DOI: https://doi.org/10.4314/njbmb.v40i3.14 NJBMB (2025) 40(3) 363-368



Nigerian Journal of Biochemistry and Molecular Biology

The Official Publication of the Nigerian Society of Biochemistry & Molecular Biology (NSBMB). Journal homepage: https://www.nsbmb.org.ng/journals/

e-ISSN: 2659-0042

Research Article

AI-Guided Identification of a Critical Enhancer Element Controlling Fetal Haemoglobin Genes HBG1 and HBG2: Implications for Sickle Cell Disease Therapy in Nigeria

Aimola Idowu[1,*], Firdausi S. Muhammad[1], Aliyu Salihu[1], Aliyu Muhammad[1], Idris Z. Sadiq[1], Samuel A. Obo[1], Precious N. Onuoha[1], Usman U. Umar[1], Favour O. Jaja[1], Olamide E. Abolarin[1], Ismail Adamu[1]

[1]*Department of Biochemistry, Faculty of Life Sciences, Ahmadu Bello University, Zaria, Nigeria*

OPEN ACCESS

CORRESPONDENCE
Idowu, A.
aimolaidowu@gmail.com

ARTICLE HISTORY

CITATION
Idowu A., Muhammad F.S., Salihu M., Muhammad A., Sadiq I.Z., Obo S.A., Onuoha P.N., Umar U.U., Jaja F.O., Abolarin O.E. & Adamu I. (2025). AI-guided identification of a critical enhancer element controlling fetal haemoglobin genes HBG1 and HBG2: implications for sickle cell disease therapy in Nigeria. *Nigerian Journal of Biochemistry and Molecular Biology*. 40(3), 363-368
https://doi.org/10.4314/njbmb.v40i3.14

ABSTRACT

Sickle cell disease (SCD) affects approximately 150,000 Nigerian children annually, with Nigeria bearing 75% of Africa's SCD burden. Reactivating fetal haemoglobin (HbF) genes HBG1 and HBG2 represents a promising therapeutic strategy. This study applied Enformer, a deep learning transformer model, to predict regulatory activity across a 393 kb region of chromosome 11 containing HBG1 and HBG2. The model predicted 5,313 genomic tracks including H3K27ac (active enhancer mark), DNase hypersensitivity, and CAGE (transcription) signals across 896 genomic bins. We identified a strong enhancer element at chr11:5,308,936-5,309,374 (438 bp) located 40 kb downstream of HBG1, with an H3K27ac signal of 0.661, the highest in the analysed region. This enhancer showed multiple characteristics of active regulatory DNA: high DNase accessibility (signal=15.0), CAGE peaks indicating nearby transcription, and JASPAR motif scanning (threshold ⩾4.0) confirmed binding sites for GATA1 (6 matches, top score 7.59) and TAL1::TCF3 (2 matches, top score 9.17); no KLF1 motif was detected at this locus. Additionally, we confirmed the β-globin Locus Control Region (LCR) approximately 118 kb upstream (H3K27ac=0.201). Both HBG1 and HBG2 share these regulatory elements, enabling coordinated expression. This computational analysis provides precise molecular coordinates for a candidate HbF regulatory enhancer that may contribute to BCL11A-mediated repression, pending experimental validation, establishing a foundation for developing targeted gene therapies for SCD treatment in resource-limited settings.

Keywords: *Fetal haemoglobin; HBG1; HBG2; Enformer; Sickle cell disease; Gene regulation*

INTRODUCTION

Sickle cell disease (SCD) is a hereditary blood disorder caused by a single nucleotide mutation (Glu6Val) in the adult beta-globin gene (HBB), resulting in haemoglobin S (HbS) polymerisation and red blood cell sickling (Ware et al., 2017). Nigeria has the world's largest population of SCD patient, with approximately 150,000 affected births annually and a prevalence of 2–3% (Piel et al., 2013). Despite this enormous disease burden, access to advanced therapeutics remains severely limited. Fetal haemoglobin (HbF), composed of two alpha-globin and two gamma-globin chains ($\alpha_2\gamma_2$), does not undergo sickling and remains functional throughout life if maintained (Wilber et al., 2011). The gamma-globin genes HBG1 (Gγ) and HBG2 (Aγ) are naturally expressed at high levels before birth but are silenced shortly after, replaced by adult haemoglobin (HbA: $\alpha_2\beta_2$). Individuals with hereditary persistence of fetal haemoglobin (HPFH) or elevated HbF

Copyright © 2025 *Idowu et al. This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

levels experience milder SCD symptoms, establishing HbF reactivation as a validated therapeutic strategy (Akinsheye et al., 2011).

