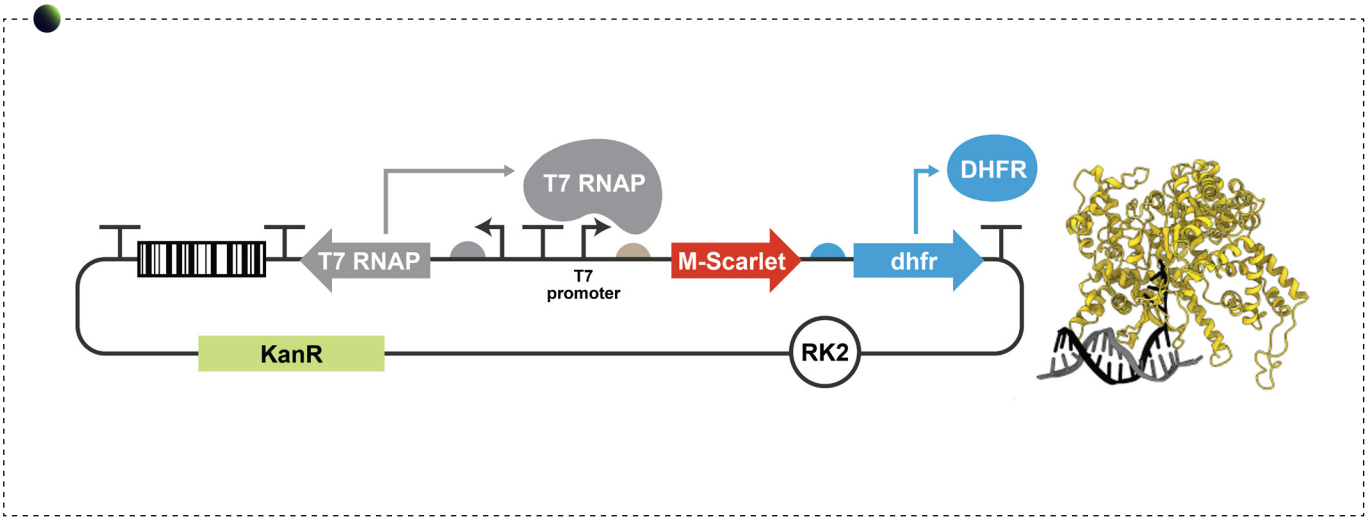
GROQ-seq Platform Expansion: Design of growth-coupled measurements of T7 RNA Polymerase



A proposal for onboarding T7 RNA polymerase to the protein sequence-to-function measurement platform.

- Uses a gene circuit to tie T7 RNA polymerase activity to bacterial cell growth.
- Calibration variants span the dynamic range of activity, enabling quantitative function measurements via **GROQ-seq**.
- The first dataset will include measurements of all single point amino acid substitutions and a library of recombined homologs.

DOI: DOI: 10.5281/zenodo.15025331

Contributors

Align to Innovate:

Dana Cortade - Technical Project Manager, Open Datasets Initiative
Peter Kelly - Co-Founder & Head of Science, Open Datasets Initiative

Proposal Leaders:

Erika DeBenedictis, The Francis Crick Institute

Additional Proposal Authors:

Justin Booth, The Francis Crick Institute
Ragnor Comerford, The Francis Crick Institute
Lily Nematollahi, The Francis Crick Institute

Reviewers:

Bryan C. Dickinson, Department of Chemistry & Chan Zuckerberg Biohub, University of Chicago
Philip Romero, Duke University
Ariel Tennenhouse, Weizmann Institute of Science

DOI: DOI: 10.5281/zenodo.15025331

Overview

T7 RNA polymerase (T7 RNAP) is a cornerstone of synthetic biology and is widely used for high-specificity transcription. However, despite its widespread use, predicting how its sequence variations influence function remains challenging, limiting its full potential in bioengineering applications.

We propose to apply growth-based assays to systematically investigate the sequence-function relationship of T7 RNAP and develop predictive models. We will onboard a T7 RNAP circuit to the newly named growth-based quantitative sequencing (**GROQ-seq**) platform using a set of calibration variants that span the desired dynamic range.

