Development of single-cell massively parallel reporter assays for measurement of disease-associated variant effects.

Mackenzie C. Noon, Erin N. Gilbertson, Steven K. Reilly

BEST

Boehringer

Emerging

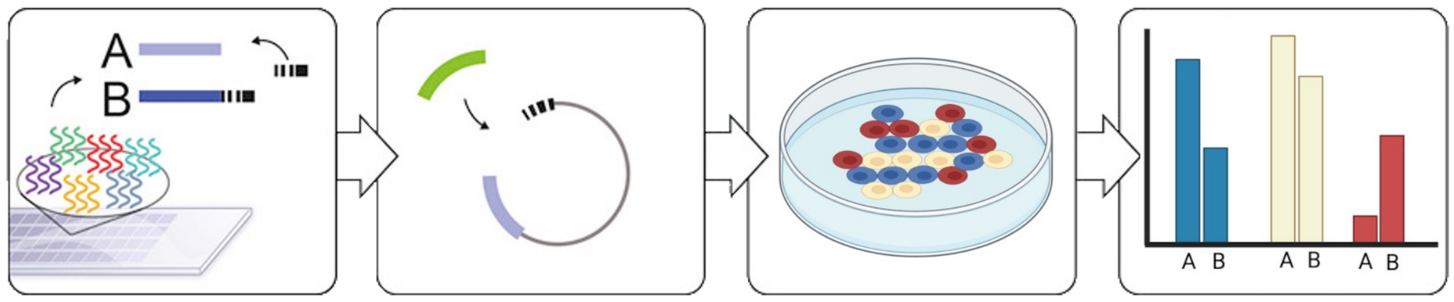
Scientist

Talent

Introduction

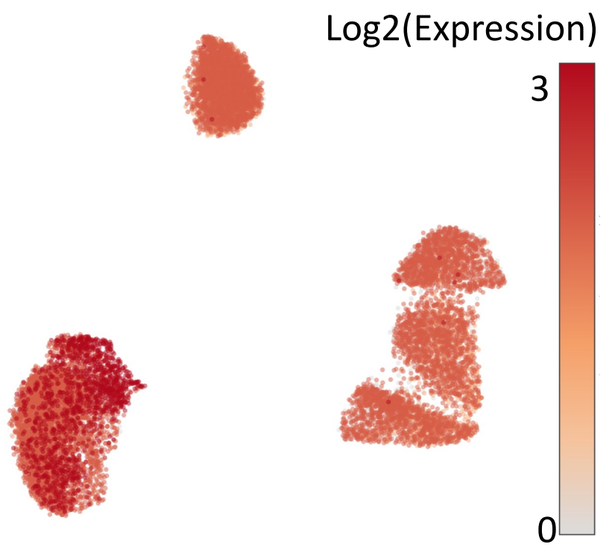
Understanding the contribution of human genetic variation to disease is essential for developing therapeutics to treat those diseases; Drug mechanisms which have genetic support are about 2.6 times more likely to succeed (Minikel et al. 2024). Genome-wide association studies (GWAS) are a popular approach for finding genetic variants associated with human disease. Though powerful, GWASes cannot demonstrate causality and exclude confounders like linkage disequilibrium. GWAS hits also reveal that the majority of trait associated variants lie in noncoding DNase hypersensitivity sites, like enhancers and promoters, which are much more difficult to interpret than protein coding mutations (Maurano et al. 2012). This project develops a new class of assay, single cell massively parallel reporter assays, and associated statistical techniques to functionally validate GWAS hits in a high-throughput, cell-type and state aware fashion.

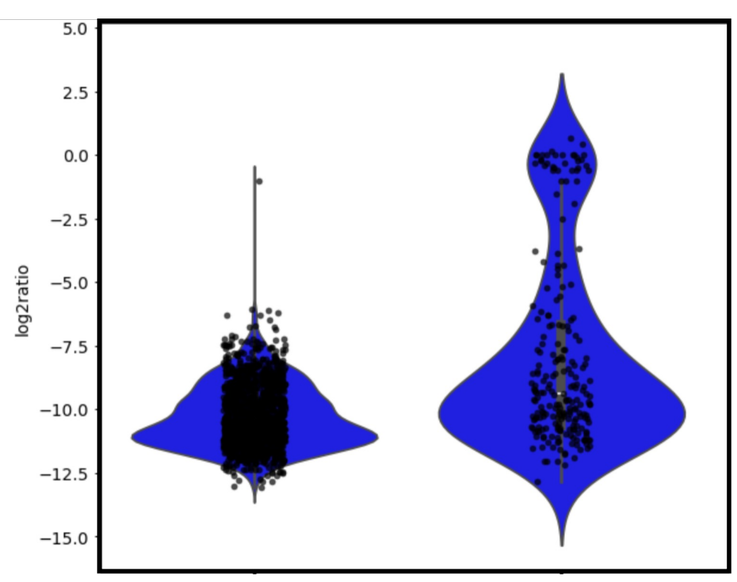
Methods



Massively Parallel Reporter Assays facilitate high-throughput functional validation of GWAS hits (Tewhey et al. 2016). Noncoding elements are synthesized with patient variants, cloned into a library with random barcodes, and sequenced. This library is then transformed into cells, allowed to express, and transcript is sequenced. The ratio of transcript RNA to library DNA reports on the strength of a variant-containing CRE. Many noncoding elements act in cell-type or cell-state specific fashions, and this is not captured by traditional MPRA, so we have extended the technique by combining it with combinatorial barcoding based single cell RNA sequencing : “scMPRA”. MPRA and endogenous gene expression (GEX) transcripts are PCRed and sequenced in separate libraries.

