ML-GUIDED MINING OF AN EXTENSIVELY VALIDATED SCFV LIBRARY FOR OPEN-SOURCE ENZYMES FOR DIAGNOSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cost and IP barriers limit access to high-performance enzyme start/stop modifiers such as hot-start systems that suppress premature activity during reaction setup. We combine an accessible 10^{10} human scFv phage-display library with activity-linked screening to engineer open, recombinant enzyme regulators. Using low-cost fluorescence workflows (in-house dye synthesis with 67-86× cost reduction), we identify scFv inhibitors that convert standard polymerases into hot-start formulations. In head-to-head benchmarking against commercial hot-start enzymes, scFv-regulated polymerases achieve commercial-grade suppression during setup with heat-triggered recovery and robust amplification. To scale beyond individual hits, we outline a data-centric pipeline: NGS-tracked selections yielding more than 10^4 binder/non-binder sequences per target and deep mutational scanning of lead scFvs (5,000-20,000 variants) to map CDR-level inhibitory fitness landscapes for predictive design. We highlight prospective extensions to ligases, restriction enzymes, and CRISPR-Cas systems.

1 INTRODUCTION

Hot-start polymerases improve PCR reliability by suppressing activity during reaction setup, reducing mis-priming and primer-dimer formation that erode sensitivity in diagnostics. Current high-performance hot-start systems remain proprietary and enzyme-specific, limiting reproducible engineering and local manufacturing. We pursue an open alternative using single-chain variable fragments (scFvs): compact (25 kDa), bacterially expressible binders engineered for reversible inhibition. Leveraging the McCafferty human scFv library (10^{10} diversity), we build an experimental-to-computational pipeline for open enzyme regulation, validating it on DNA polymerases with paired binding and functional benchmarking that naturally generates ML-ready datasets.

2 APPROACH

Conserved, surface-exposed epitopes within thermostable Family A and B polymerases were identified by sequence/structure analysis and used as peptide baits for phage display selection from the McCafferty library. Enriched pools were screened by ELISA against full-length enzymes (Bst, Bsu, KOD, DeepVent, phi29), yielding four lead scFvs (A7B, D8B, F4A, A1A) expressed in *E. coli* and purified. Binding specificity was profiled by ELISA across polymerases. Hot-start function was quantified using an optimized isothermal EvaGreen fluorescence assay that reports polymerase activity without thermal cycling. In-house EvaGreen synthesis reduced dye cost 67-86×.

3 RESULTS

ELISA screening revealed family-structured binding: A7B/D8B preferentially recognized Family A enzymes (Bst/Bsu), while F4A/A1A recognized Family B enzymes (KOD/DeepVent/phi29) (Figure 1C). Functional benchmarking showed scFv-mediated temporal control comparable to commercial hot-start polymerases: under setup-temperature pre-incubation, scFv-regulated polymerases suppressed baseline activity similarly to commercial formulations, while non-hot-start enzymes

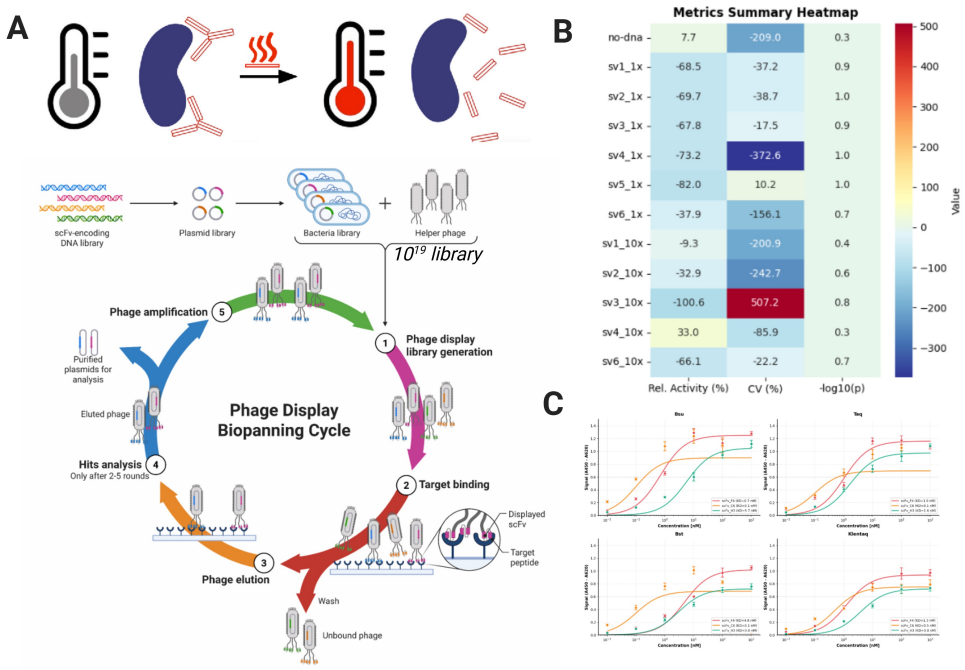

Figure 1: **scFv discovery and polymerase binding characterization.** (A) Schematic of scFv-enabled hot-start regulation and phage display biopanning from 10^{10} library. (B) Activity assay metrics heatmap. (C) ELISA dose-response curves for lead scFvs.

showed immediate activity (Figure 1B). Following activation, scFv-regulated polymerases recovered activity and enabled robust amplification.

ML-Ready Data Generation The wet-lab pipeline produces ML supervision: ELISA provides binding targets, while the isothermal assay provides functional inhibition labels. To scale beyond individual leads, we implement phage display coupled to deep sequencing to label more than 10,000 binder and non-binder sequences per target from enrichment/depletion trajectories. For validated inhibitors, deep mutational scanning (5,000-20,000 variants per scFv) will map inhibitory fitness landscapes, supporting models that predict inhibitory function from sequence and propose optimized CDRs for validation.

4 Discussion and Outlook

These results establish scFv-based hot-start regulation as an open, engineerable capability and define a scalable route to predictive inhibitor design. Next targets prioritize temporal control in ligation/assembly, restriction digestion, and CRISPR-Cas systems. All sequences, protocols, and trained models will be released under open licenses to support distributed reagent development for diagnostics in resource-limited settings.

References

[1] D. J. Schofield et al. Application of phage display to antibody generation. *Genome Biology*, 8:R254, 2007.

[2] J. McCafferty et al. Phage antibodies displaying antibody variable domains. *Nature*, 348:552-554, 1990.

[3] D. M. Fowler and S. Fields. Deep mutational scanning. *Nature Methods*, 11:801-807, 2014.

[4] Q. Chou et al. Prevention of pre-PCR mis-priming improves amplifications. *Nucleic Acids Research*, 20:1717-1723, 1992.