Can we conduct CRISPR tiling deletion screens *in silico?*

Gloria Grama, McVicker Lab
August 7th, 2024

Watch talk here: https://watch.salk.edu/media/t/1_u5yxkpgm/350492942

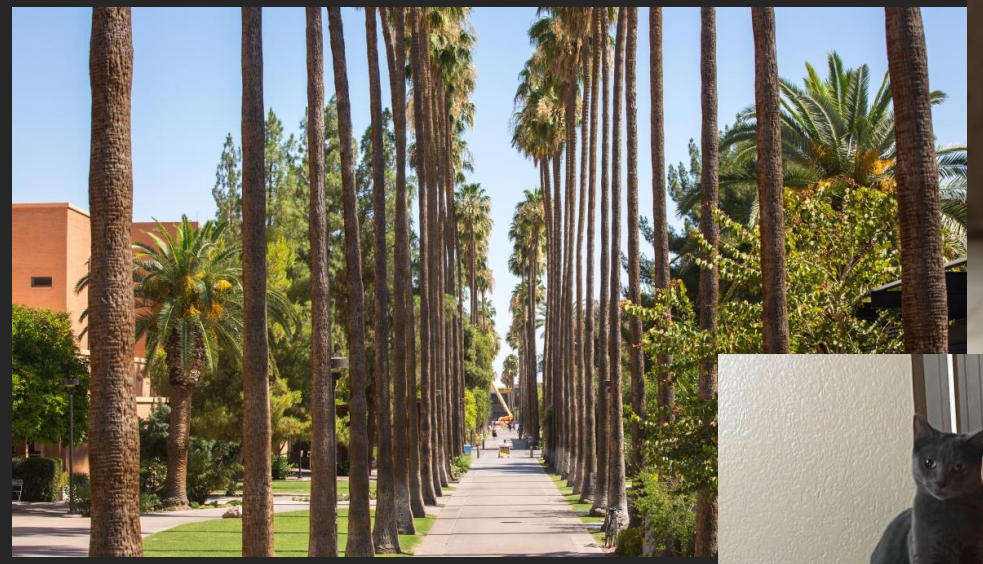
A little about me

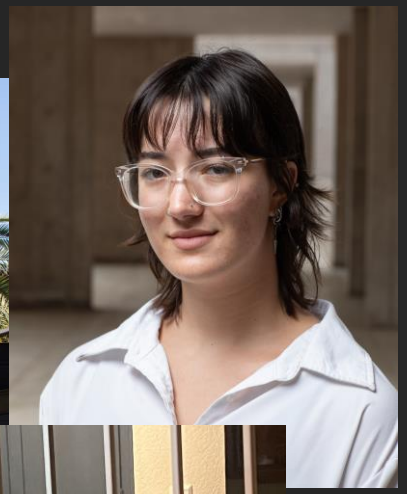
Hometown: Phoenix, AZ

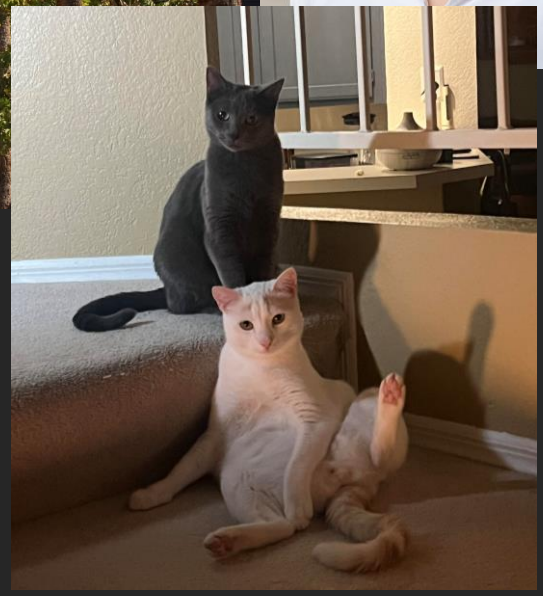
Undergraduate Education: Computational Biology student at Arizona State University

Salk Position: SURF intern in McVicker Lab

Research: Identifying genetic sequences and variants associated with changes in gene expression





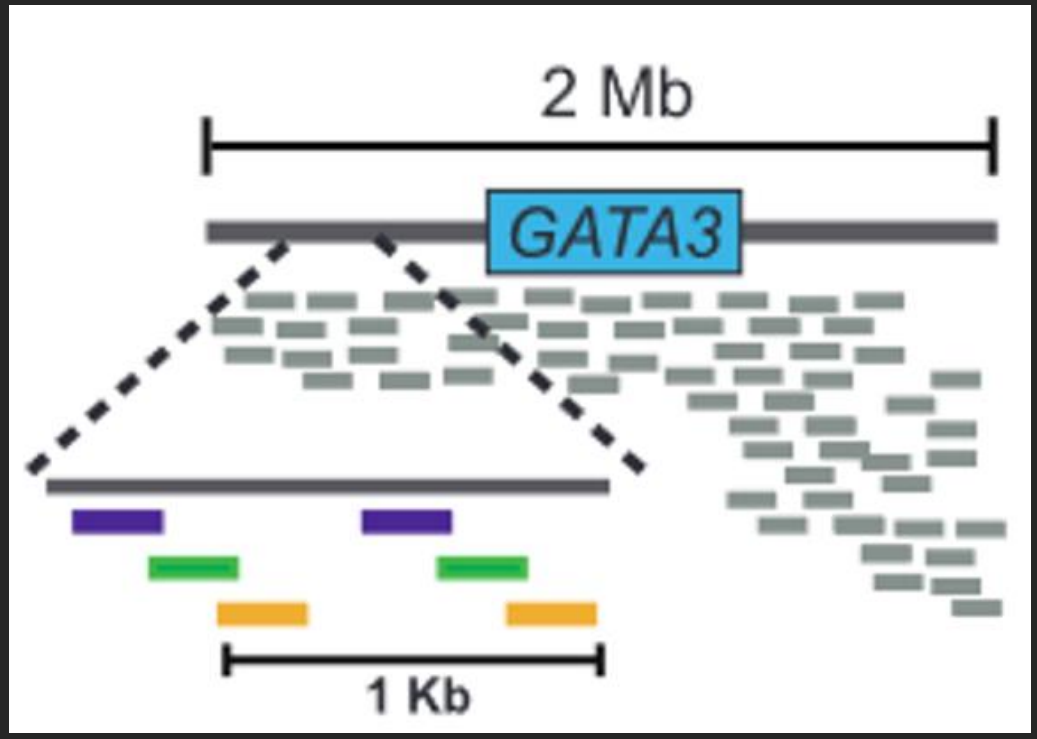


What is CRISPR?

- CRISPR is a powerful and precise tool for genetic editing

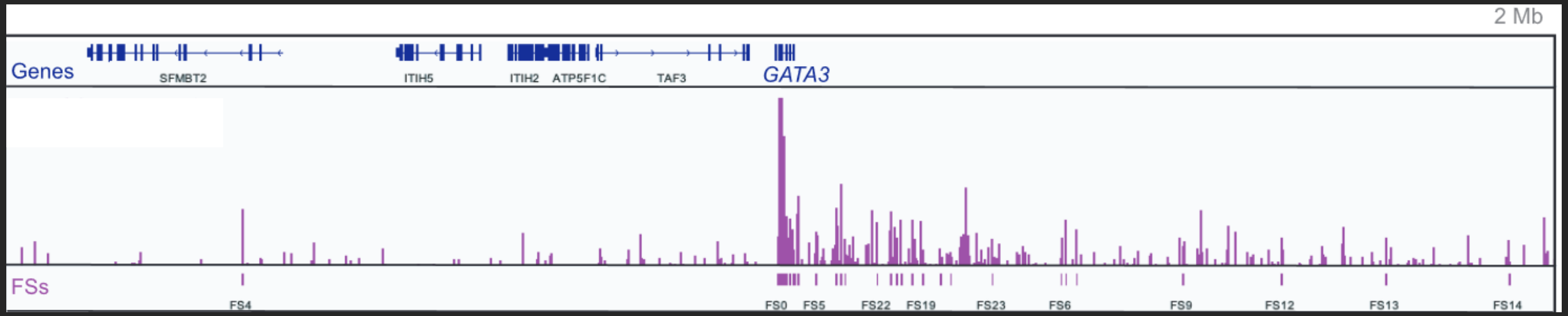
What is a CRISPR tiling screen?

