ISSN:2394-2371
CODEN (USA):IJPTIL

RESEARCH PAPER

Functional analysis of variants associated with thalassemia using in silico tools

Aastha Gaur, Sudarshan Singh Lakhawat, Sunil Kumar*
[1]Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur 303002, Rajasthan, India

*Corresponding Author: **Dr. Sunil Kumar**

ABSTRACT

Genetic Diseases are one of the leading causes of mortality around the world. Mutations in the genetic composition can alter the sequences and end up being the cause of disorders like Thalassemia, Cancer, Down 's syndrome, etc. The diagnosis of such disorders use conventional methods, which are time consuming and often skip some results. In our study, we have used in-silico tools based on next-generation sequencing for the functional analysis of variants involved in Thalassemia for rapid results. The conventional methods of diagnosis have been in use since a long time, but they are time consuming and not too accurate. Next-generation sequencing is a healthier and better way for diagnosing and analysing different disorders (both genetic and non-genetic). Not only is it faster, but it can also process a huge amount of data in very less time. The accuracy shown by next-gen sequencing is commendable, and this technique is more reliable than the conventional methods of diagnosis. The future of this technique is definitely brighter and promising in the health-sector.

Keywords: - *In-silico tools, thalassemia, analysis, next-gen sequencing, prenatal diagnosis, clinical management.*

INTRODUCTION

Clearly, understanding genetics and the genome as a whole and its variation in the human population, are integral to understanding disease processes and this understanding provides the foundation for curative therapies, beneficial treatments and preventative measures [1]. Alpha-thalassemia is caused most frequently by deletions involving one or both alpha globin genes. The most common deletions remove a single alpha globin gene, resulting in the mild alpha^{+}-thalassemia phenotype (- alpha/alpha alpha). Reciprocal recombination between highly homologous regions called (Z boxes) results in a chromosome with a 3.7-kb deletion containing only one alpha gene (-alpha$^{3.7}$), whereas recombination between mis – paired homologous X boxes produces a 4.2-kb deletion (-alpha$^{4.2}$). These recombinational events also result in the production of chromosomes containing three alpha globin genes. The -alpha$^{3.7}$ and -alpha$^{4.2}$ deletions are the most common alpha^{+} alpha-

***CORRESPONDING AUTHOR**
Dr Sunil Kumar
Assistant Professor, Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur, Rajasthan, 303002, India
E.Mail: skumar6@jpr.amity.edu
Article Published: April – June 2024
CITE THIS ARTICLE AS
Gaur A., et al. Functional analysis of variants associated with thalassemia using in silico tools. Int. J. Pharm. Technol. Biotechnol. 2024; 11(2):45-56.

thalassemia defects. Other rare deletions totally or partially remove one of the two alpha globin genes. Extended deletions, varying from 100 to >250 kb, removing all or part of the cluster including both alpha globin genes and sometimes the embryonic zeta2 gene, result in the complete absence of alpha chain synthesis (alpha°-thalassemia). Such deletions are the result of several molecular mechanisms including illegitimate recombination, reciprocal translocation, and truncation of chromosome 16. More than 40 different alpha°-thalassemia deletions have been described, the most common being the Southeast Asian, Filipino, and Mediterranean types. Two deletions [-(alpha)$^{5.2}$ and -(alpha)$^{20\text{-}5}$] removing the alpha-2 and partially the alpha1 globin gene also result in alpha°-thalassemia. Rare large deletions extending from 100 to >200 kb and removing the entire alpha globin cluster, and other genes that flank the cluster, including a DNA repair enzyme (methyladenine DNA glycosylase), and inhibitor of GDP dissociation from Rho (Rho GDI γ), a protein disulfide isomerase (PDI-R) and other anonymous housekeeping genes, have been reported in single families. Despite the removal of several genes, such patients seem to have a normal phenotype apart from having alpha-thalassemia. A deletion removing the alpha1 gene, the theta gene, and extending downstream centromeric from the alpha cluster results in alpha°-thalassemia. The silencing of intact alpha2 gene is related to an antisense RNA transcribed from the widely expressed *LUC7L* gene, becoming juxtaposed to the normal alpha2 gene by the deletion and running through the alpha2 gene sequences [2-3]. Several different deletions involving the MCS-R regulatory regions, but leaving both alpha genes intact, have also been reported, and all result in alpha°-thalassemia [2,4].

