Identification of genomic signatures in idiopathic mental retardation cases by whole exome sequencing

A Thesis Submitted to

NIRMA UNIVERSITY

In partial fulfillment of the award of the Degree of

MASTERS OF SCIENCE

IN

BIOTECHNOLOGY

By

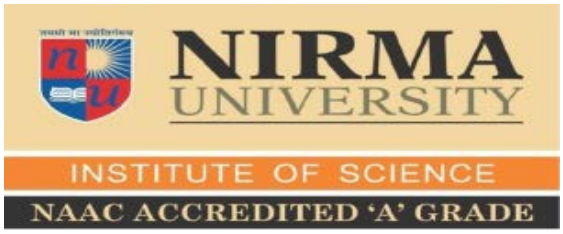
Silkey Mehta (15MBT020)

Shweta Patel (15MBT030)

Under the guidance of

Dr. Sonal Bakshi

(Asst. Professor)



ACKNOWLEDGMENT

We thank Almighty God, for it is He who began this work in us and carried it to completion. It is He who has blessed us with the people whose names we feel privileged to mention here.

This work has been possible due to motivation and support of various individuals who gave us sound advice and guidance at numerous occasions throughout our tenure as a postgraduate student at the ***Institute of Science, Nirma University, Ahmedabad****. It is with immense pleasure to offer our heartfelt thanks to our parents who have always been the rock strong support through this journey. It also gives us immense pleasure today to acknowledge all those personalities who contributed directly or indirectly to our project.*

We express our sincere gratitude to ***Prof. Sarat Dalai,*** *Director, Institute of Science, Nirma University for giving us this invaluable opportunity of carrying out our dissertation project at the Institute of Science, Nirma University. We wish to express our sincere thanks, with a deep sense of gratitude, to our respected guide* ***Dr. Sonal R. Bakshi,*** *Assistant Professor, Institute of Science, Nirma University for initiating and suggesting the theme of work, for her valuable guidance, supervision, creative suggestions, meticulous attention, sustained interest, and support she has bestowed upon us for the timely completion of this work.*

We are highly grateful to ***Gujarat State Biotechnology Mission (GSBTM), Gandhinagar*** *and* ***Dr. Subhash Soni,*** *Mission Director, GSBTM for providing adequate facilities to conduct our work. We convey our humblest way of expressing our sincere obligation to* ***Dr. Jayashankar Das,*** *Sector Specialist, GSBTM for giving us chance to undertake this work at OMICS laboratory, Gandhinagar. Without his precious support it would not be possible to conduct this dissertation project. We are immensely thankful to* ***Dr. Shivarudrappa Bhairappanavar,*** *Scientist B for his scholarly guidance. We owe our sincere thanks to* ***Mr. Inayat Shaikh,*** *Research Associate, GSBTM for providing his insight and expertise which greatly assisted this project. This dissertation work would not have been possible without his master advice, supervision, valuable guidance and generous help. We would like to extend our wholehearted thanks to* ***Ms. Garima Ayachit,*** *Research Associate, GSBTM and other staff members for helping us throughout the project.*

*We are highly obliged to **Mr Alok Patel,** Business Development Manager, Theragen Etex Bio Institute for setting up the milestone and developing our interest in Bioinformatics. Also, we would like to thank **Theragen Etex Bio Institute** for providing the whole exome data to conduct our dissertation project. Our special and deep thanks to **Mr. Praful Joshi,** Computer facility coordinator, Institute of Technology, Nirma University for his technical support in downloading the data and troubleshooting the problems.*

*We are grateful to **Dr. Shalini Rajkumar, Dr. Shriram Seshadr , Dr. Nasreen Munshi , Dr. Vijay Kothari, Dr. Amee Nair, Dr. Rajeev Tyagi, Dr. Heena Dave,** and **Dr. Pranati Sar** for helping us in one or the other ways.*

*We are deeply indebted to Ph.D students and mentors, **Fulesh Kunwar, Shikha Tiwari,** and **Nazia Saiyad** for stimulating suggestions and their encouragement which helped us to conduct the project and writing of this thesis.*

*We are thankful to **Mr. Hasit Trivedi, Mr. Sachin Prajapati, Mrs. Sweta Prajapati** and other non teaching staff members of Institute of Science, Nirma University for their appreciable constant support and kind cooperation.*

*Words are short to express our deep sense of gratitude towards our friends **Aakansha Shah, Chakshu Rathi, Kiran Lalwani, Krupa Desai, Prapti Purohi, Ruapl Jatvada** and **Shivani Sheth** for all their moral support and helping hands. Last but not the least, our thanks extends to **Aishwarya Joshi, Shivani Patel** and **GSBTM friends** who were the most cheerful and amazing people we found.*

Index

List of figures

List of tables

I. ABBREVIATIONS

ACMG- American College of Medical Genetics and Genomics

BAM- Binary Alignment Mapping

chr- Chromosome

CLNSIG- Clinical Significance

CNV- Copy Number Variation

HGP- Human Genome Project

ID- Intellectual Disability

IGV- Integrative Genomic Viewer

IMR- Idiopathic Mental Retardation

INDEL- Insertion and Deletions

IQ- Intelligence Quotient

MR- Mental Retardation

NGS- Next Generation Sequencing

OMIM- Online Mendelian Inheritance in Man

rsID- Reference Sequence Cluster ID

SAM- Sequence Alignment Mapping

SNP- Single Nucleotide Polymorphism

VCF- Variant Call Format

WES- Whole Exome Sequencing

WGS-Whole Genome Sequencing

hg19- Human Genome 19

II. ABSTRACT

Intellectual disability is the most common developmental disorder characterized by a congenital limitation in intellectual functioning and adaptive behavior. Idiopathic mental retardation refers to the individuals who show no evidence of gross chromosomal defects or single-gene anomalies. Genetic factors play a major part in intellectual disability (ID), so far very few studies attempted to reveal genetic profile of the subjects, limited to few genes. The purpose of this study focused on the identification of genomic signatures in idiopathic mental retardation cases through whole exome sequencing (WES) studies.

In the current study, exome sequencing of two cases revealed 29 pathogenic variations occurring in 29 genes, of which 4 variations associated with IMR were found in Case1 and 8 variations were found in Case 2. The comparison of two cases resulted in 4 common variations in both the probands. These 8 variations c.85C>T, c.185T>C, c.399C>T, c.1250A>G, c.2353C>T, c.247C>A, c.1237A>T and, c.693C>T in *DPYD, HEXB, MAG, SYCE, C5/F42, PTH, CLPB* and, *SPATA* genes respectively, were associated with intellectual disability and developmental delay. Remaining variations were associated with muscular, vision, hyperthyroidism, jaundice, etc. problems which were not shown in any of the proband of the two cases, hence suggests that probands may develop them in future.

Therefore, exome sequencing could be a potential tool for unfolding causative genetic variations in non-syndromic intellectual disability.

III. Introduction And Review of literature

Intelligence is not a thorough characteristic but is determined on the basis of different numbers of more, or less specific skills. Each individual generally develops a similar level of skills but some large discrepancies can be seen in mentally retarded person. Severe impairments can be seen in such people who may be on a particular area (e.g. language) or on area of higher skills (e.g. visual-spatial tasks). The assessment of intelligence level has been done on the basis of available information which includes clinical findings, behaviour of adaptation and performance of psychometric tests (ICD-10, 1996).

Mental retardation is hence considered as a condition of incomplete and arrested mind development due to impairment of skills expressed during the developmental period of overall intelligence level which include social, language, cognitive and motor abilities. Mental retardation can be the result of genetic disorders. The diagnostic category should not be based on a single impairment area or skill, but should take global areas into consideration (ICD-10, 1996). The earliest case of mental disorder was reported in 1552 B.C. Prior to it, people having unusual social skills and mild IQ probably receive less or no special attention and care. Jean-Marc Itard created the first systematic intervention program for intellectually disable people in the late 18^{th} century. Psychological tests for exploring the intelligence level were developed in 20^{th} century. This helped to increase the number of cases relating mental retardation. Many international, national and local policies have now been made for securing their civil rights and giving them quality services for bringing them parallel to the normal people (Maulik et al., 2010).

3.1 Classification of mental retardation: Mental retardation can be generally diagnosed on the presence of IQ (Intelligence Quotient) score. IQ is a number used to measure the intelligence in relation to person's age group. Average IQ of a person ranges between 90 and 100. Over 120 are considered to be superior. Statistics shows that 68% of the population has an IQ level ranging between 85 and 115. Score less than 70 indicate some kind of disability present in that person. An IQ test generally consists of tasks for measuring the measure of intelligence which include analytical thinking, short-term memory, spatial recognition and mathematical ability. The result of an individual is

compared with the people of same age group. Mental retardation is usually categorized under four levels: mild, moderate, severe and profound (Table 1).

1. **Mild retardation:** People with IQ level ranging between 55 to 69 are considered as mildly retarded. Such people are generally slow in walking and talking. Prevalence of mildly retarded people in ID population is 85%. These people can learn reading and writing and can live independently. They can have social and job skills too.
2. **Moderate retardation:** People with IQ level ranging between 40 to 54 falls under this category. They show noticeable delay in speech and motor skill development. They can learn basic communication and safety skills but cannot gain beneficial academic skills. They face problems in learning, reading and performing mathematical tasks. Prevalence of moderately retarded people amongst ID population is 10%. Such people cannot live or travel alone.
3. **Severe retardation:** People with IQ level ranging between 20 to 39 are considered as severely retarded. Their conditions are usually diagnosed at the time of birth or soon after it. They have little or no communication ability and shows motor developmental delays. They may learn basic walking and talking skills with training and age. Prevalence of such people is 5% amongst ID population. They need a good supervision and protected environment for living.
4. **Profound retardation:** Very few people have their IQ levels below 20. Their conditions can also be diagnosed at birth and requires a good nursing care. These children show developmental delays in all aspects. They are unable to take care of themselves so requires a continuous supervision and a complete support for daily living. 1% of ID population shows profound retardation (Kaufman et al, 1988).

Class	IQ
Borderline intellectual functioning	70 to 80
Mild retardation	55 to 69
Moderate retardation	40 to 54
Severe retardation	20 to 39
Profound retardation	Below 20

Table 1: Classification of mental retardation (AAMR, 2002)

3.2 Prevalence of different disabilities in India: According to the census in 2011, mental retardation cases are less than the other cases of disability (Figure 1 & Figure 2).

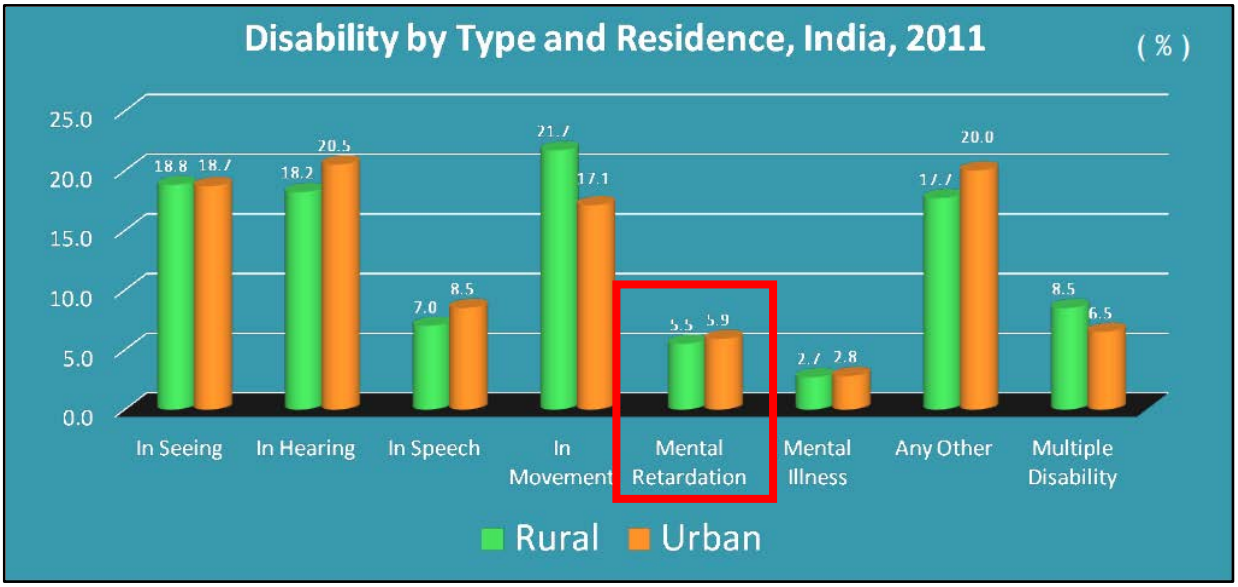


Figure 1: Disabled population by type and residence (%) India **(**Census of India, 2011**)**

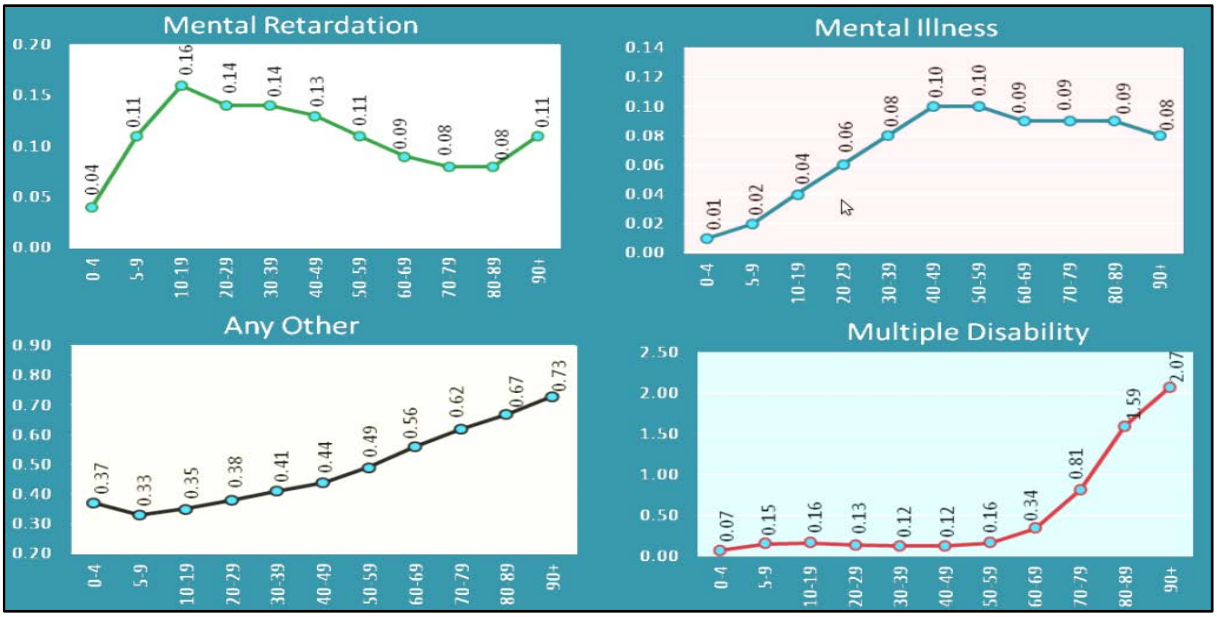


Figure 2: Percentage of disabled population in India **(**Census of India, 2011**)**

3.3 Causes of mental retardation: Mental retardation can be the result of interaction of many factors. But in approx. 75% of cases, the exact causative reasons are unknown this type of cases are known as idiopathic mentally retarded cases. Some of the causes for ID

include defects in the genes or chromosomal aberrations which took place at the time of foetus development or while developing in the womb and some of the environmental influences.