- Systematic deletions of small overlapping segments
- Deletions are measured against their effect on gene expression



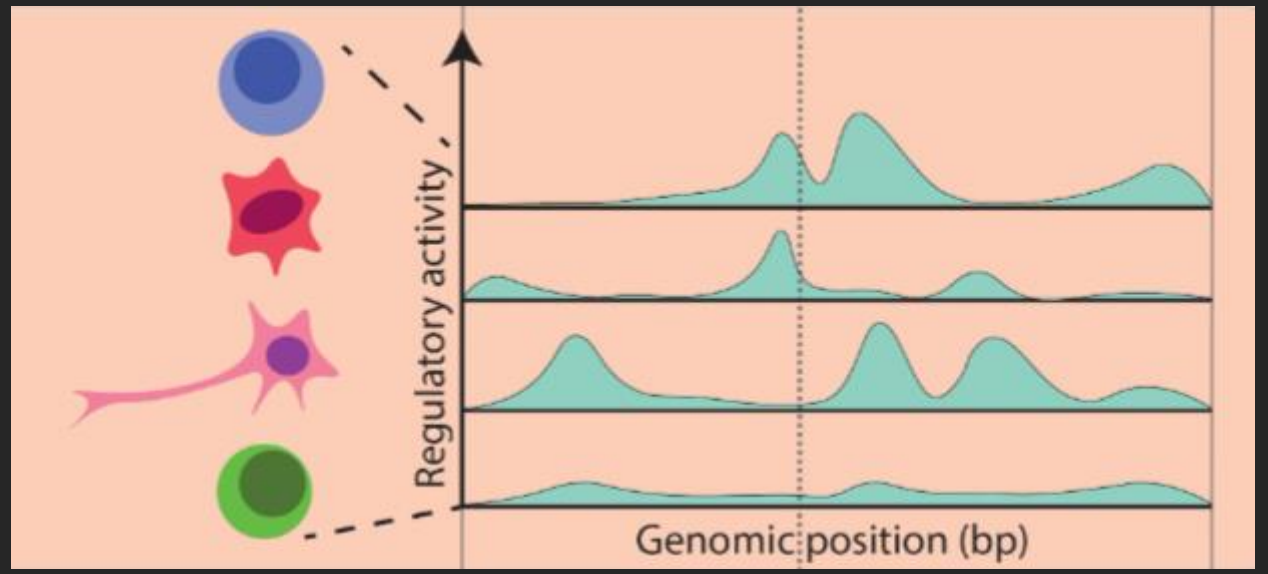
Previous CRISPR deletion tiling experiment

- CRISPR tiling on GATA3 in Jurkat T cells has identified functional sequences (Chen et. al 2023)
- Only some parts of the genome play a role in how genes are expressed



Problem

- It’s impractical to perform CRISPR tiling screens on ~30,000 protein coding genes in the genome
- Effects differ across different cell types
- Experiments are costly (time and money)



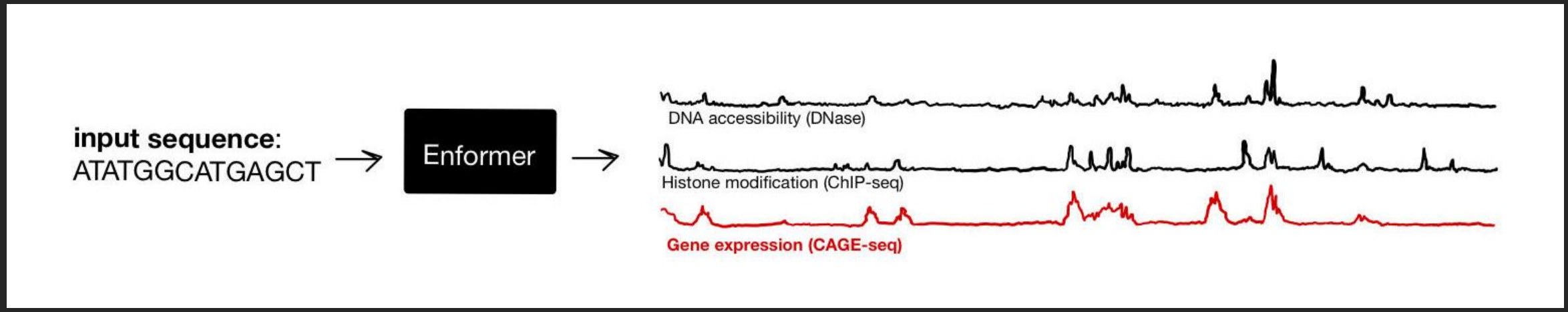
Hypothesis:

We can iteratively tile through a sequence *in silico* and measure functional sequences using deep learning predictions.

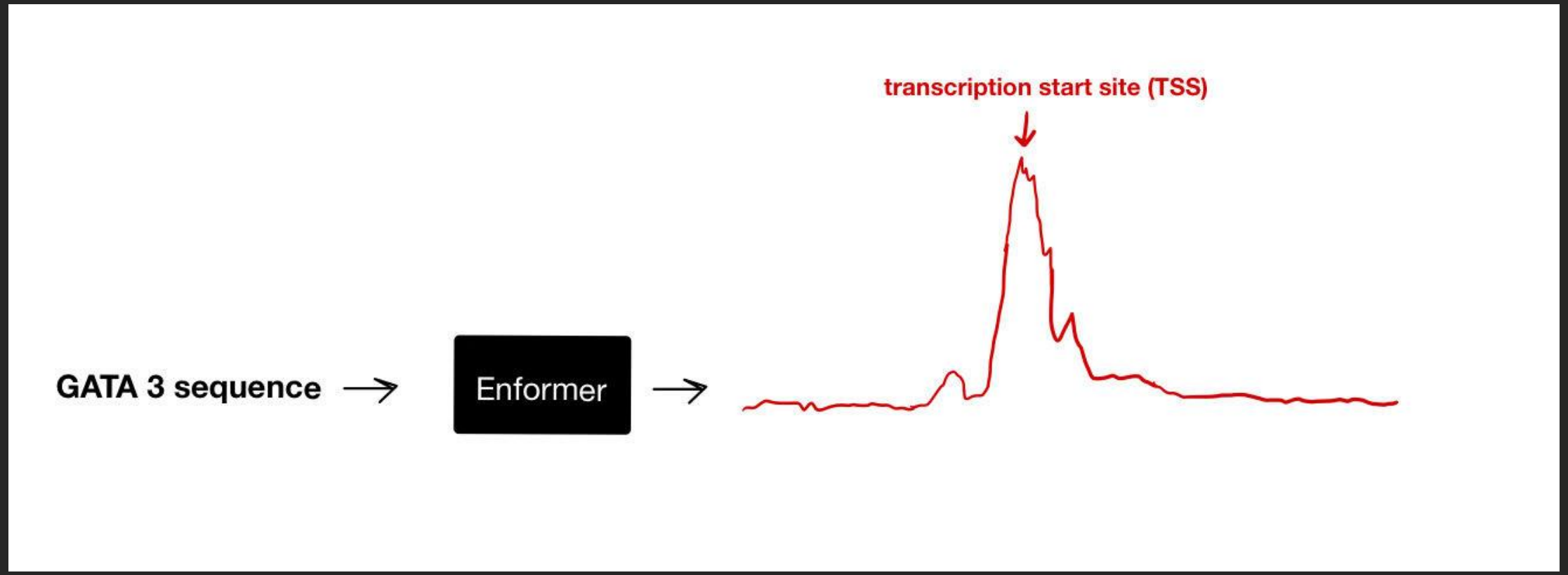
in silico = experimentation conducted within a computer

What can we use machine learning for?

- Deep learning models are being used to predict gene expression (they're pretty good at it too)

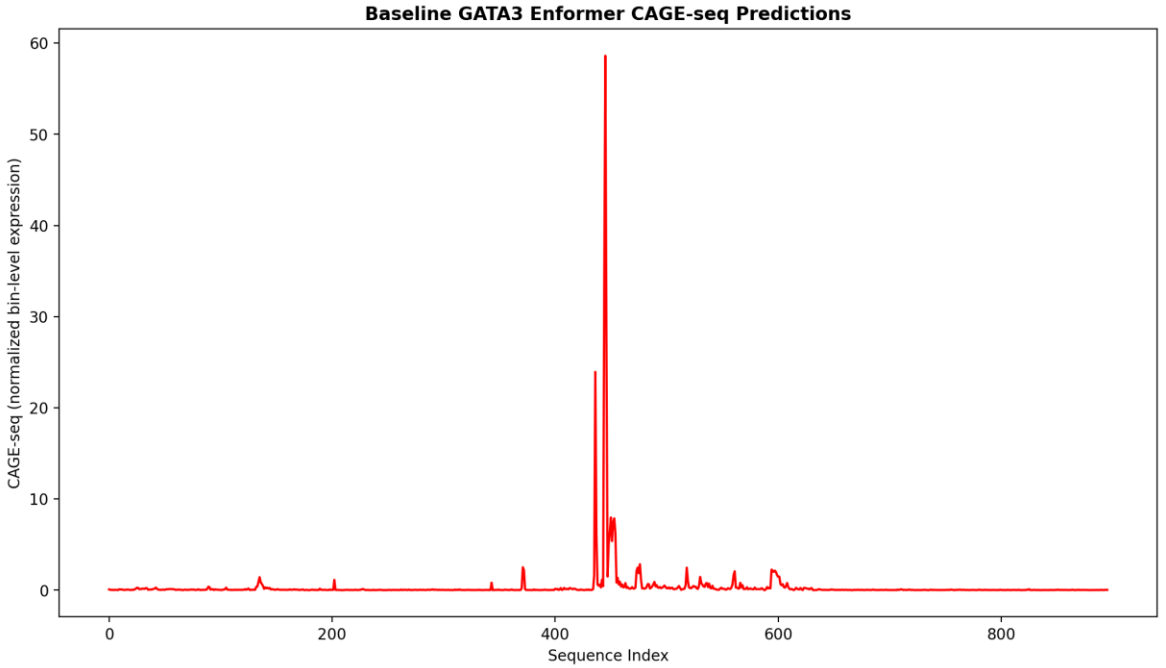


CAGE-seq track predicts gene expression



CAGE-seq track predicts gene expression

GATA 3 sequence → Enformer →



Recreating CRISPR tiling screen *in silico*

- Iteratively replace sequence with N's in 128bp bins

baseline:

AATGCCCTGACTGACGTAC … GATCAGTTTAGCCAAAAA

iteration 1:

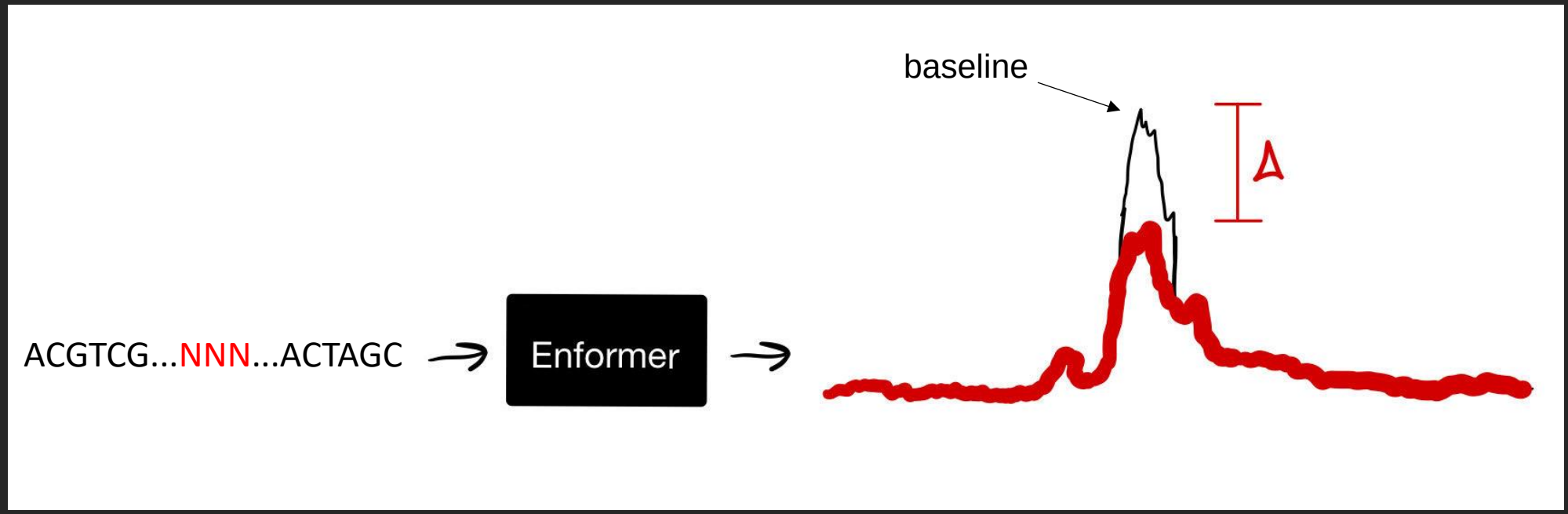
NNNNNNNNNNNNNNNNNN … GATCAGTTTAGCCAAAAA

iteration 3072:

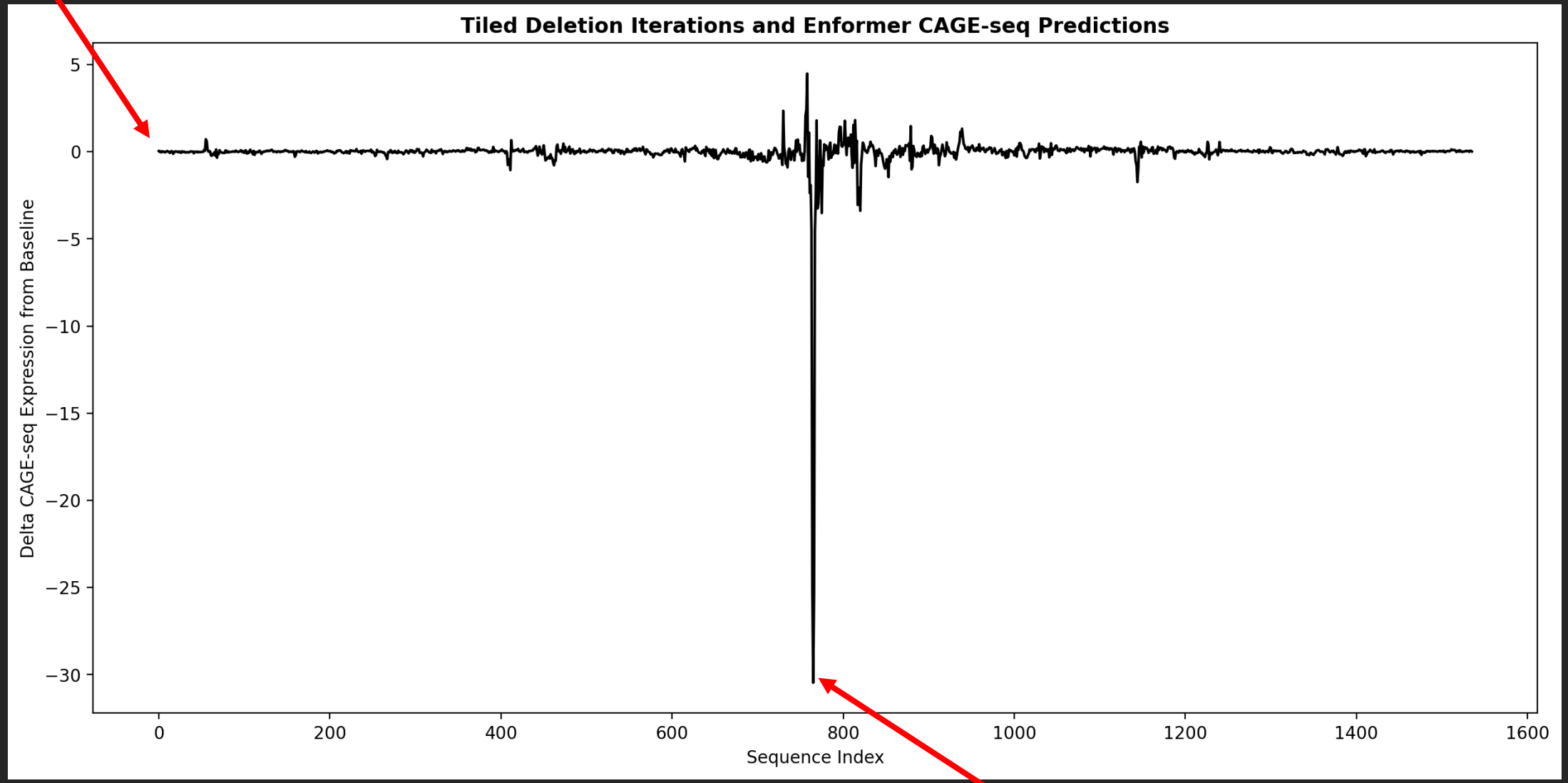
AATGCCCTGACTGACGTAC … NNNNNNNNNNNNNNNNN

Recreating CRISPR tiling screen *in silico*

- Input iterations into Enformer and output the predicted CAGE-seq expression at the TSS
- Compute change in expression (Δ) from baseline TSS