NONDELETION ALPHA-THALASSEMIA

Nondeletion defects less frequently cause alpha-thalassemia. These defects include single nucleotide substitutions or oligonucleotide deletions/insertions in regions critical for alpha globin gene expression. Several molecular mechanisms (abnormalities of RNA splicing and of initiation of mRNA translation, frameshift and nonsense mutations, in-frame deletions, and chain termination mutations) have been described, the majority occurring in the predominant alpha2 gene and producing alpha^{+}-thalassemia. The most common nondeletional variants are the T→C initiation codon mutation and the -5nt alpha -IVS1 deletion in Mediterranean, polyadenylation site mutations in Mediterranean and Middle East populations, stop codon mutations resulting in elongated alpha globin variants, including the T>C (stop→glu) of the alpha2 gene that results in Hb Constant Spring, and other elongated variants (Hb Icaria, Hb Seal Rock, and Hb Koya Dora) found in Mediterranean, middle East Asia, and Southeast Asia. Hb Constant Spring, the most common (up to 4%) nondeletion defect present in Southeast Asian population, is an alpha chain variant elongated by 31 amino acids, which is produced in a very low

amount (~1%). The instability of the mRNA, due to disruption of the untranslated region may be the reason for the reduced production of Hb Constant Spring. As for beta globin gene, mutations of alpha genes, which result in the production of hyperunstable globin variants, such as Hb Quong Sze, (alpha 109 Leu→Pro), Hb Heraklion (alpha 137 pro→0), and Hb Agrinio (alpha 29 Leu→Pro), unable to assemble in stable tetramers and thus rapidly degraded, might produce the phenotype of alpha-thalassemia. At present, about 30 alpha globin chain hyperunstable variants have been described [5].

MATERIALS AND METHODS

GeneCards

Studying Thalassemia is a very in-depth process, which demands precise knowledge of the affected genes in the body of the patient. To know more about the genes involved, and their basic information, **GeneCards** (Version 4.14), was used. It is an integrative database with wide variety of information on all annotated and human-predicted genes. This user friendly and comprehensive database gives a complete insight about the genomic, transcriptomic, proteomic, genetic, clinical and functional information.

Genes like HBA1, HBA2, HBB, ATRX, BCS1L, HBE1, HBD, BCL11A and SCN2A was identified as the main genes undergoing genetic changes, resulting in Thalassemia, and/or its side effects.