Gene expression is controlled by enhancer elements, DNA sequences that can be located thousands of base pairs from their target genes yet regulate transcription through three-dimensional chromatin looping (Schoenfelder and Fraser, 2019). The BCL11A transcription factor has been identified as the primary developmental repressor of HBG1 and HBG2 in adults (Sankaran et al., 2008), but the precise genomic location where BCL11A acts to silence these genes remains incompletely characteried.

Enformer, developed by DeepMind and published in Nature Methods (2021), represents a breakthrough in computational prediction of gene regulatory elements (Avsec et al., 2021). This transformer-based deep learning model was trained on massive datasets from the ENCODE and Roadmap Epigenomics projects, learning to predict regulatory activity from DNA sequence alone. Enformer achieves 66% correlation with experimental measurements, far exceeding previous methods, and can identify enhancers, promoters, and other regulatory elements without requiring expensive experimental validation.

Given Nigeria's disproportionate SCD burden and limited access to existing therapies, identifying precise molecular targets for simpler, more affordable interventions is critical. This study leverages artificial intelligence to map the regulatory landscape of HBG1 and HBG2 genes, with specific objectives to: (1) identify enhancer elements controlling HBG1 and HBG2 expression; (2) characterise the regulatory signals (histone marks, chromatin accessibility, transcription) at these elements; (3) determine the spatial relationship between regulatory elements and target genes; and (4) provide precise genomic coordinates for future therapeutic development.

MATERIALS AND METHODS

Genomic region selection

We analysed the beta-globin gene cluster on chromosome 11, focusing on a 393,216 bp region (chr11:5,072,392–5,465,608) centred on the HBG1 gene (chr11:5,269,000) and encompassing the nearby HBG2 gene (chr11:5,276,000). This window size was chosen to match Enformer's input requirements while capturing both proximal and distal regulatory elements.

All analysis code and scripts used in this study are publicly available at https://doi.org/10.5281/zenodo.18824086 and https://github.com/firdausicopywriter-crypto/hbg-enformer-analysis

Sequence retrieval

Genomic sequences were retrieved from the Ensembl Genome Browser (https://www.ensembl.org), using the hg38/GRCh38 human reference assembly. DNA sequences were extracted in FASTA format and one-hot encoded (A= [1,0,0,0], C= [0,1,0,0], G= [0,0,1,0], T= [0,0,0,1]) for model input.

Enformer model architecture and predictions

Enformer is a transformer-based neural network with the following characteristics: Input: 393,216 bp DNA sequence (one-hot encoded); Architecture: Convolutional layers + transformer blocks with attention mechanisms; Output: 5,313 genomic tracks predicted across 896 bins (439 bp per bin); Effective receptive field: 200 kb, enabling long-range interaction modelling. The model was accessed via TensorFlow Hub (https://tfhub.dev/deepmind/enformer/1) and run using Python 3.8 with TensorFlow 2.13. Key predicted tracks analysed included CAGE (Cap Analysis of Gene Expression): transcription start site activity (Track indices 10 and 50); DNase-seq: chromatin accessibility (Track index 100); H3K27ac: histone modification marking active enhancers and promoters (Track index 200); and H3K4me3: histone modification specific to active promoters (Track index 500). Track indices refer to Enformer's output ordering as accessed via TensorFlow Hub and are provided here to enable exact reproduction of the tracks analysed.

