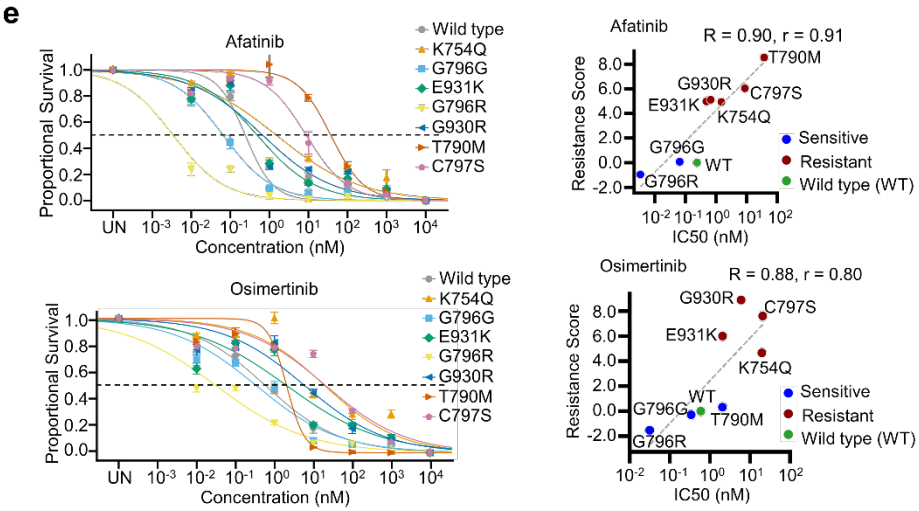


Figure 20. Evaluation of TKI resistance in a conventional cell-based and murine models. (a), Location of the G930 residue within the crystal structure of the EGFR tyrosine kinase domain bound to osimertinib (Protein Data Bank: 6JXT, EGFR complex with osimertinib). The different colors indicate amino acid regions corresponding to the coding sequence of each exon. **(b)**, Percentage of reads containing intended edits before (Day 0, blue) and after (Day 11) treatment with osimertinib (red) or control solvent (blue). The data include biological replicates for treatment with osimertinib (top) and afatinib (bottom), with n = 3 replicates. (**c**), Schematic overview of the strategy for functional evaluation of SNVs in EGFR exon 20 generated by prime editing in PC-9 cells, followed by transplantation into mice. A total of 28 pegRNAs were selected based on SynPrime evaluation results, including pegRNAs inducing 3 resistant, 7 intermediate, and 18 sensitive SNVs. (NOG, NOD/Shi-scid/IL-2Rγ-null). **(d)**, Correlation between resistance scores (adjusted LFCs) in cultured PC-9 cells and in tumors formed by transplantation of PC-9 cells into NOG mice. Pearson correlation coefficient (r) is provided. The number of SNVs is n = 28. (**e**), Dose-dependent sensitivity to afatinib and osimertinib in PC-9 cells containing specific variants (Left), and correlations between resistance scores determined by SynPrime experiments and half-maximal inhibitory concentrations (IC50) determined by the MTT assay (Right).

We conducted a comprehensive assessment of SNVs through high-throughput analysis in murine models. Following the transduction of the selected pegRNA library Syn-vivo-exon20 (refer to Methods) into PEmax-expressing MLH1-knockout PC-9 cells, these cells were subcutaneously transplanted into immunocompromised NOD-SCID mice to induce tumor formation (**Figure 20c**

Subsequently, the mice were treated with either osimertinib or a vehicle solution (serving as a negative control) for a duration of two weeks. Post-treatment, the transplanted tumors were harvested and subjected to deep sequencing analysis. Notably, we observed a robust correlation between the resistance scores obtained in cultured cells and those observed in tumors within the murine models (**Figure 20d**). This finding strongly indicates that the resistance scores determined using the SynPrime approach are reproducible in in vivo mouse models, thus bolstering the reliability of our methodology in assessing TKI resistance in a physiological context.