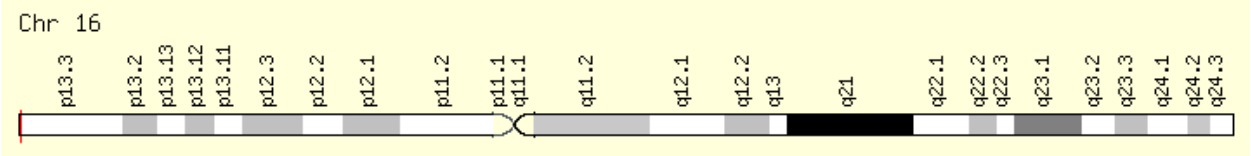

Figure 1: Genomic view for HBA1 gene

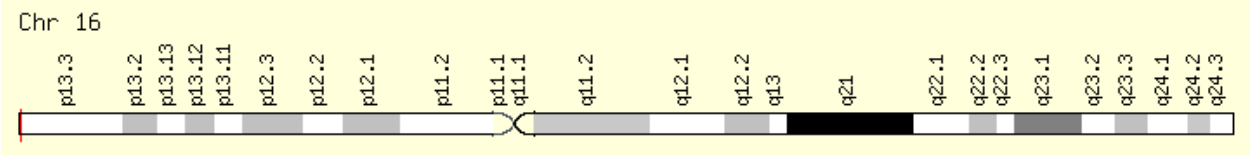

Figure 2: Genomic view of HBA2 gene

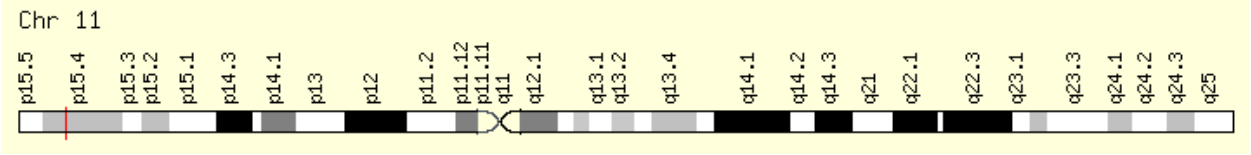

Figure 3: Genomic view for HBB gene

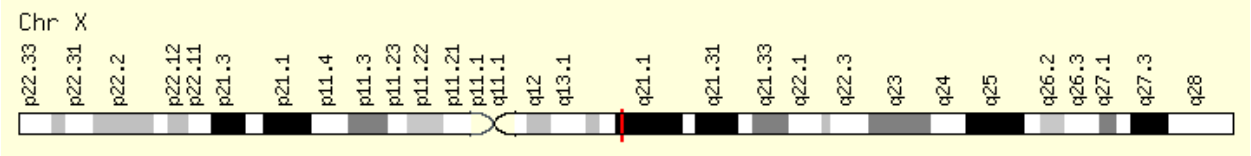

Figure 4: Genomic view for ATRX gene

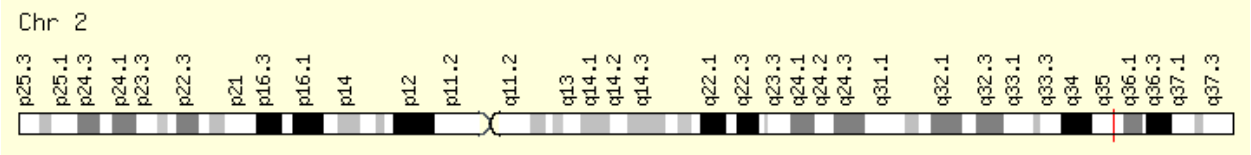

Figure 5: Genomic view for BCS1L gene

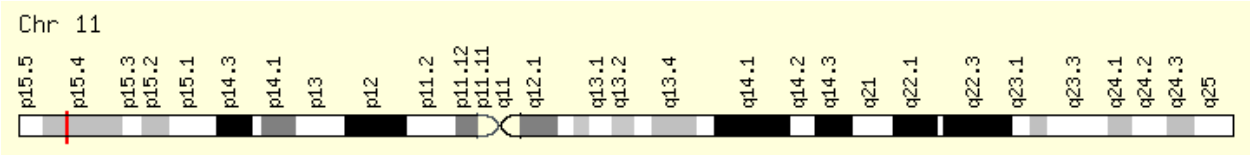

Figure 6: Genomic view for HBE1 gene

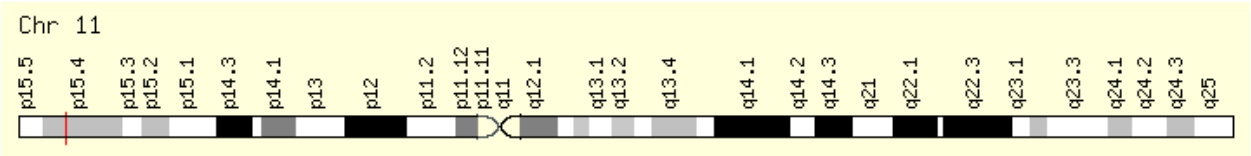

Figure 7: Genomic view for HBD gene

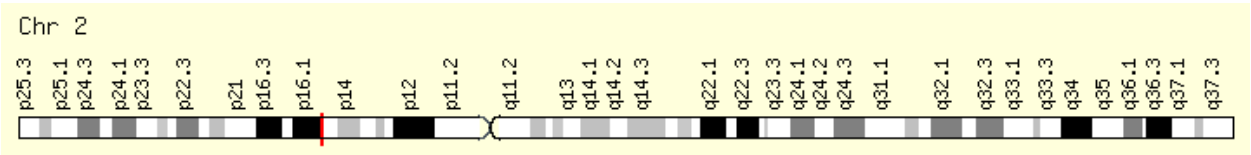

Figure 8: Genomic view for BCL11A gene

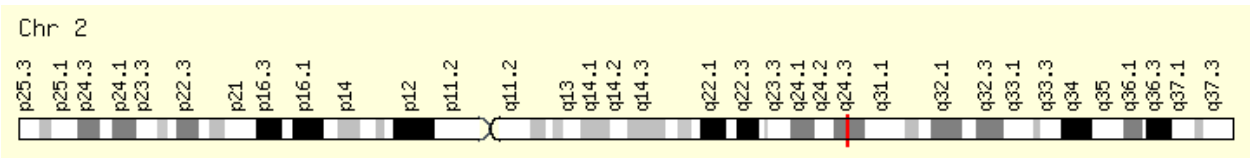

Figure 9: Genomic view for SCN2A gene

Ensembl Genome Browser

For the retrieval of ensembl transcript of gene sequence, **Ensembl Genome Browser** was used. This database gives access and information about gene sequence, splice variants and other related variants. Once downloaded, the transcript gives various information like Variant ID, alleles, chromosome

position/no. location, etc. The most important and useful information are the Variant ID's which are later used to estimate the linkage disequilibrium. Gene sequence transcripts of all the above-mentioned genes were downloaded and the variant ids were copied.

Single Nucleotide Polymorphisms Annotator (SNiPA)

