Debunking the “junk”: Unraveling the role of lncRNA–miRNA–mRNA networks in fetal hemoglobin regulation

Motiur Rahaman^a, Chiranjib Bhowmick^a, Jaikrishna Komanapalli^b, Mandrita Mukherjee^a,
Prasanna Kumar Byram^a, Praphulla Chandra Shukla^a, Tuphan Kanti Dolai^c, Nishant
Chakravorty^{a*}

*^aSchool of Medical Science and Technology, IIT Kharagpur, Kharagpur, Paschim Medinipur,
West Bengal – 721302, India*

*^bDepartment of Biotechnology, IIT Kharagpur, Kharagpur, Paschim Medinipur, West Bengal –
721302, India*

*^cDepartment of Hematology, Nil Ratan Sircar Medical College and Hospital, Kolkata, West
Bengal- 700014, India*

*Address of corresponding author:

Dr. Nishant Chakravorty

Assistant Professor

School of Medical Science and Technology

Indian Institute of Technology, Kharagpur

Paschim Medinipur, West Bengal -721302, India

Email: nishant@smst.iitkgp.ac.in

Contact no: +91 9880188877

Running head title- Non-coding RNA mediated HbF regulation

Keywords: β -hemoglobinopathies; Fetal hemoglobin; non-coding RNAs; Machine learning;

lncRNA-miRNA-mRNA networks

Abstract

Fetal hemoglobin (HbF) induction is considered to be a promising therapeutic strategy to ameliorate the clinical severity of β -hemoglobin disorders, and has gained a significant amount of attention in recent times. Despite the enormous efforts towards the pharmacological intervention of HbF reactivation, progress has been stymied due to limited understanding of γ -globin gene regulation. In this study, we intended to investigate the implications of lncRNA-associated competing endogenous RNA (**ceRNA**) interactions in HbF regulation. Probe repurposing strategies for extraction of lncRNA signatures and subsequent *in silico* analysis on publicly available datasets (**GSE13284**, **GSE71935** and **GSE7874**) enabled us to identify 46 differentially expressed lncRNAs (**DElncRNAs**). Further, an optimum set of 11 lncRNAs that could distinguish between high HbF and normal conditions were predicted from these DElncRNAs using supervised machine learning and a stepwise selection model. The candidate lncRNAs were then linked with differentially expressed miRNAs and mRNAs to identify lncRNA-miRNA-mRNA ceRNA networks. The network revealed that 2 lncRNAs (UCA1 and ZEB1-AS1) and 4 miRNAs (hsa-miR-19b-3p, hsa-miR-3646, hsa-miR-937 and hsa-miR-548j) sequentially mediate cross-talk among different signaling pathways which provide novel insights into the lncRNA-mediated regulatory mechanisms, and thus lay the foundation of future studies to identify lncRNA-mediated therapeutic targets for HbF reactivation.

1. Introduction:

Fetal hemoglobin ($\alpha_2\gamma_2$) is the predominant form of hemoglobin at the time of birth. The switching from γ -globin to β -globin gene expression occurs during ontogeny, leading to the replacement of fetal hemoglobin (HbF) to adult hemoglobin (HbA) in a gradual manner (Bank 2006). The complex globin gene switching mechanisms are fundamentally important to understand higher eukaryotic transcriptional regulation. Fetal hemoglobin represents less than 1% of total hemoglobin in adult individuals; and has low clinical relevance in normal physiology (Sankaran and Orkin 2013). However, congenital, acquired, and drug-induced strategies for HbF induction are reported to ameliorate the clinical manifestations of β -hemoglobin disorders such as Sickle cell anemia (SCA) and β -thalassemia disease by reducing the α -globin chain precipitation and providing an alternate form of functional hemoglobin. Therefore, understanding the intricate mechanisms of γ -globin gene regulation are expected to have direct translational clinical relevance and hence lead to potential therapeutic advancements in β -hemoglobinopathies (Demirci, Leonard, and Tisdale 2020).

Numerous factors, covering both genetic and epigenetic alterations are known to influence the regulation of γ -globin genes in mature adults. Although certain hereditary conditions such as Hereditary persistence of fetal hemoglobin (HPFH), delta-beta thalassemia ($\delta\beta$ -thalassemia) and pathological conditions like Juvenile myelomonocytic leukemia (JMML) provide us with natural platforms to gain useful insights regarding γ -globin gene regulation in adulthood (Fluhr et al. 2017) (Wienert et al. 2015); however, a detailed understanding of the mechanisms to elevate γ -globin expression are in still primitive stage and attracts prime attention of the researchers throughout the world.

As the search for methods to elevate HbF levels to treat β -hemoglobin disorders such as sickle cell anemia (SCA), beta-thalassemia etc. continues, growing evidence suggests a major role of non-coding RNAs as modifiers of fetal hemoglobin expression (Sumera et al. 2020). Long non-coding RNAs (lncRNAs) - a major class of non-coding RNAs (ncRNAs) longer than 200 nucleotides, have become the focus of study in the recent years. Previously thought as transcriptional noise, these long non-coding RNAs are reported to exert their regulatory role at different levels including transcription, post-transcription and translation, and can modulate protein expression, thus playing pivotal roles in various physiological and pathological processes like cell division, cellular differentiation, embryonic development, carcinogenesis and neurodegenerative diseases (Kazemzadeh, Safaralizadeh, and Orang 2015). In the last decade, many studies have shown that lncRNAs have important implications in hematopoiesis and blood-related pathophysiologies (Xueqing Zhang, Weissman, and Newburger 2014)(Alvarez-Dominguez et al. 2014)(Nobili, Lionetti, and Neri 2016). Recently, it has reported that lncRNAs can interact with microRNAs (miRNA). miRNAs are major subsets of ncRNAs which belong to the category of small non-coding RNAs (sncRNAs) class (<200 nucleotides in length). They are known to regulate target gene expression by inhibiting mRNA translation or promoting mRNA degradation (Gebert and MacRae 2019). LncRNAs can function as competing endogenous networks (ceRNA) with mRNAs and small non-coding RNAs. They are reported to act as miRNA sponges that can sequester miRNAs and consequently inhibit their interactions with target mRNAs. LncRNAs can directly bind to the 3'UTR of mRNA, the binding site of miRNA, and thus preventing miRNA-mediated repression or degradation (López-Urrutia et al. 2019)(Paraskevopoulou and Hatzigeorgiou 2016). On the contrary, some miRNAs are reported to modulate lncRNA expression (Huang 2018). In addition, lncRNAs have been shown to act as

precursors of the miRNAs and thus can regulate miRNA biogenesis in different pathophysiological conditions (Fernandes et al. 2019).

