

2008

Creating an internet-based database of beta thalassemia mutations

Shalu Susan George
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

George, Shalu Susan, "Creating an internet-based database of beta thalassemia mutations" (2008). *Master's Theses*. 3616.
DOI: <https://doi.org/10.31979/etd.u62g-re2k>
https://scholarworks.sjsu.edu/etd_theses/3616

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

CREATING AN INTERNET-BASED DATABASE
OF
BETA THALASSEMIA MUTATIONS

A Thesis

Presented to

The Faculty of the Department of Biological Sciences
San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Shalu Susan George

May 2008

UMI Number: 1458150

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 1458150

Copyright 2008 by ProQuest LLC.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

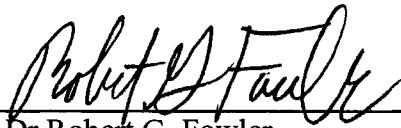
ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

© 2008

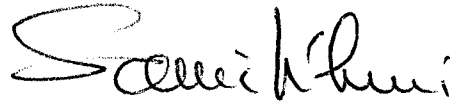
Shalu Susan George

ALL RIGHTS RESERVED

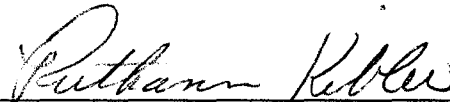
APPROVED FOR THE DEPARTMENT OF
BIOLOGICAL SCIENCES



Dr Robert G. Fowler



Dr Sami Khuri



Dr Ruthann Kibler

APPROVED FOR THE UNIVERSITY



ABSTRACT

CREATING AN INTERNET-BASED DATABASE OF BETA THALASSEMIA MUTATIONS

by Shalu Susan George

Beta thalassemia, one of the most common single-gene disorders world-wide, is caused by some 200 mutations in the beta globin gene. The disease, in its most severe form, can result in a chronic anemia which can only be treated with repeated blood transfusions. In order to inform and educate students, patients, the general public, and possibly the medical community, about this incurable but preventable disease, an internet-based database was created containing information regarding various mutations that result in beta thalassemia. MySQL was used to create the database back-end, while PHP, JavaScript, and Dreamweaver were used to make the front-end. The database currently contains 106 mutations, drawn from 139 literature references. The user-friendly database also provides the entire sequence of the beta globin gene, in a clickable-map format, as well as the ability to generate filtered mutation lists. Many future improvements have been planned for this database.

ACKNOWLEDGEMENTS

For this project, I would first and foremost like to thank my thesis committee member Dr. Khuri, for introducing me to this fascinating disease - beta thalassemia, for encouraging me to take up this project, for sharing my interest in the subject, and for permitting me to use his brilliant idea of creating a mutation map. Thanks are also due to Dr. Khuri for providing me with excellent guidance during this project, enabling me to complete right on schedule, and for giving me numerous opportunities to present my research to both my peers and the faculty.

A huge debt of thanks is owed to Dr. Fowler, my graduate advisor and thesis committee chair. Throughout this project he has been a constant source of support and encouragement. It was always very comforting to know that Dr. Fowler's invaluable advice was just an e-mail away. I have also been extremely fortunate to have him as my advisor at both the undergraduate and graduate level at San Jose State University (SJSU). I am grateful to Dr. Fowler for always being so readily available and for responding promptly whenever help or guidance was needed.

I would like to express my gratitude to my thesis committee member Dr. Kibler who took time off her busy schedule to go through my thesis.

My appreciation is extended to all the great instructors at SJSU who made my stay not only enlightening but also enjoyable.

I would also like to thank my peers who came up with some great suggestions and ideas, many of which were implemented in this project. Thanks are also due to all my classmates and friends who made my student career so much more memorable.

Last, but surely not the least, I would like to thank my husband, Naveen, for his belief in my capabilities and for his unconditional support and constant encouragement, my sister, Nisha, for standing by me, and my late parents, especially my mother, for instilling in me the importance of education, dedication, perseverance, and hard-work.

Table of Contents

List of Figures.....	ix
Chapter 1 Introduction	1
1.1 <i>The Thalassemias</i>	1
1.2 <i>Beta Thalassemia</i>	2
1.3 <i>Genetic Cause of Beta Thalassemia</i>	3
1.4 <i>The Link between Beta Thalassemia and Malaria</i>	3
1.5 <i>Distribution of Beta Thalassemia</i>	4
1.6 <i>Phenotype of Beta Thalassemia</i>	4
1.7 <i>Treatment for Beta Thalassemia</i>	6
1.8 <i>Diagnosis and Prevention of Beta Thalassemia</i>	7
1.9 <i>Purpose of the Database</i>	8
Chapter 2 Creation of the Database Back-end.....	12
2.1 <i>Data Sources</i>	12
2.2 <i>Data Collection</i>	15
2.3 <i>Software Used for Database Back-end</i>	15
2.4 <i>Designing the Database</i>	17
2.5 <i>Implementing the Database Design</i>	20
2.6 <i>Population of the Database</i>	24
Chapter 3 Creation of the Database Front-end	26
3.1 <i>Software Used for Front-end</i>	26
3.2 <i>Establishing a Connection between the Database and the Front-end</i>	30
3.3 <i>The Mutation Map</i>	31
3.4 <i>The Mutation Details Page</i>	36
3.5 <i>The New Mutation Input Form</i>	38
3.6 <i>The Mutation List</i>	42
Chapter 4 The Database Content	47
Chapter 5 Plans for Future Improvements	56

References	59
Appendix A Code for Automated Underlining of Hyperlinks	62
Appendix B List of Mutations in the Beta Thalassemia Database	64

List of Figures

<i>Figure 2.1.</i>	The Entity-Relationship Database Schema	18
<i>Figure 2.2.</i>	The Relational Database Schema	21
<i>Figure 3.1.</i>	Three-tier Architecture	30
<i>Figure 3.2.</i>	The Mutation Map	33
<i>Figure 3.3.</i>	The Structure of the Beta Globin Gene	34
<i>Figure 3.4.</i>	A Customized Tooltip	34
<i>Figure 3.5.</i>	The Mutation Details Page	37
<i>Figure 3.6.</i>	The New Mutation Input Form	38
<i>Figure 3.7.</i>	The Comments Text Area with a Character Counter	40
<i>Figure 3.8.</i>	Dynamically Generated Error Messages	42
<i>Figure 3.9.</i>	The Mutation List Query Page	43
<i>Figure 3.10.</i>	Options Available for Filtering Mutations by Ethnicity	44
<i>Figure 3.11.</i>	A Sample Mutation List	45

Chapter 1 Introduction

The thalassemias, including alpha and beta thalassemia, are the most common monogenic diseases in humans. Thalassemia was first recognized as a disorder by researchers in the United States and Italy, in 1925 (Weatherall, 2004). The name of the disease was coined by George Whipple and is derived from the Greek words meaning “sea” and “blood” due to the mistaken notion that the disease was restricted to individuals of Mediterranean origin (Weatherall, 2004). Over the years, however, it became apparent that the disease was equally, if not more prevalent, in many parts of the world.

1.1 *The Thalassemias*

The thalassemias are disorders of hemoglobin. Hemoglobin is the major protein found in red blood cells and is responsible for binding and carrying oxygen from the lungs to the various body tissues. Each hemoglobin molecule is a tetramer, made up of two alpha-like globin chains and two beta-like globin chains. Each globin chain has a heme group attached (Weatherall, 2004). Oxygen is bound and transported through the heme-groups.

A heme group is a prosthetic group, that is, it is a non-protein group bound to a protein. The heme group consists of an iron atom contained in the center of a large heterocyclic organic ring called a porphyrin (Casiday & Frey, 1998). The heme group is bound to a histidine residue in the globin protein. Both the heme group and the hemoglobin protein undergo conformational changes upon the binding and unbinding of

oxygen. These changes in shape improve the efficiency of the hemoglobin molecule, enabling it to bind or release more oxygen (Casiday & Frey, 1998)

Abnormalities in hemoglobin, resulting from the deficiency or complete absence of one or more of the globin chains, give rise to the thalassemias. Patients with thalassemia suffer from varying degrees of anemia, that is, a shortage of red blood cells (Weatherall & Clegg, 2001).

1.2 *Beta Thalassemia*

One major type of thalassemia is beta thalassemia, which results from a disorder in the production of beta globin. In the fetus, the primary beta-like globin is gamma-globin; however, soon after birth, a switch in globins takes place such that beta globin replaces gamma-globin (Weatherall, 2004). The major adult hemoglobin, called HbA, thus contains alpha globin and beta globin. The symptoms of beta thalassemia become apparent in a patient, only a few months following birth, when the switch to adult hemoglobin occurs (Weatherall, 2004).

Beta thalassemia poses the greatest burden on health care, compared to the other thalassemias, where the most severe forms of the disease often lead to death in-utero. Beta thalassemia symptoms develop within the first two years of birth and without treatment a patient, with the most severe form of the disease, would die by 20 years of age. The clinical description of beta thalassemia was first published in the US by two Detroit based physicians – Thomas B. Cooley and Pearl Lee. As a result, beta thalassemia is often also referred to as Cooley's anemia (Weatherall, 2004).

1.3 Genetic Cause of Beta Thalassemia

Beta thalassemia is caused by mutations in the beta globin gene. The gene is part of the beta globin locus, located on chromosome 11, in humans (Nienhuis, Anagnou, & Ley, 1984). Mutations in the gene result in either a deficiency of beta globin, referred to as beta⁺ thalassemia, or a complete lack of beta globin, referred to as beta⁰ thalassemia (Weatherall, 2004). The disease is an autosomal recessive disease, meaning that individuals that have the most severe form of the disorder carry two mutations – one in their paternal gene and one in their maternal gene. The disorder varies in severity. Individuals with beta thalassemia minor carry a mutation in only one of their two genes and have little to no symptoms. Those with beta thalassemia intermedia carry a mutation in either one or both of their genes and have moderate symptoms. Patients with beta thalassemia major have a mutation in both their genes and are the worst affected, requiring blood transfusions every 8-12 weeks or more frequently (Weatherall, 2004). Some 200 different mutations give rise to beta thalassemia. Most thalassemia major patients are compound heterozygotes and carry a different mutation in each gene (Weatherall, 2004).

1.4 The Link between Beta Thalassemia and Malaria

J. B. S. Haldane, in 1945, was the first one to suggest a connection between the high frequencies of thalassemia, seen in some parts of the world, and the prevalence of malaria in these regions. He believed that individuals that were heterozygous for globin mutations were somehow protected against a deadly case of malaria. This idea came to

be known as the Haldane hypothesis. The mechanism of this protection, in the case of thalassemia, is not clear (Kwiatkowski, 2005). Protection has been attributed to the inability of the parasite to invade or grow in the distorted red blood cells or the inability of the parasite to complete its lifecycle due to premature destruction of the mutant red blood cells (Ayi, Turrini, Piga, & Arese, 2004). Malaria, a deadly killer responsible for millions of deaths annually, has acted as one of the strongest evolutionary forces for the selection and maintenance of protective globin mutations in human populations (Kwiatkowski, 2005).

1.5 Distribution of Beta Thalassemia

Beta thalassemia has been prevalent in many parts of the Old World, especially in the so called malaria belt. The disease is prevalent, not only in the regions surrounding the Mediterranean, but also in Africa, the Middle East, the Indian subcontinent, Southeast Asia, and Southern China (Vichinsky, MacKlin, Waye, Lorey, & Olivieri, 2005; Weatherall, 2004). A few mutations dominate in each region while the rest of the mutations are not so commonly seen (Weatherall, 2004).

Beta thalassemia was brought to the New World by immigrants from the Mediterranean region, Africa, and, more recently, Asia (Vichinsky et al., 2005).

1.6 Phenotype of Beta Thalassemia

Beta thalassemia affects multiple organs but the primary phenotype of the disorder is hemolytic anemia (Cunningham, Macklin, Neufeld, & Cohen, 2004). Each red blood cell contains millions of hemoglobin molecules. Beta thalassemia patients, due

to their deficiency of beta globin, have abnormal hemoglobin, and, as a result, their red blood cells are microcytic, that is, they are smaller than normal and they have lower hemoglobin content. These mutated red blood cells are recognized as defective by the spleen, and are destroyed prematurely (Cunningham et al., 2004). The premature destruction of red blood cells leads to a severe case of anemia.

The spleen, in the process of destroying large numbers of red blood cells, becomes hyperactive and enlarged, a condition referred to as splenomegaly (Cunningham et al., 2004; Weatherall & Clegg, 1996). The iron, from the destroyed red blood cells, is normally stored in the liver and spleen, so that it can be used later on to make new cells. However, in patients, the rapid destruction of high numbers of red blood cells leads to an iron overload in the liver and spleen, as well as in many other organs, leading to organ failure. Patients often suffer from hepatitis or liver carcinoma. Iron overload in the ovaries, testes, and hypothalamus leads to infertility, in the pancreas it leads to diabetes, in the gall bladder it leads to gall stones, while in the heart it leads to irregular heart-beat and heart failure (Cunningham et al., 2004).

