# WHOLE EXOME SEQUENCING TO IDENTIFY MUTATIONS IN GENES IN NEPALESE PATIENTS WITH RARE BLEEDING DISORDERS

**M.Sc. Thesis**

2017

Submitted to

**Central Department of Biotechnology**

**Tribhuvan University**

**Kirtipur, Kathmandu, Nepal**


By

**Binod Neupane**

Registration No: 5-2-37-963-2009


**Supervisors**


**Dr. Tilak R. Shrestha**
Professor
Central Department of Biotechnology,
Tribhuvan University, Kirtipur,
Nepal

**Dr. Sridhar Sivasubbu**
Senior Scientist
CSIR-Institute of Genomics and
Integrative Biology, New Delhi,
India

# Acknowledgement

*Though only my name appears on the cover of this dissertation, a lot of guidance and assistance from many great people have contributed to its success and final outcome. I owe my gratitude to all those people who have made this dissertation possible and because of whom my dissertation experience has been one that I will cherish forever. Whatever I have done is only due to such guidance and assistance and it's my great pleasure to thank them.*

*First and foremost, I would like to express my profound gratitude to my thesis supervisor Prof. Dr. Tilak R. Shrestha for his eternal support, proper guidance and constant feedback without which this research would not have been done. The present research work on rare bleeding disorders is based on his ideas, preconception and collaborative research agreement done with Dr. Shridhar of IGIB, New Delhi, India.*

*I am equally indebted to my another supervisor, Dr. Sridhar Sivasubbu for accepting me as a training fellow and providing me an opportunity to access the world class laboratory and research facilities including Next Generation Sequencing and Zebra fish facilities. His guidance, encouragement and dedication has not only helped me at the time of this research but also has helped me in personality development.*

*My deep appreciation goes out to Prof. Dr. Rajani Malla, former HOD of CDBT-TU (HOD at our time) for her kind support my thesis work.*

*I am extremely thankful to NGS teammates of SSB lab (Lab No. 123) of CSIRI-IGIB viz. Shamsudheen K Vellarikkal, Ankit Verma, Rijith Jayarayan, Rowmika Ravi and Anoop Kumar for sharing expertise, sincere and valuable guidance and encouragement extended to me.*

*My deep appreciation goes out to members of Nepal Hemophilia Society (NHS) especially President Mr. Bed Raj Dhungana, Secretory Mr. Ujjawal K.C., Program manager Ms. Laxmi Karki and nurses working at Hemophilia Care Unit, Bir Hospital. Similarly, all the patients with rare bleeding disorders and their family members who participated in this research without whom my dream to work in molecular characterization of Nepalese Hemopiliacs would have never been completed, owe my deepest acknowledgement.*

*I owe my special thanks to all my batch mates (CDBT 5$^{th}$ Batch); professors and lecturers at CDBT-TU; my seniors and my juniors. Medha K.C., Nutan Thakur, Sujan Biswakarma and Gauri Thapa deserve more gratitude for their kind support throughout my whole M.Sc. including this research work.*

*Last but not the least, I feel very happy to thank my ever inspiring mom-dad, sisters and brothers for their eternal love, care and encouragement. They have been supporting me all through the thick and thin from my birth and always have kept me in high spirit throughout the entire period of my academic as well as personal life.*

# TABLE OF CONTENTS

## List of Abbreviations

| | | |
|---|---|---|
| % | - | Percentage |
| 2N VWD | - | Type 2 Normandy von Willebrand disease |
| AD | - | Autosomal Dominant |
| ADAMTS13 | - | A Disintegrin And Metalloproteinase with a Thrombospondin Type 1 Motif, Member 13 |
| APTT | - | Activated Partial Thromboplastin Time |
| AR | - | Autosomal Recessive |
| B | - | Benign |
| BAM | - | Binary Alignment/Map |
| bcl | - | base call |
| BT | - | Bleeding Time |
| BWA | - | Burrow Wheeler Aligner |
| BWA-MEM | - | Burrows-Wheeler Alignment- Maximal Exact Matches |
| CASAVA | - | Consensus Assessment of Sequence And Variation |
| Chr | - | Chromosome |
| CSGE | - | Conformation Sensitive Gel Electrophoresis |
| D | - | Deleterious or Probably Damaging |
| ddNTPs | - | DiDeoxy Nucleotides |
| DDW | - | Double Distilled Water |
| DGGE | - | Denaturing Gradient Gel Electrophoresis |
| DHPLC | - | Denaturing High Pressure Liquid Chromatography |
| EDTA | - | Ethylene Diamine Tetra-Acetic Acid |
| ERGIC-53 | - | Endoplasmic Reticulum-Golgi Intermediate Compartment 53 Kda Protein |
| EtOH | - | Ethyl Alcohol |
| ExAC | - | Exome Aggregation Consortium |
| F | - | Phenylalanine |
| *F10* | - | Coagulation Factor X gene |
| *F11* | - | Coagulation Factor XI gene |
| *F12* | - | Coagulation Factor XII gene |
| *F13A1* | - | Coagulation Factor XIII A Chain gene |

| | | |
|---|---|---|
| *F13B* | - | Coagulation Factor XIII B Chain gene |
| *F2* | - | Coagulation Factor II gene |
| *F5* | - | Coagulation Factor V gene |
| *F7* | - | Coagulation Factor VII gene |
| *F8* | - | Coagulation Factor VIII gene |
| *F9* | - | Coagulation Factor IX gene |
| FC | - | Flow Cell |
| *FGA* | - | Fibrinogen Alpha Chain |
| *FGB* | - | Fibrinogen Beta Chain |
| Fig | - | Figure |
| FII | - | Factor II |
| FIX | - | Factor IX |
| FV | - | Factor V |
| FVIII | - | Factor VIII |
| FX | - | Factor X |
| FXD | - | Factor X Deficiency |
| FXI | - | Factor XI |
| FXII | - | Factor XII |
| gDNA | - | Genomic Deoxy Ribonucleic Acid |
| *GGCX* | - | γ- Glutamyl Carboxylase gene |
| GWAS | - | Genome Wide Association Studies |
| HA | - | Hemophilia A |
| HB | - | Hemophilia B |
| HMWK | - | High Molecular Weight Kininogen |
| IGV | - | Integrative Genomics Viewer |
| Indels | - | insertions deletions |
| IU/ml | - | International Units per Millilitre |
| LMAN1 | - | Lectin Mannose Binding 1 |
| mg | - | Miligram |
| mL | - | Mililitre |
| MCFD2 | - | Multiple Coagulation Factor Deficiency 2 |

| | | |
|---|---|---|
| mRNA | - | Messenger Ribonucleic Acid |
| NCBI | - | National Center for Biotechnology Information |
| NFW | - | Nuclease Free Water |
| NGS | - | Next Generation Sequencing |
| NHS | - | Nepal Hemophilia Society |
| PCR | - | Polymerase Chain Reaction |
| Polyphen2 | - | Polymorphism Phenotyping v2 |
| PT | - | Prothrombin Time |
| Q | - | Phred Quality Score |
| R | - | Arginine amino acid |
| RBDs | - | Rare Bleeding Disorders |
| RSB | - | Resuspension Buffer |
| RT | - | Room Temperature |
| S | - | Serine |
| SAM | - | Sequence Alignment/Map |
| SDS | - | Sodium Dodecyl Sulfate |
| SIFT | - | Sorting Intolerant From Tolerant |
| SNVs | - | Single Nucleotide Variations |
| SSCP | - | Single Strand Conformation Polymorphism |
| TAE | - | Tris-Acetate-EDTA |
| TE | - | Tris-EDTA |
| UCSC | - | University of California, Santa Cruz |
| UTR | - | Untranslated Region |
| vcf | - | variant call format |
| *VKORC1* | - | Vitamin K Epoxide Reductase 1 gene |
| VWD | - | von Willebrand Disease |
| VWF | - | von Willebrand Factor |
| W | - | Tryptophan amino acid |
| WES | - | Whole Exome Sequencing |
| WGS | - | Whole Genome Sequencing |
| WFH | - | World Federation of Hemophilia |

## List of Figures

## List of Tables

# Abstract

Rare bleeding disorders (RBDs) are among the oldest described genetic diseases, generally leading to lifelong hemorrhagic complications. These are monogenic in nature and are inherited in Mendelian patterns. The genetic cause of RBDs is the defect(s) in gene(s) coding or regulating various clotting factor(s). RBDs manifest themselves in the form of either severe or moderately severe or mild and have affected approximately 400,000 individuals worldwide. Von Willebrand disease and hemophilia A are the most common type of RBDs. Since the clinical presentations of various types of RBDs intersect with each other, only the laboratory studies may not be sufficient for the accurate diagnosis of the RBDs. Genetic studies are required in such cases. Moreover, genetic studies allow better understanding of the biology of rare bleeding disorders and the genetic information can be used for the translational application, prenatal diagnosis and the detection of carrier status, prediction of development of inhibitors and can also assist in genetic counseling. However, traditional molecular techniques have shown limitations in efficient characterization of mutations causing RBDs. In present era of high through-put sequencing, Next Generation Sequencing (whole genome sequencing and whole exome sequencing) which has emerged as a gold-standard for the identification of disease-causing mutations in various other rare Mendelian diseases has also shown a convincing potential to explore the underlying genetic lesions in the patients with rare bleeding disorders. In our current study whole exome sequencing has been used for the screening of mutations in patients suffering from two rare bleeding disorders viz. Type 2 Normandy von Willebrand disease (2N VWD) and Factor X Deficiency (FXD). Sequencing was performed in Illumina platform (HiSeq 2500). We developed our own bioinformatics analysis pipeline for WES data and ended up with only one causative mutation in both the RBDs following rigorous prioritization of the variants. The causative mutation identified in FXD, c.T212C:p.F71S, which is reported as a founder effect in Algerian population has not yet been reported from the other parts of the world. In case of 2N VWD, the causative mutation identified, c.C2446T:p.R816W is one of a very common variant reported all over the world. Both the causative mutations were validated by capillary sequencing and also the carrier status among the family members was checked. We found two daughters of male patient of 2N VWD are carrier for the disorder.

Key words: rare bleeding disorders, 2N VWD, FXD, whole exome sequencing, bioinformatics analysis of WES data, validation of WES results, detection of carrier status

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Higher organisms possess a highly conserved machinery which tightly regulates the processes of blood clotting, platelets activation and vascular repair; known as hemostatic system (Hoffbrand *et al*., 2016). The human hemostatic system provides a natural balance between procoagulant force (include platelet adhesion and aggregation and fibrin clot formation) and anticoagulant force (include the natural inhibitors of coagulation and fibrinolysis (Kasper *et al*., 2015). Under normal physiological conditions, hemostatic system is regulated to promote blood flow; however, in case of any vascular injury it engages a plethora of vascular and extravascular receptors that act in concert with blood components to seal off the damage inflicted to the vasculature and the surrounding tissue (Versteeg *et al.*, 2013). The first important component that contributes to hemostasis is the coagulation system comprising various plasma proteins, while the second important component starts with platelet activation, which not only contributes to the hemostatic plug, but also accelerates the coagulation system. Coagulation and platelet activation are switched off by blood-borne inhibitors and proteolytic feedback loops. The system remodels the damaged vessel after bleeding is successfully halted and restores normal blood flow (Versteeg *et al.*, 2013).

The events that occur during hemostasis are broadly classified as following:

1. Primary hemostatic response –which comprises vasospasm (Vasoconstriction) and platelet plug formation
2. Secondary hemostatic response –which consists of simultaneous activation of the coagulation system.
3. Tertiary hemostatic response –which includes fibrinolytic system

Hemostasis is accomplished through a network of processes that include the platelet system, coagulation system, and anticoagulant and fibrinolytic pathways, which all support the dynamic equilibrium that provides proper blood flow. Disruption of this well-regulated balance leads to pathologic conditions, such as uncontrolled bleeding and thrombosis. Bleeding disorders can be broadly classified as (Hoffbrand *et al*., 2016):

1. Primary Hemostatic Defects – consist of quantitative and qualitative platelet disorders, von Willebrand Disease (VWD) and
2. Secondary Hemostatic Defects – deficiencies of components of the coagulation cascade. For e.g. Hemophilia A (HA), Hemophilia B (HB), Rare Factor Deficiencies

Inherited deficiencies or dysfunction of procoagulant proteins generally lead to lifelong bleeding disorders. Hemophilia A and B, inherited as X-linked recessive traits, along with von Willebrand disease (VWD) are the most common hereditary hemorrhagic disorders comprising 95% to 97% of all the inherited deficiencies of coagulation factors (Peyvandi *et al.*, 2006).

## 1.2 Rare Bleeding Disorders

It is an umbrella term for a wide range of medical problems that lead to nonstop bleeding following any trauma, surgery and even spontaneously. The condition arises due to impaired blood clotting mechanism. Blood clotting mechanism functions improperly due to defects in gene(s) encoding and/or regulating the various proteins and elements present in the blood plasma generally known as clotting factors. These clotting factors acts in cascade to cease bleeding by forming a stable blood clot at the site of an injury. Thus both quantitative and qualitative deficiencies of such factors exhibit life-long recurrent bleeding episodes.

### 1.2.1 Types

The rare bleeding disorders are categorized into following types based on the affected clotting factor:

I). **Hemophilias** –These are rare hemorrhagic conditions resulted from deficiency of clotting factors VIII (Hemophilia A) or IX (Hemophilia B). These are X-linked recessive disorders and hence are manifested almost exclusively in males. Based on the residual factor concentration each hemophilia can be divided into severe, moderately severe and mild form as shown in Table 1.1.

Table 1.1: Severity classification of Hemophilia (Graw et al., 2005)

| Type | Clotting factor level % activity (IU/ml) | Usual Age of diagnosis | % Prevalence |
|---|---|---|---|
| Severe | <1% (<0.01) | 1st year of life | 40 |
| Moderately severe | 1-5% (0.01-0.05) | Before age 5-6 years | 50 |
| Mild | 6-30% (0.06-0.3) | Often later in life | 10 |

ii). **Rare Factor Deficiencies** –These are called as rare factor deficiencies due to their ultra-rare prevalence in general population. These conditions arise due to deficiency of clotting factors I, II, III, V, VII, X, XI, and XIII. These disorders are autosomal recessive with equal chances of affecting both males and females.

iii). **von Willebrand disease (VWD)** –VWD is one of the most common bleeding disorder resulting from a deficiency or dysfunction of von Willebrand factor (VWF). VWF is a large multimeric glycoprotein which mediates platelet adhesion/aggregation at the site of injury and stabilizes factor VIII (FVIII) in blood circulation (Kasper *et al*., 2015). VWD is also called "Pseudo-Hemophilia" because it is associated with the decreased plasma concentration of FVIII. Based on the mode of deficiency of VWF, VWD is divided into 3 types (James and Goodeve, 2011); type 1- partial quantitative deficiency of VWF, type 3- virtual or complete quantitative deficiency of VWF and type 2- qualitative deficiency of VWF. Type 2 is further divided into 2A, 2B, 2M and 2N.

## 1.2.2 Diagnosis

Diagnosis of bleeding disorders is performed as follows:

### 1.2.2.1 Laboratory Studies

Laboratory studies done for the diagnosis of individuals with bleeding disorders include the following:

i). Platelets count and Platelet Function Analyzer (PFA) –It is done for the differential diagnosis of hemophilia and rare factor deficiencies from platelets related bleeding disorders.

ii). Coagulation screening tests –These include tests like Bleeding Time (BT), Prothrombin Time (PT), Activated Partial Thromboplastin Time (APTT).

iii). Coagulation Factor assay –Bleeding disorders are associated with decrease in concentration or plasma level ($F_c$) of corresponding clotting factors in the blood circulation. Factor assay is essential for diagnosis of quantitative deficiencies of the clotting factors.

There are some tests specific for the factor deficiency. For e.g. Russell Viper Venom Time (RVVT) for FX deficiency (Uprichard and Perry, 2002), Ristocetin Cofactor Activity Assay (VWF:RCo) and Collagen Binding Assay (VWF:CB) for VWD (Goodeve, 2010).

Table 1.2: Summary of coagulation screening tests

| Test | Normal Range | Range in case of bleeding disorders | Interpretation |
|---|---|---|---|
| Platelets count | 150,000 -450,000 /ml blood | Normal | No platelets related bleeding disorders |
| BT | Around 8 min | >10 min | Bleeding disorders |
| PT | 11- 14 sec | >25 sec | Deficiency of factors related to extrinsic pathway- FVII, FV and FX |
| APTT | 25-39 sec | >70 sec | Deficiency of factors related to intrinsic pathway- FIX, FVIII and FX |

BT- bleeding time, PT- prothrombin time, APTT- activated partial thromboplastin time, F- factor

### 1.2.2.2 Molecular Diagnosis

Molecular diagnosis is based on identification of putative mutations in genes encoding corresponding coagulation factors. There are two different approaches to the genetic evaluation of various bleeding disorders (Peyvandi *et al.*, 2006):

i). Linkage analysis –It involves the analysis of single nucleotide polymorphism (SNP) or microsatellite markers in the genes related to bleeding disorders to track the defective chromosome in the family. This means linkage analysis is useful only in familial cases where more than one individuals are affected by same inherited bleeding disorder.

ii). Direct mutation detection –This approach involves the direct identification of the disease causing mutation in the defective gene by sequencing. It is equally efficient and sensitive in detecting mutations in both familial and sporadic cases of bleeding disorders. Hence it is the most popular approach used to unravel the genetic lesions underlying the diseases. We can use two strategies for direct detection of the putative mutations associated with inherited bleeding disorders as follows:

a). Targeted mutation analysis -This strategy utilizes the target regions of the genes such as mutation hotspot for the detection of putative variants. For example in patients of Hemophilia A, almost 50% of the severe cases are reported due to intron 22 inversion of *F8* (Jayandharan *et al.*, 2012). Similarly, in case of type 2 Normandy VWD (2N VWD) almost 85% of the disease causing mutations are reported to be harbored within exons 18-20 of *VWF* (Goodeve, 2010).

b). Mutation scanning or whole gene sequence analysis -Under this approach, the complete gene related to the phenotype is sequenced using standard capillary sequencing approach.

Inherited bleeding disorders generally leads to lifelong hemorrhagic complications. The severity of such disorders is determined by the concentration of corresponding clotting factor in blood plasma. These disorders affects almost 400,000 individuals worldwide and deteriorates their living standards. The severe factor deficiencies in children are more tragic, developing the economical, mental and physical burdens to the family. Many females suffering from rare factor deficiencies like FXD and 2N VWD, undergo miscarriages which develops social burdens to themselves. In such scenario, genotyping of the patients with inherited bleeding disorders to unravel the underlying genetic lesions efficiently, holds great importance. One can make the translational application of the genetic information in genetic characterization of disease causing variants, carrier testing and prenatal diagnosis. Furthermore, information from genotyping can also help in (1) confirming uncertain phenotypical diagnosis such as in the case of multiple factor deficiencies (FV+ FVIII), (2) supplementary genotype-phenotype information such as FVIII mutations and risks of inhibitors formation and (3) solving the dilemma of differential diagnosis (differentiation of phenocopies) as in the case of Hemophilia A and 2N VWD (Peyvandi *et al.,* 2013). Following the cloning and sequencing of genes encoding the coagulation factor proteins in early 1980s (Peyvandi *et al*., 2013), significant progress has been made in the translational application of this genetic information especially in the two most common severe inherited bleeding disorders, hemophilia A and B. Mutations are detected directly by using either targeted mutation analysis or whole sequence analysis of affected gene. The basic strategy includes amplification the gene (exonic and their flanking intronic regions, the 5'UTR and 3'UTR) by PCR followed by detection of mutations using various screening methods example Single-Strand Conformation Polymorphism (SSCP), Denaturing Gradient Gel Electrophoresis (DGGE), Conformation Sensitive Gel Electrophoresis (CSGE), Denaturing High Pressure Liquid Chromatography (DHPLC)

(Peyvandi et al., 2006) and traditional capillary sequencing. Direct mutation detection approach has a near 100% accuracy and is informative in over 95% of families with hemophilia A and almost 100% of families with hemophilia B (Peyvandi *et al.*, 2006). However such genetic analysis techniques have failed to identify the genetic lesions in 5% to 10% of patients affected with severe clotting factor deficiencies. Moreover, in case of VWD, genetic mutations have been identified only in approximately 65% of index cases (Peyvandi *et al*., 2013) and rest remains undetected. Furthermore, in patients with rare factor deficiencies, variable genotype-phenotype correlations is a serious challenge for traditional approaches to decipher the underlying defects accurately.

Genetic testing using traditional molecular diagnostic techniques such as linkage analysis, karyotyping etc. have limitations as it will not be able to cover all the genes responsible for rare genetic diseases (Yang *et al.*, 2013). Therefore certain rare genetic diseases remains undiagnosed due to lack of mutation in the known gene. In case of rare bleeding disorders, 5% to 10% of total patients lack the mutation in the known gene which increase the demand of new molecular diagnostic tool in clinical setups (Peyvandi *et al.*, 2013). The lack of accurate diagnosis can have considerable adverse effects for patients and their families, including failure to identify potential treatments, failure to recognize the risk of recurrence in subsequent generations, and failure to provide anticipatory guidance and prognosis. In present era of high through-put sequencing, NGS (whole genome sequencing and whole exome sequencing) has emerged as a gold-standard for the identification of disease-causing mutations in rare Mendelian diseases (Bamshad *et al*., 2011; Shin *et al*., 2014). NGS has shown a convincing potential to explore the underlying genetic lesions in the patients with rare bleeding disorders (Peyvandi *et al*., 2013) which in near future may have translational applications.

## 1.3 Lacunae of the Study

Inherited bleeding disorders are among the oldest described genetic diseases, however no molecular studies regarding such diseases are done till date in Nepal. Government till now does not have facility for genomic diagnosis and mutational database for prevention, treatment and management of hemophilia and its patients. Poor socio-economic status of Nepal plays an additive role that prevents the disease to identify early and go for the further treatment procedures. There are requirements of sophisticated molecular diagnostic tools and effective government policies to address the issues of rare bleeding disorders in Nepalese population.

## 1.4 Rationale

This molecular genetics study of the related genes will allow better understanding of rare bleeding disorders and their diagnosis. In addition, the detection of carrier status, prediction of development of inhibitors and genetic counseling can be done. The present research study will also provide deeper insights for the disease pattern which is shared by the patients with similar phenotypic complications. This kind of mutational

characterization and sequence variation structure will contribute in the field of pharmacogenomics.

## 1.5 Hypothesis

Rare bleeding disorders, especially those which are autosomal recessive, are reported to have a founder effect. Certain mutations are reported to be confined within a specific ethnic community. In other word they exhibit identity by descent phenomenon. In such background, because of the genetic and population diversity of Nepalese community, we assume and thus hypothesize that mutations causing various rare bleeding disorders have a founder effect. This means de novo mutation may exist in this research.

## 1.6 Objective

### 1.6.1 General Objectives

1. To screen the DNA mutations in genes related to rare bleeding disorders in patients of Nepalese origin by Whole Exome Sequencing
2. To validate the results of Whole Exome Sequencing by capillary sequencing.

### 1.6.2 Specific Objectives

1. To identify causative mutations in patients with rare bleeding disorders of Nepalese origin.
2. To relate type of the mutation and severity of the bleeding disorder.
3. To identify carrier status of family members of the affected patients.

# CHAPTER 2

# REVIEW OF LITERATURE

## 2.1 Historical Background of Rare Bleeding Disorders

*"The history of haemophilia shows the human mind attempting to define and encompass a mysterious yet fascinating phenomenon"* G. I. C. Ingram, 1976

Incidences of excessive or abnormal bleeding what may have been human rare bleeding disorders are found to be recorded hundreds of years ago. The earliest written reference suggestive of hemophilia dates back to the 2nd century AD (Ingram, 1976). The Babylonian Talmud, a collection of Jewish rabbinical writings on laws and traditions, stated that male boys should not be circumcised if two brothers had already died owing to excessive bleeding from the procedure. In the 10th century Al-Zahrawi (also known as Abulcasis), an Arabian physician, described a family where males died of excessive bleeding following even a trivial injury (Kaadan and Angrini, 2010). However the first modern description of hemophilia appeared in 1803, when a Philadelphian physician, Dr. John Conrad Otto, described an inheritable bleeding disorder in several families in which only males – who he called "bleeders"– were affected. Otto noted that disorder was transmitted by unaffected females to a proportion of their sons (Schramm, 2014). To describe such hemorrhagic disorders various names such as *haemorrhoea, idiosyncrasia haemorrhagica, haematophilia, bleeding disease, hereditary haemorrhagic diathesis* etc. were used (Ingram, 1976). The recent, rather strange name *"haemophilia"* which means 'love of blood', is first appeared in the dissertation of Friedrich Hopff, student at University of Zurich, Switzerland, in 1828. His supervisor, Dr. Johann Lukas Schönlein, was probably not in favour of the term "*haemophilia*" as he himself preferred to use the term "*haemorraphilia*" which means 'affinity to bleed'. Descriptions regarding the genetics of hemophilia were published by Nasse for the first time in 1820. It culminated in Nasse's law, which states that hemophilia is transmitted entirely by unaffected females to their sons (Schramm, 2014). In 1926 Dr. Eric von Willebrand, a Finnish physician, described a hereditary bleeding disorder affecting both the sexes equally and called it "*hereditary pseudohemophilia"*, with characteristic features suggesting that this was distinct from Hemophilia (James and Lillicrap, 2013). It was later named by the name of the discoverer as von Willebrand disease (Kaushansky *et al*., 2015). The clotting defect associated with hemophilia was first believed to be due to calcium deficiency or fragility of vessels or a platelet defect(s). It was in 1936, when two Harvard doctors viz. Patek and Taylor found that the clotting defect could be corrected by adding a fraction precipitated from normal plasma, which they called "*antihemophilic globulin"(*AHG*)* (Ingram, 1976). While the pathophysiology of hemophilia was becoming increasingly clear, doubts were raised as to whether or not hemophilia was a single entity. In 1947, Pavlosky from Buenos Aires, Argentina observed that transfusion of one hemophiliac with the blood from another hemophiliac, temporarily but completely, normalized the clotting time of the recipient (Franchini and Mannucci, 2014). This suggested the existence of at least two different

types of hemophilias. Eventually Biggs et al. from Oxford in 1952 established a disease entity that was different to already identified hemophilia, and called "*Christmas disease*" (Schramm, 2014). Subsequently, the term hemophilia A was proposed for the more common form associated with FVIII deficiency, while hemophilia B was proposed for the less common type (Christmas disease) which was later identified to be associated with FIX deficiency (Franchini and Mannucci, 2014). Factor I deficiency was first described in 1920. Factors II and V deficiency were identified in the 1940s. The 1950s saw an explosion of work on rare factor deficiencies, as deficiencies of FVII, XI and XII were first recognized. Factor X was originally reported in the late 1950s as the "Stuart-Prower Factor," named after the first two identified factor X–deficient patients. In 1960, Duckert described patients who had a bleeding disorder and characteristic delayed wound healing. This fibrin stabilizing factor was called factor XIII (Kaadan and Angrini, 2010).

Hemophilia has often been called "*the royal disease*" – the disease of kings, as it affected several members of the European royal families through the descendants of Queen Victoria (1837–1901), queen of England. She was a carrier of hemophilia and through her two daughters, Alice and Beatrice, hemophilia spread to the royal families in Germany, Spain and Russia as shown in Fig. 2.1. Her son Leopold – Duke of Albany – had frequent bleeds and died of a brain hemorrhage following a minor injury at the age of 31 years. Alexandra – Alice's daughter – married the Tsar of Russia Nicholas and was the mother of Alexis – the Tsarevich – who is one of the most famous as well as one of the most tragic hemophiliac.

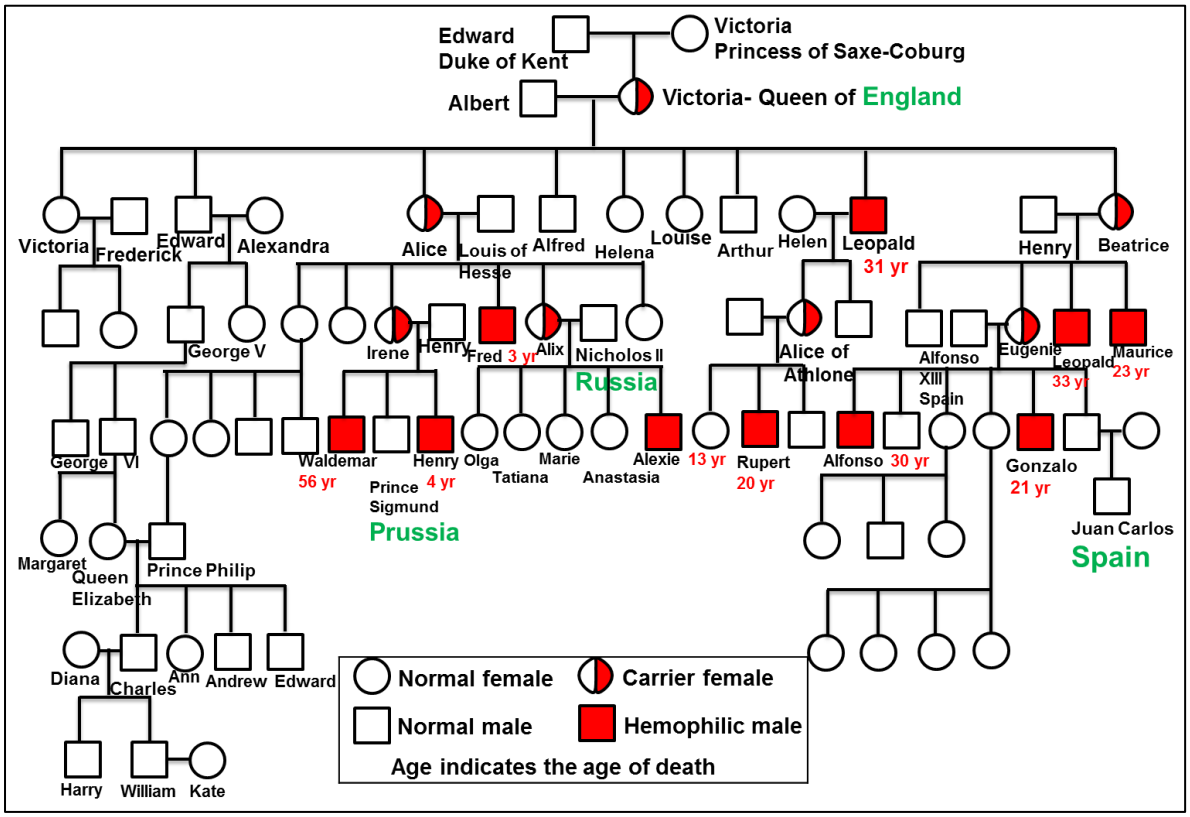


**Fig. 2.1: Pedigree of Queen Victoria** depicting the spread of hemophilia in European royal families through her descendants.

Regarding the treatment of these rare bleeding disorders; as hemophilia being often associated with fatal bleeding it is unsurprising that hemophilia became closely linked with transfusion medicine. Schönlein in 1832 proposed the use of blood transfusion (Schramm, 2014). However even allogeneic blood transfusions were reported to have "disadvantageous outcomes" in about half of recipients. Addis in 1911 prepared a very crude fraction of plasma by its acidification. This is regarded as the first "*antihemophilic factor*". In 1934 Robert Macfarlane showed that the venom of Russell's Viper clotted hemophilic blood rapidly and so was regarded eminently suitable as a local application (Ingram, 1976). It became commercially available as *"Stypen*". In 1946, Cohn et al described the Cohn Fraction, which was found to possess antihaemophilic activity besides fibrinogen (Schramm, 2014). This lead to the development of human preparations of factor VIII in 1950s. Judith Pool in 1965 reported that on slowly thawing frozen plasma, much of the FVIII activity remained within the fibrinogen "sludge" that was slow to re-dissolve. This so-called "*cryoprecipitate"* could be spun down and so re-frozen for storage (Ingram, 1976). This discovery of cryoprecipitate revolutionized the treatment of hemophilia. However, it was in 1960s that the introduction of FVIII and FIX concentrates derived from plasma made the clotting factor replacement therapy possible (Schramm, 2014). The introduction of pasteurized FVIII/VWF concentrate Haemate® P (in 1981), Beriate® P (in 1990) and later recombinant FVIII products (in 2004) resulted in a dramatic increase in both the quality of life and life expectancy of patients of hemophilia A. The discovery of desmopressin in 1977 provided a new, inexpensive and safe way of treating patients with certain types of VWD and mild hemophilia A. Recombinant concentrates were introduced in 1992 which enabled "mass production" of clotting factor concentrates.

## 2.2 Pathophysiology of Blood Clot Formation

Hemostasis in mammals is maintained by the combined action of clotting factors, platelets and fibrinolytic agents. Under normal physiological condition, any damage to blood vessels leads to blood clot formation. In response to injury, platelets begin to (1) adhere to substances such as collagen and von Willebrand factor (VWF) outside the interrupted endothelium, (2) activate to release chemicals such as thromboxane A2 (TXA2) and (3) aggregate to form platelet plug at the site of injury. VWF plays a crucial role on adhesion and aggregation of platelets to form platelets plug. This plug is stabilized by a fibrin mesh formed through cascade of reactions involving various clotting factors. There are two different pathway namely extrinsic and intrinsic pathways which occur simultaneously, leading to formation of cross linked fibrin polymer as shown in Fig 2.2.

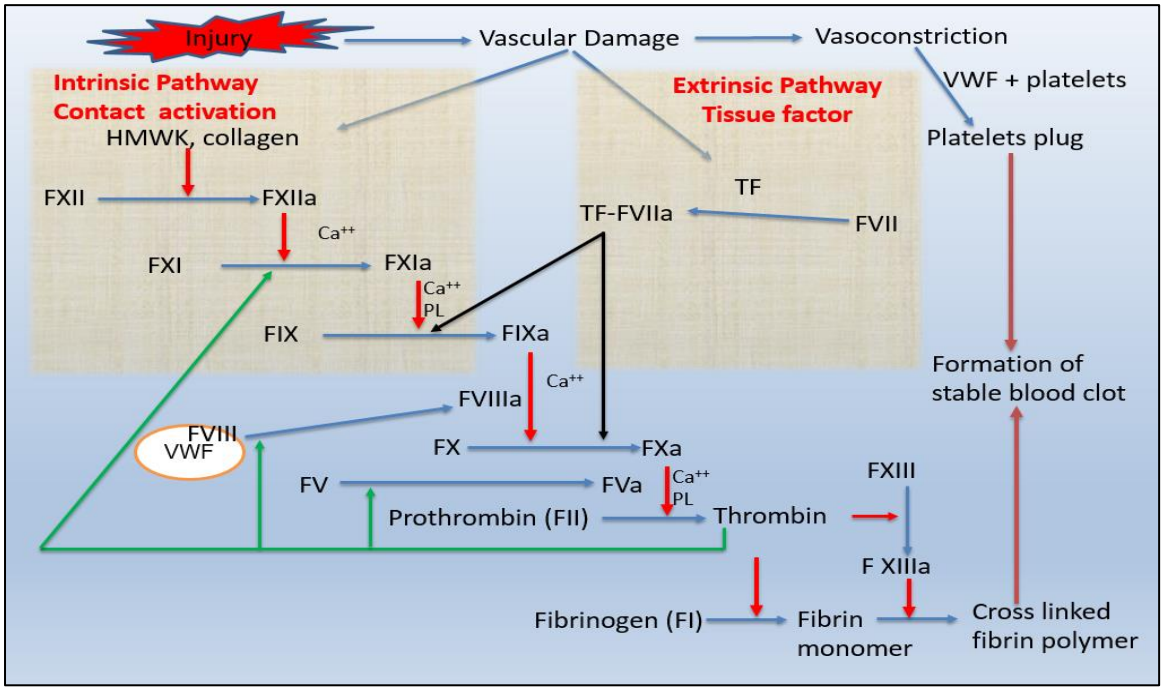


**Fig 2.2: Mechanism of blood coagulation process.** TF- Tissue factor, F- Factor, VWF- von Willebrand factor, HMWK- high molecular weight kininogen, a- activated form of corresponding factor, $Ca^{++}$-Calcium ion, PL- phospholipid. Adopted from (Kasper *et al.*, 2015)

Extrinsic pathway requires an external factor (Tissue Factor from subendothelial cells like smooth muscle cells and fibroblasts) to generate thrombin burst essential positive feed-back. The process initiates with the exposure of TF which activates factor VII to FVIIa and binds to its activated form. TF-VIIa complex then activates FIX to FIXa and FX to FXa. On the other hand, intrinsic pathway utilizes components present in the blood. This pathway begins with formation of complex on exposed collagen by plasma proteins like high molecular weight kininogen (HMWK), prekallikrein, and FXII. Once FXII is auto-activated, it activates FXI. FXIa in turn activates FIX in the presence of $Ca^{++}$ as shown in Fig 2.2. Activated FIX (FIXa) combines with VIIIa to form tenase complex which is essential for activating FX. Once activated, FXa forms a complex with its cofactor FVa known as prothrombinase (FXa-FVa) complex which converts prothrombin to active thrombin. Thrombin finally results in the formation of cross linked fibrin polymer as shown in Fig 2.2.

