Bioinformatic analysis of Whole Exome Sequence data

Project work submitted to the Central University of Punjab

For the award of

Master of Science

Life science with Specialization in Human Genetics

In

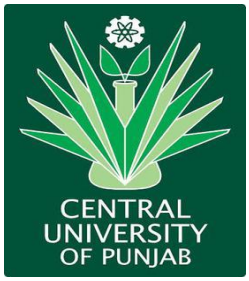
Department of Human Genetics and Molecular Medicine

By

Md Momin Ali

Supervisor

Dr. Preeti Khetarpal



Department of Human Genetics and Molecular Medicine

School of Health Science

Central University of Punjab, Bathinda

May, 2018

Central University of Punjab, Bathinda

Declaration

I declare that all the changes suggested by the external examiner in the Research project entitled " **Bioinformatic Analysis of Whole Exome Sequence data** " submitted by me for the award of degree of Master of life science with specialization in Human Genetics in the Department of Human Genetics and Molecular Medicine has been incorporated in the research project.

Md Momin Ali

Department of Human Genetics and Molecular Medicine

School of Health Science

Date: May, 2018

Dr. Preeti Khetarpal

Department of Human Genetics and Molecular Medicine

School of Health Science

Date: May, 2018

Certificate

I declare that the project entitled "**Bioinformatic analysis of Whole Exome Sequence data**" has been prepared by me under the guidance of Dr. Preeti Khetarpal assistant professor, department of Human Genetics and Molecular Medicine, Central University of Punjab, Bathinda. No part of this project work has formed the basis for the award of any degree or fellowship previously.

Md Momin Ali

Department of Human Genetics and Molecular Medicine

School of Health Science

Central university of Punjab,

Bathinda – 151001

Date: May, 2018

Certificate

I certify that Md. Momin Ali has prepared his project entitled "**Bioinformatic analysis of Whole Exome Sequence data**", for the award of M.Sc. degree from Central University of Punjab, Bathinda under my guidance. He has carried out this work at the Department of Human Genetics and Molecular Medicine, School of Health Science, Central University of Punjab, Bathinda.

Dr. Preeti Khetarpal

Professor, Department of Human Genetics and Molecular Medicine,

School of Health Science,

Central University of Punjab,

Bathinda – 151001

Date: May, 2018

ABSTRACT

Bioinformatic Analysis of Whole Exome Data

Name of the Student: Md Momin Ali

Registration number: 16mslshg13

Degree for which submitted: M.Sc. Life Sciences with specialization in Human Genetics

Name of Supervisor: Dr. Preeti Khetarpal

Name of Department: Department of Human Genetics and Molecular Medicine

Name of School: School of Health Sciences

Key words: Whole Exome Sequencing, Next Generation Sequencing, *in silico,* SIFT, PolyPhen, ANNOVAR.

Whole Exome Sequencing (WES) is a capture-based method developed to sequence exome and to identify variants in the coding region of genes that affect protein function. Understanding the exomes of individuals at single base resolution allows the identification of pathogenic variants responsible for causing genetic disease. In this paper we mentioned all the object of the project steps and bioinformatic computational tools involved step by step. The objective of the study was to find all the disease causing heterozygous variants in the proband from the WES data. Using *in silico* tools SIFT, PolyPhen and ANNOVAR. All the annotated non-synonymous SNPs was arranged and filtered according to the pathogenicity of variants. All the variants were narrowed down to five matching variants. associated with clinical phenotype of the proband. As a conclusion Bioinformatic analysis of WES data proved to be a good tool for finding disease causing variants. Majority of the tools used in the analysis are generally linux based with poor user interface which makes it a challenging experience for a non-computational individual. Future of this field is dependent on software developers, so that more people can understand and help in the progress.

Md Momin Ali Dr. Preeti Khetarpal

ACKNOWLEDGEMENTS

I express my intense sense of gratitude to my supervisor **Dr. Preeti Khetarpal**, assistant professor, department of Human Genetics and Molecular Medicine, Central University of Punjab, Bathinda whose guidance was always available to me. Whenever I ran into a trouble spot or had a question about my research or writing, she steered me in the right direction. I am grateful to her for her resourceful guidance, constant encouragement, imperative suggestion and support.

I am extremely thankful to **Dr. Anjana Munshi**, professor and COC & Dean of department of Human Genetics and Molecular Medicine, Central University of Punjab, Bathinda for her constant encouragement.

I also extend my thanks to all the faculty members of center for human genetics and molecular medicine namely **Dr. Sandeep Singh, Dr. Harish Chander, Dr. Sabyasachi Senapati,** for their kind cooperation and support whenever required. I am tremendously grateful to our laboratory assistant **Mr. Roshan Kumar**, whose help is completely available to me.

At last I must express my very Profound thankfulness to my parents my dad and mom for providing me with unfailing support and continuous encouragement throughout my two years of study and through the process of research and writing this Project Work. This accomplishment would not have been possible without them. Thank you

Md Momin Ali

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

Sr. No.	Full Form	Abbreviation
1.	Autism spectrum disorder	ASD
2.	Congenital adrenal hyperplasia	CAH
3.	Cholinergic receptor nicotinic delta subunit	CHRND
4.	Copy Number Polymorphisms	CNP
5.	Copy number variation	CNV
6.	Cerebroocular dysplasia-muscular dystrophy syndrome	COD-MD
7.	Cytotoxic T-lymphocyte protein 4 precursor	CTLA4
8.	Common variable immune deficiency	CVID
9.	Cytochrome P450 family 21 subfamily A member 2	CYP21A2
10.	Database of Genotypes and Phenotypes	dbGaP
11.	FYVE, RhoGEF And PH Domain Containing 5	FGD5
12.	Fukuyama Type Congenital Muscular Dystrophy	FKTN
13.	Immunoglobulin superfamily member 3	IGSF3
14.	Ikaros family zinc finger 1	IKZF1
15.	Light chain deposition disease	LCDD
16.	Lipopolysaccharide-responsive and beige-like anchor protein	LRBA
17.	Methicillin-resistant Staphylococcus aureus	MRSA
18.	Nuclear factor NF-kappa-B p105	NFKB1
19.	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta	PIK3CD
20.	Slow Channel Myasthenic Syndrome	SCCMS

21.	Solute Carrier Family 26 (Anion Exchanger)	SLC26A3
22.	Single nucleotide polymorphism	SNP
23.	Supported Oligo Ligation Detection	SOLiD
24.	Signal transducer and activator of transcription 3	STAT3
25.	Syntaxin-Binding Protein 2	STXBP2
26.	Singular Value Decomposition	SVD
27.	Tumor necrosis factor receptor superfamily member	TNFRSF
28.	Walker–Warburg syndrome	WWS

INTRODUCTION

1. Introduction

Each molecule of the Deoxyribonucleic acid (DNA), genetic material in human cell is approximately 3-meter-long and consists of 3.0 $\times 10^9$ bp (base pairs). The DNA consists of coding and non-coding region. Approximately **1.5** percentage of DNA constitutes exonic region. Mutation which is responsible for causing genetic disorders. More than 4,000 Mendelian disorders known 2,937 genes has been discovered but the genes causing 50% of all known Mendelian phenotypes are still unknown, and many more Mendelian conditions have yet to be recognized (*Chong et al, 2015)*. In some cases when it becomes difficult to associate a gene with a phenotype; WES can be an important diagnostic tool to identify those variants.

The processes of determining the exact sequence of nucleotides within a DNA molecule is getting advance and cheaper. There are different choices for sequencing, namely Whole Genome Sequencing (WGS) or Whole Exome Sequencing (WES). Because WES is precise, less time consuming and more cost effective it is often preferred. Whole exome sequencing plays a very important role in identifying new mendelian disorder genetic basis of many disease, and unknown congenital abnormalities.

After announcement of Human Genome project in 2003 many researchers published papers on Sequencing. In 2005 the first article was released on whole exome sequencing which emerged as a successful diagnostic tool in the study of genetic disease and has proven to be particularly effective in identifying disease-associated genes that are refractory to linkage analysis.

According to Ku *et al*, 2012 commonly adopted approaches to identify causal mutation from the WES data are removal of common variants using public databases such as the 1000 Genome Project and then identification of deleterious variants i.e nonsense and missense mutation, splice site variants and coding indels followed by step to predict the functional effects of variants using *in silico* tools such as SIFT, PolyPhen and ANNOVAR. For unrelated cases common variants are found in a single gene shared by all the patients

and for trios, identifying *de novo* variants that might be causative for sporadic cases of dominant disorders. If there is multiple affected siblings intersection of data is needed to identify common shared variants. Compound heterozygous variant is usefull in predicting recessive disorder and heterozygous variants are dominant disorders.

There are different techniques and tools are available important for exome sequencing analysis. In recent years there is a huge advancement in the field of omics, Platform like NGS are providing the fastest sequencing technique. And the cost of sequencing is decreasing day by day. A Proband was identified with multiple congenital malformation and unexplained genetic causes

Availability of open source software in the internet plays a very significant role in the advancement of this field. And also, the software developers with better GUI are very less in this field.

Objectives

Identification of pathogenic heterozygous variant of a Whole exome sequenced data.

REVIEW OF LITERATURE

2. Review of literature

In the development of WES technique Next generation sequencing (NGS) plays an important role. In 2005 the first commercial NGS instrument was released, 454 Life Sciences released the 454 sequencer, followed by Solexa Genome Analyzer and SOLiD (Supported Oligo Ligation Detection) by Agencourt in 2006. NGS is a massively parallel or deep sequencing are related terms that describe a DNA sequencing technology which has revolutionized genomic research. Using NGS an entire human genome can be sequenced within a single day.

After the introduction of NGS in 2005 many articles started to came up with different ideas. Whole exome sequencing was started in 2009 for disease gene identification. The first article was (*Choi et al*, 2009) reported the clinical utility of whole exome sequencing, they used whole-exome capture on single arrays on the Roche/NimbleGen platform to the Illumina sequencing platform**.** Five patients, 4 from Turkey and 1 from Saudi Arabia suspected to have Bartter syndrome but who did not have mutations in known genes for this disease were chosen for the disease gene discovery it was reported that gene discovery and clinical diagnosis was homozygous deleterious mutations in *SLC26A3**.***

2.1 Steps involved in the WES technique

Ku et al, 2012 illustrated all the basic steps involved in the exome sequence data analysis as follows:

2.1.1 Sample preparation and Sequencing

There are many different ways to prepare sample depending on the sequencing technique. Fisher *et al (2011)* described automated Illumina DNA library construction, Agarose gel-based size selection, has now been replaced by fully automated, bead-based steps. Evenness of sequence coverage among captured targets, avoiding extremely overrepresented targets. High library quality, in order that libraries contain massive numbers of ordination fragments. Reproducibility, so that performance

of the process is highly predictable. Low cost of the targeting process relative to sequencing. Quality control checkpoints built into the process to identify poor pergilis prior to sequencing; and limited human labor.

Ku *et al, 2012* divided the sample preparation in 4 simple steps: -

1. Genomic DNA extracted from peripheral blood
2. Sequencing library preparation
3. Exome capture and enrichment
4. Sequencing to sufficient coverage

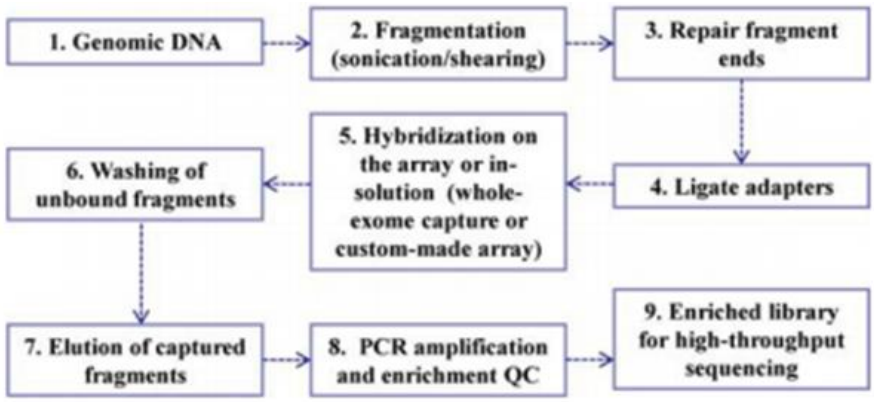


Figure 2.1: Workflow of library preparation and exome or targeted genomic regions capture (source- Ku *et al,* 2012).