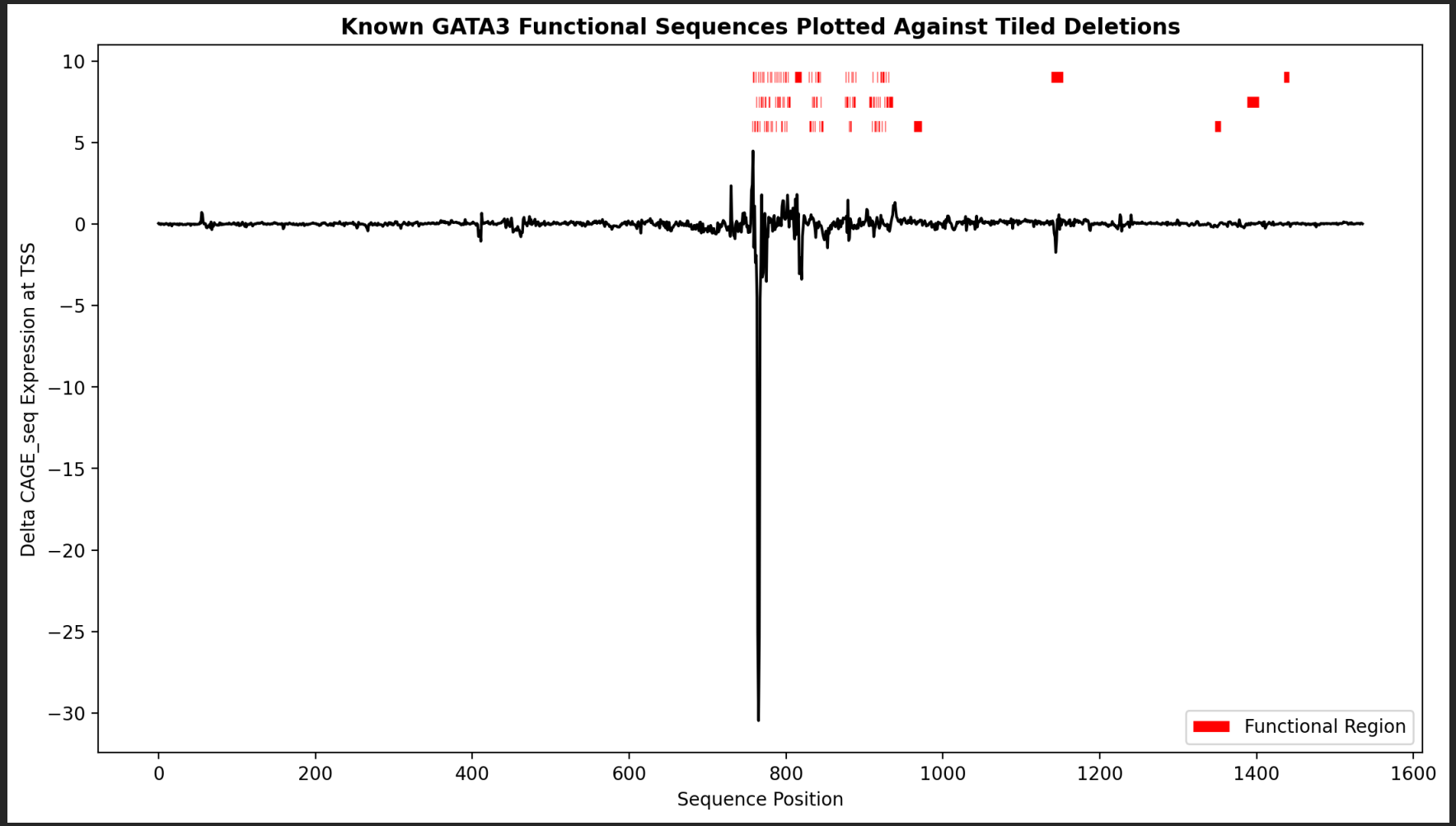


Iteration 1:
NNNNNNNNNNNNNNNNNNNNNNNN … ATGAGACCCATAGAGGGATAGACAGATG



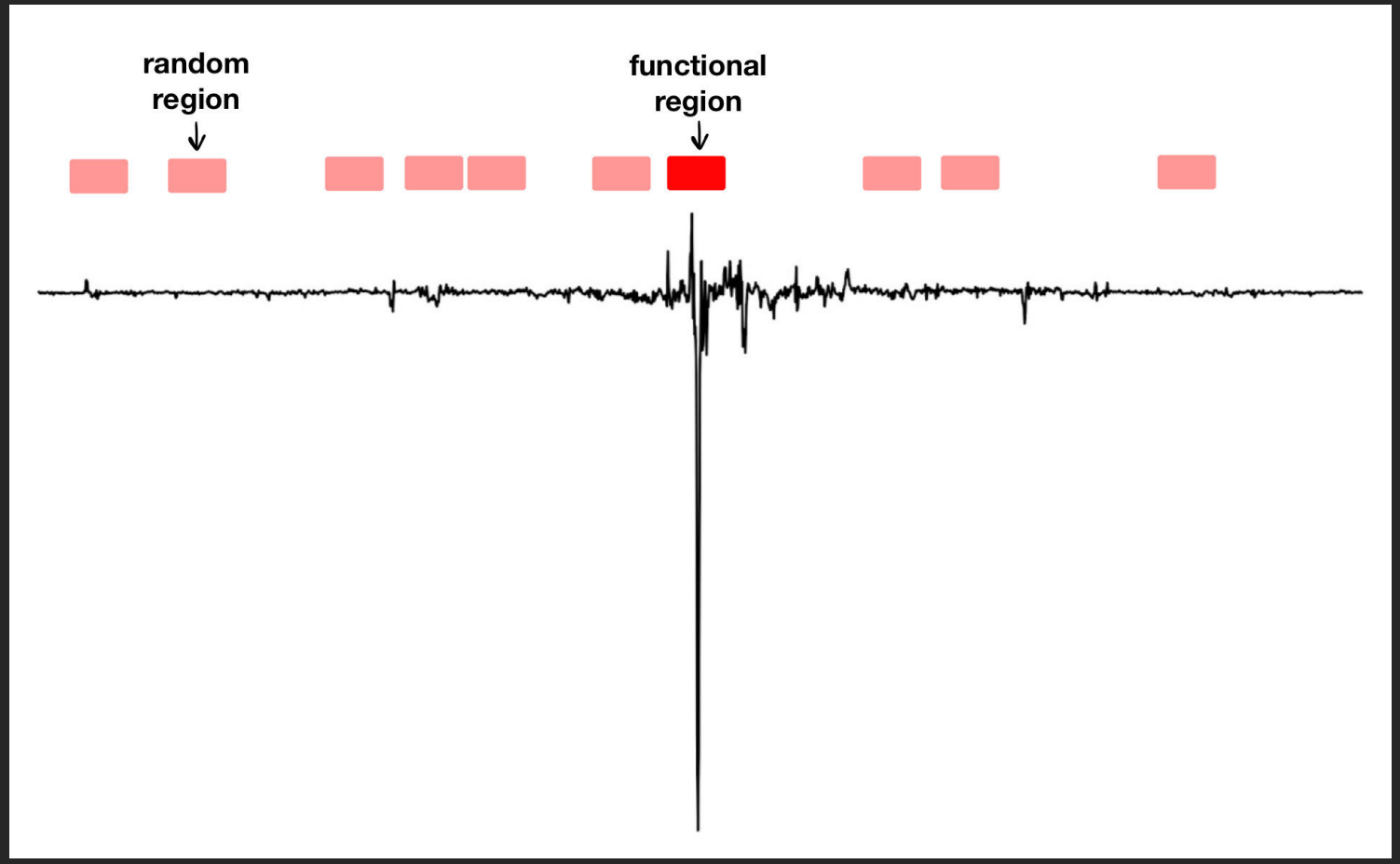
Iteration 1533 (replaced TSS with N's)**:**
AGAGTACCCAGAT… NNNNNNNNNNNNNNNNNNNNNNNN … GGGATAGACAGAT

Can *in silico* mutagenesis validate CRISPR tiling functional sequences?



Can *in silico* mutagenesis validate CRISPR tiling functional sequences?

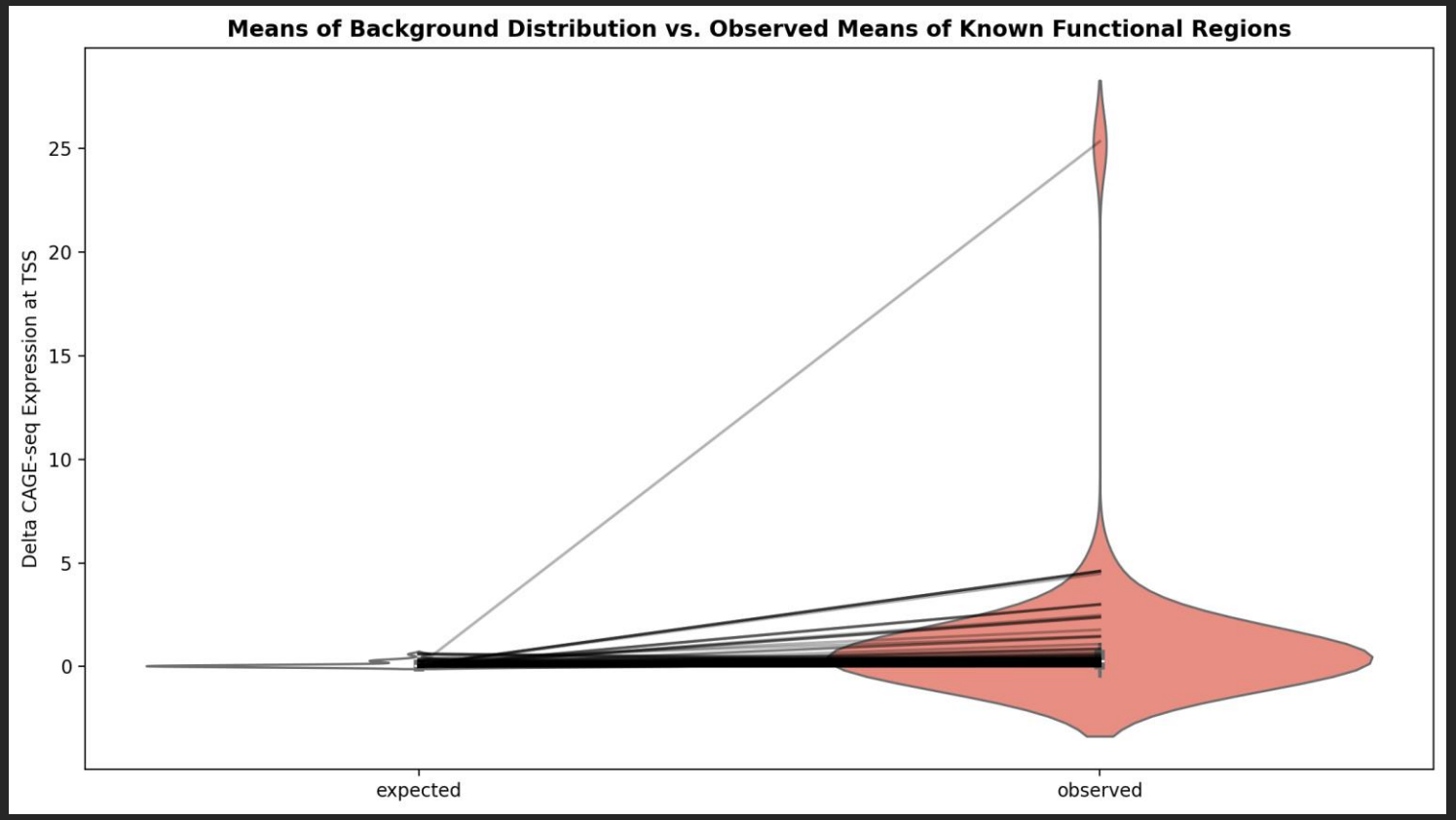
- Generate a background distribution
- This is what we would expect to see given noise of data