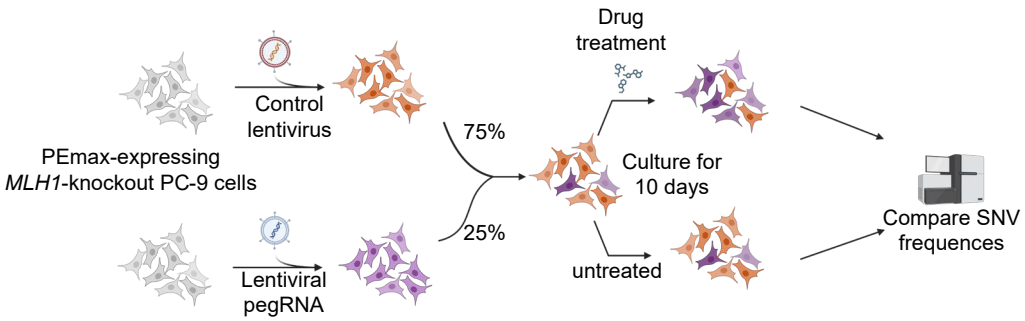


Figure 21. Schematic overview of the strategy for conventional evaluation of TKI resistance. Lentivirus encoding pegRNAs designed to induce the generation of SNVs encoding K754Q, G930R, and E931K mutations, alongside control lentivirus, were individually transduced into PEmax-expressing MLH1-knockout PC-9 cells. Ten days post-transduction, the two cell populations (i.e., cells transduced with pegRNA and those with empty vector) were mixed and divided into untreated and drug-treated conditions. Frequencies of SNVs were subsequently measured by targeted deep sequencing, ten days after initiation of drug treatment.

IV. DISCUSSION

In this investigation, we analyzed 2,476 (95%) of the single nucleotide variants (SNVs), corresponding to 1,726 (95%) of 1,817 potential protein-altering variants within the EGFR tyrosine kinase domain. Accurate drug resistance profiling is pivotal for informed clinical decision-making. To refine this accuracy, we implemented several methodologies: i) analysis was conducted directly on endogenous sequences rather than deriving inferences from guide RNA abundance[22,23,40,42] or surrogate reporter targets[24,25,41], ii) we notably reduced both sequencing and PCR-associated artifacts relative to the signals of targeted SNVs through the implementation of an additional synonymous editing strategy termed SynPrime (double-hit approach), iii) prime editing efficiency was enhanced via the DeepPrime-engineered pegRNAs, coupled with MLH1 knockout and the introduction of synonymous mutations to amplify the signal, iv) classification of SNVs as either significant or negligible was based on stringent filtering criteria using both P-values and odds ratios, rather than merely excluding reads below a certain frequency threshold[27], and v) SNVs were categorized as sensitive or resistant only when corroborated by findings from two distinct biological replicates.

Given that SNVs exert their functional effects through mutations within native gene sequences, it can be postulated that classifications derived directly from these sequences are more accurate compared to those obtained through indirect evaluations such as guide RNA abundance or surrogate reporter assays. However, direct empirical support for this hypothesis remains outstanding. This study furnishes evidence suggesting that direct interrogation of native sequences offers enhanced precision in the functional classification of SNVs, as demonstrated across two different genes: RPL15 and EGFR.

A primary challenge associated with direct analysis of native sequences is the low occurrence rate of intended SNVs. For instance, in high-throughput prime editing, the detectability of specific SNVs is complicated by the dilution effect caused by the extensive size of the pegRNA library. Assume a library comprises 600 pegRNA sequences, each designed to induce a distinct edit within a 200-bp target region, with an editing efficiency of 60% for each intended edit. The resulting frequency of any specific intended SNV would be a mere 0.1% (= 60%/600), a value comparable to the error rates typical in deep sequencing. The frequency diminishes further with larger libraries or lower editing efficiencies. In this study, the observed median frequencies of SNVs ranged from 0.009% to 0.012% across the analyzed EGFR exons. To enhance the accuracy of SNV detection, we

incorporated an additional substitution strategy that decreased the incidence of false-positive reads by a median factor of 626, thus improving the signal-to-noise ratio by approximately 443-fold.

