Genetic Analysis of Mendelian Disorders by Next Generation Sequencing

Dissertation submitted in partial fulfilment for the degree of

Master of Science in Applied Microbiology

Submitted By

Reema Kar

Roll No. 1662013

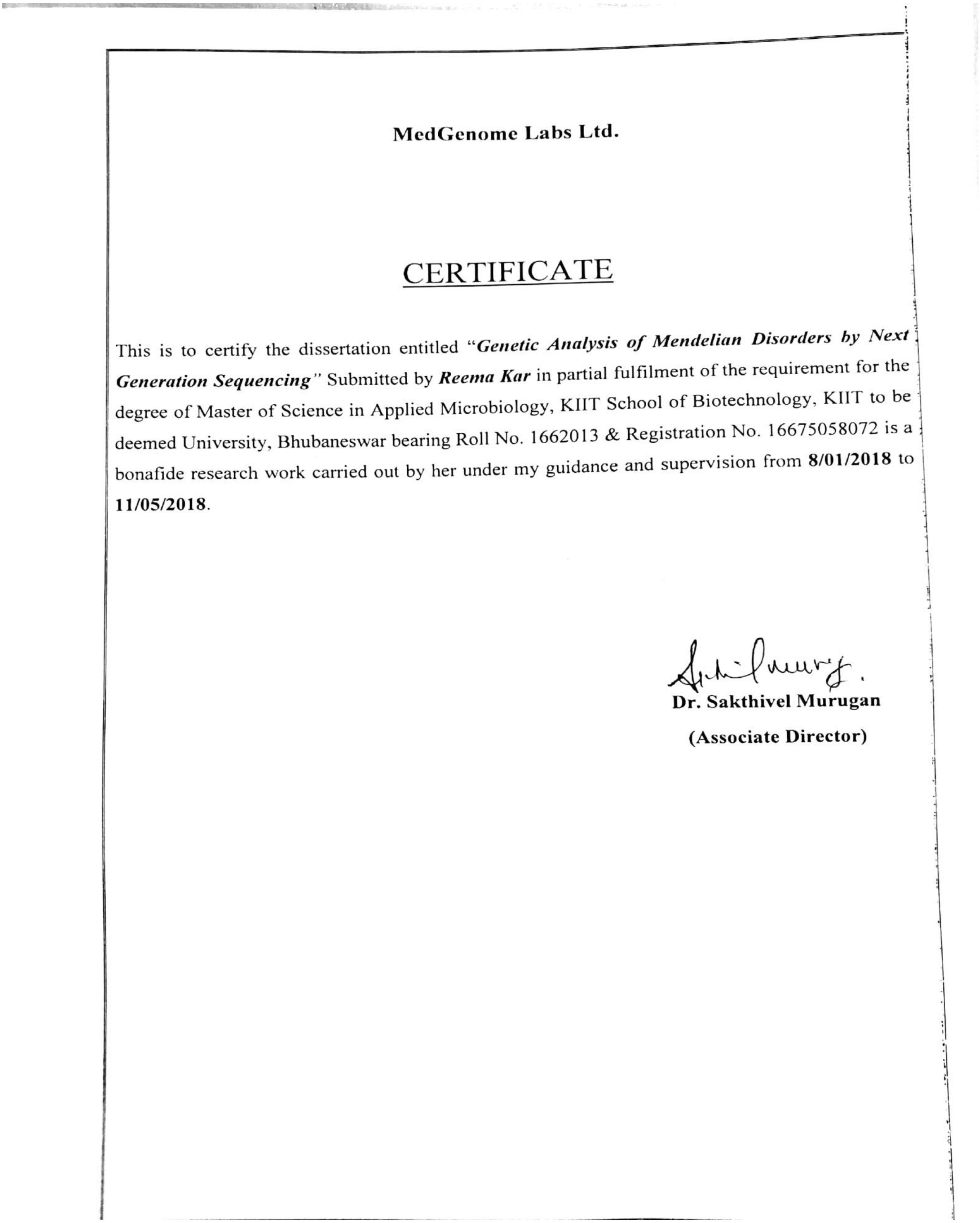
KIIT School of Biotechnology, Campus- 11
KIIT to be deemed University
Bhubaneswar, Odisha, India

Under the Supervision of

Dr. Lova Satyanarayana Matsa

MedGenome Labs Ltd.

3rd Floor, Narayana Nethralaya Building, Narayana Health City,
#258/A, Bommasandra, Hosur Road, Bangalore - 560 099, India.



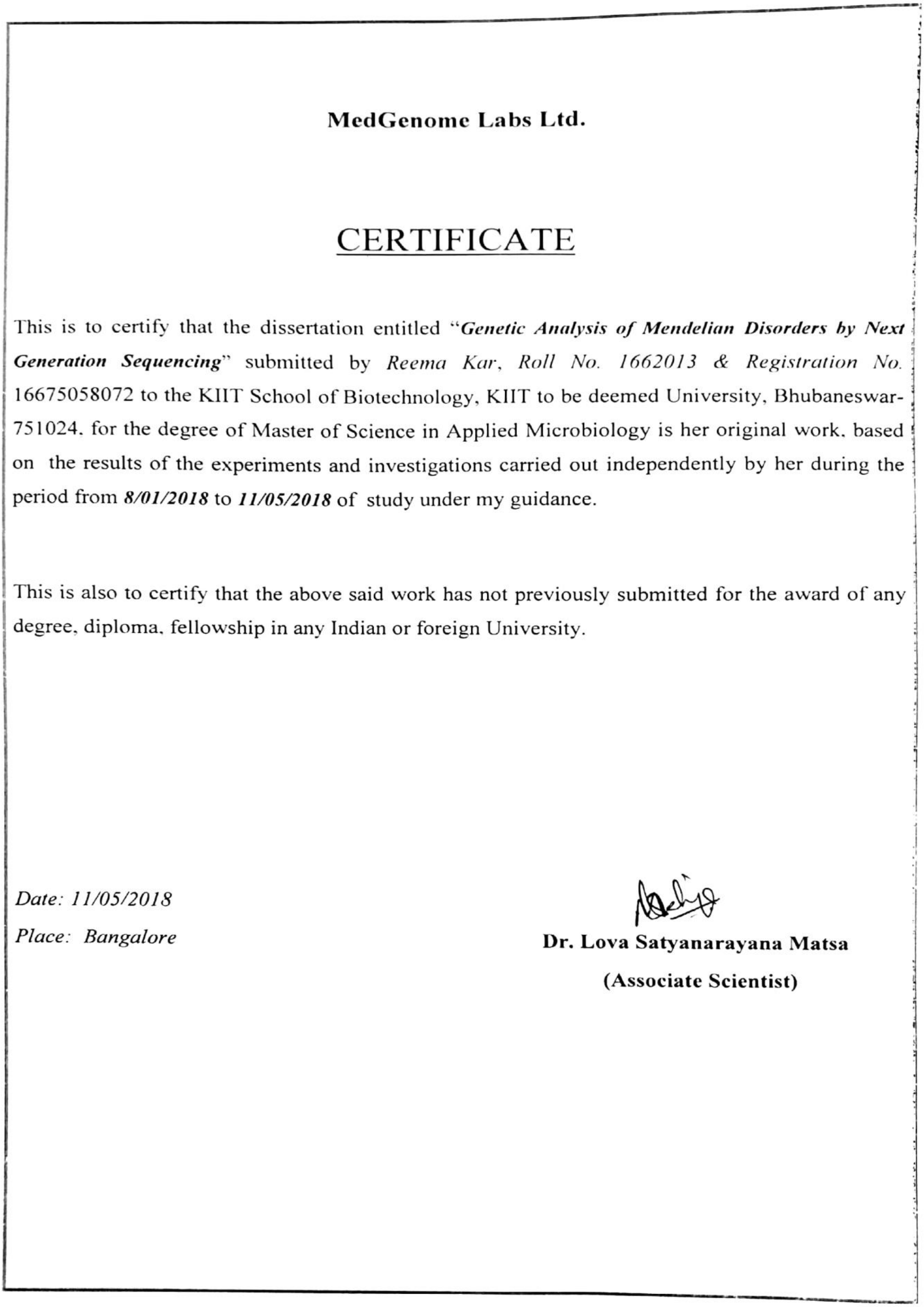
MedGenome Labs Ltd.

CERTIFICATE

This is to certify the dissertation entitled "***Genetic Analysis of Mendelian Disorders by Next Generation Sequencing***" Submitted by ***Reema Kar*** in partial fulfilment of the requirement for the degree of Master of Science in Applied Microbiology, KIIT School of Biotechnology, KIIT to be deemed University, Bhubaneswar bearing Roll No. 1662013 & Registration No. 16675058072 is a bonafide research work carried out by her under my guidance and supervision from **8/01/2018** to **11/05/2018**.

Dr. Sakthivel Murugan

(Associate Director)



MedGenome Labs Ltd.

CERTIFICATE

This is to certify that the dissertation entitled "***Genetic Analysis of Mendelian Disorders by Next Generation Sequencing***" submitted by *Reema Kar, Roll No. 1662013 & Registration No.* 16675058072 to the KIIT School of Biotechnology, KIIT to be deemed University, Bhubaneswar-751024. for the degree of Master of Science in Applied Microbiology is her original work, based on the results of the experiments and investigations carried out independently by her during the period from ***8/01/2018*** to ***11/05/2018*** of study under my guidance.