The genomic DNA is used for sequencing library preparation involving several steps such as DNA fragmentation and adapter ligation (steps 1 and 2). Exome capture and enrichment is usually performed using commercial kits such as the Agilent, Illumina, and Nimblegen human exome capture kits. Exome sequencing is then performed using high-throughput next-generation sequencing (NGS) technologies such as Illumina Genome Analyzer or HiSeq, Life Technologies SOLiD, and Roche 454 Genome Sequencer. These steps are performed according to the manufacturer's recommended protocols. An average coverage of approximately 30-fold is deemed sufficient for the accurate calling of variants (steps 3 and 4).

2.1.2 Primary DATA processing

According to *Ku et al, 2012* primary data processing includes 3 steps: -

1. Data processing to raw sequence reads
2. Reads alignment to reference genome using alignment tools like ELAND and BWA
3. PCR duplicates removal

2.1.3 Secondary DATA processing

The analytical pipeline is then followed by variant calling and filtering to remove false positives according to commonly applied quality control criteria. Finally, the variants are annotated to obtain information such as genomic position and functional effect missense and nonsense variants, before they are examined to identify the causal mutations.

Sequencing is done using High throughput illumine Hi-seq2500 via Medgenome, library preparation is done using SureSelectXT Human All Exon V5 kit, fragments were them end paired and ligated with capture adpaters. Then the library was amplified with adapter specific primers, the product is then purified using Agencourt AMPure XP beads and then fragment distribution is checked using DNA D1000 screen tapes. gDNA hybridization to SureSelect all exon V5 cRNA Biotin for 16 – 24 hours at 65°C, after post incubation the PCR product is purified using Agencourt AMPure XP beads, the final library was assessed for fragment size, the quantified library is amplified on cBOT and sequencedon the HiSeq 4000 with 100bp paired end.

2.2 Computational tools

Tools – Hintzsche *et al* (2016) reviewed several computational tools for the analysis and interpretation of WES data with an elaborate diagram given below: -

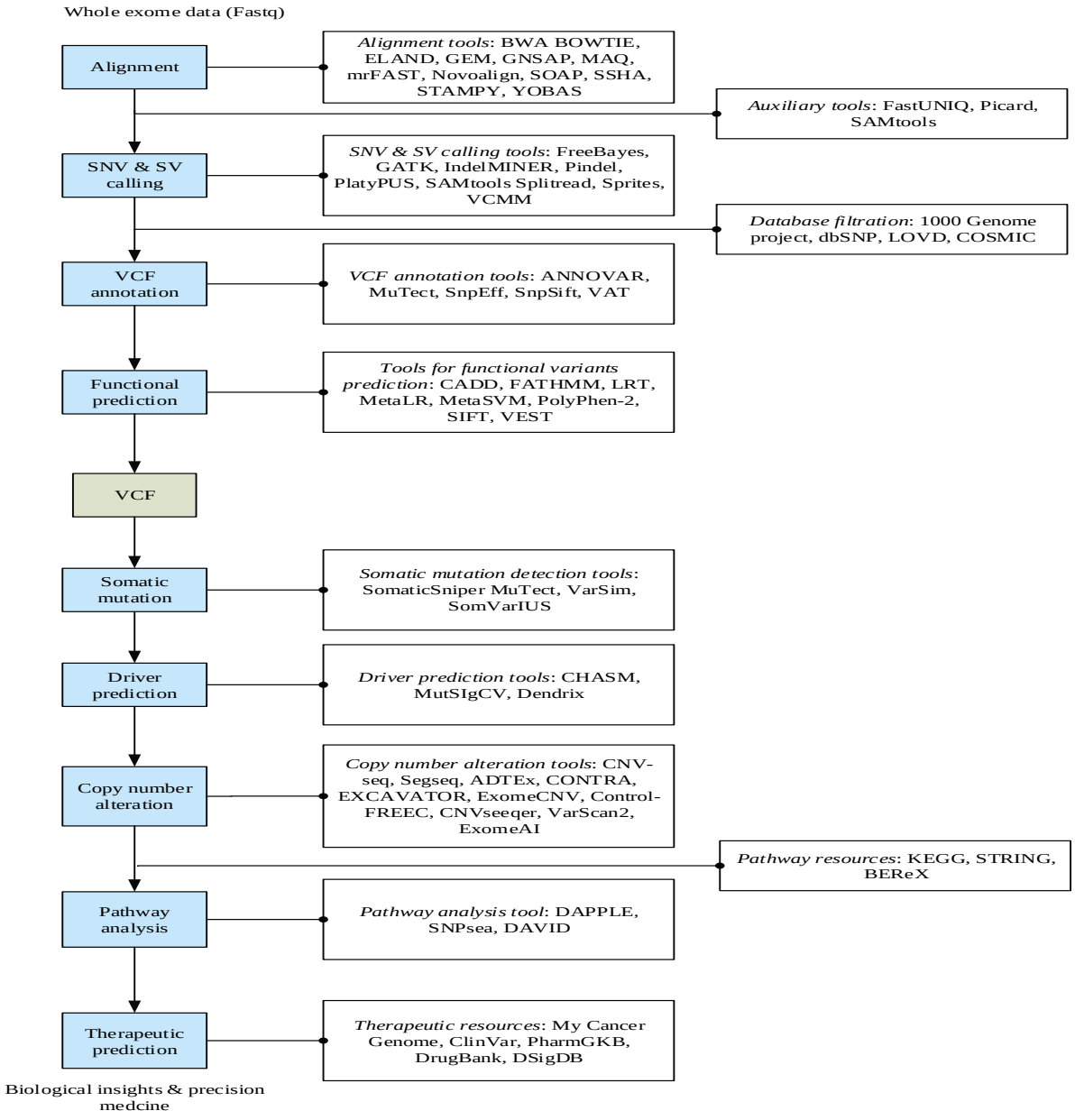


Figure 2.2: All the tools involved step wise (Source -Hintzsche *et al*., 2016)

2.3 Challenges and Drawbacks of the technique

WGS availability is increased due to its vast usage, demand and improved cost efficiency. WGS is predicted to be more economical than WES because the capture process is skipped entirely**.** WES has the advantage of capturing all of the exome (as some can be missed by the exome capture process) and can provide information on variants in highly evolutionary conserved non-coding regions and other variants throughout the genome. The bioinformatics filtering techniques, storage facilities, software and hardware for data analysis will be a challenge and most ongoing projects initially focus on the exome for the first analysis. There is variability in sequencing efficiency across the genome and the fluctuations in coverage will result in many regions of interest being missed, repetitive regions exonic and others are difficult to align in either case and can result in missed variants or an excess of variant calls. Uniform library construction, higher sequencing depth and longer reads and paired end fragment sizes are of low accuracy which in future may increase the accuracy. In comparison to WGS, WES provides most of the benefits of WGS but with lower costs, both for sequencing and for storage and analysis of the data

Ku et al, 2012 mentioned there will be a major challenge to analyze and manage the large amounts of sequencing data generated. Exome sequencing typically generates >10,000 genetic variants that must be carefully filtered to identify the causal mutations. Therefore, the analytical pipeline must be simple enough for the analysis to be handled by the nonspecialist, requiring the development of robust user-friendly software, usable in the clinical setting.

Second, the sensitivity and specificity of the detection of microlesions (i.e. point mutations, small indels, and tandem repeats) and macrolesions (i.e. deletions and duplications) need to be improved to standard clinical diagnostic tests. NGS technologies have higher raw base-calling error rates than Sanger sequencing, although this should be remediable by increasing coverage to achieve higher consensus accuracy. Depth of sequencing coverage is additionally vital for characteristic heterozygous (whether inflicting dominant Mendelian disorders or occurring de novo) or compound heterozygotes

Currently, the causative mutations detected by exome sequencing are typically valid by Sanger sequencing. However, for a clinical diagnostic application, this may increase prices. Although these 2 parameters should be optimized further, false positives are less of a critical challenge for the technology, because variant calling criteria can be made more stringent, the depth of coverage can be increased, and final findings can be validated by alternative means. However, false positives could potentially mask the identification of true disease-causing mutations if not allowed for properly. False negatives must be managed more carefully, and it will be essential to provide clinicians and medical geneticists with a clear idea of how effective the procedure has been for each patient prior to communicating the diagnostic results.

Third, the success of exome sequencing depends upon the complete capture of the exome followed by an adequate depth of sequencing. Thus, further improvements in the efficiency of exome capture methods are needed to ensure that all the target exons are captured. Incomplete capture and uneven coverage may be due to technical limitations or the nature of the DNA sequences.

Fourth, the time-to-diagnosis, once exploitation exome sequencing as a diagnostic application, time is additionally a very important issue to think about, particularly within the case of disorders that a fast andand accurate genetic diagnosis would significantly influence clinical management. Currently available sequencing instruments take several days for the sequencing to be completed. However, the recently launched IonTorrent sequencing platform achieved the fastest sequencing time (completion in <2 hours). This sequencing speed, together with a streamlined and robust analytical pipeline, should ensure that both clinicians and patients obtain the requisite results in a timely manner.

In case of whole genome analysis, it takes so much time and since the exome sequence is only 1% it takes less time depending on the hardware being used.

2.4 Applications of WES

"Clinical application of whole-exome sequencing (WES) represents succeeding step within the progression to complete elucidation of the genomic determinants of a patient's hereditary make-up.". Using ever-evolving technologies, WES can provide insights into known and unknown variations in approximately 95% of the individual patient's genome. Until very recently, the time, cost, and technical expertise required to generate and analyze WES data largely precluded serious consideration of its use outside of research settings. This can no longer be assumed to be the case

While some applications, such as single cell microbiology and clinical metagenomics, are currently confined to translational research. Whole exome sequencing is increasingly being employed for outbreak investigation and bacterial genomic epidemiology, species identification, and to a lesser extent, culture independent microbiology and susceptibility testing. In particular, WGS has been instrumental in tracking outbreaks and unraveling the epidemiology of carbapenem-resistant *Klebsiella pneumoniae*, methicillin-resistant *Staphylococcus aureus* (MRSA) difficile enter hemorrhagic *Escherichia coli*. These studies have been invaluable in expanding knowledge of bacterial epidemiology and provide striking examples of the transformative potential of WGS.

2.4.1 Genetic Diagnosis by WES

Choi *et al* (2009) demonstrated the utility of whole-exome sequencing for identification of rare variants and disease-causing mutations. They captured and sequenced samples from 5 Caucasian subjects with suspected diagnosis of Bartter syndrome, a renal salt-wasting disease and reported who had homozygous mutations in *SLC26A3.*

Niklas Krumm *et al (*2015) used exome sequencing data from eight HapMap individuals and exomes from 122 mother-father-proband ASD trios (for 366 total individuals). They outlined a method for making read-depth data from exomes amenable to rare CNV discovery, as well as copy number genotyping of CNP loci. They used SVD standardization to beat a number of coverage biases introduced by the capture and sequencing of exomes. This method allows for differing sample preparations and capture reactions to be integrated into the same experiment, provided each ‘‘batch’’ is sufficiently

large enough. This includes correct normalization of the X chromosome, such that deletions and duplications can be assayed regardless of the sample's sex this method can integrate exomes captured with different exome capture target designs.

Stark, Z. *et al.* (2016) (2016) did smart analysis of whole exome sequencing as a primary tier molecular take a look at in new born with suspected heritable disorders. DNA was extracted from peripheral blood, and exome sequencing used Nextera v1.2 Rapid Exome Capture Kit and TruSeq Rapid SBS chemistry on a HiSeq. The mean coverage obtained was 145.38. Variants were characterized using the Melbourne Genomics Health Alliance shared bioinformatics pipeline, Cpipe11 version 0.9.8 Cpipe, a clinically oriented pipeline built on Bpipe, includes the Burrows-Wheeler Aligner and the Genome Analysis. Variants were initially filtered against an internal database of variant counts to exclude technical and locally recurrent variants.