Recently, some studies have demonstrated the emerging role of lncRNAs in the regulation of fetal hemoglobin (HbF) expression (Ivaldi et al. 2018) (Jia et al. 2019). It is expected that identification of new lncRNAs would contribute substantially to our understanding of the regulatory mechanisms of the γ -globin genes and hence bring us closer to developing new therapeutic strategies to treat beta-hemoglobinopathies like β -thalassemia and SCA disease. Genome-wide association studies (GWAS) have found that three major gene loci – *HBB*, *HBSIL-MYB* intergenic region, and *BCL11A*, account for more than 20-45% of the HbF variations amongst various population (Uda et al. 2008). In the *HBSIL-MYB* locus, a 3bp deletion polymorphism in an enhancer sequence leads to the synthesis of a 1283bp transcript HMI-lncRNA. Upon knockdown of this lncRNA, gamma globin expression increased 200 folds. The binding site of this transcript and downstream mechanisms are yet to be elucidated but it has been hypothesized to interact with the *MYB* promoter to regulate HbF levels (Morrison et al. 2018). Another lncRNA encoded downstream of γ -globin, known as BGLT3, actively contributes to the increased expression of HbF. The precise mechanism for this action is by looping between the γ -globin gene and BGLT3 sequences and interaction with chromatin mediator complex (Ivaldi et al. 2018). These significant studies conducted till now are suggestive of the pivotal role played by long non-coding RNAs in facilitating the transcriptional control of γ -globin synthesis. However, to best of our knowledge no study has focused on the possible involvement of the lncRNA-miRNA-mRNA-signaling pathway network in HbF regulation and their effects on HbF induction so far.

In the present study, we've explored the possible molecular regulation of γ -globin expression by predicting intricate networks of lncRNA-miRNA-mRNA-signaling pathway axis involved in HbF regulation. We have obtained datasets GSE13284, GSE71935 and GSE7874 (Sankaran et al. 2008)(Helsmoortel et al. 2016)(Tanno et al. 2007) from the publicly available Gene Expression Omnibus (GEO) database and carried out *in silico* analysis to identify the differentially expressed lncRNAs. Supervised Machine Learning methods were employed to construct the lncRNA-based classifier to distinguish between high and normal HbF condition. Additionally, we performed a feature selection procedure to identify optimal set of lncRNAs that are possibly associated with high HbF level. We analyzed Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and constructed the ceRNA network by connecting lncRNA-miRNA-mRNA-signaling pathway axis to gain new insights of intricate mechanisms of HbF regulation which could suggest novel targets of HbF reactivation in β -hemoglobinopathies. The overall workflow is schematically represented in Figure 1.

Figure 1. Flowchart depicting methodological workflow of predicted lncRNA- miRNA crosstalk in HbF regulation.

2. Results

2.1. LncRNA expression profiles on Affymetrix arrays

To extract lncRNA expression profiles, all datasets were transformed into probeID-centric data. A total of 54,675 entries were obtained for each dataset. By probe-repurposing and annotating

the probe-sets (RefSeq and Ensemble annotations), we have obtained 2248 unique lncRNAs in our datasets (Supplementary file 1) which were utilized for further analysis. The probe sets with ambiguous annotations were excluded from our study.

2.2. Distinctive lncRNA expression between high HbF and normal HbF condition

The lncRNA expression patterns were compared between high HbF and normal HbF condition to identify the lncRNAs which are intricately associated with fetal hemoglobin regulation. Differentially expressed lncRNAs between two study groups were identified by performing Significant Analysis of Microarrays (SAM) analysis in training dataset, GSE13284. We obtained a total of 46 differentially expressed lncRNAs that had significant differential expression between the studied groups (Figure 2). Among the 46 deregulated lncRNAs, 20 lncRNAs were found to be up-regulated and the rest, 26 lncRNAs were identified to be down-regulated in patients (Supplementary file 2).

Figure 2. Differential expression of lncRNA probe sets between high HbF and normal condition. A total of 46 differentially expressed lncRNAs were identified to be significant in SAM analysis. **A.** The hierarchical clustering of lncRNA probe sets in training dataset, GSE13284. **B.** Validation of lncRNA probe sets in the discovery or test dataset, GSE7874. T-statistics with FDR < 25% and Permutation of 1000 were performed to determine significant genes. Red = high expression; Green = low expression. Bar colors indicate sample type: yellow, high HbF; blue, normal HbF.

2.3. Supervised machine learning approach to determine optimal set of lncRNA, which could be involved in fetal hemoglobin regulation

In order to identify optimal combination of lncRNAs which could play a crucial role in fetal hemoglobin regulation, we performed an lncRNA-based classification and utilized feature selection methods with the 46 differentially expressed lncRNAs using a supervised machine learning algorithm, SVM with radial basis function (RBF) kernel and stepwise selection model (Figure 3).

Figure 3. Showing features selection procedure to identify optimal set of lncRNA from differentially expressed lncRNAs. Stepwise selection method, involving forward selection was used.

A total of 11 (**C6orf223**, **AP000679.1**, **GSEC**, **UCA1**, **CT66**, **MIR4453HG**, **AL135925.1**, **ZEB1-AS1**, **AC091057.1**, **PPM1F-AS1** and **LINC02249**) out 46 differentially expressed lncRNAs were identified to be optimal combination which could form the lncRNA-based

network for fetal hemoglobin regulation. Three machine learning classifiers, i.e. Random Forest, Support Vector Machine with linear kernel (SVM-L) and Support Vector Machine with RBF kernel (SVM-RBF), were tried, and we have calculated the performance metrics using the optimal set of 11 lncRNAs. The best results for the test set data were obtained for the SVM-RBF classifier with Accuracy and Precision of 89.47% and 93.33 %, respectively. The performance metrics for the test set data using all the classifiers is shown in Table 1.

Table 1. Performance matrix of different ML classifiers

ML Classifiers	Accuracy	Precision	Recall	F1 score	AUC
Random Forest	0.8421	0.8285	0.7948	0.808	0.8461
SVM(Linear)	0.7368	0.7023	0.7179	0.7076	0.7564
SVM(RBF)	0.8947	0.9333	0.8333	0.8642	0.8076

The Receiver Operating Characteristics (ROC) for all the classifiers employed in our study were plotted to show the discriminative power of each of this classifier. A comparison of the ROC curves computed for all the three classifiers is shown in Figure 4.

Figure 4. ROC curves of lncRNA-based classifiers in the discovery cohort.

2.4. Correlated differentially expressed genes which are important in lncRNA-mRNA networks construction

To investigate the functional implications of the optimal set of 11 lncRNAs, we performed a correlation test to identify the association between each of these 11 lncRNAs and mRNAs in GEO dataset, GSE13284. A list of 16,383 correlated genes was obtained for each lncRNAs. Of these co-expressed mRNAs, top (1%) positively and negatively correlated genes were selected for further study. In addition, 82 genes were found to be differentially expressed between high HbF and control group in the GSE13284 dataset (Sankaran et al. 2008) (Figure 5). The lists of differentially expressed genes are provided as Supplementary file 3.

Figure 5. The Hierarchical clustering of differentially expressed mRNAs between high HbF and control group in GSE13284. Expression heatmap of differentially expressed genes was generated using R v3.6.3. “Red” and “Blue” colors represent above and below expression, respectively, relative expression.

Common genes between top correlated genes and differentially expressed genes were used to construct lncRNA-mRNA networks.

