Genomic Patterns in Sickle Cell Modifier Variants: Mapping and Predicting Variation Across Key Modifier Loci

Aba Pobee

Department of Data Science, The George Washington University

DATS 4001: Data Science Capstone

Dr. Sushovan Majhi

December 12, 2025

Sickle cell disease (SCD) is a hereditary blood disorder caused by a single point mutation in the beta-globin gene, which leads to the production of misshapen red blood cells and chronic complications such as pain crises, anemia, and organ damage. Although all individuals with SCD carry the same primary mutation, the severity of symptoms varies widely. Some patients experience relatively mild symptoms, while others face life-threatening complications. This variability can't be explained by the HbS mutation alone; severity of sickle cell is influenced by additional genetic factors known as modifier genes.

Modifier genes and regulatory regions affect key biological processes relevant to SCD, including fetal hemoglobin (HbF) production, erythropoiesis (red blood cell production), oxidative stress responses, and inflammation. Four of the most well-studied modifier loci are BCL11A, HBS1L–MYB, KLF1, and HMOX1, each of which has been shown to influence clinical outcomes by altering hematologic or regulatory pathways. For example, variants in the HBS1L–MYB intergenic enhancer region are associated with increased HbF levels, while variants in the BCL11A gene are strong repressors of HbF and contribute to baseline differences in disease severity across individuals.

This project focuses on analyzing known single nucleotide polymorphisms (SNPs) within the HBS1L–MYB intergenic regulatory region and the BCL11A, KLF1, and HMOX1 genes. Using variant annotations and allele frequency data from large public genomic resources (NIH's dbSNP and Ensembl), this project explores how these modifier variants are distributed across population-coded cohorts. Additionally, a simple predictive model is developed to determine whether basic genomic features, such as allele frequency, variant type, and predicted functional annotations, can distinguish intergenic regulatory variants from gene-body variants across the

four loci. This combined descriptive and predictive approach provides a deeper look into the genome-level circumstances that shape clinical variability within sickle cell disease.

Methods

Data Sources and Loci Selection

Four established sickle-cell disease (SCD) modifier loci were included in this analysis: the HBS1L–MYB intergenic region, and the BCL11A, KLF1, and HMOX1 genes. These loci were selected based on prior scientific literature showing their influence on fetal hemoglobin (HbF) regulation, erythropoiesis, and oxidative stress responses. For each locus, variant-level information was extracted from dbSNP and the Ensembl Variant Browser, including genomic coordinates, rsID identifiers, gene annotations, variant type, clinical significance, and predicted functional consequences.

Population-coded allele frequencies were also extracted from dbSNP/Ensembl. These frequency fields are recorded across multiple genomic cohorts and biobanks, including 1000 Genomes, ALFA, ALSPAC, Estonian Biobank, TWINSUK, Vietnamese samples, and others. These sources represented diverse population backgrounds, but they don't map cleanly onto classic global ancestry categories. Therefore, my analysis focuses on differences across the population-coded data sources as provided by the databases.

Frequency Table Processing

Allele frequency information appears as multiple rows per variant: one row per population-coded cohort. In order to generate a single allele frequency value per SNP in a way that would be suitable for modeling, allele frequencies were collapsed by averaging across all

available sources. For each rsID, the mean alternate-allele frequency was computed. This approach provided a stable, comparable estimate of allele frequency while avoiding duplication of SNP rows in the modeling dataset.

Merging Variant Annotations and Frequency Data

For each modifier gene, the collapsed allele-frequency summary was left-joined onto the annotation table using the rsID as the primary key, with chromosome and position used as secondary checks for alignment. This generated a unified SNP-level dataset containing both functional descriptors and population-coded allele frequency information. A feature representing the number of predicted variant consequences was created by counting the number of functional annotations listed for each SNP. Additional categorical predictors (variant type, clinical significance, gene label) were saved.

Each locus was assigned a categorical label ("HBS1L-MYB," "BCL11A," "KLF1," or "HMOX1"). Based on genomic context, variants in HBS1L–MYB were classified as intergenic, while variants in the three remaining loci were classified as genic. A binary outcome variable (genic) was created, where 0 = intergenic variant and 1 = gene-body variant. The four locus-specific datasets were then concatenated to form a final dataset of 787 SNPs.

