



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, Secretary 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 5th 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.

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

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE NO.
	Acknowledgment	i
	Table of Contents	ii
	List of Abbreviations	v
	List of figures	viii
	List of Tables	ix
	Abstract	x
1	Introduction	
1.1	Background	1
1.2	Rare Bleeding Disorders	2
1.2.1	Types	2
1.2.2	Diagnosis	3
1.2.2.1	Laboratory Studies	3
1.2.2.2	Molecular Diagnosis	3
1.3	Lacunae of the study	5
1.4	Rationale	5
1.5	Hypothesis	5
1.6	Objectives	
1.6.1	Specific Objectives	6
1.6.2	General Objectives	6
2	Review of Literature	
2.1	Historical Background of Rare Bleeding Disorders	7
2.2	Pathophysiology of Blood Clot Formation	9
2.3	Characteristics of Rare Bleeding Disorders	11
2.3.1	Epidemiology	12
2.3.2	Clinical Presentation	13
2.3.3	Genetics	14
2.4	Next Generation Sequencing (NGS) and Whole Exome Sequencing (WES)	16

3	Materials and Methodology	
3.1	Samples' Selection Criteria	18
3.1.1	Inclusion Criteria	18
3.1.2	Exclusion Criteria	18
3.2	Sample Collection	18
3.3	Genomic DNA Extraction	18
3.4	DNA Quality Check and Quantification	19
3.5	Library Preparation and Sequencing	20
3.5.1	Library Preparation	20
3.5.2	Cluster Generation and Sequencing	27
3.6	Bioinformatics Analysis of Whole Exome Sequencing Data	28
3.6.1	Raw Sequencing Reads	29
3.6.2	Data Quality Check	29
3.6.3	Data Trimming	30
3.6.4	Alignment	31
3.6.5	Preprocessing and Variant Calling	31
3.6.6	Variant Annotation	31
3.6.7	Variant Prioritization	31
3.6.8	Validation of Putative Variants	33
4	Results	
4.1	Clinical Presentation and Family Pedigree	34
4.2	Genomic DNA Extraction and Quality Check	35
4.3	Exome Library Preparation	36
4.4	Bioinformatics Analysis of Sequenced Data of 2N VWD cases	37
4.4.1	Quality Check and Trimming	37
4.4.2	Alignment	39
4.4.3	Variant Calling	39
4.4.4	Variant Annotation and Prioritization	40
4.4.5	Validation by Capillary Sequencing	46
4.5	Bioinformatics Analysis of Sequenced Data of FXD cases	48
4.5.1	Quality Check and Trimming	48
4.5.2	Alignment	50
4.5.3	Variant Calling	50
4.5.4	Variant Annotation and Prioritization	50

	4.5.5	Validation by Capillary Sequencing	57
5		Discussion	59
	5.1	Whole Exome Sequencing (WES)	59
		5.1.1 WES in Illumina Platform	61
	5.2	Bioinformatics Analysis of WES Data	61
		5.2.1 Data QC and Trimming	62
		5.2.2 Sequence Alignment	62
		5.2.3 Variant Calling	62
		5.2.4 Variant Annotation	63
		5.2.5 Variant Prioritization	63
	5.3	Mutation in Family 1: 2N VWD cases	64
		5.3.1 WES and Bioinformatics Analysis	64
		5.3.2 Prevalence of R816W	65
		5.3.3 Biology of R816W	65
	5.4	Mutation in Family 2: FXD cases	66
		5.4.1 WES and Bioinformatics Analysis	66
		5.4.2 Prevalence of F71S	67
		5.4.3 Biology of F71S	67
		5.4.4 Evolutionary Conservation of F71S	68
6		Summary	70
7		Conclusion	71
		Appendices	73
		References	85

