GROQ-seq: A Collaborative, Open Data Approach to Addressing Protein Function Prediction

David Ross[1], Aviv Spinner[2], Simon d'Oelsnitz[3], Svetlana Ikonomova[1], Olga Vasilyeva[1], Nina Alperovich[1], Kristen Sheldon[5], Courtney Tretheway[5], Anjali Chadha[4] Dana Cortade[2], Erika DeBenedictis[2, 4], Peter Kelly[2]

[1]Living Measurement Systems Foundry, National Institute for Standards and Technology (NIST)
[2]Align to Innovate
[3]Harvard Medical School, Harvard University
[4]Biodesign Lab, The Francis Crick Institute
[5]The DAMP Lab, Boston University

Abstract

We have developed an experimental platform and unified data ontology to facilitate the collection and sharing of open-access data on diverse protein functions, with the goal of enabling the development of predictive models that link sequence to function. Our data generation strategy utilizes the **gro**wth-based **q**uantitative **seq**uencing (GROQ-seq) platform, a simple yet adaptable system that can be easily extended to encompass new functions. This high-throughput platform allows for the quantitative functional characterization of hundreds of thousands of proteins per experiment at a cost of approximately $0.05 per sequence. By combining scalability with extensibility, GROQ-seq enables the generation of the diverse functional datasets needed to develop a generalizable model that quantitatively predicts sequence-to-function relationships. Currently, GROQ-seq assays for seven protein functions are being developed by six academic teams, complemented by orthogonal measurements and reproducibility studies across automation sites.

Introduction

Protein functions, such as enzymatic activities and binding interactions, exist as isolated islands within the broader protein function landscape, whose relationship to the corresponding sequence landscape remains incompletely understood[1,2]. While machine learning (ML) algorithms have made strides in bridging this gap, current methods still lack a generalizable solution for predicting a protein's function directly from its DNA sequence[3–5]. Building such a model requires large, high-fidelity datasets encompassing diverse protein functions. In response, we have developed GROQ-seq[2], a standardized and extensible platform for sequence-to-function data generation that can be implemented at multiple sites. GROQ-seq builds upon previous work using quantitative sequencing-based fitness proxies, aligning data with well-characterized variants to convert read counts into direct functional measurements[6]. By providing quantitative, scalable, and reproducible datasets, GROQ-seq lays the foundation for the next generation of sequence-to-function ML models and serves as a continuously expanding resource to which researchers can contribute over time (Fig. 1).

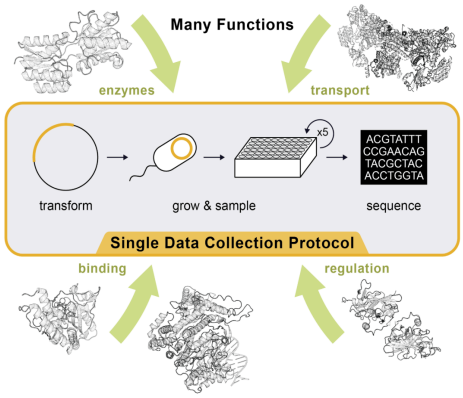

Figure 1: GROQ-seq is an extensible platform for protein sequence-to-function dataset generation.

Results

Currently, seven functional assays are being onboarded to the GROQ-seq platform[2], measuring transcription factors[7], proteases[8], aminoacyl tRNA synthetases[9], RNA polymerases[10], histidine kinases[11], single-chain antibody fragments[12], and dihydrofolate reductases[13]. To demonstrate the GROQ-seq onboarding process, we present results from our small-molecule-responsive transcription factor GROQ-seq assay which measures repression and inducibility of three transcription factors: LacI, RamR, and VanR. A 22-member calibration ladder was developed to densely span the four-order-of-magnitude dynamic range of the optimized assay (Fig. 2A). Using the calibration ladder, the robustness of our assay was validated by comparing the pooled GROQ-seq measurements with singleplex quantitative protein function measurements obtained via flow cytometry (Fig. 2A). This 22-member calibration ladder will be included in subsequent measurements to enable quantitative calibration of variant function, including the first large-scale measurement (~300,000 pooled sequences) set to be measured in early 2025. In parallel, we also measured the reproducibility of data collection across multiple automation sites. Initial function measurements of the transcription factor ladder were repeated at both the National Institute of Standards and Technology (Maryland, USA) and Boston University's DAMP Lab (Massachusetts,

