QATAR UNIVERSITY

COLLEGE OF HEALTH SCIENCES

INVESTIGATING THE VALIDITY AND SIGNIFICANCE OF VARIANT CALLS BY NEXT GENERATION SEQUANCING (NGS)

BY

YASMIN WALID JAMIL ABU AQEL

A Thesis Submitted to the Faculty of

College of Health Sciences

In Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in

Biomedical Science

June 2016

COMMITTEE PAGE

The members of the Committee approve the thesis of Yasmin Walid Abu Aqel defended on 26th of May 2016.

Dr. Ahmed Malki

Thesis Supervisor

Dr. Hatem el Shanti

Committee Member

Dr. Nasser Rizk

Committee Member

Dr. Zafar Nawaz

External Examiner, HMC

Approved:

5/6/2016

Dr. Asma Al-Thani Dean, College of Health Sciences

Abstract

Whole genome or exome sequencing enables the fast generation of large volumes of data and is currently a hot topic in research. This technique has the subject of extensive research and has vast applications in healthcare and medicine. Next generation sequencing (NGS) has the advantage of providing large length reads when compared to the traditional method of Sanger Sequencing. NGS enables the identification of genetic disease-causing variants, thus, improving the quality of healthcare, diagnostics and biomedical research.

One of the major challenges of NGS is the analysis of large data outcomes. The diversity in DNA library preparation methods for various available platforms may result in data inaccuracies. Furthermore, the disparity in variant calling accuracies as a result of using diverse algorithms complicates the process of NGS data analysis. As a result, there is a large possibility for false positive and/or false negative results due to alignment and/or chemistry errors.

In this project, we utilized the MiSeq platform that was selected based on its cost effective properties and ability to provide rapid genetic analysis. The autism panel is used in this study to assist the investigation of genomic features associated with autism by targeting 101 genes linked specifically to Autism. Here, we hypothesized that we could devise NGS analysis criteria to distinguish false positive and/or false negative sequencing calls to improve the quality of the generated sequencing data. Four Autism patients cohort of Arab descent have been used as a model for this research. We were able to prove our hypothesized criteria by validating the detected variances by Sanger Sequencing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation

t RNA	: Transfer RNA
NGS	: Next generation sequencing
MPSS	: Massively parallel sequencing platforms
SOLiD	: Sequencing by Oligo Ligation Detection
ASD	: Autism Spectrum Disorder
dNTP's	: deoxynucleoside triphosphate
dATP	: deoxyadenosine triphosphate
dGTP	: deoxyguanosine triphosphate
dCTP	: deoxycytidine triphosphate
dTTP	: deoxythymidine triphosphate
APS	: Adenosine 5′ phosphosulfate
PPi	: Pyrophosphate
GA	: Genome Analyzer
SBS	: Sequencing By Synthesis
CCD	: Charge-coupled device
In-del	: Insertion and Deletion
PCR	: Polymerase Chain Reaction
CNVs	: Copy number variations
OISM	: Oligo *in-silico* synthesized microarray
NGSTs	: Next-generation sequencing technologies
ADOS	: Autism Diagnostic Observation Schedule

ADI	: Autism Diagnostic Interview
SCQ54	: Social Communication Questionnaire
BAC	: Bacterial artificial chromosome
bp	: base pair
NLT	: Nextera Library Tagment
NEH1	: Nextra Enrichment hyb1
SMB	: streptavidin magnetic beads
VCF	: Variant Call Format
BWA	: Burrows-Wheeler Aligner
GATK	: Genome Analysis Toolkit
SIFT	: Sorting Intolerant From Tolerant
PolyPhen-2	: Polymorphism Phenotyping v2
TM	: Melting Temperature
IDT	: Integrated DNA Technologies
SAP	: Shrimp Alkaline Phosphatase
IGV	: Integrative genomic viewers
SNV	: Single nucleotide variants
SINEs	: Short interspersed nuclear elements
LINEs	: Long interspersed nuclear elements
Mins	: Minutes

Acknowledgments

Praise is to Allah, his Majesty of his uncountable blessings, and best prayers and peace be upon his best messenger Mohammed, his pure descendant, and his family and his nobel companions.

First I would like to thank my supervisors Dr. Ahmed Malki, Dr. Hatem el Shanti, and Dr. Nasser Rizk for their help and continuous support. Dr. Dina Ahram, and Yasser who helped me a lot during these years in understanding and performing well.

Next, I would like to thank my family; my parents, my parents in law, my sisters and brothers, my great husband and my lovely daughter for their patience and support. Without their love and support over the years none of this would have been possible. They have always been there for me and I am thankful for everything they have helped me achieve.

Special thank to Dr. Fatma Abdallah who helped me through every step and kept encouraging me to do my best, words cannot describe her support, love, and assistance. Dr. Nurra Abdi my great friend who assist me in reviewing the thesis. My colleagues, Hanan , Karim, Wesal, Khawla , Hiba, Eman, and Mohammed without their incredible help and assistance noting will be achieved.

I would like to express my thanks to my collage (college of Health Sciences) and all my doctors and professors for their assistance. Also, I'd like to thank my institute Qatar Biomedical Research Institute (QBRI) for their support.

1. Introduction:

DNA sequencing is used to determine the exact order of nucleotide within the DNA and/or RNA molecules. This technique provides scientists in the biological fields with access to molecular cloning and breeding which helps them in not only detecting diseases but also identifying genes behind these diseases. Most importantly this technique is capable of increasing life quality by decoding some of the life ambiguities [1]. The first and most significant achievement of the DNA sequencing project was the human genome project that has a cost of three billion dollars. This project started 13 years ago and was finalized in 2003. Human genome project was established to study the genetic structure of human as well as clarifying the human development[2]. This project was developed with excellent capabilities to illustrate lots of human disease mysteries by understanding the disease genotype therefor precisely describing the phenotypic structure of these diseases. Upon discovery this achievement showed great potential in improving disease prevention and treatment [3]. The power of DNA sequencing is to have a method that is easy to access, fast, accurate and most importantly cost-effective. This is what lead scientist to think of such a method that could aid obtaining this goal.

Sequencing technology was first developed between 1964 and 1965, by an American biochemist, Robert Holley and his colleagues, who sequenced a nucleic acid making them the first in the field to do sequencing methods for tRNA [4]. However, in 1975, Frederick Sanger thought of sequencing the DNA by using the chain-termination method, since then this gold method has been known as Sanger sequencing (first sequencing generation)[5]. The strength behind Sanger sequencing is

the fact that it is (i) very accurate, (ii) highly efficient, (iii) reduces the radioactivity and (iv) can perform long read length (800-1000bp) [6]. Despite being a robust sequencing method with high throughput data, Sanger sequencing was limited by the fact that it is very expensive and takes a long time to performs sequencing. These obstacles lead the scientist to think of a new generation of sequencing that could overcome these mentioned problems. This paves the way for invention the Next Generation Sequencing (NGS).

The first encounter with NGS or the so-called second-generation sequencing was between 1994-1998, however it is first official existence commercially was not until 2005 [7]. Lynx Therapeutics (USA) Company produced the first NGS product in 2000 now these products are owned by Illumina. NGS is also known as the Massively Parallel Signature Sequencing platforms (MPSS) for its ability to perform the sequencing using massively parallel procedure; which facilitates the sequencing with 1 million to 43 billion short readouts (50-400 bases each) per instrument run [8]. After discovering MPSS and specifically by 2004, 454 Life Sciences (Branford, CT, USA) took the advantage of producing a new version of pyrosequencing that was able to reduce the costs of sequencing to 6-fold less when compared to automated Sanger sequencing (Life Sciences, a Roche Company). In 2006, after NGS was officially commercially available, Genome Analyzer was released by Solexa, followed by the establishment of Sequencing by Oligo Ligation Detection (SOLiD), which was provided by Agencourt [1, 9].

Nowadays, NGS is highly used in the clinical and medical field specifically in the area of identification of genetic diseases, which will improve the quality of life and healthcare. That is due to its cost-effectiveness, time efficiency, and provision of

high throughput data properties in comparison to Sanger Sequencing. Using NGS brought lots of challenges; one of the greatest challenges was the analysis of large data outcomes drawing attention to the fact that the diversity of the methods used to prepare DNA library for various available platforms may result in data inaccuracies [10]. Furthermore, the disparity in variant calling accuracies as a result of using various algorithms complicates the process of NGS data analysis[11]. As a result, there is a larger possibility for false positive and/or false negative results due to alignment and/or chemistry errors [12]. Therefore, in this project, we focused in utilizing the MiSeq platform. This platform was selected for its cost effective properties as well as its ability to provide rapid genetic analysis [13]. In this project, I will be using the autism panel to assist in the investigation of genomic features associated with Autism Spectrum Disorders (ASD) by targeting 101 genes linked specifically to ASD [14-16].

Herein, we hypothesize that devising NGS analysis criteria to identify false positive and/or false negative sequencing calls is possible. By doing so, we can improve the quality of the generated sequencing data. In this project we will be utilizing four autism patient cohort of Arab descent as a research model. The findings of this research can serve in improving the identification of disease causing variants and increasing the power of diagnosing not only ASD but also other diseases using this technique.

1.1 Hypothesis and Aims:

1.1.1 Hypothesis:

"To evaluate the autism panel's genes, certain parameters and thresholds will be applied on the MiSeq NGS output data as indicated by quality indices. We hypothesize a significant increase in the reliability of the generated sequencing data and expect to narrow down the amount of data generated by using a specific analysis pipeline to identify and minimize the amount of false positives and/or false negatives in the results".

1.1.2 Aims

To prove this hypothesis, following aims were indicated:

- **Aim 1**: To produce MiSeq NGS data for four Arab patients with autism utilizing a commercially provided autism panel.
- **Aim 2**: Identify the variants that are called in each patient and determine the quality indices.
- **Aim 3:** To validate these variant calls using targeted Sanger sequencing.

2. Literature Survey:

NGS is a new technology in the genetics filed, which promises a new, fast, and cost-effective way in diagnosis and selection of treatment in the future. This could be used for many genetic diseases in detecting the gene or genes causing the disease, which will lead to decode the mysteries of life and increase the quality of life [1]. It helps mainly in the heterogeneous diseases [17]. The significance of using NGS, is that it could provide more accurate result due to deep coverage in the sequencing the data [17]. There are different types of NGS, each performing a specific function and has specific characteristic features. So the type or platform to be used depends on the research question and the research application. Below is a brief description of the major NGS systems:

2.1Roche 454 System:

Roche 454 System was the first NGS system that was commercially produced. It depends on the use of pyrosequencing technology that detect releasing of pyrophosphate during nucleotide incorporation [18]. In this system the DNA samples are first prepared in the library with 454-specific adaptors. Then they are denatured into single strand and captured by amplification beads which are then followed by PCR emulsion [19]. To complement the bases of the template strand on a picotiter plate, one of the following nucleotides (dNTP's) dATP, dGTP, dCTP and dTTP is added with the assistance of sets of enzymes such as; ATP sulfurylase, luciferase, luciferin, DNA polymerase, and adenosine 5′ phosphosulfate (APS) that will lead to release pyrophosphate (PPi) equals to the amount of the incorporated dNTP's. The release of PPi results in ATP production

that converts the luciferin into oxyluciferin, which then results in generating visible light, however, in the case of unmatched bases, apyrase will degrade it. After that, the reaction system will be completed by the addition of another dNTP, and then the pyrosequencing reaction is repeated [20].

The main three advantages of Roche 454 System are (a) the length of its readouts which is about 700 bp with 99.9%, accuracy, (b) it is quite fast requiring only 10 hours to finish and (c) it provides a cost-effective sequencing system [21].

2.2ABI SOLiD System:

In 2006, Applied Biosystems purchased a sequencing system called Sequencing by Oligo Ligation Detection (SOLiD). SOLiD technology depends on the ligation sequencing technique that relies on using a specific flowcell to do sequencing of the DNA libraries. This requires having 8 base-probe ligation sites; the first base which is the ligation site, the fifth base which is cleavage site, and in the last base there are 4 different fluorescent dyes linked to it .The fluorescent signal will be detected when the probes bind to the template strand and then removed by the cleavage of the last 3 bases. Then the sequencing of the fragment will be gathered using ladder primer sets [9].

2.3Illumina GA/HiSeq System:

Solexa first produced the Genome Analyzer (GA) in 2006. After that, Illumina bought it in 2007. GA sequencing system uses the sequencing by synthesis (SBS) technique. In this system, fixed adaptors are first attached to the DNA libraries, which are then denatured into single strands and attached to the flowcell part of the machine. Using the method of bridge amplification forms then clusters of DNA clones, linearization enzyme is next used to splice the library into single strands before proceeding into the sequencing steps. To complete the single strand template that will be used in the sequencing step, four kinds of dNTP's (ddATP, ddGTP, ddCTP and ddTTP) containing different cleavable fluorescent dye and a removable blocking group are added one base each time. Which enables signal detection by a (charge-coupled device) CCD [9].

2.4 Compact PGM Sequencers:

2.4.1 Ion PGM from Ion Torrent

Ion Personal Genome Machine (PGM) launched by Ion Torrent is the first sequencing machine that does not use fluorescence and camera scanning for sequencing because it is based on using semiconductor-sequencing technology. Protons are released when the polymerase adds dNTP's to the DNA strand. PGM detect the change in pH to recognize if a nucleotide is added or not. Each time the chip of the system adds a nucleotide after another and the detected signal will be in the form of voltages. If the nucleotide did not bind there will be no voltage and

therefore no signal, on the other hand, if there are two nucleotides bound, then two voltages will be detected [22].

2.4.2 MiSeq:

MiSeq, also produced by Illumina, depends on the sequencing by synthesis (SBS) technology, which is similar to the GA/HiSeq. Unlike the GA/HiSeq, MiSeq does not sequence large fragments making it much faster than the GA/HiSeq with a processing time of around 8 hours. MiSeq is used in sequencing smaller fragment sizes such as sequencing of amplicons, and small genome; making it perfect for sequencing targeted genes, metagenomics, targeted gene expression, and HLA typing [23]. MiSeq has the ability to detect different mutations, including the point mutation and any small In-dels, making it able to perform the same function of Sanger sequencing but with high efficiency[24]. The read length that can be carried out by MiSeq ranges between 36 bp single reads which are about (120 MB output) up to 2 × 150 paired-end reads which are about (1–1.5 GB output). Because of this improvement of the read length, the resulting data from MiSeq will act better in contig assembly if compared to other systems using the same method technology such as HiSeq [1]. Comparison between the different NGS platforms based on different features is summarized on (Figure 1).

High-end sequencing- Platform†	Sequencing chemistry	Read lengths/ through put	Run time	Template prep	Application
Roche 454 -Titanium FLX	Pyrosequencing	400 bp 400 Mb/run	10 hours	Emulsion PCR	Denovo WGS of microbes, pathogen discovery, Exome seq
Illumina/Solexa -HiSeq 2000	Reversible terminator chemistry	2×100bp 600 GB/ run (dual cell)	11.5 days	Solid-phase	Human WGS, exome seq, RNA-seq, Methylation
ABI/LifeTechnology-SOLiD 5550XL	Sequencing by ligation	2×60bp 15 GB/day	8 days	Emulsion PCR	Human WGS, exome seq, RNA-seq, Methylation
HelicosBiotechnologies	Reversible Terminator chemistry	25-55 bp 28 GB/run (avg)	>1 GB/hour	Single molecule	Human WGS, exome seq, RNA-seq, Methylation
Roche 454- GS Junior	Pyrosequencing	400 bp 50 Mb/run	10 hours	Emulsion PCR	Denovo WGS of microbes, pathogen discovery, Exome seq
Illumina/Solexa- MiSeq	Reversible terminator chemistry	2×150bp 1.0-1.4 Gb	26 hours	Solid-phase	Microbial discovery, Exome seq, Targeted capture
ABI/ Lifetechnology- Iontorrent	H+ Ion sensitive transistor	320 Mb/run	8 hours*	Emulsion PCR	Microbial discovery, Exome seq, Targeted capture

*Sample preparation – 6 hours, sequencing time – 2 hours, †Data shown here represent the highest figures currently available on the company website and is highly likely to change by the time this article is published

Figure 1. Comparison between the different types of the Next generation sequencing (NGS). [25]

The impact of using MiSeq appears clearly in many aspects, especially in clinical fields. MiSeq depends on mainly in the targeted genome methods, which aids in targeting only the anticipated region of the genome that could cause the disease. This occurs, by designing specific panel that targets hundreds of genomic regions, which are expected to cause the investigated disease. Targeted sequencing panels could help in disease diagnosis and most importantly identifying the mutations causing that disease. In addition it could assist in the process of choosing the best therapeutic decision in many diseases like cancer [26], and acute myeloid leukemia [27].

In the vaccination field, MiSeq system could be used widely for improving the identification of epitopes of phage-displayed antigen-specific libraries. This could be

done by targeting the model antigen regions by serum antibodies in immunized (vaccinated) people and then checking the immunoreactivity developed in the hundreds of antigenic fragments [28]. MiSeq is also used in immunology; nowadays it has a great role in detecting sickle cell anemia in the fetus during pregnancy, by obtaining maternal plasma and checking the fractional fetal DNA concentration. This is achieved by applying the targeted sequencing property of the MiSeq for the amplicons of the fetus that contain single-nucleotide polymorphisms [29].

The agricultural field was also hit by MiSeq system in many aspects all-aiming in increasing the food safety and quality criteria. One of the most important aspects to do so is to detect specific taxa of microorganism from the complex environment that could affect the growth of some vegetables. Detecting Salmonella strains in tomatoes is a known example of the benefits of applying MiSeq system to target a specific organism from a complex microenvironment and identify where it came from and whether it was during enrichment or before enrichment in order to know the type of microorganism that affects tomato and in which stage [30].

MiSeq has been widely used in oncology, specifically in the aspects of (i) cancer treatments, (ii) drug designing, (iii) the use of cancer drugs, (iv) patients response to the drug, (v) in the classification of tumors and (vi) in the identification of the rare somatic mutation that may cause the tumors. Also MiSeq could be used in differentiation between somatic and germline mutation which may help in identifying the causing mutation for the disease. In the field of cancer research, although the usefulness of MiSeq in the evolution of personalized medicine and finding the suitable biomarkers, still there are some limitations for its using as capturing

techniques in cancer, one of these limitations is the quality of DNA which is mainly low [31].

Despite the fact that MiSeq has great impact in the clinical field and it could help humanity in diagnosing mysterious diseases, this system is capable of generating huge amounts of data, which necessitates the need of using filtering criteria in order to get the wide-ranging benefits out of this important sequencing system. These criteria depend on the quality of the data delivered and the target depth of the coverage data, i.e. how often a specific target is read and sequenced because DNA is sequenced multiple times during the sequencing process, which increases the rate of error that may occur. The median coverage differs among different users, some users rely on 10-fold (10X) others rely on 20-fold (20X) or 30- folds (30X) at > 90% of bases covered as minimum coverage in quality assemblies [32]. The importance of these folds is to assemble properly and find the true variants with excluding the encountered errors which increases the mutation detection rate [31].

In comparison between different platforms which are based on targeted genome like Ion PGM and MiSeq (Figure 2), it was reported that the score of mean base-call quality gained from MiSeq was higher through the entire reads and the scatterings within the reads were small at specific positions; interestingly these reads were the opposite when Ion PGM was used. The read lengths obtained by MiSeq were consistent in lengths, while with Ion PGM they were widely distributed. Therefore it was concluded that the base call quality is higher when using MiSeq [17].

