Analyzing Modifier Gene Variants That Influence Sickle Cell Severity Across Populations

DATS 4001, Aba Pobee

Motivation & context

- Sickle Cell Disease (SCD) is caused by a *single* genetic mutation
- But symptoms range from mild to severe
- Why? Because **modifier genes** change how the disease behaves
- Studying these modifiers can explain:
  - Population differences
  - Treatment responses
  - Biological resilience

Research Question

How are important sickle-cell modifier variants spread across populations, and can simple genomic features like allele frequency help tell apart intergenic variants from gene variants?

Data Collection & Cleaning

Collection

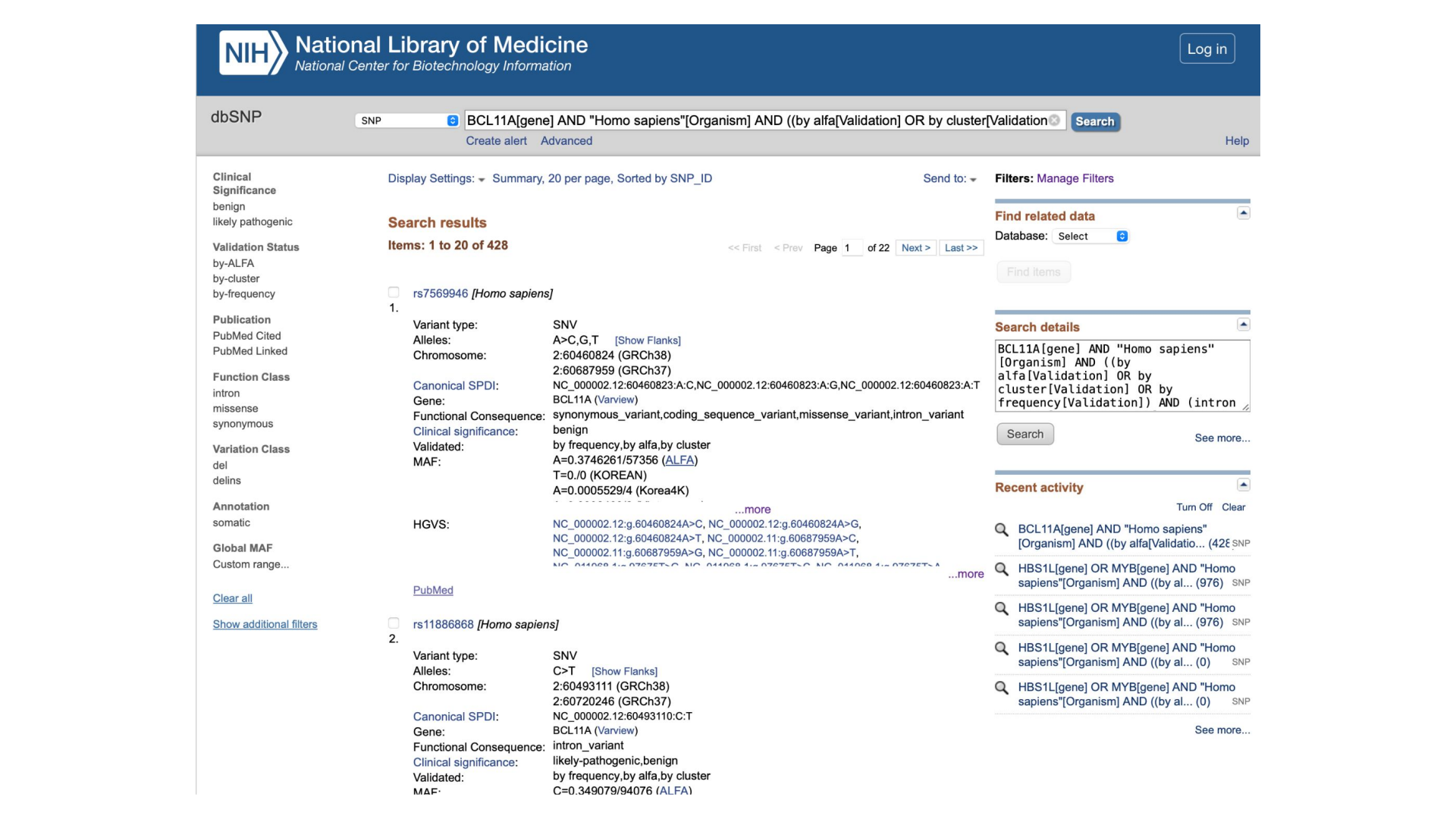
Extracted genetic variants from public databases:

- **dbSNP** (variant IDs, locations, annotations)
- **NIH's Variant Viewer** (functional consequences and visual understanding)

- Focused on four sickle-cell modifier regions:
  BCL11A, HBS1L–MYB, KLF1, HMOX1
- Combined raw data into variant-level tables
  (SNP ID, chromosome, position, allele frequency, annotation)

Cleaning

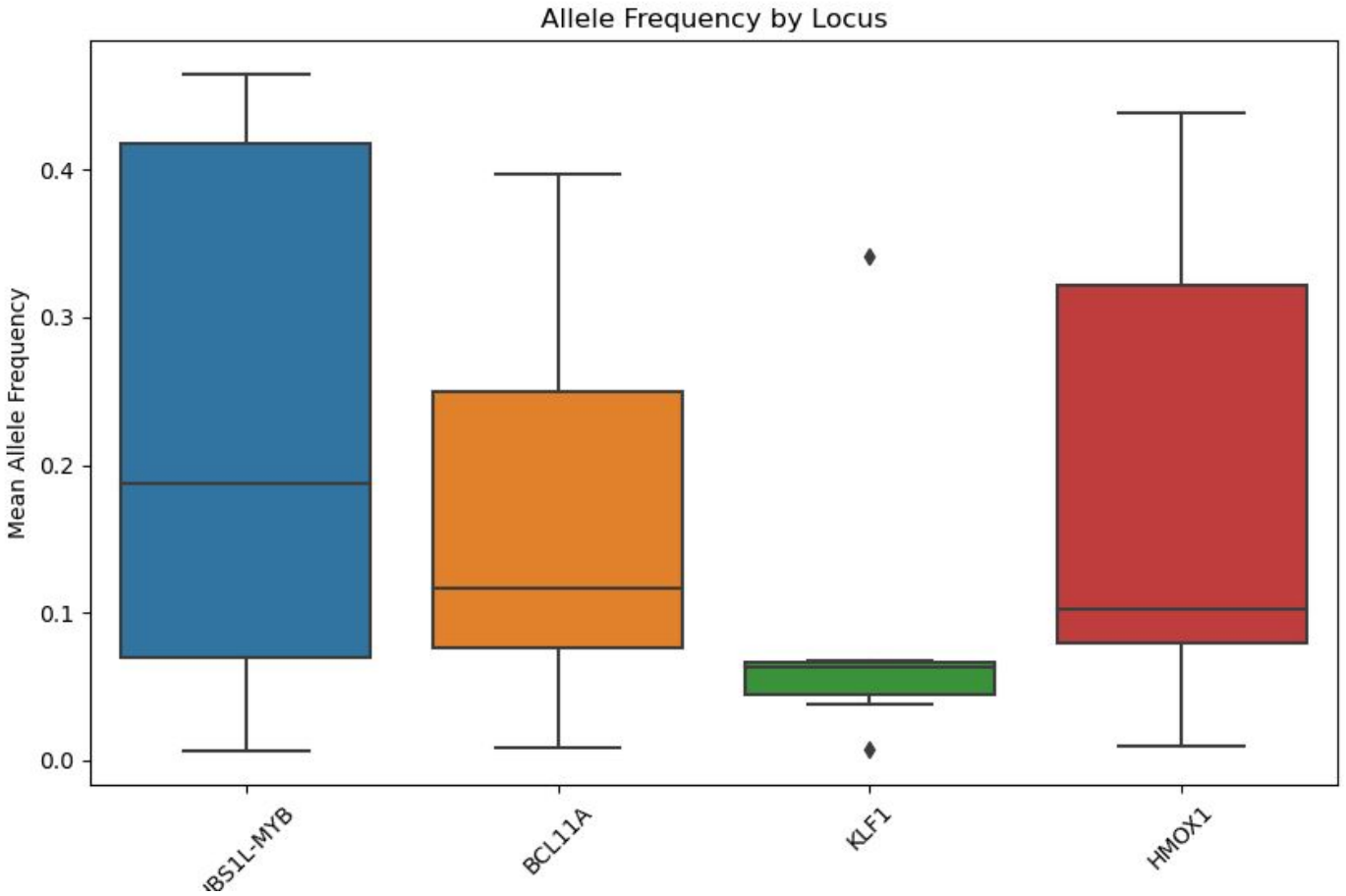
- Removed duplicates and variants lacking essential fields
- Converted functional consequences into a **numerical count**
  (e.g., 2 overlapping regulatory roles → count = 2)
- Standardized allele frequencies across sources
- Labeled variants as **genic (1)** vs **intergenic (0)**
- Merged all loci into one unified modeling dataset (≈787 variants)

