GEMS: Final Presentation

Operationalizing a ‘functional potential score’ (FPS) to prioritize GWAS variants for experimental studies

Dhruv Jain (jaind3@mskcc.org)
Mentor: Dr. Matthew F. Buas,
Date: August-10-2023

Memorial Sloan Kettering Cancer Center

Outline

- Introduction
- Approach
- Results
- Conclusion

Introduction

Background of Research

Genome-wide association studies (GWAS)

- GWAS identifies genetic links to diseases like cancer, diabetes, cardiovascular disease.
- Researchers analyze markers in individuals, revealing genetic variants tied to susceptibility.
- This aids risk assessment, prevention, and treatment strategies.

What is problem?

- Statistical associations in GWAS do not reveal biological mechanisms

- Most GWAS signals are located in “non-coding regions”, suggesting regulatory functions.

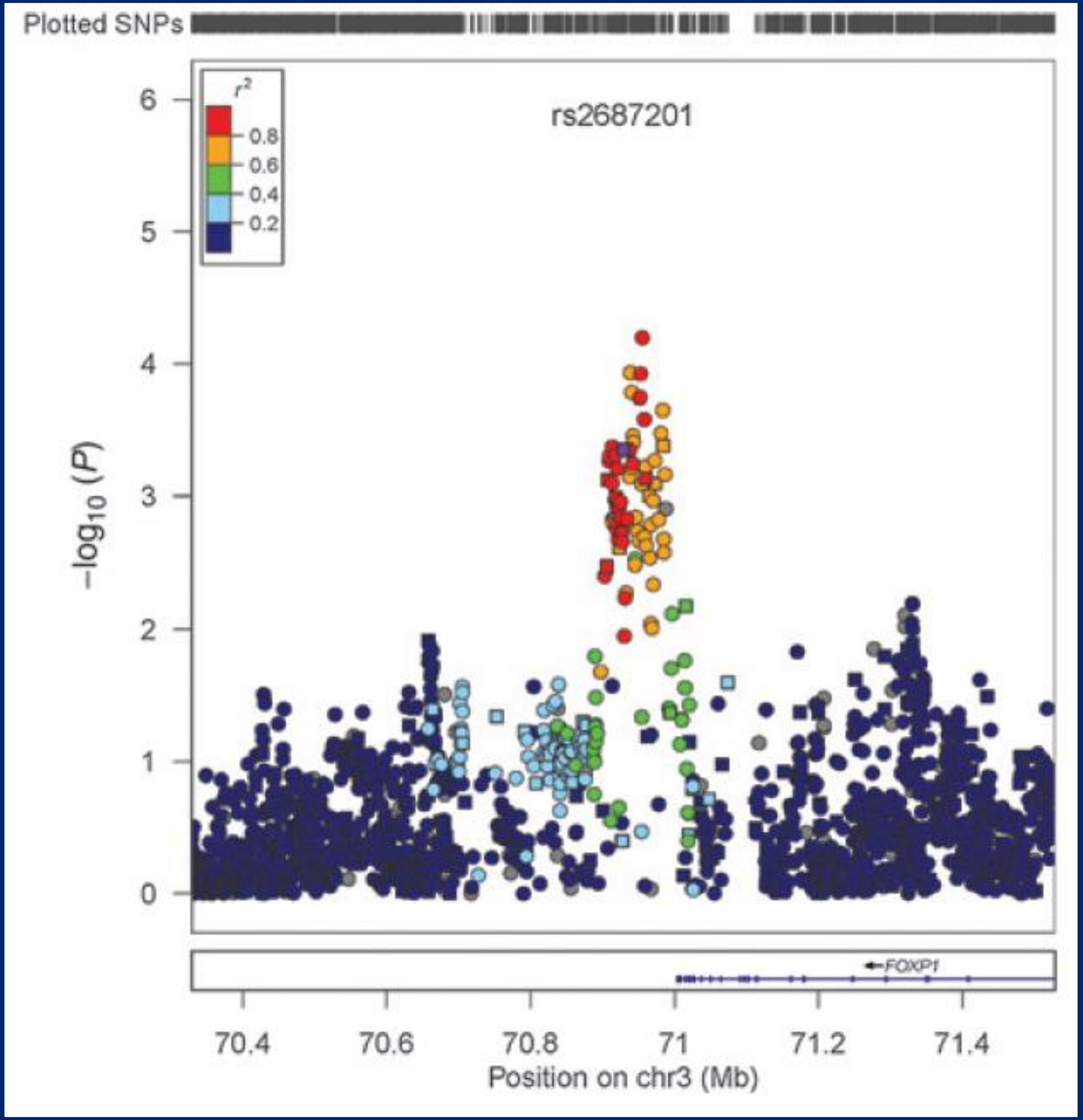
- The lead SNP in each GWAS locus is often correlated with multiple other variants, making it difficult to determine the true functional or causal signal.

- Laboratory follow-up is time-consuming, labor-intensive, and costly

Tools Exists like Haploreg, RegulomeDB, FUMA, & FORGE
Limitations:

- Incorporation of new or updated data inputs
- Customization for prioritizing annotation calls in specific tissues.
- Weighting and aggregation of inputs into a composite score.

Associated variant set (AVS)