Mutations in EGFR other than the well-documented variants (e.g., exon 19 in-frame deletions, L858R, T790M, and C797S) constitute up to 18% of all EGFR mutations observed in patients. The absence of comprehensive resistance profiles for these diverse mutations has impeded the selection of optimal therapeutic agents for affected individuals. This study contributes a detailed resistance profile for EGFR variants against tyrosine kinase inhibitors (TKIs), significantly advancing the pursuit of personalized medicine. Additionally, the extension of the SynPrime methodology to other genes like BRCA1, BRCA2, and TP53 could prove beneficial in addressing the challenges associated with variants of uncertain significance (VUSs). This approach might also be judiciously applied to genes characterized by prevalent hotspot mutations, such as HER2 and ESR1. Through careful selection of methodologies and anticancer agents, this approach can further the study of these genes. Collectively, these methodologies promise to initiate a new era in precision medicine.

V. CONCLUSION

In conclusion, our SynPrime methodology, characterized by the direct analysis of endogenous sequences and sophisticated noise-reduction techniques, delivers a highly precise and comprehensive evaluation of single-nucleotide variants (SNVs) in genes such as EGFR and RPL15. Employing deep sequencing paired with a double-hit strategy, this approach provides dependable drug resistance profiles that surpass those derived from traditional methods dependent on guide RNA abundance. The adaptability of SynPrime extends to a variety of genes, addressing challenges associated with variants of uncertain significance (VUSs), thus positioning it as a transformative tool for precision medicine and enabling the customization of therapeutic strategies tailored to individual patients.

References

1 Russo, A. et al. (2019). Heterogeneous Responses to Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Inhibitors (TKIs) in Patients with Uncommon EGFR Mutations: New Insights and Future Perspectives in this Complex Clinical Scenario. Int. J. Mol. Sci. 20.

2 Passaro, A. et al. (2021). Recent Advances on the Role of EGFR Tyrosine Kinase Inhibitors in the Management of NSCLC With Uncommon, Non Exon 20 Insertions, EGFR Mutations. J. Thorac. Oncol. 16, 764-773.

3 Yang, J. C. et al. (2020). Afatinib for the Treatment of NSCLC Harboring Uncommon EGFR Mutations: A Database of 693 Cases. J. Thorac. Oncol. 15, 803-815.

4 Janning, M. et al. (2022). Treatment outcome of atypical EGFR mutations in the German National Network Genomic Medicine Lung Cancer (nNGM). Ann. Oncol. 33, 602-615.

5 Yang, J. C. et al. (2012). Afatinib for patients with lung adenocarcinoma and epidermal growth factor receptor mutations (LUX-Lung 2): a phase 2 trial. Lancet Oncol. 13, 539-548.

6 Soria, J. C. et al. (2018) Osimertinib in Untreated EGFR-Mutated Advanced Non-Small-Cell Lung Cancer. N. Engl. J. Med. 378, 113-125.

7 Yun, C.-H. et al. (2007) Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity. Cancer Cell 11, 217-227.

8 Lynch, T. J. et al. (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N. Engl. J. Med. 350, 2129-2139.

9 Paez, J. G. et al. (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304, 1497-1500.

10 Campo, M. et al. (2016) Acquired Resistance to First-Line Afatinib and the Challenges of Prearranged Progression Biopsies. J. Thorac. Oncol. 11, 2022-2026.

11 Thress, K. S. et al. (2015) Acquired EGFR C797S mutation mediates resistance to AZD9291 in non-small cell lung cancer harboring EGFR T790M. Nat. Med. 21, 560-562.

12 Pretelli, G. et al. (2023) Overview on Therapeutic Options in Uncommon EGFR Mutant Non-Small Cell Lung Cancer (NSCLC): New Lights for an Unmet Medical Need. Int. J. Mol. Sci. 24.

13 Kohsaka, S. et al. (2017) A method of high-throughput functional evaluation of EGFR gene variants of unknown significance in cancer. Sci. Transl. Med. 9 eaan6566.