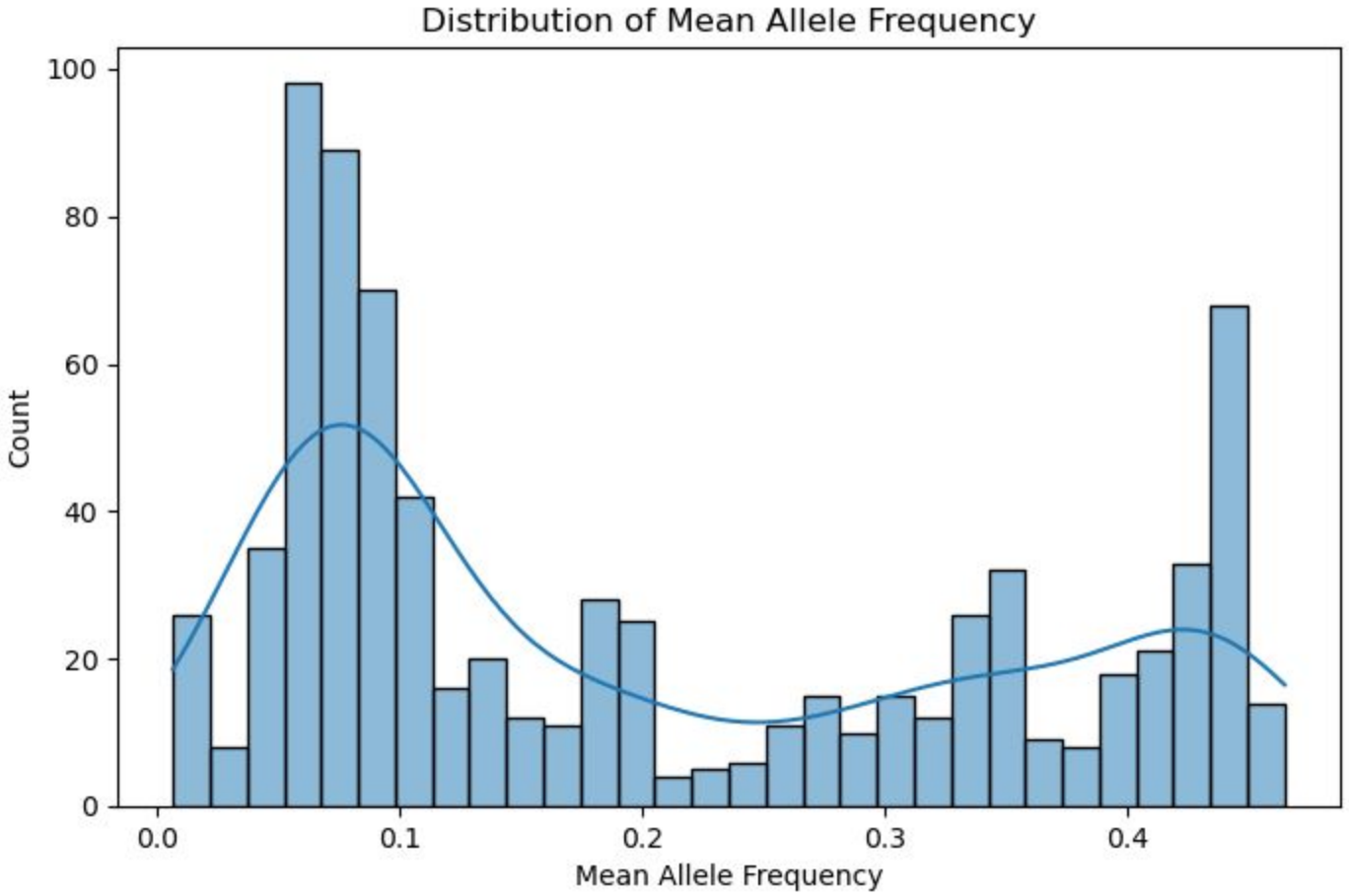

National Library of Medicine
National Center for Biotechnology Information
Log in
dbSNP
SNP
BCL11A[gene] AND "Homo sapiens"[Organism] AND ((by alfa[Validation] OR by cluster[Validation
Search
Create alert
Advanced
Help
Clinical Significance
benign
likely pathogenic
Validation Status
by-ALFA
by-cluster
by-frequency
Publication
PubMed Cited
PubMed Linked
Function Class
intron
missense
synonymous
Variation Class
del
delins
Annotation
somatic
Global MAF
Custom range...
Clear all
Show additional filters
Display Settings: Summary, 20 per page, Sorted by SNP_ID
Send to:
Search results
Items: 1 to 20 of 428
<< First
< Prev
Page 1 of 22
Next >
Last >>
rs7569946 [Homo sapiens]
1.
Variant type: SNV
Alleles: A>C,G,T [Show Flanks]
Chromosome: 2:60460824 (GRCh38)
2:60687959 (GRCh37)
Canonical SPDI: NC_000002.12:60460823:A:C,NC_000002.12:60460823:A:G,NC_000002.12:60460823:A:T
Gene: BCL11A (Varview)
Functional Consequence: synonymous_variant,coding_sequence_variant,missense_variant,intron_variant
Clinical significance: benign
Validated: by frequency,by alfa,by cluster
MAF: A=0.3746261/57356 (ALFA)
T=0./0 (KOREAN)
A=0.0005529/4 (Korea4K)
...more