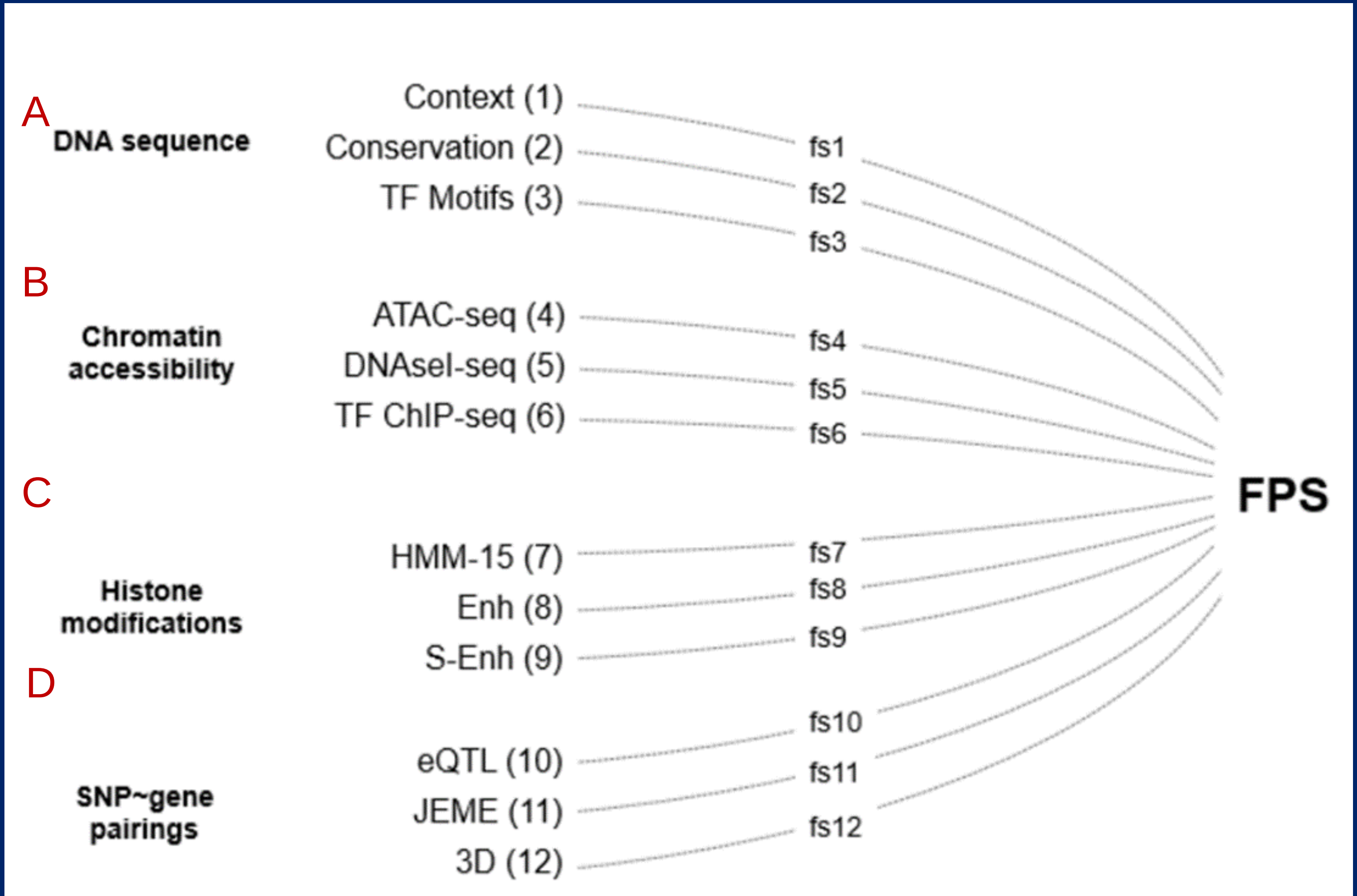


- Lead SNP (peak) often correlated with multiple other neighboring SNPs
- Any of these variants could be the functional /causal signal underlying the association

Full form
SNPs → Single Nucleotide Polymorphism

Approach

Function Potential Score (FPS)



A DNA sequence

Context (1) — fs1
Conservation (2) — fs2
TF Motifs (3) — fs3

B Chromatin accessibility

ATAC-seq (4) — fs4
DNAseI-seq (5) — fs5
TF ChIP-seq (6) — fs6

C Histone modifications

HMM-15 (7) — fs7
Enh (8) — fs8
S-Enh (9) — fs9

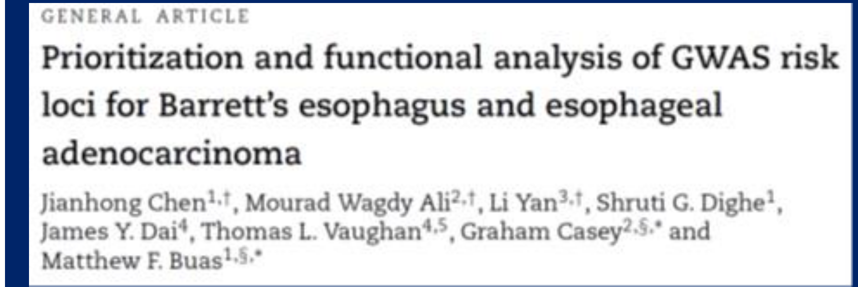
D SNP~gene pairings

eQTL (10) — fs10
JEME (11) — fs11
3D (12) — fs12

FPS

A composite SNP-level metric to summarize overall evidence for functional/regulatory potential.

A. DNA sequence
B. Chromatin Accessibility
C. Histone Modification
D. SNP ~ Gene Paring



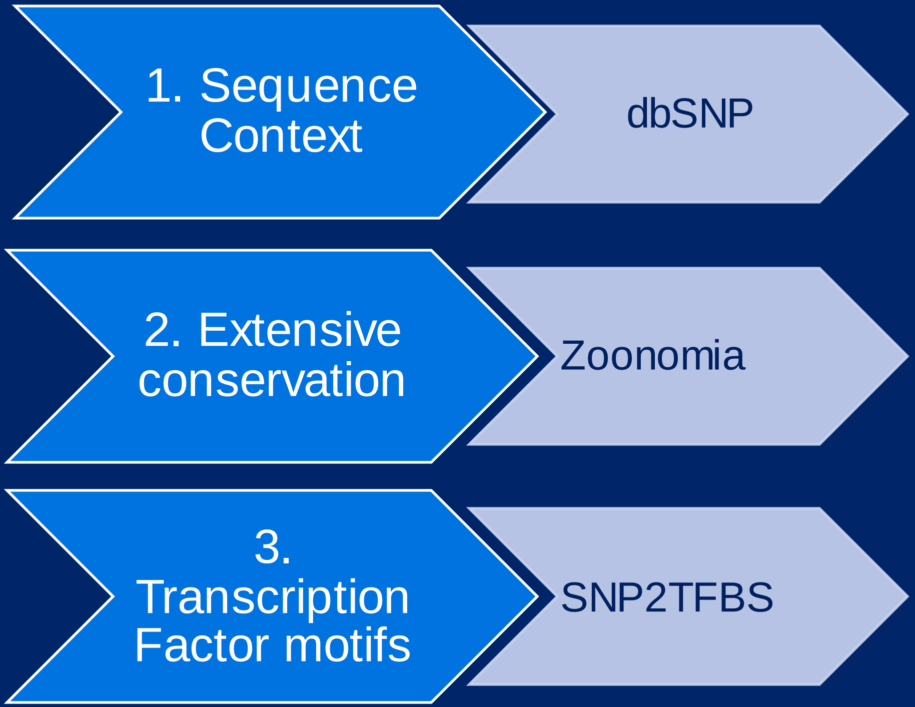
GENERAL ARTICLE

Prioritization and functional analysis of GWAS risk loci for Barrett's esophagus and esophageal adenocarcinoma

