Multiplexed, trackable CRISPR-based genome engineering for optimization of heterologous protein expression and activity in *E. coli*

Introduction

Heterologous protein expression underpins many advancements in modern molecular biology, including protein engineering, production of bioindustrial enzymes and therapeutics, and structure-function discovery. Heterologous proteins are commonly expressed from plasmids to facilitate cloning and subsequent genetic manipulation. However, plasmid-based expression has many drawbacks, including reduced stability, population drift, and limited control over copy number variations. Combining CRISPR editing tools (1, 2) and advancements in gene sequence engineering (3, 4), heterologous gene integration into chromosomal DNA has become more accessible, offering genomic stability with consistent, high expression (5). These advances facilitate in-depth examination of sequence-function relationships and systematic control of protein activity from the genome.

Thorough interrogation of gene sequence variants remains difficult, even across a single locus. To do this efficiently and cost-effectively, libraries of genetic edits must be designed and generated in multiplex, which necessitates trackability of designs to connect phenotype to genotype. An innovative genome engineering technology - the Onyx™ Digital Genome Engineering Platform - performs genome-wide and trackable CRISPR editing at scale in an automated benchtop device. This approach significantly reduces the time and resources spent on constructing variant libraries, as demonstrated in this app note. Here we used Inscripta's open-sourced MAD7™ CRISPR nuclease to integrate a green fluorescent protein (GFP) gene into the *E. coli* MG1655 chromosome. We then used the Onyx platform to design and generate GFP site-saturation libraries, with 723 designs targeting functional residues intended to modify fluorescence intensity and spectral characteristics (Figure 1A). Additionally, we generated a library of 8,284 distinct promoter and ribosome binding site (RBS) sequences extracted from the *E. coli* genome, inserted it 5' of the GFP gene and screened this library to quantify the strength of native *E. coli* promoters and RBS sites in a massively parallel experiment.

Saturation mutagenesis library design and construction

The protein sequence of superfolder GFP (sfGFP) derived from *Aequorea victoria* (6) was codon-optimized to enhance A/T richness, remove repetitive and common restriction enzyme sequences, and increase the number of accessible MAD7 editing sites. In addition to the refactored gene sequence, the integration construct contained a strong constitutive J23102 promoter, a strong synthetic RBS, and a T0 terminator (Figure 1B). This construct was inserted into the intergenic region between *ygcE* and *queE* genes using MAD7 nuclease, lambda red recombination machinery, and a double-stranded, linear DNA template containing 60-nucleotide homology arm sequences. The edited genome was sequenced, and chromosomal sfGFP expression was assessed to establish the baseline expression range.

Using an approach common to protein engineering (6), known evolutionary trees were used to identify residues of interest for modifying the sfGFP spectral characteristics. We used the FPbase database (https://www.fpbase.org/) to identify mutational hotspots by aligning variants within the GFP clade that exhibit excitation/emission wavelength shifts or changes in other fluorescence characteristics, such as lifetime and quantum yield (7). These hotspots are typically positioned near the fluorophore forming site or structurally important regions of the protein (Figure 1C). A saturation mutagenesis library was then generated for the hotspot residues, with a total of 723 variants. The entire library was created in just a few minutes using the InscriptaDesigner™ (Designer) software. The edited cell