HGVS: NC_000002.12:g.60460824A>C, NC_000002.12:g.60460824A>G,
NC_000002.12:g.60460824A>T, NC_000002.11:g.60687959A>C,
NC_000002.11:g.60687959A>G, NC_000002.11:g.60687959A>T,
...more
PubMed
rs11886868 [Homo sapiens]
2.
Variant type: SNV
Alleles: C>T [Show Flanks]
Chromosome: 2:60493111 (GRCh38)
2:60720246 (GRCh37)
Canonical SPDI: NC_000002.12:60493110:C:T
Gene: BCL11A (Varview)
Functional Consequence: intron_variant
Clinical significance: likely-pathogenic,benign
Validated: by frequency,by alfa,by cluster
MAF: C=0.349079/94076 (ALFA)
Filters: Manage Filters
Find related data
Database: Select
Find items
Search details
BCL11A[gene] AND "Homo sapiens"
[Organism] AND ((by
alfa[Validation] OR by
cluster[Validation] OR by
frequency[Validation]) AND (intron
Search
See more...
Recent activity
Turn Off Clear
BCL11A[gene] AND "Homo sapiens" [Organism] AND ((by alfa[Validatio... (428 SNP
HBS1L[gene] OR MYB[gene] AND "Homo sapiens"[Organism] AND ((by al... (976) SNP
HBS1L[gene] OR MYB[gene] AND "Homo sapiens"[Organism] AND ((by al... (976) SNP
HBS1L[gene] OR MYB[gene] AND "Homo sapiens"[Organism] AND ((by al... (0) SNP
HBS1L[gene] OR MYB[gene] AND "Homo sapiens"[Organism] AND ((by al... (0) SNP
See more...