However, there is also a new concept of coagulation. According to this new concept, upon vascular damage platelets adhere to the damaged site and aggregate through interactions with exposed endothelium's components namely collagen and von Willebrand factor (VWF) and forms a platelets' plug at the site of the injury. The TF-VIIa complex enables subsequent activation of factor X and prothrombin, which results in the formation of small amount of thrombin. This slowly accumulating thrombin will further activates Factor VIII (FVIIIa) and Factor V (FVa) along with factor XI (FXIa). Thus activated molecules will amplify the production of thrombin as shown in Fig.2.2. The tenase complex (FVIIIa-FIXa) catalyzes the conversion of FX to FXa, which in turn interacts with FVa and forms the prothrombinase complex (FXa-FVa). FXa-FVa complex produces large amounts of

thrombin which subsequently forms fibrin fibers from fibrinogen (FI). Finally these fibers forms a crosslinked fibrin polymer yielding a mesh like structure which stabilizes the platelets plug formed at the site of the injury as shown in Fig.2.2 (Versteeg *et al.*, 2013).

Since coagulation process occurs in a sequential manner as depicted in Fig 2.2, defects in synthesis of any of the clotting factors affect the functioning of downstream clotting factors which ultimately impairs the efficient blood clot formation and results in nonstop bleeding. Inherited deficiencies of these factors generally lead to lifelong bleeding disorders, whose severity is inversely proportional to the level of deficient factor. Defects of FVIII, FIX and VWF are among the commonest bleeding disorders representing 95% to 97% of all the inherited deficiencies of coagulation factors (Peyvandi *et al*., 2006).

The plasma concentration and half-life of clotting factors differs from one another. Factor VIII is among those clotting factors which have very low concentration in blood plasma. FII, FX are generally present for longer time in plasma due to relatively high half-life as shown in Table 2.1.

Table 2.1: Plasma concentration and half-life of various clotting factors (Hoffbrand *et al*., 2016) (Kasper *et al.*, 2015)

| **Clotting Factor** | **Half-life, $T_{1/2}$ (hr)** | **Normal Plasma Concentration (µg/ml)** | **Minimum Hemostatic level** |
|---|---|---|---|
| Prothrombin (FII) | 65 | 90 | 20-30% |
| Factor V | 15 | 10 | 15-20% |
| Factor VII | 3 | 0.5 | 15-20% |
| Factor VIII | 10 | 0.1 | 30% |
| Factor IX | 25 | 5 | 30% |
| Factor X | 40 | 8 | 15-20% |
| VWF | 12 | 10 | 30% |

## 2.3 Characteristics of Rare Bleeding Disorders

Various rare bleeding disorders viz. hemophilias, rare factor deficiencies and von Willebrand disease differ significantly among themselves in different parameters such as mode of inheritance, age-sex-race based prevalence, clinical presentations etc. Characteristics features of major rare bleeding disorders are tabulated in Table 2.2.

Table 2.2: Characteristics of rare bleeding disorders (Lee *et al.*, 2014), (Goodeve, 2010), (Peyvandi *et al.,* 2013)

| **Bleeding Disorder** | **Factor/protein affected** | **Mode of inheritance / Gene locus** | **Gene, size (kb) & exons** | **Prevalence /$10^6$** |
|---|---|---|---|---|
| HA | FVIII | X-linked recessive, Xq28 | *F8,* 186 kb, 26 | 100 |
| HB | FIX | X-linked recessive, Xq27 | *F9,* 34kb, 8 | 17 |
| Factor XI deficiency | FXI | AR, 4q35.2 | *F11,* 23kb, 15 | 1 |

| | | | | |
|---|---|---|---|---|
| Factor II deficiency | FII | AR, 11p11-q12 | *F2,* 21kb, 14 | 0.5 |
| Factor V deficiency | FV | AR, 1q24.2 | *F5,* 80kb, 25 | 1 |
| Factor X deficiency | FX | AR, 13q34 | *F10,* 27kb, 8 | 1 |
| VWD | VWF | 1,2A,2B,2M-AD; 2N, 3-AR, 12p13.3 | *VWF,*178 kb ,52 | 100 |

HA- Hemophilia A, HB- Hemophilia B, VWD- von Willebrand disease, F- Factor, VWF- von Willebrand factor, AR- Autosomal Recessive, AD- Autosomal Dominant.

### 2.3.1 Epidemiology

The prevalence of various bleeding disorders differs significantly as shown in Table 2.2. In general, von Willebrand disease (VWD) and hemophilia A are the most common bleeding disorders. These include 95% to 97% of all the inherited deficiencies of coagulation factors (Peyvandi *et al*., 2006). Although both the hemophilias are equally prevalent among various races and ethnic groups all over the world, rare factor deficiencies are more prevalent among the races where consanguineous marriages are common. For example factor X deficiency is 10 times more common in Iran (Akhavan *et al.*, 2007). The current worldwide incidence of bleeding disorders is estimated as more than 400,000 (National Hemophilia Foundation, USA, 2015). In the context of Nepal, approximately 3000 patients with RBDs are expected out of which only 536 cases are reported till date (Nepal Hemophilia Society, Nepal, 2015).

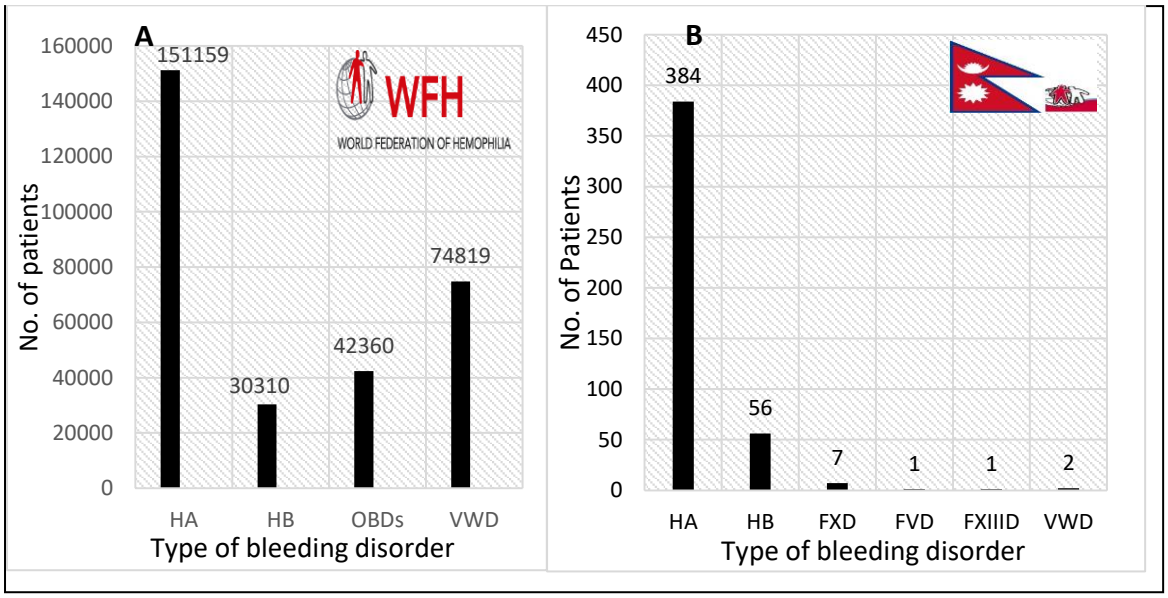


**Fig 2.3: Graph representing total number of patients with bleeding disorders.** (A) Plot of number of patients with bleeding disorders worldwide. Hemophilia A shows the highest prevalence followed by VWD. Other bleeding disorders (OBDs) includes rare factor deficiencies. (Source: Report on annual survey of World Federation of Hemophilia (WFH), 2015) (B) Plot of number of patients with bleeding disorders in Nepal. Hemophilia A shows the highest prevalence followed by Hemophilia B. Rare factors deficiencies shows vey less

prevalence. (Source: Nepal Hemophilia Society (NHS) report, 2015). HA- Hemophilia A, HB-Hemophilia B, OBDs-other bleeding disorders, VWD-von Willebrand disease, FXD-Factor X deficiency, FVD-Factor V deficiency, FXIIID- Factor XIII deficiency.

### 2.3.2 Clinical Presentation

The presentation of the rare bleeding disorders vary accordingly to their types. In general severity and frequency of bleeding are inversely correlated with the residual factor level in blood plasma as shown in Table 1.1. Also the type and site of mutation causes the variation in severity. Patients with mild bleeding disorders may present with easy bruising, inadequate clotting of traumatic injury whereas those with of severe bleeding disorders may manifest spontaneous hemorrhage. The various clinical presentations of rare bleeding disorders are as follows:

- Weakness, orthostasis, tachycardia, tachypnea
- Excessive bleeding even following a minor injury, cut.
- Uncontrolled bleeding during tooth extraction, surgery, circumcision.
- Musculoskeletal bleeding causing tingling, cracking, warmth, pain, stiffness, and refusal to use joint (children), hemarthroses and pseudotumors in joints and hematomas in muscles.
- Intercranial bleeding causing headache, stiff neck, vomiting, lethargy, irritability, and spinal cord syndromes
- Mucocutaneous bleeding such as gum bleeding, epistaxis, are more common in von Willebrand disease.
- Genitourinary bleeding causing hematuria, renal colic, post circumcision bleeding and menorrhagia in females.
- Gastrointestinal bleeding causing hematemesis, melena, frank red blood per rectum, and abdominal pain.
- Other bleeding – oral mucosal hemorrhage, hemoptysis, dyspnea (hematoma leading to airway obstruction), compartment syndrome symptoms, and contusions.

Note: Epistaxis is defined as more than five periods occurring spontaneously and lasting for more than 10 min. Menorrhagia is defined as menstrual periods lasting for a minimum of 6 days and requiring hormone therapy or causing iron deficiency. Oral bleeds are identified if they lasted for more than 10 min or required an intervention of an oral surgeon (Peyvandi *et al.,* 2002).Menorrhagia is also best defined quantitatively as a loss of >80 mL of blood per cycle, based on the quantity of blood loss required to produce iron-deficiency anemia (Kasper *et al.*, 2015).

### 2.3.3 Genetics

Rare bleeding disorders are monogenic in nature and follow Mendelian pattern of inheritance. Hemophilia A (HA) and Hemophilia B (HB are X-linked recessive disorders whereas rare factor deficiencies are autosomal recessive. VWD can be inherited differently

based on its types as shown in Table 2.2. About 70% of the rare bleeding disorders are familial whereas 30% are sporadic cases resulting from de novo mutations.

Both the hemophilias (A and B) are X-linked recessive disorders in which males are affected and females are typically asymptomatic carriers (Bowen 2006). The female carrier transmits the disorder to half their sons and the carrier state to half her daughters. All female offsprings born to a hemophilic father are obligatory carriers. Hemophilia A and B are caused due to defect on the genes encoding clotting factor VIII – *F8* and clotting factor IX – *F9* respectively. Both the human *F8* and *F9* map to the long arm of X chromosome at Xq28 and Xq27 respectively, separated by 35cM (Jayandharan *et al*., 2012). *F8* comprises 26 exons spanning 186 kb whereas *F9* consists of 8 exons spanning 34 kb. Mutations in both these genes (*F8*, N~ 2932; *F9*, n~ 1134) (https://www.cdc.gov/ncbddd /hemophilia/ champs.html) include a variety of deletions, insertions, missense, nonsense, and splice-site mutations. Apart these, intron 1 and intron 22 inversions, which occurs due to homologous recombination during crossing over, in *F8* have been reported to cause the clinical phenotype in approximately 50% severe hemophilia A. This benefits an easier and economic process for the identification of the genetic cause. Steve Sommar in 1998 devised a long range PCR protocol (Liu *et al*., 1998) which was later modified by Bagnall *et al*. for the identification of intron 22 inversion (Bagnall *et al*., 2006). Similarly, Rossetti *et al.* has also devised an inverse PCR protocol for this (Rossetti *et al*., 2005).

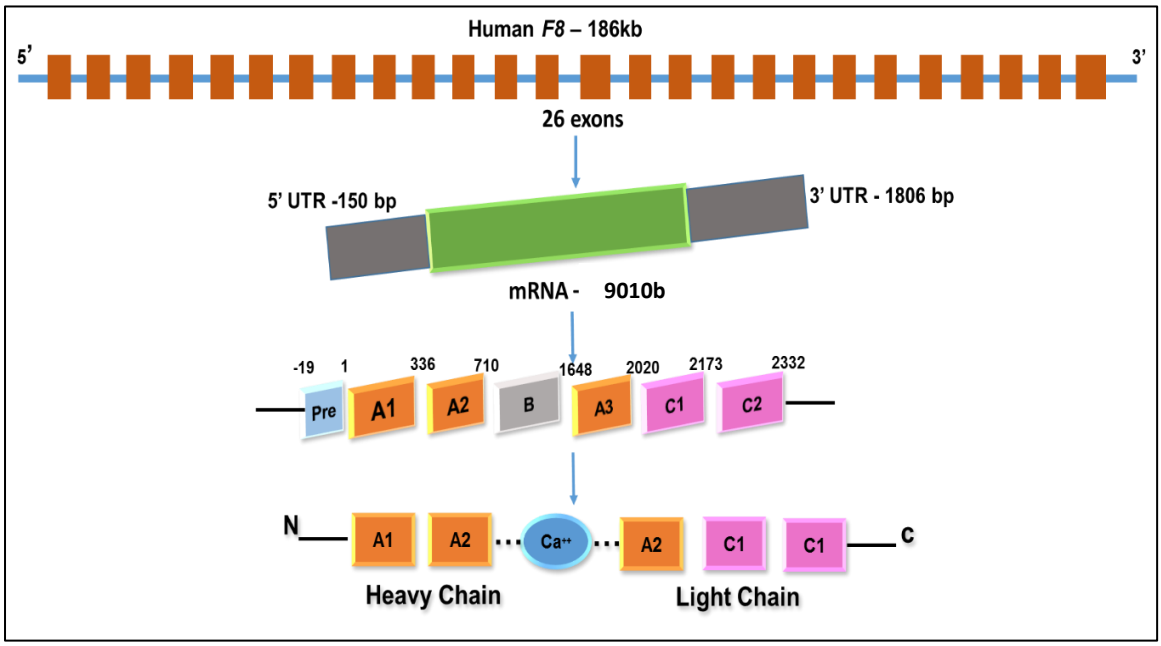


**Fig. 2.4: Schematic of the human factor VIII gene (*F8*), the messenger RNA, and the protein:** *F8* is 186 kb in length and encodes a messenger RNA of ~9 kb. The newly synthesized factor VIII protein consists of a pre-sequence of 19 amino acids and a mature peptide of 2332 amino acids. The mature multidomain factor VIII protein contains triplicated A domains, duplicated C domains, and a single B domain. The arginine residues, which are the sites for proteolytic activation, are R372, R740 and R1689. Activated factor VIII is a heterotrimer in which the dimeric N-terminal heavy chain (740-1648 aa) is held together with the monomeric C-terminal light chain (684 aa) by a metal ion bridge ($Ca^{++}$)

(Thim *et al.*, 2010). In circulation FVIII is stabilized and protected from proteolysis by von Willebrand factor's binding (Jayandharan *et al.*, 2012).

Human clotting factors II, VII, IX, X, protein C and protein S are vitamin K dependent serine proteases. All these factors share a significant homology in genomic organization as well as in structure (Kasper *et al.*, 2015). They possess a similar domain structure of a C-terminal serine protease domain and an N-terminal γ-carboxy glutamic acid (GLA) domain, which are connected by two epidermal growth factor (EGF)-like domains or kringle domains. The GLA domain indicates to the 42-residues region that nests 9 to 12 glutamic acid residues which are post translationally γ-carboxylated into GLA residues. Vitamin K is cofactor for carboxylation of the gamma carbon of the glutamic acid residues in the vitamin K–dependent factors, a critical step for calcium and phospholipid binding of these proteins (Kaushansky *et al*., 2015). A typical example; factor X gene (*F10)* comprises 8 exons spreading over 26731 nucleotides. Each exon encodes for a specific domain of the FX protein. Exon 1 encodes the signal peptide, exon 2 encodes the propeptide and γ-carboxyglutamic acid rich (GLA) domain, exon 3 encodes a short linking segment of aromatic amino acid residues termed the "aromatic stack", exons 4 and 5 encode regions homologous to epidermal growth factor (EGF), exon 6 encodes the activation peptide at the amino terminus of the heavy chain and exons 7 and 8 together encode the active serine protease domain (Brown and Kouides, 2008) as shown in Fig. 2.5.

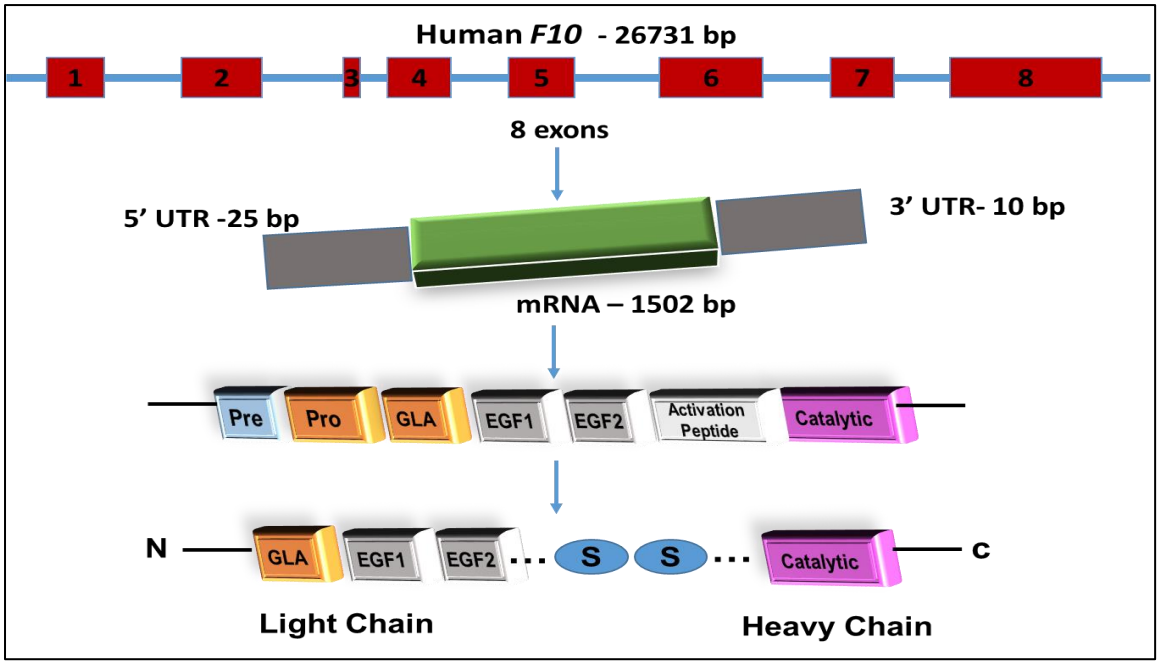


**Fig. 2.5: Schematic of the human factor X gene (*F10*), the messenger RNA, and the protein:** FX is synthesized as a single-chain precursor and is activated by the excision of tribasic peptide Arg140-Lys141-Arg142 resulting into two–chain disulphide-linked (Cys132- Cys302) heterodimer. The light chain is of 139 amino acids (residues 41-179/MW 17 kDa) comprising comprises the GLA domain with 11 GLA residues and the EGF domains; and a heavy chain is of 306 amino acids (residues 183-488/MW 45 kDa) that comprises a 52-residue activation peptide and the serine protease domain.

The genetic cause for von Willebrand disease (VWD) is the genetic lesions in the gene which codes Von Willebrand Factor (VWF) viz. *VWF* (James and Lillicrap, 2013). VWF, an adhesive glycoprotein synthesized by endothelial cells and megakaryocytes, mediates platelet adhesion/aggregation and stabilizes factor VIII (FVIII) in the circulation (Lenting *et al*., 2012; Kasper *et al*., 2015). Patients who lacks VWF either quantitatively or qualitatively manifest either primary (originating from defective formation of platelet-rich thrombi) or secondary (deficiency of FVIII due to its rapid clearance) defect of hemostatic system (Kasper *et al*., 2015).

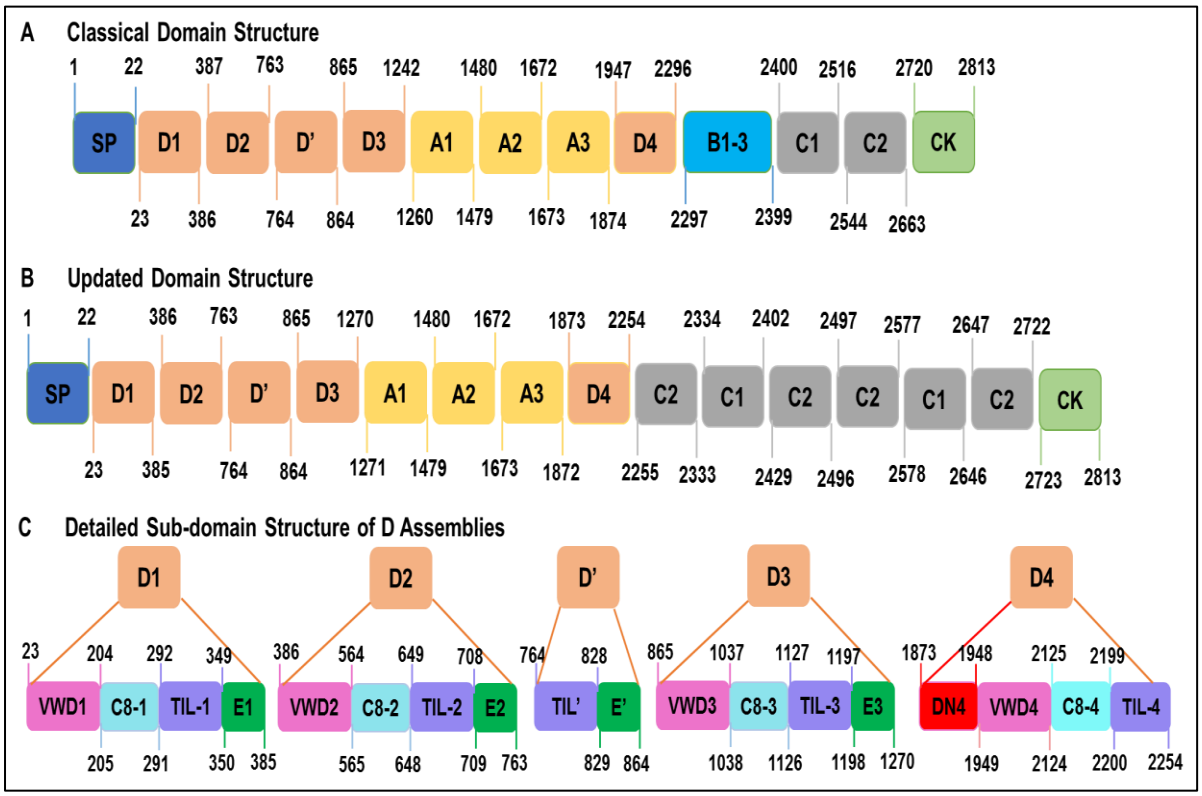


**Fig.2.6: Schematic representation of the old and new domain arrangement of VWF.** (A) Arrangement of 5 different domain according to the original analysis of the VWF sequence. (B) Domain organization as has recently been proposed by Zhou et al. (C) The D1, D2 and D3 domains each contain a VW-domain, a trypsin inhibitor-like (TIL)-structure, a C8 fold and an E module. The D' region lacks the VW domain and TIL-structure. The D4 domain lacks the E module, but instead comprises a unique sequence designated D4N (Rauch and Lenting, 2013).

*VWF* located at 12p13.3, comprises 52 exons spreading over 178 kb of genomic sequence and is transcribed into an 8.8 kb mRNA (Goodeve, 2010). *VWF* has a partial pseudogene (*VWFP*) present on chromosome 22 (22q11.22–11.23) with 97% sequence similarity to that of exons 23–34 of *VWF* which complicates the molecular analysis (James and Lillicrap, 2013). VWF is produced as a single chain pre-pro-protein consisting of a 22-amino acid (aa) signal peptide, a 741-aa propeptide and a mature subunit of 2050 aa (Sadler, 1998). It is a mosaic protein comprising different types of domains as shown in Fig.2.6, each having specific functions. The D'-D3 domains are essential in binding FVIII to VWF  The A1

domain is essential in binding VWF to platelets through the platelet receptor glycoprotein GPIbα (Zhou *et al.*, 2012). Thus, mutations within these domains causes significant loss in the FVIII levels presenting a rare hemostatic disorder named 2N VWD. Similarly, A2 domain contains the cleavage site for post-secretion processing of VWF by a protease named ADAMTS13 (A Disintegrin And Metalloproteinase with a Thrombospondin Type 1 Motif, Member 13) (Goodeve, 2010; Haberichter, 2015). The C4 domain contains an Arg-Gly-Asp-(RGD) sequence that recognize integrin αIIbβ3 which is a platelet receptor (Zhou *et al*., 2012). The VWF-αIIbβ3 interaction involves in the enforcement of platelet-platelet interactions (Rauch and Lenting, 2013). Through A3 domain VWF binds to collagen in exposed subendothelium which results in exposure of GpIbα binding sites in the A1 domain. This initiates platelet tethering at sites of vascular damage. The mature VWF protein exists as a heterologous series of covalently-linked mature subunits ranging from dimers to large polymers consisting of over 40 subunits (Rauch and Lenting, 2013).

## 2.4 Next Generation Sequencing (NGS) and Whole Exome Sequencing (WES)

Next generation sequencing, also known as deep sequencing (Behjati and Tarpey, 2013), in simple, is a catch-all term for all those sequencing platforms that are able to perform massively parallel sequencing (up to millions and even billions reads at a time) with an immensely high throughput in gigabase (Gb) or even terabase (Tb) scale (Rizzo and Buck, 2012). Although the advent of NGS has open the door for sequencing the whole genome of organisms, whole genome sequencing (WGS) is still not feasible due to the time, cost (Biesecker *et al.*, 2011) and data (Boycott *et al.*, 2013) constraints. In fact, only less than 10% (~3 Mb) of the human whole-genome sequence is characterized (Rabbani *et al.*, 2014). In such scenario targeted sequencing, for example whole exome sequencing, has evolved as a boon in the field of genetics.

Human genome is composed of over $3.3x10^{10}$ bases among which only a minuscule proportion, ~1% ($3x10^7$, 30 Mb), has protein coding potential which are called exons (Rabbani *et al*., 2014). However, this minuscule proportion harbors approximately 85% of the disease-causing mutations (Choi *et al*., 2009). This implies an efficient strategies for selectively sequencing complete coding regions (i.e., ‘‘whole exome’’) have a huge potential to contribute to the understanding the biology of rare and common human diseases. Targeted sequencing of the complete exons set of human (exome) is known as Whole Exome Sequencing (WES). WES is already emerged as a powerful and cost-effective tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies (Bamshad *et al*., 2011). That’s why sequencing of the complete coding regions (exome), exon flanking regions as well as conserved (3’UTR, 5’UTR) regions of the gene has the potential to uncover the causes of large number of rare, mostly monogenic, genetic disorders as well as predisposing variants in common diseases and cancers (Rabbani *et al*., 2014). In addition, exomes are ideal to help us understand high-penetrance allelic variation and its relationship to phenotype and

most importantly whole exome is 1/6$^{th}$ the cost of whole genome and 1/15$^{th}$ the amount of data (Biesecker *et al.*, 2011).

Different sequencing platforms, for example Illumina, Pacific Bioscience, Ion torrents etc. have their own protocols for whole exome sequencing (Mardis, 2008). However, the core basic is same for all. The basic steps in WES involves lysing cells to extract DNA, library preparation, cluster generation, high through-put massively parallel sequencing and computational data analysis (Metzker, 2010) (Shendure and Ji, 2008). Library preparation involves optimal fragmentation of the DNA, end repair of the fragments, adapter ligation, PCR enrichment, probes (baits against exome) hybridization and capture and finally PCR enrichment and size selection (Bamshad *et al*., 2011; Metzker, 2010). Those captured exome are then clonally amplified to generate millions of clusters which are then subjected for sequencing. The raw sequence is then processed through various quality check softwares such as FASTQC and Trimmomatic (Rimmer *et al.,* 2014). The quality passed data are then aligned against reference genome and variants are called and annotated. Finally the causal mutation is identified among the numerous variations by adopting different approaches such as removing common variations using public databases, focusing on deleterious variants (Ku *et al.*, 2012), predicting the functional effect of the variants by in silico tools like SIFT scores and Polyphen score, allele frequency of less than 0.01% etc. (Scaria and Sivasubbu, 2015). The causative mutation is then validated by capillary sequencing and its inheritance pattern is observed among the family members of the index case.

# CHAPTER 3

# MATERIALS & METHODOLOGY

## 3.1 Samples' Selection Criteria

We recruited patients registered as Hemophiliacs under Nepal Hemophilia Society (NHS), Bagdole, Lalitpur, Nepal who were included or excluded in this research study based on the following criteria:

### 3.1.1 Inclusion Criteria

The inclusion criteria were as follows:

- Individuals registered as hemophiliacs under Nepal Hemophilia Society (NHS), Bagdole, Nepal.
- Individuals suffering from severe bleeding disorders,
- Familial cases of severe bleeding disorders,
- Nepalese origin with different ethnicity, varying disorder and severity
- Individuals willing to give informed consent.

### 3.1.2 Exclusion Criteria

- Sporadic cases of rare bleeding disorders and
- Individuals not willing to give informed consent.

Considering the above mentioned criteria a familial case of Factor X Deficiency (FXD) and a familial case of Type 2 Normandy von Willebrand Disease (2N VWD) were selected. A written informed consent was obtained from all the participants and this genetic research had also been approved to carry out by Nepal Health Research Council, Kathmandu, Nepal.

## 3.2 Sample Collection

Approximately 5mL peripheral blood was drawn from antecubital vein of each patient and was collected into an EDTA-vacutainer tubes. Those tubes were shipped to CSIR- Institute of Genomics and Integrative Biology (CSIR-IGIB), New Delhi, India where further work was carried out.

## 3.3 Genomic DNA Extraction

Genomic DNA of the samples were extracted from their blood using Salting Out method (Miller *et al.*, 1988).

1. 2-5 ml of EDTA treated blood was transferred to well-labelled 15 mL tube, to which ~10mL of RBC Lysis Buffer (RLB) was added.
2. The mixture was inverted several times until it became translucent.
3. Then it was incubated at room temperature for about 20 minutes on a shaker (70rpm).
4. Following the incubation the tubes were then centrifuged at 2500 rpm for 10 mins at room temperature (RT).
5. The supernatant was discarded in 20% hypochlorite solution.

6. 15 ml of RLB was added to the pellet and mixed by brief vortexing and pellet was broken down thoroughly.
7. The samples were incubated again at RT for about 20 minutes on a shaker (70rpm) and re-centrifuged at 1000 rpm for 10 min at RT.
8. The supernatant was discarded in 20% hypochlorite solution.
9. The pellet was dissolved in 4 ml of Nucleus lysis buffer (NLB) by vortexing and 85 µl of 20% SDS and 55 µl of proteinase-k (20mg/ml) were added and mixed well.
10. The tubes were then incubated at $37^0$C temperature overnight in a water bath.
11. After overnight incubation 1.5 ml of 6M saturated NaCl solution was added, shaken vigorously and centrifuged at RT at 3500 rpm for 35 min.
12. The supernatant was transferred carefully without disturbing pellet into new 15mL tube.
13. The double volume of absolute ethanol (i.e. 100%) was added slowly to precipitate the DNA.
14. The precipitated DNA forms a globular white pellet or thread like structure which was transferred to 1.5mL micro centrifuge tube.
15. The pellet was washed with 70% ethanol and pelleted down at 13000 rpm for 10 minutes.
16. The supernatant was discarded and pellet was subjected to air dry at $37^0$C.
17. Later it was re-suspended in 200 µl of Tris-EDTA (TE) buffer and kept for dissolving by incubating overnight at $37^0$C.

## 3.4 DNA Quality Check and Quantification

Quality of the DNA extracted from the samples was checked by agarose gel electrophoresis. 0.8 % agarose gel was prepared by dissolving 0.8gm of agarose in 100 mL of 1X TAE buffer and heated for two minutes. 5µl of ethidium bromide (EtBr) from stock solution of 10mg/mL was added to the solution and mixed uniformly. Gel was let to cool down, poured onto a gel tray and was allowed to set. DNA samples were diluted with Nuclease Free Water (NFW) and stained with 6X gel loading dye in the ratio of 1:8:1, to final volume of 10µl. Samples were loaded into their respective wells and 2µl of 1kb DNA ladder (GeneRuler, Thermo Scientific, USA) was used as a marker to compare the size of the DNA. Electrophoresis was carried out at a constant voltage of 100V for 45 minutes. Gel was documented in GeneFlash (Syngene Bio Imaging, USA).

DNA Quantification was done by using Nanodrop ND-1000 (Thermo Scientific, USA). With the sampling arm opened, 1µl of NFW was pipetted onto the lower measurement pedestal. The sampling arm was closed both the lower and upper measurement pedestals were wiped properly by piece of lint free paper. Again 1µl of NFW was loaded and was set as blank. Both the measurement pedestals were wiped out again and 1µl DNA samples were loaded and concentration of DNA samples were noted down along with the $A_{260/280}$ ratio.

## 3.5 Library Preparation and Sequencing

The genomic DNA were made ready for whole exome sequencing. This process comprised of dilution of DNA samples to final concentration of 5ng/µl and the diluted samples were subjected to quality check. Approximately 100ul of 5ng/µl final concentration of all the samples was prepared with Tris-EDTA (TE) buffer.

### 3.5.1 Library Preparation

The library preparation of the samples for whole exome sequencing was performed by using manufacturer provided protocol of TruSeq Exome Library Prep Illumina, USA.
100ng of gDNA was used for preparing library for whole exome sequencing. The brief protocol for whole exome library preparation is shown in Fig 3.1.