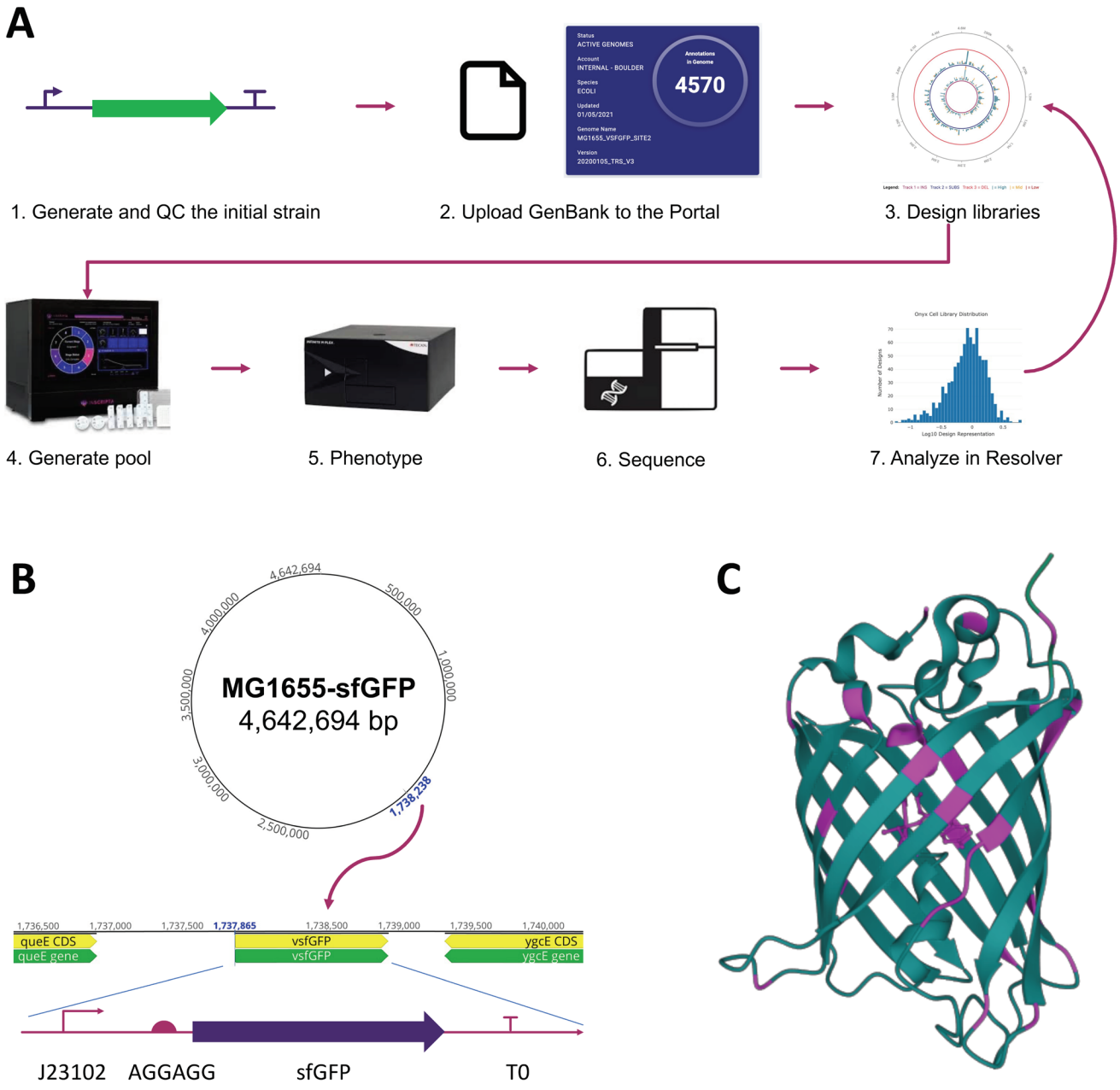

Figure 1. A) Protein library design, construction, phenotyping and genotyping workflow. B) Chromosomal integration construct, with sfGFP, J23102 promoter, RBS and T0 termination elements inserted into the genome as one cassette; C) sfGFP structural model: mutational hotspot residues indicated in magenta.

population was generated using the Onyx workflow. Edit metrics and performance were tracked using the InscriptaResolver™ (Resolver) software, which automates the library assessment and isolate identification analyses. The library showed 66% edited cell population and was subsequently prepared for phenotyping.

Individual colonies from the edited cell libraries were screened, with a total of 192 isolates selected for spectral analysis. After 24 hours of growth in 96-deep-well plates in terrific broth under dual antibiotic selection, cells were washed in minimal media and diluted 1:50 from the original volume in minimal media and arrayed in optical reader plates. Fluorescence intensity was measured for the 192 isolates using a Tecan plate reader (Infinite M Nano) at three excitation and emission wavelengths based on blue florescent protein (BFP, 380nm:445nm), green fluorescent protein (GFP, 480nm:508nm), and yellow fluorescent protein (YFP, 510nm:530nm) peaks. A subset of the edit library resulted in disrupted fluorescence on all channels, indicating residues essential for function. The remaining screened colonies demonstrated a range of fluorescence intensities along the GFP and YFP axes (Figure 2A). Colonies showing fluorescence profiles distinct from the wild-type were additionally assayed by collecting excitation-emission spectra to identify any peak shifts.