Peak detection and enhancer identification

We identified candidate regulatory elements using the following criteria: (1) H3K27ac signal > 0.15 (indicating active enhancer/promoter status); (2) DNase signal > 5.0 (indicating open, accessible chromatin); and (3) presence of multiple signal types co-localising. Peaks were ranked by H3K27ac signal strength, as this mark strongly correlates with enhancer activity in erythroid cells (Creyghton et al., 2010).

Transcription factor binding site (Motif) analysis

The 438 bp sequence of the primary enhancer (chr11:5,308,936-5,309,374, hg38) was retrieved from NCBI RefSeq (chromosome 11, NC_000011.10) using Biopython's Entrez module. Position weight matrices for GATA1 (JASPAR matrix MA0035.5 and MA2388.1), KLF1, and TAL1: TCF3 (MA0091.2) were obtained from the JASPAR 2024 database (Castro-Mondragon et al., 2022) via the pyjaspar Python package. Each matrix was scanned against the enhancer sequence on both strands using Biopython's PSSM search function at a log-odds threshold of 4.0. No KLF1 motif matches were detected even when the threshold was relaxed to 2.0.

Computational resources and code availability

All analyses were performed using Python 3.8 on standard computing hardware. Analysis code and processed data are publicly available at https://github.com/firdausicopywriter-crypto/hbg-enformer-analysis (DOI: 10.5281/zenodo.18824086), ensuring full reproducibility of our findings.

RESULTS

Overview of regulatory landscape

Enformer prediction across the 393 kb HBG1-centred region revealed multiple peaks of regulatory activity. The strongest signals clustered in three main regions: (1) Upstream region

(~118 kb before HBG1): Locus Control Region (LCR); (2) Gene bodies (HBG2 and HBG1): Promoter activity; and (3) Downstream region (40 kb after HBG1): Enhancer cluster.

Primary enhancer discovery

The strongest regulatory element was identified at chr11:5,308,936–5,309,374, a 438 bp region located 40 kb downstream of the HBG1 transcription start site (Figure 2). This primary enhancer exhibited H3K27ac: 0.661 (3.3-fold higher than LCR); DNase: 15.0 (indicating highly open chromatin); CAGE: Moderate signal (2.5–10.0) suggesting nearby transcriptional activity; and H3K4me3: 0.6 (indicating some promoter-like features). (Figure 1).

Transcription factor motif analysis

To evaluate the reviewer-raised concern regarding motif enrichment, we scanned the full 438 bp primary enhancer sequence against JASPAR position weight matrices for GATA1, KLF1, and TAL1::TCF3 (log-odds threshold ⩾4.0). This identified nine total matches: six GATA1 matches (scores 4.54-7.59) and two TAL1::TCF3 matches (scores 8.15-9.17); no KLF1 match was detected, even at a relaxed threshold of 2.0 (Table 1). These results confirm GATA1 and TAL1::TCF3 binding potential at this locus but do not support a role for KLF1 at this specific enhancer.

Table 1: JASPAR-Predicted Transcription Factor Binding Sites within the 438 Bp Primary Enhancer (Chr11:5,308,936-5,309,374, Hg38)

Factor	JASPAR Motif ID	Position (bp)	Score	Strand
TAL1::TCF3	MA0091.2	326	9.17	-
TAL1::TCF3	MA0091.2	373	8.15	+
GATA1	MA0035.5	200	7.59	+
GATA1	MA0035.5	120	7.19	-
GATA1	MA0035.5	55	6.95	-
GATA1	MA0035.5	346	6.89	-
GATA1	MA0035.5	312	4.96	+
GATA1	MA0035.5	69	4.62	+
GATA1	MA2388.1	67	4.54	-

Shared regulatory architecture between HBG1 and HBG2

Comparison of HBG1-centred and HBG2-centred analyses revealed that both genes respond to the same regulatory elements: both genes are ~60–67 kb downstream from the LCR; the primary enhancer is located 40 kb from HBG1 and 33 kb from HBG2 (Figure 3).; and promoter activity signals are similar at both gene bodies.