2.5. Functional implications of optimal set of 11 lncRNAs

We performed GO term and functional enrichment analysis of shared mRNAs to gain insights regarding biological functions and cellular signaling pathways using DAVID resources. GO analysis revealed that the most enriched biological process (bp) was the glycolytic process (GO:

Figure 7. KEGG pathway analysis of common identified mRNAs. The 6 significant ($p < 0.05$) signaling pathways are shown. All values are negative log transformed.

2.6. Construction of miRNA-mRNA regulatory networks

Based on differential miRNA expression analysis performed in GEO2R, 9 miRNAs (5 up-regulated and 4 down-regulated) were differentially expressed in high HbF vs. normal HbF conditions (Supplementary file 4). Out of these 9 differentially expressed miRNAs, 4 miRNAs (3 up-regulated and 1 down-regulated) had their targets common with list of differentially expressed common correlated mRNAs. A total of 14 genes and 4 miRNAs were found to have mutually opposite expression patterns, which were utilized for miRNA-mRNA networks construction (Figure 8).

Figure 8. The figure is showing predicted miRNA-mRNA network. The network contains 19 nodes and 17 edges. The elliptical and diamond shapes represent mRNA and miRNA, respectively. Red and green colors indicate up-regulation and down-regulation, respectively. The directionality of the arrow is from the source to the target.

2.7. Prediction of lncRNA-miRNA networks

The interactions between identified lncRNAs and miRNAs were predicted using DIANA-lncBase v2. The differentially expressed miRNAs, which had shared targets with common correlated mRNAs and interacted with the identified optimal set lncRNAs, were selected for lncRNA-miRNA network construction (Supplementary file 5). Here, lncRNA-miRNA interactions with opposite expression are shown (Figure 9).

Figure 9. The figure is showing predicted lncRNA-miRNA network. The network consists of 6 nodes and 5 edges. The V-shapes and diamond represent lncRNA and miRNA, respectively. The red and green colors indicate up-regulation and down-regulation. Each arrow represents the interaction between lncRNA and miRNA, and the arrow is pointed from the source to the target.

2.8. Construction of lncRNA-miRNA-mRNA-signaling pathway network

Based on the predicted miRNA-mRNA, lncRNA-miRNA networks, we constructed lncRNA-miRNA-mRNA networks which were further linked with signaling pathways of the identified mRNA in the networks. In the figure 10, all the possible interactions among studied lncRNA, miRNAs and mRNAs are shown.

Based upon the common link between lncRNA-miRNA and miRNA-mRNA networks, we have identified the lncRNA-miRNA-mRNA-signaling pathway axis which is predicted to be important for HbF up-regulation (Figure 10).

Figure 10. The interactions among lncRNA, miRNA and mRNA in high HbF vs. normal HbF condition.

A. All possible networks among lncRNA, miRNA and mRNA in the studied groups. **B.** Predicted lncRNA-miRNA-mRNA-signaling pathway axis, which is important in HbF regulation. The V, Diamonds and elliptical shapes indicate lncRNA, miRNA and mRNA, respectively. Red= Up-regulation; Green= Down-regulation. The arrows show interactions and pointed from the source to the target.

3. Discussion

Advancements in genomic and transcriptomic studies suggest towards the critical regulatory functions of non-coding RNAs in complex human diseases (Kazemzadeh, Safaralizadeh, and Orang 2015). Recent studies have highlighted the crucial roles of lncRNA in normal and malignant hematopoiesis (Ghafouri-Fard, Niazi, and Taheri 2021; Nobili, Lionetti, and Neri 2016). A study on the construction of an lncRNA-associated ceRNA network revealed two candidates, NR_001589 and uc002fcj.1, that are significantly expressed in individuals with high HbF levels. The study predicted that these lncRNAs promote the expression of HBG1/2 by sponging miRNA and induce HbF production (Jia et al. 2019). Microarray analysis in another study to profile lncRNAs, miRNAs and mRNAs in individuals having hereditary persistence of fetal hemoglobin has elucidated six lncRNAs that competitively sponge miR-486-3p, therefore elevating HbF level (Lai et al. 2017). Although few studies have performed the lncRNA, miRNA expression profiles, and attempted to provide preliminary evidences for the involvement of lncRNA-miRNA interactions in HPFH and β -thalassemia disease (Lai et al. 2017) (Fakhr-Eldeen, Toraih, and Fawzy 2019), the decisive role of lncRNA-miRNA-mRNA networks in HbF regulation remains largely obscure. Recent developments in high-throughput technologies and advancements in bioinformatics analysis, fueled by powerful Machine learning tools, have revolutionized the field of medical informatics. Here, we have explored microarray datasets

involving high HbF vs. normal HbF condition, deposited in the GEO database, NCBI (Clough and Barrett 2016), and analyzed them using standard bioinformatics tools and machine learning algorithms to discern the complex regulatory networks involved in HbF induction. We have analyzed three independent microarray datasets (**GSE13284**, **GSE71935** and **GSE7874**) and identified 46 differentially expressed lncRNAs between high HbF and normal condition. To obtain optimal set of lncRNA that could distinguish between high HbF and normal HbF condition, we used different machine learning algorithms (Random Forest, Linear SVM and SVM-RBF), which further narrow down the numbers of lncRNA candidates. Comparative analysis revealed that the best results were obtained from the SVM-RBF model. A total of 46 differentially expressed lncRNAs were subjected to SVM-RBF and stepwise selection model, and we found a panel of 11 lncRNAs (**C6orf223**, **AP000679.1**, **GSEC**, **UCA1**, **CT66**, **MIR4453HG**, **AL135925.1**, **ZEB1-AS1**, **AC091057.1**, **PPM1F-AS1** and **LINC02249**) with accuracy of 0.8947 for distinguishing high HbF and normal control in the discovery cohort.

To identify the underlying molecular mechanisms and possible biological implications of these 11 candidate lncRNAs in HbF regulation, we attempted to correlate them with co-expressed mRNAs in GSE13284 dataset. We performed differential expression analysis (DEA) in the same dataset and correlated differentially expressed genes with 11 candidate lncRNAs. Functional enrichment analysis of the correlated genes revealed that these lncRNAs are involved in protein-binding (GO: 0005515), oxidoreductase activity (GO: 0016702), erythrocyte development (GO: 0046854), negative regulation of apoptotic process (GO: 0043066), etc. In an earlier study, Gabbianelli et al. 2008, have already reported that HbF reactivation inhibits cellular apoptosis in adult erythrocytes (Gabbianelli et al. 2008). Ma et al. demonstrated lncRNA HBBP1 increased γ -globin expression by activating ETS transcription factor ELK1 in HUDEP-2 cells (S.-P. Ma et al.