14 Chakroborty, D. et al. (2019) An unbiased in vitro screen for activating epidermal growth factor receptor mutations. J. Biol. Chem. 294, 9377-9389.

15 Robichaux, J. P. et al. (2021) Structure-based classification predicts drug response in EGFR-mutant NSCLC. Nature 597, 732-737.

16 An, L. et al. (2023) Defining the sensitivity landscape of EGFR variants to tyrosine kinase inhibitors. Transl. Res. 255, 14-25.

17 Gaudelli, N. M. et al. (2017) Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature 551, 464-471.

18 Komor, A. C. et al. (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 533, 420-424.

19 Nishida, K. et al. (2016) Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science 353, aaf8729.

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

21 Findlay, G. M. et al. (2018) Accurate classification of BRCA1 variants with saturation genome editing. Nature 562, 217-222.

22 Hanna, R. E. et al. (2021) Massively parallel assessment of human variants with base editor screens. Cell 184, 1064-1080.e1020.

23 Cuella-Martin, R. et al. (2021) Functional interrogation of DNA damage response variants with base editing screens. Cell 184, 1081-1097 e1019.

24 Kim, Y. et al. (2022) High-throughput functional evaluation of human cancer-associated mutations using base editors. Nat. Biotechnol. 40, 874-884.

25 Sánchez-Rivera, F. J. et al. (2022) Base editing sensor libraries for high-throughput engineering and functional analysis of cancer-associated single nucleotide variants. Nat. Biotechnol. 40, 862-873.

26 Perner, F. et al. (2023) MEN1 mutations mediate clinical resistance to menin inhibition. Nature 615, 913-919.

27 Erwood, S. et al. (2022) Saturation variant interpretation using CRISPR prime editing. Nat. Biotechnol. 40, 885-895.

28 Sharma, S. V. et al. (2007) Epidermal growth factor receptor mutations in lung cancer. Nat. Rev. Cancer 7, 169-181.

29 Nelson, J. W. et al. (2022) Engineered pegRNAs improve prime editing efficiency. Nat. Biotechnol. 40, 402-410.

30 Kim, H. K. et al. (2017) In vivo high-throughput profiling of CRISPR-Cpf1 activity. Nat. Methods 14, 153-159.

31 Dang, Y. et al. (2015) Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome Biol. 16, 280.

32 Sanjana, N. E. et al. (2014) Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods 11, 783-784.

33 Briggs, K. J. et al. (2016) Paracrine Induction of HIF by Glutamate in Breast Cancer: EglN1 Senses Cysteine. Cell 166, 126-139.

34 Kim, S. et al. (2017) Rescue of high-specificity Cas9 variants using sgRNAs with matched 5' nucleotides. Genome Biol. 18, 218.

35 Koblan, L. W. et al. (2018) Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. Nat. Biotechnol. 36, 843-846.

36 Heckl, D. et al. (2014) Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing. Nat. Biotechnol. 32, 941-946.

37 Shalem, O. et al. (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343, 84-87.

38 Landrum, M. J. et al. (2020) ClinVar: improvements to accessing data. Nucleic Acids Res. 48, 835-844.

39 Masago, K. et al. (2015) Next-generation sequencing of tyrosine kinase inhibitor-resistant non-small-cell lung cancers in patients harboring epidermal growth factor-activating mutations. BMC Cancer 15, 908.

40 Cross, D. A. et al. (2014) AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. Cancer Discov. 4, 1046-1061.

41 Zhang, Q. et al. (2018) EGFR L792H and G796R: Two Novel Mutations Mediating Resistance to the Third-Generation EGFR Tyrosine Kinase Inhibitor Osimertinib. J. Thorac. Oncol. 13, 1415-1421.

42 Findlay, G. M. et al. (2014) Saturation editing of genomic regions by multiplex homology-

directed repair. Nature 513, 120-123.

43 Meitlis, I. et al. (2020) Multiplexed Functional Assessment of Genetic Variants in CARD11. Am. J. Hum. Genet. 107, 1029-1043.