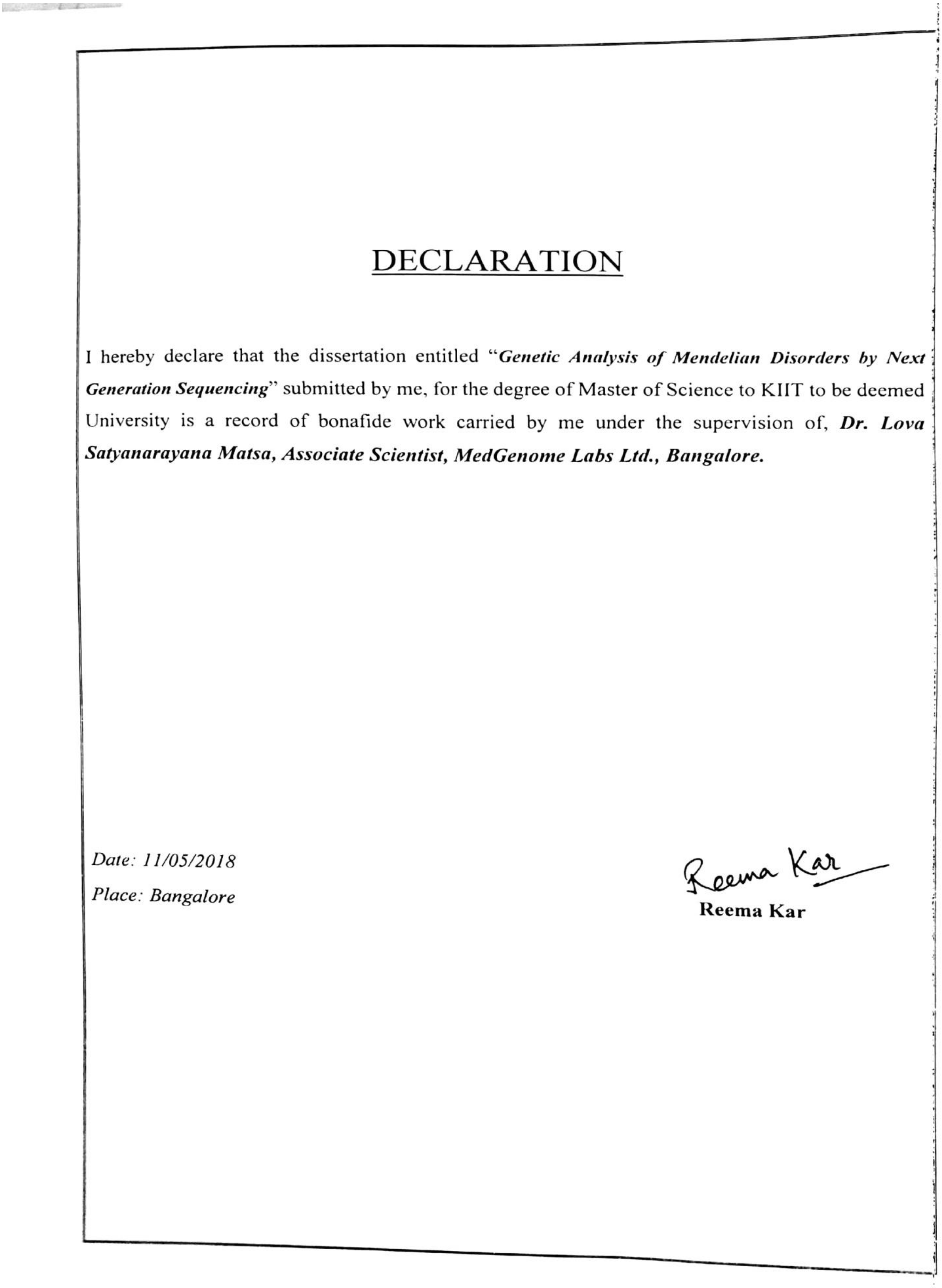
This is also to certify that the above said work has not previously submitted for the award of any degree, diploma, fellowship in any Indian or foreign University.

Date: 11/05/2018

Place: Bangalore

Dr. Lova Satyanarayana Matsa

(Associate Scientist)



DECLARATION

I hereby declare that the dissertation entitled "***Genetic Analysis of Mendelian Disorders by Next Generation Sequencing***" submitted by me, for the degree of Master of Science to KIIT to be deemed University is a record of bonafide work carried by me under the supervision of, ***Dr. Lova Satyanarayana Matsa, Associate Scientist, MedGenome Labs Ltd., Bangalore.***

Date: 11/05/2018

Place: Bangalore

Reema Kar

Reema Kar

Abstract

Over the past decade, next-generation sequencing (NGS) has led to a great increase in our understanding of the genetic basis of Mendelian diseases. NGS allows for the analysis of multiple regions of the genome in one single reaction and has been shown to be a cost-effective method, along with it also an efficient tool in investigating patients with Mendelian diseases. More recently, NGS has successfully been deployed in the clinics, with a reported diagnostic yield of ~25 %. However, recommendations on clinical implementation of NGS are still evolving with numerous key challenges that impede the widespread use of genetics in everyday medicine. These challenges include when to order, on whom to order, what type of test to order, and how to interpret and communicate the results, including incidental findings, to the patient and family.

Acknowledgements

I would like to express my sincere gratitude towards Dr. Sakthivel Murugan, Associate Director, MedGenome Labs Limited, Bangalore.

I am sincerely grateful to Dr. Lova Satyanarayana Matsa, Associate Scientist, MedGenome Labs Limited, Bangalore, for his constant encouragement, support, guidance and valuable suggestions which have helped me to complete my project and sharpen my knowledge and analytical skills.

No formal words are sufficient to express my cordial thanks to my team mates Balraj M., Chandru J., Pranav Chand Rayabarapu, Rashmi Chandra and Sakina Aamir for rendering their guidance and cooperation.

And sincerely I want to thank Dr. Mrutyunjay Suar, Director KIIT School of Biotechnology, for his constant support and guidance.

Contents

Abbreviations

DNA	Deoxyribonucleic Acid
NGS	Next-generation sequencing
PCR	Polymerase Chain Reaction
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
FISH	Fluorescence In Situ Hybridisation
CGH	Comparative Genomic Hybridisation
ACMG	American College of Medical Genetics and Genomics
CVS	Copy Number Variation
HGMD	Human Gene Mutation Database
VUS	Variant of Uncertain Significance

List of Figures Page No.

1. Introduction

Mendelian diseases or monogenic diseases are disorders caused by mutations in one gene and include diseases like cystic fibrosis, thalassemia among others. Mendelian diseases are considered to be rare individually but collectively it occurs at a rate of 40 to 82 per 1000 live births, with an estimated 7.9 million children being born annually with a serious birth defect of genetic or partially genetic origin [1]. These disorders tend to run in families, although it has been found that a significant number are caused by de novo events [2]. Depending on the pathomechanism, the phenotype is manifested in a dominant (where one allele is mutated) or recessive (where both alleles are mutated) manner. Of the estimated 20,000–25,000 protein-coding genes in the human genome, mutations in 3348 genes have been associated with Mendelian diseases [3].

Sanger sequencing, also known as the dideoxy method, has been the gold standard in molecular diagnostics in Mendelian diseases and remains the test of choice for clinical genetic testing the purpose of this is to confirm a suspected diagnosis and allow more accurate genetic counseling. However, Sanger sequencing can only analyze one DNA segment at a time and is thus laborious and time consuming. For diseases with genetic heterogeneity, like retinitis pigmentosa, cardiomyopathy, and deafness, a gene-by-gene Sanger sequencing approach has not been shown to be economical or efficient.

Recent advances over the past decade have allowed for high-throughput sequencing, and these advances are collectively referred to as next-generation sequencing (NGS). NGS has allowed for substantial increase in sequencing content while dramatically reducing the cost of sequencing per base. This allows for simultaneous interrogation of multiple genes through one single reaction and has been proven to be an effective alternative for establishing the genetic basic of Mendelian diseases in the research setting [4], and more recently in the clinical setting [5].

2. Background and Rationale

Next-generation sequencing (NGS) was developed in the last decade and allows simultaneous sequencing of millions of DNA fragments without previous sequence knowledge. This advanced technology has been a true revolution compared with the traditional sequencing methods, in which one or a few relatively short fragments of DNA, previously amplified by PCR, could be sequenced per tube. Due to the high costs and intensive work required, traditional sequencing was only performed on specific DNA regions and for specific samples. For instance, genetic screening of heterozygous mutations, such as in the case of breast/ovarian cancer or Lynch syndromes, was previously based on the screening of DNA heteroduplexes through different non-sequencing methods. Only selected samples from subjects with a strong indication for further DNA analysis would then be sequenced. Meanwhile, the Human Genome Project, which was launched in 1990, required 13 years and billions of euros in order to sequence the complete human genome.

With NGS, the promise of today is that a complete genome can be sequenced in a few days for less than $1000 per genome. Even though we are not there yet, the implications and the impact of NGS in understanding the biological processes of diseases like cancer and in personalising patient care are unprecedented.

NGS technology describes the major milestones in the technical and developments in the field of disease diagnosis.

3. Genomic Analysis

Genomic analysis is the identification, measurement or comparison of genomic features such as DNA sequence, structural variation, gene expression, or regulatory and functional element annotation at a genomic scale. Methods for genomic analysis typically require high-throughput sequencing or microarray hybridization and bioinformatics.

Genome projects typically involve three main phases: DNA sequencing, assembly of DNA to represent original chromosome, and analysis of the representation.

DNA sequencing is the process to determine the nucleotide order in a specific DNA molecule, which is useful when attempting to understand its function and consequent effects in the organism it resides in. DNA sequence assembly involves the alignment and merging of DNA fragments to reconstruct the DNA so that smaller sections of the genome can be analyzed.

The analysis of DNA phase is the final step in genome analysis. It brings together the discoveries from the previous phases of the project to form conclusions, which can offer true value to further our knowledge of the genome and be applied in relevant situations.

Annotation of DNA

DNA annotation is the process of identifying the locations of genes and coding regions in a genome to create ideas about the possible functions of the genes. There are three main steps to annotate the genome, which include to:

- Identify the portions of the genome that are not involved in coding proteins
- Identify the main elements of the genome (gene prediction)
- Connect the main elements of the genome with biological information

It is important to consider how the genome is similar to other genomes that are already known, as this can help when establishing the role of the gene. Additionally, the plasmids, phages and resistance genes of the genome can reveal information about the nature of the genome.

The traditional method of curation method uses the Basic Local Alignment Search Tool (BLAST) algorithm to find similarities to annotate the genome. However, this approach involves expert knowledge and experimental verification to be carried out.