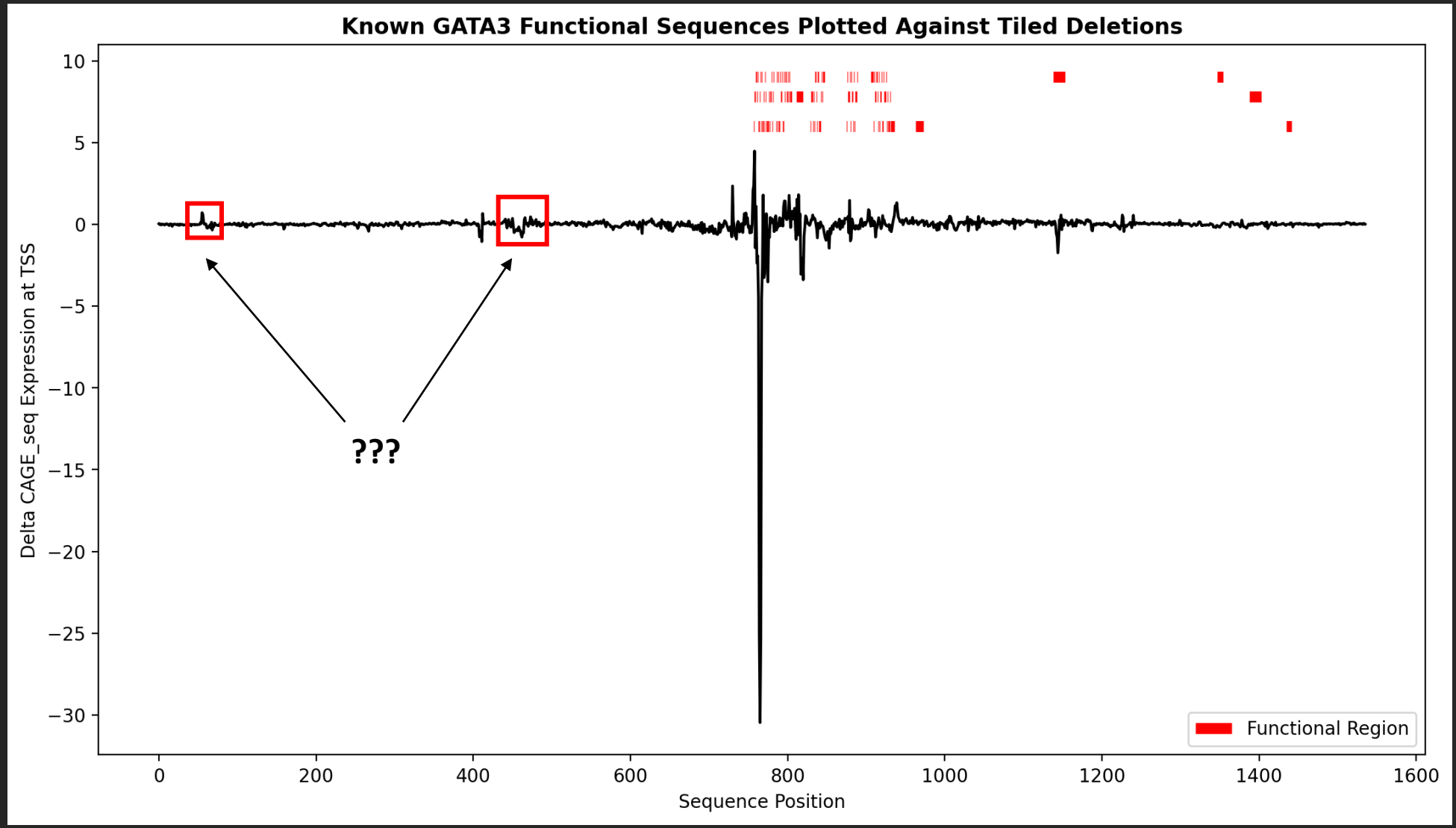
Hypothesis testing

Null: Background and functional sequences have equal effects on gene expression
Alternative: Background and functional sequences have significantly different effects on gene expression

- Paired t-test
- p value = 0.00556

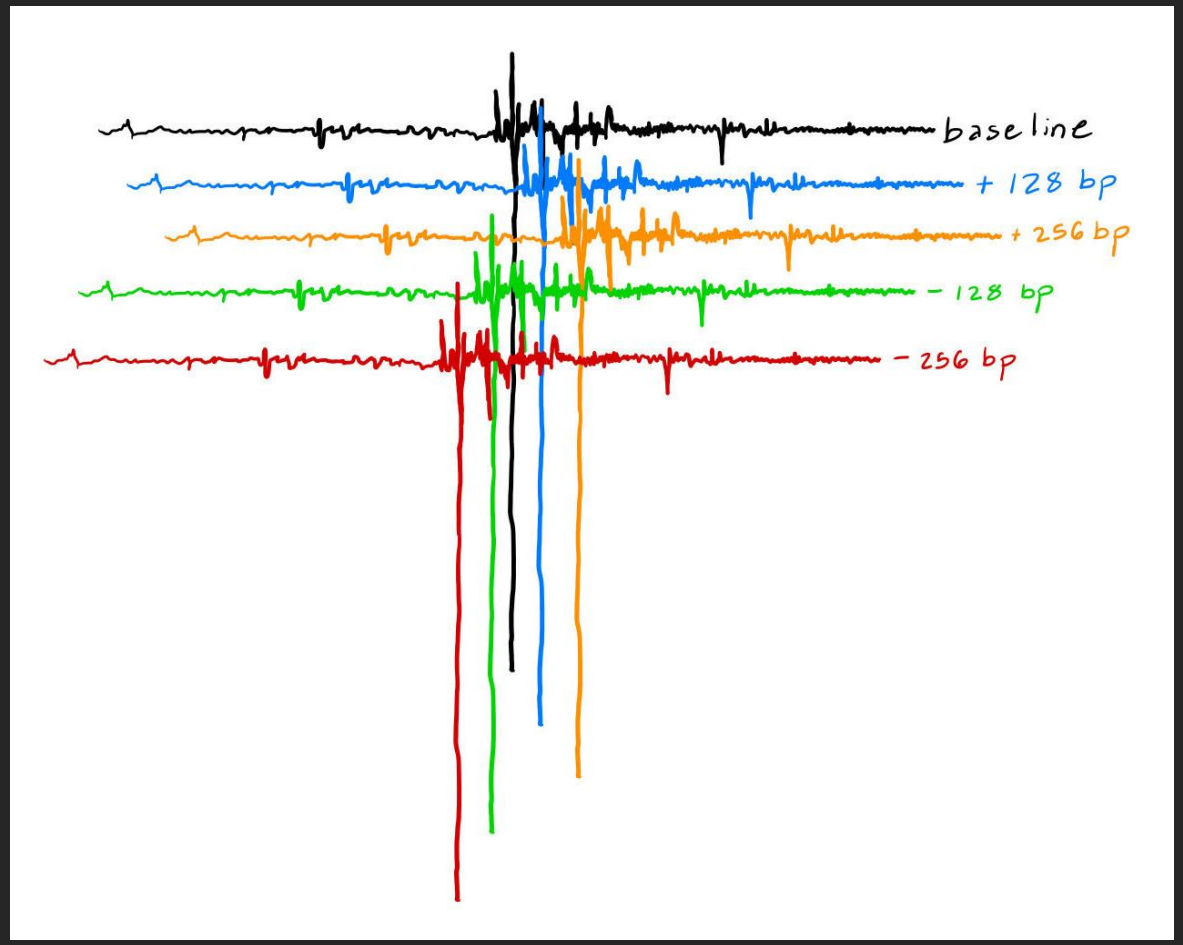


Are there uncharacterized functional sequences?