	PGM		MiSeq
	TMAP	Novoalign	
Average total number of bases (Mb)	295.97	201.73	469.42
Average read length (base)	116	116	150
% mapped on human genome	96.8%	78.8%	75%
% on target regions	26.7%	28.3%	22.7%
Mean depth of coverage	63	57	95
% of target regions at >10-fold coverage	93.7%	92.1%	96.8%
% of target regions at >20-fold coverage	85.9%	82.0%	93.2%

doi:10.1371/journal.pone.0074167.t002

Figure 2.Comparison between PGM and MiSeq sequencing performance [17].

Like any new technology, NGS has many advantages and disadvantaged. Starting with the encountered advantages, first seen an advantage in the throughput data they generated that are much higher relative to the classical sequencing. For example, massive parallel sequencing has a capacity of hundreds of thousands or millions of templates. The second advantage is reducing of the representation bias in template libraries during the PCR amplification step. The third and the most important advantage is that NGS is both cost effective and time efficient, because it needs fewer reagents than the needs of a classical way of sequencing.

For the disadvantages and limitations that are faced while using the NGS, the first point is the fact that it generates short reads. This necessitates the need for an earlier genome sequenced for the new strains to be re-sequenced to detect the point

mutations. The second point is the repetition of regions, which is difficult to be gathered if the size of the fragments exceeds the average, read length. Third point is the preparation of the libraries, which needs lots of controlled steps like fragmentation, adaptor ligation and PCR amplification. This will need to control the size of the fragments initially to avoid any pitfalls that they could face. Finally, the errors and artifacts that are produced by NGS because of the high throughput data generated [33].

In our project we used one of the genetic diseases that could be diagnosed by NGS, which is Autism Spectrum Disorders (ASD). ASD is a group of neurodevelopmental disorders that share common features such as social and language impairment, repetitive and stereotypic behaviors. It had been found first by Kanner and then Asperger in 1944 and 1943 [34, 35]. The prevalence of ASD worldwide is significantly increasing with having a 1 in 68 children occurrence nowadays; also ASD is more common in males than females. ASD is a complex inheritance and genetic heterogeneity, and frequently it is consistent with other diseases such as intellectual disability, seizure disorders, and Fragile-X [36]. NGS have been used broadly as a genetic technology in diagnosis and identification of genetic causes of ASD and expand the understanding of the disease, by recognizing novel genetic loci and risk factors, as well as chromosomal changes, which is known as copy number variations (CNVs). Starting by using Oligo *in-silico* synthesized microarray (OISM) platforms [37], then by using next-generation sequencing technologies (NGSTs) which have been employed in the last two to three decades. Although NGS has been used in high genome coverage and sensitivity to study ASD intensively [38], the genetic cause of the majority of the ASD is elusive [39].

ASD diagnosis is still a challenge because there is no specific medical laboratory test such as blood or urine analyses that could be used to identify it. Until today the gold standard diagnosis methods for ASD relies on the Autism Diagnostic Observation Schedule (ADOS), which includes behavioral evaluations and the Autism Diagnostic Interview (ADI). In addition, Social Communication Questionnaire (SCQ54) can be used, which is a survey test used as an initial screening tool. Therefore, using platforms such as microarrays and sequencing will help more in illustrating the copy number variations (CNV) that have a high impact on shaping our individuality. However, the problem with using microarray platforms seen in encompassing only a few hundreds to thousands of large BAC (Bacterial Artificial Chromosome) clones [40]. Consequently, the critical genome locations are not covered because there is no enough BAC coverage. While recently, NGS became wildly used because of its fast sequencing ability of the whole genome as well as costing less by sequencing many individuals DNA in one single platform in few days. CNV is used to define structural gain or loss of genomic segments that fall into a group between 1,000 base pairs to 5 megabases [41]. However, with the discovery of NGS the definition has become wider to include smaller variations that help in the identification of several CNVs at different scales and distinguishing polymorphic CNVs from harmful ones.

3. Materials and Methods:

3.1 Patients and Ethics statement:

QBRI institutional review board has approved the project. In this study a total of 4 independent ASD patients were used. DNA was obtained from peripheral blood leukocytes for all investigated patients after informed consent or informed parental consent for all minors.

3.2 Next Generation Sequencing using the MiSeq platform:

Library preparation and sequencing were performed on the MiSeq platform (illumina, San Diego, CA, USA) according to the manufacturer's specifications and guidelines of Autism Panel-TruSight Rapid Capture Enrichment kit (illumina) as follows:

3.2.1 Library preparation and DNA enrichment:

For sample preparation, TruSight Rapid Capture sample preparation kit used (illumine). 50ng/ µl of DNA was used as a starting genomic material and was quantified for each sample. This followed by DNA shearing process step by using an enzymatic DNA fragmentation.

3.2.1.1 Tagment Genomic DNA:

The sheared DNA will be used for the tagmentation step by adding 25µl of Tagment DNA buffer (illumina) to each well of the NLT (Nextera Library Tagment) plate containing the samples, then 15µl of Tagment DNA Enzyme 1 (illumina) was added also to the samples. After that, the plate was sealed and put

in the centrifuge at 280g for 1 minute. Then the plate was pre-heated by using microheating system and is incubated at 58 °C for 10 minutes. After that, 15µl of Stop Tagment buffer (illumina) was added to the samples on the plate and the plate was shaken by using microplate shaker at 1800rpm for 1 minute. This step was followed by centrifugation of the plate at 280g for 1 minute then incubation at room temperature for 4 minutes. Samples were then tagged by unique barcode indices and common adapter sequences were added to the ends for each fragmented sample.

3.2.1.2 Purification of the tagmented DNA:

The tagmented DNA is purified from the unbound DNA fragments and adaptors which can bind tightly to DNA ends, in which might interfere with downstream processes if not removed. This was done by adding 65 µl of well-resuspended sample purification beads to each well of the NLT plate. The plate was shaken on a microplate shaker at 1800 rpm for 1 minute, and then incubated at room temperature for 8 minutes, followed by centrifugation at 280g for 1 minute. The plate was then placed on magnetic stand for 2 minutes, after that 200µl of freshly prepared 80% ethanol (EMSURE) was added to each well without disturbing the beads. After 30 seconds incubation, the 80% ethanol was removed from each well. Then the plate was incubated at room temperature for 10 minutes to dry on the magnetic stand. After the incubation time completed, 22.5µl of suspension buffer was added to each well of NLT plate.

3.2.1.3 PCR clean up:

After that, this step was subsequently followed by library pooling step (combine multiple libraries with different indices into a single pool) and the first amplification via limited PCR cycle program. The PCR step is critical since it is also adds index 1 (i7) and index 2 (i5) that are sequencing primers needed for sequencing, as well as common adapters (P5 and P7) that required for cluster generation and sequencing. After that, PCR purification was performed by using AMPure XP Beads (illumina) to purify the library DNA, and provides a size selection step that removes very short library fragments from the population.

3.2.1.4 Hybridization Step:

Consecutively, this was followed by hybridization process, which was known as enrichment capture of targeted regions. This was performed by using TruSight Rapid capture kit of Autism panel where the purified DNA library was mixed with the capture probes of targeted regions (TruSight-Autism panel) for the recommended hybridization time by using HSP plate NEH1 (Nextra Enrichment hyb1, illumina). This should be done to makes sure that targeted regions bind to the capture probes thoroughly (Table 1). The TruSight Autism covers the genomic landscapes of known reported or related genes (101) that linked to Autism. Immediately, this step followed by addition of 250µl streptavidin magnetic beads (SMB)(illumina) to capture enriched probes that contain the regions of interest. Then, three washing steps were performed by using 200µl of Enrichment wash solution (illumina) to remove the non-specific binding from the beads.

Later on, the enriched libraries were then eluted from the beads and prepared for a second hybridization to allow for further targeted regions enrichment by using Enrichment elution buffer 1 and 2N NaOH (Table 2). In this step, the library elution was mixed again with the capture probes of target regions and was also followed by second washing and capturing buffers. Finally, second PCR was performed to amplify and prepare the enriched DNA library for sequencing process in MiSeq machine.

Table 1. Reagents for library preparation for MiSeq that added to each well for hybridization step.

Reagent	**Volume (μl)**
DNA Library sample or library pool from NLS plate	40
Enrichment Hybridization Buffer	50
TrueSight Content Set CSO	10
Total Volume per Sample	100

Table 2. Reagents for enrichment step in MiSeq that added to each well for Elution step.

Reagent	**Volume (μl)**
Enrichment Elution Buffer 1	28.5
2N NaOH	1.5
Total Volume per Sample	30

3.2.2 Clustering generation:

After library preparation, the enriched libraries were loaded into the MiSeq flowcell chip (illumina) according to the manufacture guidance of MiSeq machine preparation where the cluster generation was processed. The flowcell is a glass slide with lanes, each lane is a channel coated with two types of oligos in which each fragment/molecule was isothermally attached to these oligos to allow fragments amplification later on. The attachment of these fragments is known as the hybridization step that enriches the fragments attachment to the oligos on the chip surface. These oligos are complementary to the adapter region on one of fragment strand. Continuously, the amplification step was followed where the polymerase creates a complement on the hybridized fragment. After that, the double stranded molecule is denatured and the original strand is washed away. Then, the strand is clonally amplified to form bridge amplification. These steps were repeated for the second adapter region that is hybridized and amplified to the second type of oligos on the flowcell, resulting in two strands attached to flow cell. Repeating the process to form the clonal amplification, which contains millions of clusters. Finally, the revers strand at each oligos type was cleft away.

3.2.3 Sequencing:

After cluster generation, sequencing begins with the extension of sequencing primer to produce the first read. With each cycle, four fluorescently tagged nucleotides compete for the addition of growing chain and one that is complementary to sequence the template will be incorporated. After addition of each nucleotide, the clusters are excited by a light source, and characteristic

fluorescents signals are emitted, this process is called sequencing by synthesis. The number of cycles determines the length of the reads. The emission wavelength along with the signal intensity determines the base calls. For a given clusters, all identical strands are read simultaneously. Hundreds of millions of clusters are sequenced in a massively parallel process. Once the first read finishes, the product is washed away and the index1 primer will hybridize to the template. The second read is generated similar to the first read. After completing the second read, the product will be washed away and the template then folded over and binds a second Oligo on the flow cell. Then index2 is read as the same manner as index1 then the product will be washed away and the polymerases will extend the template and form paired-end sequencing. This double stranded DNA is denatured and the forward strand is washed away, leaving the reverse strand where the second read start by the introduction of read 2 primer sequencer, in the same manner as read 1. This entire process generates billions of short reads representing all the fragments that will be saved in FASTQ format. (Figure 3) Summarizing the workflow of the MiSeq.

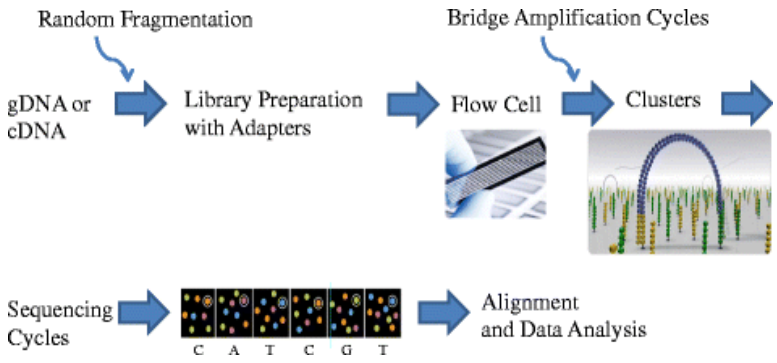


Figure 3: Workflow of next generation sequencing that based on the mechanism of sequencing by synthesis like the MiSeq [42]. It shows the steps starting from DNA till the sequencing alignment and data analysis.

3.2.4 Bioinformatics:

After the sequencing process, sequences from pooled library are separated according to unique indices introduced during the sample preparation. The enrichment workflow of DNA demultiplexes the indexed reads, will generate a FASTQ files which is a text-based format containing both nucleotide sequence and its consistent quality scores. For each sample, reads with similar stretches of base calls are locally cluster. Forward and reverse reads were paired, which will create contiguous sequences. These contiguous sequences were aligned and mapped to the human genome reference by the Burrows-Wheeler Aligner (BWA) for variants identification. Variants will be identified, and the output files will be transcribed to the alignment folder by using the Genome Analysis Toolkit (GATK). This alignment folder is called variant call format file (VCF). These

steps will be directly performed by the MiSeq platform (MiSeq Reporter 2.5)[43]. All the following steps are a default setting in the MiSeq platform we used it as it is without changing any of the parameters.

The VCF file contains an enormous amount of information needs filtration, using the specific pipeline that aid in doing this purpose. The pipeline that we used (hypothesized) depends on different parameters, quality of the sequencing coverage and target coverage, passing variants, allele frequency, and mutation type. For the quality of sequencing, we depend on the 20X (Q20) coverage as the depth of sequencing coverage with target coverage that ranges between (95.7% to 97.5%). Also we choose only the passing variants because we want to filter those passing variants to exclude the false positive.in order to end up with a list of most rare variants, we choose the allele frequency which is <1%. Moreover, to detect the mutation type, we depend on the pathological mutation only.

These mutations have been checked by using different software such as SIFT (Sorting Intolerant From Tolerant); this database distinguishes the tolerant mutation form in tolerant one. The other website that used for detecting the severity of the mutations is PolyPhen-2 (Polymorphism Phenotyping v2). This website used as a tool for expecting the effect of changing in amino acid on the structure and function of the protein. The other software that we used is MutationAssessor that predict the influence of substation of amino acid in the function of the protein (Figure 4).

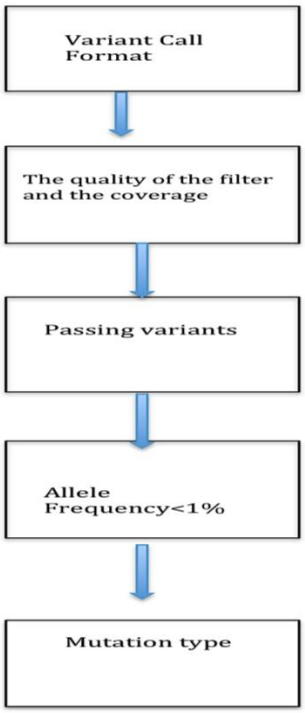


Figure 4. Pipeline of the hypothesized criteria. This pipeline containing filtering parameters that used in the filtering the data of VCF.

This pipeline could be implemented on all kind of autistic conditions, and this appears clearly in which we depend on affected patients only who include the homozygous and heterozygous, on other word, the dominant, recessive, and de novo mutations.

3.3 Evaluation of Variants in the four patients:

3.3.1 Primer design:

After variant list generation according to the applied quality test criteria, we designed primers for the 18 variants that passing the hypothesized criteria and we chose 4 random variants that were failed by the hypothesized criteria while they passed according to MiSeq reporter criteria. Using UCSC genome bioinformatics did the designing of the primers and Primer 3 input (version 0.4.0) software. By using UCSC genome, we were able to find the human reference for each gene in order to use it in primer designing of the desired variants. We used (feb.2009 (GRCH37/hg19)) assembly, after we the gene name was entered; the proper isoform was chosen based on the (NM) number that provided by VCF for each variant, which distinguishes the isoforms of the same gene. Then the mRNA/Genomic alignment was checked to design the primer based on the sequence of the mRNA, then the sequence of the gene was taken and paste it into the Primer 3 software to pick up the suitable primers.

The properties of the picked primers were checked by using Amplifix software, these properties including Tm (Melting Temperature), GC content, 3'end stability, polyX, self-dimer, and self-end dimer. All those properties must be optimized and checked well before choosing the primer; this will help in ensuring the efficiency of the primer work. For more confirmation of the quality of the picked primers, we check them by using tools in the UCSC genome such as in-silico PCR, and BLAT search genome. These tools help us in confirming the specificity of the picked primers; that means, these primers could align only to the

desired gene on its chromosome and not with other regions, genes, or another chromosome. They run in-silico PCR and check all these properties.

3.3.2 Optimization of the Primers:

The designed primers had been ordered by using Integrated DNA Technologies (IDT), after receiving we started the optimization process to check the proper condition for using these primers with the patient's samples.

First of all, we checked the (23) pairs of primers on a control sample, by doing PCR with different temperature degree to check the best annealing temperature. We choose three different temperatures, which are (52°C, 58°C, and 60°C), and do PCR reaction with 20ng/µl DNA concentration (Table 3). Due to high DNA concentration in almost all of the samples when they run on an agarose gel, we decided to repeat the same mentioned conditioned with changing the DNA concentration, by using 10ng/µl instead of 20ng/µl (Table 4) (Table 5). Then we did the PCR using the optimized temperature, which is (58°C) for the 23 pairs of primers by using the same PCR condition for all reaction. (Table 6) All the ingredients are ready made except the dNTP, which we prepared it (Table 7). The taq polymerase enzyme used is HotStar (Qiagen), and the dNTPs set is PCR grade (Qiagen).

Table 3. PCR Ingredients and the concentrations with DNA concentration of 20ng/µl.

Ingredients	**Concentration**	**1X**
Sterile distilled water	__	17.2ml
PCR Buffer	10 X	2.5ml
dNTPs	10mM	2ml
Forward Primer	10µM	0.5ml
Reverse Primer	10µM	0.5ml
Taq Polymerase (HotStar)	5000U/ml	0.3ml
DNA Sample	20ng/µl	2µl
Total volume per tube		25µl

Table 4. PCR Ingredients and the concentrations with DNA concentration of 10ng/µl.

Ingredients	**Concentration**	**1X**
Sterile distilled water	__	17.2ml
PCR Buffer	10 X	2.5ml
dNTPs	10mM	2ml
Forward Primer	10µM	0.5ml
Reverse Primer	10µM	0.5ml
Taq Polymerase (HotStar)	5000U/ml	0.3ml
DNA Sample	10ng/µl	2µl
Total volume per tube		25µl

Table 5. PCR Condition that used for optimization of the primers.

Step	Temperatures	Time	
1	95°C	15min	
2	94°C	60sec	
3	(52,58,60)°C	30sec	35 Cycles
4	72°C	60sec	
5	72°C	10min	
6	4°C	∞	

Table 6. Optimized Condition of PCR. This condition applied for the four patients to detect the correct variation.

Step	Temperatures	Time	
1	95°C	15min	
2	94°C	60sec	
3	58°C	30sec	35 Cycles
4	72°C	60sec	
5	72°C	10min	
6	4°C	∞	

Table7. dNTP preparation. It shows the concentrations that used for each dNTP preparation.

Reagent	**Concentration**	**1X**
dATP	10 mM	2.5 µl
dCTP	10 mM	2.5 µl
dGTP	10 mM	2.5 µl
dTTP	10 mM	2.5 µl
Autoclaved distilled water	____	90 µl
Total volume per tube		100 µl

3.4 Polymerase Chain Reaction of the four patient's samples:

Primers of the variants were optimized and checked for the proper condition of the PCR, then PCR for each patient with the detected variants was carried out by using annealing temperature of 58°C.

3.5 Gel Electrophoresis:

The entire PCR product was checked by using (2% agarose gel) to detect desired bands at specific sites. 100 bp Plus Ladder (Gel Pilot Qiagen) was used as a ladder marker for the patient's samples bands.

3.6 Validation by Sanger Sequencing:

Sanger sequencing was used for variants validation following checking the right size of the band of interest on the gel for the patient's samples. Sanger sequencing plays an important role in variants validation that although the next generation sequencing is more useful in the long read fragment, but Sanger sequencing is more helpful for the small-scale projects and for validation of Next Generation sequencing results. First of all, mixture of 2.5µL of the PCR product in the Optical plate, 96 well (Applied Biosystems Inc) and 1µl of 3.2pmol of primer for each reaction in Fast Reaction plate, 96 well (Applied Biosystems Inc) was submitted to the Sanger sequencing laboratory for sequencing .

After that, 1µl of ExoSAP-IT that stored at -20°C (USB/Affymetrix) will be added to 2.5µl of the submitted post-PCR reaction product to be a total of 3.5µl reaction volume. ExoSAP-IT, aimed to do a fast and efficient purification of Polymerase Chain Reaction (PCR) products for downstream applications such as sequencing. It is made up of having two hydrolytic enzymes, Exonuclease I and Shrimp Alkaline Phosphatase (SAP), which is formulated in an specially optimized buffer aimed to the removal of unwanted

primers and dNTPs from a PCR product mixture. Then the plate was sealed with MicroAmp Optical Adhesive Film (Applied Biosystems). Which followed by Spinning down for 2 seconds, mixing by vortex for 2 seconds, then spinning down to collect the reaction mix in the bottom of the well. After that, the plate was put in the thermal cycler to start the purification step and the condition was used as follows in (Table 8).

Table 8. Thermal cycler program that performed after addition of ExoSAP-IT to the post-PCR reaction product.

Temperature	**Time (minutes)**	**Remarks**
37°C	15	Hydrolysis of single-stranded DNA and dNTPS
80°C	15	Inactivate the enzymes
4°C	∞	

Once the reaction is done, the plate was briefly centrifuged to collect each sample at the bottom of its well. Then the plate Proceeded to Cycle Sequencing reaction, this step was done by transferring 2µl of purified sample to the submitted 1µl specific 3.2pmol reaction primer wells plate, and place 1µl of Control DNA template (pGEM-3Zf+(0.2 µg/µl)) in in the last reaction well of the plate run to be processed in parallel with samples. The DNA control template was added to 4µl of M13 Control Primer (0.8pmol). After that Ready

Reaction Mix (Applied Biosystem Inc.) and 5X Sequencing Buffer (Applied Biosystem) were added to the plate as follows; a total of 7µl of Master Mix is added to the sample and 5µl to the Control (M13). Then the plate had been sealed with MicroAmp Optical Adhesive Film. After that, it was spin down for 2 seconds, mixed by vortex for 2 seconds, and spin down to collect the reaction mix in the bottom of the well (Table 9). After that, the plate was put in the thermal cycler (the Fast 9800 Thermal cycler) for doing the PCR to start the sequencing reaction (Table 10). When the reaction finished, the plate was spin done by using a centrifuge to ensure the total volume was in the bottom of the wells and not evaporated.

Table 9. Master Mix reagent that added to the post-PCR reaction product.

Reagents	**PCR Product**	**M13 Control**
PCR grade distilled water (Millipore Filtration System)	4µL	2µL
5X Sequencing Buffer	1µL	1µL
Ready Reaction Mix	2µL	2µL
Total of Master Mix	7µL	5µL

Table 10. PCR condition that set in the Fast 9800 thermal cycler.

Temperature	Time	Cycle
96°C	1min	1
96°C	10sec	25
50°C	5sec	
60°C	75sec	
4°C	∞	

After the reaction had been done, it stopped by using a BigDye XTerminator Purification Kit (Applied Biosystem Inc) that composed of SAM solution and Xterminator solution. First of all, all the solution must be put in the room temperature before starting the reaction, then vortex them well to avoid any precipitation of crystals (if crystals observed in the SAM solution, it must be placed the SAM bottle in 37°C water bath for 10 minutes or till the crystals completely dissolved). For each reaction, the Dye-Terminator premix solution had been prepared by adding 45µL of SAM solution with 10µL of Xterminator solution and multiplied by the number of reactions to know the exact volume to prepare as a premix. After that, the premix had been placed on the shaker to keep solution homogenous and it must always agitate before each aspiration.

55µl of premix was aspirated and loaded to each well at the plate. After adding to all wells on the plate, the plate had been sealed properly to avoid cross contamination with MicroAmp adhesive foil. The plate was put in the vortex machine for 30 minutes at 1800rpm to allow the reagent to be mixed properly with

the samples, and then it was centrifuged for 2 minutes at 2,800 rpm. Once the centrifuge step completed, 5µL of the supernatant of each well were transferred to a new optical plate, added to 10µL of HiDi (Applied Biosystem Inc.) for each reaction sample. Then the plate was covered with septa. After that, the plate had been mixed well by using a vortex, and then it was spin down by using centrifuge for 20 seconds. For Assembling the plate for loading in the sequencing machine, the plate was placed into the plate base, and checking that the holes of the plate retainer and the septa strips are aligned, if not, the plate was reassembled. Then the plate was ready for putting into the sequencing machine. The Sanger sequencing machine that was used: ABI 3130XL, and 3730XL Genetic Analyzers.

3.7 Calculation:

(Sock Concentration X volume) before = (concentration X volume) after

$$\text{Needed volume} = \frac{(concentration\ X\ volume)after}{\text{Sock Concentration}}$$

$$= \frac{2.5\ X\ 10}{10}$$

$$= 2.5$$

3.8 List of Materials:

NO	Materials	Manufacturer	Storage Temperature	Catalogue number
A	**MiSeq NGS Materials:**			
A.1	Flow Cell	Illumina	2°C to 8°C	RH-102-1001
A.2	Optical plate, 96 well	Applied Biosystems Inc	Room temperature	N8010560
A.3	Fast Reaction plate, 96 well	Applied Biosystems Inc	Room temperature	4346907
B	**PCR Material:**			
B.1	Optical Adhesive Covers	Applied Biosystems Inc	Room temperature	4311971
B.2	MicroAmp Clear Adhesive Films	Applied Biosystems Inc	Room temperature	4306311
B.3	Optical plate, 96 well	Applied Biosystems Inc	Room temperature	N8010560
B.4	Fast Reaction plate, 96 well	Applied Biosystems Inc	Room temperature	4346907
C	**Agarose Gel Electrophoresis Material**			
C.1	Mini-Sub® Cell GT Horizontal Electrophoresis System, 7 x 7 cm tray, with casting gates	Bio-Rad	Room temperature	1704406

D	**Sanger Sequencing Material:**			
D.1	Optical Adhesive Covers	Applied Biosystems Inc	Room temperature	4311971
D.2	MicroAmp Clear Adhesive Films	Applied Biosystems Inc	Room temperature	4306311
D.3	Optical plate, 96 well	Applied Biosystems Inc	Room temperature	N8010560
D.4	Fast Reaction plate, 96 well	Applied Biosystems Inc	Room temperature	4346907
D.5	Plate septa, 96 well	Applied Biosystems Inc	Room temperature	0410065073
C.6	Plate base, 96 well 3130xl	Applied Biosystems Inc	Room temperature	4317237
D.7	Plate base, 96 well 3730xl	Applied Biosystems Inc	Room temperature	4334873
D.8	Plate retainer, 96 well 3130xl	Applied Biosystems Inc	Room temperature	4317241
D.9	Plate retainer, 96 well 3730xl	Applied Biosystems Inc	Room temperature	4334868
D.10	MicroAmp 8-Cap Strip	Applied Biosystems Inc	Room temperature	N801-0535
D.11	MicroAmp Fast Reaction Tubes	Applied Biosystems Inc	Room temperature	4358293
D.12	MicroAmp 8-Tube Strip (0.2mL)	Applied Biosystems Inc	Room temperature	N8010580

3.9 List of Reagents:

NO	Reagent and Kits	Manufacturer	Storage Temperature	Catalogue number
A	**MiSeq NGS Reagents and Kits**			
A.1	TruSight Rapid Capture kit (48 samples)	Illumina	2°C to 8°C	FC-140-1104
A.2	Nextera Rapid Capture Custom Enrichment Kit (96 Samples)	Illumina	-15°C to -25°C	FC-140-1008
A.3	Trusight Rapid Capture sample preparation kit	Illumina	Room temperature	FC-1409003DOC
A.4	AMPure XP Beads	Beckman Coulter	2°C to 8°C	A63880
A.5	2N NaOH	Illumina	-15°C to -25°C	SS264-1
A.6	MiSeq Reagent Kit v3, 150 Cycles, Box1 and Box 2	Illumina	Box1: -15°C to -25°C. Box2: 2°C to 8°C	MS-102-3001
A.7	Ethanol	EMSURE	Room Temperature	K43504883223
A.8	TruSight Autism sequencing panel	Illumina	2°C to 8°C	FC-121-0203
A.9	TruSeq Dual Index Sequencing Primer Box, Single Read	Illumina	15°C to 30°C	FC-121-1003
B	**PCR Reagents and Kits**			
B.1	PCR grade distilled water	Roche	Room temperature	03315932001
B.2	dNTPs set (PCR grade)	Qiagen	-20°C	201912
B.3	taq polymerase enzyme (Hotstar)	Qiagen	-20°C	145041263
C	**Agarose Gel Reagents**			
C.1	UltraPure10x TBE Buffer	Gibco by life technologies	Room Temperature	15581-044
C.2	Agarose powder	Research Products International Corp.	Room temperature	9012-36-6
C.3	Ethidium Bromide	Gene Choice	Room temperature	5450
C.4	5x loading Dye, Gel Pilot	Qiagen	4 °C	142346877
C.5	100bp Plus Ladder, Gel Pilot	Qiagen	4 °C	19.4

D				
	Sanger Sequencing Reagents and kits			
D.1	M13 Template (pGM3Zf (+))	Applied Biosystem Inc.	-20°C	1307127
D.2	M13 Primer	Applied Biosystem Inc.	-20°C	182871673
D.3	Hi-Di formamide	Applied Biosystem Inc.	-20°C	1508392
D.4	BigDye XTerminator Purification Kit	Applied Biosystem Inc.	4°C	1509095
D.5	SAM solution	Applied Biosystem Inc.	4°C	509095
D.6	ExoSAP-IT (ExoSAP-IT™)	USB/Affymetrix	-20°C	4227968
D.7	BigDye Terminator v3.1 Cycle Sequencing Kit	Applied Biosystem Inc.	-20°C	4336943
D.8	Genetic Analyzer 10X Running Buffer with EDTA	Applied Biosystem Inc.	4°C	1506252
D.9	Polymer POP-7™ 3130 xl POP-7™ 3730 xl	Applied Biosystem Inc.	4°C	 4352759 4363929
D.10	Ready Reaction Mix	Applied Biosystem Inc.	-20°C	1309144
D.11	Sequencing Buffer (5X)	Applied Biosystem Inc.	4°C	1004119

3.10 List of Primers:

NO	Primer Name	Manufacturer	Storage Temperature	Reference number
1	FMR1_17F	IDT	Room temperature	70016950
2	FMR1_17R	IDT	Room temperature	70016951
3	RAI1_3F	IDT	Room temperature	70016952
4	RAI1_3R	IDT	Room temperature	70016953
5	ZNF804A_4F	IDT	Room temperature	70016954
6	ZNF804A_4R	IDT	Room temperature	70016955
7	CHD7_2F	IDT	Room temperature	70016956
8	CHD7_2R	IDT	Room temperature	70016957
9	RELN_25GF	IDT	Room temperature	70016958
10	RELN_25GR	IDT	Room temperature	70016959
11	PCDH9 _2F	IDT	Room temperature	70016960
12	PCDH9 _2R	IDT	Room temperature	70016961
13	EHMT1_5F	IDT	Room temperature	70016962
14	EHMT1_5R	IDT	Room temperature	70016963
15	Dmd_17R	IDT	Room temperature	70016964
16	Dmd_17F	IDT	Room temperature	70016965
17	MET_16F	IDT	Room temperature	70016966
18	MET_16R	IDT	Room temperature	70016967
19	Tsc2_41-42F	IDT	Room temperature	70016968
20	Tsc2_41-42R	IDT	Room temperature	70016969
21	ANKRD11_9F	IDT	Room temperature	70016970
22	ANKRD11_9R	IDT	Room temperature	70016971
23	CREBBP_31R	IDT	Room temperature	70016972
24	CREBBP_31F	IDT	Room temperature	70016973
25	SHANK2_10F	IDT	Room temperature	70016974
26	SHANK2_10R	IDT	Room temperature	70016975

27	AUTS2_19F	IDT	Room temperature	70016976
28	AUTS2_19R	IDT	Room temperature	70016977
29	RELN_13F	IDT	Room temperature	70016978
30	RELN_13R	IDT	Room temperature	70016979
31	LAMC3_8F	IDT	Room temperature	70016980
32	LAMC4_8R	IDT	Room temperature	70016981
33	CHD7_2F*	IDT	Room temperature	70016982
34	CHD7_2R*	IDT	Room temperature	70016983
35	SHANK2-3F	IDT	Room temperature	70016984
36	SHANK2-3R	IDT	Room temperature	70016985
37	CDKL5_12F	IDT	Room temperature	70312521
38	CDKL5_12R	IDT	Room temperature	70312522
39	AUTS2_8F	IDT	Room temperature	70312523
40	AUTS2_8R	IDT	Room temperature	70312524
41	PON3_9F	IDT	Room temperature	70312525
42	PON3_9R	IDT	Room temperature	70312526
43	KIRREL3_3F	IDT	Room temperature	70312527
44	KIRREL3_3R	IDT	Room temperature	70312528
45	ATRX_8F	IDT	Room temperature	70312529
46	ATRX_8R	IDT	Room temperature	70312530

4. Results:

4.1 Analysis of the four patients:

Four ASD patients have been analyzed by using MiSeq platform according to the pipeline that have been described before, and result in having enrichment report for each sample as well as VCF file. The enrichment report containing all details related to the quality of the data released from the MiSeq and the coverage depth. We depend on the 20X coverage with assembly range of (95.7-97.5%) among the four patients. Within 20X coverage, we compare the four patients in the case of targeted bases covered at depth, total targeted based covered, and the target coverage. For the number of targeted bases covered at depth, it was (911, 580, 332, and 437) respectively among the four patients. While the Total targeted based covered for each patient it was as follows respectively (316061, 320822, 322027, and 320756). The target coverage for the patients was as follows: 95.7% for the first patient, 97.1% for the second patient, 97.5% for the third patient, and 97.1% for the fourth patient (Table 11).

In addition, we compared between the four patients in the level of the targeted fragment. The fragment length median varies among the patients as follows: 242bp, 196bp, 226bp, and 208bp respectively. The minimum and the maximum length of the targeted fragment in each patient were as follows: (95bp, 714bp) for the first patient, (55bp, 494bp) for the second patient, (83bp, 641bp) for the third patient, and (74bp, 563bp) for the fourth patient. Moreover, the enrichment report calculated the Standard deviation for each patient, 95, 60, 83, and 69 respectively for the patients (Table 12).

Table 11. Comparison among the four patients in the coverage of the variants.

Patients	Depth of sequencing coverage	Number of targeted bases covered at depth	Total targeted based covered	Target coverage
Patient #1	20X	911	316,061	95.7%
Patient #2	20X	580	320,822	97.1%
Patient #3	20X	332	322,027	97.5%
Patient #4	20X	437	320,756	97.1%

Table12. Comparison between the four patients in the targeted fragment length, minimum, maximum and standard deviation.

Patients	Fragment Length Median	Minimum	Maximum	Standard deviation
Patient #1	242 bp	95 bp	714 bp	95 bp
Patient #2	196 bp	55 bp	494 bp	60 bp
Patient #3	226 bp	83 bp	641 bp	83 bp
Patient #4	208 bp	74 bp	563 bp	69 bp

4.2 Applying the hypothesized pipeline:

For each patient, there is a list of passing and failing variants according to the MiSeq reporter criteria in the VCF. The hypothesized filtering method was applied to the passing variants (high quality variants) of the VCF for the each patient in order to be able to identify the variants. After using the hypothesized filtering criteria, the number of passing variants reduced dramatically in each patient as follows: in the first patient, the number of passing variants was 100 variants, while after filtering it become 5 variants only. The passing variants encountered in the second patient were 83 variants that become 4 variants after filtering. For the third patient, it was 110 then after filtering became 4 variants, while the number of passing variants for the last patient was 108 variants, which become 5 variants after filtering (Table 13). The list of passing variants in each patient are as follows: (FMR1, RAI1, ZNF804A, CHD7, CREBBP) in patients #1, (RELN-25e, CHD7*, SHANK2, and PCDH9) in patient #2, (TSC2, DMD, MET, SHANK 2-3e) in patient#3, and (AUTS2, RELN-13e, LAMC3, ANKRD11, and EHMT1) in patient #4. (Table 14,Table15, Table 16, Table 17). For the variants that was chosen randomly among the 4 patients which failed based on the hypothesized criteria while passed based on the MiSeq criteria (low quality variants), were (CDKL5, and AUTS2 in patient #1, PON3 in patient #2, KIRREL3 in patient #3,and ATRX, in patient #4) (Table 18).

Table 13. Comparison among the four patients in the pass and fail mutation before and after filtering.

Patients	No# of mutated genes before filtering	Pass variants	Fail variants	No# of passing mutated genes after filtering	Fail variants after filtering
Patient #1	121	100	21	5	95
Patient #2	93	83	10	4	79
Patient #3	122	110	12	4	106
Patient #4	129	108	11	5	103

Table 14. List of passing variants of the patient #1. The list explains the chromosomes, type of variant, genotyping, and the allele frequency.

Genes	Variant	Chromosome	Type	Genotype	Allele Frequency
FMR1	c.*746T>C	X	snv	Homozygous	0.18
RAI1	c.840delG	17	Deletion	Heterozygous	0
ZNF804A	c.2088_2089insACA	2	Insertion	Heterozygous	0
CHD7	c.389A>G	8	snv	Heterozygous	0
CREBBP	c.6542A>G	16	snv	Heterozygous	0

Table 15. List of passing variants of the patient #2. The list explains the chromosomes, type of variant, genotyping, and the allele frequency.

Genes	Variant	Chromosome	Type	Genotype	Allele Frequency
RELN-25e	c.3477C>A	7	snv	Heterozygous	0.14
PCDH9	c.2866G>C	13	snv	Homozygous	0
CHD7*	c.127A>G	8	snv	Heterozygous	0
SHANK2	c.2397_2402dupGCCATT	11	Insertion	Heterozygous	0

Table 16. List of passing variants of the patient #3. The list explains the chromosomes, type of variant, genotyping, and the allele frequency.

Genes	Variant	Chromosome	Type	Genotype	Allele Frequency
TSC2	c.5389A>T	16	snv	Heterozygous	0
SHANK2_3e	c.332 G>A	11	snv	Heterozygous	0
DMD	c.2143A>T	X	snv	Homozygous	0.36
MET	c.3497G>T	7	snv	Heterozygous	0

Table 17. List of passing variants of the patient #4. The list explains the chromosomes, type of variant, genotyping, and the allele frequency.

Genes	Variant	Chromosome	Type	Genotype	Allele Frequency
AUTS2	c.3374_3375insCCACCA	7	Insertion	Heterozygous	0
RELN_13e	c.1483A>G	7	snv	Heterozygous	0
LAMC3	c.1414C>A	9	snv	Heterozygous	0
EHMT1	c.905A>G	9	snv	Heterozygous	0
ANKRD11	c.1130C>T	16	snv	Heterozygous	0

Table 18. Low quality list of variants in all patients that were chosen randomly.

Patients	Genes	Variant	Chromosome	Type	Genotype
Patient# 1	AUTS2	c.1343T>C	7	snv	Heterozygous
	CDKL5	c.1324C>A	X	snv	Heterozygous
Patient# 2	PON3	c.939G>T	7	snv	Heterozygous
Patient# 3	KIRREL3	c.263G>T	11	snv	Heterozygous
Patient# 4	ATRX_8	c.957G>T	X	snv	Heterozygous

4.3 The effect of the hypothesized filtering criteria:

The effect of hypothesized filtering criteria shown clearly in reducing the number of passing variants that needs validation by Sanger sequencing. This will lead to decrease the error rate and help in distinguishing the real variants from false positive and/or false negative (Figure 5).

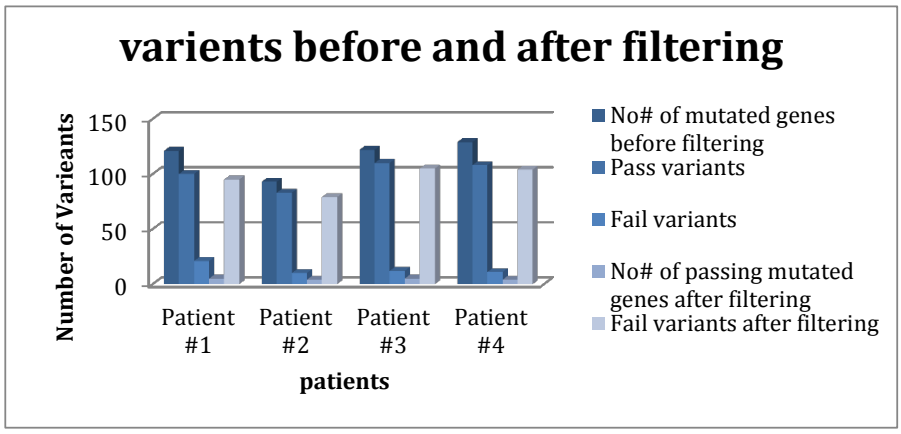


Figure 5. Variants that resulted before and after filtering criteria. It shows the effect of filtering criteria, which reduce the number of the passing variants significantly. (100 passing variants become 5 for patient#1, 83 passing variants become 6 after filtering for patients #2, 110 passing variants become 4 after filtering for patient#3, and 108 passing variants become 6 after filtering for patient#4).

After filtering we end up with a total of 18 variants among the four patients (5 variants in the first patient, 4 variants in the second patient, 4 variants in the third patient, and 5 variants in the fourth patient).

4.4 Evaluation of Variants in the four patients:

4.4.1 Primer Design:

Once the primers had been designed, 2 lists of variant primers were created. The first list includes the high quality variants (Table 19) and the second list includes the low quality variants (Table 20). Each list includes some properties such as: the primer name, the sequence of each primer, product size, Tm, and primer length.

Table 19. Primers list for the high quality variants among the 4 patients.

Primer Name (gene_exon_amplicon)	Primer Sequence	TM (°C)	Length (bp)	Product size (bp)
FMR1_17F	TGTAGCAAACCCTGTCAAAC	52	20	218
FMR1_17R	GCCTCTCTCAGATGAAGATAGTT	53	23	
RAI1_3F	CTCAGCATTCCCAGTCCTTC	54	20	496
RAI1_3R	TGCTGGCTGTAGGGAAAGTT	55	20	
ZNF804A_4F	AAAGGCCCAAATCAGAATCC	52	20	447
ZNF804A_4R	TGCAGGGAGCAACTGAAGTA	55	20	
CHD7_2F	CCGAACAGAATGATGAGCAA	52	20	440
CHD7_2R	GGTCCTGAGGTGGCAATAAA	54	20	
RELN_25GF	ggacgagaccatttccacat	54	20	633
RELN_25GR	cctgcaaatgggaaaatgag	51	20	
PCDH9 _2F	CCTGGCCAAGCACTACAAAT	55	20	190
PCDH9 _2R	ACTGCACTCTGAGGCACTGA	57	20	
EHMT1_5F	ggtttgctggtgattttggt	53	20	423
EHMT1_5R	acgaacactctgcccctatg	56	20	
Dmd_17R	CCACCACTCAGCCATCACTA	55	20	232
Dmd_17F	caccaccaacaaaactgct	53	19	
MET_16F	ggtggcatcattcactcaga	54	20	646
MET_16R	gtgcacagtttctggcacat	56	20	
Tsc2_41-42F	CCCTGCACGCAAATGTGA	55	18	788
Tsc2_41-42R	tggaggaaggtgactgtgtg	55	20	
ANKRD11_9F	AAGAACCCAGAGCCACAGAA	55	20	343
ANKRD11_9R	TGGCATTAGAAGGCTCTCGT	55	20	
CREBBP_31R	AGCCTGCAGAACCTGAATG	54	19	295
CREBBP_31F	CTTGAGGCTGCTGGAACTG	55	19	
SHANK2_10F	TGCAGGAAGAGGACGAGAAG	56	20	384
SHANK2_10R	GTTGGCTGGAGTTCAACGAAG	56	21	
AUTS2_19F	GATCCTTACCGAGAACTTGACATT	55	24	409
AUTS2_19R	GAGGGGTCTTGTTGAGGAGT	55	20	
RELN_13F	aatagctactttggccttcac	53	21	553
RELN_13R	tcactcttccttcctgataactc	53	23	
LAMC3_8F	ctgtggtaatgtcagatgtca	52	21	687
LAMC4_8R	ctgaatccccagtgtttagac	53	21	
CHD7_2F*	TGAAGTGAAGCACAGGCAA	54	19	642
CHD7_2R*	GCCATATAGCTGCCCATCTG	54	20	
SHANK2-3F	gcgtcctctgtcactcatca	56	20	250
SHANK2-3R	tacCTCCAGGGAAGGAAC	51	18	

Table 20. Primers list for the low quality variants among the 4 patients.

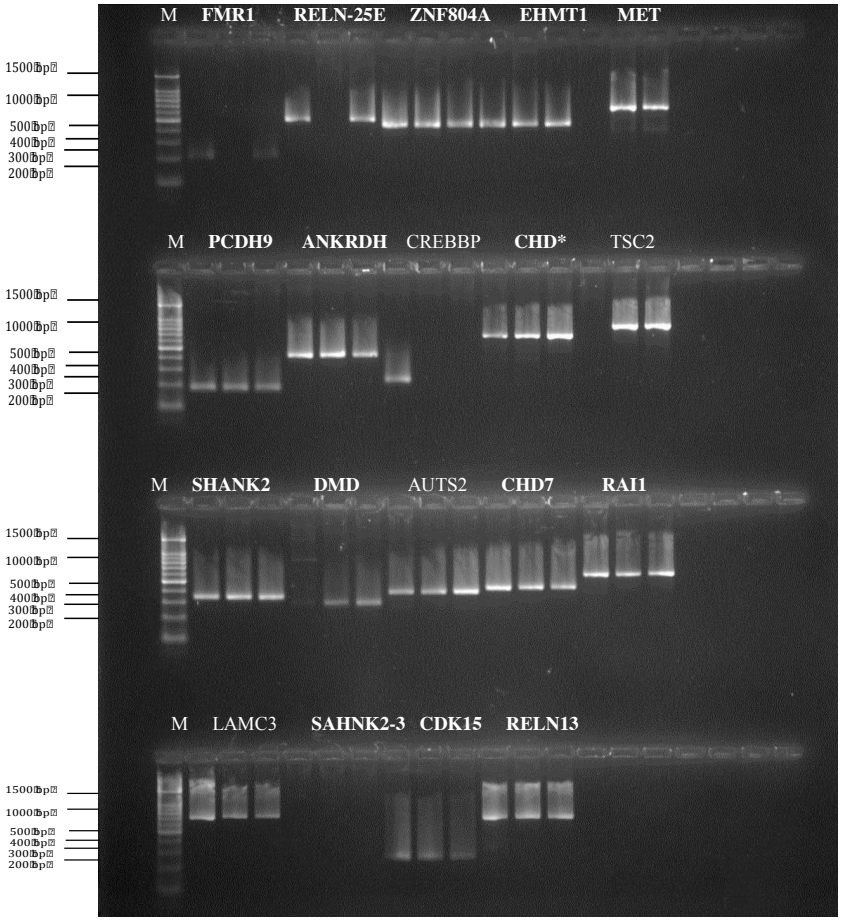
Primer Name (gene_exon_amplicon)	Primer Sequence	TM (°C)	Length (bp)	product size (bp)
CDKL5_12F	CCACACCTTCTTAGCCCAAA	20	54	195
CDKL5_12R	ATGCCGACTCTGCTTTTCA	19	54	
AUTS2_8F	gatagCAGCAGCAGAAGCAG	20	55	229
AUTS2_8R	CCACCTGCAGACTTCCTGAT	20	55	
PON3_9F	gaatgtttgggaaggacatca	21	52	210
PON3_9R	TGTGAAATACGGTGCCTATGA	21	53	
KIRREL3_3F	CAGCCAGTGACGCTACTTTG	20	55	168
KIRREL3_3R	ctaggaaggtggatgggt	18	51	
ATRX_8F	GAACAGTTGTTGCAGCAAA	19	51	233
ATRX_8R	TTGGCTGTGGTCTCAATCA	19	53	

4.4.2 Results of Primers optimization:

After receiving the primers, we did optimization of them by using different temperature degree in order to see the best condition to run the PCR for the patient, which will be proceeded to Sanger sequencing. The optimization condition was done on the temperatures degrees (52°C, 58°C, and 60°C) with 20ng/ml DNA (Figure 6). We choose these specific temperatures because they are the recommended temperatures that were given by the Amplifix software that was used when designing the primers. All the prepared primers lay within this range. After doing the first optimizing PCR, we found lots of non-specific binding bands, and this could be due to high concentration of the DNA template added. Because of that we reduced the concentration of the DNA to the half (10ng/ml) of its original concentration and repeat the PCR with the same mentioned condition except the concentration of the DNA that was reduced (Figure 7).

After that we choose the 58°C as the recommended annealing temperature in PCR that gave the band of the gene of interest shown and repeat the picture for the 23 genes (18 variants and 5 variants) (Figure 8). Once the optimized condition has been done, PCR had been done on the four patients and a normal control sample by using the previous optimized condition for each variant detected in each patient (Figure 9).

a.



b.

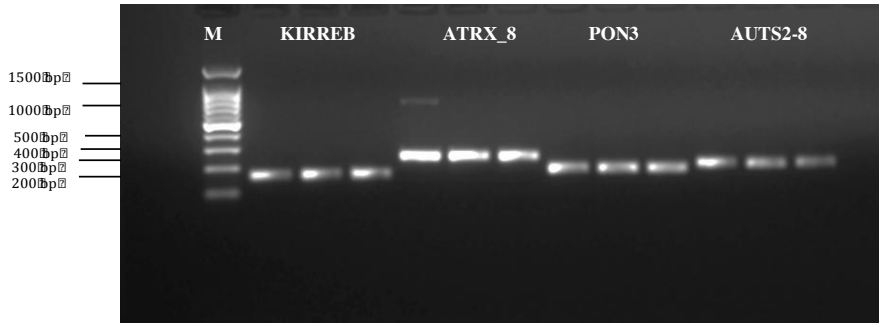
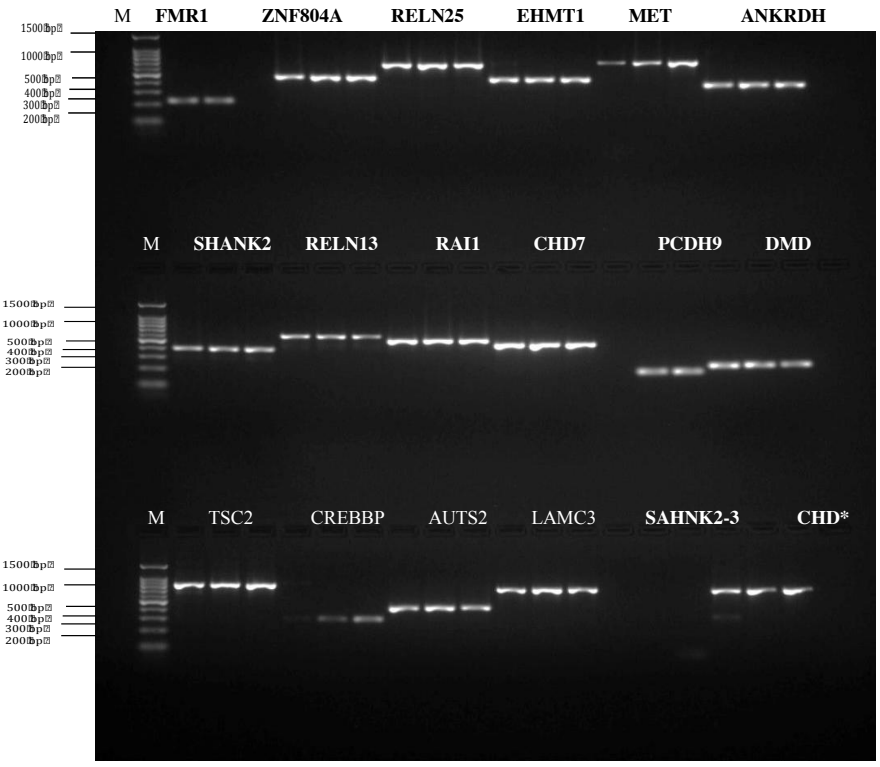


Figure 6. Agarose gel electrophoresis of PCR product represent the optimization of the variant's primers using 20ng/µl DNA concentration with gradient temperatures degrees (52°C, 58°C, and 60°C) respectively starting from the DNA ladder (M), GelPilot 100bp Plus Ladder (100-1500 bp) had been used as a DNA ladder, each 3 lanes represent one variant. a. PCR product of optimized condition of high quality variants among the 4 patients. **b.** PCR product of the optimized condition of low quality variants among the 4 patients.

a.



b.

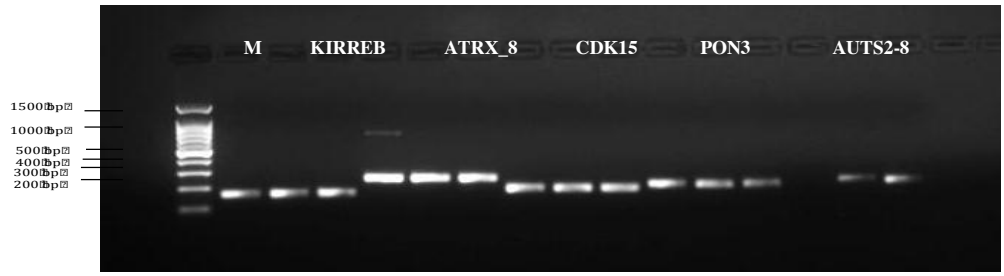
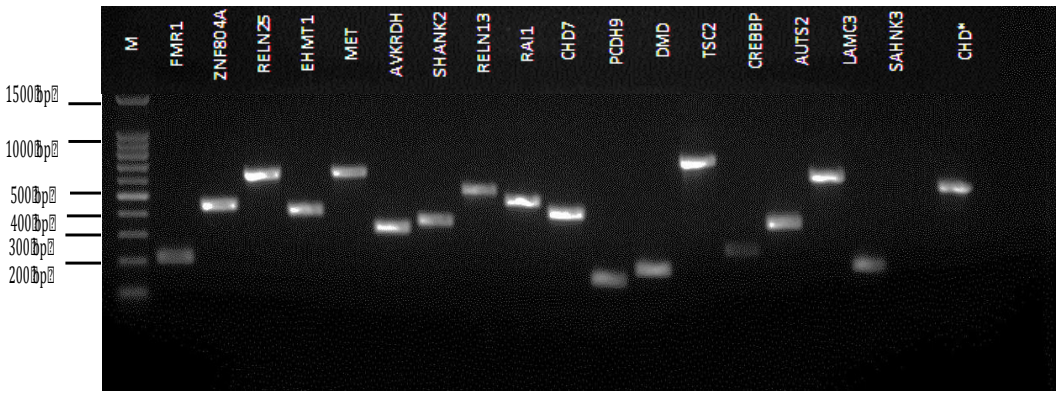


Figure 7. Agarose gel electrophoresis of PCR product represent the optimization of the variant's primers using 10ng/µl DNA concentration with gradient temperature degrees (52°C, 58°C, and 60°C) respectively starting from the DNA ladder (M), GelPilot 100bp Plus Ladder (100-1500 bp) had been used as a DNA ladder, each 3 lanes represent one variant. a. PCR product of optimized condition of high quality variants among the 4 patients. b. PCR product of the optimized condition of low quality variants among the 4 patients.

a.



b.

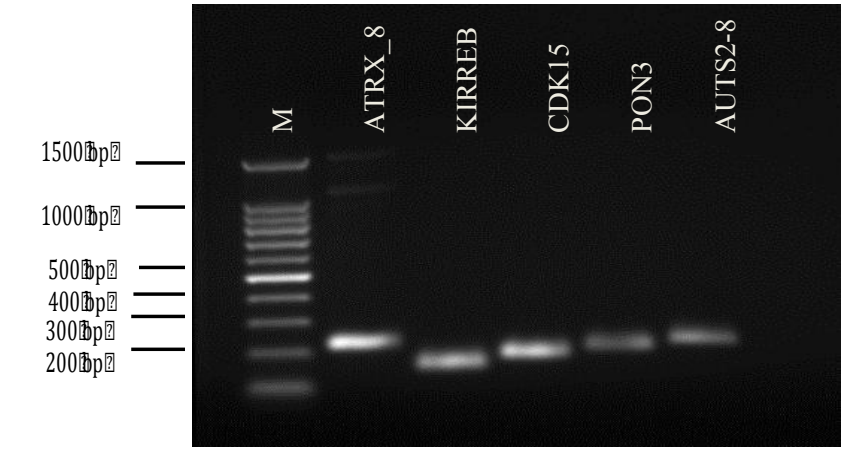
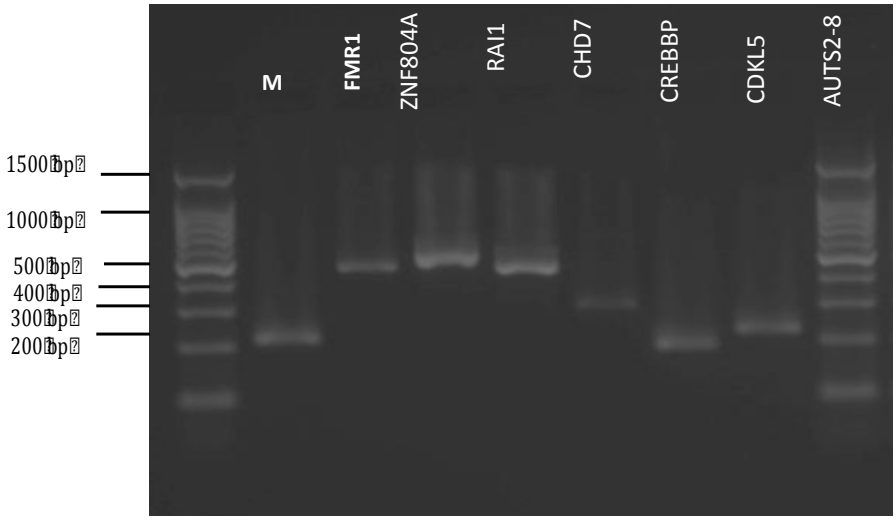
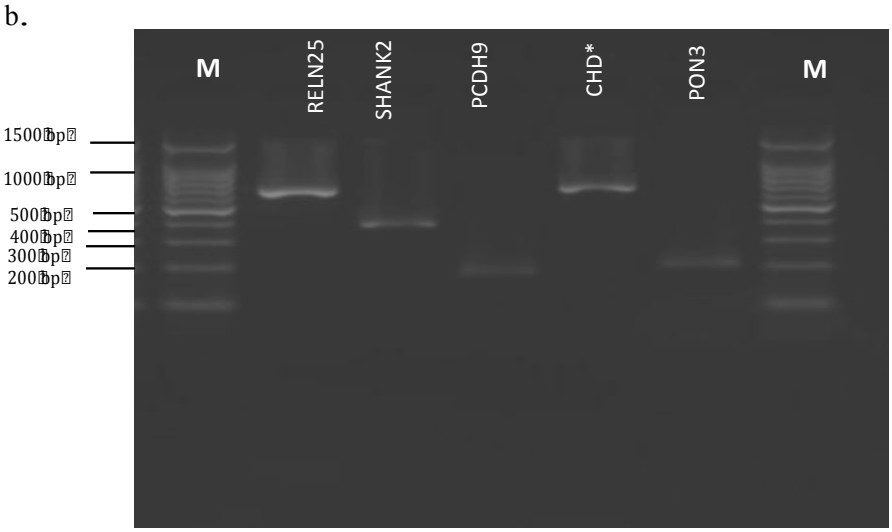


Figure 8. Agarose gel electrophoresis of PCR product represent the optimized condition of the variant's primers using 10ng/µl DNA concentration with temperature degree 58°C. (M) Represent DNA ladder GelPilot 100bp Plus Ladder (100-1500 bp), and each lanes represent one variant. **a**. PCR product of optimized condition of high quality variants among the 4 patients. **b**. PCR product of the optimized condition of low quality variants among the 4 patients.

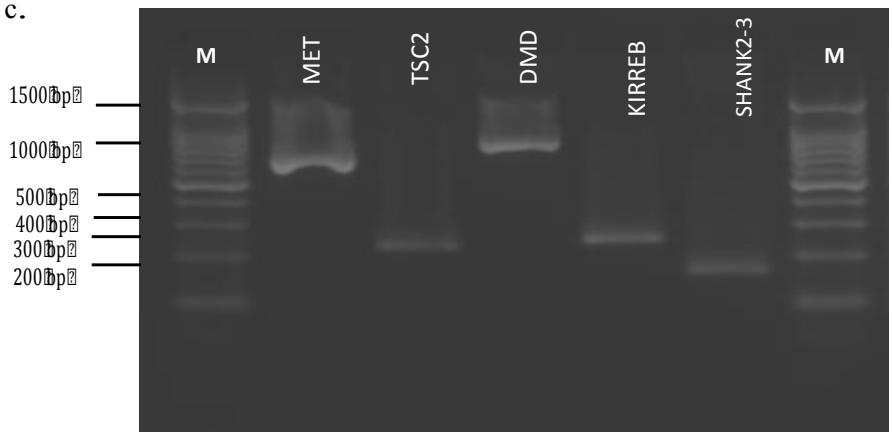
a.



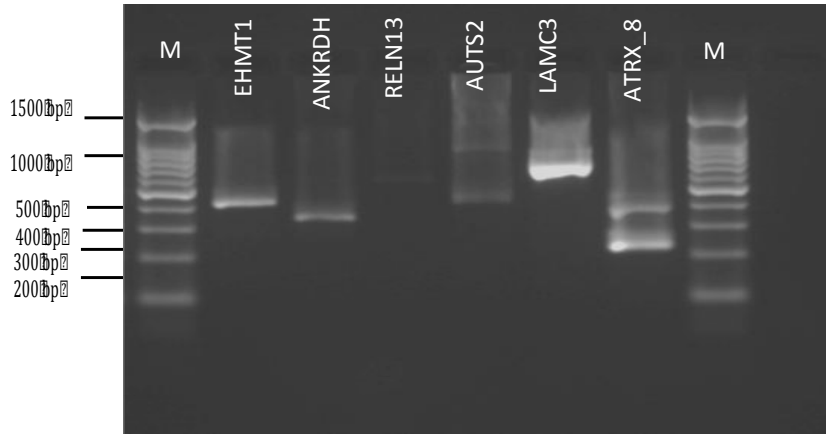
b.


M
RELN25
SHANK2
PCDH9
CHD*
PON3
M
1500bp
1000bp
500bp
400bp
300bp
200bp

c.


M
MET
TSC2
DMD
KIRREB
SHANK2-3
M
1500bp
1000bp
500bp
400bp
300bp
200bp

d.


M
EHMT1
ANKRDH
RELN13
AUTS2
LAMC3
ATRX_8
M
1500bp
1000bp
500bp
400bp
300bp
200bp

e.

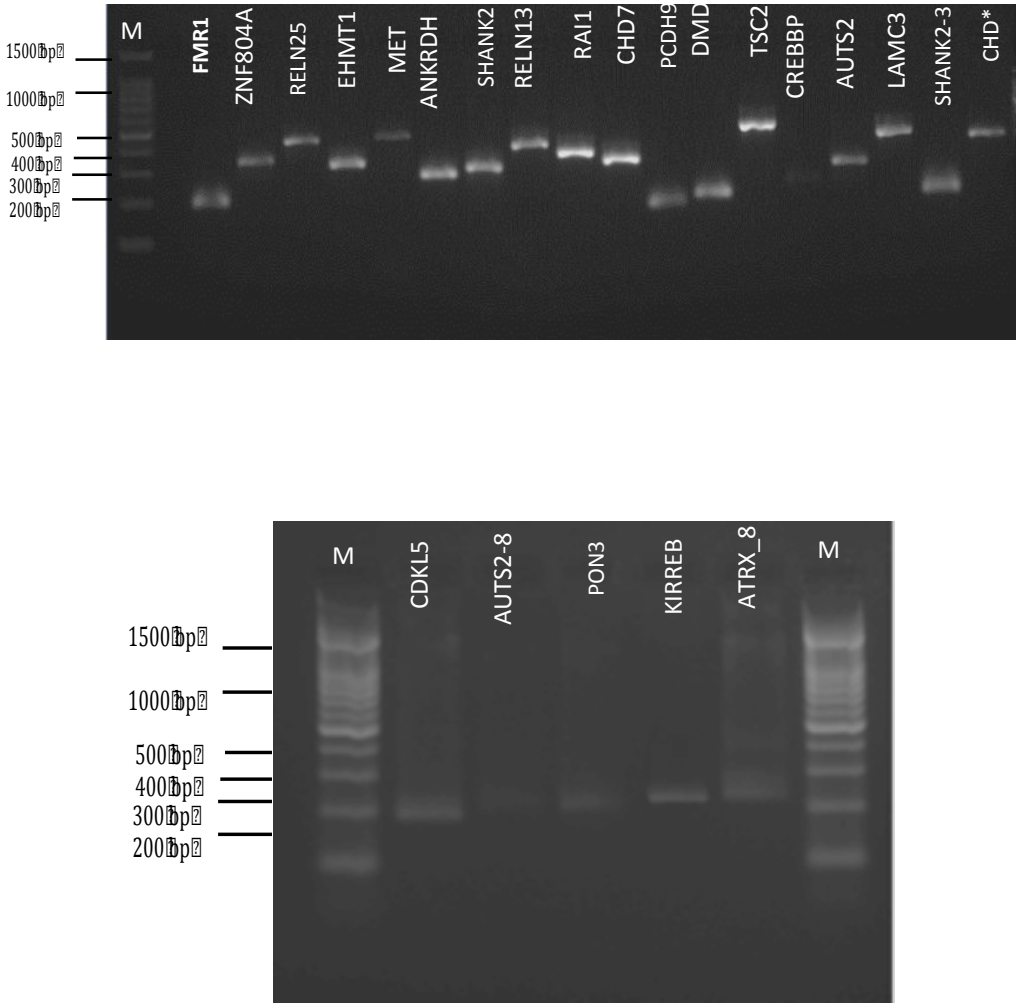


Figure 9. Agarose gel electrophoresis of PCR product represent the variant's detected among the 4 patients using 10ng/µl DNA concentration with temperature degree 58°C. (M) Represent DNA ladder GelPilot 100bp Plus Ladder (100-1500 bp), and each lanes represent one variant. **a**. PCR product for the variants that detected in patient #1. **b**. PCR product for the variants that detected in patient #2.c. PCR product for the variants that detected in patient #3. d. PCR product for the variants that detected in patient #4.e. PCR product for the variants in the normal control sample.

4.5 Identification of SNVs and In-dels in each patient:

After applying the hypothesized pipeline, we end up with a list of SNVs and In-dels in each patient. Patient#1 had 3 SNVs (point mutation) in (FMR1, CHD7, and CERBBP) and 2 In-dels (Insertion) in (ZNF804A, and RAI1), while there are 3 SNVs (point mutation) in (RELN-25e, PCDH9, and CHD*) and 1 In-del (SHANK2) detected in patient #2 .in patient #3 and #4 there was no any In-dels detected, on the other hand , all the variants that were detected was SNVs (point mutation) that are present as follows; (MET, Dmd, Tsc2, and SHANK2-3) in patient #3, and (RELN-13e, LAMC3, EHMT1, and ANKRD11) in patient #4 .(Figure 10)

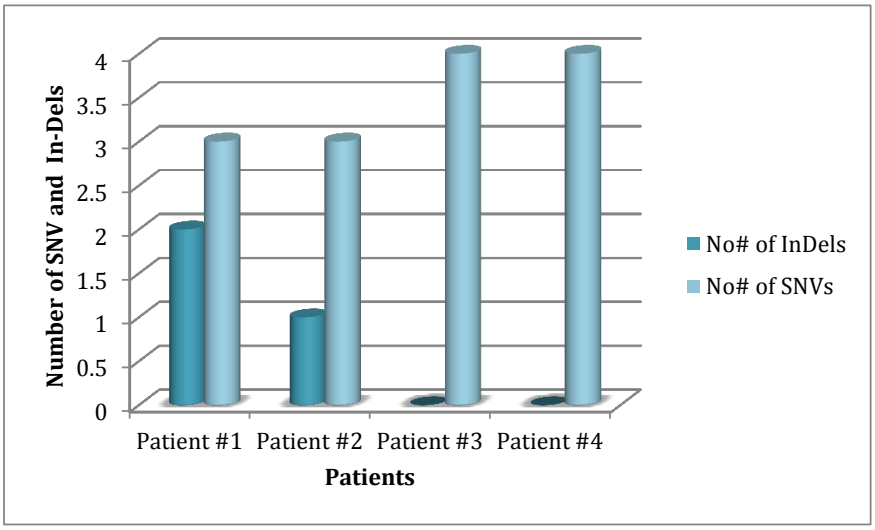


Figure 10. Chart that shows the numbers and types (SNVs or In-dels) of the variants that were detected after applying the hypothesized criteria on the four patients. Patient #1 (3 SNV and 2 In-Dels), patient #2(3 SNV and 1 In-Del), patient #3 (4 SNV), and patient #4 (4 SNV).

4.6 Validation by Sanger sequencing:

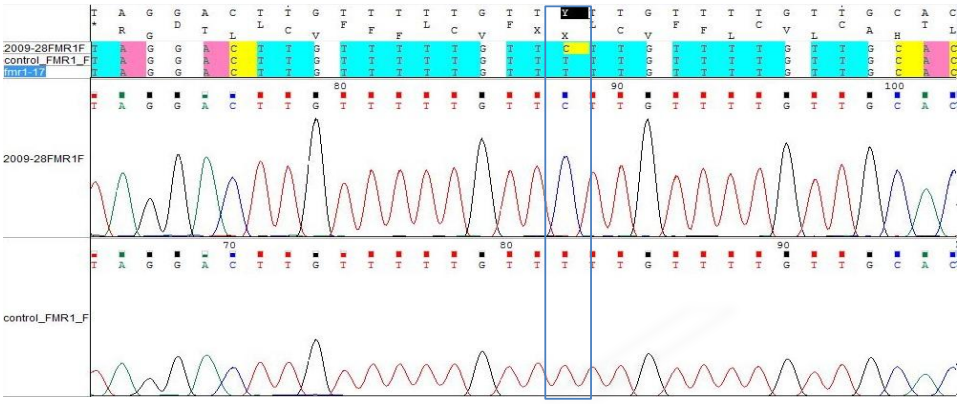
Once the condition has been optimized for the primers detecting the variants calls, Sanger sequencing analysis had been used in order to validate the presence or absence of the variants according to the hypothesized criteria.

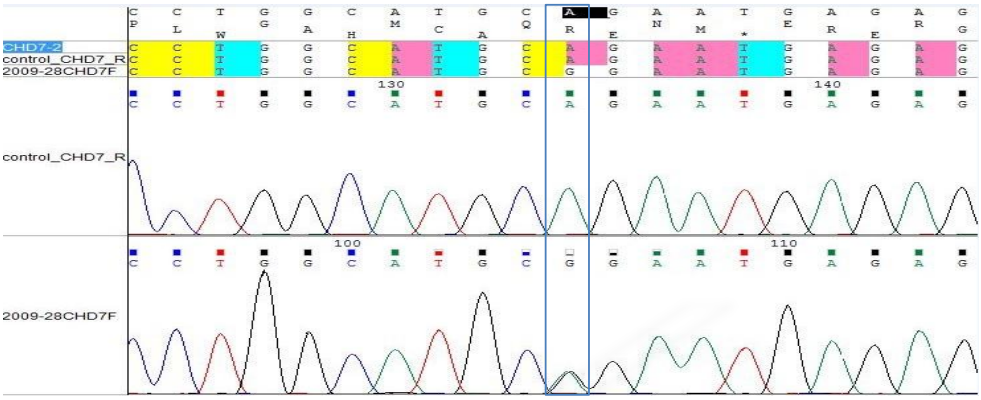
Each patient had a list of variants that had been processed by Sanger sequencing, some variants could not be sequenced due to some sequencing problems, while most of them were validated. For validation, a control sample was sequenced to be used as a reference in population along with the reference used by the software (UCSU genome browser) for more conformational analysis. Although Forward and reverse primers for each variant were employed in the sequencing to give high chance of detection and conformation of the presence or absence of the variation that had been called, one direction was used as a detection method for alignment to the reference and control sample.

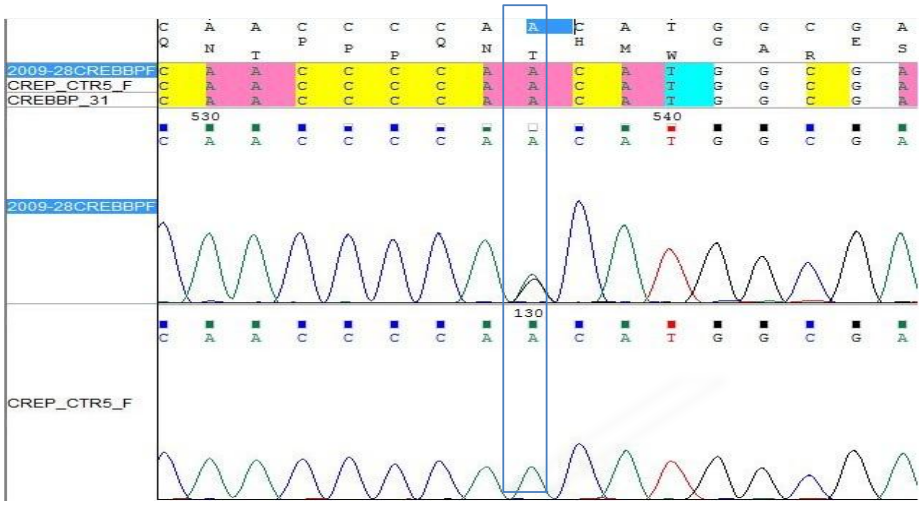
The list of variants that had been generated after doing the sequencing for the first patient was (FMR1, RAI1, ZNF804A, CREBBP, and CHD7) as the high quality variant while (CDKL5) was the only variant that had been sequenced as the low quality variant (Figure 11). (CHD7*, PCDH9, RELN-25e, and SHANK2) sequencing analysis was generated for the second patient as the high quality variant, and (PON3) as the low quality variant (Figure 12). For the third patient, (Tsc2, Dmd, SHANK2-3, and MET) were generated as the high quality variant, on the other hand, (KIRREL3) was sequenced as low quality variant (Figure 13). The variants that were analyzed by Sanger sequencing for the fourth patient were (RELN-13e, LAMC3, EHMT1, and ANKRD11) as high quality variants, while (ATRX_8) was analyzed as low quality variant (Figure 14). So, (AUTS2, and AUTS2_8) were excluded from patient #4,

patient# 1 respectively because we did not get the correct sequence.

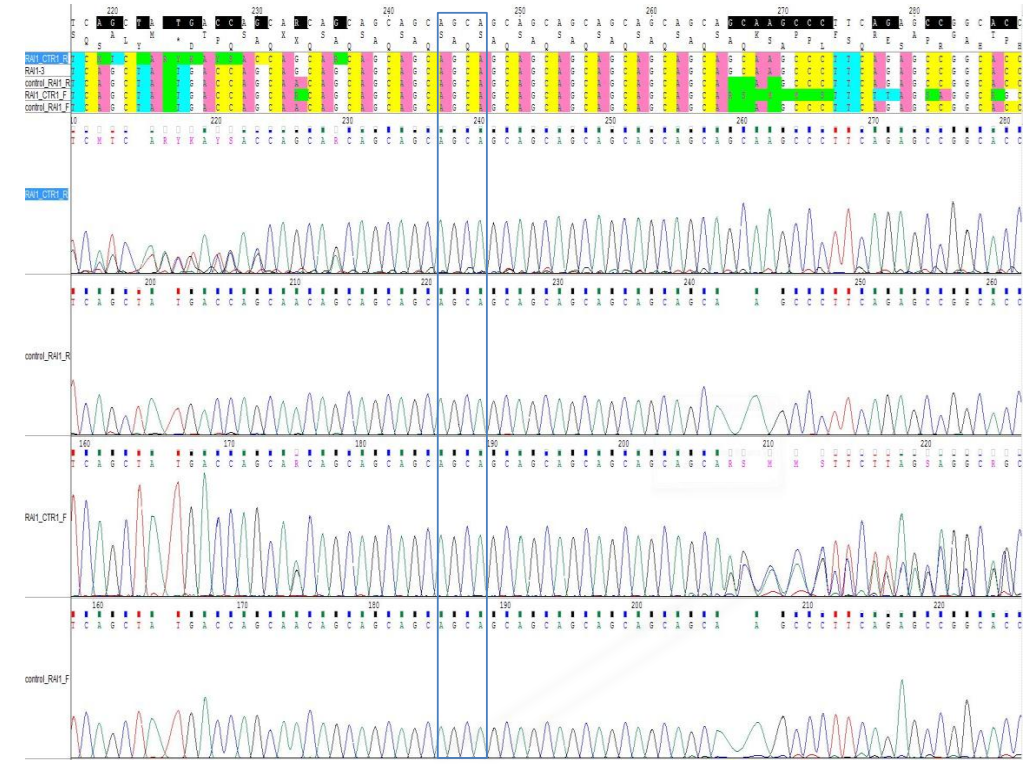
a.

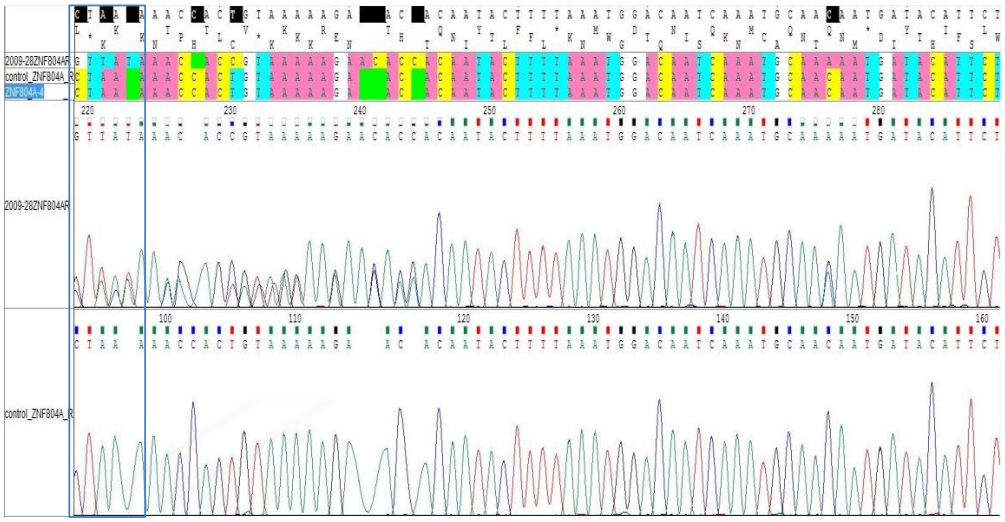






b.


R4/1_CTR1_R
R4/1-3
control_R4/1_R
R4/1_CTR1_F
control_R4/1_F
control_R4/1_R
R4/1_CTR1_F
control_R4/1_F


2009-28ZNF804AR
control_ZNF804A_R
ZNF804A-4
2009-28ZNF804AR
control_ZNF804A_R

c.

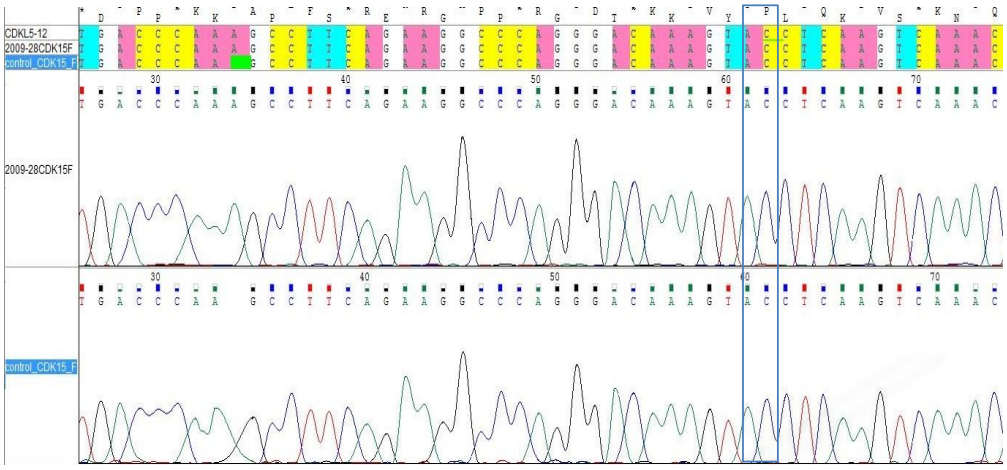
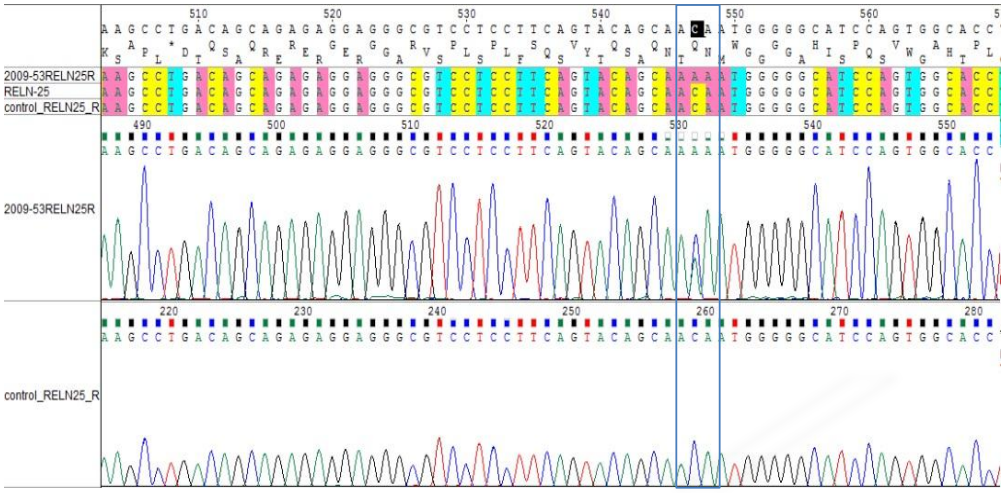
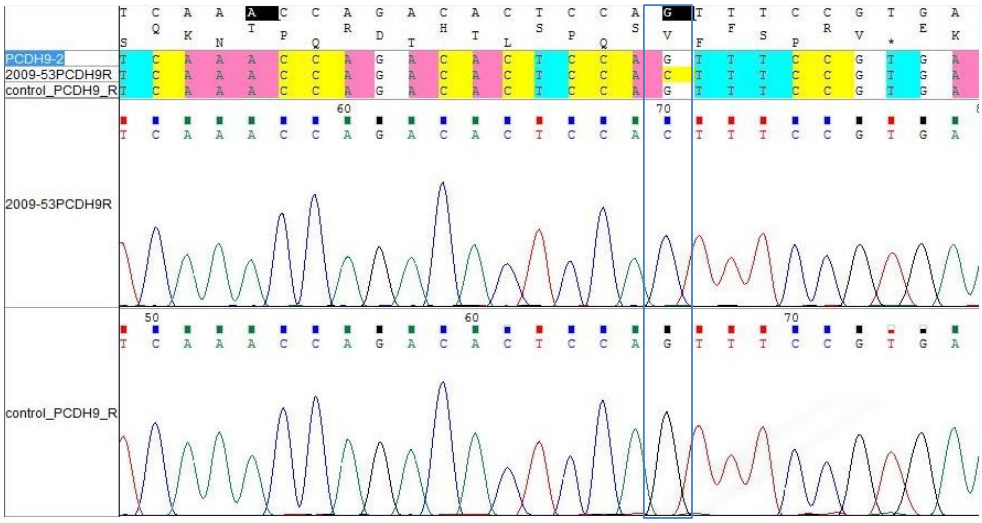
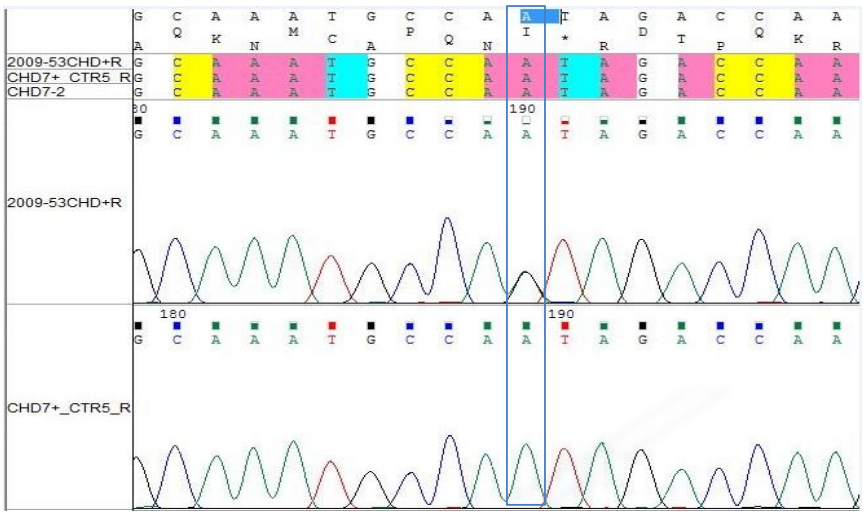


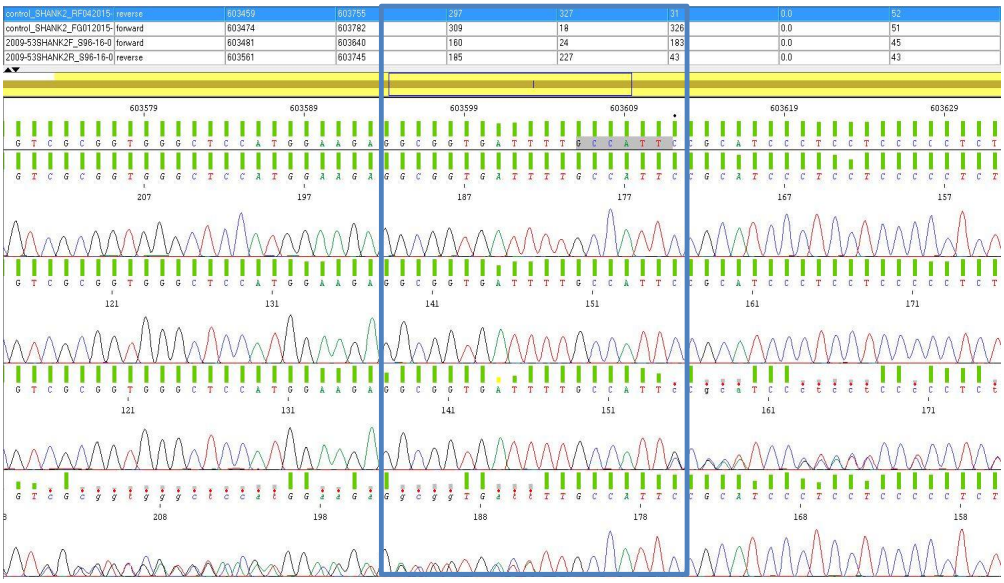
Figure 11. Chromatogram pattering of the variants called in patient #1 done through Sanger sequencing aligned with reference of each variant taken from (UCSC genome browser) and a one direction (forward or reverse) normal control sample. a. Chromatogram pattern for the variant called as high quality with point mutation (FMR1, CHD7, and CREBBP) respectively. b. Chromatogram pattern for the variant called as high quality with in-dels (insertion or deletion)(RAI1, and ZNF804A) respectively. c. Chromatogram pattern for the variant called as low quality without mutation (normal) (CDKL5).

a.


2009-53RELN25R
RELN-25
control_RELN25_R
2009-53RELN25R
control_RELN25_R


PCDH9-2
2009-53PCDH9R
control_PCDH9_R
2009-53PCDH9R
control_PCDH9_R


2009-53CHD+R
CHD7+_CTR5_R
CHD7-2
2009-53CHD+R
CHD7+_CTR5_R
180
190



c.

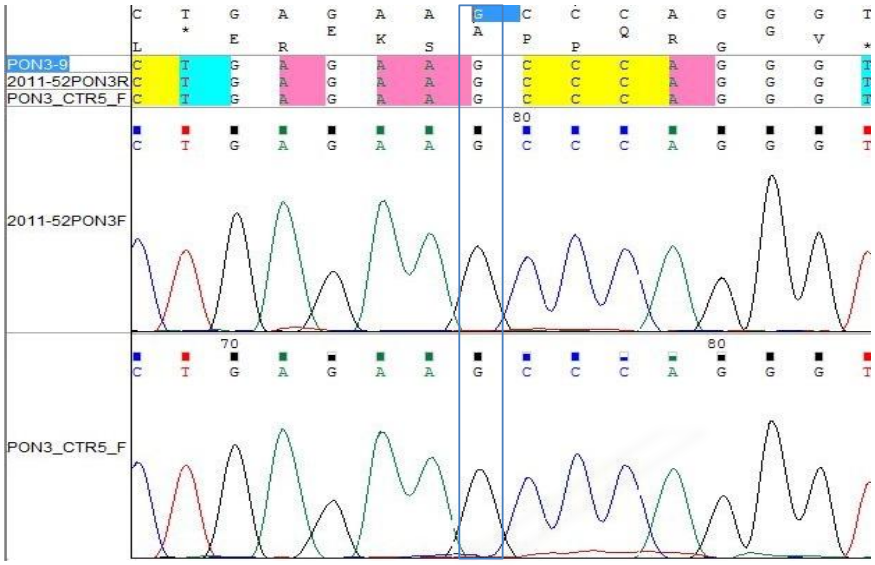
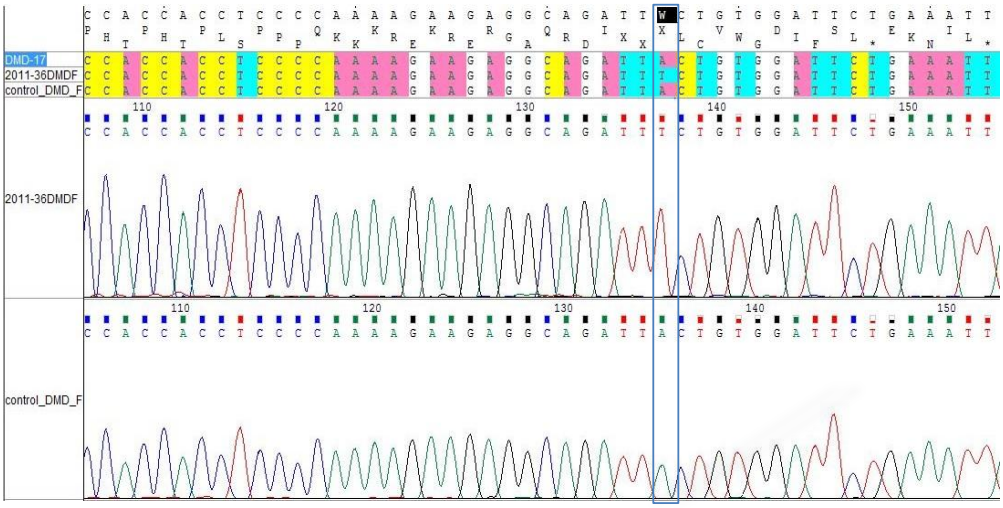
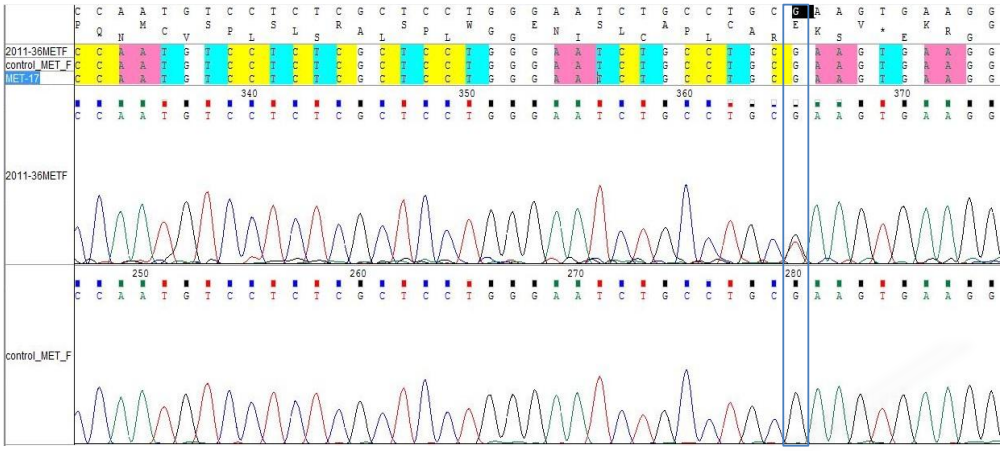
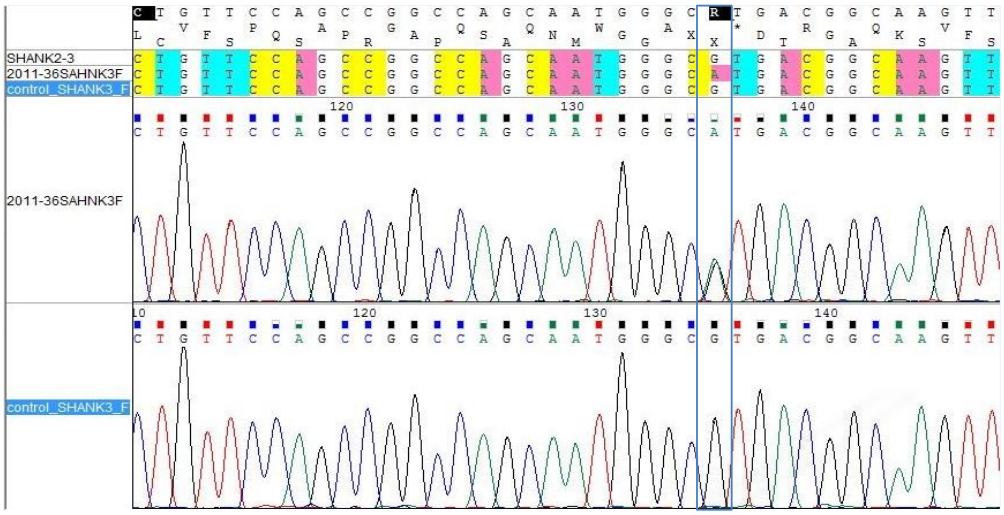


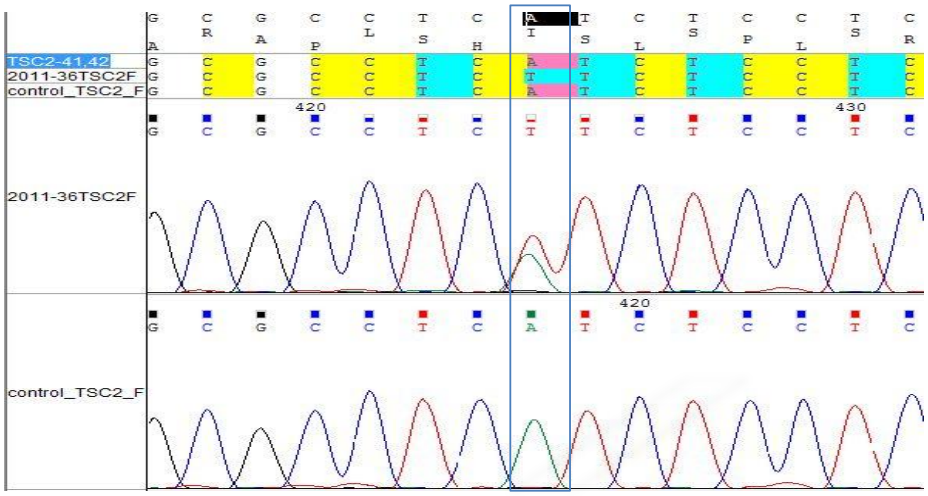
Figure 12. Chromatogram pattering of the variants called in patient #2 done through Sanger sequencing aligned with reference of each variant taken from (UCSC genome browser) and a one direction (forward or reverse) normal control sample. a. Chromatogram pattern for the variant called as high quality with point mutation (RELN-25e, CHD7*, and PCDH5) respectively. b. Chromatogram pattern for the variant called as high quality with in-dels (insertion or deletion)(SHANK2). c. Chromatogram pattern for the variant called as low quality without mutation (normal) (PON3).

a.




2011-36METF
control_MET_F
MET-17
2011-36METF
control_MET_F


SHANK2-3
2011-36SAHNK3F
control_SHANK3_F
2011-36SAHNK3F
control_SHANK3_F


TSC2-41,42
2011-36TSC2F
control_TSC2_F
2011-36TSC2F
control_TSC2_F

b.

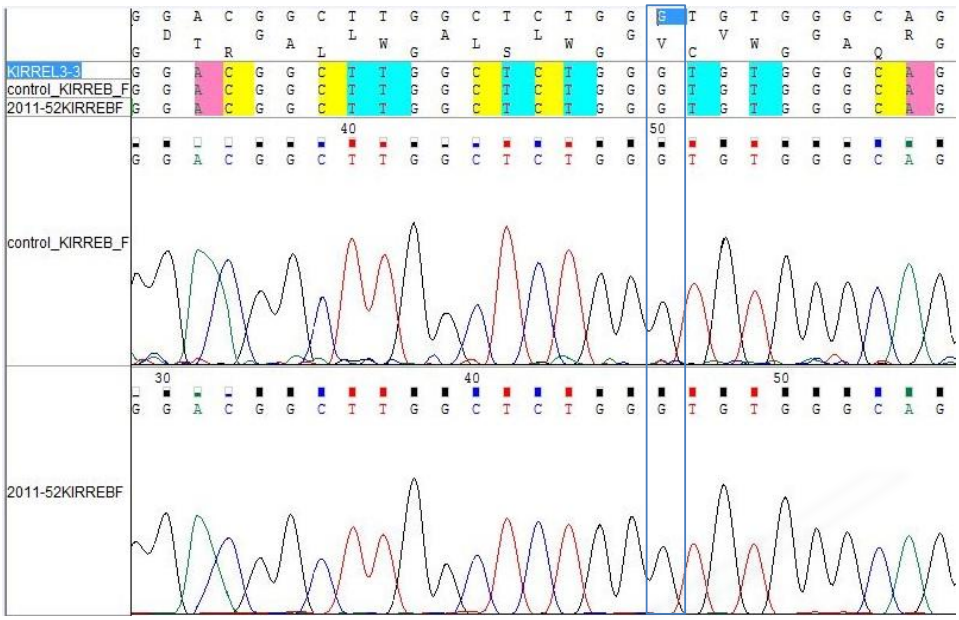
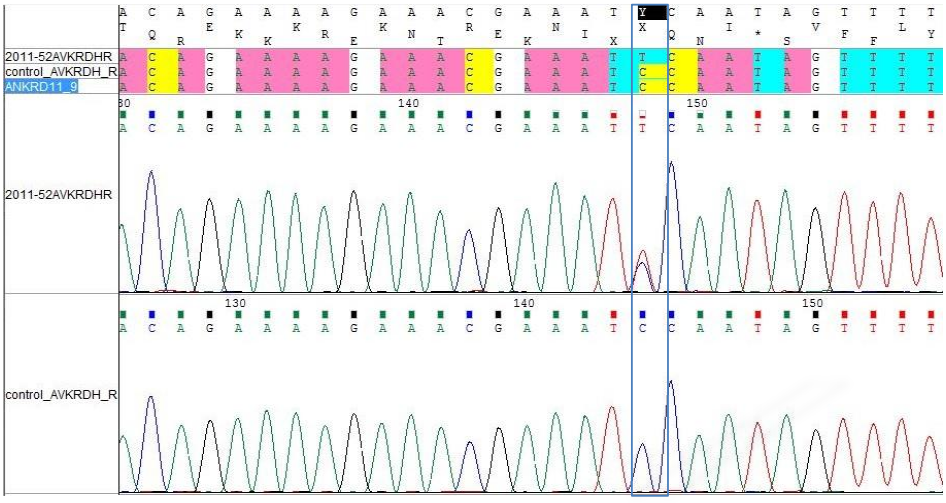
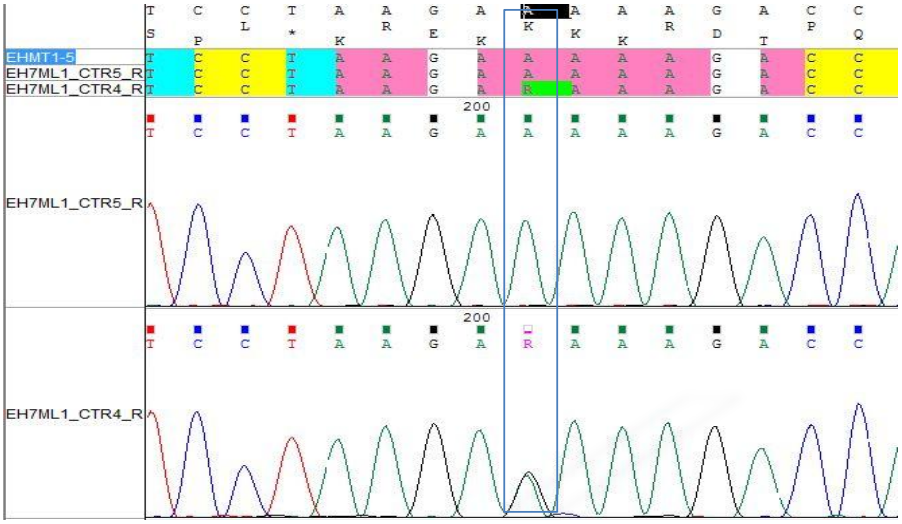
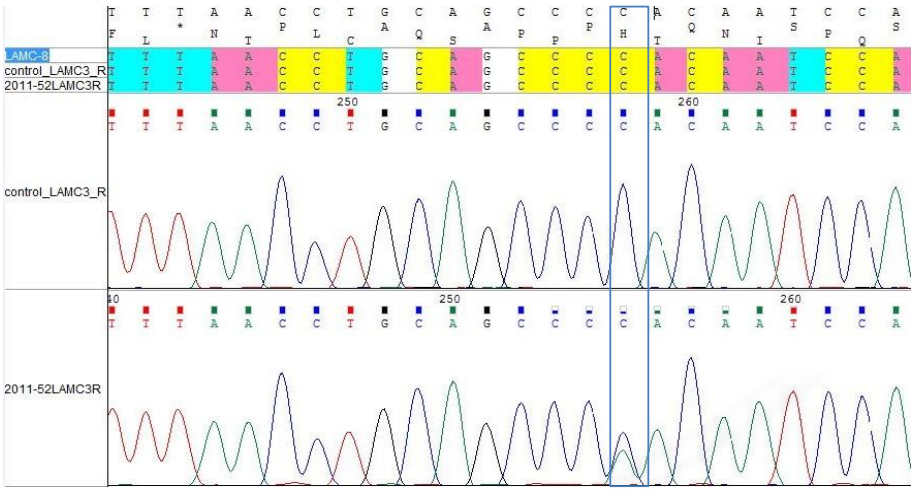


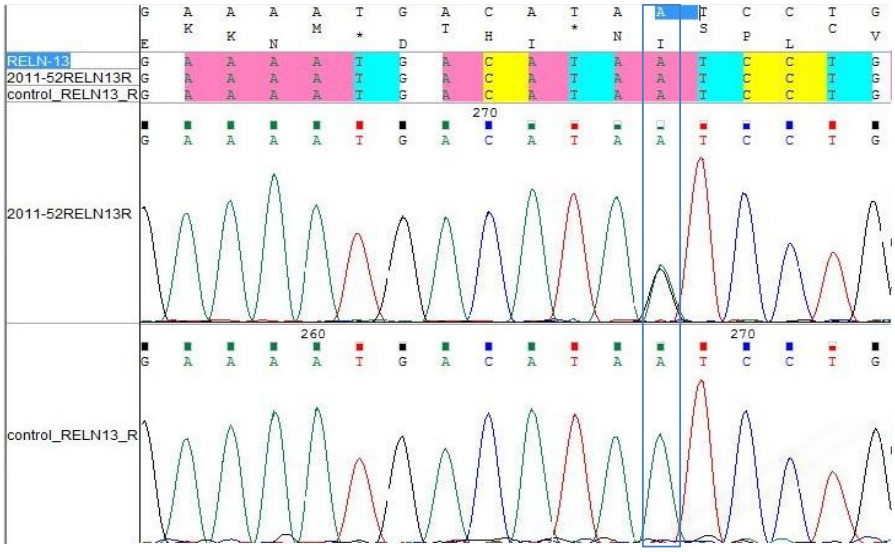
Figure 13. Chromatogram pattering of the variants called in patient #3 done through Sanger sequencing aligned with reference of each variant taken from (UCSC genome browser) and a one direction (forward or reverse) normal control sample. a. Chromatogram pattern for the variant called as high quality with point mutation (MET, Dmd, Tsc2, and SHANK2-3) respectively. b. Chromatogram pattern for the variant called as low quality without mutation (normal) (KIRRELE3).

a.




EHMT1-5
EH7ML1_CTR5_R
EH7ML1_CTR4_R
200
EH7ML1_CTR5_R
200
EH7ML1_CTR4_R


LAMC-8
control_LAMC3_R
2011-52LAMC3R
250
260
control_LAMC3_R
250
260
2011-52LAMC3R



b.

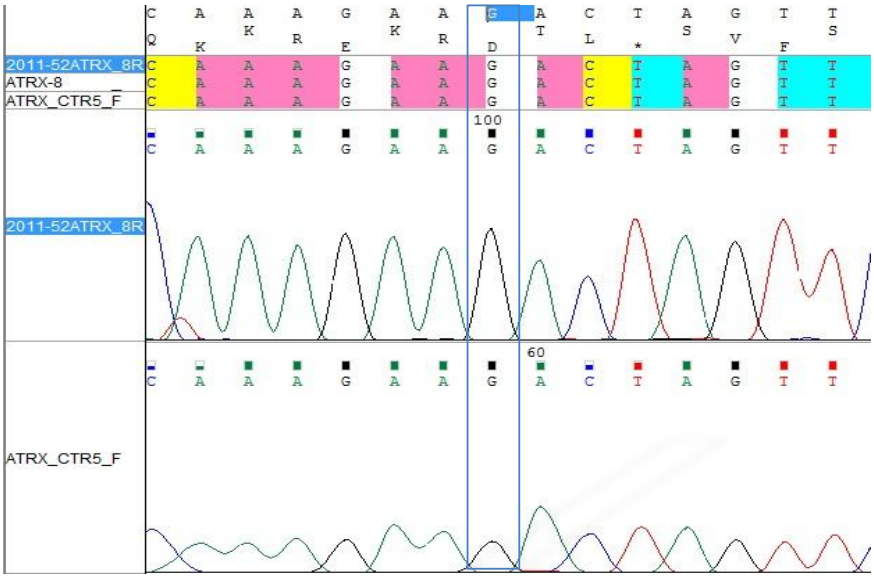


Figure 14. Chromatogram pattering of the variants called in patient #4 done through Sanger sequencing aligned with reference of each variant taken from (UCSC genome browser) and a one direction (forward or reverse) normal control sample. a. Chromatogram pattern for the variant called as high quality with point mutation (RELN-13e, LAMC3, EHMT1, and ANKRD11) respectively. b. Chromatogram pattern for the variant called as low quality without mutation (normal)(ATRX_8).

5. Discussion:

DNA Sequencing techniques are the key tools promoting new discoveries in many fields. These techniques have large benefits in different sciences (genetics, biotechnology, molecular biology, forensic sciences, and others) (Lilian *et al*., 2002)[44]. The importance of using DNA sequencing methods is clearly established in some genetic diseases; that cannot be clinically identified, or when there is no correlation between the used test-result for specific genes detection and the known associated genes of some genetic diseases. In addition, they are very useful in the case of disorders that are genetically heterogeneous, which reflect the cases of more than 100 genes [45]. Relative to the public health and safety, DNA sequencing methods showed chronological development.

Recently, Next generation sequencing (NGS) is developed worldwide as a routine diagnostic method that is used in some diseases like cancer to detect some markers like BRAC1/2[46]. First, it is used for sequencing high throughput data makes great revolution in the field of biomedical researches and genetic diagnostic fields. The second significant advantage of NGS appears significantly in the cost. The cost of running samples by using next generation platforms like MiSeq depends on the number of samples that are pooled in one run. Nowadays, the scientists reach to the point with MiSeq platform is that it could reduce the cost of sequencing to 10 folds in comparison with Sanger sequencing if we try to sequence up to 10000 amplicons per one run[45].

All of these advantages open the window of trying to find a proper way to use that magnificent machine and to get the maximum benefit from it. However, due to the sequencing of large genomes and the vast amount of data (that are generated by alignment of the short reads with their references) [47] , there is a high probability of producing false positive and false negative results and it become difficult to distinguish these from the true variants detected. Because of that, we try to develop a criteria that could help in this purpose by reducing the rate of false positive/negative and at the same time, this can help the researchers to get the benefit from the huge amount of the derived data, and manipulate these data for identifying the correct variants that are already diffuse in many biomedical research fields that the researcher is interested in.

In this research, we used one of the great vital platforms of NGS which is MiSeq platform that is known to be faster, cost-efficient, give higher resolution, and has easier workflow when compared with other platforms of NGS [48].

Using Sanger sequencing method (the gold standard method) will not be time and cost effective in generating this huge amount of data, but it could help in validating the variants detected by next generation sequencing because although it is more accurate in detecting single variant in a single exome, but it is very costly ineffective and take longer time if it is used to sequence the whole genome for many patient or different exons in a single patient. So overlapping the two methods will be more efficient specifically in the diagnosis of a disease or in the research that conduct genetic diseases and need sequencing as a major method in their research.

However, there were lots of challenges that scientists face. The major challenge is the presence of false positive/negative and the difficulty in distinguishing them from real variant. In this research, we depended particularly on the Q20 as minimum quality coverage accepted at 20X (Qscore) that reflect the probability of having error base call with the accuracy of sequencing coverage that range between (97.5%-95.7%) among the four patients, we chosen this specific Qscore based on the enrichment MiSeq report that we have for the four patient and to cover as much as we could of correct variant calls.

The obtained data results in our study are consistent with the results previously reported by Pagi *et al*., 2016 that shows consistency with the use of Q20 as minimum acceptance coverage in the criteria that used for filtering the data [49]. However, in their case they used 2 Bioinformatics pipelines and made a comparison between them, the 2 Bioinformatics pipelines were: SeqNext and MiSeq Reporter to compare the coverage values between both of them.

Although our study was consistencey with Pagi *et al*. studies in the target coverage, it showed inconsistency with the data that was published by Diociaiuti *et al*., 2016 [50]. In their defense Diociaiuti *et al*. used the Q30 as a minimum coverage (Qscore), because they mentioned that they had accuracy reads with 99.9% for all variants that are >30, but they chosen this criterion regardless of the number of variants that might be missed when using Q30. Because of that we took the broader range (Q20).

A cohort of 4 ASD patients was studied in this research by using a NGS approach from the Illumina MiSeq platform. The DNA library preparation was done by using TruSight Autism Panel illumina Kit and completed in two working days. After that, the samples were run on the MiSeq. MiSeqReporter, AV and IGV2.3 software were used for interpretation of the result. A hypothesized filtering pipeline was used for filtering the data resulted from the MiSeqReporter and displayed as VCF file. This hypothesized filtering criteria depends on many categories to be considered when doing the filtering based on Q20 (20X), passing variants, allele frequency <1%, and mutation type (choosing the pathogenic mutations only). Each of these parameters was been chosen for specific reason. Q20 was used because it gave good quality as over all depth coverage which range between (95.7% and 97.5%). In addition, to avoid the false negative where some variants were missed when we applied the Q30 (data not shown) after Sanger sequencing validation.

Passing variants depends on sample type, and sample quality. Therefore, choosing passing variants done because it is suggested that highly confident variants will be in the passing part, however, less variants will be encountered as false negatives at the same time. Moreover, this assumption is supported by the data reported by Huilei *et al*., 2014 [51], they used the passing variants to be analyzed that are post-calling filters. The second reason for not using the not passing variants is due to the limited time of our study, therefore we can do it as a future work.

Minor allele frequency was chosen to be <1%, the thresholds differs from one study to another, because most of the rare variants present in the population with the percentage of <1% [52] and this is consistent with the study done by Frazer *et al*., 2009 they mentioned that the solid frequency boundary for a rare variant one of less

than 1% , and in addition we want to have larger number of variants to be validated in order to support our hypothesized pipeline [53]. However, some recent studies like the study done by Alkes *et al*., 2010, stated that there could be less than this percentage, for example they found that between 0.5% and 2% allele frequency. Also there were about one-eighth of genes could affect the trait and has at least one working allele despite depending on the property of the selected strength [54]. Other study used the 0.5% as threshold cutoff like the work done by Tennessen *et al*., 2012 [55]. For mutation type, we choose only the pathogenic mutation because we were concerned of the mutations that result in alteration in the DNA level or even the protein level, this can open the door for further studies like molecular characterization of the newly identified mutated gene, in addition, investigate the affect of mutation in its corresponding protein structure, function and protein –protein interaction for better understanding of its pathophysiologic role in the developed disease which is our plan for the future study.

We used Sanger Sequencing to validate the filtered variants. A total of 23 variants with different sizes resulted from the 4 patients after applying the filtering criteria, we found that, Sanger sequencing validated 21 of them was validated by Sanger sequencing and the remaining 2 (AUTS2, and AUTS2-8) were excluded from the study.

The resulted in 21 variants not all related to the same variant type; they include SNVs and In-dels (short Insertion and Deletion). Patient#1 showed 3 SNVs (point mutation) in (FMR1, CHD7, and CERBBP) and 2 In-dels (Insertion) in (ZNF804A, and RAI1), while there were 3 SNVs (point mutation) in (RELN-25e, PCDH9, and CHD*) and 1 In-del (SHANK2) detected in patient #2. In patient #3 and

#4 there was no any In-dels detected, on the other hand, all the variants that were detected were SNVs (point mutation) and were verified as follows; (MET, Dmd, Tsc2, and SHANK2-3) in patient #3, and (RELN-13e, LAMC3, EHMT1, and ANKRD11) in patient #4.

Among the 4 patients, all the SNVs were validated by Sanger sequencing except AUTS2-8 and AUTS2 genes, on the other hand, only RAI1 was not validated by Sanger sequencing in the case of In-dels. While other 2 genes (ZNF804A, and SHANK2) were validated by Sanger sequencing completely, on other word, Sanger sequencing did not give the correct and exact variation in the exact position that was reported by MiSeq but it showed the correct position which guide the researcher to the position where there is a variation that needs a manual study.