The code and computational pipeline used to generate all results in this study are available on GitHub at https://github.com/firdausicopywriter-crypto/hbg-enformer-analysis and archived on Zenodo at https://doi.org/10.5281/zenodo.18824086

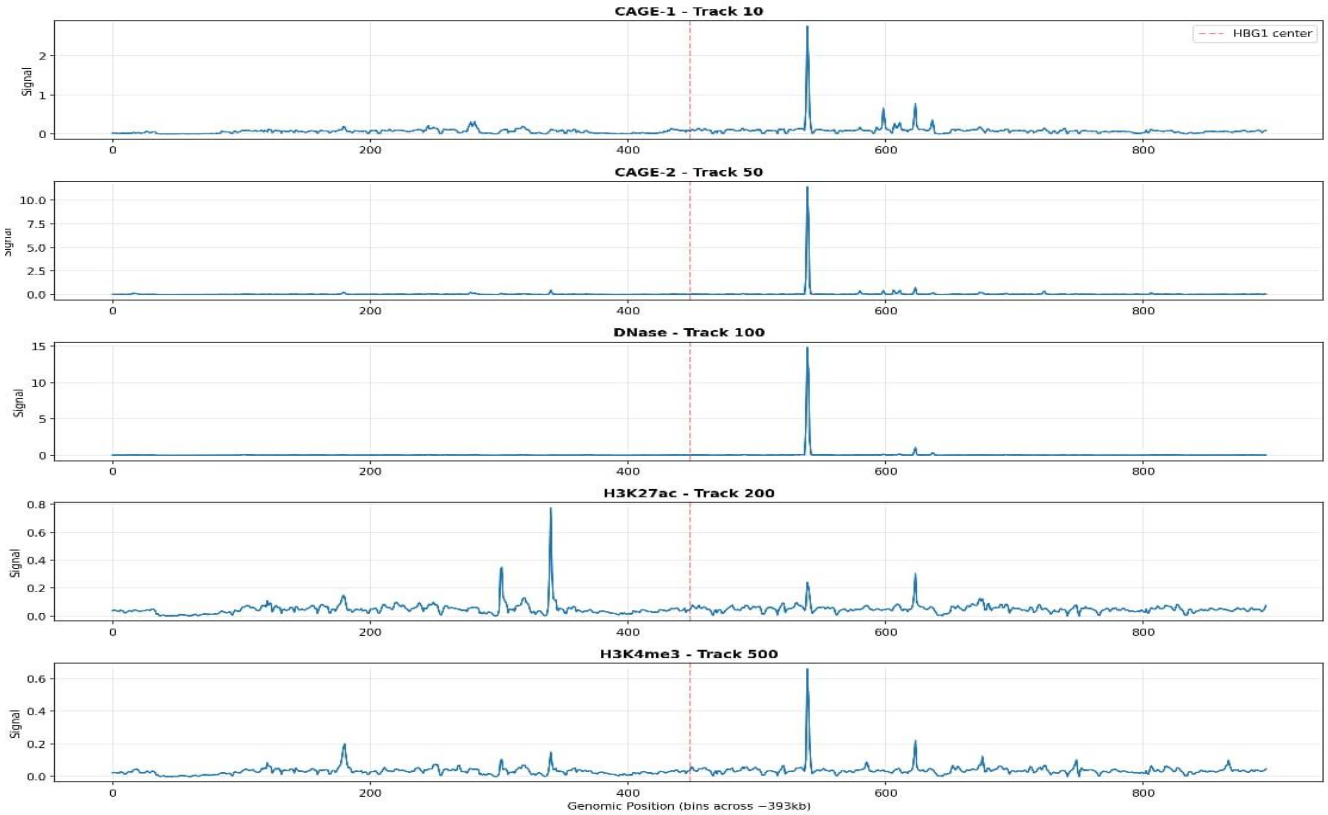


Figure 1: Enformer-predicted regulatory activity across the HBG1 genomic locus. Five epigenomic tracks are shown spanning approximately 393kb of the chromosome 11 locus centred on HBG1: CAGE-1 (Track 10), CAGE-2 (Track 50), DNase hypersensitivity (Track 100), H3K27ac (Track 200), and H3K4me3 (Track 500). The red dashed vertical line marks the HBG1 gene centre (bin 448). Prominent peaks downstream of HBG1 indicate predicted regions of active transcription and enhancer activity.