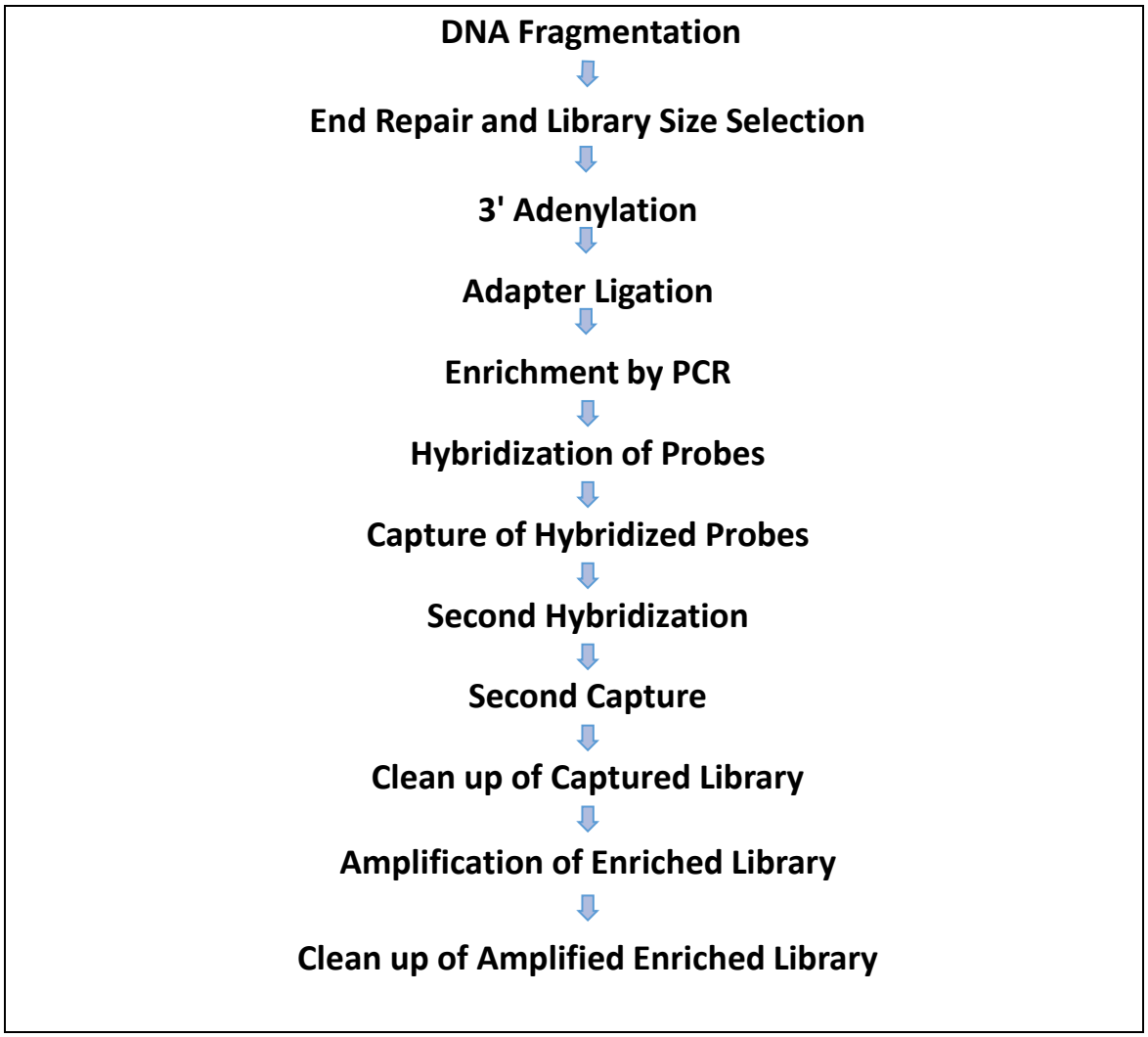


**Fig 3.1: Library Preparation Workflow of TruSeq Exome Library Preparation.**

#### 3.5.1.1 DNA Fragmentation

The genomic DNA samples were optimally fragmented to 150 bp insert size using Covaris S220 (Thermo Fisher Scientific, USA). The fragmentation comprised of following steps
i). Normalization of gDNA

1. 100ng of each gDNA samples were normalized with Resuspension Buffer (RSB) which acts as shearing buffer premix for fragmentation, by mixing 20ul each of 5ng/µl samples with 40ul of RSB.

ii). Fragmentation

2. The mixture was pipetted to the covaris tubes.
3. Fragmentation was done by using Covaris S220 (Thermo Fisher Scientific, USA) with the parameters as shown in Table 3.1
4. The fragmented DNA was transferred to 8-tube strip.

Table 3.1: Covaris parameter setting to fragment insert size of 150 bp.

| **Parameters** | **Set value** |
|---|---|
| Duty Factor (%) | 10 |
| Peak Power (W) | 175 |
| Cycles/Burst | 200 |
| Duration (seconds) | 280 |
| Temperature (°C) | 7 |
| Water Level | 12 |

iii). Clean up fragmented DNA

1. 100µl of Sample Purification Beads (SPB) was added to the tube and mixed thoroughly by pipetting.
2. It was incubated at RT for 5 minutes.
3. The mixture was then placed on a magnetic stand and waited until the liquid was clear (~8 minutes).
4. Supernatant was discarded and the DNA was washed two times with freshly prepared 80% ethanol (200µl for 30 seconds incubation on magnetic stand).
5. Incubated on the magnetic stand for 30 seconds.
6. Residual EtOH from each tube was removed by using a 20µl pipette.
7. Air-dried on the magnetic stand until the DNA pellet was dried (~5 minutes).
8. 62.5µl RSB was added to the tube containing dried DNA pellet and the tube was removed from the magnetic stand and then mixed thoroughly by pipetting up and down.
9. Incubated at room temperature for 2 minutes.
10. Again it was place on a magnetic stand and waited until the liquid was clear (~5 minutes).
11. 60µl of the supernatant was transferred to a new 8-tube strip

### 3.5.1.2 End Repair and Library Size Selection

The covaris fragmentation generates dsDNA fragments with 3' or 5' overhangs. End Repair Mix (ERP3) was used to convert sticky ends of the DNA fragments into blunt ends. The 3' to 5' exonuclease activity of this mix removes the 3' overhangs and the 5' to 3' polymerase activity fills in the 5' overhangs of the DNA fragments. The end repair was performed as follows:

1. 40µl ERP3 was added to the tube and was mixed thoroughly by pipetting up and down.
2. The tube was placed on thermal cycler and following program was set and run

- The preheat lid option set to 100°C
- 30°C for 30 minutes
- Hold at 4°C
- Each tube contains 100µl.

After the end repair, the fragment length of the library was optimized as follows:

1. 90µl of SPB was added to the tube mixed thoroughly by pipetting. It was incubated at RT for 5 minutes.
2. The mixture was then placed on a magnetic stand and waited until the liquid was clear (~8 minutes).
3. 185µl supernatant was transferred to the corresponding tube and 125µl SPB was added to each tube. They were mixed thoroughly by pipetting.
4. Further processing was done following the step 2 to 7 of clean up fragmented DNA.
5. Then, 20µl RSB was added to the tube.
6. The tube was removed from the magnetic stand and then mixed thoroughly by pipetting up and down.
7. Incubated at room temperature for 2 minutes.
8. Again it was place on a magnetic stand and waited until the liquid was clear (~5 minutes).
9. Finally, 17.5µl of the supernatant was transferred to a new 8-tube strip.

#### 3.5.1.3 3'- Adenylation

After end repair and optimization of library size, a single 'A' nucleotide was added to the 3' ends of the blunt fragments to prevent them from ligating to each other during the adapter ligation reaction as follows:

1. 12.5µl of A Tailing Mix (ATL2) to each tube and was mixed thoroughly.
2. The tube was placed on thermal cycler and following program was set and run
    - The preheat lid option set to 100°C
    - 37°C for 30 minutes
    - 70°C for 5 minutes
    - Hold at 4°C
    - Each tube contains 30µl

#### 3.5.1.4 Adapter Ligation

Following the 3’ end Adenylation, indexing adapters were ligated to the DNA fragments, which prepared them for hybridization onto a flow cell. Each DNA fragment was ligated with a unique index which so that various DNA samples can be pooled into one pool. Dual indexed adapter were ligated as follows:

1. To the tube containing DNA fragments 2.5µl of LIG2 (Ligation Mix 2) and 2.5µl of DNA adapters (DAP) were added and mixed thoroughly.
2. The tube was placed on thermal cycler and following program was set and run
    - The preheat lid option set to 100°C

- 30°C for 10 minutes
- Hold at 4°C
- Each tube contains 37.5µl

3. 5µl STL (Stop Ligation Buffer) was added to the tube and mixed thoroughly

After the indexing adapter ligation the ligated fragments are cleaned by using SPB as follows:

1. 42.5µl SPB was added to the tube and mixed thoroughly by pipetting.
2. Further processing was done following the step 2 to 7 of clean up fragmented DNA.
3. 52.5µl RSB was added to the tube and the tube was removed from the magnetic stand and then mixed thoroughly by pipetting up and down.
4. Incubated at room temperature for 2 minutes.
5. Again it was place on a magnetic stand and waited until the liquid was clear (~5 minutes).
6. 50µl of the supernatant was transferred to a new 8-tube strip.
7. Again this whole process from 1 to 5 was repeated by using 50µl of SPB and 27.5µl of RSB
8. Finally 25µl of the supernatant was transferred to new 8-tube strip.

#### 3.5.1.5 Enrichment of DNA fragments by PCR

Eight cycles of PCR was used to selectively enrich those DNA fragments that have adapter molecules on both ends and to amplify the amount of DNA in the library. PCR was performed with PPC (PCR Primer Cocktail) that anneals to the ends of the adapters as follows:

1. The tube was placed on ice and 5µl PPC was added.
2. 20 µl EPM (Enhanced PCR Mix) was added tube and mixed thoroughly.
3. The tube was placed on thermal cycler and following program was set and run
    - The preheat lid option set to 100°C
    - 98°C for 20 minutes
    - 60°C for 15 seconds
    - 72°C for 30 seconds
    - 72°C for 5 minutes
    - Hold at 4°C
    - Each tube contains 50µl

Amplified DNA was then cleaned by using SPB as follows:

1. 35µl SPB was added to the tube and was mixed thoroughly.
2. It was incubated at RT for 5 minutes.
3. The mixture was then placed on a magnetic stand and waited until the liquid was clear (~8 minutes).
4. 82µl supernatant was transferred to the corresponding tube and 82µl SPB was added to each tube. They were mixed thoroughly by pipetting.

5. Further processing was done following the step 2 to 7 of clean up fragmented DNA.
6. Then, 17.5µl RSB was added to the tube.
7. The tube was removed from the magnetic stand and then mixed thoroughly by pipetting.
8. The tube was incubated at RT for 2 minutes.
9. Again it was place on a magnetic stand and waited until the liquid was clear (~5 minutes).
10. Finally 15µl of the supernatant was transferred to a new 8-tube strip.

The enriched libraries were then validated qualitatively and quantitatively.
Enriched libraries were quantified using Qubit 2.0 Fluorometer (Thermo Fisher Scientific, USA)
Quality of the library was checked by agarose gel electrophoresis.

### 3.5.1.6 Hybridization of Probes

Validated DNA libraries containing unique indexes were then combined into a single pool and the targeted regions of the DNA libraries were hybridized with capture probes (Coding Exome Oligos). Following steps were performed for this:

1. The libraries were pooled into one pool in such a way that the pool contains 100ng of each library.
2. Final volume of the pool was adjusted to 40µl with RSB.

The probes were in turn hybridized to the pooled libraries as follows:

1. Following reagents were added in the order listed to a new 8-tube strip and were mixed by pipetting.
    - DNA library pool- 40µl
    - CT3 (Capture Target Buffer 3)- 50µl
    - CEX (Coding Exome Oligos)- 10µl (CEX are the probes to capture exons)
2. The tube was placed on thermal cycler and following program was set and run
    - The preheat lid option set to 100°C
    - 95°C for 10 minutes
    - 18 cycles of 1 minute each, starting at 94°C, then decreasing 2°C per cycle
    - 58°C for 90 minutes
    - Hold at 58°C for approximately 20 hours.
    - Each tube contains 100µl.

### 3.5.1.7 Capture of Hybridized Probes

The probes hybridized to the targeted regions of the DNA libraries were in turn captured using SMB (Streptavidin Magnetic Beads) as follows:

1. The SMB was incubated at RT for approximately 30 minutes before use.
2. 250µl SMB was added to a new 1.5ml micro centrifuge tube (MCT).

3. The total sample volume (~100µl) from the thermal cycler was directly transferred to the MCT containing SMB with the 8-tube strip still on thermal cycler. It was pipetted to mix.
4. The MCT was incubated at RT for 25 minutes with gentle tapping done at an interval of around 7 minutes to ensure that the beads did not get settled at the bottom of the tube.
5. It was centrifuged briefly and was incubated on a magnetic stand until the liquid was clear (2–5 minutes).
6. Supernatant was discarded into new MCT.
7. The tube was removed from the magnetic stand.

Nonspecific binding from the beads were removed by two heated washes with SWS (Streptavidin Wash Solution) as follows:

1. 200µl SWS was added to the tube and tube was tapped to mix the beads and SWS properly.
2. The tube was placed on the 50°C heat block for 30 minutes with 800 rpm shaking at an interval of 7 minutes.
3. After the heat incubation, the tube was immediately placed on a magnetic stand until the liquid was clear (~2 minutes).
4. Supernatant was discarded.
5. The tube was removed from the magnetic stand.
6. Steps 1–5 were repeated for a total of 2 washes**.**

The enriched library was then eluted from the beads and prepared for a second round of hybridization as follows:

1. The elution premix was prepared by mixing 28.5µl EE1 (Enrichment Elution Buffer 1) and 1.5µl HP3 (2N NaOH) in a 1.5 ml MCT and then was tapped and vortexed.
2. 23µl elution premix was added to the tube that contains the beads and was tapped to mix.
3. The MCT was incubated at RT for 2 minutes and was given a short spin.
4. Placed on a magnetic stand until the liquid was clear (~5 minutes).
5. 21µl supernatant was transferred to a new 8-tube strip.
6. To the tube 4µl ET2 (Elute Target Buffer 2) was added and was pipetted to mix.
7. Short spin was given.

#### 3.5.1.8 Second Hybridization

This step bound targeted regions of the enriched DNA with capture probes to ensure the high specificity of the captured regions a second time as follows:

1. To the 8-tube strip followings were added and mixed by pipetting.
    - RSB- 15µl
    - CT3 (Capture Target Buffer 3)- 50µl
    - CEX (Coding Exome Oligos)- 10µl

2. A short spin was given to the tube and was place on the thermal cycler with the same PCR programme used for first
3. The tube was kept at the 58°C holding temperature for approximately 24 hours.

#### 3.5.1.9 Second Capture

The same procedure of the first capture was repeated to ensure the high specificity of the captured regions.

#### 3.5.1.10 Clean up of Captured Library

The captured library was again purified using SPB as follows:

1. 45µl of the well-dispersed SPB was added to the tube and pipetted to mix.
2. Further processing was done following the step 2 to 7 of clean up fragmented DNA.
3. The library was then eluted on 27.5µl of RSB.
4. The tube was incubated at RT for 2 minutes.
5. Again it was place on a magnetic stand and waited until the liquid was clear (~5 minutes).
6. Finally 25µl of the supernatant was transferred to a new 8-tube strip.

#### 3.5.1.11 Amplification of Enriched Library

Enriched library was then amplified by an 8-cycle PCR program using PPC as follows:

1. To the tube 5µl PPC and 20µl NEM (Enrichment Amp Mix) were added and pipetted to mix.
2. A short spin was given to the tube and it was placed on thermal cycler and following program was set and run
    - The preheat lid option set to 100°C
    - 98°C for 10 seconds
    - 8 cycles of:
        - 98°C for 10 seconds
        - 60°C for 30 seconds
        - 72°C for 30 seconds
    - 72°C for 5 minutes
    - Hold at 4°C
    - Each well or tube contains 50µl

#### 3.5.1.12 Clean up of Amplified Enriched Library

Following the amplification using PCR program, the amplified enriched library was purified by using SPB removing the unwanted products as follows:

1. 45µl SPB was added to the tube and pipetted to mix.
2. Further processing was done following the step 2 to 7 of clean up fragmented DNA.
3. The library was then eluted on 22µl of RSB.
4. The tube was incubated at RT for 2 minutes.
5. Again it was place on a magnetic stand and waited until the liquid was clear (~5 minutes).
6. Finally 20µl of the supernatant was transferred to a new 8-tube strip.

Again validation of the amplified enriched library was done quantitatively and qualitatively as before.

### 3.5.2 Cluster Generation and Sequencing

The library was then prepared for cluster generation and finally sequencing. cBot (Illumina, USA) was used for cluster generation. This process comprised dilution and denaturation of the library as tabulated in Fig 3.3. To convert the ng/µl to nM a formula was used as depicted in Fig 3.2.

**Formula to convert ng/µL to nM**

$$\text{Concentration in nM} = \frac{\text{Concentration in ng/µL}}{660\ \text{g/mol} * \text{average library size}} * 10^6$$

Mol. Wt. of 1 base pair= 660 Dalton

**Fig 3.2: Formula to convert ng/µl to nM.**

The workflow for appropriate dilution of library for cluster generation is depicted and described in Fig 3.3.

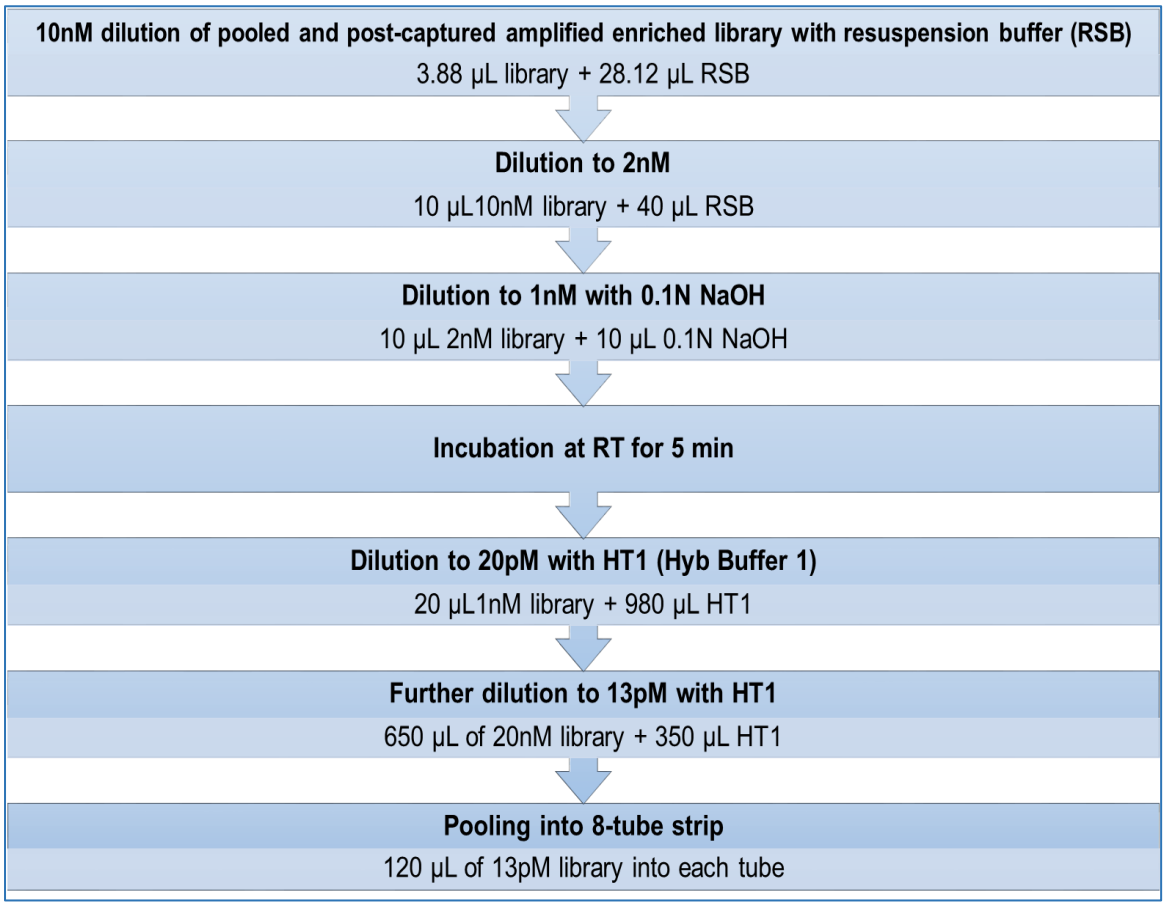


**Fig 3.3: Preparation of library for cluster generation and Sequencing.** The pool of library was diluted to 10nM with Resuspension Buffer (RSB) which was again diluted to 2nM. Then the library was denatured with equal volume of freshly prepared 0.1N sodium hydroxide

(NaOH) in a 1.5ml micro centrifuge tube incubating at room temperature for about 5 minute. Denatured library was further diluted to 20pM by adding 980µl of hybridization buffer (HT1). Finally, it was diluted to 13pM with HT1 and 120µl was added to 8-tube strip. The 8-tube strip with 13pM library was used for cluster generation on a cBot (Illumina, USA). The setup and operation were performed and monitored by cBot software interface using the touch screen monitor. After the completion of cluster generation in cBot, the flow cell was loaded into the flow cell compartment of the HiSeq 2500 (Illumina, USA) for high-throughput sequencing run. The setup and operation were performed and monitored by HiSeq Control Software (HCS) interface. After the completion of the sequencing the raw reads were extracted and proceeded for data analysis.

## 3.6 Bioinformatics Analysis of Whole Exome Sequencing Data

The raw sequencing data was subjected to extensive computational data analysis pipelines using sophisticated bio-informatic tools as described in Fig 3.4.

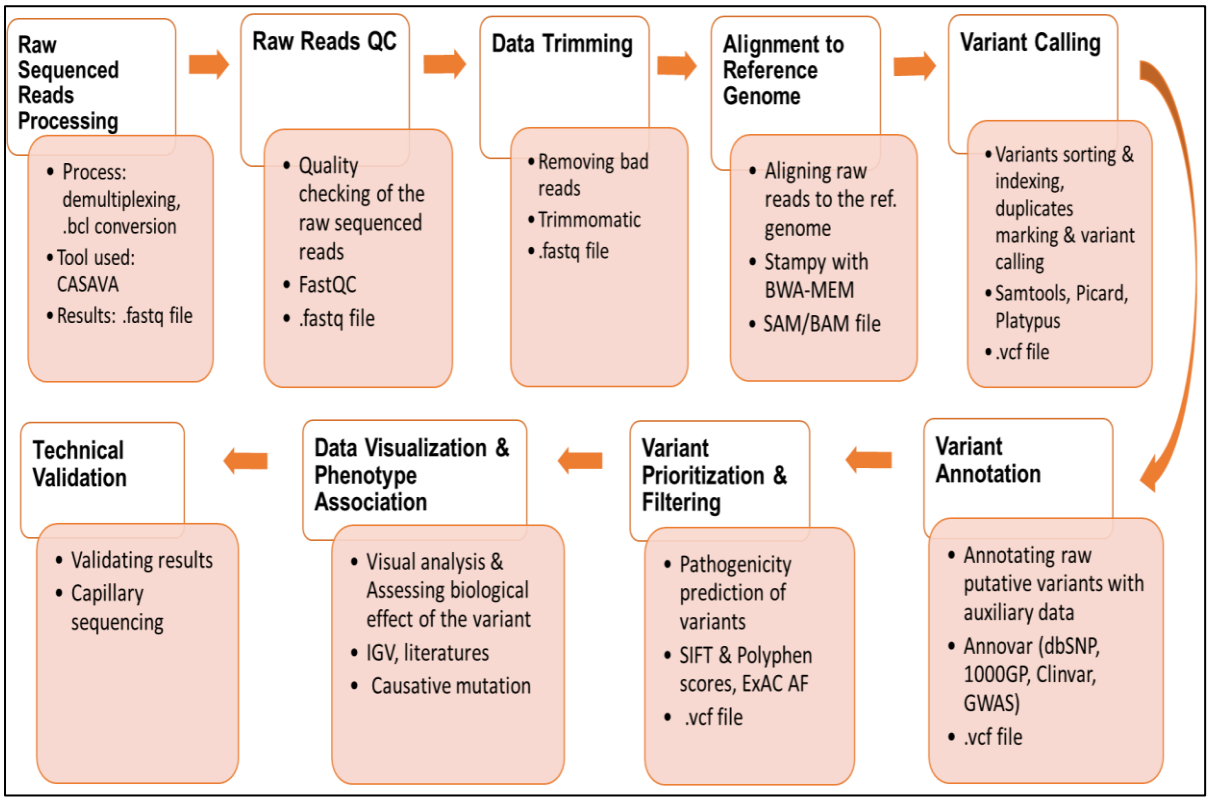


**Fig 3.4: Flow chart of bioinformatics analysis of the sequenced data.**

The analysis was carried out using the following steps:

### 3.6.1 Raw Sequence Reads

The output raw file of the sequencer were in .bcl (base call) format. This .bcl format is not compatible with downstream analysis processes and thus was converted to *.fastq.gz (compressed FASTQ files) format by using CASAVA (Consensus Assessment of Sequence And VAriation) which is a part of Illumina's sequencing analysis software.

## 3.6.2 Data Quality Check

The raw sequencing reads consists of bases with varying qualities. Variation in the base quality can occur either due to poor DNA quality or library contamination or sequencing errors. The quality of the raw sequencing data was checked by FastQC 0.11.5 (Andrews S. 2010). FastQC provides a summary report of data quality. An example of BoxWhisker type plot representing the base quality in data is shown in Fig 3.5.

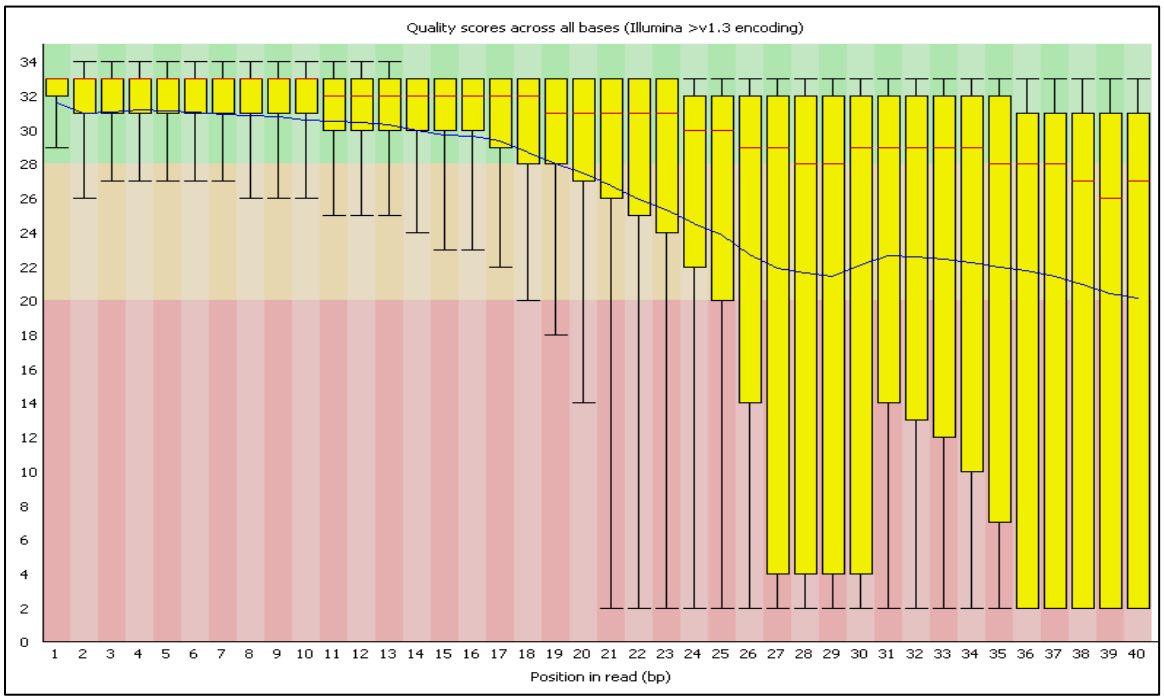


**Fig 3.5: Quality plot of the raw sequencing reads**. The y-axis on the graph shows the quality scores of the bases and x-axis shows position of the base in the read. Higher the quality score better is the quality of the base. Background color of the graph reflect the quality of the base; green, orange and red color represent bases of very good quality scores, reasonable quality scores and poor quality scores respectively. Each base in a raw sequencing read is represented by a BoxWhisker type plot. The central red line of the plot represents the median value whereas blue line represents the mean quality. The yellow box represents the inter-quartile range (25-75%) which represents the middle 50 percent of the distribution. The upper and lower whiskers represent the 10% and 90% points respectively. Adopted from Andrews S. 2010 with permission.

## 3.6.3 Data Trimming

Based on the FastQC report generated by FastQC, the raw sequencing reads were subjected to trimming. Raw sequenced reads were trimmed for bases with low phred quality scores (<20), read length below 36 bases and adapters sequences with the help of Trimmomatic 0.36 (Bolger *et al.*, 2014) as **AVGQUAL:20 ILLUMINACLIP:adapters .fa:2:30:10 SLIDINGWINDOW:5:20 MINLEN:36.** The brief description of trimming criteria used is as follows.

- AVGQUAL:20 - the read was dropped if the average phred quality score was below 20.

- ILLUMINACLIP:adapters.fa:2:30:10 - Illumina adapters were removed provided in the file adapters.fa. seedMismatches was set to 2 which means Trimmomatic will look for seed matches (16 bases) allowing maximally 2 mismatches. These seeds will be extended and clipped if; in the case of paired end reads a score of 30 is reached (about 50 bases), or in the case of single ended reads a score of 10, (about 17 bases). Alignment score is calculated as each matching base increases the alignment score by 0.6, while each mismatch reduces the alignment score by Q/10.
- SLIDINGWINDOW:5:20 – the reads were trimmed in the window of 5 bases if the average quality within the window falls below 20.
- MINLEN:36 – the reads length below 36 bases were dropped.

### 3.6.4 Alignment

The trimmed sequences reads were alighned to human reference genome (hg19/GRCh37) using Stampy 1.0.28 (Lunter and Goodson, 2011) with BWA 0.7.15 (Li and Durbin, 2009) in hybrid mode.

### 3.6.5 Pre processing and Variant calling

Following the alignment, the output SAM (Sequence Alignment/Map) file was converted to a compressed BAM (Binary Alignment/Map) file format and its sorting was done according to its position in the reference genome (coordinates) using Samtools 1.3.1 (Li *et al.*, 2009b). The PCR duplicates were marked by Picard 1.128. Such duplicates need to be removed because they could result in false variant detection. Samtools index command was used to create an index file of the output file which can be used for further downstream analysis. The variants were called using Platypus 0.8.1 (Rimmer *et al*., 2014) to give an output variant file in variant call format (.vcf).

### 3.6.6 Variant Annotation

The variants were annotated by using ANNOVAR 2.4 (Wang *et al.*, 2010). ANNOVAR draws information about the variants from various public databases such as GWAS, Clinvar and 1000 genomes, and present the information in an integrated summary.

### 3.6.7 Variant Prioritization

Following annotation, variants were prioritized based on pathogenicity, clinical phenotype, familial segregation and rare allele frequency as depicted in Fig 3.6.

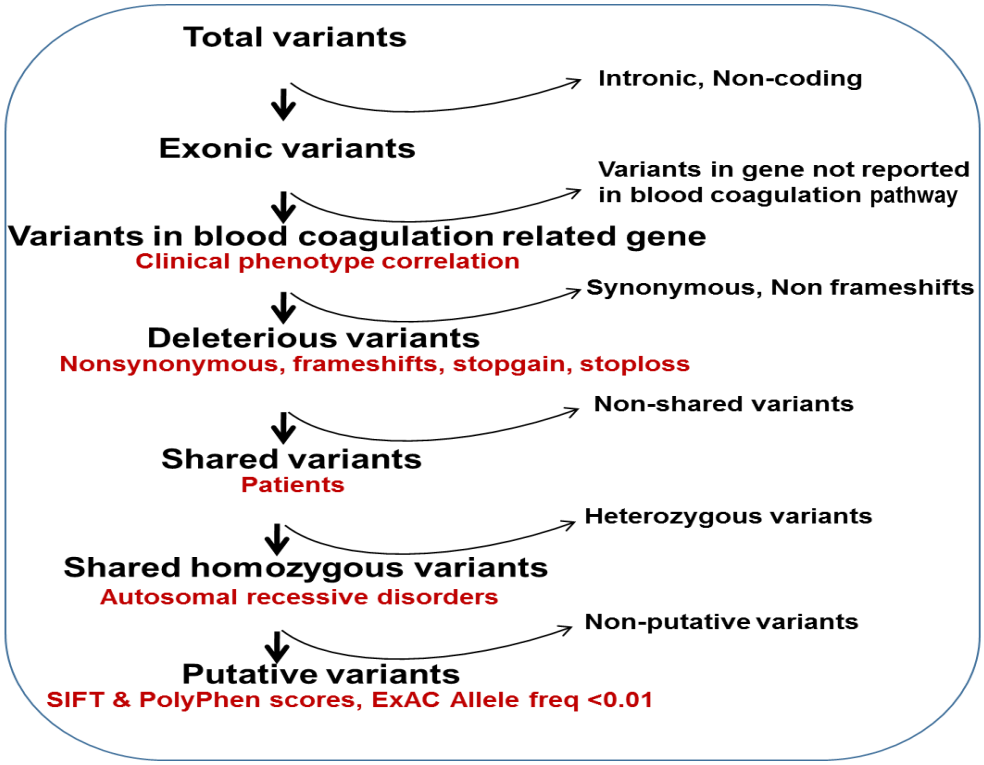


**Fig 3.6: Variant prioritization strategy.**

The exonic variants in the genes previously reported to be involved in causing various inherited bleeding disorders (shown in Table 3.2) were filtered. Then, among these variants deleterious variations which affect the protein sequences (nonsynonymous, frameshifts, stopgain and stoploss) were selected.

Table 3.2: Summary of genes related to bleeding disorders

| S. No. | Gene Name | Function of gene | Associated disorder |
|---|---|---|---|
| 1 | *FGA, FGB, FGG, F2, F3, F5, F7, F8, F9, F10, F11, F12, F13A1* and *F13B* | Encode various blood coagulation factor | Factor deficiency |
| 2 | *GGCX* and *VKORC1* | Encode Vitamin K regulating enzymes (γ- Glutamyl Carboxylase) and (Vitamin K Epoxide Reductase) | Vitamin k dependent clotting factor (FII, FVII, FIX, FX) deficiency |
| 3 | *ERGIC-53* and *MCFD2* | Encode proteins responsible for inter-organelle transport of FV and FVIII | Multiple factor deficiency (FV and FVIII deficiency) |
| 4 | *VWF* | Encodes von Willebrand factor (VWF) | Von Willebrand disease |

Following approaches were adopted to identify the putative mutation as described in Fig 3.6.