Specificity of the primers was not of great issue to obtain positive results of sequencing by using Sanger sequencing. As in our data, we found that AUTS2-8 (AUTS2 gene-exone8) was identified and detected in patient #1 and characterized under the low quality variant (not real variant), it was detected by PCR (expected product size) but when we tried to validate it by using Sanger sequencing it did not work. Although the specificity of the designed primers based on using Amplyfix software , the designed primers (specifically the forward primer) for AUTS2 gene did not give a sharp sequence. This pitfall was verified by using National Center for Biotechnology Information- primers blast (www.ncbi.nlm.nih.gov) which clarified the non specificity of these primers, as it can align with different DNA sites on the same chromosome (chromosome 7) and/or in the different chromosomes (like chromosome 1 and 9) which may indicate that the designed primers represent conserved sequence of genes of the same family. Therefore other specific primers for that gene should be

designed and prepared to overcome this defect. However, due to the restricted time of the project we exclude it.

While in the case of AUTS2 that was identified in the patient #4 and was characterized under the high quality variants (real variant), it was detected by doing PCR but did not detected by Sanger sequencing and the sequence was not clear at all. There is no specific explanation for that except the sequence technical error problem because we repeat the sequence for this sample one time and we repeat the PCR then repeating the sequencing again and always it give not clear sequence and it could not be read at all. Because of that we exclude it as well.

Regarding In-dels, ZNF804A that was detected by MiSeq platform on patient #1, Sanger sequencing validated it with the exact variation and position. Also, Sanger sequencing was able to validate SHANK2 that was detected by MiSeq platform on patient #2 which have duplication and could be difficult to be detected which is consistency with (Zhang *et al*., 2005) [56] that explained the duplication in human genome in details and the difficulty of detecting duplication repeats by sequencing, and showed a great advantage for NGS platforms that was able to detect the In-dels variation correctly although it is known to be great challenge for sequencing methods in general and NGS in particular [57].

Sanger sequencing did not validate RAI1 that was identified on patient #1 but it gave the right position of variation. In MiSeq VCF file, the variation detected in RAI1 is the deletion of G (c.840delG), however, in the Sanger sequencing it was validated to be the insertion of CAG (c.838_840CAG) in the same region.

Actually, this issue is common in the case of in-dels that sequenced by NGS; and this data is consistent with the data published by Yuen *et al*., 2015 [58] that reported the inability of detecting all the In-dels by Sanger sequencing, on the other hand, the ability to detect only 64.3 % of all detected In-dels. The major cause of this problem with In-dels is DNA repeats which means the repetition of the same sequence many times within the same genome [59]. This repetition is present in most of the large genome in almost half of human genome that is in a range of 200–500 bp [60, 61], some of these repeats are normal (not functional) that means do not have any effect on the human evolution, while other repeats could play a key role in the evolution of the human [62, 63] although they are independent element [64, 65]. These repeats result during some biological mechanism that lead to creation and insertion of these repeats within the genome, it could be ranged from 1-2 base (mono- and dinucleotide) up to million of bases. They are separated in to three categorizes: short tandem repeats (microsatellites) and longer interspersed repeats (called short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs)) [61].

Because of that, repeats make it really difficult in the computational perspective when doing the alignment with references and in genome assembly because it is aligned with multiple location, which will not have direct effect on the read mapping program, while it affect the SNPs calling. Because when alignment program face repeats it goes within three options; the first option is to ignore all repeats (multi reads), the second option is to go with the best match attitude by alignment to the less mismatch. If they are equal multiple best match alignment, the aligner will randomly pick one of them or all of them to be reported. The last choice is

to report the maximum number of the alignment regardless of the total number reported. That will result in those false positives/ negatives interpretation [57].

The first option will result in missing some important variation because it leads to discard the repeats in many multi genes families. The other two options will be able to pick the repetitive regions through the best match approach will be the only option that postulates a more realistic coverage. After that, these reads could be resolved manually by tools like IGV [66]and SAMtools [67], which give the researchers the choice of keeping or discarding the reads. However, this way is not a practical and useful in the case of large data. SAMtools could also be used for SNVs calling, that will allow finding SNVs within the repetition region which is consistent with our finding when did NGS for RAI1, we will be able to find point mutation A>G within the repetitive region.

The most common repeats problems depend on the length of the repeats that are usually longer than the read length that will lead to providing gaps in the assembly. The second problem of repeats is that the assembler sometimes failed to distinguish them so the regions that are closest to them will be missed as well. To overcome the repeats problem, we could apply the hypothesize methods on the VCF file that generated by MiSeq to detect the SNVs, then after that we use other post processing software to detect the misassemblies such as Integrated Genomic Viewer (IGV) software [68]. Also for the read length problem, the alignment methods will depend on the depth of the coverage and the paired-end data to be able to decide if the variant detected is repeated region or not [69]. Also, detection of the repeats could be done by computational statistics on the depth of coverage for each contig, that means

if the desired genome is covered 20X for example, and then most contigs will be covered by default at 20x. So any repetitive region will be deeply covered and the algorithm will detect these repetitions and deal with it in different way[57].

In our study, there are 4 genes that were chosen randomly from the list of each patient to prove our hypothesized criteria in detecting the correct variants and we called them low quality variants. Those genes are (AUTS2-8, CDKL5) in patient #1 , (PON3) in patient #2, (KIRRELE3) in patient #3, and (ATRX_8) in patient #4. All of them are validated by Sanger sequencing except AUTS 2_8 (the reason described before) and give the exact sequence when aligned with the reference that were taken from UCSC genome browser and the normal control sample that were run along with the patients samples. Which prove that our criterion is validated and could be applied to any VCF file list that generated from MiSeq for any genetic diseases, taking the Autism as an example.

In our study, we Used Sanger sequencing analysis as a gold method in the validation of all the variants resulted in from NGS which is consistence with many published papers that depend on Sanger sequencing as the known and common validation method. One of these studies was used Sanger sequencing in validating the selected samples that have novel mutation based on the data called by NGS, is the report published by Shao *et al*., 2016 [70]. Another example is the data published by Jiang *et al*., 2013 [71] that used Sanger sequencing for validation of 36 exonic variants detected as de novo and ASD-relevant variants in some Autism patients and all of them confirmed as true variants. In addition, Michaelson *et al* published a study in 2012 that supported our data of using Sanger sequencing for validation of total of

668 putative germline DNMs and the result was the validation of only 652 sites while the remaining sites gave incomplete data [72].

There is some limitation that we faced during our study, first, the difficulty in understanding the bioinformatics tool that used for the analysis in order to develop the hypothesized criteria. Second, the limited amount of samples provided for us that didn't allow us to repeat any PCR with the low quality result.

For future studies, there is some recommendation that could be done. We could study the failing or low quality variants in order to have a complete pipeline that we can relay on and be more confident that we will not going to miss any variant. Also some molecular studies could be done for the detected variants, as well as, study the protein structure for the detected mutations and do some functional studies for better understanding of its pathophysiologic role in the developed disease.

6. Conclusion:

Next generation sequencing (NGS) and specifically targeted capture exome or genes were established as a time efficient and cost effective tool for sequencing and detecting the genetic causes for multifaceted genetic diseases and simple genetic disease.[73, 74]. NGS used for sequencing of high throughput data, these data could be up to kilobases or megabases. NGS not only used for diagnosis of genetic diseases, but also it could have a great application in the field of personalized medicine, that help in understanding the response of specific individuals to infectious diseases, to vaccinations, and particular immunosuppressive drugs[48].

Targeted region capture methods, is a direct method that aid in escalating the ability to detect the rare variants that will be difficult and costly detected by normal and common methods like Sanger sequencing. Although Targeted region capture methods are considered as a great discovery in the field of sequencing, but this will not reduce the impact of using Sanger sequencing as a gold standard method for validation of variants or mutations that are generated from NGS. This result from the challenge that NGS faced in specificity and sensitivity of data generated, which is still no clear cut criteria are available to be used by NGS to ensure that all the variants that are generated from any platform of the NGS are true and there are no false positives or false negatives are included in the list.

In our research, we were able to establish a filtering pipeline that was able to reduce the false positive significantly. Also we were able to validate the filtered list resulted by the hypothesized pipeline and prove it is efficacy in detecting rare variants correctly with reducing the cost of validating all the list generated by MiSeq platform. Also, we chosen some random variants that were passed according to MiSeq platform

but failed according to our hypothesized criteria and we were able to prove that our criteria were correct and worked well and all these detected variants were normal and gives the same sequence as a control. Our future plane is to check the failing variants to evaluate the false negatives variants that may be missed due to an analysis by MiSeq platform; also we could solve the detection of In-dels by using more specific criteria that work specifically on In-dels to ease detecting them. After establishing such a criteria we could claim that this criterion good and effective to be used for filtering data of VCF file generated by MiSeq platform.

References:

1. Liu, L., et al., *Comparison of next-generation sequencing systems*. J Biomed Biotechnol, 2012. **2012**: p. 251364.
2. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome*. Nature, 2004. **431**(7011): p. 931-45.
3. Drmanac, R., et al., *Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays*. Science, 2010. **327**(5961): p. 78-81.
4. Holley, R.W., et al., *Structure of a Ribonucleic Acid*. Science, 1965. **147**(3664): p. 1462-5.
5. Sanger, F., et al., *Nucleotide sequence of bacteriophage phi X174 DNA*. Nature, 1977. **265**(5596): p. 687-95.
6. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
7. Ronaghi, M., et al., *Real-time DNA sequencing using detection of pyrophosphate release*. Anal Biochem, 1996. **242**(1): p. 84-9.
8. Grada, A. and K. Weinbrecht, *Next-generation sequencing: methodology and application*. J Invest Dermatol, 2013. **133**(8): p. e11.
9. Mardis, E.R., *The impact of next-generation sequencing technology on genetics*. Trends Genet, 2008. **24**(3): p. 133-41.
10. Durtschi, J., et al., *VarBin, a novel method for classifying true and false positive variants in NGS data*. BMC Bioinformatics, 2013. **14 Suppl 13**: p. S2.

11. Koboldt, D.C., et al., *Challenges of sequencing human genomes*. Brief Bioinform, 2010. **11**(5): p. 484-98.

12. Shen, Y., et al., *A SNP discovery method to assess variant allele probability from next-generation resequencing data*. Genome Res, 2010. **20**(2): p. 273-80.

13. Junemann, S., et al., *Updating benchtop sequencing performance comparison*. Nat Biotechnol, 2013. **31**(4): p. 294-6.

14. Marshall, C.R., et al., *Structural variation of chromosomes in autism spectrum disorder*. Am J Hum Genet, 2008. **82**(2): p. 477-88.

15. Talkowski, M.E., et al., *Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries*. Cell, 2012. **149**(3): p. 525-37.

16. Vandeweyer, G. and R.F. Kooy, *Balanced translocations in mental retardation*. Hum Genet, 2009. **126**(1): p. 133-47.

17. Koshimizu, E., et al., *Performance comparison of bench-top next generation sequencers using microdroplet PCR-based enrichment for targeted sequencing in patients with autism spectrum disorder*. PLoS One, 2013. **8**(9): p. e74167.

18. King, C. and T. Scott-Horton, *Pyrosequencing: a simple method for accurate genotyping*. J Vis Exp, 2008(11).

19. Zheng, Z., et al., *Titration-free massively parallel pyrosequencing using trace amounts of starting material*. Nucleic Acids Res, 2010. **38**(13): p. e137.

20. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors*. Nature, 2005. **437**(7057): p. 376-80.

21. Huse, S.M., et al., *Accuracy and quality of massively parallel DNA pyrosequencing*. Genome Biol, 2007. **8**(7): p. R143.

22. Rohde, H., et al., *Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4*. N Engl J Med, 2011. **365**(8): p. 718-24.

23. Loman, N.J., et al., *Performance comparison of benchtop high-throughput sequencing platforms*. Nat Biotechnol, 2012. **30**(5): p. 434-9.

24. Xuan, J., et al., *Next-generation sequencing in the clinic: promises and challenges*. Cancer Lett, 2013. **340**(2): p. 284-95.

25. Gullapalli, R.R., et al., *Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics*. J Pathol Inform, 2012. **3**: p. 40.

26. Rehm, H.L., *Disease-targeted sequencing: a cornerstone in the clinic*. Nat Rev Genet, 2013. **14**(4): p. 295-300.

27. Luthra, R., et al., *Next-generation sequencing-based multigene mutational screening for acute myeloid leukemia using MiSeq: applicability for diagnostics and disease monitoring*. Haematologica, 2014. **99**(3): p. 465-73.

28. Domina, M., et al., *Rapid profiling of the antigen regions recognized by serum antibodies using massively parallel sequencing of antigen-specific libraries*. PLoS One, 2014. **9**(12): p. e114159.

29. Barrett, A.N., et al., *Digital PCR analysis of maternal plasma for noninvasive detection of sickle cell anemia*. Clin Chem, 2012. **58**(6): p. 1026-32.

30. Ottesen, A.R., et al., *Co-enriching microflora associated with culture based methods to detect Salmonella from tomato phyllosphere*. PLoS One, 2013. **8**(9): p. e73079.

31. Ulahannan, D., et al., *Technical and implementation issues in using next-generation sequencing of cancers in clinical practice*. Br J Cancer, 2013. **109**(4): p. 827-35.

32. Tang, M., et al., *Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics*. Nucleic Acids Res, 2014. **42**(22): p. e166.

33. Delseny, M., B. Han, and Y. Hsing, *High throughput DNA sequencing: The new sequencing revolution*. Plant Sci, 2010. **179**(5): p. 407-22.

34. Kanner, L., *Irrelevant and metaphorical language in early infantile autism*. Am J Psychiatry, 1946. **103**(2): p. 242-6.

35. Asperger, H., *[On the differential diagnosis of early infantile autism]*. Acta Paedopsychiatr, 1968. **35**(4): p. 136-45.

36. AlSagob, M., D. Colak, and N. Kaya, *Genetics of autism spectrum disorder: an update on copy number variations leading to autism in the next generation sequencing era*. Discov Med, 2015. **19**(106): p. 367-79.

37. Maskos, U. and E.M. Southern, *Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ*. Nucleic Acids Res, 1992. **20**(7): p. 1679-84.

38. Shattuck, P.T., et al., *Participation in social activities among adolescents with an autism spectrum disorder*. PLoS One, 2011. **6**(11): p. e27176.

39. Herman, G.E., et al., *Genetic testing in autism: how much is enough?* Genet Med, 2007. **9**(5): p. 268-74.

40. Snijders, A.M., et al., *Assembly of microarrays for genome-wide measurement of DNA copy number*. Nat Genet, 2001. **29**(3): p. 263-4.

41. Itsara, A., et al., *Population analysis of large copy number variants and hotspots of human genetic disease*. Am J Hum Genet, 2009. **84**(2): p. 148-61.

42. Ku, C.S., et al., *Gene discovery in familial cancer syndromes by exome sequencing: prospects for the elucidation of familial colorectal cancer type X*. Mod Pathol, 2012. **25**(8): p. 1055-68.

43. Long, S.W., et al., *A genomic day in the life of a clinical microbiology laboratory*. J Clin Microbiol, 2013. **51**(4): p. 1272-7.

44. Franca, L.T., E. Carrilho, and T.B. Kist, *A review of DNA sequencing techniques*. Q Rev Biophys, 2002. **35**(2): p. 169-200.

45. De Leeneer, K., et al., *Flexible, scalable, and efficient targeted resequencing on a benchtop sequencer for variant detection in clinical practice*. Hum Mutat, 2015. **36**(3): p. 379-87.

46. Endris, V., et al., *NGS-based BRCA1/2 mutation testing of high-grade serous ovarian cancer tissue: results and conclusions of the first international round robin trial*. Virchows Arch, 2016.

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

48. Vogiatzi, P., *Some considerations on the current debate about typing resolution in solid organ transplantation*. Transplant Res, 2016. **5**: p. 3.

49. Pagin, A., et al., *Applicability and Efficiency of NGS in Routine Diagnosis: In-Depth Performance Analysis of a Complete Workflow for CFTR Mutation Analysis*. PLoS One, 2016. **11**(2): p. e0149426.

50. Diociaiuti, A., et al., *Role of molecular testing in the multidisciplinary diagnostic approach of ichthyosis*. Orphanet J Rare Dis, 2016. **11**(1): p. 4.

51. Xu, H., et al., *Comparison of somatic mutation calling methods in amplicon and whole exome sequence data*. BMC Genomics, 2014. **15**: p. 244.

52. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data*. Am J Hum Genet, 2008. **83**(3): p. 311-21.

53. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits*. Nat Rev Genet, 2009. **10**(4): p. 241-51.

54. Price, A.L., et al., *Pooled association tests for rare variants in exon-resequencing studies*. Am J Hum Genet, 2010. **86**(6): p. 832-8.

55. Tennessen, J.A., et al., *Evolution and functional impact of rare coding variation from deep sequencing of human exomes*. Science, 2012. **337**(6090): p. 64-9.

56. Zhang, L., et al., *Patterns of segmental duplication in the human genome*. Mol Biol Evol, 2005. **22**(1): p. 135-41.

57. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions*. Nat Rev Genet, 2012. **13**(1): p. 36-46.

58. Yuen, R.K., et al., *Whole-genome sequencing of quartet families with autism spectrum disorder*. Nat Med, 2015. **21**(2): p. 185-91.

59. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping*. Nat Rev Genet, 2011. **12**(5): p. 363-76.

60. Schmid, C.W. and P.L. Deininger, *Sequence organization of the human genome*. Cell, 1975. **6**(3): p. 345-58.

61. Batzer, M.A. and P.L. Deininger, *Alu repeats and human genomic diversity*. Nat Rev Genet, 2002. **3**(5): p. 370-9.

62. Jurka, J., et al., *Repetitive sequences in complex genomes: structure and evolution*. Annu Rev Genomics Hum Genet, 2007. **8**: p. 241-59.

63. Britten, R.J., *Transposable element insertions have strongly affected human evolution*. Proc Natl Acad Sci U S A, 2010. **107**(46): p. 19945-8.

64. Hua-Van, A., et al., *The struggle for life of the genome's selfish architects*. Biol Direct, 2011. **6**: p. 19.

65. Kim, P.M., et al., *Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history*. Genome Res, 2008. **18**(12): p. 1865-74.

66. Robinson, J.T., et al., *Integrative genomics viewer*. Nat Biotechnol, 2011. **29**(1): p. 24-6.

67. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.

68. Phillippy, A.M., M.C. Schatz, and M. Pop, *Genome assembly forensics: finding the elusive mis-assembly*. Genome Biol, 2008. **9**(3): p. R55.

69. Gnerre, S., et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data*. Proc Natl Acad Sci U S A, 2011. **108**(4): p. 1513-8.

70. Shao, D., et al., *A targeted next-generation sequencing method for identifying clinically relevant mutation profiles in lung adenocarcinoma*. Sci Rep, 2016. **6**: p. 22338.

71. Jiang, Y.H., et al., *Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing*. Am J Hum Genet, 2013. **93**(2): p. 249-63.

72. Michaelson, J.J., et al., *Whole-genome sequencing in autism identifies hot spots for de novo germline mutation*. Cell, 2012. **151**(7): p. 1431-42.

73. Bonnefond, A., et al., *Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome*. PLoS One, 2010. **5**(10): p. e13630.

74. Choi, M., et al., *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing*. Proc Natl Acad Sci U S A, 2009. **106**(45): p. 19096-101.

75. Life Sciences, a Roche Company [(accessed on 22 October 2012)]. Available online: http://www.454.com/