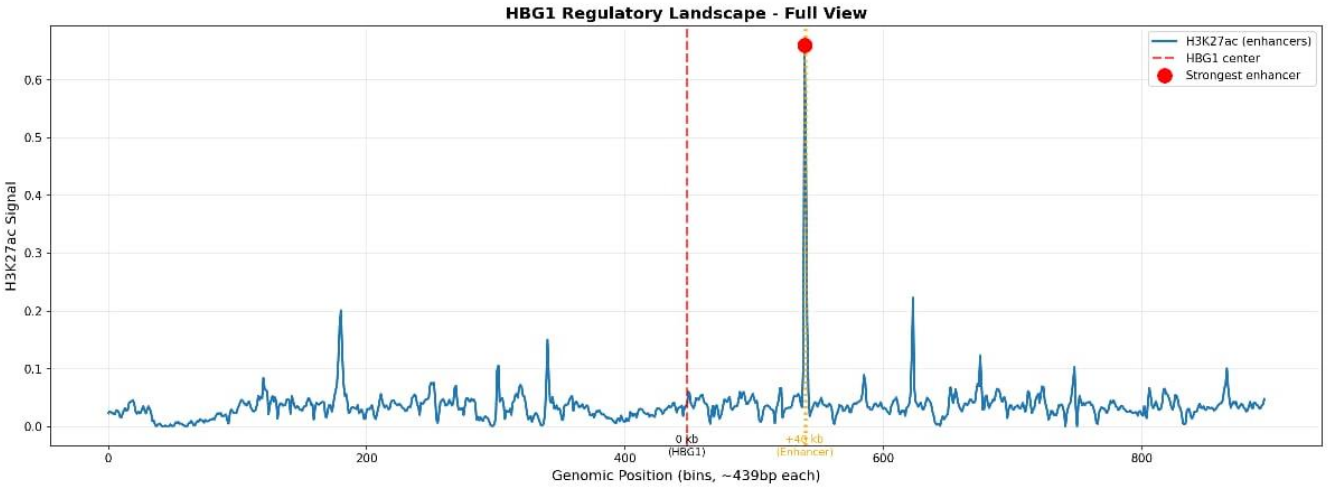


Figure 2: Full HBG1 regulatory landscape showing H3K27ac signal across the 393kb genomic window. The strongest predicted enhancer signal (0.661) is marked by a red dot at genomic coordinates chr11:5,308,936–5,309,374 (hg38), located +40kb downstream of HBG1. The red dashed line marks the HBG1 gene centre. This region represents the highest confidence predicted active enhancer element in the locus.