Pilot scMPRA experiments





- Mixed HepG2, K562, and SK-N-SH to simulate a heterogeneous environment
- Performed scMPRA
- Recovered MPRA transcripts from GEX

- Recapitulated differences in controls observed in bulk MPRA.

Bibliography

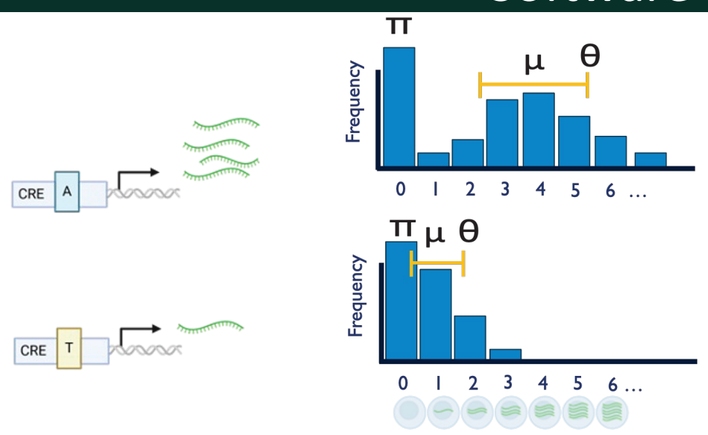
Lalanne, Jean-Benoît, Samuel G. Regalado, Silvia Domcke, Diego Calderon, Beth Martin, Tony Li, Chase C. Suiter, Choli Lee, Cole Trapnell, and Jay Shendure. 2022. “Multiplex Profiling of Developmental Enhancers with Quantitative, Single-Cell Expression Reporters.” bioRxiv. https://doi.org/10.1101/2022.12.10.519236.

Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” Science (New York, N.Y.) 337 (6099): 1190–95.

Minikel, Eric Vallabh, Jeffery L. Painter, Coco Chengliang Dong, and Matthew R. Nelson. 2024. “Refining the Impact of Genetic Evidence on Clinical Success.” Nature 629 (8012): 624–29.

Tewhey, Ryan, Dylan Kotliar, Daniel S. Park, Brandon Liu, Sarah Winnicki, Steven K. Reilly, Kristian G. Andersen, et al. 2016. “Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay.” Cell 165 (6): 1519–29.

Software for simulation and evaluation of scMPRA



$$\mathbf{X}^{(C1)} = \begin{bmatrix} \text{T} & \text{B} \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \texttt{UMI\_count}^{(C1)} = \begin{bmatrix} 12 \\ 7 \\ 10 \end{bmatrix}$$

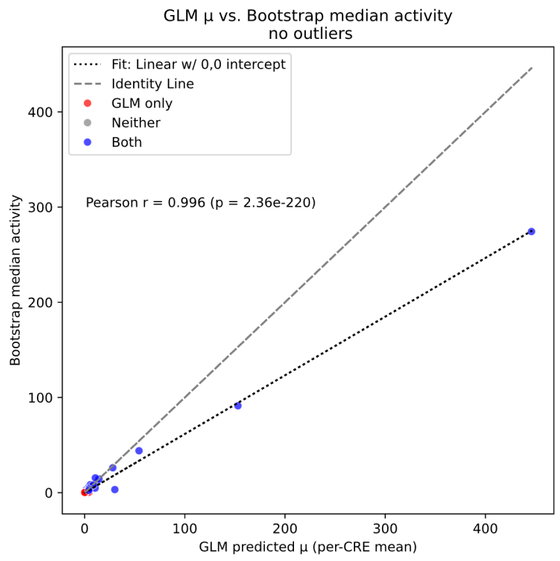
$$\mathbf{X}^{(C2)} = \begin{bmatrix} \text{T} & \text{B} \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \texttt{UMI\_count}^{(C2)} = \begin{bmatrix} 18 \\ 3 \\ 2 \\ 15 \end{bmatrix}$$

$$\mathbf{X}^{(T)} = \begin{bmatrix} \text{C1} & \text{C2} & \text{C3} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \texttt{UMI\_count}^{(T)} = \begin{bmatrix} 12 \\ 18 \\ 9 \\ 10 \\ 15 \end{bmatrix}$$

$$\mathbf{X}^{(B)} = \begin{bmatrix} \text{C1} & \text{C2} & \text{C3} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \texttt{UMI\_count}^{(B)} = \begin{bmatrix} 7 \\ 3 \\ 4 \\ 2 \end{bmatrix}$$

- Modeled MPRA data with zero-inflated negative binomial model distributions

- Employed a stratified modeling approach for computationally tractable specificity and activity testing.



Used modeling approach to perform Wald hypothesis testing. Compared estimates to existing permutation tests for scMPRA (Lalanne et al.) Recapitulated measurements & identified a superset of previously called active CREs.

Data drawn from model of type	Data then modeled with	Mean squared error	Mean biased error
By cre	By cre	2.212600	0.106597
By cre	By cell type	5.059360	-0.207537
By cell type	By cre	1.423526	0.165734
By cell type	By cell type	2.229020	0.116279

- Evaluated simulation of realistically sized scMPRA dataset with parameters from Lalanne et al.

Acknowledgements

Boehringer Ingelheim
REILLY LAB
TewheyLab
LUX ET VERITAS
Y|CRC
The Jackson Laboratory
GRUBER FOUNDATION
bio RENDER
NIH NHGRI