Jianhong Chen[1,†], Mourad Wagdy Ali[2,†], Li Yan[3,†], Shruti G. Dighe[1], James Y. Dai[4], Thomas L. Vaughan[4,5], Graham Casey[2,§,*] and Matthew F. Buas[1,§,*]

Chen, J., Ali, M. W., Yan, L., Dighe, S. G., Dai, J. Y., Vaughan, T. L., Casey, G., & Buas, M. F. (2021). Prioritization and functional analysis of GWAS risk loci for Barrett's esophagus and esophageal adenocarcinoma. In Human Molecular Genetics (Vol. 31, Issue 3, pp. 410–422). Oxford University Press (OUP). https://doi.org/10.1093/hmg/ddab259

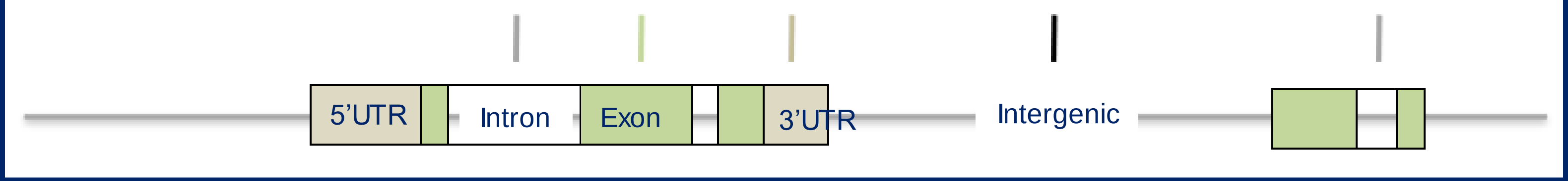
Part A ~ SEQUENCE

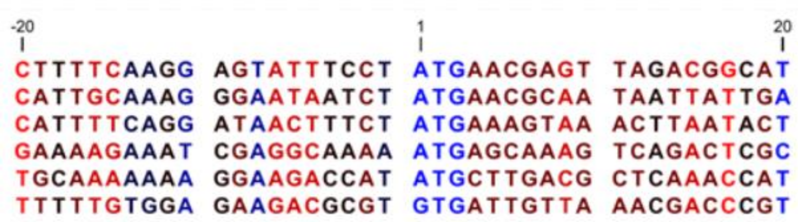


1. Sequence Context — dbSNP

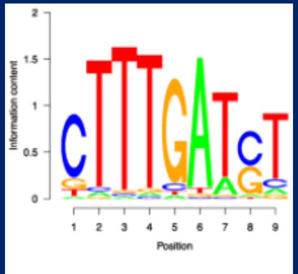
2. Extensive conservation — Zoonomia

3. Transcription Factor motifs — SNP2TFBS

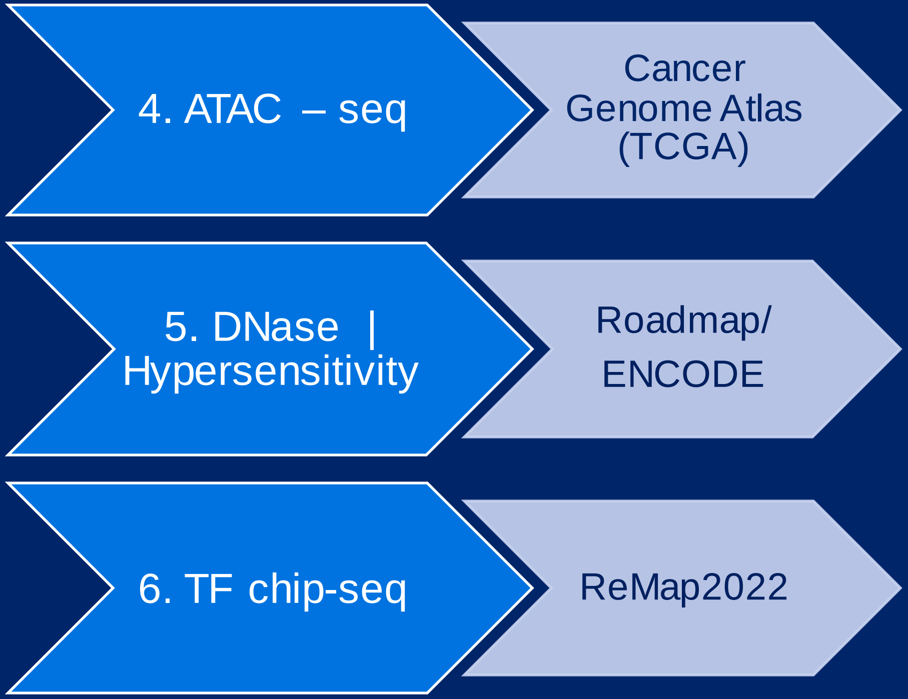




```
-20                                    1                                     20
CTTTTCAAGG  AGTATTTCCT  ATGAACGAGT  TAGACGGCAT
CATTGCAAAG  GGAATAATCT  ATGAACGCAA  TAATTATTGA
CATTTTCAGG  ATAACTTTCT  ATGAAAGTAA  ACTTAATACT
GAAAAGAAAT  CGAGGCAAAA  ATGAGCAAAG  TCAGACTCGC
TGCAAAAAAA  GGAAGACCAT  ATGCTTGACG  CTCAAACCAT
TTTTTGTGGA  GAAGACGCGT  GTGATTGTTA  AACGACCCGT
```



Part B ~ CHROMATIN ACCESSIBILITY



4. ATAC – seq

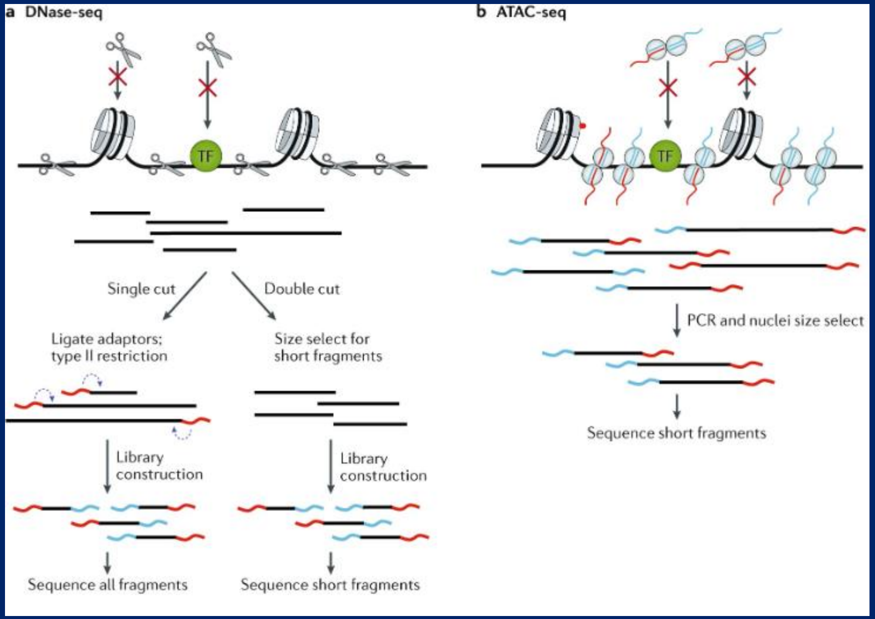
Cancer Genome Atlas (TCGA)

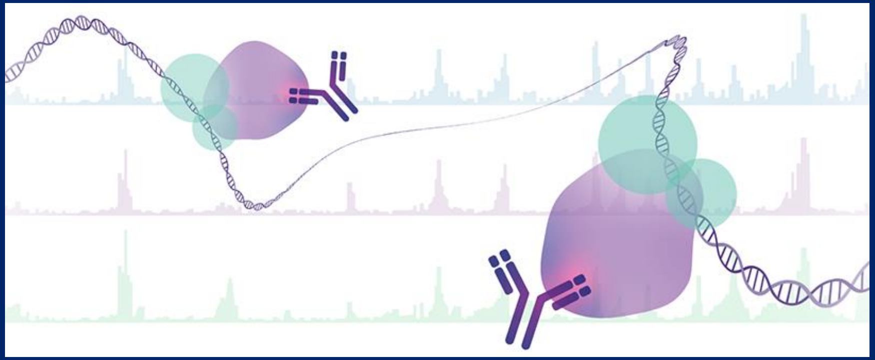
5. DNase | Hypersensitivity

Roadmap/ ENCODE

6. TF chip-seq

ReMap2022



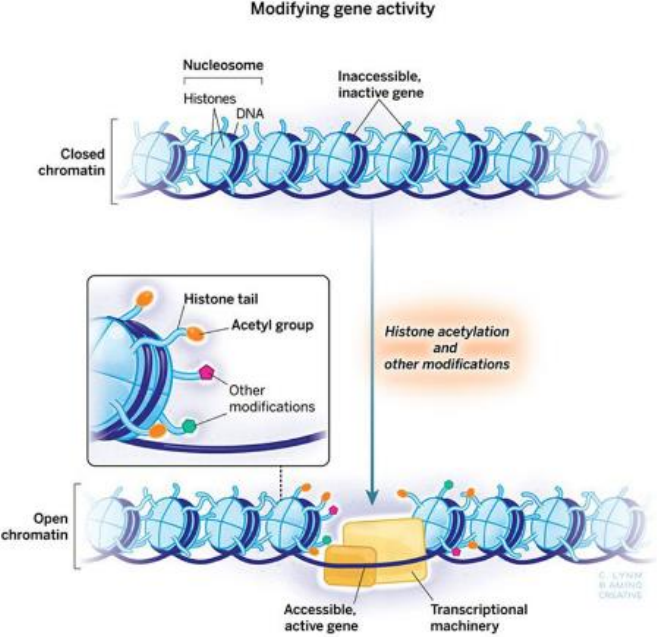


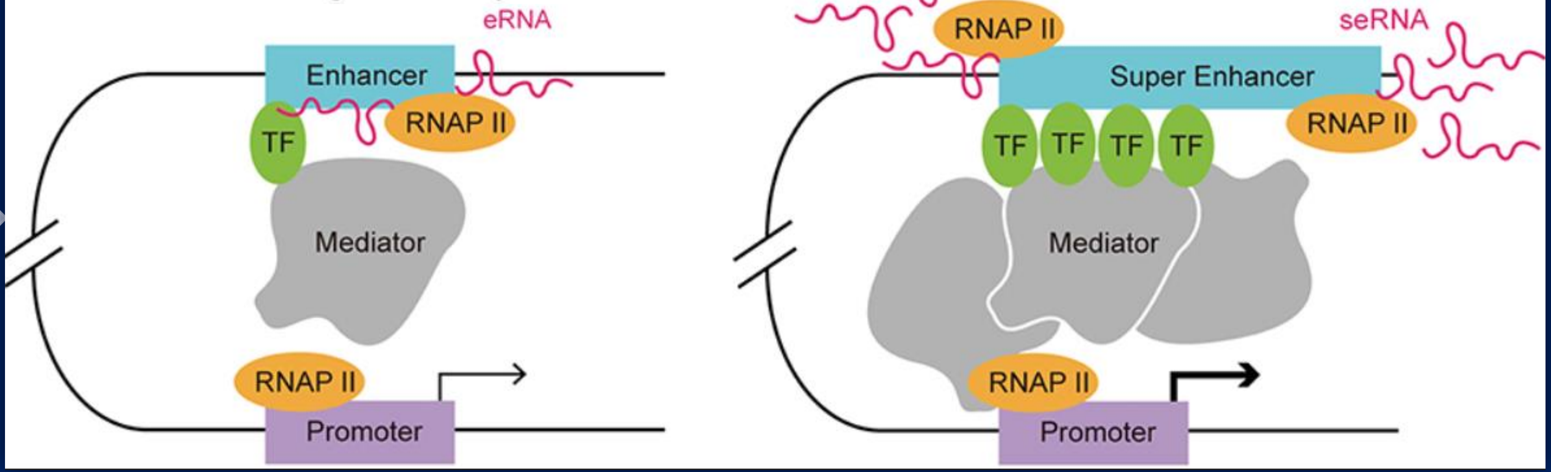
Part C ~ HISTONE MODIFICATIONS

7. HMM -15 — Roadmap/ ENCODE

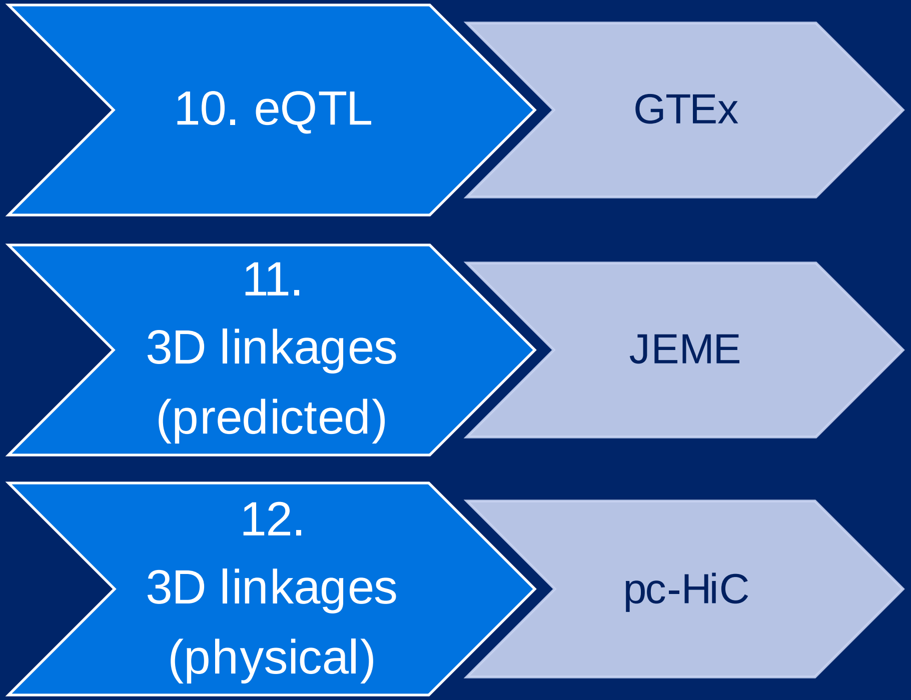
8. Enhancer — Roadmap/ ENCODE

9. Super Enhancer — Roadmap/ ENCODE


Modifying gene activity



Part D SNP ~ GENE PAIRINGS



10. eQTL

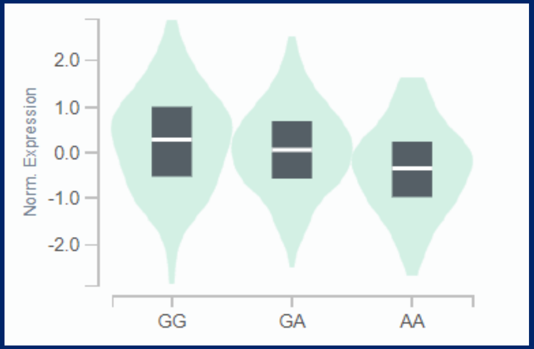
GTEx

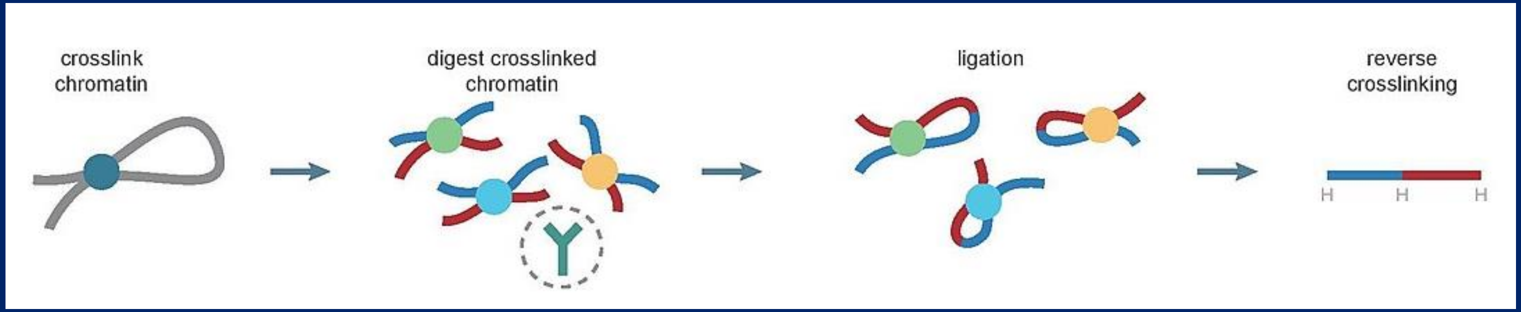
11.
3D linkages
(predicted)

JEME

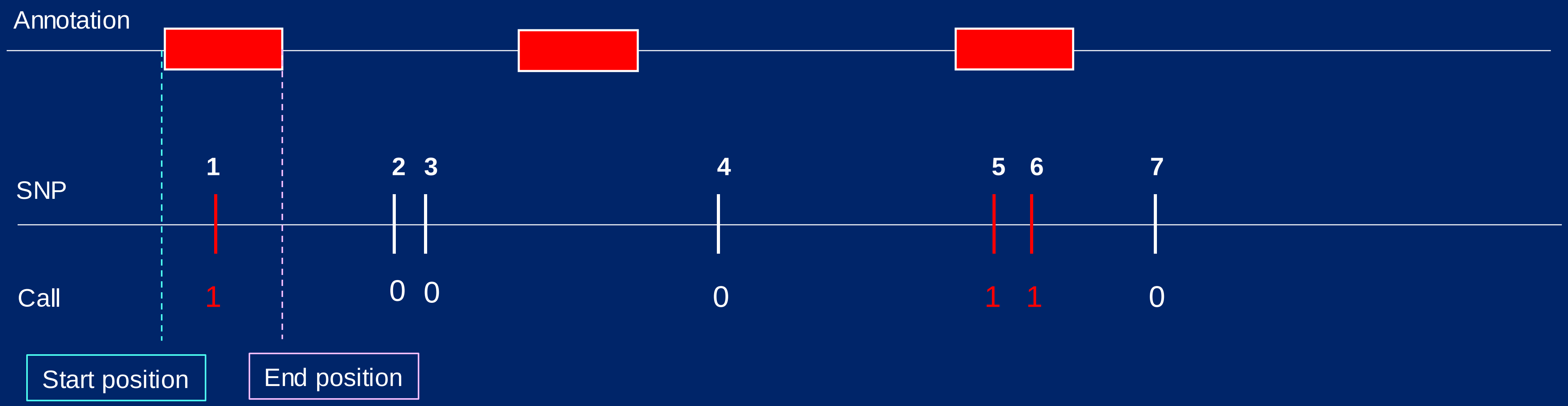
12.
3D linkages
(physical)

pc-HiC





Assigning 0/1 calls to SNPs ~ overlap with annotation windows



SNP	Call
1	1
2	0
3	0
4	0
5	1
6	1
7	0

- Browser Extensible data .BED
- Chr, start, end position.

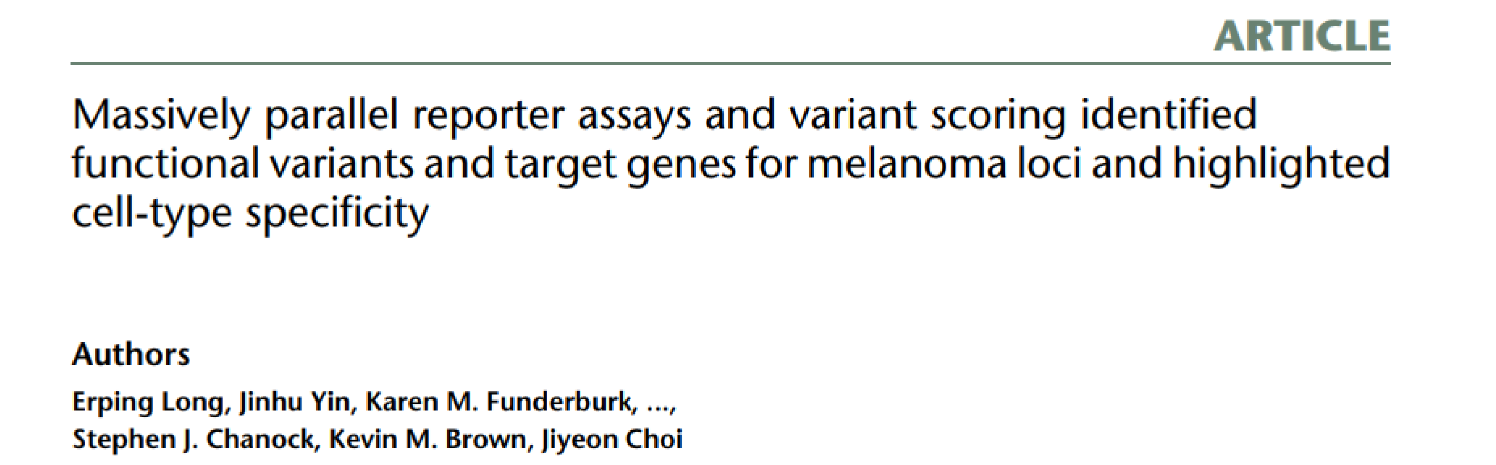
Theoretical Framework

Feature	Call	Value	Weight	Feature Score (FS)
1	c1	0/1	W1	W1 × c1
2	c2	0/1	W2	W2 × c2
3	c3	0/1	W3	W3 × c3
4	c4_1	0/1	W4_1	W4_1 × c4_1 + W4_2 × c4_2 + W4_3 × c4_3
	c4_2	0/1	W4_2	
	c4_3	0/1	W4_3	
5	c5_1	0/1	W5_1	W5_1 × c5_1 + W5_2 × c5_2 + W5_3 × c5_3
	c5_2	0/1	W5_2	
	c6_3	0/1	W5_3	
.. .. ..				
12	c12_1	0/1	W12_1	W12_1 × c12_1 + W12_2 × c12_2 + W12_3 × c12_3
	c12_2	0/1	W12_2	
	c12_3	0/1	W12_3	

[K,O]	Call	Value
[1,0]	ci_1	1
	ci_2	0
	ci_3	0
[1,1]	ci_1	0
	ci_2	1
	ci_3	0
[0,1]	ci_1	0
	ci_2	0
	ci_3	1
[0,0]	ci_1	0
	ci_2	0
	ci_3	0

$$\boxed{FPS = \sum_{j=0}^{j=3} W_j \cdot c_j + \sum_{j=4}^{j=12} \{(W_{j\_1} \cdot c_{j\_1}) + (W_{j\_2} \cdot c_{j\_2}) + (W_{j\_3} \cdot c_{j\_3})\}}$$

Experimental data for model building



ARTICLE

Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity

Authors

Erping Long, Jinhu Yin, Karen M. Funderburk, ..., Stephen J. Chanock, Kevin M. Brown, Jiyeon Choi

1. Massively parallel reporter assay (MPRA) data on 1,992 variants from 54 melanoma GWAS loci
2. 285 SNPs showed allelic-specific transcriptional activity in melanoma/melanocyte cell lines (FDR<0.01)
3. Smaller subsets of the 285 SNPs showed stronger allelic differences (%): 114 (≥10%) and 65 (≥15%)
4. SNPs randomly split to training (2/3) & testing sets (1/3)
5. MPRA functional calls for each SNP were merged with FPS calls and used for model building (training set)

Long, E., Yin, J., Funderburk, K. M., Xu, M., Feng, J., Kane, A., Zhang, T., Myers, T., Golden, A., Thakur, R., Kong, H., Jessop, L., Kim, E. Y., Jones, K., Chari, R., Machiela, M. J., Yu, K., Iles, M. M., Landi, M. T., … Choi, J. (2022). Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity. In The American Journal of Human Genetics (Vol. 109, Issue 12, pp. 2210–2229). Elsevier BV. https://doi.org/10.1016/j.ajhg.2022.11.006

Flow chart

1

Step 1:

Make feature calls for all SNPs

2

Step 2:

Use Training set to select optimal λ

(5-fold CV)

3

Step 3:

Build LASSO model using selected λ* (entire Training set)

4

Step 4:

Extract model coefficients

5

Step 5:

Run model using Testing data and assess performance

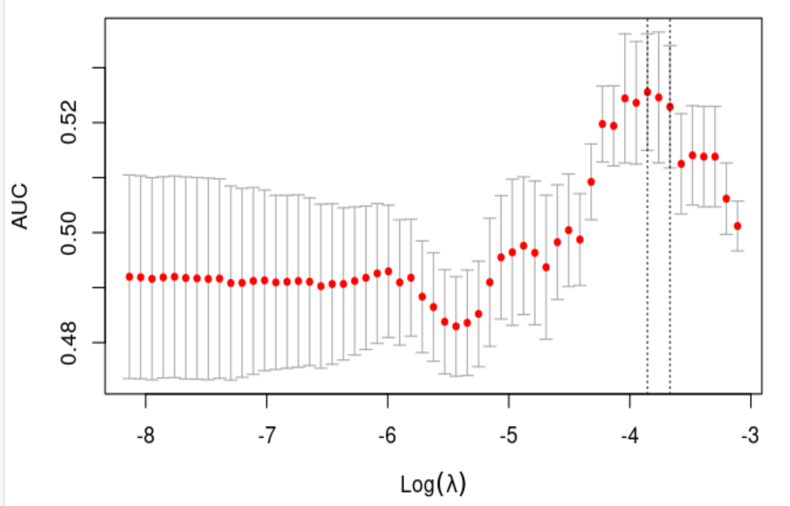
- λ (lambda) is the regularization parameter that controls the amount of shrinkage applied to the coefficients.
- Higher values of lambda lead to more aggressive shrinkage and a sparser model

- Use cross-validation to select optimal lambda
- Goal ~ reduce overfitting

Conclusion

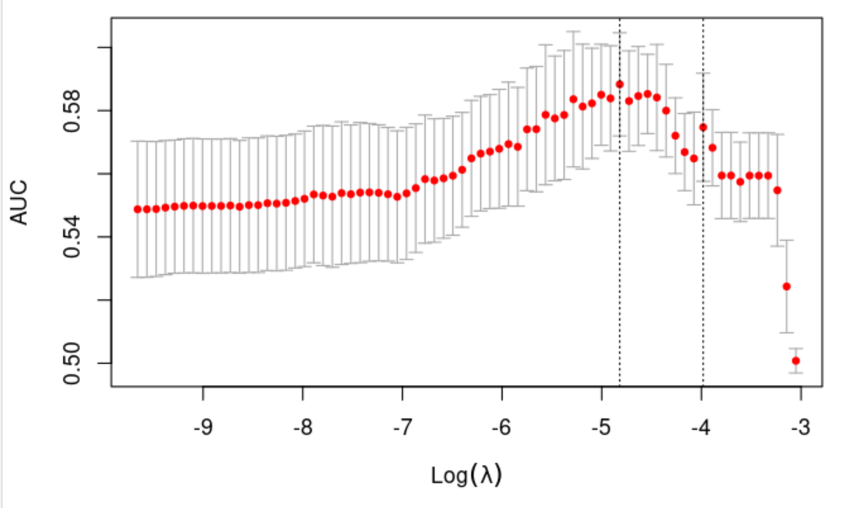
Lambda Selection

1



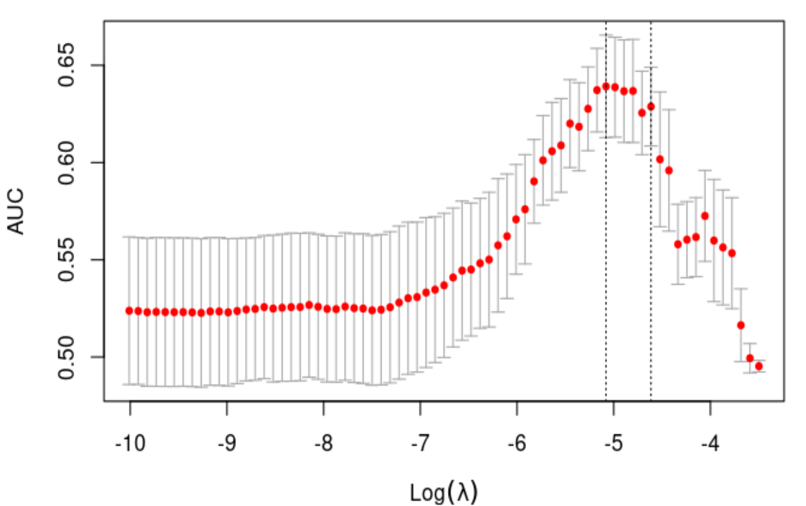
Optimal λ → 0.02118

2



Optimal λ → 0.0080

3

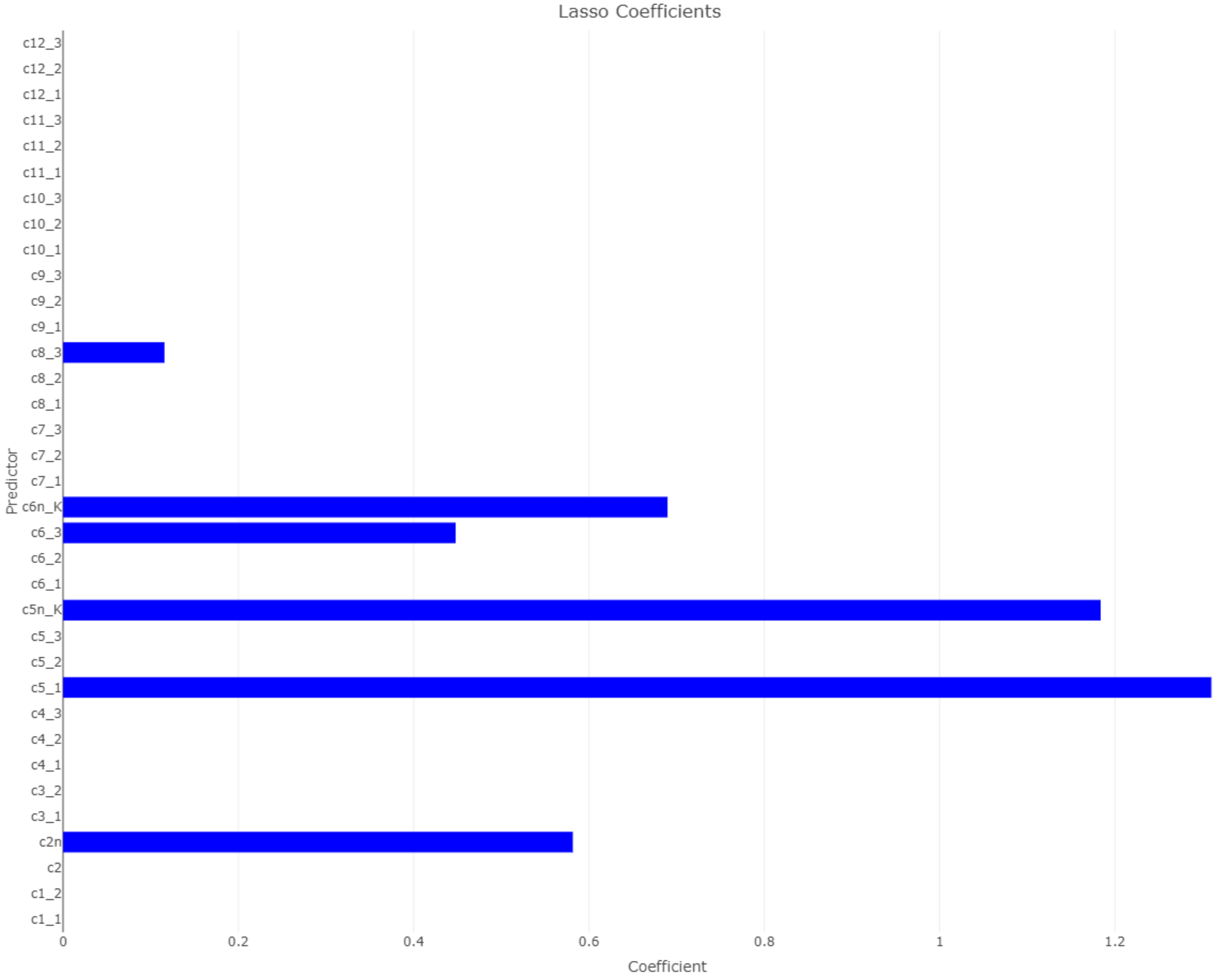


Optimal λ → 0.00621

Lasso Coefficients

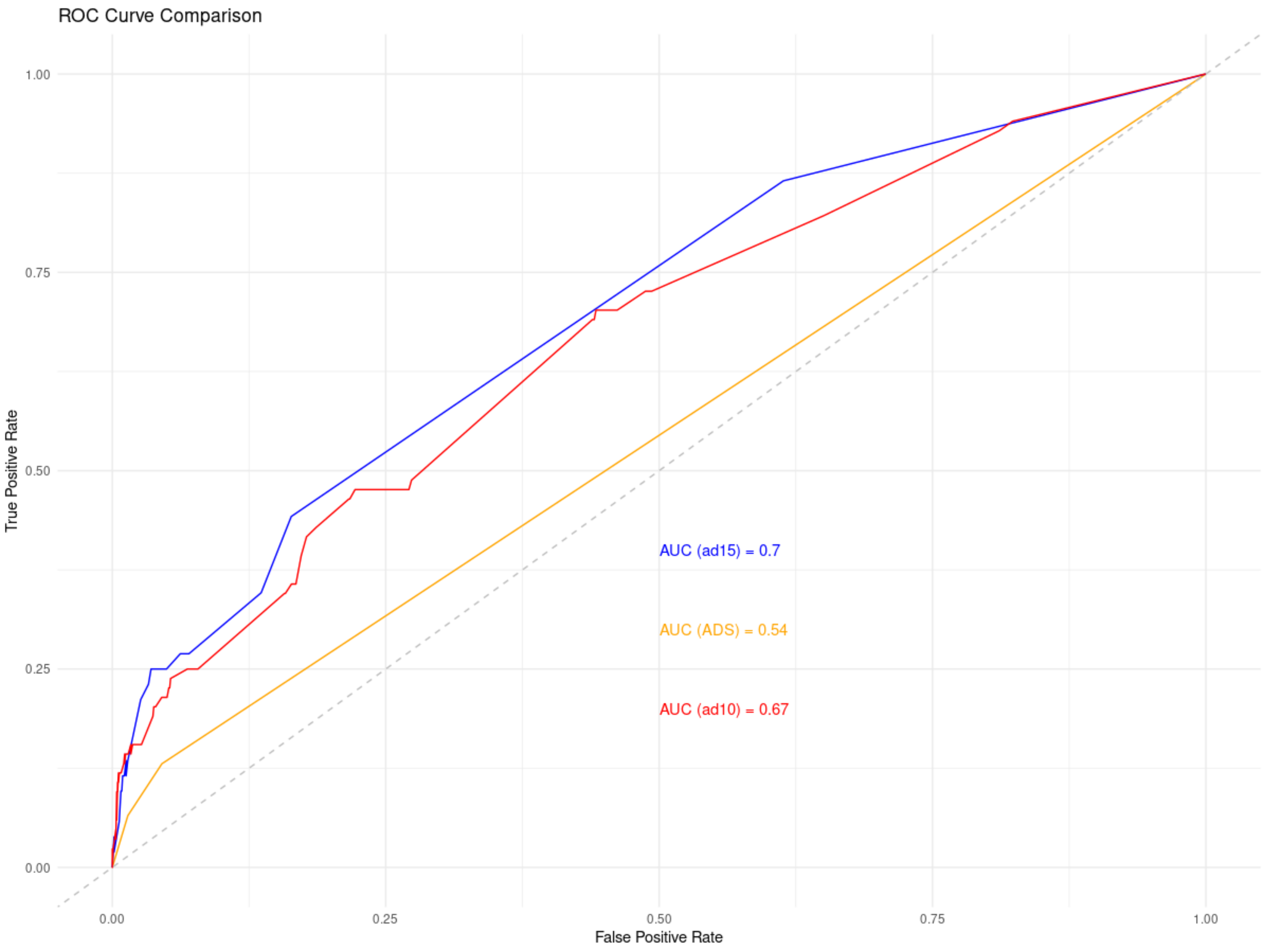
- LASSO model indicates that certain annotation data inputs were most predictive of experimentally-determined SNP regulatory function:

- Seq conservation (c2n)
- Dnase Seq (c5_1, c5n_K)
- TF-ChIP-seq (c6n_K, c6_3)



ROC CURVE (Receiver operating characteristic)

- AUC (ad15) = 0.71
- AUC (ad10) = 0.67
- AUC (ADS) = 0.54



Future Work

- Lab work can use this SNPs tell scientists whether you will react positively or not to a specific treatment.

- Identifying functional/causal variants and risk genes at GWAS loci remains a critical challenge for downstream clinical translation

- Collaborate with clinicians to access clinical samples and validate the associations between genomic regions and clinical outcomes in real patient populations.

Expressing Gratitude

Mentor: Ds. Matthew Buas

- Your guidance has been my guiding light, aiding me through challenges and growth.

Lab Assistant: Dr. Shiv Verma Prakash

- Your expertise and patience have been invaluable in my lab experiences.

Fellow Interns and their mentors

- Your camaraderie and support have enriched my perspectives and journey."

Thank you for your unwavering encouragement and contributions.

1884

Memorial Sloan Kettering
Cancer Center