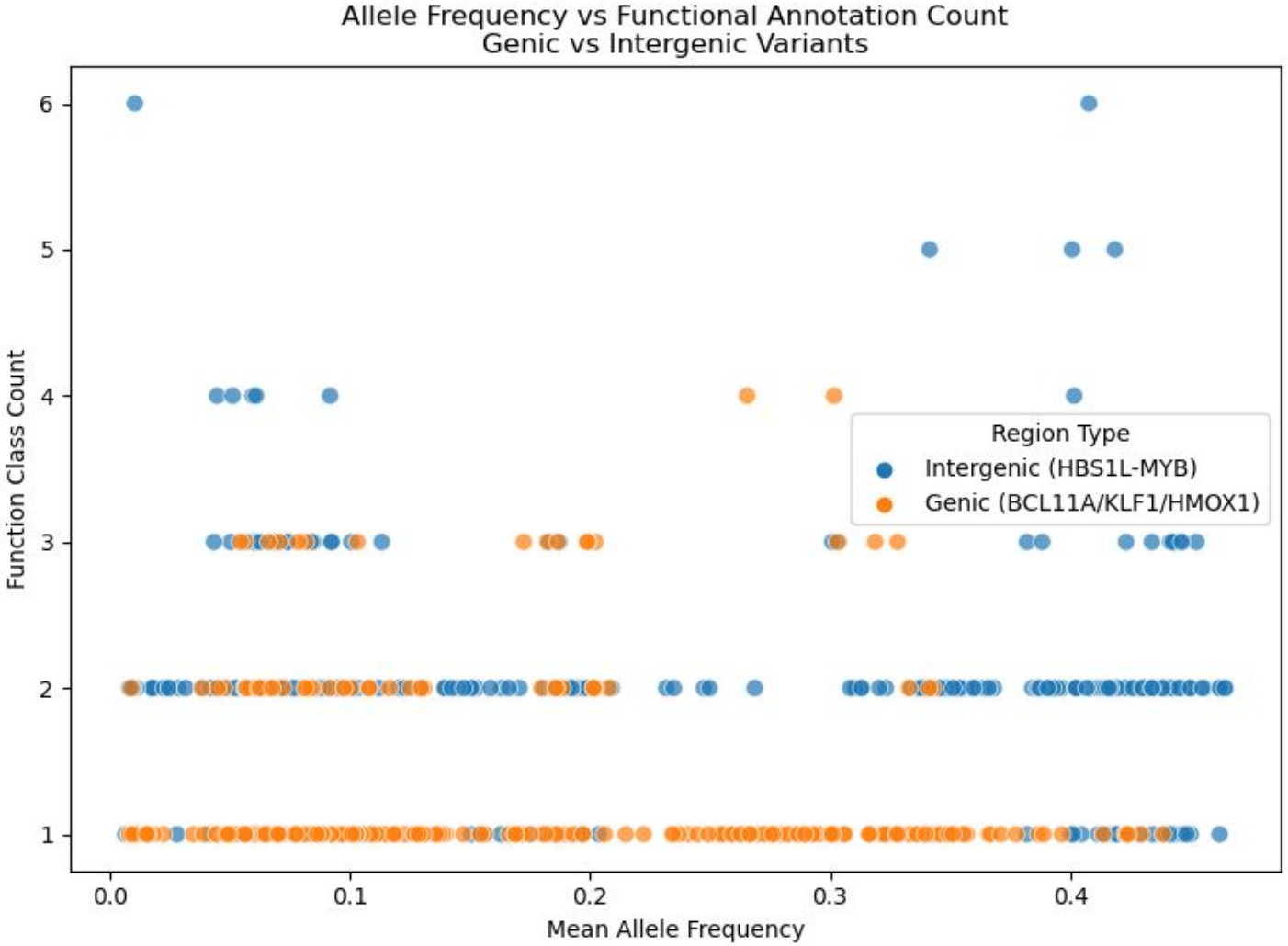
Exploratory Data Analysis

Main Findings:

- Most variants are **rare**, but some are surprisingly common
- Intergenic region (**HBS1L–MYB**) shows the **widest variation**
- Gene regions (**BCL11A, KLF1, HMOX1**) contain mostly **rare variants**
- Functional annotations differ:
  - Intergenic variants often have **multiple regulatory roles**
  - Gene variants have **fewer, simpler annotations**


Allele Frequency by Locus
Mean Allele Frequency
0.0
0.1
0.2
0.3
0.4
BCL11A
KLF1
HMOX1


Distribution of Mean Allele Frequency
Count
Mean Allele Frequency
0
20
40
60
80
100
0.0
0.1
0.2
0.3
0.4


Allele Frequency vs Functional Annotation Count
Genic vs Intergenic Variants
Function Class Count
Mean Allele Frequency
Region Type
Intergenic (HBS1L-MYB)
Genic (BCL11A/KLF1/HMOX1)
1
2
3
4
5
6
0.0
0.1
0.2
0.3
0.4