1. **Genetic conditions:** Mental retardation can be the result of single-gene disorders. Some of these disorders are associated with atypical or dysmorphic morphological characteristics. There is a great chance of children to suffer from this condition when one or both parents are carrier. Such defects are the result of inherited genetic material passing from parent to child. Down syndrome, which is caused due to presence of an extra chromosome and fragile X syndrome, are the major examples in this type of causing condition.

   About one fourth of people suffering from mental retardation have detectable chromosomal abnormality. Other genetic changes may be the result of small insertions or deletions and duplications. Such changes are rarely reported; hence their phenotype is still undetermined. Some of the abnormalities are inherited and some are *de novo.* Other genetic causes include Phelan-McDermid syndrome caused due to 22q13del. Abnormalities in X and Y chromosome are seen rarest. 48, XXXX and 49, XXXX syndrome are seen in girls, while 47, XYY, 49, XYYYY are seen in boys worldwide.
2. **Prenatal problems:** If the development of foetus inside the mother does not take place properly, the child may suffer from mental illness after the birth. Congenital infections, prolonged maternal fever and untreated maternal phenylketonuria are the main causes of prenatal problems.
3. **Perinatal problems:** Problems during late pregnancy and during delivery can cause changes which may lead to mental retardation.
4. **Exposure to toxins and certain kind of diseases:** Measles, whooping cough or meningitis can cause mental disability if specific care is not taken. Also exposure to certain poisons containing lead and mercury has severe effect on the mental status.
5. **Malnutrition:** It is a common cause of reduction in the intelligence level in most of the parts of the world.
6. **Few unknown factors:** The aetiology of mental retardation is heterogeneous and unfortunately, in about one-half of such cases the causative agents are still elusive

in nature. Such kind of disorders is called as “Idiopathic Disorders” which still needs to be discovered (Armatas et al, 2009).

3.4 Idiopathic Mental Retardation (IMR): Mental impairment is a major significant factor in the field of psychiatric, mental and education as the finding of its cause and dealing with the disability to get some meaningful conclusion is very difficult. Identifying the cause provides information of the genetic risks in present and also in future. With the discovery of novel genetic techniques, many new mysterious chromosomal aberrations have been identified in last few years (Bodensteiner et al., 2006). Idiopathic mental retardation is the condition in which individuals show no evidence of gross chromosomal changes or single-gene mutations and has low IQ. Recent MMR studies show the high chromosomal abnormalities rates which increase the possibility of proportion of people suffering from idiopathic mental retardation (Ahuja et al., 1995). About 3% of the population with IQ <70 have unknown factors affecting behind it. Chromosomal aberrations identified by cytogenetic analysis accounts for 40% severe MR cases and 10 to 20% of mild MR cases. Subtle chromosomal abnormalities that are not listed by routine testing can be one of the responsible reasons for some of the patients whose cause for MR is unknown. Recent development in molecular genetics technology allows the detection of extremely little portions of chromosomal changes and hence helps to detect the “mysterious” chromosomal defects. Chromosomal aberrations are, therefore, the common cause for many anomalies syndromes which include developmental dysmorphism and growth delays. Many new high resolution techniques including whole genome and whole-exome techniques have been developed to improve the better understanding of causative agents of Idiopathic mental retardation. Such patients so show learning disabilities and have borderline IQ scores ranging between 71-85 (Maulik et al., 2010).

3.5 Signs and symptoms shown by IMR patients: This typically involves cognitive and adaptive skills delays. A developmental delay depends on the level of MR and its aetiology. Some of the common features are as follow:-

- **Language delay:** This is one of the first signs shown by MR patients including expressive (speech) and receptive (understanding) delays. Sometimes parents may interpret it as the child is born deaf.

- **Fine adaptive/motor delay:** Delays in common activities like self-feeding, toileting, and playing are generally reported for mentally retarded children. Drooling is the sign of oral-motor incoordination.
- **Cognitive delay:** MR children face problems with memory and logical reasoning. They may have difficulties in even preacademic learning sessions.
- **Social delay:** MR children may show lack of interest in playing and making friends. Their play may reflect their developmental level.
- **Gross motor:** Delays in gross motor development may result in cerebral palsy and MR.
- **Behavioural changes:** Apart from the actual age a person is having, MR patient may be hyperactive and may have difficult temperament and sleep disorders. They may be aggressive in nature.
- **Neurological and physical abnormalities:** Seizure disorders, macrocephaly, postnatal growth retardation, and congenital anomalies are high in MR children (Zeldin et al., 2016).

Physical examination:

- **Head circumference:** All growth parameters are measured. Cognitive deficits may be correlated with microcephaly and hydrocephalus indicates macrocephaly. Macrocephaly can be associated with some metabolic inborn errors.
- **Height:** Genetic disorder may lead to short stature and hypothyroidism. Tall stature is associated with fragile X syndrome and other overgrowth syndromes.
- **Neurologic:** Assessment of head growth, muscle tone, deep tendon reflexes, strength and coordination, and some abnormal movements are done here.
- **Sensory:** Suspected children are assessed for vision and hearing (Galasso et al, 2010)

3.6 Case study in India: 1-3% of the population is suffering from mental retardation in India. Standard guidelines for evaluating an individual have been demonstrated by American College of Medical Genetics, American Academy of Neurology and The Practice Committee of the Child Neurology Society and American Academy of Paediatrics (Aggarwal et al, 2012).In the study of 338 patients (110 females and 228 males) having

mental retardation, 85 people were suffering from idiopathic mental retardation condition. The average age of the studied group was 4.75 yr. The average IQ score was 39.8. 38.4% patients were having mild MR, 17% were having moderate MR, 36.4% were having severe and 7.7% were having profound MR.

These 85 patients had no definite diagnosis for the cause. 58 patients showed dysmorphic features, 22 showed various other malformations, 29 showed positive familial history and 35 showed neurological findings (Table 2).

Aetiological categories	No. of patients (%)
Chromosomal syndromes	112 (33.1)
Non chromosomal syndromes	32 (9.5)
Neurometabolic disorders	34 (10.1)
Central nervous system structural defects	25 (7.4)
Cerebral palsy	43 (12.7)
Environmental insult	7 (2.0)
Idiopathic mental retardation	85 (25.2)
Total	338

Table 2: Various aetiological categories and number of suffering patients (Aggarwal et al., 2012)

3.7 Criteria for diagnosing IMR cases: Selection of an appropriate criterion is important for better understanding of the disease.

- **Clinical approach:** Clinical approach provides particular diagnosis and requires a comprehensive and exhaustive study of the patient. Firstly, a pedigree study of 3 generations is done which include detailed study of pre-, peri- and postnatal history. This is because the child may be suffering this condition from the time of birth. A deep developmental history and behaviour is mandatory. Medical records can help in validation in diagnosis of changes. An accurate brain MRI is sufficient at times in suspecting the presence of common disorders.

  Physical examination of the patient is important for a “gestaltic” analysis. But, the phenotypic variations among the patients with microdeletions and microduplications

may vary depending upon the genomic alterations of different sizes. In many cases of MR, unique and unspecific changes are present in the patient. For such patients, minor anomalies of the face, hands, skin, and genitalia should be reported. Also head size abnormality and growth parameters should be carefully examined.

- **Genetic approach:** This approach is generally undertaken for identifying unknown cause of mental retardation, such as in the case of idiopathic conditions. Fluorescent *in situ* hybridization (FISH) is used for detecting small chromosomal abnormalities and confirms microdeletions/microduplications syndromes. This study also determines the subtelomeric regions of all the chromosomes. The integrated study of karyotyping and subtelomeric regions helped in the detection of chromosomal anomalies in 5-10% of idiopathic patients. Some novel approaches are discovered in this area which includes chromosomal microarray technique or comparative genomic hybridization technique (array-CGH). This new technique has detected the submicroscopic chromosomal aberrations with the detection rate of 5-20%. But the problem encountered in this technique is that, array-CGH is not able to detect balance rearrangements of chromosome which are believed to occur in 0.75% of overall MR patients (Cinzia et al., 2010).

Nowadays researchers are focusing on the more advance method of genetic identification using Next Generation Sequencing (NGS) and its associated tools. This is the most recent technology developed to reduce the time and to get the accurate reason of the idiopathic condition. This technique identifies the single nucleotide polymorphisms (SNP), copy number variations (CNV) and other structural variations (SV) playing role in patient's mental retardation.

3.8 Role of sequencing techniques in identification of genetic alterations

Human Genome Project: The Human Genome Project (HGP) was the collaborative research program aiming for mapping of complete genome and understanding all the genes of human beings. It is a 13 years long project initiating in 1990s. The International Human Genome Sequencing Consortium published its first draft on human genome in journal *Nature* in February 2001. But the full sequence was published in April 2003 which revealed that there are probably about 20,500 genes present in humans (Table3).The main

goal of HGP was to supply the powerful tools to the researchers for understanding of genetic factors responsible for human disease, its diagnosis, treatment and preventions. Today, HGP has already discovered 1,800 diseased genes (https://www.genome.gov).

Release name	Date of release	Equivale UCSC version
GRCh38	Dec 2013	hg38
GRCh37	Feb 2009	hg19
NCBI Build 36.1	March 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2004	hg16

Table 3: Reference Human Genome (www.genome.ucsc.edu, 2016)

3.9 Sequencing Technology: The technology came a long way back since the time of two-dimensional chromatography in the 1970s. Scientist learnt reliable sequencing method after the invent of Sanger chain termination method in 1977. The first automated based sequencing method came after a decade later by Applied Bio systems. Capillary electrophoresis became workhorses for Celera-led Human genome Project. While these "first generation sequencing" were considered high throughput for their time, there was a great need of automated systems for sequencing which could decrease the cost and reduce the duration of sequencing period. Due to the brisk decline of cost per base and duration, Next Generation Sequencing came into role (https://www.illumina.com).

Over the past few years, the strategies for DNA sequencing have emerged on mainly four levels. First, with the introduction of Human Genome Project, the reduction in the cost of conventional sequencing was achieved. Second, the utility of short-read sequencing potentially strengthened the availability of whole genome assemblies for *Homo sapiens* and all major model organisms. Third, a variety of molecular methods were developed to gain the high-throughput results of sequencing. And fourth, the development in

technologies across different fields which includes microscopy, polymerase engineering, computation and data storage made DNA sequencing more practical to use and analyse (Shendure et al., 2008).

3.9.1 Sanger sequencing method: The first genome to be sequenced was achieved by Fred Sanger and his group in 1977 on phiX174 bacteriophage. The "Classical" Sanger sequencing method was based on base-specific chain termination carried out in four specific reaction tubes (labelled as A, G, T and C). A specific 2', 3'-dideoxyneucleotide triphosphate (ddNTP) was added in all the four reaction tubes containing 2'-deoxynucleotide triphosphate (dNTP). This method of using ddNTPs in sequencing reactions proved to be novel and effective as compared to "plus and minus method" developed by the same group. The newly synthesised DNA extension terminates every time its corresponding ddNTP incorporates (Figure 3).

The second novelty used by them was the use of radioactive phosphorous or sulphur isotopes which incorporates to the newly synthesised strand through a labelled precursor and hence make it easy to analyse by radiography (Men et al., 2008).

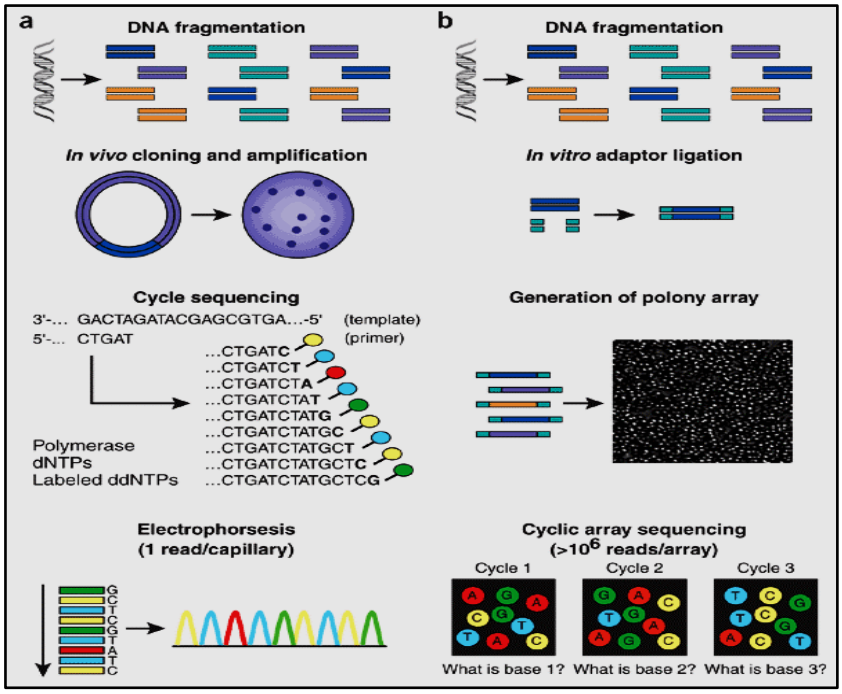


Figure 3: Sanger chain termination method protocol (Shendure et al, 2008)

3.9.2 Next Generation Sequencing: It is a non-Sanger based DNA sequencing technology. It is also known as high throughput sequencing methods. It covers the broader range of mutations than Sanger. Sanger sequencing was mostly restricted to finding of substitution, small insertions and deletions while the other are mostly discovered by NGS methods. Such mutations include submicroscopic chromosomal copy number variations (CNV), example microdeletions, single nucleotide polymorphism (SNP), structural variations (SV), etc. The increased sensitivity of NGS allows the detection of mosaic mutations. Global merits of next generation sequencing include *(i) in vitro* construction of genomic sequence library. *(ii)* Array-based sequencing which increases the effective size due to which hundreds of millions of sequencing reads can be potentially gained along with images of reasonably sized surface areas (Shendure et al., 2008)**.** These are different modern sequencing technologies: -

1. Illumina (Solexa) sequencing
2. Roche 454 sequencing
3. Ion torrent: Proton/PGM sequencing
4. SOLiD sequencing (Shendure et al, 2008)

- **Next generation sequencing applications:** Because of production of large numbers of low-cost reads, the usefulness of NGS is diverse. These include discovery of variants by resequencing the interested targeted regions or whole genome, *de novo* collection of bacterial and lower eukaryotic genomes, classification of species and discovery of genes by metagenomics studies. Different platforms of NGS have different applications of its own too. For example, Illumina/Solexa is mostly used for discovery of variants by resequencing human genomes. Illumina produces high-quality bases per run hence are best for resequencing (Metzker et al., 2010).