We will use this assay to study the relationship between polymerase sequence and promoter specificity, generating a dataset that will enable precise protein engineering for synthetic biology applications such as mRNA therapeutics and biosensors.

DOI: DOI: 10.5281/zenodo.15025331

Significance and Impact

RNA polymerases (RNAPs) form the integral bridge between DNA sequence and protein expression, implementing the sophisticated demands of transcriptional control that link environment to gene expression. In most organisms, the bulk of transcription is performed by multi-subunit RNAP complexes, a feature which allows many points of modulation to facilitate this complexity. Single-subunit RNA polymerases (ssRNAPs) form a fascinating exception to this rule, as they contain all the machinery needed to bind DNA, initiate transcription, polymerize RNA, and ensure accuracy with a single polypeptide sequence. This makes them a unique microcosm for protein modeling, as they involve DNA, RNA, and small-molecule binding, enzymatic catalysis, and sophisticated allosteric interactions between the enzyme and its substrates—all within a single protein sequence that can be readily diversified and measured using a simple reporter architecture. We propose linking the activity of an emblematic ssRNAP—that of the bacteriophage T7—to cell growth, enabling quantitative measurements of the sequence-to-fitness relationship of T7 RNAP and its promoter.

From an engineering perspective, T7 RNAP's capacity to act as a signal transducer and amplifier, along with its remarkable processivity in RNA synthesis, makes it a particularly attractive target. It forms the basis for a wide variety of synthetic biology projects, such as *in vitro* mRNA production[1], biosensor creation[2], and gene-specific continuous mutagenesis[3]. The ability to accurately predict how mutations affect critical parameters of its function would be highly valuable across diverse synthetic biology disciplines. Due to its reasonably well-studied mutational landscape, this system benefits from direct comparison to reported benchmarks, enabling particularly high confidence in future measurements[4].

To achieve this, we propose applying the recently named growth-based quantitative sequencing (**GROQ-seq**)[5] data acquisition platform to measure T7 RNAP function, beginning with an initial design task in the first full-scale library.

We aim to accurately design T7 RNAP variants with defined promoter specificity. Highly processive and promoter-specific single-subunit RNAPs like T7 RNAP are rare in nature. Designing a library of robustly orthogonal RNAPs with independently tunable activity levels would provide a valuable tool for bioengineers and enable the design of polymerases for non-canonical promoters or even non-promoter DNA sequences.

Beyond this initial application, the data generated by this work will be exceptionally valuable for protein modeling more broadly. Our proposed **GROQ-seq** T7 RNAP system offers an extremely simple and robust readout for a deeply functionally complex protein. A model that effectively leverages this data to make accurate inferences about intricate functional effects would represent a significant developement in computational protein understanding.

DOI: DOI: 10.5281/zenodo.15025331

Gene of Interest (GOI) Region

The GOI region contains an operon in which the protein sequence is expressed in the opposite orientation to the genes in the circuit region (Fig. 1). The constitutive promoter upstream of T7 RNAP will be selected from the Anderson promoter collection[5] and will vary in strength (e.g., **J23106**, strength 0.47; **J23114**, strength 0.1) for tuning purposes.

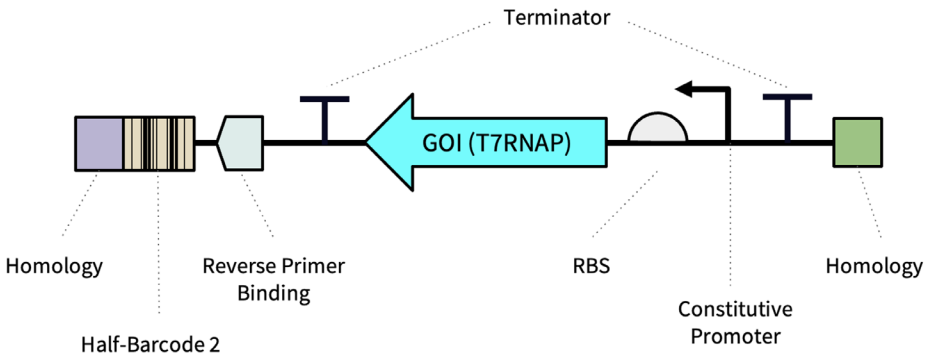


Figure 1: Scheme for the gene of interest region.