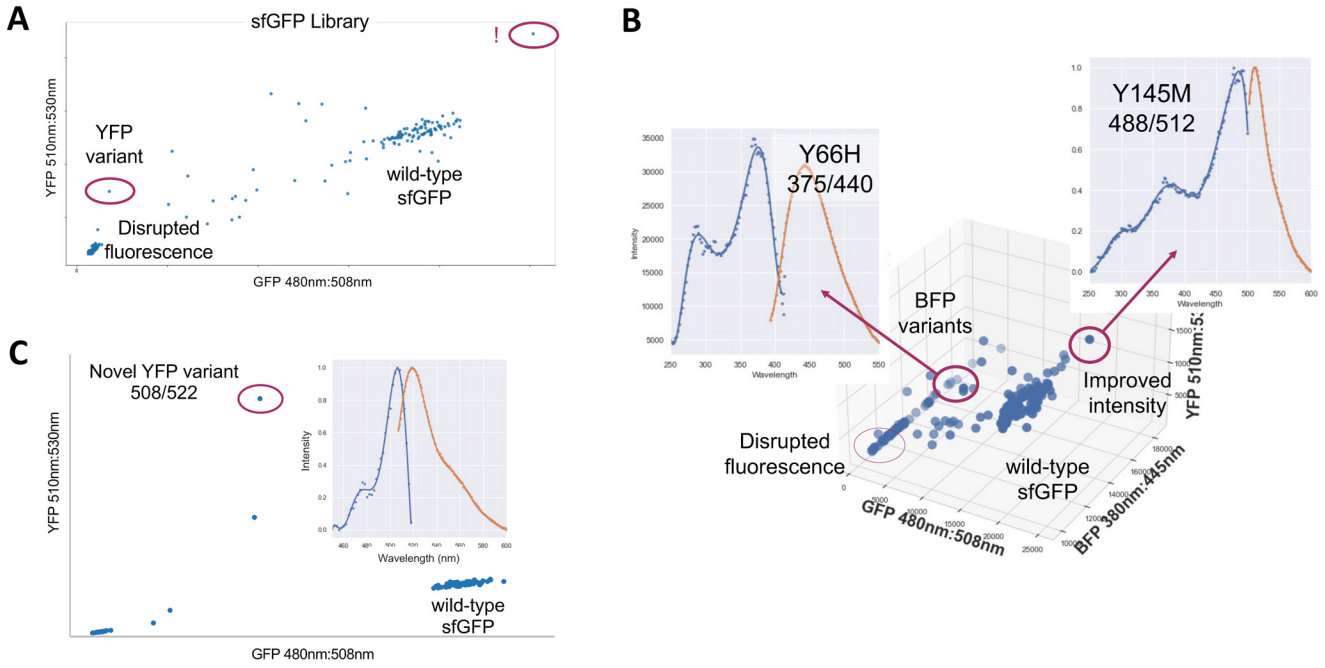

Figure 2. A) Fluorescence intensity of the screened variants along the GFP and YFP excitation-emission spectra compared to the wild-type sfGFP and YFP. B) Fluorescence profile of the screened library, including variants displaying improved intensity and yellow and blue spectral shift. C) Identification of a novel YFP variant with significantly improved intensity obtained after the second editing cycle based on the identified highly fluorescent sfGFPF$_{145M}$ variant.

One variant showed a 1.5-fold increase in intensity compared to wild-type sfGFP and several displayed emission spectrum shift toward YFP and BFP. These variants were subjected to the edit isolate identification analysis workflow in Resolver to link the phenotype back to the genotype. The analysis identified the F145M edit being responsible for increased fluorescence intensity (Figure 2B). This position has been previously implicated in improving the folding kinetics that contribute to faster maturation of sfGFP (8). A previously known variant Y66H that demonstrated a shift to the BFP spectrum was also identified in the screen. It is notable that Y66 is part of the chromophore, formed by

the modified tripeptide Thr65–Tyr66–Gly67 (9). Residue Y66 specifically is responsible for absorption at 470 nm; substituting it for histidine, another aromatic amino acid, does not disrupt fluorescence but results in a blue shift in excitation and emission wavelengths (10).

The highly fluorescent variant sfGFPF$_{145M}$ was propagated for a second round of editing to further diversify the spectral properties of this variant. The selected isolate was subjected to a curing process to remove the editing plasmids in order to enable subsequent editing. The second editing cycle was completed using the same site saturation library, resulting in stacked double-edit variants. We screened 96 of these colonies and identified two novel variants with significantly improved yellow fluorescence intensity that exhibited a yellow shift in both excitation and emission wavelength (Figure 2C).