  NGS technologies have wide applications in functional genomics research. These include gene expression profiling, annotation of genes, small RNA identification and profiling and aberrant transcription detection. The major functional genomics application is associated with determination of DNA sequences and identification of genes responsible for epigenetic modification of histones and DNA. Determination

of epigenetic modification is also done by Sanger methods but NGS improved the process by high throughputing the Sanger sequencing results with increased depth of coverage and resolution. To date, these technologies have been widely used in a numerous way, such as whole genome sequencing, targeted sequencing, whole exome sequencing and gene expression profiling which helped in the identification of disease related mutations and factors (Morozova et al., 2008).

3.9.2.1 Whole genome sequencing: With the growing need of information of DNA by scientists and researchers, there was a rapid shift from conventional Sanger sequencing method to automated analysis by Next Generation Sequencing methods. This high throughput method generates data of whole genome which can be used by scientists and researchers for further resequencing and comparing data with previous cases already reported. Whole genome sequencing gives the most comprehensive information about person's genetic variations. WGS include determination of complete DNA sequence of an organism's genome done at a single time. This helps to understand the expression of genes regulating the particular environment. WGS also helps to find the correlation between genome information causing development of cancer, disease susceptibility and metabolism of drug. The genetic information identifies the mutation driving cancer progression and tracking disease outbreaks. WGS is commonly associated with human genome sequencing, but the flexibility of NGS technology makes it easy to sequence any species which can include agriculturally associated livestock, plants and microbes which are responsible of causing diseases (https://www.illumina.com).

Whole genome sequencing helps in the detection of single nucleotide variants, INDELs, copy number variants and large structural variants. Genome sequencing projects is mainly subdivided into two broad groups: *(i) de novo* high quality genome sequence assembly which can be further used as reference for variety of species and *(ii)* resequencing, for mapping of individual sequence variations, when reference is already available. General workflow includes four steps: *(i)* collection and extraction of DNA, *(ii)* sequence library preparation, *(iii)* sequencing, and *(iv)* analyzing data bioinformatically. The processed data is further interpretated and additional analysis is done if needed (Pabinger et al, 2014).WGS also facilitates the identification of virulence factors of the pathogens and

identifies the path of transmission of disease within a population. Molecular tools provide the robust and higher-resolution processes for identification, comparison and classification of pathogenic organisms. Characterization of pathogens can help in further studies of disease and its causing factors (Gilchrist et al., 2015).

Limitations of WGS: Even though the production capabilities are highly increased, the construction of accurate genome assemblies and annotating them correctly still remains a challenge. This is more challenging when the genome has higher repeats and duplicated contents. WGS typically produces shorter sequence along with higher error rates from comparatively short insert libraries. Another limitation which should be taken into consideration is the presence of contamination while sequencing process. De novo sequence assemblies can be considered good source for the discovery of mutations like insertions and polymorphism, but it requires a particular scrutiny and supported validations for enriching the contamination artifacts. This comparison of required and not-required sequence assembly is the most problematic task. The production of large number of reads results in the increase of cost. Due to sequencing of whole genome, they may produce information which a patient doesn't even need when focusing on a single disease and its causing factors (Alkan et al., 2010).

3.9.2.2 Whole Exome Sequencing: Despite of improvement in the sequencing technology, the depth which is needed to identify the variants affecting the phenotypic expression is comparatively expensive than whole exome sequencing. Exomes may also target functional nonprotein coding genes i.e. mRNA, long intrinsic noncoding RNA, etc. (Figure 4). This cost-effective tool is used for dissecting the genetic basis responsible for diseases and helps in discovery of *de novo* causative genes. This allows exploring the rare alleles taking role in heritability of complex diseases and facilitates the diagnosis of personalized disease-risk profiling. WES can be used to identify pathogenic/susceptible genes responsible for human diseases like cancer, mental retardation, diabetes, etc. (Warr et al., 2017).

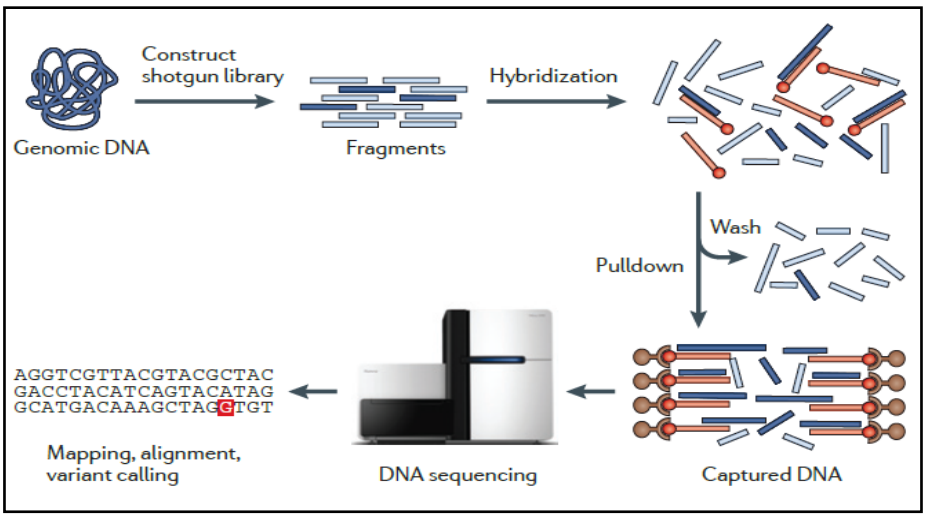


Figure 4: Whole Exome Sequencing Technology (Bamshad et al., 2011)

Two main exome capturing techniques include solution-based and array-based techniques:

- In solution-based, whole-exome sequencing technologies, DNA samples are first fragmented and then probes of biotinylated oligonucleotides are used for selectively hybridizing the target genome regions. Magnetic streptavidin beads bind to the biotinylated probe and the nontargeted regions of the genome is washed away. Enrichment of DNA sample from the target region is done by amplifying it using polymerase chain reaction (PCR). The sample is then sequenced and preceded for bioinformatics analysis (Amanda et al., 2017) (Figure 5)

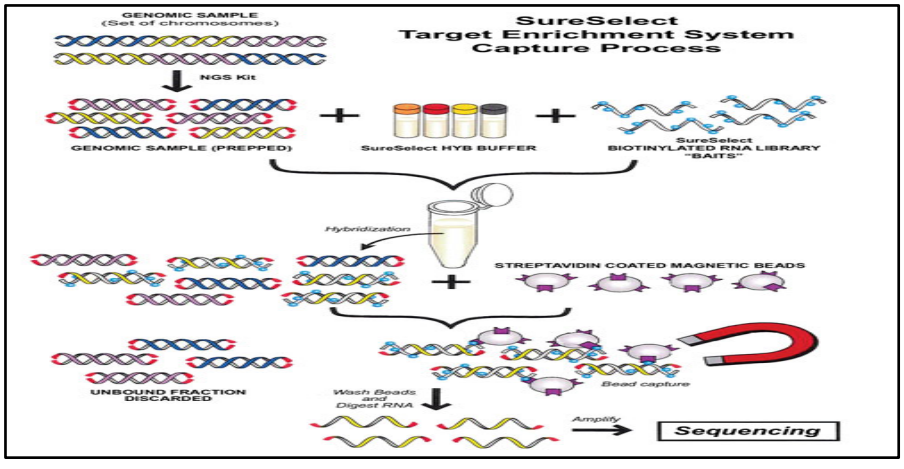


Figure 5: Solution-based method (http://www.bgi.com/)

- In array-based, whole-exome sequencing technology, the probes are bound to a high-density microarray and the rest of the steps are almost same (Figure 6). This method was first used for capturing exomes but later was taken over by solution-based methods due its more efficiency and less requirement of input DNA. However, some studies suggested that array-based methods are better than solution-based when the GC content is less. Single-nucleotide polymorphism (SNP) identification gained by this method is more specific to the targeted region. Hence, array-based methods have been successfully used to find rare and common variants

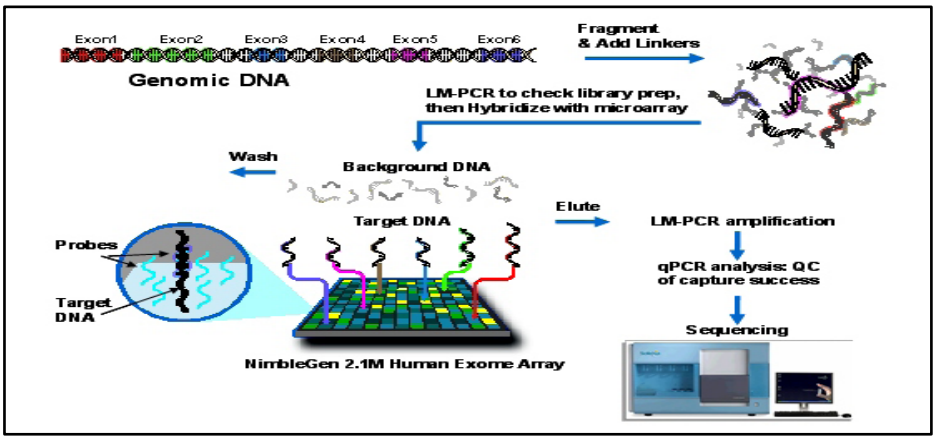


Figure 6: Array-based method (www.nimblegen.com)

- **Advances of WES over WGS**: Although WES covers less coverage than WGS; it has many benefits which invite researchers to use it for identifying mutated genes. Primarily, the first benefit lies in the decreased time and money. WES samples majorly sequence higher depths of 100X to 30Xand focuses on only ~2% of the whole genome. This is gained by pulling down or enriching at the step of DNA where RNA baits are used for hybridizing at the protein-coding genome region, making it separate from the non-coding protein regions. Only ~5-6 GB sequence needed for 100X exome sample rather than ~90Gb required for WGS. This lowers down the data storage space and cost and makes it easy for analysis. This increases the accuracy of variant discovery and interpretation (http://allseq.com/kb/wgsvswes/) (Table 4).

Properties	Whole genome sequencing	Whole exome sequencing
Target size (bp)	$3x10^9$	$5x10^7$
Sample/sequencing run	6	120
Depth of coverage	X30	X100
Data analysis time (hr.)/sample	>48	4

Table 4: WGS vs WES (* Using Illumina HiSeq 2000)

3.10 Contribution of WES technique for identification of genetic alterations: The beginning of 2009 brought the advent of whole-exome sequencing which contributed in the discovery of germline and *de novo* mutations for rare Mendelian disorders. It has also succeeded in discovering somatic mutations for variety of diseases such as cancer studies. The rare SNPs associated with complex diseases can be anticipated by sequencing studies. A new gene, named *TBK1*, was identified when exome sequence was analysed for 2,869 amyotrophic lateral sclerosis cases and 6,405 controls. For age-related mascular degenration, novel missense SNP in *UBE3D* was found. SNP in *LDLR* and *APOA5* were discovered in pateints suffering from myocardial infarction (Ku et al., 2016)

The second most common cancer found in men is the prostate cancer. It causes over 2,50,000 deaths per year. To find the genetic alterations responsible for this cancer, scientists sequenced the exomes of 112 patients with prostate tumor, keeping normal tissue pair as a control. They were able to identify mutation in multiple genes, including *MED12*, *FOXA1*and *SPOP*. A mean coverage depth to be achieved was 118X per sample. Somatic copy-number alterations were also detected by analysing both tumor containing DNA and normal sample's DNA. 5,764 somatic copy-number mutations were gained of which 997 were frameshift mutations of the mismatch repair genes *MSH6*. Along with this, some silent and non-sense mutations were also seen (Barbieri et al., 2012)**.**

Transitional cell carcinoma (TCC) is also listed in the category of the most common cancer prevailing worldwide. To get the better understanding of the etiology of bladder cancer, whole-exome sequence analysis of 99 patients were taken into consideration. This analysis resulted in the identification of 37 significantly mutated genes out of which 7 were well-

known bladder cancer genes recognised. This include *TP53*, *HRAS*, *FGFR3*, *PIK3CA*, *RB1*, *KRAS*, and *TSC1*. In addition to 7 well-known genes, 13 new mutated genes were also discovered. These contained a greater number of nonsynonymous mutations (De Ligt et al., 2012).

Exome sequencing are now yielding promising results for several intellectual disabilities. A good success rate of ~25% was reported. When exome sequence analysis of 2,000 patients were performed, it was found that rate of 'neurological-related conditions' which inlude speech delay, developmental delay, intellectual disability, and autism spectrum disorder was higher (~27%) than 'non-neurological conditions' (~20%) (Ku et al., 2016).

It is believed that *de novo* point mutations are the major responsible cause of intellectual disability. Intellectual disability can be caused by mutation in more than 1000 genes. Evaluation of 100 patients (53 females and 47 males) with unexplained intellectual disability was performed. Mutations in about 400 genes have been found out of which most of these changes have a very low prevalence and not much effect on the phenotype. *De novo* mutations gained by exome analysis was found in 53% of the patients , with an extra 3% of X-linked inherited mutations. The identification of genetic causes lead to the improvement of specific treatment and dietary advice (De Ligt et al., 2012).

3.11 American College of Medical Genetics and Genomics (ACMG) Guidelines

The past decade has brought the rise in next generation sequencing techniques with its high throughput efficiencies. A large-scale analysis of single gene, gene panel, exome, transcriptomes, epigenetic assays and other genetic tests has been started which proved as a challenge in terms of result interpretations. To solve this problem, ACMG summon a workgroup in 2013, comprising the Association for Molecular Pathology (AMP) and the College of American Pathologists to revise and revisit the guidelines and standards for variant interpretation. This workgroup consisted of directors of clinical laboratories and clinicians.

There has been rapid increase in the detection of novel sequence variants associated with genetic disorders. Some of the phenotypes are associated with single gene while others may be caused due to multiple gene variations. The clinical significance of these variants

can range from those variants which are certainly pathogenic for causing a disorder to those which are certainly benign and don't have much role as a causative agent. ACMG provides algorithm and standard terminologies for interpretation and categorization of these variants. For evaluation of different views of clinical laboratory community, various surveys were organized in United States and Canada which were listed in *GeneTests.org*, requesting the inputs and changes that should be made. Both pharmacogenomics and rare disease identification were included in Laboratory testing experience.

The first survey was sent in February 2013 which aimed on assessing terminology preference. The results were presented in an open forum at annual meeting of ACMG in 2013. The outcome of both the survey and the open forum indicated the following levels on which the variation detection and interpretation should be made:-

1. Categorization of variants under five-tier standard terminology which include **'pathogenic'**, **'likely pathogenic'**, **'uncertain significance'**, **'likely benign'**, and **'benign'** (Table 5).
2. Work group must make their first efforts on Mendelian and mitochondrial variants.