Nonsynonymous variants, splice-site variants (up to 10 base pairs), and small insertions and deletions present at a frequency of less than 1% in the Exome Sequencing Project and 1000 Genomes Project repositories were assessed further. Only variants in HUGO Gene Nomenclature Committee genes associated with Mendelian disease. The analysis pipeline excluded variants in highly penetrant genes predictive of adult-onset diseases such as familial cancer risk genes, cardiovascular disease risk genes, and late-onset neurological disease genes (122 genes). Analysis of one or more of these genes was available only when relevant to the participant's phenotype

Maffucci P *et al* (2016) did genetic diagnosis using whole exome sequencing in common variable immunodeficiency on 50 patients revealed 433 variations in 38 patients (76% of the cohort), of which 64 (in 38 genes) were either private or rare. While CVID (Common variable immune deficiency) has long been considered a clinically heterogeneous group of PIDs, the gene defects that underlie the complex syndrome have been difficult to identify, particularly in non-familial cases. In this study, Maffucci P *et al* collected complete exome data for 50 selected CVID subjects, most of whom were sporadic, who had severe phenotypes found in about half of all CVID cases. Goal was to identify variants leading to the manifestations, filtering results demonstrated likely disease-causing mutations

in *NFKB1*, *STAT3*, *CTLA4*, *PIK3CD*, *IKZF1*, *LRBA*, and *STXBP2* in 15 patients (30%) and disease-associated mutations in TNFRSF13B in 8 patients (16%).

Yang *et al.* (2017) reported data on the first 250 probands for whom referring physicians ordered whole-exome sequencing. Patients presented with a range of phenotypes suggesting potential genetic causes. Approximately 80% were children with neurologic phenotypes. 86 mutated alleles that were extremely seemingly to be anorectic in sixty two of the 250 patients was known, achieving a twenty fifth molecular diagnostic rate (95% confidence interval, 20 to 31). Among the 62 patients, 33 had autosomal dominant disease, 16 had autosomal recessive disease, and 9 had X-linked disease. A total of 4 probands received a pair of non-overlapping molecular diagnoses, that most likely challenged the clinical identification that had been created on the thought of history and physical examination. A total of 83% of the total autosomal dominant mutant alleles and 40% of the X-linked mutant alleles occurred de novo. Recurrent clinical phenotypes occurred in patients with mutations that were extremely probably to be present within the same genes and in numerous genes accountable for genetically heterogeneous disorders

Swaminathan *et al (2017)* mentioned a typical clinical exome sequencing workflow from patient first presenting symptoms at the clinic to care management based on the sequencing report

2.4.2 Exome sequencing studies can be applied to WGS studies: Biesecker *et al* (2011) Exomes offer us with low hanging fruit – to dissect the genetic design of an attribute, culling out potential high penetrance variants from exome, assessing the remaining heritability, and then tackling that remainder (assuming it is significant) with WGS would be a practical and economical approach. But in this specific example of finding breakpoints, deep WGS isn't necessary, one simply wants deep physical coverage with massive spanning paired-end reads. If they can build robust annotation pipelines for a WES sequence, they can extend and generalize the lessons learned from that activity into interpretation of intronic and intergenic variation. The methods used to identify these variants are fully transferable to working with whole genome sequencing data. The exome sequencing approach has been a cost-effective option for sequencing

the human genome and has resulted in the identification of many disease-causing variants

2.5 Advantages of whole exome over whole genome sequencing: To compare this level of coverage to what can be obtained using exome sequencing, J Clark *et al* (2011) normalized the exome sequencing data to 50M reads for each platform because this level allows multiplexing at least 3 and up to 6 exomes per lane. To do such a comparison, they performed WGS to high read depth on an Illumina HiSeq 2000 on a blood sample from the same individual analyzed for the exome sequencing comparison. WGS requires a much greater amount of sequencing to achieve equivalent coverage as exome sequencing, but its performance relative to exome sequencing has not been well described.

According to Kiezun *et al* (2012) the effect size is not only a property of an individual variant but rather a reflection of the distribution of effects coupled with how those effects are interpreted via the test. Therefore, power is higher for genes with more variants, for example, larger genes or genes in regions with elevated mutation rate. Additionally, genes in which most variants are causal are easier to identify than those in which few variants are causal. In individual candidate gene sequencing studies, estimates of this proportion ranged from 30–70%. The power of an exome sequencing study is limited by the amount of variation in genes. Enthusiasm for exome sequencing studies stems, in part, from successful candidate gene sequencing studies, and, so, they sought to test whether exome sequencing would be expected to have sufficient power to detect genes discovered by the candidate gene approach. Given the sample sizes, the likely effect sizes and frequencies of causal variants and the proportion of causal variants in a gene, do current exome sequencing studies have sufficient power to detect genes underlying complex phenotypes. This is particularly notable, because some candidate gene studies used much larger sample sizes (thousands of individuals) than ongoing exome sequencing studies (hundreds of individuals).

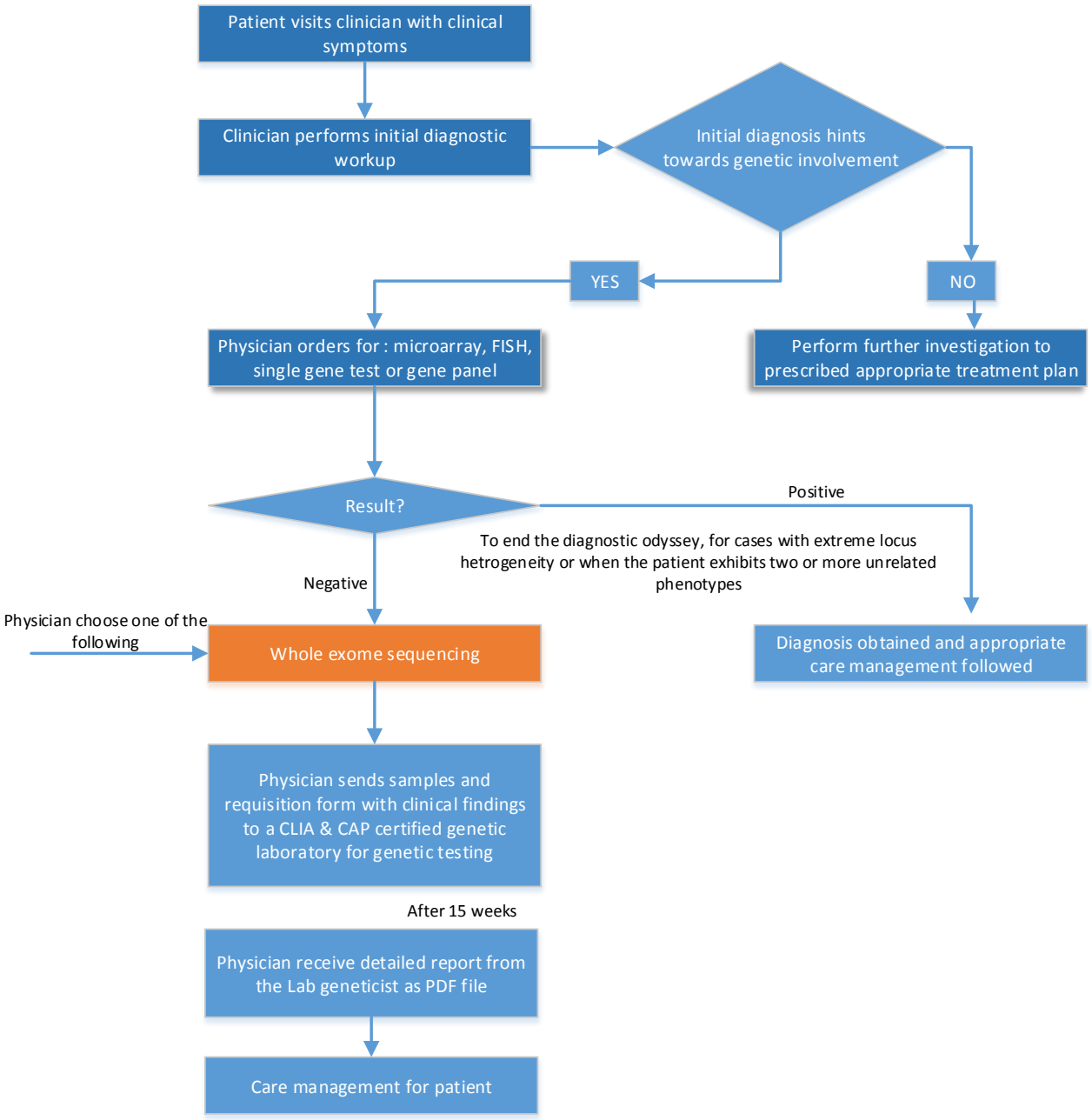


Figure 2.3: Exome sequencing workflow

2.6 Important Things to Be Considered

Kwong *et al*. (2015) mentioned all the important things to be considered before sequencing

Cost- The cost of implementation of equipment set up, routine sequencing costs for reagents and post processing bioinformatics costs is an obvious, but significant factor. This price can be measured in cost per sequencing run, cost per organism genome sequenced, or cost per megabase of output data.

Sequencing capacity- Some available technologies allow sequencing a handful of bacterial genomes in a few hours, while others have capacity to sequence 50–100 bacterial genomes in a single run that may take between 1 and 3 days.

Adaptability- Adaptability of the sequencing platform to upgrades and changing sequencing practices is another factor, with sequencing technology rapidly evolving.

Data quality- The quality of a sequence result can be reported using a score to indicate the quality and accuracy of each nucleotide base call.

Usability- Although Linux-based tools will continue to predominate due to the ability and ease in customizing analyses, tools that can be operated through a GUI may be preferred by those unfamiliar with bioinformatics.

Automation- Another key advantage of Linux-based tools, although often requiring some initial work to establish, is a ‘pipeline’ for specific types of analyses. These pipelines enable ‘batching’ or sequential running of a number of processes on multiple genome.

Speed- In a clinical setting, the ability to obtain a result quickly is often a priority over correcting minor inaccuracies in single nucleotide polymorphism (SNP) calls that do not change the overall result. Bioinformatic tools that are able to analyze multiple samples together and utilize the processing power and resources of modern computers to split large complex processes into smaller processes running in parallel hyperthreading.

Accuracy and detail- the accuracy of the analysis is important for clinical microbiology, particularly for organism identification, typing, and resistance detection.

Documentation and support- An advantage of commercial software is the availability of user manuals and professional support for troubleshooting. In contrast, while there is usually some documentation for use and limited support available from open-source software developers, many issues require local computing expertise for implementation and troubleshooting.

MATERIALS AND METHODS

3. Materials and Methods

3.1. Whole Exome Sequencing data

Files for all R1 and R2 parameter of the sample was available. The short sequence had been aligned against Hg19 built from NCBI and the alignment report using alignment score >=90 and gaps <= 5

Clinical features of proband and the following features were observed:

i. Poor cry at the time of birth with low set ears
ii. Microcephaly and Clinodactyly
iii. Dilated bowels loops
iv. Abnormal facies and hypertoxia in blood
v. Arythrogryposis and Mild retardation
vi. Craniosynostosis and clubbed foot
vii. Urinary bladder and pelvis were partially clustered

The annotated data can be of different file extension on the basis of tool used. In case of NGS strand the extension can be vCards (.vcf files) or Microsoft Excel Worksheet (.xlsx) or Microsoft Excel Binary Worksheet (.xlsb). File extensions are always convertible depending on the program/tool used for filtering. We used LibreOffice in Ubuntu 16.04 LTS on iMac system. There was total 506616 SNPs found in the whole exome data of proband as shown in figure 3.1

3.2. Filtering of the aligned data

Filtering was processed through different parameters, namely

i. Filter by Read metrices
ii. Filter using illumine QC plots
iii. Filter by duplicate reads
iv. Filter by region list.

Read metrices filtering will remove the reads with average base quality below 20. The permitted score for duplicate read filtering was permitted up to 2. This filter

helps in determining specific number of duplicates to be present. Only reads with better quality score were passed.

3.3. Nucleotide polymorphism preprocessing

Preprocessing was done involving namely three basic steps:

i. Split read realignment
ii. Local realignment
iii. Base quality recalibration

3.4. Analysis of the filtered list

For the SNP detection the input values were selected with confidence cut off score 50 and targeted region padding 100.