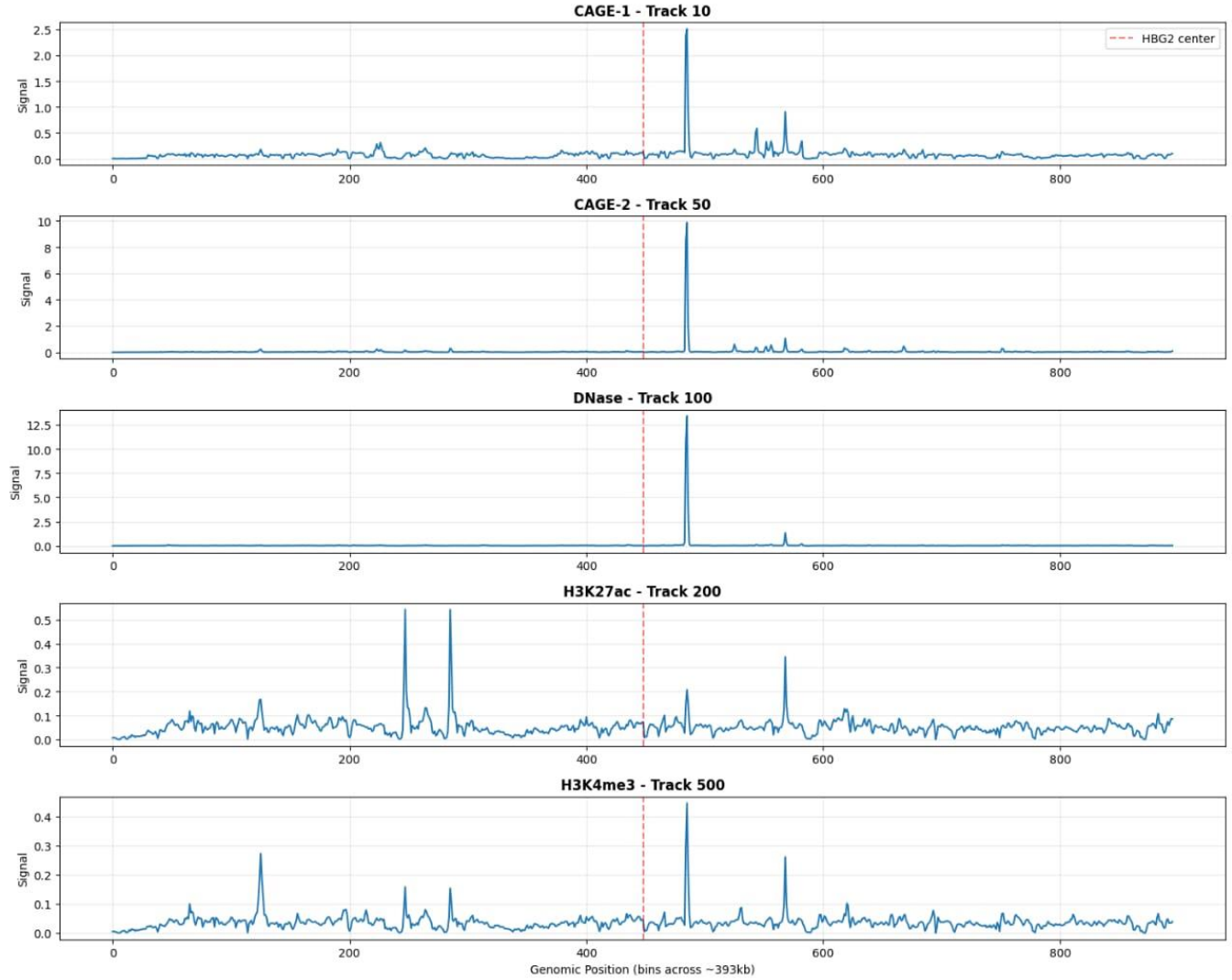


Figure 3: Enformer-predicted regulatory activity across the HBG2 genomic locus. Five epigenomic tracks are shown spanning approximately 393kb centred on HBG2: CAGE-1 (Track 10), CAGE-2 (Track 50), DNase hypersensitivity (Track 100), H3K27ac (Track 200), and H3K4me3 (Track 500). The red dashed vertical line marks the HBG2 gene centre. The shared downstream regulatory peaks observed in both HBG1 and HBG2 analyses suggest coordinated enhancer control of both fetal haemoglobin genes.

DISCUSSION

This study provides an AI-based mapping of regulatory elements controlling fetal haemoglobin genes HBG1 and HBG2 in a Nigerian context. We identified a candidate enhancer 40 kb downstream of HBG1/HBG2 with the strongest predicted activity in the region (H3K27ac = 0.661) and confirmed GATA1 and TAL1::TCF3 binding motifs within this element by JASPAR scanning. This genomic neighbourhood, downstream of HBG2 near BGLT3/HBBP1, has been previously implicated in HbF regulation: Grevet et al. (2021) identified a CTCF binding site, 3'HS1, in this region and showed that its deletion induces HbF expression through a mechanism that is independent of BCL11A protein levels. Our computational finding of a strong H3K27ac/DNase signal in this neighbourhood is therefore consistent with prior experimental evidence that this region is regulatory, though the precise mechanism, whether BCL11A-mediated, CTCF-mediated, or both, requires direct experimental testing and cannot be resolved from sequence-based prediction alone.