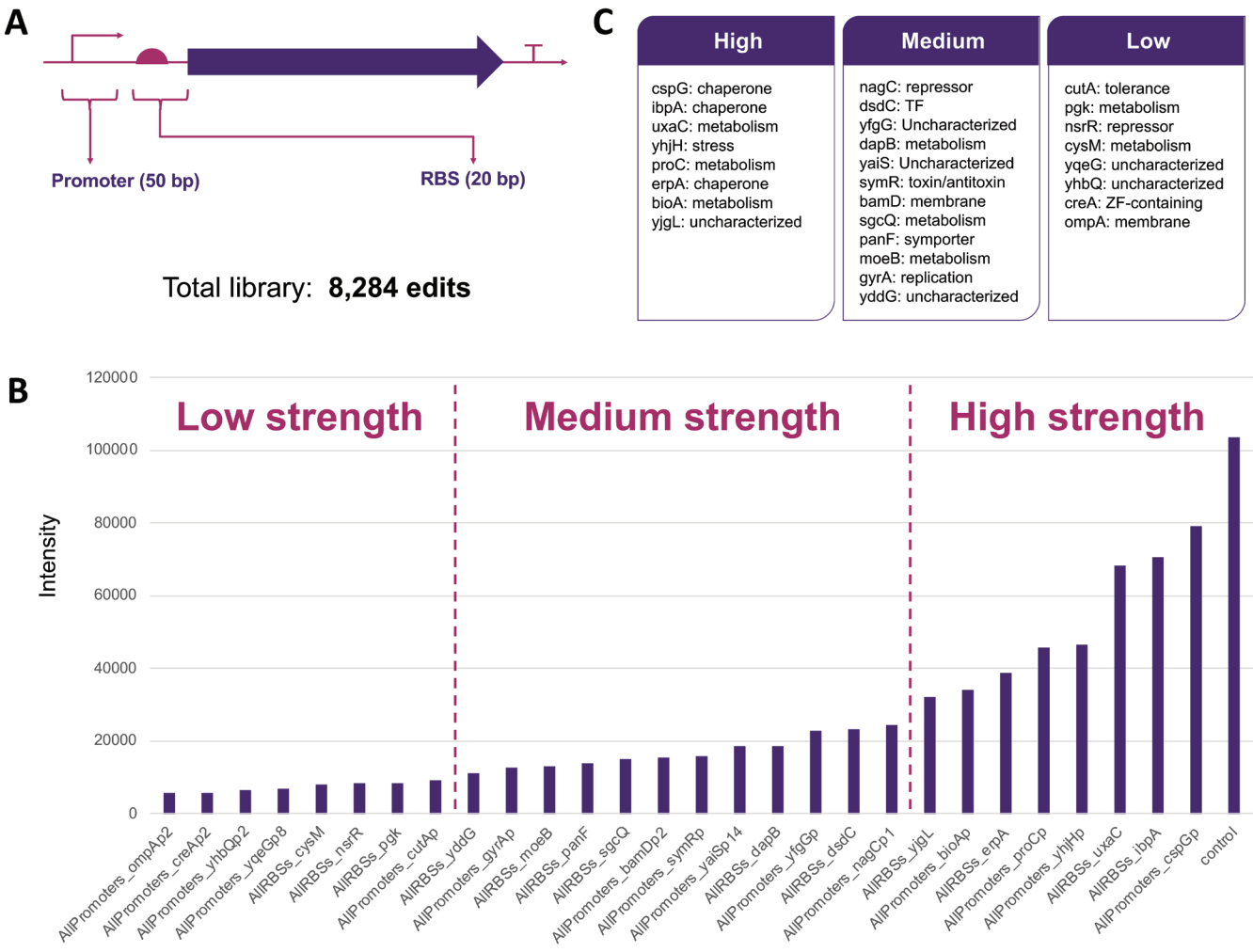

Figure 3. A) Promoter and RBS library design. Promoter sequences 50bp prior to each transcription start site were included (based on TSS annotations from EcoCyc). The RBS sequences were taken 20bp prior to every coding sequence. B) Regulatory elements grouped into categories based on the measured expression. C) Annotation of screened promoters and RBS sites, organized by strength and function

Promoter-RBS substitution library probes genome-wide promoter strengths

Leveraging the library variant sfGFPF$_{145M}$ displaying significantly improved intensity, we next set out to design a method for investigating the expression range of native *E. coli* promoters and RBS sequences. Using the Designer software, we first extracted the promoter sequences of every transcription start site in the genome and designed a comprehensive substitution library against the original J23102 promoter (Figure 3A). Similarly, we extracted all native RBS sequences by building a list of 20-nt sequences 5' of the start codon of each coding sequence, substituting these in place of the original RBS and spacer sequence 5' of sfGFP$_{F145M}$. The complete library consisted of 8,284 promoter and RBS sequences. The library was constructed using the Onyx workflow and assessed using the Onyx Genotyping Assays.