3.5. Finding the damaging variants

In order to understand annotation database for non-synonymous SNPs, dbNSFP annotation database was employed (Liu *et* al, 2016) table 1. For finding the pathogenic variants the non-synonymous and missense SNPs were targeted. Steps –

i) Select the variant allele list and the parameters for filtering based on dbSNP V2 b3 annotation.
ii) Selecting the dbSNP filters to be applied.
iii) Check for known variants.
iv) Check for structural variant detection.
v) Import utilities
vi) Filter by zygosity- All the heterozygous annotated SNP data was filtered first.
vii) Out of total SNP in proband data, 438 SNPs are found to be heterozygous pathogenic variants.
viii) Filter by deleterious nature of predication – the sources of functional prediction and the corresponding prediction categories were selected i.e. SIFT- damaging; LRT- deleterious; mutation taster- disease_causing_automatic and disease_causing.

ix) Filter by conservation status of amino acid- The following predication categories were selected for filtering the amino acid Polyphen2_HDIV- probably_damaging and possibly_damaging; Polyphen2_HVAR- probably_damaging and possibly_damaging; Phylo_P- conserved; GERP++_RS- constrained.

x) Filter by variant allele frequency in the population- the population(1000Gpl_AF) and the allelic frequency cutoff (<0.1) was selected.

xi) For checking the known variant, the variant allele list (SNP multi sample report) containing known or identified variants and was annotated with SNP region list/ read list. Read quality less than 10 were ignored.

3.6. Post filter processes

After applying all the filters there were around 20-30 variants every filter cycle, which linked to the phenotypes. Rs ID, gene and possible expression of the variants in different variant databases (dbVar, NCBI) was checked. If the Rs ID, gene or expression of the variant matched the phenotype of the proband then the variant was noted. If there was no match in the database for the variants, it indicated the *de novo* variants. Our focus was only heterozygous variant. So, we had to repeat the filtering process again and again to remove the false negative and find the exact targeted exome variant.

Table 3.1 Showing the Field Interpretation

Sr. no.	Fields	Comments
1	chr	Chromosome number
2	ref	Reference nucleotide allele
3	cds_strand	coding sequence (CDS) strand (+ or -)
4	SLR_test_statistic	SLR test statistic for testing natural selection on codons. A negative value indicates negative selection, and a positive value indicates positive selection. Larger magnitude of the value suggests stronger evidence
5	fold_degenerate	degenerate type (0, 2 or 3)

6	Ancestral_allele	Ancestral allele (based on 1000 genomes reference data). The following comes from its original README file: ACTG - high-confidence call, ancestral state supported by the other two sequences actg - low-confidence call, ancestral state supported by one sequence only N - failure, the ancestral state is not supported by any other sequence - - the extant species contains an insertion at this position. -no coverage in the alignment
7	SIFT_score	SIFT score, If a score is smaller than 0.05 the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)"
8	Polyphen2_HDIV_score	Polyphen2 score based on HumDiv, i.e. hdiv_prob. The score ranges from 0 to 1, and the corresponding prediction is "probably damaging" if it is in [0.957,1]; "possibly damaging" if it is in [0.453,0.956]; "benign" if it is in [0,0.452]. Score cutoff for binary classification is 0.5, i.e. the prediction is "neutral" if the score is smaller than 0.5 and "deleterious" if the score is larger than 0.5. Multiple entries separated by ";".
9	MutationTaster_pred	Mutation Taster prediction, "A" ("disease_causing_automatic"), "D" ("disease_causing"), "N" ("polymorphism") or "P" polymorphism_automatic")
10	PhyloP_score	PhyloP score, the larger the score, the more conserved the site.
11	mg29way_pi	The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes.
12	hg18_pos	Physical position on the chromosome as to hg19 (hg = the names/versions of human genome references as used by UCSC browser.
13	cds_strand	Coding sequence (CDS) strand (+ or -)

RESULTS

4. Result

4.1 Identification of heterozygous SNPs

All the heterozygous variant was filtered as shown in the figure 4.1

The proband data was annotated again for gene name followed by Rs ID, as shown in figure 4.2

Gene was named by cross checking the gene id with NCBI database and Expression Atlas (EMBL-EBI) the resultant expression can be checked and the tissue where the genes are expressing itself. Expression Atlas provides a well explained diagram to elaborate the expression of genes in different tissue. It is a resource to query gene and protein expression data across species and biological conditions and to visualize down-stream analysis results to explore co-expression. It contains thousands of selected microarray and RNA-indexing data, which are manually annotated with curated and autologous words, have been checked for high quality and re-analyzed using standardized approaches. Some of the variants may show novel characteristic and sometimes the expression and phenotype are not found in the proband, in that case the filtering processes was repeated with different parameters.

4.2 Identification of disease causing variant

In our study we found five polymorphisms present on chromosome number 1,2,3,6 and 9 (IGSF3, CHRND, FGD5, CYP21A2 and FKTN respectively) in proband, The molecular diagnosis was reported for the proband with the genes which has the disease-causing variants. We found almost 37 different disease-causing variants in the exome sequence, of which only few are reported and found in the database. To know all the affected variants *de novo* gene prediction has to be applied to the study. Because many novel variants are also found which is not reported previously in the databases. By using all the tools and technique mentioned in the paper confirms the presence of disease related to variants with the diseased phenotype. The list of variants in a particular gene with the diseases associated are mentioned in the table 4.1.

Table 4.1 List of disease causing variants found

S.No	Chr	Gene	rs ID	Tissue expression	Pathway involved	Disease associated
1	1	IGSF3	rs76417519	Broad expression in esophagus (RPKM 8.2), skin (RPKM 7.8) and 19 other tissues.	Associated with bilateral nasolacrimal duct obstruction.	LCDD, Congenital nasolacrimal duct obstruction.
2	2	CHRND	rs2245601	Prostate, esophagus, testis.	Acetylcholine binding receptor undergoes an extensive conformation change that affects all subunits and leads to opening of an ion-conducting channel across the plasma membrane.	Multiple pterygium syndrome, congenital myasthenic syndrome slow-channel type (SCCMS)
3	3	FGD5	rs777253442	Lung, spleen, placenta.	Activates CDC42, a member of the Ras-like family of Rho- and Rac proteins, by exchanging bound GDP for free GTP. Mediates VEGF-induced CDC42 activation. May regulate proangiogenic action of VEGF in vascular endothelial cells, including network formation, directional movement and proliferation. May play a role in regulating the actin cytoskeleton and cell shape.	Proangiogenic Action of Vascular Endothelial Growth Factor in Human Vascular Endothelial Cells.
4	6	CYP21A2	rs149143853	Adrenal	Encodes cytochrome P450 enzymes. Metabolism and synthesis of cholesterol, steroids and other lipids. protein localizes to the endoplasmic reticulum and hydroxylates steroids at the 21 position	Congenital adrenal hyperplasia (CAH)
5	9	FKTN		Ubiquitous expression in thyroid, brain, spleen.	The protein is involved in the pathway protein glycosylation, which is part of Protein modification.	Walker–Warburg syndrome (WWS),cerebroocular dysplasia-muscular dystrophy syndrome (COD-MD).

4.3 Expression of genes in different tissue with gene tree

The tissue which are found to be malformed in the proband are related with the expression of the gene containing pathogenic variants, the expressions are checked in expression atlas:

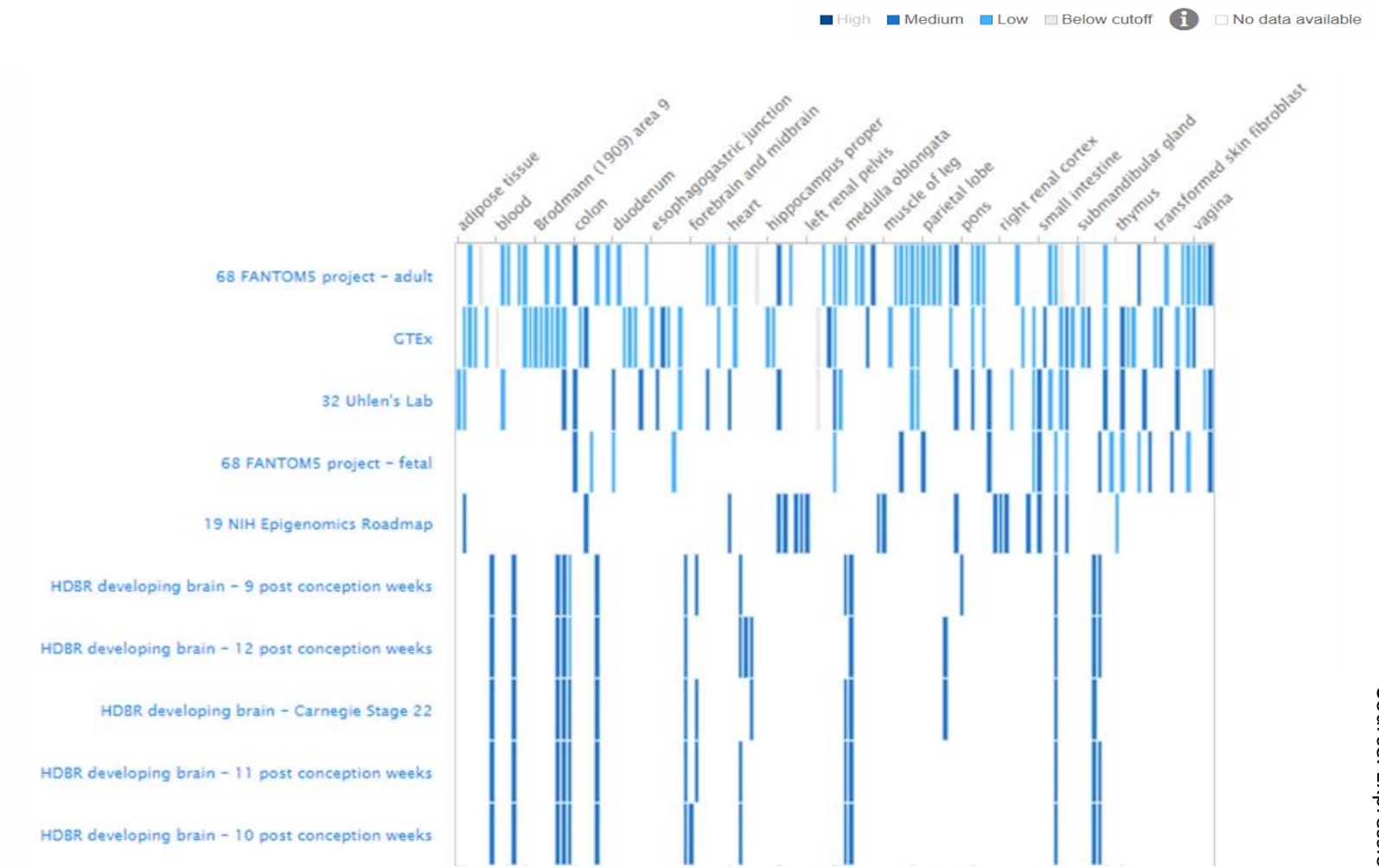
FIGURES

Figure 4.1: Snapshot of all non-synonymous heterozygous variants

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Chromosom	Start	End	Reference	Variant Allel	Variant Type	Sample Cou	Experiment /	Supporting F	Variant Reac	Total Reads	Percent Stra	Zygosity (BD3T_1)	Supporting F	Variant Reac	Total Reads
2	chr1	1178854	1178854	T	C	Substitution	4	0.33333334	0	0	57	0	Heterozygous	0	0	62
3	chr1	7797503	7797503	C	G	Substitution	4	0.33333334	0	0	61	0	Heterozygous	0	0	73
4	chr1	9324213	9324213	C	T	Substitution	4	0.33333334	0	0	46	0	Heterozygous	0	0	38
5	chr1	14107135	14107135	C	G	Substitution	4	0.33333334	0	0	88	0	Heterozygous	0	0	83
6	chr1	24664239	24664239	C	G	Substitution	4	0.33333334	0	0	45	0	Heterozygous	0	0	56
7	chr1	33160644	33160644	T	G	Substitution	4	0.33333334	0	0	28	0	Heterozygous	0	0	23
8	chr1	36235452	36235452	G	A	Substitution	4	0.33333334	0	0	61	0	Heterozygous	0	0	97
9	chr1	43223489	43223489	C	T	Substitution	4	0.33333334	0	0	31	0	Heterozygous	0	0	76
10	chr1	52828384	52828384	A	G	Substitution	4	0.33333334	0	0	39	0	Heterozygous	0	0	41
11	chr1	52850232	52850232	T	C	Substitution	4	0.33333334	0	0	109	0	Heterozygous	0	0	97
12	chr1	54433542	54433542	C	T	Substitution	4	0.33333334	0	0	67	0	Heterozygous	0	0	99
13	chr1	78511960	78511960	C	T	Substitution	6	0.5	55.172413	55.172413	29	8.620691	Heterozygous	45.16129	45.16129	31
14	chr1	87045696	87045696	G	C	Substitution	4	0.33333334	0	0	27	0	Heterozygous	0	0	38
15	chr1	117142641	117142641	G	A	Substitution	6	0.5	53.22581	53.22581	62	24.04692	Heterozygous	47.5	48.75	80
16	chr1	117142700	117142700	C	A	Substitution	6	0.5	50	50	86	13.953489	Heterozygous	38.095238	38.095238	84
17	chr1	117142736	117142736	A	G	Substitution	6	0.5	47.761192	47.761192	67	4.011196	Heterozygous	33.333336	33.333336	54
18	chr1	117146504	117146504	G	A	Substitution	6	0.5	52.808987	52.808987	89	1.6734362	Heterozygous	36.25	36.25	80
19	chr1	117146563	117146563	G	A	Substitution	6	0.5	36.363636	36.363636	44	18.181818	Heterozygous	25.641027	25.641027	39
20	chr1	117150617	117150617	A	G	Substitution	6	0.5	25	25	52	19.23077	Heterozygous	22.222223	22.222223	90
21	chr1	117156459	117156459	C	T	Substitution	6	0.5	46.341465	46.341465	41	11.553272	Heterozygous	47.457626	47.457626	59
22	chr1	118657952	118657952	C	T	Substitution	6	0.5	45.27027	45.27027	148	6.6559076	Heterozygous	58	58	150
23	chr1	145281408	145281408	C	T	Substitution	6	0.5	11.335012	11.335012	397	39.272324	Heterozygous	12.747875	12.747875	353
24	chr1	159890163	159890163	T	A	Substitution	4	0.33333334	4.5454545	4.5454545	22	18.181818	Heterozygous	0	0	19
25	chr1	171162583	171162583	T	C	Substitution	4	0.33333334	0	0	94	0	Heterozygous	0	1.8867924	53
26	chr1	171168515	171168515	A	G	Substitution	4	0.33333334	2.0408163	2.0408163	49	142.85715	Heterozygous	0	0	56
27	chr1	172356367	172356367	G	A	Substitution	4	0.33333334	0	0	62	0	Heterozygous	0	0	36
28	chr1	183495812	183495812	A	G	Substitution	4	0.33333334	0	0	79	0	Heterozygous	0	0	105
29	chr1	186063487	186063487	T	C	Substitution	4	0.33333334	0	0	90	0	Heterozygous	0	0	66
30	chr1	204587193	204587193	G	A	Substitution	4	0.33333334	0	0	28	0	Heterozygous	0	0	28
31	chr1	209791929	209791929	G	T	Substitution	4	0.33333334	0	0	51	0	Heterozygous	0	0	59
32	chr1	216052233	216052233	G	T	Substitution	4	0.33333334	0	0	37	0	Heterozygous	0	0	69
33	chr1	226019646	226019646	A	G	Substitution	6	0.5	42.105263	42.105263	57	9.21053	Heterozygous	31.578945	31.578945	57

	A	B	C	D	E	F	G	H	I	J	K	L	
1	Chrom	Start	BaseChange	VarType	Zygosity	ReadDepth	Gene	Rs-ID	InsideGeneInfo	cDNA_Change	Protein_Change	Codon_Change	VarC
2	chr1	13273	G>C	SNP	HET	54	All_NonCodingGenes - DDX11L	None	Exonic-ncGene	NA	NA	NA	NA
3	chr1	15211	T>G	SNP	HOM	3	All_NonCodingGenes - WASH7I	rs78601809	Intronic-NonCodin	NA	NA	NA	NA
4	chr1	15274	A>T	SNP	HOM	2	All_NonCodingGenes - WASH7I	rs201931625	Intronic-NonCodin	NA	NA	NA	NA
5	chr1	16288	C>G	SNP	HOM	2	All_NonCodingGenes - WASH7I	rs200736374	Intronic-NonCodin	NA	NA	NA	NA
6	chr1	16298	C>T	SNP	HOM	2	All_NonCodingGenes - WASH7I	rs200451305	Intronic-NonCodin	NA	NA	NA	NA
7	chr1	66162	A>T	SNP	HET	3	Intergenic	rs62639105	Intergenic	NA	NA	NA	NA
8	chr1	69270	A>G	SNP	HOM	3	OR4F5 (+)	rs201219564	Exonic-CDS	c.180A>G	p.60	>G	Silent
9	chr1	69511	A>G	SNP	HOM	42	OR4F5 (+)	rs75062661	Exonic-CDS	c.421A>G	p.141	>G	Silent
10	chr1	131552	G>T	SNP	HET	3	All_NonCodingGenes - RP11-3	rs371080566	Exonic-ncGene	NA	NA	NA	NA
11	chr1	133483	G>T	SNP	HET	7	All_NonCodingGenes - RP11-3	rs369820305	Exonic-ncGene	NA	NA	NA	NA
12	chr1	137923	C>T	SNP	HOM	3	AL627309.1 (-)	None	Exonic-3UTR	NA	NA	NA	NA
13	chr1	537496	C>A	SNP	HOM	2	All_NonCodingGenes - RP5-857	None	Intronic-NonCodin	NA	NA	NA	NA
14	chr1	726859	A>G	SNP	HOM	3	All_NonCodingGenes - RP11-2(	rs139100483	Intronic-NonCodin	NA	NA	NA	NA
15	chr1	726887	A>G	SNP	HET	3	All_NonCodingGenes - RP11-2(	None	Intronic-NonCodin	NA	NA	NA	NA
16	chr1	726901	G>C	SNP	HET	3	All_NonCodingGenes - RP11-2(	None	Intronic-NonCodin	NA	NA	NA	NA
17	chr1	726924	T>C	SNP	HOM	3	All_NonCodingGenes - RP11-2(	None	Intronic-NonCodin	NA	NA	NA	NA
18	chr1	726942	A>T	SNP	HOM	3	All_NonCodingGenes - RP11-2(	None	Intronic-NonCodin	NA	NA	NA	NA
19	chr1	752566	G>A	SNP	HOM	2	All_NonCodingGenes - RP11-2(	rs3094315	Intronic-NonCodin	NA	NA	NA	NA
20	chr1	752721	A>G	SNP	HET	17	All_NonCodingGenes - RP11-2(	rs3131972	Intronic-NonCodin	NA	NA	NA	NA
21	chr1	752894	T>C	SNP	HOM	7	All_NonCodingGenes - RP11-2(	rs3131971	Exonic-ncGene	NA	NA	NA	NA
22	chr1	753425	T>C	SNP	HOM	1	All_NonCodingGenes - FAM87B	rs3131970	Exonic-ncGene	NA	NA	NA	NA
23	chr1	755274	C>T	SNP	HOM	1	Intergenic	rs78408995	Intergenic	NA	NA	NA	NA
24	chr1	761732	C>T	SNP	HOM	1	All_NonCodingGenes - LINC001	rs2286139	Exonic-ncGene	NA	NA	NA	NA
25	chr1	761752	C>T	SNP	HOM	1	All_NonCodingGenes - LINC001	rs1057213	Exonic-ncGene	NA	NA	NA	NA
26	chr1	761811	G>A	SNP	HOM	2	All_NonCodingGenes - LINC001	rs55884753	Exonic-ncGene	NA	NA	NA	NA
27	chr1	761957	A>AT	INDEL	HOM	41	All_NonCodingGenes - LINC001	rs59038458	Exonic-ncGene	NA	NA	NA	NA
28	chr1	762273	G>A	SNP	HOM	119	All_NonCodingGenes - LINC001	rs3115849	Exonic-ncGene	NA	NA	NA	NA
29	chr1	762320	C>T	SNP	HET	96	All_NonCodingGenes - LINC001	rs75333668	Exonic-ncGene	NA	NA	NA	NA
30	chr1	762589	G>C	SNP	HOM	23	All_NonCodingGenes - LINC001	rs71507461	Exonic-ncGene	NA	NA	NA	NA
31	chr1	762592	C>G	SNP	HOM	20	All_NonCodingGenes - LINC001	rs71507462	Exonic-ncGene	NA	NA	NA	NA
32	chr1	762601	T>C	SNP	HOM	18	All_NonCodingGenes - LINC001	rs71507463	Exonic-ncGene	NA	NA	NA	NA
33	chr1	762632	T>A	SNP	HOM	16	All_NonCodingGenes - LINC001	rs61768173	Exonic-ncGene	NA	NA	NA	NA
34	chr1	768448	G>A	SNP	HOM	2	All_NonCodingGenes - LINC011	rs12562034	Intronic-NonCodin	NA	NA	NA	NA
35	chr1	770345	A>C	SNP	HOM	1	All_NonCodingGenes - LINC011	rs111352149	Intronic-NonCodin	NA	NA	NA	NA
36	chr1	770735	A>G	SNP	HOM	1	All_NonCodingGenes - LINC011	rs139531586	Intronic-NonCodin	NA	NA	NA	NA
37	chr1	770850	G>T	SNP	HET	4	All_NonCodingGenes - LINC011	None	Intronic-NonCodin	NA	NA	NA	NA

Figure 4.2: Snapshot of the annotated exome sequence



Source: Expression atlas

Figure 4.3: Showing the expression of *IGSF3* in tissue, the expression is high on brain during conception weeks.