Modeling and Predictive Findings

Two Models Tested:

- Logistic Regression
- Decision Tree Classifier

Performance:

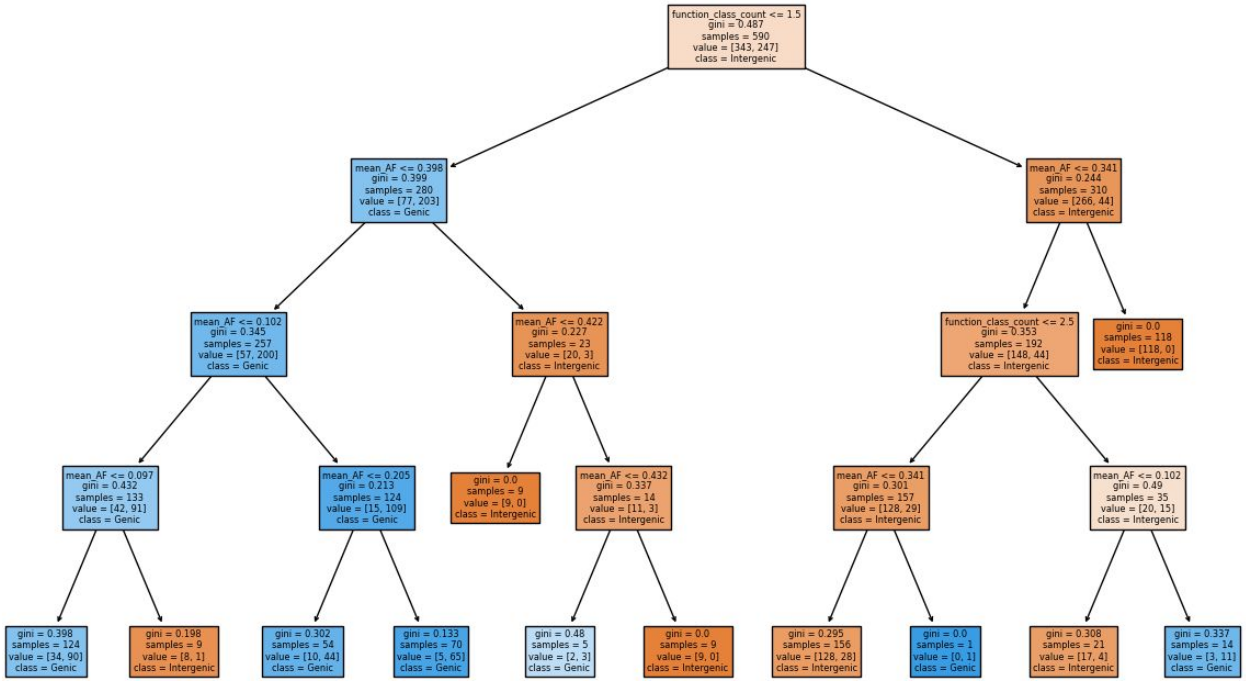
- **~80% accuracy** for both models
- Functional annotation count = **most important feature**
- High allele frequency = **more likely intergenic**
- Rare variants = **more likely gene-based**

```
Logistic Regression Accuracy: 0.7969543147208121
[[91 23]
 [17 66]]
              precision    recall  f1-score   support

           0       0.84      0.80      0.82       114
           1       0.74      0.80      0.77        83

    accuracy                           0.80       197
   macro avg       0.79      0.80      0.79       197
weighted avg       0.80      0.80      0.80       197

                Feature  Coefficient
0               mean_AF    -0.479466
1  function_class_count    -1.286006
```