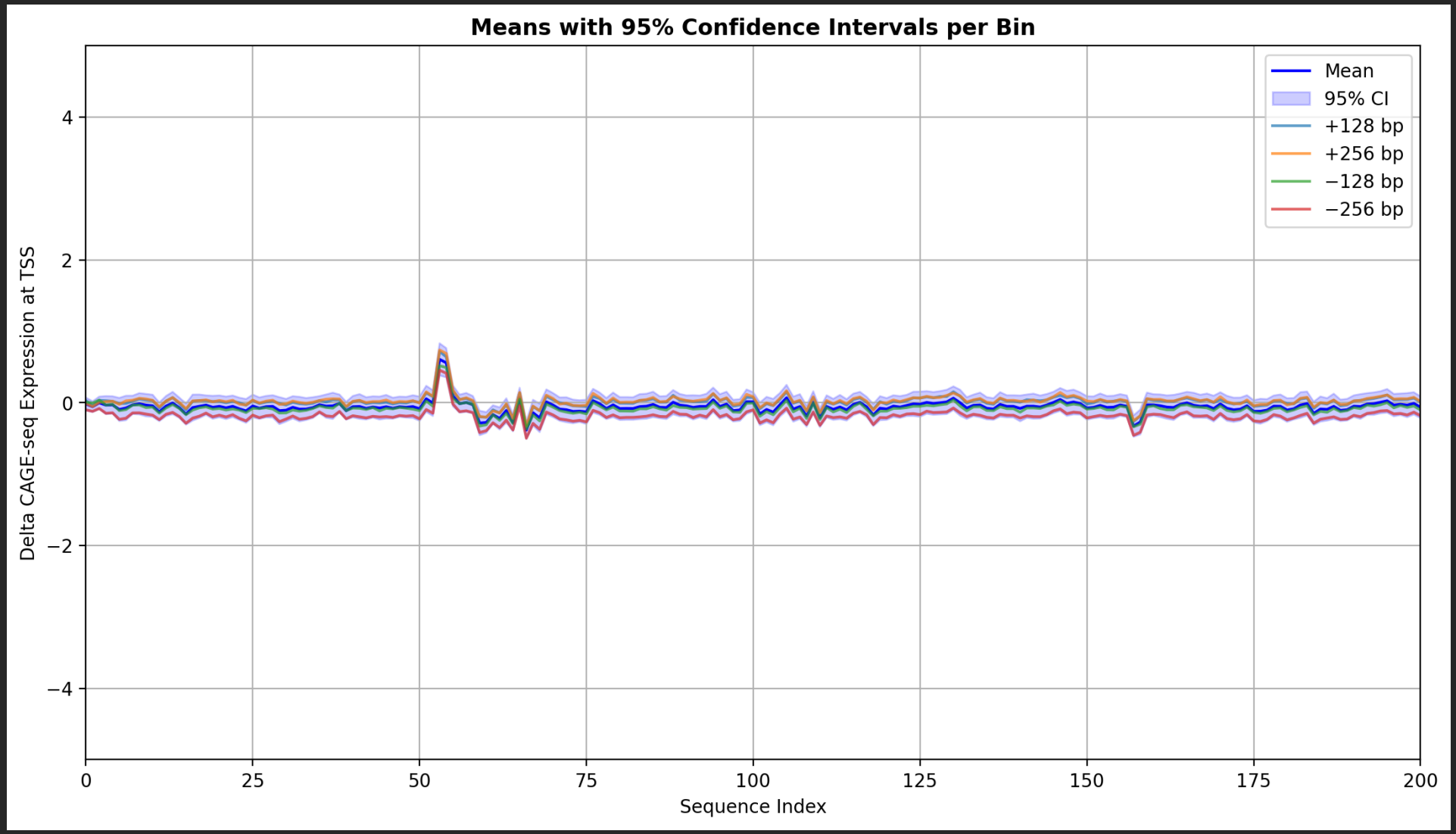


Jittered data

- Variation in context quantifies uncertainty in model's predictions

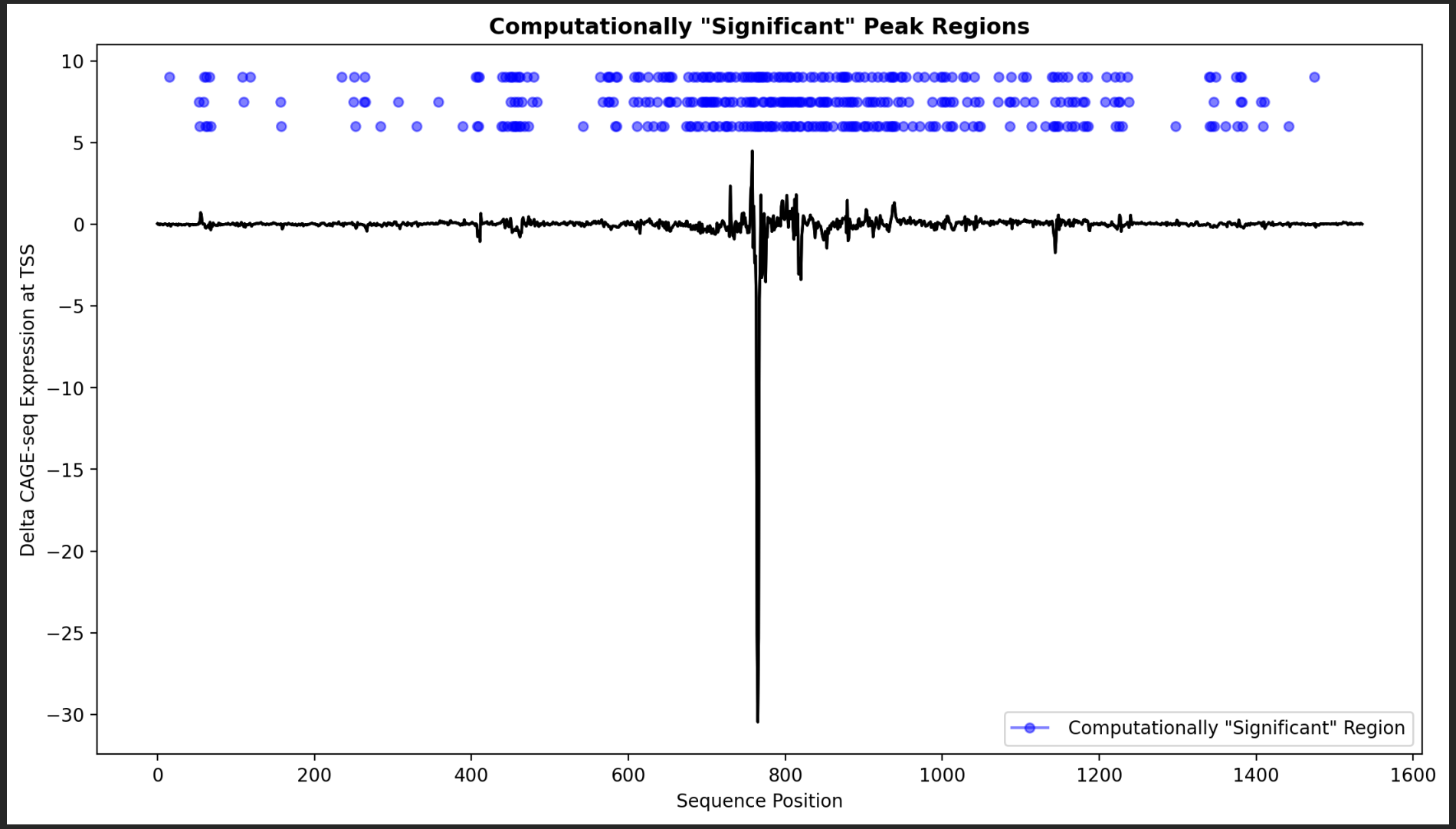


Computing significance of peaks



Null = 0

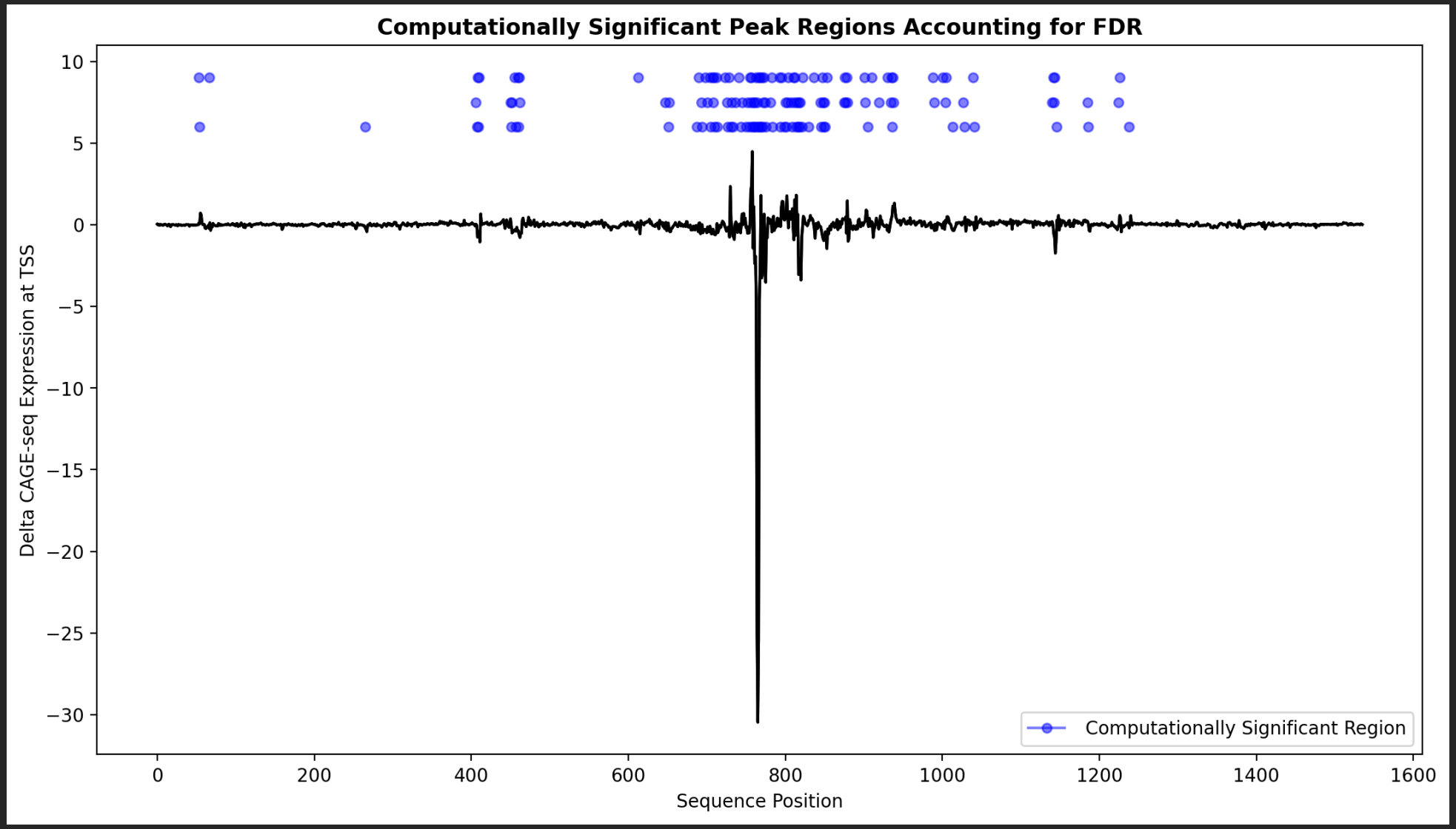
Computing significance of peaks