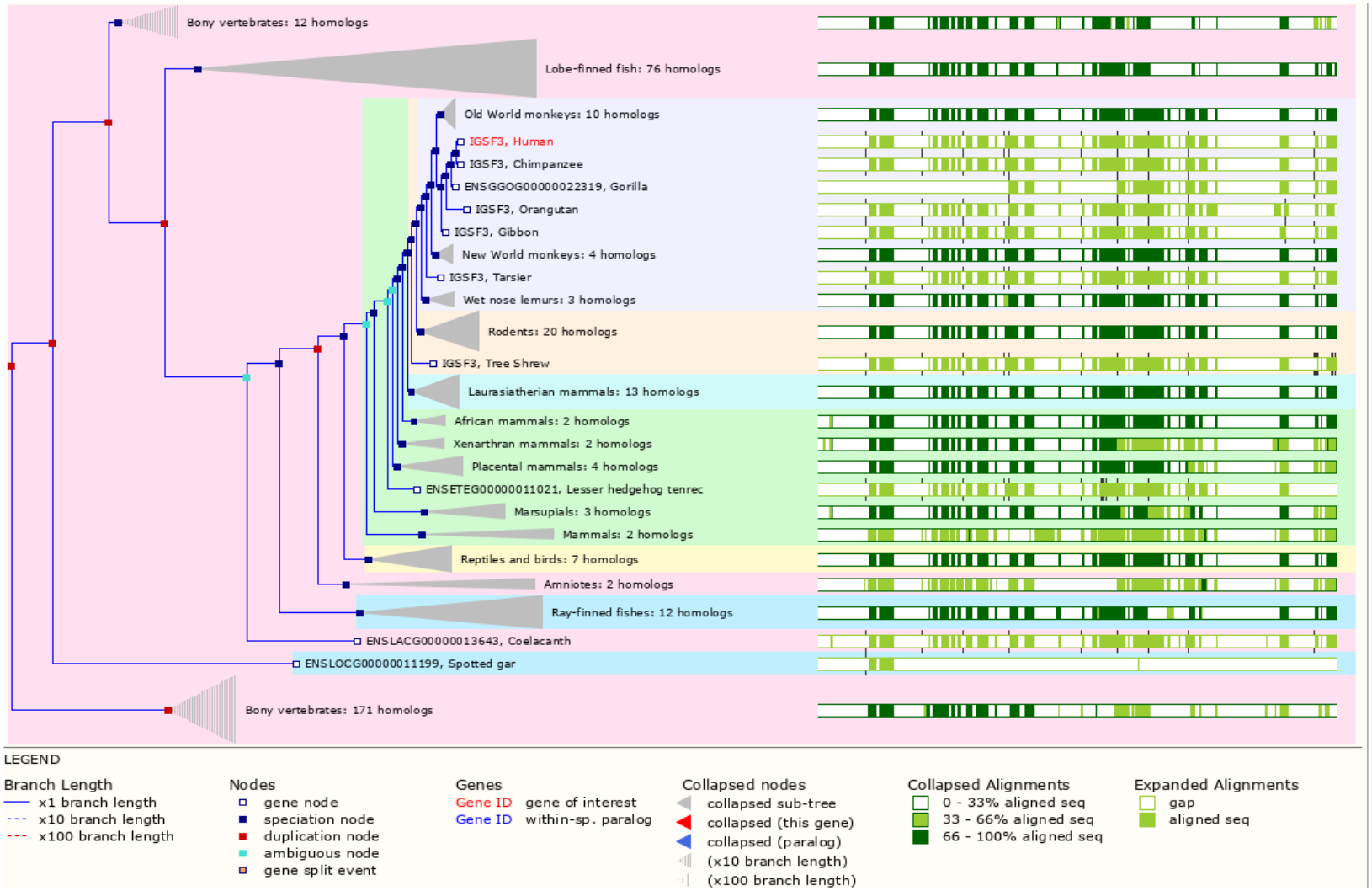


Figure 4.4: Showing the phylogenetic tree of gene *IGSF3*

Figure 4.5: Showing the expression of *CHRND*

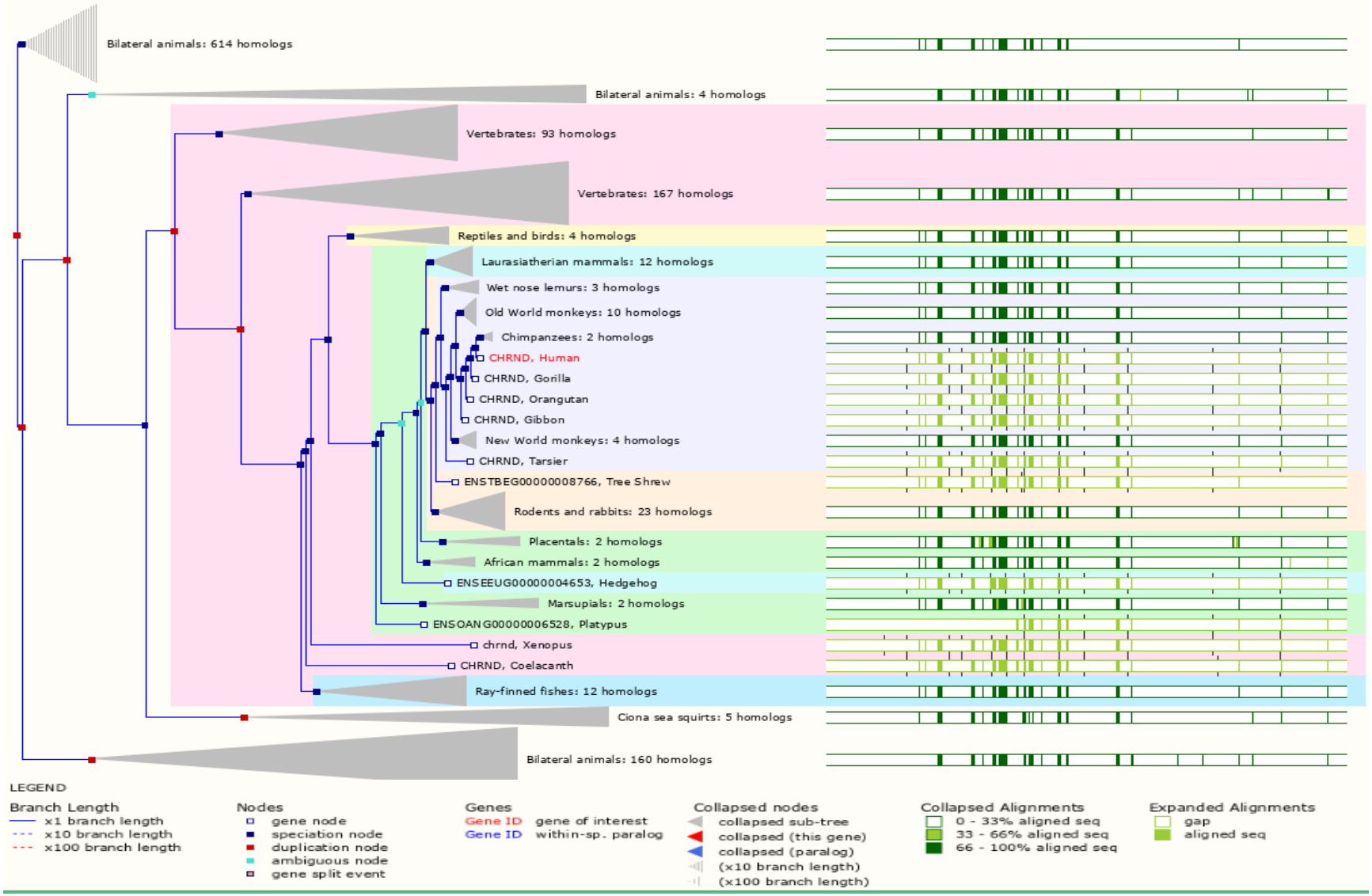


Figure 4.6: Showing the phylogenetic tree of gene

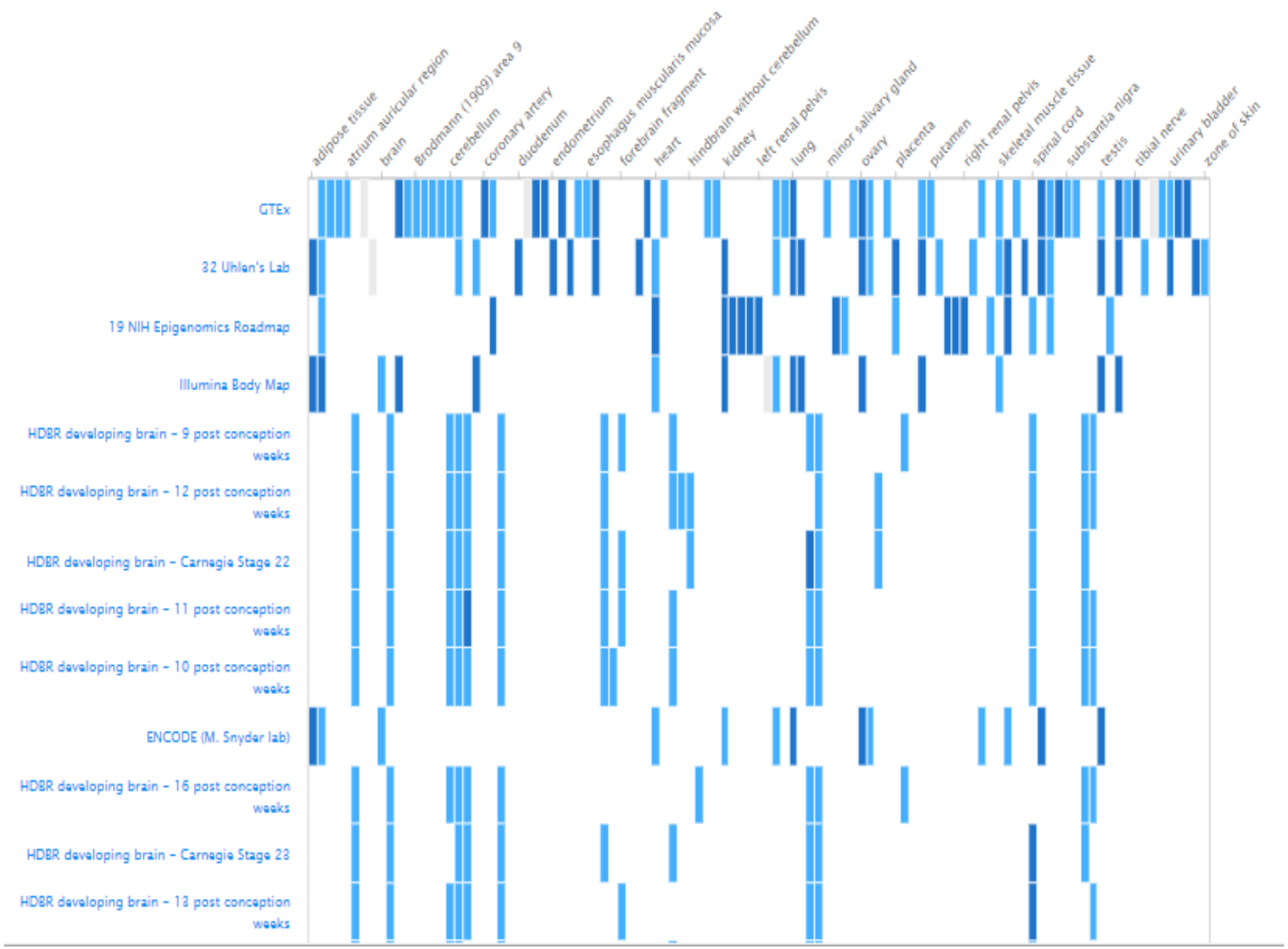


Figure 4.7: Showing the expression of *FGD5*

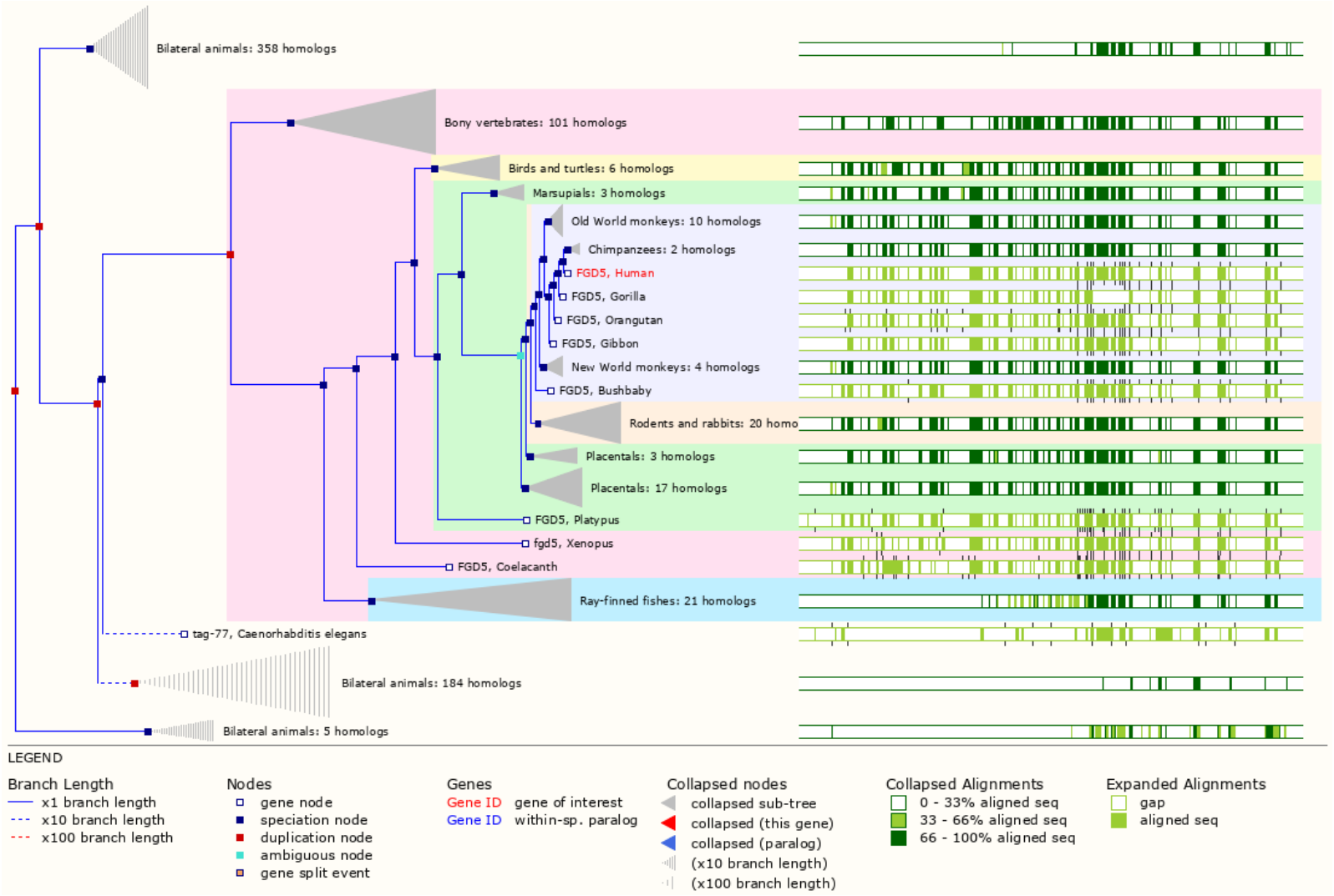


Figure 4.8: Showing the phylogenetic tree of gene *FGD5*

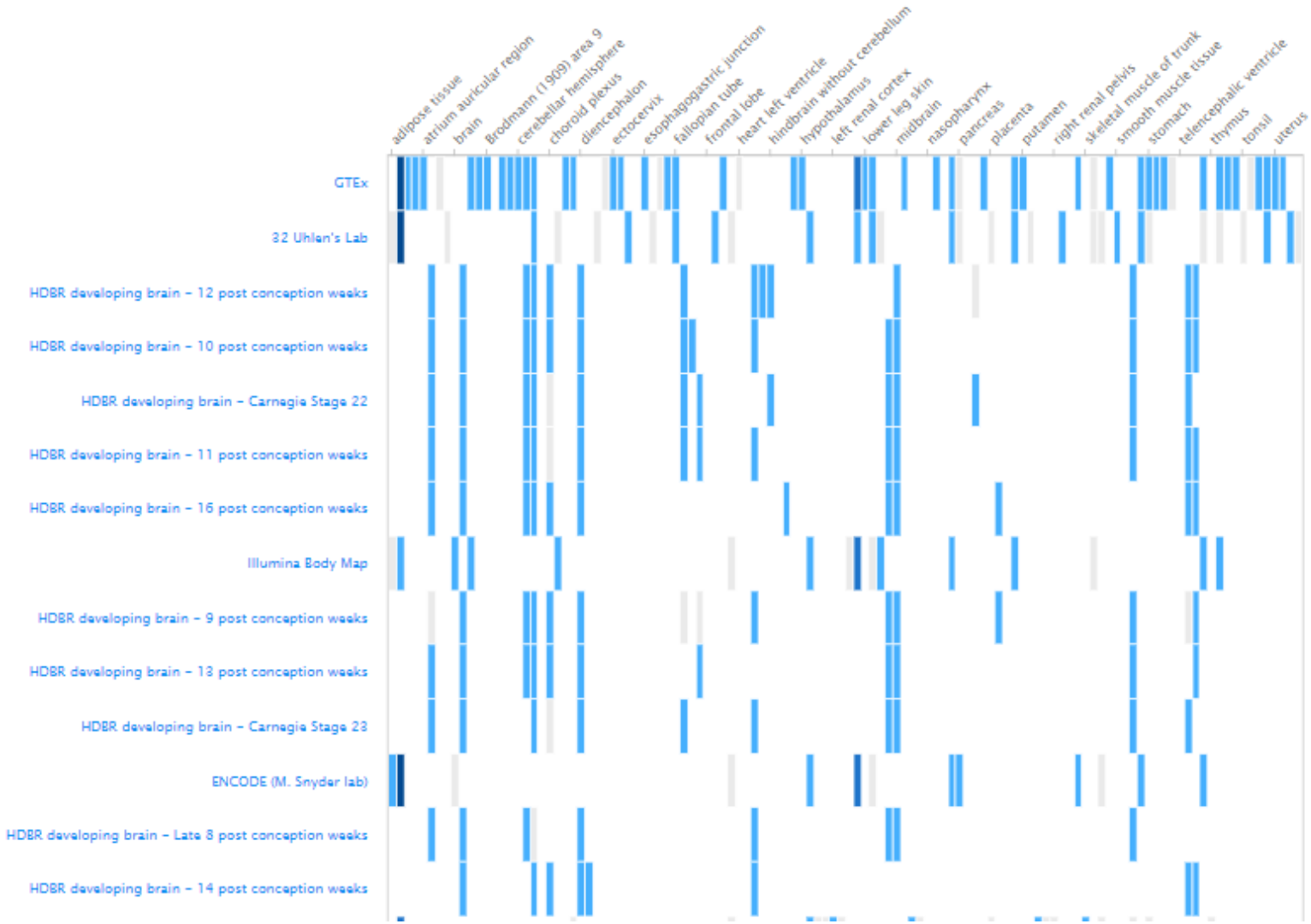


Figure 4.9: Showing the expression of *CYP21A2*

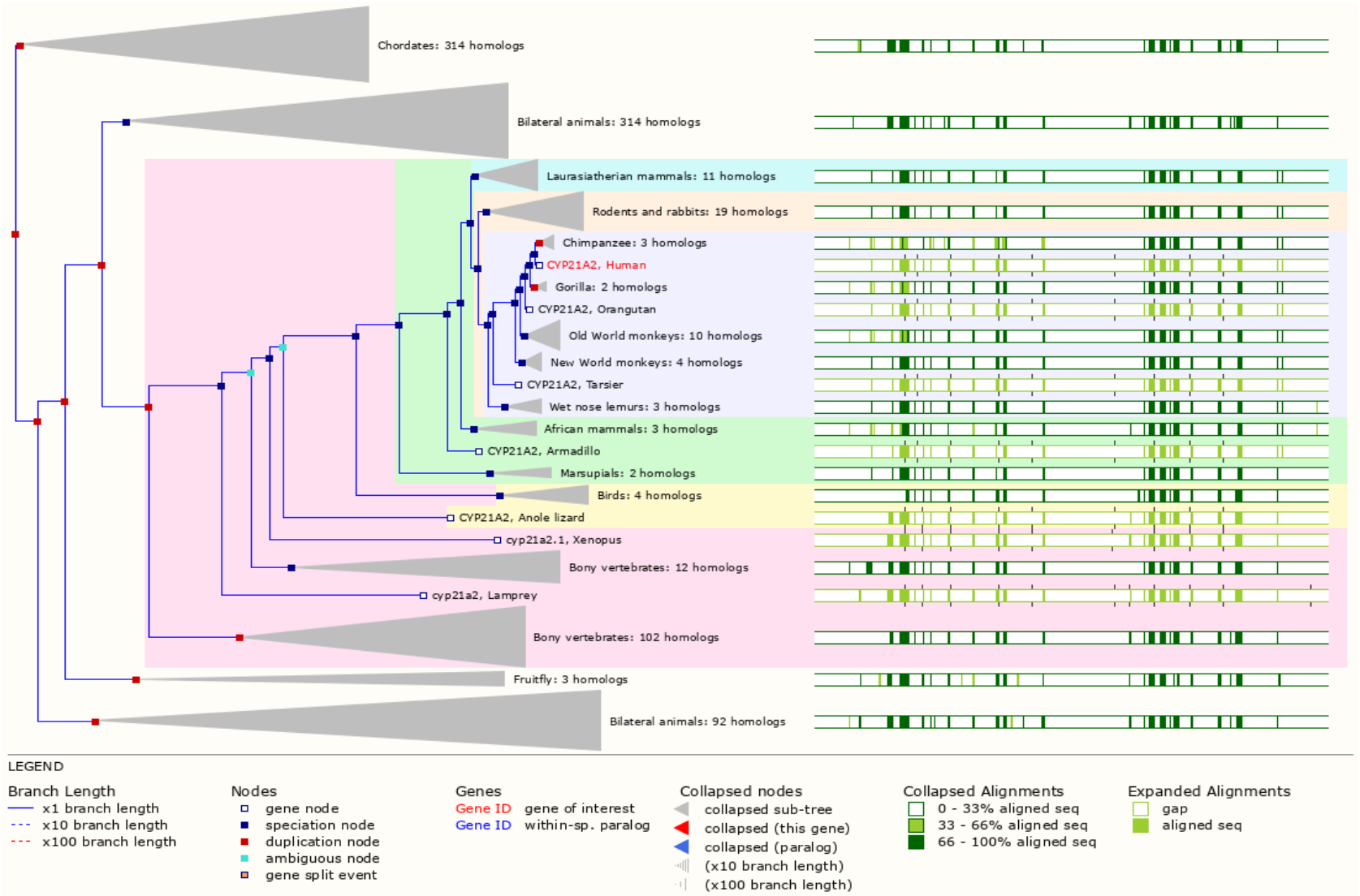


Figure 4.10: Showing the phylogenetic tree of gene

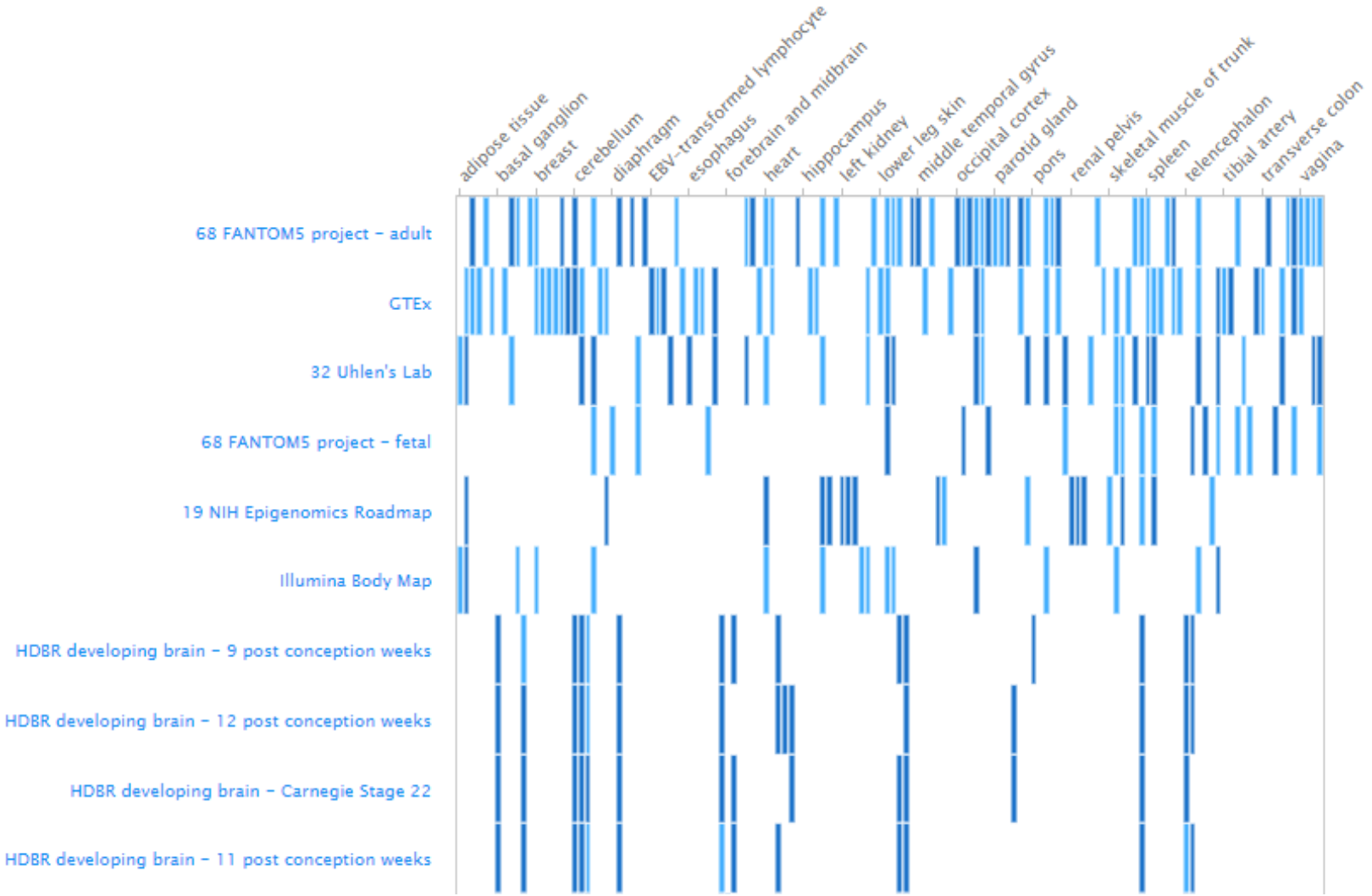


Figure 4.11: Showing the expression of *FKTN*

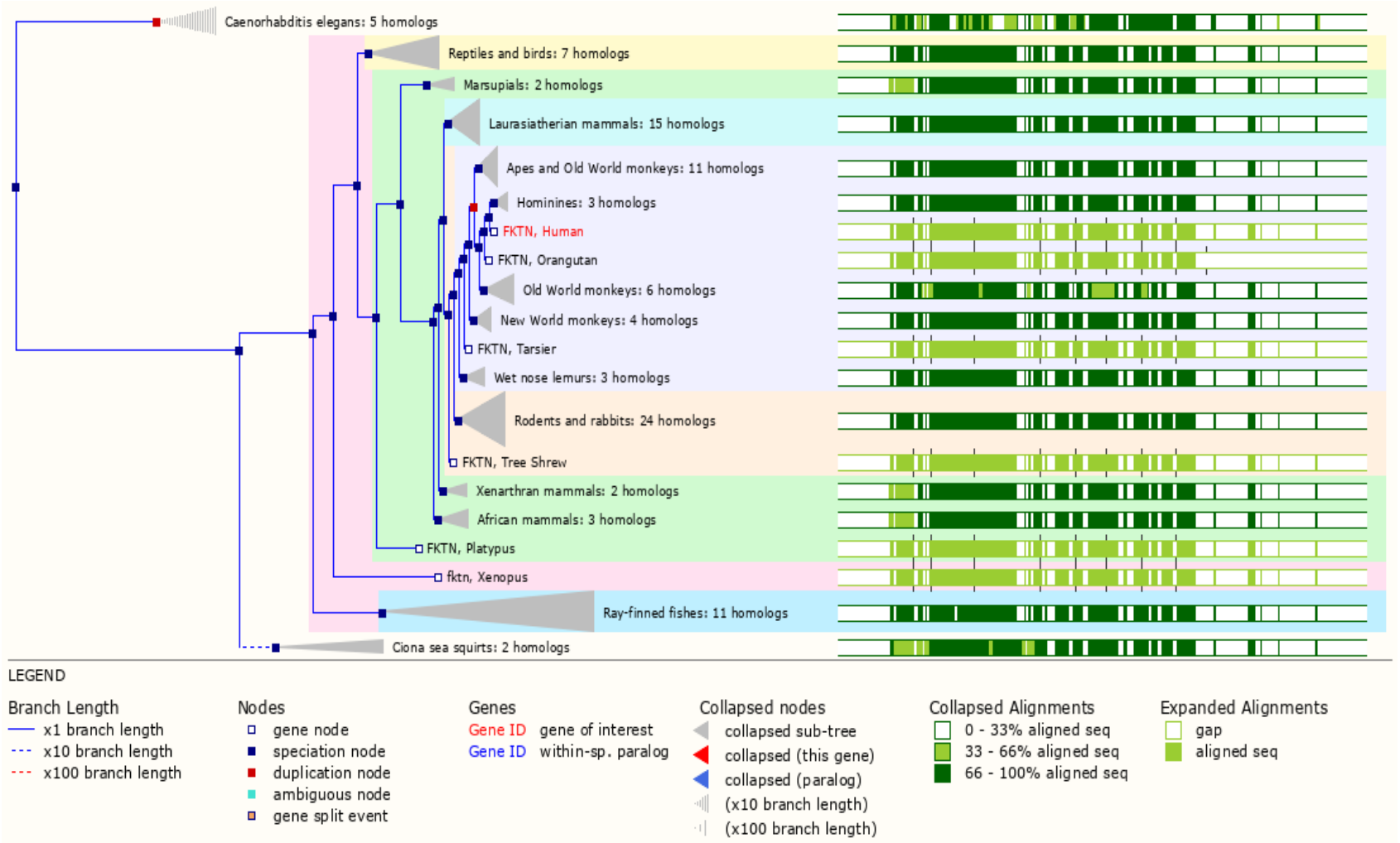


Figure 4.12: Showing the phylogenetic tree of gene *FKTN*

DISCUSSION

5. Discussion

5.1 Strategies for finding the rare variants from WES data

The origin for the congenital anomalies can be said to be multifactorial that involved identifiable environment exposures and polygenic candidate gene variation. An individual genome contains variants which may protect or increase susceptibility to environment and the multiple stress factors encountered during lifespan (Majewski *et al,* 2011).

Whole exome sequencing for the proband with a wide range of structural abnormalities diagnosed post birth, along with the parent sample was performed. The data was further subjected to alignment and variant calling. Variant calling involved the identification of the pathogenic *non-synonymous* variants using heterozygous hit strategy (Gilissen *et al.,2012). G*enes carrying the heterozygous variants were selected and filtered. Pathogenic heterozygous variants had been identified in five genes namely IGSF3, CHRND, FGD5, CYP21A2 and FKTN

5.2 Current state of WES

In the recent years, amazing progress has been made in the ability to identify the difference in the human genome faster and impartially. (Feero *et al* 2014) mentioned genotyping along with large-scale parallel recognition on genotyping data exploded, launching the era of Genome-Wide Association Studies (GWAS) and genetic testing of suspected clinical value. Large reliable evidence has been gathering for the clinical and research value of whole exome sequencing (WES) for situations like cancer, developmental delay and mendelian disorders.

Private companies like Direct-to-consumer (DTC) exome sequencing are getting developed and offering very modest price. Knome was the first company developed in 2007 to offer exome sequencing services directly to consumers, 23andMe and DNADTC are the 2012 companies which offers exome sequencing at 80X coverage and initial price

lower than $700. Helix is the most recent company developed in 2016 offering “exome plus" sequencing.

5.3 Conclusion

The clinical malformation of the proband was targeted using Whole Exome data with the reference disease and it was found that the disease-causing variants are present in many of the genes with respect to chromosome. Analysis indicates that the disease is not monogenic, and there is involvement of multiple genes which are found to be lethal and pathogenic noted by various databases. The aligned exome data contains huge number of *de-novo* pathogenic variants among which we found 5 most pathogenic non-synonymous heterozygous variants present in IGSF3, CHRND, FGD5, CYP21A2 and FKTN genes. These genes are identified to be associated with the clinical features of proband.

5.4 Future Directions

The intronic variants and transcription factors that plays a crucial role in early embryonic pathways also can be explored. For the validation of de novo variants within the proband a lot of such affected subjects could require to be recruited. As the parents of proband are phenotypically normal there must be environmental factors which might be causing congenital anomalies so there is a need for methylation analysis (Protein - DNA interactions or Methylation of cytosine (C) nucleotides in the context of a CpG dinucleotide)