2021). A recent study has suggested the possible role of mRNA-binding protein in HbF reactivation in adult erythroid cells (Chambers et al. 2020). The fetal to adult hemoglobin switching is an important event in erythrocyte development (Sankaran, Xu, and Orkin 2010). KEGG pathway analysis revealed that major enriched pathways are involved in metabolic processes such as Glycolysis/Gluconeogenesis (hsa00010), Metabolic pathways (hsa01100), Biosynthesis of amino acids (hsa01230), Carbon metabolism (hsa01200) and Blood-cell development-related processes such as Hematopoietic cell lineage (hsa04640). Alterations in metabolic processes have already been reported in β -thalassemia patients with high HbF (Musharraf et al. 2017). In 2009, Testa et al. showed the role of stem cell factors in the reactivation of HbF in adults (Gabbianelli and Testa 2009).

Furthermore, we have also analyzed a separate miRNA expression dataset, **GSE93973**. Nine differentially expressed (adjacent p-value < 0.1, Log2FC > |0.8|) miRNAs (5 up-regulated and 4 down-regulated) were obtained, and their targets were predicted. Of these 9 deregulated miRNAs, 4 miRNAs shared the targets with common correlated differentially expressed mRNAs. Results revealed that hsa-miR-548ap-5p interacts with 9 out of 14 identified mRNAs in the network. Another miRNA, hsa-miR-933, is associated with 4 mRNAs in the network. Previous studies have demonstrated that intronic hsa-miR-933 plays a pivotal role in type II diabetes mellitus and neurodegenerative disease development (Bashar et al. 2020). To the best of our knowledge, no previous study has reported the possible link of this miRNAs to fetal hemoglobin development. Studies have identified that the loss of hsa-miR-638 facilitates cancer progression by targeting *SOX2*, an important transcriptional regulator of *BCL11A* (K. Ma et al. 2014) (Lazarus et al. 2018). In 2008, Sankaran et al. showed the role of *BCL11A* as a potential regulator of HbF expression (Sankaran et al. 2008). Another study investigated the effects of

different deleterious SNPs on BCL11A and Fetal hemoglobin reactivation (Das and Chakravorty 2020). Furthermore, we observed hsa-miR-3646 targets two mRNA, WNK3 and ALOX5, in a separate network. Although, previous reports have suggested the role of hsa-miR-3646 in breast cancer progression (Tao et al. 2016); however no studies have been found which indicate the role of hsa-miR-3646 in HbF regulation. Our study suggests a possible link of this miRNA with HbF regulation.

Using DIANA-LncBase v2, we have predicted possible interactions between miRNAs and the 11 candidate lncRNAs identified. Finally, based upon the data from lncRNA-miRNA, miRNA-mRNA, lncRNA-mRNA networks, we have constructed the lncRNA-miRNA-mRNA regulatory axis of HbF expression, which was further linked with signaling pathways. The networks demonstrate the potential mechanism by which lncRNAs and miRNAs are involved in HbF regulation. As Fig 8 indicates, 2 lncRNAs (UCA1 and ZEB1-AS1) and 4 miRNAs (hsa-miR-19b-3p, hsa-miR-3646, hsa-miR-937 and hsa-miR-548j) have been identified as critical players in lncRNA-miRNA-mRNA networks. Recently, a study by Liu et al. has reported the role of lncRNA-UCA1 in heme biosynthesis and erythroid development (Liu et al. 2018). Another study reported that lncRNA-ZEB-AS1 promotes cell proliferation by regulating the Wnt signaling pathway, a crucial pathway in embryonic development (Lv et al. 2018). In a study, Mnika et al. demonstrated that hsa-miR-19b is associated with HU-induced HbF production in SCA patients (Mnika et al. 2019). All these reports are in accordance with our findings. Furthermore, by linking signaling pathways with the lncRNA-miRNA-mRNA networks, we have identified ZEB-AS1-hsa-miR-3646-ALOX5-signaling pathway axis, which might play an important role in HbF regulation and thus, could potentially contribute to therapeutic developments in β -hemoglobinopathies. Previous reports suggested the role of *ALOX5* in modulating the fetal

inflammatory response in newborns (Costa and Castelo 2016). WNK signaling has also been reported to be associated with the deoxygenation of erythrocytes in normal physiology (Zheng et al. 2019). Infections pose a serious threat, and is reported to be the major cause of mortality and morbidity in thalassemic patients (Vento, Cainelli, and Cesario 2006). Although few studies have already reported the seroprevalence of *Toxoplasma* infection in β -hemoglobinopathies (Ferreira et al. 2017) (El-Tantawy, Darwish, and Eissa 2019), no study has reported the possible association of HbF regulation with toxoplasmosis. In our study, we have identified the possible link of ZEB-AS1-has-miR-3646-ALOX5 axis with toxoplasmosis. *ALOX5* gene has previously been reported to be involved in clearance of *Toxoplasma gondii* infection (Mashima and Okuyama 2015) (Aliberti, Serhan, and Sher 2002). Our study has linked all these signaling pathways, mRNA, miRNAs and lncRNAs, and suggested their possible role in HbF regulation.

In summary, this study represents a comprehensive analysis of possible molecular interactions among lncRNAs, miRNAs and mRNAs involved in HbF regulation. In this study, we have constructed an lncRNA-based classifier which could distinguish high HbF and normal condition with high specificity and sensitivity in discovery cohort. The candidate lncRNAs were then linked with differentially expressed miRNAs and mRNAs to predict lncRNA-miRNA-mRNA regulatory networks. Based on these networks, 2 lncRNAs (UCA1 and ZEB1-AS1) and 4 miRNAs (hsa-miR-19b-3p, hsa-miR-3646, hsa-miR-937 and hsa-miR-548j) identified to be crucial in HbF regulation. This analysis also revealed the ZEB1-AS1-hsa-miR3646-ALOX5-signaling pathways axis, which provides new insights about the molecular mechanisms of HbF regulation and suggests novel therapeutic targets to ameliorate disease severity in β -hemoglobinopathies.

4. Methods and materials:

4.1. Data acquisition and processing

In the present study, three independent microarray datasets, **GSE13284** (Sankaran et al. 2008), **GSE71935** (Helsmoortel et al. 2016) and **GSE7874** (Tanno et al. 2007), which include high fetal hemoglobin and control group, were retrieved from the publicly available Gene Expression Omnibus (GEO) database. In addition, an independent miRNA expression dataset, **GSE93973** (Lai et al. 2017) concerning hereditary persistence of fetal hemoglobin (HPFH) and β -thalassemia minor with high HbF and control group was accessed from the NCBI GEO dataset for miRNA-mRNA network construction (Lai et al. 2017). For lncRNA-based classification, datasets were selected according to the following criteria: (1) dataset having high HbF vs. normal HbF expression profiles ;(2) dataset with sample no.>3; and (3) only Affymetrix HG-U133 plus 2.0 arrays expression profiles were included in our study. The raw (.CEL) files were downloaded, quantile normalized, Log2transformed and background corrected by using Robust Multichip Average (RMA, Windows Version) (Sahlabadi et al. 2018). Finally, we obtained probe ID-centric gene expression datasets. These datasets were further analyzed for lncRNA-based classification as described elsewhere (Xiaoqin Zhang et al. 2012).

