OPERATIONALIZING A "FUNCTIONAL POTENTIAL SCORE" (FPS) TO PRIORITIZE GWAS VARIANTS FOR EXPERIMENTAL STUDIES

Memorial Sloan Kettering Cancer Center

Dhruv Jain (jaind3@mskcc.org)
Mentor: Dr. Matthew F. Buas

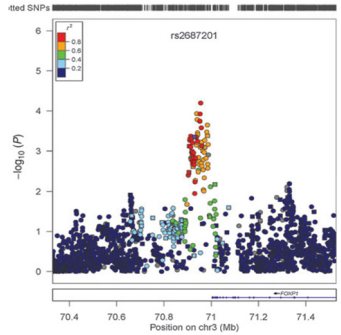
1. Introduction

A genome-wide association study (GWAS) is a powerful method that enables the identification of genetic variations linked to complex diseases such as cancer, diabetes, cardiovascular disease. By scanning a comprehensive set of markers across numerous individuals, researchers can pinpoint associations between specific genetic variants and disease susceptibility, paving the way for improved strategies in risk assessment, prevention, and treatment.

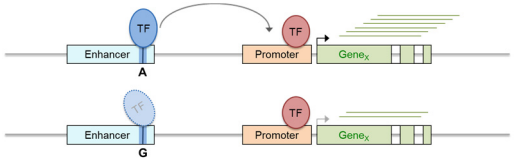
1.1 The Challenge:

While GWAS has proven successful in revealing statistical associations between genetic variants and risk of diseases, progress has been slower in identifying the functional/causal variants, risk genes, and molecular mechanisms underlying such associations. Most GWAS index variants reside in non-coding regions of the genome and are correlated with many other SNPs in linkage disequilibrium (LD), adding complexity to the search for true functional/causal signals. Laboratory-based experimental validation studies remain essential but time-consuming, labor-intensive, and costly.

1.2. Defining the "Associated Variant Set"



1.3. Non-coding regulatory SNPs ~ mechanisms



1.4. Existing Tools:

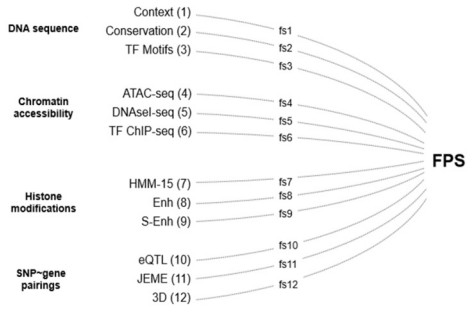
Several informatics resources have been developed to prioritize variants based on functional annotations (e.g., Haploreg, RegulomeDB, FORGE). However, these tools provide limited flexibility to end-users for:

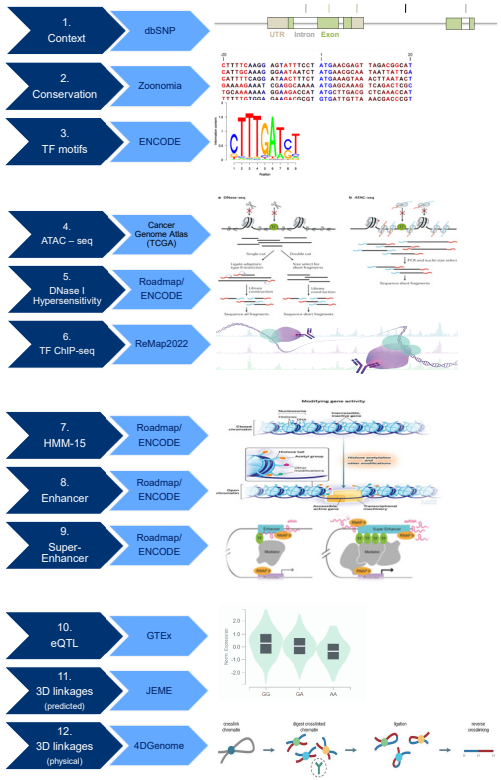
- Incorporating new or updated data inputs
- Prioritizing annotation calls in specific tissues
- Customizing weighting and aggregation of data inputs