SUMMARY

6. Summary

Whole Exome Sequencing (WES) is a capture-based methodology developed to sequence exome and to spot variants within the coding region of genes that have an effect on protein function. Understanding the exomes of people at single base resolution permits the identification of pathogenic variants liable for causing congenital disease. In this report, we had discussed the various steps involved in bioinformatic analysis of whole exome sequence data. The objective of the study was to identify pathogenic heterozygous variants in the proband. In silico tools like SIFT, PolyPhen and ANNOVAR had been utilized. The parents belonged to low income family sub-group. Mother was a housekeeper and was suffering from hypothyroidism and the father was a labor. The sample collection and sequencing had been done previously (Kaul S, 2017) after sequencing all the annotated non-synonymous SNPs was organized and filtered according to the pathogenicity of variants. All the variants were narrowed down to 5 pathogenic heterozygous variants. The variants were present in the genes which play important developmental roles. In this study, only exons were targeted and sequenced, the involvement of introns is not known. As a conclusion bioinformatic analysis of WES information is a decent tool for locating disease causing variants among patients with unidentified genetic cause. In future, non-coding DNA sequence could be studied to identify the genetic cause in the proband affected with multiple congenital malformations.

.

REFERENCES

References

ACMG Board of Directors. (2012). Points to consider in the clinical application of genomic sequencing. *Genetics in Medicine*, **14**(8), 759-761.

Biesecker, L. G., Shianna, K. V., and Mullikin, J. C. (2011). Exome sequencing: the expert view. *Genome biology*, **12**(9), 128.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., and Nelson-Williams, C. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, **106**(45), 19096-19101.

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., and Akdemir, Z. H. C. (2015). The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *The American Journal of Human Genetics*, **97**(2), 199-215.

Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Chen, R., Euskirchen, G., and Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature biotechnology*, **29**(10), 908.

Feero, W. G. (2014). Clinical application of whole-genome sequencing: proceed with care. *Jama*, **311**(10), 1017-1019.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T. M., and Berlin, A. M. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology*, **12**(1), R1.

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

Grada, A., and Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *The Journal of investigative dermatology*, **133**(8), e11.

Hintzsche, J. D., Robinson, W. A., and Tan, A. C. (2016). A survey of computational tools to analyze and interpret whole exome sequencing data. *International journal of genomics*, 2016.

Kaul Shweta (2017)., whole exome sequencing to understand genetic basis of unknown medical condition: a case study, M.sc Project work, Central University of Punjab.

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., and NHLBI Exome Sequencing Project. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome research*, **22**(8), 1525-1532.

Ku, C. S., Polychronakos, C., Tan, E. K., Naidoo, N., Pawitan, Y., Roukos, D. H., and Cooper, D. N. (2013). A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. *Molecular psychiatry*, **18**(2), 141.

Kwong, J. C., McCallum, N., Sintchenko, V., and Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology*, **47**(3), 199-210.