USA). After applying a scaling factor to account for differences in flow cytometer and plate reader measurement scales, dose-response curves demonstrate good reproducibility across sites(Fig. 2B).

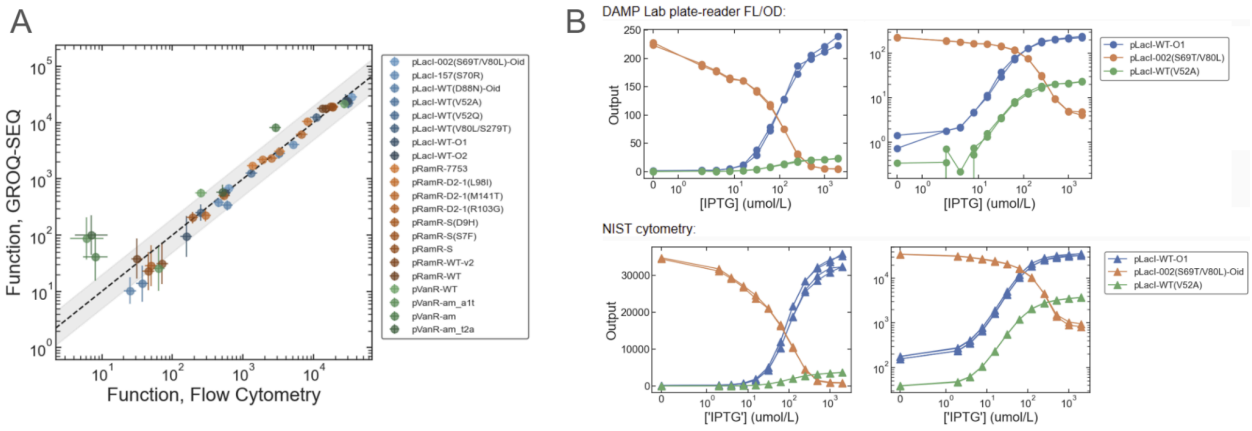

Figure 2: Results from the GROQ-seq transcription factor assay showing A) the optimized calibration ladder and B) initial results for reproducibility across two sites.

Methods

The GROQ-seq platform enables pooled libraries of gene variants to be measured via automation-enabled growth-based assays linking the function of a gene to cell growth. Diverse protein variant libraries are built from a combination of site saturation of all single amino acid substitutions, insertions, and deletions, spatially-constrained combinatorial saturation, ePCR, and ML-guided designs. These libraries are transformed, challenged under selective pressure, and then sequenced to measure the abundance of individual barcoded variants. A set of calibration variants is included in each pool, with well-characterized function values. The read count data for calibration variants can then be used as a “ladder”, to produce a quantitative function value for all variants in the pool, with data analysis performed using computation tools expanding upon the work of Tack et al[6] (soon to be open sourced). GROQ-seq data is obtained using cells grown in liquid culture to mid-log or lower density. This, combined with the use of protein-function ladders, enables the calibration of the pooled fitness measurements to function values. When possible, orthogonal in vitro protein function assays are used to characterize the calibration ladder, further translating the quantitative GROQ-seq functional measurements into biophysical property values. For example, in collaboration with Ginkgo Bioworks (Massachusetts, USA), we are currently measuring enzyme kinetics of our protease calibration ladder using an in vitro assay. The limiting factor on the measurement scale of variant pool size is the cost of sequencing. All method details are published in our original GROQ-seq proposal[2] and all associated protocols are hosted on Protocols.io[14–18].