Selection Cassette

The selection cassette is a function-specific region of the plasmid (Fig. 2) that contains circuit components with which the GOI will interact (e.g., operators or substrates) as well as all necessary reporters produced by this interaction. For fitness measurements, the cassette includes an antibiotic resistance gene (Ab^R), while a fluorescent reporter is used to assess function during development. T7 RNAP binds to the T7 promoter upstream of the mScarlet gene and drives the production of the **mScarlet-DHFR** operon for both fitness and function measurements.

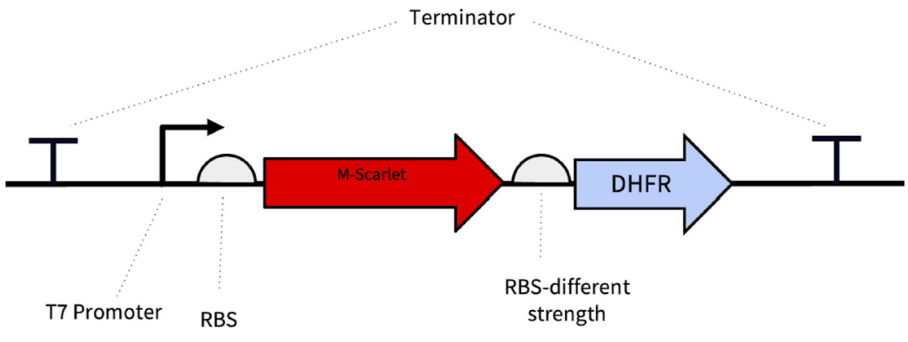


Figure 2: Scheme for the selection cassette region with mScarlet (RFP) as the reporter.

DOI: DOI: 10.5281/zenodo.15025331

Proposed Hosts and Plasmids

Two *E. coli* strains are proposed for this dataset:

- **DH10B** was chosen for this T7 RNAP dataset due to its high transformation efficiency and widespread use in synthetic biology.
- **S2060** is an *E. coli* strain with a reduced propensity to biofilm, making it valuable for continuous directed evolution experiments[6]. Some singleplex measurements are conducted in **S2060**.

Plasmid designs for fluorescence measurements:

The one-plasmid system variants are designed with two different origins of replication (**p15a** or **RK2**) and two distinct Anderson promoters of varying relative strengths positioned upstream of the reporter cassette (**mScarlet-DHFR**). Two different RBS strengths are also used for tuning purposes upstream of the **DHFR** gene. The links for both example Benchling maps to measure the activity of wild-type (WT) T7 RNAP are provided below. Note that the barcode region is highlighted, and unique barcodes should be used for each calibration variant. We are planning to use **pBD8x3** and **pBD8x6** constructs for high and low activity level, for tuning purposes. **pBD8x3** contains **p15a** origin of replication (20-30 copies), a weak promoter upstream of **DHFR** gene and **J23106** Anderson promoter with a strength level of 0.46. **pBD8x6** contains RK2 origin of replication (4-5 copies), a strong promoter upstream of **DHFR** gene and **J23114** Anderson promoter with a strength level of 0.1.

Plasmid maps of key constructs:

- pBD8x3-mscarlet-I-DHFR map
- pBD8x6-mscarlet-I-DHFR map

Proposed Controls for Assay Development

Stage 1 - Development and Tuning

We will use genotypes sourced from the literature and from our own experiments to establish a calibration ladder of polymerase variants that span the desired dynamic range. WT T7 RNAP exhibits high specificity for its promoter sequence, while evolved variants are available that recognize the T3 promoter sequence or other variations[4,7,8] (Table 1). We have confirmed that these sequences exhibit highly variable function using an existing two-plasmid luciferase reporter to measure function[9].