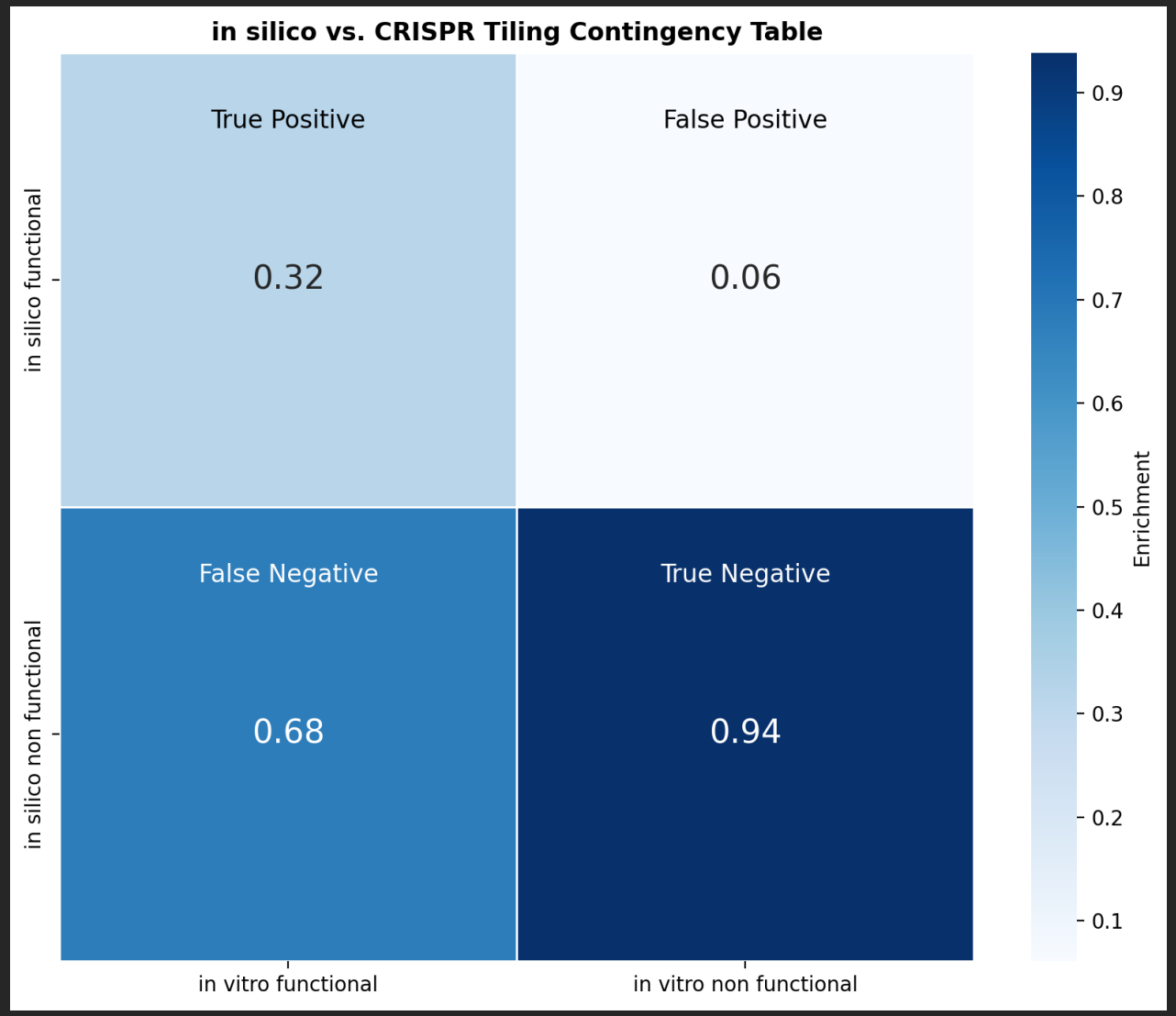
False discovery rate

- Conducting multiple hypothesis tests increases the likelihood of false positives
- FDR accounts for the number of significant values we would expect to see

False discovery rate



Did we recreate the CRISPR tiling experiment?