Discussion

Our results demonstrate that the GROQ-seq platform is a scalable and reproducible platform capable of generating high-throughput, quantitative protein function measurements. The successful demonstration of the transcription factor pooled measurements and the establishment of a calibration variant ladder highlight the platform’s potential to provide standardized, quantitative, and high-throughput functional data. Furthermore, our ability to reproduce these measurements across independent laboratories underscores the robustness of our approach and its suitability for broad adoption.

As we continue to expand our dataset to include additional protein functions outside of the initial seven, we anticipate that the resulting large-scale, high-quality functional data will be foundational to the creation of improved predictive models of sequence-function relationships. Once the first GROQ-seq large-scale dataset is generated, we will be able to evaluate which library design techniques enable the most efficient mapping of the functional landscape and enable the creation of both task-specific, and eventually, generalizable models of function prediction. By providing a unified data ontology and an open-access repository, we aim to facilitate collaboration across the scientific community, enabling iterative improvements in model accuracy and generalizability. Ultimately, our vision is to expand GROQ-seq to establish a comprehensive platform that bridges the gap between protein sequence and function, accelerating advances in protein engineering, synthetic biology, and computational modeling.

References

1. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**, 866–876 (2009).
2. 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.13909104 (2024).
3. Notin, P., Rollins, N., Gal, Y., Sander, C. & Marks, D. Machine learning for functional protein design. *Nat Biotechnol* **42**, 216–228 (2024).
4. Lin, B., Luo, X., Liu, Y. & Jin, X. A comprehensive review and comparison of existing computational methods for protein function prediction. *Brief Bioinform* **25**, (2024).
5. Kulmanov, M., Khan, M. A., Hoehndorf, R. & Wren, J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660–668 (2018).
6. Tack, D. S. *et al.* The genotype-phenotype landscape of an allosteric protein. *Mol Syst Biol* **17**, e10847 (2021).
7. d'Oelsnitz, S., Taghon, G., Cortade, D., Kelly, P. & Ross, D. Design of growth-coupled measurements of transcription factor function. Preprint at https://doi.org/10.5281/ZENODO.13286832 (2024).
8. Chadha, A., Hayes, O., Cortade, D., Kelly, P. & DeBenedictis, E. Design of growth-coupled measurements of protease function. Preprint at https://doi.org/10.5281/ZENODO.13286759 (2024).
9. Kelly, P., Cortade, D., Ross, D. & Thyer, R. Design of growth-coupled measurements of orthogonal aminoaeynolds, K. A. & Filipowska, K. Design of growth-coupled measurements of Dihydrofolate Reductase *in vivo* biochemistry. *Zenodo (manuscript in preparation)*.
14. Ross, D. Singleplex Assay for Function Measurements. (2025).
15. Alperovich, N. & Ross, D. Bar-Seq Library Preparation and Pooling: Preparation of Sera-Mag SpeedBeads. (2025).
16. Ross, D. & Alperovich, N. Automated Bar-Seq Library Preparation and Pooling. (2024).
17. Ross, D. Pooled, Growth-Based Assays. (2024).
18. Ross, D. Singleplex Assay for Fitness Measurements. (2024).cyl-tRNA synthetase function. Preprint at https://doi.org/10.5281/ZENODO.13338159 (2024).

10. Kelly, P. *et al.* Design of growth-coupled measurements of T7 RNA Polymerase function. *Zenodo (manuscript in preparation)*.
11. Kelly, P., Cortade, D., Ross, D., DeGrado, W. & Hatstat, K. Design of growth-coupled measurements of histidine kinase function. Preprint at https://doi.org/10.5281/ZENODO.13793799 (2024).
12. Kelly, P., Cortade, D., Koder, R. & Ross, D. Design of growth-coupled measurements of single-chain antibody fragment function. Preprint at https://doi.org/10.5281/ZENODO.14502177 (2024).
13. Kelly, P., Cortade, D., R