function_class_count <= 1.5
gini = 0.487
samples = 590
value = [343, 247]
class = Intergenic
mean_AF <= 0.398
gini = 0.399
samples = 280
value = [77, 203]
class = Genic
mean_AF <= 0.341
gini = 0.244
samples = 310
value = [266, 44]
class = Intergenic
mean_AF <= 0.102
gini = 0.345
samples = 257
value = [57, 200]
class = Genic
mean_AF <= 0.422
gini = 0.227
samples = 23
value = [20, 3]
class = Intergenic
function_class_count <= 2.5
gini = 0.353
samples = 192
value = [148, 44]
class = Intergenic
gini = 0.0
samples = 118
value = [118, 0]
class = Intergenic
mean_AF <= 0.097
gini = 0.432
samples = 133
value = [42, 91]
class = Genic
mean_AF <= 0.205
gini = 0.213
samples = 124
value = [15, 109]
class = Genic
gini = 0.0
samples = 9
value = [9, 0]
class = Intergenic
mean_AF <= 0.432
gini = 0.337
samples = 14
value = [11, 3]
class = Intergenic
mean_AF <= 0.341
gini = 0.301
samples = 157
value = [128, 29]
class = Intergenic
mean_AF <= 0.102
gini = 0.49
samples = 35
value = [20, 15]
class = Intergenic
gini = 0.398
samples = 124
value = [34, 90]
class = Genic
gini = 0.198
samples = 9
value = [8, 1]
class = Intergenic
gini = 0.302
samples = 54
value = [10, 44]
class = Genic
gini = 0.133
samples = 70
value = [5, 65]
class = Genic
gini = 0.48
samples = 5
value = [2, 3]
class = Genic
gini = 0.0
samples = 9
value = [9, 0]
class = Intergenic
gini = 0.295
samples = 156
value = [128, 28]
class = Intergenic
gini = 0.0
samples = 1
value = [0, 1]
class = Genic
gini = 0.308
samples = 21
value = [17, 4]
class = Intergenic
gini = 0.337
samples = 14
value = [3, 11]
class = Genic

Conclusions and Impact

Conclusions:

- Modifier regions show clear, distinct genetic patterns
- Simple features can reliably classify variant context
- Enhancer region HBS1L–MYB is highly variable and functionally rich

Why It Matters:

- Helps explain population differences in SCD severity
- Supports future genomic research and clinical prediction models

Future Directions:

- Add population-specific analyses
- Incorporate more functional scores
- Explore links to clinical severity

Thank you!