- Focusing on deleterious variants i.e., nonsynonymous SNVs, coding indels, frameshifts,
- Intersection of data- since sibs were affected so common shared variants were filtered,
- Mode of inheritance- since most of the inherited bleeding disorders have autosomal recessive mode of inheritance, homozygous and compound heterozygous variants were focused,
- Rare variants with ExAC allele frequency <0.01 taken from Exome Aggregation Consortium,
- Predicting and retaining variants with functional effects using in silico tools like SIFT, Polyphen ,
- Visualization of the variants in Integrative Genomics Viewer (IGV) (Robinson *et al*., 2011)
- Assessing biological effect of the variant based on literature reviews and online tools.

### 3.6.8 Validation of Putative Variation

The identified putative variants associated with the bleeding disorders were validated by capillary sequencing. The primers were designed against the genetic region which encompasses the putative variant by using NCBI Primer Blast tool. The genetic region encompassing the putative variants were PCR amplified using the designed primers and the PCR product was column purified. Then sequencing PCR was carried out using both reverse and forward primers. It was again purified and finally was subjected to capillary sequencing. The detailed information of all these processes is in appendix section.

# CHAPTER 4

# RESULTS

## 4.1 Clinical Presentation and Family Pedigree

The individuals were recruited from Nepal Hemophilia Society (NHS), Bagdole, Nepal. The family 1 (Fig 4.1) consists of two siblings (brother and sister) who were clinically diagnosed as the patients with Type 2 Normandy von Willebrand Disease (2N VWD). There was no history of any bleeding disorders in other family members. The clinical manifestation of the affected brother (II 3) includes uncontrolled bleeding even after minor injuries and easy bruising. His affected sister (II 4) was previously misdiagnosed as hemophilia A patient and presents complications like muscular hematomas, heavy menorrhagia, life threatening post-operative bleeding and even miscarriages. In both the siblings bleeding history was reported from the childhood.

Family 2 (Fig 4.2) consists of two siblings (sister and brother) who were clinically diagnosed as the patients with Factor X Deficiency (FXD). Other family members were asymptomatic for any kind of bleeding disorders. Phenotypical complications of the affected sister includes heavy menorrhagia and even miscarriages while brother presented easy bruising and hematomas.

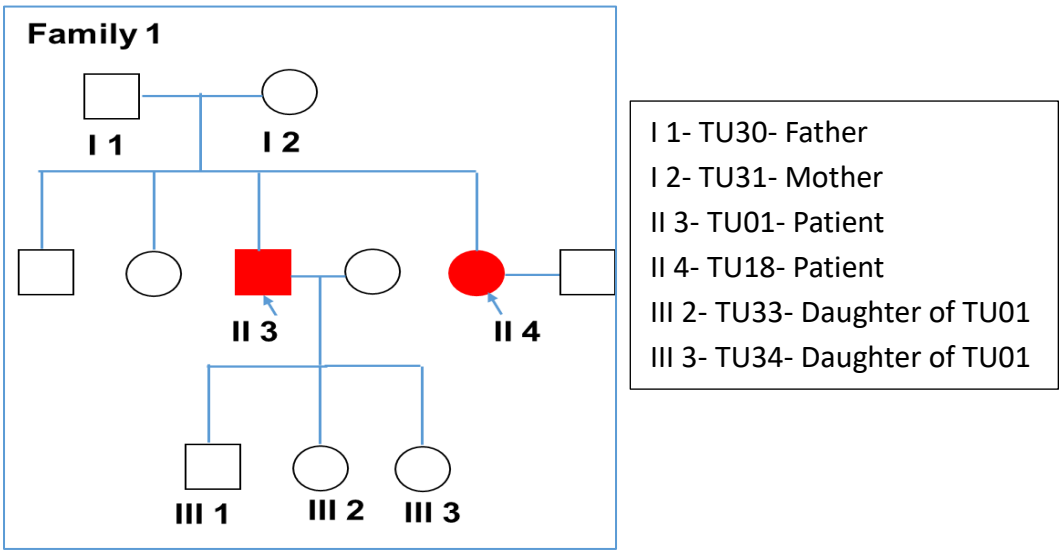


**Fig 4.1: Family Pedigree of patients with 2N VWD**. Patients are shown by arrow. Roman letter are indicating the generation. The arrow marked individuals were subjected to whole exome sequencing.

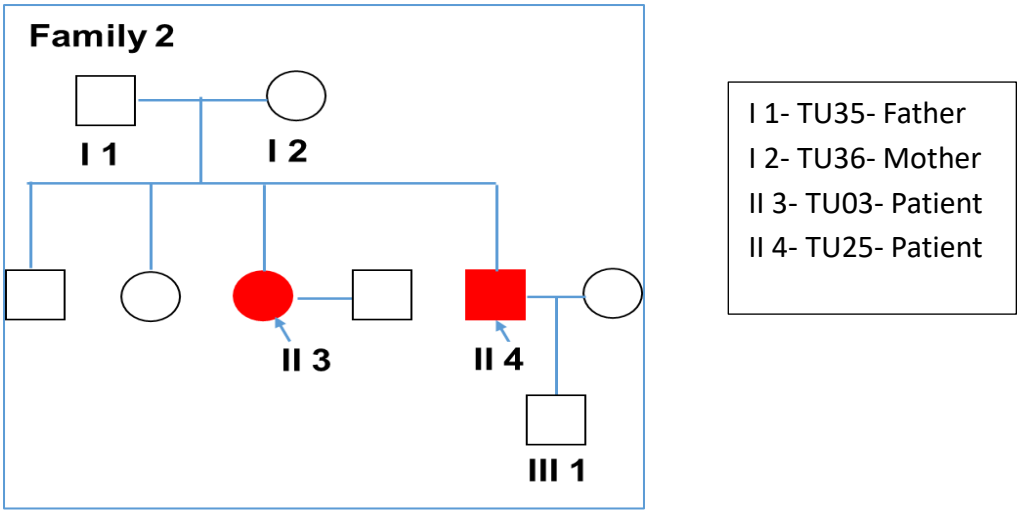


**Fig 4.2: Family Pedigree of patients with FXD**. Patients are shown by arrow. Roman letter are indicating the generation. The arrow marked individuals were subjected to whole exome sequencing.

## 4.2 Genomic DNA Extraction and Quality Check

The genomic DNA of the samples were extracted by Salting Out method (Miller et al., 1988). Extracted DNA samples were checked qualitatively on 0.8% agarose gel electrophoresis and quantitatively by Nanodrop ND-1000 (Thermo Scientific, USA) as described in materials and methodology section.

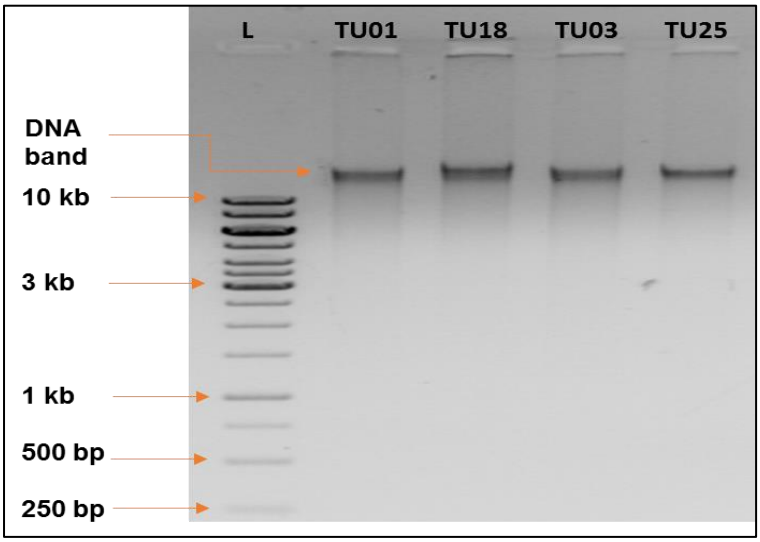


**Fig 4.3: Gel image of the extracted genomic DNA.** The product size of the extracted DNA is indicated by arrow. DNA extracted from all the 4 samples were with good yields and integrities. L- 1kb DNA ladder.

The concentration of the extracted DNA samples was measured by using Nanodrop ND-1000 (Thermo Scientific, USA) which is depicted in Table 3.1.

Table 4.1: NanoDrop quantification of DNA samples

| DNA Sample | Concentration ng/µl | $A_{260/280}$ ratio |
|---|---|---|
| TU01 | 89.6 | 1.87 |
| TU18 | 204.2 | 1.80 |
| TU03 | 313.0 | 1.9 |
| TU25 | 116.9 | 1.85 |

The extracted DNA was in good yield and quality. The ratio of absorbance at 260nm to that at 280nm was found to be within range of relatively pure DNA. (1.80-1.90).

## 4.2 Exome Library Preparation

The library preparation was performed using TruSeq Exome Library Prep Reference Guide as described in the materials and methodology section. The quality of the library was checked by agarose gel electrophoresis (Fig 4.4).

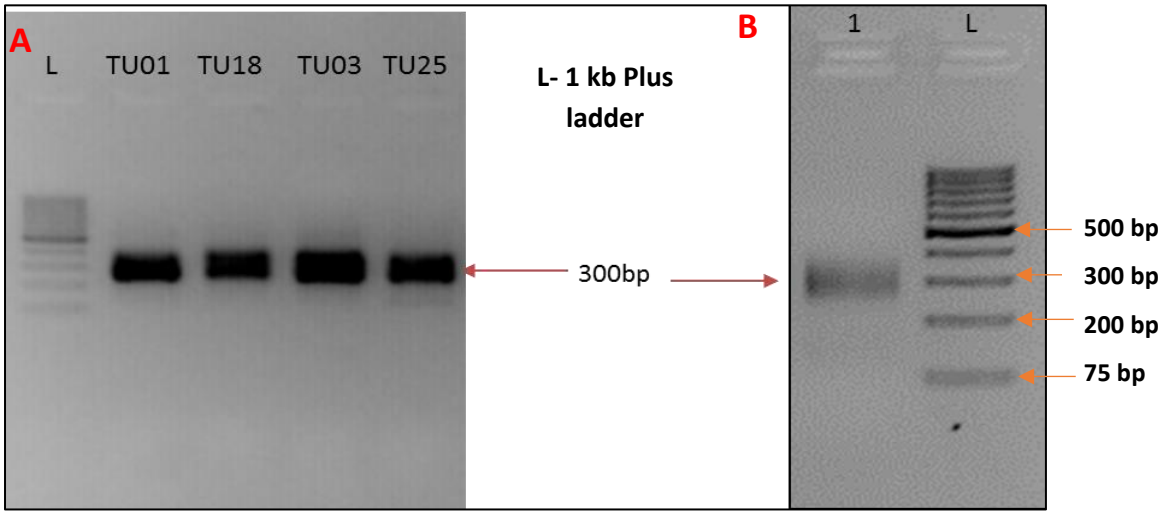


**Fig 4.4: Gel image of the pre- and post-captured library: (A)** Gel image of the library before capturing the exome. Gel image shows good amplification of the adapters-ligated libraries. The size of the library is shown by an arrow. **(B)** Gel image of library after capturing the exome and pooling multiple uniquely indexed samples into one pool represented as 1**.** Size of the library is ~300bp. L-1kb Plus ladder.

The quantification of the PCR enriched library before exome capture and after exome capture was performed by using Qubit 2.0 Fluorometer (Thermo Fisher Scientific, USA) as described in materials and methodology section. The concentration of the pre- and post-captured library is shown in Table 4.2.

Table 4.2: Quantification of pre- and post-captured library

| Sample Code | Concentration ng/µl | Remarks |
|---|---|---|
| TU01 | 37.1 | Pre-captured library |
| TU18 | 28.3 | Pre-captured library |
| TU03 | 38.2 | Pre-captured library |
| TU25 | 29.4 | Pre-captured library |
| 1 | 15.3 | Post-captured library |

1- Library after exome capture and pooling.

The whole exome sequencing of the samples was performed in HiSeq 2500 (Illumina, USA) using high-output run mode. The parameters set for sequencing was paired end reads with dual index and the sequencing was carried out for 296 cycles (140, 8, 8, 140). Size of the raw reads obtained after the whole exome sequencing is shown in Table 4.3. TU25 was sequenced twice of and the sequenced raw file was merged and was proceeded for further analysis. It was done just to check whether there will be any significant difference during variant analysis or not. Thus the size of raw file of TU25 is almost double of others.

Table 4.3: Size of raw sequencing data

| Sample Code | Size of raw sequencing file |
|---|---|
| TU01 | 4.24 GB |
| TU18 | 5.17 GB |
| TU03 | 4.92 GB |
| TU25 | 8.35 GB |

## 4.4 Bioinformatics Analysis of Sequencing Data of 2N VWD Cases (TU01 and TU18)

### 4.4.1 Quality Check and Trimming

The quality of the raw sequencing data was checked before aligning to the reference genome by using FastQC tool. It provided a summary report of the input file including information about base quality, content and read length as described in the Fig 4.5A. The FastQC report of the trimmed data is shown in Fig 4.5B. The interpretation of data quality was made based on the BoxWhisker plot as described in materials and methodology section.

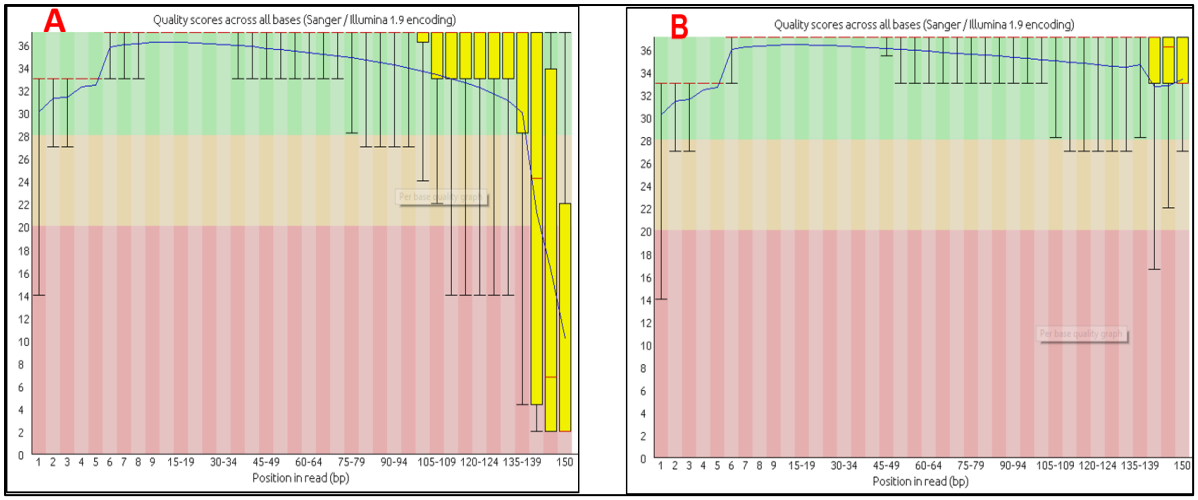


**Fig 4.5: Quality plot of the raw sequencing reads of TU01.** (A) The Box Whisker plot representing the data quality before trimming. The Y-axis represents the Phred quality scores and X-axis represents the position of base in the read. The bases towards the end of the reads were having poor quality as represented by low Phred scores (red region). The bases present on the middle region (green and orange region) in the read are of good and reasonable qualities. (B) The BoxWhisker plot representing the data quality after trimming**.** Poor quality scores towards the end of the reads were removed after trimming. All the bases with good and reasonable qualities were retained.

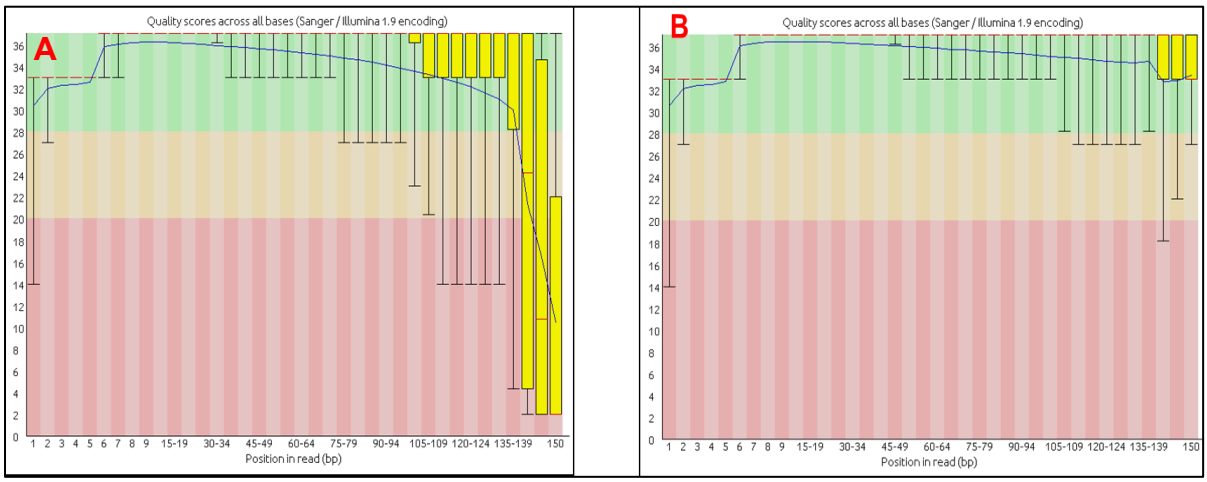


**Fig 4.6: Quality plot of the raw sequencing reads of TU18.** (A) The Box Whisker plot representing the data quality before trimming. The Y-axis represents the Phred quality scores and X-axis represents the position of base in the read. The bases towards the end of the reads were having poor quality as represented by low Phred scores (red region). The bases present on the middle region (green and orange region) in the read are of good and reasonable qualities. (B) The BoxWhisker plot representing the data quality after trimming**.** Poor quality scores towards the end of the reads were removed after trimming. All the bases with good and reasonable qualities were retained.

The information obtained from the FastQC report is summarized in Table 4.4. Approximately 4% of the total reads were trimmed in both the cases and further subjected

to alignment. The mean sequence quality score of the reads was 36 which implies low error rate in base calling.

Table 4.4: Summary of FastQC report of TU01 and TU18

| **Parameter** | **TU01** | | **TU18** | |
|---|---|---|---|---|
| | **Before Trimming** | **After Trimming** | **Before Trimming** | **After Trimming** |
| Total Sequences | 48233624 | 46570583 (96.55%) | 58853788 | 56785779 (96.49%) |
| Mean sequence Quality score | 36 | 36 | 36 | 36 |
| Sequences flagged as poor quality | 0 | 0 | 0 | 0 |
| Sequence length | 140-150 | 36-150 | 140-150 | 36-150 |
| %GC | 45 | 44 | 45 | 45 |

### 4.4.2 Alignment

After trimming the sequences with bad qualities and those without the adapters were then subjected to alignment against human reference sequence (hg19/GRCh37) using BWA-Stampy (Stampy with BWA-MEM) hybrid mode. Burrows-Wheeler Alignment tool (BWA) is based on Burrows-Wheeler Transformation (BWT) algorithm and allow efficient alignment of short sequencing reads and thus in turn allows locating mismatches and gaps arising due to indels (insertions/deletions) (Li and Durbin, 2009). Although it has an advantage of speed but it is less sensitive. Therefore to complement this limitation of BWA, another tool called Stampy, which is based on hash based algorithm (Lunter and Goodson, 2011) was used in conjunction with BWA. This will speed up the alignment process without compromising on the sensitivity of alignment. Since the average read length of our data after the pre-processing step was above 70 bases (36-140 bp), the BWA-MEM (Burrows-Wheeler Alignment- Maximal Exact Matches) was run before Stampy. The brief description of the alignment is summarized in Table 4.5. The very high mapping percentage of 99.98% implies that data has reliability.

Table 4.5: Mapping summary of TU01 and TU18

| | **QC Passed Reads** | **Mapped Reads** | **Mapped %** |
|---|---|---|---|
| **TU01** | 46570583 | 46559754 | 99.98 % |
| **TU18** | 56785779 | 56774234 | 99.98 % |

### 4.4.3 Variant Calling

Each human genome comprises a huge number of variations compared to the reference genome. Among these variations, lies the disease causing mutations. After aligning the reads against the reference genome, variants has to be called efficiently, a process known as variant calling. We used Platypus 0.8.1 which provides a fast integrated algorithm for detecting indels, SNPs and variants with high sensitivity and specificity (Rimmer *et al*., 2014) for this purpose. The brief description of the variants is shown in Table 4.6.

### 4.4.4 Variant Annotation and Prioritization

Generally a large number of variant calls are generated after variant detection and hence it is essential to annotate them. Also the variants detected may be either true variants or just sequencing artefacts arising due to sequencing errors, sample contamination or insufficient variant coverage. Thus variants needs to be functionally annotated which is done by various annotation tools. The auxiliary information of the variants are combined with the raw putative variant calls and the annotation of the variant is performed. We have used ANNOVAR 2.4 (Wang *et al*., 2010) to functionally annotate the detected putative variations.

Following annotation, variants were prioritized based on pathogenicity, clinical phenotype, familial segregation and rare allele frequency as described in details in in materials and methodology section. Genes encoding various clotting factors were set as the genetic loci for the putative variations associated. Deleterious variations which affect the protein sequences (nonsynonymous, frameshifts, stopgain and stoploss) were selected. The sorting of the variants based on the pipeline followed for putative variant associated with the bleeding disorders as described in details in materials and methodology section, is depicted in Fig 4.7.

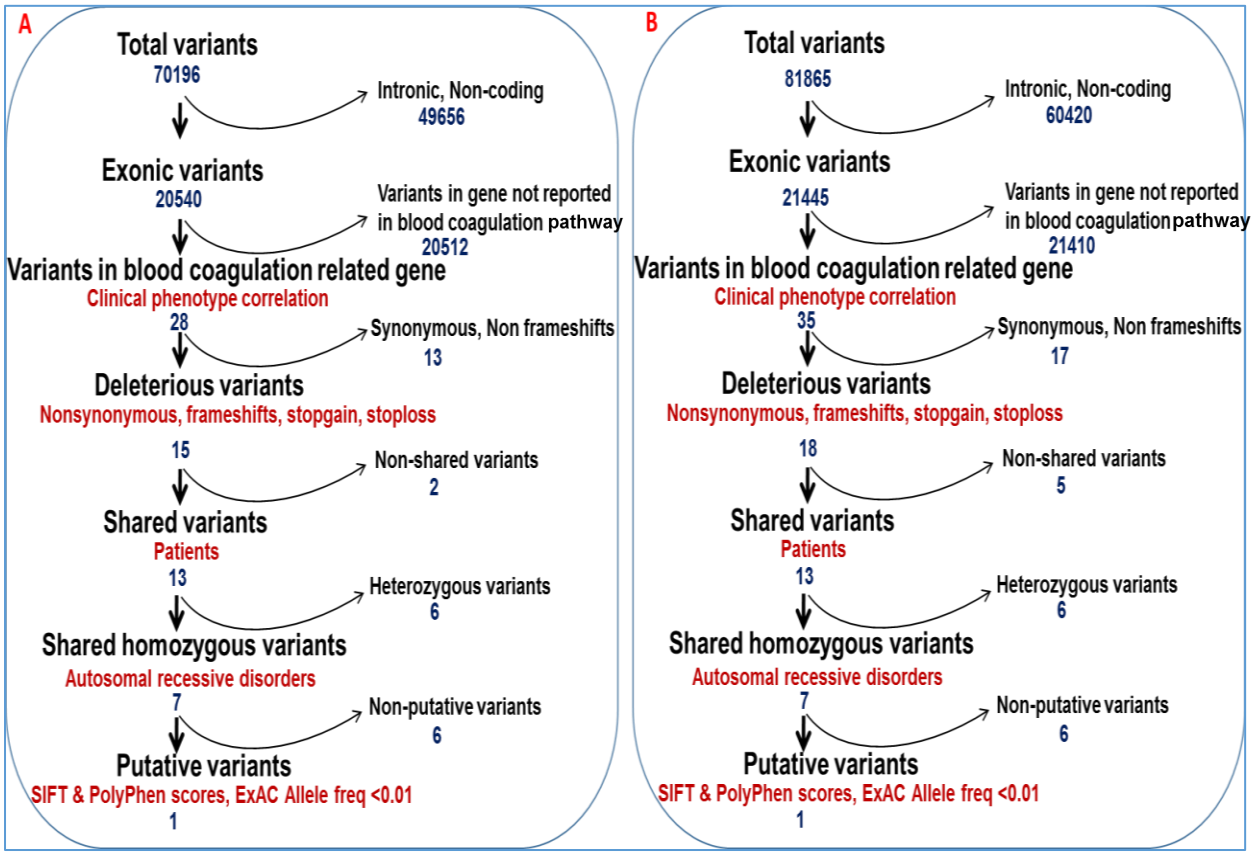


**Fig 4.7: Pipeline used for variants sorting.** (A) Variants sorting in TU01 and (B) variants sorting in TU18. Different variations show equivalent occurrence in both the siblings. Finally only one variant was prioritized in both the siblings which was the putative variant associated with the disorder.

Only 29.26% and 26.20% of the total variants were found to be present in the exonic region for TU01 and TU18 respectively as shown in Table 4.6.

Table 4.6: Summary of total variants found in TU01 and TU18

| | **TU01** | **TU18** |
|---|---|---|
| **Total Variants** | 70,196 | 81,865 |
| **Exonic Variants** | 20,540 | 21,445 |

Details of such exonic variants are also presented on Table 4.7 and also in Fig 4.8.

Table 4.7: Summary of various exonic mutations found in TU01 and TU18

| **Type of Mutation** | **No. of mutations in TU01** | **No. of mutations in TU18** |
|---|---|---|
| Frameshift deletion | 70 | 79 |
| Frameshift insertion | 57 | 62 |
| Frameshift substitution | 10 | 6 |
| Non frameshift deletion | 156 | 180 |
| Non frameshift insertion | 129 | 143 |
| Non frameshift substitution | 603 | 644 |
| Non synonymous SNVs | 8931 | 9263 |
| Synonymous SNVs | 10059 | 10510 |
| Stopgain | 67 | 67 |
| Stoploss | 7 | 7 |
| Unknown | 451 | 484 |

These different variations are presented in pie-chart as shown in Fig 4.8.

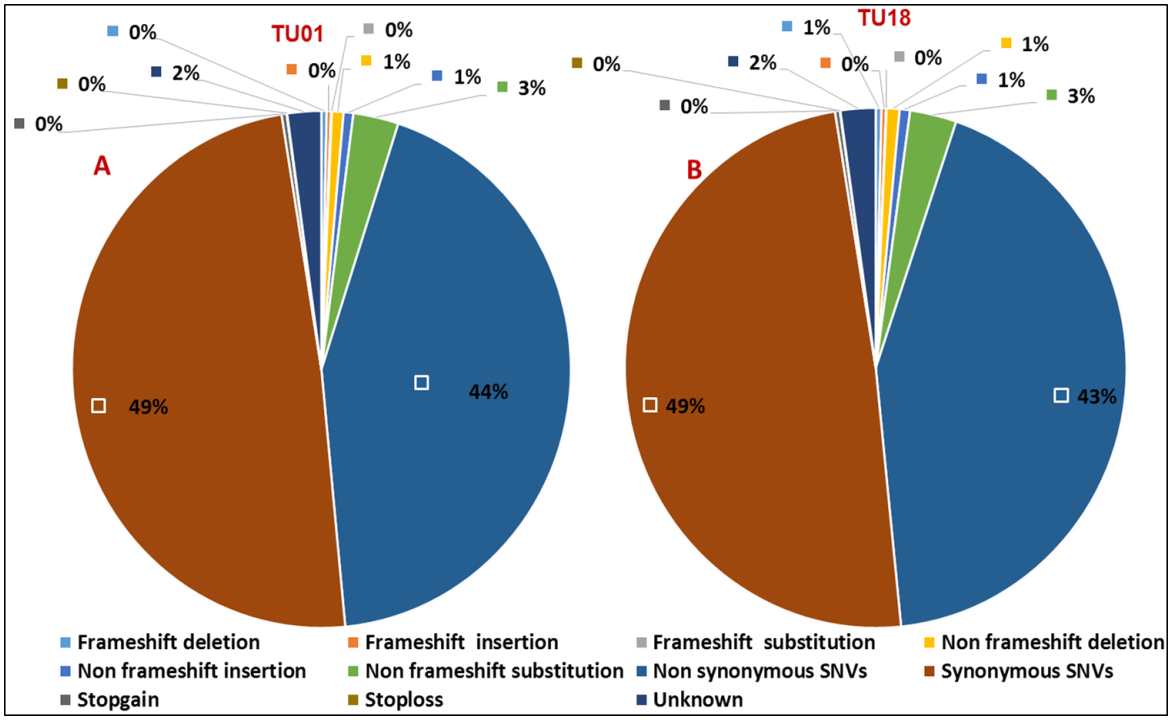


**Fig 4.8: Pie chart representing the relative percentage of various types of variations.** (A) In TU01 and (B) in TU18. Figures show both the siblings bear different variations with almost same proportion. The synonymous SNVs cover the highest percentage (49% in both) followed by nonsynonymous SNVs (44% in TU01 and 43% in TU18). Frameshift (deletion, insertion and substitution), stopgain and stopless mutations show very less prevalence whereas 2% of total mutations were unknown.

### 4.4.4.1 Variants Related to Rare Bleeding Disorders

In total 28 and 35 variations were found to be present in genes reported to cause rare bleeding disorders in TU01 and TU18 respectively. Among the total deleterious variations only nonsynonymous SNVs were found to be present in both the siblings as shown in Table 4.8.

Table 4.8: Total number of types of variations related to inherited bleeding disorders in TU01 and TU18

| Mutation Type | Mutations in TU01 | Mutations in TU18 |
|---|---|---|
| Non synonymous SNVs | 15 | 18 |
| Synonymous SNVs | 13 | 17 |
| Total | 28 | 35 |

Further sorting revealed a variation shared between the siblings (highlighted in red in Table 4.9 and 4.10) which was deleterious and probably damaging as predicted by SIFT and Polyphen2_HVAR scores.

Table 4.9: Details of Nonsynonymous SNVs of TU01

| Chr | Position | Ref | Alt | Gene.refGene | AAChange.refGene | SIFT_score_pred | Polyphen2_HVAR_score_pred | Hom/Het |
|---|---|---|---|---|---|---|---|---|
| chr1 | 169498975 | T | C | *F5* | F5:NM_000130:exon16:c.A5290G:p.M1764V | 1 T | 0 B | Het |
| chr1 | 169510475 | G | T | *F5* | F5:NM_000130:exon13:c.C3853A:p.L1285I | 0.41 T | 0.002 B | Het |
| chr1 | 169511555 | T | C | *F5* | F5:NM_000130:exon13:c.A2773G:p.K925E | 1 T | 0.001 B | Het |
| chr1 | 169511734 | T | C | *F5* | F5:NM_000130:exon13:c.A2594G:p.H865R | 0.4 T | 0 B | Het |
| chr1 | 169511755 | T | C | *F5* | F5:NM_000130:exon13:c.A2573G:p.K858R | 0.54 T | 0 B | Het |
| chr1 | 169519049 | T | C | *F5* | F5:NM_000130:exon10:c.A1601G:p.Q534R | 1 T | -- | Hom |
| chr1 | 197031021 | C | T | *F13B* | F13B:NM_001994:exon3:c.G344A:p.R115H | 0.22 T | 0.002 B | Hom |
| chr2 | 85780536 | C | T | *GGCX* | GGCX:NM_001142269:exon7:c.G803A:p.R268Q | 0.67 T | 0.004 B | Het |
| chr5 | 176831826 | C | G | *F12* | F12:NM_000505:exon7:c.G619C:p.A207P | 0.4 T | 0 B | Hom |
| chr6 | 6174866 | G | A | *F13A1* | F13A1:NM_000129:exon12: c.C1694T:p.P565L | 0.74 T | 0.002 B | Hom |
| chr12 | 6128443 | T | C | *VWF* | VWF:NM_000552:exon28:c.A4141G:p.T1381A | 1 T | 0.001 B | Hom |
| chr12 | 6143984 | T | C | *VWF* | VWF:NM_000552:exon20:c.A2555G:p.Q852R | 1 T | 0 B | Hom |
| chr12 | 6145654 | G | A | *VWF* | VWF:NM_000552:exon19:c.C2446T:p.R816W | 0.02 D | 1 D | Hom |
| chr12 | 6172202 | T | C | *VWF* | VWF:NM_000552:exon13:c.A1451G:p.H484R | 1 T | 0.025 B | Hom |
| chr18 | 57000469 | T | A | *LMAN1* | LMAN1:NM_005570:exon11:c.A1228T:p.M410L | 0.44 T | 0 B | Het |

Table 4.10: Details of Nonsynonymous SNVs of TU18

| Chr | Position | Ref | Alt | Gene.RefGene | AAChange.refGene | SIFT_score_pred | Polyphen2_HVAR_score_pred | Hom/Het |
|---|---|---|---|---|---|---|---|---|
| chr1 | 169498975 | T | C | *F5* | F5:NM_000130:exon16:c.A5290G:p.M1764V | 1 T | 0 B | Het |
| chr1 | 169510139 | G | A | *F5* | F5:NM_000130:exon13:c.C4189T:p.L1397F | 0.71 T | 0.007 B | Het |
| chr1 | 169511555 | T | C | *F5* | F5:NM_000130:exon13:c.A2773G:p.K925E | 1 T | 0.001 B | Het |
| chr1 | 169511734 | T | C | *F5* | F5:NM_000130:exon13:c.A2594G:p.H865R | 0.4 T | 0 B | Het |
| chr1 | 169511755 | T | C | *F5* | F5:NM_000130:exon13:c.A2573G:p.K858R | 0.54 T | 0 B | Het |
| chr1 | 169519049 | T | C | *F5* | F5:NM_000130:exon10:c.A1601G:p.Q534R | 1 T | - - | Hom |
| chr1 | 169519112 | C | T | *F5* | F5:NM_000130:exon10:c.G1538A:p.R513K | 0.32 T | 0.382 B | Het |
| chr1 | 197031021 | C | T | *F13B* | F13B:NM_001994:exon3:c.G344A:p.R115H | 0.22 T | 0.002 B | Hom |
| chr2 | 85780536 | C | T | *GGCX* | GGCX:NM_001142269:exon7:c.G803A:p.R268Q | 0.67 T | 0.004 B | Het |
| chr4 | 155491759 | G | A | *FGB* | FGB:NM_001184741:exon8:c.G1256A:p.R419K | 1 T | 0.002 B | Het |
| chr5 | 176831826 | C | G | *F12* | F12:NM_000505:exon7:c.G619C:p.A207P | 0.4 T | 0 B | Hom |
| chr6 | 6174842 | G | A | *F13A1* | F13A1:NM_000129:exon12:c.C1718T:p.T573M | 0.07 T | 0.795 P | Het |
| chr6 | 6174866 | G | A | *F13A1* | F13A1:NM_000129:exon12:c.C1694T:p.P565L | 0.74 T | 0.002 B | Het |
| chr12 | 6128443 | T | C | *VWF* | VWF:NM_000552:exon28:c.A4141G:p.T1381A | 1 T | 0.001 B | Hom |
| chr12 | 6143984 | T | C | *VWF* | VWF:NM_000552:exon20:c.A2555G:p.Q852R | 1 T | 0 B | Hom |
| chr12 | 6145654 | G | A | *VWF* | VWF:NM_000552:exon19:c.C2446T:p.R816W | 0.02 D | 1 D | Hom |
| chr12 | 6172202 | T | C | *VWF* | VWF:NM_000552:exon13:c.A1451G:p.H484R | 1 T | 0.025 B | Hom |
| Chr18 | 57000469 | T | A | *LMAN1* | LMAN1:NM_005570:exon11:c.A1228T:p.M410L | 0.44 T | 0 B | Het |

The putative variant was visualized in Integrative Genomics Viewer (IGV) as shown in Fig 4.9.