Our first goal in development was to transition from the existing luciferase-based two-plasmid system to the newly designed single-plasmid system that is compatible with the **GROQ-seq** platform[5]. Converting from a two-plasmid system to a one-plasmid system is desirable because it allows both the polymerase ORF and the target promoter to be encoded on the plasmid with the barcode, enabling measurement of libraries containing both protein and substrate variants.

The primary consideration for this transition was to ensure that similar dynamic range could be obtained using the single-plasmid architecture. Establishing a sufficient dynamic range is essential for effectively pooling and assaying large and diverse variant libraries. To begin this conversion, we measured WT T7 RNAP in a one-plasmid system containing either a **p15a** or **RK2** origin, along with two distinct Anderson promoters of varying relative strengths positioned upstream of the reporter cassette (**LuxAB**) (Fig. 3). We compared these results to those from the same WT T7 RNAP measured in the two-plasmid system. We assessed how replicable the results were day-to-day for each reporter variant (Fig. 4). From this, we selected two plasmid backbones, **pBD8x3** and **pBD8x6**, for further testing. These contain a combination of origin of replication and promoter strength and are likely to be tuned either 'low' or 'high', respectively (Fig. 4).

Our next goal is to switch from luminescence to fluorescence-based readouts. We have found that luminescence is excellent for measuring low-activity targets. However, luminescence signals can be significantly affected by *E. coli* metabolism, making it less suitable for measuring high-activity target proteins. We are currently switching to using mScarlet for measurement and will re-measure calibration variants with this new reporter. In this new system, we will further tune the circuit with high-activity, zero-activity, and low-activity variants to improve dynamic range and confirm that circuit readout is not binary.

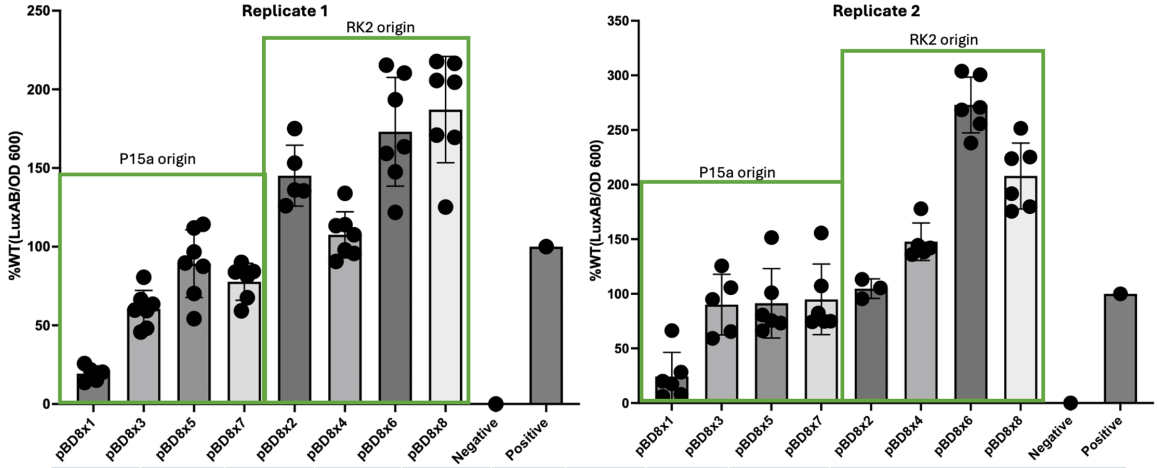
Stage 2 - Normalization and Calibration Controls

We propose to use the following eight calibration variants sourced from the literature, which we anticipate will show a range of function (Table 1). In addition, a larger diversity of variants can be investigated by using other variants available in the Crick team's collection of variants[6]. If additional controls are needed, we will perform site saturation mutagenesis to generate more variants in the hairpin loop region of T7 RNAP that makes contact with the promoter sequence[4], screen the resulting colonies using plate reader assay, and pick additional variants that build out the standard curve.