We propose a hierarchical regulatory model, to be tested experimentally, where the Locus Control Region provides constitutive activation signal to the beta-globin locus, whilst the +40 kb enhancer may act as a selector element influencing whether LCR activity reaches HBG1/HBG2. Our motif analysis found that this enhancer contains predicted binding sites for the erythroid activators GATA1 and TAL1::TCF3, but not KLF1, so any model invoking KLF1 occupancy at this specific element is not supported by our data. Whether BCL11A itself binds this enhancer in adults to block LCR-gene communication (Sankaran et al., 2008) remains a hypothesis requiring direct validation (e.g., ChIP-qPCR), as Enformer predicts chromatin activity rather than the identity of the bound factor. The stronger signal at our +40 kb enhancer (0.661) compared to the LCR (0.201) reflects Enformer's training on adult cell lines where the LCR actively regulates adult beta-globin rather than fetal genes.

With 150,000 affected births annually, Nigeria urgently needs scalable SCD therapies. Current gene therapies (Casgevy, Lyfgenia) require ₦1.2–1.8 billion per patient and are completely inaccessible in Nigerian healthcare contexts (FDA, 2023). Our identification of the exact 438 bp enhancer enables several new approaches: minimal base editing (changing <10 nucleotides instead of deleting large regions), small molecule development (drugs preventing BCL11A binding, potentially delivered orally), and antisense oligonucleotides (monthly/quarterly injections). These strategies could cost ₦50,000–200,000 annually instead of ₦1.2 billion one-time, making them scalable to Nigeria's primary healthcare system.

This study demonstrates that cutting-edge genomic research can be conducted using publicly available AI models, open-access databases, and standard computing resources without laboratory wetwork. This computational approach is particularly suited to resource-limited settings, enabling Nigerian researchers to contribute to global therapeutic development.

Whilst Enformer predictions correlate well with experimental data (r=0.66), our findings require experimental validation through ChIP-qPCR to confirm GATA1 and TAL1::TCF3 (and, if hypothesised, BCL11A) occupancy at this locus; CRISPR deletion of the 438 bp enhancer to measure HBG1/HBG2 expression changes; and reporter assays testing enhancer activity. Additionally, Enformer was trained primarily on non-African reference genomes (GRCh38), which may not capture population-specific regulatory variation. This is a particular limitation for translational claims in a Nigerian context: SCD is most prevalent in Yoruba and other West African populations, and non-coding variants segregating in these populations at the 438 bp enhancer, or elsewhere in the 393 kb window, could alter the regulatory predictions reported here. We have previously used 1000 Genomes Yoruba (YRI) personalised sequences with Enformer to study regulatory variation at the PKLR locus in sickle cell individuals; applying the same population-specific variant approach to this HBG1/HBG2 enhancer is an important direction for follow-up work, and would allow direct assessment of whether known African-population SNPs map to this 438 bp region.

CONCLUSION

We have identified and characterised a candidate 438 bp enhancer element 40 kb downstream of HBG1/HBG2, with confirmed GATA1 and TAL1::TCF3 binding motifs, situated in a genomic neighbourhood with prior experimental evidence of regulatory activity (Grevet et al., 2021). Whether this element acts through BCL11A, CTCF-dependent mechanisms, or both remains to be resolved experimentally. This enhancer shows the strongest predicted regulatory activity in the region and is shared between both fetal haemoglobin genes, enabling coordinated developmental regulation. Our findings provide precise molecular coordinates (chr11:5,308,936–5,309,374, hg38) for targeted therapeutic interventions that could be more affordable and accessible than current gene therapies. Given Nigeria's overwhelming SCD burden, these findings establish a foundation for developing the next generation of HbF-inducing therapies that could realistically reach the 150,000 Nigerian children born with SCD each year.