The hyperactive spleen not only destroys red blood cells, but also some white blood cells, which in turn causes immunosuppression (Cunningham et al., 2004). The bone marrow, which is the site of red blood cell production, also expands, to keep up with the large requirement of new cells, needed to replace the destroyed ones. This expansion of the bone marrow causes the bones to become fragile and results in an

enlargement of the skull and facial deformities in untreated children (Cunningham et al., 2004; Weatherall & Clegg, 1996).

1.7 Treatment for Beta Thalassemia

There are no cures for beta thalassemia and the disease can only be managed through frequent blood transfusions, iron chelation therapy, and splenectomy (surgical removal) in case of an enlarged spleen (Cunningham et al., 2004). Chronic cases are treated with monthly transfusions of “washed, filtered, or frozen red cells” (Marengo-Rowe, 2007). Most of the mortality associated with beta thalassemia results from iron-overload and hence all long-term transfusion programs are accompanied by treatment with an iron-chelating agent. The most widely used iron-chelator is deferoxamine (Cunningham et al., 2004; Marengo-Rowe, 2007). As it is poorly absorbed from the gastrointestinal tract, it has to be administered intravenously or subcutaneously, through a metering pump, 4-5 times a week, and for a period of 8 hrs (Marengo-Rowe, 2007).

Alternative therapies that are being considered include reactivation of fetal hemoglobin, bone marrow transplantation, and somatic gene therapy (Marengo-Rowe, 2007). The levels of fetal hemoglobin (HbF) generally drop soon after birth. However, it has been found that continuing to maintain high levels, using various therapeutic agents like erythropoietin, hydroxyurea, and cytarabine, can decrease the frequency of transfusions. HbF reactivation therapy only works for some patients (Marengo-Rowe, 2007).

Bone marrow transplantations from a human leukocyte antigen (HLA) matched donor have been performed for about a thousand patients in the US. This therapy works only for children and has many associated risks (Marengo-Rowe, 2007; Weatherall, 2004). Gene therapy to treat beta thalassemia has not been very successful as it is difficult to achieve the fine-tuned balance between alpha and globin chain production seen in normal cells (Marengo-Rowe, 2007).

1.8 Diagnosis and Prevention of Beta Thalassemia

As beta thalassemia requires life-long management, it puts a great deal of pressure on patients and the public health system. Prevention, through carrier screening and prenatal diagnosis, thus becomes vitally important. Primary screening involves various hematological tests, such as measurements of mean corpuscular volume (MCV), which is a measure of the size of a red blood cell, and mean corpuscular hemoglobin (MCH), which is a measure of the hemoglobin content of a red blood cell (Old, 2003). High Pressure Liquid Chromatography (HPLC) is used to look at the hemoglobin profile (Bhardwaj, Zhang, Lorey, McCabe, & McCabe, 2005; Lorey, 2000; Old, 2003). Patients with beta thalassemia have elevated levels of HbA2 (the secondary adult hemoglobin) and HbF (fetal hemoglobin) (Old, 2003).

If a patient is found to be a carrier, further testing is carried out to identify the mutation. Tests such as reverse dot blot or various PCR based methods are used to test for specific mutations (Bhardwaj et al., 2005; Old, 2003). 90% of the mutations can

usually be identified in this manner. For the remaining 10%, rare mutations, the entire beta globin gene is sequenced (Old, 2003).

When both parents are carriers, they have a 25% probability of having a thalassemia major child. In such cases, prenatal diagnosis is highly recommended. Amniocentesis or chorionic villus sampling (CVS) is commonly used for this purpose (Marengo-Rowe, 2007; Old, 2003). Amniocentesis involves using a needle to withdraw amniotic fluid, which contains the cells of the fetus. These cells can then be tested to determine the fetus's genotype. Alternatively, CVS can be carried out earlier in the pregnancy. CVS involves inserting a catheter through the cervix and withdrawing a piece of tissue from the outer layer of the embryo sac. This tissue has the same genotype as the fetus. If the fetus is found to have beta thalassemia major, the parents can opt to terminate the pregnancy.

In order to avoid the trauma of pregnancy termination, parents can also opt for an expensive procedure called pre-implantation diagnosis. In this procedure a number of embryos are prepared in-vitro from which a healthy embryo is then chosen and implanted (Marengo-Rowe, 2007).

1.9 Purpose of the Database

The beta thalassemia database, created as part of this project, will help increase awareness about the disease amongst the general public. The database could also be used as a teaching and learning tool in academic settings as well as in hospitals to educate patients and those in the allied medical professions.

The disease database could be utilized as an awareness tool as it contains information about the causative mutations and the affected populations. In the future, the database will include disease related information and links, as well as the basics of genetics and hematology. Thus, users will not only learn about the disease, but also will become aware of whether they belong to a high-risk community.

In the United States (US), thalassemia was previously considered as a disease only afflicting Mediterranean immigrants. Since the 1980's, however, the US has seen a sharp increase in immigration from various Southeast Asian countries where beta thalassemia is found at a higher frequency (Vichinsky et al., 2005; Lorey, 2000). As a result, beta thalassemia is rapidly emerging as "a new minority disease" in this country (Vichinsky et al., 2005). As beta thalassemia has no cure, management of the disease involves providing monthly blood transfusions to an ever growing number of patients; this in turn increases the pressure on blood banks and burdens the health care system. Under these circumstances, spreading awareness about the disease and promoting prevention, through carrier screening and prenatal diagnosis programs, becomes imperative. The database will contribute towards this effort in the US where knowledge of the disease has been found to be poor, even in highly susceptible communities (Armeli, Robbins, & Eunpu, 2005).

The only database for the hemoglobinopathies currently available to medical professionals, patients, researchers, and students alike, is called HbVar (described in detail in Chapter 2 Section 2.1), based at Penn State University (Patrinos, Giardine,

Riemer, Miller, & Chui, 2004). This database includes all globin mutations rather than just beta thalassemia mutations. The database is not very user-friendly, especially for non-professionals, and assumes considerable background knowledge of the disease and its genetics. Also, no web site or database on the internet, including HbVar, currently contains the entire sequence of the beta globin gene. The beta thalassemia database, created as part of this project, overcomes these deficiencies. The database not only provides the entire beta globin gene sequence, in an easy-to-follow format, but also is intuitive to use and hence can be easily utilized as a learning tool.

The database could also be employed as a teaching aid in courses of genetics, bioinformatics, and hematology. The database provides many tools to allow students to learn more about beta thalassemia, its distribution, and the mutations that give rise to it. The references provided for each mutation, with links to the article abstracts, give users an opportunity to further explore the disease and its many interesting variations. Since the beta globin gene is a fairly small gene, and is not alternatively spliced, it is ideal for studying genetic concepts like transcription, splicing, and translation. The size of the gene also makes it easier to understand gene structure. Since beta thalassemia involves some 200 mutations, including insertions, deletions, and substitutions, the database could educate students about the different types of mutations and their consequences. Students also have the option of learning the effect of a base insertion, deletion, or substitution by manipulating the beta globin sequence provided by the web site.

Finally, the database could be utilized in a hospital setting, to educate nurses and those in the allied medical fields. It could also be used by genetics counselors to help patients and their families better understand the disease.

The following chapters describe the process involved in the creation of the database as well as the database content. Chapter 2 deals with the design and implementation of the database back-end. Chapter 3 provides the details of the various web pages that make up the database front-end. Chapter 4 gives an overview of the records in the database. Finally, some ideas for future improvements of the database are presented in Chapter 5.

Chapter 2 Creation of the Database Back-end

The back-end is that portion of software that is not directly accessible to an end-user. It generally includes the Database Management System (DBMS) which is used to organize and store data.

The creation of the database back-end involved identifying the data sources, collecting mutation-related data from the sources, installing the software needed for database creation, creating and implementing the database design, and, finally, populating the created database with the gathered data.

2.1 Data Sources

The mutations for the database were gathered from three sources (a) HbVar, an online database of hemoglobin variants and thalassemia mutations (Wajcman, Patrinos, & Anagnou, 2008), (b) *The Thalassaemia Syndromes*, the book by Weatherall and Clegg (2001), and (c) PubMed, an online database of citations and abstracts of biomedical research articles (“NCBI: PubMed,” 2008).

The HbVar Database. The HbVar database is a “publicly available database” based at Penn State University, PA, USA (Patrinos et al., 2004) and was created through an association between Penn State University, INSERM Creteil, France and Boston University Medical Center, MA, USA (Wajcman et al., 2008). The initial data for HbVar was drawn from the books *A Syllabus of Human Hemoglobin Variants* (Huisman, Carver, & Efremov, 1998) and *A Syllabus of Thalassaemia Mutations* (Huisman, Carver, & Baysal,

1997). The database is maintained and updated with new mutations by three curators – Dr. Henri Wajcman, Dr. George Patrinos, and Dr. Nick Anagnou (2008).

The database claims to provide the most current information on different types of mutations that result in thalassemia (Wajcman et al., 2008). The information provided for a mutation includes sequence changes, the hematological and pathological effects of the mutation, the population-wise distribution of the mutation, and list of references (Patrinos et al., 2004). The database is primarily aimed at researchers and those in the medical community.

The HbVar database served as one of the sources for mutations and their references. The mutation description was taken from HbVar and each of the cited references was referred to for further information about the mutation. The information obtained from the references was then cross-checked with what had been recorded by HbVar.

“The thalassaemia syndromes” by Weatherall and Clegg (2001). Sir David J. Weatherall, the primary author of the book *The thalassaemia syndromes*, is a distinguished and multi-award-winning geneticist who has made outstanding contributions to the field of thalassemia research (Kan, 2004). He published his first paper on thalassemia as early as 1960 and spent nearly 40 years pioneering countless discoveries and authoring as many as 14 books and 600 papers on the subject (Kan, 2004). His above mentioned book, which is currently in its 4th edition, is unarguably

regarded as “the Bible” of thalassemia and is a “must read” for anyone in this line of research (Kan, 2004).

Dr. Weatherall’s book was used as an added data source. Additional mutations, not included in the HbVar database, were obtained from the book. For mutations common to both the book and HbVar, extra references, cited in the book but not in HbVar, were noted.

The PubMed Database. PubMed, a database maintained by the National Library of Medicine, is a collection of over 16 million abstracts and citations of articles from almost every well known biomedical journal (“NCBI: PubMed,” 2008). The article citations also include a link to the full-text article.

The PubMed database was used to access the reference articles for each mutation. In addition, the database was used to search for articles reporting any newly discovered mutations. A few new mutations were found in articles published in late 2007 and early 2008.

Some mutation entries in the HbVar database are based on personal communications, from researchers, received by the curators of HbVar. Some of these researchers, at a later date, publish a journal article based on the same findings. However, the HbVar entries are not updated with the journal citations. Hence, in cases where citations were absent, the PubMed database was searched for such a possible article by the researcher.

2.2 Data Collection

The gathered information for each mutation was initially recorded in a paper-based form. The mutation description and disease type (beta0, beta+) were noted from the data source. The mutation type (substitution, insertion or deletion) was derived from the mutation description. The starting position of the mutation, the mutation length, and the mutation location were deduced from the sequence of the beta globin gene. The effect of the mutation was either obtained from the references or was inferred on the basis of basic genetic principles. The disease severity (thalassemia major, intermedia or minor), the disease distribution (by nationality and ethnicity), and frequency of occurrence (common, rare, novel) were drawn from the references. Additional noteworthy facts, found in the references, were recorded as comments. The reference citations and PubMed identification numbers (PMIDs) were taken from the PubMed database.

2.3 Software Used for Database Back-end

Virtual Machine – Vmware. In order to create the back-end, a virtual machine was first setup using VMware. VMware is software that aids in the creation of an abstract machine within a machine (“VMware: Virtualization Basics,” 2008). It creates the effect of having two separate, independent computers within a physical computer. The virtual machine is referred to as the guest system, while the physical computer is referred to as the host system (“VMware: Virtualization Basics,” 2008).

A VMware virtual machine provides many advantages. It allows the use of multiple operating systems on the same computer (“VMware: Virtualization Basics,” 2008). It enables users to switch between operating systems without actually re-booting the computer. The virtual machine is also portable (“VMware: Virtualization Basics,” 2008). All the contents of the VMware virtual machine are stored in a single folder and hence can easily be moved or copied onto a different computer. The virtual machine can, in addition, acquire its own IP address and use it to communicate with the host system over a network (“VMware: Virtualization Basics,” 2008).

Since it was decided to create the database in a Linux operating system environment, a virtual machine was required to enable easy switching between the Windows XP and Linux operating systems.

Operating System – CentOS 5. CentOS is a “Community ENTERprise Operating System,” which means that it is a freely available, non-commercial software that is maintained and updated by a community of open source contributors (“CentOS: Home page”). CentOS is based on Red Hat Enterprise Linux which is a commercial product.

The advantages of CentOS are that it is a robust and reliable operating system. It also comes with the latest versions of various server and application software, including MySQL 5.0.

CentOS was chosen as the operating system because it was freely available and included the latest version of the database management system MySQL. CentOS also provided a stable development environment.