Linnarsson, S. (2010). Recent advances in DNA sequencing methods–general principles of sample preparation. *Experimental cell research*, **316**(8), 1339-1343.

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human mutation*, **32**(8), 894-899.

Maffucci, P., Filion, C. A., Boisson, B., Itan, Y., Shang, L., Casanova, J. L., and Cunningham-Rundles, C. (2016). Genetic diagnosis using whole exome sequencing in common variable immunodeficiency. *Frontiers in immunology*, 7, 220.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., and Jabado, N. (2011). What can exome sequencing do for you? *Journal of medical genetics*, **48**(9), 580-589.

SAMBROOK, H. (1989). Molecular cloning: a laboratory manual. Cold Spring Harbor, NY.

Stark, Z., Tan, T. Y., Chong, B., Brett, G. R., Yap, P., Walsh, M., and Amor, D. J. (2016). A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genetics in Medicine*, **18**(11), 1090.

Swaminathan, R., Huang, Y., Astbury, C., Fitzgerald-Butt, S., Miller, K., Cole, J., and Lin, S. (2017). Clinical exome sequencing reports: current informatics practice and future opportunities. *Journal of the American Medical Informatics Association*, **24**(6), 1184-1191.

van der Zwaag, P. A., Jongbloed, J. D., and van Tintelen, J. P. (2014). Genetic diagnosis through whole-exome sequencing. N Engl J Med, 370, 1067.

van den Berg, S., Shen, Y., Jones, S. J., & Gibson, W. T. (2014). Genetic counseling in direct-to-consumer exome sequencing: a case report. *Journal of genetic counseling*, ***23*****(5),** 742-753.

Student approval form

Name of the author	Md Momin Ali
Department	Department of Human Genetics and Molecular Medicine
Degree	M.Sc. life sciences with specialization in Human Genetics
University	Central University of Punjab
Guide	Dr. Preeti Khetarpal
Project title	Bioinformatic analysis of Whole Exome Sequence data
Year of award	2018

Agreement

1. I hereby certify that, if appropriate, I have obtained and attached here to a written permission / statement from the owner (s) of each third party copyrighted matter to be included in my dissertation/thesis, allowing distribution as specified below.

2. I hereby grant to the university and its agents the non-exclusive license to archive and make accessible, under the conditions specified below, my dissertation/thesis, in whole or in part in all forms of media, now or hereafter known. I retain all other ownership rights to the copyright of the dissertation/thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report.

Conditions:

1. Release the entire work for access worldwide	Not applicable
2. Release the entire work for Central University of Punjab only for:	After 03 years

1 year 2 years 3 years and after this time release the work for access worldwide.	
3. Release the entire work for Central University of Punjab only, while at the same time releasing the following parts of the work (e.g. because other parts relate to publications) for worldwide access: a) Bibliographic details and synopsis only b) Bibliographic details, synopsis and the following chapters only c) Preview / Table of contents / page only	Not applicable
4. View only (No downloads) (world-wide)	Not applicable

Signature of the candidate **Signature and seal of the supervisor**

Place: Bathinda

Date: May,2018

URKUND

Urkund Analysis Result

Analysed Document:	momin plagiarism.docx (D39310176)
Submitted:	5/28/2018 9:24:00 AM
Submitted By:	arora1205@gmail.com
Significance:	5 %

Sources included in the report:

https://www.ircm.qc.ca/Etudier/cours/Documents/Brais_article.pdf
http://europepmc.org/articles/PMC4127531
http://ucgd.genetics.utah.edu/wp-content/uploads/2015/11/1525.pdf
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4127531/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3727622/

Instances where selected sources appear:

13