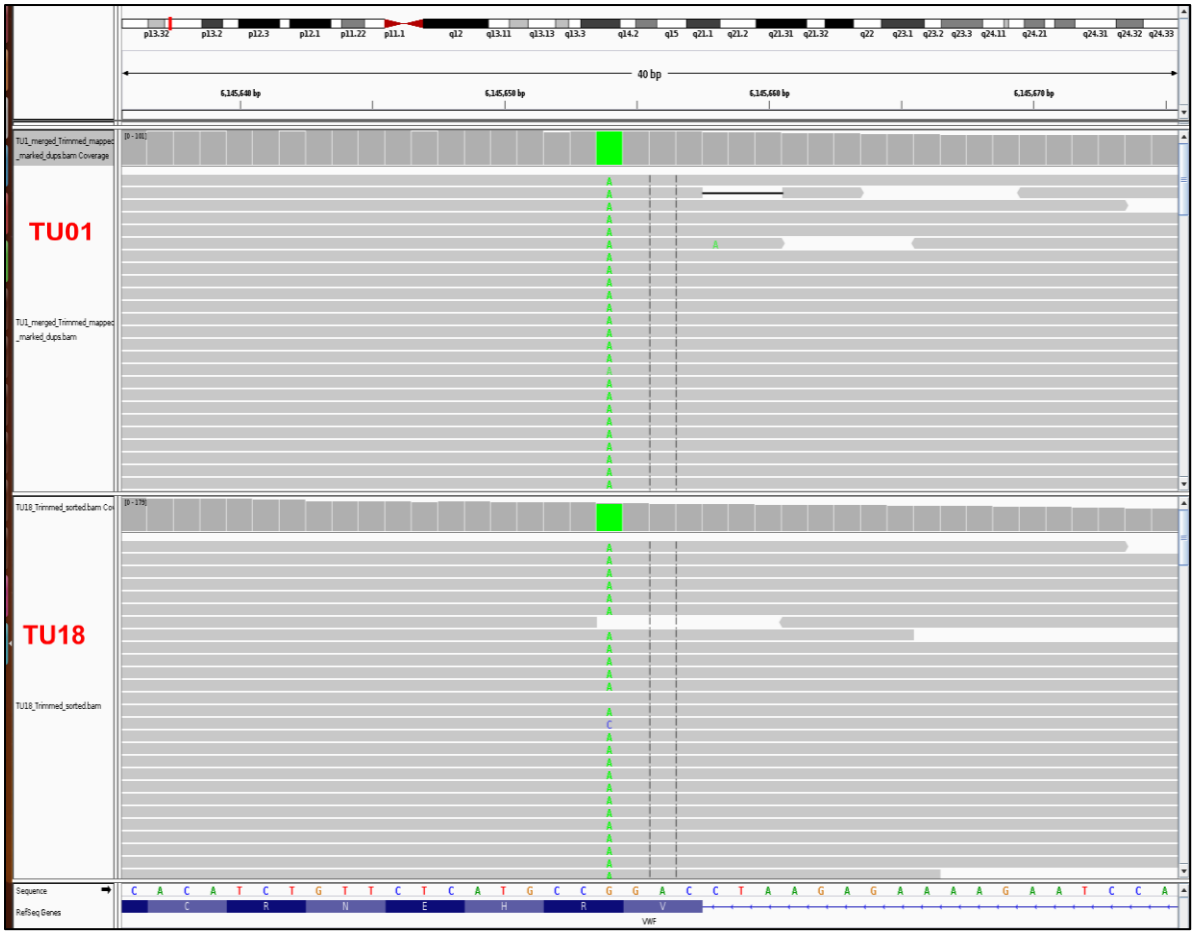


**Fig 4.9: IGV Snapshot of the variant p.R816W.** The IGV snapshot shows both the sibling (upper –TU01 and lower –TU18) being homozygous for the mutation (Chr12: 6145654G>A, c.C2446T). Coverage for the variant was 95X and 142X respectively in TU01 and TU18. Such a high coverage proves that the variant is not any sequencing artefact.

#### 4.4.4.1.1 Pathogenicity Prediction of Variants

We considered all the parameters as described in details in methods and methodology section and finally identifying one variant (c.C2446T:p.R816W in exon 19 of *VWF*) with deleterious and/or damaging effect on the protein structure or function in both the siblings as shown in above Tables 4.11. The variant was inherited in homozygous form in both the siblings. The allele frequency for the variant was found to be 0.000002489 in general population in the Exome Aggregation Consortium database. The SIFT and Polyphen2_HVAR scores were found to be 0.02 and 1 respectively for the R816W mutation.

Table 4.11: Putative mutation in TU01 and TU18

| Chr | Position | Ref | Alt | Gene.refGene | AAChange.refGene | Exonic funct change | SIF_pred | Polyphen2_HVAR_pred | hom/het | ExAC AF |
|---|---|---|---|---|---|---|---|---|---|---|
| Chr 12 | 6145654 | G | A | *VWF* | VWF:NM_000552:exon19:c.C2446T:p.R816W | Nonsynonymous SNV | 0.02 D | 1 D | hom | 2.489 E-5 |

The variant R816W is present in TIL subdomain of D' domain of VWF protein as shown in Fig 4.10. The D' domain participates in FVIII binding.

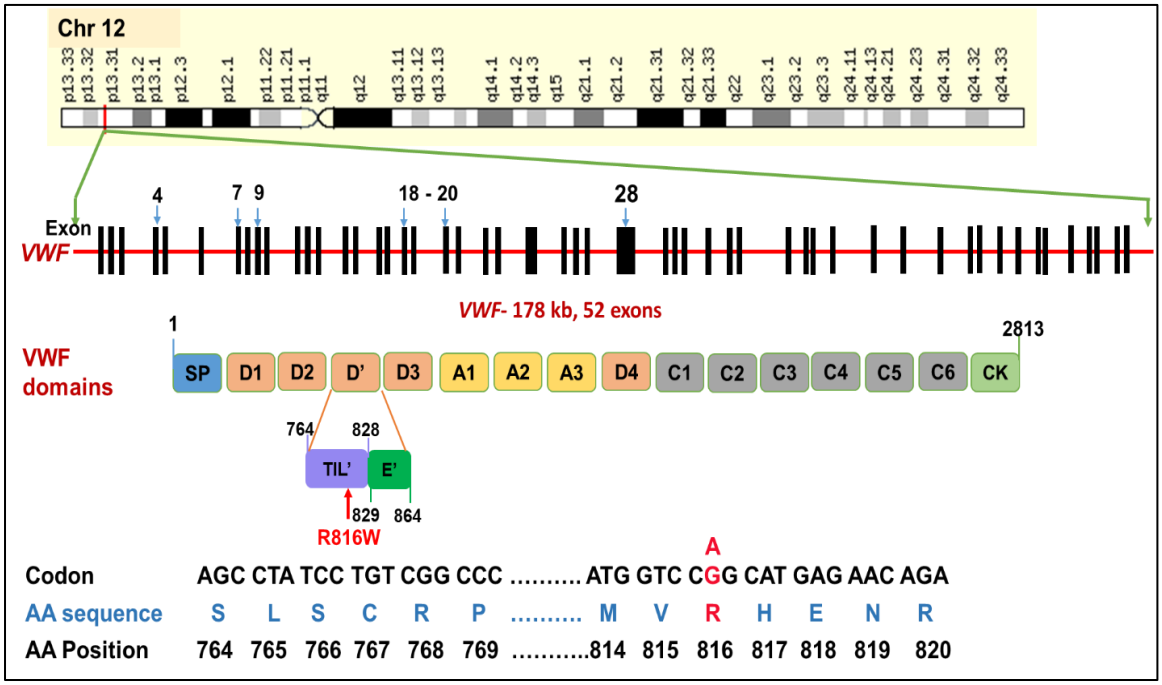


**Fig 4.10: Localization of variant R816W in VWF protein and gene.** The variant is localized in TIL' subdomain of D' domain in the mosaic VWF protein as shown by arrow. *VWF* is composed of 52 exons as shown by black bold lines in second row. The exons indicated by arrow are reported to possess mutations causing 2N VWD in literature.

### 4.4.4.1.2 Allele Frequency of the Variant R816W in Different Population

Further, allele frequency of the variant was checked separately in different population in ExAC. Only 3 alleles were found to be reported till date in ExAC. This variant seems to be present only in European (Non-Finnish) and Latino population as presented in Table 4.12.

Table 4.12: Allele Frequency of the variant R816W in different population (source ExAC, accessed on 11-07-2016)

| Population | Allele Count | Allele Number | No. of Homozygotes | Allele Frequency |
|---|---|---|---|---|
| Latino | 1 | 11472 | 0 | 8.717E-5 |
| European (Non-Finnish) | 2 | 66358 | 0 | 3.014E-5 |
| African | 0 | 10268 | 0 | 0 |
| East Asian | 0 | 8622 | 0 | 0 |
| European (Finnish) | 0 | 6500 | 0 | 0 |
| Other | 0 | 902 | 0 | 0 |
| South Asian | 0 | 16416 | 0 | 0 |
| Total | 3 | 120538 | 0 | 2.489E-5 |

This shows the novelty of the variant in Asian population. However, the same variant is reported from China (Qin *et al.*, 2014).

### 4.4.5 Validation by Capillary Sequencing

The putative variant G>A at 6145654 position (c.C2446T:p.R816W) in exon 19 of *VWF* associated with 2N VWD prioritized from the computational analysis of the whole exome sequencing data of the affected individuals was technically validated by targeted capillary sequencing as described in materials and methodology section. The genetic regions encompassing the variant was PCR amplified by using targeted primers (details in appendix section). The segregation of the variant was checked among the family members of the patients.

For this validation the genomic DNA samples of both the patients (TU01 & TU18), their father (TU30), mother (TU31), wife of TU01 (TU32) and two daughters of TU01 (TU33 & TU34) were used. All the DNA samples were PCR amplified by using primers designed to amplify the genetic region encompassing the variant. The PCR results is depicted in Fig 4.11.

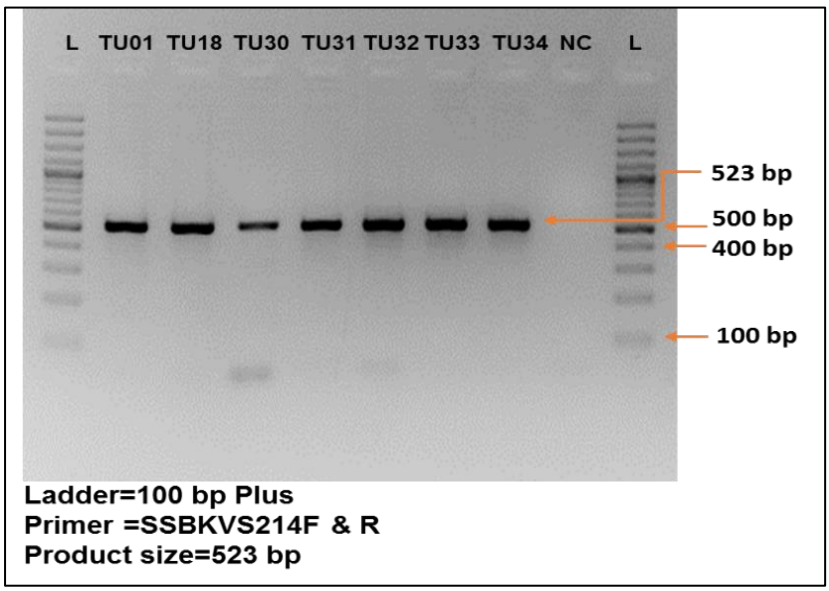


**Fig 4.11: PCR amplification of genetic region encompassing the putative variant Chr12: 6145654 G>A, c.C2446T:p.R816W present in exon 19 of *VWF*.** Product size of the amplified region was 523 bp. L- 100 bp Plus ladder (GeneRuler, Thermo Scientific, USA).

The results obtained from the targeted capillary sequencing showed that the variant was present in homozygous from in both the affected siblings and both the parents carried the heterozygous form of the variant. This validates our exome sequencing results. Also, both the daughter of the male patient were found to be the carrier of the disorder as they bear the heterozygous form of the variant. The results are depicted in Fig 4.12. The genetic region encompassing the variants were amplified by using both the forward and reverse primers separately.

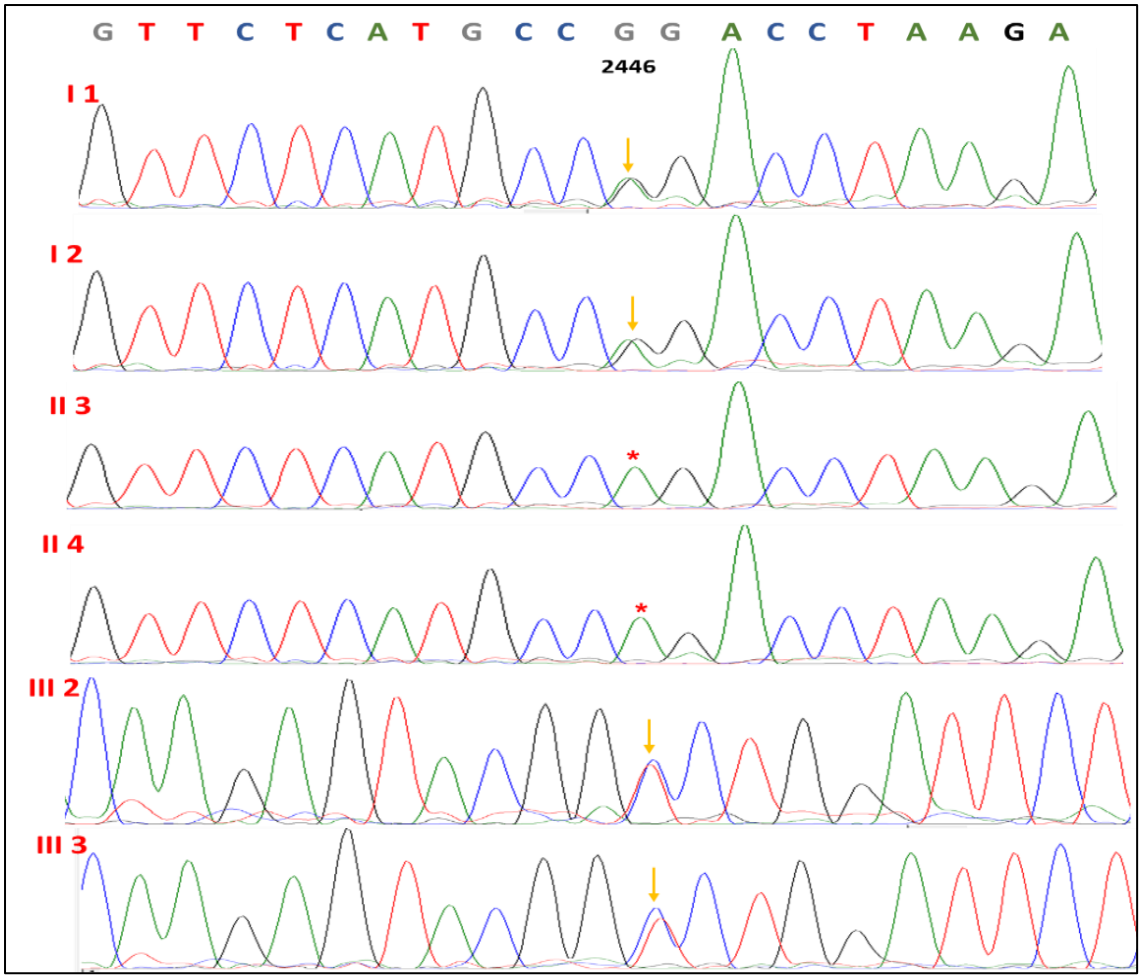


**Fig 4.12: Chromatogram derived from targeted capillary sequencing of family members of 2N VWD case.** The mutation **Chr12: 6145654 G>A, c.C2446T:p.R816W** is present in exon 19 of *VWF*. Affected individuals (II 3 & II 4) are marked with asterisks. The variant is indicated by the arrow. The variant is found to be in heterozygous form in both the father and mother of the affected siblings and also in both the daughter of II 3 as depicted in III 2 & III 3. The shown fragments were amplified by forward primer in I 1, I 2, II 3 & II 4 whereas III 2 & III 3 were amplified by reverse primer. I 1- TU30- Father, I 2- TU31- Mother, II 3- TU01- Patient, II 4- TU18- Patient, III 2- TU33- Daughter of TU01, III 3- TU34- Daughter of TU01

## 4.5 Bioinformatics Analysis of Sequencing Data of Factor X Deficiency Cases (TU03 and TU25)

The bioinformatics analysis of the whole exome sequencing data of the siblings affected with factor X deficiency was performed following the similar approaches as done for 2N VWD cases.

### 4.5.1 Quality Check and Trimming

The quality of the raw sequencing data was done by using FastQC tool and Trimmomatic was used to trim the raw data.

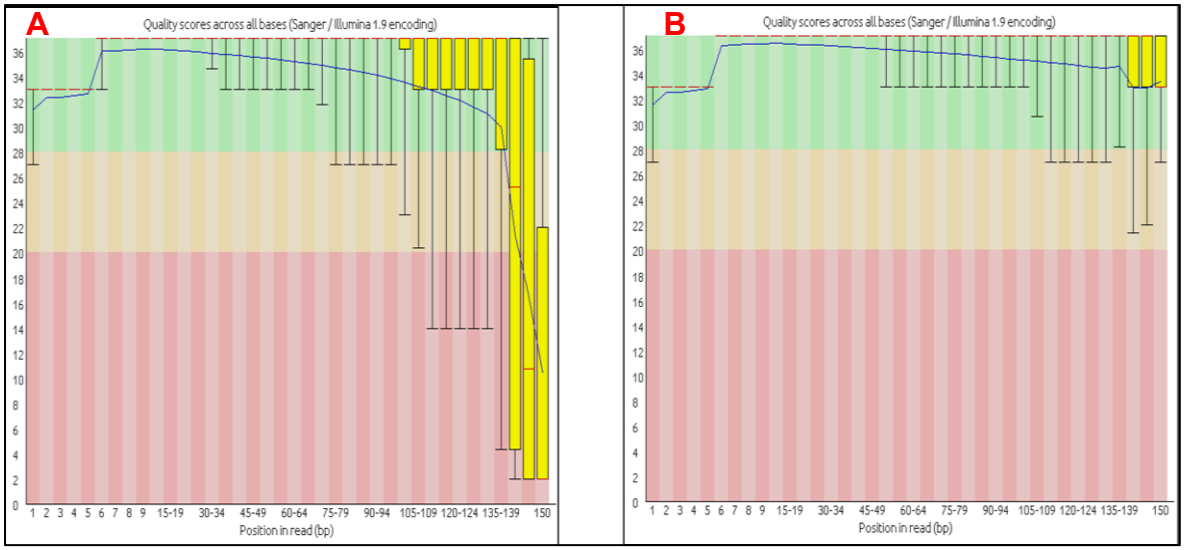


**Fig 4.13: Quality plot of the raw sequencing reads of TU03.** (A) The Box Whisker plot representing the data quality before trimming. The Y-axis represents the Phred quality scores and X-axis represents the position of base in the read. The bases towards the end of the reads were having poor quality as represented by low Phred scores (red region). The bases present on the middle region (green and orange region) in the read are of good and reasonable qualities. (B) The BoxWhisker plot representing the data quality after trimming**.** Poor quality scores towards the end of the reads were removed after trimming. All the bases with good and reasonable qualities were retained.

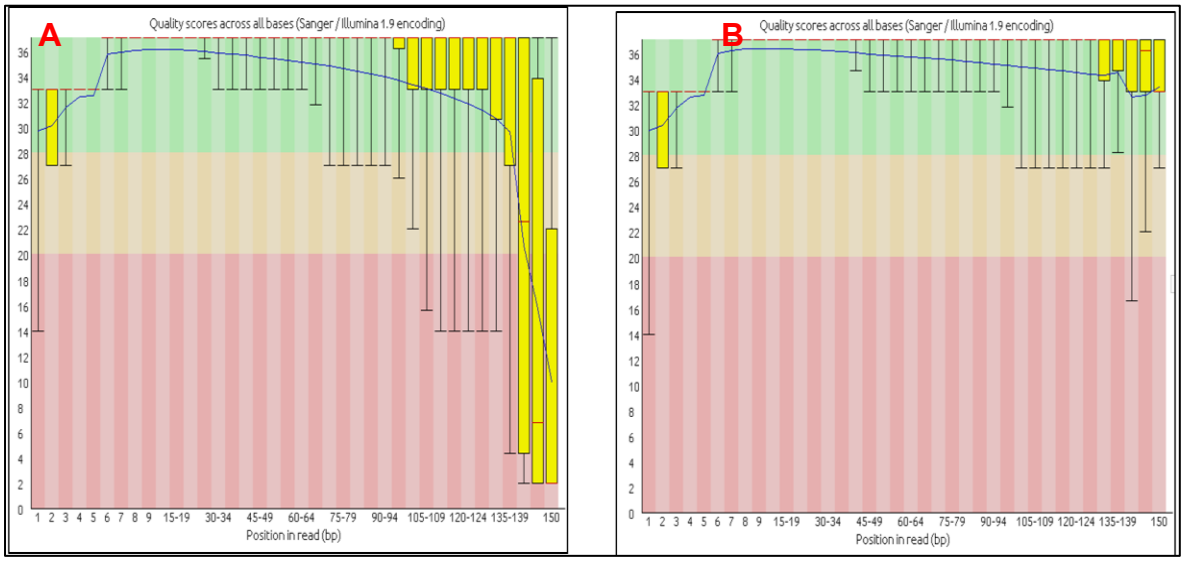


**Fig 4.14: Quality plot of the raw sequencing reads of TU25.** (A) The Box Whisker plot representing the data quality before trimming. The Y-axis represents the Phred quality scores and X-axis represents the position of base in the read. The bases towards the end of the reads were having poor quality as represented by low Phred scores (red region). The bases present on the middle region (green and orange region) in the read are of good and reasonable qualities. (B) The BoxWhisker plot representing the data quality after trimming**.** Poor quality scores towards the end of the reads were removed after trimming. All the bases with good and reasonable qualities were retained.

The information obtained from the FastQC report is summarized in Table 4.13.

Approximately 4% of the total reads were trimmed in both cases and further proceeded for alignment. The mean sequence quality score of the reads was 36 which implies low error rate in base calling.

Table 4.13: Summary of FastQC report of TU03 and TU25

| Parameter | TU03 | | TU25 | |
|---|---|---|---|---|
| | **Before Trimming** | **After Trimming** | **Before Trimming** | **After Trimming** |
| Total Sequences | 56429198 | 54362016 (96.34%) | 93526058 | 89979253 (96.21%) |
| Mean sequence Quality score | 36 | 36 | 36 | 36 |
| Sequences flagged as poor quality | 0 | 0 | 0 | 0 |
| Sequence length | 140-150 | 36-150 | 140-150 | 36-150 |
| %GC | 46 | 45 | 45 | 44 |

## 4.5.2 Alignment

The brief description of the alignment done by using Stampy with BWA-MEM is summarized in Table 4.14. The very high mapping percentage implies that sequencing was happened perfectly without any errors.

Table 4.14: Mapping summary of TU03 and TU25

| | **QC Passed Reads** | **Mapped Reads** | **Mapped %** |
|---|---|---|---|
| **TU03** | 54362016 | 54354053 | 99.99 % |
| **TU25** | 89979253 | 89965185 | 99.98 % |

## 4.5.3 Variant Calling

The brief description of variants called by using Platypus is shown in Table 4.15. Only 28.64% and 22.5% of the total variants represented the variations present within the exome in TU03 and TU25 respectively.

Table 4.15: Summary of total variants found in TU03 and TU25

| | **TU03** | **TU25** |
|---|---|---|
| **Total variants** | 74740 | 96213 |
| **Exonic Variants** | 21405 | 21669 |

## 4.5.4 Variant Annotation and Prioritization

Variant annotation and prioritization for FXD cases was done by considering the same parameters as in 2N VWD case which is described in detail in materials and methodology section. The variants which were prioritized are presented on Fig 4.15.

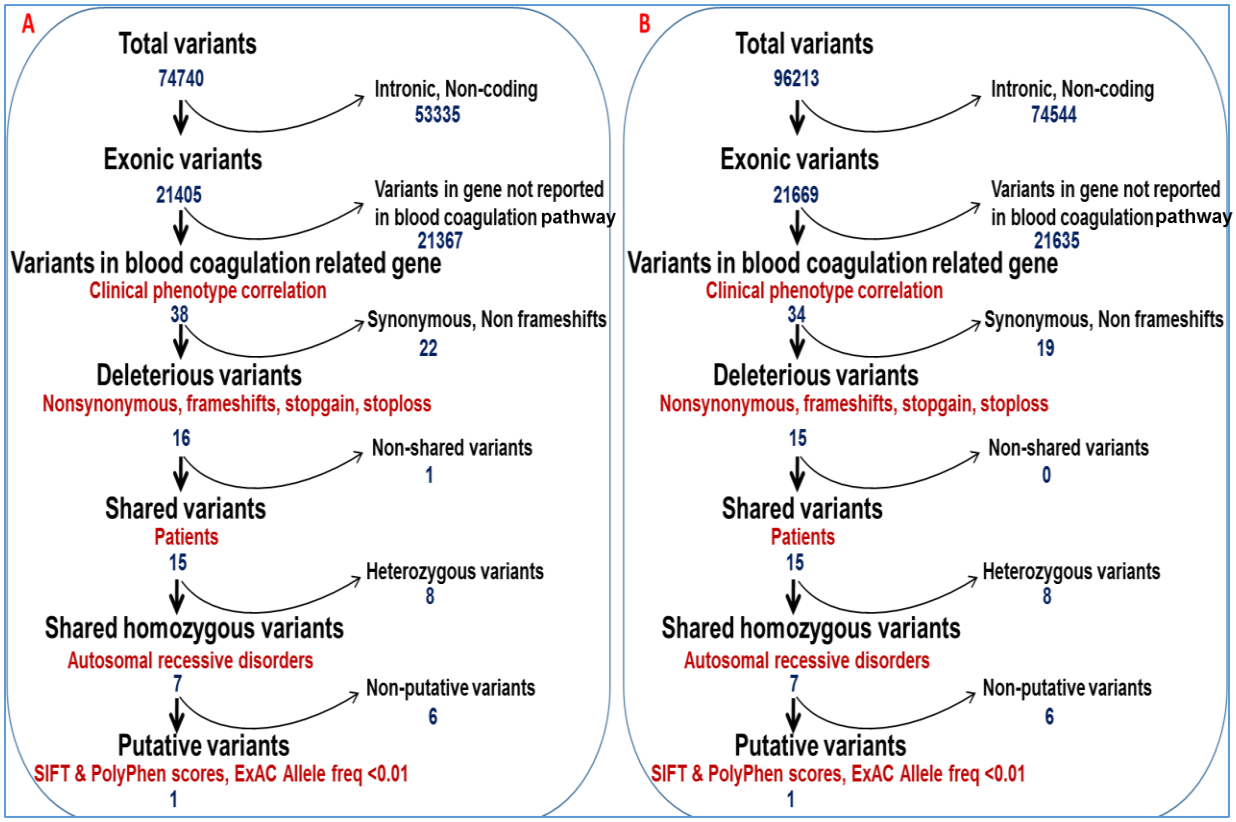


**Fig 4.15: Pipeline used for variants sorting.** (A) Variants sorting in TU03 and (B) variants sorting in TU25. Different variations show almost same proportion of prevalence in both the siblings. Finally only one variant was prioritized in both the siblings which was the putative variant associated with the disorder.

The exonic variants were of different types as depicted in Table 4.16.

Table 4.16: Summary of various exonic mutations found in TU03 and TU25

| **Types of mutation** | **Mutations in TU03** | **Mutations in TU25** |
|---|---|---|
| Frameshift deletion | 99 | 103 |
| Frameshift insertion | 64 | 68 |
| Frameshift substitution | 7 | 9 |
| Non frameshift deletion | 191 | 195 |
| Non frameshift insertion | 138 | 138 |
| Non frameshift substitution | 638 | 674 |
| Non synonymous SNVs | 9245 | 9327 |
| Synonymous SNVs | 10474 | 10606 |
| Stopgain | 71 | 66 |
| Stoploss | 6 | 4 |
| Unknown | 472 | 479 |

These different variations can be presented in pie-chart as shown in Fig 4.16.

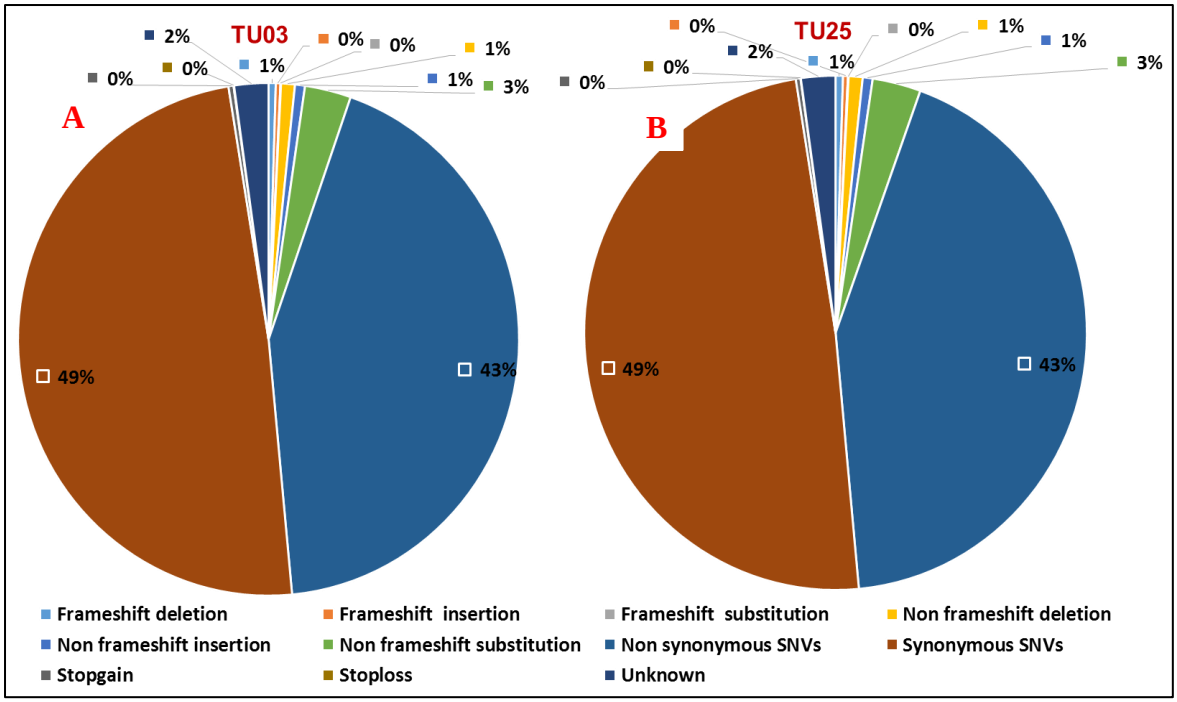


**Fig 4.16: Pie chart representing the relative percentage of various types of mutations.** (A) In TU03 and (B) in TU25. Figures show both the siblings bear different mutations with almost same proportion. Synonymous SNVs cover the highest percentage (49% in both) followed by nonsynonymous SNVs (43% in TU03 and 43% in TU25). Frameshift (deletion, insertion and substitution), stopgain and stopless mutations show very less prevalence whereas 2% of total mutations are unknown.

### 4.5.4.1 Variants Related To Rare Bleeding Disorders

Variant prioritization was done by considering the exonic variants present in the genes which encode various clotting factors and elements, as mentioned in methods and methodology chapter. In total 38 and 24 variations respectively in TU03 and TU25 which may be associated with the phenotype of the patients.

Table 4.17: Types of mutations present on genes associated with inherited bleeding disorders in TU03 and TU25

| Mutation Type | Mutations in TU03 | Mutation in TU25 |
|---|---|---|
| Non synonymous SNVs | 16 | 15 |
| Synonymous SNVs | 22 | 19 |
| Total | 38 | 34 |

Further sorting revealed a variation shared between the siblings (highlighted in red in Table 4.18 and 4.19) which was deleterious and probably damaging as predicted by SIFT and Polyphen2_HVAR scores.

Table 4.18: Details of Nonsynonymous SNVs of TU03

| Chr | Position | Ref | Alt | Gene. Ref Gene | AAChange.refGene | SIFT_score_pred | Polyphen2_HVAR_score_pred | Hom/Het |
|---|---|---|---|---|---|---|---|---|
| chr1 | 169498975 | T | C | *F5* | F5:NM_000130:exon16:c.A5290G:p.M1764V | 1 T | 0 B | Het |
| chr1 | 169510118 | G | A | *F5* | F5:NM_000130:exon13:c.C4210T:p.P1404S | 0.1 D | 0 B | Het |
| chr1 | 169511555 | T | C | *F5* | F5:NM_000130:exon13:c.A2773G:p.K925E | 1 T | 0 B | Het |
| chr1 | 169511734 | T | C | *F5* | F5:NM_000130:exon13:c.A2594G:p.H865R | 0.4 T | 0 B | Het |
| chr1 | 169511755 | T | C | *F5* | F5:NM_000130:exon13:c.A2573G:p.K858R | 0.5 T | 0 B | Het |
| chr1 | 169519049 | T | C | *F5* | F5:NM_000130:exon10:c.A1601G:p.Q534R | 1 T | -- | Hom |
| chr1 | 197031021 | C | T | *F13B* | F13B:NM_001994:exon3: c.G344A:p.R115H | 0.2 T | 0 B | Hom |
| chr5 | 176831826 | C | G | *F12* | F12:NM_000505:exon7:c.G619C:p.A207P | 0.4 T | 0 B | Hom |
| chr6 | 6174866 | G | A | *F13A1* | F13A1:NM_000129:exon12: c.C1694T:p.P565L | 0.7 T | 0 B | Het |
| chr11 | 46745003 | C | T | *F2* | F2:NM_000506:exon6:c.C494T:p.T165M, | 0 D | 0.1 B | Het |
| chr12 | 6127891 | C | A | *VWF* | VWF:NM_000552:exon28: c.G4693T:p.V1565L | 0.3 T | 0 B | Het |
| chr12 | 6128443 | T | C | *VWF* | VWF:NM_000552:exon28: c.A4141G:p.T1381A | 1 T | 0 B | Hom |
| chr12 | 6143984 | T | C | *VWF* | VWF:NM_000552:exon20: c.A2555G:p.Q852R | 1 T | 0 B | Hom |
| chr12 | 6172202 | T | C | *VWF* | VWF:NM_000552:exon13: c.A1451G:p.H484R | 1 T | 0 B | Hom |
| chr13 | 113783907 | T | C | *F10* | *F10*:NM_000504:exon2:c.T212C:p.F71S | 0 D | 1 D | Hom |
| chr18 | 57000465 | C | T | *LMAN1* | LMAN1:NM_005570:exon11: c.G1232A:p.S411N | 0.5 T | 0 B | Het |

Table 4.19: Details of Nonsynonymous SNVs in TU25

| **Chr** | **Start** | **Ref** | **Alt** | **Gene. Ref Gene** | **AAChange.refGene** | **SIFT_ score _pred** | **Polyphen2 _HVAR_ score_ pred** | **Hom /Het** |
|---|---|---|---|---|---|---|---|---|
| chr1 | 169498975 | T | C | *F5* | F5:NM_000130:exon16: c.A5290G:p.M1764V | 1 T | 0 B | Het |
| chr1 | 169511555 | T | C | *F5* | F5:NM_000130:exon13: c.A2773G:p.K925E | 1 T | 0.001 B | Het |
| chr1 | 169511734 | T | C | *F5* | F5:NM_000130:exon13: c.A2594G:p.H865R | 0.4 T | 0 B | Het |
| chr1 | 169511755 | T | C | *F5* | F5:NM_000130:exon13: c.A2573G:p.K858R | 0.5 T | 0 B | Het |
| chr1 | 169519049 | T | C | *F5* | F5:NM_000130:exon10: c.A1601G:p.Q534R | 1 T | - - | Hom |
| chr1 | 197031021 | C | T | *F13B* | F13B:NM_001994:exon3: c.G344A:p.R115H | 0.2 T | 0.002 B | Hom |
| chr5 | 176831826 | C | G | *F12* | F12:NM_000505:exon7: c.G619C:p.A207P | 0.4 T | 0 B | Het |
| chr6 | 6174866 | G | A | *F13A1* | F13A1:NM_000129:exon 12:c.C1694T:p.P565L | 0.7 T | 0.002 B | Het |
| chr11 | 46745003 | C | T | *F2* | F2:NM_000506:exon6: c.C494T:p.T165M | 0 T | 0.08 B | Het |
| chr12 | 6127891 | C | A | *VWF* | VWF:NM_000552:exon 28:c.G4693T:p.V1565L | 0.3 T | 0.002 B | Het |
| chr12 | 6128443 | T | C | *VWF* | VWF:NM_000552:exon 28:c.A4141G:p.T1381A | 1 T | 0.001 B | Hom |
| chr12 | 6143984 | T | C | *VWF* | VWF:NM_000552:exon 20:c.A2555G:p.Q852R | 1 T | 0 B | Hom |
| chr12 | 6172202 | T | C | *VWF* | VWF:NM_000552:exon 13:c.A1451G:p.H484R | 1 T | 0.025 B | Hom |
| chr13 | 113783907 | T | C | *F10* | *F10*:NM_000504:exon2: c.T212C:p.F71S | 0 D | 0.998 D | Hom |
| chr18 | 57000465 | C | T | *LMAN1* | LMAN1:NM_005570:exon 11: c.G1232A:p.S411N | 0.5 T | 0.005 B | Het |

Finally, the putative variant was visualized in Integrative Genomics Viewer (IGV) as shown in Fig 4.17.