AUTHORS' CONTRIBUTIONS

Dr Aimola Idowu: Supervision, Resources, Writing – review and editing, Project administration. Firdausi Suleiman Muhammad: Conceptualisation, Methodology, Software,

Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualisation, Project leadership. Samuel Agida Obo: Main manuscript body and Abstract preparation. Precious Nkechi Onuoha: Keywords development and Introduction. Favour Omojo Jaja: Results section and presentation; Usman Usman Umar: Discussion and Conclusion sections. Olamide Emmanuel Abolarin: Acknowledgement and contribution statement section. Ismail Adamu: Title page preparation and formatting. All authors contributed to manuscript revision,approved the final manuscript, and agreed to be accountable for all aspects of the work

CONFLICT OF INTEREST

The authors declare no conflict of interest

ACKNOWLEDGEMENT

The authors gratefully acknowledge the developers of Enformer for making the model publicly available for genomic research applications. We also acknowledge publicly accessible genomic resources, including the human reference genome (hg38) and datasets generated by the ENCODE Project Consortium, which supported comparative interpretation of regulatory elements in this study.

We extend special Dr Aliyu Muhammad and Mr Idris Zubairu for their unwavering support and guidance.

REFERENCES

Akinsheye, I., Alsultan, A., Solovieff, N., Ngo, D., Baldwin, C. T., Sebastiani, P., Chui, D. H. K., and Steinberg, M. H. (2011). Fetal hemoglobin in sickle cell anemia. *Blood*, 118(1), 19–27. https://doi.org/10.1182/blood-2011-03-325258

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10), 1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50), 21931–21936. https://doi.org/10.1073/pnas.1016071107

FDA. (2023). *FDA approves first gene therapies to treat patients with sickle cell disease*. U.S. Food and Drug Administration. Retrieved from https://www.fda.gov/news-events/press-announcements/fda-approves-first-gene-therapies-treat-patients-sickle-cell-disease

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Perez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., and Mathelier, A. (2022). JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1), D165-D173. https://doi.org/10.1093/nar/gkab1113

Grevet, J. D., Lan, X., Hamagami, N., Edwards, C. R., Sankaranarayanan, L., Ji, X., Bhardwaj, S. K., Face, C. J., Posocco, D. F., Abdulmalik, O., Keller, C. A., Giardine, B., Sidoli, S., Garcia, B. A., Chou, S. T., Liebhaber, S. A., Hardison, R. C., and Blobel, G. A. (2021). 3′HS1 CTCF binding site in human beta-globin locus regulates fetal hemoglobin expression. *eLife*, 10, e70557. https://doi.org/10.7554/eLife.70557

Muhammad, F. (2026). *HBG1/HBG2 Fetal Hemoglobin Regulatory Analysis Using Enformer* (Version 1.0.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.18824086

Piel, F. B., Hay, S. I., Gupta, S., Weatherall, D. J., and Williams, T. N. (2013). Global burden of sickle cell anaemia in children under five, 2010–2050: Modelling based on demographics, excess mortality, and interventions. *PLOS Medicine*, 10(7), e1001484. https://doi.org/10.1371/journal.pmed.1001484

Sankaran, V. G., Menne, T. F., Xu, J., Akie, T. E., Lettre, G., Van Handel, B., Mikkola, H. K., Hirschhorn, J. N., Cantor, A. B., and Orkin, S. H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science*, 322(5909), 1839–1842. https://doi.org/10.1126/science.1165409

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, 20(8), 437–455. https://doi.org/10.1038/s41576-019-0128-0

Ware, R. E., de Montalembert, M., Tshilolo, L., and Abboud, M. R. (2017). Sickle cell disease. *The Lancet*, 390(10091), 311–323. https://doi.org/10.1016/S0140-6736(17)30193-9

Wilber, A., Nienhuis, A. W., and Persons, D. A. (2011). Transcriptional regulation of fetal to adult hemoglobin switching: New therapeutic opportunities. *Blood*, 117(15), 3945–3953. https://doi.org/10.1182/blood-2010-11-316893

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher. The publisher remains neutral with regard to jurisdictional claims.

Submit your next manuscript to NJBMB at https://www.nsbmb.org.ng/journals