Dataset Preparation for Exploratory Analysis and Modeling

Variables used for exploratory data analysis included the locus label, allele frequency, variant type, clinical significance, and functional annotation count. Prior to modeling, rows missing allele frequency were removed (none in this dataset), and all categorical variables were encoded as needed. The final modeling dataset consisted of one row per SNP, containing:

- Predictors: allele frequency, variant type, clinical significance, function_class_count, locus
- Outcome: genic vs intergenic classification

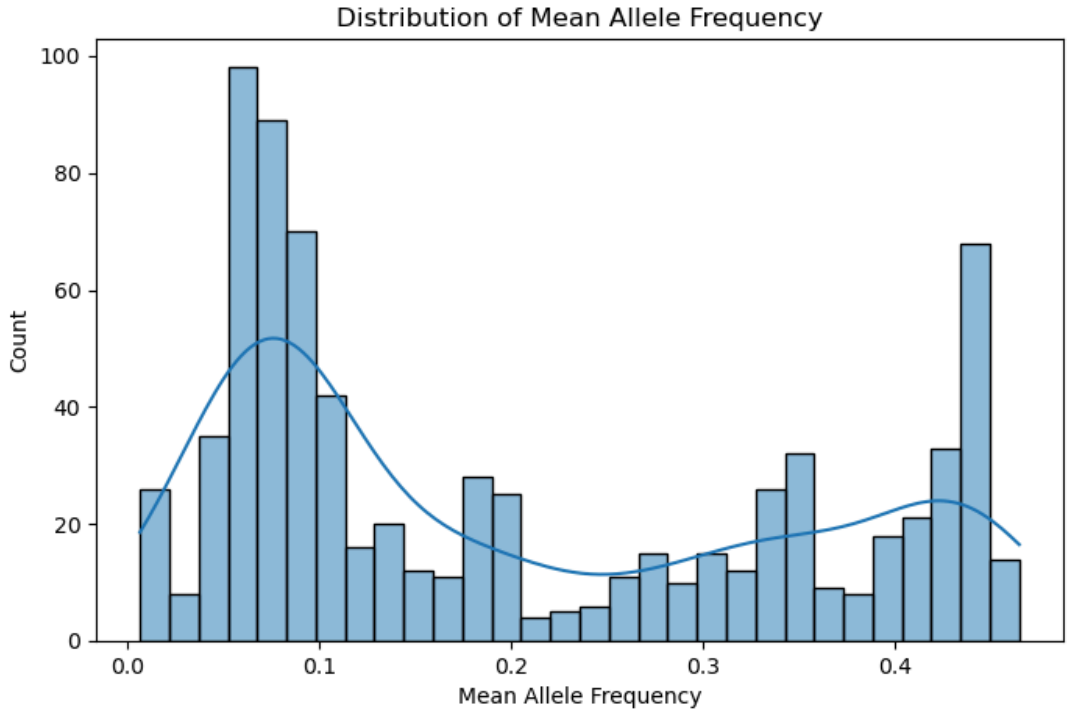
This dataset was used both for exploratory visualizations and for building predictive models in my further analysis.

Exploratory Data Analysis

Exploratory data analysis was conducted to characterize allele frequency patterns, functional annotation structure, and locus-specific variation across the four modifier regions. These analyses provide the foundation for evaluating existing differences and assessing the feasibility of predicting variant context based on genomic features.