Linkage disequilibrium is defined as the non-random association of alleles at different loci. Many genetic studies of disease association rely heavily on linkage disequilibrium (LD) patterns between pairs of markers to detect susceptibility markers. SNiPA (single nucleotide polymorphisms annotator) was used to estimate the LD. Pairwise LD was calculated, by feeding in the variant IDs of individual genes. LD r^2 was seen after setting the LD threshold to a lower value of around 0.6-0.8 for better results.

Reactome (Pathway Browser v3.7)

Pathway analysis is a very informative way to gain access to the intricacies of genes, their regulation and effect. It is done to know the pathways where the genes are involved, and what effect they put on the body. Expression or non-expression of gene can contribute to diseases and disorders within the body. For pathway analysis in the case of thalassemia, databases **REACTOME** (Pathway browser version 3.7) and **GeneCards** were used. The information given below depicts the pathways where the above-mentioned genes were involved:

GENE: HBA1

PATHWAY INVOLVED

- Erythrocyte take up carbon dioxide and release oxygen.
- Folate metabolism.
- Binding and uptake of ligands by scavenger receptors. (Heme from plasma)

GENE: HBA2

PATHWAYS INVOLVED

- Erythrocyte take up carbon dioxide and release oxygen.
- Folate metabolism.
- Binding and uptake of ligands by scavenger receptors. (Heme from plasma)

GENE: HBB

PATHWAYS INVOLVED

- Erythrocyte take up carbon dioxide and release oxygen.
- Folate metabolism.
- Binding and uptake of ligands by scavenger receptors (Heme from plasma).
- Innate immune system.
- Glucose/energy metabolism.