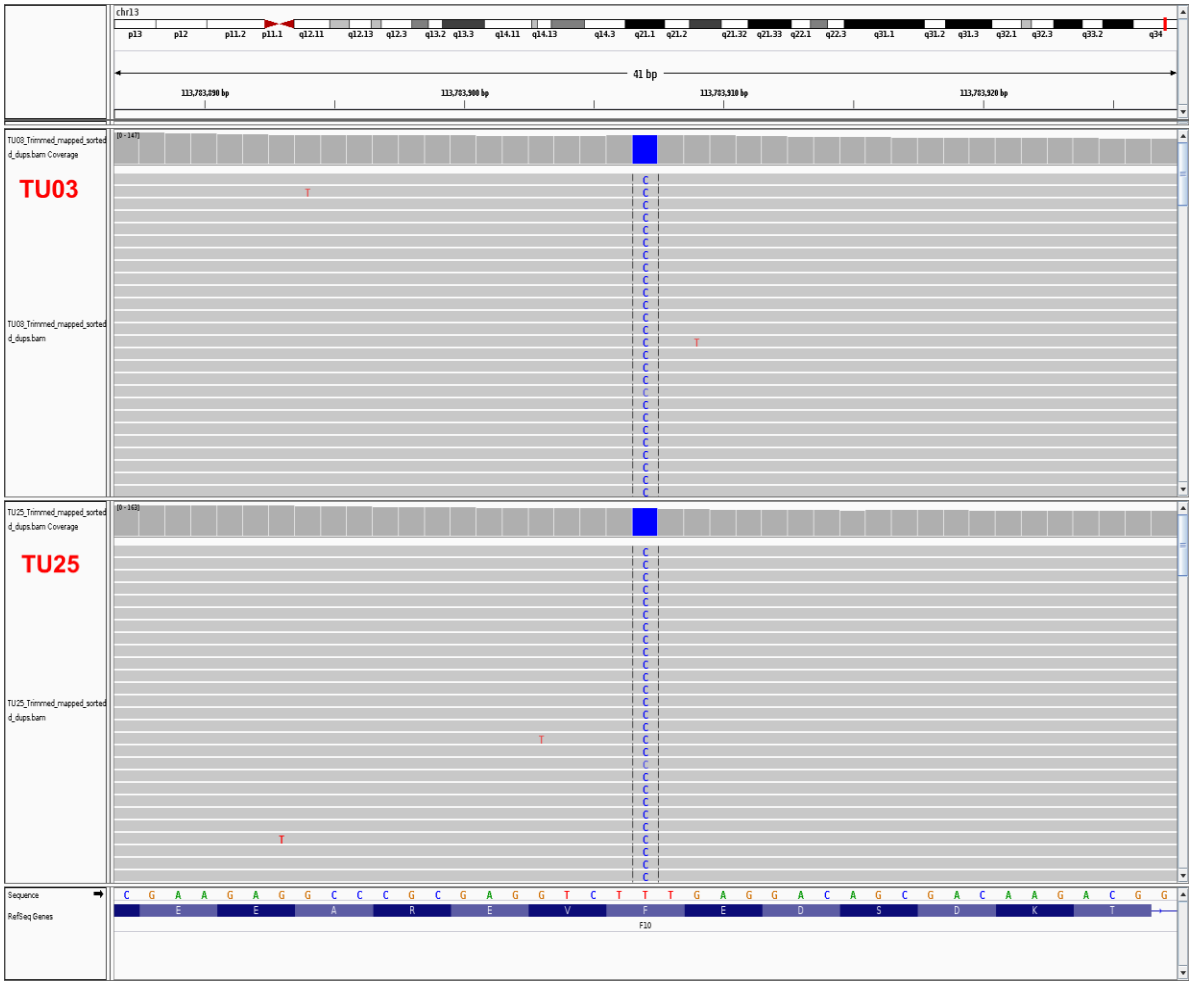


**Fig 4.17: IGV Snapshot of the variant p.F71S.** It shows both the sibling (upper –TU03 and lower –TU25) are homozygous for the mutation (Chr13: 113783907 T>C, c.T212C). Coverage for the variant was 119X and 127X respectively in TU03 and TU25.

### 4.5.4.1.1 Pathogenicity Prediction of the Variants

We considered all the parameters as described in details in methods and methodology section and finally identified one variant (c.T212C:p.F71S in exon 2 of *F10*) with deleterious and/or damaging effect on the protein structure or function in both the siblings as shown in above Tables 4.20. The variant was inherited in homozygous form in both the siblings. The allele frequency for the variant was found to be 0.0000008242 in general population in the Exome Aggregation Consortium. The SIFT and Polyphen2_HVAR scores were found to be 0 and 0.998 respectively for the substitution of phenylalanine by serine at 71 aa position in FX.

Table 4.20: Putative mutation based on SIFT and Polyphen2 score in TU03 and TU25

| Chr | Position | Ref | Alt | Gene.refGene | AAChange.refGene | Exonic function change | SIF_pred | Polyphen2_HVAR_pred | hom/het | ExAC AF |
|---|---|---|---|---|---|---|---|---|---|---|
| chr13 | 113783907 | T | C | *F10* | *F10*:NM_000504:exon2:c.T212C:p.F71S | Nonsynonymous SNV | 0 D | 0.998 D | hom | 8.242E-6 |

The variant is localized in GLA domain of the FX protein as shown in Fig 4.18. GLA domain participates in calcium and phospholipid membrane binding of the FX protein.

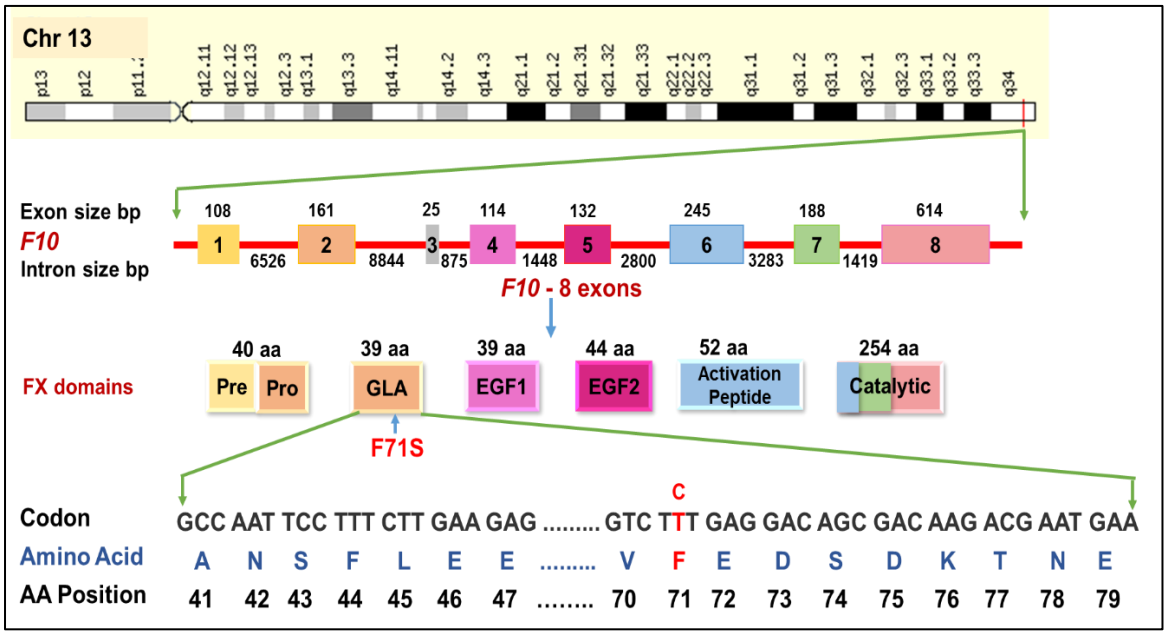


**Fig 4.18: Localization of variant F71S in FX protein.** The eight exons are represented by respective numerals. The variant F71S is localized in GLA domain of the FX protein as shown by an arrow.

#### 4.5.4.1.1.1 Allele Frequency of the Variant F71S in Different Population

The allele frequency of the variant F71S was checked separately in different population in ExAC. Only 1 allele was found to be reported till date in Latino population in ExAC as presented in Table 4.21. This reflects the fact that F71S is a very rare variant found in patients of factor X deficiency.

Table 4.21: Allele Frequency of the variant F71S in different population (soure ExAC, accessed on 11-07-2016)

| Population | Allele Count | Allele Number | No. of Homozygotes | Allele Frequency |
|---|---|---|---|---|
| Latino | 1 | 8648 | 0 | 0.0001156 |
| European (Non-Finnish) | 0 | 10384 | 0 | 0 |
| African | 0 | 6614 | 0 | 0 |
| East Asian | 0 | 66692 | 0 | 0 |
| European (Finnish) | 0 | 11574 | 0 | 0 |
| Other | 0 | 908 | 0 | 0 |
| South Asian | 0 | 16512 | 0 | 0 |
| Total | 1 | 121332 | 0 | 8.242E-6 |

#### 4.5.4.1.1.2 Interspecies Analysis of F71 Residue

The conservation of F71 residue in Factor X protein among the vertebrates was checked using an online tool (NCBI HomoloGene tool) in order to observe the evolutionary significance of the F71 in FX protein. It was found F71 is a conserved amino acid among the various vertebrates. This evolutionary conservation of F71 reflects the fact that F71 has a vital role in maintaining protein structure and function.

Table 4.22: Conservation of F71 residue in FX protein among the vertebrates (NCBI HomoloGene tool). **F** shows the conserved position of Phe at 71 position except for *D. rerio* which is at 68 position.

| Organisms | Amino acid sequences within the range mentioned | |
|---|---|---|
| *Homo sapiens* | 51 GHLERECMEETCSYEEAREV**F**EDSDKTNEFWNKYKDG-DQCETSPCQNQG | 99 |
| *Pan troglodytes* | 51 GHLERECMEETCSYEEAREV**F**EDSDKTNEFWNKYKDG-DQCETSPCQNQG | 99 |
| *Macaca mulatta* | 51 GNLERECMEETCSYEEAREVLEDSDKTNEFWNKYKDG-DQCETSPCQNEG | 99 |
| *Canis lupus* | 51 GNLERECMEETCSFEEAREV**F**EDTAKTMEFWNKYKDG-DQCESSPCQNQG | 99 |
| *Bos Taurus* | 51 GNLERECLEEACSLEEAREV**F**EDAEQTDEFWSKYKDG-DQCEGHPCLNQG | 99 |
| *Mus musculus* | 51 GNLERECMEEICSYEEVREI**F**EDDEKTKEYWTKYKDG-DQCESSPCQNQG | 99 |
| *Rattus norvegicus* | 51 GNLERECVEEICSFEEAREV**F**EDNEKTTEFWNKYEDG-DQCESSPCQNQG | 99 |
| *Gallus gallus* | 51 GNIERECNEERCSKEEAREA**F**EDNEKTEEFWNIYVDG-DQCSSNPCHYGG | 99 |
| *Danio rerio* | 48 GNMERECIEERCNYEEAREI**F**EDVKKTDEFWHKYVDGKNACLSHPCVNGG | 97 |
| *Xenopus tropicalis* | 50 GNLERECYEERCSLEEAREV**F**ENEEQTREFWSKYFDG-DQCQSNPCQYGG | 98 |

## 4.5.5 Validation by Capillary Sequencing

The putative variant associated FXD, prioritized from the computational analysis of the whole exome sequencing data of the affected individuals was technically validated by targeted capillary sequencing as described in materials and methodology section. The genetic regions encompassing the variant were PCR amplified by using targeted primers (details in appendix section). The segregation of the variant was checked among the family members of the patients.

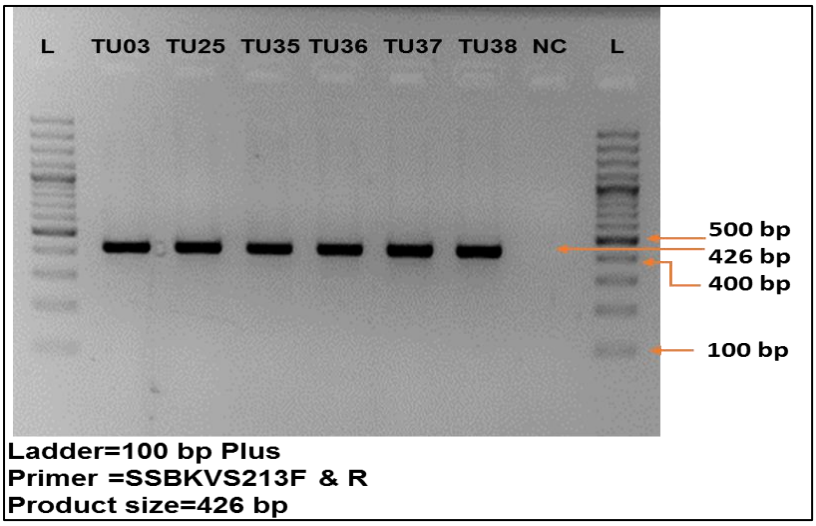


**Fig 4.19: PCR amplification of genetic region encompassing the putative variant Chr13: 113783907 T>C, c.T212C:p.F71S present in exon 2 of *F10*.** Product size of the amplified region was 426 bp. L- 100 bp Plus ladder (GeneRuler, Thermo Scientific, USA).

The results obtained from the targeted capillary sequencing showed that the variant p.F71S was present in homozygous from in both the affected siblings. Both the parents carried the heterozygous form of the variant. This validates our exome sequencing results. The results are depicted in Fig 4.20. The genetic region encompassing the variants were amplified by using both the forward and reverse primers separately. Thus, in some cases the variants are in reverse complement forms such as in Fig 4.20 II 4.

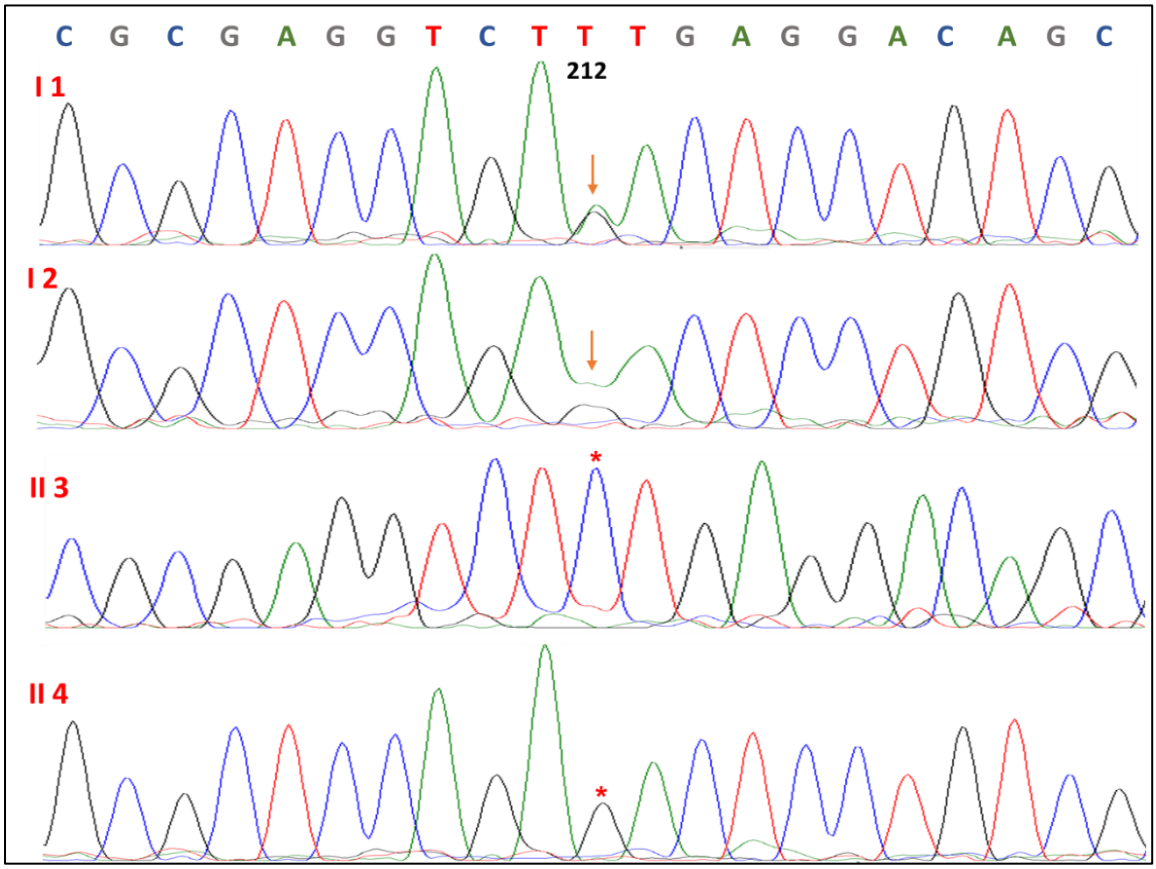


**Fig 4.20: Chromatogram derived from targeted capillary sequencing of family members of FXD case.** The mutation **Chr13: 113783907 T>C, c.T212C:p.F71S** in present in exon 2 of *F10*. Affected individuals (II 3 & II 4) are marked with asterisks. The variant is indicated by the arrow. The fragments were amplified by forward primer in II 3 whereas I 1, I 2 and II 4 were amplified by reverse primer. The variant is found to be in heterozygous form in both the father and mother of the affected siblings as depicted in I 1 & I 2 respectively. I 1- TU35- Father, I 2- TU36- Mother, II 3- TU03- Patient, II 4- TU25- Patient.

# CHAPTER 5
# DISCUSSION

Inherited bleeding disorders are life threatening conditions arising from genetic abnormalities in genes encoding and regulating the clotting factors (Peyvandi *et al*., 2002). The condition is manifested in severe, moderate and mild forms depending upon the loci of genetic lesions with in the genes (Graw *et al*., 2005) which determine the fate of protein (clotting elements) encoded. Generally laboratory based studies are used for the diagnosis such bleeding disorders. However, discrepancies on laboratory based diagnosis have been well reported. To confirm the laboratory results and to decipher the genetic cause of the disorder, the genetic screening of the genes encoding and regulating various coagulation factors is essential. Till date traditional approaches for genetic screening has been employed for diagnosis and prognosis of bleeding disorders. Recently, next generation sequencing has been emerged as an excellent tool for the cost-effective and timely diagnosis for the bleeding disorders (Peyvandi *et al.*, 2013). The availability of conventional and next generation sequencing allows one to direct the clinical treatment in right direction by translational application of the genetic information in genetic characterization of disease causing variants, carrier testing and prenatal diagnosis. The information from genotyping can also help in differentiation of phenocopies (Peyvandi *et al.*, 2013).

The main objective of this study was to identify the genetic lesions in the genes encoding and/or regulating various clotting factors/elements in the Nepalese patients with rare bleeding disorders (RBDs) by whole exome sequencing (WES). The information obtained from such genetic study can be used for the translational applications such as prenatal diagnosis, carrier detection, genetic counselling etc.

## 5.1 Whole Exome Sequencing (WES)

Whole exome sequencing (WES) of four rare bleeding disorders' patients of Nepalese origin was performed. The results obtained from the WES data analysis were then validated by automated capillary (Sanger) sequencing. This is actually the technical validation of Next Generation Sequencing by First Generation Sequencing. Besides the major limitations like low throughput, very high cost and too much time consuming, Sanger Sequencing is the most accurate method for sequencing, accuracies as high as 99.999% (Shendure and Ji, 2008). Because of its maximum accuracy and easy availability, Sanger Sequencing is still used as a gold standard for sequencing in many labs especially for small genomics researches (Rizzo and Buck, 2012). In the counterparts, Next Generation Sequencing has the major advantages of high throughout, comparatively very lower cost and much less time requirement. But these NGS platforms are associated with the higher error rate (~0.1-15%) than the standard Sanger Sequencing (Goodwin *et al*., 2016). This prompted us for the validation of NGS results by capillary sequencing.

The samples utilized were from siblings clinically diagnosed with Factor X deficiency and Type 2 Normandy von Willebrand disease. The unavailability of custom panel for genetic screening of bleeding disorders and lack of specialized diagnostic setups for rare bleeding disorders prompted us to use whole exome sequencing as a tool for genetic screening in the present cases of rare bleeding disorders.

### 5.1.1 WES in Illumina Platform

WES was performed in Illumina platform in HiSeq 2500. Illumina works under Sequencing by Synthesis (SBS) with cyclic reversible termination (CRT) technology (Morozova and Marra, 2008). The principle of sequencing is: Optimally fragmented DNA tagged with indexed adapters are loaded onto a glass flow-cell (FC) containing specific oligos which act as baits for the DNA fragments to attach to and are then amplified using bridge PCR (Bridge Amplification) in a Cluster Station (cBot) (Liu *et al*., 2011). After the amplification step, generally 35 cycles of isothermal amplification (Buermans and Den Dunnen, 2014), the flow cell with more than 40 million clusters is produced (Morozova and Marra, 2008), wherein each cluster is composed of approximately 1000 clonal copies of a single template molecule known as polonies (Buermans and Den Dunnen, 2014). The flow-cell, containing the clonally amplified DNA fragments (clusters), is then loaded to the sequencing machine where the fragments are sequenced using Sequencing by Synthesis technology. Each cycle of sequence interrogation consists of single-base extension with a modified DNA polymerase and a mixture of four modified nucleotides. The nucleotides are modified in two ways viz. they are 'reversible terminators', bearing a chemically cleavable moiety at the 3' hydroxyl position which allows only a single-base incorporation to occur in each cycle; and they are 'fluorescently labelled' each of which corresponds to the identity of different nucleotide (Shendure and Ji, 2008). CRT uses reversible terminators in a cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage (Metzker, 2010). During each cycle of sequencing, a mixture of all four individually fluorescently labelled and 3'- end blocked deoxynucleotides (dNTPs) are added. The polymerase then incorporates a single dNTP to each elongating complementary strand and the unbound dNTPs are washed away. Since all four reversible terminator bound-dNTPs are present as single, separate molecules, natural competition minimizes incorporation bias. The surface is imaged to identify which dNTP was incorporated at each cluster. Base calls are made directly from signal intensity measurements during each cycle, which greatly reduces raw error rates associated with the base calling even within repetitive sequence regions or homopolymers. The fluorophore and blocking group can then be chemically removed and a new cycle of sequencing can begin (Fuller et al., 2009; Goodwin et al., 2016; Liu et al., 2011; Mardis, 2008; Kircher et al., 2011) .

## 5.2 Bioinformatics Analysis of WES Data

Illumina sequencing instruments generate raw sequenced reads in *.bcl (base call) format as primary sequencing output. As *.bcl files are not compatible with downstream analysis

applications these per-cycle *.bcl files are converted to compressed FASTQ files as *.fastq.gz format by CASAVA (Consensus Assessment of Sequence And VAriation) which is a part of Illumina's sequencing analysis software. FASTQ files, simply are the FASTA files with associated Quality scores. Also, the multiplexed samples are demultiplexed at this stage. The reads are then proceeded for the bioinformatics analysis and interpretation. Ironically, the greatest strength of NGS – the huge volume of sequence data – has also emerged as the key limitation as the interpretation of the sequence data of gigabase order is not a trivial task (Rizzo and Buck, 2012).

### 5.2.1 Data QC and Trimming

Own data analysis pipeline for the bioinformatics analysis of the sequenced data has been developed. The raw sequence data might contain sequence artifacts like poor quality reads, adapter contamination, over-representation of sequence or sequence bias, ambiguous sequences etc. due to instrument failures or chemical process errors (Gandhi and Scaria, 2016). Hence it is very essential for the quality check of the raw data before analysis and interpretation. Thus, at first the raw sequenced files in FASTQ format were proceeded for quality check using FastQC (Andrews S., 2010). Based on the report generated by FastQC; the reads with Phred quality score less than 20, read length below 36 and the adapters sequence were removed by using Trimmomatic (Bolger et al., 2014). The base calling accuracy is measured by Phred quality score (Q). Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P); $Q = -10 \log_{10} P$. It is assumed that when sequence quality reaches Q30, virtually all of the reads will be perfect, having zero errors and ambiguities. This is why Q30 is considered a benchmark for quality in NGS. After trimming the reads with bad quality scores and removing the adapter sequences, the filter passed reads were then subjected to alignment against the human reference genome (hg19/GRCh37).

### 5.2.2 Sequence Alignment

Alignment is the step of matching each of the short nucleotide reads to positions on a reference genome (Wang *et al*., 2013). The hybrid mode of BWA-Stampy was used for the efficient and accurate mapping as Stampy and BWA (Burrows–Wheeler Alignment) are complementary to each other. BWA is a well-known short reads alignment tool (Bao *et al*., 2014) which locates mismatches and gaps quickly (Li and Durbin, 2009) whereas Stampy is known for its sensitivity (Lunter and Goodson, 2011). The resulting sequence alignment is stored in a SAM (sequence alignment/map) file (Li *et al.,* 2009a).

### 5.2.3 Variant Calling

After mapping the sequenced reads to the reference genome, post-alignment processing steps are necessary to minimize the artifacts that may affect the quality of downstream processes. The output SAM file must first be converted to a compressed BAM (Binary Alignment/Map) file format, followed by its sorting according to the coordinates and

marking PCR duplicates which could lead to dubious results. Samtools was used for converting .sam file to .bam file and sorting and indexing of the .bam file (Li *et al.*, 2009a). The index file was prepared so that one can have a fast look-up of data in a SAM/BAM file by various softwares. PCR duplicates were marked by Picard which were later removed by Platypus. Next step is variant calling – comparing the aligned sequences with known sequences to determine which positions deviate from the reference position. This process produces a list of positions or calls recorded in a VCF (variant call format) file. We used Platypus which is reported for high sensitivity and specificity for SNPs, indels and complex polymorphisms (Rimmer *et al*., 2014) for variant calling. Variants can be simply be defined as the deviations from the reference genome and the variants detected could be either true variants or just sequencing artefacts arising due to sequencing errors, sample contamination or insufficient variant coverage (Gandhi and Scaria, 2016). Variants may be in the form of single nucleotide variants, smaller insertions or deletions (indels), or larger structural variants such as transversions, trans-locations, and copy number variants (CNV) (Rudy, 2011).

### 5.2.4 Variant Annotation

Once the variants are called then they are annotated and prioritized. This reduces the tens of thousands of variants to a smaller set (Bao *et al*., 2014). Annotation involves querying known information about each variant that is detected. Annotation may reveal information associated with the variants such as whether the variant is an already-known single nucleotide polymorphism, the functional effect of the variant has already been predicted, the function or activity of the gene in question is already known, or even whether an associated disease has been identified (Dolled-Filhart *et al.*, 2013). It is important to annotate variants with attributes such as genomic feature, exonic function and amino acid changes (Gandhi and Scaria, 2016). We used ANNOVAR (ANNotate VARiation) (Wang *et al*., 2010) to functionally annotate the variants. During annotation, considering the causative variants to be rare, variants present in public databases such as HapMap, 1000 Genomes Project, GWAS, Clinvar, dbSNP etc. were excluded.

### 5.2.5 Variant Prioritization

Following the annotation the variants were prioritized based on the pathogenicity, clinical phenotype of the samples, familial segregation (as both the disorders were familial cases) and rare allele frequency. For variants prioritization we devised our own strategy as shown in Fig. 3.6. First of all, considering the putative mutation to be exonic, only the exonic variants were prioritized. Then since the samples were rare bleeding disorders patients, the putative mutation must be present within the genetic loci of those genes which codes the various blood coagulation factors and the elements which regulate those clotting factors. By literature search, we found total of 19 genes those are associated with rare bleeding disorders. So, we searched for the variants within those genes. Next, since all the variants are not disease causing, only the deleterious variants such as nonsynonymous,

frameshifts, stopgain and stoploss variants were considered. Then we looked for the shared homozygous and compound homozygous variants among the siblings as both the RBDs were familial cases and are autosomal recessive disorders (Bamshad *et al*., 2011; Gilissen *et al.*, 2012). Finally putative variant was identified by assessing the biological effect of variant based on literatures and in silico tools viz. SIFT and Polyphen2 (Ku *et al.*, 2012) and by visualizing the variant. SIFT (Sorting Intolerant From Tolerant) predicts relying on the presumption that amino acid residues that are essential for protein function should be evolutionally conserved by natural selection and thus, mutation resulting in alteration on the conserved residues are more likely to be deleterious (Kumar *et al*., 2009). On the other hand PolyPhen2 (Polymorphism Phenotyping v2) algorithm predicts the potential impact of the alteration on the structure and function of human protein based on protein sequence, phylogenetic and structural information (Adzhubei *et al*., 2013). Data visualization is very important as it differentiates the true variants and sequencing artifacts based on the coverage (minimum of 30X to be a true variant) (Gandhi and Scaria, 2016).

## 5.3 Mutation in Family 1: Type 2 Normandy von Willebrand disease

The first family consisted of two siblings affected from a rare bleeding disorder named 2N VWD. 2N (N stands for Normandy – the birth place of first patient described) VWD is the subtype of VWD in which binding of FVIII to VWF is impaired (Lillicrap, 2007). Thus, the clinical presentation of 2N VWD mimics those of hemophilia A. However, 2N VWD is an autosomal recessive disorder which affects male and female equally (Goodeve, 2010) in contrast to X-linked HA which exclusively affects males.

### 5.3.1 WES and Bioinformatics Analysis of 2N VWD

Whole exome sequencing of the siblings designated as TU01 (brother) and TU18 (sister) was performed. The raw reads were subjected to FASTQC for quality checks and then reads with bad quality were trimmed along with the removal of adapters by using Trimmomatic. Total reads sequenced for TU01 and TU18 were 48 and 58 million respectively. Among these reads approximately 2 million in each case were trimmed. The criteria for trimming were: Q <20 and read length <36 bp. The mean sequence quality score was found to be 36 in both cases which was quite above the benchmark (Q30) in NGS. After removing the reads shorter than 36 bp, only the reads of length ranging from 36 bp to 150 bp were retained. These QC passed reads were then subjected to alignment against the human reference genome. 99.98% of the reads were mapped in both the cases, which implies the reliability of the sequencing performed and data generated. Following the alignment, variants were called, functionally annotated and finally prioritized considering various approaches. Total variants identified in TU01 and TU18 were 70,196 and 81,865 respectively. We developed a strategy for variant prioritization as shown in Fig. 4.7. First, out of the total variants, under the consideration that the putative mutation in our research are exonic, we only took the exonic variants excluding the intronic and non-coding region variants for further prioritization. A total of 20,540 (29.26%) and 21,445

(26.20%) variants were identified in the exonic regions of TU01 and TU18 respectively. The expected number of exonic variants produced of WES ranges from 20,000 to 24,000 (Bamshad *et al.*, 2011). Out of these total exonic variants, the variants which possessed deleterious effect on the protein (frameshifts, nonsynonymous, stopgain and stoploss) were approximately 44.5% and 44.25% in TU01 and TU18 respectively. Correlating the clinical phenotype, then we looked for the prevalence of deleterious mutations in the 19 genes associated with rare bleeding disorders as depicted in Table 3.2. 15 and 18 nonsynonymous SNVs were identified as the deleterious variants in TU01 and TU18 respectively. Nonsynonymous SNVs alter the amino acid sequence and hence generally have a deleterious effect on the function of the protein (Majewski *et al*., 2011). Thus, we suspected the causative mutation in our case to be among those mentioned nonsynonymous SNVs tabulated in Table 4.9. Since this was a familial case, so we checked for the shared nonsynonymous variants. 13 variants were found to be shared by these siblings. Next, we looked for the homozygous and compound heterozygous variants among these 13 ones because 2N VWD is an autosomal recessive disorder. 7 homozygous shared nonsynonymous variants were identified. Finally we were able to deduce the putative variant to 1 among those 7 based on the rarity (ExAC allele frequency <0.01) and deleterious effect of the variant (SIFT & PolyPhen2 scores) as depicted in Table 4.9 and 4.10. The putative variant identified was a homozygous C to T transition in the *VWF* (c.C2446T:p.R816W) in exon 19 of both the affected siblings. This mutation was also validated by capillary sequencing as shown in Fig 4.17. Both the parents were found to carry the variant in heterozygous form. More importantly, the capillary sequencing data revealed that both the daughters of the affected male patients possess the heterozygous form of the variant p.R816W.

### 5.3.2 Prevalence of R816W

The missense mutation R816W is reported to be one of the most common mutation in 2N VWD cases (Shiltagh *et al.*, 2014). The prevalence of R816W mutation is well reported in France (Bowen *et al*., 1998) (Gaucher *et al*., 1991) (Veyradier *et al.*, 2016), Spain (Corrales *et al.* 2009), Netherland (Michiels *et al.,* 2009), China ((Qin *et al*., 2014)), Germany and Brazil (Simon and Roisenberg, 2004). The identified mutation R816W is also reported to be one of the most accurate mutation of the FVIII binding site, other being E787K and T791M, since these mutations are associated with a more severe type 2N phenotype (Shiltagh *et al.*, 2014).