Laboratories were requested to make the use of above scheme and to describe how effective this scheme helps to classify the variants and recognizing the causative agents. Feedbacks given by them were taken very seriously and amendments were made time to time. Responses submitted by over 33 laboratories specify the majority of support for this approach and feedbacks further helped in development of the standards and guidelines. In November 2013, workshop was held at the AMP meeting for the presentation of revised criteria for classification. The approach for variant classification is applicable to all Menedelian gene variants, whether recognized by single gene tests or multiple gene panels and whether based on exome sequencing or genome sequencing. This approach helps to improve the understanding of variated genes in causing diseases and disorders and to increase the knowledge about it.

General terminologies description in ACMG guidelines: ACMG provides the list of standard terminologies to annotate and evaluate a variant. A mutation is defined as a change in the nucleotide sequence which is permanent in nature, while a polymorphism is

defined as a variation with frequency more than 1%. In this case study, a proband is suffering from epilepsy which is being inherited from generation to generation. These terms often lead to incorrect assumptions between pathogenic and benign effects. Some of the laboratories have some additional tiers too which include sub classification of variants categorized under uncertainly significant. The utilization of the term ‘likely’ is being restricted to the data that support high probability of pathogenic and benign variants by some workgroups. However, no quantitative definition of the term ‘likely’ has been proposed. ‘Likely pathogenic’ and ‘Likely benign’ is generally allotted to the variant which have more than 90% certainty of being disease-causing. To explain the unclear designation of a variant, a uniform nomenclature to needed. This improves the effective sharing and downstream use of genomic information. A standard nomenclature of variant gene is maintained by the Human Genome Variation Society (HGVS). Tools which provide correct nomenclature is listed under HGVS guidelines (Richards et al., 2015).

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Table 5: Strength for ‘benign’ and ‘pathogenic’ assertion (Richards et al., 2015)

Abbreviations: -

BS: benign strong
BP: benign supporting
FH: family history
LOF: loss of function
MAF: minor allele frequency
path: pathogenic
PM: pathogenic moderate
PP pathogenic supporting
PS: pathogenic strong
PVS: pathogenic very strong

Objectives

1. To identify genomic signatures involved in Idiopathic Mental Retardation
    a. SNP, INDELS
2. To identify candidates genes associated with the disease development and progression

IV. Materials And Methods

4.1 Clinical Samples

Exome sequencing was performed on 2 patient's blood samples using the Illumina Hiseq2000 platform and the Agilent SureSelect V5 + UTR kit.

TN1604D0102& TN1604D0103 Datasets were downloaded from ftp://bioftp.org as .fastq files

4.2 NGS pipeline (Figure 7)

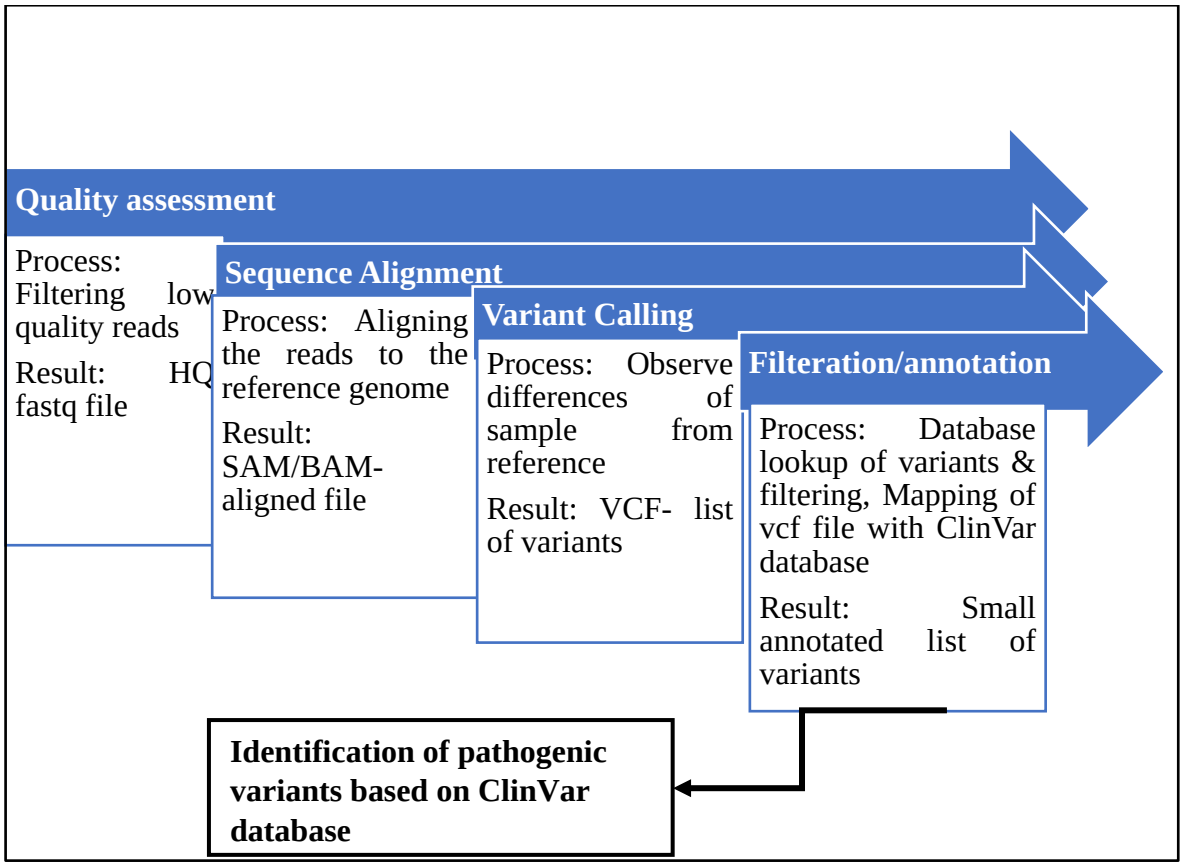


Figure 7: NGS bioinformatics workflow

4.3 Basic Exome sequence and analysis list:

1. Quality control for sequencing data
2. Sequence alignment
3. Variant calling
4. Variant filtering and annotation
5. Identification of pathogenic variations
6. *In silico* confirmation of variations

1. **Quality Assessment and Quality Control:** Raw data quality control should be the initial step of data analysis for any successful study (Urmi et al., 2014). A more advanced tool dealing with raw data quality control is the FastQC package developed by the Babraham Institute bioinformatics group. FastQC offers some additional quality control parameters like the average base quality score per read, the GC content distribution and identification of the most duplicated reads. More importantly FastQC can use aligned bam files instead of fastq files to assess the quality control of raw data. This tool filter reads shorter than given length at several steps, trim reads containing homopolymer and filter HQ reads based on Phred quality score (Patel et al., 2012). In this study, Quality control was done by **FastQC V0.11.3**.

Command: FastQC input file .fastq

2. **Sequence Alignment:** Alignment is the process of mapping short reads to a reference genome because each of the millions of short reads must be compared to the 3 billion possible positions within the human genome (Warden et al., 2014). For this study, hg19 reference genome from UCSC Genome bioinformatics browser was used.

a. **Map to Reference:** As described earlier, the sequencing was done by Illumina sequencing platform that processes both forward and reverse reads simultaneously and exports paired-end sequences in separate files. Paired-end sequences were mapped with most recent version of human genome (GRCh37/hg19) sequence and

it produced a file in SAM format using **BWA v0.7.12** software. Before this step, indexing of reference sequence was done using **SAMtools.**

- **SAMtools:** SAM (Sequence Alignment Mapping) format is a generic format for storing large nucleotide sequence alignments. SAMtools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format. It is flexible in style, compact in size, efficient in random access **(**http://samtools.sourceforge.net/)

  Command: samtools faidx hg19.fa

- **Burrows-Wheeler Aligner (BWA):** BWA is a software package for mapping low-divergent sequences against a large human reference genome, allowing mismatches and gaps. BWA-MEM is usually the preferred algorithm (Li et al., 2009)

  Command: bwa mem –M –p hg19.fasta inputfile.fastq > aligned_reads.sam

b. SAM to BAM Conversion: A BAM (Binary Alignment Mapping) file is just a SAM file but stored in Binary, so SAM file was converted into BAM using **Picard-tools-1.119** to reduce storage space and for better manipulation.

- **Picard-tools:** Picard is a set of command-line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. The Picard command-line tools are provided as a single executable .jar file (https://broadinstitute.github.io/picard/). For the tools to run properly, **Java 1.8** was installed

  Command: picard-tools SortSam INPUT= aligned_reads.sam OUTPUT= sorted_reads.bam SORT_ORDER= coordinate

3. **Variants Calling:** After alignment of short reads to the reference genome, the next step was processing of variants calling. These variants may be responsible for disease, or they may simply be genomic noise without any functional effect. Variant calling format (VCF) is the standardized generic format for storing sequence variation including SNPs, INDELs, larger structural variants and annotations. The ability to detect SNPs with both high sensitivity and specificity is a key step in identifying sequence variants associated with disease, detection of rare variants and assessment of

allele frequencies in populations (Dolled-Filhart et al., 2014). Variant calling of SNPs/INDELs was done using **SAMtools/BCFtoolsv1.3.1**.

- **SAMtools** collects summery information in the input BAMs, computes the likelihood of data by giving each possible genotype and stores the likelihoods in the format of BCF
- **BCFtools** does the actual calling. It can also concatenate BCF files, index BCF for fast random access and convert BCF to VCF. In addition, this tool is a set of utilities that manipulate variant calls in the VCF and its binary counterpart BCF. All commands work transparently with both VCFs and BCFs, both uncompressed and BGZF-compressed (https://samtools.github.io/bcftools**)**

Command: samtools mpileup –uf hg19.fasta sorted_reads.bam | bcftools view –vcg - > sorted_reads.vcf

4. **Variant Filtering and Annotation:** It is the process to predict the effect or function of individual SNP using SNP annotational tools. The variant annotation process places mutations identified by the variant calling step into their biological context. This step is required for the identification of variants of interest based upon a combined filtration of the collected data from one or multiple samples. This will include the variants data quality, their localization at the genomic, gene and transcripts levels, their genotype, their frequency in the general population, their impact at the mRNA and protein levels, and the conservation among species of the affected protein residues, the variant pathogenicity prediction, and reported associations with diseases. At the gene level, they include the gene function, its spatiotemporal expression pattern, its involvement in various pathways, and its involvement in various phenotypes/diseases (David et al., 2015).Variant filtration and annotation step was performed using **SnpSift v4.3k and SnpEff**, respectively.

- **SnpSift:** Sorting Intolerant from Tolerant (SIFT) prediction is based on conserved amino acid residues through different species using comparative sequencing analysis through PSI-BLAST. SnpSift is a toolbox that allows filtering and manipulating annotated files (http://snpeff.sourceforge.net/SnpSift.html). Once genomic variants

have been annotated, we filtered them out in order to find the "interesting/relevant variants". Filteration was performed keeping the quality filter as "(QUAL≥200)"

Command: cat sorted_reads.vcf | java –jar SnpSift.jar filter " (QUAL >= 200)" > filtered_reads.vcf

- **SnpEff:** SnpEff is an open source, Java-based program that rapidly categorizes SNP, INDELs variants in genomic sequences as having either high, medium, low or modifier functional effects. The program may find several different functions for a single variant due to competing predictions based on alternative transcripts. SnpEff uses vcf input and output styles. SnpEff is compatible with GATK and Galaxy, which are popular variant-calling toolkits. This software currently supports 260 genome versions and can be used with custom genomes and annotations (http://snpeff.sourceforge.net/SnpEff_manual.html)

➢ In the Annotation step using SnpEff, 3 types of files have been generated: HTML, .txt file and .vcf file

Command: java –Xmx4g –jar snpEff.jar –v –stats filtered_reads.ann.html GRCh37.75 filtered_reads.vcf > filtered_reads.ann.vcf

- **Mapping with clinical database:** Raw vcf file was mapped with **ClinVar** database. **ClinVar** provides a freely available archive of reports of relationships among medically important variants and phenotypes. ClinVar accessions submit reported human variation, interpretations of the relationship of that variation to human health and the evidence supporting each interpretation. The database is tightly coupled with dbSNP and dbVar, which maintain information about the location of variation on human assemblies. Variation is thus reported as the sequence at one location or as a combination of sequence changes at multiple locations. In other words, ClinVar can represent the interpretation of a single allele, compound heterozygote, haplotypes and combinations of alleles in different genes (Landrum et al., 2014) (Table 6)

Command: java –jar SnpSift.jar annotate clinvar.vcf filtered_reads.ann.vcf > variant_output.vcf

Resources	Basis of the link	Where the link appears
dbSNP	rs# representd in ClinVar	Clinvar link in the Allele section of Cluster report
dbVar	nsv represented in ClinVar	ClinVar link in the links to other resources section
GTR	Genes in which variation has been reported in ClinVar	Gene-specific link under Molecular resources
NCBI gene	Humans genes in which variation has been reported in ClinVar	ClinVar link under related information See variants in ClinVar in the Variation section
MedGen	Conditions or findings reported in ClinVar	Gene-specific link under Molecular Resources
OMIM	Allelic variants accessioned in ClinVar	Allelic variant section
PubMed	Citations provided in a ClinVar submission	ClinVar link under related information

Table 6: Some Resources linking to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)

5. **Identification of pathogenic variations:** Manual filteration was performed on ClinVar mapped vcf file. ClinVar database contains information about **CLNSIG**. It is a string that describes the variant's clinical significance (Table 7). Filteration of pathogenic variants from all the variants and categorized them on the basis of nucleotide alterations, changes in amino acid, type of mutations, disease type, allele frequency and etc. The rsID (Reference SNP ID) was checked in ClinVar database and the disease information that was reported in MedGen and OMIM was noted down.