44 Yu, G. et al. (2023) Prediction of efficiencies for diverse prime editing systems in multiple cell types. Cell 186, 2256-2272.

45 Miller, S. M. et al. (2020) Continuous evolution of SpCas9 variants compatible with non-G PAMs. Nat. Biotechnol. 38, 471-481.

46 Kim, N. et al. (2023) Deep learning models to predict the editing efficiencies and outcomes of diverse base editors. Nat. Biotechnol., doi:10.1038/s41587-023-01792-x.

47 Chen, P. J. et al. (2021) Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. Cell 184, 5635-5652.

48 Salk, J. J. et al. (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat. Rev. Genet. 19, 269-285.

49 Stoler, N. et al. (2021) Sequencing error profiles of Illumina sequencing instruments. NAR Genom. Bioinform. 3, doi:10.1093/nargab/lqab019.

50 Hart, T. et al. (2015) High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell 163, 1515-1526.

51 Doench, J. G. et al. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. 34, 184-191.

52 Doench, J. G. et al. (2018) Am I ready for CRISPR? A user's guide to genetic screens. Nat. Rev. Genet. 19, 67-80.

53 Ren, X. et al. (2023) High throughput PRIME editing screens identify functional DNA variants in the human genome. bioRxiv.

54 Gould, S. I. et al. (2023) High throughput evaluation of genetic variants with prime editing sensor libraries. bioRxiv.

55 Chardon, F. M. et al. (2023) A multiplex, prime editing framework for identifying drug resistance variants at scale. bioRxiv.

56 Yu, H. A. et al. (2013) Analysis of tumor specimens at the time of acquired resistance to EGFR-TKI therapy in 155 patients with EGFR-mutant lung cancers. Clin. Cancer Res. 19, 2240-2247.

57 Mok, T. S. et al. (2016) Osimertinib or Platinum–Pemetrexed in EGFR T790M–Positive

Lung Cancer. 376, 629-640.

58 Tan, C.-S. et al. (2018) Third generation EGFR TKIs: current data and future directions. Mol. Cancer 17, 29.

59 Kobayashi, Y. et al. (2017) Characterization of EGFR T790M, L792F, and C797S Mutations as Mechanisms of Acquired Resistance to Afatinib in Lung Cancer. Mol. Cancer Ther. 16, 357-364.

60 Liu, Y. et al. (2018) Acquired EGFR L718V mutation mediates resistance to osimertinib in non-small cell lung cancer but retains sensitivity to afatinib. Lung Cancer 118, 1-5.

61 Yang, Z. et al. (2018) Investigating Novel Resistance Mechanisms to Third-Generation EGFR Tyrosine Kinase Inhibitor Osimertinib in Non-Small Cell Lung Cancer Patients. Clin. Cancer Res. 24, 3097-3107.

62 Li, M. et al. (2022) L718Q/V mutation in exon 18 of EGFR mediates resistance to osimertinib: clinical features and treatment. Discover Oncology 13, 72.

63 Sueoka-Aragane, N. et al. (2021) The role of comprehensive analysis with circulating tumor DNA in advanced non-small cell lung cancer patients considered for osimertinib treatment. Cancer Med. 10, 3873-3885.

64 Carlo, D. E. et al. (2021) Acquired EGFR C797G Mutation Detected by Liquid Biopsy as Resistance Mechanism After Treatment With Osimertinib: A Case Report. In Vivo 35, 2941-2945.

65 Nie, K. et al. (2018) Mutational Profiling of Non-Small-Cell Lung Cancer Resistant to Osimertinib Using Next-Generation Sequencing in Chinese Patients. Biomed. Res. Int. 2018, 9010353.

66 Bean, J. et al. (2008) Acquired resistance to epidermal growth factor receptor kinase inhibitors associated with a novel T854A mutation in a patient with EGFR-mutant lung adenocarcinoma. Clin. Cancer Res. 14, 7519-7525.