GENE: HBD

PATHWAYS INVOLVED

- Factors involved in megakaryocyte development and platelet production
- Response to elevated platelet cytosolic Ca+2.

GENE: ATRX

PATHWAYS INVOLVED

- Chromatin regulation/ Acetylation
- Pathways affected in adenoid cystic carcinoma.

GENE: BCS1L

PATHWAYS INVOLVED

- Metabolism of proteins.
- Mitochondrial protein import.

GENE: HBE1

PATHWAYS INVOLVED

- Factors involved in megakaryocyte development and platelet production
- Response to elevated platelet cytosolic Ca+2.

GENE: BCL11A

PATHWAYS INVOLVED

- Transcriptional regulator.

GENE: SCN2A

PATHWAYS INVOLVED

- Activation of cAMP-dependent PKA cAMP pathway.
- Axon guidance.
- Cardiac conduction.

Due to the involvement of the genes in the mentioned pathways of the human body, the changes/ mutation in these gene and their functions result in the following diseases along with their role in thalassemia:

RESULTS

Thalassemia is a blood disorder which involves the role of many genes. Since it is a genetic disease, it is passed on from the parents to the off springs depending on the pattern of inheritance. Traditional methods were used till now for the diagnosis and analysis of thalassemia, which involved inaccuracies and are time consuming. Since genes make up an integral part of thalassemia, the diagnosis and analysis of the genes involved along with their function, should be accurate. The critical difference between the conventional methods and next generation sequencing is the Sequence Volume. NGS shows rapid result with a huge amount of data, whereas Sanger's method is comparatively slow and can process very less data at a time.

In our study, we have thrown light on the in-silico aspect of analysis which proves to be a quick and accurate method to dig information on the genes involved. The online databases like GeneCards, Ensembl, SNiPA and Reactome provide rapid results involving a huge amount of data. Large sequences are processed in no time with accurate results. The main genes responsible for hemoglobinopathies (specifically Thalassemia), and the side effects are: HBA1, HBA2, HBB, ATRX, HBD, HBE1, BCS1L, BCL11A and SCN2A. These genes affect the erythrocyte formation, induce epileptic encephalopathies, anaemia, mental retardation, along with other mitochondria-related ailments, along with Thalassemia. These genes alter the pathways where they are involved, as a result of which, a side effect arises. Like, ATRX is involved in chromatin modelling, so if this gene gets mutated, it affects the chromatin model of the gene and hence gives rise to problems. Similarly, other genes like HBA1, HBA2, HBB, HBD are involved in the transportation of oxygen, through hemoglobin. All these genes are related to hemoglobin function directly or indirectly, and the alteration/ mutation in these genes' sequences is the major reason for the disease. Once the gene is altered, the pathway linked through it is also affected adversely and ultimately results into genetic disorder.

Genes HBA1, HBA2, HBB are involved in the oxygen transport and also, the binding of Heme. If there is mutation in these genes, the uptake and binding does not take place properly and the oxygen carrying capacity of cells is drastically reduced.

Similarly, the gene ATRX plays a major role in chromatic modelling. It is a helicase: an enzyme that catalyzes the unwinding of double-stranded nucleic acids; and is a member of a family of proteins involved in DNA recombination and repair, chromatin remodeling, chromosome segregation, and regulation of transcription. Mutation in ATRX restricts the efficient and proper replication of subset of genomic loci. This is turn affects the cell growth and structure, ultimately affecting the oxygen carrying capacity due to improper structure of cell. The ATRX syndrome, also called alpha-thalassemia X-linked mental retardation, is an inherited disorder that affects many parts of the body. This condition occurs almost exclusively in males. Most affected individuals have mild signs of alpha thalassemia, thus reducing the production of hemoglobin.