Operator (action site)	T7 RNAP variants (GOI)	Measured activity (in % WT)	Type of Control	Design Source
T7 polymerase	T7 RNAP-WT	Positive control	Calibration	Ref[10]
T7 polymerase	T7 RNAP-E207K	Low expression	Calibration	Ref[10]
T7 polymerase	T7 RNAP-N748D	Low expression	Calibration	Ref[10]
T7 polymerase	T7 RNAP-D240S	Medium expression	Calibration	Ref[11]
T7 polymerase	T7 RNAP- F21Y	Medium expression	Calibration	Ref[11]
T7 polymerase	T7 RNAP-D240G	Medium/high expression	Calibration	Ref[11]
T7 polymerase	T7 RNAP-P266L	High expression	Calibration	Ref[12]
T7 polymerase	T7 RNAP-D240E	High expression	Calibration	Ref[11]

Table 1: T7 RNAP variants to be used during stage 1.

DOI: DOI: 10.5281/zenodo.15025331



	pBD8x1	pBD8x3	pBD8x5	pBD8x7	pBD8x 2	pBD8x4	pBD8x6	pBD8x8	Positive two plasmid system
Origin	p15a	p15a	p15a	p15a	RK2	RK2	RK2	RK2	RK2
Promoter	J23106	J23106	J23114	J23114	J23106	J23106	J23114	J23114	J23114
RBS	Strong	Weak	Strong	Weak	Strong	Weak	Strong	Weak	Strong

Figure 3: Results from the **pBD8x2** one-plasmid system are comparable to those from the two-plasmid systems used as the positive control, as seen in two biological replicates.

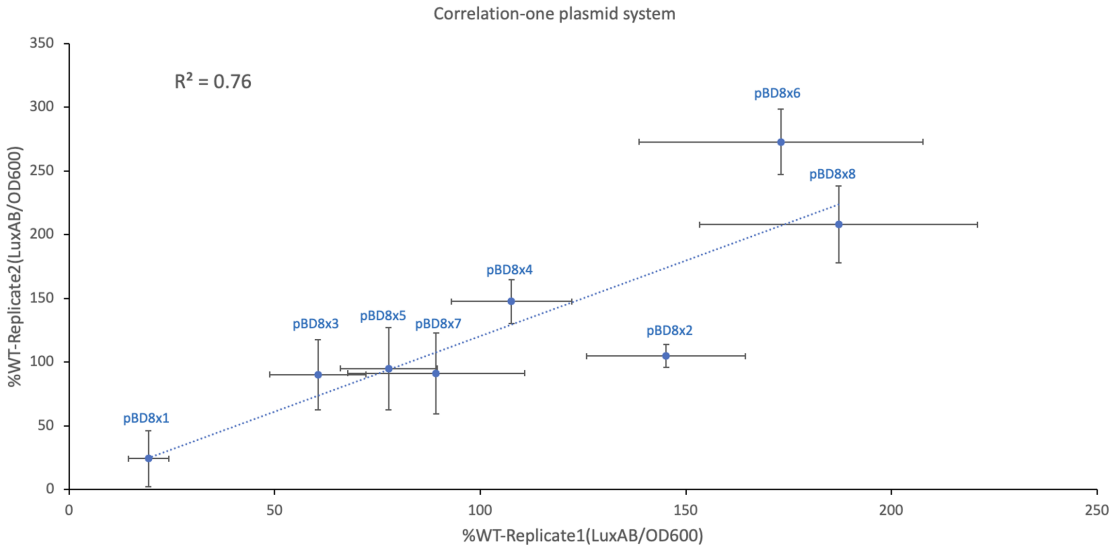


Figure 4: Luminescence results from the one-plasmid system. Two biological replicates were measured for each origin/promoter/RBS combination. **pBD8x3** and **pBD8x6** were selected as the plasmid backbones for further experiments, as they cover the desired dynamic range for the assay. Crosshairs indicate in-plate standard deviation.