### 5.3.3 Biology of the Mutation (R816W)

The residue R816 is located in the region of positive charge density on TIL' subdomain of D' domain of VWF (as shown in Fig 4.8) into which the negatively charged domain a3 of FVIII binds (Shiltagh *et al*., 2014). The TIL' subdomain harbors a cluster of putative mutations which lead to impaired FVIII binding affinity causing 2N VWD. The principle FVIII binding site is nested within a flexible, positively charged region of TIL', which is assisted

by a rigid scaffold of the TIL' and E' subdomain β sheets (Shiltagh *et al.*, 2014). Almost 85% of the total type 2N VWD reported mutations are found to be present in the D' domain of VWF protein (Michiels *et al.*, 2009). Remaining causative mutations of 2N VWD are reported to lie in exons 4,9 (Mazurier *et al.*, 2002), 17 (Casonato *et al.*, 2003), 24 (Hilbert *et al*., 2004), 25 and 27 (Goodeve, 2010). The deleterious effect of the R816W mutation is likely due to the introduction of a large and hydrophobic side chain and/or loss of positive charge (Shiltagh *et al*., 2014). The substitution of positively charged Arg to Trp results in the loss of conserved positive charge in putative FVIII a3 binding region, and also introduction of a large and hydrophobic side chain of the Trp may perturb the original conformation of the binding pocket within in TIL' thus making the domain incompatible for binding the FVIII.

Additionally the conservation of the residue R816 in VWF protein is supported by the results of two widely used in silico prediction tools viz. SIFT and PolyPhen2. SIFT (Sorting Intolerant from Tolerant) uses evolutionary conservation of the amino acid at the specific position in the protein to predict whether the variation is deleterious (D: sift<0.05) or tolerated (T: sift>0.05). The SIFT prediction is under the basic assumption that if an evolutionarily conserved amino acid at a particular position in the protein is altered, the alteration could be functionally deleterious. On the other hand, PolyPhen-2 (Polymorphism Phenotyping v2) predicts possible impact of an amino acid substitution on the structure and function of the protein and based on the score the variant can be predicted either as Probably damaging (D: >=0.909), possibly damaging (P: 0.447<=pp2_hdiv<=0.909); benign (B: pp2_hdiv<=0.446) (Scaria and Sivasubbu, 2015). The SIFT score of 0.02 and PolyPhen2_HVAR score of 1 predicts the deleterious and probably damaging effect of the mutation in the VWF respectively. The prediction scores suggests the vital role of R816 in maintaining the protein structure and function. The allele frequency of the variant R816W is found to be 0.000002489 in general population in ExAC (Exome Aggregation Consortium) browser. On an average more than 100X coverage was observed for both the siblings supporting the mutation is inherited in homozygous form.

The homozygous R816W mutation is well reported to be associated with severe recessive VWD type Normandy due to a severe FVIII binding defect of VWF (Michiels *et al.*, 2009). R816W can also be associated with another substitution mutation R924Q as a compound heterozygous for a type 2N VWD allele (R816W) with unexpectedly low FVIII level (Berber *et al.*, 2009). Also this mutation is found in compound heterozygous state with c.1911delC in a Chinese patient (Qin *et al.,* 2014).

## 5.4 Mutation in Family 2: Factor X deficiency (FXD)

The second family consists of two siblings affected from factor X deficiency (FXD), one of the rarest bleeding disorder.

### 5.4.1 WES and Bioinformatics Analysis of FXD

For the identification of causative mutation in this case, we followed the same strategy followed for 2N VWD case earlier because both of these disorder are rare familial cases that are inherited in autosomal recessive fashion. High mapping percentage of 99.99 & 99.98 implied the reliability of the reads. Around 21,000 of exonic variants were identified in both the cases. In this case also 7 homozygous shared nonsynonymous variants were identified out of which only one variant passed the criteria we had set for the variant to be a causative mutation. We pinpointed a putative homozygous T to C transition (c.T212C:p.F71S) in exon 2 of *F10* of both the affected siblings. This mutation was also validated by capillary sequencing as presented in Fig 4.18. Both the parents of the affected siblings were found to carry the heterozygous form of the variant.

### 5.4.2 Prevalence of F71S

The F71S missense mutation has been reported as a founder effect in Algerian population of Kabyle origin because of its redundant occurrence among the patients of five unrelated families sharing from same geographical origin (Akhavan *et al.*, 2007). From other parts of the world, this mutation has not been reported yet.

### 5.4.3 Biology of the Mutation F71S

Mutations present in *F10* are thought to be rare because of the central role of FX in blood coagulation mechanism. Report of FX knockout mice showing embryonic or perinatal lethality strongly suggests that a complete absence of FX is incompatible with life (Brown and Kouides, 2008; Tai *et al.*, 2008). FXD along with other recessively inherited bleeding disorders are reported to be 3 to 7 times more frequent in communities where consanguineous marriages are common such as in Iran (Peyvandi *et al*., 2002). We also have done the curation of FXD mutations reported till date and developed an in-house built mutation database of FXD. Total 130 FXD mutations were found to be reported in different papers/literatures. These mutations were located along all 8 exons except exon 3. In short 11, 18, 6, 11, 13, 10 and 59 mutations were found nesting in exon 1, 2, 4, 5, 6, 7 and 8 respectively. Remaining 2 mutations were deletions of exon 7-8. This shows numbers of mutations located in each exon are proportional to the length of the exon suggesting the fact that there are no mutation hotspots in human *F10*. However, mutations affecting exon 8 –the largest exon which codes for catalytic domain of FX –are particularly common (Lee *et al.*, 2014). A proportion of FXD mutations are reported as founder effect (identity by descent). In particular; Arg40Thr which is associated with a severe phenotype is reported in the homozygous state in four unrelated patients from Iran, Pro343Ser is reported in more than 10 patients from northern Italy (Hoffbrand, 2011), Gly380Arg which is associated with intracranial hemorrhages (Menegatti and Peyvandi, 2009) was identified in six homozygous patients from Costa Rica (Herrmann *et al.*, 2005). Similarly, as mentioned above already F71S has been reported as a founder

effect in Algerian Kabyle population (Akhavan *et al*., 2007), Gly420Arg in Costa Rica, Gly21Arg in Venezuela and in Italy with differing haplotypes (Herrmann *et al*., 2006). Jayandharan *et al*. has reported a FXD patient of Nepali origin with triple compound heterozygous mutations –a heterozygous T>C transition in exon 2 resulting a Phe71Ser substitution and 2 heterozygous 514delT and 514T>G mutations in exon 6 within a single codon Cys172 (Jayandharan *et al*., 2005).

The residue F71 in which the mutation is identified in our case, is located in the region of Gla domain of FX. The Gla domain is encoded almost by exon 2 of *F10 (*as shown in Fig 4.13*).* There are 11 are glutamic acid residues within the Gla domain that are modified by vitamin K-dependent γ-carboxylation to form γ-carboxyglutamic acid i.e. Gla residues. These γ-carboxyglutamic acid residues are essential for the binding to phospholipids on platelet membranes and also in calcium binding (Hertzberg, 1994).

The severe phenotype of the patients give a hunch that the Phe71Ser substitution besides compromising the protein function could also affect its secretion or stability. A similar reduction was also described for other mutations (Glu44Lys, Glu44Gly, Glu65Lys, Glu59Ala) affecting glutamic residues in the FX the Gla domain (Pinotti *et al*., 2002; Akhavan *et al*., 2007).Similarly, other mutations flanking Phe71 namely Glu66Asp (Wallmark *et al.*, 1991) and Glu72Gln (Factor X Tokyo) (Zama *et al.*, 1999) have been shown to be pathognomic.

Interestingly, the Phe71Ser mutation is to date the only missense mutation in the GLA domain that does not involve a GLA residue. Phe71Ser substitutes the hydrophobic phenylalanine, sandwiched by two GLA residues at position 69 and 72, with the hydrophilic serine. The substitution of this aromatic side chain by the hydroxyl group of serine may open a cavity inside the hydrophobic core to change the secondary structure or result in partial misfolding of GLA domain which could probably affect conformation dependent GLA domain binding to phospholipid membrane and normal activation (Jayandharan *et al*., 2005).

FX shares strong structural homology with other vitamin K dependent serine proteases related to coagulation mechanism (Hoffbrand, 2011) hence, the disrupted function of the Phe71Ser mutant might be surmised from similar alterations in those protein. In fact this phenylalanine residue is strongly conserved and spaced between GLA residue of factors IX, FX, protein C and protein. The presence of a topologically equivalent mutation in the FIX gene in two phenylalanine residues (Phe71 and Phe78) in the *F9*, which are sandwiched by GLA residues, causing Hemophilia B shows the importance of this codon in GLA domain. Based on structural homology between FIX and FX, a similar genotype phenotype relationship might also exist for the Phe71Ser variant (present in FX at the same relative position of the GLA domain as Phe71 in FIX), leading to conformation changes with subsequent defective carboxylation and reduced protein stability (Akhavan *et al*., 2007).

### 5.4.4 Evolutionary Conservation of F71S

Furthermore, interspecies analysis of amino acid residues by using the NCBI HomoloGene tool, an automated system for constructing putative homology groups from the complete gene sets of a wide range of eukaryotic species, indicated that in FX the Phe71 residue is highly conserved among different species. This suggests the vital role of F71 in maintaining protein structure and function. This is also supported by the results of two widely used prediction tools viz. SIFT and PolyPhen2. The SIFT score of 0 and PolyPhen2_HVAR score of 0.998 respectively predicts the deleterious and probably damaging effect of the substitution of phenylalanine by serine at 71 aa position in FX. Allele frequency of the mutation Phe71Ser is found to be 0.0000008242 in general population ExAC (Exome Aggregation Consortium). On an average more than 100X coverage was observed for both the siblings supporting the mutation is inherited in homozygous form.

# Chapter 6

# Summary

Rare bleeding disorders (RBDs), though being rare, are substantially compromising the living standard of a large number of individuals all around the world. In the context of our country, there are 536 cases of RBDs reported to Nepal Hemophilia Society. This includes the patients of Hemophilia A & B, Factor II Deficiency, Factor V Deficiency, Factor VII Deficiency, Factor XI Deficiency, Factor XIII Deficiency and 2N VWD. The molecular cause for RBDs such as hemophilia, rare factor deficiencies, VWD is the genetic lesions on the genes encoding or regulating various clotting factors. Whole exome sequencing (WES) on the Illumina Platform (HiSeq 2500) was performed for the screening of mutations in 4 Nepalese patients with 2 RBDs viz. Factor X Deficiency (FXD) and Type 2 Normandy von Willebrand disease (2N VWD). Prior to that gDNA was extracted by Salting Out method and library preparation was done by following TruSeq Exome Library Prep Illumina, USA. For the bioinformatics analysis of the sequenced data, own bioinformatics analysis strategy was developed. One very rare causative mutation was found in the exon 2 of *F10* (c.T212C:p.F71S) in the patients with FXD. This mutation has been reported as a founder effect in Algerian population and has not yet been reported from the other parts of the world. In case of 2N VWD, the causative mutation identified, c.C2446T:p.R816W, although being rare (with ExAC allele frequency of 0.000002489) is one of a common variant reported from different parts of the world such as Spain, Brazil, China, Germany etc. On performing the validation of these causative mutations by capillary sequencing, the carrier status among the family members was identified. Both the parents of the patients of both the RBDs were found to be heterozygous for the respective mutations. Whereas in the case of 2N VWD, daughters of the male patient were identified as the carrier for the 2N VWD. Thus, genetic counseling should be done to the family.

# CHAPTER 7
# CONCLUSION

Inherited bleeding disorders are of severe concerns especially in those regions where consanguineous marriages are common. Recessive forms of such disorders are life threatening. Literature reflects that many rare hemorrhagic disorders have the higher chances of misdiagnosis. von Willebrand disease is often initially misdiagnosed as Hemophilia A. Similarly, rare factors deficiencies can also occur in combined form because of their similar physiological processing and structure homology. The best example of this is multiple factor deficiency (FV and FVIII) which occurs due to defect in *ERGIC-53* Endoplasmic Reticulum/Golgi Intermediate Compartment) and *MCFD2* (Multiple Coagulation Factor Deficiency 2). These genes are responsible for encoding the proteins responsible for inter-organelle transport of FV and FVIII. Also, if there is defect in vitamin k regulating genes (*GGCX* (γ- Glutamyl Carboxylase) and *VKORC1* (Vitamin K Epoxide Reductase)), factor deficiencies of those factor which are vitamin k dependent namely FII, FVII, FIX and FX may occur in combined form. As an instance, in a report by Menegatti *et al.,* the initial phenotypical analysis in two severely affected siblings led to the diagnosis of mild FXD for which the genotyping revealed a homozygous missense mutation Ser3Cys in *F10*. Further phenotypical analysis showcased the additional presence of severe FVII deficiency with homozygous Cys310Phe in *F7* (Menegatti *et al*., 2004). This demands for a comprehensive and critical analysis of clinical phenotypes, laboratory results and genotype in patients of RBDs. Genetic screening of the genes encoding and regulating various coagulation factors is imperative for the accurate diagnosis of the hemorrhagic disorders. It is very crucial to diagnose the rare bleeding disorders accurately so that patients could be benefited with correct factor replacement which will minimize the chances of suffering from extreme phenotypical complications such as damaged joints especially elbow and knee, pseudo tumor in joints etc. Also this will lead to the correct prenatal diagnosis. For this whole exome sequencing can be the most efficient technique, both on economic aspect as well as on time factor. In our current study, the molecular diagnosis of the two severe rare bleeding disorders viz. Factor X Deficiency (FXD) and type 2 Normandy von Willebrand disease (2N VWD) was done by using whole exome sequencing approach. Those results obtained from the exome sequencing were validated by targeted capillary sequencing. The accurate identification of the underlying genetic lesions causing the bleeding diseases enabled us for the carrier screening within the family members. Furthermore, this genetic study also enables for the prenatal testing in the respective families. This research opens a new avenue for the genetic research of other rare bleeding disorders such as Hemophilia A & B, rare factor deficiencies such as Factor II deficiency, Factor V deficiency, Factor VII deficiency, Factor XIII deficiency etc. prevalent in Nepal. The finding of this research is very interesting especially in FXD the mutation identified F71S was reported as a founder effect in an Algerian population. Further genetic analysis are required to confirm whether the Nepalese FXD patients and the Algerian FXD patients are of same origin or not. The mutation identified in this research, although being

extremely rare couldn't be the de novo one as we had hypothesized. One can expect the de novo mutations in the other Nepalese patients with RBDs belonging to different ethnic community.

## Appendix 1: Reagents & Composition

**1. RBC lysis buffer (10X)**

8.26 g of $NH_4Cl$

1.19 g of NaHCO3

200 µL of EDTA [0.5 M, pH8]

Adjust pH to 7.3

Add DDW until 100 mL

Filter sterilized

**2. Nuclei lysis buffer (NLB):**

50 ml of 2 M Tris-HCL, pH 8.0

200ml of 0.5M EDTA, pH8.0

2 ml of 5 M NaCl

Add DDW until 975 mL

**3. 6M saturated NaCl**

35.064 gm NaCl dissolved in DDW and final volume made upto 100mL.

**4. Prepare 20% SDS**

20 gram SDS in 100mL autoclaved distilled water

**5. Proteinase K (20 mg/ml)**

20 mg proteinase K in 1 mL of DDW.

# Consent form of the patients

## Consent Form — Code: TU03

सिक्वेन्स भेरीएसन & मिऊटेशन एनलाइसिस अफ हिमोफिलिया जिन

**(Sequence variation & Mutation Analysis of Hemophilia gene )**

सहभागीको मन्जूरीनामा फारम

**Study No :** TU03 मिति : 9th Jan 2016

तपाईलाई यस अध्ययनमा सहभागी हुनलाई आह्वान गर्नुको कारण रगतको नमुना जांचवाट Hemophilia gene को **Sequence variation** र **Mutation** पत्ता लगाउनको लागि हो । यस मन्जूरीनामा पत्रको उद्देश्य तपाईलाई यस अध्ययन सम्बन्धि पर्याप्त जानकारी दिनु हो ताकी तपाई यस अध्ययनमा सहभागी हुने नहुने महि निर्णय लिन सक्नु हुनेछ । तपाईलाई प्रष्ट नभएका कुनै पनि प्रश्नहरु सोध्न सक्नुहुने छ । तपाईवाट ५ मि.लि. रगत लिइने छ । यस प्रक्रियामा करिव १० मिनेट समय लाग्नेछ । यी संकलित नमुनाहरुवाट भारतको **CSIR-Institute of Genomics and Integrative Biology,Delhi** मा विश्लेशण गरिने छ । संकलित नमुनाहरुलाई गोप्य संख्यामा लेखिनेछ ताकी तपाईको व्यक्तिगत परिचय गोप्य रहन सकोस् । अध्ययनमा प्रयोग गरिसके पछि बाकी रहेको नमूनाहरुलाई र तपाईवाट लिईएको जानकारीलाई भविष्यमा यस रोगको उपचारमा प्रयोग गर्न सकिनेछ, तथापि तपाईको व्यक्तिगत परिचयलाई गोप्य राखिने छ । रगत निकाल्नाले तपाईलाई कुनै किसिमको हानि हुने छैन ।

**फाइदा :**

यस अनुसन्धानमा सहभागी भएर तपाईलाई प्रत्यक्ष रुपमा कुनै फाइदा छैन तापनि, यस अध्ययनवाट प्राप्त जानकारीवाट नेपालमा हिमोफिलियाको Data Base तयार गर्न सकिनेछ र साथै यस रोगको उपचारमा प्रयोग गर्न सकिनेछ ।

**गोपनीयता :**

यस अनुसन्धान-अध्ययनको नतीजा प्रकाशित गर्न सकिनेछ, तर त्यसमा तपाईको नाम तथा परिचय उल्लेख हुनेछैना

**स्वेच्छिक सहभागिताको बयान :**

यस अनुसन्धानमा मेरो सहभागिता स्वेच्छिक हो । मैले आफ्नो इच्छाले विना जरिवाना वा डर, तल उल्लेख गरीएका अनुसन्धानकर्ता समक्ष पूर्व सुचना बिनाने कुनै पनि वेला यस अनुसन्धानवाट सहभागिता परित्याग गर्न सक्नेछु । मैले माथि लेखिएका कुराहरु पढेको छु अथवा मलाई पढेर सुनाइएको छ , मेरो प्रश्नहरुको जवाफ दिइएको छ र आफ्नो इच्छाले यस फाराममा सही गरी मेरो शरीरवाट ५ मि.लि. रगत झिक्न अनुमति दिएको छु ।

सहभागीको हस्ताक्षर

Brief

समस्या अथवा प्रश्नहरु

यस अनुसन्धानसंग सम्बन्धित समस्या, प्रश्न अथवा तपाईको अधिकारको बारेमा कुनै प्रश्न भएमा निम्न लिखित व्यक्तिहरुसंग सम्पर्क गर्न सक्नुहुनेछ ।

| | |
|---|---|
| १ प्रा. डा. तिलक राम श्रेष्ठ | २. प्रा.डा. रर्जान मल्ल |
| जैविक प्रविधि केन्द्रिय विभाग, | विभाग प्रमुख, जैविक प्रविधि केंन्द्रिय विभाग |
| त्रि.वि. कीर्तिपुर | त्रि.वि. कीर्तिपुर |
| कार्यालय : ४३३६२२१ | कार्यालय : ४३३६२२१ |
| मो.न. ९८४९२८५४४ | मो.न. : ९८५१०९३०३४ |

सहभागीको नाम [redacted]

उमेर 32 yrs

ठेगाना Butwal, Rupendehi

मिति 9th Jan 2016

सम्पर्क न. [redacted]

.....

लेखपढ गर्न नजान्नेहरुको लागि तल भनुंसेस ।

म यो प्रमाणित गर्दछु कि मैले माथि उल्लेखित सूचना सहभागी श्रीमान .....लाई सवै पढेर सुनाएको छु तथा व्याख्या गरेको छु । मैले भनेको कुरा उहांले बुझेको कुरा दर्शाएको छ तथा प्रश्न सोध्नको लागी पुरा मौका दिएको छ र उहांले अनुसन्धानमा सहभागी हुन मन्जुरी जनाउनु भएको छ । यो सही/चिन्ह/ओठाको छाप उहांको हो भनि म प्रमाणित गर्दछु ।

फिल्डवर्करको हस्ताक्षर Bishnu

फिल्डवर्करको नाम Bishnu Neupane

मिति 9th Jan 2016

सहभागीको सही/चिन्ह/ओठाको छाप

Central Department of Biotech Kritipur-64

Referred to IGIB, Delhi for genetic test.

**RESEARCH CONSENT FORM**

Title of the Project: Centre for Cardiovascular and Metabolic Disease Research (BSC0122)

**Institute address:** CSIR-Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research, New Delhi, India.

Reference Code: TU03

I, [redacted]

Son/daughter/wife of [redacted] aged 32 yrs

Residing at Butwal, Nepal

.....

hereby consent to freely participate in the genetic study aimed at understanding the human gnome. I have been informed about the implications of my personal genome data being made publicly available through public databases as well as scientific communications. I have been advised to discuss my participation in this study with my family members. I have been provided written information that may be circulated to them, if necessary. I have been further informed that personal and medical data collected during this study will be associated with my publicly available genome and may be used for scientific analysis. My participation in this study is entirely voluntary and I am free to withdraw from this study as and when I feel so inclined.

1. **I choose to disclose / not to disclose my identity** (select one option).
2. **I choose to be / not to be informed of the results of the analysis that may impact my health** (Applicable only to those who have chosen to disclose their identity - select one option).
3. **I choose to exclude the information attached on the "Exclusion Form" from analysis / public disclosure** (Applicable only to those who have chosen to disclose their identity).

Tribhuvan University consent form has been signed by the patient and is attached herewith.

(Signature/ Thumb impression of volunteer) Date: 8 July 2016

Certified that the above consent has been signed in my presence. The purpose for which the sample will be used has been explained to the above volunteer. The individual is free to withdraw from the study as and when he/she feels so inclined.

(Signature of the investigator) Date: 8 July 2016

Central Department of Biotechnology Kritipur-64

## Appendix 2: Table 1: Details of variations found in genes related to RBDS in TU01

| Chr | Position | Ref | Alt | Gene.refGene | ExonicFunc.refGene | AAChange.refGene |
|---|---|---|---|---|---|---|
| **chr1** | 95001600 | A | C | *F3* | synonymous SNV | F3:NM_001178096:exon3:c.T333G:p.P111P |
| **chr1** | 169498975 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon16:c.A5290G:p.M1764V |
| **chr1** | 169510380 | G | A | *F5* | synonymous SNV | F5:NM_000130:exon13:c.C3948T:p.L1316L |
| **chr1** | 169510475 | G | T | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.C3853A:p.L1285I |
| **chr1** | 169510524 | A | G | *F5* | synonymous SNV | F5:NM_000130:exon13:c.T3804C:p.S1268S |
| **chr1** | 169511555 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2773G:p.K925E |
| **chr1** | 169511734 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2594G:p.H865R |
| **chr1** | 169511755 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2573G:p.K858R |
| **chr1** | 169512027 | T | C | *F5* | synonymous SNV | F5:NM_000130:exon13:c.A2301G:p.S767S |
| **chr1** | 169512093 | A | G | *F5* | synonymous SNV | F5:NM_000130:exon13:c.T2235C:p.N745N |
| **chr1** | 169512120 | G | A | *F5* | synonymous SNV | F5:NM_000130:exon13:c.C2208T:p.I736I |
| **chr1** | 169519049 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon10:c.A1601G:p.Q534R |
| **chr1** | 197009798 | A | G | *F13B* | synonymous SNV | F13B:NM_001994:exon11:c.T1806C:p.N602N |
| **chr1** | 197030201 | T | C | *F13B* | synonymous SNV | F13B:NM_001994:exon4:c.A456G:p.T152T |
| **chr1** | 197031021 | C | T | *F13B* | nonsynonymous SNV | F13B:NM_001994:exon3:c.G344A:p.R115H |
| **chr2** | 85780131 | G | A | *GGCX* | synonymous SNV | GGCX:NM_001142269:exon8:c.C1047T:p.R349R |
| **chr2** | 85780536 | C | T | *GGCX* | nonsynonymous SNV | GGCX:NM_001142269:exon7:c.G803A:p.R268Q |
| **chr2** | 85781318 | T | G | *GGCX* | synonymous SNV | GGCX:NM_001142269:exon6:c.A666C:p.G222G |
| **chr4** | 187201211 | A | G | *F11* | synonymous SNV | F11:NM_000128:exon8:c.A801G:p.T267T |
| **chr4** | 187205301 | T | C | *F11* | synonymous SNV | F11:NM_000128:exon11:c.T1191C:p.G397G |
| **chr5** | 176831826 | C | G | *F12* | nonsynonymous SNV | F12:NM_000505:exon7:c.G619C:p.A207P |
| **chr6** | 6174866 | G | A | *F13A1* | nonsynonymous SNV | F13A1:NM_000129:exon12:c.C1694T:p.P565L |
| **chr12** | 6128443 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon28:c.A4141G:p.T1381A |
| **chr12** | 6143984 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon20:c.A2555G:p.Q852R |
| **chr12** | 6145654 | G | A | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon19:c.C2446T:p.R816W |
| **chr12** | 6172202 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon13:c.A1451G:p.H484R |
| **chr13** | 113801737 | C | T | *F10* | synonymous SNV | *F10*:NM_000504:exon7:c.C792T:p.T264T |
| **chr18** | 57000469 | T | A | *LMAN1* | nonsynonymous SNV | LMAN1:NM_005570:exon11:c.A1228T:p.M410L |

## Appendix 3: Table 2: Details of variations found in genes related to RBDs in TU18

| Chr | Position | Ref | Alt | Gene.refGene | ExonicFunc.refGene | AAChange.refGene |
|---|---|---|---|---|---|---|
| **chr1** | 95001600 | A | C | *F3* | synonymous SNV | F3:NM_001178096:exon3:c.T333G:p.P111P |
| **chr1** | 169498975 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon16:c.A5290G:p.M1764V |
| **chr1** | 169510139 | G | A | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.C4189T:p.L1397F |
| **chr1** | 169510380 | G | A | *F5* | synonymous SNV | F5:NM_000130:exon13:c.C3948T:p.L1316L |
| **chr1** | 169510524 | A | G | *F5* | synonymous SNV | F5:NM_000130:exon13:c.T3804C:p.S1268S |
| **chr1** | 169511555 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2773G:p.K925E |
| **chr1** | 169511734 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2594G:p.H865R |
| **chr1** | 169511755 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2573G:p.K858R |
| **chr1** | 169512027 | T | C | *F5* | synonymous SNV | F5:NM_000130:exon13:c.A2301G:p.S767S |
| **chr1** | 169512093 | A | G | *F5* | synonymous SNV | F5:NM_000130:exon13:c.T2235C:p.N745N |
| **chr1** | 169512120 | G | A | *F5* | synonymous SNV | F5:NM_000130:exon13:c.C2208T:p.I736I |
| **chr1** | 169519049 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon10:c.A1601G:p.Q534R |
| **chr1** | 169519112 | C | T | *F5* | nonsynonymous SNV | F5:NM_000130:exon10:c.G1538A:p.R513K |
| **chr1** | 169551682 | T | C | *F5* | synonymous SNV | F5:NM_000130:exon2:c.A237G:p.Q79Q |
| **chr1** | 197009798 | A | G | *F13B* | synonymous SNV | F13B:NM_001994:exon11:c.T1806C:p.N602N |
| **chr1** | 197030201 | T | C | *F13B* | synonymous SNV | F13B:NM_001994:exon4:c.A456G:p.T152T |
| **chr1** | 197031021 | C | T | *F13B* | nonsynonymous SNV | F13B:NM_001994:exon3:c.G344A:p.R115H |
| **chr2** | 85780131 | G | A | *GGCX* | synonymous SNV | GGCX:NM_001142269:exon8:c.C1047T:p.R349R |
| **chr2** | 85780536 | C | T | *GGCX* | nonsynonymous SNV | GGCX:NM_001142269:exon7:c.G803A:p.R268Q |
| **chr2** | 85781318 | T | G | *GGCX* | synonymous SNV | GGCX:NM_001142269:exon6:c.A666C:p.G222G |
| **chr4** | 155488821 | C | T | *FGB* | synonymous SNV | FGB:NM_001184741:exon4:c.C390T:p.S130S |
| **chr4** | 155490832 | C | T | *FGB* | synonymous SNV | FGB:NM_001184741:exon7:c.C948T:p.Y316Y |
| **chr4** | 155491759 | G | A | *FGB* | nonsynonymous SNV | FGB:NM_001184741:exon8:c.G1256A:p.R419K |
| **chr4** | 187209702 | G | T | *F11* | synonymous SNV | F11:NM_000128:exon15:c.G1812T:p.R604R |
| **chr4** | 187209729 | G | A | *F11* | synonymous SNV | F11:NM_000128:exon15:c.G1839A:p.E613E |
| **chr5** | 176831826 | C | G | *F12* | nonsynonymous SNV | F12:NM_000505:exon7:c.G619C:p.A207P |
| **chr6** | 6174842 | G | A | *F13A1* | nonsynonymous SNV | F13A1:NM_000129:exon12:c.C1718T:p.T573M |
| **chr6** | 6174866 | G | A | *F13A1* | nonsynonymous SNV | F13A1:NM_000129:exon12:c.C1694T:p.P565L |
| **chr12** | 6127943 | A | G | *VWF* | synonymous SNV | VWF:NM_000552:exon28:c.T4641C:p.T1547T |
| **chr12** | 6128443 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon28:c.A4141G:p.T1381A |
| **chr12** | 6143984 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon20:c.A2555G:p.Q852R |
| **chr12** | 6145654 | G | A | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon19:c.C2446T:p.R816W |
| **chr12** | 6172202 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon13:c.A1451G:p.H484R |
| **chr13** | 113801737 | C | T | *F10* | synonymous SNV | *F10*:NM_000504:exon7:c.C792T:p.T264T |
| **chr18** | 57000469 | T | A | *LMAN1* | nonsynonymous SNV | LMAN1:NM_005570:exon11:c.A1228T:p.M410L |

## Appendix 4: Table 3: Details of variations found in genes related to RBDs in TU03

| Chr | Start | Ref | Alt | Gene. | ExonicFunc.refGene | AAChange.refGene |
|---|---|---|---|---|---|---|
| **chr1** | 95001600 | A | C | *F3* | synonymous SNV | F3:NM_001178096:exon3:c.T333G:p.P111P |
| **chr1** | 169498975 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon16:c.A5290G:p.M1764V |
| **chr1** | 169500210 | T | C | *F5* | synonymous SNV | F5:NM_000130:exon15:c.A5022G:p.G1674G |
| **chr1** | 169510118 | G | A | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.C4210T:p.P1404S |
| **chr1** | 169510233 | G | A | *F5* | synonymous SNV | F5:NM_000130:exon13:c.C4095T:p.T1365T |
| **chr1** | 169510380 | G | A | *F5* | synonymous SNV | F5:NM_000130:exon13:c.C3948T:p.L1316L |
| **chr1** | 169510524 | A | G | *F5* | synonymous SNV | F5:NM_000130:exon13:c.T3804C:p.S1268S |
| **chr1** | 169511555 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2773G:p.K925E |
| **chr1** | 169511734 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2594G:p.H865R |
| **chr1** | 169511755 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon13:c.A2573G:p.K858R |
| **chr1** | 169512027 | T | C | *F5* | synonymous SNV | F5:NM_000130:exon13:c.A2301G:p.S767S |
| **chr1** | 169512093 | A | G | *F5* | synonymous SNV | F5:NM_000130:exon13:c.T2235C:p.N745N |
| **chr1** | 169512120 | G | A | *F5* | synonymous SNV | F5:NM_000130:exon13:c.C2208T:p.I736I |
| **chr1** | 169519049 | T | C | *F5* | nonsynonymous SNV | F5:NM_000130:exon10:c.A1601G:p.Q534R |
| **chr1** | 169551682 | T | C | *F5* | synonymous SNV | F5:NM_000130:exon2:c.A237G:p.Q79Q |
| **chr1** | 197009798 | A | G | *F13B* | synonymous SNV | F13B:NM_001994:exon11:c.T1806C:p.N602N |
| **chr1** | 197030201 | T | C | *F13B* | synonymous SNV | F13B:NM_001994:exon4:c.A456G:p.T152T |
| **chr1** | 197031021 | C | T | *F13B* | nonsynonymous SNV | F13B:NM_001994:exon3:c.G344A:p.R115H |
| **chr2** | 85781318 | T | G | *GGCX* | synonymous SNV | GGCX:NM_001142269:exon6:c.A666C:p.G222G |
| **chr4** | 187201211 | A | G | *F11* | synonymous SNV | F11:NM_000128:exon8:c.A801G:p.T267T |
| **chr4** | 187205301 | T | C | *F11* | synonymous SNV | F11:NM_000128:exon11:c.T1191C:p.G397G |
| **chr4** | 187209702 | G | T | *F11* | synonymous SNV | F11:NM_000128:exon15:c.G1812T:p.R604R |
| **chr4** | 187209729 | G | A | *F11* | synonymous SNV | F11:NM_000128:exon15:c.G1839A:p.E613E |
| **chr5** | 176831826 | C | G | *F12* | nonsynonymous SNV | F12:NM_000505:exon7:c.G619C:p.A207P |
| **chr6** | 6174866 | G | A | *F13A1* | nonsynonymous SNV | F13A1:NM_000129:exon12:c.C1694T:p.P565L |
| **chr11** | 46745003 | C | T | *F2* | nonsynonymous SNV | F2:NM_000506:exon6:c.C494T:p.T165M, |
| **chr12** | 6091000 | A | G | *VWF* | synonymous SNV | VWF:NM_000552:exon42:c.T7239C:p.T2413T |
| **chr12** | 6105387 | G | A | *VWF* | synonymous SNV | VWF:NM_000552:exon35:c.C5844T:p.C1948C |
| **chr12** | 6127891 | C | A | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon28:c.G4693T:p.V1565L |
| **chr12** | 6127943 | A | G | *VWF* | synonymous SNV | VWF:NM_000552:exon28:c.T4641C:p.T1547T |
| **chr12** | 6128443 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon28:c.A4141G:p.T1381A |
| **chr12** | 6143984 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon20:c.A2555G:p.Q852R |
| **chr12** | 6167196 | A | G | *VWF* | synonymous SNV | VWF:NM_000552:exon14:c.T1548C:p.Y516Y |
| **chr12** | 6172202 | T | C | *VWF* | nonsynonymous SNV | VWF:NM_000552:exon13:c.A1451G:p.H484R |
| **chr13** | 113783907 | T | C | *F10* | nonsynonymous SNV | *F10*:NM_000504:exon2:c.T212C:p.F71S |
| **chr13** | 113801737 | C | T | *F10* | synonymous SNV | *F10*:NM_000504:exon7:c.C792T:p.T264T |
| **chr18** | 57000465 | C | T | *LMAN1* | nonsynonymous SNV | LMAN1:NM_005570:exon11:c.G1232A:p.S411N |
| **chrX** | 154158201 | T | G | F8 | synonymous SNV | F8:NM_000132:exon14:c.A3864C:p.S1288S |

## Appendix 5: Table 4: Details of variations found in genes related to RBDs in TU25

| Chr | Start | Ref | Alt | Gene. | ExonicFunc.refGene | AAChange.refGene |
|---|---|---|---|---|---|---|
| **chr1** | 95001600 | A | C | F3 | synonymous SNV | F3:NM_001178096:exon3:c.T333G:p.P111P |
| **chr1** | 169498975 | T | C | F5 | nonsynonymous SNV | F5:NM_000130:exon16:c.A5290G:p.M1764V |
| **chr1** | 169510233 | G | A | F5 | synonymous SNV | F5:NM_000130:exon13:c.C4095T:p.T1365T |
| **chr1** | 169510380 | G | A | F5 | synonymous SNV | F5:NM_000130:exon13:c.C3948T:p.L1316L |
| **chr1** | 169510524 | A | G | F5 | synonymous SNV | F5:NM_000130:exon13:c.T3804C:p.S1268S |
| **chr1** | 169511555 | T | C | F5 | nonsynonymous SNV | F5:NM_000130:exon13:c.A2773G:p.K925E |
| **chr1** | 169511734 | T | C | F5 | nonsynonymous SNV | F5:NM_000130:exon13:c.A2594G:p.H865R |
| **chr1** | 169511755 | T | C | F5 | nonsynonymous SNV | F5:NM_000130:exon13:c.A2573G:p.K858R |
| **chr1** | 169512027 | T | C | F5 | synonymous SNV | F5:NM_000130:exon13:c.A2301G:p.S767S |
| **chr1** | 169512093 | A | G | F5 | synonymous SNV | F5:NM_000130:exon13:c.T2235C:p.N745N |
| **chr1** | 169512120 | G | A | F5 | synonymous SNV | F5:NM_000130:exon13:c.C2208T:p.I736I |
| **chr1** | 169519049 | T | C | F5 | nonsynonymous SNV | F5:NM_000130:exon10:c.A1601G:p.Q534R |
| **chr1** | 197009798 | A | G | F13B | synonymous SNV | F13B:NM_001994:exon11:c.T1806C:p.N602N |
| **chr1** | 197030201 | T | C | F13B | synonymous SNV | F13B:NM_001994:exon4:c.A456G:p.T152T |
| **chr1** | 197031021 | C | T | F13B | nonsynonymous SNV | F13B:NM_001994:exon3:c.G344A:p.R115H |
| **chr2** | 85781318 | T | G | GGCX | synonymous SNV | GGCX:NM_001142269:exon6:c.A666C:p.G222G |
| **chr4** | 187201211 | A | G | F11 | synonymous SNV | F11:NM_000128:exon8:c.A801G:p.T267T |
| **chr4** | 187205301 | T | C | F11 | synonymous SNV | F11:NM_000128:exon11:c.T1191C:p.G397G |
| **chr5** | 176831826 | C | G | F12 | nonsynonymous SNV | F12:NM_000505:exon7:c.G619C:p.A207P |
| **chr6** | 6174866 | G | A | F13A1 | nonsynonymous SNV | F13A1:NM_000129:exon12:c.C1694T:p.P565L |
| **chr11** | 46745003 | C | T | F2 | nonsynonymous SNV | F2:NM_000506:exon6:c.C494T:p.T165M |
| **chr12** | 6091000 | A | G | VWF | synonymous SNV | VWF:NM_000552:exon42:c.T7239C:p.T2413T |
| **chr12** | 6105387 | G | A | VWF | synonymous SNV | VWF:NM_000552:exon35:c.C5844T:p.C1948C |
| **chr12** | 6127891 | C | A | VWF | nonsynonymous SNV | VWF:NM_000552:exon28:c.G4693T:p.V1565L |
| **chr12** | 6127919 | T | G | VWF | synonymous SNV | VWF:NM_000552:exon28:c.A4665C:p.A1555A |
| **chr12** | 6127943 | A | G | VWF | synonymous SNV | VWF:NM_000552:exon28:c.T4641C:p.T1547T |
| **chr12** | 6128443 | T | C | VWF | nonsynonymous SNV | VWF:NM_000552:exon28:c.A4141G:p.T1381A |
| **chr12** | 6143984 | T | C | VWF | nonsynonymous SNV | VWF:NM_000552:exon20:c.A2555G:p.Q852R |
| **chr12** | 6167196 | A | G | VWF | synonymous SNV | VWF:NM_000552:exon14:c.T1548C:p.Y516Y |
| **chr12** | 6172202 | T | C | VWF | nonsynonymous SNV | VWF:NM_000552:exon13:c.A1451G:p.H484R |
| **chr13** | 113783907 | T | C | *F10* | nonsynonymous SNV | *F10*:NM_000504:exon2:c.T212C:p.F71S |
| **chr13** | 113801737 | C | T | *F10* | synonymous SNV | *F10*:NM_000504:exon7:c.C792T:p.T264T |
| **chr18** | 57000465 | C | T | LMAN1 | nonsynonymous SNV | LMAN1:NM_005570:exon11:c.G1232A:p.S411N |
| **chrX** | 154158201 | T | G | F8 | synonymous SNV | F8:NM_000132:exon14:c.A3864C:p.S1288S |

## Appendix 6: Validation of Putative Variants by Capillary Sequencing

The putative mutations associated with both bleeding disorders (FXD and 2N VWD) were technically validated by capillary sequencing. The segregation of the variant was checked among the family members of the patients. Primers were designed to amplify the genetic loci containing the putative variant by Primer Blast (NCBI).