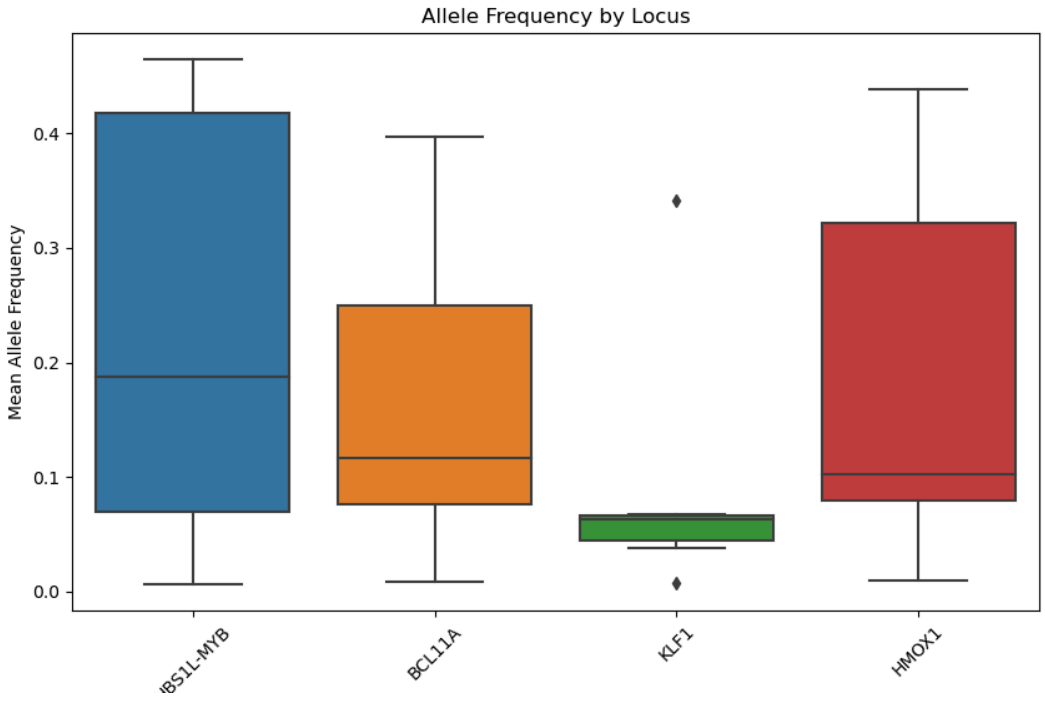
Distribution of Allele Frequencies

Across all 787 variants, allele frequency displayed a multimodal distribution with a strong peak at low frequencies (0.02–0.10) and secondary peaks near 0.30 and 0.45. With that being said, the majority of variants are therefore rare or uncommon, while a smaller subset occurs at moderate or high frequencies across cohorts. This pattern is consistent with human genomic variation in a broader sense, where most variants arise from recent mutations and fewer alleles become common through drift or selection.



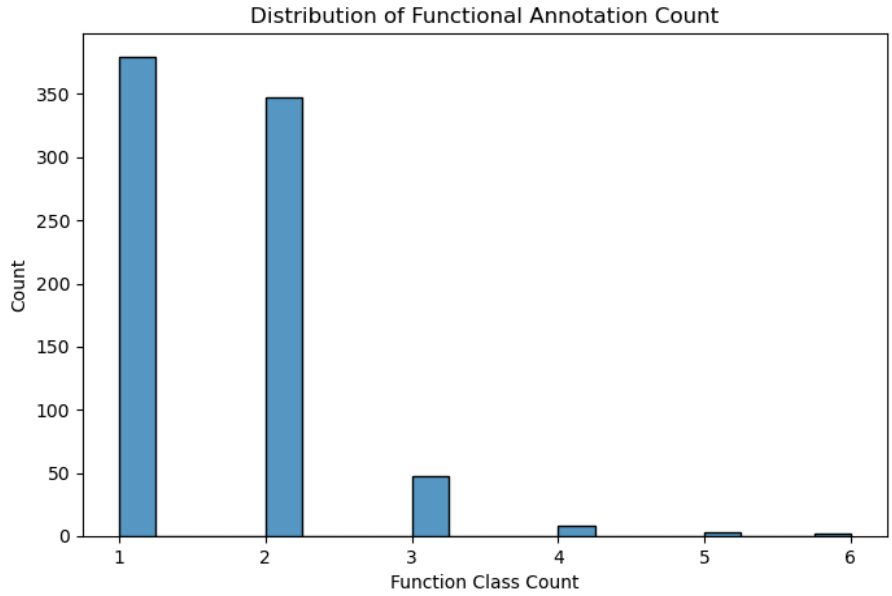
Allele Frequency Variation Across Modifier Loci

Boxplots showed that allele frequency distributions differed substantially by locus. The HBS1L–MYB intergenic region displayed the widest range and highest median allele frequency, reflecting the greater mutational flexibility often seen in regulatory regions. In contrast, KLF1 variants were universally rare, remaining consistent with its known role as a conserved transcription factor where disruptive variants are strongly selected against. BCL11A and HMOX1 displayed intermediate ranges, aligning with their mixed regulatory and coding functions. These differences shine light on locus-specific evolutionary constraints that shape population-level variation.