DOI: DOI: 10.5281/zenodo.15025331

Pilot-Scale Collection

As part of a previous project, we designed and cloned a combinatorial library of T7 RNAP variants and used them as input to a continuous directed evolution experiment. We propose repurposing this library as the first to be measured by the T7 RNAP **GROQ-seq** platform.

We previously developed a computational pipeline designed to maximize the diversity of protein homologs sampled within the protein landscape. This approach is inspired by previous efforts to create combinatorial libraries of protein homologs[13]. In this approach, the target protein is broken into structurally distinct sub-domains. Many versions of each sub-domain are synthesized, and pooled cloning generates a library containing combinations of the sub-domain variants (Fig. 5a). The library design pipeline has the following steps:

1. **Homolog search.** We employed two complementary approaches for deep homology searches: **BLAST** for sequence similarity and **Foldseek** for structural similarity. This process yielded a shortlist of 10,000 protein sequences.
2. **Filter for known structure.** From these, we filtered down to sequences with structures present either in the Protein Data Bank (PDB) or the AlphaFold database, which enables structure-based scoring. Structures were not filtered based on confidence or experimental structure quality.
3. **Identify the optimal set of homologs.** We then applied 100 rounds of Monte Carlo optimization to select 10 sequences. These sequences were chosen to maximize both sequence and structural similarity to WT T7 RNAP, as measured by sequence identity and Template Modeling (TM) score, while maintaining minimal sequence identity among themselves (Fig. 5b).
4. **Identify split sites.** Once these sequences were selected, we performed structural alignments to identify three highly conserved residues to serve as split sites for recombination. This process breaks the full-length T7 RNAP into four sub-domains, each of which is small enough to be synthesized separately. Note that some homologous sequences lacked one or both final two domains. We truncated these at the nearest split site, producing a total of 34 fragments, thereby creating a library with an approximate size of 10,000 variants. Additionally, for the WT T7 RNAP sequence, we split it into four fragments and identified mutations of interest by leveraging information encoded by the probability distributions over sequence space learned by protein language models.
5. **DNA synthesis and library generation.** We synthesized all 34 fragments and cloned the full combinatorial library, along with sub-libraries composed of a smaller selection of fragments. We measured coverage of the assembled library using **Nanopore sequencing** (Fig. 5c).

More specifically, we posit that targeting residues at sites where the model demonstrates significant entropy in its distribution - signifying a degree of uncertainty in prediction—may correspond to less deleterious mutations. To test this hypothesis, we sampled the three lowest- and three highest-entropy positions within each fragment, generating 24 fragments. Each fragment represented both low- and high-entropy positions. The goal was to cover a wide range of the sequence-structure similarity space, ensuring a diverse mix of variants, including positive and negative controls, to thoroughly explore the protein landscape.

In addition to this library, we will assay a site-saturation mutagenesis library of WT T7 RNAP, including all single amino acid mutations, as well as insertions and deletions (~40,000 variants). We will also assay an error-prone PCR (ePCR) library of WT T7 RNAP, aiming for an average of four amino acid mutations on average across the variants.