The library metrics were analyzed in Resolver and used to inform the screening depth of the population. The selected isolates showed a very broad intensity range; however, the highest intensity among the selected isolates was observed with the original J23102 + RBS combination (Figure 3B). The promoter and RBS libraries were analyzed and ranked, revealing that the highest promoter strengths are associated with diverse gene categories, including genes involved in metabolism, stress response, chaperones and some uncharacterized genes (Figure 3C). The entire workflow was carried out in a span of two weeks and allowed us to catalog and organize a large number of different promoters and RBS sequences. This data can potentially help uncover feedback mechanisms, transcription responses to small molecules and stress conditions, and generally improve annotation of native genomic regulatory elements.

Conclusion

This work demonstrates the use of the Onyx platform for interrogation and optimization of heterologous proteins in *E. coli*. Compared to traditional methods for protein engineering, such as random mutagenesis or targeted site-directed mutagenesis approaches, the Onyx technology dramatically increases the throughput of editing and allows testing of many different hypotheses simultaneously in one experiment. There is no more trade-off between covering a large scope of sequence positions and making comprehensive edits. As demonstrated here, this approach allows researchers to discover new protein functions and improve activity by changing protein residues directly in the chromosomally integrated gene in a fast and easy workflow. Additionally, the Onyx platform enabled us to probe an unprecedented range of genomic promoters and RBS sites and quantify their strength in a high-throughput, massively parallel experiment. This method can be used to characterize native regulatory elements, discover feedback mechanisms, investigate responses to chemical inducers and environmental conditions, and to fine-tune chromosomal expression of heterologous proteins in *E. coli*. These types of deep studies can be carried out in a matter of weeks with Onyx, presenting a significant advantage over the current methods.

REFERENCES

1. Pyne, M. E., Moo-Young, M., Chung, D. A., & Chou, C. P. (2015). Coupling the CRISPR/Cas9 system with Lambda Red recombineering enables simplified chromosomal gene replacement in Escherichia coli. Applied and Environmental Microbiology, 81(15), 5103-5114.
2. Garst, A.D., Bassalo, M.C., Pines, G., Lynch, S.A., Halweg-Edwards, A.L., Liu, R., Liang, L., Wang, Z., Zeitoun, R., Alexander, W.G., et al. (2017). Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat. Biotechnology 35, 48–55.
3. Hossain, A., Lopez, E., Halper, S. M., Cetnar, D. P., Reis, A. C., Strickland, D., ... & Salis, H. M. (2020). Automated design of thousands of nonrepetitive parts for engineering stable genetic systems. Nature biotechnology, 38(12), 1466-1475.
4. Burgess-Brown, N. A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., & Gileadi, O. (2008). Codon optimization can improve expression of human genes in *Escherichia coli:* A multi-gene study. Protein expression and purification, 59(1), 94-102.
5. Bassalo, M. C., Garst, A. D., Halweg-Edwards, A. L., Grau, W. C., Domaille, D. W., Mutalik, V. K., ... & Gill, R. T. (2016). Rapid and efficient one-step metabolic pathway integration in *E. coli*. ACS synthetic biology, 5(7), 561-568.
6. Li, C., Zhang, R., Wang, J., Wilson, L. M., & Yan, Y. (2020). Protein engineering for improving and diversifying natural product biosynthesis. Trends in biotechnology 36.7, 729-744.
7. Kashimoto, R., et al. (2021). Expansion and Diversification of Fluorescent Protein Genes in Fifteen Acropora Species during the Evolution of Acroporid Corals. Genes 12.3, 397.
8. Pédelacq, J.-D., et al. (2006). Engineering and characterization of a superfolder green fluorescent protein." Nature biotechnology 24.1, 79-88.
9. Tsien, R. Y. (1998). The green fluorescent protein. Annual review of biochemistry, 67(1), 509-544.
10. Heim, R., Prasher, D. C., & Tsien, R. Y. (1994). "Wavelength mutations and posttranslational autoxidation of green fluorescent protein." Proceedings of the National Academy of Sciences 91.26, 12501-12504.

Learn more at **INSCRIPTA.COM**