4.2. Retrieval and analysis of lncRNA expression profiles

LncRNA expression profiles were obtained by probe repurposing of Affymetrix HG-U133 Plus 2.0 array based upon the NetAffx annotations following a standard lncRNA based classification pipeline (Xiaoqin Zhang et al. 2012). The annotations include gene name, gene symbol, Ensemble and RefSeq IDs for every probe set ID present in the dataset. LncRNAs extractions were based

upon the Ensemble and RefSeq IDs of the probes. In case of RefSeq annotations, we extracted the probes labeled with 'NR' (NR stands for non-coding RNA in RefSeq database). Similarly, for probes with Ensemble gene IDs, we retrieved the probes labeled with 'linc RNA', 'non-coding RNA', 'miscRNA', etc. Following this, the other ncRNAs (miRNAs, snoRNAs, snRNAs, etc.) and pseudogenes were filtered out manually, retaining probes labeled with long non-coding RNAs only. Finally, 2248 unique lncRNAs corresponding to 4254 probe ids were obtained and used for downstream analysis.

4.3. Significance Analysis of Microarrays (SAM) analysis

The Significance Analysis of Microarrays (SAM) was performed using the R package (Chu et al. 2011) to evaluate the differentially expressed lncRNAs between the study groups (High HbF vs. Normal HbF). SAM analysis was carried out using stringent filters with a false discovery rate (FDR) <25% and permutations of 1000 to reduce ambiguity and increase accuracy. LncRNA probe sets with fold change (FC) > 1.5 and delta value 0.26 were defined as significantly different in SAM analysis. The differentially expressed lncRNAs were imported in Cluster view 3.0 (de Hoon et al. 2004) to carry out the Hierarchical Cluster Analysis (HCA).

4.4. LncRNA-based classification and prediction

Machine Learning Algorithms were utilized to build the lncRNA based classifier to distinguish between physiologically high HbF and normal HbF level in python 3.7 using scikit-learn package. All the three datasets, i.e. GSE13284, GSE7874 and GSE71935, were combined and implemented training-validation approach, in which 75% of the data was employed for training and the rest 25% was used for testing. The list of lncRNAs obtained after Significance Analysis

of Microarray (SAM) was fed to Sequential Forward Selector (SFS) using a supervised machine learning algorithm, i.e. Support Vector Machine (SVM) with radial basis function (RBF) kernel and 5 fold cross-validation to obtain the optimal combination of lncRNAs. Three machine learning classifiers, i.e. Random Forest, Support Vector Machine with linear kernel (SVM-L) and Support Vector Machine with RBF kernel (SVM-RBF), were used to determine the performance metrics using the optimal set of lncRNAs. To determine the discriminative power of the lncRNA-based classifier, a receiver operating characteristics (ROC) curve was plotted, and the area under the curve (AUC) was calculated.

4.5. Construction of lncRNA-mRNA networks

To gain insights regarding the functional spectrum of lncRNAs, we performed a Pearson correlation test and calculated paired correlation coefficient between lncRNAs and mRNAs using GSE13284 dataset to construct the lncRNA-mRNA network. Differential expression analysis was performed for the same dataset to obtain the differentially expressed genes between the studied groups. Correlated genes were then mapped against differentially expressed genes (DEGs), and common genes were extracted for further analysis.

4.6. Functional enrichment analysis

To evaluate the specific biological functions of these differentially expressed common correlated genes, we carried out Gene enrichment analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID) tool (Dennis et al. 2003). Additionally, enriched pathways associated with these genes were identified and visualized using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of DAVID software (Kanehisa and Goto 2000).

4.7. Construction of lncRNA-miRNA-mRNA regulatory networks

Differential expression analysis (DEA) was performed to obtain differentially expressed miRNAs (adjusted p-value < 0.1 and log2FC > |0.8|) between high HbF and normal conditions using publicly available miRNA expression dataset, GSE93973 (Lai et al. 2017), and the targets of these differentially expressed miRNAs were identified by using multiMiR R package and database (Ru et al. 2014). The multiMiR is a comprehensive collection of predicted and validated miRNA target prediction database which compiled data from 14 different databases. Differentially expressed miRNAs that have shared targets with common identified mRNA with opposite expression pattern were selected to construct miRNA-mRNA network. To obtain interactions among common lncRNAs and miRNAs, we used DIANA-LncBase v2 (Karagkouni et al. 2020). The common differentially expressed miRNAs were mapped against the optimal set of selected 11 lncRNAs for miRNA-lncRNA network constructions. Finally, based on miRNA-mRNA, lncRNA-miRNA and miRNA-mRNA-signaling pathway networks, we constructed lncRNA-miRNA-mRNA-signaling pathway network. All networks were visualized using Cytoscape v3.1 (Shannon et al. 2003).

Acknowledgements

Authors would like to thank Indian Institute of Technology Kharagpur for providing the infrastructural support. Mr. Motiur Rahaman would like to thank the Ministry of Education for financial support.

References

- Aliberti, Julio, Charles Serhan, and Alan Sher. 2002. “Parasite-Induced Lipoxin A4 Is an Endogenous Regulator of IL-12 Production and Immunopathology in Toxoplasma Gondii Infection.” *Journal of Experimental Medicine* 196 (9): 1253–62.
<https://doi.org/10.1084/jem.20021183>.
- Alvarez-Dominguez, Juan R, Wenqian Hu, Bingbing Yuan, Jiahai Shi, Staphany S Park, Austin A Gromatzky, Alexander van Oudenaarden, and Harvey F Lodish. 2014. “Global Discovery of Erythroid Long Noncoding RNAs Reveals Novel Regulators of Red Cell Maturation.” *Blood* 123 (4): 570–81. <https://doi.org/10.1182/blood-2013-10-530683>.
- Bank, Arthur. 2006. “Regulation of Human Fetal Hemoglobin: New Players, New Complexities.” *Blood* 107 (2): 435–43. <https://doi.org/10.1182/blood-2005-05-2113>.
- Bashar, Abul, Mir Khademul, Eusra Mohammad, and Abdullah-al-kamran Khan. 2020. “Aberration of the Modulatory Functions of Intronic MicroRNA Hsa-MiR-933 on Its Host Gene ATF2 Results in Type II Diabetes Mellitus and Neurodegenerative Disease Development,” 1–11.
- Chambers, Christopher B, Jeffrey Gross, Katherine Pratt, Xiang Guo, Colleen Byrnes, Y Terry Lee, Donald Lavelle, Ann Dean, Jeffery L Miller, and Andrew Wilber. 2020. “The MRNA-Binding Protein IGF2BP1 Restores Fetal Hemoglobin in Cultured Erythroid Cells from Patients with β -Hemoglobin Disorders.” *Molecular Therapy. Methods & Clinical Development* 17 (June): 429–40. <https://doi.org/10.1016/j.omtm.2020.01.011>.
- Chu, Gil, Jun Li, Balasubramanian Narasimhan, Robert Tibshirani, and Virginia Tusher. 2011.

“SAM - Significance Analysis of Microarrays - Users Guide and Technical Document.”

Policy, 1–42.

Clough, Emily, and Tanya Barrett. 2016. “The Gene Expression Omnibus Database.” *Methods in Molecular Biology (Clifton, N.J.)* 1418: 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.

Costa, Daniel, and Robert Castelo. 2016. “Umbilical Cord Gene Expression Reveals the Molecular Architecture of the Fetal Inflammatory Response in Extremely Preterm Newborns.” *Pediatric Research* 79 (3): 473–81. <https://doi.org/10.1038/pr.2015.233>.

Das, Sankha Subhra, and Nishant Chakravorty. 2020. “Identification of Deleterious SNPs and Their Effects on BCL11A, the Master Regulator of Fetal Hemoglobin Expression.” *Genomics* 112 (1): 397–403. <https://doi.org/https://doi.org/10.1016/j.ygeno.2019.03.002>.

Demirci, Selami, Alexis Leonard, and John F Tisdale. 2020. “Genome Editing Strategies for Fetal Hemoglobin Induction in Beta-Hemoglobinopathies.” *Human Molecular Genetics* 29 (R1): R100–106. <https://doi.org/10.1093/hmg/ddaa088>.

Dennis, Glynn, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. 2003. “DAVID: Database for Annotation, Visualization, and Integrated Discovery.” *Genome Biology* 4 (9): R60. <https://doi.org/10.1186/gb-2003-4-9-r60>.

El-Tantawy, Nora, Ahmad Darwish, and Eman Eissa. 2019. “Seroprevalence of Toxoplasma Gondii Infection Among B-Thalassemia Major Pediatric Population: Implications for Transfusion Transmissible Toxoplasmosis.” *The Pediatric Infectious Disease Journal* 38

(3): 236–40. <https://doi.org/10.1097/INF.0000000000002111>.

Fakhr-Eldeen, Abeer, Eman A Toraih, and Manal S Fawzy. 2019. “Long Non-Coding RNAs MALAT1, MIAT and ANRIL Gene Expression Profiles in Beta-Thalassemia Patients: A Cross-Sectional Analysis.” *Hematology* 24 (1): 308–17.
<https://doi.org/10.1080/16078454.2019.1570616>.

Fernandes, Juliane C R, Stephanie M Acuña, Juliana I Aoki, Lucile M Floeter-Winter, and Sandra M Muxel. 2019. “Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease.” *Non-Coding RNA* 5 (1). <https://doi.org/10.3390/ncrna5010017>.

Ferreira, Marina Neves, Claudia Regina Bonini-Domingos, Isabeth Fonseca Estevão, Clarice Lopes De Castro Lobo, Gisele Cristina Souza Carrocini, Aparecida Perpétuo Silveira-Carvalho, Octávio Ricci, Luiz Carlos De Mattos, and Cinara Cássia Brandão De Mattos. 2017. “Anti- Toxoplasma Gondii Antibodies in Patients with Beta-Hemoglobinopathies: The First Report in the Americas.” *BMC Research Notes* 1 (1): 1–7.
<https://doi.org/10.1186/s13104-017-2535-7>.

Fluhr, Silvia, Christopher Felix Krombholz, Angelina Meier, Thomas Epting, Oliver Mücke, Christoph Plass, Charlotte M Niemeyer, and Christian Flotho. 2017. “Epigenetic Dysregulation of the Erythropoietic Transcription Factor KLF1 and the β -like Globin Locus in Juvenile Myelomonocytic Leukemia.” *Epigenetics* 12 (8): 715–23.
<https://doi.org/10.1080/15592294.2017.1356959>.

Gabbianelli, Marco, Ornella Morsilli, Adriana Massa, Luca Pasquini, Paolo Cianiulli, Ugo Testa, and Cesare Peschle. 2008. “Effective Erythropoiesis and HbF Reactivation Induced by Kit Ligand in β -Thalassemia.” *Blood* 111 (1): 421–29.

<https://doi.org/https://doi.org/10.1182/blood-2007-06-097550>.

Gabbianelli, Marco, and Ugo Testa. 2009. “Role of f Stem Cell Factor in the Reactivation of f Human Fetal Hemoglobin.” <https://doi.org/10.4084/MJHID.2009.009>.

Gebert, Luca F R, and Ian J MacRae. 2019. “Regulation of MicroRNA Function in Animals.” *Nature Reviews Molecular Cell Biology* 20 (1): 21–37. <https://doi.org/10.1038/s41580-018-0045-7>.

Ghafouri-Fard, Soudeh, Vahid Niazi, and Mohammad Taheri. 2021. “Role of MiRNAs and LncRNAs in Hematopoietic Stem Cell Differentiation.” *Non-Coding RNA Research* 6 (1): 8–14. <https://doi.org/10.1016/j.ncrna.2020.12.002>.

Helsmoortel, Hetty H, Silvia Bresolin, Tim Lammens, Hélène Cavé, Peter Noellke, Aurélie Caye, Farzaneh Ghazavi, et al. 2016. “LIN28B Overexpression Defines a Novel Fetal-like Subgroup of Juvenile Myelomonocytic Leukemia.” *Blood* 127 (9): 1163–72. <https://doi.org/10.1182/blood-2015-09-667808>.

Hoon, M J L de, S Imoto, J Nolan, and S Miyano. 2004. “Open Source Clustering Software.” *Bioinformatics* 20 (9): 1453–54. <https://doi.org/10.1093/bioinformatics/bth078>.

Huang, Ying. 2018. “The Novel Regulatory Role of LncRNA-MiRNA-MRNA Axis in Cardiovascular Diseases.” *Journal of Cellular and Molecular Medicine* 22 (12): 5768–75. <https://doi.org/10.1111/jcmm.13866>.

Ivaldi, Maria Soledad, Luis Francisco Diaz, Lyubomira Chakalova, Jongjoo Lee, Ivan Krivega, and Ann Dean. 2018. “Fetal G-Globin Genes Are Regulated by the BGLT3 Long Noncoding RNA Locus.” *Blood* 132 (18): 1963–73. <https://doi.org/10.1182/blood-2018-07->

862003.

Jia, Wenguang, Siyuan Jia, Ping Chen, and Yunyan He. 2019. “Construction and Analysis of a Long Non-Coding RNA (LncRNA)-Associated CeRNA Network in β -Thalassemia and Hereditary Persistence of Fetal Hemoglobin.” *Medical Science Monitor* 25: 7079–86. <https://doi.org/10.12659/MSM.915946>.

Kanehisa, M, and S Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30. <https://doi.org/10.1093/nar/28.1.27>.

Karagkouni, Dimitra, Maria D. Paraskevopoulou, Spyros Tastsoglou, Giorgos Skoufos, Anna Karavangeli, Vasilis Pierros, Elissavet Zacharopoulou, and Artemis G. Hatzigeorgiou. 2020. “DIANA-LncBase v3: Indexing Experimentally Supported MiRNA Targets on Non-Coding Transcripts.” *Nucleic Acids Research* 48 (D1): D101–10. <https://doi.org/10.1093/nar/gkz1036>.

Kazemzadeh, Mina, Reza Safaralizadeh, and Ayla Valinezhad Orang. 2015. “LncRNAs: Emerging Players in Gene Regulation and Disease Pathogenesis.” *Journal of Genetics* 94 (4): 771–84. <https://doi.org/10.1007/s12041-015-0561-6>.

Lai, Ketong, Siyuan Jia, Shanjuan Yu, Jianming Luo, and Yunyan He. 2017. “Genome-Wide Analysis of Aberrantly Expressed LncRNAs and MiRNAs with Associated Co-Expression and CeRNA Networks in β -Thalassemia and Hereditary Persistence of Fetal Hemoglobin.” *Oncotarget* 8 (30): 49931–43. <https://doi.org/10.18632/oncotarget.18263>.

Lazarus, Kyren A, Fazal Hadi, Elisabetta Zambon, Karsten Bach, Maria-Francesca Santolla, Julie K Watson, Lucia L Correia, et al. 2018. “BCL11A Interacts with SOX2 to Control the

Expression of Epigenetic Regulators in Lung Squamous Carcinoma.” *Nature*

Communications 9 (1): 3327. <https://doi.org/10.1038/s41467-018-05790-5>.

Liu, Jinhua, Yapu Li, Jingyuan Tong, Jie Gao, Qing Guo, Lingling Zhang, Bingrui Wang, et al.

2018. “Long Non-Coding RNA-Dependent Mechanism to Regulate Heme Biosynthesis and Erythrocyte Development.” *Nature Communications* 9 (1): 4386.

<https://doi.org/10.1038/s41467-018-06883-x>.

López-Urrutia, Eduardo, Lilia P Bustamante Montes, Diego de Guevara Cervantes, Carlos Pérez-

Plasencia, and Alma D Campos-Parra. 2019. “Crosstalk Between Long Non-Coding RNAs, Micro-RNAs and MRNAs: Deciphering Molecular Mechanisms of Master Regulators in Cancer.” *Frontiers in Oncology* 9: 669. <https://doi.org/10.3389/fonc.2019.00669>.

Lv, Shao-Yan, Ti-Dong Shan, Xin-Ting Pan, Zi-Bin Tian, Xi-Shuang Liu, Fu-Guo Liu, Xue-Guo

Sun, et al. 2018. “The LncRNA ZEB1-AS1 Sponges MiR-181a-5p to Promote Colorectal Cancer Cell Proliferation by Regulating Wnt/ β -Catenin Signaling.” *Cell Cycle*

(*Georgetown, Tex.*) 17 (10): 1245—1254. <https://doi.org/10.1080/15384101.2018.1471317>.

Ma, Kelong, Xiaorong Pan, Pingsheng Fan, Yinghua He, Jun Gu, Wei Wang, Tengyue Zhang,

Zonghai Li, and Xiaoying Luo. 2014. “Loss of MiR-638 in Vitro Promotes Cell Invasion and a Mesenchymal-like Transition by Influencing SOX2 Expression in Colorectal

Carcinoma Cells.” *Molecular Cancer* 13 (1): 1–13. <https://doi.org/10.1186/1476-4598-13-118>.

Ma, Shuang-Ping, Hai-Rui Xi, Xu-Xia Gao, Jing-Min Yang, Ryo Kurita, Yukio Nakamura,

Xian-Min Song, Hong-Yan Chen, and Da-Ru Lu. 2021. “Long Noncoding RNA HBBP1

Enhances γ -Globin Expression through the ETS Transcription Factor ELK1.” *Biochemical*

and *Biophysical Research Communications* 552: 157–63.

<https://doi.org/https://doi.org/10.1016/j.bbrc.2021.03.051>.

Mashima, Ryuichi, and Torayuki Okuyama. 2015. “The Role of Lipoxygenases in Pathophysiology; New Insights and Future Perspectives.” *Redox Biology* 6: 297–310.

<https://doi.org/https://doi.org/10.1016/j.redox.2015.08.006>.

Mnika, Khuthala, Gaston K Mazandu, Mario Jonas, Gift D Pule, Emile R Chimusa, Neil A Hanchard, and Ambroise Wonkam. 2019. “Hydroxyurea-Induced MiRNA Expression in Sickle Cell Disease Patients in Africa.” *Frontiers in Genetics* 10: 509.

<https://doi.org/10.3389/fgene.2019.00509>.

Morrison, Tasha A., Ibifiri Wilcox, Hong Yuan Luo, John J. Farrell, Ryo Kurita, Yukio Nakamura, George J. Murphy, Shuaiying Cui, Martin H. Steinberg, and David H.K. Chui. 2018. “A Long Noncoding RNA from the HBS1L-MYB Intergenic Region on Chr6q23 Regulates Human Fetal Hemoglobin Expression.” *Blood Cells, Molecules, and Diseases* 69 (November 2017): 1–9. <https://doi.org/10.1016/j.bcmed.2017.11.003>.

Musharraf, Syed Ghulam, Ayesha Iqbal, Saqib Hussain Ansari, and Sadia Par. 2017. “ β - Thalassemia Patients Revealed a Significant Change of Untargeted Metabolites in Comparison to Healthy Individuals,” no. February: 1–10.

<https://doi.org/10.1038/srep42249>.

Nobili, Lucia, Marta Lionetti, and Antonino Neri. 2016. “Long Non-Coding RNAs in Normal and Malignant Hematopoiesis.” *Oncotarget* 7 (31): 50666–81.

<https://doi.org/10.18632/oncotarget.9308>.

Paraskevopoulou, Maria D, and Artemis G Hatzigeorgiou. 2016. “Analyzing MiRNA-LncRNA Interactions.” *Methods in Molecular Biology (Clifton, N.J.)* 1402: 271–86.

https://doi.org/10.1007/978-1-4939-3378-5_21.

Ru, Yuanbin, Katerina J Kechris, Boris Tabakoff, Paula Hoffman, Richard A Radcliffe, Russell Bowler, Spencer Mahaffey, et al. 2014. “The MultiMiR R Package and Database:

Integration of MicroRNA-Target Interactions along with Their Disease and Drug Associations.” *Nucleic Acids Research* 42 (17): e133. <https://doi.org/10.1093/nar/gku631>.

Sahlabadi, Amirhossein, Ravie Chandren Muniyandi, Mahdi Sahlabadi, and Hossein

Golshanbafghy. 2018. “Framework for Parallel Preprocessing of Microarray Data Using Hadoop.” Edited by Florentino Fdez-Riverola. *Advances in Bioinformatics* 2018: 9391635.

<https://doi.org/10.1155/2018/9391635>.

Sankaran, Vijay G, Tobias F Menne, Jian Xu, Thomas E Akie, Guillaume Lettre, Ben Van Handel, Hanna K A Mikkola, Joel N Hirschhorn, Alan B Cantor, and Stuart H Orkin. 2008.

“Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor BCL11A.” *Science (New York, N.Y.)* 322 (5909): 1839–42.

<https://doi.org/10.1126/science.1165409>.

Sankaran, Vijay G, and Stuart H Orkin. 2013. “The Switch from Fetal to Adult Hemoglobin.”

Cold Spring Harbor Perspectives in Medicine 3 (1): a011643.