2. Approach

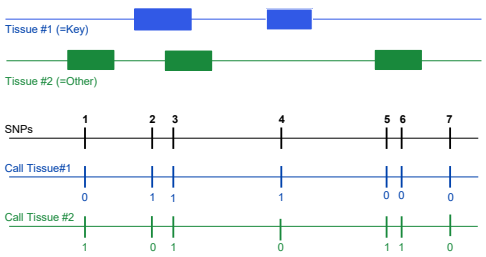
2.1 Function Potential Score (FPS):

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





2.2 Assigning SNP calls based on annotation overlap



2.3 LASSO Regression:

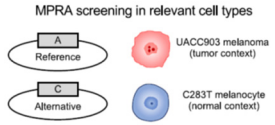
LASSO regression incorporates a penalty term (λ) to control the amount of shrinkage applied to model coefficients. Higher λ values result in more aggressive shrinkage and a sparser model. The optimal λ value is selected to maximize model performance and reduce overfitting using 5-fold cross-validation in the training set.

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	

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

2.4 Training / Testing Data

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

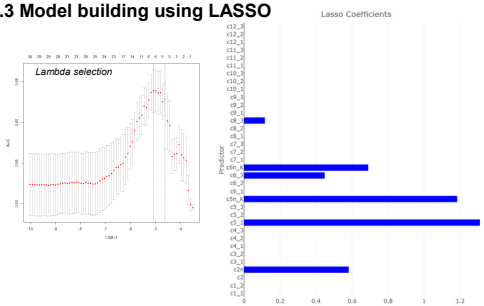


3. Results

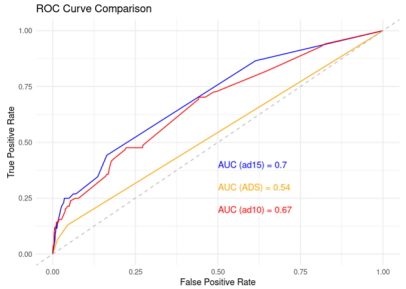
3.1 Assembly of FPS Feature Call Variables

	raid	c11_k	c11_o	c11_1	c11_2	c11_3
1	rs2409718	1	0	1	0	0
2	rs4642600	1	0	1	0	0
3	rs10813910	1	0	1	0	0
4	rs67418928	1	0	1	0	0
5	rs74832541	1	0	1	0	0
6	rs75754122	1	0	1	0	0
7	rs76826248	1	0	1	0	0
8	rs11306	0	1	0	0	1
9	rs281510	0	1	0	0	1
10	rs716427	0	1	0	0	1
11	rs716741	1	1	0	1	0
12	rs716742	1	1	0	1	0

3.3 Model building using LASSO



3.4 Assessment of model performance



4. Conclusions

- Identifying functional/causal variants and risk genes at GWAS loci remains a critical challenge for downstream clinical translation
- Integration of tissue-specific annotations into a "functional potential score" (FPS) can improve variant prioritization for experimental studies
- LASSO model indicates that certain annotation data inputs were most predictive of experimentally-determined SNP regulatory function: **DHS** (c5_1, c5n_K), **TF-ChIP-seq** (c6n_K, c6_3), **Seq conservation** (c2n).

References

- Chen, J., ..., Buas, M. F. (2022). Prioritization and functional analysis of GWAS risk loci for Barrett's esophagus and esophageal adenocarcinoma. *Hum Mol Genet.* 31: 410–422.
- Ali MW, ..., Buas M.F. (2022) A risk variant for Barrett's esophagus and esophageal adenocarcinoma at chr8p23.1 affects enhancer activity and implicates multiple gene targets. *Hum Mol Genet.* 31:3975-3986.
- Long, E., et al. (2022) Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity. *Am J Hum Genet.* 109:2210–2229.
- Cano-Gamez, E., & Trynka, G. (2020) From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet.* 11:424.

Acknowledgements

I gratefully acknowledge support from the National Cancer Institute (R25CA272282), the National Institute of Diabetes and Digestive and Kidney Diseases (R01DK128615), the MSK Society, and my dedicated mentor, Dr. Buas.