Functional Annotation Patterns

Most variants had one or two predicted functional consequences, with smaller subsets annotated with three or more. When compared across loci, gene-body regions (BCL11A, KLF1, HMOX1) tended to show modestly higher functional annotation counts, whereas HBS1L–MYB variants displayed broader variability. This indicates that regulatory-region SNPs span a diverse range of predicted impacts, while coding regions exhibit more constrained functional categories.



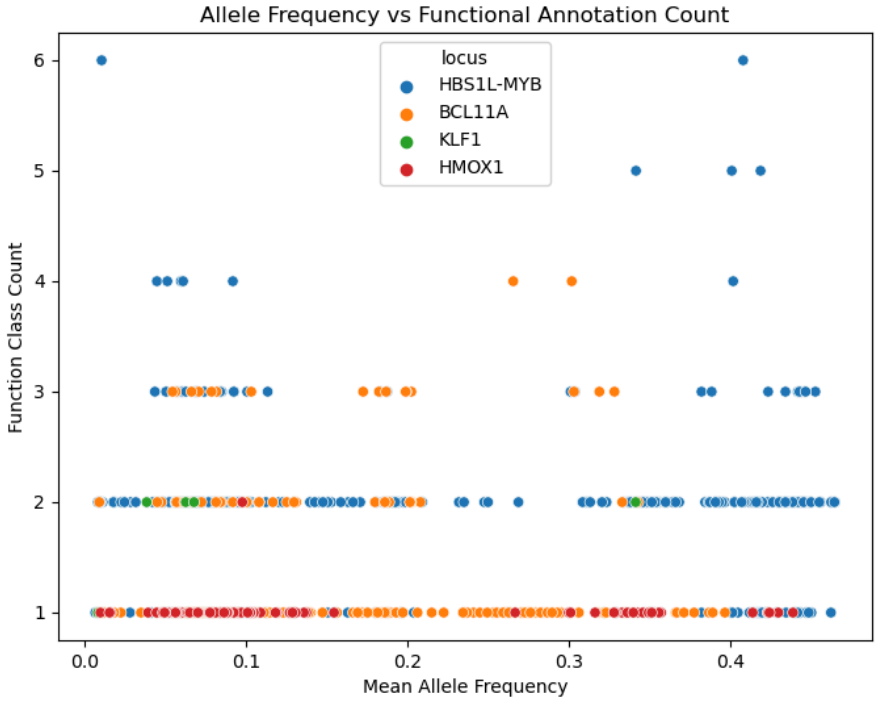
Variant Type Distributions

Single nucleotide variants (SNVs) dominated the dataset, making up the majority of variants. Small insertions, deletions, and delins (deletion-insertion) variants were present but rare. This distribution is typical of human genetic variation and reflects known biologic principles.

Relationship Between Allele Frequency and Functional Complexity

Scatterplots comparing allele frequency to functional annotation count revealed distinct clustering patterns across loci. KLF1 variants were rare and functionally simple, while HBS1L–MYB variants spanned the full range of allele frequencies and functional counts. BCL11A and HMOX1 showed intermediate profiles. These patterns suggest that both allele frequency and functional annotation show valuable information about the genomic context of variants. Notably, intergenic variants had greater spread and heterogeneity, whereas gene-body

variants were more tightly clustered, supporting the expectation that these features may aid my classification in predictive modeling.



Modeling

To assess whether simple genomic features could distinguish intergenic regulatory variants from gene-body variants, I constructed predictive models. The primary outcome variable was a binary indicator of genomic context (genic vs intergenic), where variants from the HBS1L–MYB region were labeled as intergenic and variants from BCL11A, KLF1, and HMOX1 were labeled as genic. Predictor variables included mean allele frequency and functional annotation count, selected based on their biological relevance and observed separation during the exploratory data analysis I conducted.