<https://doi.org/10.1101/cshperspect.a011643>.

Sankaran, Vijay G, Jian Xu, and Stuart H Orkin. 2010. “Advances in the Understanding of Haemoglobin Switching.” *British Journal of Haematology* 149 (2): 181–94.

<https://doi.org/10.1111/j.1365-2141.2010.08105.x>.

- Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504. <https://doi.org/10.1101/gr.1239303>.
- Sumera, Afshan, Ammu K. Radhakrishnan, Abdul Aziz Baba, and Elizabeth George. 2020. “A Link between Long Non-Coding RNA (LncRNA) and Thalassaemia: A Review.” *Malaysian Journal of Pathology* 42 (3): 323–32.
- Tanno, Toshihiko, Natarajan V Bhanu, Patricia A Oneal, Sung-Ho Goh, Pamela Staker, Y Terry Lee, John W Moroney, et al. 2007. “High Levels of GDF15 in Thalassemia Suppress Expression of the Iron Regulatory Protein Heparin.” *Nature Medicine* 13 (9): 1096–1101. <https://doi.org/10.1038/nm1629>.
- Tao, Shuang, Yao-Bang Liu, Zhi-Wei Zhou, Bin Lian, Hong Li, Jin-Ping Li, and Shu-Feng Zhou. 2016. “MiR-3646 Promotes Cell Proliferation, Migration, and Invasion via Regulating G2/M Transition in Human Breast Cancer Cells.” *American Journal of Translational Research* 8 (4): 1659–77.
- Uda, Manuela, Renzo Galanello, Serena Sanna, Guillaume Lettre, Vijay G Sankaran, Weimin Chen, Gianluca Usala, et al. 2008. “Genome-Wide Association Study Shows BCL11A Associated with Persistent Fetal Hemoglobin and Amelioration of the Phenotype of α -Thalassemia” 105 (5).
- Vento, Sandro, Francesca Cainelli, and Francesco Cesario. 2006. “Infections and Thalassaemia.” *Lancet Infectious Diseases* 6 (4): 226–33. [https://doi.org/10.1016/S1473-3099\(06\)70437-6](https://doi.org/10.1016/S1473-3099(06)70437-6).

- Wienert, Beeke, Alister P W Funnell, Laura J Norton, Richard C M Pearson, Lorna E Wilkinson-White, Krystal Lester, Jim Vadolas, et al. 2015. “Editing the Genome to Introduce a Beneficial Naturally Occurring Mutation Associated with Increased Fetal Globin.” *Nature Communications* 6 (1): 7085. <https://doi.org/10.1038/ncomms8085>.
- Zhang, Xiaoqin, Stella Sun, Jenny Kan Suen Pu, Anderson Chun On Tsang, Derek Lee, Venus On Ying Man, Wai Man Lui, Stanley Thian Sze Wong, and Gilberto Ka Kit Leung. 2012. “Long Non-Coding RNA Expression Profiles Predict Clinical Phenotypes in Glioma.” *Neurobiology of Disease* 48 (1): 1–8. <https://doi.org/10.1016/j.nbd.2012.06.004>.
- Zhang, Xueqing, Sherman M. Weissman, and Peter E. Newburger. 2014. “Long Intergenic Non-Coding RNA HOTAIRM1 Regulates Cell Cycle Progression during Myeloid Maturation in NB4 Human Promyelocytic Leukemia Cells.” *RNA Biology* 11 (6). <https://doi.org/10.4161/rna.28828>.
- Zheng, Suilan, Nathan A. Krump, Mary M. McKenna, Yen Hsing Li, Anke Hannemann, Lisa J. Garrett, John S. Gibson, David M. Bodine, and Philip S. Low. 2019. “Regulation of Erythrocyte Na/K/2Cl Cotransport by an Oxygen-Switched Kinase Cascade.” *Journal of Biological Chemistry* 294 (7): 2519–28. <https://doi.org/10.1074/jbc.RA118.006393>.

Tables

Table 1. Performance matrix of different ML classifiers

ML Classifiers	Accuracy	Precision	Recall	F1 score	AUC
Random Forest	0.8421	0.8285	0.7948	0.808	0.8461
SVM(Linear)	0.7368	0.7023	0.7179	0.7076	0.7564
SVM(RBF)	0.8947	0.9333	0.8333	0.8642	0.8076

Figure Legends

- Figure 1.** Flowchart depicting methodological workflow of predicted lncRNA- miRNA crosstalk in HbF regulation.
- Figure 2.** Differential expression of lncRNA probe sets between high HbF and normal condition. A total of 46 differentially expressed lncRNAs were identified to be significant in SAM analysis. **A.** The hierarchical clustering of lncRNA probe sets in training dataset, GSE13284. **B.** Validation of lncRNA probe sets in the discovery or test dataset, GSE7874. T-statistics with FDR<25% and Permutation of 1000 were performed to determine significant genes. Red= high expression; Green=low expression. Bar colors indicate sample type: yellow, high HbF; blue, normal HbF.
- Figure 3.** Showing features selection procedure to identify optimal set of lncRNA from differentially expressed lncRNAs. Stepwise selection method, involving forward selection was used.
- Figure 4.** ROC curves of lncRNA-based classifiers in the discovery cohort.

- **Figure 5.** The Hierarchical clustering of differentially expressed mRNAs between high HbF and control group in GSE13284. Expression heatmap of differentially expressed genes was generated using R v3.6.3. “Red” and “Blue” colors represent above and below expression, respectively, relative expression.
- **Figure 6.** Gene Ontology (GO) annotation of differentially expressed correlated mRNAs, with a list top enriched genes ($p < 0.05$), covering the domain of biological processes, cellular components and molecular functions. Enrichment values were $-\log_{10}$ (p-value) transformed.
- **Figure 7.** KEGG pathway analysis of common identified mRNAs. The 6 significant ($p < 0.05$) signaling pathways are shown. All values are negative log transformed.
- **Figure 8.** The figure is showing predicted miRNA-mRNA network. The network contains 19 nodes and 17 edges. The elliptical and diamond shapes represent mRNA and miRNA, respectively. Red and green colors indicate up-regulation and down-regulation, respectively. The directionality of the arrow is from the source to the target.
- **Figure 9.** The figure is showing predicted lncRNA-miRNA network. The network consists of 6 nodes and 5 edges. The V-shapes and diamond represent lncRNA and miRNA, respectively. The red and green colors indicate up-regulation and down-regulation. Each arrow represents the interaction between lncRNA and miRNA, and the arrow is pointed from the source to the target.
- **Figure 10.** The interactions among lncRNA, miRNA and mRNA in high HbF vs. normal HbF condition. **A.** All possible networks among lncRNA, miRNA and mRNA in the studied groups. **B.** Predicted lncRNA-miRNA-mRNA-signaling pathway axis, which is important in HbF regulation. The V, Diamonds and elliptical shapes indicate lncRNA, miRNA and mRNA,

respectively. Red= Up-regulation; Green= Down-regulation. The arrows show interactions and pointed from the source to the target.