Since Thalassemia is a genetic disorder and mutated genes are present at birth also, genes like HBE1 plays a major role in the embryonic sac. If this gene is altered, a condition known as Hypotonia-Cystinuria Syndrome develops in the patient wherein the amino acid absorption is affected. Patient suffers from anomalies of the kidney and urinary tract. The diagnosis of this disorder is otherwise difficult as neurological signs are aspecific, but by using next-gen sequencing, we can test the alterations in this gene and the probability of developing this disease, in case. A better understanding of the complete clinical scenario associated with HCS can help clinicians suspect, diagnose and treat HCS earlier with a positive influence on both neurological and renal outcome.

BCS1L, on mutation gives rise to some disorders. These are autosomal recessive in nature, and are related to the mitochondrial activity and function disruption. BCS1L syndromes consist of: Gracile Syndrome, Complex III deficiency and Björnstad Syndrome. These syndromes lead to growth retardation, aminoaciduria, cholestasis, iron overload; muscele movement may also be hindered in children suffering from complex III deficiency, and hearing loss and brittle hair are often seen in patients suffering from Björnstad Syndrome. Children might also fall prey to early deaths due to these diseases.

The HBD gene if mutated, gives rise to an abnormal form of hemoglobin called Hemoglobin-D. It is a hemoglobin variant. It forms a significant percentage of the hemoglobinopathies. Mutations in the delta-globin gene are associated with Delta-thalassemia. It is mostly seen in the Punjab region, (of India) and is also sometimes known as HBD Punjab.

Similarly, mutations in the BCL11A gene also gives rise to intellectual developmental disorder with persistence of fetal hemoglobin and corpus callosum. characterized by intellectual disability of variable degree, microcephaly, distinctive but variable facial characteristics, behavior problems, and asymptomatic persistence of fetal hemoglobin. Growth delay, seizures, and autism spectrum disorder can also be seen in some affected individuals.

And lastly, the SCN gene that mediates the voltage-dependent sodium ion permeability of excitable membranes, can also be mutated during Thalassemia. Once mutated, it can result into infantile epileptic encephalopathy (IEE). This disorder is characterized by seizures beginning during infancy followed by developmental delay. In a nutshell. *SCN2A* is one of the most common causes of neurodevelopmental disease, because, *SCN2A* encodes an alpha subunit in a voltage-gated sodium channel and is pivotal for neuronal signaling. If this gene is mutated, the pathway is also affected which ultimately leads to neurodevelopmental disorders.

Along with the gene, the variants also undergo changes and come up with different side effects in the patients. Seizures, mental retardation and Sickle Cell Anemia are the most common side effects seen in patients with Thalassemia. All these side effects result through the alteration of variants' pathway in the body. Using online databases have enabled us to come up with this information quickly and with minimal inaccuracies. Traditional methods, that have been used till now also display results but can also miss out on a lot of changes in genes. Next-generation sequencing (in silico) on the other hand, minutely tests each chromosome number and analyzes any mutation taking place. This, hence, results in accurate outcomes. We have thus used in silico tools to scrutinize the gene variants and have come to know about the various side-diseases these result in, along with thalassemia. Pathways are altered, which are not treatable by diseases. The sole cure for this disease is either bone marrow transplant or gene therapy, both of which require minute inspection and expertise. Thalassemia has been one of the most silent yet fatal diseases in India, and parts of Southern Asia, African regions. This study resulted in quick output of information on the disease, and therefore, it can be said that using next generation sequencing can be more beneficial for the analysis of genetic variants and information in diseases like these for accurate results. Next generation sequencing technologies have already been widely investigated and are increasingly being applied to many research areas, including de novo sequencing of bacterial and viral genomes; searching for genetic variants by resequencing whole genome or targeted genome region; understanding the genetic mechanisms underlying human gene expression variation and characterizing the transcriptomes of cells, tissues and organisms by RNA-Seq; and genome-wide profiling of DNA-

binding proteins and epigenetic marks by ChIP-Seq. With the cost of genome sequencing falling sharply, more and more personal genomes for individuals will become available.