Technology for genome analysis

The recent advances in technology that allow high throughput genomic sequencing to be undertaken quickly and relatively cheaply has propelled the work of genome analysis forward.

However, this progression also places a large demand for efficient and robust tools of analysis to interpret the data into a form that can be utilized in practice. The massive sets of data that have been produced by projects, such as the Human Genome Project, remain largely under-utilized, despite the fact that the project concluded more than a decade ago.

It is important to develop techniques to both analyze the information that we currently have available and the level of data that we are generating.

4. Next Generation Sequencing (NGS)

High-throughput massively parallel DNA sequencing more commonly termed “next-generation sequencing,” is an innovation in sequencing that emerged during the past decade. Next-generation sequencing (NGS) is not a single technology, rather several different technologies that share a common feature of huge parallel sequencing of amplified or single DNA molecules in a flow cell or chip. NGS technologies are unique sequencing chemistries that differ from the Sanger dideoxynucleotide chain termination chemistry. NGS can generate, in a single instrument run, hundreds of millions to gigabases of nucleotide sequence data depending upon the platform configuration, chemistry and flow cell or chip capacity. Although several NGS technologies have been commercialized and technologies finding greatest adoption into clinical laboratories are emphasized. Current clinical testing, its applications includes multigene panels and exome and genome sequencing for candidate and causal gene identification. While the examples are primarily based on analyses for inherited disorders, the principles described are applicable to oncology and other infectious diseases, with certain modifications mostly specific to the specimen characteristics and sensitivity requirements for these other applications [6].

Next generation methods of DNA sequencing have three general steps:

- Library preparation: libraries are created using random fragmentation of DNA, followed by ligation with custom linkers
- Amplification: the library is amplified using clonal amplification methods and PCR
- Sequencing: DNA is sequenced using one of several different approaches

4.1 Library Preparation

Firstly, DNA is fragmented either enzymatically or by sonication (excitation using ultrasound) to create smaller strands. Adaptors (short, double-stranded pieces of synthetic DNA) are then ligated to these fragments with the help of DNA ligase, an enzyme that joins DNA strands. These adaptors enable the sequence to become bound to a complementary counterpart.

Adaptors are synthesised so that one end is 'sticky' whilst the other is 'blunt' (non-cohesive) with the view to joining the blunt end to the blunt ended DNA. This could lead to the potential problem of base pairing between molecules and therefore dimer formation. To prevent this, the chemical structure of DNA is utilised, since ligation takes place between the 3′-OH and 5′-P ends. By removing the phosphate from the sticky end of the adaptor and therefore creating a 5′-OH end instead, the DNA ligase is unable to form a bridge between the two termini.

The library fragments need to be spatially clustered in PCR colonies or 'polonies' as they are conventionally known, which consist of many copies of a particular library fragment. Since these polonies are attached in a planar manner, the features of the array can be manipulated enzymatically in parallel.

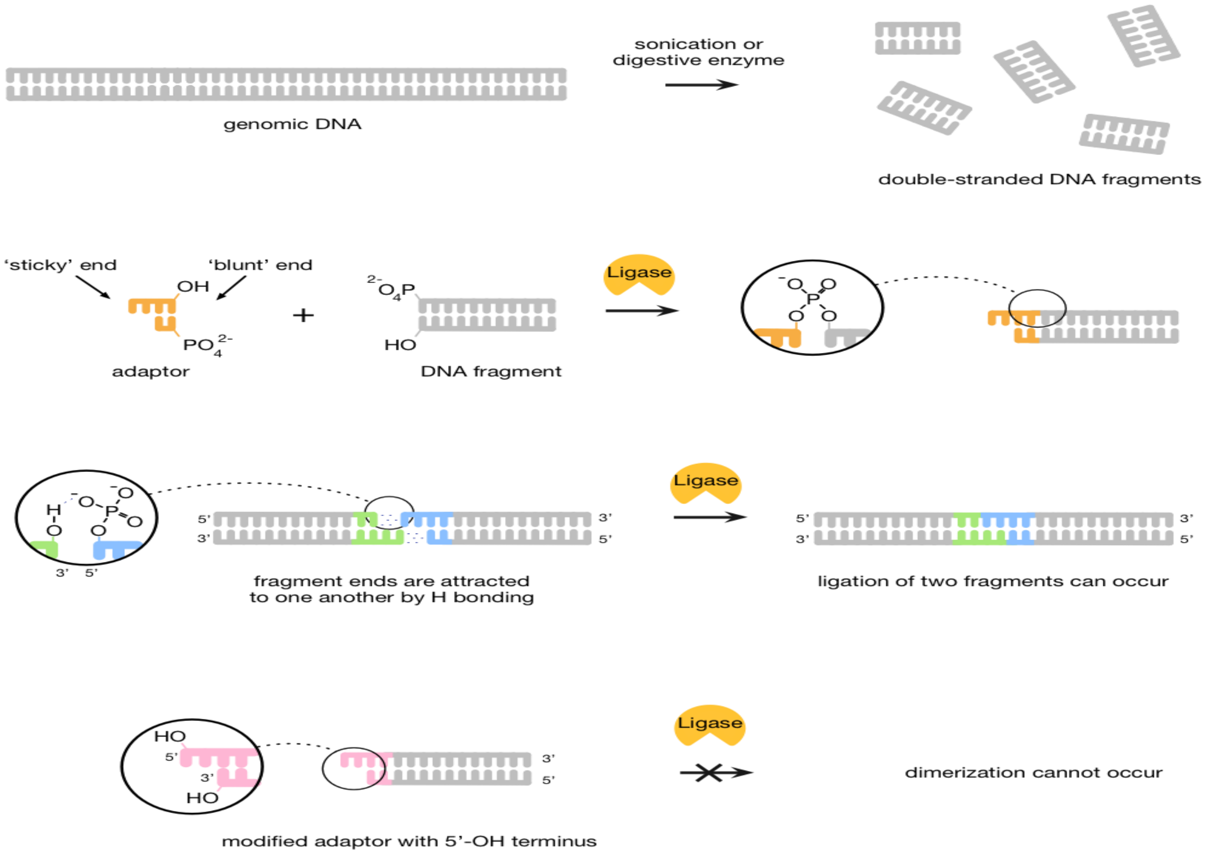

genomic DNA
sonication or
digestive enzyme
double-stranded DNA fragments
'sticky' end
'blunt' end
OH
PO$_4^{2-}$
adaptor
+
$^{2-}O_4P$
HO
DNA fragment
Ligase
fragment ends are attracted
to one another by H bonding
ligation of two fragments can occur
5'
3'
HO
modified adaptor with 5'-OH terminus
dimerization cannot occur

Fig 1: Library preparation of Next-generation sequencing

4.2 Amplification

Library amplification is required so that the received signal from the sequencer is strong enough to be detected accurately. With enzymatic amplification, phenomena such as 'biasing' and 'duplication' can occur leading to preferential amplification of certain library fragments. Instead, there are several types of amplification process which use PCR to create large numbers of DNA clusters.

4.3 Bridge PCR

The surface of the flow cell is densely coated with primers that are complementary to the primers attached to the DNA library fragments. The DNA is then attached to the surface of the cell at random where it is exposed to reagents for polymerase based extension. On addition of nucleotides and enzymes, the free ends of the single strands of DNA attach themselves to the surface of the cell via complementary primers, creating bridged structures. Enzymes then interact with the bridges to make them double stranded, so that when the denaturation occurs, two single stranded DNA fragments are attached to the surface in close proximity. Repetition of this process leads to clonal clusters of localised identical strands. In order to optimise cluster density, concentrations of reagents must be monitored very closely to avoid overcrowding.

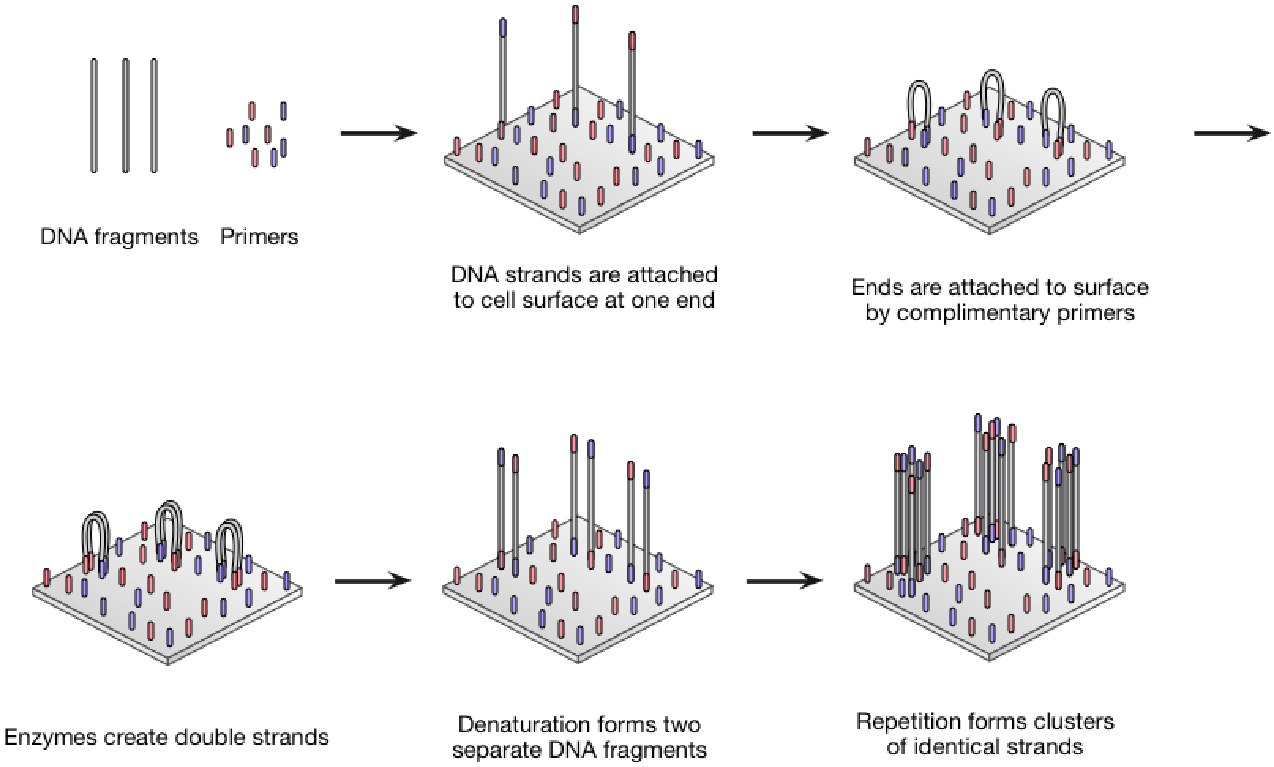

DNA fragments
Primers
DNA strands are attached to cell surface at one end
Ends are attached to surface by complimentary primers
Enzymes create double strands
Denaturation forms two separate DNA fragments
Repetition forms clusters of identical strands

Figure 2: Bridging PCR

4.4 Sequencing-by-Synthesis (Illumina Sequencing Technology)

The Illumina sequencing method is similar to Sanger sequencing, but it uses modified dNTPs containing a terminator which blocks further polymerization- so only a single base can be added by a polymerase enzyme to each growing DNA copy strand. The sequencing reaction is conducted simultaneously on a very large number (many millions in fact) of different template molecules spread out on a solid surface. The terminator also contains a fluorescent label, which can be detected by a camera. Only a single fluorescent color is used, so each of the four bases must be added in a separate cycle of DNA synthesis and imaging. Following the addition of the four dNTPs to the templates, the images are recorded and the terminators are removed. This chemistry is called “reversible terminators”. Finally, another four cycles of dNTP additions are initiated. Since single bases are added to all templates in a uniform fashion, the sequencing process produces a set of DNA sequence reads of uniform length.

Although the fluorescent imaging system used in Illumina sequencers is not sensitive enough to detect the signal from a single template molecule, the major innovation of the Illumina method is the amplification of template molecules on a solid surface. The DNA sample is prepared into a “sequencing library” by the fragmentation into pieces each around 200 bases long. Custom adapters are added to each end and the library is flowed across a solid surface (the “flow cell”) and the template fragments bind to this surface. Following this, a solid phase “bridge amplification” PCR process (cluster generation) creates approximately one million copies of each template in tight physical clusters on the flowcell surface. Illumina has improved its image analysis technology dramatically which allows for higher cluster density on the surface of the flowcell.

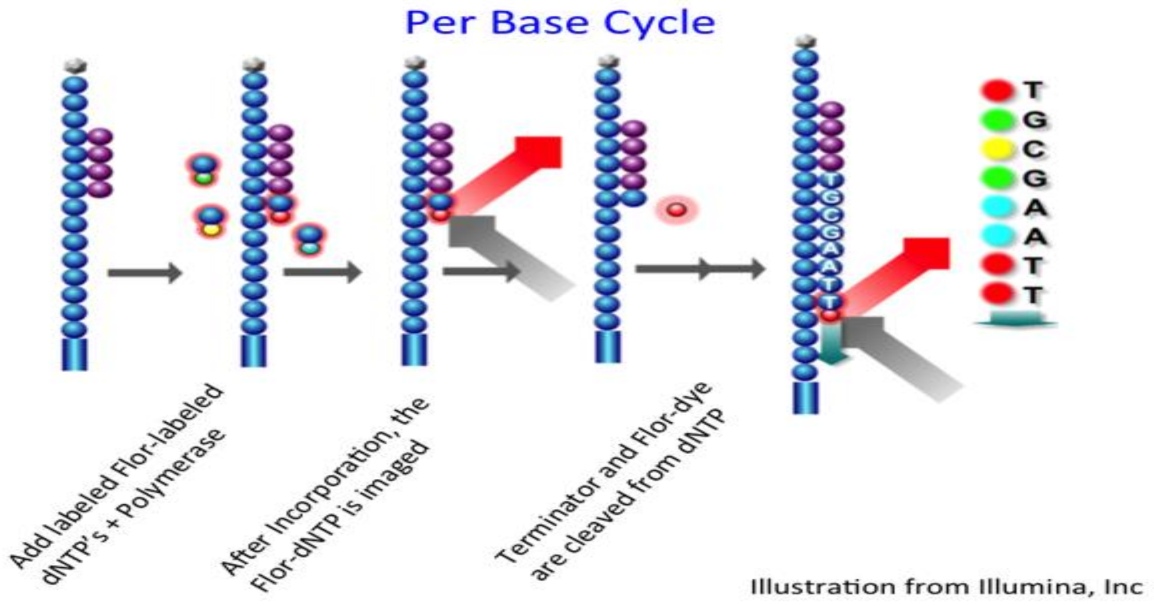

Per Base Cycle
Add labeled Flor-labeled dNTP's + Polymerase
After Incorporation, the Flor-dNTP is imaged
Terminator and Flor-dye are cleaved from dNTP
T
G
C
G
A
A
T
T
Illustration from Illumina, Inc

Figure 3: Sequencing-by-Synthesis

4.5 Applications of NGS:

Clinical Genetics

NGS captures a broader spectrum of mutations than Sanger sequencing. The spectrum of DNA variation in a human genome comprises small base changes (substitutions), insertions and deletions of DNA, large genomic deletions of exons or whole genes and rearrangements such as inversions and translocations. Traditional Sanger sequencing is restricted to the discovery of substitutions and small insertions and deletions. For the remaining mutations dedicated assays are frequently performed, such as fluorescence in situ hybridisation (FISH) for conventional karyotyping, or comparative genomic hybridisation (CGH) microarrays to detect submicroscopic chromosomal copy number changes such as microdeletions. However, these data can also be derived from NGS sequencing data directly, obviating the need for dedicated assays while harvesting the full spectrum of genomic variation in a single experiment. The only limitations reside in regions which sequence poorly or map erroneously due to extreme guanine/cytosine (GC) content or repeat architecture, for example the repeat expansions underlying Fragile X syndrome or Huntington's disease.

Microbiology

The main utility of NGS in microbiology is to replace conventional method of characterisation of pathogens by morphology, metabolic criteria and staining properties with a genomic definition of pathogens. The genomes of pathogens defines about them, may harbour information about drug sensitivity and inform the relationship of different pathogens with each other which can be used to trace sources of infection and outbreaks. The last recently received media attention, when NGS was used to reveal and trace an outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) on a neonatal intensive care unit in the UK [7]. The most remarkable was that routine microbiological surveillance did not show that the cases of MRSA that occurred over several months were related. NGS of the pathogens, however allowed precise characterisation of the MRSA isolates and revealed a protracted outbreak of MRSA which could be traced to a single member of staff.

Oncology

The fundamentals of cancer genomics is that cancer is caused by somatically acquired mutations, and consequently it is a disease of the genome. Although capillary-based cancer sequencing has been ongoing for over a decade, these investigations were limited to relatively very few samples and small numbers of candidate genes. With the advent of NGS, cancer genomes can now be systemically studied in their entirety, an endeavour ongoing via several large scale cancer genome projects around the world, including a dedicated paediatric cancer genome project. For the child suffering from cancer this may provide many benefits including a more precise diagnosis and classification of the disease, more accurate prognosis and potentially the identification of ‘drug-able’ causal mutations. Individual cancer sequencing may therefore, provide the basis of personalised cancer management. Currently pilot projects are underway using NGS of cancer genomes in clinical practice and mainly aiming to identify mutations in tumours that can be targeted by mutation-specific drugs.

It is hard to overstate the importance of DNA sequencing to biological research; at the most fundamental level it is how we measure one of the major properties by which terrestrial life forms can be defined and differentiated from each other. Therefore over the last half century many researchers from around the globe have invested a great deal of time and resources to developing and improving the technologies that underpin DNA sequencing. At the genesis of this field, working primarily from accessible RNA targets, researchers would spend years laboriously producing sequences that might number from a dozen to a hundred nucleotides in length. Over the years, innovations in sequencing protocols, molecular biology and automation increased the technological capabilities of sequencing while decreasing the cost, allowing the reading of DNA hundreds of base pairs in length, massively parallelised to produce gigabases of data in one run. Researchers moved from the lab to the computer, from pouring over gels to running code. Genomes were decoded, papers published, companies started and often later dissolved – with repositories of DNA sequence data growing all the while. Therefore DNA sequencing – in many respects a relatively recent and forward focussed research discipline – has a rich history. An understanding of this history can provide appreciation of current methodologies and provide new insights for future ones, as lessons learnt in the previous generation inform the progress of the next.

5. Genetics

Genetics is the study of genes, genetic variation, and heredity in living organisms. It is generally considered a field of biology, but intersects frequently with many other life sciences and is strongly linked with the study of information systems.

The father of genetics is Gregor Mendel, a late 19th-century scientist and Augustinian friar. Mendel studied "trait inheritance", patterns in the way traits are handed down from parents to offspring. He observed that organisms (pea plants) inherit traits by way of discrete "units of inheritance". This term, still used today, is a somewhat ambiguous definition of what is referred to as a gene. Trait inheritance and molecular inheritance mechanisms of genes are still primary principles of genetics in the 21st century, but modern genetics has expanded beyond inheritance to studying the function and behavior of genes. Gene structure and function, variation, and distribution are studied within the context of the cell, the organism (e.g. dominance), and within the context of a population. Genetics has given rise to a number of subfields, including epigenetics and population genetics. Organisms studied within the broad field span the domains of life (archaea, bacteria, and eukarya).

Genetic processes work in combination with an organism's environment and experiences to influence development and behavior, often referred to as nature versus nurture. The intracellular or extracellular environment of a cell or organism may switch gene transcription on or off. A classic example is two seeds of genetically identical corn, one placed in a temperate climate and one in an arid climate. While the average height of the two corn stalks may be genetically determined to be equal, the one in the arid climate only grows to half the height of the one in the temperate climate due to lack of water and nutrients in its environment.