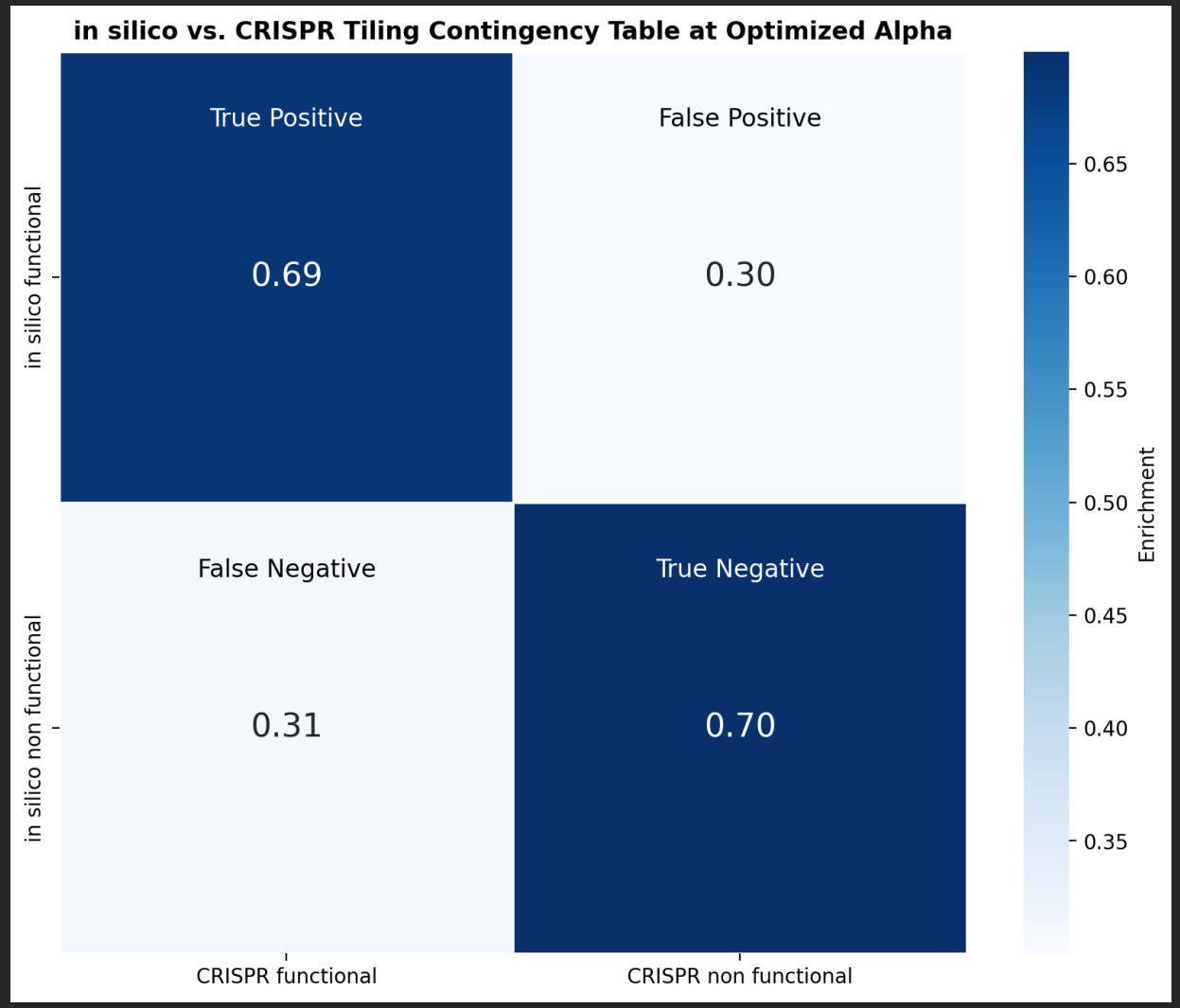


- 7-fold enrichment in determining functional sequences under *in silico* conditions vs. CRISPR tiling conditions

  p-value = 1.9×10^{-21}

- Recreated 32% of true positives at an FDR cutoff of 0.05

Intersection of sensitivity and specificity

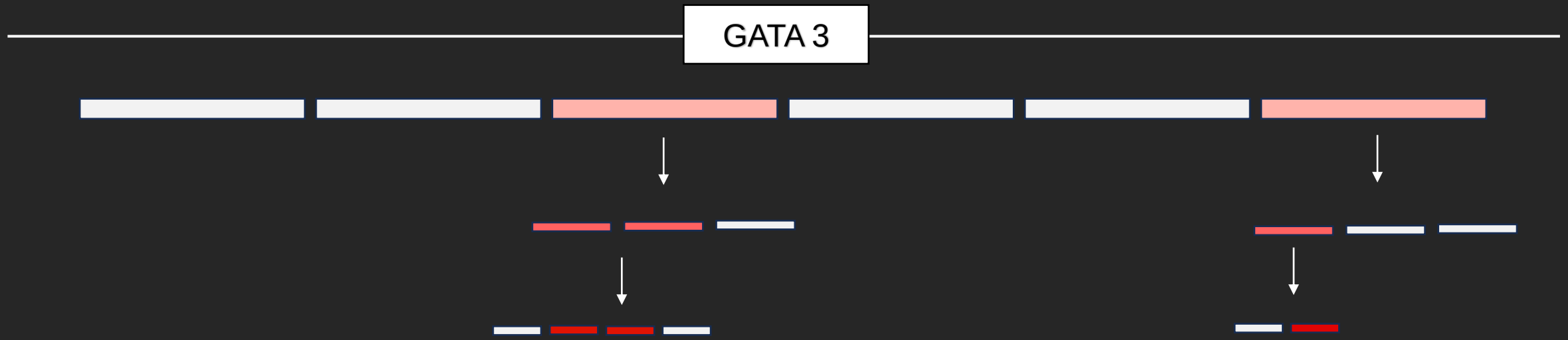


- 5-fold enrichment at $\alpha = 0.3$

  p-value = 7.7×10^{-24}

Future work

- Experimentally validate computationally determined functional sequences using CRISPR tiling
- Incorporate other prediction tracks for a more nuanced analysis of gene expression
- Incorporate smaller tiling bins and increase resolution of significant regions to eventually get single base pair measurements



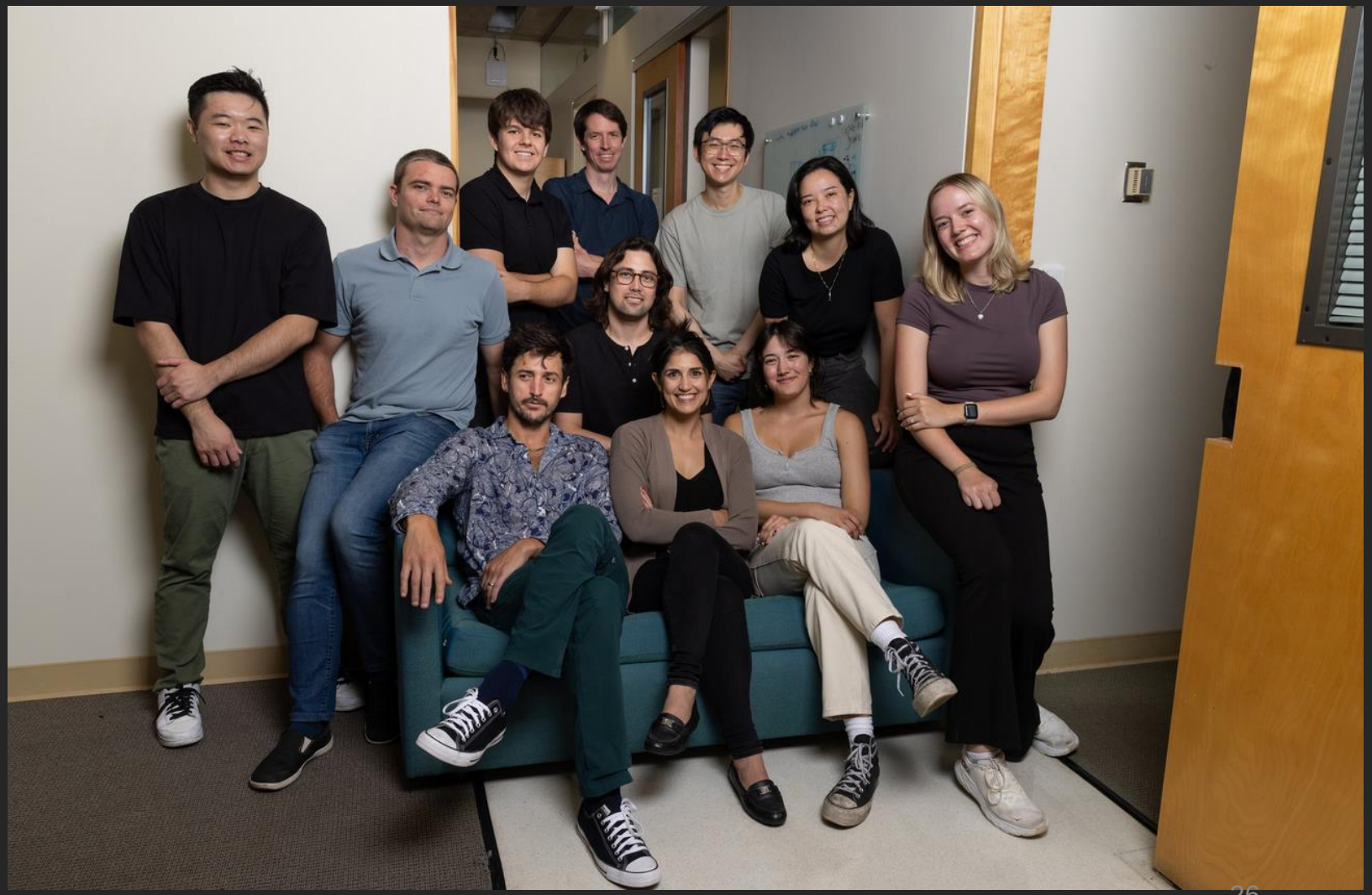
Acknowledgments

McVicker Lab

- **Graham McVicker (PI, PhD)**
- **Brad Balderson (PhD)**
- Jeff Jaureguy
- Shanna Lavalle (PhD)
- Mickey Lorenzini
- Elise Marvin
- Ariana Fonseca
- Aaron Ho
- Chris Duroiu
- Han Chen
- Kiki Spaulding

SURF Coordinators

- Sophie Bales
- Monica Miller



Extra slides

CAGE-seq

In vitro identification of a gene's transcription start site (TSS)

1. Isolate mRNA (transcribed sequence)
2. Biochemically modify the 5' cap (site of protein synthesis initiation)
3. Modified RNA is reverse transcribed → cDNA
4. Sequence is aligned to reference genome

Hypothesis testing

Null: Background and functional sequences excluding TSS will have equal effects on gene expression
Alternative: Background and functional sequences excluding TSS will have significantly different effects on gene expression

- Paired t-test
- p value = 3.49 e^{-7}