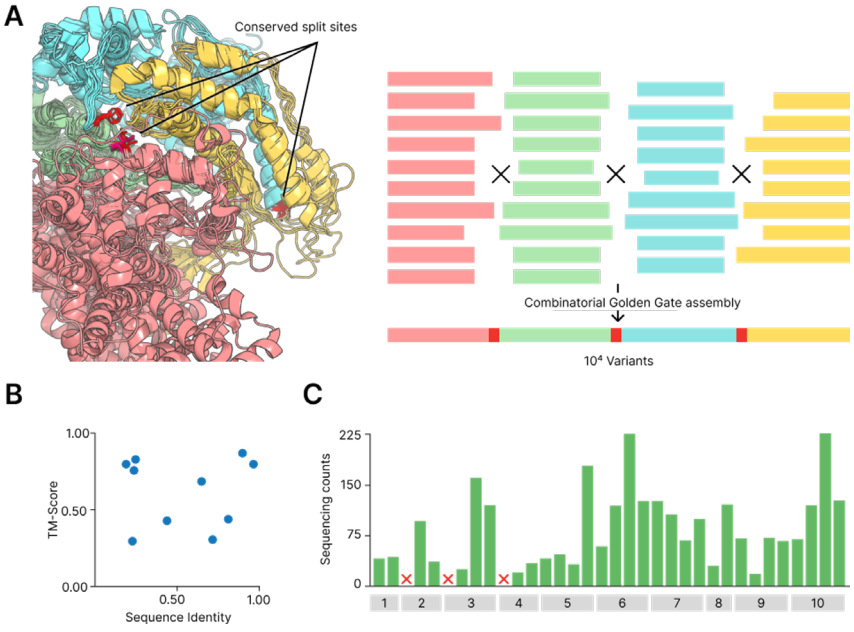


Figure 5: **(A)** Overlay of crystal and AlphaFold structures for the 10 chosen homologs, and a schematic of the library generation approach. **(B)** The 10 selected T7 RNAP homologs plotted by TM score (a measure of structural similarity) and sequence identity. **(C)** Nanopore sequencing counts of the plasmid library, showing nearly complete coverage of all fragments. Among the 10 homologs, only three subdomains from three homologs are absent in the final library.

DOI: DOI: 10.5281/zenodo.15025331

Large-Scale Collection

In our pilot-scale data collection, we diversified by sampling sequences of 10 different T7 RNAP homologs. At a larger scale, we will combine this with a promoter library consisting of the WT T7 RNAP promoter, randomized from 10% to 90% sequence identity, along with all base combinations at the -8, -9, and -11 positions, which are key determinants of specificity among the closely related but orthogonal T7, T3, and SP6 promoters[8]. We aim to sample the WT T7 RNAP activity on 500,000 such promoter sequences.

To further push the depth to which we can unpick specific sequence features conferring different aspects of T7 RNAP function, we will use results from the pilot study to build a computational model for T7 RNAP activity based on sequence data. This model will cluster the protein sequence into regions of functional importance. Each region will then be computationally mutagenized, assessing multiple generative models to fill them in with plausible protein sequences and maximize the space of interesting variants we sample while minimizing the number which are "dead on arrival". Additionally, we may layer on random mutagenesis with ePCR or phage-assisted continuous evolution (PACE) to sample point mutant variation on top of homolog shuffling. The 500,000 sequences, with varying levels of homology to WT T7 RNAP, will be chosen and assayed against the WT T7 RNAP promoter sequence.

Finally, we propose to use data from the first two large-scale data collection rounds to choose a set of 1,000 promoters and polymerase sequences, assaying them combinatorially for a total of 1,000,000 RNAP-promoter pairs. By co-sampling differences in both the promoter and the polymerase, this dataset should be highly informative in dissecting the sequence determinants of promoter specificity and orthogonality.

Current Developmental Stage

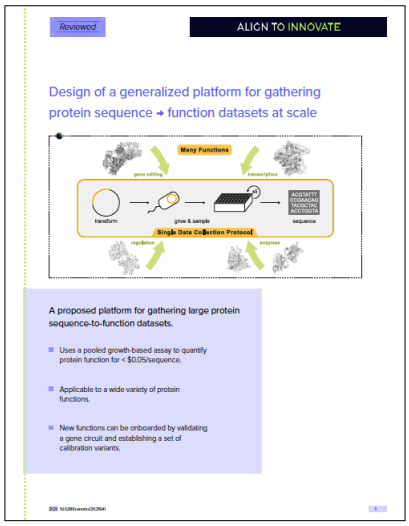
We have identified variants with activities ranging from 9% to 400% of wild-type activity using a luminescence reporter in a two-plasmid system, where one plasmid expresses the variant and the other encodes the luciferase reporter. We are now converting this system into a one-plasmid measurement assay by creating plasmids that contain both protein expression and selection cassettes (Figs. 1-2). To establish calibration controls, we will measure the same variants in this new system. Additionally, we are conducting experiments to replace **LuxAB** with mScarlet as a fluorescent reporter.