67 Avizienyte, E. et al. (2008) Comparison of the EGFR resistance mutation profiles generated by EGFR-targeted tyrosine kinase inhibitors and the impact of drug combinations. Biochem. J. 415, 197-206.

68 Zhang, L. et al. (2022) Molecular Characteristics of the Uncommon EGFR Exon 21 T854A Mutation and Response to Osimertinib in Patients With Non-Small Cell Lung Cancer. Clin. Lung Cancer 23, 311-319.

69 Xing, P. et al. (2019) Co-mutational assessment of circulating tumour DNA (ctDNA)

during osimertinib treatment for T790M mutant lung cancer. J. Cell. Mol. Med. 23, 6812-6821.

70 Malapelle, U. et al. (2017) Profile of the Roche cobas® EGFR mutation test v2 for non-small cell lung cancer. Expert Rev. Mol. Diagn. 17, 209-215.

71 Yu, Z. et al. (2007) Resistance to an irreversible epidermal growth factor receptor (EGFR) inhibitor in EGFR-mutant lung cancer reveals novel treatment strategies. Cancer Res. 67, 10417-10427.

72 Zhang, X. et al. (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. Cell 125, 1137-1149.

Abstract in Korean

프라임 편집 대량 스크리닝을 이용한 약물감수성과 관련된 새로운 표피성장인자수용체 불확실성 변이형 발견

단일염기변이의 표현형을 이해하는 것은 종양과 더불어 다양한 질병에서 해결해야 할 과제로 남아있음. 하지만, 이러한 유전변이를 적절히 유발하고 평가하는 방법에 있어 기술적인 한계가 있는 상황임. 본 연구에서는 프라임 편집기를 이용하여 단일염기변이를 대량으로 유발하면서 그 표현형을 정확도 높게 평가할 수 있는 기술인 '신프라임'을 최초로 개발하였음. 이는 평가하고자 하는 단일염기변이 인접코돈에 동의 돌연변이 (Synonymous mutation)을 추가적으로 도입함으로써 프라임편집기 교정 효율을 높이고 시퀀싱 기법상에서의 에러와의 구분을 높은 정확도로 해주는 방법임.

본 연구에서는 해당 기술을 이용하여 표피성장인자수용체 (EGFR) 유전자의 티로신키나제 (Tyrosine kinase) 도메인에 대해 2,476개의 단일염기변이를 대량으로 유발하였고, 2세대 및 3세대 티로신키나제억제제에 대한 암변이의 저항성 여부를 평가함. 또한, T790M 변이를 동반한 단일염기변이를 대량으로 유발하여 암변이의 항암제 저항성을 평가함. 스크리닝 결과가 가이드라인 수준의 결과와 일치하며, 개별테스트와 동물실험에서도 재현되는 것을 확인함.

본 연구에서 제시한 스크리닝 방법은 단일 염기 수준의 세포변이의 기능변화를 높은 정확도로 평가함으로써 다양한 질병과 유전자에서 변이의 표현형을 평가하는 데에 활용할 수 있을 것으로 기대됨.

핵심되는 말: 크리스퍼 스크리닝, 프라임 편집기, 불확실형 변이형, 항암제 저항성변이 스크리닝

PUBLICATION LIST

Oh, H. *et al.* Combined effects of niclosamide and temozolomide against human glioblastoma tumorspheres. *J Cancer Res Clin Oncol.* **146(11)**, 2817-2828 (2020). (first author)

Oh, H. *et al.* Apparent diffusion coefficient as a prognostic factor in clival chordoma. *Sci Rep.* **11**(486) (2021). (first author)

Oh, H. *et al.* The role of apparent diffusion coefficient as a predictive factor for tumor recurrence in patients with cerebellopontine angle epidermoid tumor. *Neurosurg Rev.* **45**, 1383-1392 (2022). (first author)

Oh, H. *et al.* The effect of hematoma evacuation with decompressive craniectomy on clinical outcomes in patients with parenchymal hematoma type 2 of hemorrhagic transformation after middle cerebral artery infarction. *Neurol Res.* **44**, 894-901 (2022). (first author)