CLNSIG#	ClinVar Definition
0	Uncertain
1	Not Provided
2	Benign
3	Likely Benign
4	Likely Pathogenic
5	Pathogenic
6	Drug-response related
7	Histocompatibility-related
255	Other (Conflicts, flips, etc)

Table 7: Variant Clinical Significance
(https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/)

6. ***In silico* confirmation of variation:** Data visualization is an essential component of genomic data analysis. Visualization was done using **IGV v2.3.x** (Figure 8)

- **Integrative Genomics Viewer (IGV):** It is a high-performance desktop tool for interactive visual exploration of diverse genomic data. IGV supports real-time interaction at all scales of genome resolution, from whole genome to base pairs. IGV is designed to be accessible to a wide range of users, including bench biologists and bioinformaticians. Many sequencing protocols produce reads from both ends ('paired ends') of genomic fragments of known size distribution. IGV uses this information to color-code paired ends if their insert sizes are larger than expected, fall on different chromosomes or have unexpected pair orientations. Those misalignments, particularly in repeat regions, can also yield unexpected insert sizes and can be diagnosed with the IGV (Robinson et al., 2012)

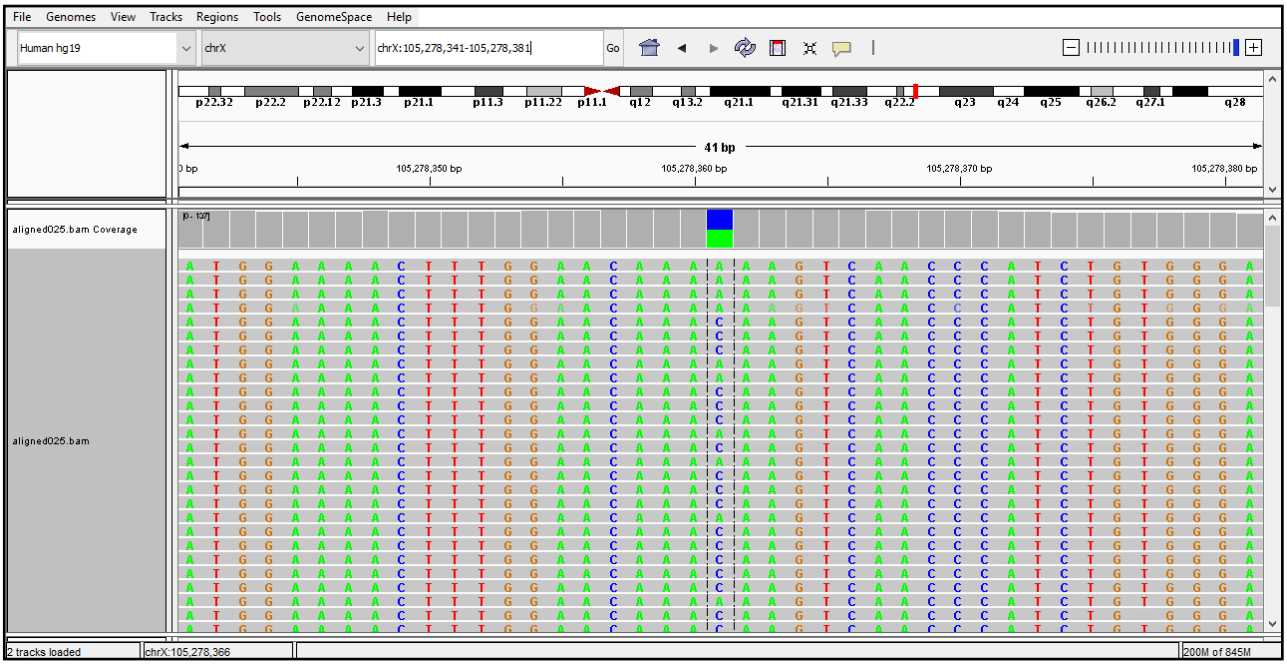


Figure 8: Integrative Genomics Viewer (IGV)

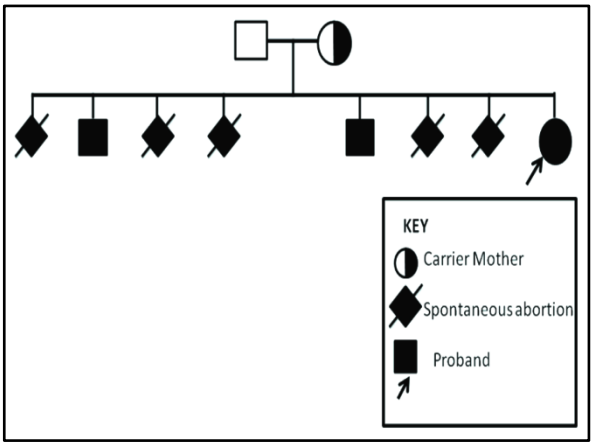
V. Results

In our study, we took two cases of idiopathic mental retardation into considerations with normal karyotype. The report of two of the siblings in Case 1 has already been published (Kunwar et al., 2016) and we took the third sibling for our study. The details of cases are as follows: -

5.1 Case 1 (Kunwar et al., 2016)

Balanced chromosomal translocation may not show morphologically appearing phenotypic expressions but may have increased risk in genetic expressions and can result in spontaneous abortion and serious birth defects in children. The level of expression can vary from person to person. This is the case of a **9-year-old, female** suffering from moderate intellectual disability and minor dysmorphic features of limbs and face

5.1.1 Familial history: The parent of the child had non-consanguineous marriage and the mother had suffered 5 miscarriages. No notable evidence of exposure to toxic elements or illness during pregnancy and problems after post birth was observed. Along with the proband, the other two siblings also have variable intellectual disability (23/M, 19/M, 9/F) (Figure 9).



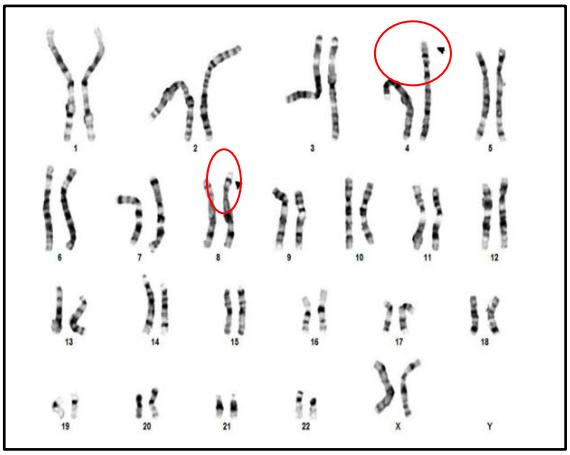


Figure 9: Pedigree chart of the family **Fig 10:** Karyotyping of mother showing t(4p;8p)

The high-resolution karyotype was found to be normal in all the three siblings and father while the mother showed balanced translocation at 4p and 8p arm i.e. t(4p;8p) (Figure 10).

5.1.2 Clinical details: Some of the phenotypic traits of the proband are listed in the following Table 8: -

Phenotype	Proband (09/F)
Hair pattern	Normal
Skull	High anterior hairline, Broad and large forehead
Shape of face	Square face, Pointed chin
Tongue	Microglossia
Nose	Wide nose, Long nose, Wide nasal bridge
Lips	Thin vermilion border of upper lip, Absent cupid's bow, Inverted lower lip vermilion
Ears	Prominent antitragus, Broad width of superior crus of antihelix, Macrotla
Eyes	Strabismus
Periorbital region	Telecanthus
Philtrum	Broad philtrum
Palate	Normal
Neck	Short
Palm	Broad palm
Fingers	Clinodactyly of the 5^{th} finger
Nails	Leukonychia
Legs	Normal
Feet	Normal
Toes	Normal
Chest	Barrel-shaped chest

Table 8: Phenotypic traits and minor dysmorphism observed in proband

Genomic imbalance is responsible for many morphologic variations which are listed in Table 9: -

Body parts	Proband (09/F)
Head: Skull	V
Hair	N
Nose	V
Neck	V
Chest appearance	V
Nails: Finger	V
Toes	N
Head circumference	N
Chest circumference	V
Shoulder width	V
Upper limbs	N
Full hand length	N
Palm length	N
Middle finger length	N
Foot length	N
Outer canthal	V
Inner canthal	V
Inter papillary	V
Palpebral fissure	V
Philtrum	N
Ear length	V

Table 9: Standard measurement of variated body parts (N= Normal, V= Variation)

Figure 11: Phenotypic features of Proband (Case1)

5.2 Case 2

Long contiguous stretches of homozygosity (LCSH) are detected repeatedly by single nucleotide polymorphism (SNP) microarray of chromosomes. Neurodevelopment disorders are the result of genetic variations which include rearrangement of chromosome, single gene mutation and copy number variations (CNV). Along with SNPs and CNVs, epigenetic factors also play major role in neuropsychiatric diseases (Patel et al., 2012). This is the case of a **9-year-old, male**, suffering from idiopathic mental retardation with IQ level less than the normal IQ level of person.

5.2.1 Familial history: The proband was second child born to a non consanguineous healthy parent. There was no remarkable family history of exposure to any drug or toxic during pre and postnatal period. (Figure 12).

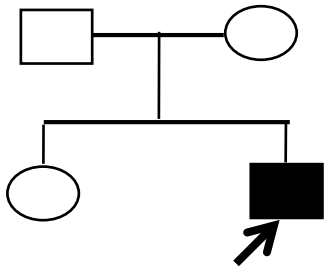


Figure 12: Pedigree chart of the family along with proband

5.2.2 Karyotype: The karyotype result of proband was found to be normal.

5.2.3 Chromosomal Microarray: The microarray technique result showed a long contiguous stretches of homozygosity (LCSH) and a variant of uncertain significance on chromosome 15q26.3. The analysis showed the gain of 405 Kb on the long arm of chromosome 15. Three OMIM-diseases annotated genes were discovered on duplicated region: *ADAMTS17, CERS3* and *LINS*.

5.2.4 Clinical details: The clinical details of proband include poor personal eye contact, intellectual disability and global developmental delay.

5.3 Quality Control: Many tools have been developed to check the quality of the sequencing data. In our experiment, FastQC tool have been selected for Illumina platform generated sequence (Patel et al., 2012). Modern high throughput sequencers (Illumina platform used in our case) have the ability for the generation of tens of millions of sequences in a single run. So before drawing any biological conclusion, simple quality check was done for checking the biasness of the data as it may affect the results. The results obtained are as follows: -

1) **Basic Statistics:** Some simple statistics were generated which contained some of the information (Table 10 and 11) : -
    - Filename: The original filename information
    - File type: To describe whether the file contains actual base calls or other data which had to be transformed to base calls
    - Encoding: To describe which ASCII encoding had been used
    - Total Sequencers: To describe the number of processed sequences
    - Filtered Sequences: Filtering of flagged sequences in Casava mode
    - Sequence Length: Only one value is reported in the report if the length of longest and the shortest sequence are equal
    - %GC: The overall GC content of bases in the sequence

Measure	Value
Filename	TN1604D0102_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	54701680
Sequence length	101
%GC	47

Table 10: Basic statistics of Case 1

Measure	Value
Filename	TN1604D0103_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	51242224
Sequence length	101
%GC	47

Table 11: Basic Statistics of Case 2

2) **Per Base Sequence Quality:** An overview of the range of quality values at each position of the bases in fastq file have been described here. BoxWhisker type of plot was generated for each position. The elements of the plot are described as follow: - (Fig 13)

- The median value is shown by the central red line
- The inter-quartile range is represented by the yellow box
- 10% and 90% points are represented by the lower and upper whiskers
- The mean quality is shown by the blue line

The scores obtained must be high for the better base calls. The y-axis represents the quality scores of the bases. The background of y-axis is divided into three regions which are *(i) green region* represents very good quality scores, *(ii) orange region* represents the reasonable quality calls, and *(iii) red region* represents poor quality calls. For the lower quartile less than 10 will be issued as a warning**.** As all the base calls in both the cases lie in the *green region*, hence it can be considered as a good quality sequence.

Case 1 **Case 2**

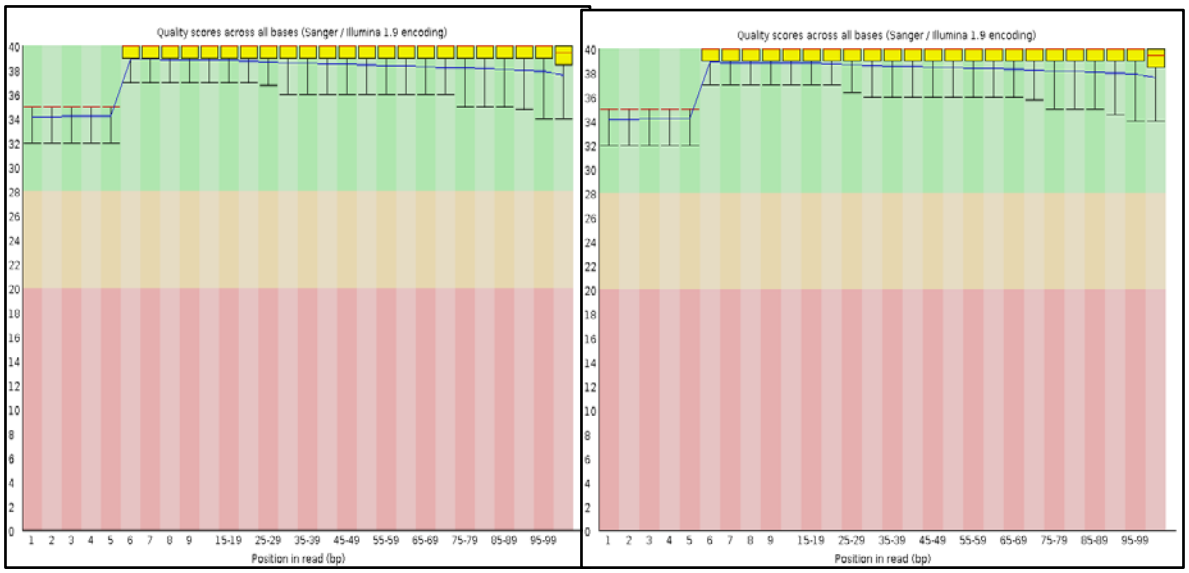


Figure 13: Per base quality

3) **Per Base Sequence Content:** This graph plots out the total proportion of position of base which has been called for four normal DNA bases. The presence of very less differences between different bases depicts the less number of overrepresented sequences in the library (Figure 14).

Case 1 **Case2**

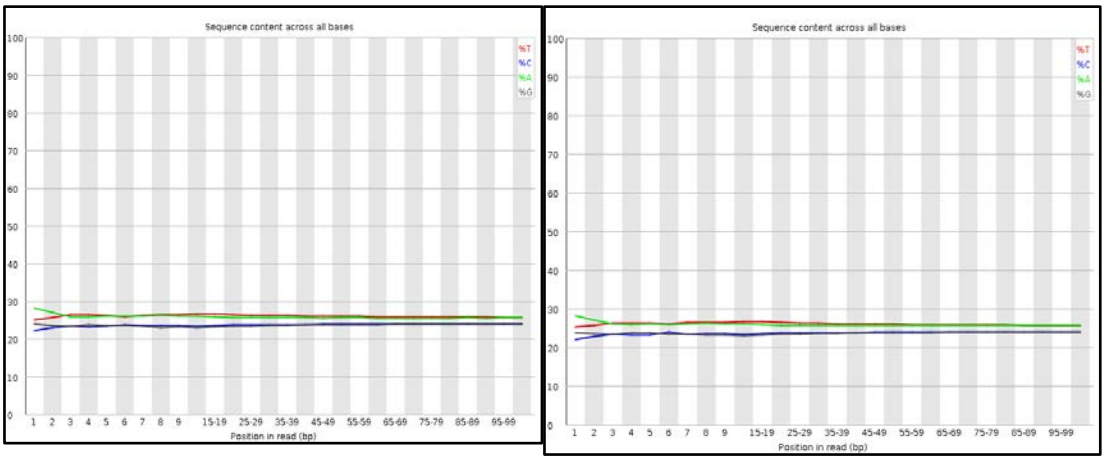


Figure 14: Per base seq. content

4) **Per Base GC Content:** This gives the information about the GC content at each position of base in a file. The overall GC content indicates the percentage of GC in the genome. The less change from the expected result indicates the small amount of overrepresented sequences in the library (Figure 15).

Case 1 **Case 2**

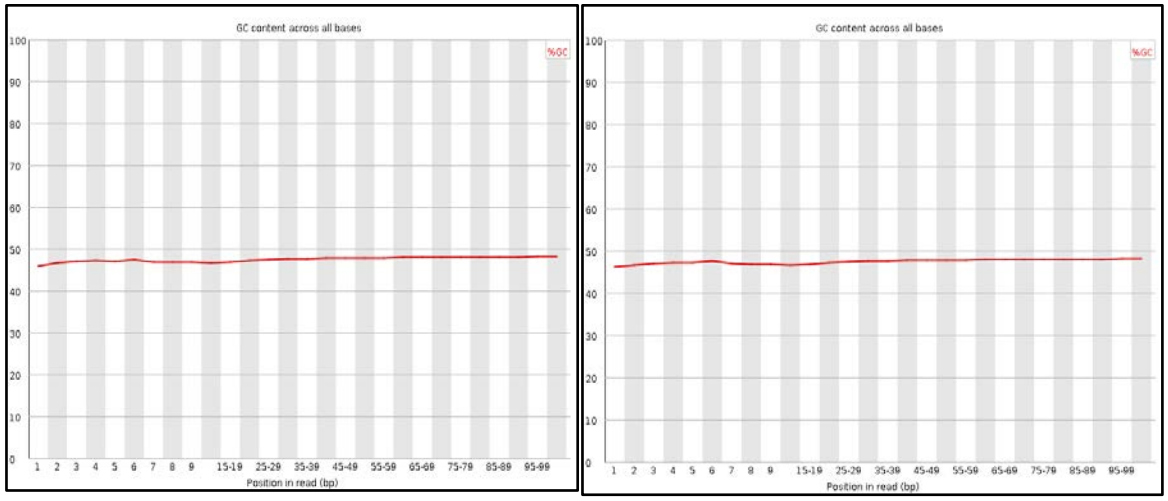


Figure 15: Per base GC content

5) **Per Sequence GC Content:** The GC content present across the whole length of each sequence was measured. This data was then compared with the modelled distribution of GC content. The overlapping of the two central peaks represented the overall GC content in the genome. It also confirms about less contamination while preparing the library (Figure 16).

Case 1 **Case 2**

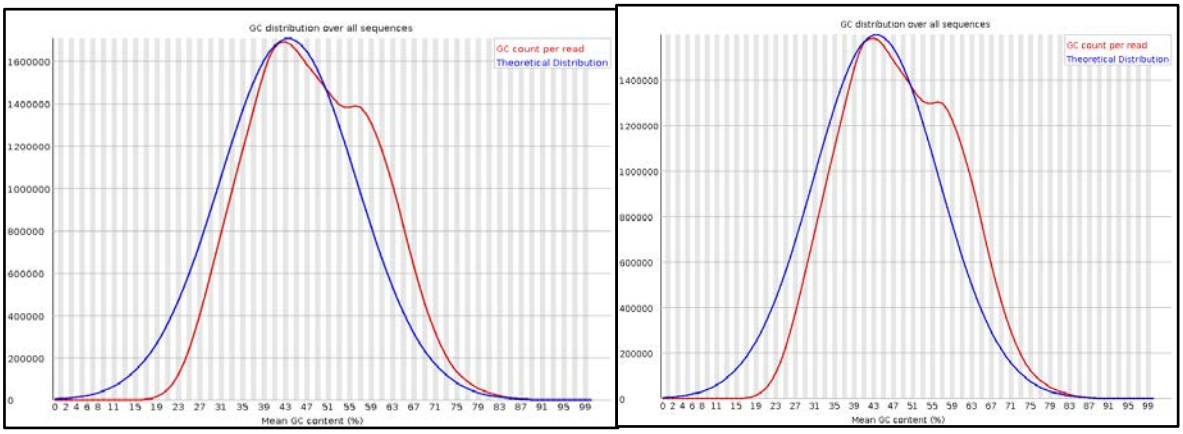


Figure 16: Per sequence GC content

5.4 Variant identification and filteration: The DNA was sequenced with an average sequence length of 101 nucleotides keeping the average depth of 100X coverage.

Case no.	Variants before Filteration	Variants after filteration	Variant rate
Case 1	122,396	65,796	1 variant at every 46,861 bases
Case 2	117,289	62,871	1 variant at every 49,048 bases

Table 12: Number of variants before and after filteration

5.4.1 Chromosome based distribution: The whole exome sequence analysis showed different variation rates on different chromosomes (Figure 17). Number of variations on each chromosome varies with cases. Although the effect of these changes in the variation rates on chromosome is still unknown.

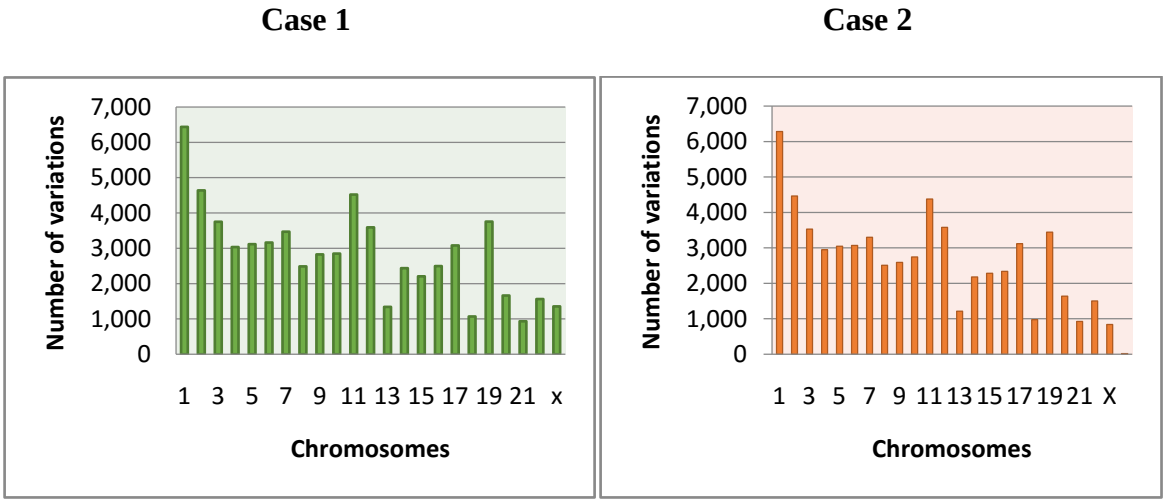


Figure 17: Chromosome wise variation

5.4.2 SNPs and INDELs: Genetic alterations can play crucial role in a person's physical and mental status. SNPs are those genetic markers which helps the researchers to diagnose the causative genes of complex diseases (Jiao et al., 2017). It can be listed amongst the simplest form of DNA variation that can be responsible to cause changes both phenotypically and genotypically. SNPs can be in the form of transition and transversion and can be seen throughout the genome. However, the frequency of occurrence of SNPs can vary from person to person. These SNPs are generally responsible to bring the diversity amongst individual (Shastry et al., 2009).

INDELs are genomic insertions and deletions. They can vary from single nucleotide change to a fewer large region. They can be caused by unequal crossing over, tandem duplication and are hence responsible for frameshift mutation. Presences of intergenic INDELs are more than intragenic INDELs (https://carta.anthropogeny.org).

The numbers of SNPs were found to be more than number of insertions and deletions (Figure 18). However, these variations are yet not classified into the 5 categories listed under ACMG guidelines; hence all variations cannot be considered as having equal impact on causing the idiopathic mental retardation in these cases.

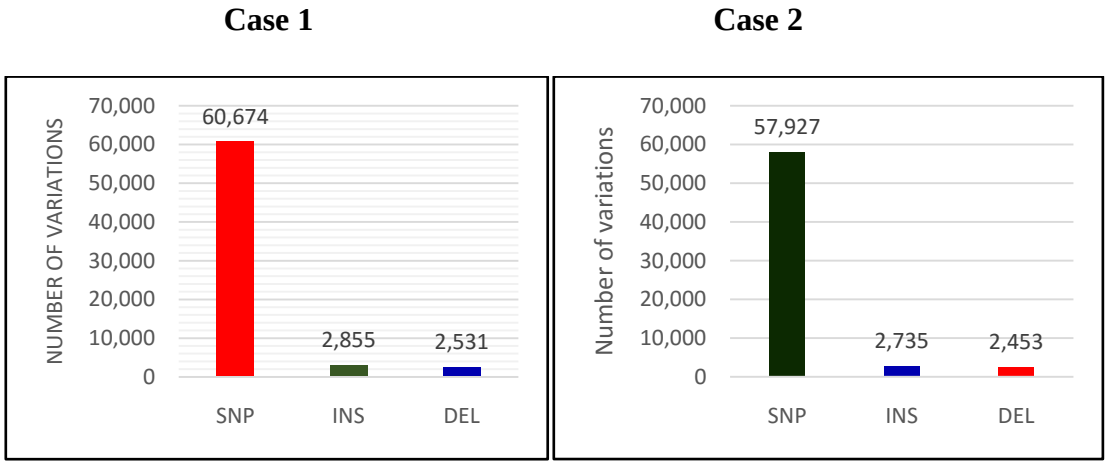


Figure 18: No. of SNPs and INDELs

5.4.2 Transitions and Transversions: Transition (Ts) is defined as change between purine-purine nucleotides and pyrimidine-pyrimidine nucleotides, while transversion (Tv) is defined as change between purine-pyrimidine nucleotides and vice versa. The numbers of transitions were found to be more than number of transversions (Figure 19). The ratio of transition and transversion should be in the range of 2-4. This ratio also plays a major role in evaluation of reported SNP and novel SNPs. The value of Ts/Tv ratio in our case was found out to be 2.4 which lie in the ideal range of the ratio.

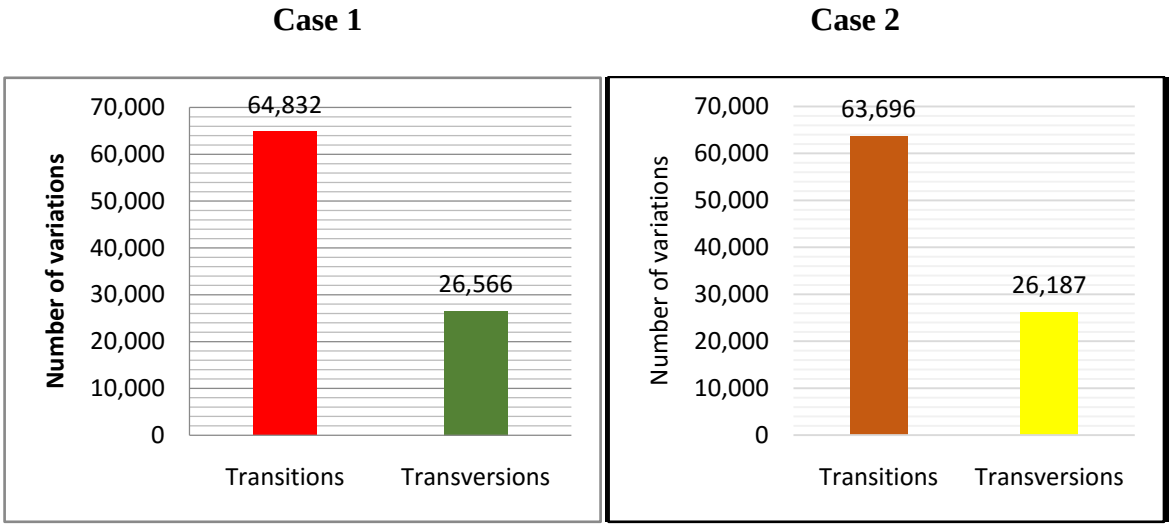


Figure 19: Number of **Ts and Tv**

5.4.4 Missense and Nonsense Variations: Missense mutations are generally those kinds of mutations in which a change in a single nucleotide in a codon results in change in codon that now codes for another amino acid. Nonsense mutations are those mutations in which a change in a codon, turns it into a chain-terminating codon. In our result, we haven't taken silent mutations into consideration because of its less significance (Figure 20).

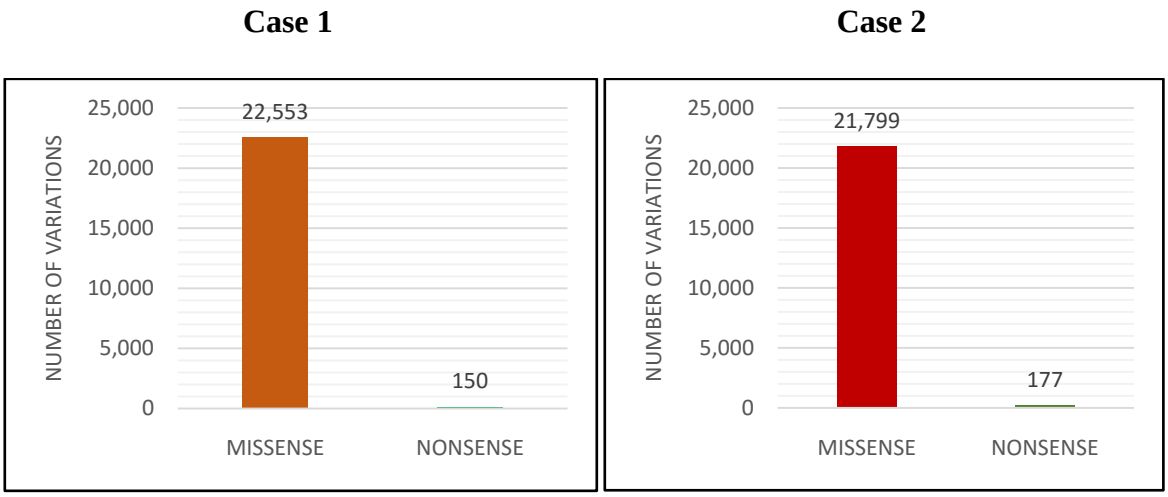


Figure 20: Number of effects by functional type

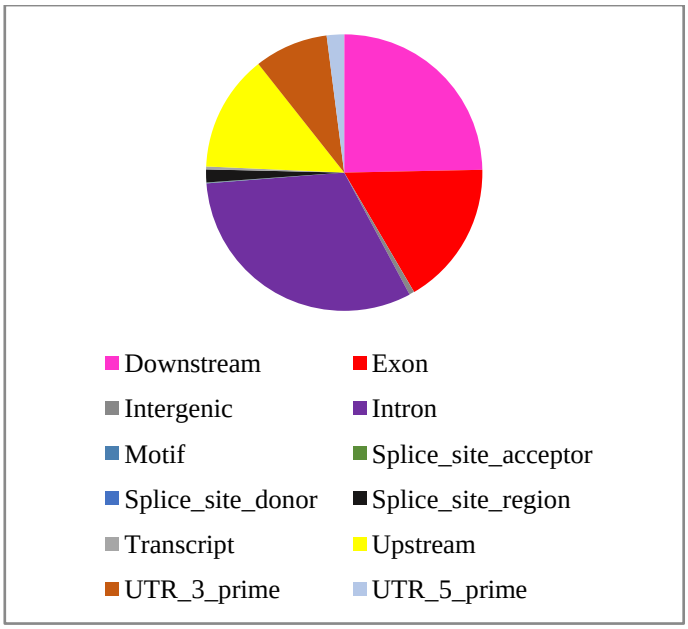


Figure 21(a): Number of effects by type (Case 1)

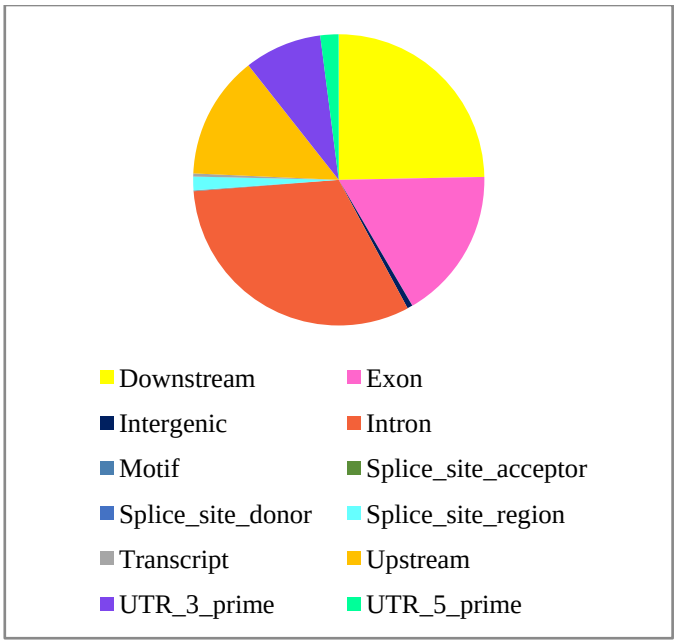


Figure 21 (b): Number of effects by type (Case 2)

Numbers of variations found in intronic regions were found to be more than number of variations in exons (Figure 21). However, the exonic variations are more responsible for phenotypic and genotypic changes.

5.5 Clinical mutation Identification (After mapping with ClinVar)

The annotated file produced using SnpEff tool was further mapped by ClinVar database which have already published variants and its information. Gross filteration helped in the interpretation of genes which are responsible for causing mental retardation in these cases. Variants which were tagged by rsID were filtered out from the file generated after mapping our raw vcf file with ClinVar vcf file. The rs number refers to the accession number which is used by the researchers for referring the specific SNP. It stands for Reference SNP cluster ID (rsID). The standard terms which are used to denote any clinical significance of the genes responsible for disease or disorder is provided by ClinVar. These five standard terms are already being described under ACMG Guidelines. Some of the non-standard terms are also included in the list of clinical significances in ClinVar which include risk factors, Histocompatibility, drug response, etc (https://www.ncbi.nlm.nih.gov/clinvar).

CLNSIG=5,4,0 are the most important terms for whole exome analysis so we filtered out all the three parameters manually and obtained following results in **Case 1**: -

1. 29 Pathogenic genes were detected using CLNSIG=5 as a filter
2. 5 Likely Pathogenic genes were detected using CLNSIG=4 as a filter
3. 78 Uncertain Significant genes were detected using CLNSIG=0 as a filter

These three terms are most important to consider while others are less clinically significant so were not included into the results. All the Pathogenic genes were further well studied under MedGene and OMIM databases. Following table describes the details of these pathogenic genes in Case 1 (Table 13):

chromosome	rsID	Gene	DNA change	Protein change	Clinical effect	Diseases
chr1	rs1801265	*DPYD*	c.85C>T	p.Arg29Cys	Pathogenic	DPYD_deficiency
chr1	rs12021720	*DBT*	c.*4597A>G	p.Ser384Gly	Pathogenic	Intermediate_maple_syrup_urine_disease_type_2
chr1	rs17602729	*AMPD*	c.133C>T	p.Gln45*	Pathogenic	Muscle_AMP_deaminase_deficiency
chr1	rs6025	*F5*	c.1601A>G	p.Gln534Arg	Pathogenic	Thrombophilia_due_to_factor_V_Leiden
chr2	rs1805086	*C2 or F88*	c.458A>G	p.Lys153Arg	Pathogenic	Muscle_hypertrophy
chr3	rs1008642	*SSUH*	c.99C>T	p.Asn33Asn	Pathogenic	Distal_myopathy
chr3	rs696217	*GHRLOS*	c.214C>A	p.Leu72Met	Pathogenic	Metabolic_syndrome
chr3	rs1805124	*SCN5A*	c.1673A>G	p.His558Arg	Pathogenic	Progressive_familial_heart_block_type_1A
chr3	rs696217	*GHRLOS*	c.214C>A	p.Leu72Met	Pathogenic	Obesity
chr3	rs1805124	*SCN5A*	c.1673A>G	p.His558Arg	Pathogenic	Progressive_familial_heart_block_type_1A
chr3	rs3732378	*CX3CR1*	c.935C>T	p.Thr312Met	Pathogenic	Human_immunodeficiency_virus_type_1
chr3	rs3732379	*CX3CR1*	c.745G>A	p.Val249Ile	Pathogenic	Human_immunodeficiency_virus_type_1

chr4	rs36094464	*DSPP*	c.202A>T	p.Arg68Trp	Pathogenic	Dentinogenesis_imperfecta_-_Shield's_type_I
chr4	rs3733402	*KLKB1*	c.428G>A	p.Ser143Asn	Pathogenic	Prekallikrein_deficiency
chr5	rs1494558	*IL7R*	c.197T>C	p.Ile66Thr	Pathogenic	Severe_combined_immunodeficiency
chr5	rs1494555	*IL7R*	c.412G>A	p.Val138Ile	Pathogenic	Severe_combined_immunodeficiency
chr5	rs820878	*HEXB*	c.185T>C	p.Leu62Ser	Pathogenic	Sandhoff_disease
chr6	rs2815822	*F13A1*	c.-19+12A>C	Not Provided	Pathogenic	Deficiency_of_Factor_xiii
chr6	rs1799945	*HFE*	c.187C>G	p.His40Asp	Pathogenic	Hemochromatosis_type_1
chr10	rs10509305	*STOX1*	c.1824A>C	p.Glu608Asp	Pathogenic	Preeclampsia/eclampsia_4
chr10	rs61751507	*CPN1*	c.-533G>A	p.Gly178Asp	Pathogenic	Anaphylotoxin_inactivator_deficiency
chr11	rs3740955; rs886041745	*RAG1*	c.746A>G	p.His249Arg	Pathogenic	Not Provided
chr12	rs1169305	*HNF1A*	c.1813A>G	p.Ser605Gly	Pathogenic	Maturity-onset_diabetes_of_the_young
chr14	rs142777869	*TINF2*	c.734C>A	p.Ser245Tyr	Pathogenic	Revesz_syndrome|Dyskeratosis_Congenita
chr16	rs2238472	*ABCC6*	c.3803G>A	p.Arg1268Gln	Pathogenic	Pseudoxanthoma_elasticum
chr16	rs4784677	*BBS2*	c.209G>A	p.Ser70Asn	Pathogenic	Bardet-biedl_syndrome_2
chr18	rs2269219	*FECH*	c.68-23C>T	Not Provided	Pathogenic	Erythropoietic_protoporphyria|Erythema|Jaundice
chr19	rs2301600	*MAG*	c.399C>T	p.Ser133Ser	Pathogenic	Spastic_paraplegia_75
chr22	rs450046	*PRODH*	c.1562G>A	p.Arg521Gln	Pathogenic	Proline_dehydrogenase_deficiency|Schizophrenia_4
chr22	rs6151429	*ARSA*	c.*96A>G	Not Provided	Pathogenic	Arylsulfatase_A_pseudodeficiency|Metachromatic_leukodystroph

						y
chrx	rs1804495	*SERPINA*	c.909G>T	p.Leu303 Phe	Pathogenic	Thyroxine-binding_globulin
chr1	rs190139590	*SDHB*	n.*3950T>A	Not Provided	Likely pathogenic	Pheochromocytoma
chr6	rs6929137	*CCDC*	c.1831G>A	p.Val611Ile	Likely pathogenic	Estrogen_resistance
chr7	rs146281367	*SLC*	c.1001G>T	p.Gly334 Va	Likely pathogenic	Pendred's_syndrome
chr13	rs9536062	*THSD*	c.670C>G	p.Arg224 Gly	Likely pathogenic	Non-immune_hydrops_fetalis
chr19	rs9384	*SYCE*	c.*288G>T	Not Provided	Likely pathogenic	Glutaric_aciduria

Table 13: Characterization of variants (Case 1)

Maroon- Likely pathogenic variants; Black- Pathogenic variants

- *De novo* mutations in **Case 1:** The variants with rsID were the one which are already linked to some of the diseases and the disorders. But there are some variants which were identified without rsID and hence these can be listed under *de novo* genes. These genes might be also responsible for the disease causing. So, the variants without rsID, i.e. the one denoted as '.' were filtered. Out of 60,931 total variants with ID '.', 8002 were found to be synonymous. These synonymous variants were removed to gain only non-synonymous variants as these are for our use. Numbers of non-synonymous variants were found to be 52,929 out of which 30,695 intronic variants were removed. Hence, the final file contained only 22,234 exonic variants which were then further analyzed to gain following useful variants in Case 1: -

1. Missense variants= 5089
2. Non-sense variants= 4968
3. Frameshift variants= 58

- Numbers of variations found in **Case 2** are as follows: -

1. 29 Pathogenic genes were detected using CLNSIG=5 as a filter
2. 1 Likely Pathogenic genes were detected using CLNSIG=4 as a filter
3. 79 Uncertain Significant genes were detected using CLNSIG=0 as a filter

All the Pathogenic genes were further well studied under ClinVar and OMIM databases. Following table describes the details of these pathogenic genes in Case 2 (Table 14):

chromosome	rsID	Gene	DNA change	Protein change	Clinical effect	Diseases
chr1	rs1801265	*DPYD*	c.85C>T	p.Arg29Cys	Pathogenic	Dihydropyrimidine_dehydrogenase_deficiency
chr1	rs12021720	*DBT*	c.1150A>G	p.Ser384Gly	Pathogenic	Intermediate_maple_syrup_urine_disease_type_2
chr1	rs6025	*F5*	c.1601A>G	p.Gln534Arg	Pathogenic	Thrombophilia_due_to_factor_V_Leiden|Ischemic_stroke
chr1	rs2266782	*FMO3*	c.472G>A	p.Glu158Lys	Pathogenic	Trimethylaminuria
chr2	rs144467873	*APOB*	c.10579C>T	p.Arg3527Trp	Pathogenic	Hypercholesterolemi
chr3	rs1805124	*SCN5A*	c.1673A>G	p.His558Arg	Pathogenic	Progressive_familial_heart_block_type_1A
chr4	rs36094464	*DSPP*	c.202A>T	p.Arg68Trp	Pathogenic	Dentinogenesis_imperfecta_-_Shield's_type_II
chr4	rs3733402	*KLKB1*	c.569G>A	p.Ser190Asn	Pathogenic	Prekallikrein_deficienc
chr5	rs1494558	*IL7R*	c.197T>C	p.Ile66Thr	Pathogenic	Severe_combined_immunodeficiency
chr5	rs1494555	*IL7R*	c.412G>A	p.Val138Ile	Pathogenic	Severe_combined_immunodeficiency
chr5	rs820878	*HEXB*	c.185T>C	p.Leu62Ser	Pathogenic	Sandhoff_disease
chr5	rs351855	*FGFR4*	c.1162G>A	p.Gly388Arg	Pathogenic	Cancer_progression_and_tumor_cell_motility
chr5	rs8632251	*C5/F42*	c.2353C>	5_prime	Pathogenic	Joubert_syndrome

	63		T	-UTR_variant		_17 (Developmental delay)
chr6	rs2815822	*F13A1*	c.-19+12A>C	Not Provided	Pathogenic	Deficiency_of_Factor_xii
chr9	rs1800435	*ALAD*	c.177G>C	p.Lys59Asn	Pathogenic	AMINOLEVULINATE_DEHYDRATASE, Porphobilinogen_synthase_deficiency
chr9	rs2301612	*ADAMTS*	c.1342C>G	p.Gln448Glu	Pathogenic	Upshaw-Schulman_syndrome
chr10	rs10509305	*STOX1*	c.1824A>C	p.Glu608Asp	Pathogenic	Preeclampsia/eclampsia_4
chr11	rs6256	*PTH*	c.247C>A	p.Arg83Arg	Pathogenic	Primary_hyperparathyroidism
chr11	rs144078282	*CLPB*	c.1237A>T	p.Arg413Trp	Pathogenic	neutropenia
chr11	rs1800497	*ANKK1*	c.2137G>A	p.Glu713Lys	Pathogenic	Dopamine_receptor_d2
chr12	rs16910526	*CLEC7A*	c.714T>G	p.Tyr238*	Pathogenic	Familial_chronic_mucocutaneous_candidiasis|Aspergillosis
chr12	rs1169305	*HNF1A*	c.1813A>G	p.Ser605Gly	Pathogenic	Maturity-onset_diabetes_of_the_young
chr12	rs1154510	*HPD*	c.97A>G	p.Thr33Ala	Pathogenic	4-Alpha-hydroxyphenylpyruvate_hydroxylase_deficiency
chr16	rs4784677	*BBS2*	c.209G>A	Not provided	Pathogenic	Bardet-biedl_syndrome_2
chr17	rs12948217	*SPATA22*	c.693C>T	p.Tyr231Tyr	Pathogenic	Spongy_degeneration_of_central_nervous_system
chr17	rs2229989	*SOX9*	c.507C>T	p.His169His	Pathogenic	Camptomelic_dysplasia
chr18	rs2269219	*FECH*	c.68-23C>T	Not Provided	Pathogenic	Erythropoietic_protoporphyria|Erythema|Jaundice

chr19	rs2301600	*MAG*	c.399C>T	p.Ser133Ser	Pathogenic	Spastic_paraplegia_75
chr20	rs73598374	*ADA*	c.22G>A	p.Asp8Asn	Pathogenic	Adenosine_deaminase_2_allozyme|Severe_combined_immunodeficiency_due_to_ADA_deficiency
chr22	rs450046	*PRODH*	c.1562G>A	p.Arg521Gln	Pathogenic	Proline_dehydrogenase_deficiency|Schizophrenia_4
chr19	rs8012	*SYCE2*	c.1250A>G	p.Gln417Arg	Likely pathogenic	Glutaric_aciduria

Table 14: Characterization of variants (Case 2)

Maroon- Likely pathogenic variants; Black- Pathogenic variants

- *De novo* mutations in **Case 2**: The variants without rsID, i.e. the one denoted as '.' were filtered. Out of 58,254 total variants with ID '.', 7,567 were found to be synonymous. These synonymous variants were removed to gain only non-synonymous variants as these are for our use. Numbers of non-synonymous variants were found to be 50,687, out of which 29,696 intronic variants were removed. Hence, the final file contained only 20,991 exonic variants which were then further analyzed to gain following useful variants in Case 2: -

4. Missense variants= 5089
5. Non-sense variants= 2277
6. Frameshift variants= 58

VI. Discussion And Conclusion

The above study helped in highlighting the opportunities and challenges encountered in the diagnosis of intellectual disability by using next generation sequencing technique i.e., Whole Exome Sequencing (Need et al., 2012). The data gained here helped in revealing the dominancy of WES over other technologies in detection of the genetic variants in intellectual disability. The cause of mental retardation generally remains unknown in up to 80% of the patients and out of that, 29% are due to the chromosomal abnormalities which are not cytogenetically visible and are cryptic (Lenhard et al., 2005). A wide variety of causes being listed, 17.4% to 47.1% being the genetic cause of mental retardation (Moeschler et al., 2006). According to the previous studies, Next-generation sequencing can be helpful for molecular diagnosis of the intellectual disability. It provided the diagnosis of 31% patients with nonsyndromic (16 of 51 patients) and 13% with severe intellectual disability (13 of 100 patients) (Need et al., 2012).

The diagnosis of above cases helped in detection of some of the genes which might be responsible for phenotypic changes in Case 1 and Case 2. These are listed in Table 15 and Table 16: -

Chromosome	rsID	Gene	DNA change	Protein change	Clinical effect	Diseases
chr1	rs1801265	*DPYD*	c.85C>T	p.Arg29Cys	Pathogenic	**Dihydropyrimidine_dehydrogenase_deficiency (Intellectual disability)**
chr5	rs820878	*HEXB*	c.185T>C	p.Leu62Ser	Pathogenic	**Sandhoff_disease (Intellectual disability)**
chr19	rs2301600	*MAG*	c.399C>T	p.Ser133Ser	Pathogenic	**Spastic_paraplegia_75 (Intellectual disability)**
chr19	rs8012	*SYCE2*	c.1250A>G	p.Gln417Arg	Likely pathogenic	**Glutaric_aciduria (Intellectual disability)**

Table 15: Clinically significant variations in Case 1

Red- Primary conditions; Blue- Secondary conditions

Out of 29 pathogenic variations gained in both the cases, 4 variations were found to be common in both the probands. These variations may have direct or indirect impact on the characters shown by both the probands.

Chromosome	rsID	Gene	DNA change	Protein change	Clinical effect	Diseases
chr1	**rs1801265**	***DPYD***	**c.85C>T**	**p.Arg29Cys**	**Pathogenic**	**Dihydropyrimidine_dehydrogenase_deficiency (Intellectual disability)**
chr5	**rs820878**	***HEXB***	**c.185T>C**	**p.Leu62Ser**	**Pathogenic**	**Sandhoff_disease (Intellectual disability)**
chr19	**rs2301600**	***MAG***	**c.399C>T**	**p.Ser133Ser**	**Pathogenic**	**Spastic_paraplegia_75 (Intellectual disability)**
chr19	**rs8012**	***SYCE2***	**c.1250A>G**	**p.Gln417Arg**	**Likely pathogenic**	**Glutaric_aciduria (Intellectual disability)**
chr5	rs863225163	*C5/F42*	c.2353C>T	p.Arg785*	Pathogenic	Joubert_syndrome_17 **(Developmental delay)**
chr11	rs6256	*PTH*	c.247C>A	p.Arg83Arg	Pathogenic	Primary_hyperparathyroidism **(Intellectual disability)**
chr11	rs144078282	*CLPB*	c.1237A>T	p.Arg413Trp	Pathogenic	Neutropenia **(Developmental delay)**
chr17	rs12948217	*SPATA22*	c.693C>T	p.Tyr231Tyr	Pathogenic	Spongy_degeneration_of_central_nervous_system **(Developmental delay)**

Table 16: Clinically significant variations in Case 2

Red- Primary conditions; Blue- Secondary conditions; **Bold- common variants of both cases**

Although the variation in genes associated with intellectual disability are very prone to cause phenotypic changes, not all genes have the same intensity of effects in all the persons suffering from this disease. Also, an uncertainty is faced while relating the expression of pathogenic genes in phenotypic changes of the patients but not all variations have the same impact on the expression of phenotype, which confirms that only unbiased sequencing results are capable in characterization of clinical range associated with specific genetic mutations (Rauch et al., 2012). Such genetic mutations may show their effect in future.

Hence, the comparison of genotype-phenotype is an important step while reporting the causative genes and variations.

The number of SNPs in Case 1 was found out to be more than number of SNPs in Case 2. But the number of genes responsible for phenotypic changes are more in Case 2, which predicts the presence of more number of *de novo* variations in Case 1. There is a wide similarity in chromosome wise variations in both cases. These variations do not have direct relation with the length of chromosome. For example, in the case of chr19 and chr3, the length of chr19 is smaller than length of chr3, but the number of variations is almost similar in both the chromosomes. The number of intronic variations was found to be more than exonic variation. But there is no significant role of intronic variants is seen till now.

This result mainly focuses on those potential pathogenic genes which are at present responsible for intellectual disability. There are several genes which do not have direct association with a specific phenotype (Rauch et al., 2012). This finding is also seen in our study. For example, *DPYD* gene variations have been reported to cause abnormality of the nervous system, skeletal system and epilepsy (Willemsen et al., 2011). But none of them were seen in both the probands. Instead, both had the same problem of intellectual disability. The other 7 genes along with *DPYD* which shows intellectual disability and developmental delay as their secondary condition include: *HEXB* (Ebrahimzadeh-Vesal et al., 2017), *MAG* (Webber et al., 2009), SYCE2 (Jorge et al., 2015), *C5/F42* (Strømme et al., 2000), *PTH* (Roberts et al., 2014), *CLPB* (Wartman et al., 2015), SPATA22 (Serikawa et al., 2015). This observation shows that one need to study the secondary functions too while studying the genetic variations.

The diagnosis of variations would help in prenatal analysis, chances of recurrence of the variations in future generations and genetic counseling. It may also provide knowledge for future studies related to intellectual disability. The familial study could add additional information to the current study. There can be a possibility of alteration in the phenotypic expression with time (Need et al., 2012). *De novo* mutations in these patients can be identified by using a family-based exome sequencing approach (Gilissen et al., 2012). Knowledge about pathogenicity and frequency of sequence variant can also help in better understanding and diagnosis of unknown diseases (Rauch et al., 2012).

VII. References

1. Aggarwal, Shagun, et al. "Aetiologic spectrum of mental retardation & developmental delay in India." *Indian Journal of Medical Research* 136.3 (2012): 436.
2. Alkan, Can, Saba Sajjadian, and Evan E. Eichler. "Limitations of next-generation genome sequence assembly." *Nature methods* 8.1 (2011): 61-65.
3. Aneek, Bhowmik, et al. “Exome data analysis for clinicians: how and why.” *“Genetic Clinics 10”* (2017)
4. Armatas, Vasilios. "Mental retardation: definitions, etiology, epidemiology and diagnosis." *Journal of Sport and Health Research* 1.2 (2009): 112-122.
5. Bamshad, Michael J., et al. "Exome sequencing as a tool for Mendelian disease gene discovery." *Nature Reviews Genetics* 12.11 (2011): 745-755.
6. Barbieri, Christopher E., et al. "Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer." *Nature genetics* 44.6 (2012): 685-689.
7. Bodensteiner, John B., and G. Bradley Schaefer. "Evaluation of the patient with idiopathic mental retardation." *Journal of Neuropsychiatry and Clinical Neurosciences* 7 (1995): 361-361.
8. De Ligt, Joep, et al. "Diagnostic exome sequencing in persons with severe intellectual disability." *New England Journal of Medicine* 367.20 (2012): 1921-1929.
9. Dolled-Filhart, Marisa P., et al. "Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing." *The Scientific World Journal* 2013 (2013).
10. Ebrahimzadeh-Vesal, Reza, et al. "Identification of novel missense HEXB gene mutation in Iranian-child with juvenile Sandhoff disease." *Meta Gene* 12 (2017): 83-87.
11. file:///C:/Users/Acer/Downloads/FastQC_Manual.pdf
12. Galasso, Cinzia, et al. "" Idiopathic" mental retardation and new chromosomal abnormalities." *Italian journal of pediatrics* 36.1 (2010): 17.
13. Gilchrist, Carol A., et al. "Whole-genome sequencing in outbreak analysis." *Clinical microbiology reviews* 28.3 (2015): 541-563.

14. Gilissen, Christian et al. “Disease Gene Identification Strategies for Exome Sequencing.” *European Journal of Human Genetics* 20.5 (2012): 490–497. *PMC*. Web. 21 Apr. 2017.
15. http://allseq.com/kb/wgsvswes/
16. http://enabled.in/wp/census-of-india-2011-disabled-population/
17. http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties
18. http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties
19. http://samtools.sourceforge.net/mpileup.shtml
20. http://www.bgi.com/applications/mouse-exome-sequencing/#tab-id-2
21. https://carta.anthropogeny.org/moca/topics/genomic-insertions-and-deletions-indels
22. https://carta.anthropogeny.org/moca/topics/genomic-insertions-and-deletions-indels
23. https://link.springer.com/protocol/10.1007/978-1-60327-411-1_1
24. https://www.biosciencetechnology.com/white-papers/2013/12/targeted-resequencing-streamlining-ngs-clinical-research
25. https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/
26. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf
27. https://www.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html
28. https://www.ncbi.nlm.nih.gov/clinvar/
29. https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/#clinsig_scv
30. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3244507/
31. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3510890/
32. ICD-10 GUIDE FOR MENTAL RETARDATION
33. Jiao, Yun et al. “Single Nucleotide Polymorphisms Predict Symptom Severity of Autism Spectrum Disorder.” *Journal of Autism and Developmental Disorders* 42.6 (2012): 971–983. *PMC*. Web. 20 Apr. 2017.
34. Jorge, Rita, et al. "Intellectual disability and overgrowth—A new case of 19p13. 13 microdeletion syndrome with digital abnormalities." *American Journal of Medical Genetics Part A* 167.11 (2015): 2839-2843.

35. Kaufman, Sandra Z., and Nicole Kaufman. *Retarded isn't stupid, Mom!*. PH Brookes, 1988.
36. Ku, C-S., D. N. Cooper, and G. P. Patrinos. "The Rise and Rise of Exome Sequencing." *Public Health Genomics* (2016).
37. Kunwar, Fulesh, and Sonal R. Bakshi. "Familial Constitutional Rearrangement of Chromosomes 4 & 8: Phenotypically Normal Mother and Abnormal Progeny." *Journal of clinical and diagnostic research: JCDR* 10.4 (2016): GD01.
38. Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics* 25.14 (2009): 1754-1760.
39. Maulik, Pallab K., and Catherine K. Harbour. "Epidemiology of intellectual disability." *International encyclopedia of rehabilitation* (2010): 1-12.
40. Men, Artem E., et al. "Sanger DNA sequencing." *Next-Generation Genome Sequencing: Towards Personalized Medicine* (2008): 1-11.
41. Metzker, Michael L. "Sequencing technologies—the next generation." *Nature reviews genetics* 11.1 (2010): 31-46.
42. Moeschler, John B., and Michael Shevell. "Clinical genetic evaluation of the child with mental retardation or developmental delays." *Pediatrics* 117.6 (2006): 2304-2316.
43. Morozova, Olena, and Marco A. Marra. "Applications of next-generation sequencing technologies in functional genomics." *Genomics* 92.5 (2008): 255-264.
44. Need, Anna C., et al. "Clinical application of exome sequencing in undiagnosed genetic conditions." *Journal of medical genetics* (2012): jmedgenet-2012.
45. Pabinger, Stephan, et al. "A survey of tools for variant analysis of next-generation genome sequencing data." *Briefings in bioinformatics* 15.2 (2014): 256-278.
46. Patel, Ravi K., and Mukesh Jain. "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data." *PloS one* 7.2 (2012): e30619.
47. Rauch, Anita, et al. "Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study." *The Lancet* 380.9854 (2012): 1674-1682.
48. Richards, Sue, et al. "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical

Genetics and Genomics and the Association for Molecular Pathology." *Genetics in Medicine* 17.5 (2015): 405-423.

49. Roberts, Jennifer L., et al. "Chromosomal microarray analysis of consecutive individuals with autism spectrum disorders or learning disability presenting for genetic services." *Gene* 535.1 (2014): 70-78.
50. Robinson, James T. et al. “Integrative Genomics Viewer.” *Nature biotechnology* 29.1 (2011): 24–26. *PMC*. Web. 20 Apr. 2017.
51. Serikawa, Tadao, et al. "Advances on genetic rat models of epilepsy." *Experimental animals* 64.1 (2015): 1-7.
52. Shastry, Barkur S. "SNPs: impact on gene function and phenotype." *Single Nucleotide Polymorphisms: Methods and Protocols* (2009): 3-
53. Shendure, Jay, and Hanlee Ji. "Next-generation DNA sequencing." *Nature biotechnology* 26.10 (2008): 1135-1145.
54. Stavropoulos, Dimitri J., et al. "Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine." *Npj Genomic Medicine* 1 (2016): 15012.
55. Strømme, Petter, and Trond H. Diseth. "Prevalence of psychiatric diagnoses in children with mental retardation: data from a population-based study." *Developmental Medicine & Child Neurology* 42.04 (2000): 266-270.
56. Thapar, Anita, et al. "The genetics of mental retardation." *The British Journal of Psychiatry* 164.6 (1994): 747-758.
57. Trivedi, Urmi H., et al. "Quality control of next-generation sequencing data without a reference." *Frontiers in genetics* 5 (2014): 111.
58. Warden, Charles D., et al. "Detailed comparison of two popular variant calling packages for exome and targeted exon studies." *PeerJ* 2 (2014): e600.
59. Warr, Amanda, et al. "Exome sequencing: current and future perspectives." *G3: Genes| Genomes| Genetics* 5.8 (2015): 1543-1550.
60. Webber, Caleb, et al. "Forging Links between Human Mental Retardation–Associated CNVs and Mouse Gene Knockout Models." *PLoS Genet* 5.6 (2009): e1000531.

61. Willemsen, Marjolein H., et al. "Chromosome 1p21. 3 microdeletions comprising DPYD and MIR137 are associated with intellectual disability." *Journal of medical genetics* (2011): jmedgenet-2011.
62. Wortmann, Saskia B., et al. "CLPB mutations cause 3-methylglutaconic aciduria, progressive brain atrophy, intellectual disability, congenital neutropenia, cataracts, movement disorder." *The American Journal of Human Genetics* 96.2 (2015): 245-257.
63. www.nimblegen.com