5.1 Mendelian Genetics/ Disorders

Mendel’s studies of inheritance patterns in pea plants are a solid foundation for our current understanding of single-gene diseases in humans. Also called Mendelian or monogenic diseases, these kinds of diseases are caused by mutations in one gene, and they sometimes run in families. Pedigree analyses of large families with many affected individuals can be used to determine whether a disease-associated gene is located on an autosome or on a sex chromosome, and whether the related disease phenotype is dominant or recessive.

5.2 Inheritance pattern

The basic laws of inheritance are important in understanding patterns of disease transmission. The inheritance patterns of single gene diseases are often referred to as Mendelian since Gregor Mendel first observed the different patterns of gene segregation for selected traits in garden peas and was able to determine probabilities of recurrence of a trait for subsequent generations. If a family is affected by a disease, an accurate family history will be important to establish a pattern of transmission. In addition, a family history can even help to exclude genetic diseases, particularly for common diseases where behavior and environment play strong roles.

Most genes have one or more versions due to mutations or polymorphisms referred to as alleles. Individuals may carry a ‘normal’ allele and/or a ‘disease’ or ‘rare’ allele depending on the impact of the mutation/polymorphism (e.g., disease or neutral) and the population frequency of the allele. Single-gene diseases are usually inherited in one of several patterns depending on the location of the gene and whether one or two normal copies of the gene are needed for the disease phenotype to manifest.

The expression of the mutated allele with respect to the normal allele can be characterized as dominant, co-dominant, or recessive. There are five basic modes of inheritance for single-gene diseases: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive, and mitochondrial.

Autosomal Dominant Inheritance

- In autosomal dominant inheritance, only one copy of a disease allele is necessary for an individual to be susceptible to expressing the phenotype.
- With each pregnancy, there is a one in two (50%) chance the offspring will inherit the disease allele.
- Unless a new mutation has occurred, all affected individuals will have at least one parent who carries the disease allele.
- Autosomal dominant inheritance is often called vertical inheritance because of the transmission from parent to offspring.
- Across a population, the proportion of affected males should be equal to the proportion of affected females.
- Male-to-male transmission can be observed.

Examples:

Huntington's disease: This is a progressive neurodegenerative disorder, is a well-known example of an autosomal dominant single-gene disease; most individuals with a single copy of the mutant huntingtin gene (*HTT*) will have Huntington's disease later in life. Typically, autosomal dominant diseases affect individuals in their early years and prevent them from living past infancy or childhood, which in turn precludes these individuals from reproducing and potentially passing on the mutation to their offspring. In the case of Huntington's disease, however, the late onset of the disorder means that many affected individuals have already had children before they are even aware that they carry the mutation. Disease-associated changes in the huntingtin gene consist of a special type of mutation called triplet repeats; these mutations are simply extra repetitions of the three-base DNA sequence CAG. The number of CAG repeats in a mutated huntingtin gene determines the age at which a person will develop Huntington's disease, as well as how severe the condition will be. Genetic tests can be used to determine how many CAG repeats are in an individual's huntingtin gene, thereby providing a highly accurate assessment of the individual's disease risk. Because affected parents have a 50% chance of passing a mutant copy of the huntingtin gene on to each of their offspring, children of people with Huntington's disease are often faced with the dilemma of whether to undergo such testing. Genetic testing can either provide immediate relief in knowing that one is free from the disease, or the confirmation that one will certainly suffer from the condition at some point in the future.

Myotonic dystrophy, familial hypercholesterolemia, neurofibromatosis, and polycystic kidney disease serve as additional examples of autosomal dominant single-gene diseases. Myotonic dystrophy is associated with dominant mutations in the dystrophia myotonica protein kinase (*DMPK*) gene; familial hypercholesterolemia is associated with dominant mutations in both the low-density lipoprotein receptor (*LDLR*) gene and the apolipoprotein B (*APOB*) gene; and neurofibromatosis is associated with dominant mutations in the neurofibromin (*NF1*) gene. Autosomal dominant polycystic kidney disease can be caused by mutations in either the polycystic kidney disease 1 (*PKD1*) gene or the polycystic kidney disease 2 (*PKD2*) gene; the *PKD1*gene is located on human chromosome 16, whereas the *PKD2* gene is located on human chromosome 4.

Autosomal Recessive Inheritance

- In autosomal recessive inheritance, two copies of a disease allele are required for an individual to be susceptible to expressing the phenotype.
- Typically, the parents of an affected individual are not affected but are gene carriers.
- With each pregnancy of carrier parents:
    - There is a one in four (25%) chance the offspring will inherit two copies of the disease allele and will therefore have the phenotype.
    - There is a one in two (50%) chance the offspring will inherit one copy of the disease allele and will be a carrier.
    - There is a one in four (25%) chance the offspring will inherit no copies of the disease allele and will not express the phenotype or be a carrier. This individual would not be at risk for passing the disorder on to his/her offspring.
- As with autosomal dominant inheritance, the proportion of affected males should be equal to the proportion of affected females in a given population.

Examples:

Phenylketonuria (PKU): PKU is a prominent example of a single-gene disease with an autosomal recessive inheritance pattern. PKU is associated with mutations in the gene that encodes the enzyme phenylalanine hydroxylase (PAH); when a person has these mutations, he or she cannot properly manufacture PAH, so he or she is subsequently unable to break down the amino acid phenylalanine, which is an essential building block of dietary proteins. As a result, individuals with PKU accumulate high levels of phenylalanine in their urine and blood, and this build-up eventually causes mental retardation and behavioural abnormalities.

Several other human diseases, including cystic fibrosis, sickle-cell anemia, and oculocutaneous albinism, also exhibit an autosomal recessive inheritance pattern. Cystic fibrosis is associated with recessive mutations in the *CFTR* gene, whereas sickle-cell anemia is associated with recessive mutations in the beta hemoglobin (*HBB*) gene. Interestingly, although individuals homozygous for the mutant *HBB* gene suffer from sickle-cell anemia, heterozygous carriers are resistant to malaria. This fact explains the higher frequency of sickle-cell anemia in today's African Americans, who are descendants of a group that had an advantage against endemic malaria if they carried *HBB* mutations. Finally, oculocutaneous albinism is associated with autosomal recessive mutations in the *OCA2* gene.

This gene is involved in biosynthesis of the pigment melanin, which gives color to a person's hair, skin, and eyes.

X-Linked Dominant Inheritance

- As in autosomal dominant inheritance, only one copy of a disease allele on the X chromosome is required for an individual to be susceptible to an X-linked dominant disease.
- Both males and females can be affected, although males may be more severely affected because they only carry one copy of genes found on the X chromosome. Some X-linked dominant disorders are lethal in males.
- When a female is affected, each pregnancy will have a one in two (50%) chance for the offspring to inherit the disease allele. When a male is affected, all his daughters will be affected, but none of his sons will be affected.
- Examples of diseases with X-linked dominant inheritance are hypophosphatemic ricketsm, oral-facial-digital syndrome type I, and Fragile X syndrome.

Examples:

Examples of X chromosome-linked dominant diseases are rare, but several do exist. For instance, dominant mutations in the phosphate-regulating endopeptidase gene (*PHEX*), which resides on the X chromosome, are associated with X-linked dominant hypophosphatemic rickets. Similarly, Rett syndrome, a neurodevelopmental disease, is associated with dominant mutations in the methyl-CpG-binding protein 2 gene (*MECP2*). Rett syndrome almost exclusively affects females, because male embryos with a dominant mutation in the *MECP2* gene rarely survive.

X-Linked Recessive Inheritance

- As in autosomal recessive inheritance, two copies of a disease allele on the X chromosome are required for an individual with two X chromosomes (a female) to be affected with an X-linked recessive disease.
- Since males are hemizygous for X-linked genes (they have only one X chromosome), any male with one copy of an X-linked recessive disease allele is affected.
- Females are usually carriers because they only have one copy of the disease allele. Affected males are related through carrier females.

- For a carrier female, with each pregnancy there is a one in two (50%) chance her sons will inherit the disease allele and a one in two (50%) chance her daughters will be carriers.
- Affected males transmit the disease allele to all of their daughters, who are then carriers, but to none of their sons.
- Women are affected when they have two copies of the disease allele. All of their sons will be affected, and all of their daughters will be unaffected carriers.

Examples:

Single-gene diseases that involve genes found on the sex chromosomes have somewhat different inheritance patterns than those that involve genes found on a person's autosomes. The reason for these differences lies in the genetic distinction between males and females. Recall that females have two copies of the X chromosome, and they receive one copy from each parent. Therefore, females with an X chromosome-linked recessive disease inherit one copy of the mutant gene from an affected father and the second copy of the mutant gene from their mother, who is most often a carrier (heterozygous) but who might be affected (homozygous). Males, on the other hand, have only one copy of the X chromosome, which they always receive from their mother. Therefore, males with an X chromosome-linked disease always receive the mutant copy of the gene from their mother. Moreover, because men don't have a second copy of the X chromosome to potentially "cancel out" the negative effects of X-linked mutations, they are far more likely than women to be affected by X chromosome-linked recessive diseases.

The blood-clotting disorder hemophilia A is one of several single-gene diseases that exhibit an X chromosome-linked recessive pattern of inheritance. Males who have a mutant copy of the factor VIII gene (*F8*) will always have hemophilia. In contrast, women are rarely affected by this disease, although they are most often carries of the mutated gene. Duchenne muscular dystrophy is another example of a single-gene disease that exhibits an X chromosome-linked recessive inheritance pattern. This condition is associated with mutations in the dystrophin gene (*DMD*).

Y–Linked Diseases:

Like X-linked dominant diseases, Y chromosome-linked diseases are also extremely rare. Because only males have a Y chromosome and they always receive their Y chromosome from their father, Y-linked single-gene diseases are always passed on from affected fathers to their sons. It makes no difference whether the Y chromosome-linked mutation is dominant or recessive, because only one copy of the mutated gene is ever present; thus, the disease-associated phenotype always shows.

One example of a Y-linked disorder is nonobstructive spermatogenic failure, a condition that leads to infertility problems in males. This disorder is associated with mutations in the ubiquitin-specific protease 9Y gene (*USP9Y*) on the Y chromosome.