Database Management System – MySQL 5.0. MySQL, which stands for Structured Query Language, is a very popular, openly available, DBMS. MySQL is popularly used for web applications. It is used by many well known web sites such as YouTube, Flickr, Friendster, Wikipedia, Adobe, and Nokia. It can store large amounts of data, is flexible, and is fast at both storing and retrieving data (“Top reasons to use MySQL”).

2.4 Designing the Database

Database designing involves firstly, identifying the data to be stored in the database and secondly, choosing appropriate structures to represent and store the data. Database designing is the first step in creating a database and it is done before the actual database is created and populated with data.

The description of a database is called the database schema. The schema is defined during the designing phase and is not expected to change frequently. Different types of database schemas, based on different sets of rules, can be made. For this database, an Entity-Relationship (ER) schema was created (shown in Figure 2.1).

DISTRIBUTION is a weak entity that depends on the entity MUTATION or, in other words, a distribution does not make sense till it is tied to a mutation.

Ovals are used to represent the attributes of an entity. An attribute is a property that describes an entity. The attributes that are primary keys are underlined by a solid line (e.g., MutID, RefID). A primary key is an attribute that can be used to uniquely identify each instance of an entity. Weak entities do not have complete primary keys; instead they have partial keys. Partial keys are underlined by a broken line (e.g., Nationality, Ethnicity). The primary key of a weak entity is constructed by combining its partial key with the primary key of the related strong entity (also called the owner entity). In this case Nationality and Ethnicity need to be combined with MutID in order to construct a primary key for DISTRIBUTION. An attribute that can have multiple values is surrounded by a double oval.

Diamonds are used to represent the relationships between the entities. A relationship is the association between two or more entities. Double diamonds are used to depict the relationships of weak entities.

Lines connecting the diamonds to the boxes stand for the participation of an entity in a relationship. Partial participation is represented by a single line. A partial participation, for instance, is seen between the entity MUTATION and the relationship HAS, since a mutation may or may not have a user comment. Total participation, on the other hand, is represented by double lines. Total participation is seen between the entity

USER_COMM and the relationship HAS, as a user comment must have a mutation associated with it.

Furthermore, the type of relationship is given alongside the lines. A relationship could be one-to-one (1:1), one-to-many (1:N) or many-to-many (M:N). The entities MUTATION and REFERENCES have a many-to-many relationship because one mutation can have multiple references and, at the same time, one reference can have multiple mutations associated with it. On the other hand, entities MUTATION and DISTRIBUTION have a one-to-many relationship because one mutation can have multiple populations associated with it.

An ER schema is best suited for end users and helps them perceive data (Elmasri & Navathe, 2007). However, a schema that can be both understood by end users and, at the same time, can easily be implemented on a computer is the Relational schema. The ER schema can easily be mapped to a Relational schema.

2.5 Implementing the Database Design

A Relational schema was created to enable the implementation of the database design. An ER-to-Relational Mapping Algorithm from Elamsri and Navathe (2007, pp. 224-30) was used. The Relational schema is shown in Figure 2.2.

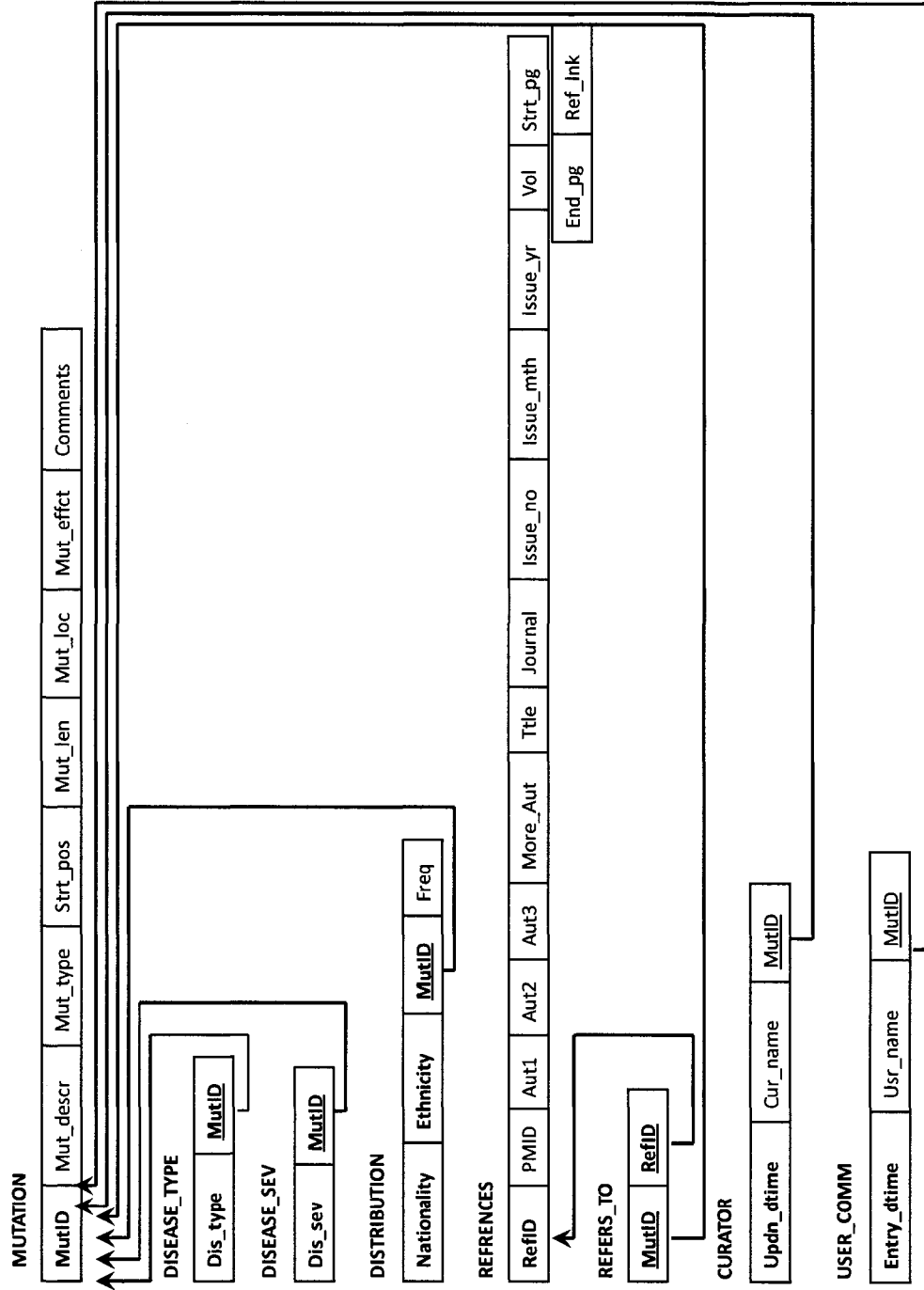


Figure 2.2. The Relational database schema.

The mapping steps used to create the schema are described below:

Step 1: Mapping of Regular Entity Types. For each regular (strong) entity type in the ER schema, a *relation* was created that included all the attributes of the entity (a relation is synonymous to a table of the database). A primary key was also chosen for each relation. Thus, the relations MUTATION, REFERENCES, CURATOR, and USER_COMM were created. The primary key for each of these relations is depicted in bold in Figure 2.3. The primary key for MUTATION is MutID (mutation ID), for REFERENCES is RefID (reference ID), for CURATOR is Updn_dtime (update date and time), and for USER_COMM is Entry_dtime (entry date and time).

Step 2: Mapping of Weak Entity Types. For each weak entity type in the ER schema, a relation was created that included all the attributes of the entity. In addition, the primary key of the owner entity was included as a foreign key attribute. Thus, the relation DISTRIBUTION was created and MutID, from the relation MUTATION, was included as a foreign key. The foreign key, MutID, is depicted as underlined in Figure 2.3. The source of the foreign key is indicated with an arrow. The primary key of DISTRIBUTION is a combination of the partial keys Nationality, Ethnicity, and the primary key of its owner MutID.

Step 3: Mapping of Binary 1:1 Relationship Types. For each 1:1 relationship type in the ER schema, a corresponding relationship was established between the relations in the Relational schema. The primary key of one participant in a 1:1 relationship was included as a foreign key of the other participant. Thus, to map the 1:1 relationship between the

entities MUTATION and CURATOR, the primary key of the MUTATION relation was included as a foreign key in the CURATOR relation. The foreign key, MutID, is depicted as underlined in Figure 2.3. The source of the foreign key is indicated with an arrow.

Step 4: Mapping of Binary 1:N Relationship Types. For each 1:N relationship type in the ER schema, a corresponding relationship was established between the relations in the Relational schema. The primary key of the participant on the 1-*side* of the relationship was included as a foreign key of the participant on the N-*side*. Thus, to map the 1:N relationship between the entities MUTATION and USER_COMM, the primary key of the MUTATION relation was included as a foreign key in the USER_COMM relation. The foreign key, MutID, is depicted as underlined in Figure 2.3. The source of the foreign key is indicated with an arrow. The 1:N relationship between the entities MUTATION and DISTRIBUTION was mapped in a similar manner.

Step 5: Mapping of Binary M:N Relationship Types. For each M:N relationship type in the ER schema, a new relation was created to represent the relationship between the participating entities. The primary keys of the participating relations were included as attributes in the new relation. Thus, to map the M:N relationship between the entities MUTATION and REFERENCES, a new relation REFERS_TO was created. RefID from REFERENCES and MutID from MUTATION were included as attributes of REFERS_TO. The primary key of REFERS_TO is a combination of the foreign keys MutID and RefID. The primary-cum-foreign keys, MutID and RefID, are depicted as

bold and underlined in Figure 2.3. The source of the foreign keys is indicated with arrows.

Step 6: Mapping of Multi-valued Attributes. For each multi-valued attribute in the ER schema, a new relation was created. This relation included its own attribute plus the primary key of the entity of which the relation is an attribute. Thus, for the multi-valued attribute Dis_sev of MUTATION, a new relation DISEASE_SEV was created that contained the attribute Dis_sev and the primary key, MutID, of MUTATION. The primary key of DISEASE_SEV is a combination of Dis_sev and the foreign key MutID. The primary key is shown in bold and the foreign key is underlined in Figure 2.3. The source of the foreign key is indicated with arrows. The multi-valued attribute Dis_type was mapped in a similar manner.

Step 7: Mapping of N-ary Relationship Types. As there were no n-ary relationship types in the ER schema, this step was omitted.

2.6 Population of the Database

Manual Population. The first 70 records in the database were entered using batch files. A batch file is a text file containing a series of commands that can be executed in a sequence, one after the other, just as if they were typed in manually (Weddell, 2003). MySQL commands, to insert records into the database, were written into a text file and the text file was then executed. Each batch file inserted 10 records into the database. The database entries were verified against the paper-based records.

This method of data entry was used at the start because the front-end data entry form had not as yet been constructed. The use of batch files was also advantageous as it enabled the modification and fine-tuning of the database schema. Following changes to the schema, batch files allowed easy deletion and reentry of records.

Automated Population. Once the mutation data entry form was constructed, as part of the front-end, mutations were entered one at a time using the browser-based form. Thirty-six mutations were entered in this manner. This method of data entry was also helpful in testing and de-bugging the underlying front-end code. The data entry form will be discussed in greater detail in Chapter 3.

The next chapter describes the web pages that make up the front-end of the database, the software that was used to generate the pages, as well as the three tier architecture that was used to establish communication between the front-end and the database.

Chapter 3 Creation of the Database Front-end

The front-end is that portion of software or a web site that is directly accessible to the end-user and is used to interact with or access other inaccessible parts of the software or web site. In this case, the database forms the back-end, while a web site forms the front-end. The web site, through web pages, allows the end-user to view the contents of the database, to enter data into the database and to query (request specific data from) the database. What access and how much access an end-user has to the database are determined by the web page designer.

The creation of the database front-end involved installing the needed software, establishing a connection between the database and the front-end, and writing the scripts to generate the required web pages such as the mutation map, the mutation details page, the new mutation input form, and the mutation list query page.

3.1 *Software Used for Front-end*

XAMPP for Windows. The name XAMPP is an acronym for X (the X stands for the many operating systems it supports) Apache MySQL PHP Perl (“XAMPP,” 2008). XAMPP is an installer that helps in the easy installation of Apache, PHP, and MySQL on a computer (McFarland, 2007).

XAMPP was used as it is freely available. It is also very easy and fast to download and install (Seidler, 2007). It provides an Apache HTTP server and an interpreter for the PHP scripting language (McFarland, 2007), both of which were

required to generate dynamic pages. XAMPP also enables testing of front-end scripts without actually connecting to the internet (McFarland, 2007).

Web Server – Apache 2.2. A web server is a program that accepts HTTP (hyper text transfer protocol) requests for web pages from a client (Brookshear, 2007). The client sending the HTTP request is a web-browser such as Internet Explorer, Mozilla Firefox, Safari or Opera. The request could be sent when, for instance, a user types a web address (URL) into the address bar of the browser or when a hyperlink is clicked on or a “Submit” button is pressed. The server responds with a web page (HTML document) which is then displayed by the browser (Brookshear, 2007).

The content that is sent by the server is static if it comes from an existing file, such as a text file or HTML page (McFarland, 2007). However, the content is said to be dynamic if it comes from another program. Data that is pulled from a database by a program or script is considered dynamic (McFarland, 2007).

Apache was used as it is one of the most popular HTTP servers on the World Wide Web (Pfaffenberger, 2003), is free, provides support for the server-side scripting language of choice PHP, and can be locally installed to test code during development.

Server-side Scripting – PHP 5.1. Server-side scripting is used to generate dynamic HTML pages. A request sent to the web server, by the browser, is fulfilled by executing a script at the web server (Brookshear, 2007). Server-side scripting enables the user to access data stored in the database. It can also be used to validate data prior to its input into the database.

PHP is a widely used server-side scripting language (McFarland, 2007). PHP is a recursive acronym for PHP Hypertext Preprocessor (Pfaffenberger, 2003). PHP scripts can easily be embedded into any HTML document by simply surrounding the code with the delimiters `<? and ?>`.

PHP was used because it is freely available, is easy to learn, and works well with the Apache web server and the MySQL database management system.

Web Development Tool – Dreamweaver CS3. Adobe Dreamweaver is a web development tool. It is one of the most popular What-You-See-Is-What-You-Get (WYSIWG) editors (McFarland, 2007). A WYSIWG editor is especially helpful for those who have very little familiarity with web-development technologies. Dreamweaver automatically generates the HTML code required to create a web page. However, for those with some familiarity, it also provides a “code view” where the developer can access and modify the automatically generated code.

Dreamweaver is suitable for generating dynamic web pages as, in addition to automatically generating HTML code, it can generate PHP and JavaScript (used for client-side scripting, explained in greater detail further in this section). It also provides support for Cascading Style Sheets (CSS), which is a language that allows a developer to specify how a web page will be presented; it allows the definition of style elements like color, font-size, font-style, and layout and thus helps in the separation of a document’s content, given as HTML, from a document’s style (Pfaffenberger, 2003).

The new version of Dreamweaver includes a Spry framework for Ajax, developed by Adobe. The Spry framework helps in the development of applications that use Ajax (McFarland, 2007). Ajax, an acronym for Asynchronous JavaScript and XML, is a collection of technologies that help in the creation of interactive web pages (Eichorn, 2006). Dreamweaver's Spry framework allows the developer to create dynamic pages on the client-side; it allows the inclusion of animation effects, stylized menus and tabbed panels, and form validation.

Client-side Scripting – JavaScript. Client-side scripting refers to programs that are executed by the web browser as opposed to the web server. It allows the creation of dynamic HTML pages where the content of the page doesn't remain fixed but changes depending on the web site user's input (Brookshear, 2007).

JavaScript is one of the languages used to write client-side scripts. The code can either be embedded in an HTML document or it can be included from a separate file. JavaScript can be used to open popup windows with messages, to disable or enable the elements in an input form based on user input, to validate form input, and to detect user actions like clicking a mouse button or moving the cursor over a link or image. As JavaScript is run by the browser, it can respond faster to user actions and hence can give the web page an interactive feel (McFarland, 2007). However, JavaScript will only work if the user's browser supports this scripting language and has it enabled.

3.2 Establishing a Connection between the Database and the Front-end

In order to establish a connection between the database and the front-end, a private network was set up between the host machine and the guest virtual machine (both set up on the same computer). Dreamweaver was used to help connect to the database.

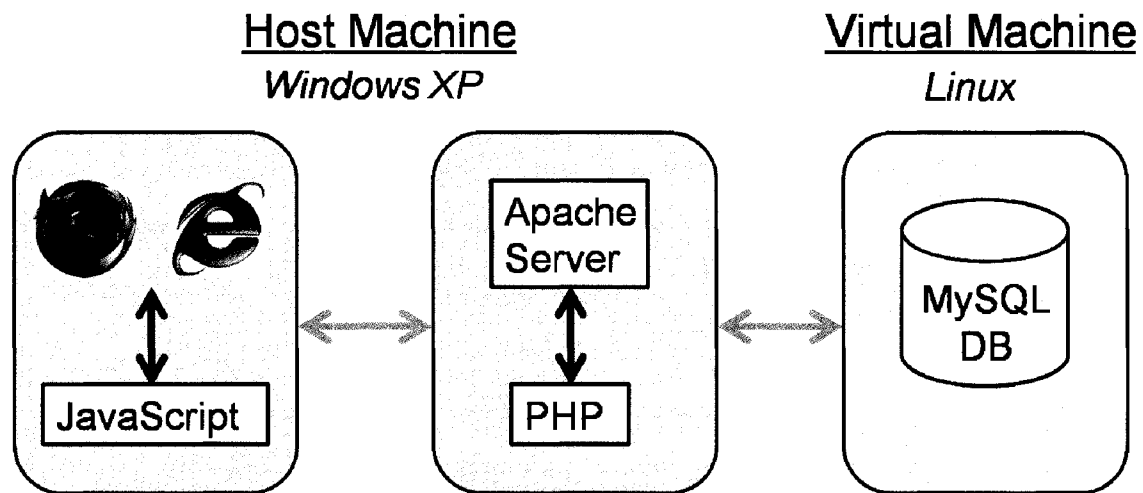


Figure 3.1. Three-tier architecture. The web browser forms the first tier, the application server forms the second tier, while the database server forms the third tier.

As described in Chapter 2, for development and testing purposes, the MySQL database was located on the virtual machine, while the Apache web server and PHP were located on the host machine. The guest operating system was Linux, while the host operating system was Windows XP. Mozilla Firefox and Internet Explorer were used for testing code. The PHP script is executed by the Apache web server, while JavaScript is executed by the web browser being used. The three-tier setup can be seen in Figure 3.1.

3.3 The Mutation Map

Obtaining the Sequence of the Beta Globin Gene. The sequence for the beta globin gene was obtained from the book by Weatherall and Clegg (2001). The BLAST (Basic Local Alignment and Search Tool) program, provided online by NCBI (National Center for Biotechnology Information), allows comparison of sequences, to check for sequence similarity. Nucleotide blast, also called blastn, was used to compare the sequence obtained from Weatherall and Clegg with all the human genomic sequences in the NCBI database. The sequence provided by the book and that stored by the NCBI database were found to be completely identical.

Design of the Mutation Map. The idea of a mutation map is credited to Dr. Sami Khuri, Professor (and thesis committee member for this project), Department of Computer Science, San Jose State University. The original mutation map was designed by Dr. Khuri and Keith Callenberg (2007), an undergraduate Computer Science major at San Jose State University.

The mutation map created for this project is shown in Figure 3.2. It extends from nucleotide -120 to +1706 of the beta globin gene. Nucleotides -120 to -1 are shown in lower case and represent the promoter of the gene. The important regulatory elements in the promoter, such as the two CACCC boxes, the CCAAT box, and the TATA box, are depicted in red. These regulatory elements are required for the optimal binding of transcription factors and the efficient initiation of transcription (Basran, Reiss, & Luo, 2008).

The transcribed portion of the beta globin gene is 1606 bases long and is shown in upper case letters. The gene has three exons and two introns. Nucleotides +1 to +50, shown in black, represent the 5'UTR (untranslated region), which forms part of exon 1. Nucleotide regions shown in blue represent the open reading frame of the gene. The three nucleotides forming each codon are grouped together. The open reading frame begins at the start codon (ATG) at position +51, in exon 1, and ends at the stop codon (TAA) at position +1472, in exon 3. The two introns, also called intervening sequences (IVS), are shown in black. Nucleotides +1475 to +1606, shown in black, represent the 3'UTR (untranslated region), which forms part of exon 3. The 3'UTR contains the polyadenylation signal sequence (AATAAA), depicted in red, that signals the 3' cleavage of the primary mRNA transcript. The 3'flanking region of the gene, extending from nucleotide +1607 to +1706, is shown in lower case. The overall structure of the beta globin gene can be seen in Figure 3.3.

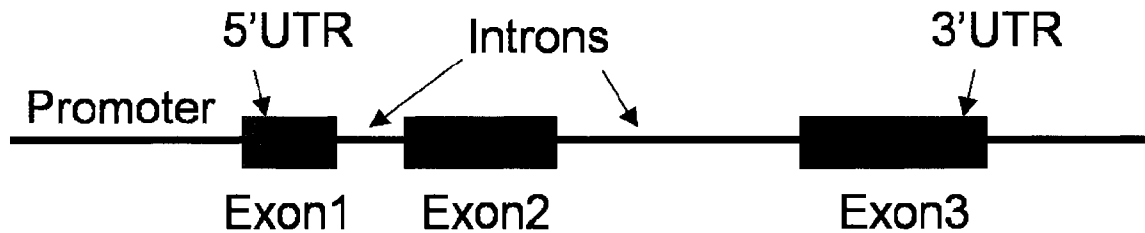


Figure 3.3 The structure of the beta globin gene. (UTR-Untranslated region)

As the cursor is moved over different bases, a hand tooltip pops up to give more information about the related map position. The tooltip can be seen in Figure 3.4. The tooltip includes information such as the position of the base, relative to the start point of transcription, and the structural location of the base. If the base is part of a codon, the codon number and the amino acid encoded by the codon are included in the tooltip. If the base is part of an intron, the position of the base, relative to the 5' exon-intron junction, is given.

```

AAG GTG AAG GCT CAT GGC AAG AAA GTG CTC
GGT GCC TTT AGT GAT GGC CTG GCT CAC CTG
GAC AAC CTC AAG GAT GGC AAG GCC ACA CTG
AGT GAG CTG CAC TGT GAC AAG CTG CAC GTG

```

A tooltip box is positioned over the 'A' in the 10th column of the 3rd line, containing the text: "401 Exon 2, Codon 73-Asp". An arrow points from the tooltip to the 'A' base.

Figure 3.4. A customized tooltip. On the map page the tooltip provides detailed information about every map position.

Base positions that have one or more mutations associated with them are underlined. These underlined bases are hyperlinks that can be clicked on to open a mutation details page (described in Section 3.4). The details page contains the

information about the associated mutations. Base positions that have no attached mutations are not underlined. Clicking on such bases produces a popup box that informs the user that the position has no mutations recorded.

When mutations, related to a new position, are added to the database, the corresponding base position is automatically underlined in the mutation map when the web page reloads.

Underlying Code for the Mutation Map. The background code that generates the mutation map contains a series of anchor tags, each of which references the file that contains the code for generating the mutation details page. The map position of each base is passed to the program file so that only mutation details related to that particular map position will be displayed.

A combination of PHP and JavaScript were used to produce both the automated underlining of links and the popup box, generated for map positions that have no associated records. The code has been included in Appendix A. The pseudocode is given below:

```
establish a connection with the database
declare an array
for each map position
    determine if there are corresponding mutations (records) in the database
    if records are found
        assign the map position to the array
    else
        if map position is clicked on
            show a popup box with “no records” message
        endif
```

```

        endif
    endfor

    if page loads
        for each map position in the array
            underline the corresponding map position
        endfor
    endif

```

3.4 The Mutation Details Page

Design of the Mutation Details Page. The mutation details page contains all the information pertaining to the mutations that were found associated with a particular map position. A sample page, for the +140 map position, is shown in Figure 3.5. Currently, in the database, map positions have between 0-3 mutations associated with them.

The titles of the references (seen underlined in Figure 3.5) are hyperlinks. Clicking on the title opens up the PubMed page with the abstract for the related journal article. Hovering over the hyperlink shows a tooltip with the PubMed identification number (PMID) of the article.

Underlying Code for the Mutation Details Page. Primarily PHP and HTML were used. The high-level code is given below:

```

establish a connection with the database
select all records from all tables where mutation position = map position passed by
                                                    hyperlink (on clicking)
if records are not found
    display a message saying no records found

```

```

else
  for each record found
    appropriately format and display the record
  endfor
endif

```

Mutation Data

Mutation ID:	80
Mutation Description:	Codon 29(GGC->GGT)
Mutation Type:	Substitution - Transition, Silent
Starting Position of Mutation:	140
Mutation Length:	1
Mutation Location:	Exon 1 - Codon 29
Mutation Affects:	mRNA processing - Creates cryptic splice site in exon
Thalassemia Type:	beta+
Thalassemia Severity:	Unknown
Population Found in:	Lebanese (<i>Rare mutation</i>)
Comments:	Mutation can also be referred to as IVS-I(-3)(C->T). Change in amino acid code Gly->Gly. The substitution occurs 2 nucleotides upstream of the GT splice donor site creating a new, competing, splice donor site (Normal splicing - GGC AG:GTTGGT -> Abnormal splicing - G:GTAGGTTGGT). As some normal splicing of the mRNA still occurs the mutation results in beta+ thalassemia.
References:	<ul style="list-style-type: none"> • Chehab FF, Der Kaloustian V, Khouri FP, et al. <u>The molecular basis of beta-thalassemia in Lebanon: application to prenatal diagnosis</u>. <i>Blood</i>. Apr 1987; 69(4):1141-5.
Record Added/ Last Updated By:	Shalu on <i>February 21, 2008 at 11:00:50 am</i>

Figure 3.5. The mutation details page.

3.5 The New Mutation Input Form

Design of the New Mutation Input Form. The mutation input form allows the curator of the database to add new mutation records to the database. The input form can be seen in Figure 3.6.

Mutation ID: 107

Mutation Description:

Mutation Type: or

Starting Position:

Mutation Length:

Mutation Location:

Mutation Effect: or

Disease Type: beta+ (silent)
 beta+
 beta0
 beta(0 or + unclear)

Disease Severity: Thalassemia Major
 Thalassemia Intermedia
 Thalassemia Minor
 Unknown

Comments:

Distribution:

Nationality:	Ethnicity:	Frequency:
<input type="text" value="--Choose one--"/> or <input type="text" value="Enter new Nationality."/>	<input type="text"/>	<input type="text" value="--Choose one--"/>

References:

PMID:	Authors:	Title:	Journal:	Vol:	Iss.No.:
<input type="text"/>	<input type="text" value="Γ et al."/>	<input type="text"/>	<input type="text" value="--Choose one--"/> or <input type="text"/>	<input type="text"/>	<input type="text"/>

Figure 3.6. The new mutation input form.

The Mutation ID is automatically generated. An example is provided on how to enter the Mutation Description. An existing Mutation Type can either be chosen from a drop-down box or a new Mutation Type can be entered into a text box. If a Mutation Type is selected from the drop-down box, the text box is disabled, and vice-versa, to ensure that just one Mutation Type is entered for a mutation. The selection list of the drop-down box is drawn from the existing values in the database. The Starting Position of the mutation refers to the map position of the first base that is mutated. In case of an insertion, the base position of the first base in the insertion is taken as the Starting Position. For instance, if a base C is inserted between an A (position +40) and a T (position +41), the inserted base C is now at position +41 and hence +41 is taken as the Starting Position of the insertion.

The Mutation Length specifies the number of bases involved in the mutation. Examples for the Starting Position and Mutation Length are included in their respective text boxes. Clicking inside the text box removes the example, thus emptying the text box for data entry. An example is provided, in a similar manner, for Mutation Location. Mutation Effect is dealt with much like Mutation Type where both a drop-down box, for existing values, and a text box for the entry of new values is provided. Just one Mutation Effect, an existing one or a new one, can be added. Radio buttons are provided for Disease Type where only one selection can be made. On the other hand checkboxes are provided for Disease Severity where multiple choices can be made. The same mutation can result in different disease severities depending on whether the individual has just one mutated allele, has two mutated alleles, or has other modifying factors present. If the

Unknown option is selected for Disease Severity the other options are disabled and vice-versa.

The Comments text area has a character counter on the side which can be seen in Figure 3.7. If the character count exceeds the limit, an error message is displayed. Dreamweaver's Spry framework was used to add the character counter.

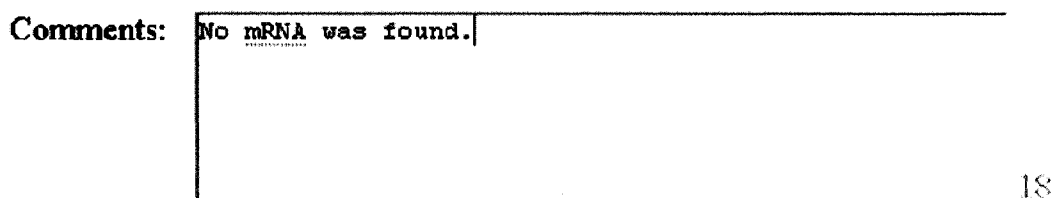


Figure 3.7. The Comments text area with a character counter (seen on the bottom right).

The Distribution includes the Nationality, Ethnicity, and Frequency. The Nationality can be selected from a drop-down box of existing nationalities or a new nationality can be entered into the provided text box. The Ethnicity can be typed into a text box. The Frequency can be selected from one of 4 fixed values – Common, Rare, Novel, and Unknown. As a single mutation can be found in multiple populations, an Add Row button is provided to enable the addition of multiple distributions. The Delete Row button can be used to delete an unwanted row or record.

The References were handled in a similar manner where Add Row and Delete Row buttons were provided to enable the addition of multiple records. Each reference

has a Reference ID that is automatically generated. Text boxes are provided for the PubMed ID (PMID), the names of the first three authors, the title of the article, the volume number, the issue number, the year of publication, and the page numbers. Drop-down boxes are provided for the name of the journal and the month of publication. While the Month drop-down box contains a fixed set of values, the Journal drop-down box contains a selection list based on the existing journal names in the database. A check-box is provided to indicate whether the article has more than 3 authors. The page numbers can be entered by the curator as abbreviated numbers (301-10) or as complete numbers (301-310). If the pages are entered as complete numbers, the underlying code converts the numbers to the abbreviated format that is used in reference citations.

Finally, two buttons are available at the bottom of the page, the Insert Record button allows the entered values to be submitted to the database, while the Clear Form button resets all the fields enabling values to be re-entered.

The contents of a field are checked dynamically and errors messages are generated, without refreshing the page, using Dreamweaver's Spry framework. Sample error messages can be seen in Figure 3.8.

Starting Position: The map begins at -120.

Mutation Length: A value is required.

Figure 3.8. Dynamically generated error messages. The first message prompts the user to enter a valid value for the starting position while the second message warns the user about a blank field.

Underlying Code for the New Mutation Input Form. HTML form elements were used to create the page. In addition, a combination of PHP, JavaScript, and Dreamweaver's Spry framework was used to make the page dynamic. PHP was used to populate the drop-down boxes with existing database values and to insert the entered values into the appropriate tables and columns. The curator name and the date and time of record entry are automatically added to the database using PHP. PHP was also used to abbreviate the page numbers prior to entry into the database. On the other hand, JavaScript was used to disable or enable form elements, based on changes made by the user. JavaScript was also used to reset the form, and to add and delete rows to tables. Dreamweaver was used for form validation and for generating error messages.

3.6 The Mutation List

Design of the Mutation List Query Page. The query page allows database users to generate mutation lists. The users can either choose to look at a list of all the mutations

in the database or they can create a mutation list that is filtered based on various user defined criteria. The query page can be seen in Figure 3.9.

Mutation List - Query Page

Show unfiltered mutation list

Show filtered mutation list

Filter by Nationality:

Filter by Mutation Frequency:

Filter by Ethnicity:

Filter by Disease Type:

Filter by Disease Severity:

Figure 3.9. The mutation list query page. This page allows web site users to create filtered or unfiltered mutation lists.

The web site user can filter the mutations based on Nationality. The drop-down box provides a list of the nationalities currently available in the database. The user can also filter the mutation list by disease severity, disease type, and mutation frequency. Some of the ethnicities that frequently occur in the database are provided in the Ethnicity drop-down box. The list of ethnicities is a static list and can be seen in Figure 3.10.

Filter by Ethnicity:

Filter by Disease Type:

Filter by Disease Severity:

- Black
- Chinese
- Greek
- Italian
- Jew
- Maharashtrian
- Muslim
- Punjabi
- Southern Chinese
- Southern Italian

Figure 3.10. Options available for filtering mutations by ethnicity.

Underlying Code for the Mutation List Query Page. HTML form elements were used to create the page. PHP was used to generate the dynamic list for the Filter by Nationality drop-down box.

Design of the Mutation List Page. The list page displays a filtered or unfiltered list of records from the database. The list page can be seen in Figure 3.11. The number of mutations and associated references found is displayed above each table. The only fields excluded from the mutation list are Comments and Updation Date and Time. The reference titles, as in the mutation details page, link to the article abstract in PubMed. The Reference ID, which was omitted from the details page, can be seen here.

Mutation List

1 mutation(s) found.

Mut. ID	Description	Type	Starting Position	Length	Location	Effect	Disease Type	Disease Severity	Distribution	Entered By
81	IVS-I, 3' end; -17bp	Deletion	256	17	Intron 1 - Position +114 to +130	mRNA processing - Abolishes splicing at 3' splice site	beta0	Unknown	Kuwaiti	Shalu

Reference List

1 reference(s) found.

Ref. ID	PubMed ID	Authors	Title	Journal	Volume	Issue No	Issue Date	Pages
106	3048433	Kazazian HH Jr , Boehm CD	<u>Molecular basis and prenatal diagnosis of beta-thalassemia.</u>	Blood	72	4	Oct 1988	1107-16

Figure 3.11. A sample mutation list. This list is generated when the option Filter by Nationality: Kuwaiti is chosen.

Underlying Code for the Mutation List Page. When the user clicks on the Submit button of the query page, the user's selections are passed, as variables, to the file containing the code for generating the list page. PHP and HTML elements were used to generate the list page. Based on the user's choice of filter, a query is constructed. PHP is then used to connect to the database and obtain the relevant records from the database. HTML is used to generate the table of records.

The following chapter provides some statistics based on records currently in the database. It also enlists some conclusions that could be made by examining the database contents.

Chapter 4 The Database Content

As of now, the database contains 106 mutations. These mutations are located in the promoter, the 5' and 3' UTRs, the open reading frame and the first intron. A listing of all 106 mutations can be seen in Appendix B. Given below are some statistics based on the records in the database:

Mutation Locations

- Twenty-eight of the mutations are located in the promoter of the gene. Nearly all of these mutations are found in the regulatory elements. Most of the promoter mutations are found in the TATA box, followed by the proximal CACCC box. Fewer mutations are based in the CCAAT box and the distal CACCC box.
- Eight of the mutations are located in the 5'UTR, while nine mutations fall in the 3'UTR. Most of the 3'UTR mutations are located in the polyadenylation signal sequence, AATAAA.
- Most of the open reading frame mutations are found in Exon 1 (40 of a total of 55 exonic mutations).
- Nine mutations are based in Intron 1. The majority of these mutations affect the intronic 5' and 3' splice consensus sequences.
- Most base positions have one or two mutations associated with them, however, bases -87, +52, +53, and +144 have three attached mutations each.

-87 is part of the proximal CACCC box, +52 and +53 are part of the start codon (ATG), while +144 is part of the important GT dinucleotide of the intron 1-5' splice consensus sequence.

Mutation Effects

- The processes most often affected by the mutations are transcription followed by mRNA translation.

Mutation Types

- Substitutions are the most frequently seen type of mutations (68 of 106) amongst which transversions are slightly more common than transitions. Deletions (30) are seen far more often than insertions (8).
- The longest mutation in the database is a 105 base pair deletion, extending from base positions -25 to +80. However, overall, there are fewer insertions or deletions involving more than two bases. Large deletions, that involve more than the beta globin gene, have been excluded from the database.

Mutation References

- Almost all the 106 mutations in the database, with the exception of 4, have curator comments attached that are drawn from the references.
- The mutations reference some 139 different sources.

- The largest number of references are drawn from the journal Hemoglobin followed by the British Journal of Haematology.
- The oldest article in the database is from June 1979, while the latest article referenced is from February 2008.
- Two of the mutations have six associated references. The article that is referenced the most (six times) by various records is *beta-thalassemia mutations in Japanese and Koreans* (Ohba, Hattori, Harano, Harano, Fukumaki, & Ideguchi, 1997).

Disease Types

- Most of the mutations result in a disease type of beta0 (56 of 106).

Disease Distribution

- The database includes 46 nationalities and 45 different ethnicities.
- The nationality with the most associated mutations are Asian Indians (13 mutations), followed by Turks (12).
- The ethnicities that have the most mutations in the database are US-Blacks (10), followed by Asian Indian – Punjabis (4). Ashkenazi, Sephardic, and Kurdish Jews (belonging to different countries like Iran, Israel, Russia, and Canada) have 6 associated mutations in the database.

- Twelve of the mutations have a high frequency (are common mutations) in the population in which they are found. The remaining are rare (28) or novel (46) or have an unknown frequency (42).

Most of the above statistics are reflective of what one would expect to find with beta thalassemia mutations. For instance, most beta thalassemia mutations are known to be substitutions (Giardine et al, 2007), a few common mutations are found in each population while the majority of mutations are rare or novel (Weatherall, 2004).

By studying and comparing the properties of the various mutations in the database, some interesting conclusions can be drawn. Furthermore, information can be obtained about the disease and its variations by looking at the given curator comments. Some examples of what one could learn about thalassemia, by examining the database contents, are given below:

Patient Genotype

- Most severely affected patients of beta thalassemia are compound heterozygotes for two beta thalassemia alleles or a beta thalassemia allele and a hemoglobin variant. Hemoglobin variants are beta globin mutations that affect the quality of hemoglobin as opposed to the quantity.
- Homozygous beta thalassemia results from consanguineous marriages, that is, marriages between relatives such as first-cousins. This is seen amongst the Jewish Kurds, who are a highly in-bred ethnic group.

Beta Thalassemia in Different Populations

- Black patients, from the US, tend to have a milder form of thalassemia. The mutations they carry usually lead to beta⁺ thalassemia and thalassemia intermedia. Black patients, also, often have a thalassemia mutation associated with HbS, the sickle-cell anemia causing hemoglobin variant.
- Patients from Northern Africa (Algeria and Tunisia) also show compound heterozygosity with HbS.
- Most Thai patients carry a thalassemia mutation combined with the hemoglobin variant HbE.
- The island of Corsica is the only place in France where beta thalassemia is endemic.
- A promoter mutation (-29 (A->G)) is the most common mutation found in both US-Blacks and Chinese. However, the same mutation may have likely arisen independently in the two populations as no major interactions, between the two populations, have occurred.
- Mutations found in New World patients (US, Canada, Mexico, Surinam) and Northern European patients are seen as a result of immigration from regions where thalassemia has been prevalent such as Africa, Southern Europe, the Middle-east, and India. A few mutations in these regions are the result of novel mutations.
- Beta thalassemia is a severe disorder amongst Bulgarians and Sardinians. On the island of Sardinia beta thalassemia arose as a single mutational event.

Disease Type and Severity

- Most of the promoter and UTR mutations tend to result in a milder disease, as they cause only beta⁺ or silent beta⁺ thalassemia. Silent beta thalassemia is a very mild form of beta thalassemia; patients heterozygous for such mutations are hematologically normal and the condition is only detected when combined with another more severe beta thalassemia mutation. Promoter and UTR mutations frequently result in thalassemia intermedia.
- Exonic mutations usually affect the open reading frame and hence result in beta⁰ thalassemia and thalassemia major. The mRNA transcript carrying a frame-shift mutation is generally destroyed prior to translation.
- Intronic mutations result in beta⁰ thalassemia and thalassemia major as they tend to affect splicing. The improperly spliced mRNA transcripts are destroyed prior to translation.

Mutation Locations

- The majority of the disease-causing promoter mutations are located in the regulatory elements of the promoter, as these regions are important for the binding of transcription factors and the efficient initiation of transcription. The distal CACCC box has fewer mutations than the proximal CACCC box,

although the same factors bind to both, suggesting that the proximal box plays a more important role in transcription initiation.

- When a region or a base does not have mutations associated with it, it just means that no patients have been found carrying a mutation at that position or within that region, most likely because the mutation has little or no effect on the individual. There is also a small likelihood that the region in question is somewhat less prone to mutations, for instance, due to the presence of fewer repeats. As such, overall, the beta globin gene is considered a hot-spot for mutations (Das & Talukder, 2001).

Mutation Effects

- Silent mutations, that do not change the amino acid encoded by a codon, may not seem to affect the polypeptide sequence, but may actually cause disease by disrupting a splice site.
- Some seemingly harmless mutations in the exons or introns may give rise to cryptic splice sites which may compete with the actual splice sites.
- The destruction of a splice site due to a mutation may at times give rise to multiple cryptic splice sites that are then utilized as alternatives.
- Some mutations create or disrupt the recognition site for a restriction enzyme. This fact may help in the identification of a mutation.

Disease Variants

- If a beta thalassemia mutation is simultaneously inherited with an alpha thalassemia, the carrier of the mutations may have a milder disease or may be totally asymptomatic. The presence of both types of thalassemias helps reduce the imbalance between the alpha and globin chains thus reducing disease severity.
- If a beta thalassemia mutation is simultaneously inherited with a mutation in the gamma globin gene, a less severe form of beta thalassemia is seen. The gamma globin gene encodes gamma globin, a component of the fetal hemoglobin, HbF. Certain mutations in the gamma globin gene cause fetal hemoglobin to persist into adulthood, a condition referred to as the Hereditary Persistence of Fetal Hemoglobin (HPFH). Patients that simultaneously inherit beta thalassemia and HPFH usually suffer from thalassemia intermedia instead of major.
- Promoter mutations in the beta globin gene sometimes lead to an indirect increase in the production of delta globin, a component of the minor adult hemoglobin HbA₂. As beta and delta globin have a close evolutionary relationship, the similar promoter regulatory elements in the delta globin gene may compete for the common transcription factors. An increase in delta globin production leads to a less severe disease and likely results in thalassemia intermedia.

- The insertion of a codon may give rise to a dominant beta⁰ thalassemia. The abnormal mRNA is translated into a beta chain with an extra amino acid which does not bind to the alpha-chain, resulting in dominant beta⁰ thalassemia. This condition results in the destruction of beta chains and an accumulation of alpha chains.

Mutation Types

- Beta thalassemia is mostly caused by substitutions. Also, substitutions involving transversions are more likely to cause disease than those involving transitions. This can be explained by the fact that transversions involve a dramatic change in chemical structure.
- Most disease causing insertions and deletions are those that lead to frameshifts. This is expected as frameshifts would majorly alter the open reading frame of the gene.
- Most insertions and deletions are found in regions of short repeats. The repeats could lead to slipped strand mispairing, during DNA replication, which would in turn result in an insertion or deletion.

The following chapter puts forward some ideas for the further enhancement of the created database.

Chapter 5 Plans for Future Improvements

Given below are some of the suggestions for improvement of the Beta Thalassemia Database:

- Input Form
 - Users can be provided with a guide on how to enter a mutation using the form.
 - The functionality of the form can be improved. For instance, along with the delete row option the users can be allowed to choose the row to be deleted (as of now only the last row can be deleted).
 - More Ajax driven error checks can be added to the form.
- Mutation List
 - The user can be allowed to choose which columns are to be included in the mutation list.
 - A filter allowing display of mutations by continent or world region could be added.
 - More complex compound queries, combining two or more filters can be added. For example, users could be allowed to filter mutations by nationality and disease severity.
- Overall Improvements
 - More mutations could be added.

- Users could be allowed to login, leave comments for a mutation or leave suggestions. A user forum could also be provided.
- Thalassemia information and resources for patients could be included.
- Links to the latest thalassemia news could be provided.
- Information could be provided to students about
 - the functions of various parts of a gene,
 - mutation nomenclature,
 - the causes of different types of mutations,
 - the effects of different mutations,
 - the inheritance of thalassemia,
 - the relation between malaria and thalassemia,
 - thalassemia treatment and prevention,
 - the spread of thalassemia mutations within populations,
 - haplotypes and their use in the study of thalassemia mutations,
 - restriction enzymes and their use in detecting mutations,
 - information about beta thalassemia variants (e.g., thalassemia intermedia, silent beta⁺ thalassemia, dominant beta⁰ thalassemia),
 - the structure of hemoglobin,
 - processes involved in the formation of hemoglobin from the gene,
 - the different types of globins, and

- the evolution of the globins.
- Disease related information specific to a nationality or ethnicity could be given. Measures adopted by certain countries, like Cyprus, to successfully reduce beta thalassemia rates could also be included.
- Authentication for the curator should be provided.

These improvements, if implemented, will make the web site easier to use as well as more comprehensive and informative.

References

- Armeli, C., Robbins, S., & Eunpu, D. (2005). Comparing knowledge of β -thalassemia in samples of Italians, Italian-Americans, and non-Italian-Americans. *Journal of Genetic Counseling*, 14(5), 365-76.
- Ayi, K., Turrini, F., Piga, A., & Arese, P. (2004). Enhanced phagocytosis of ring-parasitized mutant erythrocytes: A common mechanism that may explain protection against falciparum malaria in sickle trait and beta-thalassemia trait. *Blood*, 104(10), 3364-71.
- Basran, R., Reiss, U., & Luo, H. (2008). Beta-thalassemia intermedia due to compound heterozygosity for two beta-globin gene promoter mutations, including a novel TATA box deletion. *Pediatric Blood Cancer*, 50(2), 363-6.
- Bhardwaj, U., Zhang, Y., Lorey, F., McCabe, L., & McCabe, E. (2005). Molecular genetic confirmatory testing from newborn screening samples for the common African-American, Asian Indian, Southeast Asian, and Chinese β -thalassemia mutations. *American Journal of Hematology*, 78, 249-55.
- BLAST. (n.d.). Retrieved November 29, 2007, from NCBI: <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>
- Brookshear, J. (2007). *Computer science: An overview* (9th ed.). Boston, MA, USA: Pearson Education.
- Callenberg, K. (2007). *Map: Beta thalassemia mutation map*. Retrieved September 18, 2007, from Beta thalassemia mutation map: <http://beta.sproutcasa.com/node/10>
- Casiday, R., & Frey, R. (1998). *Hemoglobin and the heme group: Metal complexes in the blood for oxygen transport*. Retrieved March 28, 2008, from Department of Chemistry, Washington University, St. Louis, MO: http://www.chemistry.wustl.edu/~courses/genchem/Tutorials/Hemoglobin/151_T3_hemoglobin.htm
- CentOS: Home page. (n.d.). Retrieved March 3, 2008, from CentOS: <http://www.centos.org/>
- Cunningham, M., Macklin, E., Neufeld, E., & Cohen, A. (2004). Complications of β -thalassemia major in North America. *Blood*, 104(1), 34-9.
- Das, S., & Talukder, G. (2001). A review on the origin and spread of deleterious mutants of the beta-globin gene in Indian populations. *HOMO - Journal of Comparative Human Biology*, 52(2), 93-109.

- Eichorn, J. (2006). *Understanding AJAX*. Upper Saddle River, NJ, USA: Prentice Hall.
- Elmasri, R., & Navathe, S. (2007). *Fundamentals of database systems* (5th ed.). Boston, MA, USA: Addison Wesley.
- Giardine, B., van Baal, S., Kaimakis, P., Riemer, C., Miller, W., Samara, M., et al. (2007). HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. *Human Mutation*, 28(2), 206.
- Haldane, J. (1949). Disease and evolution. *La Ricerca Scientifica Supplement*, 19, 3–10.
- Huisman, T., Carver, M., & Baysal, E. (1997). *A syllabus of thalassemia mutations*. Augusta, GA, USA: The Sickle Cell Anemia Foundation.
- Huisman, T., Carver, M., & Efremov, G. (1998). *A syllabus of human hemoglobin variants* (2nd ed.). Augusta, GA, USA: The Sickle Cell Anemia Foundation.
- Kan, Y. (2004). 2003 William Allan Award address - Introductory speech for Sir David Weatherall. *American Journal of Human Genetics*, 74, 382–84.
- Kwiatkowski, D. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics*, 77, 171–90.
- Lorey, F. (2000). Asian immigration and public health in California: Thalassemia in newborns in California. *Journal of Pediatric Hematology/Oncology*, 22(6), 564–66.
- Marengo-Rowe, A. (2007). The thalassemias and related disorders. *Baylor University Medical Center Proceedings*, 20, 27–31.
- McFarland, D. (2007). *Dreamweaver CS3: The missing manual*. Sebastopol, CA, USA: O'Reilly Media.
- NCBI: *PubMed*. (2008). (U.S. National Library of Medicine) Retrieved February 21, 2008, from NCBI: www.pubmed.gov
- Nienhuis, A., Anagnou, N., & Ley, T. (1984). Advances in thalassemia research. *Blood*, 63(4), 738-58.
- Ohba, Y., Hattori, Y., Harano, T., Harano, K., Fukumaki, Y., & Ideguchi, H. (1997). Beta-thalassemia mutations in Japanese and Koreans. *Hemoglobin*, 21(2), 191-200.
- Old, J. (2003). Screening and genetic diagnosis of haemoglobin disorders. *Blood Reviews*, 17, 43-53.

- Patrinos, G., Giardine, B., Riemer, C., Miller, W., & Chui, D. (2004). Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. *Nucleic Acids Research*, 32(Database issue), D537-41.
- Pfaffenberger, B. (2003). *Webster's new world computer dictionary* (10th ed.). Indianapolis, IN, USA: Wiley.
- Seidler, K. (2007, December 15). *Index page*. Retrieved December 27, 2007, from Apache Friends: <http://www.apachefriends.org/en/index.html>
- Top reasons to use MySQL*. (n.d.). Retrieved September 24, 2007, from MySQL: <http://www.mysql.com/why-mysql/topreasons.html>
- Vichinsky, E., MacKlin, E., Wayne, J., Lorey, F., & Olivieri, N. (2005). Changes in the epidemiology of thalassemia in North America: A new minority disease [Electronic article]. *Pediatrics*, 116(6), e818-25. Retrieved April 13, 2008, from Pediatrics: <http://pediatrics.aappublications.org/cgi/content/full/116/6/e818>
- VMware: Virtualization Basics*. (2008). Retrieved March 28, 2008, from VMware: <http://www.vmware.com/virtualization/>
- Wajcman, H., Patrinos, G., & Anagnou, N. (2008, February 5). *Menu*. Retrieved February 21, 2008, from HbVar: A Database of Human Hemoglobin Variants and Thalassemias: <http://globin.bx.psu.edu/hbvar/menu.html>
- Weatherall, D. (2004). 2003 William Allan Award address - The thalassemias: The role of molecular genetics in an evolving global health problem. *American Journal of Human Genetics*, 74, 385–92.
- Weatherall, D., & Clegg, J. (1996). Thalassemia – a global public health problem. *Nature*, 2(8), 847–9.
- Weatherall, D., & Clegg, J. (2001). *The thalassaemia syndromes* (4th ed.). Oxford, UK: Blackwell Scientific.
- Weddell, P. (2003, March 25). *Glossary*. Retrieved March 3, 2008, from Paul Weddell's Computing Guide: <http://www.mtw.pwp.blueyonder.co.uk/weddell/guide/gloss1.html>
- XAMPP*. (2008, February 25). Retrieved March 6, 2008, from Wikipedia - The Free Encyclopedia: <http://en.wikipedia.org/wiki/XAMPP>

Appendix A

Code for Automated Underlining of Hyperlinks

The following code detects which map positions in the mutation map have associated mutations in the database and then proceeds to underline those map positions. Underlining indicates that the map position is a hyperlink that can be clicked on to display the details of the associated mutations. This piece of code overcomes the problem of having to manually convert map positions into hyperlinks every time the database is updated and mutations are added to new positions.

```
<?php
require_once('Connections/conn2ThalassemiaDB.php');
//Declaring a PHP array.
$Und_pos_arr= array();

//For each map position, connecting to the database and determining if there are mutation
records associated with the position.
for ($Map_pos=-120; $Map_pos <=+1706; $Map_pos++)
{
    mysql_select_db($database_conn2ThalassemiaDB, $conn2ThalassemiaDB);
    $query_Recordset1 = sprintf ("SELECT MUTATION.MutID FROM MUTATION
WHERE Strt_pos = %s", $Map_pos);
    $Recordset1 = mysql_query($query_Recordset1, $conn2ThalassemiaDB) or
die(mysql_error());
    $row_Recordset1 = mysql_fetch_assoc($Recordset1);
    $num_rows=mysql_num_rows($Recordset1);
//If mutations are found associated with a position, the map position is assigned to the
declared PHP array.
//If no associated mutations are found, when a user clicks on the map position, a popup
box is displayed.
```

```

if ($num_rows > 0)
{
    $Und_pos_arr[]=$Map_pos;
}
else
{?>
<script type="text/javascript">
    <!--
        var No_recs = document.getElementById(<?php echo $Map_pos ?>);
        No_recs.onclick= function showBox()
        {
            alert ("Sorry, no mutations recorded for this position.");
            return false;
        }
    -->
</script>
<?php
}
}
?>
<!--When the page reloads, each map position stored in the PHP array is underlined-->
<script type="text/javascript">
<!--
window.onload= function Changes()
{
<?php foreach ($Und_pos_arr as $Und_pos)
{?>
    document.getElementById(<?php echo $Und_pos ?>).style.textDecoration="underline";
<?php
}?>
}
-->
</script>

```

Appendix B

List of Mutations in the Beta Thalassemia Database

Given below is the list of 106 mutations currently in the database:

Mut. ID	Description	Type	Starting Position	Length	Location	Effect	Disease Type	Disease Severity	Distribution
1	-101(C->T)	Substitution - Transition	-101	1	Promoter regulatory element, distal CACCC	Transcription	beta+ (silent)	Intermedia	Turkish (<i>Rare mutation</i>); Bulgarian (<i>Rare mutation</i>); Italian - Southern Italian (<i>Rare mutation</i>)
2	-92(C->T)	Substitution - Transition	-92	1	Promoter regulatory element, proximal CACCC	Transcription	beta+ (silent)	Intermedia	Sicilian (<i>Rare mutation</i>)
3	-90(C->T)	Substitution - Transition	-90	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Unknown	Portuguese (<i>Rare mutation</i>)
4	-88(C->T)	Substitution - Transition	-88	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Intermedia	US - Black (<i>Common mutation</i>); Asian Indian - Punjabi (<i>Rare mutation</i>)

5	-88(C->A)	Substitution - Transversion	-88	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Intermedia	Iranian - Kurdish Jew
6	-87(C->G)	Substitution - Transversion	-87	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Intermedia	Sardinian (<i>Rare mutation</i>), Turkish (<i>Rare mutation</i>); Italian (<i>Rare mutation</i>)
7	-87(C->T)	Substitution - Transition	-87	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Intermedia	German, Italian - German- Italian (<i>Rare mutation</i>)
8	-87(C->A)	Substitution - Transversion	-87	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Intermedia	US - Black (<i>Novel mutation</i>)
9	-86(C->G)	Substitution - Transversion	-86	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Major	Thai (<i>Rare mutation</i>); Lebanese
10	-86(C->A)	Substitution - Transversion	-86	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Intermedia	Italian (<i>Rare mutation</i>)
11	-32(C->A)	Substitution - Transversion	-32	1	Promoter regulatory element, TATA box- CATATAA	Transcription	beta+	Minor	Taiwanese - Fukien Taiwanese (<i>Novel mutation</i>)

12	-31(A->G)	Substitution - Transition	-31	1	Promoter regulatory element, TATA box-CATAAAA	Transcription	beta+	Intermedia	Japanese (<i>Common mutation</i>)
13	-31(A->C)	Substitution - Transversion	-31	1	Promoter regulatory element, TATA box-CATAAAA	Transcription	beta+	Unknown	Italian (<i>Novel mutation</i>)
14	-30(T->A)	Substitution - Transversion	-30	1	Promoter regulatory element, TATA box-CATAAAA	Transcription	beta+	Intermedia	Macedonian (<i>Rare mutation</i>) ; Turkish (<i>Rare mutation</i>) ; Tunisian (<i>Rare mutation</i>)
15	-30(T->C)	Substitution - Transition	-30	1	Promoter regulatory element, TATA box-CATAAAA	Transcription	beta(0 or + unclear)	Unknown	Chinese - South-eastern Chinese (Fujian province) (<i>Novel mutation</i>)
16	-29(A->G)	Substitution - Transition	-29	1	Promoter regulatory element, TATA box-CATAAAA	Transcription	beta+	Minor	US - Black (<i>Common mutation</i>) ; Chinese (<i>Common mutation</i>)
17	-28(A->C)	Substitution - Transversion	-28	1	Promoter regulatory element, TATA box-CATAAAA	Transcription	beta+	Major	Israeli - Kurdish Jew
18	-28(A->G)	Substitution - Transition	-28	1	Promoter regulatory element, TATA box-CATAAAA	Transcription	beta+	Intermedia	Chinese (<i>Common mutation</i>)

19	-27(A->T)	Substitution - Transversion	-27	1	Promoter regulatory element, TATA box- CATAAAA	Transcription	beta+	Minor	French - Corsican (<i>Novel mutation</i>)
20	-25(G->C)	Substitution - Transversion	-25	1	Promoter regulatory element, TATA box- CATAAAA	Transcription	beta(0 or + unclear)	Unknown	US - Black
21	-93(C->G)	Substitution - Transversion	-93	1	Promoter regulatory element, proximal CACCC	Transcription	beta+	Unknown	Surinamese
22	-56(G->C)	Substitution - Transversion	-56	1	Promoter	Transcription	beta(0 or + unclear)	Unknown	Moroccan
23	-32(C->T)	Substitution - Transition	-32	1	Promoter regulatory element, TATA box- CATAAAA	Transcription	beta+	Unknown	Canadian - Hispanic (<i>Novel mutation</i>)
24	-27(-AA)	Deletion	-27	2	Promoter regulatory element, TATA box- CATAAAA	Transcription	beta(0 or + unclear)	Intermedia	US - Black (<i>Novel mutation</i>)
25	-102(C->A)	Substitution - Transversion	-102	1	Promoter regulatory element, distal CACCC	Transcription	beta+ (silent)	Intermedia	French - Tunisian(North African) (<i>Novel mutation</i>)

26	-73(A->T)	Substitution - Transversion	-73	1	Promoter regulatory element, CCAAT box	Transcription	beta+ (silent)	Intermedia	Chinese - Southern Chinese (<i>Novel mutation</i>)
27	-101(C->G)	Substitution - Transversion	-101	1	Promoter regulatory element, distal CACCC	Transcription	beta+ (silent)	Intermedia	Russian - Ashkenazi Jew
28	105 bp del (-25 to 80)	Deletion	-25	105	Promoter-5'UTR-Exon1(Cd 9)	Transcription	beta0	Unknown	Thai
29	CAP +1(A->C)	Substitution - Transversion	1	1	5'UTR (1st base to be transcribed)	Transcription	beta+ (silent)	Intermedia, Major, Minor	Asian Indian - Punjabi (<i>Rare mutation</i>)
30	CAP +8(C->T)	Substitution - Transition	8	1	5'UTR	Transcription	beta+ (silent)	Unknown	Chinese (<i>Novel mutation</i>)
31	CAP +10(-T)	Deletion	10	1	5'UTR	Transcription, mRNA stability	beta+ (silent)	Intermedia	Greek (<i>Novel mutation</i>)
32	CAP +20(C->T)	Substitution - Transition	20	1	5'UTR	Transcription, mRNA stability	beta+	Intermedia	Turkish, Bulgarian
33	CAP +22(G->A)	Substitution - Transition	22	1	5'UTR	Transcription, mRNA stability	beta+	Intermedia	Canadian - Italian (<i>Novel mutation</i>)
34	CAP +33(C->G)	Substitution - Transversion	33	1	5'UTR	Transcription, mRNA stability	beta+ (silent)	Intermedia	Cypriot - Greek
35	CAP +40(-AAAC)	Deletion	40	4	5'UTR	Transcription, mRNA stability	beta+	Unknown	Chinese (<i>Novel mutation</i>)
36	+1480(C->G)	Substitution - Transversion	1480	1	3'UTR	mRNA stability	beta+ (silent)	Intermedia	Greek (<i>Rare mutation</i>)

37	+1565(- GCATCTGGATTCT)	Deletion	1565	13	3'UTR	mRNA stability	beta(0 or + unclear)	Unknown	Turkish (<i>Novel mutation</i>)
38	+1570(T->C)	Substitution - Transition	1570	1	3'UTR	mRNA stability	beta+	Minor	Canadian - Irish (<i>Novel mutation</i>)
39	+1584(T->C)	Substitution - Transition	1584	1	3'UTR, PolyA signal AATAAA	mRNA cleavage and polyadenylation	beta+	Intermedia, Major	Turkish (<i>Rare mutation</i>); US - Black (<i>Rare mutation</i>)
40	+1521(C->G)	Substitution - Transversion	1521	1	3'UTR	Unknown	beta+	Unknown	Armenian
41	+1585(A->G)	Substitution - Transition	1585	1	3'UTR, PolyA signal AATAAA	mRNA cleavage and polyadenylation	beta+	Unknown	Macedonian (<i>Rare mutation</i>); Bulgarian (<i>Rare mutation</i>); Greek (<i>Rare mutation</i>)
42	+1586(A->G)	Substitution - Transition	1586	1	3'UTR, PolyA signal AATAAA	mRNA cleavage and polyadenylation	beta+	Unknown	Malay (<i>Rare mutation</i>)
43	+1587(A->G)	Substitution - Transition	1587	1	3'UTR, PolyA signal AATAAA	mRNA cleavage and polyadenylation	beta+	Intermedia, Major	Israeli - Kurdish Jew (<i>Common mutation</i>)
44	+1582(-AATAA)	Deletion	1582	5	3'UTR, PolyA signal AATAAA	mRNA cleavage and polyadenylation	beta+	Major	UAE, Israeli - Arab Muslim (Gaza) (<i>Novel mutation</i>)
45	+1584(-TA)	Deletion	1584	2	3'UTR, PolyA signal AATAAA	mRNA cleavage and polyadenylation	beta+	Unknown	US - Black, French - Normandy (<i>Novel mutation</i>)

46	Start codon(ATG->GTG)	Substitution - Transition	51	1	Exon 1, Start codon	mRNA translation	beta0	Unknown	Japanese (<i>Rare mutation</i>)
47	Start codon(ATG->AAG)	Substitution - Transversion	52	1	Exon 1, Start codon	mRNA translation	beta(0 or + unclear)	Intermedia, Minor	Canadian - North European (<i>Novel mutation</i>), French - Caucasian(South-eastern French)
48	Start codon(ATG->ACG)	Substitution - Transition	52	1	Exon 1, Start codon	mRNA translation	beta0	Unknown	Croatian (<i>Rare mutation</i>); Swiss - Bern (<i>Novel mutation</i>); Belgian
49	Start codon(ATG->AGG)	Substitution - Transversion	52	1	Exon 1, Start codon	mRNA translation	beta0	Unknown	Chinese (<i>Novel mutation</i>); Korean (<i>Common mutation</i>)
50	Start codon(ATG->ATC)	Substitution - Transversion	53	1	Exon 1, Start codon	mRNA translation	beta0	Unknown	Japanese (<i>Novel mutation</i>)
51	Start codon(ATG->ATA)	Substitution - Transition, Missense	53	1	Exon 1 - Start codon	mRNA translation	beta0	Minor	Italian, Swedish - Northern Swedish
52	Start codon(ATG->ATT)	Substitution - Transversion, Missense	53	1	Exon 1 - Start codon	mRNA translation	beta0	Unknown	Iranian
53	Codon 1(-G), (GTG->IG)	Deletion - Frameshift	54	1	Exon 1 - Codon 1	mRNA translation	beta0	Unknown	Sardinian (<i>Novel mutation</i>)
54	Codons 2/3/4(-9bp; +31bp)	Deletion & Insertion - Frameshift	57	9	Exon 1 - Codons 2,3,4	mRNA translation	beta0	Unknown	Algerian (<i>Novel mutation</i>)
55	Codon 4(ACT->ACA), Codon 5 (CCT->TCT), Codon 6(GAG->TAG)	Substitution - Transversion (Codon 4,6), Transition (Codon 5), Silent(Codon 4), Missense (Codon 5), Nonsense (Codon 6)	65	5	Exon 1 - Codons 4,5,6	mRNA translation	beta0	Major, Minor	Asian Indian - North Indian (Lucknow)

56	Codon 5(-CT), (CCT->C)	Deletion - Frameshift	67	2	Exon 1 - Codon 6	mRNA translation	beta0	Major	Greek (<i>Rare mutation</i>)
57	Codon 6(-A), (GAG->GG)	Deletion - Frameshift	70	1	Exon 1 - Codon 6	mRNA translation	beta0	Major	Bulgarian, US - Black(South-eastern US), US - Southern Italian, Italian - Southern Italian (Calabria) (<i>Rare mutation</i>)
58	Codon 7(GAG->TAG)	Substitution - Transversion, Nonsense	72	1	Exon 1 - Codon 7	mRNA translation	beta0	Unknown	British - English
59	Codon 8(-AA), (AAG->G)	Deletion - Frameshift	75	2	Exon 1 - Codon 8	mRNA translation	beta0	Unknown	Turkish - South-eastern Turkish (Gaziantep) ; Israeli
60	Codon 8/9(+G), (AAG TCT -> AAG G TCT)	Insertion - Frameshift	78	1	Exon 1 - Codons 8/9	mRNA translation	beta0	Intermedia	Japanese, Asian Indian
61	Codons 9/10(+T), (TCT GCC -> TCT T GCC)	Insertion - Frameshift	81	1	Exon 1 - Codons 9/10	mRNA translation	beta0	Unknown	Greek (<i>Novel mutation</i>)
62	Codon 10(GCC->GCA)	Substitution - Transversion, Silent	83	1	Exon 1 - Codon 10	mRNA processing - Creates cryptic splice site in exon	beta+	Major	Asian Indian - Central Indian (Madhya Pradesh) (<i>Novel mutation</i>)
63	Codon 11(-T), (GTT->GT)	Deletion - Frameshift	86	1	Exon 1 - Codon 11	mRNA translation	beta0	Unknown	Mexican - Mestizo (<i>Novel mutation</i>)
64	Codons 14/15(+G), (CTG TGG -> CTG G TGG)	Insertion - Frameshift	96	1	Exon 1 - Codons 14/15	mRNA translation	beta0	Major	Chinese - Southern Chinese (Guangdong province) (<i>Novel mutation</i>)
65	Codon 15(TGG->TAG)	Substitution - Transition, Nonsense	97	1	Exon 1 - Codons 15	mRNA translation	beta0	Intermedia	Asian Indian (<i>Common mutation</i>) ; Turkish (<i>Rare mutation</i>) ; Japanese
66	Codon 15(TGG->TGA)	Substitution - Transition, Nonsense	98	1	Exon 1 - Codon 15	mRNA translation	beta0	Major, Minor	Japanese, Portuguese - Central Portuguese (<i>Common mutation</i>)

67	Codon 15(-T), (TGG->GG)	Deletion - Frameshift	96	1	Exon 1 - Codon 15	mRNA translation	beta0	Major	Malay (Novel mutation)
68	Codons 15/16(-G), (TGG GGC -> TGG GC)	Deletion - Frameshift	99	1	Exon 1 - Codons 15/16	mRNA translation	beta(0 or + unclear)	Minor	German (Novel mutation)
69	Codon 16(-C), (GGC->GG)	Deletion - Frameshift	101	1	Exon 1 - Codon 16	mRNA translation	beta0	Unknown	Asian Indian (Rare mutation)
70	Codon 17, (AAG->TAG)	Substitution - Transversion, Nonsense	102	1	Exon 1 - Codon 17	mRNA translation	beta0	Unknown	Chinese
71	Codon 22(GAA->TAA)	Substitution - Transversion, Nonsense	117	1	Exon 1 - Codon 22	mRNA translation	beta0	Unknown	Reunion Island (Novel mutation)
72	Codons 22/23/24(-7 bp), (-AAGTTGG)	Deletion - Frameshift	118	7	Exon 1 - Codons 22/23/24	mRNA translation	beta0	Major	Turkish (Novel mutation)
73	Codon 24(-G; +CAC), (GGT -> CAC GT)	Deletion & Insertion - Frameshift	123	1	Exon 1 - Codon 24	mRNA translation	beta0	Major	Egyptian (Novel mutation)
74	Codons 25/26(+T), (GGT GAG -> GGT T GAG)	Insertion - Frameshift	129	1	Exon 1 - Codons 25/26	mRNA translation	beta0	Major	Tunisian (Novel mutation)
75	Codon 26(+T), (GAG -> GTA G)	Insertion - Frameshift	130	1	Exon 1 - Codon 26	mRNA translation	beta0	Unknown	Japanese (Novel mutation)
76	Codon 26(GAG->TAG)	Substitution - Transversion, Nonsense	129	1	Exon 1 - Codon 26	mRNA translation	beta0	Major	Thai - North-eastern Thai (Novel mutation)
77	Codons 27/28(+C), (GCC CTG -> GCC C CTG)	Insertion - Frameshift	135	1	Exon 1 - Codons 27/28	mRNA translation	beta0	Major	Chinese (Novel mutation), Taiwanese - Hakka Taiwanese (Novel mutation)
78	Codon 28(-C), (CTG->TG)	Deletion - Frameshift	135	1	Exon 1 - Codon 28	mRNA translation	beta0	Unknown	Egyptian (Novel mutation)

79	Codons 28/29(-G), (CTG GGC -> CT GGC or CTG GC)	Deletion - Frameshift	137	1	Exon 1 - Codon 29	mRNA translation	beta0	Unknown	Japanese (<i>Rare mutation</i>), Egyptian (<i>Rare mutation</i>)
80	Codon 29(GGC- >GGT)	Substitution - Transition, Silent	140	1	Exon 1 - Codon 29	mRNA processing - Creates cryptic splice site in exon	beta+	Unknown	Lebanese (<i>Rare mutation</i>)
81	IVS-1, 3' end; -17bp	Deletion	256	17	Intron 1 - Position +114 to +130	mRNA processing - Abolishes splicing at 3' splice site	beta0	Unknown	Kuwaiti
82	Codon 24(GGT- >GGA)	Substitution - Transversion, Silent	125	1	Exon 1 - Codon 24	mRNA processing - Creates cryptic splice site in exon	beta+	Intermedia, Minor	US - Black(South-eastern US) (<i>Rare mutation</i>); Japanese
83	Codons 38/39(-CC), (ACC CAG -> A CAG)	Deletion - Frameshift	296	2	Exon 2 - Codons 38/39	mRNA translation	beta0	Unknown	Belgian (<i>Novel mutation</i>)
84	Codon 39(CAG- >TAG)	Substitution - Transition, Nonsense	298	1	Exon 2 - Codon 39	mRNA translation	beta0	Intermedia, Major	Sardinian (<i>Common mutation</i>); Italian (<i>Common mutation</i>); Turkish
85	Codon 40(-G), (AGG- >AG)	Deletion - Frameshift	302	1	Exon 2 - Codon 40	mRNA translation	beta0	Unknown	Japanese
86	Codon 30(AGG- >GGG)	Substitution - Transition, Missense	141	1	Exon 1 - Codon 30	mRNA processing - Abolishes splicing at 5' splice site	beta0	Unknown	Canadian - Sephardic Jew (<i>Novel mutation</i>)

87	Codon 30(AGG->AAG)	Substitution - Transition, Missense	142	1	Exon 1 - Codon 30	mRNA processing - Abolishes splicing at 5' splice site	beta0	Unknown	Bulgarian, UAE
88	Codon 40(+86 bp)	Insertion - Frameshift	303	86	Exon 2 - Codon 40	mRNA translation	beta0	Minor	Portuguese - Northern Portuguese (<i>Novel mutation</i>)
89	Codon 30(AGG->AGC)	Substitution - Transversion, Missense	273	1	Exon 2 - Codon 30	mRNA processing - Abolishes splicing at 3' splice site	beta0	Unknown	UAE
90	Codon 31(-C), (CTG->TG)	Deletion - Frameshift	274	1	Exon 2 - Codon 31	mRNA translation	beta0	Unknown	Taiwanese - Chinese (<i>Novel mutation</i>)
91	Codons 30/31(+CGG)	Insertion	274	3	Exon 2 - Codons 31/32	Assembly of hemoglobin molecule	beta0	Minor	Spanish (<i>Novel mutation</i>)
92	44 bp deletion	Deletion - Frameshift	126	44	Exon 1 - Codon 25 to Intron 1 - Position +27	mRNA processing - Abolishes splicing at 5' splice site	beta0	Major	Macedonian, Greek
93	Codon 35(-C), (TAC->TA)	Deletion - Frameshift	288	1	Exon 2 - Codon 35	mRNA translation	beta0	Unknown	Malay - Western Malay (<i>Rare mutation</i>)
94	Codon 35(TAC->TAA)	Substitution - Transversion, Nonsense	288	1	Exon 2 - Codon 35	mRNA translation	beta0	Major	Thai - core-Thai (<i>Rare mutation</i>)
95	Codons 36/37(-T), (CCT TGG -> CCT GG)	Deletion - Frameshift	291	1	Exon 2 - Codons 36/37	mRNA translation	beta0	Major	Iranian - Kurdish Jew(Western Iran), Iranian - Muslim

96	Codon 37(TGG->TGA)	Substitution - Transition, Nonsense	294	1	Exon 2 - Codon 37	mRNA translation	beta0	Major	Spanish - Catalanian(Ebro Delta) (<i>Common mutation</i>); Saudi Arabian, Jordanian.
97	25 bp deletion	Deletion - Frameshift	252	25	Intron 1 - Position +110 to Exon 2 - Codon 31	mRNA processing - Abolishes splicing at 3' splice site	beta0	Unknown	Asian Indian
98	Codon 37(TGG->TAG)	Substitution - Transition, Nonsense	293	1	Exon 2 - Codon 37	mRNA translation	beta0	Intermedia, Major	Afghan (<i>Novel mutation</i>)
99	Codons 37/38/39(-7 bp), (TGG ACC CAG -> TG)	Deletion - Frameshift	294	7	Exon 2 - Codons 37/38/39	mRNA translation	beta0	Major	Turkish (<i>Novel mutation</i>)
100	Codons 38/39(-C), (ACC CAG -> ACC AG)	Deletion - Frameshift	296	1	Exon 2 - Codons 38/39	mRNA translation	beta0	Minor	Czech (<i>Novel mutation</i>)
101	IVS I-1(G->A), (AG/GTTGGT -> AG/ATTGGT)	Substitution - Transition	143	1	Intron 1 - Position +1	mRNA processing - Abolishes splicing at 5' splice site	beta0	Major	Cypriot - Greek (<i>Rare mutation</i>); Spanish - South-western Spanish (Huelva province) (<i>Common mutation</i>); Turkish, Czech
102	IVS I-1(G->T), (AG/GTTGGT -> AG/TTGGT)	Substitution - Transversion	143	1	Intron 1 - Position +1	mRNA processing - Abolishes splicing at 5' splice site	beta0	Unknown	Asian Indian - Maharashtra, Asian Indian - Punjabi
103	IVS I-2(T->G), (AG/GTTGGT -> AG/GGTGGT)	Substitution - Transversion	144	1	Intron 1 - Position +2	mRNA processing - Abolishes splicing at 5' splice site	beta0	Unknown	Tunisian

104	IVS 1-2(T->C); (AG GT GGT -> AG GCT GGT)	Substitution - Transition	144	1	Intron 1 - Position +2	mRNA processing - Abolishes splicing at 5' splice site	beta0	Major	US - Black, Algerian (<i>Rare mutation</i>)
105	IVS 1-2(T->A); (AG GT GGT -> AG GAT GGT)	Substitution - Transversion	144	1	Intron 1 - Position +2	mRNA processing - Abolishes splicing at 5' splice site	beta0	Intermedia, Major	Italian - Southern Italian (<i>Novel mutation</i>); Algerian
106	IVS 1-5(G->C); (AG GT GGT -> AG GT GCT)	Substitution - Transversion	147	1	Intron 1 - Position +5	mRNA processing - reduces splicing at 5' splice site	beta+	Intermedia, Major	Asian Indian - Punjabi (<i>Common mutation</i>); Asian Indian - Maharashtrian (<i>Common mutation</i>); Asian Indian - Oriya, Chinese - Southern Chinese (Guangdong province) (<i>Rare mutation</i>); Vanuatu - Maewo island (<i>Common mutation</i>); German, Papua New Guinean - Coastal (<i>Common mutation</i>)