The dataset was split into training and test sets using a 75/25 split. Two interpretable models were evaluated: logistic regression and a decision tree classifier. Logistic regression was chosen for its ability to estimate directional relationships between predictors and the outcome, while the decision tree was selected to capture potential nonlinear interactions and threshold effects without requiring feature scaling. Both models were also chosen for their interpretability.

Logistic Regression

Logistic regression was trained using standardized predictor variables. Model performance was evaluated on the held-out test set using accuracy, precision, recall, and a confusion matrix. The model achieved an accuracy of approximately 0.80, with balanced recall across both classes.

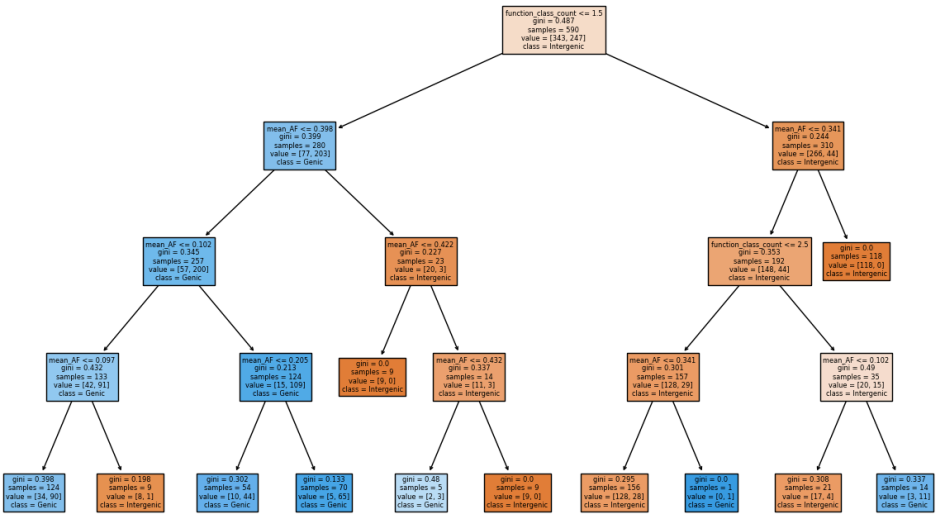
```
Logistic Regression Accuracy: 0.7969543147208121
[[91 23]
 [17 66]]
              precision    recall  f1-score   support

           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
```

Coefficient estimates indicated that mean allele frequency had a negative association with being genic, meaning that higher-frequency variants were more likely to be intergenic. Functional annotation count also showed a negative association, suggesting that variants with multiple predicted functional consequences were enriched in the intergenic HBS1L–MYB

region. These relationships align with known biological patterns, as regulatory enhancer regions often accumulate common variants with overlapping functional roles, whereas gene-body variants are more constrained and tend to remain rare.

Decision Tree

A decision tree classifier with limited depth was trained to explore nonlinear relationships between predictors. The decision tree achieved a slightly higher accuracy of approximately 0.81 on the test set. Feature importance analysis revealed that functional annotation count was the strongest predictor, followed by allele frequency.



The structure of the tree was biologically interpretable. The first split was based on functional annotation count, separating variants with a single predicted consequence from those with multiple annotations. Subsequent splits relied on allele frequency thresholds, with

high-frequency variants overwhelmingly classified as intergenic. These decision boundaries closely mirrored the patterns observed during EDA and supported the plausibility of the model.

Discussion

This project demonstrates that sickle cell disease modifier loci contain distinct population-genomic and functional significance that reflect their underlying biological roles. Variants in the HBS1L–MYB intergenic enhancer region showed broad allele-frequency distributions and high functional annotation complexity, consistent with the flexibility of enhancer elements involved in fetal hemoglobin regulation. Conversely, variants within BCL11A, KLF1, and HMOX1 were predominantly rare and functionally constrained, reflecting selective pressures acting on coding regions critical to erythropoiesis and oxidative stress response.

The predictive modeling results show that even minimal genomic features can capture meaningful biological differences. Both logistic regression and decision tree models achieved strong performance using only allele frequency and functional annotation count, suggesting that these features encode substantial information about genomic context. Importantly, the interpretability of both models allows direct biological insight, rather than relying on black-box prediction alone.