Mitochondrial Inheritance:

- Can affect both males and females, but only passed on by females
- Can appear in every generation

5.3 Disease mechanism:

Loss-of-function mutations

Generally, loss-of-function (null) mutations are found to be recessive. In a wild-type diploid cell, there are two wild-type alleles of a gene, both making normal gene product. In heterozygotes (the crucial genotypes for testing dominance or recessiveness), the single wild-type allele may be able to provide enough normal gene product to produce a wild-type phenotype. In such cases, loss-of-function mutations are recessive. In some cases, the cell is able to “upregulate” the level of activity of the single wild-type allele so that in the heterozygote the total amount of wild-type gene product is more than half that found in the homozygous wild type. However, some loss-of-function mutations are dominant. In such cases, the single wild-type allele in the heterozygote cannot provide the amount of gene product needed for the cells and the organism to be wild type. Example- Sickle cell anemia caused by *HBB* gene, cystic fibrosis caused by *CFTR* gene and Duchenne muscular dystrophy caused by *DMD* gene.

Gain-of-function mutations:

Because mutation events introduce random genetic changes, most of the time they result in loss of function. The mutation events are like bullets being fired at a complex machine; most of the time they will inactivate it. However, it is conceivable that in rare cases a bullet will strike the machine in such a way that it produces some new function. So it is with mutation events; sometimes the random change by pure chance confers some new function on the gene. In a heterozygote, the new function will be expressed, and therefore the gain-of-function mutation most likely will act like a dominant allele and produce some kind of new phenotype. Example-

Haploinsufficiency:

Haploinsufficiency is the phenomenon where a diploid organism has only a single functional copy of a gene (with the other copy inactivated by mutation) and the single functional copy of the gene does not produce enough gene product (typically a protein) to bring about a wild-type condition, leading to an abnormal or diseased state. It is responsible for some but not all autosomal dominant disorders. Haploinsufficiency is therefore an example of incomplete or partial dominance, as a heterozygote (with one mutant and one normal allele) displays a phenotypic effect. It can be contrasted with haplosufficiency, where the single, unmutated allele can produce enough gene product to maintain the wild-type phenotype without the other allele.

Dominant negative:

A mutation whose gene product adversely affects the normal, wild-type gene product within the same cell. This usually occurs if the product can still interact with the same elements as the wild-type product,but block some aspect of its function. Examples: 1. A mutation in a transcription factor that removes the activation domain, but still contains the DNA binding domain. This product can then block the wild-type transcription factor from binding the DNA site leading to reduced levels of gene activation.2. A protein that is functional as a dimer. A mutation that removes the functional domain,but retains the dimerization domain would cause a dominate negative phenotype, because some fraction of protein dimers would be missing one of the functional domains. Example- Marfan syndrome is caused by mutations in the *FBN1* gene.

6. How NGS helps in disease diagnosis

With the complete sequence of the human genome in hand, scientists are now poised to match monogenic disease phenotypes to their corresponding genes. By analysing complex pedigrees, geneticists can correlate changes in gene sequence with particular disease states. After all, once a disease-associated change in the DNA sequence of a gene is identified, it is much easier to determine how the structure of the corresponding gene product (protein) might be changed in a manner that alters its biological function. The nature of disease-associated changes in protein structure and function can in turn enhance our ability to design drugs that effectively and specifically target mutant proteins [9].

Recent estimates predict that the human genome includes 25,000 protein-encoding genes. Although 1,822 of the protein-encoding genes in humans are estimated to be associated with monogenic disease, the identities of more than 1,500 of these genes remain unknown, largely because many of these single-gene diseases are rare and occur in small numbers of families [8]. Referred to as "orphan" diseases, these relatively uncommon disorders receive much less research funding than more common diseases, which are often considered a better investment by funding agencies and pharmaceutical companies. However, many of the common diseases exhibit a more complex inheritance pattern and are associated with mutations in multiple genes (in other words, these conditions are polygenic) [9]. As a result, research efforts have begun to shift from a focus on monogenic disease to a focus on polygenic disease, which can involve complex interactions between genes and the environment that are not easily interpreted.

Mendelian diseases, also known as monogenic diseases, are disorders caused by mutations in one gene and include diseases like thalassemia, cystic fibrosis, among others. Mendelian diseases are considered to be rare individually but collectively occur at a rate of 40 to 82 per 1000 live births, with an estimated 7.9 million children being born annually with a serious birth defect of genetic or partially genetic origin. These disorders tend to run in families, although it has been found that a significant number are caused by de novo events. Depending on the pathomechanism, the phenotype is manifested in a dominant (where one allele is mutated) or recessive (where both alleles are mutated) manner. Of the estimated 20,000–25,000 protein-coding genes in the human genome, mutations in 3348 genes have been associated with Mendelian diseases.

Sanger sequencing, also known as the dideoxy method, has been the gold standard in molecular diagnostics in Mendelian diseases and remains the test of choice for clinical genetic testing; the purpose of which is to confirm a suspected diagnosis and allow more accurate genetic counseling. However, Sanger sequencing can only analyze one DNA segment at a time and is thus laborious and time consuming. For diseases with genetic heterogeneity, like retinitis pigmentosa, cardiomyopathy, and deafness, a gene-by-gene Sanger sequencing approach has not been shown to be economical or efficient.

Recent advances over the past decade have allowed for high-throughput sequencing, and these advances are collectively referred to as next-generation sequencing (NGS). NGS has allowed for substantial increase in sequencing content while dramatically reducing the cost of sequencing per base. This allows for simultaneous interrogation of multiple genes through one single reaction and has been proven to be an effective alternative for establishing the genetic basic of Mendelian diseases in the research setting, and more recently, in the clinical setting.

In this review, we shall aim to discuss the clinical utility of NGS, including the opportunities and challenges that arise from the clinical standpoint. We will also focus our discussion on ways of implementing NGS in the routine clinical workflow.

6.1 Next-generation sequencing: brief overview

The process of NGS starts with extraction of DNA of an individual, most commonly from peripheral leukocytes obtained from blood sample but can be from another tissue such as buccal swab or saliva. The DNA is then broken down into short fragments and amplified using polymerase chain reaction (PCR) or hybridization-based approaches. The regions that are amplified could include either a subset of genes (targeted approach) or all the genes in the genome. When sequencing all the genes, if only the protein-coding regions are amplified, the method is referred to as whole exome sequencing (WES). However, if the target is the entire genome, then it is known as whole genome sequencing (WGS).

These amplified products are then sequenced with the use of one of the various sequencing technologies (e.g., Illumina’s sequencing by synthesis, Life Technologies’ sequencing by ligation, or Ion semiconductor sequencing) to generate millions of short sequence reads. These sequences are then processed bioinformatically. First, they are aligned to a reference genome (assembly) and then compared for similarities and differences at each target position. A list of variants (or differences in the sequence) is then generated, which is then filtered further, to determine the significance of each of the variant. Common filters include rare or previously unreported variants, variants that lead to altered function of the protein, and variants previously reported to cause disease. More recently, algorithms that include the phenotypic information in the variant analysis have been developed that aid the clinician/researcher in narrowing down the candidate gene list.

6.2 Next-generation sequencing in the clinical setting

Targeted vs whole exome sequencing

One of the biggest challenges clinicians are facing is deciding between using targeted versus whole exome sequencing. As the cost of sequencing continues to decrease, WES appears to be a more cost-effective approach. However, there are certain considerations before embarking on one over the other.

Although exomes are supposed to cover the protein-coding regions of the genome, the overall coverage tends to be between 85–95 % only. This means that a particular gene of interest with respect to a specific phenotype may not be covered, either completely or partially. Reasons include poorly performing capture probes due to high GC content, sequence homology, or repetitive sequences. A targeted approach, on the other hand, has a much higher or even complete coverage of all the phenotype-specific genes by filling in the gaps with complementary technologies such as Sanger sequencing or long range PCR. For example, 4 of the 73 genes in a hearing loss panel are inadequately covered on WES but are completely covered in a targeted panel [10].

Besides offering a more comprehensive coverage of the "known" phenotype-specific genes, a targeted approach also allows for deeper coverage of these genes compared to WES, which provides greater confidence in the variants detected. However, both are still prone to sequencing artifacts, and Sanger sequencing of candidate variants is recommended in both approaches before returning the results to the patients.

Lastly, laboratories that offer targeted testing may have expert knowledge for the given phenotype and may be in a better position to prioritize variants detected through NGS. They may also be able to recommend specific evaluations to determine the significance of certain variants; for example, temporal bone evaluation for *SLC26A4* and otoacoustic emission testing for *OTOF* when variants are detected on a targeted hearing loss panel.

Indications and clinical usefulness

NGS is currently indicated for the detection of rare variants in patients with a phenotype suspected to be due to a Mendelian disease. This is done either after single gene testing for candidates has returned negative, or as first line, if there exists genetic heterogeneity where multi-gene Sanger sequencing would be costly and time consuming. Diseases could either include a specific phenotype such as cardiomyopathy, deafness, retinitis pigmentosa, intellectual disability, or be part of a multiple congenital anomaly spectrum (where two or more organs are affected) [11].

The clinical usefulness of performing NGS in the clinical setting varies for different disorders. In the majority of cases, the finding does not alter the clinical management, treatment or prognosis. However, it does allow an end to an expensive and stressful diagnostic odyssey. Reaching a molecular diagnosis allows the clinician to allay the guilt that parents face in the absence of a firm diagnosis and helps them to accept the child's condition.

Identification of the causative variant allows for gene-specific prognostication, based on cases reported in the literature and, in some cases, family support groups. It can also allow for anticipatory management of other comorbidities that an individual may be susceptible to. These include assessing other organs (such as performing blood tests and imaging to assess the heart and kidneys) for possible complications. In some cases, it may redirect therapy to treat the underlying genetic etiology as illustrated by the case of a 15-month-old child with clinical presentation mimicking Crohn's disease. WES revealed a pathogenic variant in *XIAP*, affecting the proinflammatory response and bacterial sensing, predisposing the individual to developing hemophagocytosis. Based on these findings, the child was successfully treated with hematopoietic bone marrow transplantation which cured him of his gastrointestinal disease as well [12].