**Primer design for FXD case**

AGAGCTTTTAACCCTGTCCTCCCTGCCTTCCAGTGTTCATCCGCAGGGAGCAGGCCAACAACATCCTGGCGAGGGTCAC
GAGGGCCAATTCCTTTCTTGAAGAGATGAAGAAAGGACACCTCGAAAGAGAGTGCATGGAAGAGACCTGCTCATACG
AAGAGGCCCGCGAGGTCTTTGAGGACAGCGACAAGACGGTAAGGGCTGGGGATAGCCTGGCTGTTGGTAAGGAGCTC
AGGCCACAGCGCCCTCGCTGGCCCCGCTGCTCCGTCCATCCAGGGGGGCGGCCTGGAGGAAGGGGCAGCGTGCGCGA
AGGCTTTCAGGGGCGGGGCCCAGCAAATCGAGGCCTCGGCGGAGTCCTGCCCACAGGGACATCAGTGCCGCCCCCGC
GCTGACTCCTTCCCGGCGAGGACTCAGCGGGGAGGGATGCGCCCAAGTCCCTTGAGGGTCACAGGGCTTCTGCCAGA

| Primer | Sequence 5'-3' | Length bp | Tm °c | GC% |
|---|---|---|---|---|
| Forward | **CCTTCCAGTGTTCATCCGCA** | 20 | 62 | 55 |
| Reverse | **CTGTGACCCTCAAGGGACTT** | 20 | 62 | 55 |

**Product size= 426 bp**

**Fig 1: Primer designed for FXD case.** Primer was designed against c.T212C:p.F71S present in exon 2 of *F10*. The variant is marked in red. Primer sequences are highlighted in yellow and shown by arrow and. The exon 2 of the *F10* is shown in blue color.

**Primer design for 2N VWD**

GTCTATGCAGGATGGACACAGGTGATGAAAAAGATTCTCCCCATTTTACAGATGGAGAAACACAGGCACGGAGGAAAGGCAGAGAGTA
ACCAGGTTCCCACAGGGGGCTGGAGGCAAGTGCGGAAGGTCCTGTGGCCGCGTGCACCCTCACTCCACCCGCAGGGCCTGGGTCCCC
GGCGCAGCCCCTCACTGAGCCTCACCAAGTGTTGCAGCCAATCTTCACTGTTTCTCCAGGGGCATACTCCTTGCCCTGATGGAAGCAGG
GACACCTTTCCAGGGCCACACATCTGTTCTCATGCCGGACCTAAGAGAAAAGAATCCAAAAGTCCTCAGGGCCACAGTGACTGATCTAA
AGCCCTCCTCCAGCCCGCTCCAGGAAGTGTTGTCTGAGACTTGAGCCTACGGGTAAGCCCAGCACTGGGACTGGCACCAGGACCAGGG
CTGAGCCAGGGGGACAGCTTCCACTGCAGGCCAAGCCAGGACAGGGATCCTATGCCCAGGCCAGGGCTTGGCAAACATGGGTTCCCA

| Primer | Sequence 5'-3' | Length bp | Tm °c | GC% |
|---|---|---|---|---|
| Forward | **CTATGCAGGATGGACACAGGT** | 21 | 64 | 52 |
| Reverse | **GAACCCATGTTTGCCAAGCC** | 20 | 62 | 55 |

**Product size= 523 bp**

**Fig 2: Primer designed for 2N VWD case.** Primer was designed against c.C2446T:p.R816W present on exon 19 of *VWF*. The variant is marked in red. Primer sequences are highlighted in yellow and shown by arrow and. The exon 19 of the *VWF* is shown in blue color.

## PCR MasterMix

All the PCR components (Invitrogen) were mixed in appropriate volume according to the manufacturer provided protocol. PCR master mix was prepared by adding the various PCR components for total number of samples. For factor X deficiency we had 6 samples hence master mix was prepared for 7 samples (one negative control) s shown in Table 5. Similarly, for 2N VWD we had 7 samples hence PCR master mix was prepared for 8 samples (one negative control) as shown in Table 6.

Table 5: PCR master mix for FXD case

| **Components** | **Volume for 1 sample** | **Volume for 7 samples** |
|---|---|---|
| 10X PCR buffer minus Mg | 5µl | 35µl |
| 50 mM MgCl2 | 2µl | 14µl |
| 10 mM dNTP mixture | 2µl | 14µl |
| 0.1 µM Forward primer | 1µl | 7µl |
| 0.1 µM Reverse primer | 1µl | 7µl |
| Taq DNA Polymerase (U/µl) | 0.5µl | 3.5µl |
| 10 ng DNA template | 3µl | - |
| Nuclease free water | 35.5µl | 248.5µl |
| Total | 50µl | 329µl |

Table 6: PCR mastermix for 2N VWD case

| **Components** | **Volume for 1 sample** | **Volume for 8 samples** |
|---|---|---|
| 10X PCR buffer minus Mg | 5µl | 40µl |
| 50 mM MgCl2 | 2µl | 16µl |
| 10 mM dNTP mixture | 2µl | 16µl |
| 0.1 µM Forward primer | 1µl | 8µl |
| 0.1 µM Reverse primer | 1µl | 8µl |
| Taq DNA Polymerase (U/µl) | 0.5µl | 4µl |
| 10 ng DNA template | 3µl | - |
| Nuclease free water | 35.5µl | 284µl |
| Total | 50µl | 376µl |

## PCR Program

PCR conditions were set according to the PCR components' manufacturer provided protocol. The melting temperature of both sets of primer was calculated using following formula:

$$T_m = 2\,(A+T) + 4\,(G+C)$$

The accordingly calculated melting temperature of the primers is shown in Table 7.

Table 7: Primers and their melting temperature

| Primer | | Sequence | $T_m$= 2 (A+T) + 4 (G+C) |
|---|---|---|---|
| Primer for FXD | $F_p$ | CCTTCCAGTGTTCATCCGCA | 62°C |
| | $R_p$ | CTGTGACCCTCAAGGGACTT | 62°C |
| Primer for 2N VWD | $F_p$ | CTATGCAGGATGGACACAGGT | 64°C |
| | $R_p$ | GAACCCATGTTTGCCAAGCC | 62°C |

Generally the annealing temperature of the primer is 2-5°C less than melting temperature ($T_m$) of the primer. Thus, the annealing temperature is was set as 60°C for both sets of the primers and PCR conditions were set as depicted in Table 8.

Table 8: PCR Programme to amplify the targeted genetic region.

| Temperature | Duration | No. of cycles | Conditions |
|---|---|---|---|
| 94°C | 3 min | 1 | Initial denaturation |
| 94°C | 45 sec | 35 | Denaturation |
| 60°C | 30 sec | | Annealing |
| 72°C | 45 sec | | Extension |
| 72°C | 10 min | 1 | Extension |
| 4°C | ∞ | 1 | Hold |

## PCR Products Purification

After the amplification of genetic region of interest, the PCR amplified products were purified by column purification method using QIAquick PCR Purification Kit (Qiagen, USA). This is based on silica-membrane-based purification of PCR products of size >100 bp. The PCR product is purified using a simple and fast binding-washing-elution procedure using the corresponding reagents provided. The purification protocol is as follows:

1. 300µl of Buffer PB (binding buffer) was taken in 1.5ml micro centrifuge tube (MCT).
2. The PCR product (50µl) was mixed with the binding buffer by pipetting and the mixture was transferred into a column.
3. The column was spun at 13000 rpm for 90 seconds at RT.
4. Flow through was discarded and the column was washed with 500µl of Buffer PW (wash buffer).
5. It was again spun at 13000 rpm for 90 seconds at RT.
6. The flow through was discarded and a dry spin was given at 13000 rpm for 3 minutes at RT
7. The column was transferred to 1.5ml MCT and the DNA bound in the column was eluted by 12.5µl of NFW.
8. It was incubated at RT for 2 minutes and was spun at 13000 rpm for 90 seconds at RT.

9. The remnant DNA was re-eluted with same volume of NFW.

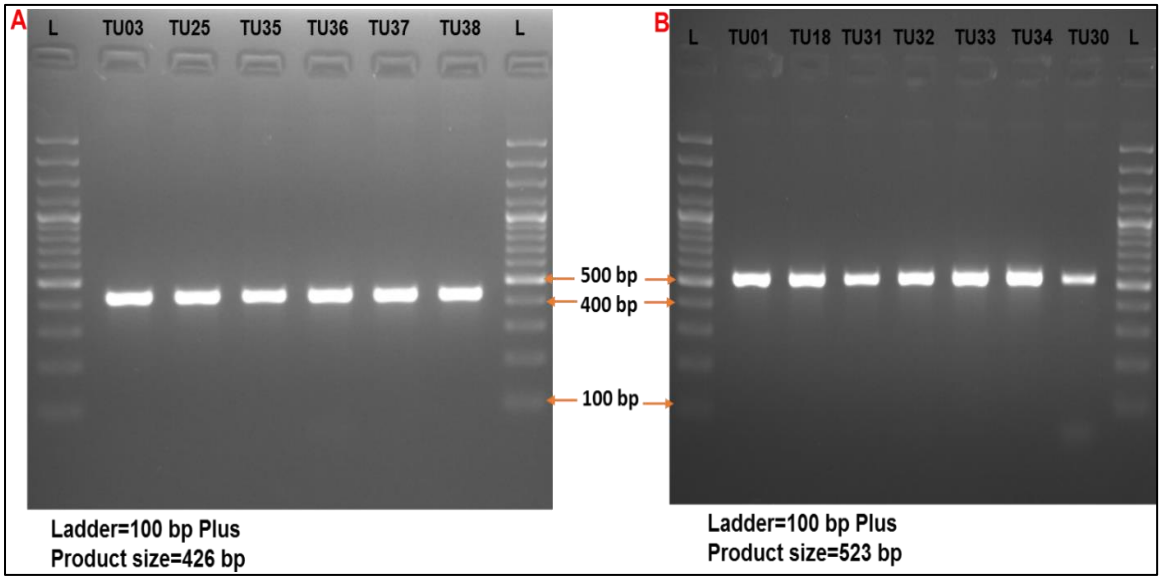


**Fig 3: Gel image of purified PCR products.** (A) Purified PCR product with putative variant c.T212C:p.F71S present in exon 2 of *F10*. Product size of the amplified region was 426 bp. (B) Purified PCR product with putative variant c.C2446T:p.R816W present on exon 19 of *VWF.* Product size of the amplified region was 523 bp. L- 100 bp Plus ladder (GeneRuler).

The purified PCR products were then subjected to sequencing PCR prior to the capillary sequencing. The components of the PCR are tabulated in Table 9 and PCR conditions are tabulated in Table 10.

Table 9: Components for Sequencing PCR

| Components | Volume (µl) |
|---|---|
| 2µM primer (Forward or Reverse) | 1 |
| 100ng purified PCR product | 5 |
| RM | 3 |
| NFW | 2 |

Table 10: Sequencing PCR conditions

| Temperature | Duration | No. of cycles | Conditions |
|---|---|---|---|
| 96°C | 3 min | 1 | Initial denaturation |
| 96°C | 10 sec | 35 | Denaturation |
| 60°C | 05 sec | | Annealing |
| 60°C | 5 min | | Extension |
| 60°C | 7 min | 1 | Extension |
| 4°C | ∞ | 1 | Hold |

## Sequencing PCR products purification:

The PCR products after sequencing reaction may have impurities such as unused primers, dNTPs, ddNTPs, salts etc. In order to remove these unused products, purification is necessary prior to keep the samples for the sequencing. The purification is done as follows:

1. 45µl of AMPure XP beads (Beckman Coulter, USA) was added to the sequencing PCR product in a 96 well plate and was mixed well by pipetting.
2. Incubated at RT for 10 minutes.
3. Then the plate was kept on magnetic plate for 5 minutes.
4. Supernatant was discarded from the plate.
5. The beads was washed with 200µl freshly prepared 80% ethanol. Mixing of beads and the ethanol was done by sliding the 96 well plate over the magnetic plate for about 30 seconds.
6. Ethanol was removed.
7. The washing of the beads was again repeated.
8. Ethanol was completely removed by using 20µl pipette.
9. The beads was air dried on the magnetic plate for 15 minutes.
10. 12µl HIDI-Formamide was added to the dried beads and was mixed by pipetting.
11. Plate was sealed and given a short spin.
12. The DNA was denatured at 94ºC for 5 minutes on a PCR machine and was immediately chilled at -20ºC for 5 minutes.
13. The seal was removed and the mixture was again mixed by pipetting.
14. Then the plate was kept on magnetic plate for 3 minutes.
15. 10µl of the supernatant was transferred to a new 96 well plate and given a short spin.
16. The sample sheet was prepared and the samples were transferred into 384 well plate for capillary sequencing.
17. Plate was then linked into ABI 3130Xl Genetic Analyzer (Applied Biosystems, USA) for targeted capillary sequencing.

# References


Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics,* 7.20. 1-7.20. 41.

Akhavan, S., Chafa, O., Obame, F. N., Torchet, M. F., Reghis, A., Fischer, A. M. & Tapon-Bretaudiere, J. 2007. Recurrence of a Phe31Ser mutation in the Gla domain of blood coagulation factor X, in unrelated Algerian families: a founder effect? *Eur J Haematol,* 78(5)**,** pp 405-9.

Bagnall, R., Giannelli, F. & Green, P. 2006. Int22h-related inversions causing hemophilia A: A novel insight into their origin and a new more discriminant PCR test for their detection. *Journal of Thrombosis and Haemostasis,* 4(3)**,** pp 591-598.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A. & Shendure, J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet,* 12(11)**,** pp 745-55.

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. A., Jiang, H. & Feng, G. 2014. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer informatics,* 67-83.

Behjati, S. & Tarpey, P. S. 2013. What is next generation sequencing? *Archives of disease in childhood-Education & practice edition,* edpract-2013-304340.

Berber, E., James, P. D., Hough, C. & Lillicrap, D. 2009. An assessment of the pathogenic significance of the R924Q von Willebrand factor substitution. *J Thromb Haemost,* 7(10)**,** pp 1672-9.

Biesecker, L. G., Shianna, K. V. & Mullikin, J. C. 2011. Exome sequencing: the expert view. *Genome Biology,* 12(9)**,** pp 1-3.

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* btu170.

Bowen, D. J., Standen, G., Mazurier, C., Gaucher, C., Cumming, A., Keeney, S. & Bidwell, J. 1998. Type 2N von Willebrand disease: rapid genetic diagnosis of G2811A (R854Q), C2696T (R816W), T2701A (H817Q) and G2823T (C858F)–detection of a novel candidate type 2N mutation: C2810T (R854W). *Thrombosis and haemostasis,* 80(1)**,** pp 32-36.

Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics,* 14(10)**,** pp 681-691.

Brown, D. & Kouides, P. 2008. Diagnosis and treatment of inherited factor X deficiency. *Haemophilia,* 14(6)**,** pp 1176-1182.

Buermans, H. & Den Dunnen, J. 2014. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease,* 1842(10)**,** pp 1932-1941.

Casonato, A., Sartorello, F., Cattini, M. G., Pontara, E., Soldera, C., Bertomoro, A. & Girolami, A. 2003. An Arg760Cys mutation in the consensus sequence of the von Willebrand factor propeptide cleavage site is responsible for a new von Willebrand disease variant. *Blood,* 101(1)**,** pp 151-156.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S. & Sanjad, S. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences,* 106(45)**,** pp 19096-19101.

Dolled-Filhart, M. P., Lee, M., Ou-yang, C.-w., Haraksingh, R. R. & Lin, J. C.-H. 2013. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *The Scientific World Journal,* 2013(

Franchini, M. & Mannucci, P. M. The history of hemophilia. Seminars in thrombosis and hemostasis, 2014. Thieme Medical Publishers, 571-576.

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., Jovanovich, S. B., Nelson, J. R., Schloss, J. A. & Schwartz, D. C. 2009. The challenges of sequencing by synthesis. *Nature biotechnology,* 27(11)**,** pp 1013-1023.

Gandhi, S. & Scaria, V. 2016. *The Hitchhiker's Guide to Whole Exome Analysis*.

Gaucher, C., Mercier, B., Jorieux, S., Oufkir, D. & Mazurier, C. 1991. Identification of two point mutations in the von Willebrand factor gene of three families with the 'Normandy'variant of von Willebrand disease. *British journal of haematology,* 78(4)**,** pp 506-514.

Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. 2012. Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics,* 20(5)**,** pp 490-497.

Goodeve, A. C. 2010. The genetic basis of von Willebrand disease. *Blood Rev,* 24(3)**,** pp 123-34.

Goodwin, S., McPherson, J. D. & McCombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17(6)**,** pp 333-351.

Graw, J., Brackmann, H.-H., Oldenburg, J., Schneppenheim, R., Spannagl, M. & Schwaab, R. 2005. Haemophilia A: from mutation analysis to new therapies. *Nature Reviews Genetics,* 6(6)**,** pp 488-501.

Haberichter, S. L. 2015. von Willebrand factor propeptide: biology and clinical utility. *Blood,* 126(15)**,** pp 1753-61.

Herrmann, F. H., Auerswald, G., Ruiz-Saez, A., Navarrete, M., Pollmann, H., Lopaciuk, S., Batorova, A., Wulff, K. & Greifswald Factor, X. D. S. G. 2006. Factor X deficiency: clinical manifestation of 102 subjects from Europe and Latin America with mutations in the factor 10 gene. *Haemophilia,* 12(5)**,** pp 479-89.

Herrmann, F. H., Navarette, M., Salazar-Sanchez, L., Carillo, J. M., Auerswald, G. & Wulff, K. 2005. Homozygous Factor X gene mutations Gly380Arg and Tyr163delAT are associated with perinatal intracranial hemorrhage. *J Pediatr,* 146(1)**,** pp 128-30.

Hertzberg, M. 1994. Biochemistry of factor X. *Blood reviews,* 8(1)**,** pp 56-62.

Hilbert, L., Jorieux, S., Fontenay-Roupie, M., Guicheteau, M., Fressinaud, E., Meyer, D. & Mazurier, C. 2004. Expression of two type 2N von Willebrand disease mutations identified in exon 18 of von Willebrand factor gene. *British journal of haematology,* 127(2)**,** pp 184-189.

Hoffbrand, A. 2011. *Postgraduate haematology*, Sixth: A John Wiley & Sons Ltd.

Hoffbrand, V., Higgs, D. R., Keeling, D. M. & Mehta, A. B. 2016. *Postgraduate haematology*: John Wiley & Sons.

Ingram, G. 1976. The history of haemophilia. *Journal of clinical Pathology,* 29(6)**,** pp 469-479.

James, P. & Lillicrap, D. 2013. The molecular characterization of von Willebrand disease: good in parts. *British journal of haematology,* 161(2)**,** pp 166-176.

James, P. D. & Goodeve, A. C. 2011. von Willebrand disease. *Genet Med,* 13(5)**,** pp 365-76.

Jayandharan, G., Viswabandya, A., Baidya, S., Nair, S., Shaji, R., George, B., Chandy, M. & Srivastava, A. 2005. Six novel mutations including triple heterozygosity for Phe31Ser, 514delT and 516T→ G factor X gene mutations are responsible for congenital factor X deficiency in patients of Nepali and Indian origin. *Journal of Thrombosis and Haemostasis,* 3(7)**,** pp 1482-1487.

Jayandharan, G. R., Srivastava, A. & Srivastava, A. Role of molecular genetics in hemophilia: from diagnosis to therapy. Seminars in thrombosis and hemostasis, 2012. Thieme Medical Publishers, 64-78.

Kaadan, A. N. & Angrini, M. 2010. Who discovered hemophilia? *J Int Soc Hist Islamic Med,* 8-9.

Kasper, D. L., Fauci, A. S., Hauser, S. L., Longo, D. L., Jameson, J. L. & Loscalzo, J. 2015. *Harrison's Principles of Internal Medicine*, 19th: McGraw-Hill Education.

Kaushansky, K., Lichtman, M. A., Prchal, J., Levi, M. M., Press, O., Burns, L. & Caligiuri, M. 2015. *Williams Hematology, 9E*: McGraw-Hill Education.

Kircher, M., Heyn, P. & Kelso, J. 2011. Addressing challenges in the production and analysis of illumina sequencing data. *BMC genomics,* 12(1)**,** pp 382.

Ku, C. S., Cooper, D. N., Polychronakos, C., Naidoo, N., Wu, M. & Soong, R. 2012. Exome sequencing: dual role as a discovery and diagnostic tool. *Annals of neurology,* 71(1)**,** pp 5-14.

Kumar, P., Henikoff, S. & Ng, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols,* 4(7)**,** pp 1073-1081.

Lee, C. A., Berntorp, E. E. & Hoots, W. K. 2014. *Textbook of Hemophilia*, Third: John Wiley & Sons.

Lenting, P., Casari, C., Christophe, O. & Denis, C. 2012. von Willebrand factor: the old, the new and the unknown. *Journal of thrombosis and haemostasis,* 10(12)**,** pp 2428-2437.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25(14)**,** pp 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. 2009a. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25(16)**,** pp 2078-2079.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. 2009b. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25(16)**,** pp 2078-9.

Lillicrap, D. 2007. von Willebrand disease—phenotype versus genotype: deficiency versus disease. *Thrombosis research,* 120(S11-S16.

Liu, L., Hu, N., Wang, B., Chen, M., Wang, J., Tian, Z., He, Y. & Lin, D. 2011. A brief utilization report on the Illumina HiSeq 2000 sequencer. *Mycology,* 2(3)**,** pp 169-191.

Liu, Q., Nozari, G. & Sommer, S. S. 1998. Single-tube polymerase chain reaction for rapid diagnosis of the inversion hotspot of mutation in hemophilia A. *Blood,* 92(4)**,** pp 1458-1459.

Lunter, G. & Goodson, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res,* 21(6)**,** pp 936-9.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. 2011. What can exome sequencing do for you? *Journal of medical genetics,* jmedgenet-2011-100223.

Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Trends in genetics,* 24(3)**,** pp 133-141.

Mazurier, C., Parquet-Gernez, A., Gaucher, C., Lavergne, J. M. & Goudemand, J. 2002. Factor VIII deficiency not induced by FVIII gene mutation in a female first cousin of two brothers with haemophilia A. *British journal of haematology,* 119(2)**,** pp 390-392.

Menegatti, M., Karimi, M., Garagiola, I., Mannucci, P. & Peyvandi, F. 2004. A rare inherited coagulation disorder: combined homozygous factor VII and factor X deficiency. *American journal of hematology,* 77(1)**,** pp 90-91.

Menegatti, M. & Peyvandi, F. 2009. Factor X deficiency. *Semin Thromb Hemost,* 35(4)**,** pp 407-15.

Metzker, M. L. 2010. Sequencing technologies—the next generation. *Nature reviews genetics,* 11(1)**,** pp 31-46.

Michiels, J. J., Gadisseur, A., Vangenegten, I., Schroyens, W. & Berneman, Z. 2009. Recessive von Willebrand disease type 2 Normandy: variable expression of mild hemophilia and VWD type 1. *Acta Haematol,* 121(2-3)**,** pp 119-27.

Miller, S., Dykes, D. & Polesky, H. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids research,* 16(3)**,** pp 1215.

Morozova, O. & Marra, M. A. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics,* 92(5)**,** pp 255-264.

Peyvandi, F., Duga, S., Akhavan, S. & Mannucci, P. 2002. Rare coagulation deficiencies. *Haemophilia,* 8(3)**,** pp 308-321.

Peyvandi, F., Jayandharan, G., Chandy, M., Srivastava, A., Nakaya, S., Johnson, M., Thompson, A., Goodeve, A., Garagiola, I. & Lavoretano, S. 2006. Genetic diagnosis of haemophilia and other inherited bleeding disorders. *Haemophilia,* 12(s3)**,** pp 82-89.

Peyvandi, F., Kunicki, T. & Lillicrap, D. 2013. Genetic sequence analysis of inherited bleeding diseases. *Blood,* 122(20)**,** pp 3423-31.

Pinotti, M., Marchetti, G., Baroni, M., Cinotti, F., Morfini, M. & Bernardi, F. 2002. Reduced activation of the Gla19Ala FX variant via the extrinsic coagulation pathway results in symptomatic CRMred FX deficiency. *Thrombosis and haemostasis,* 88(2)**,** pp 236-241.

Qin, H. H., Xing, Z. F., Wang, X. F., Ding, Q. L., Xi, X. D. & Wang, H. L. 2014. Similarity in joint and mucous bleeding syndromes in type 2N von Willebrand disease and severe hemophilia A coexisting with type 1 von Willebrand disease in two Chinese pedigrees. *Blood Cells Mol Dis,* 52(4)**,** pp 181-5.

Rabbani, B., Tekin, M. & Mahdieh, N. 2014. The promise of whole-exome sequencing in medical genetics. *J Hum Genet,* 59(1)**,** pp 5-15.

Rauch, A. & Lenting, P. 2013. On the versatility of von Willebrand factor. *Mediterranean journal of hematology and infectious diseases,* 5(1)**,** pp 2013046.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., Consortium, W. G. S., Wilkie, A. O., McVean, G. & Lunter, G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet,* 46(8)**,** pp 912-8.

Rizzo, J. M. & Buck, M. J. 2012. Key principles and clinical applications of" next-generation" DNA sequencing. *Cancer prevention research,* canprevres. 0432.2011.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nature biotechnology,* 29(1)**,** pp 24-26.

Rossetti, L. C., Radic, C. P., Larripa, I. B. & De Brasi, C. D. 2005. Genotyping the hemophilia inversion hotspot by use of inverse PCR. *Clinical chemistry,* 51(7)**,** pp 1154-1158.

Rudy, G. 2011. A Hitchhiker's Guide to Next-Generation Sequencing. *URL: http://www. goldenhelix. com/pdfs/whitepapers/Hitchhikers-Guide-to-NGS. pdf*.

Sadler, J. E. 1998. Biochemistry and genetics of von Willebrand factor. *Annual review of biochemistry,* 67(1)**,** pp 395-424.

Scaria, V. & Sivasubbu, S. 2015. *Exome Sequence Analysis and Interpretation*: Research in Genomics.

Schramm, W. 2014. The history of haemophilia–a short review. *Thrombosis research,* 134(S4-S9.

Shendure, J. & Ji, H. 2008. Next-generation DNA sequencing. *Nature biotechnology,* 26(10)**,** pp 1135-1145.

Shiltagh, N., Kirkpatrick, J., Cabrita, L. D., McKinnon, T. A., Thalassinos, K., Tuddenham, E. G. & Hansen, D. F. 2014. Solution structure of the major factor VIII binding region on von Willebrand factor. *Blood,* 123(26)**,** pp 4143-51.

Shin, J., Ming, G.-l. & Song, H. 2014. Decoding neural transcriptomes and epigenomes via high-throughput sequencing. *Nature neuroscience,* 17(11)**,** pp 1463-1475.

Simon, D. & Roisenberg, I. 2004. Type 2N von Willebrand disease mutations in Brazilian individuals. *Haemophilia,* 10(5)**,** pp 473-6.

Tai, S., Herzog, R., Margaritis, P., Arruda, V., Chu, K., Golden, J., Labosky, P. & High, K. 2008. A viable mouse model of factor X deficiency provides evidence for maternal transfer of factor X. *Journal of Thrombosis and Haemostasis,* 6(2)**,** pp 339-345.

Thim, L., Vandahl, B., Karlsson, J., Klausen, N., Pedersen, J., Krogh, T., Kjalke, M., Petersen, J., Johnsen, L. & Bolt, G. 2010. Purification and characterization of a new recombinant factor VIII (N8). *Haemophilia,* 16(2)**,** pp 349-359.

Uprichard, J. & Perry, D. J. 2002. Factor X deficiency. *Blood reviews,* 16(2)**,** pp 97-110.

Versteeg, H. H., Heemskerk, J. W., Levi, M. & Reitsma, P. H. 2013. New fundamentals in hemostasis. *Physiol Rev,* 93(1)**,** pp 327-58.

Veyradier, A., Boisseau, P., Fressinaud, E., Caron, C., Ternisien, C., Giraud, M., Zawadzki, C., Trossaert, M., Itzhar-Baikian, N., Dreyfus, M., d'Oiron, R., Borel-Derlon, A., Susen, S., Bezieau, S., Denis, C. V., Goudemand, J. & French Reference Center for von Willebrand, d. 2016. A Laboratory Phenotype/Genotype Correlation of 1167 French Patients From 670 Families With von Willebrand Disease: A New Epidemiologic Picture. *Medicine (Baltimore),* 95(11)**,** pp e3038.

Wallmark, A., Larson, P., Ljung, R., Monroe, D. & High, K. 1991. Molecular defect (Gla26Asp) and its functional consequences in a hereditary factor X deficiency (Factor X Malmo 4). *Blood,* 78(suppl 1)**,** pp 60.

Wang, K., Li, M. & Hakonarson, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research,* 38(16)**,** pp e164-e164.

Wang, Z., Liu, X., Yang, B.-Z. & Gelernter, J. 2013. The role and challenges of exome sequencing in studies of human diseases. *Frontiers in genetics,* 4(160.

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., Braxton, A., Beuten, J., Xia, F. & Niu, Z. 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine,* 369(16)**,** pp 1502-1511.

Zama, T., Murata, M., Watanabe, R., Yokoyama, K., Moriki, T., Ambo, H., Murakami, H., Kikuchi, M. & Ikeda, Y. 1999. A family with hereditary factor X deficiency with a point mutation Gla32 to Gln in the Gla domain (factor X Tokyo). *British journal of haematology,* 106(3)**,** pp 809-811.

Zhou, Y.-F., Eng, E. T., Zhu, J., Lu, C., Walz, T. & Springer, T. A. 2012. Sequence and structure relationships within von Willebrand factor. *Blood,* 120(2)**,** pp 449-458.

http://emedicine.medscape.com/article/779322

https://genohub.com/ngs-instrument-guide/

https://support.illumina.com/content/dam/illuminasupport/documents/documentation/software_ documentation/bcl2fastq/bcl2fastq_letterbooklet_15038058brpmi.pdf

https://support.illumina.com/content/dam/illuminasupport/documents/myillumina/a557afc4-bf0e-4dad-9e59-9c740dd1e751/casava_userguide_15011196d.pdf

http://support.illumina.com/downloads/truseq-exome-library-prep-reference-guide-15059911. html

https://www.cdc.gov/ncbddd/hemophilia/champs.html

https://www.hemophilia.org/Bleeding-Disorders/History-of-Bleeding-Disorders

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_ sequencing_introduction.pdf

https://www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf

https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

https://www.illumina.com/documents/techspotlights/techspotlight_sequencing.pdf

http://www.openbioinformatics.org/annovar/annovar_filter.html#ljb23

https://www.scigenom.com/whitepapers/exomesequencingandanalysis.pdf

http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf