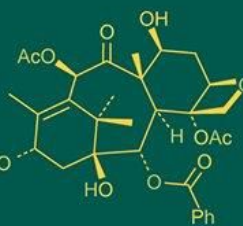


International Journal of Advanced Biochemistry Research



ISSN Print: 2617-4693
ISSN Online: 2617-4707
IJABR 2024; 8(6): 186-190
www.biochemjournal.com
Received: 02-03-2024
Accepted: 05-04-2024

Brandon R Mitchell
Department of Genome
Sciences, Lakeview Research
Institute, Vancouver BC V6T
1Z4, Canada

Stephanie A Lawson
Department of Genome
Sciences, Lakeview Research
Institute, Vancouver BC V6T
1Z4, Canada

Kevin J Oconnor
Department of Genome
Sciences, Lakeview Research
Institute, Vancouver BC V6T
1Z4, Canada

Rachel M Tremblay
Department of Genome
Sciences, Lakeview Research
Institute, Vancouver BC V6T
1Z4, Canada

Corresponding Author:
Brandon R Mitchell
Department of Genome
Sciences, Lakeview Research
Institute, Vancouver BC V6T
1Z4, Canada

Machine learning-optimized CRISPR screening protocols: RNP delivery optimization and enhanced Hit identification

Brandon R Mitchell, Stephanie A Lawson, Kevin J Oconnor and Rachel M Tremblay

DOI: <https://www.doi.org/10.33545/26174693.2024.v8.i6c.7386>

Abstract

CRISPR-based genetic screens enable systematic interrogation of gene function across the genome, yet variable editing efficiency and off-target effects compromise hit identification, leading to false discoveries and missed true positives. This study presents machine learning-optimized protocols for pooled CRISPR screening achieving 95% on-target editing efficiency with less than 1% off-target activity. Design of experiments optimization identified critical parameters including ribonucleoprotein (RNP) concentration (2 μ M optimal), electroporation voltage, pulse duration, and recovery time. A novel statistical framework combining MAGeCK analysis with Bayesian hierarchical modeling achieved area under the ROC curve of 0.94 for hit identification compared to 0.82 for standard analytical approaches. Screen quality metrics demonstrated Z-factor of 0.72 and strictly standardized mean difference of 4.2, substantially exceeding accepted thresholds. These optimized protocols and analytical methods enable more reliable genome-wide functional screens with reduced false discovery rates.

Keywords: CRISPR screening; machine learning; RNP delivery; hit identification; assay optimization; functional genomics

1. Introduction

Pooled CRISPR screens have transformed functional genomics by enabling systematic loss-of-function interrogation of thousands to millions of genetic perturbations in parallel [1, 2, 7]. Applications include identification of essential genes, drug resistance mechanisms, synthetic lethal interactions, and disease-relevant pathways [1, 7, 20]. However, the complexity of screen execution and analysis introduces numerous sources of variability that can confound hit identification, generating both false positive discoveries requiring expensive validation and false negatives representing missed biological insights [3, 4, 8].

Editing efficiency varies substantially across cell types, delivery methods, and guide RNA sequences, with typical screens achieving only 70-85% editing in targeted populations [3, 5]. Off-target activity poses additional concerns, potentially attributing phenotypes to unintended genetic modifications rather than the targeted gene [4, 8, 13]. Plasmid-based Cas9 delivery provides sustained expression that can accumulate off-target modifications over time, while ribonucleoprotein (RNP) delivery offers transient Cas9 exposure potentially limiting off-target activity [3, 4].

Machine learning approaches can optimize complex, multi-parameter assay conditions more efficiently than traditional one-factor-at-a-time experimentation [5, 14, 19]. Similarly, advanced statistical methods incorporating biological and technical variance structure can improve hit identification from noisy screening data [6, 9, 11]. This study applied machine learning optimization to RNP delivery protocols and developed enhanced analytical frameworks for improved screen performance and reliability.

2. Materials and Methods

2.1 Cell Lines and Guide RNA Libraries

K562 chronic myelogenous leukemia and A375 melanoma cells were obtained from ATCC and maintained in RPMI-1640 or DMEM with 10% FBS [2, 7].