From a population-genomic perspective, these findings help explain why modifier effects vary across cohorts and populations. Regions with higher-frequency regulatory variants may contribute to population-level differences in baseline fetal hemoglobin levels and disease severity. These insights are particularly relevant for precision medicine efforts aimed at

improving sickle cell outcomes in diverse populations historically underrepresented in genetic research.

Limitations

Several limitations should be noted. First, allele frequencies were averaged across population-coded cohorts rather than analyzed separately by ancestry, limiting the ability to draw ancestry-specific conclusions. Second, functional annotation counts rely on database-predicted consequences, which may not fully capture biological impact. Third, the binary classification of genic versus intergenic variants simplifies a more complex landscape, where gene-proximal regulatory elements might have blurred these categories. Finally, clinical severity data were not incorporated, preventing direct genotype–phenotype association analysis.

When pitching my results of this project, it was suggested to me to include a random forest model– I chose not to because the dataset contained only a small number of interpretable features, and my primary goal was biological interpretability rather than marginal gains in predictive accuracy.

Conclusion

This project provides a data-driven analysis of key sickle cell disease modifier loci using publicly available genomic resources. The results demonstrate that allele frequency and functional annotation complexity vary systematically across modifier regions and can be leveraged to distinguish intergenic regulatory variants from gene-body variants with high accuracy. The HBS1L–MYB enhancer region emerged as particularly diverse and functionally rich, underscoring its central role in modulating fetal hemoglobin and disease severity.

By integrating population genomics with interpretable modeling, this work contributes to a deeper understanding of how genetic modifiers shape the clinical heterogeneity of sickle cell disease. These findings lay the groundwork for future studies that incorporate ancestry-specific analyses, functional validation, and clinical outcomes to support equitable precision medicine approaches for individuals living with sickle cell disease.

References

Bauer, D. E., Kamran, S. C., & Orkin, S. H. (2012). Reawakening fetal hemoglobin: Prospects for new therapies for the β-globin disorders. *Blood*, 120(15), 2945–2953. https://doi.org/10.1182/blood-2012-06-292078

Bean, C. J., Boulet, S. L., Yang, G., & Goodman, D. A. (2012). Heme oxygenase-1 gene polymorphisms and sickle cell disease complications. American Journal of Hematology, 87(9), 879–881. https://doi.org/10.1002/ajh.23255

Borg, J., Papadopoulos, P., Georgitsi, M., Gutierrez, L., Grech, G., Fanis, P., … Grosveld, F. (2010). Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. Nature Genetics, 42(9), 801–805. https://doi.org/10.1038/ng.630

Cunningham, F., Allen, J. E., Allen, J., et al. (2022). Ensembl 2022. Nucleic Acids Research, 50(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Farrell, J. J., Sherva, R. M., Chen, Z. Y., Luo, H. Y., Chu, B. F., Ha, S. Y., … Steinberg, M. H. (2011). A deletion in the HBS1L–MYB intergenic region is associated with fetal hemoglobin expression. Blood, 117(18), 4935–4945. https://doi.org/10.1182/blood-2010-09-307843

Liu, N., & Orkin, S. H. (2010). Transcriptional repression of the fetal globin gene by BCL11A. Blood, 116(4), 435–444. https://doi.org/10.1182/blood-2009-12-256479

National Center for Biotechnology Information. (2024). dbSNP. https://www.ncbi.nlm.nih.gov/snp/

OpenAI. (2024). ChatGPT (GPT-4) [Large language model]. OpenAI. https://chat.openai.com/

Sankaran, V. G., Xu, J., Ragoczy, T., et al. (2009). Developmental and species-divergent globin switching are driven by BCL11A. Nature, 460(7259), 1093–1097. https://doi.org/10.1038/nature08243

Tallack, M. R., & Perkins, A. C. (2010). KLF1 coordinates erythroid gene expression. Blood, 115(5), 1154–1163. https://doi.org/10.1182/blood-2009-08-239277

Thein, S. L., Menzel, S., Peng, X., et al. (2007). A quantitative trait locus between HBS1L and MYB controls fetal hemoglobin. PNAS, 104(27), 11346–11351. https://doi.org/10.1073/pnas.0703407104