DISCUSSION

Genetics plays a major role in most of the diseases. Genetic variations, alongside the environment, including our lifestyles contribute to disease processes. Abnormalities of hemoglobin (Hb) synthesis are among the most common inherited disorders of man and can be quantitative (thalassemia syndrome) or qualitative (variant Hbs) [6]. Of these, thalassemia syndromes particularly beta thalassemia major and certain alpha thalassemia are serious and a major cause of morbidity. The frequency of β-thalassemia in India ranges from 3.5 to 15% in general population. Every year 10,000 children with thalassemia major are born in India, which constitute 10% of the total numbers in the world. India spends nearly Rs. 1,000 crore per annum in the treatment of thalassemia patients [7]. Majority of the centers in India use conventional methods for diagnosis of hemoglobinopathies, which includes clinical and family history, red cell indices, complete blood counts (CBC), HbA2, HbF estimation, sickling test, and Hb electrophoresis. Genetic disorders can range from single gene, chromosomal imbalance, cancer, epigenetics to complex disorders. By starting from the base of genetics, we might be able to figure out the pathway of diseases, their diagnosis and proper management and after care. Diagnosis till now had been possible with the help of conventional methods of testing. But, due to the advancement of technology, next-generation sequencing (NGS) has shown impressive results in the diagnosis of diseases and disorders [8].

It is a well-versed fact that the Next Generation Sequencing can map the entire genome within a very short period. The NGS technologies are promising to refine and advance scientific approaches across many fields, including molecular diagnostics. However, the dramatic increase in sequence throughput has come at a cost of much lower read accuracy and shorter sequence length compared with traditional Sanger sequencing [9]. Although these shortcomings can be partially mitigated through increasing sequence coverage and bioinformatics means, the realization of many promises for diagnostic applications is predicated on. progress in overcoming obstacles in handling massive datasets and in developing tools to check and assure sequence quality, conduct sequence alignment and assembly, and biologically interpret and draw inferences from the data. Keeping this is mind, the clinical aspect of diseases, involving the analysis of it and diagnosis can be carried out using this technique [10]. The positive aspect is the ease to cope with the complex diagnosis of genetically heterogeneous disorders and to identify novel disease genes. The technology of enrichment makes it possible to ease out the complexity of the variety of genes and test each one of them properly. This is being used to provide

insights into the genetics underlying Mendelian traits involved in myopathies and to set up cost-effective diagnostic tests. The challenge will soon become how to efficiently turn the large volumes of whole-genome sequence data into clinically useful information, including molecular diagnostics and treatment selection. The NGS technologies are markedly accelerating multiple research areas, making it feasible to conduct experiments that were previously not affordable or even technically possible. Novel fields and applications in biology, life sciences and medicine are becoming a reality. Being able to provide single-nucleotide resolution, as well as simplified sample preparation, has made NGS technologies promising in molecular diagnostics in which high sensitivity and specificity are required. The translation of NGS technologies into clinical diagnostics, like in the diagnosis of thalassemia, and other diseases, is also in the early stages of development and is mainly focusing on applications that require a modest amount of genomic sequence information, relative quantification and high-sensitivity detection [11,12]. It will probably take several years for NGS technologies to become a common clinical diagnostic tool. However, there is no question that the information obtained from NGS technologies will dramatically reshape our understanding of many human diseases and guide researchers to develop new clinical diagnostic tools and discover novel drugs that target the critical genes that cause diseases.

CONCLUSION

Ever since medical science has accelerated into the world of technology, it has seen many miracles which were considered impossible at first. Traditional/ conventional methods have been taken over by the next generation sequencing. This huge potential of the NGS applications makes likely that these will soon become the first approach in genetic diagnostic laboratories.

A targeted NGS can offer cost effective, safe and fairly rapid turnaround time, which can improve quality of care for patients with early onset myopathies and muscular dystrophies; in particular, hemoglobinopathies.