More importantly, it facilitates genetic counseling and allows for more accurate estimates of recurrence risk in the family. Identification of the molecular etiology allows the clinician to guide subsequent pregnancies, either through prenatal diagnostics or preimplantation genetic diagnosis. In some situations, it also allows for identification of other at-risk family members and any available treatment can be instituted in the presymptomatic phase. For example, identification of mutation in a gene causing long QT syndrome in the proband can allow identification of other at-risk family members, who can then have regular Holter monitoring and, in some instances, implantation of a cardiac pacemaker before a catastrophic event occurs.

Lastly, as our understanding of the molecular pathways and gene-gene interactions improves, it is possible that targeted molecular therapy may be available for a specific genetic mutation that helps to ameliorate the patient's symptoms. For example, in individuals with vascular malformations, somatic mutations in the *AKT3-PI3K-mTOR* pathway have been identified. Some of these patients have been successfully treated with mTOR-inhibitors such as rapamycin. Our ability to offer such targeted therapy will only improve with time as more high-throughput drug screening methods are being deployed.

Gathering information

The process of ordering NGS starts with gathering a detailed family history to determine if there are individuals with similar or related phenotypes within the same family, as well as to assess for possible inheritance pattern. For example, multiple individuals in the same generation and/or history of consanguinity will suggest a recessive pattern, while having affected individuals in each generation would suggest a dominant pattern.

The next step is to perform detailed phenotyping of the affected individual(s). These could include evaluations from other subspecialists and/or performing biochemical and/or radiological tests. For example, a child with a limb anomaly would benefit from cardiac, gastrointestinal, renal, and skeletal evaluations. With the phenotype and pedigree information, a systematic review of literature or syndrome database should be performed to exclude rare but established syndromes [13]. This can then guide the clinician on which gene(s) to test. In cases of genetic heterogeneity, targeted NGS may be the preferred approach. On the other hand, if the disease mechanism is unknown, WES may be the test of choice.

Pretest counseling and informed consent

It is imperative that a patient and his/her family are counseled appropriately by a healthcare professional (such as a clinical geneticist or genetic counselor) who is aware of the nuances of the test. It is important to maintain realistic expectations for the patient and his/her family, as it is possible that the testing may not return any positive results. The diagnostic yield would differ on the test being performed. In a targeted approach, it is possible that the causal variant is in a gene that was not in the subset of genes or regions that were targeted. On the other hand, as the coverage in WES tends to be ~85–95 %, it is possible that the causal variant lies in the region that was either poorly covered or resides outside the protein-coding region of the gene (such as the promoter or regulatory region). The sensitivity and limitations of the specific NGS test must be clearly communicated.

It is also important to emphasize to the families that a positive result may not change treatment or management decisions, and hence, may not alter the prognosis or outcome for the affected individual. But it is also possible that the test may identify targets that may be amenable to certain medications. This has been the case in the field of oncology, where specific molecular targets are used to treat certain subtypes of cancer.

Cost remains an important consideration as clinical targeted NGS or WES can cost between USD 2000 to 10,000 and USD 5000 to 15,000, respectively [13]. Insurance companies may not approve such costly tests, and in self-payer healthcare systems (such as Singapore) where the individual has to bear the cost, this may be prohibitively expensive.

Lastly, the patient should be advised that variants which are not related to the primary phenotype, also referred to as secondary or incidental findings (IFs), may be detected. Detection of IFs is more common with WES and WGS and may be less of an issue with targeted panels. This may have implications not only for the individual, but also his/her family members. For example, WES on a patient with intellectual disability may detect pathogenic variants in *BRCA1*. This mutation could be inherited from the parents, which would mean that the affected parent (and his/her sibling) is at risk of developing cancer and would require surveillance and monitoring. This family may or may not be prepared to receive such information and this should be discussed during the informed consent process. While genetic nondiscrimination act (GINA 2008) protects individuals against discrimination in the USA, similar laws are lacking in other countries. Hence, individuals with a genetic diagnosis, especially those with incidental findings, may face discrimination at work or be denied medical insurance without any avenue for legal redress.

Interpretation of results

On average, ~60,000 to 100,000 variants are detected on WES. These variants can be broadly classified into pathogenic, benign, or variants of uncertain significance (VUS). Pathogenic variants are defined as those variants that adversely alter protein function and have either been reported previously in other affected individuals or have been shown to affect protein function in cellular or animal models. These include variants such as nonsense, frameshift, splicing, small insertion-deletions (indels), or nonsynonymous missense variants. Benign variants, also known as polymorphisms, are variants that exist in a significant proportion of the population, including healthy individuals, and account for majority of the variants detected on NGS testing. These include synonymous missense variants, intronic, or intergenic variants [14].

VUS are variants that could possibly affect protein function based on *in silico* software (such as Polyphen-2 and SIFT) or other similar parameters, but either have not been described in other individuals (affected or unaffected) or do not have any functional analysis in other model systems. As there are genomic variations across different populations and ethnic groups, a database of common variants from healthy individuals is imperative in understanding the significance of a given variant. Such databases include dbSNP, 1000 Genomes project, Exome Aggregation Consortium (ExAC), and Exome Sequencing Project (ESP), but certain ethnicities are underrepresented in these databases. Interpretation of VUS in such instances requires segregation analysis (by analyzing the variant in other family members)—presence of the variant in affected but not in unaffected family members adds further evidence to the possible causal relationship of a given variant. An alternative approach includes testing the variant in cellular and/or animal models, but this is beyond the realms of a clinic or clinical laboratory and may be best addressed through a research laboratory.

There is a wide range of possible outcomes with NGS testing. Using WES, a single pathogenic variant that is likely to be the cause of the disease can be detected about 20–36 % of the time [15]. For the remainder of the cases, it is possible to either find multiple candidate variants or none at all. In the event that multiple candidate variants are detected, segregation analysis and/or functional analysis would help to determine the molecular etiology. If no candidate variants are found, possibilities include poor coverage or the mutation residing outside the protein-coding region of the gene or the defect is not due to a simple nucleotide change in a single gene.

Delivery of results

Clinicians should review the results of NGS and correlate the findings with the relevant medical information. When a pathogenic variant is detected in a gene that explains the patient’s clinical phenotype, the clinician should review the results as well as relevant clinical information, including inheritance, prognosis, complications, or management, with the patient and his/her family. Testing of at-risk individuals should be offered, when possible.

On other occasions, finding of a particular variant may prompt the clinician to look for additional history or perform tests which may or may not support the conclusion of the NGS test report. For example, in a patient with chronic diarrhea, finding biallelic mutations in *TTC37* may prompt the clinician to review the hair for specific features of trichorrhexis nodosa, suggestive of a diagnosis of tricho-hepato-enteric syndrome [16]. On the other hand, if more than one candidate variant is detected, the clinician will need to perform further evaluation(s) to determine which of the variant is causing the phenotype.

Lastly, if the test results are negative, reasons for this should be discussed with the patient (as discussed in the pretest counseling). In addition, as our understanding of the human genome improves and more similar cases are reported, there may be a possibility of associating the patient's phenotype to a newly described genetic syndrome. This requires reanalysis of the both clinical reports and genetic data at regular intervals. However, regulations regarding who and how to perform reanalysis are currently lacking and remain a challenge for the near future.

7. NGS based clinical diagnosis at MedGenome

Methodology-

1. Collection of the sample.
2. DNA extractions
3. Library preparation from DNA sample
4. Next generation sequencing on sequencing platform
5. Bioinformatics analysis
6. Evaluation and interpretation of result.

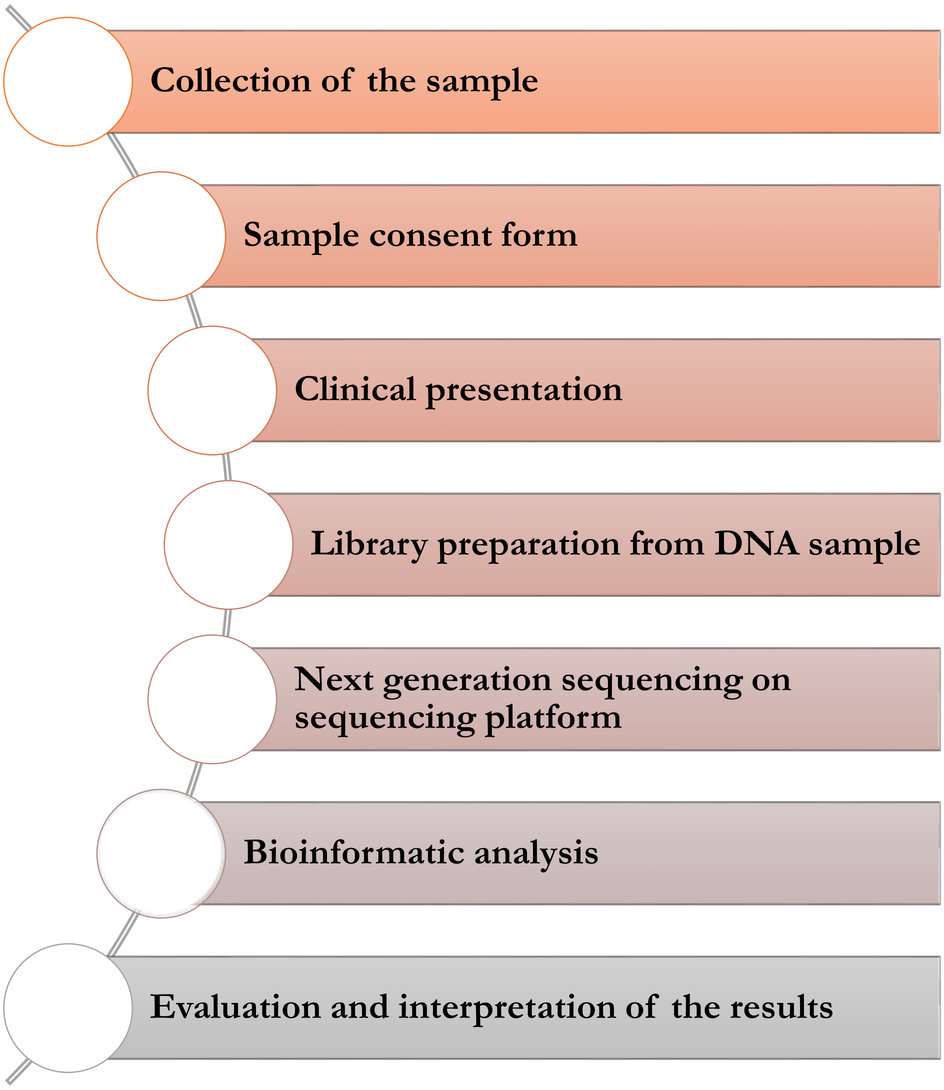

Collection of the sample
Sample consent form
Clinical presentation
Library preparation from DNA sample
Next generation sequencing on sequencing platform
Bioinformatic analysis
Evaluation and interpretation of the results

Fig 4: Flow chart of procedure

7.1 Collection of the sample

Blood sample is collected from the patient and are stored in EDTA tube or in PAX gene tube and brought into laboratory for further processing.