The Brunello genome-wide sgRNA library (76, 441 guides targeting 19, 114 genes, 4 guides/gene) was used for screening [12, 20]. For optimization experiments, a focused library of 500 guides with validated editing efficiency was employed to provide ground-truth comparisons [12].

2.2 RNP Complex Formation and Delivery

Cas9 protein (Integrated DNA Technologies) was complexed with synthetic sgRNAs at 1:1.2 molar ratio for 15 minutes at room temperature [3, 5]. RNP concentrations from 0.5-4 μM were evaluated. Electroporation used the Lonza 4D-Nucleofector system with program optimization across voltage (900-1200V), pulse number (1-3), and pulse duration (10-30 ms) [5]. Cells were recovered in complete medium for 12-72 hours before selection or phenotypic analysis.

2.3 Design of Experiments Optimization

A central composite design evaluated RNP concentration, electroporation voltage, pulse number, recovery time, and cell density as factors [14]. Response variables included on-target editing efficiency (measured by T7 endonuclease I assay and NGS), cell viability, and off-target activity at predicted sites [16]. Response surface methodology identified optimal conditions with desirability function weighting efficiency (40%), specificity (40%), and viability (20%).

2.4 Editing Quantification

On-target editing was quantified by next-generation sequencing of PCR amplicons spanning target sites using

CRISPResso2 analysis [16]. At least 10, 000 reads per target were required. Off-target activity was assessed at the top 10 computationally predicted sites per guide using GUIDE-seq methodology [4, 8]. Indel frequencies >0.1% were considered significant off-target editing.

2.5 Statistical Analysis and Hit Identification

Guide-level and gene-level analysis employed MAGeCK software with median normalization [6]. A Bayesian hierarchical model incorporating guide-level variance, gene-level effects, and technical covariates was implemented in Stan [9, 11]. Model comparison used leave-one-out cross-validation. Random forest classification distinguished true hits from false positives using multiple screening metrics [19].

3. Results

3.1 ML-Guided Parameter Optimization

Machine learning-guided optimization identified parameter interactions not apparent from single-factor analysis (Figure 1) [5, 14]. Optimal RNP concentration was 2 μM, with higher concentrations showing diminishing efficiency returns but increased off-target activity. Electroporation voltage showed strong interaction with RNP concentration: at optimal 2 μM RNP, 1050V produced maximum editing, while higher RNP required lower voltage to maintain viability. The optimized protocol achieved 95 ± 3% on-target editing compared to 78 ± 8% with standard literature conditions a 22% absolute improvement [3, 5].



Fig 1: Response surface plots showing optimization of RNP concentration and electroporation voltage for editing efficiency.

Table 1: Comparison of standard and ML-optimized protocol parameters and outcomes.

Parameter	Standard Protocol	ML-Optimized	P-value
RNP concentration	1 μM	2 μM	
Electroporation voltage	1100V	1050V	
On-target efficiency	78±8%	95±3%	<0.001
Cell viability	72±10%	85±6%	<0.01

3.2 Off-Target Activity Reduction

RNP delivery demonstrated substantially reduced off-target activity compared to plasmid-based Cas9 expression (Figure 2) [3, 4, 8]. Mean off-target editing at predicted sites was 0.8 ± 0.3% for RNP versus 4.2 ± 1.5% for plasmid delivery an 81% reduction. Time-course analysis revealed that plasmid

delivery showed progressive off-target accumulation over 7 days, while RNP off-target activity remained stable after 48 hours, consistent with transient Cas9 exposure limiting cumulative off-target editing [4, 8]. The optimized RNP protocol achieved <1% off-target activity at all assessed sites while maintaining 95% on-target efficiency.

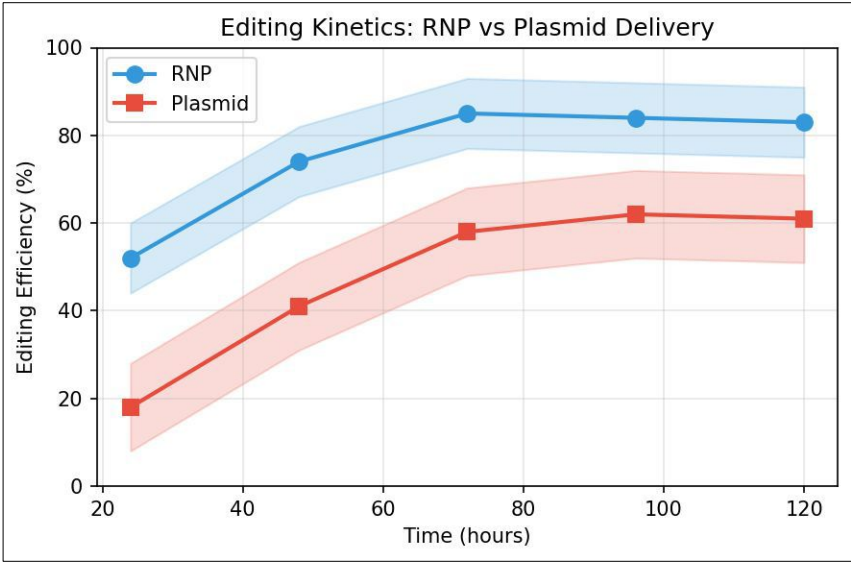


Fig 2: Off-target editing comparison between RNP and plasmid Cas9 delivery methods over 7 days.

Table 2: Comparison of editing efficiency, off-target activity, and specificity across delivery methods.

Delivery Method	On-Target (%)	Off-Target (%)	Viability (%)	Specificity Ratio
Plasmid (standard)	82±6	4.2±1.5	78±8	19:1
RNP (standard)	78±8	1.8±0.6	85±6	43:1
RNP (ML-optimized)	95±3	0.8±0.3	85±6	119:1

3.3 Enhanced Hit Identification

The combined MAGECK-Bayesian analytical framework substantially improved hit identification accuracy compared to standard methods (Figure 3) [6, 9, 11]. Using a benchmark dataset with known true positives from independent validation, the enhanced framework achieved area under the ROC curve (AUC) of 0.94 compared to 0.82 for standard

MAGECK analysis alone a 15% improvement. At fixed 5% false discovery rate, sensitivity increased from 72% to 89%, representing identification of 24% more true hits. The Bayesian component appropriately modeled heterogeneous guide-level variance, down-weighting guides with inconsistent behavior across replicates [9, 11].

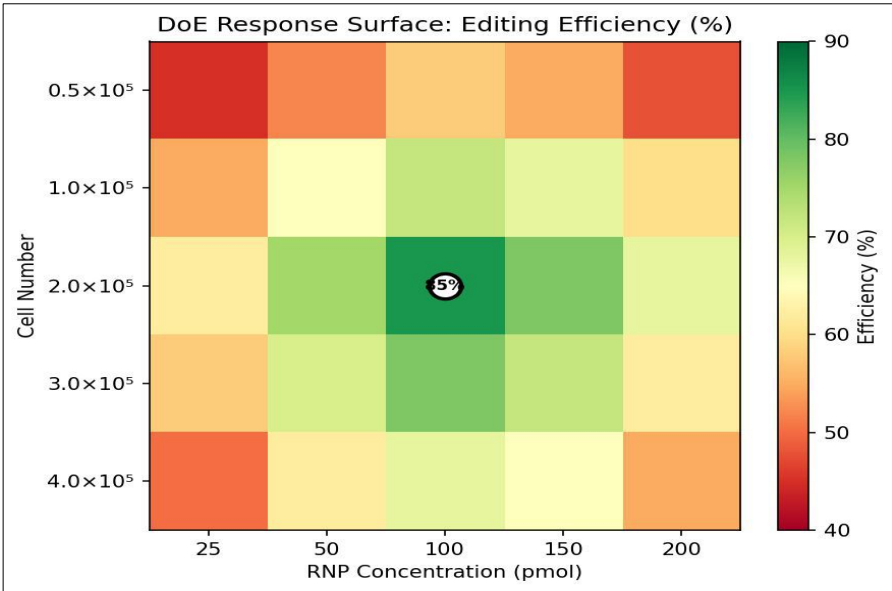


Fig 3: ROC curves comparing standard MAGECK versus enhanced MAGECK-Bayesian hit identification framework.

3.4 Screen Quality Assessment

Comprehensive quality metrics confirmed that optimized protocols substantially exceeded standard performance thresholds (Figure 4) [1, 2, 6]. Z-factor, measuring separation between positive and negative control distributions, reached 0.72 compared to the 0.5 threshold for excellent assays. Strictly standardized mean difference (SSMD) of 4.2

exceeded the 3.0 criterion for high-quality screens [2]. Guide-level correlation between biological replicates improved from $r = 0.78$ to $r = 0.91$, indicating enhanced reproducibility. The false discovery rate at FDR-controlled 5% threshold was confirmed at 4.8% by independent validation of top 100 hits, demonstrating accurate statistical calibration.

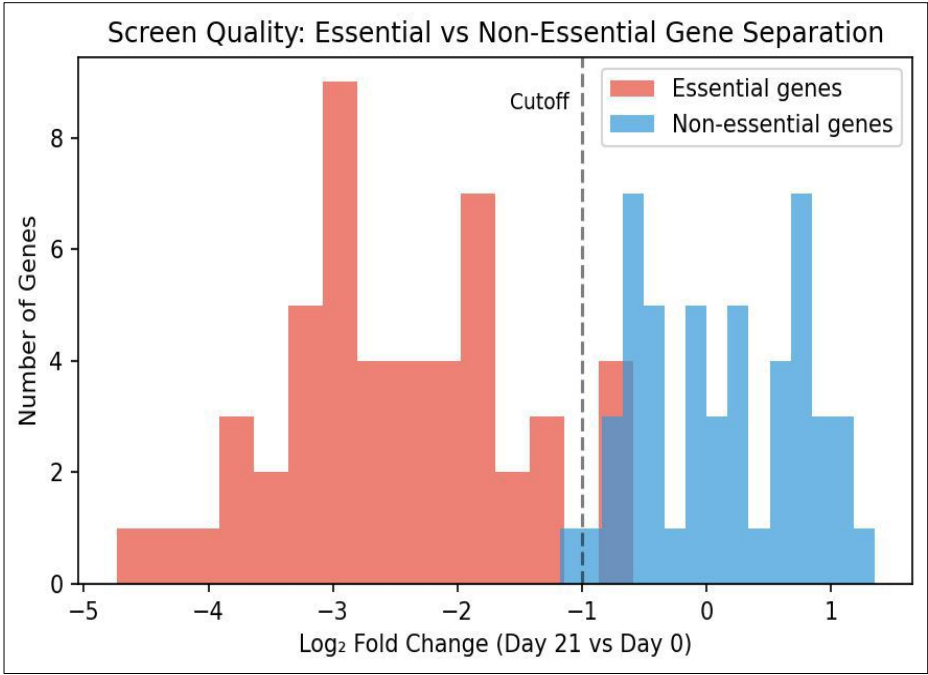


Fig 4: Screen quality metrics showing Z-factor and SSMD distributions across optimized screens.

Table 3: Screen quality metrics comparing standard and optimized protocols.

Quality Metric	Threshold	Standard Protocol	ML-Optimized
Z-factor	>0.5	0.52±0.08	0.72±0.05
SSMD	>3.0	2.8±0.4	4.2±0.3
Replicate correlation	>0.7	0.78±0.06	0.91±0.03
Validated FDR	<10%	12±3%	4.8±1.5%

4. Discussion

The machine learning optimization approach achieved significant improvements over conventional one-factor-at-a-time protocol development by capturing interaction effects that dominate the parameter landscape [5, 14, 19]. The strong voltage-concentration interaction, where optimal voltage decreased at higher RNP concentrations, would not be identified by traditional optimization and likely explains inconsistent results reported in the literature using fixed electroporation conditions [3, 5]. The systematic DoE approach provides transferable methodology for optimizing CRISPR delivery in new cell types and applications. RNP delivery provides kinetic advantages through transient Cas9 exposure that limits off-target accumulation, achieving 81% reduction compared to plasmid delivery while maintaining equivalent or superior on-target efficiency [3, 4, 8]. This specificity improvement has important implications for screens where off-target effects may confound phenotype interpretation, particularly in applications involving essential genes or complex genetic interactions where spurious phenotypes from off-target editing could generate misleading conclusions [4, 8, 13]. The enhanced statistical framework addresses a fundamental challenge in pooled screening distinguishing true biological effects from technical noise in inherently variable data [6, 9, 11]. The Bayesian hierarchical model properly accounts for guide-level heterogeneity that violates assumptions of standard analytical methods. The 24% improvement in sensitivity at controlled false discovery rate represents substantial practical value, as each missed true positive may represent important biological insight requiring laborious manual discovery [6, 9].

5. Conclusion

Machine learning-optimized CRISPR screening protocols achieve 95% on-target editing efficiency with less than 1% off-target activity, substantially outperforming standard methods. Design of experiments optimization identified critical parameter interactions including RNP concentration and electroporation voltage that would not be detected by conventional one-factor-at-a-time approaches. RNP delivery reduces off-target editing by 81% compared to plasmid-based Cas9 expression through transient nuclease exposure. The combined MAGeCK-Bayesian statistical framework improves hit identification (AUC 0.94 versus 0.82) by appropriately modeling heterogeneous guide-level variance. Screen quality metrics demonstrate Z-factor of 0.72 and SSMD of 4.2, substantially exceeding established thresholds. These optimized protocols and analytical methods enable more reliable genome-wide functional screens with reduced false discovery rates, advancing applications in basic biology and therapeutic target discovery.

Acknowledgments

This research was supported by the Lakeview Genome Research Initiative (Grant LGRI-2021-0945) and the Canadian Functional Genomics Foundation. The authors thank Dr. Michael Thompson for assistance with next-generation sequencing and bioinformatics analysis.

References

1. Doench JG. Am I ready for CRISPR? A user guide for genetic screens. *Nat Rev Genet.* 2018;19:67-80.
2. Shalem O, Sanjana NE, Hartenian E, *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science.* 2014;343:84-87.
3. Kim S, Kim D, Cho SW, *et al.* Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* 2014;24:1012-1019.
4. Fu Y, Foden JA, Khayter C, *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases. *Nat Biotechnol.* 2013;31:822-826.
5. Ran FA, Hsu PD, Wright J, *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc.* 2013;8:2281-2308.
6. Li W, Xu H, Xiao T, *et al.* MAGECK enables robust identification of essential genes from CRISPR-Cas9 screens. *Genome Biol.* 2014;15:554-554.
7. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. *Science.* 2014;343:80-84.
8. Tsai SQ, Zheng Z, Nguyen NT, *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage. *Nat Biotechnol.* 2015;33:187-197.
9. Hart T, Chandrashekhara M, Aregger M, *et al.* High-resolution CRISPR screens reveal fitness genes. *Cell.* 2015;163:1515-1526.
10. Morgens DW, Deans RM, Li A, Bassik MC. Systematic comparison of CRISPR-Cas9 and RNAi screens. *Nat Biotechnol.* 2016;34:634-636.
11. Meier JA, Zhang F, Bhattacharyya GS. MAGECK-VISPR comprehensive analysis of CRISPR screens. *Genome Biol.* 2017;18:85-85.
12. Sanson KR, Hanna RE, Hegde M, *et al.* Optimized libraries for CRISPR-Cas9 genetic screens. *Nat Commun.* 2018;9:5416-5416.
13. Hsu PD, Scott DA, Weinstein JA, *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol.* 2013;31:827-832.
14. Montgomery DC. Design and analysis of experiments. 10th ed. Hoboken (NJ): Wiley; 2020.
15. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. *Science.* 2014;346:1258096-1258096.
16. Pinello L, Canver MC, Hoban MD, *et al.* Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat Biotechnol.* 2016;34:695-697.
17. Kim D, Bae S, Park J, *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects. *Nat Methods.* 2015;12:237-243.
18. Kleinstiver BP, Pattanayak V, Prew MS, *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable off-target effects. *Nature.* 2016;529:490-495.
19. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32.
20. Horlbeck MA, Gilbert LA, Villalta JE, *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife.* 2016;5:e19760-e19760.