But with these advantages come some limitations too. The first one being the use of NGS approach in clinical diagnostic is the management of the amount of data generated. Indeed generation, analysis and storage of NGS data require sophisticated bioinformatics infrastructure.

A skilled bioinformatics staff is needed to manage and analyze NGS data, and so both computing infrastructure and manpower impact on costs of NGS applications in clinical diagnostics. Bioinformaticians are to be mandatory in the organization chart of clinical laboratories in the NGS era, where they must closely collaborate with clinicians and laboratory staff to optimize the panel testing and the NGS data analyses.

Along with hemoglobinopathies, other genetic disorders can also be diagnosed and analyzed with the help of next generation sequencing and the in-silico tools. This not only makes the process easier but also accurate and less time consuming. This project is an evidence of the kind of information in silico tools provides us with, and the depth that they touch in respect to various diseases. Thalassemia is one such disorder which involves alteration in the pathways of the human body, at genetic level. The genes involved in Thalassemia had been identified, along with the pathways they were associated with, and the linkage disequilibrium was also measured. All this was possible because of the databases used. This proves that Next Generation Sequencing does save time, money and efforts along with providing us with accurate results.

References

1. Wadia, M.R. Phanasgaokar, S.P., Nadkarni, A.H., Surve, R.R., Gorashankar, A.C., Colah, R.B., & Mohanty, D.(2002). Usefulness of automated chromatography for rapid fetal blood analysis for second trimester prenatal diagnosis of β-thalassemia. *Prenatal Diagnosis*, 22(2), 153-157.
2. Amid, A., Chen, S., Brien, W., Kirby-Allen, M., Odame, I.(2016). Optimizing chronic transfusion therapy for survivors of hemoglobin Barts hydrops fetalis. *Blood*.127:1208–11.
3. Coelho, A., Picanço, I., Seuanes, F., Seixas, M.T., Faustino, P.(2010). Novel large deletions in the human alpha-globin gene cluster: clarifying the HS-40 long-range regulatory role in the native chromosome environment. *Blood Cells, Molecules and Diseases*. 45, 147–153.
4. Origa, R., Moi, P., Pagon, R.A., Adam, M.P., Ardinger, H.H., Wallace, S.E., Amemiya, A., Bean, L.J.H., Bird, T.D., Ledbetter, N., Mefford, H.C., Smith, R.J.H., Stephens, K. (2016). Alpha-Thalassemia. GeneReviews.
5. Harteveld, C.L., Higgs, D.R. (2010) . Alpha-thalassaemia. Orphanet Journal of Rare Diseases. 28, 5–13.
6. Weatherall, D. J., & Clegg, J. B. (2001). The thalassemia syndromes (4th ed.). Blackwell Science.
7. Rund, D., & Rachmilewitz, E. (2005). Beta-thalassemia. The New England Journal of Medicine, 353(11), 1135-1146. https://doi.org/10.1056/NEJMra050436
8. Verma, I. C., & Saxena, R. (2005). India: Priorities for the management of thalassemia. Indian Journal of Pediatrics, 72(10), 865-867. https://doi.org/10.1007/BF02724006
9. Steinberg, M. H., & Forget, B. G. (2009). Hemoglobinopathies in humans: Molecular genetics and clinical management (2nd ed.). John Wiley & Sons.
10. Roberts, I. A., & Weatherall, D. J. (2008). The hereditary anaemias: A laboratory guide to diagnosis and management. Wiley-Blackwell.
11. Gahl, W. A., & Markello, T. C. (2003). Genomics and medicine: Genetics of rare diseases. The New England Journal of Medicine, 349(3), 2527-2538. https://doi.org/10.1056/NEJMra012418
12. Green, E. D., & Guyer, M. S. (2011). Charting a course for genomic medicine from base pairs to bedside. Nature, 470(7333), 204-213. https://doi.org/10.1038/nature09764.