DOI: DOI: 10.5281/zenodo.15025331

Suggested Further Reading

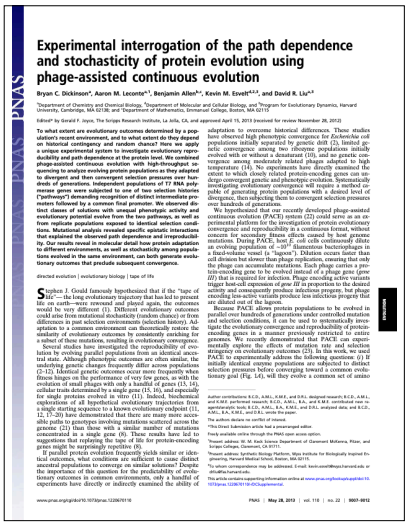
For more details on the recently named GROQ-seq platform, read this platform design document

"Design of a generalized platform for gathering protein sequence-to-function datasets at scale", 2024, Zenodo



Read more about the history of studying T7 RNAP's promoter specificity through directed evolution experiments.

"Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution", 2013, PNAS



Learn about combinatorial library assembly as as cost-effective strategy for generating diverse protein libraries.

"Combinatorial assembly and design of enzymes", 2023, Science

DOI: DOI: 10.5281/zenodo.15025331

References

1. Dousis, A., Ravichandran, K., Hobert, E. M., Moore, M. J. & Rabideau, A. E. An engineered T7 RNA polymerase that produces mRNA free of immunostimulatory byproducts. *Nat. Biotechnol*. **41**, 560–568 (2023).
2. Pu, J., Zinkus-Boltz, J. & Dickinson, B. C. Evolution of a split RNA polymerase as a versatile biosensor platform. *Nat. Chem. Biol*. **13**, 432–438 (2017).
3. Moore, C. L., Papa, L. J., 3rd & Shoulders, M. D. A processive protein chimera introduces mutations across defined DNA regions in vivo. *J. Am. Chem. Soc*. **140**, 11560–11564 (2018).
4. Carlson, J. C., Badran, A. H., Guggiana-Nilo, D. A. & Liu, D. R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat. Chem. Biol*. **10**, 216–222 (2014).
5. Cortade, D. *et al*. Design of a generalized platform for gathering protein sequence ➔ function datasets at scale. Preprint at https://doi.org/10.5281/ZENODO.12521641 (2024).
6. Hubbard, B. P. *et al*. Continuous directed evolution of DNA-binding proteins to improve TALEN specificity. *Nat. Methods* **12**, 939–942 (2015).
7. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499–503 (2011).
8. Dickinson, B. C., Leconte, A. M., Allen, B., Esvelt, K. M. & Liu, D. R. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc. Natl. Acad. Sci. U. S. A*. **110**, 9007–9012 (2013).
9. DeBenedictis, E. A. *et al*. Systematic molecular evolution enables robust biomolecule discovery. *Nat. Methods* **19**, 55–64 (2022).
10. E, C., Dai, L. & Yu, J. Switching promotor recognition of phage RNA polymerase in silico along lab-directed evolution path. *Biophys. J*. **121**, 582–595 (2022).
11. Stano, N. M. & Patel, S. S. The intercalating beta-hairpin of T7 RNA polymerase plays a role in promoter DNA melting and in stabilizing the melted DNA for efficient RNA synthesis. *J. Mol. Biol*. **315**, 1009–1025 (2002).
12. Guillerez, J., Lopez, P. J., Proux, F., Launay, H. & Dreyfus, M. A mutation in T7 RNA polymerase that facilitates promoter clearance. *Proc. Natl. Acad. Sci. U. S. A*. **102**, 5958–5963 (2005).
13. Lipsh-Sokolik, R. *et al.* Combinatorial assembly and design of enzymes. *Science* **379**, 195–201 (2023).

DOI: DOI: 10.5281/zenodo.15025331