List of Figures

Figure	Title	Page No.
2.1	Pedigree of Queen Victoria	8
2.2	Mechanism of blood coagulation process	10
2.3	Graph representing total number of patients with bleeding disorders	13
2.4	Schematic of the human factor VIII gene (<i>F8</i>), the mRNA, and the protein	15
2.5	Schematic of the human factor X gene (<i>F10</i>), the mRNA, and the protein	16
2.6	Schematic representation of old and new domain arrangement of VWF	
3.1	Library Preparation Workflow of TruSeq Exome Library Preparation	20
3.2	Formula to convert ng/ μ L to nM	27
3.3	Preparation of library for cluster generation and Sequencing	28
3.4	Flow chart of bioinformatics analysis of the sequenced data	29
3.5	Quality plot of the raw sequencing reads	30
3.6	Variant prioritization strategy	32
4.1	Family pedigree of patients with 2N VWD	34
4.2	Family Pedigree of patients with FXD	35
4.3	Gel image of the extracted genomic DNA	35
4.4	Gel image of the pre- and post-captured library	36
4.5	Quality plot of the raw sequencing reads of TU01	38
4.6	Quality plot of the raw sequencing reads of TU18	38
4.7	Pipeline used for variants sorting	40
4.8	Pie chart representing the relative percentage of each variation	42
4.9	IGV Snapshot of the variant p.R816W	45
4.10	Localization of variant R816W in VWF protein	46
4.11	PCR amplification of genetic region encompassing the putative variant c.C2446T:p.R816W present on exon 19 of <i>VWF</i>	47
4.12	Chromatogram derived from targeted capillary sequencing of family members of 2N VWD case	48
4.13	Quality plot of the raw sequencing reads of TU03	49
4.14	Quality plot of the raw sequencing reads of TU25	49
4.15	Pipeline used for variants sorting	51
4.16	Pie chart representing the relative percentage of each variation	52
4.17	IGV Snapshot of the variant p.F71S	54
4.18	Localization of variant F71S in FX protein	55
4.19	PCR amplification of genetic region encompassing the putative variant c.T212C:p.F71S present in exon 2 of <i>F10</i>	57
4.20	Chromatogram derived from targeted capillary sequencing of family members of FXD case	58

List of Tables

Table	Title	Page No.
1.1	Severity classification of Hemophilia	2
1.2	Summary of coagulation screening tests	3
2.1	Plasma concentration and half-life of various clotting factors	11
2.2	Characteristics of bleeding disorders	12
3.1	Covaris parameter setting to fragment insert size of 150 bp	21
3.2	Summary of genes related to bleeding disorders	32
4.1	NanoDrop quantification of DNA samples	36
4.2	Quantification of pre- and post-captured library	37
4.3	Size of raw sequencing data	37
4.4	Summary of FastQC report of TU01 and TU18	39
4.5	Mapping summary of TU01 and TU18	39
4.6	Summary of total variants found in TU01 and TU18	41
4.7	Summary of various exonic mutations found in TU01 and TU18	41
4.8	No. of various mutations related to inherited bleeding disorders in TU01 and TU18	42
4.9	Details of Nonsynonymous SNVs of TU01	43
4.10	Details of Nonsynonymous SNVs in TU18	44
4.11	Putative mutation based on SIFT and Polyphen2 score in TU01 and TU18	46
4.12	Allele Frequency of the variant R816W in different population	46
4.13	Summary of FastQC report of TU03 and TU25	50
4.14	Mapping summary of TU03 and TU25	50
4.15	Summary of total variants found in TU03 and TU25	50
4.16	Summary of various exonic mutations found in TU03 and TU25	51
4.17	Types of mutations present on genes associated with inherited bleeding disorders in TU03 and TU25	52
4.18	Details of Nonsynonymous SNVs of TU03	53
4.19	Details of Nonsynonymous SNVs in TU25	54
4.20	Putative mutation based on SIFT and Polyphen2 score in TU03 and TU25	56
4.21	Allele Frequency of the variant F71S in different population	56
4.22	Conservation of F71 residue in FX protein among the vertebrates	57

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
<i>F10</i>	-	Coagulation Factor X gene
<i>F11</i>	-	Coagulation Factor XI gene
<i>F12</i>	-	Coagulation Factor XII gene
<i>F13A1</i>	-	Coagulation Factor XIII A Chain gene

<i>F13B</i>	-	Coagulation Factor XIII B Chain gene
<i>F2</i>	-	Coagulation Factor II gene
<i>F5</i>	-	Coagulation Factor V gene
<i>F7</i>	-	Coagulation Factor VII gene
<i>F8</i>	-	Coagulation Factor VIII gene
<i>F9</i>	-	Coagulation Factor IX gene
FC	-	Flow Cell
<i>FGA</i>	-	Fibrinogen Alpha Chain
<i>FGB</i>	-	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
<i>GGCX</i>	-	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
<i>VKORC1</i>	-	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