Innovations in Programmable Nucleic Acid Libraries and CRISPR Enrichment for Molecular Biology Applications

by

Natanya Kaitlin Villegas

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Biology

Dissertation Committee:

Dr. Diane Hawley, Chair

Dr. Calin Plesa, Advisor

Dr. Kryn Stankunas, Core Member

Dr. Matthew Barber, Core Member

Dr. Victoria DeRose, Institutional Representative

University of Oregon

Spring 2025

DISSERTATION ABSTRACT

Natanya Kaitlin Villegas

Doctor of Philosophy in Biology

Title: Innovations in Programmable Nucleic Acid Libraries and CRISPR Enrichment for Molecular Biology Applications

Synthetic gene libraries are pivotal in advancing protein engineering, functional genomics, and synthetic biology, yet their quality is hindered by errors in oligonucleotide synthesis. These errors pose significant challenges for large-scale gene synthesis of long genes due to excessive imperfect assemblies. This dissertation presents innovative methods that leverage CRISPR-Cas9 technologies for targeted retrieval of perfect gene assemblies, aiming to increase their length and quality.

A major contribution of this work is the development of Barcode Assisted Retrieval-CRISPR Activated Targeting (BAR-CAT), a method that uses deactivated Cas9 (dCas9) to selectively enrich perfect synthetic genes from complex libraries. By tagging genes with unique DNA barcodes and targeting these barcodes with *in vitro* transcribed sgRNAs, we successfully enriched three targeted barcodes by up to 1,094-fold. However, BAR-CAT scalability was limited beyond 12 targeted barcodes due to challenges with excessive library diversity and competition among sgRNAs for dCas9 binding, which will require further method optimization.

Parallel to the development of BAR-CAT, I led the development of a scalable sgRNA synthesis workflow that reduces costs by over 70% by harnessing large pools of microarray-derived oligos. These oligos are assembled into dsDNA templates and *in vitro* transcribed to generate sgRNA libraries. Despite optimizations, RNA-seq analysis revealed biases in spacer representation driven by guanine-rich sequences near the T7 promoter. We mitigated these biases by padding sgRNA spacers with a guanine tetramer and evaluating alternative approaches, such as compartmentalization in emulsions. These strategies improved sgRNA library uniformity, with broad implications for CRISPR-Cas9 screens and sgRNA design.

Taken together, these advances in targeted DNA enrichment and sgRNA library production are advances that will contribute to improving the scalability, affordability, and fidelity of synthetic gene libraries. BAR-CAT, in particular, offers a promising approach for multiplexed retrieval of perfect genes that could benefit applications in synthetic biology, ancient DNA

analysis, diagnostics, and targeted sequencing. Future refinements to both BAR-CAT and sgRNA synthesis methods will further extend their utility, enabling high-throughput exploration of protein function and genome biology. This dissertation includes unpublished co-authored material.

PUBLICATIONS:

Villegas,N.K., Gaudreault,Y.R., Keller,A., Kearns,P., Stapleton,J.A. and Plesa,C. (2025) Optimizing *in vitro* transcribed CRISPR-Cas9 single-guide RNA libraries for improved uniformity and affordability. bioRxiv, 10.1101/2025.03.24.644170.

Villegas,N.K., Tran,M.H., Keller,A., and Plesa,C. (2025) Barcode-Assisted Retrieval-CRISPR Activated Targeting (BAR-CAT) is Method for Enriching Synthetic Genes. In preparation.

ACKNOWLEDGMENTS

I give special thanks to the University of Oregon Biology Department, Institute of Molecular Biology, and the Phil and Penny Knight Campus for Accelerating Scientific Impact for providing me with academic and institutional support. Additionally, this work was supported by the National Science Foundation MCB-2032259 grant. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM149387 (to Natanya Kaitlin Villegas). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Special thanks to Yukiko R. Gaudreault for assisting with some of the experimental work in chapter 3 of this dissertation.

DEDICATION

This dissertation is dedicated to my grandparents, who would be very proud of my accomplishments. My maternal grandparents, Walter and Dora Araujo, immigrated from Uruguay in the 1970s and instilled in their children and grandchildren a deep appreciation for education. In addition to my parents, they contributed to my homeschool education by teaching me foundational subjects, helping me learn to read in Spanish, and showing me the importance of kindness. They were my best friends, and I know they would have loved to read this dissertation. My paternal grandparents, Robert and Carmela Villegas, would also be proud of me. Although I never had the chance to meet Carmela, as she passed away before I was born, I believe she too would have shared in this pride. I hope this work honors them all.

TABLE OF CONTENTS

Chapter Page

1. Introduction 15
1.1 The Essential and Ever-Evolving Roles of DNA Libraries in Molecular Biology 15
1.2 Synthetic Oligonucleotides are Vehicles of Programmability with Constraints on Accuracy and Length 19
1.3 Gene Synthesis Enables Large-Scale Synthetic Biology but Remains Limited by Errors in Oligo Inputs 26
1.4 Programmable sgRNA Libraries Drive Large-Scale CRISPR-Cas9 Applications 38
1.5 Expanding CRISPR-Cas9 Targeted DNA Enrichment from Diagnostics to Multiplexed Synthetic Gene Enrichment 51
1.6 References 57
2. Barcode-Assisted Retrieval-CRISPR Activated Targeting (BAR-CAT) is a Method for Enriching Synthetic Genes 72
2.1 Author contributions 72
2.2 Introduction 72
2.3 Materials and Methods 79
2.4 Results and Discussion 101
2.4.1 Proof-of-Concept Enrichment of 18 Barcodes from a Barcoded *mcherry* Library (BAR-CAT v0.1) 101
2.4.2 Bead Wash Optimization and the Development of BAR-CAT v0.2 108
2.4.3 Approaches to Denature dCas9 and Reduce Off-Target Barcodes for the Development of BAR-CAT v0.3 110
2.4.4 Assessing DNA Input Amount and Incubation Time on Enrichment Performance to Develop BAR-CAT v1.0 113
2.4.5 Performance of BAR-CAT v1.0 in the Scale-Up of Target Enrichment from DropSynth Libraries 121
2.4.6. Evaluating sgRNA Performance, sgRNA Library Bias, and Barcode Target Selection to Improve Future Large-Scale BAR-CAT Studies 129
2.5 Conclusions 136
2.6 Conflicts of Interest 138
2.7 Bridge 138
2.8 References 138

3. Optimizing *in vitro* Transcribed CRISPR-Cas9 Single-Guide RNA Libraries for Improved Uniformity and Affordability 146
3.1 Author contributions 146
3.2 Introduction 147
3.3 Materials and Methods 152
3.4 Results and Discussion 164
3.4.1 Feasibility and Quality Assessment of sgRNA Library Synthesis 164
3.4.2 Reducing PCR Cycles in Microarray Oligo Amplification Fails to Improve sgRNA Library Uniformity 170
3.4.3 T7 RNA Polymerase Favors Spacers Starting with Four Guanines 176
3.4.4 Reducing sgRNA Spacer Bias by Adding 5′ Guanine Tetramers 179
3.4.5 Effect of Oligo Type on DNA Library Quality for IVT 183
3.4.6 Emulsion IVT (eIVT) Enhances sgRNA Library Uniformity 185
3.4.7 Evaluation of High Molecular Weight Byproducts in sgRNA Libraries 194
3.4.8 Future Directions for Optimizing eIVT 197
3.4.9 Effects of IVT Reaction Conditions on sgRNA Library Uniformity 198
3.5 Conclusions 202
3.6 Conflicts of Interest 205
3.7 Data and Materials Availability 205
3.8 References 205
4. General Conclusions 211
APPENDIX A. Supplementary Material for Chapter 2 214
A1. Figures 214
A2. Tables 222
A3. References 222
APPENDIX B. Supplementary Material for Chapter 3 223
B1. Figures 223
B2. Tables 235
B3. References 241

LIST OF FIGURES

Figure Page

1. Figure 1. Role of DNA libraries, from genomic to synthetic, in enabling large-scale functional screens and machine learning models. 16

2. Figure 2. Phosphoramidite Oligo Synthesis (POS) is a four-step process commonly used to synthesize oligonucleotides. 21

3. Figure 3. DropSynth enables multiplexed gene assembly via polymerase cycling assembly (PCA) but suffers from reduced fidelity due to oligo errors. 35

4. Figure 4. Mechanistic overview of CRISPR-Cas9 activation, from its inactive (apo) state to the formation of the Cas9:single-guide RNA (sgRNA) complex and DNA cleavage. 42

5. Figure 5. sgRNA format matters: From CRISPR screens to *in vitro* CRISPR assays 45

6. Figure 6. Using CRISPR-Cas9 and dCas9 to enrich targeted DNA sequences. 53

7. Figure 7. Overall workflow for BAR-CAT proof-of-concept and initial optimizations. 104

8. Figure 8. Effects of DNA input amount and incubation time on BAR-CAT enrichment. 118

9. Figure 9. BAR-CAT v1.0 performance declines with increasing enrichment scale and is further reduced by sgRNA spacers starting with 5′ guanine tetramers (5′ GGGG) compared to single 5′ guanine (5′ G) designs. 125

10. Figure 10. Experimental workflow and proof-of-concept for sgRNA library synthesis. 165

11. Figure 11. Comparison of sgRNA library metrics of two scales (389 and 1,382) with differing oligo sources and PCR cycles. 172

12. Figure 12. Influence of base composition on spacer abundance for 389-plex sgRNA libraries. 178

13. Figure 13. Evaluating *in vitro* transcription in emulsions (eIVT) as a novel approach to synthesize sgRNA libraries. 189

14. Figure 14. Gini Coefficients of sgRNA libraries across different scales, DNA input amounts, and reaction volumes. 199

15. Figure 15. Wash stringency of various buffers at different temperatures when used to wash streptavidin-coated magnetic beads. 214

16. Figure 16. Evaluating residual DNA after bead washing by performing agarose gel electrophoresis of qPCR products..... 215

17. Figure 17. Barcode distributions before and after enrichment of 18 target barcodes for various bead wash conditions. 215

18. Figure 18. Barcode distributions before and after enrichment of 18 targeted barcodes for various DNA format and dCas9 denaturation conditions. 216

19. Figure 19. Total barcode dropout counts are compared across three dCas9 denaturation methods: urea, proteinase K, and heat (boiling). 217

20. Figure 20. Log2 enrichment values for the 18 individual barcodes targeted from the *rfp* supercoiled library treated with proteinase K. 217

21. Figure 21. Barcode distributions before and after enrichment of 18 target barcodes for various DNA input amounts. 217

22. Figure 22. Change in median $\log_2$ enrichment scores as a function of sgRNA age for chemically synthesized and *in vitro* transcribed (IVT) sgRNAs used in BAR-CAT enrichment experiments. 218

23. Figure 23. sgRNA spacer selection pipeline. 218

24. Figure 24. Barcode distributions before and after singleplex and 12-plex enrichment of a DropSynth DHFR library with varying 5′ guanine additions in sgRNA spacers. 219

25. Figure 25. Barcode distributions before and after 60-, 389-, or 1,384-plex enrichment from DropSynth DHFR libraries with varying 5′ guanine additions in sgRNA spacers. 219

26. Figure 26. Correlation analysis of $\log_2$ enrichment values for targeted barcodes versus predicted sgRNA performance from the CRISPRscan algorithm (3). 220

27. Figure 27. Distribution of $\log_2$ enrichment values for targeted barcodes relative to the abundance of their corresponding spacers within transcribed sgRNA libraries. 220

28. Figure 28. Assessing the impact of varying DNA, sgRNAs, and dCas9 input amounts on the enrichment of 12 targets from the 384-gene DHFR library (library S4). 221

29. Figure 29. Barcode distributions before and after 12-plex enrichment of a 384-gene DHFR library (library S4) with varying amounts of input DNA, sgRNAs, and dCas9. 222

30. Figure 30. Trends in quality metrics for 10 microarray-derived sgRNA libraries. 223

31. Figure 31. Distribution of 18 spacers transcribed as a single sgRNA library, shown by the normalized fraction of reads for perfect sequences (y-axis, log scale). 224

32. Figure 32. Comparison of percent perfect spacer sequences across sgRNA libraries. 224

33. Figure 33. Pearson correlations of spacer reads for oPool- and microarray-derived sgRNA libraries prepared with varying PCR cycle numbers. 225

34. Figure 34. Effect of PCR cycle number on the fraction of mutant and target spacers in 389- and 1,382-plex microarray-derived sgRNA libraries. 226

35. Figure 35. Influence of base composition on spacer abundance within 389- and 1,382-plex sgRNA libraries prepared with differing oligo sources and PCR cycle numbers. 226

36. Figure 36. Distribution of 12 sgRNA spacers starting with a 5' G and 5' GGGG (12G and 12G4 libraries). 227

37. Figure 37. Pearson correlations of spacer reads for sgRNA libraries containing 389 spacers starting with 5' G and 5' GGGG (389G and 389G4 libraries). 227

38. Figure 38. Gini Coefficients of sgRNA libraries containing 389 spacers starting with 5' G and 5' GGGG (389G and 389G4 libraries) transcribed with modified IVT conditions. 228

39. Figure 39. Comparison of sgRNA yields between libraries with 5’ G and 5’ GGGG spacers. 228

40. Figure 40. Comparison of quality metrics for 135 bp DNA libraries containing 389 spacers, prepared from microarray- and oPool-derived oligos 229

41. Figure 41. Pearson correlations of spacer reads between 135 bp DNA libraries containing 389 spacers and their corresponding IVT sgRNA libraries. 229

42. Figure 42. Comparison of incubation times for transcribing a 12-plex sgRNA library in emulsions..... 230

43. Figure 43. Effect of base composition on spacer abundance in 389-plex sgRNA libraries with 5’ G (389G library), transcribed via bulk IVT or emulsion IVT (eIVT). 230

44. Figure 44. Effect of base composition on spacer abundance in 2,626-plex sgRNA libraries with 5’ G, transcribed via bulk IVT or emulsion IVT (eIVT). 231

45. Figure 45. Spacer representation changes in 389-plex sgRNA libraries with 5’ G, transcribed via eIVT with varying DNA input amounts. 232

46. Figure 46. Spacer representation changes in 2,626-plex sgRNA libraries with 5’ G, transcribed via eIVT with varying DNA input amounts. 232

47. Figure 47. Assessment of single-stranded sgRNA (100 nt) products and high molecular weight (200 bp+) RNA byproducts between 5' G and 5' GGGG sgRNA libraries transcribed in emulsions or bulk. 233

48. Figure 48. Spacer representation changes in 2,626-plex sgRNA libraries with 5’ G transcribed with 100 ng or 400 ng input DNA in a 20 µL IVT reaction volume. 234

LIST OF TABLES

Table Page

1. Table 1. Vendor-reported error rates for commercially available oligonucleotides (May 2025). 23

2. Table 2. Compositions of wash buffers tested for streptavidin bead washing stringency. 222

3. Table 3. Pearson correlations among various sgRNA library metrics for 10 microarray-derived sgRNA libraries. 235

4. Table 4. Kolmogorov–Smirnov (KS) statistical analysis of spacer distribution changes between eIVT and bulk IVT sgRNA libraries. 235

5. Table 5. Kolmogorov–Smirnov (KS) analysis of spacer distribution changes in sgRNA libraries transcribed with bulk IVT using 1 ng, 10 ng, or 100 ng input DNA. 236

6. Table 6. The 18-target single-stranded DNA oligonucleotide sequences for an 18-plex sgRNA target library. 237

7. Table 7. Primer pairs and conditions for subpooling and bulk-amplifying ten target oligo ssDNA libraries of various scales from microarray-synthesized chip9 oligonucleotides. 238

8. Table 8. Additional nucleic acid sequences that were not used for subpooling amplification or 5’ RACE in this study. 239

9. Table 9. Primer pairs and conditions are provided for subpooling four 98 nt oligo libraries of varying sizes, with either one or three guanines downstream of the T7 promoter. 240

10. Table 10. Sequences used for the 5' RACE protocol in RNA-seq of sgRNA libraries, including RT primers, template-switching oligos, and gene-specific primers. 240

1. Introduction

1.1 The Essential and Ever-Evolving Roles of DNA Libraries in Molecular Biology

DNA and cDNA libraries have historically played foundational roles in molecular biology by enabling early gene discovery, physical genome mapping, expression studies, functional analyses, and genome sequencing via shotgun approaches (Fig. 1A) (1, 2). These early libraries were derived from biological samples and reflected the natural genetic diversity of an organism (1). However, with the advent of next-generation sequencing (NGS) technologies, the roles and relevance of these libraries began to shift. Between 2001 and 2021, the cost of sequencing a human genome dropped by seven orders of magnitude, surpassing Moore's Law (Fig. 1A) (3, 4). By 2024, the Illumina NovaSeq X reduced this cost to $200 per genome, compared to one million dollars in 2007 (Illumina). As a result, DNA libraries are no longer central to basic gene discovery in model organisms or humans, since most of these genes have already been identified and characterized (5).

Despite this shift, DNA libraries remain indispensable for exploring gene function in non-model organisms and for advancing the growing field of functional metagenomics (6–8). This field has expanded in parallel with sequencing technologies, uncovering millions of uncharacterized proteins in microbial communities across diverse environments, including the human gut (9), oceans (10), and soil (7). Current research aims to characterize these proteins to discover enzymes with novel properties that are valuable for pharmaceuticals and industrial applications (11).

As traditional roles in molecular biology have diminished, DNA libraries have become central to emerging high-throughput applications such as large-scale lineage tracing (12), CRISPR-based screens (13), gene regulation studies (14), directed enzyme evolution (15), and

massively parallel assays of variant effects (MAVEs) (16). For instance, CRISPR screens facilitate genome-wide exploration of gene function (13), while directed evolution uses libraries to optimize enzymes for industrial or therapeutic use (15).

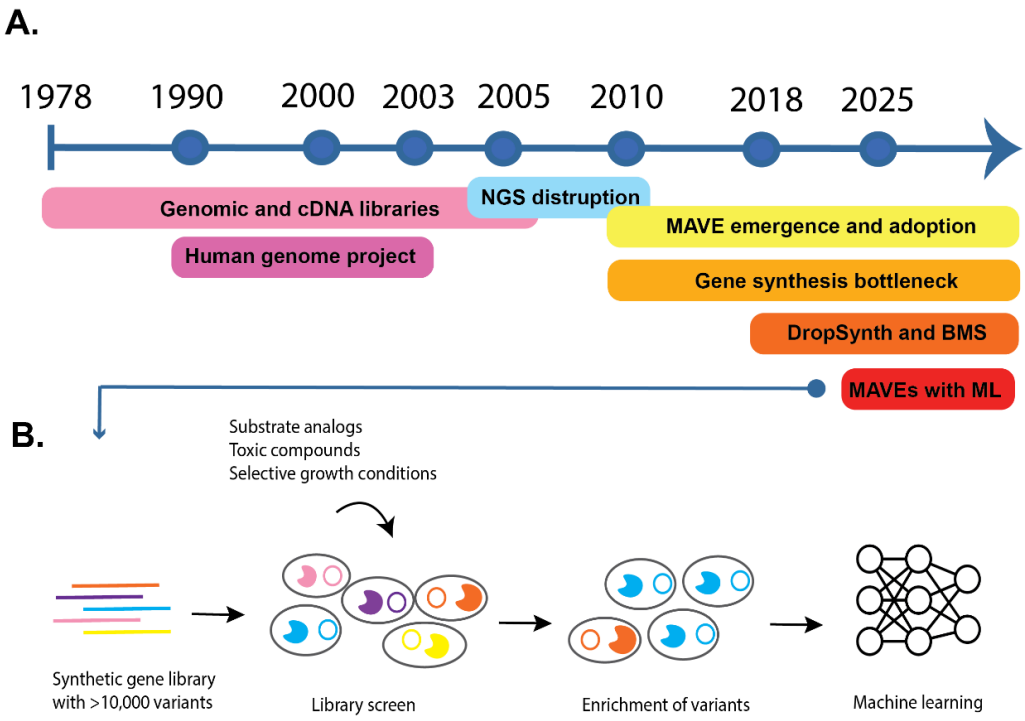


Figure 1. Role of DNA libraries, from genomic to synthetic, in enabling large-scale functional screens and machine learning models.

A. Timeline illustrating the evolution of DNA library technologies and their applications in molecular biology. Early genomic DNA libraries (circa 1979–2006) were foundational to large-scale sequencing efforts, culminating in the Human Genome Project, which enabled the assembly of reference genomes for many model organisms. The next-generation sequencing (NGS) era (2005–2010) significantly reduced sequencing costs and drove adoption of high-throughput methods. Around 2010, gene synthesis emerged as a key bottleneck in functional genomics, prompting development of strategies such as metagenomic functional screening and multiplexed assays of variant effect (MAVEs, 2010–present). The introduction of cost-effective gene synthesis approaches like DropSynth, along with broad mutational scanning (BMS, 2018), allowed scalable construction of synthetic libraries. More recently, advances in machine learning (2025–future) are driving a shift toward predictive modeling and in silico design of DNA libraries for protein engineering and functional assays. **B.** Example of how synthetic gene libraries with >10,000 variants (e.g., DropSynth-based) can support multiplexed assays of variant effects (MAVEs). These libraries, when expressed in cells and functionally screened, generate large datasets on variant fitness, which can be used to train machine learning (ML) models for predicting sequence–function relationships or guiding protein engineering efforts.

These technologies, particularly directed evolution and MAVEs, depend on libraries that encode highly tailored sequence diversity (17). MAVEs encompass a wide range of functional

screens that link DNA sequence variants to phenotypes or sequencing readouts (Fig. 1A). One subtype, the massively parallel reporter assay (MPRA), evaluates thousands of variants involved in gene regulation and other biological processes. These variants are often screened using phenotypic indicators such as cell viability, fluorescence, or antibiotic resistance, and can be used to train machine learning models (18, 19) (Fig. 1B). Another important MAVE approach, deep mutational scanning (DMS), evaluates the effects of thousands of amino acid substitutions in protein-coding genes, offering insight into protein function and the potential impact of variants on human disease (20–22).

Library construction methods such as error-prone PCR (23) and saturation mutagenesis (24), which introduce variation into natural sequences, are commonly used to generate sequence diversity. However, these approaches often generate mutations that are close to the parental sequence and result in limited diversity, a challenge that remains even as various strategies have been explored to mitigate it (25, 26). Although mutagenesis-based approaches are widely used, they remain imperfect in terms of control over sequence content and diversity.

Broad mutational scanning (BMS), a distinct category of MAVE, incorporates natural variation across entire protein families to provide a more global view of protein function and evolution. Unlike approaches that rely on mutating a single gene, BMS requires the assembly of gene homologs across many organisms, including unculturable organisms identified only through metagenomic data (27, 28). This depends on programmable gene libraries synthesized from designed oligonucleotides, enabling precise control over sequence content and expanding the accessible sequence space beyond what occurs naturally. Programmable synthetic sequences can also be codon-optimized to enhance gene expression in model organisms such as *Escherichia coli* (29). As applications like BMS and directed evolution grow in scale and

complexity and begin to inform machine learning models for predictive design, gene synthesis will become an even greater bottleneck for building libraries that span thousands of variants or orthologs (Figs. 1,2). High-quality synthesis is especially important when exact sequence identity is necessary to test biological hypotheses, such as in cross-species comparisons of protein function. The ability to design and assemble genes at scale represents a major shift in how researchers use DNA as a tool for discovery (4).

DropSynth is a promising method for scalable gene synthesis and programmable library construction (27, 30). However, like all gene synthesis technologies, its fidelity is limited by errors present in the synthetic oligonucleotides used for assembly. Improving the fidelity and scalability of synthetic gene assembly remains an ongoing challenge in the field.

To address this, my dissertation presents CRISPR-based technologies that expand the synthetic biology toolkit. These methods focus on increasing the length and quality of synthetic genes to advance the large-scale study of gene homologs, variants, and the proteins they encode, with the goal of predicting and engineering protein functions. In particular, I explore how high-quality single-guide RNA (sgRNA) libraries can support both existing CRISPR applications and new approaches to error correction in synthetic gene libraries. This work is organized around two main aims. Aim 1 develops a programmable CRISPR-based method to selectively enrich perfect synthetic gene assemblies from complex mixtures, while Aim 2 establishes a scalable, low-cost strategy for producing large, customizable sgRNA libraries (Fig. 5C, see 1.4 Programmable sgRNA Libraries Drive Large-Scale CRISPR-Cas9 Applications). Together, these advances are designed to support more accurate, scalable, and accessible tools for understanding and engineering biological systems.

1.2 Synthetic Oligonucleotides are Vehicles of Programmability with Constraints on Accuracy and Length

Synthetic oligonucleotides are fundamental to programming biological systems, enabling precise control over genetic sequences. This programmability drives advances in synthetic biology, gene synthesis, genome engineering, and molecular diagnostics. Yet, this programmability remains constrained by limitations in synthesis accuracy, maximum length, and economic cost.

The discovery of DNA's double-helix and base-pairing structure by Watson and Crick, with critical contributions from Rosalind Franklin, sparked intense interest in oligonucleotide synthesis (31). Just two years later, the first thymidine dinucleotide was synthesized in solution via phosphodiester chemistry (32). By the late 1950s, alternative coupling strategies such as H-phosphonate (33) and phosphodiester activation (34) had been developed (4).

Despite early proof-of-concept studies, oligonucleotide synthesis remained inefficient for years. A major advance came in the 1960s with the transition from solution-based to solid-phase synthesis, which significantly improved yields (35, 36). However, early solid-phase methods still suffered from low efficiency and high levels of side products, partly due to solid supports that were both reactive and prone to swelling, leading to reagent absorption and increased side reactions (37). A breakthrough came in 1981 when Beaucage and Caruthers introduced phosphoramidite oligonucleotide synthesis (POS), which employed nucleoside phosphoramidites to reduce the reactivity of intermediates and the phosphate backbone, thereby increasing yields and minimizing side reactions (38). POS also introduced a non-swelling silica gel solid support, with controlled pore glass (CPG) introduced later for its superior mechanical strength and uniformity (37–39).

POS follows a four-step cycle per nucleotide addition (Fig. 2). First, acid removes the dimethoxytrityl (DMT) protecting group to expose the 5’ hydroxyl. A DMT-protected phosphoramidite nucleotide is then added and reacts with the free 5’ hydroxyl. Unreacted 5’ ends may be capped to prevent deletions, and phosphite linkages are oxidized to form stable phosphates. This cycle repeats iteratively for each base addition, culminating in cleavage of the completed oligo from the solid support (4) (Fig. 2). Initial reports estimated 95% coupling efficiency per addition (39), with later improvements achieving 99.5%–99.8% (40). For example, at 99% efficiency, a 30-mer oligo would yield approximately 74% full-length product.

The development of PCR further underscored the importance of early efforts to synthesize DNA. By 1988, thermostable Taq polymerase enabled its widespread adoption as a powerful molecular biology tool (41). Since PCR requires short, synthetic oligonucleotide primers, the efficiency and accessibility of phosphoramidite DNA synthesis played a crucial role in successfully developing PCR (42). Recognizing the commercial potential of oligo synthesis, the first automated DNA synthesizers were developed, allowing researchers to produce oligonucleotides at the push of a button. This automation further drove the adoption of PCR and other DNA-based technologies (43).

While POS on column-based solid supports became the gold standard for producing robust nmol to mmol yields of oligonucleotides (44), its low throughput, typically yielding only 96 to 384 oligos per run (4), and high cost per oligo became a bottleneck as demand grew for custom DNA oligos in gene assembly and genomic studies (45). This demand was driven by emerging applications such as hybridization-based sequencing and large-scale gene expression profiling. In response, microarray-based synthesis platforms were developed, beginning with photolithographic masking to control oligo synthesis on glass surfaces (46, 47). Subsequent

innovations, including maskless photolithography (48) and Agilent's inkjet printing technology enabling parallel synthesis of up to 25,000 oligos (49). This dramatically improved the scalability of microarray oligo synthesis and reduced costs by two to four orders of magnitude compared to column-based oligo synthesis (4).

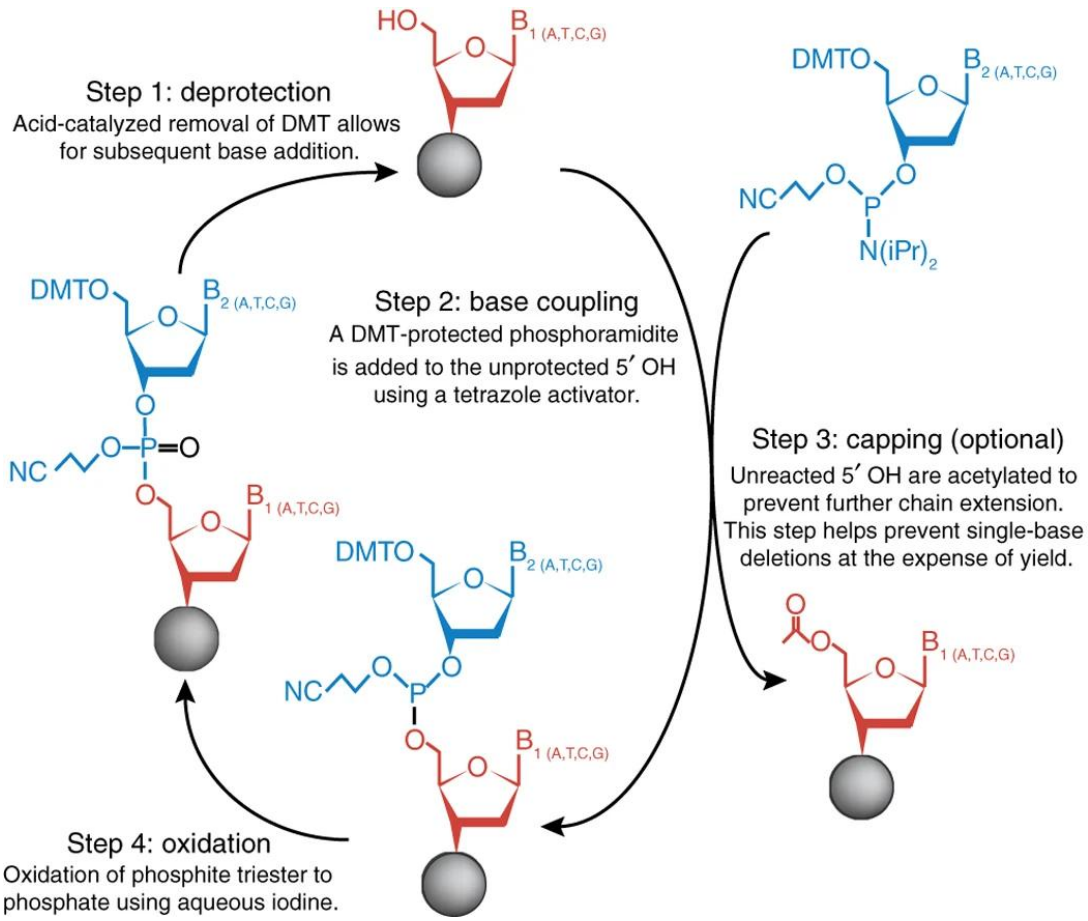


Figure 2. Phosphoramidite Oligo Synthesis (POS) is a four-step process commonly used to synthesize oligonucleotides.

The first step is deprotection and involves removing the Dimethoxytrityl (DMT) protecting group from a nucleotide on a solid support. Step 2 involves the addition of a phosphoramidite base with a DMT protecting group. Step 3 is an optional but recommended capping step that involves eliminating unreacted 5' OH groups that may participate in side reactions. Step 4 involves oxidizing the phosphite triester backbone to phosphate (4) (Adapted from Kosuri and Church, 2014. Reproduced with permission from Springer Nature).

As researchers sought to synthesize longer oligos for gene assembly, following assembly of the first synthetic gene in 1970 (50), the decrease in POS efficiency with increasing oligo length became a major constraint. Oligo length limitations are due to accumulating synthesis

errors and declining coupling efficiency (4). Specifically, harsh deprotection conditions can cause depurination of adenine and guanine residues, resulting in strand cleavage and loss of yield, particularly in oligos longer than ~100 nt (4, 40, 44).

Ciccarelli and colleagues pushed the limits of maximum oligo lengths by attempting to produce both strands of 393 bp and 655 bp genes using an automated oligo synthesizer in 1991 (51). Both full-length and truncated strands were converted into double-stranded DNA via single-primer reverse extension from a conserved 3' end. These were then hybridized, PCR-amplified, cloned, and sequence-verified. While full-length genes were successfully assembled, it was hindered by high error rates and extremely low yields, underscoring the difficulty of synthesizing long, accurate oligos (51).

In microarray-based synthesis, this issue was exacerbated by solvent evaporation, which concentrated the acid and further degraded oligo quality (44). To mitigate these effects, LeProust and colleagues introduced an oxidation step following deprotection in 2001, enabling the synthesis of longer, higher quality oligos up to 150 nt (44). Today, commercial providers have pushed these limits further. Agilent can synthesize pooled microarray-synthesized oligos up to 220 nt, and Twist Bioscience up to 300 nt (Table 1). Although longer oligos have been synthesized using specialized column-based methods, scalable and cost-effective microarray-based production of oligos at this length remains out of reach.

Another factor affecting oligo length was the POS error rate of 1 in 200 nt, primarily due to inefficiencies in coupling and deprotection (4). Most errors are deletions, while substitutions are also present but less common (52). For researchers and companies aiming to purchase as many pooled oligos as possible at the lowest price, Twist Bioscience can provide up to 240,000 oligos, each up to 300 nt in length per pool, at USD $0.17 per base. While the reported error rate

for these oligos has improved in recent years to 1 error in 2,000 to 3,000 nt (0.3–0.5 errors per kbp; Table 1), these pools contain mixtures of oligos with and without errors due to POS. Other leading vendors, such as Agilent, produce oligos with 1 error in 4,000 nt (0.25 errors/kbp), while Eurofins and Integrated DNA Technologies (IDT) demonstrate similar error rates: Eurofins at 1 error in 1,430 to 2,000 nt (0.7–0.5 errors/kbp) and IDT at 1 error in 2,000 nt (0.5 errors/kbp). However, IDT can produce much longer oligos (350 nt) compared to Eurofins (200 nt) (Table 1). Errors within synthesized oligos represent lapses in programmability—instances where the final product diverges from the user-defined sequence—creating a bottleneck in workflows such as gene synthesis, where fidelity is paramount. This is because errors in oligos are retained in the genes they are used to assemble.

Table 1. Vendor-reported error rates for commercially available oligonucleotides (May 2025).

Vendor	Average errors per kbp	Maximum oligo length (nt)
Twist Bioscience	0.3-0.5	300
Agilent Technologies	0.25	230
Eurofins	0.5-0.7	200
Integrated DNA Technologies	0.5	350

Efforts to maximize the programmability of oligo synthesis include error correction and selection steps intended to selectively retain oligos with perfect sequences. Traditional approaches to oligo quality control include polyacrylamide gel electrophoresis (PAGE), which allows size-selection of full-length products to eliminate truncated oligos (53). Several years later, pyrosequencing was used to identify perfect sequences, with a robotic system isolating validated oligos directly from the sequencing platform, achieving a 500-fold enrichment (54).

While effective, this method relied on the now-discontinued Roche 454 sequencer, which was eventually replaced by more accurate and cost-effective Illumina platforms (55).

To address these limitations, Schwartz and colleagues developed dial-out PCR, which attaches unique adaptors to oligo pools, sequences them on Illumina platforms, and selectively amplifies perfect oligos without the need for robotics (56). However, dial-out PCR lacks the ability to recover multiple desired sequences simultaneously, making it low-throughput and cost-prohibitive for highly scalable applications (4, 43). More recently, sequencing-by-synthesis (SBS) has enabled the synthesis and selection of high-quality oligos by hybridizing them to universal primers and extending until a specific length is reached, at which point a biotinylated reversible terminator is incorporated for selective capture of non-truncated sequences (57). Still, SBS extension reactions often fail to synchronize across all molecules, causing some correct oligos to drop out during selection (58).

Despite oligo length limitations, DNA libraries consisting of single oligos can be utilized in a variety of applications. One such example is their use in addressing long-standing questions in functional genomics, such as the relationship between regulatory elements and the genes they influence. The overexpression of ~66,000 peptide fragments tiling 65 cancer driver proteins in human cells, derived from a microarray-based oligo library, led to the identification of peptides that interfere with protein-protein interactions, providing potential cancer therapeutic strategies (14). In another study, an oligo library was transcribed into 22,000 unique RNA baits that captured over 15,000 protein-coding exons for targeted massively parallel sequencing of exonic sequences from genomic DNA (59). Short-insert oligo libraries have also been crucial in CRISPR research. For instance, 98,599 sgRNAs, generated from microarray-derived oligos, enabled the identification of enhancers that regulate the MYC and GATA1 transcription factors

(13). These studies underscore how relatively short, imperfect oligos can yield transformative biological insights when applied at scale.

In Chapter 3 of this dissertation, I discuss an optimized workflow for producing CRISPR-Cas9 sgRNA libraries with improved uniformity, achieving over a 70% cost reduction by bulk-ordering a single pool of microarray-derived oligos and amplifying them into separate libraries (60). These examples underscore the potential of synthetic oligos, even within their current length limitations, to drive significant advances in genomic research.

However, researchers that require DNA sequences longer than 300 nt must rely on mutagenized natural sequences or gene synthesis methods as workarounds for producing DNA assemblies that can be cloned into plasmids to generate long-insert DNA libraries. Gene synthesis methods rely on DNA assembly strategies and like for oligos, methods for selecting gene assemblies that don't contain errors from oligo POS. These methods and applications will be discussed in further depth in the next section of this introduction.

One additional limitation of POS is its reliance on hazardous solvents, which are expensive and difficult to dispose of due to environmental concerns (43). Enzymatic oligonucleotide synthesis offers a more sustainable alternative and the potential for longer oligos. Template-independent enzymatic oligonucleotide synthesis (TiEOS), which shares conceptual similarities with POS, has emerged as a promising approach. Terminal deoxynucleotidyl transferase (TdT), an enzyme used in TiEOS, adds nucleotides to the 3' end of DNA, though its random incorporation can be a challenge for controlled synthesis. Recent advancements have improved TdT's efficiency, such as the use of linker sequences to control nucleotide addition, achieving a 97.7% incorporation efficiency (61). Thermostable TdT variants and directed evolution have further improved its performance, allowing for faster incorporation, reduced

sequence bias, and the use of mild deblocking conditions (62, 63). Despite these advances, challenges remain, including the formation of secondary structures as oligos grow, and data on the maximum oligo length achievable with TdT synthesis are still limited (62).

Despite significant advancements in phosphoramidite oligo synthesis and high-throughput, miniaturized microarray methods, key limitations persist. Challenges in accuracy, oligo length, and synthesis errors still impede the full programmability of synthetic DNA, particularly for gene synthesis and multiplex functional assays, where errors can affect downstream functionality. Overcoming these obstacles through improved error correction and selection strategies is crucial for advancing synthetic biology applications.

1.3 Gene Synthesis Enables Large-Scale Synthetic Biology but Remains Limited by Errors in Oligo Inputs

From the construction of synthetic genomes to high-throughput protein engineering and multiplexed assays of variant effects (MAVEs), gene synthesis plays a central role in modern biology. These applications often require the assembly of long, precisely defined DNA sequences that far exceed the length of individual oligos. Although improvements in oligo synthesis and DNA assembly have expanded the scope of gene synthesis (4, 43), errors introduced during oligo synthesis continue to limit the accuracy and scalability of the process. Overcoming this challenge requires methods that can selectively enrich perfect full-length sequences, allowing researchers to generate high-quality synthetic genes at scale.

Innovations in gene synthesis date back to 1970, when the first synthetic gene, a 77 nt yeast alanine tRNA, was assembled (50). Since oligo synthesis was inefficient prior to phosphoramidite chemistry (38), the 17 oligo sequences, each 20–30 nt long, took five years to synthesize (50, 64). To assemble the gene, these oligos were phosphorylated and ligated using

T4 polynucleotide ligase, a process the authors later described as "welding" DNA oligos. Because PCR had not yet been invented, the gene couldn't be amplified, which complicated further analyses (50).

The emergence of PCR revolutionized the life sciences, significantly reducing the cost and time required for gene synthesis (64). Importantly, PCR enabled Polymerase Cycling Assembly (PCA), a DNA-hybridization-based approach for gene synthesis (45). While PCA shares conceptual similarities with DNA shuffling, a technique that recombines gene fragments to enhance sequence diversity for protein engineering (65, 66), it serves a distinct purpose, to construct synthetic genes from oligonucleotides rather than reshuffling existing natural sequences.

The first reported application of PCA involved the assembly of a 2.5 kb plasmid from 40 nt gel-purified synthetic oligos with 5-25 nt complementary regions (45). As thermocycling progresses, these overlapping regions anneal and are extended by a high-fidelity DNA polymerase, resulting in the formation of full-length double-stranded DNA. The assembled product is then selectively amplified using PCR primers that anneal to the 3' and 5' ends of the assembled gene (Fig. 3A) (45). This use of PCR was revolutionary, enabling the selective amplification of DNA assemblies from the reaction mixture, thereby increasing their abundance for successful cloning and downstream applications (64).

The achievements of emerging gene synthesis methods, such as PCA, are best demonstrated by the ambitious genome synthesis projects of the early 2000s. PCA was used to completely assemble the 7.5 kb poliovirus genome, the first synthetic viral genome capable of replication (67). This project was driven by the need for alternative methods to develop vaccines without relying on scarce or potentially dangerous viral samples (53, 67), an objective that had

been established by the *in vitro* construction of a ~9.6 kb hepatitis C virus genome from PCR-assembled naturally sourced sequences three years prior (68). A year later, a synthetic 5.4 kb φX174 bacteriophage genome was built in only 14 days, demonstrating a major leap in genome construction (53). Initially, ligase chain reaction (LCR), an assembly method that selects for error-free oligos by only allowing ligation when oligos hybridize to an oligo splint, was applied to assemble the φX174 genome (69)). However, LCR was deemed inefficient because variations in oligo concentrations resulted in uneven and incomplete assembly (53). Additionally, since it is presumed that they used PAGE-purified oligos (70), it is possible that PAGE resulted in variable yields of each oligo that reduced LCR effectiveness. Additionally, PCA was found to produce DNA assemblies with fewer errors compared to LCR (71). To overcome these limitations, the φX174 genome was assembled with PCA following LCR (53, 64).

These studies demonstrated both the feasibility of synthetic genome construction and the limitations of ligation-based methods, which were slow and costly due to their dependence on high-quality oligos and phosphorylation steps (4). In contrast, PCA offered a faster and more cost-effective alternative (4, 53). The φX174 project served as a compelling proof of concept for PCA's utility in genome synthesis. Still, combining LCR and PCA remained common practice for several years, especially in efforts to reduce errors and improve assembly efficiency. This approach was later adopted in gene synthesis from microarray-derived oligos, where both methods were used together to overcome error-prone inputs (72, 73).

PCA and PCR are limited by reduced efficiency when amplifying DNA longer than 5 kb, although high-fidelity or long-range polymerases can extend this limit to approximately 35 kb (4, 74). To overcome these limitations, Gibson and colleagues developed a two-step *in vitro* recombination method that successfully assembled the 538 kb *Mycoplasma genitalium* genome

(75). This work led to a simpler, one-step isothermal technique known as Gibson Assembly, which remains widely used for joining long DNA segments (76). In this approach, T5 exonuclease generates 3’ overhangs, which are then annealed, extended by Phusion high-fidelity polymerase, and ligated by Taq ligase. As a proof of concept, the researchers assembled *M. genitalium* genome quarters, reconstructing up to 318 kb at a time using a hierarchical approach (76, 77). To complete full genome assemblies, Gibson and colleagues typically used *in vivo* recombination in yeast (75, 78). Today, large-scale genome synthesis projects, such as the 19-year effort to build a synthetic yeast genome, rely on Cre-lox and CRISPR-Cas9–mediated recombination in yeast, which offer greater efficiency and precision than earlier methods (79).

Gibson Assembly is well-suited for constructing very long DNA sequences. It does not rely on restriction enzymes, eliminating the need to computationally omit restriction sites from synthetic oligos (80). Unlike PCA, it also avoids PCR amplification, allowing for the assembly of DNA fragments hundreds of kilobases in length (4). However, Gibson Assembly has several limitations. It cannot assemble DNA oligos shorter than 100 nt, making it unsuitable for assembling short oligos into full-length genes—a task typically accomplished using LCR or PCA. Additionally, its efficiency declines as the number of DNA inserts increases; while up to five inserts are generally feasible, efficiency drops beyond that (80). Each DNA fragment must include 20–40 nt homologous overlaps for proper assembly (81), which can lead to cross-hybridization if the sequences are highly similar. This drawback is not unique to Gibson Assembly; it also affects other homology-based methods such as USER cloning, In-Fusion Cloning (Takara Bio), and PCR-based methods like PCA (80).

In contrast, the Golden Gate Assembly (GGA) method uses type IIS restriction enzymes, such as BsaI, which cut at specific recognition sites (81). These cuts generate 4 nucleotide (nt)

overhangs on each DNA fragment, unlike the extensive complementary overhangs required by Gibson Assembly. T4 DNA ligase then ligates the sticky ends, enabling precise, sequential DNA assembly over multiple heating and cooling cycles (82). GGA is advantageous because its use of 256 unique overhangs with characterized ligation efficiencies simplifies design, in contrast to the complex and variable complementary overhangs required for Gibson Assembly (81, 83, 84). However, not all of these overhangs support efficient ligation, and assemblies involving more than 40 fragments often suffer from incomplete assembly and misassembly due to palindromic or inefficient junctions (85). Additionally, GGA requires exclusion of internal Type IIS restriction sites from input oligos, a constraint not shared by Gibson Assembly or PCA. Thus, the practical upper limit of GGA is likely between 50 and 70 fragments. Still, the ability to assemble dozens of fragments in a single step remains a key advantage over the hierarchical strategies required for assembling long DNA sequences with Gibson Assembly.

In the 2010s, gene synthesis became a bottleneck in synthetic biology, hindering the rapid development of synthetic pathways and genomes for programmable cells and microbes (86). By 2010, the synthetic gene market was worth billions, driven by this demand (73). Microarray-derived oligos presented a promising low-cost, high-throughput solution, though early challenges included low concentrations (fmol to pmol range) and higher error rates (70, 73), which have since improved.

There was a Cambrian explosion of gene synthesis methods between 2004 and 2020 as researchers sought to overcome the challenges of assembling genes from microarray-derived oligos. These methods introduced key innovations in synthesis and error-correction workflows, many of which still underpin modern gene synthesis. One early breakthrough by Tian and colleagues in 2004 involved bulk amplification of microarray-derived oligos to overcome their

inherently low yields (70). To reduce synthesis errors, they developed a hybridization-based selection method that enriched for perfect oligos using complementary reference sequences immobilized on streptavidin-coated beads. This technique bypassed labor-intensive approaches like manual cloning and PAGE purification, which only improved error rates from 1 in 160 bp to 1 in 450 bp. In contrast, hybridization selection reduced errors to 1 in 1,394 bp (4, 70). For gene assembly, they applied Single-Step Polymerase Assembly Multiplexing (PAM), a PCR-based method that could assemble all overlapping oligos in a single reaction. PAM enabled the parallel construction of 21 genes—a major milestone at the time—though it was eventually superseded by more scalable techniques like PCA (70, 73).

While early methods addressed yield and error rates, scalability remained a major hurdle. By 2010, gene synthesis from microarray-derived oligos was feasible, but expanding to larger, more complex pools led to oligo cross-hybridization during assembly (4, 27). To overcome this, Kosuri and colleagues leveraged newly available 150–200 nt microarray-derived oligos (87). From a pool of 13,000 oligos, they synthesized 47 genes in parallel, each encoding variable regions of single-chain antibody fragments, which are famously difficult to synthesize (88). Their approach relied on selective amplification using orthogonal primer pairs tailored to each gene. They also optimized their PCA protocol and incorporated error correction with the commercial ErrASE enzyme (88).

Another strategy to reduce oligo cross-hybridization was to compartmentalize gene assembly. One such method combined inkjet oligo synthesis with *in situ* isothermal amplification, sequestering 9 oligos per gene within discrete microarray subarrays. PCA was then performed separately in each subarray, enabling the parallel assembly of up to 30 genes, each 0.5–1 kb in length. For error correction, the assembled genes were denatured and

reannealed to form heteroduplexes, which were then treated with Surveyor nuclease, a mismatch-specific enzyme derived from a plant CEL nuclease (89, 90). Surveyor achieved error rates comparable to those obtained with ErrASE, while the entire compartmentalized workflow reduced costs by an order of magnitude compared to the subpooling strategy used by Kosuri and colleagues (88, 90).

Despite earlier advances, parallel gene assembly remained expensive and limited the scalability of synthetic gene libraries. Multiplexed assembly was needed to overcome this bottleneck (4, 27, 88). Klein and colleagues were the first to attempt it in 2016, assembling 2,271 genes across 10 subpools, with 130–252 sequences per subpool (91). Each gene was split into two fragments and assembled using 160-nt oligos. Subpools were amplified with orthogonal primer pairs containing uracil, which were later removed by the USER enzyme to enable overlap extension PCR. While subpool assembly yielded 90.5% perfect sequences, accuracy dropped to 70.6% in the fully multiplexed reaction. They used Dial-Out PCR to selectively recover correct products using unique sequence tags (56, 91). Though this established the feasibility of one-pot multiplexed assembly, the synthesized genes were short (<300 bp), and chimeric products were common (91).

DropSynth, developed in 2018 by Plesa and colleagues in the Kosuri lab, revolutionized multiplexed gene assembly. It used subpool amplification with orthologous primer sets to construct 384-plex gene libraries from hundreds of 200–230 nt oligos. Unlike earlier methods, DropSynth captured all oligos for each gene on beads with unique DNA barcodes, which were compartmentalized in water-oil emulsions for PCA-based gene assembly (Fig. 3B) (27). This approach reduced cross-hybridization, minimized template competition, and lowered costs

compared to microarray compartmentalization, while also providing modularity by separating oligo synthesis, preparation, and assembly steps (27, 88, 90).

DropSynth has since proven to be the most scalable gene synthesis method to date. In its first iteration (DropSynth 1.0), it assembled 5,775 dihydrofolate reductase (DHFR) homologs across 25 gene libraries, with 384 genes per library. The DHFR libraries, which included two-codon versions for maximized coverage, represented a major improvement in gene length (381–669 bp), a significant step up from the 200 bp genes assembled in previous methods (91). However, these libraries had a median of 3.8% perfect genes. DropSynth was also used to assemble 1,152 phosphopantetheine adenylyltransferase (PPAT) homologs in three libraries, which showed an even lower median of 2.9% perfect genes, likely due to the longer oligos used for PPAT synthesis, which are more error-prone than shorter oligos (27).

Although the PPAT gene libraries were successfully applied for a groundbreaking BMS functional study characterizing the evolution of that enzyme family, the low percent perfects of DropSynth 1.0 dramatically reduced the inherent programmability of DropSynth gene libraries (27). As a result, applying these libraries to multiplex functional assays, such as BMS, would require time-consuming oversampling to identify rare perfects. In 2020, DropSynth 2.0 was released, which increased the scale of multiplex gene assembly from 384 to 1,536 (30).

Crucially, switching to a high-fidelity DNA polymerase for PCA (Kapa HiFi) resulted in a median of 22.6% to 27.6% perfect gene assemblies out of 1,208 DHFR genes with a median length of 501 bp (30, 92). This indicates that a proportion of the errors within genes were due to PCA and PCR errors from the DNA polymerase used. However, the remainder of the errors were likely due to the inclusion of oligos with errors during assembly.

Finally, the development of DropSynth 2.0 included efforts to add an error-correction step by incubating gene assemblies with MutS, a protein involved in the mismatch repair pathway. MutS binds to mismatched base pairs in heteroduplex DNA, ranging from one to four bases, facilitating base excision and subsequent repair by other Mut enzymes (93). Beads and columns functionalized with MutS have been employed to selectively bind and remove error-containing sequences, enriching perfect gene assemblies up to 25.2-fold (4, 94). However, MutS incubation failed to improve the percentage of perfect assemblies (30), likely because the output of assembled DNA was too low for error-correction following DropSynth PCA. Given the current DropSynth 2.0 protocol, MutS treatment can only be done following PCA but prior to final PCR amplification (Fig. 3A) because during PCR, only a single strand of DNA is amplified, disrupting the heteroduplexes needed for MutS recognition (95). Therefore, MutS error correction is incompatible with the current DropSynth 2.0 method (Fig. 3B).

Overall, DropSynth has been shown to assemble genes from five oligos with a maximum length of 1 kbp, with median 8% perfect gene assemblies at the DNA sequence level (Fig. 3C) (92). The percentage of perfect assemblies is expected to decrease for genes longer than 1 kbp due to the progressive accumulation of errors as the length and number of oligos increase. The chance of incorporating oligos with errors rises as more oligos are assembled (92). Ultimately, genes about 2 kbp in length would have less than 1% perfect gene assemblies, which would require large sequencing depth in any downstream multiplex functional screens to capture functional information.

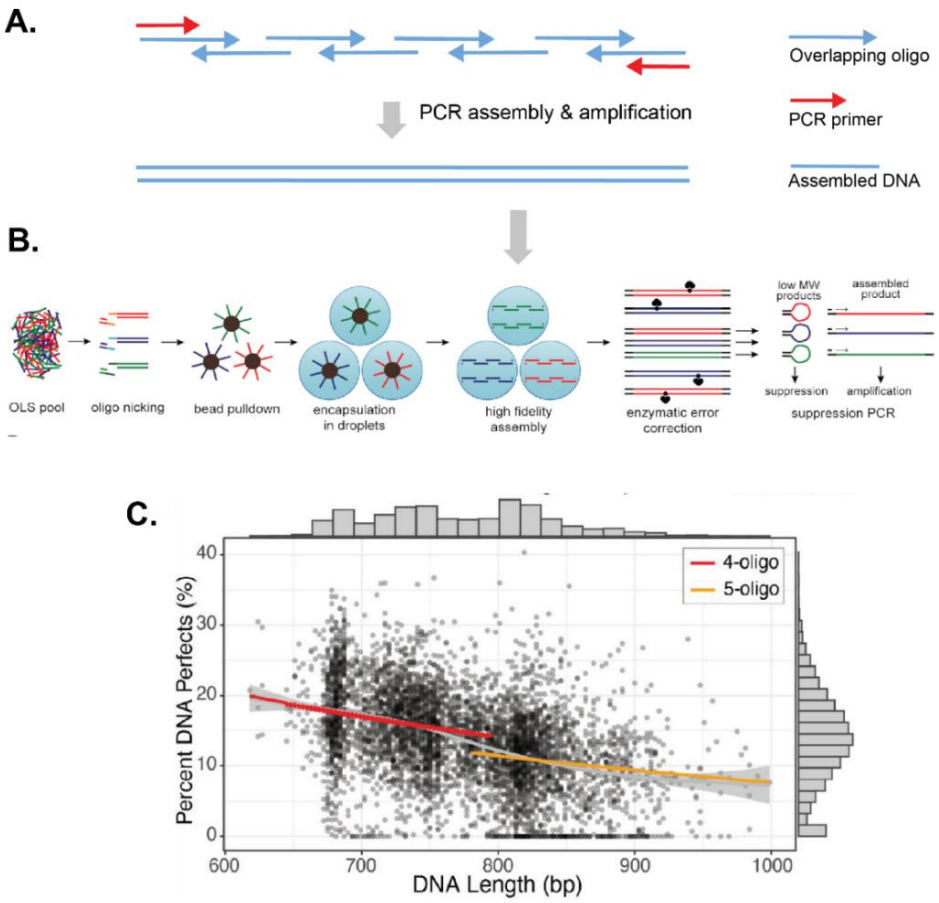


Figure 3. DropSynth enables multiplexed gene assembly via polymerase cycling assembly (PCA) but suffers from reduced fidelity due to oligo errors.

A. Polymerase cycling assembly (PCA) is a gene assembly technique in which chemically synthesized oligonucleotides (blue arrows) overlap and hybridize to form full-length genes through iterative cycles of annealing and extension by a high-fidelity DNA polymerase. PCR primers flanking the terminal oligos allow amplification of the assembled product either during or after the assembly process. Original figure from Tian *et al.* (2009) (103), re-created in vector format and reprinted with permission from the Royal Society of Chemistry due to resolution constraints. **B**. Schematic of DropSynth 2.0, adapted from Sidore *et al.* (2020) (30). As in DropSynth 1.0, a microarray-derived oligonucleotide pool encodes a multiplexed gene library. Barcoded beads, loaded with oligos through complementary overhangs, are emulsified with a high-fidelity polymerase, enabling compartmentalized PCA-based gene assembly (27). DropSynth 2.0 incorporated single-primer suppression PCR to eliminate short, misassembled products via panhandle suppression, enriching for full-length assemblies flanked by inverted terminal repeats. While enzymatic error correction using MutS was tested, it was ultimately excluded due to limited efficacy (30). Image reprinted from Sidore *et al.* (2020) under Creative Commons license (CC BY 4.0). **C**. Percent-perfect gene assemblies for 4- and 5-oligo constructs (600–1,000 bp) generated using Degenerate DropSynth, a variation of DropSynth 2.0 that allows multiple gene variants per droplet—advantageous for protein engineering applications (92). Image reprinted from Holston *et al.* (2023) under Creative Commons (CC BY-NC 4.0).

Errors in synthesized oligos and genes have long been a challenge, and few solutions are compatible with selecting perfect gene assemblies from DropSynth libraries. Traditional methods

like colony picking and Sanger sequencing are accurate but labor-intensive and costly, as they can only validate one gene at a time (4, 67, 96). While advances in nanopore long-read sequencing and bioinformatics now allow multiplex validation of up to six plasmids, these methods still require labor-intensive screening and offer insufficient throughput for efficiently locating perfect assemblies in DropSynth gene libraries (96, 97). Meanwhile, other researchers have used dial-out PCR to selectively retrieve perfect gene assemblies (56, 91). While this method has been used to retrieve small numbers of DNA sequences from DropSynth libraries (28), dial-out PCR lacks multiplexing capability, making it low-throughput and cost-prohibitive for DropSynth applications requiring high scalability (4, 43).

On the other hand, biotinylated RNA probes produced from microarray-derived oligos offer a targeted and scalable enrichment strategy, with straightforward retrieval via streptavidin-coated beads. While this hasn't been applied to enrich perfect genes, similar methods have been used for targeted NGS, such as enriching thousands of exons from sheared genomic DNA fragments to (59) and enriching ancient human DNA from contaminating environmental DNA (98). One major limitation of these strategies includes low accuracy if probes are not designed carefully, as mismatched probes can capture erroneous sequences (99). A less targeted, yet scalable alternative involves functionally selecting perfect gene assemblies that are translationally fused to fluorescent or antibiotic resistance protein sequences (100). Gene assemblies with indels that introduce premature stop codons will not fluoresce or persist following antibiotic selection. However, translation re-initiation at downstream start codons can cause N-terminal truncations, leading to false positives (101). While these methods are scalable, they often sacrifice precision, highlighting the need for more accurate and high-throughput alternatives.

As previously mentioned, error-removal enzymes like MutS have not effectively reduced errors in DropSynth gene assemblies. Another candidate, T7 endonuclease I, a bacteriophage resolvase that cleaves single-base mismatches, has improved error rates from 3.45 to 0.43 errors per kilobase in a synthetic GFP construct, representing an approximately eightfold improvement over untreated controls (102). However, the effectiveness of T7 endonuclease I, as well as other mismatch-cleaving enzymes like Surveyor and ErrASE, has not been evaluated with DropSynth because its yield of assembled DNA is too low to meet the input requirements for these correction enzymes. Furthermore, T7 endonuclease I may not be ideal for error-correction, as it does not cleave all mismatches, as evidenced by the persistence of residual errors following treatment. While site-directed mutagenesis (SDM) offers near-perfect fidelity (102), it is labor-intensive and unsuitable for large-scale applications like DropSynth.

To address the limitations of existing methods for enriching perfect gene assemblies from DropSynth libraries, we aimed to develop a scalable, cost-effective, and programmable solution. This method would enable precise retrieval of perfect assemblies identified by next-generation sequencing (NGS) and support multiplexed enrichment in a single reaction, similar to DropSynth's production process. It must be gene-length independent and avoid reliance on heteroduplex DNA, as a very low amount of assembled DNA comes out of the DropSynth reaction and heteroduplexes are lost during subsequent PCR amplification.

A promising strategy to meet these criteria is CRISPR-based enrichment, which, although widely used in nucleic acid targeting, has not been applied to enrich synthetic gene assemblies. Therefore, Aim 1 of this dissertation is to develop a programmable CRISPR-based method to selectively enrich perfect synthetic gene assemblies from mixtures of perfect and imperfect sequences. Sub-aim 1.1 will focus on producing a proof-of-concept by enriching targeted

barcodes from a low-complexity, single-gene library with hundreds of thousands of unique barcodes. Sub-aim 1.2 will optimize enrichment efficiency to ensure a robust signal before applying it to high-complexity DropSynth gene libraries. In sub-aim 1.3, targeted enrichment of perfect gene assemblies from DropSynth libraries will be demonstrated, and sub-aim 1.4 will evaluate the scalability of this method for enriching large numbers of perfect synthetic genes in a single reaction.

Chapter 2 of this dissertation, Barcode-Assisted Retrieval–CRISPR Activated Targeting (BAR-CAT) is a Method for Enriching Synthetic Genes, will present this project in detail and includes material co-authored with Mindy H. Tran, Abigail Keller, and Dr. Calin Plesa. The remainder of this introduction provides essential background on CRISPR biology, engineering principles, guide design, and library generation to contextualize the work described in this dissertation.

1.4 Programmable sgRNA Libraries Drive Large-Scale CRISPR-Cas9 Applications

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) system, a prokaryotic adaptive immune system, is inherently programmable. Despite this, its full biological function remained unclear for nearly two decades. CRISPR's direct repeat sequences were first identified by Ishino and colleagues in 1987 in *E. coli*. These 29-nucleotide repeats, located at the 3′ end of the gene of interest, were interspersed with unique 32-nucleotide spacer sequences. The spacers, which showed no homology to known sequences at the time, sparked significant interest in their origin and function (104, 105).

The biological function of CRISPR remained a mystery for many years. The advent of next-generation sequencing (NGS) enabled the sequencing of more prokaryotic genomes, revealing that CRISPR elements were exclusive to prokaryotes (105–107). Genes encoding

CRISPR-associated (Cas) proteins, such as Cas1 through Cas4, were frequently found adjacent to CRISPR loci. While some Cas proteins contained nuclease or helicase domains, their biological roles were not yet understood (108). Around the same time, researchers discovered that CRISPR loci were transcribed into small non-messenger RNAs, providing early insights into CRISPR function (109).

A key breakthrough came when spacer sequences within CRISPR arrays were found to match foreign DNA, including sequences from bacteriophages and conjugative plasmids. This observation led to the hypothesis that CRISPR RNAs (crRNAs), in conjunction with Cas proteins, comprise an adaptive immune system in prokaryotes (105, 108, 110, 111). This idea gained experimental support in 2007, when researchers observed that *Streptococcus thermophilus*, a bacterium used in cheese production, incorporated new spacers into its CRISPR array after exposure to bacteriophages (112). These newly acquired spacers conferred immunity to the phages, and this immunity was transferable: bacteria that had not been exposed to phages gained resistance when engineered to carry these spacers (113).

By 2011, researchers had elucidated a general three-stage mechanism for CRISPR immune function: adaptation, expression (also known as crRNA biogenesis), and interference. During the adaptation stage, foreign DNA segments are integrated into CRISPR loci. This process is mediated by Cas1 and Cas2, which are conserved across diverse CRISPR systems (113, 114). When a virus or other mobile genetic element infects a bacterial cell, Cas1 and Cas2 recognize the foreign DNA based on the presence of a short, species-specific nucleotide sequence known as a protospacer adjacent motif (PAM). The DNA region adjacent to the PAM, called the protospacer, is excised and inserted into the CRISPR locus alongside a repeat sequence (115). The second stage, expression, involves transcription of the CRISPR array into a

long precursor CRISPR RNA (pre-crRNA), which is then processed into individual crRNAs, each containing a single spacer (116). This processing can be carried out either by Cas proteins or by cellular ribonucleases, depending on the type of CRISPR system (112). Finally, during interference, the mature crRNAs guide Cas effector proteins to complementary sequences within invading DNA or RNA. The crRNA-Cas complex binds the foreign target and cleaves it, thereby neutralizing the threat (112).

CRISPR-Cas systems were beginning to be classified according to their expression and interference mechanisms, with some prokaryotic species containing multiple CRISPR types. These early classifications laid the foundation for gene editing and other biotechnological applications. CRISPR systems are divided into two broad classes. Class II systems carry out interference with single effector proteins and include the widely recognized Type II system, CRISPR-Cas9. In contrast, Class I systems, which include Types I and III, rely on large multi-subunit effector complexes such as the CRISPR-associated complex for antiviral defense (Cascade) (112, 117).

In 2011, researchers discovered that Cas9 alone mediates interference in many Type II CRISPR systems (118). Around the same time, trans-encoded small RNAs (tracrRNAs) were found to contain ~22-nucleotide regions complementary to the repeats in pre-crRNA transcripts, enabling duplex formation (119, 120). This duplex is processed by endogenous RNase III in the presence of Cas9 to generate mature crRNAs (119). Cas9 then forms a ribonucleoprotein complex (RNP) with the tracrRNA:crRNA duplex and locates its DNA target by recognizing PAM sites, specifically the NGG sequence on the 3′ end of the non-complementary strand in *Streptococcus pyogenes* (Fig. 4) (120). PAM recognition triggers DNA unwinding and R-loop formation, allowing the crRNA's 20-nucleotide spacer to hybridize to the complementary target

strand (120, 121). The seed region, comprising the first 7 nucleotides at the 3′ end of the spacer, must perfectly match the target adjacent to the PAM, while the 5′ end is more tolerant of mismatches (120, 122). Cas9 then introduces a blunt double-strand break via its two nuclease domains—HNH, which cleaves the complementary strand, and RuvC, which cleaves the non-complementary strand (Fig. 4). The simplicity, modularity, and programmability of this dual-RNA system, along with its demonstrated ability to cleave DNA *in vitro*, quickly established CRISPR-Cas9 as a transformative gene editing tool (120, 123).

CRISPR-Cas9 technology emerged just in time to transform the field of gene editing and genome engineering, offering a simpler method to operate while providing robust programmability. Early DNA editing strategies included triplex-forming oligonucleotides (TFOs), first conceptualized in 1958. TFOs were used to target specific chromosomal sites, binding by specific purine and pyrimidine interactions, protecting the sequence from methylation for restriction enzyme cleavage (124, 125). However, TFO binding was pH-dependent and less efficient at physiological pH (124). While synthesizing TFOs with phosphoramidite chemistry enabled the incorporation of various modifications that improved performance at physiological pH (126), TFOs were difficult to use and did not yield reliable results.

Other approaches, like homologous recombination (127) and self-splicing introns (128), showed promise but were labor-intensive and inefficient. Zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) advanced gene editing but required extensive protein engineering, making them costly for routine use (129–131). In contrast, CRISPR-Cas9, directed by RNA–DNA base pairing, was easier to design and modify, enabling multiplexed editing potential (132), though early off-target effects were a challenge (133, 134).

Despite this, CRISPR-Cas9 quickly surpassed ZFNs and TALENs in simplicity and efficiency, meeting the demand for a high-efficiency gene editing tool (131, 135).

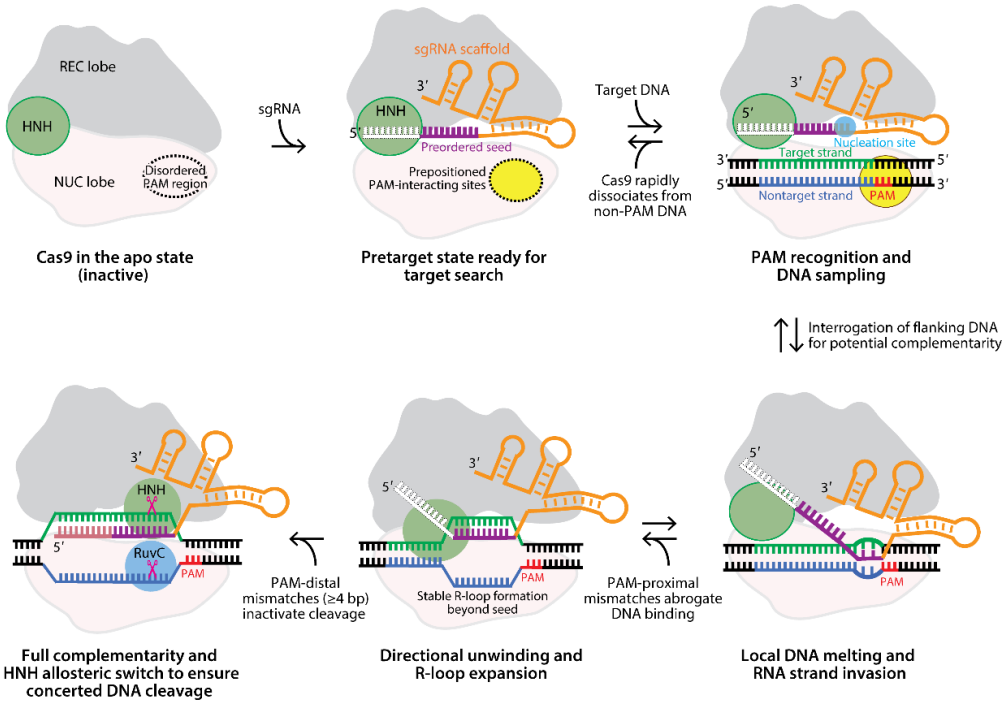


Figure 4. Mechanistic overview of CRISPR-Cas9 activation, from its inactive (apo) state to the formation of the Cas9:single-guide RNA (sgRNA) complex and DNA cleavage.

Mechanistic overview of CRISPR-Cas9 activation, from the inactive (apo) state to formation of the Cas9–sgRNA complex and targeted DNA cleavage, reprinted from Jiang and Doudna (2017) (170) with permission from Annual Reviews, Inc. For reference, this mechanism applies equally to either a dual-RNA system (tracrRNA:crRNA) or to a single-guide RNA (sgRNA), which is a chimeric fusion of tracrRNA and crRNA via a linker sequence (120).The mechanism is simplified into six discrete steps (left to right): (1) Apo-Cas9, the inactive form, diffuses on and off DNA without target search activity. Its C-terminal domain (CTD), which recognizes the protospacer adjacent motif (PAM), is disordered and nonfunctional. (2) Upon binding an sgRNA via its scaffold region, Cas9 undergoes a conformational change in its recognition (REC) lobe, activating the PAM-interacting site within the CTD and preorganizing the 5′ sgRNA "seed" for DNA interrogation. This complex forms an active ribonucleoprotein (RNP). (3) The RNP scans DNA until locating a PAM sequence (e.g., NGG) on the non-target strand. (4) PAM binding promotes local DNA unwinding, enabling the sgRNA seed to invade and pair with the complementary target strand. (5) If the sgRNA seed matches the target, an RNA–DNA hybrid (R-loop) begins to form. PAM-proximal mismatches lead to RNP dissociation. (6) If R-loop propagation continues through the full sgRNA spacer, and there are no more than ~4 mismatches in the PAM-distal region, Cas9 undergoes a final conformational shift that positions the HNH and RuvC domains for cleavage 3 bp upstream of the PAM, generating a blunt-end double-stranded break.

CRISPR-Cas9 programmability was first demonstrated through *in vitro* reconstitution studies and key optimizations to the tracrRNA:crRNA dual-RNA complex. Jinek and colleagues

streamlined this system by engineering a chimeric single-guide RNA (sgRNA) that fused the tracrRNA and crRNA with a short linker. This allowed efficient targeting of five genes using five 20-nt spacers, greatly enhancing CRISPR-Cas9's simplicity and programmability (120). The compact sgRNA design enables expression as a single transcript, eliminating the need for co-expression of separate RNA components. This was a major engineering advance in sgRNA design.

Early sgRNAs reduced editing efficiency compared to native tracrRNA:crRNA complexes, as shown in the first report of multiplexed CRISPR-Cas9 editing in mammalian cells. A single array expressing crRNAs with five unique spacers achieved robust indel formation at all five EMX1 targets, while corresponding sgRNAs did not, suggesting variable on-target editing efficiency for individual sgRNAs (132). Despite this, the study was a critical step toward multiplexed genome engineering, highlighting the importance of sgRNA design and the need for improved formats for scalable applications.

Together, these early studies revealed three key lessons: (1) sgRNAs are central to CRISPR-Cas9 programmability, (2) sgRNA design directly impacts editing efficiency, and (3) CRISPR-Cas9 is inherently compatible with multiplexed gene targeting. These insights laid the groundwork for developing sgRNA libraries—collections of hundreds to thousands of guides that differ only in their 20-nt spacer sequences. The introduction of microarray-derived oligonucleotides made sgRNA library synthesis affordable (136), catalyzing genome-wide CRISPR screens (Fig. 5A). This propelled the field in two directions: systematic studies of how spacer sequence influences efficacy, and the development of pooled libraries for large-scale functional genomics. Such screens included CRISPR knockout (CRISPRko), CRISPR interference (CRISPRi), and CRISPR activation (CRISPRa). CRISPRko uses Cas9 to disrupt

genes and reveal gene-function relationships (137), while CRISPRi employs a catalytically inactive Cas9 (dCas9) to block transcription initiation or elongation. Because dCas9 binding is reversible, CRISPRi enables tunable gene repression (136), which can be strengthened by fusing dCas9 to repressive domains like KRAB. Conversely, fusion to transcriptional activators such as VP16 enables CRISPRa to upregulate gene expression (138, 139).

Following the first CRISPRko screen, the Genome-scale CRISPR-Cas9 Knockout (GeCKO) libraries were introduced, significantly expanding the accessibility and scalability of CRISPR screening (144). Although produced using established methods (137), GeCKO libraries employed the lentiCRISPR vector, enabling co-expression of sgRNA and Cas9 from a single plasmid. This expanded the feasibility of genome-wide CRISPR screens to cell types where stable Cas9 integration was difficult to achieve (144, 145). The original GeCKO library included 64,751 unique sgRNAs targeting 18,080 protein-coding genes, increasing the number of gene targets by approximately 2.5-fold over earlier libraries (137, 144). GeCKO was first applied to both negative and positive selection screens: essential genes were identified via depletion in cancer and stem cell lines, while resistance genes were uncovered through survival in the presence of the melanoma drug vemurafenib (144). The following year, the GeCKOv2 libraries were released for both human and mouse genomes, containing 123,411 and 130,209 sgRNAs respectively (146). These libraries were accompanied by an updated lentiCRISPRv2 vector optimized for improved viral titer and expression, with a dual-vector system also available to help overcome transfection challenges (145, 146).

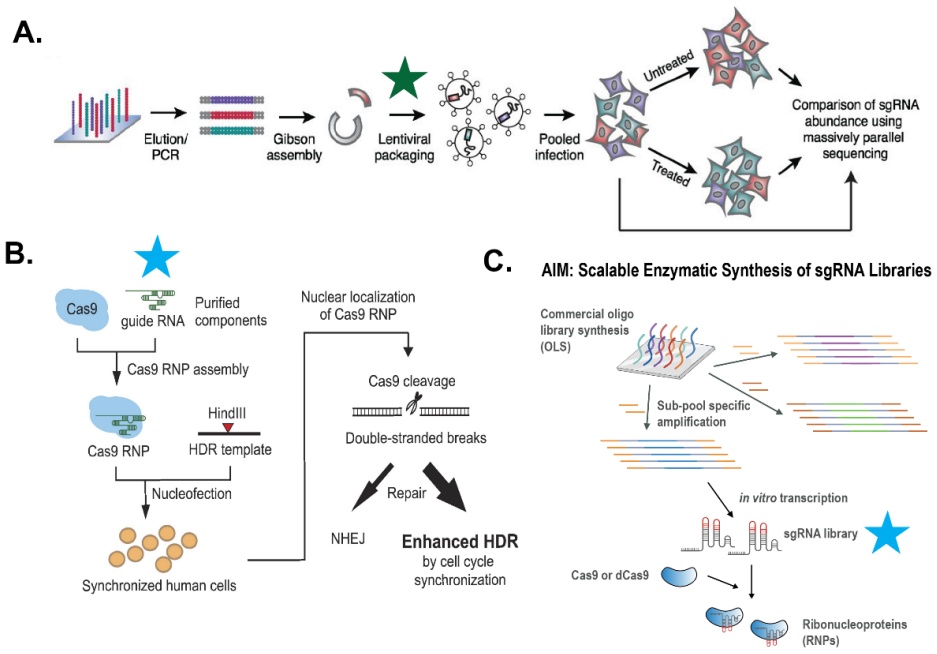


Figure 5. sgRNA format matters: From CRISPR screens to *in vitro* CRISPR assays.

Stars represent sgRNA formats: dark green for lentiviral sgRNA libraries and turquoise for IVT sgRNAs. **A**. CRISPR-Cas9 screening with lentiviral vector sgRNA libraries. Oligo sequences used to generate large-scale sgRNA libraries are typically obtained as a pool of microarray-derived oligos. Gibson assembly is used to clone these inserts into lentiviral vectors (often containing Cas9), producing an sgRNA library (dark green star). The resulting library is packaged into lentivirus and used to infect cells, where sgRNAs integrate into the genome and barcode each cell. After exposure to a drug or other selection condition, massively parallel sequencing is used to assess sgRNA enrichment (positive selection) or depletion (negative selection). From Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science, 343, 80–84., reprinted with permission from AAAS. **B**. Singleplex CRISPR-Cas9 gene editing with IVT sgRNAs. Active Cas9 ribonucleoproteins (RNPs) are assembled by complexing purified Cas9 protein with IVT sgRNAs (turquoise star). In the example shown (161), only one sgRNA is transcribed and delivered at a time. RNPs are electroporated into synchronized human cells alongside homology-directed repair (HDR) templates. Because HDR is active only during S and G2 phases (162), synchronization enhances editing precision. Transient RNP activity minimizes off-target editing and toxicity while enabling high-efficiency genome editing at single loci. Reprinted from Lin S *et al.* (2014) from eLife under CC-BY. **C**. Aim 2: Scalable enzymatic synthesis of sgRNA libraries for CRISPR RNP applications. This dissertation aims to develop a cost-effective, scalable method for enzymatically synthesizing sgRNA libraries. Similar to panel A, oligo pools encoding sgRNA sequences are generated via commercial oligo library synthesis (OLS), then amplified into subpools to reduce synthesis costs. These DNA templates are then *in vitro* transcribed by T7 RNA polymerase to produce sgRNA libraries (turquoise star) and complexed with either Cas9 or dCas9 to form RNPs. This approach was primarily developed to support *in vitro* CRISPR assays, including our method for enriching synthetic genes that lack synthesis errors. However, it could also enable CRISPR screens in primary cells such as natural killer cells, which do not support lentiviral sgRNA expression (158, 159).

The mouse GeCKOv2 library was soon applied in a genome-wide CRISPR knockout (CRISPRko) screen in dendritic cells to identify regulators of the innate immune response to bacterial lipopolysaccharide (LPS), using TNF-α production as a functional readout (147). While

the screen discovered numerous novel immune regulators, it also highlighted a key limitation of genome-wide approaches: the sheer scale of these libraries can lead to undersampling and false negatives if insufficient cell numbers are available. To address this, the authors constructed a focused library of 25,690 sgRNAs targeting 2,569 genes using the same pipeline as GeCKO and successfully validated several key hits. This study helped establish a growing consensus that genome-wide libraries, while powerful for discovery, may be impractical in certain contexts such as *in vivo* models or rare primary cell populations, where limited cell numbers reduce screening sensitivity. In such cases, smaller, more customized libraries can yield higher signal-to-noise ratios by focusing on fewer, biologically relevant targets. Additionally, genome-wide libraries often lack coverage of non-coding regions, which may be critical for certain research questions and are better addressed using targeted libraries (145).

One example of a study that benefited from a more focused CRISPR-Cas9 screen was the investigation of microRNAs (miRNAs) involved in cancer (148). miRNAs are short non-coding RNAs (~22 nt) that regulate gene expression and influence processes like metastasis by acting as oncogenes or tumor suppressors (149). Despite their critical role in cancer biology, miRNAs remain poorly understood, partly because libraries like GeCKOv2 are not optimized for studying them. Although GeCKOv2 includes 7,405 sgRNAs targeting 1,864 miRNA genes, these only account for about 6% of the total library (146). While GeCKOv2 has been used to screen miRNAs in models like non-small cell lung cancer (150) and myeloid leukemia (151), Kurata and Lin sought alternatives to large genome-wide libraries. They developed the LX-miR library, focusing exclusively on 1,594 pre-miRNA genes. Using a similar workflow to GeCKOv2, they synthesized 7,382 sgRNAs, with 4-5 guides per gene, excluding unrelated targets to optimize their screen for miRNAs (148).

A head-to-head comparison showed that GeCKOv2's miRNA-targeting sgRNAs had suboptimal activity and high off-target potential, while the LX-miR library demonstrated improved on-target editing and stronger signal-to-noise ratios. These performance improvements were evident in functional screens, where LX-miR helped identify both known and novel cancer-associated miRNAs in HeLa and NCI-N87 cell lines (148). While some differences may stem from cell line context, they likely reflect the advantages of optimized spacer selection and reduced library complexity. This example underscores how focused libraries can enhance screening efficiency and interpretability, particularly when targeting compact gene classes like miRNAs.

Although the advantages of smaller, lentiviral-based sgRNA libraries are well-established, their production has remained costly and labor-intensive, making large, commercially available libraries like GeCKOv2 the more accessible option for many researchers. To broaden access to high-quality, customizable CRISPR-Cas9 screening, innovation in sgRNA library production and delivery methods is essential. Accordingly, Aim 2 of this dissertation focuses on developing an affordable, scalable strategy for generating highly flexible sgRNA libraries (Fig. 5C).

Past efforts to build custom libraries from synthetic oligos have faced significant limitations. For instance, Henser-Brownhill and colleagues constructed a 3,150-sgRNA library targeting 450 human epigenetic regulators by cloning each guide individually into lentiviral vectors and expressing them in *E. coli*. Each bacterial culture contained a single sgRNA and had to be manually pooled to assemble both arrayed and pooled formats. This laborious process introduced cross-contamination between cultures and resulted in only 83% of sgRNAs being recovered in the final pool (152).

To overcome these challenges, we implemented a more scalable and modular approach. All sgRNA spacers are synthesized in a single microarray oligo pool, and focused subpools are selectively amplified by PCR using subpool-specific primers (Fig. 5C). This strategy builds on a method developed by Read and colleagues, who retrieved up to 24 distinct subpools from a single oligo pool for subsequent cloning into expression vectors (153). By eliminating one-by-one guide assembly and reducing the need for custom oligo synthesis, the approach described in Aim 2 significantly lowers costs and enhances scalability.

Rather than cloning sgRNA oligos into lentiviral vectors, we *in vitro* transcribed (IVT) sgRNA libraries using T7 RNA polymerase (Fig. 5C). These sgRNAs can be complexed with Cas9 protein to form ribonucleoproteins (RNPs), which are then electroporated into cells (Fig. 5B). This approach offers several advantages: RNPs minimize off-target activity, eliminate the risk of genomic integration, and improve editing efficiency in primary human cells, which are often refractory to viral transduction (154). RNP electroporation has been shown to enable high-efficiency genome editing in both mouse and human T lymphocytes (155, 156), which are partially permissive to lentiviral delivery but still benefit from RNP-based approaches (157). In contrast, natural killer (NK) cells, key effectors of innate immunity, are highly resistant to viral transduction, and electroporation of RNPs remains the only effective method for introducing CRISPR-Cas9 machinery (158, 159). Similarly, in mesenchymal stem cells (MSCs), plasmid-based delivery yielded only 9.01% indels with poor viability, whereas RNP electroporation achieved 20.21% editing efficiency with over 90% cell survival (160). Finally, one major advantage of RNPs is that they can induce editing within cells synchronized to specific stages in the cell cycle. For instance, introducing Cas9 RNPs during the S or G2 phases of the cell cycle promotes homology directed repair (HDR) of Cas9 cleavage sites, enabling precise editing (Fig.

5B) (161). On the other hand, expressing sgRNAs on lentiviral vectors could promote non-homologous end joining (NHEJ) repair of cut sites during cell cycle phases G1, S and G2 (161, 162). Therefore, RNPs enable precise knockout and knock-in insertion edits.

Despite their advantages, RNPs have a key limitation: their transient presence in cells prevents integration of selectable markers or barcodes, restricting their use in pooled CRISPR screens where linking genotype to phenotype is essential (163). To address this, the GUIDE-SWAP strategy was developed for primary T cells and hematopoietic stem cells (HSCs). In this method, cells are transduced with lentiviral sgRNA libraries while electroporating Cas9 protein complexed with inert sgRNAs to exploit the efficiency of RNP delivery (164). While effective, this approach is still not compatible with NK cells that cannot be efficiently transduced with lentivirus and rely on RNPs for CRISPR studies (158, 159). GUIDE-SWAP also involves additional steps and complexity.

Nevertheless, RNP-based delivery remains a promising strategy for conducting high-efficiency CRISPR screens in primary cells, particularly when paired with single-cell RNA sequencing or flow cytometry to capture phenotypic outcomes (145). Hybrid approaches have already emerged, such as transfecting sgRNAs while electroporating Cas9 protein, as demonstrated in primary T cells (157). Aim 2 of this dissertation seeks to support these evolving methodologies by developing programmable sgRNA libraries specifically designed for RNP delivery—filling a key gap in the current genome editing toolkit (Fig. 5C).

The second application of Aim 2 is to support Aim 1 (Chapter 2) by providing affordable and customizable sgRNA libraries. This is essential because the targeted enrichment strategy described in Aim 1 relies on *in vitro*-assembled RNPs to selectively recover correctly synthesized gene constructs from DropSynth gene libraries (Fig. 5C). Beyond this, the

programmable sgRNA libraries developed in Aim 2 are broadly applicable to both *in vitro* and *in vivo* CRISPR-Cas9 assays (Fig. 5C). For instance, catalytically inactive Cas9 (dCas9) fused to peptides or small molecules can be guided by these libraries to recruit endogenous factors, enabling precise modulation of the epigenetic landscape (165). These libraries can also be leveraged for functional *in vitro* assays or multiplexed diagnostics, including the detection of antimicrobial resistance genes or rare alleles (Fig. 6A,B) (166, 167). This versatility opens new possibilities for high-throughput detection of infectious diseases and rare genetic variants, offering powerful tools to tackle pressing biomedical challenges (168, 169).

A third utility of Aim 2 is its ability to overcome many limitations of commercially available IVT kits, which often lack support for subpooled DNA templates and are prohibitively expensive for large-scale sgRNA library generation. Aim 2 addresses these challenges by enabling the scalable production of high-quality sgRNA libraries, providing a more flexible, efficient, and cost-effective platform for CRISPR-based screens and assays, with the sgRNA libraries serving as their programmable core.

To achieve these objectives, Aim 2 is divided into three sub-aims. In sub-aim 2.1, we synthesized a small, proof-of-concept sgRNA library to demonstrate that our custom IVT protocol functions independently of commercial kits. Sub-aim 2.2 extended this approach to produce ten distinct sgRNA libraries, each containing hundreds to thousands of unique sgRNAs, all derived from a single microarray-based oligo pool, thus showcasing the scalability of our method. Finally, Sub-aim 2.3 focused on optimizing key quality metrics of the sgRNA libraries by identifying reaction conditions and sequence modifications that improve spacer representation and ensure uniformity across the libraries.

Chapter 3 of this dissertation, Optimizing *in vitro* Transcribed CRISPR-Cas9 Single-Guide RNA Libraries for Improved Uniformity and Affordability, provides a detailed exploration of the development and implementation of the sgRNA library synthesis platform. This chapter includes material co-authored with Yukiko R. Gaudreault, Abigail Keller, Phillip Kearns, James A. Stapleton, and Dr. Calin Plesa. The final section of this introduction will now shift focus to strategies beyond sgRNA design—specifically, those underlying methods in CRISPR-targeted DNA enrichment. These concepts will serve as a foundation for Aim 1, as outlined in Chapter 2.

1.5 Expanding CRISPR-Cas9 Targeted DNA Enrichment from Diagnostics to Multiplexed Synthetic Gene Enrichment

Enrichment of DNA sequences has long been a fundamental tool in molecular biology for gene cloning, NGS, and clinical diagnostics. Before the adoption of PCR, Southern blotting was a cutting-edge technique for isolating DNA fragments of interest. First published by Sir Edwin Southern in 1975, this method involved genomic DNA digestion to generate millions of fragments, followed by fragment separation through agarose gel electrophoresis, blotting onto a nitrocellulose membrane, and hybridization with radioactive or fluorescent RNA probes. The capture and visualization of hybridized DNA facilitated gene cloning and genome analysis. While revolutionary at the time for its relative simplicity and efficiency compared to manual gel excision before hybridization (171, 172), this method had limitations—most notably, its inability to reliably hybridize DNA fragments shorter than 500 base pairs.

The rapid adoption of PCR superseded Southern blotting as a method for detecting specific DNA sequences. This shift was driven by PCR's ability to enrich even a single target molecule from as little as a nanogram of genomic DNA, equivalent to approximately 143 diploid

copies of each gene (41, 173). Unlike Southern blotting, PCR did not require detailed knowledge of restriction sites or the use of multiple probes to hybridize across large regions. Multiplex PCR soon emerged as a method for detecting deletions associated with the Duchenne muscular dystrophy (DMD) locus, utilizing nine primer pairs flanking deletion-prone regions (173). The PCR amplicons were screened via mobility shift assays, reducing the need for multiple tests per patient (174). While multiplex PCR remains a powerful tool, even with optimization, most reactions can accommodate only around 12 primer pairs. Additionally, multiplex PCR is prone to mis-hybridization, off-target amplification, and biases that affect coverage (174, 175).

CRISPR-Cas9 is ideal for targeted DNA enrichment because Cas9 cuts DNA like a programmable restriction enzyme (176) and offers single-base resolution (167). Additionally, it is less prone to off-target effects, has lower reagent costs compared to hybridization-based approaches, and enables all of this in a single reaction (166). Unlike PCR, CRISPR-Cas9 methods are inherently less sensitive to cross-contamination between reactions, as they require both a guide RNA and a matching DNA sequence for cleavage, limiting false positives to rare instances where mismatched pairs co-occur. This dual dependency enhances specificity and supports greater multiplexing potential. In contrast, PCR-based methods, such as dial-out PCR (56), can be more prone to off-target amplification in complex libraries, limiting scalability and effectiveness in selective enrichment.

The application of CRISPR-Cas9 for targeted enrichment is especially useful for detecting rare alleles or low-frequency mutations. Amplifying DNA for NGS alone is often insufficient for these applications due to background sequencing errors and the high abundance of wildtype alleles (177, 178). As of 2021, Illumina MiSeq reported a mean error rate of 0.473% (SD 0.938), which is too high for reliable detection of variants below 0.5% frequency (179).

Unique molecular identifiers (UMIs) help track original molecules and identify errors introduced during amplification and sequencing. However, the use of UMIs requires high sequencing depth and complex computational pipelines, such as DELFMUT (Depth Estimation model for stable detection of Low-Frequency MUTations), to reliably detect low-frequency variants (180).

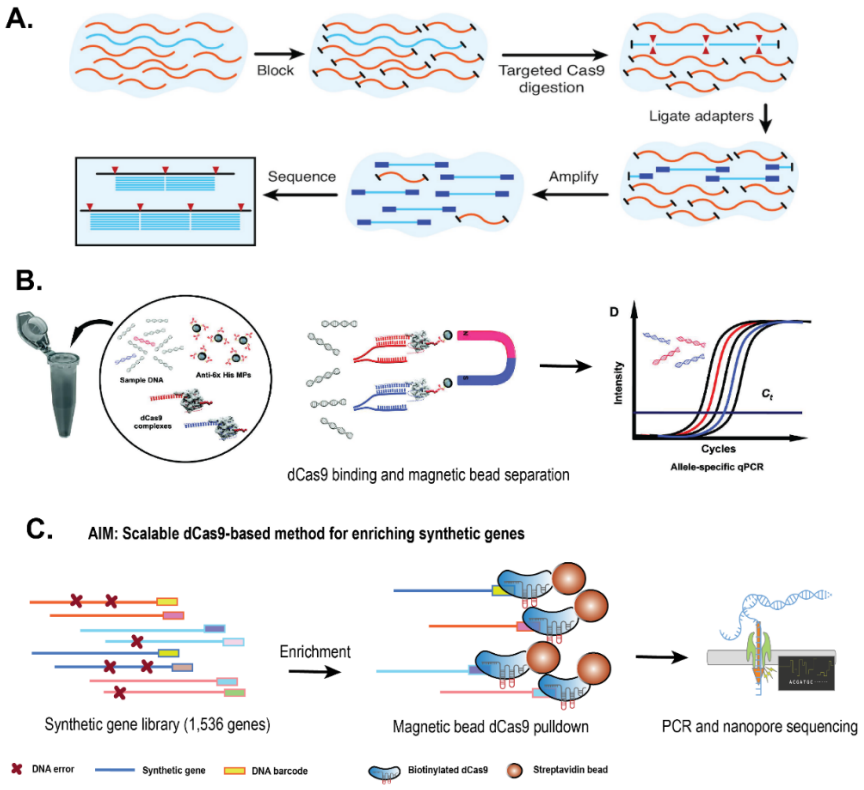


Figure 6. Using CRISPR-Cas9 and dCas9 to enrich targeted DNA sequences.

A. Multiplexed enrichment of genomic or cDNA using CRISPR-Cas9. An example of this approach is FLASH (Finding Low Abundance Sequences by Hybridization) NGS, which was developed to substitute for multiplex PCR in diagnostics (166). DNA is first dephosphorylated to prevent nonspecific ligation. Cas9 RNPs, assembled with IVT sgRNAs, cleave target sites to produce double-stranded breaks with 5′ phosphate and 3′ hydroxyl ends. These ends are ligated to Illumina adapters, selectively amplified by PCR, and sequenced. Reprinted from Quan *et al.* (2019) eLife, under CC-BY. **B.** Enrichment of minor alleles from cfDNA using CRISPR-dCas9. This singleplex or 3-plex method uses dCas9 tagged with a polyhistidine affinity tag and complexed with sgRNAs spanning target mutations. After binding to cfDNA, dCas9 RNPs are isolated using magnetic beads functionalized with anti-His antibodies. Targeted minor alleles are then quantified using qPCR (167). Reprinted from Aalipour *et al.* (2018), Clinical Chemistry, with permission from Oxford University Press. **C.** Aim 1: Scalable CRISPR-dCas9 enrichment of perfect gene variants from synthetic libraries. To enrich perfect assemblies from large synthetic gene libraries (e.g., DropSynth), we developed a scalable method using biotinylated dCas9 RNPs and IVT sgRNAs (see Fig. 5C). After RNPs bind to barcodes linked to perfect gene sequences (identified via NGS), targets are captured using streptavidin magnetic beads and amplified by PCR for nanopore sequencing verification. Nanopore sequencing icon by DBCLS (https://togotv.dbcls.jp/en/pics.html), licensed under CC-BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

One form of CRISPR enrichment is negative selection, where sgRNAs target and deplete unwanted sequences. Depletion of Unwanted Sequences by Hybridization (DASH) used 54 sgRNAs to remove abundant mitochondrial rRNA, enriching rare pathogen sequences in cerebrospinal fluid from meningitis patients and a KRAS driver mutation (176). Similarly, three sgRNAs were used to deplete wild-type alleles of somatic mutations in genes like EGFR and HBB, improving detection in NSCLC diagnostics (181). While effective, these approaches have limited scalability, typically targeting only 2–3 loci.

Although effective in clinical diagnostics, negative selection strategies like DASH are impractical for highly complex DropSynth libraries. Each gene is tagged with a unique 20-nt barcode, and imperfect assemblies often appear as single-copy variants. Depleting all error-containing sequences would require tens of thousands or more unique sgRNAs—far beyond the scale of prior DASH-style applications and infeasible in practice. In contrast, positive selection targets only the correct assemblies and can be achieved with 200–2,000 unique sgRNAs, offering a far more scalable and efficient strategy for enriching accurate constructs in synthetic libraries.

Cas9 Locus-Associated Proteome (CLASP), a CRISPR-based positive selection method, uses dCas9 to enrich DNA-bound regulatory proteins from specific loci for mass spectrometry. Unlike chromatin immunoprecipitation followed by sequencing (ChIP-seq), which captures known regulators using antibodies to locate their targets, CLASP is ideal when the DNA binding site is known but the regulatory proteins are unknown. In this method, cells are crosslinked, and chromatin is extracted and incubated with RNPs containing dCas9 fused to a 3×FLAG tag. An anti-FLAG antibody pulls down dCas9, enriching the targeted region; nuclease treatment then releases bound proteins for identification by mass spectrometry. In Drosophila melanogaster S2 cells, eight sgRNAs targeted the H2A/H2B promoter region (HisC), successfully identifying Vig

and Vig2—two previously unrecognized regulators of HisC (182). As with DASH, CLASP relies on multiple sgRNAs focused on a single locus rather than targeting diverse sequences across a complex library.

Another dCas9-based positive selection method was developed to enrich three minor alleles for diagnostic purposes, including a deletion (ELREA) and two EGFR mutations (T790M and L858R) associated with non-small cell lung cancer (NSCLC) in cell-free DNA (cfDNA) (167). Since the L858R allele lacked a PAM site, a spacer with a single mismatch was designed to introduce a PAM, preserving specificity by still discriminating against the wild-type allele. RNPs were formed with each sgRNA and polyhistidine-tagged dCas9, and enrichment at singleplex or 3-plex scales was performed using streptavidin-coated beads functionalized with a biotinylated anti-His antibody, followed by qPCR and NGS (Fig. 6B). While the method successfully enriched targets from both reference and patient samples, some limitations emerged: 13 of 34 variants identified by NGS were undetectable, although 8 of those 13 were recovered using multiplexed CRISPR-dCas9. This study demonstrated proof-of-concept for CRISPR-mediated positive selection in clinical diagnostics on a small scale. The authors proposed expanding this strategy using sgRNA libraries to enable multigene panels, highlighting the need for scalable, targeted enrichment methods (167).

Inspired by this future direction, we hypothesized that a similar approach could be adapted for Aim 1. Specifically, biotinylated dCas9 could retrieve perfect gene assemblies from barcode-labeled DropSynth libraries by pulling down dCas9-bound barcodes with streptavidin-coated beads (Chapter 2). Given that our approach aimed to use hundreds or thousands of sgRNAs for multiplexed enrichment in a single reaction, we first explored large-scale enrichment methods based on positive rather than negative selection (Fig. 6C).

In addition to dCas9, we also initially considered using catalytically active Cas9 for targeted enrichment. One method, Finding Low Abundance Sequences by Hybridization (FLASH), was developed to detect antimicrobial resistance (AMR) mutations in clinical samples, initially applied to Staphylococcus aureus and Plasmodium falciparum (166). The protocol involved dephosphorylating DNA or cDNA from respiratory fluids or dried blood spots to block unwanted ligation, then incubating with Cas9 RNPs assembled using IVT-generated sgRNAs. After Cas9 cleavage, Cas9 was removed with proteinase K, and the DNA fragments were ligated to Illumina adapters for sequencing (Fig. 6A). FLASH-NGS included a tool called FLASHit for sgRNA design, enabling the design of 5,513 sgRNAs to target 3,624 AMR sites. In pilot studies, 127 clinically relevant genes were enriched from both cultured isolates and clinical samples. A 2,226-sgRNA library achieved 90.6% recovery of target regions, demonstrating FLASH-NGS scalability (166).

A similar method, Cas12a-Capture, was developed to improve molecular diagnosis of underdiagnosed Mendelian disorders, such as Joubert Syndrome (JS), a recessive disorder affecting hindbrain development (169, 183). Like FLASH, Cas12a-Capture involved dephosphorylating genomic DNA, followed by cleavage with Cas12a RNPs. Cas12a, a Class 2 CRISPR effector, is programmable with 42-nt guide RNAs that do not require a tracrRNA. Using custom selection criteria, 7,176 guide RNAs were designed to target 21-nt protospacers adjacent to Cas12a's TTTN PAM, capturing 47 genes associated with JS. Cas12a cleavage generates 4–5 nt sticky ends, to which i5 adapters were ligated. i7 adapters were added via Tn5 tagmentation, followed by streptavidin pull-down and PCR amplification of adapter-ligated fragments. Although effective, 12.7% of reads aligned to off-target regions, likely due to oligo

synthesis errors or suboptimal guide design. A predictive model was used to design an expanded set of 11,438 guides, which modestly improved enrichment uniformity (169).

Given these considerations, we chose the CRISPR-Cas9 system for Aim 1 due to its simplicity and extensive characterization, compared to newer CRISPR systems like Cas12a. While catalytically active Cas9 has been widely used for targeted enrichment with hundreds to thousands of sgRNAs (166, 169) (Fig. 6A), we opted for dCas9 based on its proven success in targeted binding and DNA pull-down for positive selection (167, 182) (Fig. 6B). Additionally, we identified a key gap in the field: the lack of scalable *in vitro* dCas9 enrichment methods (Fig. 6C). To address this, we developed a new approach designed not only to fill this gap but also to uncover fundamental engineering principles that could guide future CRISPR-based technologies. This approach is described in detail in the following chapter (Chapter 2).

1.6 References

1. Wu,C., Xu,Z. and Zhang,H.-B. (2006) DNA Libraries. *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, 10.1002/3527600906.mcb.200300065.

2. Shizuya,H., Birren,B., Kim,U.J., Mancino,V., Slepak,T., Tachiiri,Y. and Simon,M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 8794–8797.

3. Shendure,J. and Lieberman Aiden,E. (2012) The expanding scope of DNA sequencing. *Nat. Biotechnol.*, **30**, 1084–1094.

4. Kosuri,S. and Church,G.M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*, **11**, 499–507.

5. Head,S.R., Komori,H.K., LaMere,S.A., Whisenant,T., Van Nieuwerburgh,F., Salomon,D.R. and Ordoukhanian,P. (2014) Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, **56**, 61–4, 66, 68, passim.

6. Rondon,M.R., August,P.R., Bettermann,A.D., Brady,S.F., Grossman,T.H., Liles,M.R., Loiacono,K.A., Lynch,B.A., MacNeil,I.A., Minor,C., *et al.* (2000) Cloning the soil metagenome: A strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.*, **66**, 2541–2547.

7. Negri,T., Mantri,S., Angelov,A., Peter,S., Muth,G., Eustáquio,A.S. and Ziemert,N. (2022) A rapid and efficient strategy to identify and recover biosynthetic gene clusters from soil metagenomes. *Appl. Microbiol. Biotechnol.*, **106**, 3293–3306.

8. Lema,N.K., Gemeda,M.T. and Woldesemayat,A.A. (2023) Recent advances in metagenomic approaches, applications, and challenge. *Curr. Microbiol.*, **80**, 347.

9. Piscotta,F.J., Whitfield,S.T., Nakashige,T.G., Estrela,A.B., Ali,T. and Brady,S.F. (2021) Multiplexed functional metagenomic analysis of the infant microbiome identifies effectors of NF-κB, autophagy, and cellular redox state. *Cell Rep.*, **36**, 109746.

10. Weiland-Bräuer,N., Saleh,L. and Schmitz,R.A. (2023) Functional metagenomics as a tool to tap into natural diversity of valuable biotechnological compounds. *Methods Mol. Biol.*, **2555**, 23–49.

11. Robinson,S.L., Piel,J. and Sunagawa,S. (2021) A roadmap for metagenomic enzyme discovery. *Nat. Prod. Rep.*, **38**, 1994–2023.

12. Chen,C., Liao,Y. and Peng,G. (2022) Connecting past and present: single-cell lineage tracing. *Protein Cell*, **13**, 790–807.

13. Fulco,C.P., Munschauer,M., Anyoha,R., Munson,G., Grossman,S.R., Perez,E.M., Kane,M., Cleary,B., Lander,E.S. and Engreitz,J.M. (2016) Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, **354**, 769–773.

14. Ford,K.M., Panwala,R., Chen,D.-H., Portell,A., Palmer,N. and Mali,P. (2021) Peptide-tiling screens of cancer drivers reveal oncogenic protein domains and associated peptide inhibitors. *Cell Syst.*, **12**, 716–732.e7.

15. Tan,Y., Zhang,Y., Han,Y., Liu,H., Chen,H., Ma,F., Withers,S.G., Feng,Y. and Yang,G. (2019) Directed evolution of an α1,3-fucosyltransferase using a single-cell ultrahigh-throughput screening method. *Sci. Adv.*, **5**, eaaw8451.

16. Rubin,A.F., Stone,J., Bianchi,A.H., Capodanno,B.J., Da,E.Y., Dias,M., Esposito,D., Frazer,J., Fu,Y., Grindstaff,S.B., *et al.* (2025) MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome Biol.*, **26**, 13.

17. Gasperini,M., Starita,L. and Shendure,J. (2016) The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.*, **11**, 1782–1787.

18. Melnikov,A., Murugan,A., Zhang,X., Tesileanu,T., Wang,L., Rogov,P., Feizi,S., Gnirke,A., Callan,C.G.,Jr, Kinney,J.B., *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, **30**, 271–277.

19. La Fleur,A., Shi,Y. and Seelig,G. (2024) Decoding biology with massively parallel reporter assays and machine learning. *Genes Dev.*, **38**, 843–865.

20. Fowler,D.M., Araya,C.L., Fleishman,S.J., Kellogg,E.H., Stephany,J.J., Baker,D. and Fields,S. (2010) High-resolution mapping of protein sequence-function relationships. *Nat. Methods*, **7**, 741–746.

21. Araya,C.L. and Fowler,D.M. (2011) Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.*, **29**, 435–442.

22. Fowler,D.M. and Fields,S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods*, **11**, 801–807.

23. Cadwell,R.C. and Joyce,G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl.*, **2**, 28–33.

24. Sarkar,G. and Sommer,S.S. (1990) The 'megaprimer' method of site-directed mutagenesis. *Biotechniques*, **8**, 404–407.

25. Zhao,J., Kardashliev,T., Joëlle Ruff,A., Bocola,M. and Schwaneberg,U. (2014) Lessons from diversity of directed evolution experiments by an analysis of 3,000 mutations: Lessons From Diversity of Directed Evolution. *Biotechnol. Bioeng.*, **111**, 2380–2389.

26. Yang,J., Ruff,A.J., Arlt,M. and Schwaneberg,U. (2017) Casting epPCR (cepPCR): A simple random mutagenesis method to generate high quality mutant libraries. *Biotechnol. Bioeng.*, **114**, 1921–1927.

27. Plesa,C., Sidore,A.M., Lubock,N.B., Zhang,D. and Kosuri,S. (2018) Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science*, **359**, 343–347.

28. Romanowicz,K.J., Resnick,C., Hinton,S.R. and Plesa,C. (2025) Exploring antibiotic resistance in diverse homologs of the dihydrofolate reductase protein family through broad Mutational Scanning. *bioRxivorg*, 10.1101/2025.01.23.634126.

29. Welch,M., Govindarajan,S., Ness,J.E., Villalobos,A., Gurney,A., Minshull,J. and Gustafsson,C. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One*, **4**, e7002.

30. Sidore,A.M., Plesa,C., Samson,J.A., Lubock,N.B. and Kosuri,S. (2020) DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Res.*, **48**, e95.

31. Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.

32. Michelson,A.M. and Todd,A.R. (1955) Nucleotides part XXXII. Synthesis of a dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage. *J. Chem. Soc.*, **0**, 2632–2638.

33. Hall,R.H., Todd,A. and Webb,R.F. (1957) 644. Nucleotides. Part XLI. Mixed anhydrides as intermediates in the synthesis of dinucleoside phosphates. *J. Chem. Soc.*

34. Khorana,H.G., Razzell,W.E., Gilham,P.T., Tener,G.M. and Pol,E.H. (1957) Syntheses of

dideoxyribonucleotides. *J. Am. Chem. Soc.*, **79**, 1002–1003.

35. Nishimura,S., Jones,D.S. and Khorana,H.G. (1965) Studies on polynucleotides. 48. The *in vitro* synthesis of a co-polypeptide containing two amino acids in alternating sequence dependent upon a DNA-like polymer containing two nucleotides in alternating sequence. *J. Mol. Biol.*, **13**, 302–324.

36. Letsinger,R.L. and Mahadevan,V. (1965) Oligonucleotide synthesis on a polymer support. *J. Am. Chem. Soc.*, **87**, 3526–3527.

37. Pon,R.T. (2001) Solid-phase supports for oligonucleotide synthesis. *Curr. Protoc. Nucleic Acid Chem.*, **Chapter 3**, Unit 3.1.

38. Beaucage,S.L. and Caruthers,M.H. (1981) Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.*, **22**, 1859–1862.

39. Matteucci,M.D. and Caruthers,M.H. (1992) Synthesis of deoxyoligonucleotides on a polymer support. 1981. *Biotechnology*, **24**, 92–98.

40. Efcavitch,J.W. and Heiner,C. (1985) Depurination as a yield decreasing mechanism in oligodeoxynucleotide synthesis. *Nucleosides Nucleotides*, **4**, 267–267.

41. Saiki,R., Gelfand,D., Stoffel,S., Scharf,S., Higuchi,R., Horn,G., Mullis,K. and Erlich,H. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.

42. Sandahl,A.F., Nguyen,T.J.D., Hansen,R.A., Johansen,M.B., Skrydstrup,T. and Gothelf,K.V. (2021) On-demand synthesis of phosphoramidites. *Nat. Commun.*, **12**, 2760.

43. Hoose,A., Vellacott,R., Storch,M., Freemont,P.S. and Ryadnov,M.G. (2023) DNA synthesis technologies to close the gene writing gap. *Nat Rev Chem*.

44. LeProust,E., Zhang,H., Yu,P., Zhou,X. and Gao,X. (2001) Characterization of oligodeoxyribonucleotide synthesis on glass plates. *Nucleic Acids Res.*, **29**, 2171–2180.

45. Stemmer,W.P., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.

46. Fodor,S.P., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.

47. Fodor,S.P., Rava,R.P., Huang,X.C., Pease,A.C., Holmes,C.P. and Adams,C.L. (1993) Multiplexed biochemical assays with biological chips. *Nature*, **364**, 555–556.

48. Singh-Gasson,S., Green,R.D., Yue,Y., Nelson,C., Blattner,F., Sussman,M.R. and Cerrina,F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.*, **17**, 974–978.

49. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R., *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.

50. Agarwal,K.L., Büchi,H., Caruthers,M.H., Gupta,N., Khorana,H.G., Kleppe,K., Kumar,A., Ohtsuka,E., Rajbhandary,U.L., Van de Sande,J.H., *et al.* (1970) Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature*, **227**, 27–34.

51. Ciccarelli,R.B., Gunyuzlu,P., Huang,J., Scott,C. and Oakes,F.T. (1991) Construction of synthetic genes using PCR after automated DNA synthesis of their entire top and bottom strands. *Nucleic Acids Res.*, **19**, 6007–6013.

52. Filges,S., Mouhanna,P. and Ståhlberg,A. (2021) Digital Quantification of Chemical Oligonucleotide Synthesis Errors. *Clin. Chem.*, 10.1093/clinchem/hvab136.

53. Smith,H.O., Hutchison,C.A.,3rd, Pfannkoch,C. and Venter,J.C. (2003) Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 15440–15445.

54. Matzas,M., Stähler,P.F., Kefer,N., Siebelt,N., Boisguérin,V., Leonard,J.T., Keller,A., Stähler,C.F., Häberle,P., Gharizadeh,B., *et al.* (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.*, **28**, 1291–1294.

55. Luo,C., Tsementzi,D., Kyrpides,N., Read,T. and Konstantinidis,K.T. (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One*, **7**, e30087.

56. Schwartz,J.J., Lee,C. and Shendure,J. (2012) Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nat. Methods*, **9**, 913–915.

57. Choi,H., Choi,Y., Choi,J., Lee,A.C., Yeom,H., Hyun,J., Ryu,T. and Kwon,S. (2021) Purification of multiplex oligonucleotide libraries by synthesis and selection. *Nat. Biotechnol.*, 10.1038/s41587-021-00988-3.

58. Cheng,C., Fei,Z. and Xiao,P. (2023) Methods to improve the accuracy of next-generation sequencing. *Front Bioeng Biotechnol*, **11**, 982111.

59. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C., *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.

60. Villegas,N.K., Gaudreault,Y.R., Keller,A., Kearns,P., Stapleton,J.A. and Plesa,C. (2025) Optimizing *in vitro* transcribed CRISPR-Cas9 single-guide RNA libraries for improved uniformity and affordability. *bioRxiv*, 10.1101/2025.03.24.644170.

61. Palluk,S., Arlow,D.H., de Rond,T., Barthel,S., Kang,J.S., Bector,R., Baghdassarian,H.M.,

Truong,A.N., Kim,P.W., Singh,A.K., *et al.* (2018) De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.*, **36**, 645–650.

62. Barthel,S., Palluk,S., Hillson,N.J., Keasling,J.D. and Arlow,D.H. (2020) Enhancing terminal deoxynucleotidyl transferase activity on substrates with 3' terminal structures for enzymatic DE Novo DNA synthesis. *Genes (Basel)*, **11**, 102.

63. Forget,S.M., Krawczyk,M.J., Knight,A.M., Ching,C., Copeland,R.A., Mahmoodi,N., Mayo,M.A., Nguyen,J., Tan,A., Miller,M., *et al.* (2025) Evolving a terminal deoxynucleotidyl transferase for commercial enzymatic DNA synthesis. *Nucleic Acids Res.*, **53**, gkaf115.

64. Czar,M.J., Anderson,J.C., Bader,J.S. and Peccoud,J. (2009) Gene synthesis demystified. *Trends Biotechnol.*, **27**, 63–72.

65. Stemmer,W.P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, **370**, 389–391.

66. Crameri,A., Whitehorn,E.A., Tate,E. and Stemmer,W.P. (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat. Biotechnol.*, **14**, 315–319.

67. Cello,J., Paul,A.V. and Wimmer,E. (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*, **297**, 1016–1018.

68. Blight,K.J., Kolykhalov,A.A. and Rice,C.M. (2000) Efficient initiation of HCV RNA replication in cell culture. *Science*, **290**, 1972–1974.

69. Chalmers,F.M. and Curnow,K.M. (2001) Scaling up the ligase chain reaction-based approach to gene synthesis. *Biotechniques*, **30**, 249–252.

70. Tian,J., Gong,H., Sheng,N., Zhou,X., Gulari,E., Gao,X. and Church,G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.

71. Richmond,K.E., Li,M.-H., Rodesch,M.J., Patel,M., Lowe,A.M., Kim,C., Chu,L.L., Venkataramaian,N., Flickinger,S.F., Kaysen,J., *et al.* (2004) Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.*, **32**, 5011–5018.

72. Zhou,X., Cai,S., Hong,A., You,Q., Yu,P., Sheng,N., Srivannavit,O., Muranjan,S., Rouillard,J.M., Xia,Y., *et al.* (2004) Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.*, **32**, 5409–5417.

73. Borovkov,A.Y., Loskutov,A.V., Robida,M.D., Day,K.M., Cano,J.A., Le Olson,T., Patel,H., Brown,K., Hunter,P.D. and Sykes,K.F. (2010) High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic Acids Res.*, **38**, e180.

74. Barnes,W.M. (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 2216–2220.

75. Gibson,D.G., Benders,G.A., Andrews-Pfannkoch,C., Denisova,E.A., Baden-Tillson,H., Zaveri,J., Stockwell,T.B., Brownley,A., Thomas,D.W., Algire,M.A., *et al.* (2008) Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science*, **319**, 1215–1220.

76. Gibson,D.G., Young,L., Chuang,R.-Y., Venter,J.C., Hutchison,C.A.,3rd and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.

77. Merryman,C. and Gibson,D.G. (2012) Methods and applications for assembling large DNA constructs. *Metab. Eng.*, **14**, 196–204.

78. Gibson,D.G., Glass,J.I., Lartigue,C., Noskov,V.N., Chuang,R.-Y., Algire,M.A., Benders,G.A., Montague,M.G., Ma,L., Moodie,M.M., *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, **329**, 52–56.

79. Goold,H.D., Kroukamp,H., Erpf,P.E., Zhao,Y., Kelso,P., Calame,J., Timmins,J.J.B., Wightman,E.L.I., Peng,K., Carpenter,A.C., *et al.* (2025) Construction and iterative redesign of synXVI a 903 kb synthetic Saccharomyces cerevisiae chromosome. *Nat. Commun.*, **16**, 841.

80. Roth,T.L., Milenkovic,L. and Scott,M.P. (2014) A rapid and simple method for DNA engineering using cycled ligation assembly. *PLoS One*, **9**, e107329.

81. Chao,R., Yuan,Y. and Zhao,H. (2015) Recent advances in DNA assembly technologies. *FEMS Yeast Res.*, **15**, 1–9.

82. Engler,C., Kandzia,R. and Marillonnet,S. (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One*, **3**, e3647.

83. Strzelecki,P., Joly,N., Hébraud,P., Hoffmann,E., Cech,G.M., Kloska,A., Busi,F. and Grange,W. (2024) Enhanced Golden Gate Assembly: evaluating overhang strength for improved ligation efficiency. *Nucleic Acids Res.*, **52**, e95.

84. Potapov,V., Ong,J.L., Kucera,R.B., Langhorst,B.W., Bilotti,K., Pryor,J.M., Cantor,E.J., Canton,B., Knight,T.F., Evans,T.C.,Jr, *et al.* (2018) Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. *ACS Synth. Biol.*, **7**, 2665–2674.

85. Sikkema,A.P., Tabatabaei,S.K., Lee,Y.-J., Lund,S. and Lohman,G.J.S. (2023) High-Complexity One-Pot Golden Gate Assembly. *Curr Protoc*, **3**, e882.

86. Ellis,T., Adie,T. and Baldwin,G.S. (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol.* , **3**, 109–118.

87. LeProust,E.M., Peck,B.J., Spirin,K., McCuen,H.B., Moore,B., Namsaraev,E. and Caruthers,M.H. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.*, **38**, 2522–2540.

88. Kosuri,S., Eroshenko,N., Leproust,E.M., Super,M., Way,J., Li,J.B. and Church,G.M. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.*, **28**, 1295–1299.

89. Qiu,P., Shandilya,H., D'Alessio,J.M., O'Connor,K., Durocher,J. and Gerard,G.F. (2004) Mutation detection using Surveyor nuclease. *Biotechniques*, **36**, 702–707.

90. Quan,J., Saaem,I., Tang,N., Ma,S., Negre,N., Gong,H., White,K.P. and Tian,J. (2011) Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol.*, **29**, 449–452.

91. Klein,J.C., Lajoie,M.J., Schwartz,J.J., Strauch,E.-M., Nelson,J., Baker,D. and Shendure,J. (2016) Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.*, **44**, e43.

92. Holston,A.S., Hinton,S.R., Lindley,K.A., Kearns,N.C. and Plesa,C. (2023) Degenerate DropSynth for Simultaneous Assembly of Diverse Gene Libraries and Local Designed Mutants. 10.1101/2023.12.11.569291.

93. Biswas,I. and Hsieh,P. (1997) Interaction of MutS protein with the major and minor grooves of a heteroduplex DNA. *J. Biol. Chem.*, **272**, 13355–13364.

94. Lubock,N.B., Zhang,D., Sidore,A.M., Church,G.M. and Kosuri,S. (2017) A systematic comparison of error correction enzymes by next-generation sequencing. *Nucleic Acids Res.*, **45**, 9206–9217.

95. Obmolova,G., Ban,C., Hsieh,P. and Yang,W. (2000) Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature*, **407**, 703–710.

96. Uematsu,M. and Baskin,J.M. (2023) Barcode-free multiplex plasmid sequencing using Bayesian analysis and nanopore sequencing. *bioRxivorg*.

97. Brown,S.D., Dreolini,L., Wilson,J.F., Balasundaram,M. and Holt,R.A. (2023) Complete sequence verification of plasmid DNA using the Oxford Nanopore Technologies' MinION device. *BMC Bioinformatics*, **24**, 116.

98. Carpenter,M.L., Buenrostro,J.D., Valdiosera,C., Schroeder,H., Allentoft,M.E., Sikora,M., Rasmussen,M., Gravel,S., Guillén,S., Nekhrizov,G., *et al.* (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.*, **93**, 852–864.

99. Teer,J.K. and Mullikin,J.C. (2010) Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.*, **19**, R145–51.

100. Kim,H., Han,H., Shin,D. and Bang,D. (2010) A fluorescence selection method for accurate large-gene synthesis. *Chembiochem*, **11**, 2448–2452.

101. Wang,T., Badran,A.H., Huang,T.P. and Liu,D.R. (2018) Continuous directed evolution of proteins with improved soluble expression. *Nat. Chem. Biol.*, **14**, 972–980.

102. Sequeira,A.F., Guerreiro,C.I.P.D., Vincentelli,R. and Fontes,C.M.G.A. (2016) T7 Endonuclease I Mediates Error Correction in Artificial Gene Synthesis. *Mol Biotechnol*, **58**, 573–584.

103. Tian,J., Ma,K. and Saaem,I. (2009) Advancing high-throughput gene synthesis technology. *Mol. Biosyst.*, **5**, 714–722.

104. Ishino,Y., Shinagawa,H., Makino,K., Amemura,M. and Nakata,A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.*, **169**, 5429–5433.

105. Ishino,Y., Krupovic,M. and Forterre,P. (2018) History of CRISPR-Cas from encounter with a mysterious repeated sequence to genome editing technology. *J. Bacteriol.*, **200**.

106. Mojica,F.J., Díez-Villaseñor,C., Soria,E. and Juez,G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.*, **36**, 244–246.

107. Jansen,R., van Embden,J.D.A., Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.

108. Pourcel,C., Salvignol,G. and Vergnaud,G. (2005) CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, **151**, 653–663.

109. Tang,T.-H., Bachellerie,J.-P., Rozhdestvensky,T., Bortolin,M.-L., Huber,H., Drungowski,M., Elge,T., Brosius,J. and Hüttenhofer,A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 7536–7541.

110. Mojica,F.J.M., Díez-Villaseñor,C., García-Martínez,J. and Soria,E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.*, **60**, 174–182.

111. Bolotin,A., Quinquis,B., Sorokin,A. and Ehrlich,S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, **151**, 2551–2561.

112. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J.M., Wolf,Y.I., Yakunin,A.F., *et al.* (2011) Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*, **9**, 467–477.

113. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.

114. Garneau,J.E., Dupuis,M.-È., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadán,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.

115. Mojica,F.J.M., Díez-Villaseñor,C., García-Martínez,J. and Almendros,C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, England)*, **155**.

116. Brouns,S.J.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J.H., Snijders,A.P.L., Dickman,M.J., Makarova,K.S., Koonin,E.V. and van der Oost,J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.

117. Hille,F., Richter,H., Wong,S.P., Bratovič,M., Ressel,S. and Charpentier,E. (2018) The biology of CRISPR-Cas: Backward and forward. *Cell*, **172**, 1239–1259.

118. Sapranauskas,R., Gasiunas,G., Fremaux,C., Barrangou,R., Horvath,P. and Siksnys,V. (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.*, **39**, 9275–9282.

119. Deltcheva,E., Chylinski,K., Sharma,C.M., Gonzales,K., Chao,Y., Pirzada,Z.A., Eckert,M.R., Vogel,J. and Charpentier,E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, **471**, 602–607.

120. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.

121. Pacesa,M., Loeff,L., Querques,I., Muckenfuss,L.M., Sawicka,M. and Jinek,M. (2022) R-loop formation and conformational activation mechanisms of Cas9. *Nature*, **609**, 191–196.

122. Boyle,E.A., Andreasson,J.O.L., Chircus,L.M., Sternberg,S.H., Wu,M.J., Guegler,C.K., Doudna,J.A. and Greenleaf,W.J. (2017) High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 5461–5466.

123. Gasiunas,G., Barrangou,R., Horvath,P. and Siksnys,V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, E2579–86.

124. Strobel,S.A. and Dervan,P.B. (1991) Single-site enzymatic cleavage of yeast genomic DNA mediated by triple helix formation. *Nature*, **350**, 172–174.

125. Strobel,S.A., Doucette-Stamm,L.A., Riba,L., Housman,D.E. and Dervan,P.B. (1991) Site-

specific cleavage of human chromosome 4 mediated by triple-helix formation. *Science*, **254**, 1639–1642.

126. Puri,N., Majumdar,A., Cuenoud,B., Natt,F., Martin,P., Boyd,A., Miller,P.S. and Seidman,M.M. (2001) Targeted gene knockout by 2'-O-aminoethyl modified triplex forming oligonucleotides. *J. Biol. Chem.*, **276**, 28991–28998.

127. Scherer,S. and Davis,R.W. (1979) Replacement of chromosome segments with altered DNA sequences constructed *in vitro*. *Proc. Natl. Acad. Sci. U. S. A.*, **76**, 4951–4955.

128. Yang,J., Zimmerly,S., Perlman,P.S. and Lambowitz,A.M. (1996) Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. *Nature*, **381**, 332–335.

129. Kim,Y.G., Cha,J. and Chandrasegaran,S. (1996) Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 1156–1160.

130. Boch,J., Scholze,H., Schornack,S., Landgraf,A., Hahn,S., Kay,S., Lahaye,T., Nickstadt,A. and Bonas,U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.

131. Doudna,J.A. and Charpentier,E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**, 1258096.

132. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A., *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

133. Mahfouz,M.M., Piatek,A. and Stewart,C.N.,Jr (2014) Genome engineering via TALENs and CRISPR/Cas9 systems: challenges and perspectives. *Plant Biotechnol. J.*, **12**, 1006–1014.

134. Fu,Y., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,J.K. and Sander,J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.

135. Lino,C.A., Harper,J.C., Carney,J.P. and Timlin,J.A. (2018) Delivering CRISPR: a review of the challenges and approaches. *Drug Deliv.*, **25**, 1234–1257.

136. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.

137. Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.

138. Gilbert,L.A., Larson,M.H., Morsut,L., Liu,Z., Brar,G.A., Torres,S.E., Stern-Ginossar,N., Brandman,O., Whitehead,E.H., Doudna,J.A., *et al.* (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, **154**, 442–451.

139. Konermann,S., Brigham,M.D., Trevino,A.E., Joung,J., Abudayyeh,O.O., Barcena,C., Hsu,P.D., Habib,N., Gootenberg,J.S., Nishimasu,H., *et al.* (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**, 583–588.

140. Booker,M., Samsonova,A.A., Kwon,Y., Flockhart,I., Mohr,S.E. and Perrimon,N. (2011) False negative rates in Drosophila cell-based RNAi screens: a case study. *BMC Genomics*, **12**, 50.

141. Echeverri,C.J., Beachy,P.A., Baum,B., Boutros,M., Buchholz,F., Chanda,S.K., Downward,J., Ellenberg,J., Fraser,A.G., Hacohen,N., *et al.* (2006) Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat. Methods*, **3**, 777–779.

142. Paddison,P.J., Silva,J.M., Conklin,D.S., Schlabach,M., Li,M., Aruleba,S., Balija,V., O'Shaughnessy,A., Gnoj,L., Scobie,K., *et al.* (2004) A resource for large-scale RNA-interference-based screens in mammals. *Nature*, **428**, 427–431.

143. Koike-Yusa,H., Li,Y., Tan,E.-P., Velasco-Herrera,M.D.C. and Yusa,K. (2014) Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.*, **32**, 267–273.

144. Shalem,O., Sanjana,N.E., Hartenian,E., Shi,X., Scott,D.A., Mikkelson,T., Heckl,D., Ebert,B.L., Root,D.E., Doench,J.G., *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.

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

146. Sanjana,N.E., Shalem,O. and Zhang,F. (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*, **11**, 783–784.

147. Parnas,O., Jovanovic,M., Eisenhaure,T.M., Herbst,R.H., Dixit,A., Ye,C.J., Przybylski,D., Platt,R.J., Tirosh,I., Sanjana,N.E., *et al.* (2015) A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. *Cell*, **162**, 675–686.

148. Kurata,J.S. and Lin,R.-J. (2018) MicroRNA-focused CRISPR-Cas9 library screen reveals fitness-associated miRNAs. *RNA*, **24**, 966–981.

149. Volinia,S., Calin,G.A., Liu,C.-G., Ambs,S., Cimmino,A., Petrocca,F., Visone,R., Iorio,M., Roldo,C., Ferracin,M., *et al.* (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 2257–2261.

150. Chen,S., Sanjana,N.E., Zheng,K., Shalem,O., Lee,K., Shi,X., Scott,D.A., Song,J., Pan,J.Q., Weissleder,R., *et al.* (2015) Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell*, **160**, 1246–1260.

151. Wallace,J., Hu,R., Mosbruger,T.L., Dahlem,T.J., Stephens,W.Z., Rao,D.S., Round,J.L. and O'Connell,R.M. (2016) Genome-wide CRISPR-Cas9 screen identifies MicroRNAs that regulate myeloid leukemia cell growth. *PLoS One*, **11**, e0153689.

152. Henser-Brownhill,T., Monserrat,J. and Scaffidi,P. (2017) Generation of an arrayed CRISPR-Cas9 library targeting epigenetic regulators: from high-content screens to *in vivo* assays. *Epigenetics*, **12**, 1065–1075.

153. Read,A., Gao,S., Batchelor,E. and Luo,J. (2017) Flexible CRISPR library construction using parallel oligonucleotide retrieval. *Nucleic Acids Res.*, **45**, e101.

154. Kim,S., Kim,D., Cho,S.W., Kim,J. and Kim,J.-S. (2014) Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.*, **24**, 1012–1019.

155. Schumann,K., Lin,S., Boyer,E., Simeonov,D.R., Subramaniam,M., Gate,R.E., Haliburton,G.E., Ye,C.J., Bluestone,J.A., Doudna,J.A., *et al.* (2015) Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 10437–10442.

156. Seki,A. and Rutz,S. (2018) Optimized RNP transfection for highly efficient CRISPR/Cas9-mediated gene knockout in primary T cells. *J. Exp. Med.*, **215**, 985–997.

157. Shifrut,E., Carnevale,J., Tobin,V., Roth,T.L., Woo,J.M., Bui,C.T., Li,P.J., Diolaiti,M.E., Ashworth,A. and Marson,A. (2018) Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, **175**, 1958–1971.e15.

158. Rautela,J., Surgenor,E. and Huntington,N.D. (2020) Drug target validation in primary human natural killer cells using CRISPR RNP. *J. Leukoc. Biol.*, **108**, 1397–1408.

159. Elmas,E., Saljoughian,N., de Souza Fernandes Pereira,M., Tullius,B.P., Sorathia,K., Nakkula,R.J., Lee,D.A. and Naeimi Kararoudi,M. (2022) CRISPR gene editing of human primary NK and T cells for cancer immunotherapy. *Front. Oncol.*, **12**, 834002.

160. Han,A.R., Shin,H.R., Kweon,J., Lee,S.B., Lee,S.E., Kim,E.-Y., Kweon,J., Chang,E.-J., Kim,Y. and Kim,S.W. (2024) Highly efficient genome editing via CRISPR-Cas9 ribonucleoprotein (RNP) delivery in mesenchymal stem cells. *BMB Rep.*, **57**, 60–65.

161. Lin,S., Staahl,B.T., Alla,R.K. and Doudna,J.A. (2014) Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife*, **3**, e04766.

162. Heyer,W.-D., Ehmsen,K.T. and Liu,J. (2010) Regulation of homologous recombination in eukaryotes. *Annu. Rev. Genet.*, **44**, 113–139.

163. Fallon,T.K. and Knouse,K.A. (2025) A roadmap toward genome-wide CRISPR screening throughout the organism. *Cell Genom.*, **5**, 100777.

164. Ting,P.Y., Parker,A.E., Lee,J.S., Trussell,C., Sharif,O., Luna,F., Federe,G., Barnes,S.W., Walker,J.R., Vance,J., *et al.* (2018) Guide Swap enables genome-scale pooled CRISPR-Cas9 screening in human primary cells. *Nat. Methods*, **15**, 941–946.

165. Liszczak,G.P., Brown,Z.Z., Kim,S.H., Oslund,R.C., David,Y. and Muir,T.W. (2017) Genomic targeting of epigenetic probes using a chemically tailored Cas9 system. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 681–686.

166. Quan,J., Langelier,C., Kuchta,A., Batson,J., Teyssier,N., Lyden,A., Caldera,S., McGeever,A., Dimitrov,B., King,R., *et al.* (2019) FLASH: a next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. *Nucleic Acids Res.*, **47**, e83.

167. Aalipour,A., Dudley,J.C., Park,S.-M., Murty,S., Chabon,J.J., Boyle,E.A., Diehn,M. and Gambhir,S.S. (2018) Deactivated CRISPR Associated Protein 9 for Minor-Allele Enrichment in Cell-Free DNA. *Clin. Chem.*, **64**, 307–316.

168. Gootenberg,J.S., Abudayyeh,O.O., Kellner,M.J., Joung,J., Collins,J.J. and Zhang,F. (2018) Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science*, **360**, 439–444.

169. Mighell,T.L., Nishida,A., O'Connell,B.L., Miller,C.V., Grindstaff,S., Thornton,C.A., Adey,A.C., Doherty,D. and O'Roak,B.J. (2022) Cas12a-Capture: A Novel, Low-Cost, and Scalable Method for Targeted Sequencing. *CRISPR J.*, **5**, 548–557.

170. Jiang,F. and Doudna,J.A. (2017) CRISPR-Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.*, **46**, 505–529.

171. Southern,E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.*, **98**, 503–517.

172. Tofano,D., Wiechers,I.R. and Cook-Deegan,R. (2006) Edwin Southern, DNA blotting, and microarray technology: A case study of the shifting role of patents in academic molecular biology. *Genomics Soc. Policy*, **2**, 1–12.

173. Chamberlain,J.S., Gibbs,R.A., Ranier,J.E., Nguyen,P.N. and Caskey,C.T. (1988) Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Res.*, **16**, 11141–11156.

174. Zangenberg,G., Saiki,R.K. and Reynolds,R. (1999) Multiplex PCR. In *PCR Applications*. Elsevier, pp. 73–94.

175. Kebschull,J.M. and Zador,A.M. (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.*, **43**, e143.

176. Gu,W., Crawford,E.D., O'Donovan,B.D., Wilson,M.R., Chow,E.D., Retallack,H. and DeRisi,J.L. (2016) Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.*, **17**, 41.

177. Salk,J.J., Schmitt,M.W. and Loeb,L.A. (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.*, **19**, 269–285.

178. Malekshoar,M., Azimi,S.A., Kaki,A., Mousazadeh,L., Motaei,J. and Vatankhah,M. (2023) CRISPR-Cas9 targeted enrichment and next-generation sequencing for mutation detection. *J. Mol. Diagn.*, **25**, 249–262.

179. Stoler,N. and Nekrutenko,A. (2021) Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.*, **3**, lqab019.

180. Kivioja,T., Vähärautio,A., Karlsson,K., Bonke,M., Enge,M., Linnarsson,S. and Taipale,J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.

181. Jia,C., Huai,C., Ding,J., Hu,L., Su,B., Chen,H. and Lu,D. (2018) New applications of CRISPR/Cas9 system on mutant DNA detection. *Gene*, **641**, 55–62.

182. Tsui,C., Inouye,C., Levy,M., Lu,A., Florens,L., Washburn,M.P. and Tjian,R. (2018) dCas9-targeted locus-specific protein isolation method identifies histone gene regulators. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, E2734–E2741.

183. Maria,B.L., Hoang,K.B., Tusa,R.J., Mancuso,A.A., Hamed,L.M., Quisling,R.G., Hove,M.T., Fennell,E.B., Booth-Jones,M., Ringdahl,D.M., *et al.* (1997) 'Joubert syndrome' revisited: key ocular motor signs with magnetic resonance imaging correlation. *J. Child Neurol.*, **12**, 423–430.

2. Barcode-Assisted Retrieval-CRISPR Activated Targeting (BAR-CAT) is a Method for Enriching Synthetic Genes

2.1 Author contributions

I was the primary contributor to the experimental portion of this work, while Dr. Calin Plesa was primarily responsible for the computational analysis. I prepared, cloned, and submitted all gene libraries used in this study for Illumina sequencing, including one library assembled using DropSynth. All BAR-CAT experiments were performed by me, with Abigail Keller providing hands-on assistance for the experiments shown in Figures 8 and 9.

Dr. Calin Plesa and I both contributed to optimizing the BAR-CAT protocol. Dr. Plesa proposed increasing the number of bead washes for BAR-CAT version 0.2 and linearizing gene library DNA. I proposed denaturing dCas9 prior to final BAR-CAT amplification to develop version 0.3, as well as increasing the amount of input gene library DNA and evaluating incubation time to establish BAR-CAT version 1.0.

Dr. Plesa was solely responsible for analyzing all Illumina and nanopore sequencing data. He also developed the sgRNA spacer selection pipeline in collaboration with Mindy H. Tran and selected all barcode targets from gene libraries for BAR-CAT enrichment. I interpreted most of the results, with input from Dr. Calin Plesa, and was the primary writer of this work. Dr. Plesa contributed to the writing and participated in editing and revising the manuscript. Dr. Plesa designed most of the main figures, while I prepared Figures 15, 16, and 19 (Appendix B). I initially prepared Figure 22 (Appendix B), which Dr. Plesa later revised to improve its design.

2.2 Introduction

DNA libraries have been foundational to molecular biology, supporting early gene discovery, physical genome mapping, expression studies, and functional analyses, and serving as a key format in early shotgun genome sequencing (1, 2). Between 2001 and 2021, the cost of sequencing individual human genomes decreased by seven orders of magnitude, outpacing Moore's Law (3, 4). By 2024, sequencing a human genome costs approximately $200 with the Illumina NovaSeq X, compared to one million USD in 2007 (Illumina). These advances in next-generation sequencing (NGS) technologies shifted the role of DNA libraries within molecular biology. As a result, DNA libraries have become essential for large-scale applications, such as barcode lineage tracing (5), CRISPR screens (6), gene regulation studies (7), functional assays of variant effects (8, 9) and directed evolution of enzymes (10).

Modern large-scale applications increasingly rely on synthetic rather than natural sequences. Synthetic libraries are fully customizable and reflect the vast potential design space. Depending on the application, libraries may contain short inserts (<300 bp) or long inserts (>300 bp), with methods tailored accordingly. Short-insert libraries are often generated by microarray or column-based oligo synthesis to encode sequences like peptide fragments, gene exons, DNA barcodes, or guide RNAs. These pools can be PCR-amplified, cloned, or used directly without further assembly. Such libraries have driven progress in functional genomics, for example by enabling pooled CRISPR screens to map enhancer-gene relationships (6), large-scale peptide mapping to identify protein interactions in cancer genes (7), or RNA "bait" libraries for targeted exon sequencing (11).

While synthetic short-insert DNA libraries have enabled important discoveries, some applications require longer sequences. These include large-scale protein engineering, functional assays of variant effects (MAVEs) (12), synthetic genome assembly, and synthetic

metagenomics. Directed evolution experiments often require DNA libraries exceeding 300 nucleotides (nt)to encode full-length protein variants, presenting a challenge for DNA synthesis approaches. Since this length exceeds the limit of standard synthetic oligos, researchers often rely on error-prone PCR (10) or saturation mutagenesis (13) to introduce diversity. Although straightforward, these methods produce variants close to the parental templates, limiting their diversity (13, 14). Synthetic oligo libraries offer more control and broader diversity, making them particularly valuable for broad mutation scanning (BMS), a type of MAVE that assesses the functional impact of thousands of homologs across the tree of life. BMS has been used to characterize large libraries of phosphopantetheine adenylyltransferase (PPAT) homologs involved in vitamin B5 biosynthesis (15) and dihydrofolate reductase (DHFR) homologs, uncovering gain-of-function variants that confer resistance to trimethoprim with implications for antimicrobial resistance (16).

DropSynth, a high-throughput gene synthesis method, assembles up to 1,536 genes in parallel and has made BMS studies feasible by enabling access to large, diverse libraries (15–17). It assembles user-defined, microarray-derived oligos into full-length genes within emulsion droplets (17), allowing for synthetic gene libraries that are more customizable than those based on natural DNA (15). DropSynth's scalability is driven by barcoded beads that capture specific oligos. After hybridization, these beads are emulsified with reagents for polymerase cycling assembly (PCA), a PCR-like method that joins overlapping oligos and amplifies full-length genes (18). Emulsion droplets help prevent cross-hybridization between oligos from different gene assemblies. DropSynth reduces gene synthesis costs to approximately US$0.70 per kilobase pair (kbp), compared to the US$70 per kbp cost of traditional synthesis methods.

Currently, DropSynth gene assemblies are limited to around 1,000 base pairs (bp), with only about 8% of constructs perfectly assembled at the DNA level (19). This limitation stems from errors introduced during phosphoramidite oligo synthesis (POS), originally developed by Beaucage and Caruthers, which accumulates truncations, deletions, and substitutions as oligos grow longer (4, 20, 21). POS uses a four-step cycle with roughly 99% efficiency per nt addition (21), but as oligo length increases, yields decline and error rates increase (22). As a result, assembling longer genes, such as those 2,000 bp in length, likely yields less than 1% perfect constructs due to compounded oligo synthesis errors and additional errors during polymerase cycling assembly in emulsions (19). MAVEs and other multiplex functional assays perform better with higher rates of perfect assemblies. Identifying perfect constructs with very low representation (<1%) demands substantial oversampling, increasing costs (4). While commercial vendors can synthesize oligos up to 300 nt with error rates as low as 1 in 2,000 to 1 in 3,000 nt, maintaining this rate across longer sequences remains difficult (vendor-provided rates).

To address these challenges, several strategies have aimed to enrich error-free oligos during or after synthesis. One approach uses sequencing-by-synthesis (SBS) to extend oligos hybridized to universal primers until they reach the correct length, at which point a biotinylated reversible terminator is incorporated to isolate full-length products (23). Other methods leverage next-generation sequencing (NGS) to select validated oligos before assembly. For example, an earlier approach used Roche 454 sequencing with robotic retrieval of error-free oligos (24), but this has been replaced by more accurate and affordable Illumina platforms (25). To improve scalability, Schwartz and colleagues developed dial-out PCR, which adds unique adaptors to oligo pools, sequences them, and selectively amplifies perfect sequences (26). Though effective, this method is expensive and difficult to scale (4, 22). Hybridization-based enrichment offers

better scalability by melting away mismatched oligos during annealing. However, this method only modestly improves error rates, such as to 1 in 1,394 bp (27), and requires carefully designed probes to avoid capturing erroneous sequences (28). It also lacks the resolution to discriminate highly similar sequences (11).

To enable functional studies of longer genes using DropSynth and other assembly methods, it is crucial to enrich perfect full-length gene assemblies directly. A traditional approach involves cloning and validating assembled genes by Sanger sequencing (4, 29). While accurate, this method validates only one gene at a time and is labor-intensive and costly (30, 31). Nanopore sequencing enables validation of multiple plasmids at once (32), and while the identification of perfect assemblies could be automated, it still requires substantial bioinformatic processing and remains less scalable for high-throughput library screening. Dial-out PCR can also retrieve perfect genes or oligos (26, 33) but requires independent reactions and custom primers for each targeted gene, limiting scalability.

More scalable strategies for enriching perfect gene assemblies include fusing fluorescent or antibiotic resistance markers, enabling functional selection. Assemblies with premature stop codons fail to express the marker and are removed (34), though re-initiation at downstream start codons can yield truncated proteins, leading to false positives (35). Another approach uses MutS, a mismatch repair protein that binds mismatched base pairs in heteroduplex DNA (36). Beads or columns functionalized with MutS can enrich perfect assemblies up to 25.2-fold (4, 37), though this requires large amounts of heteroduplex input DNA (17, 38). Other mismatch-recognizing enzymes like T7 endonuclease I, used for perfect gene selection also require large amounts of intact heteroduplexes and do not cleave all mismatch types, as shown by persistent errors in Sanger sequencing (39). The low quantity of heteroduplex DNA produced by multiplex gene

assembly methods like DropSynth limits the application of error-correcting enzymes such as MutS and T7 endonuclease I.

Given the lack of sufficiently targeted, affordable, and scalable methods for enriching perfect gene assemblies from synthetic gene libraries, we sought to develop a solution that could not only address this gap but also have broader applications. This method must be sufficiently versatile to enrich DNA sequences beyond perfect gene assemblies, such as selectively enriching specific gene subsets from large gene libraries. This flexibility would help save time, labor, and resources. In addition to its immediate utility for DropSynth, we envisioned it as a tool with potential applications in emerging fields such as DNA data storage (40), offering a "cut and paste" mechanism for large-scale DNA manipulation.

We selected *S. pyogenes* Cas9 (SpyCas9), a well-characterized nuclease from the CRISPR gene-editing suite, as the core of our DNA enrichment method. CRISPR-Cas9 is ideal for this application due to its precision, scalability, and ability to multiplex with minimal cross-talk among targets, all driven by programmable single-guide RNAs (sgRNAs) (41–43). Its extensive characterization and broad compatibility with various tools make it a reliable choice for diverse DNA manipulation applications (44).

To minimize off-target effects associated with CRISPR-Cas9, we chose the catalytically inactive version, dCas9, which offers reduced leakiness. In a CRISPR-based high school education kit, dCas9:sgRNA complexes, known as ribonucleoproteins (RNP), reduced gene expression by 88-fold and 10-fold compared to active Cas9 RNPs (45). This reduced leakiness is crucial for enhancing specificity in gene targeting. The versatility of dCas9 is further demonstrated by its successful use in a cell-free diagnostic test to enrich rare alleles, illustrating

its potential for multiplex DNA enrichment *in vitro* (43). These features make dCas9 an ideal candidate for precise, multiplexed DNA manipulation in our approach.

Here, we introduce Barcode-Assisted Retrieval – CRISPR-Activated Targeting version 1.0 (BAR-CAT v1.0), an *in vitro* method for selectively retrieving barcoded genes from DNA libraries. BAR-CAT v1.0 requires DNA libraries to be tagged with random 20-nt PAM-adjacent barcodes, which serve as binding sites for dCas9 complexed with sgRNAs containing matching spacers. After sequencing the barcoded libraries using Illumina, the barcodes are mapped to perfect gene assemblies or other DNA sequences of interest (Fig. 7A, see 2.4 Results and Discussion). A custom computational pipeline filters spacers that may lead to dCas9 off-target binding and organizes the remaining spacers into sub-libraries. Each sub-library is transcribed *in vitro* into an sgRNA library targeting a specific gene library using a single-pot reaction with T7 RNA polymerase (T7 RNAP) (46). The sgRNAs are incubated with biotinylated dCas9 to form RNPs, which are then pulled down using streptavidin-coated magnetic beads. Non-target sequences are washed away, and the enriched genes are PCR-amplified, sequenced, and analyzed to compare barcode frequencies before and after enrichment (Fig. 7A).

To make BAR-CAT v1.0 scalable and multiplexable, we aimed to identify and extend the limits of CRISPR-dCas9 multiplexed binding. Multiplexed CRISPR editing enables simultaneous regulation of multiple loci, facilitating studies of complex phenotypes and genetic interactions (47–49). In mammalian cells, CRISPR-Cas9 has been used for multiplex knockouts of up to three targets per cell but scaling beyond this number is limited by the need to express each sgRNA from its own promoter, complicating cloning and delivery (48, 50). In contrast, cell-free systems like BAR-CAT v1.0 bypass this limitation by using *in vitro* transcribed sgRNAs added directly to CRISPR reactions, enabling more extensive multiplexing. Cell-free CRISPR-

dCas9 with in vitro transcribed sgRNAs has shown success in diagnostic applications for enriching rare alleles (43, 51). Thus, the development of BAR-CAT v1.0 provided an opportunity to examine the constraints and parameters of CRISPR-dCas9 binding in vitro, informing future optimizations and expanding the potential of multiplexed CRISPR systems.

In this study, we demonstrate that optimizations to BAR-CAT, including increased bead washes and dCas9 denaturation prior to amplification, modestly improve enrichment of 18 targeted barcodes by approximately 3-fold. The most robust improvement (median ~600-fold for three barcodes) was achieved by increasing input gene library amounts tenfold, likely improving the DNA to dCas9 ratio and reducing off-target enrichment. When applied to DropSynth-assembled dihydrofolate reductase (DHFR) libraries containing 384 or 1,536 genes, we observed up to 149-fold enrichment for 12 targets. However, at larger library sizes, enrichment became inconsistent, and target dropout rates increased. These results suggest further optimization strategies and provide insights for scaling BAR-CAT and other *in vitro* CRISPR methods to improve multiplexing and reduce off-target effects.

2.3 Materials and Methods

Constructing the pEVBC3 barcoding plasmid for the *rfp* library

Plesa and colleagues developed pEVBC1, a pUC19-derived plasmid that adds 20-mer barcodes to cloned genes and expresses them in *E. coli* via a pLac-UV5 promoter (15). Since pEVBC1 was available in-house, a new plasmid, designated pEVBC3, was derived from pEVBC1 to develop a pilot gene library using PCR-round-the-horn. Four alternative PAM sites, an EcoRI restriction site, and the pLac-UV5 promoter were removed.

To prepare the vector backbone, the parent plasmid pEVBC1 was linearized by double digestion with XbaI and KpnI-HF (New England Biolabs, Ipswich, MA, USA), which removed the *mcherry* cargo gene. Each 50 microliter (µL) reaction contained 500 nanograms of pEVBC1, 5 microliters of 10X CutSmart Buffer (NEB), 1.0 microliter of XbaI (20,000 units per mL), 1.0 microliter of KpnI-HF (20,000 units per mL), and nuclease-free water. The digestion was incubated at 37°C for one hour and heat inactivated at 65°C for 20 minutes. The linearized plasmid was purified using DNA Clean and Concentrator columns (5 microgram capacity, Zymo Research Corp, Irvine, CA, USA). A 1,869 base pair product corresponding to the digested vector was verified by 1% agarose gel electrophoresis alongside a 1 kbp DNA ladder (New England Biolabs).

To build pEVBC3, A series of three consecutive PCR protocols was performed to complete PCR round-the-horn, following the method previously described in prior work (15). The first PCR aimed to modify the linearized pEVBC1 into pEVBC3. Eight PCR reactions were set up, each containing 0.5 ng of linearized pEVBC1, 0.5 µL each of 10 µM forward (EVBC3_FWD1) and reverse (EVBC3_REV1) primers, 25 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to a total volume of 50 µL. The thermocycler PCR program included an initial denaturation at 98°C for 30 seconds, followed by 35 cycles of 98°C for 10 seconds, annealing at 72°C for 30 seconds, extension at 72°C for 1 minute per cycle, and a final extension at 72°C for 2 minutes. The resulting 1,950 bp pEVBC3 backbone (PCR 1 product) was purified using 5 µg DNA Cleanup Columns (Zymo Research Corp, Irvine, CA).

The purpose of the second PCR was to add a randomly generated 20-mer sequence to barcode genes cloned into pEVBC3. Each individual PCR was prepared with 1 ng of PCR 1 product, 2.5 µL each of 10 µM forward (EVBC3_FWD2), and reverse (EVBC3_REV2) primers,

25 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to a final volume of 50 µL. The same thermocycler PCR program was used as in PCR 1, except that only five amplification cycles were performed. The resulting 2 kbp barcoded pEVBC3 backbone (PCR 2 product) was purified using 5 µg DNA Cleanup Columns (Zymo Research Corp, Irvine, CA).

The purpose of the third PCR was to bulk amplify the barcoded pEVBC3 backbone in preparation for generating large-scale diversity in gene libraries. Each individual PCR was prepared with 1 µL of PCR 2 product, 2.5 µL each of 10 µM forward (EVBC3_FWD3RE), and reverse (EVBC3_REV2) primers, 25 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to a total volume of 50 µL. The same thermocycler PCR program was used as in PCR 1, except that only 15 amplification cycles were performed. The amplified pEVBC3 backbone (PCR 2 produ993 bp PCR 3 product was verified with a 1% agarose gel and 1 kbp ladder (New England Biolabs, NEB, Ipswich, USA). The pEVBC3 backbone (2.7 kbp) was verified via Sanger Sequencing following ligation to an *mcherry* gene, transformation into NEB® 5-alpha Competent *E. coli*, and colony picking.

Amplification of *mcherry* red fluorescent protein (*rfp*) gene for Pilot Library Generation

Bulk amplification of a 708 bp *mcherry* gene was performed to obtain sufficient insert material for generating a single gene *rfp* library. The library was later barcoded using pEVBC3. The original *mcherry* sequence was sourced from pZS2-123, a plasmid gifted to Dr. Sri Kosuri from Michael Elowitz (Addgene plasmid #26598; http://n2t.net/addgene:26598; RRID: Addgene_26598). The Kosuri group derived pSK48, the plasmid containing amplified *mcherry* used in this work, from its parent plasmid pZS2-123.

In total, 30 PCRs were performed to obtain 5-10 µg of bulk amplified, biotinylated *mcherry* for downstream large-scale diversity generation. Each reaction contained 1.0 ng of

pSK48, 2.5 µL each of 10 µM forward (RFP_KpnI_FWD_Biotin_NV) and reverse (RFP_Ndel_REV_Biotin_NV) primers, 25 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to a total volume of 50 µL. The thermocycler PCR program included an initial denaturation at 98°C for 30 seconds, followed by 35 cycles of 98°C for 10 seconds, annealing at 72°C for 30 seconds, extension at 72°C for 30 minute per cycle, and a final extension at 72°C for 2 minutes. The *mcherry* PCR product was purified using 5 µg DNA Cleanup Columns (Zymo Research Corp, Irvine, CA).

Constructing the pEVBC8 barcoding plasmid for DropSynth dihydrofolate reductase (DHFR) libraries

The main modification made to the pEVBC3 barcoding plasmid was the addition of a protospacer adjacent motif (PAM) site to the 3′ end. To prepare the vector backbone, pEVBC3 containing an *mcherry* insert (pEVBC3-*rfp*) and the barcode TAAATATTACct) was purified using 5 µg DNA Cleanup Columns (Zymo Research Corp, Irvine, CA). The 1,CGGTCTCTTT, as verified by Sanger sequencing, was linearized with KpnI-HF® (New England Biolabs, Ipswich, USA) and NdeI (NEB) restriction enzymes. Each digestion included 666 ng of pEVBC3, 5 µL of 10X CutSmart® Buffer (NEB), 1.0 µL of KpnI-HF® (20,000 units/mL) (NEB), 1.0 µL of NdeI (20,000 units/mL) (NEB), and nuclease-free water to a total volume of 50 µL. The digest was incubated in a thermocycler at 37°C for 1.5 hours and heat inactivated at 65°C for 20 min. The products were purified using 5 µg DNA Cleanup Columns (New England Biolabs, Ipswich, USA). The 2 kbp pEVBC3 backbone was size-selected from a 1.0% agarose gel using the Monarch® DNA Gel Extraction Kit (NEB), with a 1 kbp ladder (NEB) for comparison.

To build pEVBC8, we used the same PCR-round-the-horn protocol for producing pEVBC3 but with a few modifications. The first PCR included 1 ng of linearized pEVBC3, a forward primer (EVBC8_FWD1), and 0.625 µL of each 10 µM primer (6.25 µM final concentration). The thermocycler program used was identical to that used in the first PCR for generating pEVBC3. The 1,950 bp PCR 1 product was size-selected from a 1.0% agarose gel using the Monarch® DNA Gel Extraction Kit (NEB). Modifications to the second PCR included the addition of a forward primer (EVBC8_FWD2) to eliminate a potential alternative PAM site on the 5' end of the barcode. Additionally, 0.625 µL of each 10 µM primer (6.25 µM final concentration) was added. The third PCR was performed as described for generating pEVBC3, however, 60 PCRs were prepared and 1 ng of PCR 2 product was added instead of 1 µL. The final amplified pEVBC8 backbone was validated with agarose gel electrophoresis (1.0% agarose gel) and nanopore sequencing (Plasmidsaurus).

DropSynth assembly of DHFR gene libraries and amplification

Four DHFR DropSynth libraries were assembled and amplified to obtain gene library insert material, later barcoded using pEVBC8. Library S4 consisted of 384 genes, ranging from 473 to 545 bp in length, assembled from Twist Bioscience 300-mer oligos. In contrast, library S2 and S3 contained 1,536 genes, ranging from 464 to 548 bp in length, assembled from HiFi SurePrint High Fidelity (HiFi) 230-mer oligos (Agilent, Santa Clara, USA). All libraries were synthesized using DropSynth 2.0, with library S2 assembled by a prior study (17) and library S4 assembled for this study.

Prior to bulk amplification of the 384-gene DHFR library (S4), qPCR (CFX Opus 96, Bio-Rad) was performed to determine the number of PCR cycles corresponding to the amplification plateau. Each qPCR included 1 ng of assembled library (S2), 2.5 µL of 10 µM of

biotinylated forward primer (skpp504F), 2.5 µL of 10 µM of biotinylated reverse primer (skpp504R), 0.25 µL of 100X Biotium Thiazole Green (Thermo Fisher Scientific), 12.5 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to a total volume of 25 µL. The qPCR program included initial denaturation at 95°C for 3 min, followed by 40 cycles of 98°C for 15 seconds, annealing at 65°C for 30 seconds, and extension at 72°C for 15 seconds.

In total, sufficient PCRs were prepared to yield 5-10 µg of biotinylated library S4 for downstream large-scale diversity generation. Bulk amplification PCRs were prepared using the same setup as described for preparing the qPCRs. A 1 min extension step at 72°C was added to the thermocycler PCR protocol used for qPCR and library S4 was amplified for 12 cycles. The amplified products were purified with 5 µg DNA Cleanup Columns (NEB).

The 1,536-gene DHFR (S2) fragments were PCR-amplified to add standard Illumina adapters and sequenced on a MiSeq. Amplicons were subsequently re-amplified using biotinylated P5 and P7 primers. To determine the optimal number of PCR cycles, qPCR reactions were prepared using 1 ng of library S2 amplicons, 1.25 µL of 10 µM biotinylated forward primer (P5_FWD_Biotin_NV), 1.25 µL of 10 µM biotinylated reverse primer (P7_REV_Biotin_NV), 0.25 µL of 100× Biotium Thiazole Green (Thermo Fisher Scientific), 12.5 µL of 2× Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to a final volume of 25 µL. The qPCR program consisted of an initial denaturation at 98 °C for 30 s, followed by 40 cycles of 98 °C for 10 s, 70 °C for 30 s, and 72 °C for 30 s.

Bulk amplification was then performed until 5–10 µg of biotinylated DNA was obtained for downstream large-scale diversity generation. Bulk PCR reactions were prepared identically to the qPCR reactions described for library S4, except that the thermocycler protocol included a

2 min final extension at 72°C. Library S2 was amplified for 24 cycles, and the resulting products were purified using 5 µg DNA Cleanup Columns (NEB).

Generation of large-scale diversity of barcoded gene libraries

Barcoded gene libraries were digested, ligated into barcoding plasmids, and transformed into electrocompetent *E. coli* with high efficiency to generate large-scale library diversity. This approach was adapted from the methodologies of Plesa and colleagues (15).

First, the pEVBC barcoding plasmids and gene library inserts, such as *mcherry* or DHFR libraries, were digested with KpnI-HF® and NdeI restriction enzymes to generate compatible sticky ends. Each digest contained 666 ng of pEVBC backbone or 333 ng of gene insert, 5 µL of 10X CutSmart® Buffer, 1 µL of KpnI-HF® (20,000 units/mL) (NEB), 1 µL of NdeI (20,000 units/mL) (NEB), and nuclease-free water to a total volume of 25 µL. The pEVBC digests also included 1 µL of Shrimp Alkaline Phosphatase (rSAP, 1,000 units/mL) (NEB) to prevent plasmid recircularization. Digestions were incubated at 37°C for 1.5 hours, followed by heat inactivation of NdeI and rSAP at 65°C for 20 minutes. The digested products were then purified using 5 µg DNA Cleanup Columns (NEB).

Next, uncut biotinylated molecules were removed with streptavidin magnetic beads to concentrate digested products. Streptavidin Magnetic Beads (NEB, Ipswich, USA), ranging from 30 to 60 µL at a concentration of 4 mg/mL, were prepared by removing the storage buffer and performing equilibration washes according to the manufacturer's instructions. The quantity of beads used per digest was determined based on their binding capacity of 1 mg per 500 pmol of 25 nt single-stranded DNA, as specified by the manufacturer. This amount was adjusted according to the size and quantity of the backbone or insert. The equilibrated beads were then resuspended with twice the original volume of 2X binding and wash buffer (2X B&W buffer;

2M NaCl, 1mM EDTA, 10mM Tris) and added to the digested DNA in 1.5 mL tubes. The bead capture samples were incubated on a thermomixer at RT for 30 min with shaking at 1700 RPM.

After incubation, a DynaMag-2 magnetic rack was used for 1 min to separate the beads. The supernatant fraction was transferred to a clean 1.5 mL tube, and the beads were subsequently resuspended in 50 µL of 2X B&W buffer. The magnetic bead capture process was repeated three times. The supernatant fractions containing cut non-biotinylated DNA were purified with 5 µg DNA Cleanup Columns (NEB).

The digested pEVBC backbones and gene library insert sticky ends were ligated together using T4 DNA ligase at a 1:4 plasmid to insert ratio. For each gene library, 4-6 ligation reactions were prepared. Each full ligation (FL) assembly reaction consisted of 0.05 pmol of digested pEVBC backbone combined with approximately 0.2 pmol of gene library insert, except for the no insert control (NIC). The ligation mixtures included 4 µL of 10X T4 DNA ligase reaction buffer (NEB) supplemented with a final concentration of 1 mM ATP (New England Biolabs, NEB, Ipswich, USA), 2 µL T4 DNA ligase (400,000 units/mL) (NEB, Ipswich, USA), and nuclease-free water to a final volume of 20 µL. The ligations were incubated overnight (~22 hours) at 16°C in a thermocycler, followed by purification using 5 µg DNA Cleanup Columns (NEB). The ligated products were eluted in the minimum volume of Monarch® DNA Elution Buffer (NEB) possible for the DNA columns, typically 7 µL, to achieve a final concentration between 50 and 100 ng/µL.

To remove guanidinium salt contaminants from FL and NIC products, drop dialysis was performed. Each ligated gene library was pipetted onto a 50 nm pore nitrocellulose membrane floating on deionized water contained in an 150 x 15 mm petri dish (Kord-Valmark Labware

Products, Bristol, USA). After 20 minutes of incubation, each ligation was transferred to a clean 1.5 mL tube and quantified.

The ligated products were then transformed into *E. coli* to generate high-diversity libraries. FL and NIC ligated products were initially transformed into NEB® 5-alpha Electrocompetent *E. coli,* and later into NEB® 10-beta Electrocompetent *E. coli* once the 5-alpha cells were discontinued (NEB). Electrocompetent cells were thawed and 25 µL aliquots were distributed to 1.5 mL tubes chilled on ice. One to two microliters of 50-100 ng of ligated products were added to each aliquot of cells. Three cell aliquots received each individual FL product, while NIC and the supercoiled pUC19 (Bayou Biolabs, Metairie, USA) transformation positive control were each added to separate aliquots. After flicking each 1.5 mL tube to mix the DNA and cells, the mixture was loaded onto Gene Pulser/MicroPulser Electroporation Cuvettes (Bio-Rad Laboratories Inc, Hercules, USA). After BioRad Micropulser electroporation, the cells were resuspended in Super Optimal Broth with Catabolite Repression Medium (NEB) and incubated at 37°C for 1 hour in an Innova shaking incubator.

All three recovered FL transformants, initially diluted to 1:10 from the original cell aliquots, were combined. A 30 µL aliquot of FL transformants was then diluted 10-fold to prepare a 1:100 dilution. This 1:100 dilution underwent four additional 10-fold serial dilutions to prepare colony counting plates. Similarly, a dilution series was prepared for the single aliquot of pUC19 transformants. In contrast, 100 µL of NIC transformants were serially diluted 10-fold and 100-fold to achieve 1:100 and 1:1000 dilutions, respectively. Finally, 100 µL of each prepared dilution was spread-plated in duplicate on Lysogeny Broth (LB) agar supplemented with 0.1 mg/mL carbenicillin. Additionally, large-scale transformed libraries were plated by adding 300

µL of 1:10 dilution FL transformants to 5-10 pre-warmed 150 x 15 mm plates containing LB and 0.1 mg/mL carbenicillin. All plates were then incubated overnight at 37°C.

Colonies from the dilution series were counted to determine CFU/mL for FL, NIC, and pUC19 transformations. The FL dilution series yielded at least 1.0 x 10^6 CFU/mL, with large-scale transformations showing lawn growth, indicating sufficient diversity in the transformed FL gene libraries. Additionally, the NIC CFU/mL, consistently lower by several orders of magnitude compared to FL, suggested minimal background from self-ligated pEVBC backbone.

To harvest FL gene library bacterial lawns, 5-10 mL of LB was dispensed onto each large 150 x 15 mm plate and the bacterial lawns were scraped with a sterile spreader. The scraped lawns were transferred into a 50 mL tube and the OD600 was measured. Multiple glycerol stocks of both undiluted and diluted (OD 4-7) cells were stored at -80°C. The remaining diluted cells were partitioned into 5 mL aliquots, spun down into pellets, and stored at -80°C. Supercoiled gene libraries were isolated from the scraped cells by thawing one or two pellets and using the Monarch® Plasmid Miniprep Kit protocol or a midiprep kit.

Indexed amplification of the barcoded *rfp* gene library for MiSeq

Primers designed to add Illumina adapters and a sequencing index to the barcoded library were ordered from Integrated DNA Technologies (IDT, Coralville, USA). The forward primer (mi7_FWD_Amp_NV) annealed to the 5’ end of the barcode, incorporating a sequence buffer and P5 Illumina adapter. The reverse primer (mi7_REV_Amp4_NV) annealed 458 bp downstream of the barcode, adding a P7 Illumina adapter, sequence buffer, and Illumina N701 index 1. The resulting PCR product was 586 bp and included the barcode without *rfp*.

qPCR (CFX Opus 96, Bio-Rad) was performed to determine the optimal number of cycles needed to amplify the barcode region of the *rfp* library. Each qPCR was prepared with 1

ng of scraped *rfp* library, 2 µL each of 100 nM forward and reverse primers, 10 µL of 2X Kapa Fast SYBR Green (Roche), and nuclease-free water to 20 µL. The qPCR program included initial denaturation at 98°C for 30 seconds, followed by 40 cycles of 98°C for 10 seconds, annealing at 60°C for 30 seconds, and extension at 72°C for 20 seconds.

Ten PCRs were prepared to amplify the barcode product for sequencing. Each PCR contained 1 ng of scraped *rfp* library, 2.5 µL of 10 µM forward and reverse primers, 25 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to 50 µL. The thermocycler protocol included a 2-minute extension at 72°C, with amplification proceeding for 24 cycles as determined by qPCR. The amplified products were purified with 5 µg DNA Cleanup Columns (NEB) and size-selected at the 600 bp band from a 2.0% agarose gel using the Monarch® DNA Gel Extraction Kit (NEB). Purified products were submitted to the UO Genomics & Cell Characterization Core Facility (GC3F) for sequencing. Additionally, 30 µL of 100 µM custom sequencing primers for read 1 (Mi7_R1_NV), read 2 (Mi7_R2_NV), and the index read (Mi7_Rindex1_NV) were submitted. For Miseq sequencing, 25 million paired end reads with a 75 bp read length were requested.

Indexed amplification of the barcoded DHFR gene libraries for MiSeq

We ordered primers from Integrated DNA Technologies (IDT; Coralville, USA) to append Illumina adapters and sequencing indices to two barcoded DHFR libraries assembled by DropSynth. The forward primer (mi9_FWD_Amp_NV) annealed a region adjacent to the DHFR 5’ ends, and reverse primers (mi9_REV_Amp_Index#_NV) added N701, N702, N704, and N705 indices to libraries S1, S2, S3, and S4. The resulting PCR products were between 600 and 700 bp in length since they included DHFR genes of various lengths and barcodes.

qPCR (CFX Opus 96, Bio-Rad) was performed on all four libraries to optimize cycle numbers for DHFR gene and barcode amplification. Each qPCR was prepared with 1 ng of scraped pEVBC8-DHFR library, 2.5 µL each of 10 µM forward and reverse primers, 25 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to 50 µL. The qPCR program included initial denaturation at 98°C for 30 seconds, followed by 40 cycles of 98°C for 10 seconds, annealing at 72°C for 30 seconds, and extension at 72°C for 60 seconds.

Four PCRs were prepared to bulk amplify each library for sequencing using the same protocol as the qPCR. However, library S4 was amplified for 16 cycles and library S2 for 20 cycles. Amplified products were purified using 5 µg DNA Cleanup Columns (NEB) and size-selected from a 2.0% agarose gel with the Monarch® DNA Gel Extraction Kit (NEB). Purified products and 30 µL of 100 µM custom sequencing primers (Mi9_R1_NV, Mi9_R2_NV, Mi9_EVBC8_Rindex_NV) were submitted to Admera Health for 2 x 300 bp paired-end sequencing reads. Due to low Miseq sequencing depth, samples were resubmitted for sequencing.

Illumina library sequencing

We mapped barcodes to gene variants using Illumina Miseq runs analyzed with a custom pipeline described elsewhere (16). Briefly, reads were merged using bbmerge (bbmap 38.18), and a custom python script was used to extract barcodes and variable gene regions. Barcodes were collapsed using Starcode spheres algorithm on distance of 1 (52). A majority call was used to determine the consensus sequence linked to each barcode when it had multiple reads. All required scripts are available on the lab's GitHub repository (https://github.com/PlesaLab/BC_Mapping).

sgRNA spacer selection pipeline

The computational spacer selection pipeline converts long-read amplicon sequencing data from a barcoded gene library into a set of single-guide RNAs (sgRNAs) for CRISPR-dCas9 that (i) cover every targeted gene, (ii) avoid recognizing any non-target sequence and (iii) are normalized so that each gene targeted by an sgRNA has a similar number of sequencing reads (Fig. 23, see Appendix A). The code is written in R 4.3.0 with several small helper scripts in Python. Sequencing reads are first mapped to their corresponding 20-nucleotide barcodes as described previously (REF) and collapsed with Starcode (spheres distance ≤ 1) to remove PCR and sequencing errors (52). Each mapped insert is then expanded by appending 20bp of the constant sequences that flank the variable region in the corresponding vector, recreating the entire cloning context. This allows detection of spacers spanning the junction between the constant and variable regions. The full plasmid sequence itself is also loaded so that spacers targeting backbone DNA can be excluded at a later step. Barcodes whose corresponding translated open reading frame exactly matches an entry in the gene design file are provisionally labelled good; all others, together with barcodes identified in the vector or those with undetermined N bases, are labelled bad. We also keep track of the abundance of each unique barcode in the pooled library. For every read, the code searches for all occurrences of the sequence motif N20NGG where NGG represents the canonical SpCas9 PAM. A sliding-window regular expression returns every overlapping 20-nucleotide protospacer immediately upstream of an NGG, on both strands. The corresponding barcode sequence is appended to the protospacer list and labeled as either inside the gene region or directly on the barcode. Off-target potential is judged by a 16-nucleotide (adjustable) seed defined as the 3′ end of each 20-mer spacer. We track two kinds of collisions. A seed collision occurs when the same 16-mer is found in at least one good and one bad read. A full-length collision is recorded when an entire 20-mer is shared

between the two classes. Protospacers involved in full-length collisions are immediately discarded; seed collisions can be retained or removed in later filtering steps depending on the desired stringency. Because sgRNAs are ultimately cloned by Golden-Gate assembly, any protospacer with flanking context sequence which recreates a BsaI recognition site is excluded. A barcode is designated good if it possesses at least one protospacer that is free of full-length collisions, passes the BsaI filter, and if required, is free of 16-mer seed collisions.

Barcode counts within a gene can vary by more than two orders of magnitude and could potentially bias dCas9 capture. To target molecules with similar abundance, the pipeline selects up to three barcodes per gene whose abundance is closest to a single global target value. First, the median barcode read count is calculated for every gene. The medians from all genes are then combined and the median of that distribution is taken as the desired global target count. Within each gene the absolute difference between every barcode's read count and this target is computed. Barcodes whose counts exactly match the target are accepted directly. If fewer than three such barcodes exist, the remaining positions are filled by barcodes in ascending order of their absolute deviation from the target until either three barcodes have been chosen or no candidates remain. The effectiveness of this normalization is evaluated by plotting Lorenz curves and computing the Gini Coefficient. In the test 1,384 DHFR dataset, the Gini Coefficient dropped from 0.63 before normalization to 0.14 afterwards. The complete codebase, including the small Python utilities, is available at the Plesa Lab GitHub repository (https://github.com/PlesaLab/).

Examining the Impact of Magnetic Bead Washes on Removal of Non-enriched DNA

Before enriching the pilot library (pEVBC3-*rfp*) with BAR-CAT 0.1, the effect of magnetic bead washes on DNA recovery was assessed to establish optimal washing conditions.

Due to the unconfirmed functionality of biotinylated dCas9 for enrichment, EnGen® Spy dCas9 (SNAP-tag®) (NEB) and SNAP-Capture Magnetic Beads (NEB) were used for CRISPR enrichments. Each enrichment reaction, except the no library DNA control, included 50 ng of the pilot library (pEVBC3-*rfp*). Following enrichment, beads were washed 15 times with 1 mL of immobilization buffer (IB) (Table S2, see Appendix A) according to the NEB protocol for SNAP-capture magnetic beads. Supernatant from washes 3, 6, 9, 12, and 15, including the no template control, was collected.

Quantitative PCR (qPCR) and Agarose Gel Electrophoresis Analysis

Quantitative PCR (qPCR) was performed to compare Cq values across washes. Reactions were prepared using the supernatants collected from dCas9 capture bead washes, with beads from wash 15 included as a positive control. Each qPCR reaction contained 3.75 µL of wash supernatant or SNAP-tag dCas9 capture beads, 1.25 µL each of 10 µM forward and reverse primers, 0.25 µL of 100X Biotium Thiazole Green (Thermo Fisher Scientific), 12.5 µL of 2X Q5 High-Fidelity Master Mix (NEB), and nuclease-free water to a total volume of 25 µL, following the Illumina-indexed primer amplification protocol for MiSeq of the pEVBC3-*rfp* library. Since qPCR Cq values were uninformative, the 586 bp amplified products were purified using Monarch® 5 µg DNA Cleanup Columns (NEB). Purified DNA was visualized with SYBR™ Safe DNA Gel Stain (Thermo Fisher) following agarose gel electrophoresis on a 2% gel, using a 100 bp ladder (NEB) as a reference. DNA band intensities were quantified using ImageJ (1).

qPCR Evaluation of Buffer Composition on the Stringency of Streptavidin-Coated Magnetic Bead Washes

Prior to the development of BAR-CAT 0.1, the ability of biotinylated dCas9 to mediate enrichment was unverified. To assess and improve wash stringency, four buffer compositions

compatible with streptavidin-coated magnetic beads were evaluated (see Table 2 in Appendix A). Buffers containing high salt (2 M NaCl) or an anionic detergent (10% NP-40) were predicted to enhance wash stringency.

Streptavidin-coated magnetic beads (Sigma) were equilibrated according to the manufacturer's instructions, then resuspended in 2X B&W buffer (2 M NaCl, 1 mM EDTA, 10 mM Tris-HCl, pH 7.4) at twice the original volume. Ten microliters of equilibrated beads were transferred to individual 1.5 mL tubes.

Each bead aliquot received 1 µL of 10 ng/µL pEVBC3-*rfp* plasmid containing barcode 12 (BC12). Beads were incubated for 30 minutes at 37 °C with shaking at 1700 rpm. Following incubation, beads were washed either six times (for RT and 30 °C conditions) or three times (for RT) with 50 µL of the corresponding buffer. Washes were performed using magnetic separation or vortexing followed by magnetic separation between steps. After the final wash, beads were resuspended in 10 µL of nuclease-free water.

To assess wash efficiency, qPCR was performed to quantify the remaining pEVBC3-*rfp* BC12 template in the final wash supernatants. qPCRs for each wash condition were prepared using either 3 µL or 3.3 µL of final wash supernatant as template. Each reaction included 0.4 µL of 10 µM forward (qPCR_RFP_NV_FWD) and reverse (qPCR_RFP_NV_REV) primers, 10 µL of KAPA SYBR Fast 2X Master Mix (Roche), and nuclease-free water to a final volume of 20 µL. A positive control (10 ng pEVBC3-*rfp* BC12) and no-template controls were included.

BAR-CAT v0.1: Proof-of-concept enrichment of 18 targeted barcodes from the *rfp* library

CRISPR ribonucleoproteins (RNPs) were assembled using an 18-plex sgRNA library *in vitro* transcribed elsewhere (46). Each RNP assembly reaction contained 3.0 µL of 10X NEBuffer r3.1 (New England Biolabs), 3.0 µL of 300 nM sgRNAs (final: 33.3 nM), 1.0 µL of 1

µM dCas9-3xFLAG-Biotin (Sigma-Aldrich; final: 33.3 nM), 0.75 µL of Murine RNase inhibitor (40,000 U/mL, NEB), and nuclease-free water to a final volume of 27 µL. Reagents were added in the order listed. Reactions were mixed, pulse-spun, and incubated at 25 °C for 10 min, followed by 10 min at 37 °C in a thermocycler. RNPs were used immediately for downstream enrichment.

To each RNP reaction, 1.0 µL (50 ng) of supercoiled pEVBC-ligated gene library was added, and the total volume was adjusted to 30 µL with nuclease-free water. A negative control was prepared by adding 3.0 µL of water to an RNP-only reaction. Reactions were pulse-spun and incubated at 37 °C for 15 min.

Streptavidin magnetic beads (NEB, 4 mg/mL) were equilibrated per manufacturer's protocol and resuspended in 2X B&W buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl). To each enrichment reaction, 10 µL of equilibrated beads were added. Capture of dCas9 was performed at 37 °C for 30 min in a thermomixer at 1700 rpm. Beads were collected on a DynaMag-2 rack (Thermo Fisher Scientific), and the supernatant was discarded. Beads were washed six times with 50 µL of 2X B&W buffer, transferring the beads to fresh tubes before the sixth wash. Beads were resuspended in 60 µL of nuclease-free water and stored at 4 °C prior to PCR.

To estimate optimal amplification cycles, duplicate 50 µL qPCRs were prepared for each condition. Reactions contained 10 µL of washed beads, 2.5 µL of 10 µM forward (Mi7_FWD_Amp_NV) and reverse (Mi8_REV_Amp4_Index2_NV) primers, 0.5 µL of 100X Biotium Thiazole Green (Thermo Fisher), 25 µL of 2X Q5 High-Fidelity Master Mix (NEB), and water to 50 µL. Thermocycling was performed on a CFX Opus 96 machine (Bio-Rad) using

the following program: 98 °C for 30 s; 40 cycles of 98 °C for 10 s, 72 °C for 30 s, and 72 °C for 2 min.

Bulk amplification of enriched DNA was carried out using the same conditions and number of cycles determined by qPCR. Amplified products were cleaned using Monarch® 5 µg DNA Cleanup Columns (NEB). The 586 bp enrichment products were size-selected using 0.8% SYBR Safe E-Gels (CloneWell™ II, Thermo Fisher Scientific) with a 1 kbp Plus DNA Ladder (NEB) as reference. Final gel-purified DNA was concentrated using new Monarch® 5 µg Cleanup Columns and submitted to Plasmidsaurus (Eugene, USA) for nanopore sequencing.

BAR-CAT v0.2: Enrichment of 18 targeted barcodes from the *rfp* library with optimized bead washes

To improve removal of non-enriched barcodes, BAR-CAT v0.2 incorporated increased wash volumes and additional wash steps following dCas9 capture by streptavidin magnetic beads. All reactions used the barcoded *rfp* gene library and were assembled as described for BAR-CAT v0.1, including RNP formation, target DNA addition, and bead capture.

Following bead capture, each reaction was transferred to a 5 mL tube and placed on a 15 mL tube magnetic rack (Sergi Lab Supplies, Seattle, WA, USA) for high stringency washing. After 1 minute on the rack, the supernatant was removed, and beads were resuspended in 2 mL of 2X B&W buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl). Beads were incubated for 1 minute per wash. A total of nine washes were performed per sample, with the beads transferred to fresh tubes prior to the sixth wash to reduce carryover. After the final wash, beads were resuspended in 60 µL of nuclease-free water, vortexed at 2700 rpm, and centrifuged briefly to collect the beads. Samples were stored at 4 °C prior to amplification. qPCR, bulk PCR, cleanup, gel extraction, and nanopore sequencing were performed as described for BAR-CAT

v0.1.

BAR-CAT v0.3: Denaturation and removal of dCas9 prior to amplifying enriched DNA

BAR-CAT v0.3 introduced a dCas9 denaturation step after bead washes to release enriched DNA into the supernatant for downstream amplification. This eliminated the need to amplify directly from the beads, reducing the risk of non-specific DNA carryover. We also evaluated whether DNA format affected enrichment by comparing results from linearized and supercoiled versions of the *rfp* barcode library. All enrichments targeted 18 barcodes from the barcoded *rfp* gene library. dCas9 capture by magnetic beads was performed as described for BAR-CAT v0.1.

To generate linearized *rfp* gene library DNA, we first determined the optimal amplification cycle number by qPCR using 10 µM forward (Mi9_FWD_Amp_NV) and reverse (Mi9_REV_amp_NV) primers, as outlined in the section Indexed amplification of the barcoded DHFR gene libraries for MiSeq. Bulk amplification was then performed using 14 cycles to avoid overamplification. The 827 bp PCR products were purified using Monarch® 5 µg DNA Cleanup Columns (NEB) and verified on 1% agarose gels with a 100 bp Plus DNA Ladder (NEB).
For CRISPR enrichment using either linearized or supercoiled *rfp* libraries, bead washes were performed as described for BAR-CAT v0.2: nine washes in 2 mL of 2X B&W buffer using 5 mL tubes on a 15 mL magnetic rack. After the final wash, beads were resuspended in 50 µL of nuclease-free water.

To denature dCas9 and release enriched DNA from supercoiled library samples, we incorporated a proteinase K digestion step. For each reaction, 25–50 µL of dCas9 capture beads were combined with 1 µL of proteinase K (800 U/mL, NEB) in 1.5 mL tubes. Nuclease-free water was added to a final volume of 50 µL. The mixtures were vortexed, spun down, and

incubated at room temperature for 10 minutes in a thermomixer. Samples were then placed on a DynaMag-2 magnetic rack (Thermo Fisher Scientific, Waltham, USA), and the supernatant was collected for downstream purification using Monarch® 5 µg DNA Cleanup Columns (NEB). Enriched DNA was eluted in 50 µL of DNA elution buffer and amplified by qPCR and PCR. Alternative dCas9 denaturation methods were also tested but not adopted in the final BAR-CAT v0.3 protocol due to lower performance. These included: (1) boiling beads at 65 °C for 5 minutes in a thermomixer at 1700 rpm, and (2) treatment with 200 µL of 8 M urea for 30 minutes at room temperature, as described by Aalipour and colleagues (43). For each, the supernatant was collected post-treatment. Urea-treated samples were further purified using Monarch® DNA Cleanup Columns before amplification.

To determine appropriate PCR cycle numbers for enriched DNA, qPCR was performed for all conditions, including negative controls, using Mi9_FWD_Amp_NV and Mi9_REV_amp_NV primers as described in the section Indexed amplification of the barcoded DHFR gene libraries for MiSeq. Bulk amplification was then conducted using cycle numbers derived from qPCR. Final PCR products were purified using Monarch® 5 µg DNA Cleanup Columns (NEB).

Enriched 827 bp products were size-selected using CloneWell™ II Agarose Gels with SYBR Safe 0.8% E-Gels™ (Thermo Fisher Scientific) and a 1 kbp ladder (NEB). In later experiments, 2% agarose gels and the Monarch® Plasmid Miniprep Kit were used for size selection. All enriched products were submitted to Plasmidsaurus (South San Francisco, USA) for nanopore sequencing.

BAR-CAT v1.0: Optimized enrichment of three target barcodes using increased *rfp* gene library input

BAR-CAT v1.0 achieved the best barcode enrichment by incorporating a 10-fold increase in input DNA from the barcoded *rfp* gene library, thereby providing more target molecules for RNP binding. Enrichments were performed using both the supercoiled *rfp* gene library and later barcoded DropSynth DHFR libraries.

CRISPR RNPs were prepared as previously described, using three pooled synthetic sgRNAs (Integrated DNA Technologies) for enrichment from the *rfp* gene library, or a single synthetic sgRNA or *in vitro* transcribed sgRNA libraries prepared as described by Villegas and colleagues (46) for DropSynth DHFR enrichment.

To perform BAR-CAT enrichment, 500 ng of supercoiled barcoded gene library (up to 3 µL) was added to 27 µL of RNPs in qPCR tubes, then brought to a final volume of 30 µL with nuclease-free water. A negative control reaction contained 3 µL of water in place of library DNA. All reactions were mixed, pulse-spun, and incubated at 37 °C for 15 minutes.

For dCas9 pull-down, 5 µL of streptavidin magnetic beads (per reaction) were equilibrated with 2X B&W buffer at twice the final volume. Reactions were transferred from qPCR tubes to 1.5 mL microcentrifuge tubes, and 10 µL of equilibrated beads were added to each. These dCas9-capture reactions were incubated in a thermomixer at 37 °C with shaking at 1700 rpm for 30 minutes.

Bead-bound reactions were transferred to 5 mL tubes for high-stringency washing. Each tube was placed on a 15 mL magnetic rack (Sergi Lab Supplies, Seattle, USA) for one minute to pellet the beads. Supernatant was discarded, and beads were resuspended in 2 mL of 2X B&W buffer for a one-minute incubation. Each reaction underwent nine such washes; before the sixth wash, beads were transferred to fresh 5 mL tubes. After the final wash, beads were resuspended

in 50 µL of nuclease-free water, vortexed at 2700 rpm, spun briefly, and stored at 4 °C until elution.

For proteinase K digestion, 25–50 µL of dCas9-bound beads per sample were combined with 1 µL of Proteinase K (800 U/mL, NEB) and nuclease-free water to a total volume of 50 µL. Tubes were mixed, spun down, and incubated for 10 minutes at room temperature in a thermomixer. Afterward, samples were placed on a DynaMag-2 magnetic rack (Thermo Fisher Scientific, Waltham, USA), and the supernatant containing the enriched DNA was collected and purified using either Monarch® (NEB) or GeneJET (Thermo Fisher Scientific) 5 µg DNA Cleanup Columns. DNA was eluted in 20 µL of elution buffer.

To determine the optimal number of amplification cycles, qPCR was performed on all enrichment samples, including the negative control, using 10 µM Mi9_FWD_Amp_NV and Mi9_REV_amp_NV primers and 2 µL of enriched DNA as template. Reactions followed the protocol for Indexed amplification of the barcoded DHFR gene libraries for MiSeq, with a shortened extension time of 30 seconds.

Bulk PCRs were then prepared for each condition using the same setup and cycle numbers identified by qPCR. Amplified products were purified using Monarch® 5 µg DNA Cleanup Columns and validated for correct size using 0.8%, 2%, or 4% E-Gels™ (Thermo Fisher Scientific). Final enrichment products from both the *rfp* gene library and DropSynth DHFR libraries were submitted to Plasmidsaurus (South San Francisco, USA) for nanopore sequencing.

Nanopore data analysis and enrichment scores

Nanopore sequencing was used to analyze the enrichment data. Enrichments up to and including testing the different methods to denature dCas9 were run on an R9.4.1 flowcell with

v10 chemistry called using Guppy on super-accurate mode. Subsequent enrichments were sequenced with an R10.4.1 flowcell, v14 chemistry and Guppy 6.5.7 on super-accurate mode. Raw fastq data was loaded into a custom python script which extracted the barcode region from each read. Experiments with the pEVBC3 plasmid used the motif ACCTAAGTGTCGCTGCCGAACAGG N20 GCTAGAAGAGCGCACGACGTCACG while experiments with pEVBC8 used GGTACCTAAGTGTCGCTGCCGAACAGC N20 AGGAGAAGAGCGCACGACGTCACG, where N20 is the barcode region. Extracted barcodes were collapsed using the Starcode spheres algorithm on distance of 1 (52). Barcode reads were imported into R for further analysis. We determined the fraction of the population for each barcode by normalizing reads by total sequencing depth. Log2 fold enrichment scores were calculated using the ratio of the population fraction after enrichment to their fraction in the initial population. For barcodes not in the initial population a pseudocount of 0.5 was used. Barcodes which dropped out of the population were tracked separately. The population level fold enrichment metric was determined by taking the total population fraction of all targets after enrichment divided by the total fraction of turrets before enrichment. RNA-seq data was analysed as described elsewhere (46). Significance testing was done using a Wilcoxon paired test. CRISPRscan scores were generated using crisprScore (1.4.0) (53).

2.4 Results and Discussion

2.4.1 Proof-of-Concept Enrichment of 18 Barcodes from a Barcoded *mcherry* Library (BAR-CAT v0.1)

We sought to test whether the BAR-CAT method could enrich specific barcodes of interest from a low-complexity DNA library prior to scaling up to more complex DropSynth

gene libraries. To this end, we constructed a barcoded red fluorescent protein (*rfp*) gene library by PCR-amplifying the *mcherry* gene using in-house primers and cloning it directly adjacent to 20 nucleotide (nt) unique barcodes within a barcoding plasmid, pEVBC3. This plasmid was derived from the pEVBC vector previously used to barcode DropSynth gene assemblies (15) but lacked a PAM site adjacent to the 3′ end of the barcode on the template strand, which is required for CRISPR-Cas9 to initiate R-loop formation on the non-template strand. This omission was intentional, as our goal was to rapidly develop a proof-of-concept using a single-gene library. The barcoded *rfp* library contained sufficient barcode diversity to select targets that naturally contained a 5′-NGG-3′ PAM sequence at the 3’ end of the variable barcode region. These targets were used for initial proof-of-concept experiments and subsequent optimization.

Following a standard workflow, we ligated *mcherry* amplicons into pEVBC3 to produce a 2.7 kb barcoded *rfp* plasmid, transformed it at high efficiency into E. coli DH5α, pooled the resulting colonies, isolated plasmid DNA, and sequenced the library on an Illumina MiSeq to quantify barcode diversity for BAR-CAT target selection (Fig. 7A). Sequencing revealed >300,000 unique barcodes, comparable in diversity to DropSynth gene libraries. Barcode representation varied widely, ranging from approximately 1,000 reads for the most abundant barcodes to 1 read for the least. This variability can be summarized using the Gini Coefficient, a metric of inequality where a value of zero corresponds to perfect uniformity and a value of one to perfect inequality (19, 54). The observed Gini Coefficient of 0.56 highlights the highly skewed distribution typical of synthetic DNA libraries, where biases introduced during oligonucleotide synthesis, amplification, assembly, or cloning can result in significant disparities in sequence abundance (55). In contrast, prokaryotic genomes have relatively uniform sequence composition. However, sequence features such as GC content can vary among various phyla,

indicating that GC content is specific to each phylum and is influenced by environmental selective pressures instead of changes in specific genes (56). This comparison underscores a key distinction between synthetic and natural DNA libraries: synthetic libraries require targeted enrichment methods like BAR-CAT to compensate for representation biases, particularly when downstream applications depend on uniform sequence recovery.

To minimize abundance-dependent biases during enrichment, we selected 18 target barcodes spanning this range. Two barcodes were chosen per read abundance tier, from the highest (barcodes 1 and 2, ~1,000 reads) to the lowest (barcodes 17 and 18, 1 read), producing a balanced set for this initial proof-of-concept (Fig. 7B). These barcodes were used as protospacers for CRISPR-dCas9 targeting (Fig. 7A).

To generate the corresponding sgRNA library, synthetic DNA oligos encoding the 20 nt spacer sequences were ordered with flanking PCR priming sites and a BsaI type IIS restriction site. Oligos were pooled and assembled via Golden Gate Assembly (GGA) to produce DNA templates for *in vitro* transcription (IVT). The resulting pooled template was transcribed by T7 RNAP into an 18-plex sgRNA library, which was validated by gel electrophoresis and RNA-seq (Fig. 7A). Our full IVT protocol and validation results are detailed elsewhere (46).

For targeted dCas9 enrichment, we selected the dCas9-3XFLAG™-Biotin protein (Sigma). After RNPs form with this biotinylated dCas9 protein and the18-plex sgRNA library, dCas9 should selectively bind to the 18 targeted barcodes. We reasoned that streptavidin magnetic beads would be a simple, effective, and straightforward way to capture the biotinylated dCas9 (Fig. 7A), as demonstrated by other CRISPR-Cas9 targeted DNA enrichment methods (51).

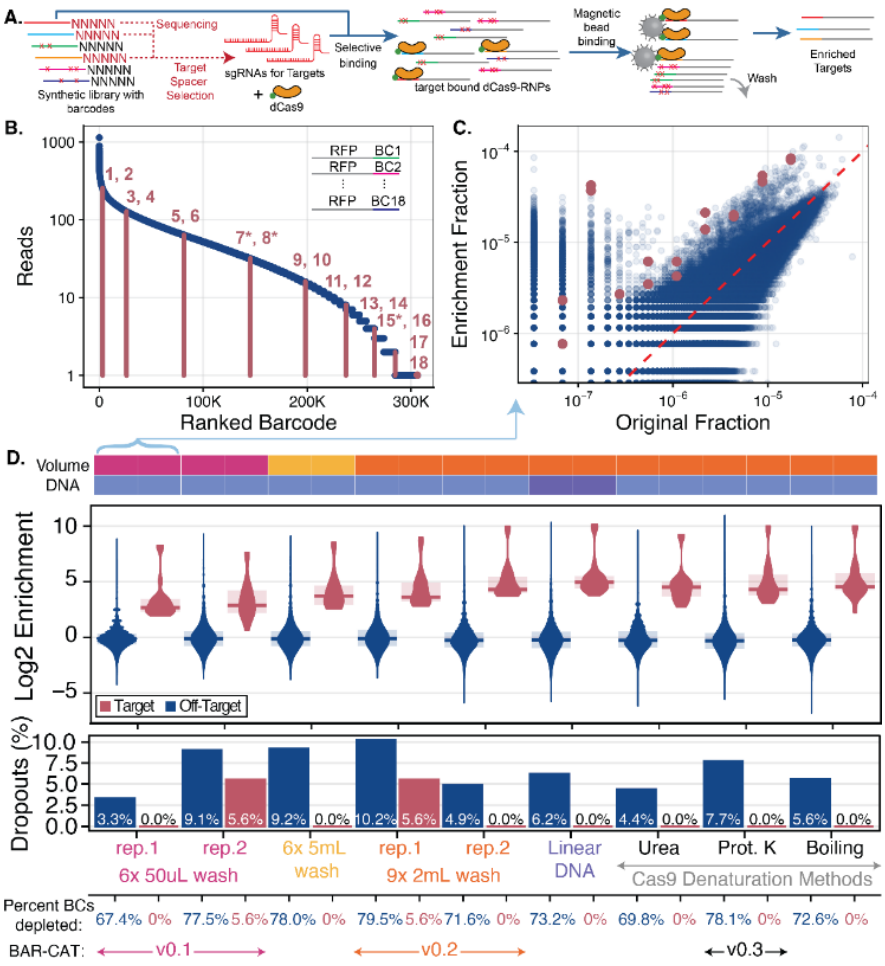


Figure 7. Overall workflow for BAR-CAT proof-of-concept and initial optimizations.

A. Synthetic gene library is first generated and sequenced to map 20-nt barcodes to their corresponding DNA molecules. Barcodes associated with perfect sequences are selected as protospacers for targeting. Spacer sequences are then *in vitro* transcribed into sgRNA libraries, which are complexed with dCas9 to form ribonucleoprotein complexes (RNPs). These RNPs are incubated with the synthetic library, allowing dCas9 to bind target barcodes. Biotinylated dCas9 is captured using streptavidin-coated magnetic beads, followed by bead pulldown and washes to remove unbound, non-target sequences. Enriched DNA is PCR-amplified and sequenced via nanopore to assess target enrichment. **B.** Rank-ordered read abundance of 18 selected target barcodes from a >300,000-barcode single-gene *rfp* library. Each dot represents a barcode, ranked by read count. Dark magenta lines denote target barcodes, with magenta numbers indicating barcode IDs. Asterisks mark the three target barcodes used in BAR-CAT optimizations shown in Fig. 8. **C.** Scatter plot comparing the fraction of barcode reads in the *rfp* library before (original) and after enrichment. Each blue dot represents a non-target barcode; each magenta dot represents a target barcode. The red dashed unity line denotes equal representation before and after enrichment, serving as a reference for assessing enrichment magnitude. **D**. Enrichment conditions are shown chronologically along the x-axis. The colored bar above the plots summarizes key iterative protocol changes, including bead wash volume, DNA input format (linear vs. supercoiled), and dCas9 denaturation method. Violin plots display the distribution of $\log_2$ enrichment scores for off-target (blue) and target (magenta) barcodes. Shaded areas represent the interquartile range (25th–75th percentile), with horizontal bars indicating median enrichment. Percent dropouts are indicated in bar plots for off-target (blue) and target (magenta) barcodes. All barcodes with reduced abundance ($\log_2$ enrichment < 0), including barcodes that drop out post-enrichment, are reported as % depleted. The pink condition (6× 50 µL) corresponds to the original proof-of-concept protocol (v0.1) and is linked to panel **C** by a blue arrow.

After the RNPs bound to barcode targets following enrichment, streptavidin-coated magnetic beads were incubated with biotinylated molecules. Subsequent washes were performed using a wash buffer to remove non-specifically bound DNA, thereby isolating the biotinylated complexes retained on the beads (Fig. 7A). To identify optimal washing conditions for enriching targeted barcodes with high sensitivity, we screened four buffer compositions for wash stringency. A sham DNA capture experiment was performed using a synthetic library containing a single barcode from the *rfp* library. This DNA was incubated with equilibrated streptavidin beads in the absence of biotinylated dCas9 or RNPs, simulating nonspecific DNA-bead interactions (Fig. 15A, see Appendix A). Beads were washed either three or six times with 50 µL of each buffer, and wash steps were carried out at either room temperature (RT, ~25 °C) or 30 °C to assess temperature effects on wash stringency.

After the final wash, supernatants were subjected to quantitative PCR alongside a 10-ng positive control and a no-template control. Amplification cycle (Cq) values were plotted to evaluate wash efficiency, where higher Cq values indicated reduced residual DNA and thus greater stringency (Fig. 15A). The buffer conditions tested included immobilization buffer (IB), 2× binding and wash buffer (2× B&W), 2× B&W supplemented with 10% NP-40, and 1× TE supplemented with 10% NP-40 (Table 2, see Appendix A). The 2× B&W buffer, containing 2 M NaCl in TE with EDTA, is recommended by the bead manufacturer. The NP-40-supplemented variant was included to test whether the non-ionic detergent would enhance stringency. In contrast, IB lacked salt but included DTT to stabilize dCas9, while the 1× TE + NP-40 formulation combined detergent without salt.

Among the tested conditions, both 2× B&W (n = 6) and 2× B&W + NP-40 (10%) (n = 4) yielded the highest Cq values relative to the no-template control, suggesting the most stringent

washes (Fig. 15B). Due to limited sample size and overlapping trends, we could not definitively distinguish between the two, so 2× B&W was selected for all subsequent BAR-CAT experiments. This buffer offered both reliable performance and ease of preparation, and its selection aligned with the manufacturer's recommendations.

We first aimed to use our 18-plex sgRNA library to enrich the 18 corresponding barcode protospacers from the *rfp* library. To form active RNPs, we combined 3 µL of 33 nM sgRNA library with 1 µL of 1 µM biotinylated dCas9, maintaining a 1.11:1 dCas9-to-sgRNA molar ratio, in a total volume of 27 µL, following NEB recommendations. Complexes were pre-incubated for 10 minutes at 25 °C, followed by an additional 10-minute incubation at 37 °C. Then, 50 ng of the *rfp* library was added as a binding substrate, corresponding to a 39.6:1 RNP-to-DNA molar ratio, and incubated for 15 minutes at 37 °C to allow selective binding.

Streptavidin magnetic beads (5 µL), sufficient to capture 25 pmol of biotinylated dCas9, approximately 25-fold excess relative to the dCas9 input, were added to the enrichment reaction and incubated for 30 minutes at 37 °C with shaking. Beads were then washed six times with 50 µL of 2× B&W buffer, as previously determined to reduce non-specific retention of off-target barcodes. Enriched barcodes were amplified directly from the beads by qPCR and subjected to nanopore sequencing.

Sequencing analysis revealed increased abundance of all 18 targeted barcodes following enrichment relative to the original *rfp* library (magenta dots), evident as a vertical shift above the red dashed unity line indicating equal abundance in both libraries (Fig. 1C). However, two distinct subpopulations of off-target barcodes (blue dots) were observed: (1) a majority that remained at moderate abundance (10^{-5} to 10^{-6} fraction), and (2) a smaller group that were unexpectedly enriched, with representation similar to that of the targeted barcodes. We expected

the population of non-enriched off-target barcodes to be lower, as bead washing with streptavidin beads should remove non-specifically bound DNA. The second subpopulation, low-abundance off-target barcodes (10^{-8} to 10^{-6} fraction) that became enriched, was attributed to known off-target binding of CRISPR-dCas9 (57–59). We identified that sequence errors arising during sgRNA spacer oligo synthesis, PCR amplification (2.8×10^{-7} errors/nt), and IVT by T7 RNAP (5×10^{-5} errors/nt) (60) likely generated mutant sgRNA spacers within the 18-plex library (46). These mutations could have especially contributed to off-target binding if they were present at the PAM distal region at the 5' end of the sgRNA spacers, known to tolerate mismatches to targets (59).

Overall, the enrichment of the 18 targeted barcodes was successful, establishing an initial proof-of-concept for BAR-CAT design. Log_2 enrichment scores, defined as the log_2 fold change in barcode abundance before and after enrichment, are shown in Fig. 1D for replicate 1 of the 6×50 µL bead wash condition (shown in pink), corresponding to the protocol used in BAR-CAT v0.1. Most off-target barcodes had log_2 enrichment scores near zero, indicating no enrichment. Barcodes with scores between −6 and 0 were considered depleted or dropped out relative to the input library. Barcodes with scores below −6 were classified as post-enrichment dropouts (Fig. 7D).

We observed that 67.4% of off-target barcodes were depleted and 3.3% dropped out, suggesting that washing under these conditions was insufficient to fully eliminate off-targets. In contrast, the 18 targeted barcodes were enriched with a median log_2 fold change of 2.7 (6.3-fold enrichment), with no evidence of target depletion. To assess the practical impact of this enrichment, we calculated the population fraction enrichment, defined as the ratio of the sum of the population fractions of all target barcodes after enrichment to the sum of their fractions

before enrichment. This metric yielded a 6.4-fold increase, indicating that approximately 6.5-fold fewer colonies would need to be screened to recover clones bearing the desired barcodes.

2.4.2 Bead Wash Optimization and the Development of BAR-CAT v0.2

Although the initial post-enrichment results were promising, they exhibited a high level of non-enriched, off-target barcodes, indicating insufficient removal of unwanted sequences during the bead wash steps. High-stringency washes are critical in hybridization-based DNA capture methods for eliminating cross-hybridized and unbound sequences (11). Drawing from this precedent, we hypothesized that increasing the number and volume of bead washes would help reduce the retention of non-target barcodes caused by weak or non-specific dCas9 interactions in the BAR-CAT protocol. Having previously optimized wash buffer stringency (Fig. 15), we next tested whether increasing wash volume and frequency could further enhance the removal of off-target barcodes.

We conducted a preliminary experiment to assess the appropriate number of washes. To do this, we incubated 50 ng of supercoiled *rfp* library DNA with magnetic SNAP-Capture bead aliquots, using these beads as a proxy for evaluating wash efficiency (not as a replacement for the streptavidin magnetic beads). The beads were washed 15 times with 1 mL of IB. Supernatants from washes 3, 6, 9, 12, and 15, along with the beads from the final wash, were collected for quantitative PCR (qPCR) and gel electrophoresis to quantify the abundance of the 586 bp *rfp* library PCR product.

We chose IB over 2X B&W buffer due to the known inhibitory effects of high salt concentrations on PCR efficiency (61, 62). The PCR product band became undetectable after wash 9, while PCR of the beads washed 15 times produced abundant product (Fig. 16A, see Appendix A). Quantification of the PCR bands using ImageJ (63) revealed that wash 3

supernatant produced a DNA peak intensity of 18,552 arbitrary units (a.u.), while the intensities for washes 9 and 12 were 5,708 a.u. and 4,005 a.u., respectively (Fig. 16B). Based on these results, we increased the cumulative 2X B&W buffer washes of streptavidin beads to 30 mL (6x 5 mL) and 18 mL (9x 2 mL). The 18-plex sgRNA library and protocol described for BAR-CAT v0.1 were used, with the only variation being the number of washes. Additionally, we included an enrichment control (BAR-CAT v0.1) washed with a cumulative wash volume of 0.3 mL (6x 50 µL, replicate 2).

A similar distribution of barcodes was observed in replicate 2 of the original 6× 50 µL bead wash condition, consistent with replicate 1 shown in Fig. 7C, validating the reproducibility of our experimental control (Fig. 17A, see Appendix A). In contrast, the 6× 5 mL (Fig. 17C, see Appendix A) and 9× 2 mL (Fig. 17B) bead wash conditions displayed a noticeable shift in target barcode enrichment compared to the original distribution. A key difference was observed in the distribution of off-target barcodes: a greater number of moderately abundant off-target barcodes from the original population (fractional abundance between 10^{-5} and 10^{-6}; Fig. 7C, Fig. 17A) showed enrichment in these higher-volume wash conditions, likely due to a reduction in background signal from non-enriched barcodes (Fig. 17B,C). We initially hypothesized that this apparent enrichment reflected increased dCas9 off-target binding that had been obscured in the 6× 50 µL control due to higher background levels. These findings suggest that increasing bead wash volume can reduce background DNA and enable finer resolution of enrichment dynamics.

The control 6× 50 µL condition (replicate 2, shown in pink) yielded a median $\log_2$ enrichment of 2.9 of targeted barcodes (7.4-fold enrichment), and a population enrichment score of 5.8. These values closely matched those observed in replicate 1, confirming protocol reproducibility. However, 1 of the targeted barcodes was depleted relative to the input library,

and one barcode dropped out (5.6% target dropouts) entirely post-enrichment, indicating reduced recovery for certain targets. Among off-target barcodes, 77.5% were depleted and 9.9% dropped out, representing a ~3-fold increase in dropouts compared to replicate 1 (Fig. 7D). These results suggest that while overall enrichment performance was reproducible, some stochastic variation in barcode retention may still occur between replicates.

The 6× 5 mL bead wash condition (shown in yellow) produced a median $\log_2$ enrichment of 3.7 of targeted barcodes (12.5-fold enrichment) and a 13.4-fold population fraction enrichment with no depletion or dropout of targets. The 9× 2 mL condition (orange, replicate 1) showed a median $\log_2$ enrichment of 3.6 (12.2-fold increase) and a 12-fold population fraction enrichment, but with 5.6% target depletion and 5.6% target dropouts. However, a greater percentage of off-target barcodes dropped out with increased washes compared to the 6x 50 µL wash control (replicate 2). More specifically, the 6x 5 mL wash condition had 78.0% depleted off-target barcodes and 9.2% dropouts, while the 9x 2 mL wash condition had 79.5% depleted off-targets and 10.2% dropouts (Fig. 7D). Overall, increasing the cumulative wash volume by approximately 100-fold resulted in a roughly two-fold improvement in target barcode enrichment while slightly reducing off-target barcodes. Based on these results, we selected the 9× 2 mL bead wash condition for all subsequent enrichments and designated this protocol as BAR-CAT v0.2, as it provided better enrichment with less buffer usage and labor compared to the 6× 5 mL condition.

2.4.3 Approaches to Denature dCas9 and Reduce Off-Target Barcodes for the Development of BAR-CAT v0.3

While optimizing bead washes yielded modest improvements, we sought to further increase enrichment of targeted barcodes and reduce off-target carryover. We hypothesized that

denaturing dCas9 to release bound DNA into the supernatant could facilitate recovery of enriched targets while simultaneously removing dCas9 and magnetic beads, which may retain non-specifically bound off-target sequences. Additionally, we tested whether using a linearized version of the *rfp* library could enhance enrichment by enabling dCas9 to more efficiently access target PAM sites.

The experiment to test these conditions was performed as follows. Eighteen targets were enriched from either the supercoiled or linearized *rfp* library using the 18-plex sgRNA pool. The resulting beads were washed using the 9× 2 mL protocol (BAR-CAT v0.2), and three dCas9 denaturation methods were tested for their ability to release bound DNA: incubation with 8 M urea (43), treatment with proteinase K (64, 65), or heat incubation at 65 °C for 5 minutes (65, 66). Following denaturation with urea or proteinase K, supernatants containing released DNA were collected, column-cleaned, PCR-amplified, and nanopore-sequenced to identify enriched barcodes. The linear *rfp* library template (827 bp) was generated by PCR amplification of the supercoiled plasmid using mi9_FWD_amp_NV and mi9_REV_amp_NV primers for 14 cycles. Compared to the 2.7 kbp supercoiled version, the shorter length of the linear *rfp* library reduced the RNP-to-DNA mass ratio by approximately 3.5-fold. However, dCas9 and sgRNA concentrations remained constant, with a ~10-fold molar excess and an RNP-to-target ratio of 1.11.

The distributions of target and off-target barcodes across the tested conditions were similar to those observed in the 9× 2 mL replicate 1 condition (orange) (Fig. 17B; see also Fig. 18A–E in Appendix A). This indicates that, at a global level, neither the use of a linear *rfp* library nor dCas9 denaturation appreciably altered the representation of target or off-target barcodes. This outcome contrasts with the substantial differences observed between the 6× 50 µL

(Fig. 7C, Fig. 17A) and 9× 2 mL (Fig. 17B) bead wash conditions, where wash stringency had a pronounced effect on barcode distributions.

Next, we examined the $\log_2$ enrichment scores for the linear and supercoiled *rfp* library conditions. The supercoiled *rfp* library control (replicate 2, 9 × 2 mL, shown in orange) yielded a median $\log_2$ 4.3-fold enrichment of the targeted barcodes (19.5-fold enrichment) and 23-fold population fraction enrichment. These values were slightly higher than those obtained in replicate 1 of the same condition, possibly due to experimental variation. Enrichment of targets from the linear *rfp* library yielded a median $\log_2$ 5-fold enrichment (31-fold enrichment) and 26-fold population fraction enrichment. This represented a ~1.6-fold increase in targeted enrichment relative to the supercoiled condition. While this difference could be attributed to the linear DNA format or experimental variation, it is also possible that the increased molar concentration of DNA molecules in the linear library due to their shorter length led to a slight increase in enrichment efficiency.

We then examined the $\log_2$ enrichment scores for the dCas9 denaturation methods of urea, proteinase K, and heat. All three conditions produced similar targeted enrichment values compared to the supercoiled and linear *rfp* library enrichment conditions. Urea treatment produced a median $\log_2$ 4.5-fold enrichment of the targeted barcodes (23-fold increase), with 14-fold population fraction enrichment. Proteinase K treatment produced a median $\log_2$ enrichment of 4.3 (20.0-fold increase), with 22-fold population fraction enrichment. Lastly, boiling yielded a median $\log_2$ enrichment of 4.5 (23-fold increase), with 22-fold population fraction enrichment. None of the tested dCas9 denaturation enrichments showed depletion or dropout of targeted barcodes (Fig. 7D).

To evaluate the impact of the conditions tested on off-target barcode representation, we compared the off-target barcode depletion and dropouts. The supercoiled DNA enrichment control showed 71.6% depletion and (83,414, 4%) dropouts. The linear DNA enrichment showed a slight increase in depletion (73.2%) and dropouts (106,119, 6%), potentially reflecting experimental variation or improved depletion due to the linear format. Urea-mediated dCas9 denaturation showed less depletion (69.8%) and dropouts (74,542, 4.4%) compared to the supercoiled and linear conditions. Boiling produced slightly increased depletion (72.6%) and dropouts (95,186, 5.6%). The highest amount of depletion (78.1%) and dropouts (131,254, 7.7%) was observed with proteinase K treatment, representing a 3.7% increase compared to the supercoiled control (Fig. 7D, Fig. 19 in Appendix A). Despite this, proteinase K did change enrichment scores for targeted barcodes compared to the supercoiled and linear conditions, suggesting it effectively removed off-targets while preserving on-target signals. We concluded that urea treatment was less effective than the other conditions, likely due to its harsh, non-specific denaturation mechanism, whereas proteinase K provided a more targeted and efficient dCas9 denaturation approach.

Based on these findings, we adopted proteinase K-mediated dCas9 denaturation for all subsequent enrichments and designated this protocol optimization as BAR-CAT v0.3. Although the linearized *rfp* library contained roughly 3.5-fold more DNA molecules than the supercoiled control and accordingly showed a modestly higher enrichment of the 18 targets, it remains unclear whether this advantage stems from the increased DNA copy number, differences in DNA conformation, or experimental variation.

2.4.4 Assessing DNA Input Amount and Incubation Time on Enrichment Performance to Develop BAR-CAT v1.0

Since targeting 18 barcodes from the linearized *rfp* library resulted in a modest increase in enrichment compared to the supercoiled version (Fig. 7D), we sought to determine whether the ~3.5-fold higher number of DNA molecules in the linear library contributed to this improvement. We hypothesized that the primary factor driving this difference was the greater molar abundance of DNA molecules available for dCas9 binding in the linear format.

For context, the least-represented target in our synthetic *rfp* library produced just 1 read out of 14,641,735 total reads, corresponding to 6.8 x 10-8 of the molecules. A single-copy gene in diploid human genomic DNA, by contrast, makes up roughly 2 copies in a 6.2 Gb genome (3.2 x 10-6 of all molecules if one assumes a 10 kb gene), making it about 47-fold more abundant than our scarcest target. While the human genomic DNA is essentially uniform, our synthetic library exhibited far higher inequality (Gini: 0.56).

This rationale is further supported by trends observed in other CRISPR-Cas9 enrichment methods, which typically require substantially higher DNA inputs. For example, RNA-guided endonuclease (RGEN) enrichment methods such as RGEN-R and RGEN-D (65) use 10–20 μg of fragmented genomic DNA per reaction. In RGEN-R, Cas9 cleaves large genomic fragments (>20 kb), which are then captured via biotinylated adaptors and streptavidin beads. In contrast, RGEN-D employs dCas9 pre-complexed with biotinylated sgRNAs for direct pulldown, followed by proteinase K treatment to release the DNA. These methods use over 200-fold more DNA than our standard BAR-CAT reactions (65). However, one important difference is that RGEN-D and RGEN-R enrich targeted regions from genomic DNA, which is more uniform than the barcodes within the *rfp* library used for BAR-CAT enrichment. Together, these comparisons support the idea that improved enrichment with linear *rfp* DNA is at least partly due to the increased molar abundance of targetable molecules.

To test this, we enriched three barcodes (7, 8, and 15) from the supercoiled rfp library (barcodes with asterisks, Fig. 7B) using two DNA input amounts: the original 50 ng and a 10-fold higher input of 500 ng. The 500-ng input DNA condition decreased the ratio of RNPs to DNA to approximately 4:1 from the 40:1 ratio used previously (50 ng input DNA). We targeted three barcodes for this experiment due to their enrichment performance across conditions when 18 barcodes were targeted. More specifically, barcodes 7 and 8 showed modest enrichment in the BAR-CAT v0.3 enrichment condition (proteinase K denaturation, Fig. 7D) while barcode 15 was one of the two most enriched targets of all 18 barcodes (Fig. 20, see Appendix A). To streamline the experiment, we used synthetic sgRNAs (Integrated DNA Technologies) due to concurrent challenges with optimizing our IVT sgRNA workflow (46).

We retained a supercoiled *rfp* format for the input DNA for several reasons. Although the linearized version showed slightly better enrichment, it had been prepared using 14 cycles of PCR, raising concerns about potential bias. PCR can skew barcode representation by preferentially amplifying GC-rich sequences and introducing mutations in spacer regions, which may increase off-target dCas9 binding (67–69). Since BAR-CAT already includes a post-enrichment PCR step, we sought to minimize additional amplification. Restriction digestion could, in principle, provide a PCR-free linearization strategy, but it would add complexity to the workflow and, crucially, many enzymes would cut within the 20 bp random barcodes embedded in our library, thereby removing or truncating a portion of uniquely identifiable molecules.

We also faced practical constraints on DNA concentration. To maintain consistent reaction conditions, input DNA had to be concentrated to ≤3 µL. Using column-based purification, we were typically able to concentrate the *rfp* library to ~167 ng/µL, making 500 ng inputs feasible. However, higher concentrations were difficult to achieve reliably with our

existing protocol, so we capped the input at 500 ng. While actual concentrations varied between preparations, this amount represented the practical upper limit. Future BAR-CAT optimizations could incorporate tRNA-assisted ethanol precipitation to further concentrate DNA, potentially enabling >5 μg of input DNA in ≤3 μL. This approach has been shown to enhance the concentration of genomic DNA libraries and improve plasmid transformation efficiency by up to 400-fold compared to conventional methods (70).

In addition to testing 500 ng and 50 ng DNA input amounts, we evaluated three enrichment incubation times at 37 °C: 15 minutes (our original condition), 1 hour, and 8 hours. The 15-minute time was based on NEB recommendations, but other Cas9 enrichment methods have various incubation times. RGEN-D uses 20 minutes (65), Aalipour and colleagues used 30 minutes (43), Kim and colleagues used 2 hours (51), and RGEN-R used 8 hours (65). While the rationale behind these incubation times is not always specified, *in vitro* Cas9 binding studies provide some context. A Cas9 binding rate constant of $0.8 \pm 0.2\ \text{min}^{-1}$ has been reported (66), implying that ~80% of target DNA can be bound within one minute under simplified conditions with a single target on a plasmid. However, BAR-CAT employs a highly multiplexed library with ~300,000 barcodes, which increases search complexity. We therefore hypothesized that longer incubation times might improve enrichment by giving dCas9 more time to locate targets. To test this, we included 1- and 8-hour incubation times to assess whether extended search duration would enhance recovery of the three selected barcodes.

Log_2 enrichment scores for all three targeted barcodes increased substantially across all tested DNA input amounts and incubation times, relative to the original 18-plex enrichment values, while no target barcodes dropped out (Fig. 7D). The 50-ng input DNA and 15 min incubation control condition (BAR-CAT v0.3) produced a median $\log_2$ 6.1-fold enrichment (70-

fold enrichment) and 135-fold population enrichment (Fig. 8A), demonstrating a median ~11-fold increase compared to its corresponding 18-plex enrichment (Fig. 7D). This suggests that dCas9 binding is more efficient when fewer targets are present, resulting in stronger signals. However, it may also reflect the effect of selecting the three high-performing barcodes to target, which could inflate enrichment scores by skewing the median upward. Nevertheless, the 500 ng, 15-minute incubation enrichment condition had a median $\log_2$ enrichment of 9.2 (600-fold increase), with a 1,094-fold population enrichment. This indicates a ~12-fold increase in median enrichment when the input DNA amount was increased from 50 ng to 500 ng (Fig. 8A).

Next, we evaluated off-target barcode depletion and dropouts for these conditions. The depletion and dropouts were similar for the 50 ng, 15 min incubation (82.2% depletion, 3.9% dropouts) and 500 ng, 15 min incubation (79.1% depletion, 3.4% dropouts) conditions (Fig. 8A). The distribution of target and off-target barcodes before and after enrichment with 500 ng DNA for 15 min showed that off-target barcodes aligned closely with the dashed red unity line, indicating that nearly all off-targets were non-enriched (Fig. 8C, see Fig. 21D in Appendix A), as opposed to the 50 ng DNA condition with a distinct population of moderately enriched, off-target barcodes (Fig. 21A). These results suggest that increasing the DNA input amount has an effect on reducing off-target enrichment, improving CRISPR-dCas9 binding specificity. However, it had no effect on eliminating off-target, non-enriched barcodes, which are likely due to the magnetic bead pull-down and washing approach overall (Fig. 8C).

Meanwhile, the 50 ng, 1-hour enrichment condition had a median $\log_2$ enrichment of 5.0 (32-fold enrichment, a ~2.2-fold decrease compared to the 50 ng, 15 min incubation condition. The 55-fold population fraction score (50 ng, 1 hr) also decreased from the 135-fold value corresponding to the equivalent condition incubated for 15 min. However, the 5.7-fold median

Log_2 enrichment for the 50 ng, 8-hour incubation condition represents a slight 1.6-fold increase in enrichment, with the population enrichment score increasing from 61-fold to 84-fold. Meanwhile, the off-target barcodes were depleted by 88.2% for the 1-hour condition and by 91.4% for the 8-hour condition. These results revealed a trend of 15-minute incubation producing the highest targeted enrichment, with a marked decrease at 1 hour, and an intermediate recovery after 8 hours, with no decrease in off-target depleted barcodes (Fig. 8A).

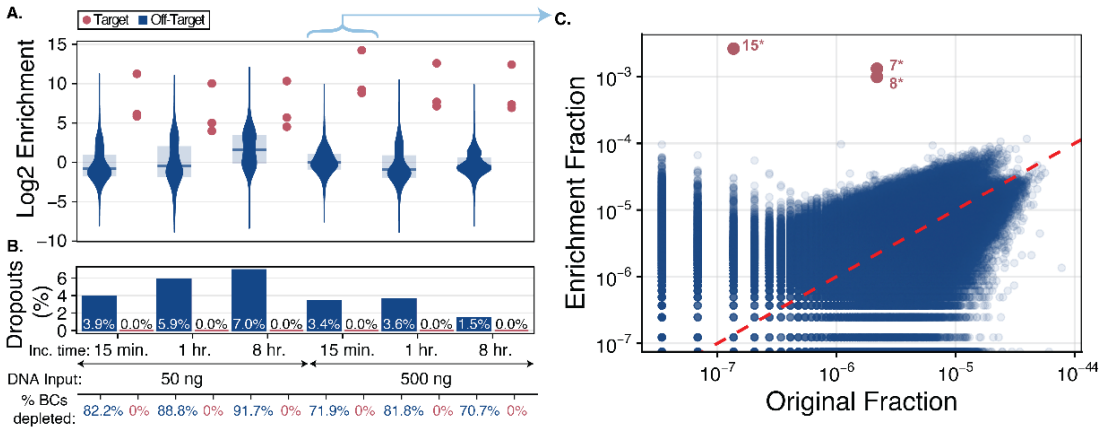


Figure 8. Effects of DNA input amount and incubation time on BAR-CAT enrichment.

A. Violin plots showing log_2 enrichment scores, calculated as the log_2 fold change in barcode abundance before and after enrichment. Each distribution compares non-target (blue) and target (magenta) barcodes across conditions varying in enrichment incubation time and DNA input amount. Shaded regions represent the interquartile range (25th–75th percentile) for non-target barcodes, with bars indicating median log_2 enrichment. The condition using 500 ng of input DNA and a 15 min incubation corresponds to the optimized BAR-CAT protocol (v1.0) and is linked to panel C by a blue arrow. **B.** Percent dropout values are listed in the bar plots for target (magenta) and non-target (blue) barcodes. Protocol variables tested, such as enrichment incubation time and DNA input amount, are indicated below the dropout bars. Barcodes with reduced abundance (log_2 enrichment < 0), including dropouts, are reported as % BCs depleted. **C.** Scatter plot comparing the fraction of barcode reads in the *rfp* library before (original) and after enrichment for the 15 min, 500 ng input condition, as indicated by the blue arrow. This condition, referred to as BAR-CAT v1.0, was used for all subsequent experiments, as it produced a 9.2-fold median log_2 enrichment (600-fold) and a 1,094-fold population fraction enrichment. Blue dots represent non-target barcodes; magenta dots represent target barcodes. Asterisks denote the three protospacers selected from the original set of 18 target barcodes (Fig. 7B). The red dashed unity line marks equal representation before and after enrichment and serves as a reference for evaluating enrichment.

We then evaluated the enrichment scores for the 50 ng and 500 ng input DNA amounts incubated for 1 hour and for 8 hours. The 50-ng enrichment condition showed decreased

enrichment of targets for these incubation times compared to the 15 min incubation time. The 1-hour enrichment produced a median $\log_2$ 5.0-fold enrichment (32-fold) of the three targets, a ~2.2 decrease compared to the 15 min incubation time. The population score between these two incubation times also decreased from 135-fold to 55-fold. For the 8-hour enrichment, the median $\log_2$ 5.7-fold (57-fold) for the three targets increased by approximately 1.6-fold compared to the 1-hour incubation median enrichment, while the population enrichment was 75-fold. Similarly, the 500 ng enrichment conditions showed similar trends. The 500 ng, 1-hour enrichment condition had a median $\log_2$ enrichment of 7.7 (205-fold), a roughly 2.9-fold decrease compared to the 500 ng, 15 min incubation, while the population score decreased from 1,094-fold to 354-fold. However, the median $\log_2$ enrichment of 7.4 for the 500 ng, 8-hour incubation condition was similar to the 1-hour condition, with the population enrichment score decreased from 205-fold to 163-fold (Fig. 8A). Additionally, none of the three targets were depleted or dropped out for any of the conditions tested (Fig. 8B). Overall, these results show that targeted enrichment decreases once incubation has proceeded 1 to 8 hours within the range of 50 ng and 500 ng DNA input amounts.

We then evaluated whether the targeted enrichment was higher with 500 ng DNA inputs across incubation times compared to 50 ng DNA inputs due to reduced off-target barcodes. For the 50 ng input conditions, the 1 hour (88.8% depletion, 5.9% dropouts) and 8 hour (91.7% depleted, 7.0% dropouts) increased compared to the 15 min incubation condition (82.8% depleted, 3.9% dropouts (Fig. 8A). Surprisingly, while the distribution of target and off-target barcodes appeared similar between the 15 min (Fig. 21A, see Appendix A) and 1 hour incubation enrichments (Fig. 21B), the 8-hour condition showed that the off-target, non-enriched barcode population had decreased (Fig. 21C). Since the 8-hour condition had a slightly higher sequencing

depth compared to the others [data not shown], we concluded that the observed decrease in non-enriched barcodes was likely due to the longer incubation time. However, since the $\log_2$ fold enrichment was lower for the 8-hour condition (Fig. 8A) compared to the other incubation times with 50 ng input DNA, we concluded that increased depletion of off-target barcodes did not improve targeted enrichment.

Next, we evaluated whether the trend in off-target barcode depletion and dropouts could be observed across incubation times for the 500 ng DNA input condition. Increased off-target depletion was not observed for the 1-hour incubation (81.8% depletion, 5.9% dropouts) compared to the 15 min incubation (71.9% depletion, 3.4% dropouts), and a decrease in depletion and dropouts was observed for the 8-hour incubation (70.7% depletion, 1.5% dropouts) (Fig. 8A). Meanwhile, the distribution of target and off-target barcodes appeared similar between 15 min (Fig. 21D), 1 hour (Fig. 21E), and 8-hour incubation enrichments with 500 ng input DNA (Fig. 21F). Based on these results, we observed decreased depletion in the 8-hour incubation time with 500 ng of input DNA compared to 15 min and 1 hr incubations with 500 ng input DNA. This is the opposite of the trend observed for incubation time with 50 ng of input DNA.

Overall, these results show that the 500-ng input DNA condition yields superior enrichment of the three targeted barcodes across all incubation times compared to the 50-ng condition. However, both inputs exhibit reduced enrichment after 1 hour of incubation. This was unexpected, as ~97% of Cas9 is estimated to be bound to target DNA at this time point based on cleavage assays (66, 71), though these estimates may not fully reflect dCas9 binding dynamics over longer timescales. One possibility is that a fraction of dCas9 molecules dissociate and fail to rebind due to sgRNA degradation, which can occur after 8 hours at 37 °C (72).

Another explanation is that the chemically synthesized sgRNAs used here may contain synthesis errors, particularly within the RDR (nucleotides 8–17), which could impair long-term binding. These sgRNAs were purified by standard desalting, and IDT estimates an error rate of ~0.05 errors per 100 nt. By comparison, our IVT sgRNAs have an estimated error rate of ~0.005 errors per 100 nt (60), suggesting they contain fewer mutations. Although the vendor-supplied sgRNAs include chemical modifications to reduce RNase degradation, we observed their median $\log_2$ enrichment dropped 3-fold over 20 weeks. This is about twice as slow as our IVT sgRNAs (12-plex), which reached the same decline after just 10 weeks (Fig. 22, see Appendix A). These results suggest that while IVT sgRNAs may degrade more rapidly, they likely contain fewer synthesis errors and could be better suited for long-term enrichment, although their performance must still be evaluated experimentally.

While it was not clear why longer incubation times reduced enrichment, we concluded that performing targeted enrichment with 500 ng of input DNA for 15 minutes was the best condition tested. Therefore, we incorporated these optimizations into BAR-CAT to produce the BAR-CAT v1.0 protocol used to test scale-up evaluation of perfect gene assemblies from DropSynth gene libraries.

2.4.5 Performance of BAR-CAT v1.0 in the Scale-Up of Target Enrichment from DropSynth Libraries

Given the promising performance of BAR-CAT v0.1 in enriching three barcodes from the supercoiled *rfp* library using synthetic sgRNAs, we next applied this method to enrich perfect gene assemblies from two DropSynth dihydrofolate reductase (DHFR) gene libraries. One library contained 1,536 DHFR genes (495–579 bp; library S2), while the other contained 384

DHFR genes (504–576 bp; library S4). These libraries were previously assembled in-house for a study investigating mutations in DHFR homologs that confer trimethoprim resistance (16). To enable barcode targeting, DHFR inserts were cloned into a pEVBC3-derived plasmid (pEVBC8), modified to include a 5′-AGG-3′ protospacer adjacent motif (PAM) at the 3′ end of each barcode. This modification was necessary because DropSynth libraries contain few perfect gene assemblies per target, limiting our ability to computationally select barcodes terminating with a GG dinucleotide as we had with the *rfp* library and the original pEVBC3 backbone, which lacks a PAM site.

Barcoding was performed via T4 DNA ligase-mediated insertion of DHFR products into barcoded pEVBC8, using the same strategy as for the *rfp* library. The resulting plasmids ranged from 2.5 to 2.57 kbp. After *E. coli* transformation, colonies were scraped, plasmid DNA was isolated, and libraries were sequenced on an Illumina MiSeq to assess barcode coverage and guide sgRNA selection. We recovered 5,526,305 unique barcodes from the 384-gene library, and 93,418 from the 1,536-gene library.

Perfect DHFR gene assemblies were mapped to barcodes using a custom computational pipeline, and barcodes were selected as protospacers based on specific criteria (Fig. 23, see Appendix A). To test BAR-CAT's sensitivity, we intentionally selected ultra-rare barcodes with 1–2 reads in the original libraries. In total, 1,384 sgRNAs were designed to target 684 perfect DHFR genes in the 1,536-gene library, and 389 sgRNAs were designed to target 149 genes in the 384-gene library.

The 1,384-plex and 389-plex sgRNA libraries were *in vitro* transcribed from a microarray-derived oligo pool (46). However, RNA-seq revealed severe spacer bias in these libraries, likely stemming from T7 RNAP preference for templates with a 5′ guanine tetramer (5'

GGGG) immediately downstream of the T7 promoter (46, 73). As the initial sgRNA libraries were synthesized with only a single 5′ G to initiate transcription, spacer representation was highly unequal, though these libraries were still used for large-scale enrichment experiments. To address this bias, we explored adding a 5′ GGGG sequence to all spacer templates and compared the resulting uniformity against the 5’ G condition. While the sgRNA libraries used here predated full optimization, we sought to determine whether the 5′ GGGG modification impacted $\log_2$ enrichment.

For this comparison, we subdivided the 389-plex library into smaller groups of 12, 60, and 389 spacers, targeting perfect assemblies from the 384-gene library. All templates were padded with a 5′ GGGG except for one 12-spacer subpool, which retained a single 5′ G to directly compare the two conditions. Libraries were *in vitro* transcribed in emulsions (46) and subsequently used for BAR-CAT enrichment.

To assess the impact of target scale on BAR-CAT performance, we enriched 1, 12, and 389 barcodes from the 384-gene DHFR library, and 1,384 barcodes from the 1,536-gene library, using sgRNAs containing either a 5′ G or 5′ GGGG. A synthetic sgRNA targeting a single barcode (Integrated DNA Technologies) served as a small-scale control.

Baseline performance was established from singleplex (n=2) and 12-plex (5′ G, n=2; 5′ GGGG, n=1) enrichments. Two replicates of single-barcode enrichment achieved 8.7-fold and 7.1-fold $\log_2$ enrichment, corresponding to 426-fold and 140-fold linear values, respectively. Enrichment of 12 barcodes with 5′ G sgRNAs resulted in a 7.2-fold median $\log_2$ enrichment (149-fold) (Fig. 9A) and 64-fold population fraction enrichment, although 8.3% of targets dropped out. A second replicate yielded a 3.8-fold median $\log_2$ enrichment (14-fold) with a 17-fold population fraction enrichment and 25% target dropout (Fig. 9B). Enrichment using a 5′

GGGG sgRNA library produced similar results (log_2 enrichment of 4.9 [31-fold] and 25% target dropout).

We observed variability in 12-plex enrichment levels, with no clear explanation. However, sgRNA age impacted performance, as indicated by a 3-fold decrease in median log_2 enrichment for the 12-plex sgRNA library after just 10 weeks. In comparison, the synthetic single sgRNA used for DHFR enrichment showed only a 1.5-fold decrease, while the 3-plex synthetic sgRNAs exhibited about a 5-fold drop in median log_2 enrichment (Fig. 22, see Appendix A). These results suggest that sgRNA degradation over time likely contributed to reduced enrichment efficiency. Nonetheless, sgRNAs with a 5′ GGGG sequence did not show lower enrichment values than those with a single 5′ G in the 12-plex context.

We next examined off-target barcode dropouts. Singleplex enrichments had 96.9% and 99.6% off-target dropout, while 12-plex enrichments showed 97.6% (5′ G), 89.0% (5′ G), and 94.4% (5′ GGGG) (Fig. 9B). These values were markedly higher than those from *rfp* library enrichments (1.5–10.2%) (Fig. 8B, Fig. 7D). Since bead washing conditions were unchanged between BAR-CAT v0.3 and v1.0, we attributed the elevated off-target dropout to the vastly larger barcode population in the 384-gene library (~5.5 million) compared to the *rfp* library (~300,000), which increased the likelihood of off-target loss during washing.

Fractional barcode distribution analysis showed that low-abundance off-target barcodes were consistently enriched from ~10^{-7}–10^{-6} to ~10^{-4}–10^{-3}, across both singleplex and 12-plex enrichments, independent of spacer design. This off-target enrichment likely masked the enrichment of true targets, especially at small scales (Fig. 24A-E, see Appendix A). Moreover, given that only ~4,302 barcodes out of 5,526,305 had original frequencies above 10^{-5}, most high-

abundance off-target barcodes likely dropped out, while low-abundance barcodes had a higher probability of enrichment simply by chance.

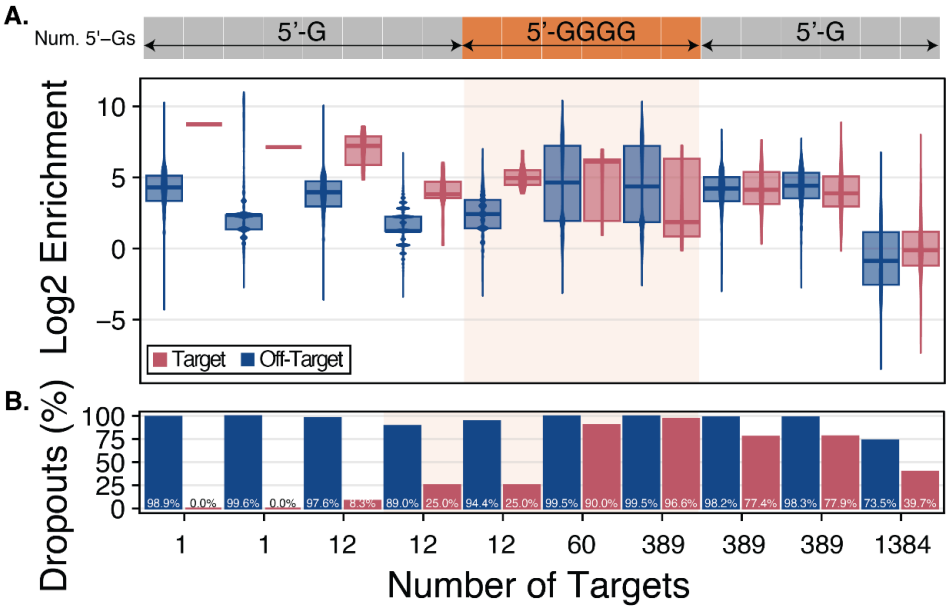


Figure 9. BAR-CAT v1.0 performance declines with increasing enrichment scale and is further reduced by sgRNA spacers starting with 5′ guanine tetramers (5′ GGGG) compared to single 5′ guanine (5′ G) designs.

A. Overlaid violin and boxplots show $\log_2$ enrichment scores ($\log_2$ fold change in barcode abundance before vs. after enrichment) for target (magenta) and non-target (blue) barcodes across different enrichment scales and sgRNA 5′ guanine modifications. Shaded areas indicate the interquartile range; bars mark the median. The colored bar above each distribution indicates sgRNA design: 5′ G (gray) or 5′ GGGG (orange). At the 12-plex scale, enrichment values were similar between the two sgRNA designs. Singleplex enrichments from the 384-gene DHFR library (library S4) achieved 7.1–8.7 median $\log_2$ enrichment (140–426-fold). At 12-plex, values decreased to 3.8–7.2 (14.2–149-fold) with 18.8–64-fold population fraction enrichment. At 389-plex, 5′ GGGG sgRNAs yielded a median $\log_2$ enrichment of only 1.8 (3.6-fold) and 0.8-fold population fraction enrichment, while 5′ G sgRNAs performed modestly better (3.9–4.1 $\log_2$ enrichment; 14.7–17.5-fold; 6.0-fold population enrichment). Enrichment at 1,384-plex scale from the 1,536-gene DHFR library (library S2) was negligible. These reductions likely stem from sgRNA competition for dCas9 at larger scales (78, 79), and, for 5′ GGGG sgRNAs, dilution of active complexes by excess high molecular weight (HMW) byproducts (46). **B.** Percent dropout values are listed in the bar plots corresponding to off-target (blue) and target (magenta) barcodes according to the scale of targeted barcodes, as indicated by the x-axis. For 12-plex enrichments, target dropouts ranged from 8.3% to 25%. At 60-plex and 389-plex, target dropouts increased substantially to 77.9–96.6%. In contrast, the 1,384-plex enrichment showed 39.7% target dropouts, reflecting improved barcode diversity in the 1,536-gene DHFR library. Off-target dropouts remained high across all scales (73.5–99.6%).

Next, we evaluated whether the trends observed for small-scale enrichment from the 384-gene DHFR library extended to larger-scale enrichment targeting 60 (5′ GGGG, n=1), 389 (5′ G, n=2), and 389 (5′ GGGG, n=1) barcodes. The 60-plex enrichment yielded a 6.1-fold median $\log_2$ enrichment (69-fold) (Fig. 9A) with a 4.4-fold population fraction enrichment and 90.0% target

barcode dropout (Fig. 9B). In contrast, the 389-plex enrichment with 5′ GGGG-modified sgRNAs resulted in a low 1.9-fold median $\log_2$ enrichment of (3.6-fold) (Fig. 9A), a negligible 0.8-fold population fraction enrichment, and 96.6% target dropouts (Fig. 9B). The poor performance of the 5′ GGGG-modified sgRNAs at large scales suggests that the 5′ GGGG modification may be detrimental compared to 5’ G sgRNAs during highly multiplexed enrichments.

In contrast, 389-plex enrichments using 5′ G-modified sgRNAs achieved slightly increased median $\log_2$ enrichments of 4.1 (17-fold) and 3.9 (15-fold) (Fig. 9A), 6.0-fold population fraction enrichment, and 77.4% and 77.9% target barcode dropouts, respectively (Fig. 9B). Although dropout rates remained high, they were ~20% lower compared to 5′ GGGG-modified sgRNAs. Together, these results indicate that poor performance at large scales is intrinsic to BAR-CAT v0.1, compounded by the sgRNA design, where 5′ GGGG-modified sgRNAs perform substantially worse than 5′ G-modified sgRNAs.

We attribute the inferior performance of 5′ GGGG-modified sgRNAs to the production of excessive high molecular weight (HMW) RNA species during IVT. Because the same mass of sgRNA was added for each enrichment condition, the presence of HMW, inactive RNA effectively diluted the concentration of functional sgRNA spacer molecules, disproportionately impacting enrichment at higher multiplexing levels (46). Based on these findings, we recommend avoiding 5′ GGGG sgRNAs for large-scale applications. Instead, sgRNAs should be transcribed with a single 5′ G preceding the spacer sequence either in emulsions or using high template input (e.g., 400 ng DNA in 20 μL IVTs), both of which improve sgRNA uniformity without introducing excessive junk RNA species (46).

We next examined the distribution of target and off-target barcode fractions before and after enrichment for the 60-plex and 389-plex DHFR library enrichments. Similar to trends observed in the singleplex and 12-plex experiments, originally low-abundance off-target barcodes were significantly enriched, increasing from initial fractions of 10^{-7}–10^{-6} to 10^{-4}–10^{-3} post-enrichment. For both the 60-plex (Fig. 25A, see Appendix A) and 389-plex (Fig. 25C) conditions, target and off-target barcode distributions overlapped after enrichment, with only a small subset of target barcodes enriched above background. Most target barcodes dropped out, while enrichment of low-abundance off-targets obscured the enrichment signal. High-abundance off-target barcodes from the original population were largely depleted after enrichment. A similar pattern was observed for the two 389-plex (5′ G) replicates (Fig. 25D,E), although slightly more target barcodes were recovered compared to 5′ GGGG conditions, consistent with reduced dropout rates.

These observations suggest that excessive barcode diversity of the 384-gene DHFR library (library S4) led to excessive low-abundance barcodes, which, upon enrichment, introduced significant noise. Applying an even more stringent bottleneck to reduce overall library diversity should therefore mitigate this issue and improve enrichment performance. While sgRNA modification (5′ GGGG vs 5′ G) impacted target dropout, the off-target enrichment behavior was primarily driven by library composition rather than sgRNA design.

Finally, we assessed the performance of the largest enrichment condition tested: enrichment of 1,384 targets from the 1,536-gene DHFR library (library S2) using 5′ G sgRNAs (n=1). This condition yielded a negative 0.11-fold median $\log_2$ enrichment value (0.92-fold) (Fig. 9A) with a population fraction enrichment of 1.2-fold and 39.8% target dropout (Fig. 9B), indicating a failure in targeted enrichment. Although enrichment was negligible at the 1,384-plex

compared to 389-plex scales, the percentage of target barcode dropouts was lower at 39.7% instead of ranging between 77.4% (389-plex, 5’ G) and 96.6% (389-plex, 5’ GGGG). The population fraction of off-target barcodes did not show enrichment (1.0-fold) for the 1,384-plex condition, similarly to the 389-plex conditions. Additionally, the 1,384-plex condition had 73.5% off-target dropouts, notably less compared to the range of 94.4% to 99.6% obtained across all enrichments prepared with the 384-gene DHFR library regardless of scale (Fig. 9B).

These results suggest that, while the 1,384-plex enrichment did not yield strong enrichment, it did result in reduced dropout of both target and off-target barcodes. We hypothesize that this improvement is due to lower diversity of the 1,536-gene DHFR library (S2), in contrast to the 384-gene DHFR library (S4), which contained a much higher proportion of high-abundance target and off-target barcodes. Supporting this, a comparison of spacer fractions in the enriched versus original populations revealed that most targeted barcodes in the 1,384-plex condition were retained and modestly enriched as a distinct subpopulation, while off-target barcodes were also retained. Both sets were largely clustered around the unity line (Fig. 25B, see Appendix A), indicating minimal change in representation. Altogether, these observations support our hypothesis that high sequence diversity in the smaller DHFR library contributed to excessive dropout of both targets and high-abundance off-targets.

However, while lower diversity likely helped retain more target barcodes, it did not fully explain the low enrichment scores for the 1,384-plex. We hypothesized that the low enrichment was partly due to dilution of sgRNA spacers per target barcode at larger scales, leading to insufficient sgRNA-mediated enrichment for most targets. Nevertheless, since some barcodes still achieved strong enrichment while most did not (Fig. 25B), we reasoned that additional factors might differentiate high-performing from low-performing targets. Therefore, we took a

closer look at the features of individual barcodes to better understand the underlying causes of variable performance.

2.4.6. Evaluating sgRNA Performance, sgRNA Library Bias, and Barcode Target Selection to Improve Future Large-Scale BAR-CAT Studies

To investigate the low or negligible $\log_2$ fold enrichment observed in large-scale DHFR gene libraries (Fig. 9A) and the variability in sgRNA spacer performance (Fig. 25B), we first examined whether the $\log_2$ enrichment of targeted barcodes in the 389-plex and 1,384-plex experiments correlated with predicted sgRNA on-target efficiencies. We used CRISPRscan, which predicts sgRNA performance based on *in vivo* editing efficiencies measured in zebrafish embryos (53). However, comparing these predicted CRISPR efficiency scores to the observed $\log_2$ enrichment values from BAR-CAT v0.1 (excluding sgRNAs with dropped-out targets) showed no significant correlation for either the 389-plex enrichments from the 384-gene DHFR library (5′ G sgRNAs, replicate 2, $R^2 = 0.009$; See Fig. 26A in Appendix A) or the 1,384-plex enrichments from the 1,536-gene DHFR library (5′ G sgRNAs, $R^2 = 0.015$; Fig. 26B). This lack of correlation is not unexpected, as most prediction tools are developed for *in vivo* genome editing rather than *in vitro* systems like BAR-CAT. Similarly, Henser-Brownhill and colleagues observed minimal correlation between sgRNA activity and predicted scores using sgRNA Scorer 2.0 (74, 75), and other studies have highlighted the limitations of predictive algorithms for *in vitro* applications (76, 77). While this mismatch between predictions and our assay provides one possible explanation, other factors beyond sgRNA efficiency may be affecting performance in our system.

Next, we explored whether the distribution of spacers within the sgRNA libraries, as determined by RNA-seq analysis from our prior work (46), influenced the $\log_2$ enrichment of

targeted barcodes in the 389-plex and 1,384-plex experiments. We hypothesized that uneven spacer abundance could hinder enrichment at larger scales. However, no correlation was found between spacer distribution and observed $\log_2$ enrichment for either the 389-plex (5′ G sgRNAs, replicate 2; see Fig. 27A in Appendix A) or the 1,384-plex (5′ G sgRNAs; Fig. 27B) experiments. This lack of correlation suggests that, despite substantial variation in sgRNA spacer distribution, other factors such as dCas9 binding dynamics or local sequence context may play a more dominant role in determining enrichment efficiency.

Additionally, as the number of sgRNAs increases, they compete for free dCas9, reducing the overall activity of all sgRNAs (78, 79). This phenomenon was highlighted in the development of a negative feedback circuit, which increases dCas9 expression when free dCas9 decreases due to the co-expression of multiple sgRNAs (80), regardless of sgRNA sequence. The impact of sgRNA competition is particularly dramatic. For instance, in one study using dCas9 to repress promoter expression, repression was 58-fold for a single sgRNA but dropped to just 10-fold when 7 sgRNAs were co-expressed (78). Furthermore, another study showed that dCas9 saturation occurs when approximately 12 sgRNAs compete for binding (81), which may explain the target dropouts observed when enriching more than 12 targets from the DHFR libraries. As a result, we predicted that increasing sgRNA, dCas9, or DNA inputs could mitigate this effect and improve enrichment scores. We wanted to increase the DNA input in particular to mitigate any increased off-target effects from increasing sgRNAs or dCas9, since it reduces the RNP to DNA ratio, as described previously (Figs. 1D,2A).

To test this, we enriched 12 targeted barcodes from the 384-gene DHFR library (library S4) while varying the sgRNA, dCas9, or DNA inputs to evaluate their impact on enrichment. We performed the standard BAR-CAT v1.0 protocol using 12-plex sgRNA pools with spacers

beginning with either a 5′ G or 5′ GGGG. Across all conditions, the percentage of target barcode dropouts was high (25–66.7%), likely due to issues with dCas9 reconstitution. Nonetheless, the results provided general preliminary trends. The 5′ G condition showed a 4.9 $\log_2$-fold median enrichment (29-fold) and a 17-fold population fraction enrichment with 41.7% target dropouts. The 5′ GGGG condition yielded a 6.1 $\log_2$-fold median enrichment (66-fold) and a 30-fold population fraction enrichment with an even higher percentage of target dropouts (66.7%). A condition using 3× more input sgRNAs (5′ G) resulted in a 4.1 $\log_2$-fold median enrichment (18-fold) and 9.4-fold population fraction enrichment, markedly lower than the corresponding 5′ GGGG or 5′ G controls. Increasing both DNA (2.2×) and dCas9 (3×) led to a 4.3 $\log_2$-fold enrichment (20-fold) and 14-fold population fraction enrichment. A condition using 2.2× DNA and 3× sgRNAs produced a 3.4 $\log_2$-fold enrichment (11-fold) and 4.2-fold population fraction enrichment. Lastly, we tested a 2× reaction volume condition, scaling sgRNAs by 3× and DNA by 5×, which resulted in a 4.6 $\log_2$-fold enrichment (24-fold) and 14-fold population fraction enrichment (Fig. 28, see Appendix A).

Overall, these results showed that increasing sgRNA or dCas9 concentrations did not substantially improve enrichment at the 12-plex scale, and increasing input DNA did not rescue this effect. In fact, enrichment scores often decreased, potentially due to greater off-target barcode enrichment due to excess active RNPs. Off-target barcode dropout ranged from 42.2% to 49.1% (Fig. 28, see Appendix A). Comparing enriched versus original barcode fractions across all conditions did not differ from those previously observed for 12-plex enrichments with the 384-gene DHFR library (library S4) (Fig. 29A–F, see Appendix A); both low-abundance off-targets and target barcodes that did not drop out were enriched. These results indicate that varying DNA, sgRNA, or dCas9 input has little effect on enrichment efficiency at small scales.

However, we anticipate that increasing dCas9 concentration will have a greater impact at larger scales (e.g., 389+ targets), where competition among sgRNAs for dCas9 is more pronounced. Thus, increasing dCas9 levels during RNP assembly, while keeping sgRNA levels constant, may reduce competition and ensure sufficient complex formation per target.

Since the enrichment of target barcodes from the DHFR gene libraries also exhibited off-target enrichment, we sought to better understand the identity of enriched off-target barcodes. We first examined whether enriched off-target barcodes contained perfect matches to the 7–10 base pair PAM-adjacent seed region of any of the 18 target spacers. However, we found no significant enrichment of off-target sequences with seed matches compared to those without seed matches (data not shown). This finding is surprising, as previous studies have shown that CRISPR off-target effects are primarily driven by seed sequence matches (41, 42, 58, 71).

One possible explanation for this lack of correlation is the promiscuity of dCas9 binding. Both dCas9 and Cas9 can bind to off-target sites in the presence of a PAM, even when mismatches occur within the seed region, and this binding may persist unless there are mismatches in the RDR as well (59). Although mismatches in the seed region generally reduce binding efficiency (42, 82), off-target enrichment could result from numerous partial sequence matches across a wide range of sequences in complex libraries like the DHFR libraries (83). While we cannot definitively explain why seed sequence similarity did not predict off-target enrichment in this case, we propose that dCas9's binding promiscuity, in combination with other experimental variables, contributes to the observed off-target enrichment trends. This suggests that factors such as the specific sequence context, spacer distribution, or RNP concentration may also influence off-target behavior in high-complexity libraries.

Since off-target enrichment does not appear to be driven by seed sequence similarity, re-designing target sequences may not significantly reduce off-target noise. Instead, we propose targeting barcodes based on their original abundance as a more effective strategy. For example, the three barcodes selected for BAR-CAT optimization (Fig. 8) were chosen from a set of 18 barcodes (Fig. 7B) spanning a range of abundance. Barcodes 14 and 15, selected from the low-to-medium abundance range, exhibited the best $\log_2$ enrichment scores following BAR-CAT v0.3 (proteinase K denaturation, Fig. 7D, Fig 20 in Appendix A). In contrast, enrichment efficiency for both the most abundant (BCs 1–13) and least abundant (BCs 16–18) barcodes declined (Fig. 20). This is because it is inherently easier for low abundance targets to achieve high fold enrichments. When fold change is normalized to the starting copy number, a barcode present at a single read needs only nine additional captures to register a 10-fold gain, whereas one that begins at 1,000 reads must accumulate 9,000 more molecules, a requirement that rapidly encounters binding and amplification saturation limits. In other words, high-abundance targets may show lower fold-enrichment because further amplification has limited impact on already abundant sequences, while targeting ultra-low-abundance barcodes could increase off-target enrichment as RNPs search for rare targets. Therefore, targeting barcodes with low-to-medium abundance may provide a "sweet spot" for efficient enrichment. The robust enrichment with reduced noise observed in Fig. 8 for barcodes 7, 8, and 15 may stem from this strategy. However, this approach may not be effective in very diverse libraries, like the 384-gene DHFR library (S4), where barcode distributions are skewed towards low abundance, leading to frequent target dropouts and higher off-target binding.

An alternative approach to improving both accuracy and scalability involves transitioning from dCas9 to catalytically active Cas9 in a modified version of BAR-CAT. Cas9 could enable

more stringent target selection by cleaving DNA, unlike dCas9, which only binds. Its ability to cleave even ultra-low abundance barcodes would facilitate their amplification by PCR, whereas dCas9-based methods rely on higher target abundance for effective enrichment. A Cas9-based strategy could linearize supercoiled targets within the gene library via cleavage, enabling adapter ligation and generation of unique PCR handles for selective amplification, similar to the FLASH method for detecting low-abundance DNA (76). This approach would likely reduce background barcodes more effectively than the dCas9 approach, which relies on bead washing which is less effective for eliminating off-targets. Similarly, Cas12a-Capture uses Cas12a to cleave targets for ligation and sequencing (84). While a Cas9-based BAR-CAT would necessitate additional steps such as cloning and ligation post-enrichment, it could also expand BAR-CAT's utility to targeted sequencing of rare alleles or pathogenic variants. However, if Cas9 exhibits non-specific cleavage, as reported in the literature (45), its off-target effects may be comparable to those observed with dCas9.

There are also distinct advantages to using dCas9 for BAR-CAT, particularly in applications involving cloning and length-independent DNA enrichment. First, increasing the multiplexity of our dCas9-based approach in future BAR-CAT versions could streamline direct cloning of supercoiled perfect genes into *E. coli*, thereby enhancing screening efficiency and functional validation compared to a Cas9-based system. Second, a key benefit of dCas9 is that it does not cleave DNA, making BAR-CAT enrichment inherently length-independent in principle. In BAR-CAT v1.0, barcodes were successfully targeted from ~2.7 kbp supercoiled gene libraries, demonstrating the method's capacity to enrich DNA of this length.

In practice, recovery of longer DNA fragments can be hindered by several factors. Early studies indicated that mechanical shear stress, such as pipetting, often fragmented DNA into

heterogeneous distributions of 0.5–10 kbp (85). However, modern protocols, including slow-pipetting techniques to reduce shearing of 100 kbp DNA (86), have improved the handling of HMW DNA, making pipetting less of a limiting factor. We predict that BAR-CAT may show reduced recovery of DNA targets >20 kbp due to slower diffusion kinetics and increased steric hindrance during bead capture of dCas9 and subsequent washing steps. These factors suggest that while enrichment of fragments up to 10 kbp is feasible, recovery efficiency may decline with increasing fragment size, depending on fragment design and reaction conditions. Direct experimentation will be required to accurately assess the DNA length limitations of BAR-CAT. Based on current predictions, BAR-CAT should be effective for enriching sequences up to 20 kbp with minimal issues.

In addition to synthetic gene libraries, dCas9-based BAR-CAT could theoretically be applied to the enrichment of ancient DNA for targeted next-generation sequencing, offering a potential alternative to RNA hybridization-based methods (87), which may be limited by mismatches or degradation. In this context, dCas9 could bind conserved housekeeping genes without cleaving the fragile, partially degraded DNA, enabling recovery of fragments even when only a few conserved targets are available. Thus, a highly multiplexed and optimized version of BAR-CAT could serve as a valuable tool for enriching both synthetic genes and ancient DNA fragments up to ~20 kbp in length, leveraging the inherent length-independence of dCas9 binding and pull-down.

Currently, BAR-CAT v1.0 successfully enriches 3–12 unique targets, demonstrating its potential and providing valuable insights for further method development. However, improving its scalability is critical to fully achieving the original goals of BAR-CAT. To address this, we plan to titrate dCas9 concentrations during enrichment reactions to better manage sgRNA

competition as library complexity increases. Additionally, we aim to selectively target barcodes with low to moderate abundance in DropSynth libraries to minimize off-target enrichment. We will also investigate whether sequential enrichment can boost the representation of targeted barcodes. For this, we hypothesize that targeting low-abundance barcodes corresponding to perfect gene assemblies will result in moderate enrichment levels. These moderate levels could then be enhanced by repeating the BAR-CAT process prior to final amplification, effectively driving moderately enriched barcodes to higher abundance and reducing off-target enrichment and noise. Finally, we will evaluate a BAR-CAT variant using Cas9 instead of dCas9 to assess whether direct cleavage can improve both enrichment specificity and scalability by enabling selective amplification of cleaved targets.

2.5 Conclusions

Although the application of DNA libraries has evolved from gene discovery and early sequencing efforts to large-scale vehicles of functional discovery, their role as drivers of molecular biology has not diminished. However, as massively parallel assays, protein engineering, functional genomics, and synthetic biology applications continue to expand, so will the demand for larger DNA libraries with longer genes. DropSynth is one notable method that aims to meet this demand, but as with other assembly methods, it suffers from the significant accumulation of oligo errors at long lengths >1 kbp. To address this issue, we developed BAR-CAT, a method that applies CRISPR-dCas9 to selectively enrich perfect gene assemblies. During its proof-of-concept, BAR-CAT enriched 18 unique barcodes distributed from high to low abundance within a barcoded *rfp* library by a median 6.3-fold enrichment value. Increasing the cumulative wash volume of the magnetic beads from 0.3 mL (6 × 50 µL) to 18 mL (9 × 2

mL) boosted the median enrichment to about 1.9-fold. Additionally, non-enriched off-targets were further eliminated following enrichment, reducing background. Meanwhile, adding a dCas9 denaturation step to minimize non-specific interactions between dCas9, DNA, and beads did not increase enrichment but contributed to greater off-target barcode dropout. Lastly, increasing the input DNA from 50 ng to 500 ng led to a median 600-fold enrichment value and an impressive 1,094-fold population fraction enrichment for three low-to-median abundance barcode targets in the *rfp* library. These improvements likely stem from adjusting the RNP-to-DNA ratio by adding more DNA, thereby limiting excess RNPs that could otherwise bind off-targets. All of these process optimizations were incorporated into BAR-CAT version 1.0, as presented in this work.

When BAR-CAT was applied to enrich ultra-low-abundance targets from 384- and 1,536-member DHFR libraries assembled by DropSynth, we observed excessive target dropouts exceeding 90% in some cases. At the 389-plex scale, some of these effects were attributable to sgRNA libraries containing a 5′ GGGG sequence, which resulted in insufficient sgRNA concentrations and excessive HMW products. However, we believe that additional factors, including the very high diversity of the DHFR library, targeting of extremely low-abundance members, and sgRNA competition for dCas9, contributed to robust off-target effects and severe dropout of targets. Future work will address these shortcomings and explore alternatives such as using Cas9 in place of dCas9 and implementing sequential enrichment strategies.

In conclusion, the development of BAR-CAT v1.0 provides a practical framework for researchers aiming to design CRISPR-based DNA enrichment methods. It offers insight into the nuances of enrichment behavior in a simplified yet still complex system. We hope that the potential utility of BAR-CAT, once further optimized, will inspire the development of additional

tools for applications such as synthetic gene enrichment, ancient DNA enrichment, diagnostics, and targeted next-generation sequencing.

2.6 Conflicts of Interest

N.V. and C.P. are named inventors on a patent based on this method. CP is a co-founder and holds equity in SynPlexity.

2.7 Bridge

This chapter presented the development of BAR-CAT, a CRISPR-dCas9-based method for enriching perfect synthetic gene sequences. Although the sgRNA libraries synthesized *in vitro*, as detailed in the following chapter, were integral to the optimization of BAR-CAT, the development of BAR-CAT and the sgRNA synthesis method proceeded largely in parallel. This was due to challenges encountered during sgRNA library optimization. While our sgRNA synthesis system is fully compatible with BAR-CAT, its applications extend beyond this method, with broader potential for CRISPR screens, *in vitro* CRISPR assays, and other research requiring scalable, high-quality sgRNA pools. Chapter 3 describes how we optimized IVT of sgRNA libraries using T7 RNAP to enhance both the uniformity and cost-effectiveness of this method.

2.8 References

1. Shizuya,H., Birren,B., Kim,U.J., Mancino,V., Slepak,T., Tachiiri,Y. and Simon,M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 8794–8797.

2. Wu,C., Xu,Z. and Zhang,H.-B. (2006) DNA Libraries. *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, 10.1002/3527600906.mcb.200300065.

3. Shendure,J. and Lieberman Aiden,E. (2012) The expanding scope of DNA sequencing. *Nat.*

Biotechnol., **30**, 1084–1094.

4. Kosuri,S. and Church,G.M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*, **11**, 499–507.

5. Chen,C., Liao,Y. and Peng,G. (2022) Connecting past and present: single-cell lineage tracing. *Protein Cell*, **13**, 790–807.

6. Fulco,C.P., Munschauer,M., Anyoha,R., Munson,G., Grossman,S.R., Perez,E.M., Kane,M., Cleary,B., Lander,E.S. and Engreitz,J.M. (2016) Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, **354**, 769–773.

7. Ford,K.M., Panwala,R., Chen,D.-H., Portell,A., Palmer,N. and Mali,P. (2021) Peptide-tiling screens of cancer drivers reveal oncogenic protein domains and associated peptide inhibitors. *Cell Syst.*, **12**, 716–732.e7.

8. Esposito,D., Weile,J., Shendure,J., Starita,L.M., Papenfuss,A.T., Roth,F.P., Fowler,D.M. and Rubin,A.F. (2019) MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.*, **20**, 223.

9. Rubin,A.F., Stone,J., Bianchi,A.H., Capodanno,B.J., Da,E.Y., Dias,M., Esposito,D., Frazer,J., Fu,Y., Grindstaff,S.B., *et al.* (2025) MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome Biol.*, **26**, 13.

10. Tan,Y., Zhang,Y., Han,Y., Liu,H., Chen,H., Ma,F., Withers,S.G., Feng,Y. and Yang,G. (2019) Directed evolution of an α1,3-fucosyltransferase using a single-cell ultrahigh-throughput screening method. *Sci. Adv.*, **5**, eaaw8451.

11. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C., *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.

12. Starita,L.M., Ahituv,N., Dunham,M.J., Kitzman,J.O., Roth,F.P., Seelig,G., Shendure,J. and Fowler,D.M. (2017) Variant interpretation: Functional assays to the rescue. *Am. J. Hum. Genet.*, **101**, 315–325.

13. Sayous,V., Lubrano,P., Li,Y. and Acevedo-Rocha,C.G. (2020) Unbiased libraries in protein directed evolution. *Biochim. Biophys. Acta Proteins Proteom.*, **1868**, 140321.

14. Lindenburg,L., Huovinen,T., van de Wiel,K., Herger,M., Snaith,M.R. and Hollfelder,F. (2020) Split & mix assembly of DNA libraries for ultrahigh throughput on-bead screening of functional proteins. *Nucleic Acids Res.*, **48**, e63.

15. Plesa,C., Sidore,A.M., Lubock,N.B., Zhang,D. and Kosuri,S. (2018) Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science*, **359**, 343–347.

16. Romanowicz,K.J., Resnick,C., Hinton,S.R. and Plesa,C. (2025) Exploring antibiotic resistance in diverse homologs of the dihydrofolate reductase protein family through broad Mutational

Scanning. *bioRxivorg*, 10.1101/2025.01.23.634126.

17. Sidore,A.M., Plesa,C., Samson,J.A., Lubock,N.B. and Kosuri,S. (2020) DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Res.*, **48**, e95.

18. Stemmer,W.P., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.

19. Holston,A.S., Hinton,S.R., Lindley,K.A., Kearns,N.C. and Plesa,C. (2023) Degenerate DropSynth for Simultaneous Assembly of Diverse Gene Libraries and Local Designed Mutants. 10.1101/2023.12.11.569291.

20. Beaucage,S.L. and Caruthers,M.H. (1981) Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.*, **22**, 1859–1862.

21. Masaki,Y., Onishi,Y. and Seio,K. (2022) Quantification of synthetic errors during chemical synthesis of DNA and its suppression by non-canonical nucleosides. *Sci. Rep.*, **12**, 12095.

22. Hoose,A., Vellacott,R., Storch,M., Freemont,P.S. and Ryadnov,M.G. (2023) DNA synthesis technologies to close the gene writing gap. *Nat Rev Chem*.

23. Choi,H., Choi,Y., Choi,J., Lee,A.C., Yeom,H., Hyun,J., Ryu,T. and Kwon,S. (2021) Purification of multiplex oligonucleotide libraries by synthesis and selection. *Nat. Biotechnol.*, 10.1038/s41587-021-00988-3.

24. Matzas,M., Stähler,P.F., Kefer,N., Siebelt,N., Boisguérin,V., Leonard,J.T., Keller,A., Stähler,C.F., Häberle,P., Gharizadeh,B., *et al.* (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.*, **28**, 1291–1294.

25. Luo,C., Tsementzi,D., Kyrpides,N., Read,T. and Konstantinidis,K.T. (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One*, **7**, e30087.

26. Schwartz,J.J., Lee,C. and Shendure,J. (2012) Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nat. Methods*, **9**, 913–915.

27. Tian,J., Gong,H., Sheng,N., Zhou,X., Gulari,E., Gao,X. and Church,G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.

28. Teer,J.K. and Mullikin,J.C. (2010) Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.*, **19**, R145–51.

29. Cello,J., Paul,A.V. and Wimmer,E. (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*, **297**, 1016–1018.

30. Uematsu,M. and Baskin,J.M. (2023) Barcode-free multiplex plasmid sequencing using Bayesian analysis and nanopore sequencing. *bioRxivorg*.

31. Slatko,B.E., Gardner,A.F. and Ausubel,F.M. (2018) Overview of next-generation sequencing technologies: Overview of next-generation sequencing. *Curr. Protoc. Mol. Biol.*, **122**, e59.

32. Brown,S.D., Dreolini,L., Wilson,J.F., Balasundaram,M. and Holt,R.A. (2023) Complete sequence verification of plasmid DNA using the Oxford Nanopore Technologies' MinION device. *BMC Bioinformatics*, **24**, 116.

33. Klein,J.C., Lajoie,M.J., Schwartz,J.J., Strauch,E.-M., Nelson,J., Baker,D. and Shendure,J. (2016) Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.*, **44**, e43.

34. Kim,H., Han,H., Shin,D. and Bang,D. (2010) A fluorescence selection method for accurate large-gene synthesis. *Chembiochem*, **11**, 2448–2452.

35. Wang,T., Badran,A.H., Huang,T.P. and Liu,D.R. (2018) Continuous directed evolution of proteins with improved soluble expression. *Nat. Chem. Biol.*, **14**, 972–980.

36. Biswas,I. and Hsieh,P. (1997) Interaction of MutS protein with the major and minor grooves of a heteroduplex DNA. *J. Biol. Chem.*, **272**, 13355–13364.

37. Lubock,N.B., Zhang,D., Sidore,A.M., Church,G.M. and Kosuri,S. (2017) A systematic comparison of error correction enzymes by next-generation sequencing. *Nucleic Acids Res.*, **45**, 9206–9217.

38. Obmolova,G., Ban,C., Hsieh,P. and Yang,W. (2000) Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature*, **407**, 703–710.

39. Sequeira,A.F., Guerreiro,C.I.P.D., Vincentelli,R. and Fontes,C.M.G.A. (2016) T7 Endonuclease I Mediates Error Correction in Artificial Gene Synthesis. *Mol Biotechnol*, **58**, 573–584.

40. Yoo,E., Choe,D., Shin,J., Cho,S. and Cho,B.-K. (2021) Mini review: Enzyme-based DNA synthesis and selective retrieval for data storage. *Comput. Struct. Biotechnol. J.*, **19**, 2468–2476.

41. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.

42. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A., *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

43. Aalipour,A., Dudley,J.C., Park,S.-M., Murty,S., Chabon,J.J., Boyle,E.A., Diehn,M. and Gambhir,S.S. (2018) Deactivated CRISPR Associated Protein 9 for Minor-Allele Enrichment in Cell-Free DNA. *Clin. Chem.*, **64**, 307–316.

44. Farasat,I. and Salis,H.M. (2016) A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLoS Comput. Biol.*, **12**, e1004724.

45. Collins,M., Lau,M.B., Ma,W., Shen,A., Wang,B., Cai,S., La Russa,M., Jewett,M.C. and Qi,L.S. (2024) A frugal CRISPR kit for equitable and accessible education in gene editing and synthetic biology. *Nat. Commun.*, **15**, 6563.

46. Villegas,N.K., Gaudreault,Y.R., Keller,A., Kearns,P., Stapleton,J.A. and Plesa,C. (2025) Optimizing *in vitro* transcribed CRISPR-Cas9 single-guide RNA libraries for improved uniformity and affordability. *bioRxiv*, 10.1101/2025.03.24.644170.

47. McCarty,N.S., Graham,A.E., Studená,L. and Ledesma-Amaro,R. (2020) Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nat. Commun.*, **11**, 1281.

48. Hsiung,C.C.-S., Wilson,C.M., Sambold,N.A., Dai,R., Chen,Q., Teyssier,N., Misiukiewicz,S., Arab,A., O'Loughlin,T., Cofsky,J.C., *et al.* (2024) Engineered CRISPR-Cas12a for higher-order combinatorial chromatin perturbations. *Nature Biotechnology*.

49. Wong,A.S.L., Choi,G.C.G., Cui,C.H., Pregernig,G., Milani,P., Adam,M., Perli,S.D., Kazer,S.W., Gaillard,A., Hermann,M., *et al.* (2016) Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 2544–2549.

50. Zhou,P., Chan,B.K.C., Wan,Y.K., Yuen,C.T.L., Choi,G.C.G., Li,X., Tong,C.S.W., Zhong,S.S.W., Sun,J., Bao,Y., *et al.* (2020) A Three-Way combinatorial CRISPR screen for analyzing interactions among druggable targets. *Cell Rep.*, **32**, 108020.

51. Kim,B., Kim,Y., Shin,S., Lee,S.-T., Cho,J.Y. and Lee,K.-A. (2022) Application of CRISPR/Cas9-based mutant enrichment technique to improve the clinical sensitivity of plasma EGFR testing in patients with non-small cell lung cancer. *Cancer Cell Int.*, **22**, 82.

52. Zorita,E., Cuscó,P. and Filion,G.J. (2015) Starcode: sequence clustering based on all-pairs search. *Bioinformatics*, **31**, 1913–1919.

53. Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.-D., Fernandez,J.P., Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*. *Nat. Methods*, **12**, 982–988.

54. Mateyko,N. and de Boer,C.G. (2024) Culture wars: Empirically determining the best approach for Plasmid library amplification. *ACS Synth. Biol.*, **13**, 2328–2334.

55. Guido,N.J., Handerson,S., Joseph,E.M., Leake,D. and Kung,L.A. (2016) Determination of a screening metric for high diversity DNA libraries. *PLoS One*, **11**, e0167088.

56. Bohlin,J., Snipen,L., Hardy,S.P., Kristoffersen,A.B., Lagesen,K., Dønsvik,T., Skjerve,E. and Ussery,D.W. (2010) Analysis of intra-genomic GC content homogeneity within prokaryotes. *BMC Genomics*, **11**, 464.

57. Mahfouz,M.M., Piatek,A. and Stewart,C.N.,Jr (2014) Genome engineering via TALENs and CRISPR/Cas9 systems: challenges and perspectives. *Plant Biotechnol. J.*, **12**, 1006–1014.

58. Fu,Y., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,J.K. and Sander,J.D. (2013)

High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.

59. Boyle,E.A., Andreasson,J.O.L., Chircus,L.M., Sternberg,S.H., Wu,M.J., Guegler,C.K., Doudna,J.A. and Greenleaf,W.J. (2017) High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 5461–5466.

60. Huang,J., Brieba,L.G. and Sousa,R. (2000) Misincorporation by wild-type and mutant T7 RNA polymerases: identification of interactions that reduce misincorporation rates by stabilizing the catalytically incompetent open conformation. *Biochemistry*, **39**, 11571–11580.

61. Wilson,I.G. (1997) Inhibition and facilitation of nucleic acid amplification. *Appl. Environ. Microbiol.*, **63**, 3741–3751.

62. Lorenz,T.C. (2012) Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *J. Vis. Exp.*

63. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.

64. Zou,R.S., Liu,Y. and Ha,T. (2021) *In vitro* Cleavage and Electrophoretic Mobility Shift Assays for Very Fast CRISPR. *Bio Protoc*, **11**, e4138.

65. Slesarev,A., Viswanathan,L., Tang,Y., Borgschulte,T., Achtien,K., Razafsky,D., Onions,D., Chang,A. and Cote,C. (2019) CRISPR/CAS9 targeted CAPTURE of mammalian genomic regions for characterization by NGS. *Sci. Rep.*, **9**, 3587.

66. David,S.R., Maheshwaram,S.K., Shet,D., Lakshminarayana,M.B. and Soni,G.V. (2022) Temperature dependent *in vitro* binding and release of target DNA by Cas9 enzyme. *Sci. Rep.*, **12**, 15243.

67. Dabney,J. and Meyer,M. (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, **52**, 87–94.

68. Aird,D., Ross,M.G., Chen,W.-S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

69. Pan,W., Byrne-Steele,M., Wang,C., Lu,S., Clemmons,S., Zahorchak,R.J. and Han,J. (2014) DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnol.*, **14**, 10.

70. Zhu,H. and Dean,R.A. (1999) A novel method for increasing the transformation efficiency of *Escherichia coli*-application forbacterial artificial chromosome library construction. *Nucleic Acids Res.*, **27**, 910–911.

71. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation

by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.

72. Liu,L., Li,W., Li,J., Zhao,D., Li,S., Jiang,G., Wang,J., Chen,X., Bi,C. and Zhang,X. (2023) Circular Guide RNA for Improved Stability and CRISPR-Cas9 Editing Efficiency *in Vitro* and in Bacteria. *ACS Synth. Biol.*, **12**, 350–359.

73. Conrad,T., Plumbom,I., Alcobendas,M., Vidal,R. and Sauer,S. (2020) Maximizing transcription of nucleic acids with efficient T7 promoters. *Commun. Biol.*, **3**, 439.

74. Henser-Brownhill,T., Monserrat,J. and Scaffidi,P. (2017) Generation of an arrayed CRISPR-Cas9 library targeting epigenetic regulators: from high-content screens to *in vivo* assays. *Epigenetics*, **12**, 1065–1075.

75. Chari,R., Yeo,N.C., Chavez,A. and Church,G.M. (2017) SgRNA Scorer 2.0: A species-independent model to predict CRISPR/Cas9 activity. *ACS Synth. Biol.*, **6**, 902–904.

76. Quan,J., Langelier,C., Kuchta,A., Batson,J., Teyssier,N., Lyden,A., Caldera,S., McGeever,A., Dimitrov,B., King,R., *et al.* (2019) FLASH: a next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. *Nucleic Acids Res.*, **47**, e83.

77. Marinov,G.K., Kim,S.H., Bagdatli,S.T., Higashino,S.I., Trevino,A.E., Tycko,J., Wu,T., Bintu,L., Bassik,M.C., He,C., *et al.* (2023) CasKAS: direct profiling of genome-wide dCas9 and Cas9 specificity using ssDNA mapping. *Genome Biol.*, **24**, 85.

78. Zhang,S. and Voigt,C.A. (2018) Engineered dCas9 with reduced toxicity in bacteria: implications for genetic circuit design. *Nucleic Acids Res.*, **46**, 11115–11125.

79. Chen,P.-Y., Qian,Y. and Vecchio,D.D. (2018) A model for resource competition in CRISPR-mediated gene repression. *bioRxiv*, 10.1101/266015.

80. Huang,H.-H., Bellato,M., Qian,Y., Cárdenas,P., Pasotti,L., Magni,P. and Del Vecchio,D. (2021) dCas9 regulator to neutralize competition in CRISPRi circuits. *Nat. Commun.*, **12**, 1692.

81. Clamons,S. and Murray,R. (2019) Modeling predicts that CRISPR-based activators, unlike CRISPR-based repressors, scale well with increasing gRNA competition and dCas9 bottlenecking. *bioRxiv*, 10.1101/719278.

82. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O., *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.

83. Tadić,V., Josipović,G., Zoldoš,V. and Vojta,A. (2019) CRISPR/Cas9-based epigenome editing: An overview of dCas9-based tools with special emphasis on off-target activity. *Methods*, **164-165**, 109–119.

84. Mighell,T.L., Nishida,A., O'Connell,B.L., Miller,C.V., Grindstaff,S., Thornton,C.A., Adey,A.C., Doherty,D. and O'Roak,B.J. (2022) Cas12a-Capture: A Novel, Low-Cost, and Scalable Method for Targeted Sequencing. *CRISPR J*, **5**, 548–557.

85. Dancis,B.M. (1978) Shear breakage of DNA. *Biophys. J.*, **24**, 489–503.

86. Prall,T.M., Neumann,E.K., Karl,J.A., Shortreed,C.G., Baker,D.A., Bussan,H.E., Wiseman,R.W. and O'Connor,D.H. (2021) Consistent ultra-long DNA sequencing with automated slow pipetting. *BMC Genomics*, **22**, 182.

87. Carpenter,M.L., Buenrostro,J.D., Valdiosera,C., Schroeder,H., Allentoft,M.E., Sikora,M., Rasmussen,M., Gravel,S., Guillén,S., Nekhrizov,G., *et al.* (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.*, **93**, 852–864.

3. Optimizing *in vitro* Transcribed CRISPR-Cas9 Single-Guide RNA Libraries for Improved Uniformity and Affordability

3.1 Author contributions

I was the primary contributor to the experimental portion of this work, while Dr. Calin Plesa led the computational analysis. Dr. Jim Stapleton provided the idea of using Golden Gate Assembly to join oligos encoding sgRNA spacers to a conserved scaffold oligo, which became the core of our scalable and affordable sgRNA library generation method. I developed the experimental protocol for bulk *in vitro* transcription, and Yukiko R. Gaudreault assisted with transcribing the proof-of-concept library containing 18 sgRNAs.

Dr. Calin Plesa designed and ordered all oligo pools for our large-scale experiment producing 10 sgRNA libraries with 206 to 2,626 unique sgRNA spacers. I sub-pooled the oligos and transcribed the resulting sgRNA libraries, while Yukiko R. Gaudreault implemented the RNA-seq workflow that we used for all subsequent sgRNA libraries. Dr. Plesa also developed a Python- and R-based pipeline for RNA-seq analysis, although I ran the Python script to analyze several datasets.

After observing severe bias in sgRNA spacer distribution, Phillip Kearns developed a machine learning model that suggested T7 RNA polymerase was the source of the bias. Dr. Plesa proposed IVT of sgRNAs in emulsions, and I developed a custom experimental protocol for this method, with Abigail Keller assisting in its execution. I also conceived and tested the idea that increasing input DNA or changing reaction volume could reduce sgRNA spacer bias.

I interpreted most of the results and was the primary writer of this work, with contributions from Yukiko R. Gaudreault. Dr. Calin Plesa, Dr. Jim Stapleton, Yukiko R. Gaudreault, and I all contributed to editing and revising the manuscript. Dr. Calin Plesa created

most of the figures based on discussions we had; I designed Figure 47 (Appendix B). We jointly calculated the statistics.

3.2 Introduction

CRISPR-Cas9 technology has revolutionized DNA manipulation, advancing high-throughput gene editing, functional genomics, genetic engineering, and large-scale functional screens. At the core of this technology are single-guide RNAs (sgRNAs) with 20-nucleotide spacer sequences that enable precise targeting of both coding and non-coding DNA, thanks to their high programmability (1). This versatility allows for multiplex functional genomics studies, where multiple genetic elements can be investigated simultaneously.

To fully harness this capability, these studies rely on sgRNA libraries containing thousands to hundreds of thousands of unique sgRNA spacers targeting diverse genomic regions (2). Libraries such as GeCKOv2 (3), Brunello (4), Sabatini human knockout (5), and human CRISPRi v2 (6) are commonly used for genome-wide CRISPR knockout (CRISPR-KO), interference (CRISPRi), and activation (CRISPRa) screens. These libraries typically feature sgRNA sequences synthesized as microarray-derived oligonucleotides and subsequently cloned into lentiviral vectors to transfect cell lines.

To reduce the costs associated with CRISPR screens using these large genome-wide sgRNA libraries such as GeCKOv2, numerous smaller, more cost-effective libraries have been developed (7). One such library contains four sgRNAs per gene (8), compared to the 10 used in the large genome-wide Sabatini library (5). Another is modular, consisting of three sublibraries with 2–4 sgRNAs per gene that are synthesized separately and can be combined to adjust sgRNA

representation (9). However, even these smaller libraries are typically produced as distinct sublibraries rather than amplified from a larger pool.

In contrast, amplifying sgRNA subpools from microarray-derived oligos provides a more efficient and cost-effective approach for targeted screening. One parallel retrieval strategy generated 24 subpools, each containing 1,000 unique sgRNA sequences, with coverage ranging from 98.8% to 99.8% (10). Beyond its efficiency, this method leverages bulk pricing advantages, as a single pool of microarray-synthesized oligonucleotides can be subdivided into multiple libraries. For example, ordering five separate microarray-synthesized pools, each containing 1,000 oligos, costs $1,120 per pool, totaling $5,600. In contrast, ordering a single pool of 5,000 oligos costs $1,680, resulting in 72% savings (Twist Bioscience). Furthermore, this approach enables the generation of multiple targeted sgRNA libraries from a single oligo pool. By subdividing microarray-derived oligo pools, researchers can design libraries tailored to specific cellular states or protein families, optimizing representation within individual subpools.

These examples outline a pathway for generating sgRNA libraries that are fully programmable and partitionable for targeted, cost-effective studies. While these methods offer substantial improvements, they typically rely on lentiviral vector delivery, posing challenges for certain cell types and contexts. While lentiviral vectors allow for stable sgRNA expression in human cell lines, their prolonged expression increases the risk of off-target editing (11). Additionally, random genomic integration can lead to mutagenesis and prolonged Cas9 expression, further exacerbating off-target effects (11, 12). In contrast, delivery of active ribonucleoprotein complexes (RNPs), comprising of synthesized sgRNAs with Cas9 or catalytically inactivated Cas9 (dCas9), has been shown to achieve higher editing efficiencies and reduced off-target effects compared to lentiviral plasmid expression, even in cells that are

notoriously difficult to transfect (11–13). Additionally, RNPs can be produced from synthesized sgRNAs and dCas9 fused to small molecules or peptides, enhancing the recruitment of epigenetic proteins to DNA sites in mammalian cells (14). RNPs are also compatible with *in vitro* CRISPR-Cas9 applications, as seen in CasKAS, a method for identifying off-target sites via N3-kethoxal binding to single-stranded DNA. Synthetic sgRNA RNPs can be electroporated into cells or applied to DNA *in vitro*, highlighting their versatility (13).

To enable the use of RNPs, sgRNAs can be synthesized chemically or enzymatically with T7 RNA polymerase (T7 RNAP), allowing flexible design. Although chemical synthesis yields high-purity products, it is expensive and results in low yields, making it impractical for large-scale sgRNA libraries (11). Enzymatic synthesis is more affordable, produces high sgRNA yields, and is accessible using commercially available kits for *in vitro* transcription (IVT) by T7 RNAP. Although commercially available IVT kits provide an affordable means to synthesize individual sgRNAs, they are not a cost-effective solution for generating large-scale, multiplex sgRNA libraries. This limitation primarily arises from the need for specific DNA oligo inputs. Microarray-derived oligos yield only femtomole (fmol) to picomole (pmol) quantities per sequence, making them insufficient for direct large-scale IVT without amplification. While pooled, column-synthesized oligos (oPools) offer greater material yields, their cost is approximately tenfold higher than microarray-derived oligos, making them less practical for high-throughput applications. That said, Cas12a-Capture successfully used 11,438 full-length crRNAs compatible with Cas12a, ordered as two oPools (15), demonstrating that oPools are a viable input source for crRNA libraries. Finally, these IVT kits typically require single-stranded DNA (ssDNA) as input and are incompatible with double-stranded DNA (dsDNA) obtained from PCR amplification of sgRNA subpools (10).

In this work, we developed a highly customizable, scalable, and programmable *in vitro* method for producing sgRNA libraries using T7 RNAP. We first amplified subpools of sequences from thousands of unique microarray-synthesized DNA target oligos containing sgRNA spacers, taking advantage of affordable bulk pricing. To further reduce costs, we minimized oligo length by excluding the conserved sgRNA scaffold sequence required for Cas9 complexation (16). We then used Golden Gate Assembly (GGA) to ligate the dsDNA spacer fragment with the scaffold sequence, efficiently generating full-length templates for transcription (Fig. 10A, see 3.4 Results and Discussion).

GGA enables the ligation of multiple unique DNA fragments in a single reaction due to the activity of its type IIs restriction enzymes, which generate precise, user-specified overhangs (17). This approach offers distinct advantages over homology-based methods such as Gibson Assembly, uracil-specific excision reagent cloning (USER), and In-Fusion Cloning (Takara Bio). The DNA target oligos include a reverse primer site for subpooling that must be removed before assembly with the scaffold oligo. If left intact, this primer site remains adjacent to the scaffold oligo, displacing the spacer sequence that should be positioned at its 5' end, thereby disrupting CRISPR targeting. To address this, a GGA restriction cut site was placed upstream of the reverse primer site, enabling a type IIs enzyme to cleave the priming site during DNA fragment assembly.

In contrast, homology-based methods would require multiple additional steps, including the digestion of unwanted primer sites, purification of DNA fragments, and hybridization of the two fragments. This increases workflow complexity and labor requirements, reducing the cost-effectiveness of adding the scaffold oligo separately. By introducing the conserved scaffold oligo after subpooling, we achieve an additional 14% cost savings. This contributes to the overall 72%

reduction in cost achieved by subpooling microarray-derived oligos, as shorter oligo pools are more affordable. An additional advantage of GGA is that assembly is guided by short restriction sticky ends, whereas Gibson Assembly requires relatively long 20–40 nt homology regions for efficient assembly. Because the spacer region varies between sequences, designing consistent homology regions for Gibson Assembly would require extending the conserved sgRNA scaffold region beyond the spacer, significantly increasing oligo length and cost while also raising the risk of mis-hybridization (18, 19). While the USER method requires shorter overhangs for assembly, it is more complex and expensive compared to Gibson Assembly or GGA (18). Overall, these modifications led to substantial cost reductions while avoiding potential complications in our sgRNA synthesis process.

Once the DNA templates were assembled, we transcribed full-length sgRNAs *in vitro* using T7 RNAP and assessed metrics like spacer uniformity and coverage via RNA-seq analysis (Fig. 10). Our results validate the viability of this sgRNA synthesis workflow and reveal how IVT reaction conditions influence spacer distributions in sgRNA libraries. Since T7 RNAP preferentially transcribes certain sequences (20), we investigate how this bias affects spacer distribution. Specifically, spacer biases result in over-representation of certain spacers while under-representing others, potentially causing important functional hits to be missed. As a result, a significantly larger number of cells must be screened to achieve robust statistical power in CRISPR-based assays (21). Moreover, uneven sgRNA library coverage and uniformity can undermine the reproducibility and accuracy of these assays, limiting the reliability of their conclusions. To address this issue, we demonstrate the effectiveness of three distinct strategies to reduce this bias within sgRNA libraries containing 389 or 2,626 unique spacers and propose future directions for further optimization.

In summary, we present a scalable enzymatic sgRNA synthesis method optimized to improve spacer uniformity. This approach leverages T7 RNAP transcription to generate user-defined sgRNA libraries with reduced bias. As a result, our method provides a robust solution for *in vivo* CRISPR screening in human or mammalian cell lines and *in vitro* CRISPR assays.

3.3 Materials and Methods

Sourcing and Pooling 18 Oligos for the Pilot sgRNA Library

Eighteen individual 98nt oligos were ordered from Integrated DNA Technologies (IDT, Coralville, USA) as single-stranded DNA (ssDNA) oligos. The oligo sequences are listed in Table 6 (see Appendix B). To pool the oligos, 2 µL of each 100 µM target oligo was combined to a total volume of 36 µL. Subsequently, at least 10 reverse single primer extension (RSPE) reactions were performed to convert the pooled target oligos into double-stranded DNA. Each RSPE reaction contained 1.25 µL of pooled target oligos, 2.5 µL of 100 µM reverse primer (skpp15-1-R filt15-1181), 12.5 µL of 2X Q5 Hot-Start High-Fidelity Master Mix (NEB), and nuclease-free water to a final volume of 25 µL. The thermocycler RSPE program consisted of an initial denaturation at 98°C for 30 sec, followed by one cycle of 98°C for 30 sec, 64°C for 30 sec, and 72°C for 60 min. DNA Cleanup Columns capable of binding five micrograms of DNA (NEB) were used to purify the double-stranded DNA oligonucleotides (dsDNA oligos).

Preparing Microarray-Derived Oligos for Large-Scale sgRNA Synthesis

The pool of 11,640 microarray-derived oligos (chip9) that was resuspended to 4.3 ng/µL following manufacturer guidelines (Twist Bioscience, South San Francisco, USA). The 98 nt

chip9 oligos included 10 unique forward and reverse primer pair sites for PCR subpooling, enabling the generation of templates for 10 sgRNA libraries.

Subpool Amplification of 10 Microarray-Derived Libraries Prior to PCR Optimization

First, chip9 OLS were diluted 10-fold to a concentration of 0.43 ng/µL. Subpooling qPCRs were prepared using library-specific primer pairs and annealing temperatures as listed in Table 7 (see Appendix B). Each qPCR included 0.1 ng of diluted chip9 OLS, 1.25 µL each of 10 µM forward and reverse primers, 0.25 µL of 100X Biotium Thiazole Green (Thermo Fisher Scientific), 12.5 µL of 2X Kapa HiFi HotStart ReadyMix (Roche Sequencing Solutions Inc, Pleasanton, USA), and nuclease-free water to a total volume of 25 µL. The qPCR program included: initial denaturation at 98°C for 45 seconds, followed by 50 cycles of 98°C for 10-15 seconds, annealing for 15 seconds (see Table 7 in Appendix B for temperatures), and extension at 72°C for 15 seconds. PCRs were prepared without Thiazole Green, and each subpool was amplified with the number of PCR cycles corresponding to the amplification plateau observed via qPCR (Table 7). The ten subpooled oligo libraries were purified using 5 µg DNA Cleanup Columns (NEB).

Bulk Amplification Prior to PCR Optimization

Bulk amplification qPCR was prepared similarly to subpooling qPCR, except that 0.01 ng of subpooled DNA oligos was added per reaction. This adjustment was made to prevent the deckchair effect observed with larger template inputs. Additionally, the qPCR initial denaturation and denaturation times were reduced to 30 and 10 seconds, respectively. PCR was performed with the number of qPCR cycles corresponding to the amplification plateau for each subpool (Table 7, see Appendix B), and a final 1-minute extension at 72°C. The ten bulk-amplified libraries were purified using 5 µg DNA Cleanup Columns (NEB).

Subpool Amplification Optimization of 389-, 1,382-, and 2,626-Plex Microarray-Derived Libraries

To minimize the number of PCR cycles required for subpooling chip9 389-, 1,382-, and 2,626-plex libraries, qPCRs were performed with varying amounts of OLS template. For the 389-plex library, qPCR reactions included 0.1 ng, 0.25 ng, 0.5 ng, or 1 ng of 0.43 ng/µL chip9 OLS, along with 1.25 µL of 10 µM forward primer (skpp15-23-F filt15-577), 1.25 µL of 10 µM reverse primer (skpp15-23-R filt15-1596), 0.25 µL of 100X Biotium Thiazole Green (Thermo Fisher Scientific), 12.5 µL of 2X Kapa HiFi HotStart ReadyMix (Roche Sequencing Solutions Inc., Pleasanton, USA), and nuclease-free water to a final volume of 25 µL.

The qPCR program consisted of an initial denaturation at 98°C for 30 seconds, followed by 50 cycles of 98°C for 10–15 seconds, annealing at 53°C for 15 seconds (for the 389-plex subpool), and extension at 72°C for 15 seconds. The lowest cycle number was achieved using 1 ng of chip9 OLS, which was selected for subpooling the 389-plex library.

To determine the optimal amplification cycles for subpooling the 1,382-plex library, qPCR was performed with 1 ng of template using skpp15-6 primers and annealing at 53°C for 15 seconds (Table 7, see Appendix B). The same approach was applied to the 2,626-plex library using the skpp15-27 primer pair with an annealing temperature of 51°C for 15 seconds. All three oligo libraries (389-, 1,382-, and 2,626-plex) were subpooled using the qPCR-determined cycle numbers (Table 7) with a 1-minute extension at 72°C. Subpooled oligos were purified using 5 µg DNA Cleanup Columns (NEB).

Bulk Amplification Optimization of 389-, 1,382, and 2,626–Plex Microarray-Derived Libraries

To reduce PCR cycles for bulk amplification of chip9 389-, 1,382-, and 2,626-plex oligo libraries, qPCR was performed with 1 ng, 10 ng, or 20 ng of subpooled DNA and the same primers as used for subpooling. Amplification with 10 ng of template resulted in the least PCR cycles. PCR was performed with the number of qPCR cycles corresponding to the amplification plateau for each library (Table 7, see Appendix B). A final 1-minute extension at 72°C was also added. All three bulk-amplified libraries were purified using 5 µg DNA Cleanup Columns (NEB).

Column-Synthesized (oPool) DNA Library Templates Prepared with One PCR Cycle

Two column-synthesized oPools containing 120-nt target oligos with a T7 promoter, 20-nt spacer, and sgRNA scaffold sequence, were obtained from Integrated DNA Technologies (IDT, Coralville, USA). The oPools were resuspended in Tris-EDTA (TE) buffer, pH 8.0 according to the manufacturer's guidelines. The S2 oPool contained oligos with 1,382 unique spacers from chip9, while S4 oPool contained oligos with 389 unique spacers from chip9.

S2 and S4 oPools were converted to dsDNA by performing RSPE as follows. Three reactions were prepared per oPool, each containing 2.5 µL of µM reverse primer (sgRNA_GGA_oligo_REV_NV; Table 8, see Appendix B), 500 ng of S2 or S4 oligos, 12.5 µL of 2X Q5. The thermocycler RSPE program consisted of an initial denaturation at 98°C for 30 sec, followed by one cycle of 98°C for 10 sec, 70°C for 30 sec, and 72°C for 60 min. DNA Cleanup Columns capable of binding five micrograms of DNA (NEB) were used to purify the dsDNA oligos.

Column-Synthesized (oPool) sgRNA library templates with 5' Guanine Tetramer Spacers

One oPool (e13sgRNA) containing 473 oligos that were all 98 nt in length was obtained from Integrated DNA Technologies (IDT, Coralville, USA) and was resuspended in Tris-EDTA

(TE) buffer, pH 8.0 according to the manufacturer's guidelines. The oligos included four unique forward and reverse primer pair sites for PCR subpooling four target oligo libraries. Three subpools included a 5' GGGG sequence at positions +1 to +4 after the T7 promoter, with 12 (12G4), 60 (60G4), or 389 (389G4) unique spacers. The fourth subpool had the same 12 spacers, but each started with 5' G (12G) at the +1 position.

Subpool Amplification of oPool Oligos

qPCR was performed with varying amounts of oPool template to optimize subpooling of the 12G and 12G4 libraries. Subpool-specific primer pairs and corresponding annealing temperatures are in Table 9 (see Appendix B). Each subpool qPCR included 0.1 ng, 1 ng, or 10 ng of oPool template, 1.25 µL of 10 µM forward and reverse primers, 0.25 µL of 100X Biotium Thiazole Green (Thermo Fisher Scientific), 12.5 µL of 2X Kapa HiFi HotStart ReadyMix (Roche Sequencing Solutions Inc, Pleasanton, USA), and nuclease-free water to a total volume of 25 µL. The qPCR program included the following steps: initial denaturation at 95°C for 3 min, followed by 40 cycles of 98°C for 15 seconds, annealing for 15 seconds (see Table 9 in Appendix B for temperatures), and extension at 72°C for 15 seconds. Since 10 ng of oPool template resulted in the least number of cycles for amplification, qPCRs were also performed for 60G4 and 389G4 with this amount.

Subpool amplification of 12G, 12G4, 60G4, and 389G4 target oligos was accomplished by preparing 32-42 reactions with 10 ng of oPool template. Each subpool was amplified with the number of PCR cycles corresponding to the amplification plateau observed via qPCR (Table 9, see Appendix B). The subpooled target oligo libraries were purified using 5 µg DNA Cleanup Columns (NEB).

Golden Gate Assembly of Double-Stranded DNA Templates for IVT

Double-stranded DNA templates for IVT were prepared using a two-fragment Golden Gate Assembly (GGA). For each sgRNA library, 3-10 GGA reactions were performed. Each assembly reaction contained 2.0 pmol of duplexed 83 bp sgRNA scaffold oligos combined with 1.0 pmol of subpooled 98 bp oligos (either oPool or microarray-derived in origin) in an approximate 2:1 ratio. The sgRNA scaffold oligo (sequence listed in Table 8, Appendix B) was initially duplexed by annealing it to its reverse complement. However, this step was streamlined by ordering the sgRNA scaffold oligo already duplexed from Integrated DNA Technologies (IDT, Coralville, USA). The reaction mixture also included 2.5 µL of 10X T4 DNA ligase reaction buffer supplemented with 2.5 µL ATP to a final concentration of 1 mM (New England Biolabs, NEB, Ipswich, USA), 0.25 µL T4 DNA ligase (400,000 units/mL) (New England Biolabs, NEB, Ipswich, USA), 0.75 µL of BsaI-HF®v2 (20,000 units/mL) (New England Biolabs, NEB, Ipswich, USA), and nuclease-free water to a final volume of 25 µL.

The thermocycler GGA protocol consisted of 100 cycles, with each cycle containing the following steps: 37°C for 5 min and 16°C for 5 min. A heat inactivation step of 80°C for 20 min was added to denature T4 DNA ligase and BsaI-HF®v2. A 5 µg DNA Cleanup Column (NEB) was used to purify the assemblies. Verification of assemblies was performed with 2% E-Gel™ EX Agarose Gels (Thermo Fisher Scientific, Waltham, USA) with a 50 bp ladder (New England Biolabs, NEB, Ipswich, USA).

For the 18-plex and microarray-derived sgRNA libraries shown in **Fig. 10** (see 3.4 Results and Discussion), the GGA products corresponding to the IVT DNA template (135 bp) were amplified with PCR using the SgRNA_GGA_oligo_FWD_NV and sgRNA_GGA_oligo_REV_NV primers (**Table 6)**. The resulting PCR product was size selected from agarose gels and used as the input DNA for IVT. However, this GGA product amplification

step was eliminated from our workflow due to spacer coverage reductions due to excessive PCR cycles (Figs. 10C and 11A, see 3.4 Results and Discussion).

IVT of Pooled sgRNAs with T7 RNA Polymerase

IVT was performed using sgRNA template dsDNA oligos (135 bp) prepared with GGA. One to three IVT reactions were prepared per individual sgRNA library or subpooled library. For each standard 20 µL reaction, components were added in the following order: 0.5 µL of Murine RNase inhibitor (40,000 units/mL) (NEB), nuclease-free water, 0.01-4.5 pmol (1-400 ng) of template DNA, 2 µL of 100 mM dithiothreitol (NEB), 2 µL of 10X RNA polymerase reaction buffer (NEB), 2 µL of 25 mM ribonucleotide solution mix (NEB), 2 µL of T7 RNA polymerase (50,000 units/mL) (NEB), and nuclease-free water to a final volume of 20 µL. To prepare 100 µL reactions, all reagent volumes were scaled up 5-fold in volume, except for the input DNA amount. A negative control lacking template DNA was also prepared for all experiments. Reactions were mixed by pipetting, spun down, and incubated in a thermocycler at 37°C for two hours.

DNase Treatment of Transcribed sgRNAs

The transcribed sgRNAs were treated with 5 µL of 10X DNase I reaction buffer (NEB), 0.5 µL of DNase I (2,000 units/mL) (NEB), and nuclease-free water to bring the total volume to 50 µL. Reactions were mixed by pipetting, spun down, and incubated at 37°C for 10 min. A 50 µg RNA Cleanup Column (NEB) was used to purify the sgRNA libraries.

Gel Verification of sgRNA Synthesis Products

The synthesized sgRNAs were evaluated by Tris-borate-EDTA 10% polyacrylamide urea denaturing gel electrophoresis (Bio-Rad Laboratories, Hercules, USA). A single-stranded low-range RNA ladder (NEB) and an *S. pyogenes* control sgRNA sequence from the NEB EnGen®

sgRNA Synthesis Kit—ordered as a synthetic sgRNA from Integrated DNA Technologies (IDT, Coralville, USA) (Table 8, see Appendix B)—were included for size comparison. Samples were initially prepared by adding 2 µL of 2X TBE-urea sample loading buffer containing 7 M urea (Boston BioProducts, Milford, USA). However, due to product discontinuation, samples were later prepared with 2X RNA loading dye containing 47.5% Formamide (New England Biolabs, NEB, Ipswich, USA).

IVT of sgRNA Libraries in Emulsions (eIVT)

First, IVT reactions were prepared using sgRNA template DNA generated through GGA. One IVT reaction was prepared per sgRNA target oligo subpool. First, components were added in the following order to prepare a 100 µL aqueous phase reaction: 2.5 µL of Murine RNase inhibitor (40,000 units/mL), nuclease-free water, 0.01-9 pmol (1-800 ng) of template DNA, 10 µL of 100 mM dithiothreitol, 10 µL of 10X RNA polymerase reaction buffer, 10 µL of 25 mM ribonucleotide solution mix, 10 µL of T7 RNA polymerase (50,000 units/mL), and 10 µL of 20 mg/mL recombinant albumin (NEB). A negative control lacking template DNA was also prepared.

An oil phase for each IVT was prepared by adding 600 µL of QX200™ Droplet Generation Oil for EvaGreen to non-stick 1.5 mL tubes (Bio-Rad Laboratories, Hercules, USA) as described by Plesa and colleagues (2018). This was later switched to Pico-Surf® (2% (w/w) in Novec™ 7500) from Sphere Bio (Cambridge, UK). A pipette was used to transfer the 100 µL aqueous phase to the bottom of the tube containing the oil phase. This process was repeated for each reaction corresponding to a unique sgRNA library. The tubes were sealed with parafilm and vortexed in Vortex Genie 2 (Scientific Industries) at 3000 rpm for 3 min. Then, the IVT reactions in emulsions (eIVT) were incubated in a thermomixer at 37°C for two hours.

To begin breaking the emulsions following incubation, eIVTs were mixed thoroughly by pipetting, and each reaction was evenly distributed between two 1.5 mL tubes, with approximately 350 µL per tube. To each tube, 175 µL of Monarch® DNA Elution Buffer (New England Biolabs, NEB, Ipswich, USA) and 612.5 µL of chloroform were added. The tubes were sealed using parafilm and then vortexed by Vortex Genie 2 at 3000 rpm for one min. Vacuum grease (phase lock gel) was applied inside the tubes, which were then centrifuged at 15,500 × g for 10 min. The 44.5 µL of upper aqueous phase pertaining to each eIVT tube was aliquoted into 0.2 mL PCR tubes and combined with 5 µL of 10X DNase I reaction buffer (NEB), 0.5 µL of DNase I (2,000 Units/mL) (NEB), and nuclease-free water to bring the total volume to 50 µL. DNase I treatment of transcribed sgRNAs was incubated at 37°C for 10 min. A 50 µg RNA Cleanup Column (NEB) was used to purify the sgRNA libraries. The sgRNA libraries were evaluated as described in the section "**Verification of sgRNA synthesis".**

RNA-Seq validation of sgRNA libraries using 5' RACE protocol

A 5' Rapid Amplification of cDNA Ends (RACE) protocol was used for RNA-seq of sgRNA libraries, consisting of first strand synthesis, template switching, quantitative PCR (qPCR) validation, bulk PCR amplification, and product verification. We selected this 5' RACE protocol to capture the variable spacer sequence located on the 5' end of the sgRNAs.

For first-strand synthesis, a custom gene-specific 49 nt RT primer (RT_primer_YG, 49 nt) was annealed to the 3' end of sgRNAs to add an overhang with a priming region and an 18 nt unique molecular identifier (UMI). Each annealing reaction included 0.5 µL Murine RNase inhibitor (40,000 units/mL) (New England Biolabs, NEB), 100 ng to 1 µg of sgRNA library, 1 µL 10 µM of RT_primer_YG (Table 10, see Appendix B), 1 µL 10 µM Deoxynucleotide (dNTP)

Solution Mix (NEB), and nuclease-free water to a total volume of 6 µL. Reactions were mixed by pipetting, spun down, and incubated at 70°C for 5 min.

RT and template switching were accomplished by combining one RT reaction per annealing reaction for a total volume of 10 µL. Each 4 µL RT reaction was prepared by combining 2.5 µL of 4X Template Switching RT Buffer (New England Biolabs, NEB, Ipswich, USA), 0.5 µL of 75 µM template switching oligo (TSO_YG; Table 10, see Appendix B), and 1 µL of 10X Template Switching RT Enzyme Mix (NEB). The combined RT and annealing reactions were mixed by pipetting, spun down, and incubated in a thermocycler for 90 min at 42 C, followed by 5 min at 85°C.

Diluted template-switched cDNA (10 µL RT reactions diluted 2-fold) was then used for qPCR. Each qPCR included 1 µL diluted cDNA, 1.25 µL 10 µM TSO-specific primer (TSO_primer_YG; Table 10), 1.25 µL 10 µM gene-specific primer (Gene_specific_primer_YG; Table 10), 0.25 µL 100X Biotium Thiazole Green (Thermo Fisher Scientific), 12.5 µL 2X Q5 Hot-Start High-Fidelity Master Mix (NEB), and nuclease-free water to a total volume of 25 µL. The thermocycler PCR program consisted of four stages: Stage one included an initial denaturation at 98°C for 30 sec. Stage two consisted of five cycles with denaturation at 98°C for 10 sec and annealing/extension at 72°C for 5 sec per kilobase (30 sec/kb). Stage three involved five cycles with denaturation at 98°C for 10 sec and annealing at 70°C for 5 sec (30 sec/kb). Stage four comprised 40 cycles with denaturation at 98°C for 10 sec, annealing at 65°C for 15 sec, and extension at 72°C for 5 sec (30 sec/kb). To prevent over-amplification of cDNA products, the number of cycles corresponding to the amplification plateau was selected for each sgRNA library. RT samples with high cDNA yields that amplified in less than 10 cycles were further diluted twofold and re-amplified via qPCR.

Bulk PCR setup followed that of the qPCR validation, excluding the addition of Thiazole Green. Amplification followed the same PCR protocol as described earlier, except that a final extension step at 72°C for 5 min was added. Additionally, the number of cycles specified for stage four of the thermocycler PCR program was determined individually for each template based on prior qPCR analysis. PCR products were cleaned and concentrated using DNA Cleanup Columns (NEB) or GeneJET PCR clean-up columns (Thermo Fisher Scientific). The amplified cDNA products were verified on 2% or 4% E-Gel™ EX Agarose Gels (Thermo Fisher Scientific) with a 50 bp ladder (NEB).

Following verification, the cDNAs were initially extended from 186 bp to 1.049 kbp (see "**Lengthening cDNAs for Nanopore Sequencing**") and submitted to Plasmidsaurus (South San Francisco, USA) for long-read sequencing. This was done only for the 18-plex and 10 microarray-derived sgRNA libraries shown in **Fig. 10 (see 3.0 Results and Discussion)**. The remaining sgRNA libraries were submitted without extension once Plasmidsaurus began offering short-read nanopore sequencing.

Lengthening cDNAs for Nanopore Sequencing

Gibson assembly was used to add a pUC19 fragment to the 186 bp cDNA products of the 18-plex and microarray-derived sgRNA libraries shown in **Fig. 10 (see 3.4 Results and Discussion)**, extending their length to over 600 bp to meet the long-read nanopore sequencing requirements at the time.

To amplify a linear fragment of pUC19 for Gibson Assembly, each PCR reaction was prepared with 200 ng of pUC19 (Bayou Biolabs, Metairie, USA), 1.25 µL of 10 µM forward primer (sgRNA_cDNA_gibson_FWD_YG; Table 10, see Appendix B), 1.25 µL of 10 µM reverse primer (sgRNA_cDNA_gibson_REV_YG; Table 10), and 12.5 µL of 2X Q5 Hot-Start

High-Fidelity Master Mix (NEB). Nuclease-free water was added to bring the total reaction volume to 25 µL. The thermocycler PCR program consisted of an initial denaturation at 98°C for 30 sec, followed by 25 cycles of 98°C for 10 sec, 67°C for 30 sec, and 72°C for 45 sec, with a final extension at 72°C for 2 min. The resulting 1.576 kilobase pairs (kbp) PCR product was purified with 5 µg DNA Cleanup Columns (NEB) and digested with DpnI to remove any remaining pUC19 templates. Each DpnI digest contained 1 µg of the amplified linear pUC19 fragment, 5 µL of 10X rCutSmart Buffer (NEB), 1 µL DpnI (20,000 units/mL) (NEB), and nuclease-free water to a total volume of 50 µL. The digests were incubated in a thermocycler at 37°C for 15 min and then purified with 5 µg DNA Cleanup Columns (NEB).

Gibson Assembly was then performed to connect the linear fragment amplified from pUC19 to cDNA inserts. Each Gibson Assembly reaction included 100 ng of 1.576 kbp linear pUC19 fragment, 42.07 ng of 186 bp cDNA insert, 10 µL 2X Gibson Assembly® Master Mix, and nuclease-free water to a total volume of 20 µL. A positive control was prepared with 10 µL of NEBuilder® Positive Control and a no-template control was prepared with nuclease-free water. The assemblies were incubated in a thermocycler at 50°C for 15 min and then purified with 5 µg DNA Cleanup Columns (NEB).

PCR Amplification of Extended cDNA Product. To prepare the extended cDNA products for nanopore sequencing, a 1.049 kbp section of the 1.762 kbp Gibson Assembly product was PCR-amplified. Each PCR was prepared with 2.97 ng of the Gibson Assembly product, 1.25 µL of 10 µM forward primer (sgRNA_cDNA_1kbp_ext_FWD_YG; Table 10, see Appendix B), 1.25 µL of 10 µM reverse primer (sgRNA_cDNA_1kbp_ext_REV_YG; Table 10), 12.5 µL of 2X Q5 Hot-Start High-Fidelity Master Mix (NEB), and nuclease-free water to a total volume of 25 µL. The thermocycler PCR program consisted of an initial denaturation at 98°C for 30 sec,

followed by 25 cycles of 98°C for 10 sec, 65°C for 30 sec, and 72°C for 25 sec, with a final extension at 72°C for 2 min. The PCR products were purified with 5 µg DNA Cleanup Columns (NEB).

3.4 Results and Discussion

3.4.1 Feasibility and Quality Assessment of sgRNA Library Synthesis

To establish a sgRNA library synthesis method independent of commercially available kits and protocols, we evaluated the feasibility of our custom workflow in two steps (Fig. 10A). First, we assessed IVT templates, produced by joining fragments containing 20-nucleotide (nt) sgRNA spacers with 83-nt scaffold fragments by Golden Gate Assembly (GGA), could reliably produce sgRNAs (Fig. 10A). This validation was crucial, as our T7 RNAP IVT protocol was developed by customizing existing methods, and its effectiveness required confirmation before scaling up to transcribe sgRNA libraries from microarray-derived subpools.

For our first proof-of-concept, we transcribed a pilot sgRNA library containing 18 unique spacer sequences. The 98-nt oligos were designed to contain a T7 promoter, 20-nt spacer, a type IIS restriction cut site (BsaI), and 15-nt forward and reverse primer sites (Fig. 10A). A single guanine was included at the 5' end of spacers, downstream of the T7 promoter. This addition was made when the first nucleotide position was adenine, cytosine, or thymine, as it is necessary for T7 RNAP transcription initiation (22, 23). The 18 oligos were ordered as individual ssDNA oligos from Integrated DNA Technologies. The oligos were manually pooled to obtain an 18-plex oligo pool and duplexed by performing reverse single primer extension (RSPE). These dsDNA fragments were joined to the sgRNA scaffold by Golden Gate Assembly, producing 135 bp templates. The 18-plex sgRNA library was transcribed alongside a singleplex IVT of sgRNAs

(sgRNA 1), also present in the 18-plex library as spacer 1. Successful synthesis of the 18-plex sgRNA library and sgRNA 1 was confirmed by 10% TBE-urea denaturing electrophoresis, as indicated by 100-nt bands corresponding to the size of our synthetic control sgRNA (Fig. 10B). We also noted the presence of 200-nt bands, indicating high molecular weight (HMW) dsRNA impurities from T7 RNAP RNA-templating activity (24). Overall, the 18-plex sgRNAs and sgRNA 1 show no differences between gel lanes, indicating successful synthesis of pooled and individual sgRNAs with our custom IVT protocol.

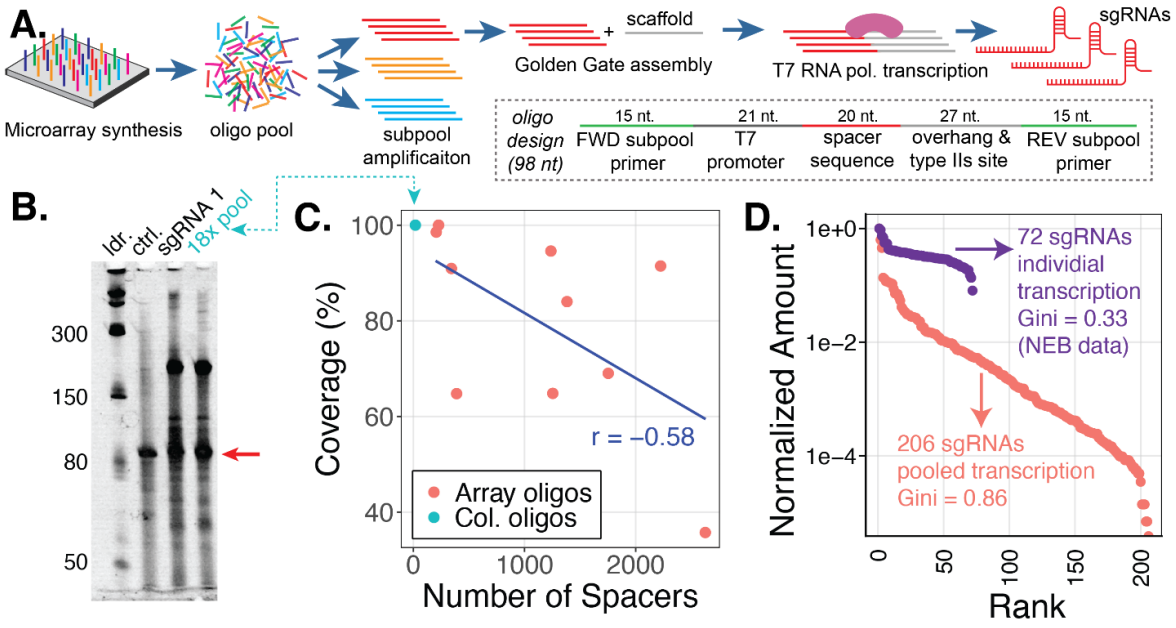


Figure 10. Experimental workflow and proof-of-concept for sgRNA library synthesis.

A. We PCR-amplified microarray-derived 98-nt oligo subpools, each containing spacers corresponding to individual sgRNA libraries. Each oligo was designed with forward and reverse PCR primer sites, a T7 promoter followed by 1-2 guanines, unique 20 nucleotide spacer sequences, and a BsaI type IIs restriction site. Golden Gate Assembly (GGA) was used to add the conserved sgRNA scaffold sequence to the oligos. The resulting GGA products were used as templates for pooled IVT of sgRNA libraries using T7 RNAP. **B.** To evaluate the initial performance of our sgRNA synthesis workflow, we manually pooled 18 column-synthesized oligos with 18 unique spacer sequences, converting them to dsDNA by extending the reverse primer, followed by GGA and IVT. Successful synthesis of pooled and individual sgRNAs (100 bases) was confirmed by 10% TBE-urea denaturing electrophoresis, followed by SYBR-Gold post-staining. **C**. Ten sgRNA libraries containing 206 to 2,626 spacers were synthesized from microarray-derived (array) oligos. The percent coverage of the microarray-derived libraries (orange dots) was compared to the percent coverage of the 18-plex sgRNA library (teal dot) with column derived oligos. We find an inverse correlation between coverage and scale (Pearson correlation analysis, r = -0.58). **D.** Comparison of spacer distribution uniformity between a microarray-derived 206-plex sgRNA library and a library of 72 independently transcribed sgRNAs from New England Biolabs (NEB). Normalized abundance is shown, with spacers ranked in descending order.

Next, we evaluated the feasibility of our complete sgRNA library synthesis workflow for generating libraries with hundreds to thousands of unique spacers. To reduce the cost of ordering microarray-derived oligos for large-scale production, we subpooled the oligo libraries (Fig. 10B). We obtained a single pool of ssDNA oligos containing 11,640 unique sgRNA spacers from Twist Bioscience. Each subpool primer pair was designed to amplify 206 to 2,626 spacer sequences corresponding to each sgRNA library.

We amplified each library using methods described in prior literature (25–27). Subpool amplification involved using quantitative PCR (qPCR) to determine the optimal number of cycles needed to reach the plateau phase while avoiding overamplification. For each qPCR and subsequent PCR amplification, we used 0.1 nanograms (ng) of diluted OLS pool. We then bulk-amplified each subpooled library using PCR, applying 0.01 ng of each subpool per reaction. We then used these amplified subpool fragments for the remainder of our sgRNA library synthesis workflow (Fig. 10A). To compensate for variable GGA product yields, we initially added a PCR and size-selection step following assembly to concentrate the transcription template for IVT. This additional step was performed before IVT to facilitate the production of large-scale sgRNA libraries and to re-transcribe the 18-plex sgRNA library for increased yield.

Following IVT, we evaluated the quality metrics of the sgRNA libraries to identify potential issues that could impact their downstream utility. We performed RNA-seq on all 10 microarray-derived sgRNA libraries and the re-transcribed 18-plex sgRNA library, based on a previous approach (28). Since the 5’ region of the sgRNAs contains the variable 20-nt spacer sequence, we employed 5’ Rapid Amplification of cDNA Ends (RACE) combined with template-switching to capture spacer diversity. The template-switched products were reverse-transcribed into cDNA, amplified, and nanopore sequenced, with a target of 1.5 million reads per

sgRNA library, though the actual depth varied. The custom primer also incorporated unique molecular identifiers (UMIs) during cDNA amplification to account for potential PCR bias, errors, and facilitate consensus calling.

The percentage of expected spacers present in each sgRNA library (percent coverage) was 100% for the pilot 18-plex sgRNA library prepared from column-derived oligos. In comparison, percent coverage for microarray-derived sgRNA libraries ranged from 100% for the 227-plex library to 36% for the 939-plex library, with an average of 80% across all 10 libraries (SD ± 20%). A moderate negative correlation (Pearson correlation: r = -0.5787; Table 3, see Appendix B) was observed between percent coverage and library scale for the microarray-derived libraries (Fig. 10C), indicating that coverage decreased as library scale increased. We also observed a strong positive correlation between coverage and nanopore sequencing depth (Pearson correlation: R^2 = 0.74, r = 0.87; Fig. 30A, Table 3; see Appendix B), suggesting that sequencing depth strongly influences observed coverage. This finding indicates a non-uniform distribution within the library, as greater sequencing depth increases the likelihood of detecting low-abundance spacers in the tail of the distribution. Since nearly all microarray-derived libraries exhibited spacer coverage below 100%, addressing this limitation was essential to generate well-represented sgRNA libraries (Fig. 10C).

We evaluated the Gini Coefficient to assess the uniformity of sgRNA library representation. The Gini Coefficient measures inequality in sequence representation, where a value of zero indicates perfect uniformity and a value of one reflects perfect inequality (27, 29). Lower Gini Coefficients indicate more uniform spacer representation, reducing biases that could affect the identification of targeted hits in CRISPR screens or other assays. The 10 microarray-derived libraries had an average Gini Coefficient of 0.90 (SD ± 0.06), with values ranging from

0.81 for the largest library (2,626-plex) to 0.97 for the 342-plex library (Fig. 30B, see Appendix B). We observed no correlation between Gini Coefficient values and library scale (Pearson correlation: $R^2 = 0.04$, $r = -0.29$; Fig. 30B, Table 3; see Appendix B). Instead, the consistently high Gini Coefficients observed across individual sgRNA libraries likely result from differences in sequence composition within each library, rather than being a systematic effect of library size.

Given the consistently high Gini Coefficient values observed, we sought to establish a baseline for comparison. We extracted publicly available yield data for 72 individually transcribed sgRNAs produced by New England Biolabs using the EnGen® sgRNA Synthesis Kit. The normalized yield distribution for these individually transcribed sgRNAs served as a proxy for the expected read abundance of individual spacers within sgRNA libraries and had a Gini Coefficient of 0.33 (Fig. 10D). The large disparity between the Gini Coefficient of individually transcribed sgRNAs (0.33) and the much higher values observed in our pooled libraries raises the question of whether this increased inequality is due to the multiplexed nature of transcription or other factors. Notably, the smallest microarray-derived library (206-plex) exhibited a 160% increase in the Gini Coefficient (0.86) compared to individually transcribed sgRNAs, suggesting that multiplex sgRNA synthesis introduces substantial biases in spacer representation (Fig. 10D).

To investigate this further, we examined the relationship between the representation of an individual spacer in a multiplexed reaction and its yield when transcribed independently. Specifically, we compared the representation of sgRNA 1 within the pilot 18-plex sgRNA library to the yields of both the entire library and the singleplex transcribed sgRNA 1 (Fig. 10B,C). The 18-plex library produced 1.7 micrograms of total sgRNA, whereas sgRNA 1 yielded 0.33 micrograms, a 5.2-fold reduction in yield in the non-multiplexed setting. RNA-seq analysis

revealed that sgRNA 1 accounted for only 0.0038 of the total reads in the 18-plex library, a 15-fold reduction from the expected 0.055 under perfect uniformity (Fig. 31, see Appendix B). The more pronounced 15-fold reduction in read abundance suggests that multiplex transcription amplifies biases in spacer distribution, likely due to competition among templates within the same reaction.

Next, we evaluated the percentage of spacers with expected (perfect) and mutant sequences. Mutations are primarily introduced during the oligo synthesis, PCR amplification (2.8 x 10^{-7} errors/nt), and T7 transcription (5×10^{-5} errors/nt) (30). The 18-plex pilot sgRNA library had a significantly lower median percent perfects (82%, SD ± 5%) compared to the microarray-derived sgRNA libraries, with median values ranging from 89% (SD ± 9%) to 95% (SD ± 3%) (Fig. 32, see Appendix B). A strong negative correlation (Pearson correlation: r = -0.90) was observed between median percent perfects and spacer coverage (Table 3, see Appendix B). A similar trend is also observed between median percent perfects and sequencing depth (Pearson correlation: R^2 = 0.58, r = -0.8; Fig. 30C, Table 3; see Appendix B). This correlation arises because spacers were included in the analysis only if they had at least 100 reads (including both perfects and mutants) in the RNA-seq data, ensuring sufficient representation. When spacer distributions are highly skewed, this threshold disproportionately includes overrepresented spacers, which tend to have a higher proportion of perfect sequences. PCR bias favors the perfect sequences which dominate, further enriching their representation. As a result, greater inequality in the distribution leads to a higher median percent perfects.

Consequently, the 18-plex sgRNA library, with 100% coverage, shows lower median percent perfects due to this dynamic (Fig. 10C). From these results, we concluded that our proof-of-concept sgRNA synthesis workflow could produce libraries with a high percentage of perfect

spacers (Fig. 32). However, the microarray-derived sgRNA libraries exhibited reduced coverage (Fig. 10C) due to severe spacer inequalities as indicated by the Gini Coefficient values close to one (Figs. 10D and S1b). These biases could lead to important functional hits being missed during CRISPR-Cas9 screens, affecting the accurate identification of functional gene targets. While oversampling during screening can improve statistical power (21), such biases may still compromise reproducibility and accuracy. Heo and colleagues (2024) demonstrated that reducing biases in the DNA used to generate sgRNA libraries decreased the number of cells required in a CRISPRi screen by 10- to 20-fold. These findings underscore the importance of improving library quality (21).

3.4.2 Reducing PCR Cycles in Microarray Oligo Amplification Fails to Improve sgRNA Library Uniformity

There are three potential sources of the observed spacer non-uniformities: oligo synthesis, PCR bias, or T7 RNA polymerase transcription bias. We first investigated how reducing the number of PCR cycles used to subpool and amplify microarray-derived libraries affected sgRNA representation, as observed through lower Gini Coefficients. PCR amplification is known to introduce biases into DNA libraries as cycle numbers increase, resulting in uneven sequence representation. These biases arise because DNA polymerase preferentially amplifies GC-rich priming motifs, leading to disproportionate enrichment of these sequences (31–34).

To test this hypothesis, we optimized the PCR cycle number for two sgRNA libraries: a large (1,382-plex) library and a smaller library (389-plex). PCR cycle optimization involved using qPCR to determine the ideal input template DNA concentration and the minimum number of PCR cycles needed for both subpool and bulk amplifications. Based on these results, we amplified the microarray-derived oligo libraries using 1 ng and 10 ng of template DNA—

representing 10-fold and 1000-fold increases, respectively, over our initial protocol. This adjustment reduced the number of PCR cycles from 53 to 22 for the 389-plex sgRNA library and from 54 to 25 for the 1,382-plex sgRNA library. We also removed the amplification step following GGA, as it doubled the number of PCR cycles in our workflow and increased the risk of introducing mutations into the spacers.

As a comparison, we also obtained 135-nt column-synthesized oligos (oPools) from Integrated DNA Technologies, which included both the spacer sequence and scaffold. These oPool libraries were duplexed using RSPE with a single PCR cycle, serving as controls to establish a baseline spacer distribution with minimal PCR-induced bias. By comparing the Gini Coefficients of the reduced-PCR microarray-derived sgRNA libraries to those of the oPool-derived libraries, we aimed to determine whether severe spacer inequalities arose from excessive PCR cycles or the IVT process itself.

After IVT of the sgRNA libraries, we performed RNA-seq and analyzed the reduced-PCR microarray-derived libraries and the oPool control libraries to evaluate spacer distribution and overall sgRNA quality (Fig. 11). First, we compared the percent coverage of expected spacers across the reduced-PCR (22, 24, and 25 cycles) and excessive-PCR (53 and 54 cycles) microarray-derived libraries, as well as the oPool libraries amplified with a single PCR cycle (Fig. 11A). The 389-plex oPool library exhibited 80% target coverage (n = 1), while the 1,382-plex oPool library, transcribed in duplicate, exhibited 80% and 85% coverage (n = 2). The reduced-PCR microarray-derived libraries showed higher coverage, with two independently prepared 389-plex libraries demonstrating 97% and 100% coverage, and the 1,382-plex library reaching 94% coverage (Fig. 11A). In contrast, the excessive-PCR microarray-derived libraries

exhibited substantially lower coverage, with only 65% for the 389-plex library and 84% for the 1,382-plex library (Figs. 10C and 11A).

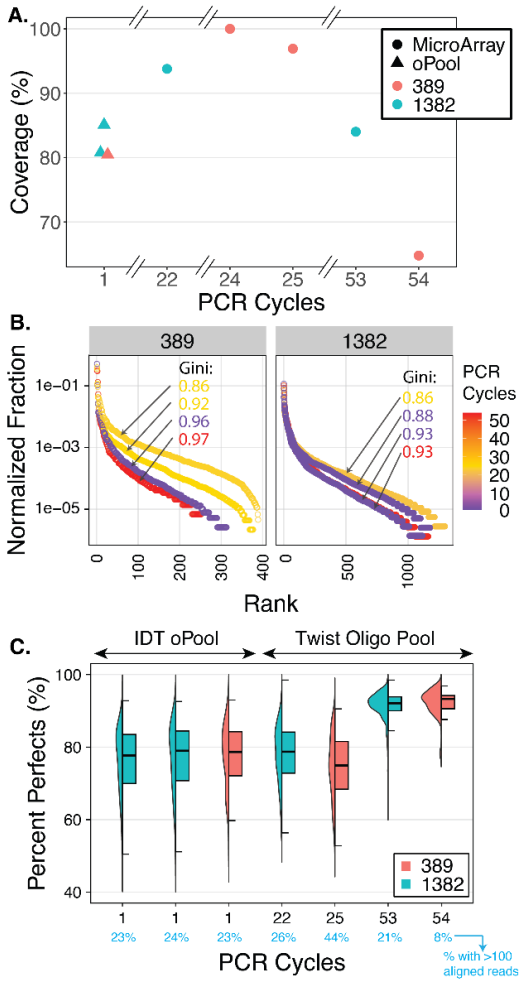


Figure 11. Comparison of sgRNA library metrics of two scales (389 and 1,382) with differing oligo sources and PCR cycles.

A. Percent coverage of spacers across 389-plex (n = 4, orange) and 1,382-plex sgRNA libraries (n = 4, teal), prepared using microarray-derived oligos (Twist oligo pool; circles) or column-synthesized oligos (IDT oPool; triangles), as a function of the number of PCR amplification cycles. The x-axis describing the number of PCR cycles is broken between 1–22, 22–24, and 25–53. B. Normalized abundance (reads per spacer, relative to total reads) for 389-plex (n = 4) and 1,382-plex (n = 4) sgRNA libraries. Spacers are ranked in descending order of abundance, with PCR cycle numbers represented by a heatmap gradient, with purple corresponding to 0 cycles and red to 50 cycles. Gini Coefficients (Gini), displayed alongside each library's rank-ordered curve, indicate the degree of inequality in spacer representation across different library scales and oligo sources. C. Percent perfects (proportion of perfectly synthesized spacer sequences) for 389-plex (orange) and 1,382-plex (teal) sgRNA libraries. Data for column-synthesized oligos (IDT oPool) and microarray-derived oligos (Twist oligo pool) are shown. Perfect spacers were only included if they had at least 100 reads in the RNA-seq data, ensuring sufficient representation for reliable analysis. The percentage of spacers with at least 100 reads is listed for each library in the x-axis (blue text). Each library is depicted using bifurcated plots: a half violin plot on the left (distribution of percent perfects) and a boxplot on the right (median percent perfect per library).

Initially, we expected oPool sgRNA libraries amplified with a single PCR cycle to achieve high coverage. We assumed that minimizing exponential PCR amplification would yield more uniform spacer representation, similar to microarray-derived libraries with low PCR cycles. However, the reduced-PCR microarray-derived libraries exhibited higher spacer coverage than the column-synthesized oPools (Fig. 11A). Despite this difference, spacer read counts were strongly correlated between the oPool and reduced-PCR microarray-derived 389- and 1,382-plex libraries (Pearson correlation: $R^2 = 0.77$, $r = 0.99$; $R^2 = 0.68$, $r = 0.86$; Fig. 33A,B see Appendix B). The strong correlation between spacer reads in two 1,382-plex oPool replicates (Pearson correlation: $R^2 = 0.89$, $r = 0.79$; Fig. 33E) supports the similarity in spacer distributions between oPool and reduced-PCR libraries. These results suggest that, despite differences in overall coverage, the relative spacer distributions remain highly similar, with variation in coverage likely arising from factors other than the oligos themselves.

We also expected the reduced-PCR libraries to achieve higher spacer coverage than the excessive-PCR libraries, and our results confirm this prediction. In the microarray-derived 389-plex sgRNA library, decreasing PCR cycles increased coverage by 32% and 35% (Fig. 11A). This increase in coverage is reflected in the moderate rather than strong correlation between spacer reads for the reduced- and excessive-PCR libraries (Pearson correlation: $R^2 = 0.44$, $r = 0.32$; Fig. 33C, see Appendix B). A similar trend was observed in the 1,382-plex library, where increasing PCR cycles resulted in a 10% decrease in coverage (Fig. 11A). Due to the smaller magnitude of this effect, the correlation between spacer reads for the reduced- and excessive-PCR libraries remained stronger (Pearson correlation: $R^2 = 0.72$, $r = 0.8$; Fig. 33D). These results demonstrate that reducing PCR cycles effectively improves spacer coverage across two

differently sized microarray-derived sgRNA libraries. We concluded that excessive PCR cycles reduce coverage due to inequality introduced by Kapa HiFi DNA polymerase bias (34).

Next, we evaluated the Gini Coefficients for the 389- and 1,382-plex sgRNA libraries, generated from DNA oligo libraries amplified with 1–54 PCR cycles. In the 1,382-plex libraries, the Gini Coefficient was 0.93 for the excessive-PCR library, 0.86 for the reduced-PCR library, and 0.93 and 0.88 for the column-derived oPool libraries (Fig. 11B). The 389-plex library followed a similar trend: the excessive-PCR library had a Gini Coefficient of 0.97, the reduced-PCR libraries yielded values of 0.92 and 0.86, and the oPool library showed a Gini of 0.96 (Fig. 11B). Overall, neither the reduced-PCR nor oPool libraries exhibited a consistent reduction in the Gini Coefficients compared to the excessive-PCR condition, suggesting that factors beyond PCR cycle number are the primary contributors to library uniformity.

In addition to percent coverage and Gini Coefficients, we evaluated the percentage of spacers with perfect sequences (percent perfects). We once again observed the trend of libraries with high inequality exhibiting higher percentage perfects. Low-PCR sgRNA libraries (1–25 PCR cycles) exhibited lower median percent perfects (mean: 78%, SD ± 1.7%, n = 5) compared to the 389- and 1,382-plex sgRNA libraries amplified with 53 and 54 PCR cycles, respectively (mean: 93%, SD ± 0.8%, n = 2; Fig. 11C). These observations suggest that DNA polymerase preferentially amplifies spacer sequences that are already overrepresented in the starting oligo libraries following excessive PCR cycles, while T7 RNA polymerase further enhances their abundance. These cumulative effects increase the representation of perfect spacers relative to less abundant mutant spacers or rare perfect spacers. While PCR, under ideal conditions, maintains the relative ratios of unique templates (in this case, spacer sequences) under ideal conditions, existing heterogeneity in the template pool may be exaggerated with increasing PCR

cycles. This occurs due to competition between templates for limited PCR reagents such as primers and dNTPs (35). Excessive PCR cycling disproportionately amplifies high-abundance spacers while causing low-abundance spacers, including mutants, to drop below sequencing depth. Consequently, perfect spacers are enriched while overall spacer diversity decreases.

Pearson correlation analysis of the excessive-PCR microarray-derived sgRNA libraries further supports this observation. As mentioned previously, there is a strong negative correlation between the median percentage of perfect spacers and sequencing depth across all 10 libraries (Pearson correlation: $R^2 = 0.58$, $r = -0.8$; Fig. 30C, see Appendix B). This indicates that perfect spacers are more abundant and more likely to be captured when sequencing depth is low, while more mutant spacers are included when sequencing depth is higher. This aligns with the strong positive correlation between spacer coverage and sequencing depth, showing that deeper sequencing captures both low-abundance perfect and mutant spacers (Peason's correlation: $R^2 = 0.74$, $r = -0.87$; Fig. 30A, Table 3; see Appendix B). Finally, our data directly support this prediction. In both the 389- and 1,382-plex sgRNA libraries, the fraction of reads for mutant spacers decreased with increasing PCR cycle number, while reads for target spacers increased (Fig. 34, see Appendix B). Based on these findings, we recommend minimizing PCR cycles during oligo preparation to improve spacer coverage, consistent with prior reports (35).

Having established the effect of PCR cycle number on percent coverage (Fig. 11A) and percent perfects (Fig. 11C), we turned our attention to improving the Gini Coefficients. Our sgRNA synthesis workflow contains two steps mediated by polymerases: PCR with DNA polymerase and IVT with T7 RNAP (Fig. 10A). In addition to DNA polymerase, other enzymes, such as DNase and Tn5 transposase, compromise molecular genomics workflows due to their sequence preferences (36–38). Given our results we reasoned that T7 RNAP was the primary

source of sequence bias in the sgRNA libraries, which would explain the persistence of spacer distribution bias following PCR cycle optimization (Fig. 11B). Since the sgRNA quality metrics were evaluated by reverse transcribing the sgRNAs into cDNA, it is possible that any reduction in the Gini Coefficient at the DNA level was masked by the subsequent T7 RNAP bias. This masking could occur because IVT is the final step in our workflow and would strongly influence the final Gini Coefficients obtained for each sgRNA library. On the other hand, improvements to the percent coverage of spacers were only due to the reduction in PCR cycles.

3.4.3 T7 RNA Polymerase Favors Spacers Starting with Four Guanines

To identify spacers preferentially transcribed by T7 RNAP and understand their impact on spacer distribution, we analyzed position-dependent biases in sgRNA library spacer sequences produced by our workflow. Specifically, we assessed the $\log_2$ fold change (FC) in the representation of each nucleotide at the first 10 positions of the 5′ end of all 20-nt spacers in each sgRNA library. We compared observed-to-expected ratios for each nucleotide at every position, visualizing the results as heatmaps for each library. The expected ratios were calculated based on the frequency of each nucleotide at each position in the designed spacer sequences, assuming a uniform distribution. Deviations from this baseline were captured by the $\log_2$ fold change metric, as determined by RNA-seq data.

We first applied this analysis to the 389- and 1,382-plex oPool, PCR-reduced, and excessive-PCR sgRNA libraries (excluding duplicates) shown in Fig. 11, as these displayed pronounced spacer biases even after PCR optimization. The heatmap of the 389-plex oPool sgRNA library showed a mean $\log_2$ fold guanine enrichment of 2.1 within the first four nucleotide positions (mean: 2.1, SD ± 0.26). Conversely, adenines (mean: -1.89, SD ± 0.56), cytosines (mean: -2.25, SD ± 1.40), and thymines (mean: -2.97, SD ± 0.60) exhibited a $\log_2$ fold

depletion of approximately 2.0 at these positions (Fig. 12A; Fig. 35A, see Appendix B). Similar biases appeared in the reduced-PCR 389-plex library (Fig. 35B, see Appendix B). However, in the excessive-PCR 389-plex library, guanine enrichment was limited to the first three positions (Fig. 35C), likely due to low spacer coverage (65%) caused by excessive PCR cycles (Fig. 11A). The 1,382-plex libraries showed similar but less pronounced bias trends (Fig. 35D-F), which may reflect library-to-library variation driven by unique spacer sequence compositions.

To investigate the guanine enrichment, we tested whether it was driven by the overall nucleotide composition of the first four positions or the presence of 5' guanine homopolymers. In the 389-plex oPool sgRNA library, spacers containing 5' homopolymers were enriched exclusively when containing 2- to 4-nt guanine homopolymer runs in the first four positions. In contrast, adenine, thymine, and cytosine 5' homopolymer containing spacers were never enriched. Notably, spacers containing 2-, 3-, and 4-nt guanine homopolymers, represented 6.6%, 17%, and 53% of the total reads after transcription (Fig. 12B). Strikingly, spacers with 5′ guanine tetramers (5′ GGGG) were enriched by 3.1-fold compared to those with guanine triplets (5′ GGG) (Fig. 12B). These findings suggest that T7 RNAP strongly prefers transcribing sequences with guanine tetramers at the 5′ end, corresponding to positions +1 to +4 of the T7 promoter.

A prior study identified 5′ guanine triplets (5′ GGG) as a key determinant of T7 RNAP transcription efficiency. Using a library of randomized T7 promoter variants with an initiating guanine at position +1, they found that sequences with guanines at positions +2 and +3 were transcribed robustly. By randomizing positions +4 to +8 while keeping guanines fixed at +1 to +3, they demonstrated that guanine triplets enhanced transcription efficiency. This effect was observed regardless of the downstream sequence, likely due to the role of guanines in stabilizing

the transcription bubble (39). However, the study did not assess the effect of guanines beyond position +3 or the impact of guanine tetramers on transcription efficiency.

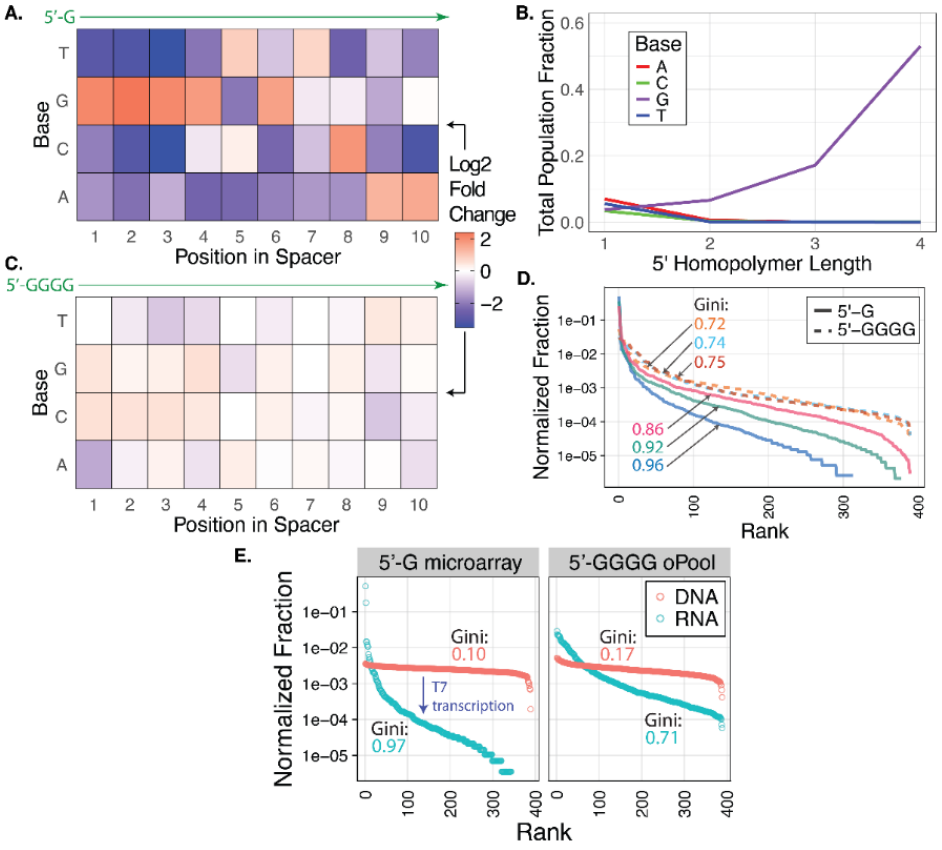


Figure 12. Influence of base composition on spacer abundance for 389-plex sgRNA libraries.

A. Log_2 fold change of observed vs. expected spacer abundance for each nucleotide at the first 10 positions of 20-nt spacers in a 389-plex sgRNA library transcribed from oPool oligos. Observed values are based on spacer frequencies from RNA-seq data, while expected values assume uniform spacer distribution. Positive log_2 fold changes (orange) indicate enrichment, negative values (blue) indicate depletion, and white indicates no change. B. Fraction of spacers containing homopolymer stretches within the first four nucleotides at the 5' end, analyzed from the same RNA-seq dataset as in A. C. Log_2 fold change of spacer abundance as in A, but for 389 spacers with a 5' guanine tetramer added at positions +1 to +4 downstream of the T7 promoter. Nucleotide positions 1-10 along the spacer are immediately downstream of the 5′ tetramer. The same color scale applies as in A. D. Normalized fraction of spacer reads relative to total reads for 389 spacers with a single guanine at the first spacer position (solid lines, n = 3) versus 389 spacers padded with 5′ guanine tetramers (dashed lines, n = 3). Spacers are ranked by descending abundance. Each library, transcribed *in vitro* from 400 ng of template DNA in a 20 µL reaction volume, is represented by a distinct color. Rank-order curves with listed Gini Coefficients (Gini) quantify the inequality in spacer representation across library types. E. Gini Coefficients for 135 bp DNA libraries (coral) containing 389 spacers and their corresponding sgRNA libraries (teal), *in vitro* transcribed by T7 RNAP using 100 ng of input DNA in a 100 µL reaction. DNA libraries were generated from microarray-derived oligos with spacers starting with a 5’ guanine (left) or oPool-derived oligos with spacers padded with a 5’ guanine tetramer (right). Spacer distributions for each library format are shown as in D.

In contrast, our study expanded the analysis to the first 10 nucleotides of sgRNA library spacers, providing a more comprehensive view of T7 RNAP's sequence preferences. The 2.7-fold increased representation of 5′ guanine tetramers compared to triplets in the sgRNA libraries suggests that the additional guanine at position +4 enhances transcription efficiency (Fig. 12B). Given that a 5′ guanine triplet stabilizes the transcription bubble and prevents abortive initiation, we propose that the fourth guanine reinforces this effect, facilitating the transition from transcriptional initiation to elongation. These findings extend the model proposed by Conrad and colleagues (2020) and provide new insights into T7 RNAP behavior during sgRNA synthesis.

Our data confirm that T7 RNAP introduces extreme spacer bias within sgRNA libraries. Consequently, we recommend RNA-sequencing all sgRNA libraries prior to their use in CRISPR-Cas9 studies. To our knowledge, no studies have directly assessed spacer distribution within sgRNA libraries before their use in CRISPR applications. Typically, DNA libraries used as transcription templates and downstream CRISPR outputs, such as gene enrichment or phenotypic changes, serve as proxies for evaluating sgRNA library quality (15, 40, 41). Incorporating RNA-seq for direct quality control could improve accuracy, identifying biases early and saving time and resources during CRISPR screen optimization.

A common strategy for improving CRISPR-Cas9 screen reliability is redundancy, using multiple sgRNAs per gene and transduction replicates to enhance signal robustness (7). This approach helps mitigate variability in spacer activity and off-target effects, while also compensating for experimental noise and the limitations of prediction tools. However, unrecognized biases in sgRNA library composition, like those introduced by T7 RNAP, could further distort representation and impact screen performance. Incorporating RNA-seq for direct quality control would allow researchers to identify underrepresented spacers early, ensuring

higher-quality libraries and potentially optimizing redundancy. Rather than using excessive sgRNA numbers to overcome unknown biases, well-characterized libraries improve redundancy, enhancing target coverage and streamlining experimental design and resource use. This approach would improve both the robustness and practicality of CRISPR screens and related assays.

3.4.4 Reducing sgRNA Spacer Bias by Adding 5′ Guanine Tetramers

Given the strong bias of T7 RNAP for spacers with guanines in the first four positions (Fig. 12A,B), we hypothesized that adding a 5′ guanine tetramer to all spacers would reduce this bias. This modification could stabilize the T7 RNAP initiation-to-elongation transition and improve transcription of spacers otherwise disfavored when preceded by a single guanine. However, this approach raised some considerations. The four guanine homopolymer, corresponding to the first four positions after the TSS of the T7 promoter, could introduce 5′ sequence heterogeneity in sgRNAs, potentially complicating the production of single-species RNA for structural studies like NMR or X-ray crystallography. Nonetheless, correctly sized species can be isolated through gel electrophoresis or size-selection methods (42). Additionally, guanine tetramer padding might interfere with sgRNA spacer formation of R-loop complexes after Cas9 binding, a critical step in CRISPR-Cas9 target recognition. However, functionality tests in a cell-free system demonstrated a 2-fold higher enrichment of 12 DNA targets using a 12-plex sgRNA library containing a 5’ guanine tetramer complexed with dCas9 compared to using a 12-plex sgRNA library containing a single 5’ guanine (43). Thus, we concluded that the benefits of improved spacer uniformity outweigh potential 5′ heterogeneity, with no observed deficiencies in CRISPR-dCas9 binding or target recognition.

To test our hypothesis, we redesigned the 389-plex spacer library by adding a guanine tetramer to the 5′ end of each spacer. To evaluate whether the effects of this modification were

dependent on scale, we grouped the 389 spacers into 12-, 60-, and 389-spacer subpools. As a control for the 12-plex 5′ GGGG subpool, we designed a second version of the 12-plex library by adding a single 5′ guanine (5′ G) only to spacers starting with A, C, or T. These were ordered as a single 473-oligo oPool from Integrated DNA Technologies. qPCR optimization showed that 10 ng of oPool input and no more than 12 PCR cycles yielded robust amplification across all subpools. Unlike microarray-derived oligos, which require separate amplification steps, a single subpool PCR generated sufficient DNA for GGA to append the sgRNA scaffold, producing 135-bp products for IVT.

Our first objective was to assess whether adding the 5′ guanine tetramer to spacers improves sgRNA library uniformity. To this end, we performed IVT with the GGA-produced DNA library containing 12 spacers padded with a 5′ GGGG sequence (library 12G4) or a single 5′ guanine (5′G, 12G). Resulting sgRNA libraries were reverse transcribed and nanopore sequenced using the 5′ RACE protocol. When comparing the spacer distributions between the two libraries, depicted as the normalized fraction of reads per spacer out of total reads, both showed similar trends. The Gini Coefficient values calculated from the spacer read distributions were 0.61 for 12G and 0.74 for 12G4 (Fig. 36A, see Appendix B), suggesting slightly increased inequality in the 12G4 library. To further evaluate whether this sequence modification influenced spacer biases, we conducted Pearson correlation analysis of spacer reads between the 12G and 12G4 libraries. A lack of correlation would indicate that the modification exaggerated or reduced biases. Instead, we observed a strong positive correlation ($R^2 = 0.76$, $r = 0.962$; Fig. 36B), indicating that the relative abundance of spacers remained consistent across both libraries.

From these data, we inferred that adding a 5′ guanine tetramer did not significantly affect spacer biases at small scales. These findings prompted us to investigate whether this trend holds

in larger sgRNA libraries containing 389 spacers padded with a 5′ guanine tetramer. To explore this, we performed IVT to generate an oPool-derived sgRNA library containing 389 spacers with a 5′ GGGG sequence (library 389G4). The resulting sgRNA library was sequenced as before and we determined the $\log_2$ FC of observed vs. expected spacer abundance for each nucleotide within the first 10 positions of the 389G4 spacers.

The heatmap of the 389G4 sgRNA library showed a mean $\log_2$ fold enrichment of 0.40 for guanines within the first four nucleotide positions (SD ± 0.15), reflecting a 5.3-fold decrease compared to the 389G library at these positions (mean $\log_2$ fold change: 2.11, SD ± 0.25). Additionally, the depletion of adenines, cytosines, and thymines at these positions was less pronounced than in the 389G library (Fig. 12A,C). Pearson correlation analysis between 389-spacer microarray-derived sgRNA libraries with a single 5′ G and oPool-derived libraries padded with a 5′ GGGG showed no correlation in read counts (R^2 = 0.04, r = 0.05; Fig. 37A, see Appendix B). A replicate of this comparison also showed no correlation (R^2 = 0.18, r = 0.14; Fig. 37C). Similarly, no correlation was observed between oPool-derived 5′ GGGG spacers and oPool-derived 5′ G spacers (R^2 = 0.02, r = 0; Fig. 37B). Analysis of multiple 389-plex sgRNA libraries showed lower Gini Coefficients for 5′ GGGG spacers (0.72, 0.74, 0.75; n = 3) compared to 5′ G spacers (0.86, 0.92, 0.96; n = 3). Overall, the 5′ GGGG libraries showed a mean 19.3% decrease in the Gini Coefficient (SD ± 2.81%) (Fig. 12D). We also evaluated the Gini Coefficients of 389-spacer sgRNA libraries transcribed with modified IVT conditions, using 100 ng input DNA in a scaled-up 100 µL reaction instead of the standard IVT setup, which used 400 ng input DNA in a 20 µL reaction. This adjustment produced a similar trend: a mean 22.7% decrease (SD ± 7.74%) in the Gini Coefficient for 5′ GGGG-padded spacers (n = 3) compared to spacers starting with a single guanine (n = 1) (Fig. 38, see Appendix B). Combined, these results

indicate that guanine tetramer padding substantially improved spacer representation, reducing inequality across the library while maintaining a more consistent and expected spacer distribution.

To determine whether the 5′ GGGG spacer modification affected overall sgRNA library yield, we analyzed the yields of multiple sgRNA libraries. A Wilcoxon rank-sum test revealed no significant difference in yield between libraries with 5′ GGGG spacers (n = 8, median: 0.805 μg) and those with 5′ G spacers (n = 13, median: 1.29 μg, p = 0.5002) (Fig. 39, see Appendix B). Although there was a slight trend toward higher yields in the 5′ G libraries, the difference was not statistically significant. These results demonstrate that the 5’ GGGG spacer modification does not affect the overall sgRNA yield.

3.4.5 Effect of Oligo Type on DNA Library Quality for IVT

We assessed whether oligo type influenced DNA library quality by sequencing the DNA templates and comparing microarray-derived oligos with 5′ G spacers and oPool-derived oligos with 5′ GGGG spacers. The percentage of perfect spacers was similar between the 389-plex microarray-derived library (median: 96.30%, n = 1) and the 389-plex oPool-derived library (median: 95.73%, n = 1). The distribution of percent-perfect spacers was also comparable between the two libraries, as shown in the violin plot in Fig. 40A (see Appendix B). These results contrast with earlier reports of higher error rates for microarray-derived oligos compared to column-derived oligos (44, 45). Notably, the microarray-derived 5′ G library underwent 24 PCR cycles, while the oPool-derived oligos were amplified with only 7 cycles. All PCR amplification occurred during the subpool and bulk amplification stages, with no additional PCR performed after GGA of oligos to the conserved sgRNA scaffold. Despite this higher cycle count, both libraries exhibited similarly high percentages of perfect spacers (Fig. 40A, see

Appendix B). Given the vendor-reported error rate for microarray oligo synthesis (one error per 3,000 nt), it is likely that microarray technology now produces oligos with error rates comparable to column-derived oligos. Furthermore, while Mighell and colleagues (2022) recommend using column-derived oligos due to their lower reported synthesis error rates, our results support the use of modern microarray-derived oligos as a high-quality and cost-effective alternative.

We also compared spacer inequality between libraries by analyzing the Gini Coefficients of normalized read fractions. The oPool-derived 389-plex 5′ GGGG DNA library showed a more skewed spacer distribution (Gini = 0.17, n = 1) compared to the microarray-derived 5′ G library (Gini = 0.10, n = 1), as determined by the Wilcoxon rank-sum test ($p = 3.97 \times 10^{-8}$; Fig. 40B). Furthermore, there was no correlation between the two libraries (Pearson correlation: $R^2 = 0$, $r = 0.04$; Fig. 40C). The observed inequalities could be due to the 5′ GGGG modification, DNA polymerase bias, or the oligo synthesis method. Yet, despite undergoing only 7 PCR cycles compared to the 24 PCR cycles for the 5'G library, the 5′ GGGG library exhibited greater skew, which could be caused by variable amplification efficiencies or PCR polymerase bias associated with the homopolymer region. Still, both libraries maintained high rates of spacer percent perfects rates and good overall quality, reinforcing that the primary source of inequality in sgRNA libraries arises from T7 RNAP sequence preferences (Fig. 12E).

We also examined the impact of oligo source (microarray vs. oPool) on the distribution of transcribed sgRNA libraries. Neither the 5′ GGGG nor 5′ G DNA library spacer reads correlated with those of their transcribed sgRNA libraries (Fig. 41A,B, see Appendix B). This indicates that T7 RNAP biases are present in both library formats and contribute to skewing. More specifically, the microarray-derived 5'G sgRNA library had a Gini Coefficient of 0.97 (n = 1), an 89.7% increase from the Gini of the 5'G DNA library that encodes it. In contrast, the

oPool-derived 5′ GGGG sgRNA library yielded a 13.6% lower Gini coefficient of 0.71 (n = 1), reflecting a 76.1% increase from its corresponding DNA library (Fig. 12D).

Notably, this difference stems from T7 RNAP sequence preferences, as the difference in Gini coefficients between the 5' G and 5' GGGG sgRNA libraries is 3.7-fold larger than the difference observed between their DNA libraries (Fig. 12D). These findings suggest that the type of oligo used to construct the DNA library had less influence on spacer distribution than the 5′ guanine modification's impact on the resulting sgRNA libraries. Together, these data demonstrate that padding the 389 spacers with a 5′ guanine tetramer effectively reduced T7 RNAP transcription biases, though this effect was not observed at the smaller, 12-spacer scale.

Both microarray-derived and oPool-derived oligos demonstrated excellent quality, with fewer than 5% mutant spacers and minimal spacer inequality. Though the oPool-derived library with 5′ GGGG spacers showed slightly greater spacer distribution skew than the 5′ G microarray-derived library, the difference was small but statistically significant. These inequalities were decoupled from the inequalities that we have observed after transcription due to the behavior of T7 RNAP. While we initially used oPool-derived oligos to quickly optimize libraries and IVT conditions with small pools, our experimental design offers a more cost-effective strategy.

3.4.6 Emulsion IVT (eIVT) Enhances sgRNA Library Uniformity

Designing sgRNA spacers with a 5′ GGGG sequence substantially improved sgRNA library uniformity. However, this approach still has substantially more inequality than the Gini (0.33) observed for the 72 sgRNAs individually transcribed by NEB (Fig. 10D). We hypothesized that modifying the IVT process itself could further improve library uniformity, either independently or in combination with the 5′ GGGG spacer design. To this end, we developed an emulsion-based IVT (eIVT) workflow that isolates sgRNA template DNA

molecules within droplet microcompartments, reducing competition for IVT reagents and T7 RNAP. This strategy draws from the successful use of oil-water emulsions in DropSynth gene synthesis, where compartmentalization prevents cross-hybridization of DNA oligos and reduces competition during both DNA assembly and amplification (25, 27). A similar approach is used in whole-genome amplification (WGA) for single-cell sequencing, where emulsions minimize competition among genome fragments during amplification, reducing bias (46) . Given these effects, we predicted that eIVT would yield sgRNA libraries with lower Gini Coefficients compared to bulk IVT. To accommodate our existing emulsification protocol, we increased IVT reagent concentrations fivefold while varying DNA input, scaling a standard 20 μL IVT reaction up to 100 μL, which is the aqueous phase volume used for emulsification.

The aqueous phase was combined with commercial fluorinated oil and emulsified by vortexing (25). To assess whether compartmentalization in emulsions slowed transcription, we prepared two 12-plex sgRNA libraries by incubating eIVT reactions at 37°C for either 2 hours, which is the standard IVT time, or 16 hours. The 2-hour incubation condition showed a 2-fold increase in yield compared to the 16-hour incubation condition [data not shown], confirming that the transcription process was complete within 2 hours within emulsions. Additionally, IVT incubation time did not affect spacer distribution for both conditions (Fig. 42A, see Appendix B), with nearly identical normalized read fractions and a strong positive one-to-one linear relationship (Pearson correlation: $R^2 = 0.99$, $r = 0.99$; Fig. 42B). Based on these results, all subsequent eIVT experiments used a 2-hour incubation at 37°C.

After incubation, we broke the emulsions with chloroform, treated the aqueous phase with DNase, and purified the sgRNAs via column purification (Fig. 13A). Since emulsions can function as compartmentalized bioreactors on a microscale, we evaluated whether eIVT affects

the total yield of sgRNA libraries compared to bulk IVT. We also assessed how library scale and spacer sequence modifications (5’ G vs. 5’ GGGG) influence IVT performance under both conditions. Bulk IVT controls were prepared identically to the aqueous phase of each eIVT, using the same DNA input and 100 μL reaction volume. By maintaining equivalent volumes, we controlled for any potential effect of reaction volume on sgRNA yield and spacer uniformity. A Wilcoxon rank-sum test showed no significant difference ($p > 0.05$) in sgRNA yields between libraries transcribed with 100 ng of input DNA using bulk IVT (median: 1.8 μg, SD = ± 0.98, n = 14) and those transcribed with eIVT (median: 1.8 μg, SD = ± 0.98 n = 13; Fig. 13A). These results indicate that eIVT is a viable method for sgRNA transcription, as it maintains comparable yields to bulk IVT.

Previously, we observed that adding a 5’ GGGG sequence to a 389-plex sgRNA library reduced the Gini Coefficient by a mean of 19.3%, improving uniformity compared to the equivalent library with 1G spacers (Fig. 12D). We hypothesized that eIVT of a library DNA template with the same 5’ GGGG modification would further enhance spacer uniformity, lowering Gini Coefficient values even more. To test this, we transcribed 60- and 389-plex sgRNA libraries with 5’ GGGG spacers (60G4 and 389G4) using eIVT and bulk IVT with either 100 ng or 800 ng of input DNA. These DNA amounts were chosen to assess the impact of DNA input on eIVT, given that DNA partitioning within emulsion droplets follows a Poisson distribution. To maximize compartmentalization efficiency, reducing this mean toward one molecule per droplet is ideal, but too little DNA may compromise sgRNA library yield for downstream applications. Therefore, we compared 800 ng (Poisson = 3,539) and 100 ng (Poisson = 442) DNA inputs. At these high DNA inputs, multiple molecules will co-localize

within droplets. For instance, in emulsions with a 5 μm mode droplet diameter (25), 500 ng of 135 bp DNA yields a Poisson mean of 942 DNA molecules per droplet.

Following RNA-seq analysis of transcribed sgRNA libraries, the 60G4 sgRNA library had median Gini Coefficients of 0.69 (SD ± 0.0068, n = 2) for 100 ng bulk IVTs, 0.70 (SD ± 0.036, n = 3) for 100 ng eIVTs, and 0.70 (n = 1) for the 800 ng eIVT. The 389G4 sgRNA library had median Gini Coefficients of 0.71 (SD ± 0.073, n = 3) for 100 ng bulk IVTs, 0.75 (SD ± 0.017, n = 3) for 100 ng eIVTs, and 0.80 (n = 1) for the 800 ng eIVT. No bulk IVT controls were included for the 800-ng input condition. Across all comparisons, eIVT did not reduce median Gini Coefficients relative to bulk IVT sgRNA libraries (Fig. 13C, upper panel). These findings suggest that a Poisson distribution of 442 DNA molecules per droplet (from 100 ng input DNA) is too high for effective template partitioning, potentially allowing intermolecular competition for reagents.

To further investigate the effects of Poisson loading in emulsions on library uniformity, we tested input DNA amounts of 100 ng (Poisson 442), 10 ng (Poisson 44), and 1 ng (Poisson 4) while performing eIVT of 389- and 2,626-plex 5’ G sgRNA libraries. The 2,626-plex library was included to assess whether eIVT offers greater benefits at larger library scales, where competition among spacers for reagents is expected to be more pronounced. No sgRNA libraries transcribed with 1 ng of input DNA, nor the two bulk IVTs with 10 ng of input DNA, yielded detectable RNA by fluorometer quantification. However, we performed 5’ RACE-RT and successfully sequenced all sgRNA libraries using nanopore sequencing, regardless of yield. Each sgRNA library condition was transcribed once (n = 1). To assess statistical significance, we applied the Kolmogorov–Smirnov (KS) test to determine whether the spacer distributions of eIVT and bulk IVT sgRNA libraries were drawn from the same population at a given scale.

While the KS test evaluates whether observed differences arise by chance or reflect a real effect, the Gini Coefficient indicates the directionality of changes in distribution uniformity.

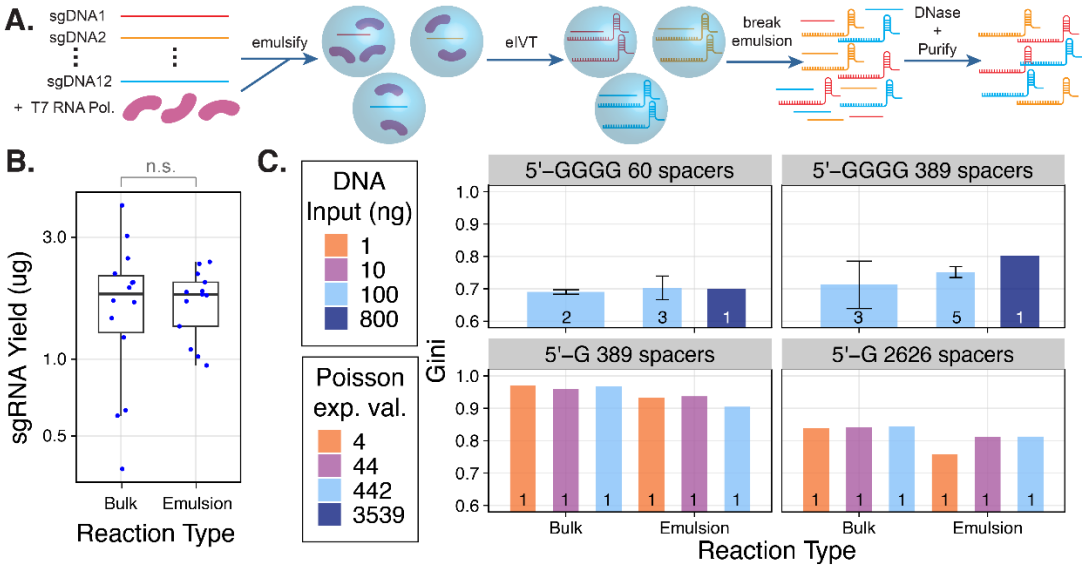


Figure 13. Evaluating *in vitro* transcription in emulsions (eIVT) as a novel approach to synthesize sgRNA libraries.

A. A 100 µL IVT reaction was prepared by scaling up a standard 20 µL IVT reaction fivefold and adding Golden Gate-assembled library DNA templates to the aqueous phase. The aqueous phase was emulsified in a fluorinated oil by vortexing, generating stable droplets that function as microreactors for IVT. After transcription, emulsions were broken, and sgRNA library products were purified via DNase treatment and column purification. B. Comparison of sgRNA yields for various sgRNA library scales transcribed with 100 ng of input DNA using bulk IVT (n = 14) or eIVT (n = 13). sgRNA libraries generated by bulk IVT serve as controls for each sgRNA library produced by eIVT. Yields (µg) are further categorized by library scale: 12-plex (n = 5), 18-plex (n = 1), 60-plex (n = 5), 389-plex (n = 12), and 2,626-plex (n = 3). A Wilcoxon rank-sum test found no significant difference in sgRNA yields between bulk IVT and eIVT ($p > 0.05$). C. Libraries contained spacers starting with either a single 5' guanine (5' G) or a 5' guanine tetramer (5' GGGG). The top panel displays the median Gini coefficients for 5' GGGG sgRNA libraries synthesized using bulk eIVT versus bulk IVT with 800 ng (navy, Poisson = 3,539) or 100 ng (aqua, Poisson = 442) of input DNA. The Poisson value represents the estimated mean number of input DNA molecules per emulsion droplet. Individual bars represent each sgRNA library type, with sample sizes labeled. Bar heights indicate the median Gini coefficient per category, and black bars represent standard deviation (SD). The bottom panel shows median Gini coefficients for 5' G sgRNA libraries synthesized using bulk eIVT versus bulk IVT with 100 ng (aqua, Poisson = 442), 10 ng (purple, Poisson = 44), or 1 ng (orange, Poisson = 4) of input DNA. Sample size is one for all sgRNA libraries in the bottom panel.

The 389-plex sgRNA library produced with eIVT and 1 ng of input DNA exhibited a significantly lower Gini Coefficient (0.93) compared to the bulk IVT control (0.97) (Fig. 13C). This difference was statistically significant according to the KS test ($D = 0.20352$, $p = 3.53 \times 10^{-7}$; Table 4, see Appendix B). To assess whether position-dependent biases of T7 RNAP persisted following eIVT, we evaluated the $\log_2$ FC for each nucleotide within the first 10

positions at the 5′ end of all 20-nt spacers. The eIVT library showed a global reduction in overall nucleotide bias compared to the bulk library, with FC values closer to the expected baseline of zero (Fig. 43A,B, see in Appendix B). This indicates that the strong bias toward guanine within the first four positions was reduced in the eIVT library relative to the bulk IVT control.

This bias reduction was even more pronounced in the 2,626-plex library, where the Gini coefficient decreased from 0.84 (bulk IVT) to 0.76 (eIVT) (Fig. 13C). The KS test confirmed the statistical significance of this result ($D = 0.22022$, $p = 2.20 \times 10^{-16}$; Table 4, see Appendix B). Additionally, the $\log_2$ FC heatmap showed a global decrease in nucleotide bias, with a distinct pattern of enrichment for guanine at position 2 and cytosine at position 5 (Fig. 44A,B see Appendix B). This pattern likely reflects library-to-library variation driven by unique spacer sequence compositions, as previously observed in the 1,382-plex library (Fig. 35D-F, see Appendix B).

With 10 ng of input DNA, the 389-plex eIVT library did not exhibit a lower Gini Coefficient (0.94) compared to the bulk IVT control (0.96) (Fig. 13C), as indicated by the lack of statistical significance (KS test: $D = 0.0977$, $p = 0.054$; Table 4, see Appendix B). However, the FC heatmap showed a global reduction in bias (Fig. 43C,D see Appendix B). In contrast, the 2,626-plex eIVT library exhibited a significantly lower Gini Coefficient (0.81) than the bulk IVT control (0.84) (Fig. 13C). The KS test confirmed significance ($D = 0.0984$, $p = 7.35 \times 10^{-10}$; Table 4), accompanied by a consistent reduction in positional bias (Fig. 44C,D see Appendix B).

With 100 ng of input DNA, the 389-plex eIVT library exhibited a lower Gini Coefficient (0.90) than the bulk IVT control (0.97) (Fig. 13C). The KS test confirmed statistical significance ($D = 0.3626$, $p = 2.20 \times 10^{-16}$; Table 4), and the FC heatmap indicated a global reduction in nucleotide bias (Fig. 43E,F see Appendix B). Similarly, the 2,626-plex eIVT library exhibited a

significantly lower Gini Coefficient (0.81) than the bulk IVT control (0.84) (Fig. 13C), with the KS test confirming significance ($D = 0.0811$, $p = 3.19 \times 10^{-7}$; Table 4). The FC heatmap reflected decreased guanine overrepresentation and a broader recovery of A, T, and C levels toward the expected $\log_2$ FC values of zero. (Fig. 44E,F see Appendix B).

Together, these results show that for both the 389-plex and 2,626-plex libraries, the eIVT method consistently reduced Gini Coefficients and improved uniformity at 1 ng and 100 ng DNA input amounts. At 10 ng input, the 389-plex library showed a decrease in Gini Coefficient under eIVT, though not statistically significant, while the 2,626-plex library showed a significant reduction. These findings suggest a possible scale-dependent effect.

We next examined how spacer representation differs between eIVT and bulk IVT. To assess whether varying the DNA input reduces guanine-driven bias at the 5' end of spacers, we analyzed guanine representation at the first five positions. The percent decrease in guanine representation served as a simple proxy for bias in spacer distribution. For each sgRNA library transcribed with 1 ng, 10 ng, or 100 ng input DNA, this value was calculated as the change in $\log_2$ FC between bulk IVT and eIVT, normalized by the bulk IVT $\log_2$ FC. This approach was based on our previous findings, which showed that guanines within the first four positions drive overrepresentation (Fig. 12).

For the 389-plex library, the percent decrease in guanine representation at spacer positions 1 to 4 remains at or below 25% across all DNA input amounts. At position 5, guanine representation decreases by 48.2%, 23.6%, and 47.6% for the 1 ng, 10 ng, and 100 ng DNA input conditions, respectively. Notably, the 1 ng and 100 ng conditions roughly twice the reduction seen at 10-ng (23.6%) (Fig. 45A, see Appendix B), consistent with the lack of a significant Gini Coefficient decrease at the intermediate input (Fig. 13C; Table 4, see Appendix B). For the

2,626-plex sgRNA libraries, the percent decrease in guanine representation at positions 1 to 3 remained below 86.3% across all DNA input amounts. At position 4, reductions were more pronounced: 113% for the 100-ng input condition, 166% for the 10-ng input, and 253% for the 1 ng input. In contrast, at position 5, guanine representation increased slightly for the 1 ng (25.8%) and 10 ng (19.8%) inputs, while the 100-ng input showed a modest 16.7%, decrease (Fig. 46A, see Appendix B).

These results show that eIVT effectively reduces the overrepresentation of guanine-containing spacers at nucleotide positions previously identified as biased toward guanine. In the 389-plex library, the greatest relative reduction in bias occurred at position 5, immediately following positions 1 to 4 (Fig. 45A, see Appendix B). In contrast, the 2,626-plex library exhibited a stronger bias toward guanine-containing spacers at positions 2 and 4, with the most substantial reduction at position 4 (Fig. 46A, see Appendix B). Across all DNA input amounts, percent decreases in guanine bias were consistently lower for the 389-plex library compared to the 2,626-plex library (Figs. 45A and 46A). For the 389-plex library, the 1 ng and 100 ng eIVT conditions produced the most notable reduction in bias, while in the 2,626-plex library, the largest decrease occurred at 1 ng, supporting the prediction that lower DNA input enhances effective compartmentalization. As expected, the 10 ng and 100 ng conditions showed more modest reductions (Fig. 46A). The smaller scale of the 389-plex library may have contributed to variability in guanine bias reduction due to stochastic effects. Finally, this analysis focuses exclusively on changes in guanine bias and does not capture broader nucleotide composition bias.

Given the narrow scope of this analysis, we evaluated overall changes in spacer representation for bulk IVT and eIVT sgRNA libraries using Pearson correlation analysis.

Correlation linear fits were compared to a unity line, which represents a perfect 1:1 relationship between spacer read fractions in each condition. Steeper linear fit slopes indicate greater reductions in bias using emulsion compartmentalization. Red dashed lines denote the median read fraction per spacer from the input DNA library transcription template, reflecting a highly uniform spacer distribution (Gini = 0.102; Fig. 40B, see Appendix B). Spacers that decreased in representation due to eIVT appear as blue dots above the unity line, while those that increased appear as red dots.

The 389-plex sgRNA libraries exhibited strong positive correlations across 1 ng (R^2 = 0.9, r = 0.95), 10 ng (R^2 = 0.85, r = 0.99), and 100 ng (R^2 = 0.79, r = 0.98) DNA input amounts (Fig. 45B-D, see Appendix B). With eIVT, overabundant spacers decreased in representation for the 1 ng and 10 ng input conditions compared to bulk IVT (Fig. 45B,C). In contrast, the 100-ng condition showed an increase in the representation of low-abundance spacers (Fig. 45D). Notably, the two most overrepresented spacers decreased in the 10 ng and 100 ng input conditions, whereas only one of these spacers showed a reduction in the 1 ng condition (Fig. 45B-D). Since these changes are plotted on a $\log_{10}$ scale, reductions in the most overrepresented spacers are substantial. The 2,626-plex sgRNA libraries also showed positive correlations across 1 ng (R^2 = 0.66, r = 0.61), 10 ng (R^2 = 0.76, r = 0.91), and 100 ng (R^2 = 0.76, r = 0.92) DNA input conditions. However, differences between individual input amounts were less pronounced. The most overrepresented spacer decreased in the 1 ng and 10 ng conditions but remained largely unchanged in the 100-ng condition (Fig. 46B-D, see Appendix B).

Overall, these findings suggest that reducing the most overrepresented spacers has the greatest impact on uniformity metrics. The more pronounced effects observed in the smaller 389-plex library likely reflect stochastic variation, whereas the larger 2,626-plex library exhibits more

consistent behavior across conditions. Paradoxically, the pronounced stochastic effects in the smaller 389-plex library offer valuable insights into how DNA input amount influences transcription efficiency in emulsions. Across both libraries, the lower Gini Coefficients and greater reduction in guanine overrepresentation support our hypothesis that the 1 ng DNA input condition (Poisson 4) enhances compartmentalization and uniformity. However, despite these promising proof-of-concept results, the 1 ng condition did not yield a quantifiable sgRNA output, limiting its practical utility for library generation. In contrast, the 100-ng input condition (Poisson 442) produced the most substantial reduction in guanine bias and a significantly lower Gini Coefficient for the 389-plex library. This suggests that higher DNA template concentrations enhance T7 RNAP transcription efficiency and improve uniformity by increasing substrate availability. The 10-ng input condition (Poisson 44) showed moderate effectiveness in reducing T7 RNAP bias in both libraries but did not significantly lower the Gini Coefficient for the 389-plex library. This may indicate that the 10-ng input condition fails to strike a balance between sufficient DNA concentration for transcription efficiency and the compartmentalization benefits observed at lower inputs.

Given the limited sample size and the application of the KS test across all conditions, our ability to fully resolve the relationship between DNA input and Gini Coefficients may be restricted. Nonetheless, these findings show that eIVT significantly reduces bias in the spacer distributions of sgRNA libraries with 389-plex or greater complexity. Further optimization will be required to achieve comparable or greater bias reduction for 5' GGGG libraries relative to bulk IVT protocols.

3.4.7 Evaluation of High Molecular Weight Byproducts in sgRNA Libraries

We investigated the presence of double-stranded RNA (dsRNA) byproducts in sgRNA libraries to better understand why eIVT improves spacer uniformity in libraries with 5' G spacers but not in those modified with a 5' GGGG sequence. We hypothesized that the 5' GGGG motif may interfere with the uniformity gains observed in eIVT for libraries containing 5' G spacers. To explore this, we performed polyacrylamide gel electrophoresis of 389-plex sgRNA libraries transcribed under various DNA input conditions using either eIVT or bulk IVT (Fig. 47A, see Appendix B). Only libraries with quantifiable yields were analyzed. The gel revealed that all sgRNA libraries contained the expected 100-nt sgRNA products, confirmed by band alignment with a synthetic sgRNA control from IDT (Fig. 47A). However, most libraries also showed a prominent high molecular weight (HMW) 200-nt dsRNA byproduct, likely generated by the residual RNA-dependent activity of T7 RNAP (47). We did not analyze these byproducts beyond this point since we cannot capture information about these dsRNA byproducts in our RNA-seq data. The residual RNA-dependent activity of T7 RNAP causes cis-primed extension from small hairpin loops within transcribed sgRNAs, producing looped-back dsRNA products (48, 49). These dsRNA impurities are known to trigger innate immune responses when IVT-produced RNA is used in therapeutic applications (50–52). As a result, they are often removed by high-performance liquid chromatography, cellulose-based chromatography, or native purification methods (53–55). The formation of these byproducts is associated with high-yield conditions during T7 RNAP IVT batch synthesis (28, 49), so their presence in both eIVT and IVT libraries is unsurprising.

Unexpectedly, the eIVT condition prepared with 10 ng of input DNA produced minimal dsRNA byproducts, as observed in the gel image (Fig. 47A). ImageJ (56) quantification of band intensities revealed that this condition showed almost no detectable dsRNA signal, with levels

even lower than the synthetic sgRNA control, while still producing clear sgRNA products. In contrast, the corresponding IVT reaction with 10 ng of input DNA displayed low band intensities for both sgRNA and dsRNA products (Fig. 47B, see Appendix B). Analysis of relative dsRNA band intensity confirmed that most of the total RNA signal in the eIVT condition corresponded to sgRNA products, whereas the IVT condition showed a higher proportion of dsRNA byproducts (Fig. 47C). These results suggest that emulsions combined with a low DNA input of 10 ng may help prevent the formation of dsRNA byproducts by promoting low-yield rather than high-yield conditions. However, we could not directly compare this observation to band intensities from eIVT prepared with 1 ng of input DNA due to insufficient yield for analysis by gel electrophoresis. These promising results warrant further replication to confirm whether eIVT with 10 ng or less input DNA enhances sgRNA production while minimizing dsRNA byproducts. If validated, this approach could offer a valuable strategy for optimizing sgRNA synthesis conditions.

In addition to the 200-nt dsRNA byproducts, the gel also revealed that all sgRNA libraries containing spacers starting with 5' GGGG had multiple HMW RNA bands longer than 300 nt (Fig. 47A). Padding spacers with a 5' GGGG sequence offers advantages, such as improved spacer uniformity compared to 5' G sgRNA libraries when both are transcribed in bulk IVT (Figs. 12D and 13D). However, this approach also produces additional HMW RNA species, which are a clear drawback. Although we do not observe a statistically significant difference in overall RNA yield between 5' G and 5' GGGG sgRNA libraries (Fig. 39, see Appendix B), a higher proportion of the quantified RNA from the 5' GGGG libraries consists of non-functional HMW RNAs rather than functional sgRNAs. These excessive HMW byproducts appear regardless of whether the libraries were transcribed using eIVT or standard IVT (Fig. 47A).

Furthermore, since we first began transcribing them, we have consistently observed greater levels of HMW byproducts in 5' GGGG libraries compared to 5' G libraries (Fig. 47D, see Appendix B). Thus, when the 5’ guanine tetramer is added, the functional sgRNA species are a smaller relative fraction of the total RNA population.

Despite the presence of these HMW products, the 5' GGGG libraries remain suitable for *in vitro* CRISPR applications (43). Moreover, these byproducts can be reduced or eliminated through purification methods and the use of engineered T7 RNAP variants (52, 57), making this approach likewise potentially viable for producing high-quality messenger RNA libraries for mammalian cells or *in vivo* applications.

3.4.8 Future Directions for Optimizing eIVT

Comparing the NEB dataset on individually transcribed sgRNAs to our eIVT results (Fig. 13C) suggests substantial room for further optimization within our compartmentalization approach. In our current method, we were only able to reduce the Poisson expectation value to 4, indicating that multiple spacers were still co-transcribed within individual droplet compartments. One potential strategy to overcome this limitation is using barcoded beads to isolate spacers into individual microreactions, similar to the DropSynth gene synthesis method (25). In this approach, beads decorated with unique barcodes corresponding to short barcodes on the 5′ end of DNA oligos could selectively capture DNA templates encoding single spacers. By scaling the number of beads with the library size, this method could improve spacer representation and reduce bias. For example, 1,536 unique beads could transcribe a 1,536-plex sgRNA library. Concentrating individual spacer sequences within droplets would likely enhance library uniformity and scalability, making this a promising direction for future development. Moreover, using barcoded beads ensures a high local concentration of template DNA in each droplet, which

should avoid the reduced yields observed when Poisson loading was minimized without bead capture.

Another potential improvement involves reducing droplet size without decreasing input DNA amounts. Our current eIVT emulsions, generated by vortexing, produce droplets with a typical diameter of 5 μm, which is relatively large compared to the DNA template input. Reducing the diameter to 700 nm while maintaining 100 ng of input DNA would yield a Poisson expectation value of 0.76, substantially decreasing the probability of multiple spacers co-localizing within a single droplet. This reduction in droplet size could improve spacer uniformity while preserving sufficient sgRNA yield for CRISPR applications. Although uniform droplets of this size can be produced using microfluidic methods (58, 59), they require specialized chips and equipment, increasing operational complexity. Similarly, generating very small vesicles to reduce the Poisson loading is technically challenging and may offer limited additional benefits. In contrast, a bead-based barcoding approach would be relatively straightforward to implement, making it the most practical option for optimizing eIVT.

3.4.9 Effects of IVT Reaction Conditions on sgRNA Library Uniformity

Due to optimizations made to IVT reactions for compatibility with an emulsions-based transcription approach, we investigated how varying DNA input and total IVT reaction volume affect sgRNA library uniformity in the absence of emulsions. To do this, we transcribed sgRNA libraries using 100 μL and 20 μL reaction volumes with varying DNA input amounts. The 100 μL reaction was included as it matched the aqueous phase conditions used in eIVT. Following RNA-seq, we assessed the Gini Coefficients of sgRNA libraries across different scales (12-, 60-, 389-, and 2,626-plex) using DNA input amounts ranging from 1 to 400 ng and spacers beginning with either a 5’ G (1G) or 5’ GGGG (4G) (Fig. 14). Given the limited sample size, with most

libraries represented by a single sample, we applied the KS test to determine whether spacer distributions deviated significantly from the input DNA population.

For the 12- and 60-plex libraries, Gini Coefficients remained consistent across different reaction volumes and DNA input amounts. In the 12-plex libraries, the median Gini Coefficient for 4G spacers was 0.59 (SD ± 0.029, n = 2) for a 100 µL reaction with 100 ng DNA, compared to 0.62 (SD ± 0.20, n = 3) in 20 µL reactions with 400 ng DNA. Similarly, the 60-plex libraries maintained Gini Coefficients between 0.69 to 0.70 across all tested conditions. Variability was more pronounced in libraries transcribed with 400 ng DNA in 20 µL reactions (Fig. 14), though this was less evident in larger libraries, likely due to limited replication conditions.

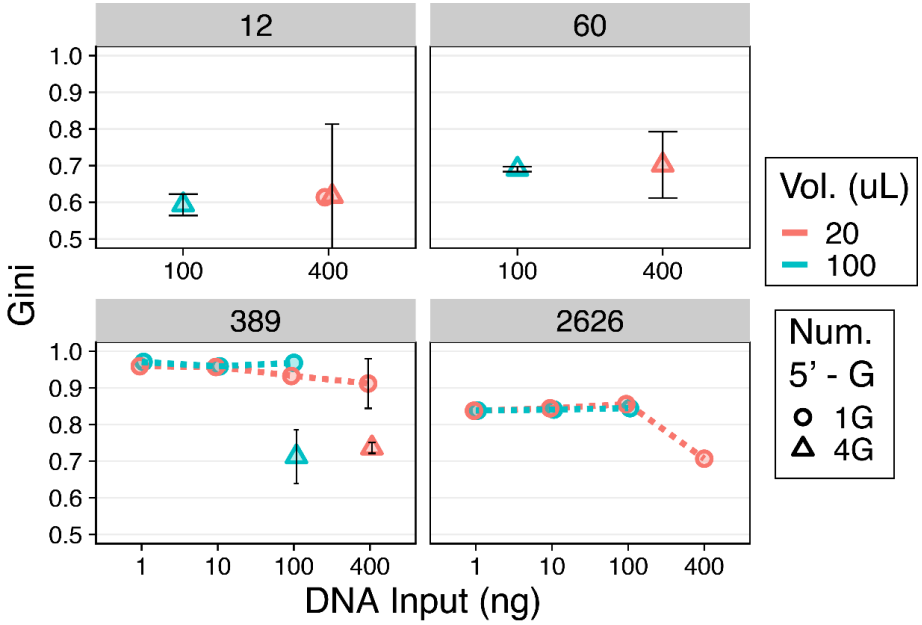


Figure 14. Gini Coefficients of sgRNA libraries across different scales, DNA input amounts, and reaction volumes.

Libraries were transcribed in either a standard 20 µL reaction volume (orange) or a scaled-up 100 µL reaction volume (blue). Spacers were designed to start with either a single 5' guanine (1G, shown as open circles) or four 5' guanines (4G, shown as open triangles), except for the 2,626-plex library, which only contains 1G spacers. Libraries were generated at four different scales: 12-plex, 60-plex, 389-plex, and 2,626-plex, using DNA input amounts ranging from 1 ng to 400 ng. Most libraries were transcribed with a sample size of one. However, the following sgRNA libraries contain a sample size of two: 12-plex (4G, 100 ng, 100 µL), 60-plex (4G, 100 ng, 100 µL), 60-plex (4G, 400 ng, 20 µL), and 389-plex (1G, 400 ng, 20 µL). The following sgRNA libraries contain a sample size of three: 12-plex (4G, 400 ng, 20 µL), 389-plex (4G, 100 ng, 100 µL), and 389-plex (4G, 400 ng, 20 µL). For libraries with n > 2, the median Gini Coefficient is plotted with bars representing standard deviation.

In the 389-plex libraries, small yet statistically significant changes in Gini Coefficients were observed across DNA input conditions in 100 μL reactions (KS test, $p < 0.05$; Table 5, see Appendix B). The 10-ng input showed a slight reduction in Gini (0.96) compared to the 1-ng and 100-ng inputs (both 0.97) (Fig. 14). More substantial changes were observed in 20 μL reactions, where increasing DNA input from 1 ng to 400 ng progressively and significantly reduced the Gini Coefficient from 0.97 to 0.91 (Fig. 14, KS test, $p < 0.05$, Table 5). This suggests that higher DNA input can improve library uniformity in smaller reaction volumes. The 4G version of the 389-plex library consistently outperformed the 1G version, with lower Gini values (0.71 to 0.74) across both reaction volumes (Fig. 14), consistent with earlier results (Fig. 12) demonstrating that the 5' GGGG modification reduces T7 RNA polymerase transcriptional bias relative to 5' G spacers.

For the 2,626-plex libraries, there was minor but statistically significant variation in Gini Coefficients across DNA input amounts in both 20 μL and 100 μL reactions (KS test, $p < 0.05$, Table 5, see Appendix B). At this larger scale, the effect of DNA input on uniformity became more pronounced. In 100 μL reactions, libraries prepared with 1 ng and 100 ng DNA showed similar Gini Coefficients (approximately 0.84), while the 10-ng condition showed a slight but statistically significant improvement (Gini = 0.81) (KS test, $p < 0.05$, Table 5). For the 20 μL reactions, the trend was clearer: increasing DNA input from 1 ng to 400 ng progressively decreased the Gini Coefficient from 0.84 to 0.71, representing the best uniformity observed in this experiment. Notably, the change from 100 ng to 400 ng DNA in the 20 μL condition resulted in a substantial drop in Gini from 0.84 to 0.71 (Fig. 14).

Given that the largest improvement in the Gini Coefficient was observed when increasing the input DNA amount from 100 ng to 400 ng in a 20 µL reaction volume for the 2,626-plex 5'G sgRNA libraries, we evaluated this condition more closely. Notably, the 400-ng library showed a clear reduction in nucleotide bias compared to the 100-ng library, with $\log_2$ fold decrease of 1.3 in guanine overrepresentation at position 2 and a broader recovery of A, T, and C levels toward baseline FC values (Fig. 48A,B see Appendix B). To further investigate whether increasing the DNA input from 100 ng to 400 ng reduces bias at the first five spacer positions which are most susceptible to bias, we examined the percent decrease in guanine representation. The percent decrease at positions 1 to 3 was at or below 53.6% and increased to 331% at position 4 (Fig. 48C, see Appendix B). This represents a 78% reduction in guanine bias at position 4 compared to the 253% decrease in guanine representation observed for the 2,626-plex library produced by eIVT with 1 ng of input DNA (Figs. 48C and 46A). However, since in the baseline condition guanines are not as overrepresented at position 4 compared to positions 1 and 2, any decrease in guanine representation at this position results in larger percent decreases. Finally, the most overrepresented spacers within the 2,626-plex sgRNA library decreased in abundance when the library was produced using eIVT with 400 ng of input DNA, further supporting these observations (Fig. 48D).

We hypothesize that the improvements in Gini Coefficient for libraries transcribed with 400 ng input DNA in 20 µL result from higher sgRNA template concentration. This likely increases T7 RNAP substrate availability, leading to more efficient and uniform transcription. While higher DNA input improves uniformity, the most consistent enhancement comes from using spacers that start with a 5′ GGGG sequence rather than 5′ G. This benefit was observed across various library sizes, including the 389-plex (Figs. 12 and 14), and likely applies to larger

libraries as well. Therefore, we conclude that incorporating a 5′ GGGG sequence before spacer regions is the most effective strategy for improving sgRNA library uniformity, especially for libraries of 389 spacers or more.

When redesigning spacers with a 5′ GGGG sequence is not feasible, we recommend using 400 ng of input DNA instead of 100 ng for IVT of large-scale sgRNA libraries in a 20 µL reaction volume. Our findings may also inform future optimization of our eIVT method, as we observed that increasing DNA input to 100 ng reduced bias in both the 389- and 2,626-plex libraries. While we tested up to 800 ng of input DNA for smaller libraries (12- and 60-plex) without observing bias reduction, we did not test this approach at larger scales, where higher DNA input may still yield additional improvements in uniformity.

3.5 Conclusions

Large-scale sgRNA libraries are essential tools in CRISPR-Cas9 research, enabling functional screens and a wide range of *in vitro* assays. However, there is a growing demand for user-defined, programmable sgRNA libraries that can be prepared using simple and cost-effective methods. In this work, we developed a workflow that meets this need while addressing a critical challenge often overlooked following IVT of heterogeneous templates: the spacer distribution biases introduced by T7 RNA polymerase. We found that T7 RNAP strongly favors spacers containing guanines in their first four positions, resulting in the systematic underrepresentation of other nucleotide sequences. This bias was further confirmed by comparing the severe spacer inequalities in our sgRNA libraries to the highly uniform DNA templates used for transcription.

The bias we identified in sgRNA libraries can compromise the accuracy and reproducibility of CRISPR-Cas9 screens by underrepresenting certain spacers, potentially leading to the loss of important functional hits. Beyond CRISPR screens, T7 RNAP biases may also affect methods that rely on IVT-produced RNA baits for hybridization to exons or ancient DNA in targeted sequencing, although this remains to be assessed (60, 61). Consequently, these biases could impact a broad range of CRISPR-based and RNA library-based studies.

In this study, we explored three independent strategies to mitigate T7 RNAP-driven bias and improve sgRNA library uniformity. The most effective was the addition of a 5' guanine tetramer to all spacer sequences, which consistently and substantially reduced bias. This simple design stage modification has the potential to enhance the accuracy and reproducibility of CRISPR-Cas9 screens by enhancing the detection of functional hits that might otherwise be missed due to poor spacer representation. However, we also found that 5'GGGG modified templates generate substantially more high-molecular-weight (HMW) RNA byproducts than those starting with a single 5'G posing a key limitation to this approach.

In addition to spacer sequence redesign, we explored two alternative strategies to reduce transcription bias: compartmentalizing transcription within emulsion droplets and varying DNA input and reaction volumes during bulk IVT. Our results demonstrate that emulsions improve library uniformity in a scale-dependent manner, particularly at lower DNA inputs. Future optimizations, such as ligating input DNA templates to barcoded beads prior to emulsification, could enhance this effect. Notably, an sgRNA library transcribed in emulsions with a low DNA input amount (e.g., 10 ng) lacked HMW RNA byproducts, which are known to trigger inflammatory responses in human cell lines and *in vivo*. Although this requires further validation, it suggests that emulsion-based transcription could offer additional, unexpected benefits. We also

found that higher DNA concentrations (20 ng/μL) significantly reduced bias compared to lower concentrations (5 ng/μL or below). When spacer redesign or emulsion-based transcription is not feasible, we recommend maximizing DNA input (400 ng) in small reaction volumes (20 μL) under standard IVT conditions to achieve optimal library uniformity.

Beyond addressing T7 RNAP biases, our results show that modern microarray-derived oligos have low mutation rates and comparable quality compared to column-derived oligos. This contrasts with earlier reports of poor microarray oligo fidelity and reflects improvements in synthesis technologies. By ordering a pool of 11,640 unique oligos and subpooling them via PCR into 10 sgRNA libraries of varying sizes, we achieved up to 72% cost savings compared to ordering individual libraries. Coupled with our IVT workflow, this approach eliminates the need for commercial kits requiring specific formats, making large-scale sgRNA library production more accessible and affordable for high-throughput studies.

Finally, this study opens avenues for future improvements in sgRNA design and synthesis. The principles used to mitigate T7 RNAP biases could be extended to guide RNA libraries for alternative Cas proteins, such as Cas12a and Cas13a. In particular, CRISPR-Cas12a guide RNA (crRNA) libraries may naturally exhibit greater spacer uniformity due to their sequence architecture where the conserved crRNA scaffold lies at the 5’ end downstream of the T7 promoter (15), unlike Cas9 sgRNAs, which place the variable spacer region at the 5’ end. By combining emulsion-based synthesis with careful reaction optimization, our approach represents an important step toward producing high-quality, uniformly distributed sgRNA libraries. We hope this work inspires new methodologies and enhances the precision and reliability of CRISPR-based applications.

3.6 Conflicts of Interest

N.V. and C.P. are named inventors on a patent based on this method. CP is a co-founder and holds equity in SynPlexity.

3.7 Data and Materials Availability

Raw nanopore sequence reads for sgRNA libraries were submitted to the NCBI Sequence Read Archive under BioProject accession PRJNA1237881 (https://www.ncbi.nlm.nih.gov/bioproject/1237881).

Supplementary Data 1 contains all microarray and oPool oligo sequences. Supplementary Data 2 contains processed data sequences and counts can be found at https://doi.org/10.6084/m9.figshare.28635950

3.8 References

1. Wang,J.Y. and Doudna,J.A. (2023) CRISPR technology: A decade of genome editing is only the beginning. *Science*, **379**, eadd8643.

2. Miles,L.A., Garippa,R.J. and Poirier,J.T. (2016) Design, execution, and analysis of pooled *in vitro* CRISPR/Cas9 screens. *FEBS J.*, **283**, 3170–3180.

3. Sanjana,N.E., Shalem,O. and Zhang,F. (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*, **11**, 783–784.

4. Sanson,K.R., Hanna,R.E., Hegde,M., Donovan,K.F., Strand,C., Sullender,M.E., Vaimberg,E.W., Goodale,A., Root,D.E., Piccioni,F., *et al.* (2018) Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat. Commun.*, **9**, 5416.

5. Wang,T., Birsoy,K., Hughes,N.W., Krupczak,K.M., Post,Y., Wei,J.J., Lander,E.S. and Sabatini,D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, **350**, 1096–1101.

6. Horlbeck,M.A., Gilbert,L.A., Villalta,J.E., Adamson,B., Pak,R.A., Chen,Y., Fields,A.P., Park,C.Y., Corn,J.E., Kampmann,M., *et al.* (2016) Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife*, **5**.

7. Quintero-Ruiz,N., Oliveira,W. de L., Esteca,M.V., Granato,D.C. and Simabuco,F.M. (2024) Uncovering the bookshelves of CRISPR-based libraries: Advances and applications in cancer studies. *Crit. Rev. Oncol. Hematol.*, **196**, 104287.

8. Hart,T., Tong,A.H.Y., Chan,K., Van Leeuwen,J., Seetharaman,A., Aregger,M., Chandrashekhar,M., Hustedt,N., Seth,S., Noonan,A., *et al.* (2017) Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)*, **7**, 2719–2727.

9. Blanck,M., Budnik-Zawilska,M.B., Lenger,S.R., McGonigle,J.E., Martin,G.R.A., le Sage,C., Lawo,S., Pemberton,H.N., Tiwana,G.S., Sorrell,D.A., *et al.* (2020) A flexible, pooled CRISPR library for drug development screens. *CRISPR J.*, **3**, 211–222.

10. Read,A., Gao,S., Batchelor,E. and Luo,J. (2017) Flexible CRISPR library construction using parallel oligonucleotide retrieval. *Nucleic Acids Res.*, **45**, e101.

11. Lin,Y., Wagner,E. and Lächelt,U. (2022) Non-viral delivery of the CRISPR/Cas system: DNA versus RNA versus RNP. *Biomater Sci*, **10**, 1166–1192.

12. Kim,S., Kim,D., Cho,S.W., Kim,J. and Kim,J.-S. (2014) Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.*, **24**, 1012–1019.

13. Marinov,G.K., Kim,S.H., Bagdatli,S.T., Higashino,S.I., Trevino,A.E., Tycko,J., Wu,T., Bintu,L., Bassik,M.C., He,C., *et al.* (2023) CasKAS: direct profiling of genome-wide dCas9 and Cas9 specificity using ssDNA mapping. *Genome Biol.*, **24**, 85.

14. Liszczak,G.P., Brown,Z.Z., Kim,S.H., Oslund,R.C., David,Y. and Muir,T.W. (2017) Genomic targeting of epigenetic probes using a chemically tailored Cas9 system. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 681–686.

15. Mighell,T.L., Nishida,A., O'Connell,B.L., Miller,C.V., Grindstaff,S., Thornton,C.A., Adey,A.C., Doherty,D. and O'Roak,B.J. (2022) Cas12a-Capture: A Novel, Low-Cost, and Scalable Method for Targeted Sequencing. *CRISPR J*, **5**, 548–557.

16. Boyle,E.A., Andreasson,J.O.L., Chircus,L.M., Sternberg,S.H., Wu,M.J., Guegler,C.K., Doudna,J.A. and Greenleaf,W.J. (2017) High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 5461–5466.

17. Engler,C., Kandzia,R. and Marillonnet,S. (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One*, **3**, e3647.

18. Chao,R., Yuan,Y. and Zhao,H. (2015) Recent advances in DNA assembly technologies. *FEMS Yeast Res.*, **15**, 1–9.

19. Casini,A., Storch,M., Baldwin,G.S. and Ellis,T. (2015) Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.*, **16**, 568–576.

20. Kennedy,W.P., Momand,J.R. and Yin,Y.W. (2007) Mechanism for de novo RNA synthesis and initiating nucleotide specificity by t7 RNA polymerase. *J. Mol. Biol.*, **370**, 256–268.

21. Heo,S.-J., Enriquez,L.D., Federman,S., Chang,A.Y., Mace,R., Shevade,K., Nguyen,P., Litterman,A.J., Shafer,S., Przybyla,L., *et al.* (2024) Compact CRISPR genetic screens enabled by improved guide RNA library cloning. *Genome Biol.*, **25**, 25.

22. Fulco,C.P., Munschauer,M., Anyoha,R., Munson,G., Grossman,S.R., Perez,E.M., Kane,M., Cleary,B., Lander,E.S. and Engreitz,J.M. (2016) Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, **354**, 769–773.

23. Kuzmine,I., Gottlieb,P.A. and Martin,C.T. (2003) Binding of the priming nucleotide in the initiation of transcription by T7 RNA polymerase. *J. Biol. Chem.*, **278**, 2819–2823.

24. Cazenave,C. and Uhlenbeck,O.C. (1994) RNA template-directed RNA synthesis by T7 RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 6972–6976.

25. Plesa,C., Sidore,A.M., Lubock,N.B., Zhang,D. and Kosuri,S. (2018) Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science*, **359**, 343–347.

26. Sidore,A.M., Plesa,C., Samson,J.A., Lubock,N.B. and Kosuri,S. (2020) DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Res.*, **48**, e95.

27. Holston,A.S., Hinton,S.R., Lindley,K.A., Kearns,N.C. and Plesa,C. (2023) Degenerate DropSynth for Simultaneous Assembly of Diverse Gene Libraries and Local Designed Mutants. 10.1101/2023.12.11.569291.

28. Gholamalipour,Y., Karunanayake Mudiyanselage,A. and Martin,C.T. (2018) 3’ end additions by T7 RNA polymerase are RNA self-templated, distributive and diverse in character-RNA-Seq analyses. *Nucleic Acids Res.*, **46**, 9253–9263.

29. Mateyko,N. and de Boer,C.G. (2024) Culture wars: Empirically determining the best approach for Plasmid library amplification. *ACS Synth. Biol.*, **13**, 2328–2334.

30. Huang,J., Brieba,L.G. and Sousa,R. (2000) Misincorporation by wild-type and mutant T7 RNA polymerases: identification of interactions that reduce misincorporation rates by stabilizing the catalytically incompetent open conformation. *Biochemistry*, **39**, 11571–11580.

31. Nichols,R.V., Vollmers,C., Newsom,L.A., Wang,Y., Heintzman,P.D., Leighton,M., Green,R.E. and Shapiro,B. (2018) Minimizing polymerase biases in metabarcoding. *Mol. Ecol. Resour.*, **18**, 927–939.

32. Dabney,J. and Meyer,M. (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, **52**, 87–94.

33. Aird,D., Ross,M.G., Chen,W.-S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

34. Pan,W., Byrne-Steele,M., Wang,C., Lu,S., Clemmons,S., Zahorchak,R.J. and Han,J. (2014) DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnol.*, **14**, 10.

35. Polz,M.F. and Cavanaugh,C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.*, **64**, 3724–3730.

36. Martins,A.L., Walavalkar,N.M., Anderson,W.D., Zang,C. and Guertin,M.J. (2018) Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Res.*, **46**, e9.

37. Sung,M.-H., Guertin,M.J., Baek,S. and Hager,G.L. (2014) DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell*, **56**, 275–285.

38. Yardımcı,G.G., Frank,C.L., Crawford,G.E. and Ohler,U. (2014) Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.*, **42**, 11865–11878.

39. Conrad,T., Plumbom,I., Alcobendas,M., Vidal,R. and Sauer,S. (2020) Maximizing transcription of nucleic acids with efficient T7 promoters. *Commun. Biol.*, **3**, 439.

40. Barber,K.W., Shrock,E. and Elledge,S.J. (2022) CasPlay provides a gRNA-barcoded CRISPR-based display platform for antibody repertoire profiling. *Cell Rep Methods*, **2**, 100318.

41. Henser-Brownhill,T., Monserrat,J. and Scaffidi,P. (2017) Generation of an arrayed CRISPR-Cas9 library targeting epigenetic regulators: from high-content screens to *in vivo* assays. *Epigenetics*, **12**, 1065–1075.

42. Pleiss,J.A., Derrick,M.L. and Uhlenbeck,O.C. (1998) T7 RNA polymerase produces 5’ end heterogeneity during *in vitro* transcription from certain templates. *RNA*, **4**, 1313–1317.

43. Villegas, N. K., Tran, M. H., Plesa, C. (2025) Barcode-Assisted Retrieval-CRISPR Activated Targeting (BAR-CAT) is method for enriching synthetic genes, In preparation.

44. Kuiper,B.P., Prins,R.C. and Billerbeck,S. (2022) Oligo pools as an affordable source of synthetic DNA for cost-effective library construction in protein- and metabolic pathway engineering. *Chembiochem*, **23**, e202100507.

45. Kosuri,S. and Church,G.M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*, **11**, 499–507.

46. Fu,Y., Li,C., Lu,S., Zhou,W., Tang,F., Xie,X.S. and Huang,Y. (2015) Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 11923–11928.

47. Arnaud-Barbe,N., Cheynet-Sauvion,V., Oriol,G., Mandrand,B. and Mallet,F. (1998) Transcription of RNA templates by T7 RNA polymerase. *Nucleic Acids Res.*, **26**, 3550–3554.

48. Gholamalipour,Y., Johnson,W.C. and Martin,C.T. (2019) Efficient inhibition of RNA self-primed extension by addition of competing 3'-capture DNA-improved RNA synthesis by T7 RNA polymerase. *Nucleic Acids Res.*, **47**, e118.

49. Triana-Alonso,F.J., Dabrowski,M., Wadzack,J. and Nierhaus,K.H. (1995) Self-coded 3'-extension of run-off transcripts produces aberrant products during *in vitro* transcription with T7 RNA polymerase. *J. Biol. Chem.*, **270**, 6298–6307.

50. Liu,L., Botos,I., Wang,Y., Leonard,J.N., Shiloach,J., Segal,D.M. and Davies,D.R. (2008) Structural basis of toll-like receptor 3 signaling with double-stranded RNA. *Science (New York, N.Y.)*, **320**.

51. Mu,X. and Hur,S. (2021) Immunogenicity of *in vitro*-transcribed RNA. *Acc. Chem. Res.*, **54**, 4012–4023.

52. Camperi,J., Roper,B., Freund,E., Leylek,R., Nissenbaum,A., Galan,C., Caothien,R., Hu,Z., Ko,P., Lee,A., *et al.* (2024) Exploring the impact of *in vitro*-transcribed mRNA impurities on cellular responses. *Anal. Chem.*, **96**, 17789–17799.

53. Weissman,D., Pardi,N., Muramatsu,H. and Karikó,K. (2013) HPLC purification of *in vitro* transcribed long RNA. *Methods Mol. Biol.*, **969**, 43–54.

54. Baiersdörfer,M., Boros,G., Muramatsu,H., Mahiny,A., Vlatkovic,I., Sahin,U. and Karikó,K. (2019) A facile method for the removal of dsRNA contaminant from *in vitro*-transcribed mRNA. *Mol. Ther. Nucleic Acids*, **15**, 26–35.

55. Katri,E., Mirka,L., Christine,C., Jafargholi,I., Karl-Heinz,K. and M,P.M. (2022) Analysis and purification of ssRNA and dsRNA molecules using asymmetrical flow field flow fractionation. *J. Chromatogr. A*, **1683**, 463525.

56. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.

57. Dousis,A., Ravichandran,K., Hobert,E.M., Moore,M.J. and Rabideau,A.E. (2023) An engineered T7 RNA polymerase that produces mRNA free of immunostimulatory byproducts. *Nat. Biotechnol.*, **41**, 560–568.

58. Magde,D., Akoopie,A., Magde,M.D.,Jr and Müller,U.F. (2021) Water/oil emulsions with controlled droplet sizes for *in vitro* selection experiments. *ACS Omega*, **6**, 21773–21783.

59. Vladisaljević,G.T. (2024) Droplet microfluidics for high-throughput screening and directed evolution of biomolecules. *Micromachines (Basel)*, **15**, 971.

60. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C., *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.

61. Carpenter,M.L., Buenrostro,J.D., Valdiosera,C., Schroeder,H., Allentoft,M.E., Sikora,M., Rasmussen,M., Gravel,S., Guillén,S., Nekhrizov,G., *et al.* (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.*, **93**, 852–864.

4. General Conclusions

This dissertation integrates DropSynth gene synthesis, CRISPR-based DNA enrichment, and scalable sgRNA synthesis technologies to develop novel tools in molecular biology. In Aim 1 (Chapter 2), I tackled the challenge of imperfect assemblies within high-diversity synthetic gene libraries produced by DropSynth. As gene lengths exceed 1 kb, the yield of error-free products becomes insufficient to support multiplexed assays of variant effects (MAVEs), which are critical for large-scale exploration of sequence-function relationships and the development of machine learning models for protein design. While PCR-based methods can isolate perfect genes from libraries also containing imperfect genes, they lack scalability. To overcome this, I developed Barcode-Assisted Retrieval using CRISPR Activated Targeting (BAR-CAT), a method that enriches perfect synthetic genes using CRISPR-dCas9.

In BAR-CAT, genes are tagged with unique PAM-adjacent 20-mer barcodes, which are mapped to error-free assemblies and selected as sgRNA spacers through a computational pipeline. These selected barcodes are transcribed into sgRNA libraries, which direct biotinylated dCas9 to the targeted barcodes. The bound dCas9 can then be pulled down with streptavidin beads, and the enriched DNA can be amplified and sequenced. As a proof of concept, BAR-CAT enriched 18 barcodes in a single-gene (*rfp*) library by a median of 6.3-fold. Subsequent protocol improvements, including increased bead wash volumes and higher DNA input, led to BAR-CAT version 1.0, achieving up to 1,094-fold enrichment of three targets. Scaling up to 384- and 1,536-gene libraries revealed new challenges, such as sgRNA competition for free dCas9, which led to target dropout beyond 12 targets. Despite these limitations, BAR-CAT provides a promising framework for multiplexed DNA enrichment with applications in synthetic biology, ancient DNA recovery, diagnostics, and targeted sequencing.

In parallel, I developed a scalable, cost-effective sgRNA synthesis workflow, reducing costs by ~70% through sub-pool amplification of microarray-derived oligos encoding unique spacers. These oligos were assembled into double-stranded DNA templates using Golden Gate Assembly and *in vitro* transcribed with T7 RNA polymerase. RNA-seq revealed significant sequence bias, with certain spacers highly overrepresented or absent, primarily due to guanine-rich sequences immediately downstream of the T7 promoter. To mitigate this, I introduced a guanine tetramer upstream of all spacers, improving uniformity by 19% in a 389-spacer library. However, this modification also produced high-molecular-weight RNA species, diluting the abundance of active sgRNAs. To address this, I tested two alternative strategies, emulsion-based *in vitro* transcription (eIVT) and DNA input optimization, which reduced bias in 2,626-plex libraries. An improved version of eIVT is currently being tested to further enhance uniformity. These advances increased both the affordability and uniformity of sgRNA pools and contributed directly to the development of BAR-CAT, with broader applications in CRISPR-Cas9 screens and guide RNA design.

Together, these efforts demonstrate how the development of new *in vitro* systems can drive both technical innovation and biological insights. For example, our observation that T7 RNA polymerase preferentially transcribes guanine tetramers has broader implications for RNA production across various systems. Similarly, in optimizing BAR-CAT, we found that barcode abundance in input libraries influences system noise and off-target enrichment—an insight likely applicable to any *in vitro* CRISPR system involving uneven input distributions. These findings underscore the importance of interdisciplinary approaches that connect tool development to biological insight.

Throughout both projects, I encountered various strategies to overcome technical bottlenecks, including enzyme engineering, reagent substitution, and process optimization. While enzyme engineering can be powerful, it is time-intensive and challenging to scale. Instead, I focused on process optimization and minor reagent changes. For instance, BAR-CAT performance improved with simple adjustments such as more stringent washes and increased DNA input. Future work will explore additional optimization strategies, including testing increased dCas9 concentrations and other effector proteins, such as wild-type Cas9. Similarly, while optimizing the sgRNA library synthesis workflow, I avoided modifying T7 RNA polymerase and instead improved performance through sgRNA sequence design and reaction format. These examples highlight the power of lightweight, scalable optimizations in advancing molecular biology tools without requiring extensive protein engineering.

Both BAR-CAT and the sgRNA synthesis method rely on microarray-derived oligos as a low-cost foundation for constructing large-scale synthetic libraries. These projects demonstrate the impact of high-throughput oligo pools in enabling novel methods. I encourage researchers to consider the potential of microarray oligos to scale their work. More broadly, this dissertation reflects on the legacy and limitations of phosphoramidite chemistry, which has shaped molecular biology for decades. Looking ahead, I advocate for expanding enzyme-based approaches to DNA and RNA synthesis. Had the field invested in enzyme-powered synthesis earlier, we might now have more robust and programmable systems for synthetic biology. With this work, I aim to contribute to a shift toward enzyme-enabled molecular biology and inspire the development of new methods for perfect gene selection and beyond.

APPENDIX A. Supplementary Material for Chapter 2

A1. Figures

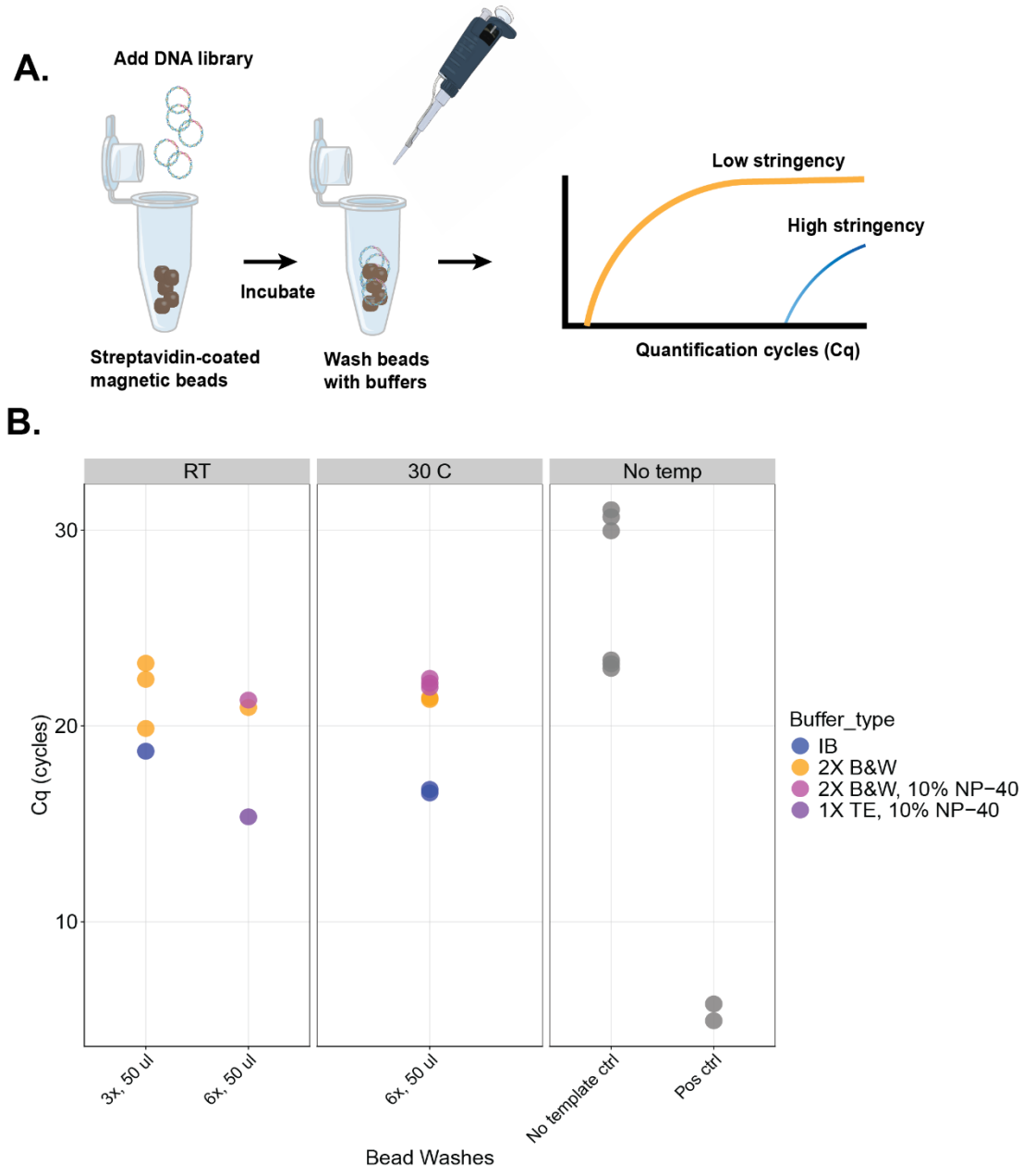


Figure 15. Wash stringency of various buffers at different temperatures when used to wash streptavidin-coated magnetic beads.

A. Schematic of the bead wash stringency test. Single-gene library DNA was incubated with streptavidin-coated beads in the absence of Cas9 or sgRNAs at 37 °C with shaking. Beads were then washed with one of four buffer formulations (see Table 2 in Appendix A). Room temperature (RT) washes were performed three times with 50 µL

buffer, while washes at 25 °C and 30 °C were performed six times with 50 µL buffer. DNA retention was quantified by qPCR to evaluate the efficiency of nonspecific DNA removal. Icons used: microtube-open-blue by Servier (CC-BY 3.0, https://smart.servier.com/), pipette icon by James-Lloyd (CC0, https://www.badgrammargoodsyntax.com/), and plasmid-l-insert-dna by DBCLS (CC-BY 4.0, https://togotv.dbcls.jp/en/pics.html). **B.** qPCR quantification of DNA remaining on beads following buffer washes under the indicated conditions.

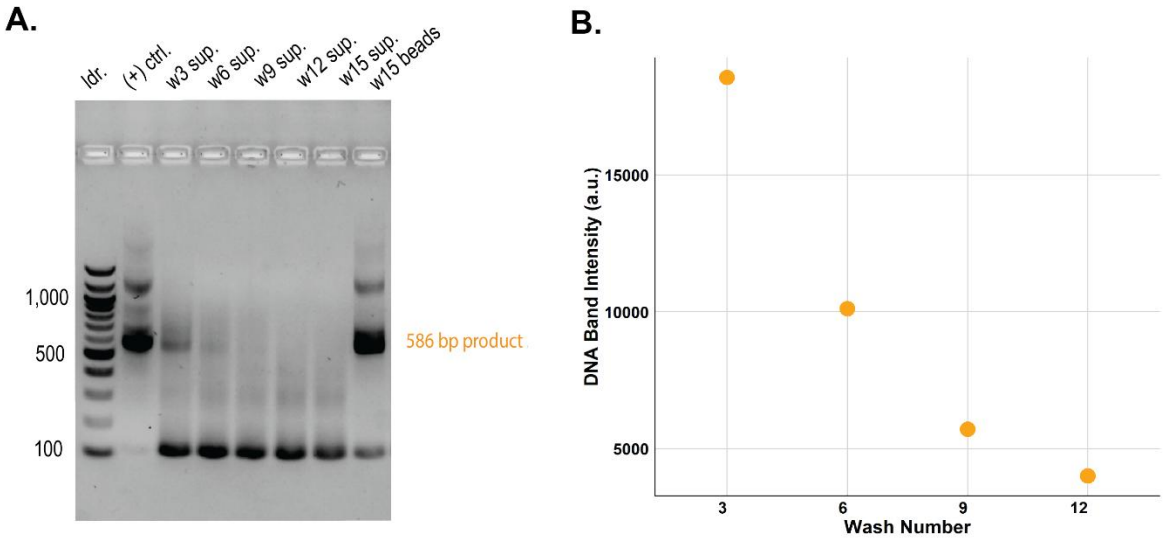


Figure 16. Evaluating residual DNA after bead washing by performing agarose gel electrophoresis of qPCR products.

A. Agarose gel showing qPCR products. A 100 bp NEB ladder was included for size reference. The positive control contained 448 ng of a 586 bp product amplified from the original *rfp* library. SNAP-capture beads were washed 15 times with 1 mL of immobilization buffer (IB, Table 2, see Appendix A). Supernatants were collected from washes 3, 6, 9, 12, and 15, and the beads were retained after wash 15. qPCR was performed on both the supernatants and the washed beads to amplify any remaining *rfp* library DNA. Bands were visualized using SYBR™ Safe DNA Gel Stain. The presence of the 586 bp product indicates residual DNA not removed by washing. **B.** Quantification of DNA band intensities from the gel in panel A using ImageJ (1), shown only for amplification of supernatant samples from washes 3, 6, 9, and 12.

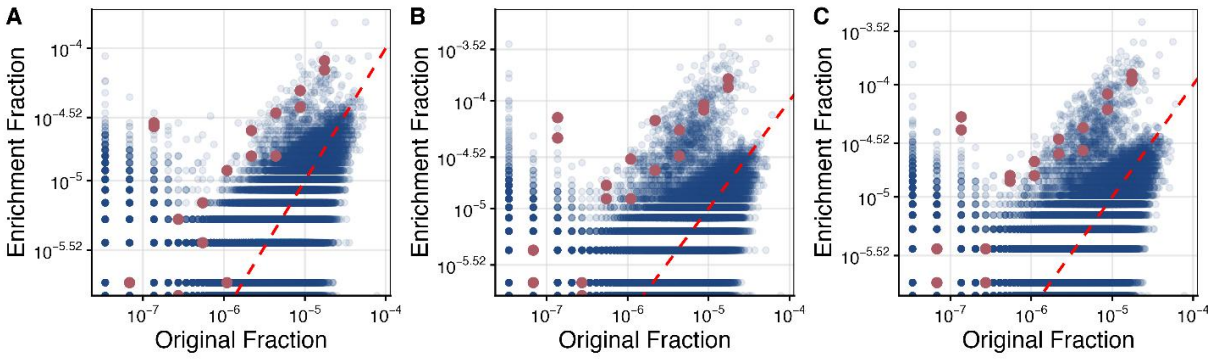


Figure 17. Barcode distributions before and after enrichment of 18 target barcodes for various bead wash conditions.

Scatter plots comparing the fractional abundance of each barcode in the *rfp* library before (original) and after enrichment under different wash conditions. Each blue dot represents a non-target barcode while each magenta dot represents a barcode targeted by the 18-plex sgRNA library. The red dashed unity line indicates equal representation before and after enrichment and serves as a reference for assessing enrichment. Three bead washing protocols were

tested: **A.** Control condition using 6 × 50 µL washes, **B.** 9 × 2 mL washes (used for subsequent enrichments), and **C.** 6 × 5 mL washes.

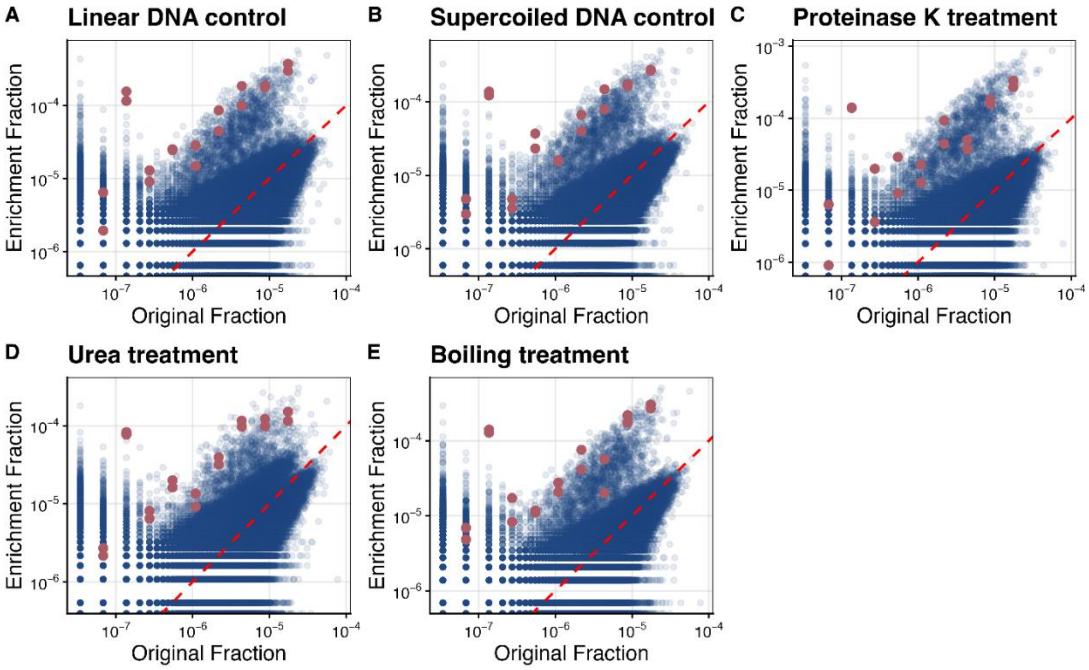


Figure 18. Barcode distributions before and after enrichment of 18 targeted barcodes for various DNA format and dCas9 denaturation conditions.

Scatter plots comparing the fractional abundance of each barcode in the *rfp* library before (original) and after enrichment under different conditions. Each blue dot represents a non-target barcode, while each magenta dot represents a barcode targeted by the 18-plex sgRNA library. The red dashed unity line indicates equal representation before and after enrichment and serves as a reference for assessing enrichment. The following conditions were tested: **A.** Linear DNA control with amplification off the beads, **B.** Supercoiled DNA control with amplification off the beads, **C.** Supercoiled DNA with proteinase K denaturation of dCas9 and amplification from the supernatant (used for subsequent enrichments), **D.** Supercoiled DNA with 8 M urea denaturation of dCas9 and amplification from the supernatant, **E.** Supercoiled DNA with boiling denaturation of dCas9 and amplification from the supernatant.

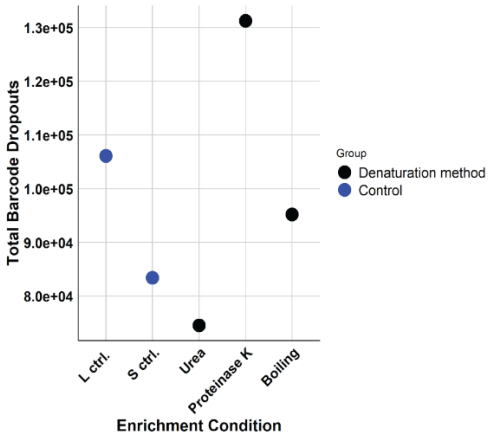


Figure 19. Total barcode dropout counts are compared across three dCas9 denaturation methods: urea, proteinase K, and heat (boiling).

Two reference controls were included, one using a supercoiled *rfp* library and another using a linearized *rfp* library.

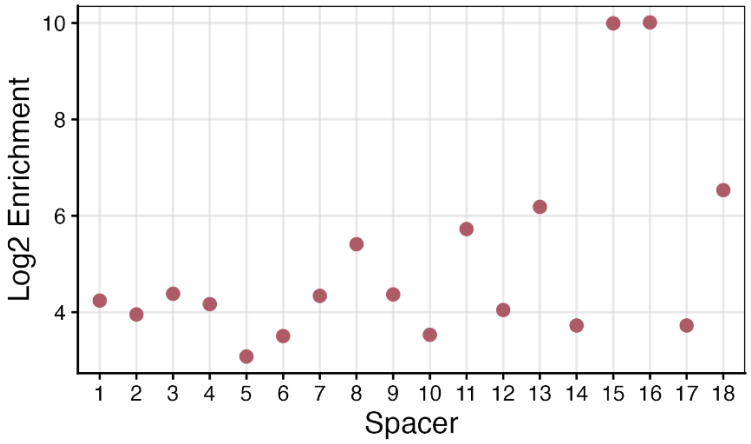


Figure 20. Log2 enrichment values for the 18 individual barcodes targeted from the *rfp* supercoiled library treated with proteinase K.

The spacers were selected from the 18 barcode protospacers in Fig. 7B, with spacer 1 corresponding to the highest abundance barcode and spacer 18 to the least abundant.

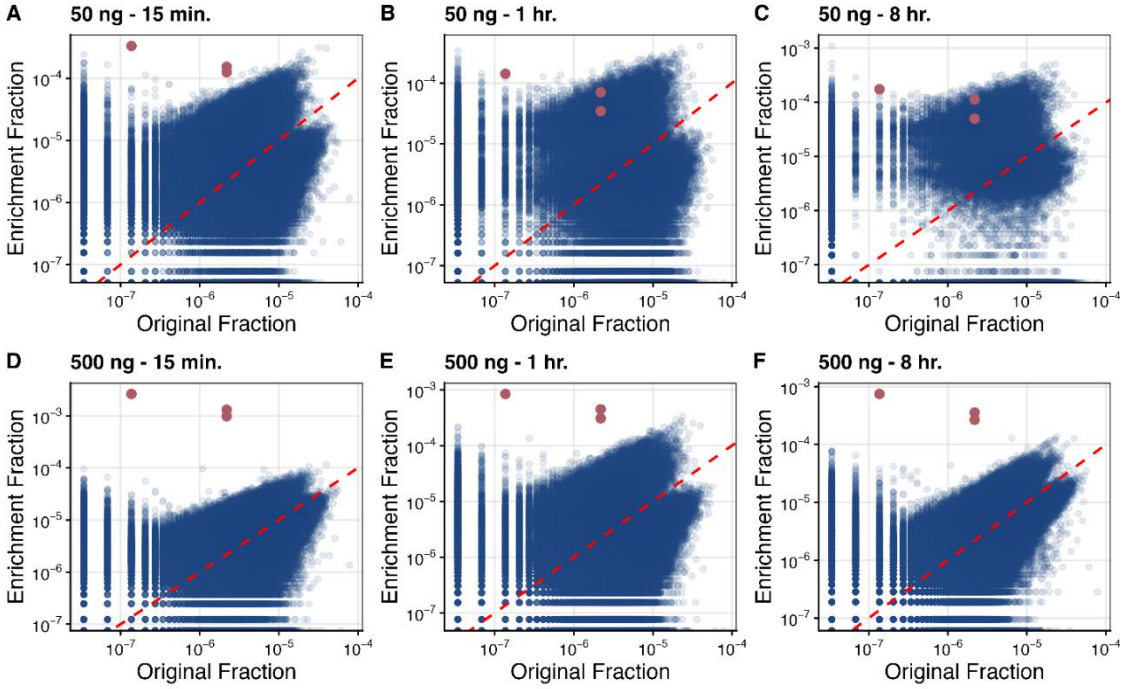


Figure 21. Barcode distributions before and after enrichment of 18 target barcodes for various DNA input amounts.

Scatter plots comparing the fractional abundance of each barcode in the *rfp* library before (original) and after enrichment under different conditions. Each blue dot represents a non-target barcode, while each magenta dot represents a barcode targeted by three unique synthetic sgRNAs. The red dashed unity line indicates equal representation before and after enrichment and serves as a reference for assessing enrichment. The following

conditions were tested: **A.** 50 ng of input DNA with 15 min of enrichment, **B.** 50 ng of input DNA with 1 hour of enrichment, **C.** 50 ng of input DNA with 8 hours of enrichment, **D.** 500 ng of input DNA with 15 min of enrichment, **E.** 500 ng of input DNA with 1 hour of enrichment, **F.** 500 ng of input DNA with 8 hours of enrichment.

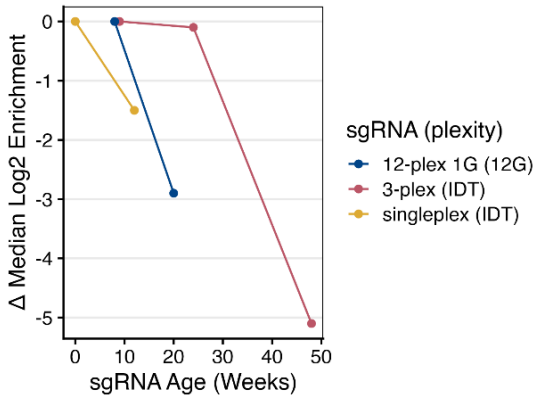


Figure 22. Change in median $\log_2$ enrichment scores as a function of sgRNA age for chemically synthesized and *in vitro* transcribed (IVT) sgRNAs used in BAR-CAT enrichment experiments.

Singleplex and 3-plex pooled sgRNAs (Integrated DNA Technologies) were chemically synthesized, while 12-plex sgRNAs with a 5′ G were generated via IVT using our protocol (2). This analysis evaluates how storage time impacts the performance of sgRNAs used in BAR-CAT.

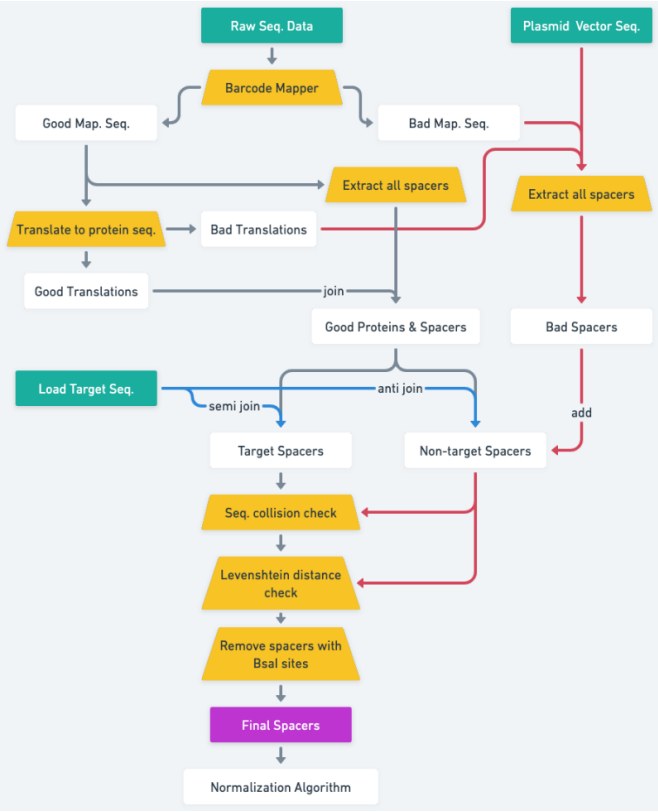


Figure 23. sgRNA spacer selection pipeline.

Long-read amplicon sequencing data from barcoded gene libraries were transformed into sgRNA libraries using a pipeline. The pipeline was designed to cover every targeted gene, avoid recognizing non-target sequences, and normalized so that the target genes selected had similar numbers of reads. The description of this entire method can be found in Section 2.3 Materials and Methods.

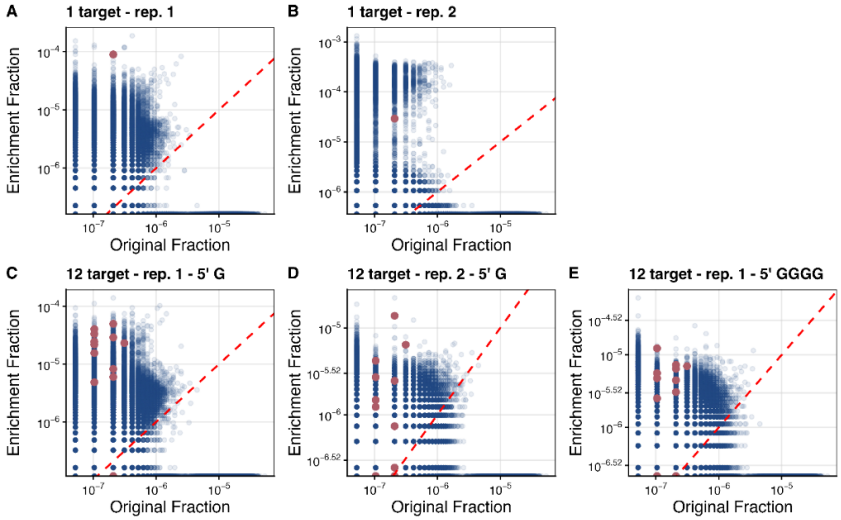


Figure 24. Barcode distributions before and after singleplex and 12-plex enrichment of a DropSynth DHFR library with varying 5′ guanine additions in sgRNA spacers.

Scatter plots compare the fractional abundance of each barcode in a 384-gene DHFR library (library S4) before (original) and after enrichment. Blue dots represent non-target barcodes; magenta dots indicate target barcodes from either a single synthetic sgRNA with a 5′ guanine (5′ G) or a 12-plex sgRNA pool with either 5′ G or a guanine tetramer (5′ GGGG) at the spacer 5′ end. The red dashed unity line denotes equal representation before and after enrichment, serving as a reference to assess barcode enrichment. Conditions shown: **A.** singleplex enrichment with synthetic 5′ G sgRNA (replicate 1), **B.** replicate 2 of panel **A**, **C.** 12-plex enrichment with 5′ GGGG sgRNAs (replicate 1), **D.** replicate 2 of panel **C**, E. 12-plex enrichment with 5′ GGGG (replicate 1).

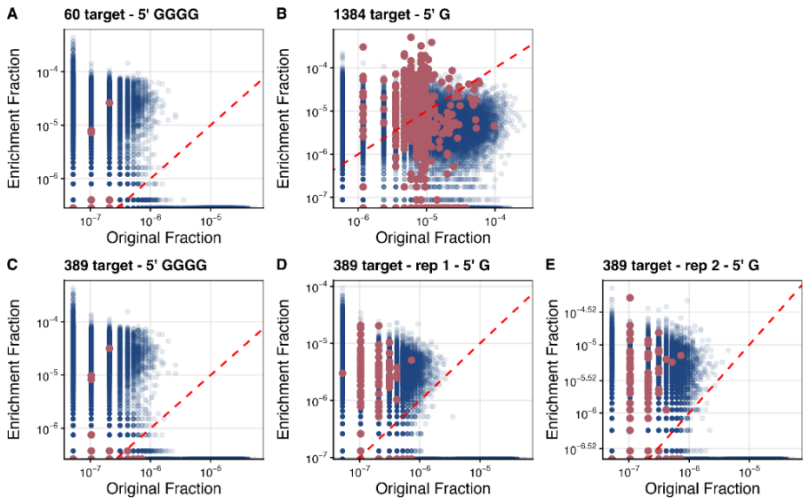


Figure 25. Barcode distributions before and after 60-, 389-, or 1,384-plex enrichment from DropSynth DHFR libraries with varying 5′ guanine additions in sgRNA spacers.

Scatter plots compare the fractional abundance of each barcode in 384-gene (S4) and 1,536-gene (S2) DropSynth DHFR libraries before (original) and after enrichment. Blue dots indicate non-target barcodes; magenta dots

represent target barcodes from IVT sgRNA libraries with spacers beginning with either a 5′ guanine (5′ G) or a guanine tetramer (5′ GGGG). The red dashed unity line marks equal abundance before and after enrichment, serving as a reference to assess enrichment or depletion. Conditions shown: **A.** 60-plex enrichment with 5′ GGGG sgRNAs from a 384 gene DHFR library (S4), **B.** 1,384-plex enrichment with 5′ G sgRNAs from a 1,536 gene DHFR library (S2), **C.** 389-plex enrichment with 5′ GGGG sgRNAs from a 384-gene DHFR library (S4), **D.** 389-plex enrichment with 5′ G sgRNAs (replicate 1) from a 384-gene DHFR library (S4), **E**. Replicate 2 of panel **C**.

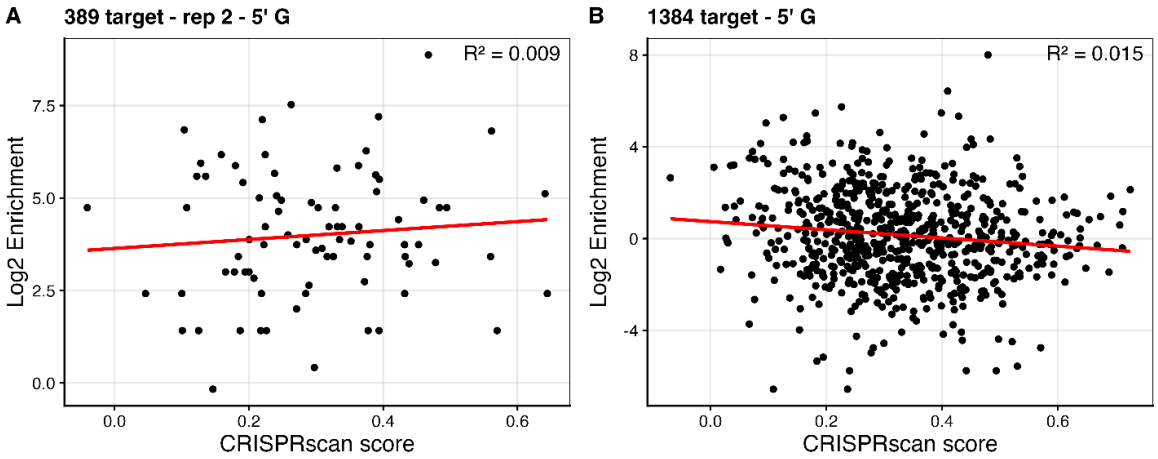


Figure 26. Correlation analysis of $\log_2$ enrichment values for targeted barcodes versus predicted sgRNA performance from the CRISPRscan algorithm (3).

Both the x and y axes are on the log scale and the red line shows the linear regression fit. **A.** Linear regression (R^2= 0.009) for 389-target barcode enrichment with 5′ G sgRNAs (replicate 2) from the 384-gene DHFR library (S4). **B.** Linear regression (R^2 = 0.015) for 1,384-target barcode enrichment with 5′ G sgRNAs from the 1,536-gene DHFR library (S2). Other scores are located in this link https://bioconductor.org/packages/release/bioc/vignettes/crisprScore/inst/doc/crisprScore.html

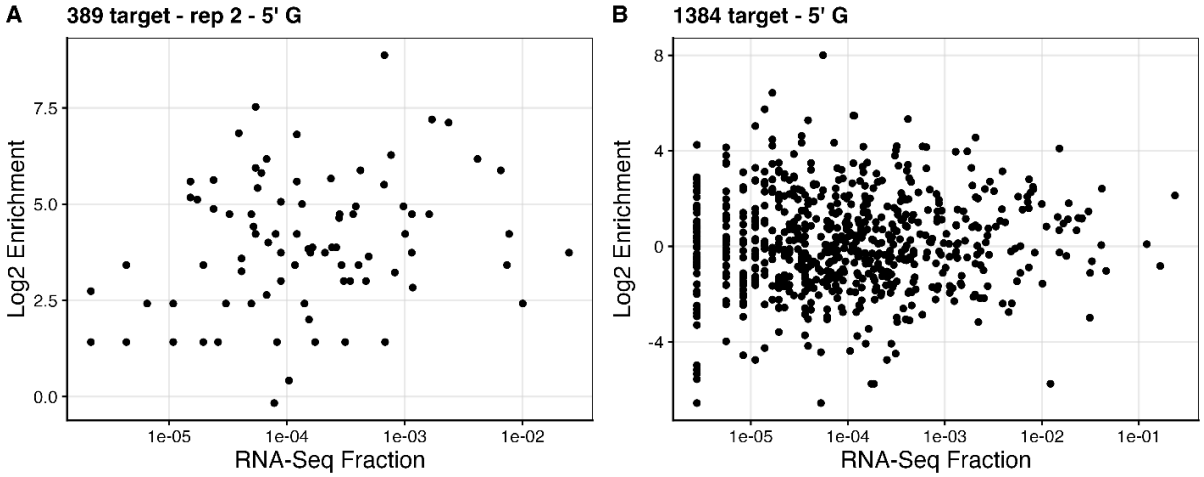


Figure 27. Distribution of $\log_2$ enrichment values for targeted barcodes relative to the abundance of their corresponding spacers within transcribed sgRNA libraries.

RNA-seq was performed on the sgRNA libraries (2), and the RNA-seq fraction for each spacer was plotted to assess the relationship between enrichment and spacer abundance. **A.** 389-plex enrichment with 5′ G sgRNAs. **B.** 1,384-plex enrichment with 5′ G sgRNAs.

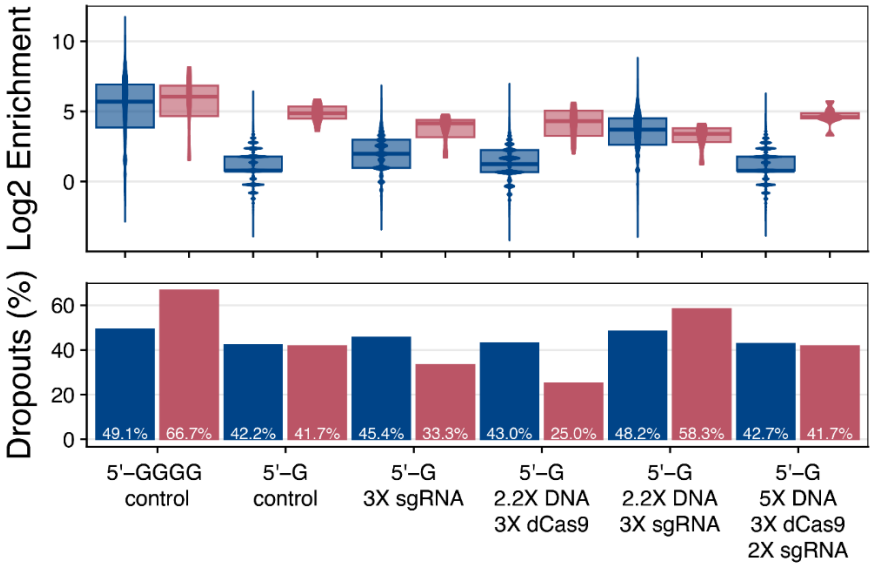


Figure 28. Assessing the impact of varying DNA, sgRNAs, and dCas9 input amounts on the enrichment of 12 targets from the 384-gene DHFR library (library S4).

Overlaid violin and boxplots showing $\log_2$ enrichment scores, calculated as the $\log_2$ fold change in barcode abundance before and after enrichment. Each distribution compares non-target (blue) and target (magenta) barcodes across the different enrichment conditions tested (x-axis). Shaded areas represent the interquartile range (25th–75th percentile); bars indicate median $\log_2$ enrichment. Percent dropout values are listed in the bar plots corresponding to off-target (blue) and target (magenta) barcodes according to the scale of targeted barcodes, as indicated by the x-axis. None of the modified sgRNA, dCas9, or DNA input amounts affected enrichment scores or barcode dropouts compared to the 5' GGGG and 5' G controls.

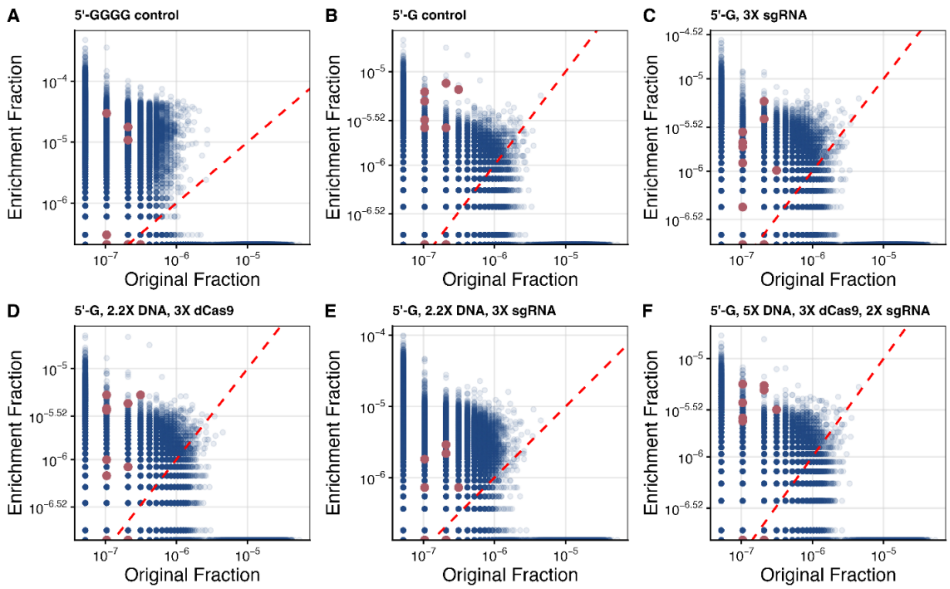


Figure 29. Barcode distributions before and after 12-plex enrichment of a 384-gene DHFR library (library S4) with varying amounts of input DNA, sgRNAs, and dCas9.

Scatter plots compare the fractional abundance of barcodes in the 384-gene DHFR library (S4) before (original) and after enrichment. Blue dots represent non-target barcodes, while magenta dots represent 12 targeted barcodes enriched under the experimental conditions listed above each plot. All enrichments used 5′ G sgRNAs, except for panel A, which used 5′ GGGG sgRNAs. The red dashed unity line indicates equal representation before and after enrichment, serving as a reference for assessing enrichment. The following conditions were evaluated: **A.** BAR-CAT v1.0 control enrichment using 5′ GGGG sgRNAs, **B.** BAR-CAT v1.0 control enrichment using 5′ G sgRNAs, **C.** Increasing sgRNAs by 3x (5' G) **D.** Increasing input DNA by 2.2x (5' G) and dCas9 by 3x , **E.** Increasing input DNA by 2.2x and sgRNAs by 3x (5' G) **F.** Increasing enrichment volume by 2x, sgRNAs by 2x, input DNA by 5x, and dCas9 by 3x.

A2. Tables

Table 2. Compositions of wash buffers tested for streptavidin bead washing stringency.

The qPCR results for bead wash stringency are shown in Fig. 15 (A1. Figures).

Wash buffer	Buffer Composition
Immobilization buffer (IB)	**1 mM DTT,** 10 mM Tris-HCl, 1 mM EDTA●Na_2, pH 8.0
2X binding and wash buffer (2X B&W)	**2M NaCl**, 1 mM EDTA, 10mM Tris-HCl pH 7.4
2X B&W, 10% NP-40	**2M NaCl**, **10% NP-40,** 1mM EDTA, 10mM Tris-HCl pH 7.4
1X TE, 10% NP-40	**10% NP-40,**1 mM EDTA, 10mM Tris-HCl pH 7.4

A3. References

1. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, 9, 671–675.

2. Villegas,N.K., Gaudreault,Y.R., Keller,A., Kearns,P., Stapleton,J.A. and Plesa,C. (2025) Optimizing *in vitro* transcribed CRISPR-Cas9 single-guide RNA libraries for improved uniformity and affordability. *bioRxiv*, 10.1101/2025.03.24.644170.

3. Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.-D., Fernandez,J.P., Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*. *Nat. Methods*, 12, 982–988.

APPENDIX B. Supplementary Material for Chapter 3

B1. Figures

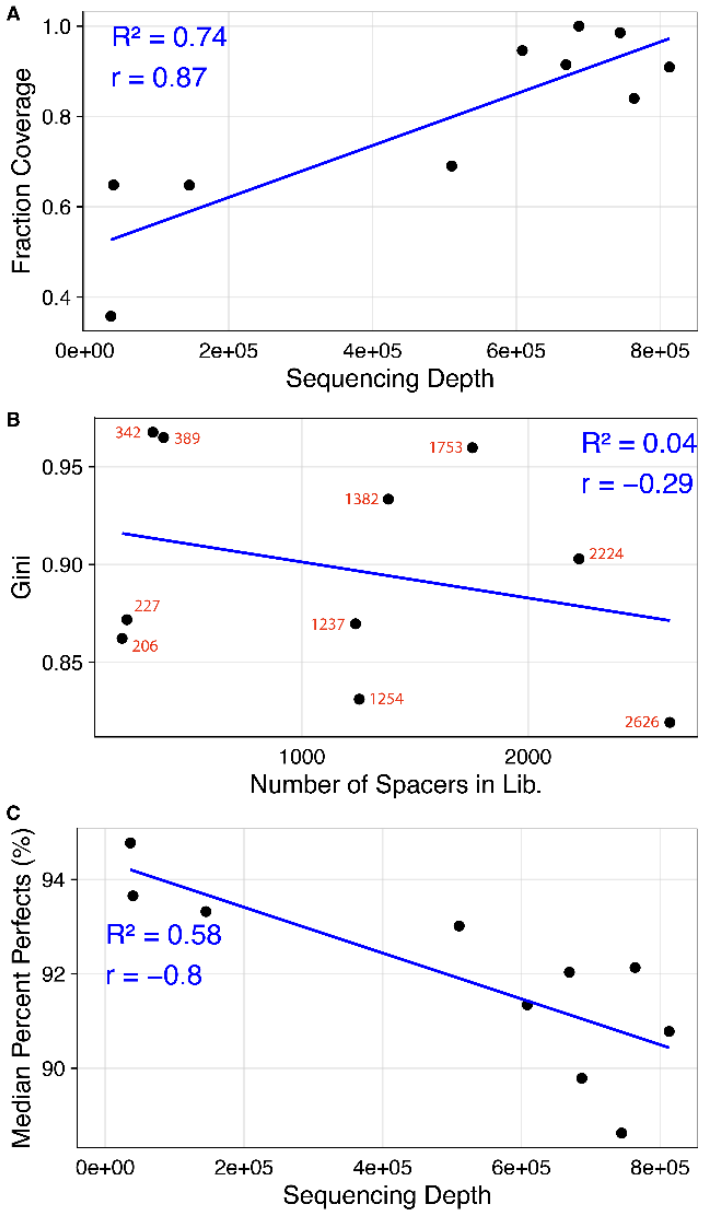


Figure 30. Trends in quality metrics for 10 microarray-derived sgRNA libraries.

A. Fraction of spacers observed out of expected (coverage) versus sequencing depth (R^2 = 0.74, r = 0.87). **B.** Gini Coefficients versus library scale (number of spacers in lib) (R^2 = 0.04, r = -0.29). The exact number of spacers per library are shown in red. **C.** Median percent of spacers with perfect or expected spacer sequences (median percent perfects) versus sequencing depth (R^2 = 0.71, r = -0.9).

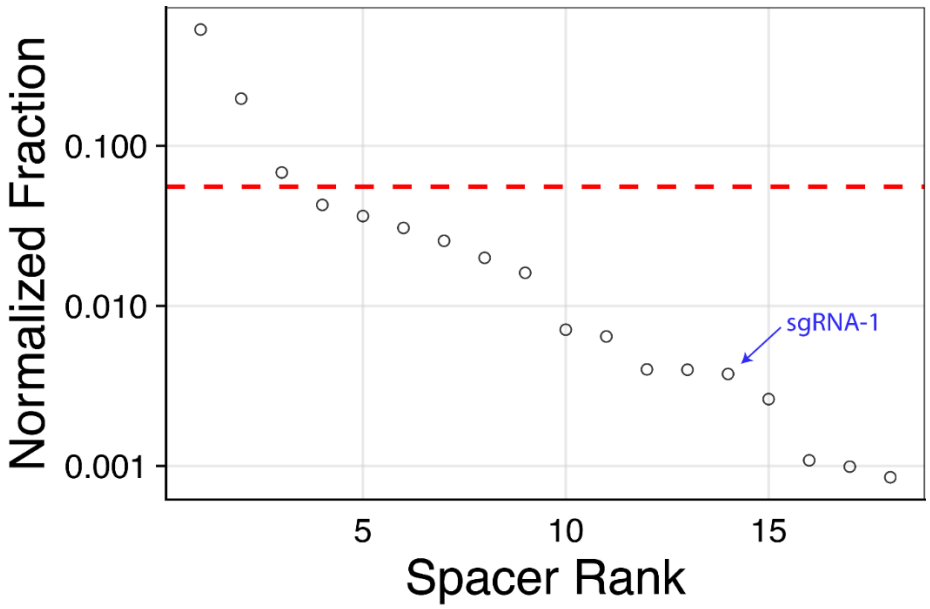


Figure 31. Distribution of 18 spacers transcribed as a single sgRNA library, shown by the normalized fraction of reads for perfect sequences (y-axis, log scale).

Each open circle represents an individual spacer, ranked by decreasing abundance. The dashed horizontal line indicates the expected read distribution under perfect uniformity.

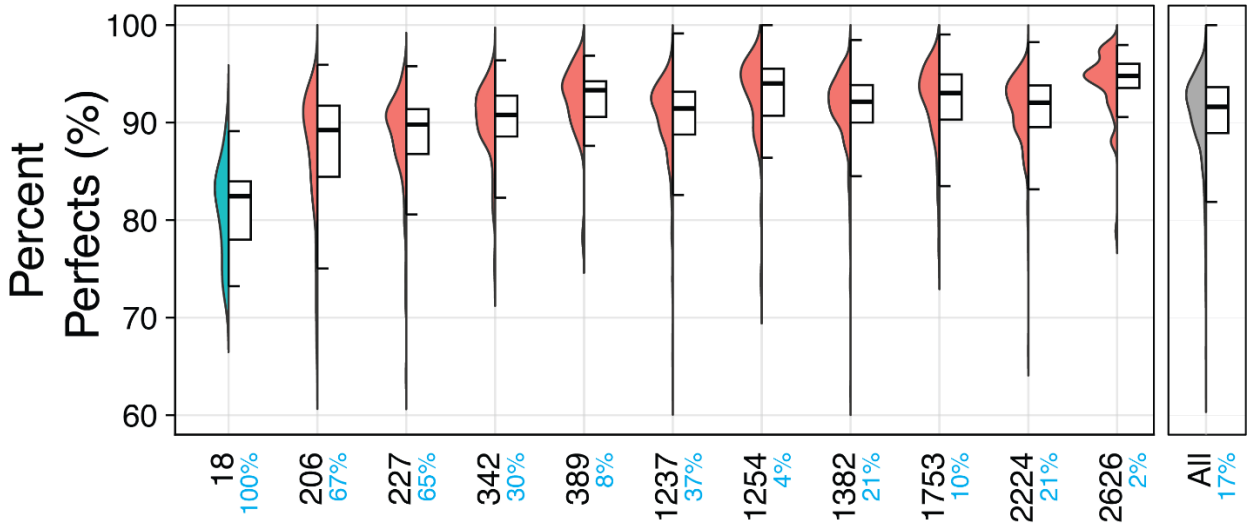


Figure 32. Comparison of percent perfect spacer sequences across sgRNA libraries.

Perfect spacer sequences correspond to the expected spacers programmed into each library. Each library is represented by a bifurcated plot: the left half shows a half-violin plot (distribution of percent perfects), and the right half displays a boxplot (median percent perfect per library). Only spacers with at least 100 reads in the RNA-seq data were included to ensure reliable analysis. The percentage of spacers meeting this threshold is indicated on the x-axis (blue text). The column-synthesized 18-plex sgRNA library is shown in teal, microarray-derived libraries in orange, and the overall median percent perfects across all 11 libraries is represented by the gray bar.

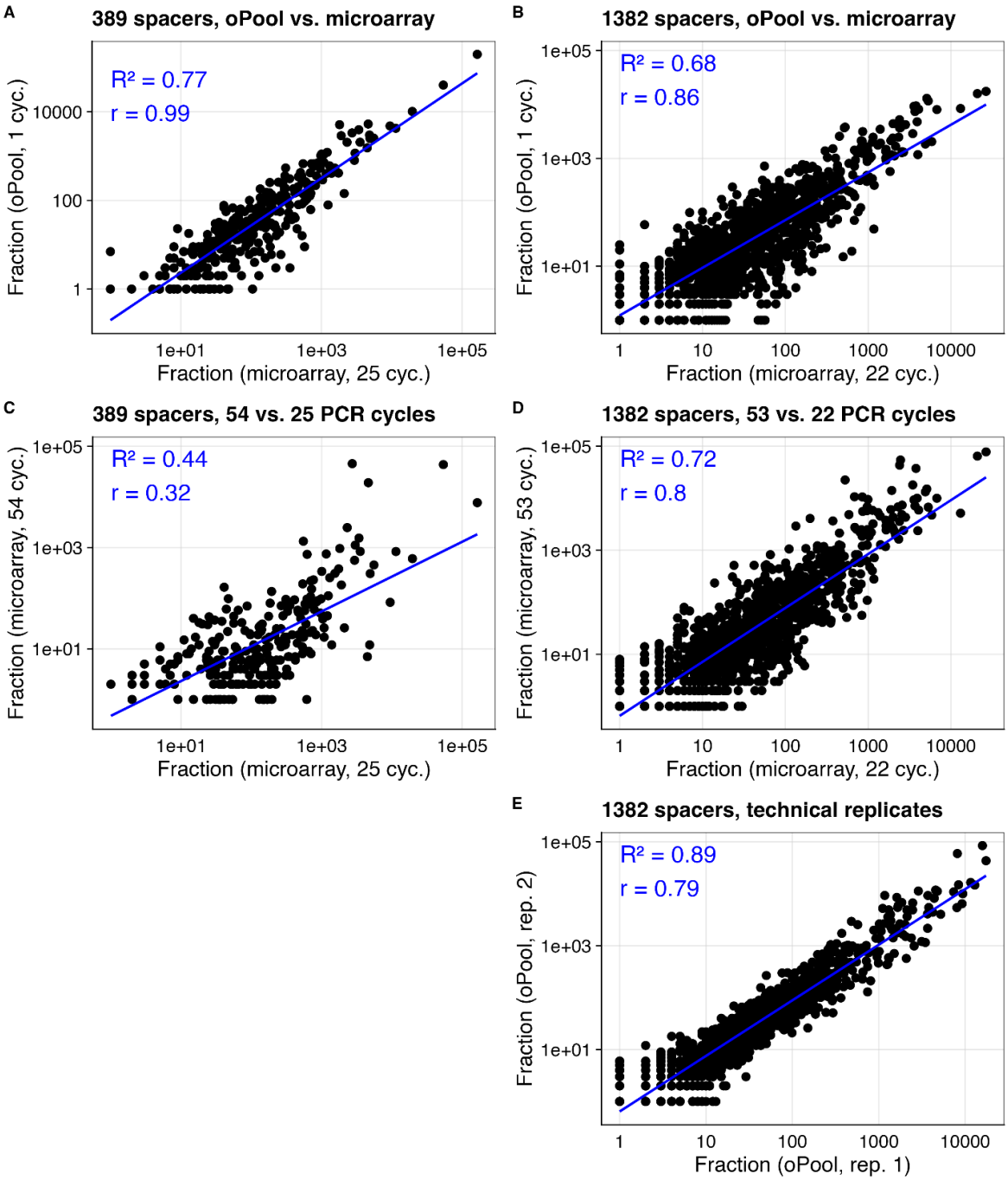


Figure 33. Pearson correlations of spacer reads for oPool- and microarray-derived sgRNA libraries prepared with varying PCR cycle numbers.

A. oPool-derived 389-plex sgRNA library spacer reads (1 cycle) versus a microarray-derived version prepared with 25 PCR cycles (R^2 = 0.77, r = 0.99). **B**. oPool-derived 1,382-plex sgRNA library spacer reads (1 cycle) versus a microarray-derived version prepared with 22 PCR cycles (R^2 = 0.68, r = 0.86). **C**. Microarray-derived 389-plex sgRNA library spacer reads (52 cycles) versus a reduced-PCR version prepared with 25 PCR cycles (R^2 = 0.44, r = 0.32). **D**. Microarray-derived 1,382-plex sgRNA library spacer reads (53 cycles) versus a reduced-PCR version prepared with 22 PCR cycles (R^2 = 0.72, r = 0.80). **E**. Comparison of oPool-derived 1,382-plex sgRNA library spacer reads between two replicates (R^2 = 0.89, r = 0.79

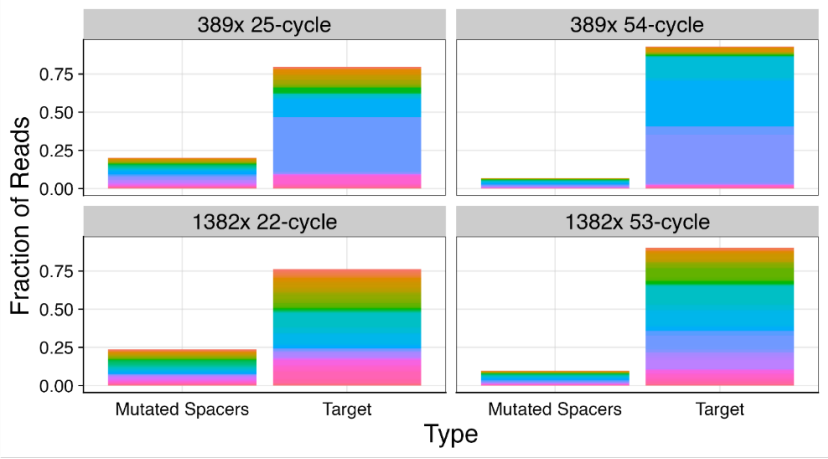


Figure 34. Effect of PCR cycle number on the fraction of mutant and target spacers in 389- and 1,382-plex microarray-derived sgRNA libraries.

Fraction of total reads, scaled to one, representing mutant spacers that deviate from the expected sequence compared to target spacers with the correct expected sequences. The top panel shows these metrics for the 389-plex sgRNA library prepared with reduced PCR using 25 cycles and excessive PCR using 54 cycles. The bottom panel shows the same for the 1,382-plex sgRNA library, comparing reduced PCR with 22 cycles to excessive PCR with 53 cycles. Each color within the rainbow pattern represents a unique spacer sequence, whether mutant or target.

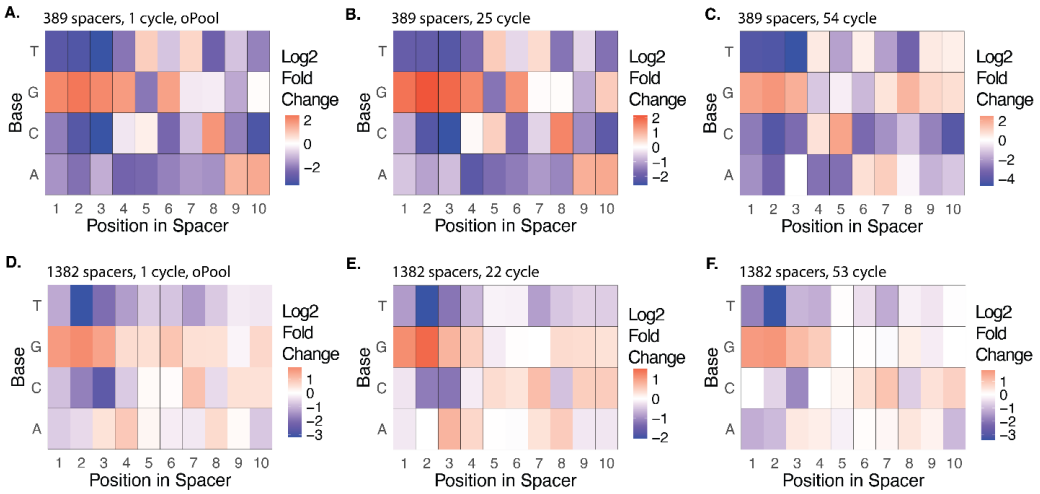


Figure 35. Influence of base composition on spacer abundance within 389- and 1,382-plex sgRNA libraries prepared with differing oligo sources and PCR cycle numbers.

Log_2 fold change (FC) of observed vs. expected spacer abundance across all nucleotide identities within the first 10 nucleotides of 20-nt spacers. "Observed" refers to the fraction of spacers identified from RNA-seq data that meet the sequence criteria, while "expected" represents the fraction assuming perfect uniformity across the population. Positive log_2 FC values (orange) within the heat maps indicate increased spacer abundance compared to expected, while negative values (blue) indicate decreased abundance. Zero change is shown in white. The log_2 fold change scale ranges from -4 to 2 for the 389-plex library and -3 to 1 for the 1,382-plex library. **A**. oPool-derived 389-plex spacer library prepared with one PCR cycle. **B**. Microarray-derived 389-plex spacer library prepared with reduced PCR cycles (25 total). **C**. Microarray-derived 389-plex spacer library prepared with excessive PCR cycles (54 total). **D**. oPool-derived 1,382-plex spacer library prepared with one PCR cycle. **E.** Microarray-derived 1,382-plex spacer library prepared with reduced PCR cycles (22 total). **F.** Microarray-derived 1,382-plex spacer library prepared with excessive PCR cycles (53 total).

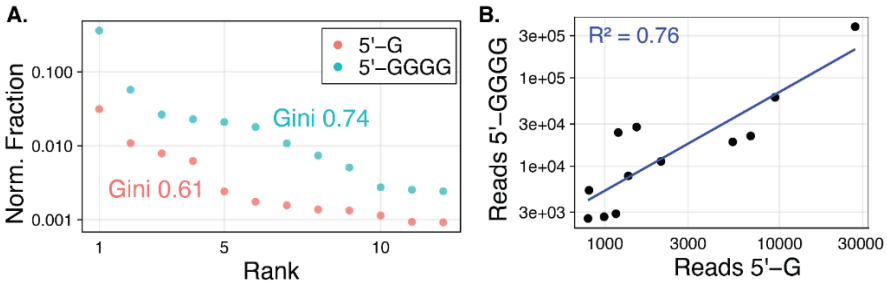


Figure 36. Distribution of 12 sgRNA spacers starting with a 5' G and 5' GGGG (12G and 12G4 libraries).

A. Comparison of spacer distribution uniformity between two 12-plex sgRNA libraries containing identical spacers, except that the spacers in one library start with a single 5′ guanine (5′ G, orange), while the spacers in the other library are padded with a 5′ guanine tetramer (5’ GGGG, teal). Normalized spacer abundance is shown, with each dot representing a unique spacer, ranked in descending order. Gini Coefficients (Gini), shown alongside each library's spacer distribution, measure the inequality in spacer representation between the two library types. **B**. Pearson correlation (R^2 = 0.76, r = 0.962) of spacer abundance between the two sgRNA libraries presented in **A**.

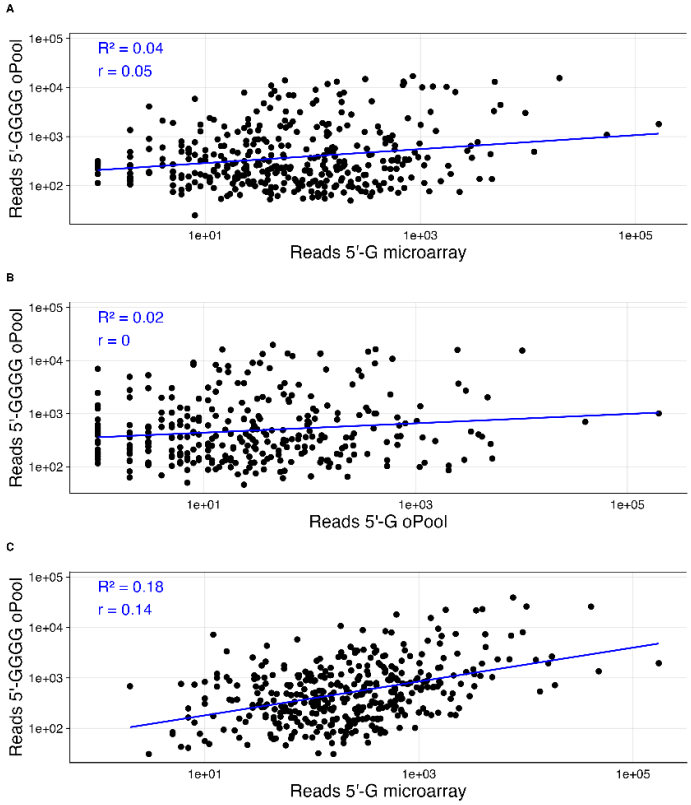


Figure 37. Pearson correlations of spacer reads for sgRNA libraries containing 389 spacers starting with 5' G and 5' GGGG (389G and 389G4 libraries).

A. oPool-derived sgRNA library (7 PCR cycles) with four guanines compared to a microarray-derived sgRNA library (25 PCR cycles) with a single guanine (R^2 = 0.04, r = 0.05). **B.** oPool-derived sgRNA library with four guanines (7 PCR cycles) compared to an oPool-derived sgRNA library with a single guanine (1 PCR cycle) (R^2 = 0.02, r = 0). **C.** Replicate of the condition shown in panel **A** using different sgRNA libraries (R^2 = 0.18, r = 0.14).

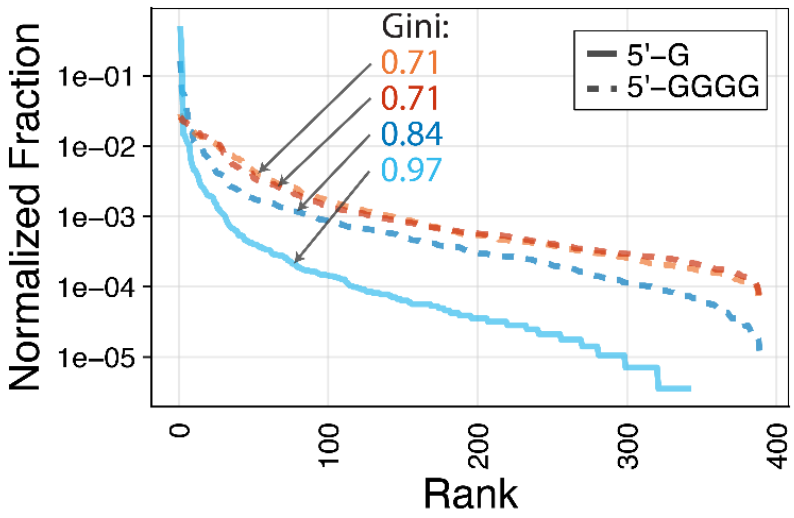


Figure 38. Gini Coefficients of sgRNA libraries containing 389 spacers starting with 5' G and 5' GGGG (389G and 389G4 libraries) transcribed with modified IVT conditions.

Normalized fraction of spacer reads relative to total reads for sgRNA libraries containing 389 spacers, either starting with 5’ G (solid line, n = 1) or 5’ GGGG (dashed lines, n = 3). Spacers are ranked by decreasing abundance. Each library, transcribed from 100 ng of template DNA in a 100 µL reaction volume, is represented by a distinct color. Gini Coefficients (Gini) are listed next to each library’s rank-ordered curve to quantify inequality in spacer representation across library types and scales.

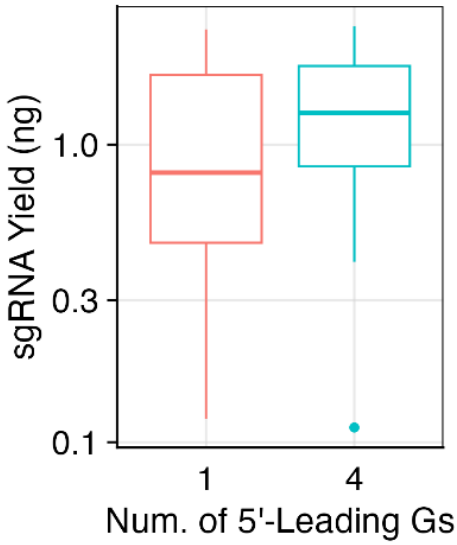


Figure 39. Comparison of sgRNA yields between libraries with 5’ G and 5’ GGGG spacers.

Comparison of sgRNA yields (µg) for sgRNA libraries containing spacers that start with 5’ G (median: 0.805 µg, n = 13) versus 5’ GGGG (median: 1.29 µg, n = 8). Differences in yield were not statistically significant (Wilcoxon rank-sum test, p = 0.5002).

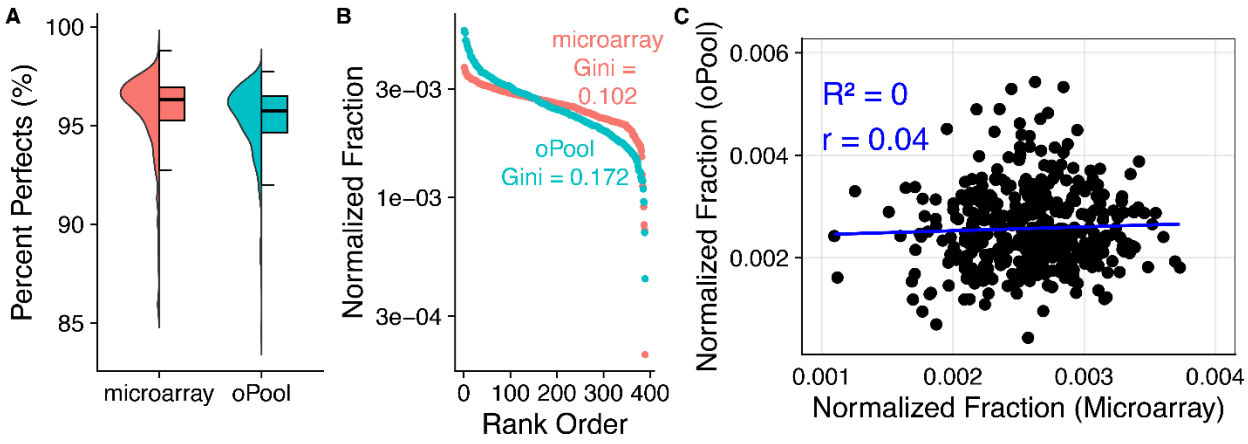


Figure 40. Comparison of quality metrics for 135 bp DNA libraries containing 389 spacers, prepared from microarray- and oPool-derived oligos

A. The percentage of perfect spacers, defined as spacers without mutations, was assessed in microarray-derived (coral, median: 96.30%) and oPool-derived (teal, median: 95.73%) DNA libraries. Each library is represented as a bifurcated plot: the left half shows a half violin plot (distribution of percent perfect spacers), while the right half displays a boxplot (median percent perfect for each library). Only spacers with at least 100 reads in the RNA-seq data were included to ensure reliable analysis. **B.** Normalized abundance (reads per spacer, relative to total reads) for the same libraries as shown in a. Spacers are ranked in descending order of abundance. Gini coefficients (Gini), listed next to each library's rank-ordered curve, indicate inequality in spacer distribution, showing a significant difference in distribution between libraries (Wilcoxon rank-sum test, $p = 3.97 \times 10^{-8}$). **C.** Pearson correlation analysis of normalized fraction of spacer reads between the microarray- versus oPool-derived DNA libraries ($R^2 = 0$, $r = 0.04$).

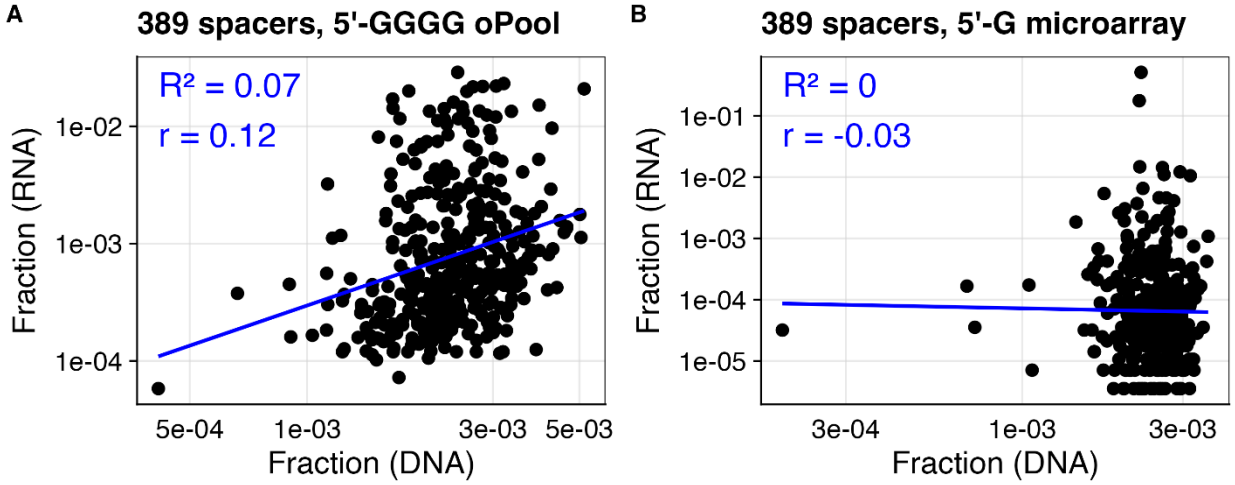


Figure 41. Pearson correlations of spacer reads between 135 bp DNA libraries containing 389 spacers and their corresponding IVT sgRNA libraries.

A. Comparison of an oPool-derived DNA library and its corresponding sgRNA library, transcribed using 100 ng input DNA in a 100 µL IVT reaction volume. Both libraries contain spacers that start with 5' GGGG ($R^2 = 0.07$, $r = 0.12$). **B.** Comparison of a microarray-derived DNA library and its corresponding sgRNA library, transcribed using 100 ng input DNA in a 20 µL IVT reaction volume. Both libraries contain spacers that start with 5' G ($R^2 = 0$, $r = -0.03$).

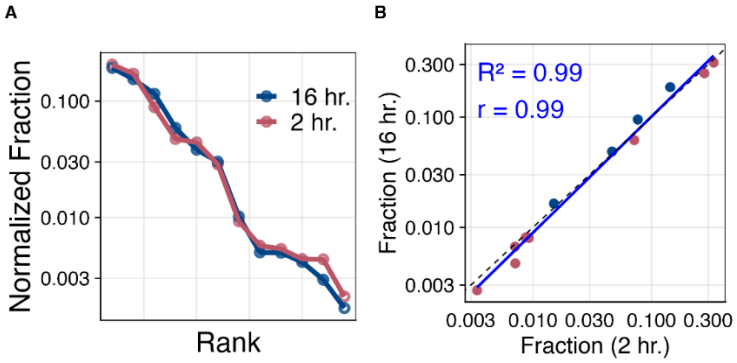


Figure 42. Comparison of incubation times for transcribing a 12-plex sgRNA library in emulsions.

A. Spacers ordered by decreasing normalized fraction of reads (rank) for libraries transcribed at 37°C for 2 hours (red) and 16 hours (blue). **B.** Pearson correlation analysis of the fraction of spacer reads for sgRNA libraries transcribed for 2 hours (red) and 16 hours (blue) (R^2 = 0.99, r = 0.99).

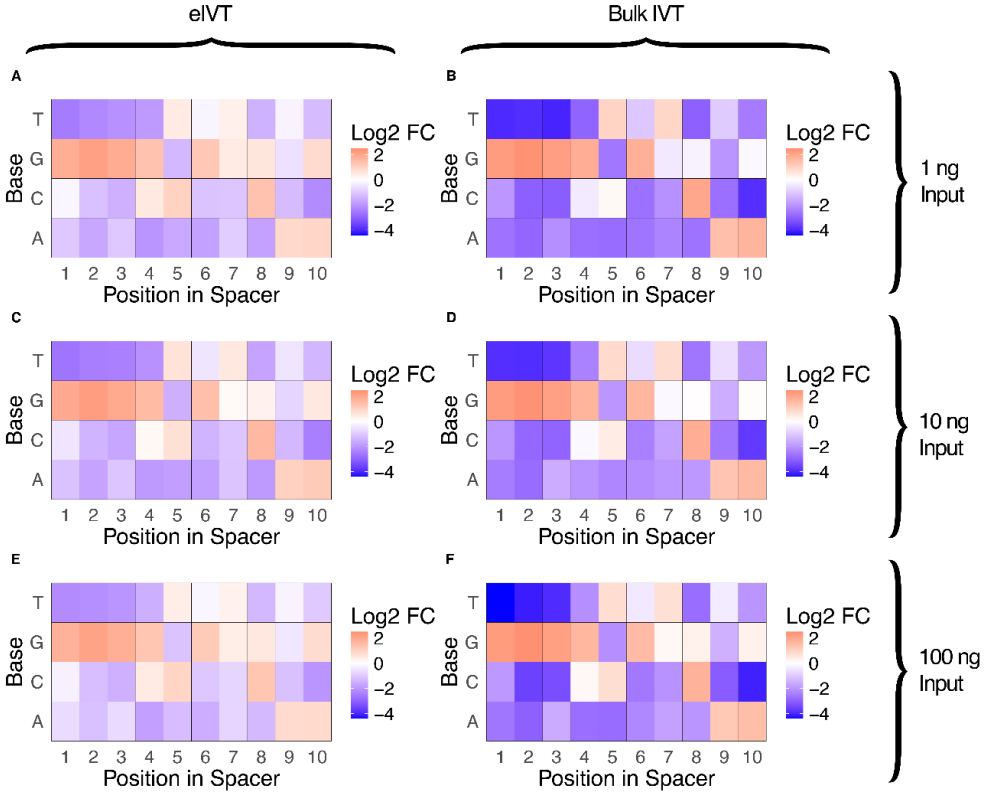


Figure 43. Effect of base composition on spacer abundance in 389-plex sgRNA libraries with 5' G (389G library), transcribed via bulk IVT or emulsion IVT (eIVT).

Log_2 fold change (FC) of observed vs. expected spacer abundance across all nucleotide identities within the first 10 nucleotides of 20-nt spacers. "Observed" refers to the fraction of spacers identified from RNA-seq data that meet the sequence criteria, while "expected" represents the fraction assuming perfect uniformity across the population. Positive log_2 FC values (orange) indicate increased spacer abundance compared to expected, while negative values (blue) indicate decreased abundance. Zero change is shown in white. The FC scale ranges from -4 to 2 across all heatmaps, which are as follows: **A.** eIVT with 1 ng input DNA, **B.** bulk IVT with 1 ng input DNA, **C.** eIVT with 10 ng input DNA, **D.** bulk IVT with 10 ng input DNA, **E.** eIVT with 100 ng input DNA, and **F.** bulk IVT with 100 ng input DNA.

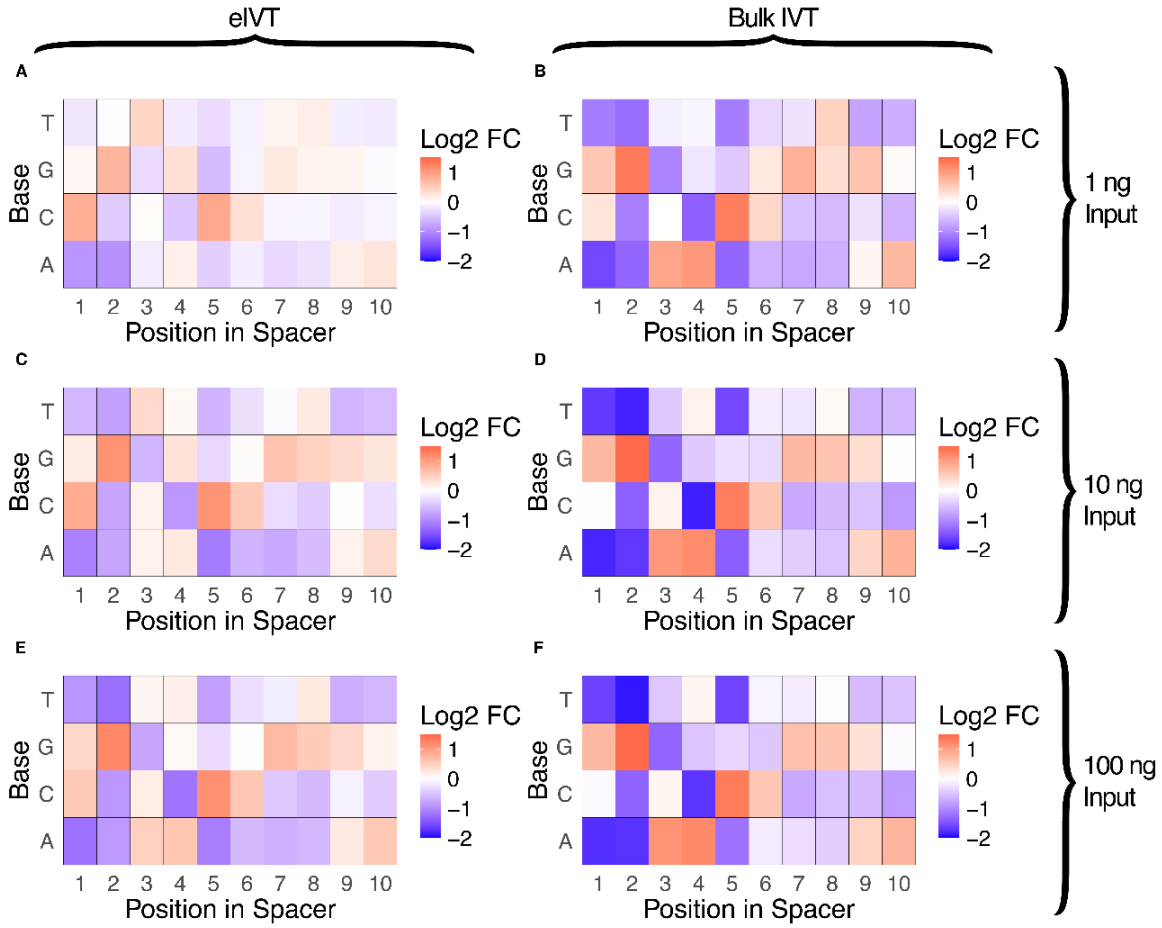


Figure 44. Effect of base composition on spacer abundance in 2,626-plex sgRNA libraries with 5' G, transcribed via bulk IVT or emulsion IVT (eIVT).

Log_2 fold change (FC) of observed vs. expected spacer abundance across all nucleotide identities within the first 10 nucleotides of 20-nt spacers. "Observed" refers to the fraction of spacers identified from RNA-seq data that meet the sequence criteria, while "expected" represents the fraction assuming perfect uniformity across the population. Positive log_2 FC values (orange) indicate increased spacer abundance compared to expected, while negative values (blue) indicate decreased abundance. Zero change is shown in white. The FC scale ranges from -2 to 2 across all heatmaps, which are as follows: **A.** eIVT with 1 ng input DNA, **B.** bulk IVT with 1 ng input DNA, **C.** eIVT with 10 ng input DNA, **D.** bulk IVT with 10 ng input DNA, **E.** eIVT with 100 ng input DNA, and **F.** bulk IVT with 100 ng input DNA.

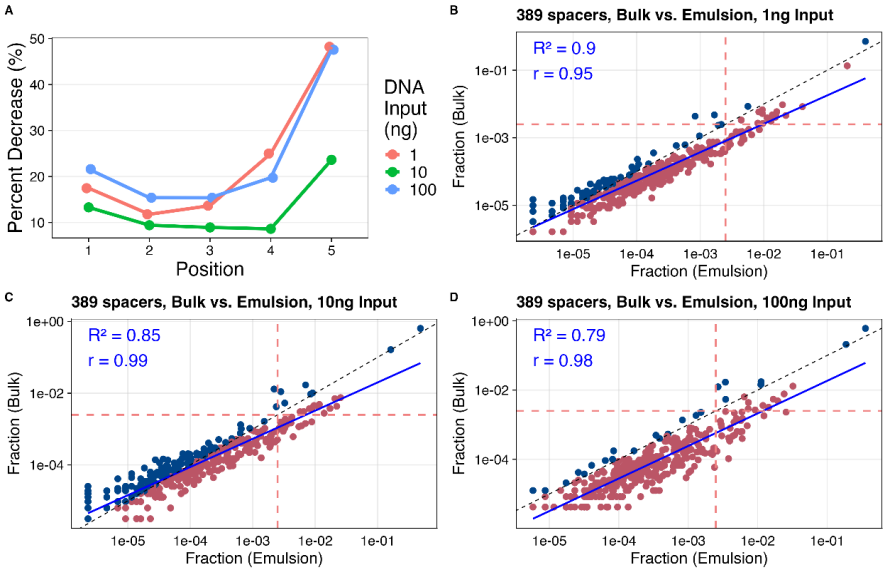


Figure 45. Spacer representation changes in 389-plex sgRNA libraries with 5' G, transcribed via eIVT with varying DNA input amounts.

A. The percentage decrease in $\log_2$ fold change (FC) of observed vs. expected abundance reflects the reduction in spacer position-dependent guanine representation in eIVT compared to bulk IVT, calculated using the formula (FC_bulk - FC_eIVT) / FC_bulk. This analysis was performed for sgRNA libraries prepared with 1 ng (coral), 10 ng (green), and 100 ng (aqua) input DNA (panels **B-D).** Pearson correlation analysis of spacer representation across input DNA amounts: **B.** 1 ng (R^2 = 0.9, r = 0.95), **C.** 10 ng (R^2 = 0.85, r = 0.99), and **D.** 100 ng (R^2 = 0.79, r = 0.98).The solid blue line represents the linear fit of spacer reads, while the dashed black line indicates the unity fit (1:1 relationship) between bulk IVT and eIVT. Blue dots represent overrepresented spacers that decrease in eIVT, while red dots represent underrepresented spacers that increase in eIVT. The red dashed lines represent the median read fraction per spacer from the DNA library input, which had a highly uniform spacer distribution and served as the template for sgRNA transcription.

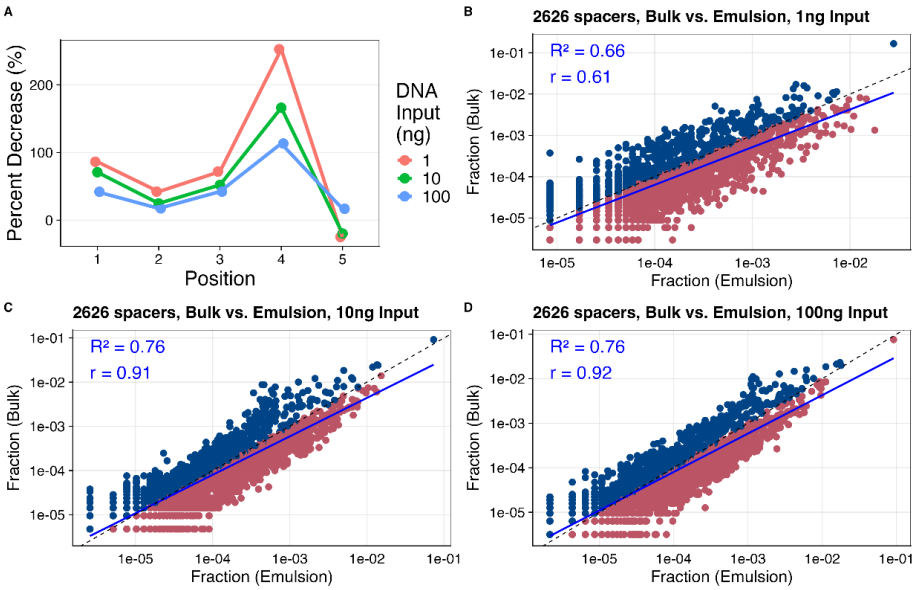
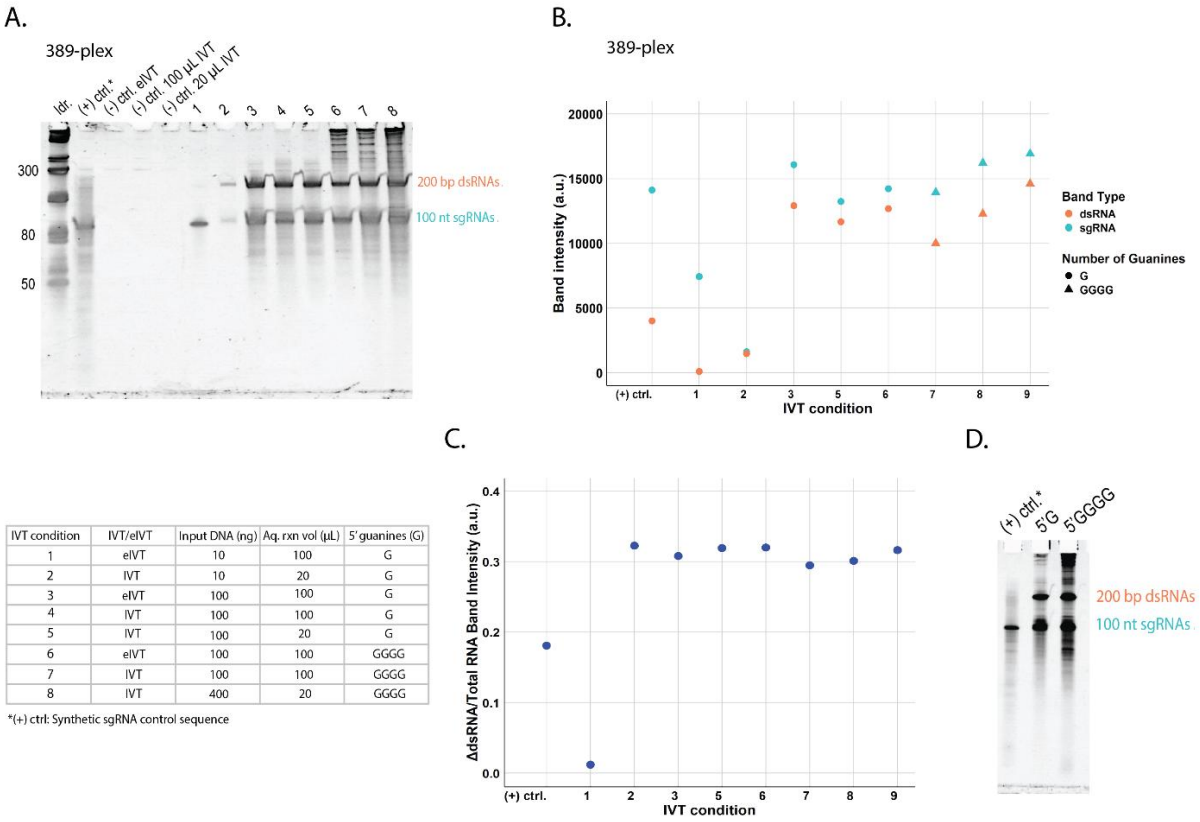


Figure 46. Spacer representation changes in 2,626-plex sgRNA libraries with 5' G, transcribed via eIVT with varying DNA input amounts.

A. The percentage decrease in $\log_2$ fold change (FC) of observed vs. expected abundance reflects the reduction in spacer position-dependent guanine representation in eIVT compared to bulk IVT, calculated using the formula (FC_bulk - FC_eIVT) / FC_bulk. This analysis was performed for sgRNA libraries prepared with 1 ng (coral), 10 ng (green), and 100 ng (aqua) input DNA (panels **B-D).** Pearson correlation analysis of spacer representation across input DNA amounts: **B.** 1 ng (R^2 = 0.66, r = 0.61), **C.** 10 ng (R^2 = 0.76, r = 0.91), and **D.** 100 ng (R^2 = 0.76, r = 0.92).The solid blue line represents the linear fit of spacer reads, while the dashed black line indicates the unity fit (1:1 relationship) between bulk IVT and eIVT. Blue dots represent overrepresented spacers that decrease in eIVT, while red dots represent underrepresented spacers that increase in eIVT.



IVT condition	IVT/eIVT	Input DNA (ng)	Aq. rxn vol (µL)	5' guanines (G)
1	eIVT	10	100	G
2	IVT	10	20	G
3	eIVT	100	100	G
4	IVT	100	100	G
5	IVT	100	20	G
6	eIVT	100	100	GGGG
7	IVT	100	100	GGGG
8	IVT	400	20	GGGG

*(+) ctrl: Synthetic sgRNA control sequence

Figure 47. Assessment of single-stranded sgRNA (100 nt) products and high molecular weight (200 bp+) RNA byproducts between 5' G and 5' GGGG sgRNA libraries transcribed in emulsions or bulk.

A. Tris-borate-EDTA-urea denaturing gel electrophoresis (10% polyacrylamide) with SYBR Gold staining was used to analyze 389-plex sgRNA libraries transcribed under different conditions. These conditions included bulk IVT (IVT), emulsion IVT (eIVT), varying DNA input DNA amounts, and spacers starting with either 5' G or a 5' GGGG. The aqueous reaction volume (aq. rxn vol) was also varied, representing the total aqueous IVT volume emulsified in eIVT or the total reaction volume in standard IVT. **B.** ImageJ (1) quantification of gel band intensities corresponding to sgRNA (100 nt) and dsRNA (200 bp) from panel **A**, plotted in arbitrary units (A.U.). **C.** Ratios of dsRNA band intensities were calculated relative to the total RNA, defined as the sum of sgRNA and dsRNA intensities, for the synthetic sgRNA control and various IVT conditions. **D.** Tris-borate-EDTA-urea denaturing gel electrophoresis (10% polyacrylamide) with SYBR Gold staining was used to assess transcribed RNA products from libraries containing 12 identical spacers. These spacers started with either 5' G or a 5' GGGG and were compared to a synthetic control sgRNA sequence (100 nt).

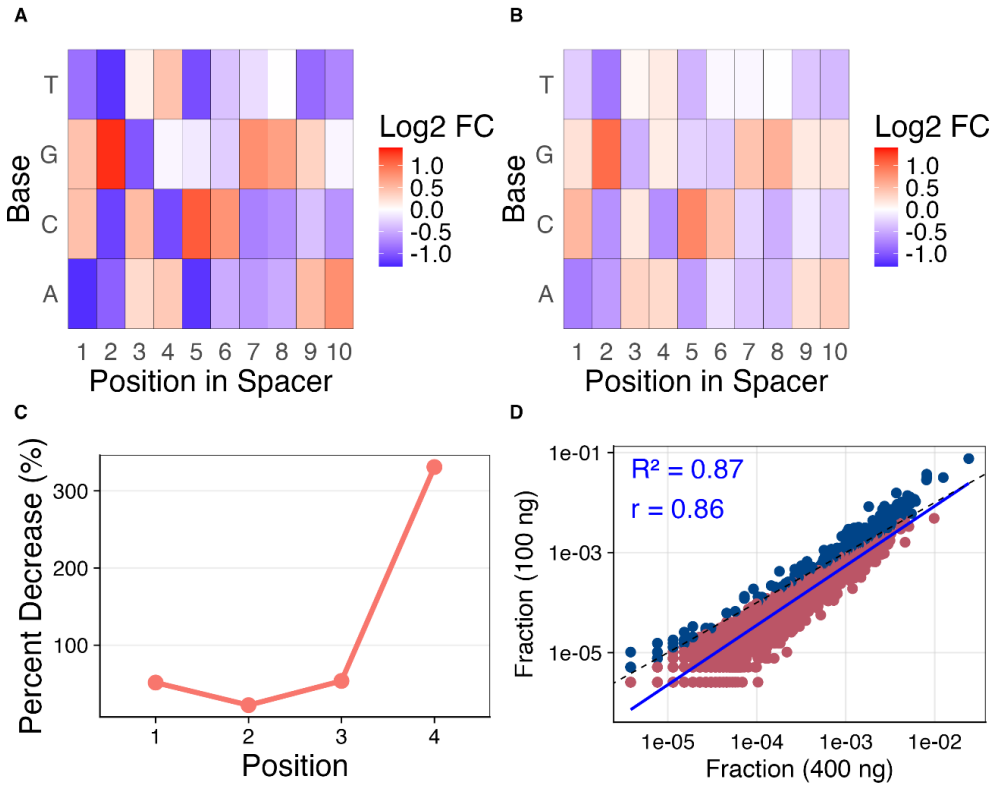


Figure 48. Spacer representation changes in 2,626-plex sgRNA libraries with 5' G transcribed with 100 ng or 400 ng input DNA in a 20 µL IVT reaction volume.

Panels **A-B**. Log_2 fold change (FC) of observed vs. expected spacer abundance across all nucleotide identities within the first 10 nucleotides of 20-nt spacers for sgRNA libraries transcribed with **A.** 100 ng and **B.** 400 ng input DNA. "Observed" refers to the fraction of spacers identified from RNA-seq data, while "expected" represents the fraction assuming perfect uniformity across the population. Positive log_2 FC values (orange) indicate increased spacer abundance compared to expected, while negative values (blue) indicate decreased abundance. **C.** Percent decrease in log_2 FC of observed vs. expected abundance, calculated separately for IVT with 100 ng and 400 ng input DNA using the formula (FC_100ng - FC_400ng) / FC_100ng. The difference (ΔFC) between the two conditions was plotted to show changes in position-dependent guanine representation across spacers. **D.** Pearson correlation analysis of the fraction of spacer reads for sgRNA libraries transcribed with 100 ng and 400 ng input DNA (R^2 = 0.87, r = 0.86). The solid blue line represents the linear fit of spacer reads, while the dashed black line indicates the unity fit (1:1 relationship) between bulk IVT and eIVT. Blue dots represent overrepresented spacers that decrease in eIVT, while red dots represent underrepresented spacers that increase in eIVT.

B2. Tables

Table 3. Pearson correlations among various sgRNA library metrics for 10 microarray-derived sgRNA libraries.

The Pearson correlation coefficient (*r*) is shown for each pairwise comparison. Selected variable name abbreviations include median percent perfects (medpperf), total PCR cycles (cyctotal), and sequencing depth (depth).

Variable 1	Variable 2	Pearson coefficient (*r*)
Gini	Scale	-0.28674
Gini	depth	0.385077
Gini	medpperf	-0.06484
Gini	cyctotal	0.022516
Gini	coverage	0.211416
Scale	depth	-0.35841
Scale	medpperf	0.694729
Scale	cyctotal	-0.55549
Scale	coverage	-0.57874
depth	medpperf	-0.80169
depth	cyctotal	0.063687
depth	coverage	0.865335
medpperf	cyctotal	-0.39103
medpperf	coverage	-0.89704
cyctotal	coverage	0.271306

Table 4. Kolmogorov–Smirnov (KS) statistical analysis of spacer distribution changes between eIVT and bulk IVT sgRNA libraries.

KS test analysis was applied to normalized sgRNA sequencing data for 389- and 2,626-plex sgRNA libraries, each containing spacers starting with 5’ G, transcribed in bulk or emulsions (see Fig. 13C in 3.4.6 eIVT Enhances sgRNA Library Uniformity). Each sgRNA library was transcribed once (n = 1) with varying reaction scales and DNA input amounts. The KS test was used to assess the maximum discrepancy between the cumulative distribution functions of sgRNAs transcribed in emulsions compared to their corresponding bulk IVT control libraries. A significant KS statistic ($p < 0.05$) indicates a substantial shift in sgRNA distribution between the two conditions, suggesting changes in library composition or uniformity.

DNA Amount (ng)	Library Scale	D value	p value
1	389	0.20352	3.53E-07
10	389	0.097695	5.40E-02
100	389	0.36263	2.20E-16
1	2626	0.22022	2.20E-16
10	2626	0.098369	7.35E-10
100	2626	0.081096	3.19E-07

Table 5. Kolmogorov–Smirnov (KS) analysis of spacer distribution changes in sgRNA libraries transcribed with bulk IVT using 1 ng, 10 ng, or 100 ng input DNA.

KS test analysis of normalized sgRNA sequencing data for libraries containing 389 and 2,626 spacers, all starting with 5' G. The KS test was used to evaluate the maximum discrepancy between the cumulative distribution functions of sgRNAs transcribed with 1-400 ng of input DNA, with IVT reaction volumes of 20 µL or 100 µL. A significant KS statistic ($p < 0.05$) indicates a substantial shift in sgRNA distribution between the two conditions, suggesting changes in library composition or uniformity.

Comparing DNA Input Amount (in ng)					
Total IVT Volume (µL)	Library Scale	DNA Amount 1	DNA Amount 2	D value	p value
20	2626	1	10	0.064611	0.000105
20	2626	10	100	0.08759	3.96E-08
20	2626	100	400	0.24361	2.20E-16
20	2626	1	400	0.21657	2.20E-16
100	2626	1	10	0.085496	1.59E-07
100	2626	10	100	0.06849	4.18E-05
100	2626	1	100	0.05971	0.000464
100	389	1	10	0.1675	4.69E-05
100	389	10	100	0.2289	3.30E-08
100	389	1	100	0.060506	0.5585

20	389	1	10	0.088837	0.112
20	389	10	100	0.18941	2.49E-06
20	389	100	400	0.23818	6.19E-10

Table 6. The 18-target single-stranded DNA oligonucleotide sequences for an 18-plex sgRNA target library.

Each sequence includes a forward primer site, T7 promoter sequence, 20 nt spacer, and reverse primer site.

Target oligo sequence name	Sequence
Pool_002_sgRNA_1	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGCGTTCAATTTAGATTAGGGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_2	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGCTATTGTTTCTAAAATGCGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_3	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGAGTGAGTTTAATCTGGGCGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_4	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGGATGATAATGGCCCTGTTGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_5	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGAAACTTAGCGTAATATAAGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_6	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGTTCTGTGCTTAACAATGAGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_7	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGAGAGCGCCTCATTCATGTGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_8	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGTATTCGTATATGTGCGAAGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_9	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGACGGTGTCGCATTTGGAAGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_10	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGTTCGGATCATGTCAACTTGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_11	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGATCTTAATGAGTATTGATGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_12	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGTAGATAGACCTTTCACCGGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC

Pool_002_sgRNA_13	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGAAGAGGTTTGCGATCTGCGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_14	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGTTATGAAAATTGTATGTCGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_15	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGATCCGATTACGACAAATAGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_16	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGTCTAATACCGCACTCTCTGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_17	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGAACGCGTCCCGTTCATCGGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC
Pool_002_sgRNA_18	GGGTCACGCGTAGGATTCTAATACGACTCACTATAGGTGACTCTGCTTAGCCTATGTTTTAGAGCTAGAGGAGACCGTTCTCGTGTGGCTGCGGAAC

Table 7. Primer pairs and conditions for subpooling and bulk-amplifying ten target oligo ssDNA libraries of various scales from microarray-synthesized chip9 oligonucleotides.

Annealing temperatures (Tm) for each subpooling primer pair, using Kapa HiFi polymerase, were obtained by gradient qPCR. The table also includes the number of PCR cycles corresponding to the qPCR amplification plateau for subpooling and bulk amplification, before and after cycle number optimization. Each target oligo subpool was *in vitro* transcribed to produce ten sgRNA libraries.

Library scale	Subpool	Forward Primer Name	Forward Primer Sequence	Reverse Primer Name	Reverse Primer Sequence	Tm* (°C)	Subpool Cycles (Original)	Subpool Cycles (Optimized)	Bulk Amp Cycles (Original)	Bulk Amp Cycles (Optimized)
342	1	skpp15-1-F filt15-1651	GGGTCACGCGTAGGA	skpp15-1-R filt15-1181	GTTCCGCAGCCACAC	51	24	N/A	17	N/A
206	4	skpp15-4-F filt15-656	GGTCGAGCCGGAACT	skpp15-4-R filt15-740	GGATGCGCACCCAGA	52	26	N/A	16	N/A
1,382	6	skpp15-6-F	CGCAGGGTCCAGAGT	skpp15-6-R	GTTCGCGCGAAGGAA	50	22	15	16	12
1,254	7	skpp15-7-F filt15-1855	AGTGACCCGTCCCTG	skpp15-7-R filt15-757	AGTCGACCTCTGCCC	51	24	N/A	17	N/A
1,753	9	skpp15-9-F filt15-453	CGATCGTGCCCACCT	skpp15-9-R filt15-1189	GTGCGGGCTCCAACT	58	24	N/A	17	N/A

1,237	13	skpp15-13-F	GGGTTCGAGCGGGAG	skpp15-13-R	TAGCGCGCAGAGAGG	56	20	N/A	16	N/A
389	23	skpp15-23-F filt15-577	AGCTGCTACACCGCC	skpp15-23-R filt15-1596	GCGCGATGGTCACAG	53	22	17	17	12
227	26	skpp15-26-F	GCGGCACCACAAACT	skpp15-26-R	CGTGGCCTCTGTCCT	52	23	N/A	17	N/A
2,626	27	skpp15-27-F	ACCTTCACGCGTCCC	skpp15-27-R	GCCCACCGACTCCAC	51	20	16	18	7
2,223	28	skpp15-28-F	GACTGCGGCGTTGGT	skpp15-28-R	TACGCCCGGGACAGA	52	22	N/A	16	N/A

*Annealing temperatures listed correspond to Kapa HiFi Polymerase

Table 8. Additional nucleic acid sequences that were not used for subpooling amplification or 5' RACE in this study.

Sequence name	Sequence	Description
ds_sgRNA_scaffold_BsaI_YG	GAGAACGGTCTCCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTT	Duplexed sgRNA scaffold DNA sequence from IDT
SgRNA_GGA_oligo_FWD_NV	GGGTCACGCGTAGGATTCTAATACG	Forward primer for amplifying sgRNA templates (GGA products) for IVT
sgRNA_GGA_oligo_REV_NV	AAAAGCACCGACTCGGTGCCAC	Reverse primer for amplifying sgRNA templates (GGA products) for IVT Also used for RSPE of S2 and S4 oPools (1 PCR cycle)
S. pyogenes control sgRNA sequence	mC*mA*mU* rCrCrU rCrGrG rCrArC rCrGrU rCrArC rCrCrG rUrUrU rUrArG rArGrC rUrArG rArArA rUrArG rCrArA rGrUrU rArArA rArUrA rArGrG rCrUrA rGrUrC rCrGrU rUrArU rCrArA rCrUrU rGrArA rArArA rGrUrG rGrCrA rCrCrG rArGrU rCrGrG rUrGrC mU*mU*mU* rU	Control sgRNA sequence from the EnGen® sgRNA Synthesis kit

Table 9. Primer pairs and conditions are provided for subpooling four 98 nt oligo libraries of varying sizes, with either one or three guanines downstream of the T7 promoter.

These libraries were subpooled from the e13sgRNA oPool, which contains 389 unique target oligos. Each 98 nt oligo includes a library-specific forward primer site, T7 promoter, 20 nt spacer, and reverse primer site. Annealing temperatures (Tm) for each subpooling primer pair, using Kapa HiFi polymerase, were obtained by gradient qPCR. It also includes the number of PCR cycles corresponding to the qPCR amplification plateau for subpooling. Each target oligo subpool was *in vitro* transcribed to produce four sgRNA libraries.

Subpool name (plexity, number of guanines)	Forward Primer Name	Forward Primer Sequence	Reverse Primer Name	Reverse Primer Sequence	Tm* (°C)	Subpool PCR Cycles
12G	skpp15-6-F	CGCAGGGTCCAGAGT	skpp15-6-R	GTTCGCGCGAAGGAA	50	11
12G4	skpp15-13-F	GGGTTCGAGCGGGAG	skpp15-13-R	TAGCGCGCAGAGAGG	56	12
60G4	skpp15-23-F filt15-577	AGCTGCTACACCGCC	skpp15-23-R filt15-1596	GCGCGATGGTCACAG	53	10
389G4	skpp15-27-F	ACCTTCACGCGTCCC	skpp15-27-R	GCCCACCGACTCCAC	51	7

Table 10. Sequences used for the 5' RACE protocol in RNA-seq of sgRNA libraries, including RT primers, template-switching oligos, and gene-specific primers.

Sequence Name	Sequence	Description
RT_primer_YG	AGCATATATCCCGGTCTGGA NNNNNNNNNNNNNNNNN AAAAGCACCGA	Reverse transcription DNA primer
TSO_YG	GCTAATCATTGCAAGCAGTGGTATCAACGCAGAGT ACATrGrGrG	Template Switching Oligo (TSO)
TSO_primer_YG	CATTGCAAGCAGTGGTATCAAC	DNA primer
gene_specific_primer_YG	AGCATATATCCCGGTCTGGA	DNA primer
sgRNA_cDNA_gibson_FWD_YG	TCCAGACCGGGATATATGCTtatgcggtgtgaaataccgcac	Forward DNA primer
sgRNA_cDNA_gibson_REV_YG	CACTGCTTGCAATGATTAGCcggtatcattgcagcactgg	Reverse DNA primer
sgRNA_cDNA_1kbp_ext_FWD_YG	AGAGTTCTTGAAGTGGTGGC	Forward DNA primer

sgRNA_cDNA_1kbp_ext_REV_YG	GTGTGGAATTGTGAGCGGAT	Reverse DNA primer

B3. References

1. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.