7.2 DNA extractions

DNA extraction involves separating the nucleic acids in a cell away from proteins and other cellular materials. Post-extraction filtration is sometimes used to concentrate low amounts of recovered DNA sample. It is important with any DNA extraction technique to remove as many substances as possible that could interfere with downstream testing and cause the extracted DNA molecules to break down over time. Although the addition of buffer components that can overcome inhibition, direct PCR now permits by-passing the DNA extraction and quantitation steps**.**

7.3 Library preparation from DNA sample

Library preparation for the major next generation sequencing platforms requires the ligation of specific adaptor oligos to fragments of the DNA to be sequenced. First, DNA is fragmented to the optimal length determined by the downstream platform. Because DNA fragmentation does not result in homogeneous, blunt-ended fragments, end repair is needed to ensure that each molecule is free of overhangs and contains 5′ phosphate and 3′ hydroxyl groups. Libraries to be used in blunt-ended adaptor ligation, including Ion Torrent or SOLiD 4 library construction, can be used directly in the ligation step. For Illumina libraries and some libraries intended for the 454 platform, incorporation of a non-templated deoxyadenosine 5′-monophosphate (dAMP) onto the 3′ end of blunted DNA fragments, a process known as dA-tailing, is necessary. dA-tails prevent concatamer formation during downstream ligation steps and enable DNA fragments to be ligated to adaptors with complementary dT-overhangs. The desired adaptor ligated DNA size for Illumina, SOLiD and Ion Torrent platforms can be selected via gel electrophoresis before amplification by the polymerase chain reaction (PCR).

7.4 Next generation sequencing on sequencing platform

The Illumina sequencing method is similar to Sanger sequencing, but it uses modified dNTPs containing a terminator which blocks further polymerization- so only a single base can be added by a polymerase enzyme to each growing DNA copy strand. The sequencing reaction is conducted simultaneously on a very large number (many millions in fact) of different template molecules spread out on a solid surface. The terminator also contains a fluorescent label, which can be detected by a camera. Only a single fluorescent color is used, so each of the four bases must be added in a separate cycle of DNA synthesis and imaging. Following the addition of the four dNTPs to the templates, the images are recorded and the terminators are removed. This chemistry is called “reversible terminators”. Finally, another four cycles of dNTP additions are initiated. Since single bases are added to all templates in a uniform fashion, the sequencing process produces a set of DNA sequence reads of uniform length.

Although the fluorescent imaging system used in Illumina sequencers is not sensitive enough to detect the signal from a single template molecule, the major innovation of the Illumina method is the amplification of template molecules on a solid surface. The DNA sample is prepared into a “sequencing library” by the fragmentation into pieces each around 200 bases long. Custom adapters are added to each end and the library is flowed across a solid surface (the “flow cell”) and the template fragments bind to this surface. Following this, a solid phase “bridge amplification” PCR process (cluster generation) creates approximately one million copies of each template in tight physical clusters on the flowcell surface. Illumina has improved its image analysis technology dramatically which allows for higher cluster density on the surface of the flowcell.

7.5 Bioinformatics analysis

The analyzed data for the sample is shared by SFTP (Secure File Transfer Protocol) and a mail intimating the same is sent to the genome analyst group. The mail contains the path of the FTP files, raw and analyzed BAM files.

Example: FTP Path: /ftp_uploads/files/54394_R760_LIB27819

This FTP folder contains

1. Quality metrics report of the sample data
2. Variant annotated file “Varimat” compatible with VarMiner (in house software tool). The sample is named with the version-id
3. Coverage of the coding region of the genes
4. Low coverage regions of the genes
5. Variants reported in HGMD
6. Suspected large indels (annotated output from Pindel)
7. Suspected copy number variants and structural variants file

Example: working directory

/MGMSTAR1/SHARED/ANALYSIS/R760_P8193_LIB27819_54394 (All analyzed files including *BAM files are present in this working directory)

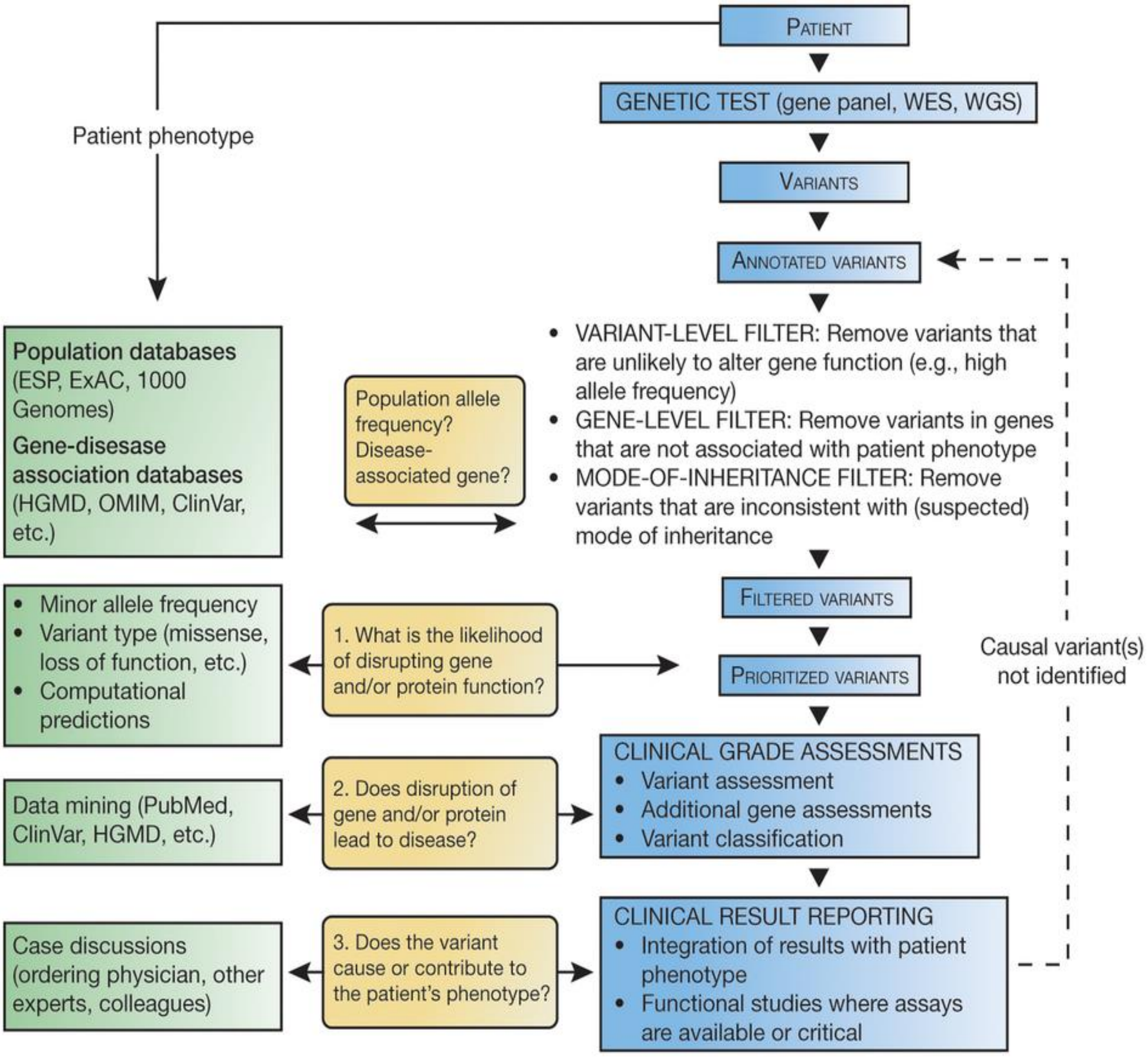
The *Bam files available in this path are used as input files by secondary visualizing software's like IGV (Integrative Genomics Viewer - Broad Institute) for checking the quality of the variants described below in section. The complete details of the files shared in this directory can be referred using the Bioinformatics SOP document.

Checking data quality for compliance:

The quality metrics shared by the bioinformatics team is thoroughly checked for compliance, following which the sample is taken forward for analysis. Given below are the parameters which are examined. It also has details of the library id, run id and software and version of the tools employed for data analysis.

1. Data generated >= expected data (The expected data varies according to the panel for eg. 4Gb for clinical exome panel, 1.5 Gb for targeted gene panel)
2. Greater than 80 % of the data to have ≥ Q30
3. Percentage of aligned Reads is greater than 95 %
4. Percentage of passed alignment is greater than 90 %
5. Overall duplicate (%) is less than 10 %
6. Panel Average Depth is greater than 80x
7. Predicted gender of the analyzed sample matches that with the clinical information
8. Total number of variants in the sample is >= 200,000 variants in clinical exome panel and >= 80,000 in case of targeted gene analysis
9. Variant summary
10. Predicted gender to be compared to that provided in the TRF

7.6 Evaluation and interpretation of result



Patient
GENETIC TEST (gene panel, WES, WGS)
Variants
Annotated variants
Patient phenotype
Population databases (ESP, ExAC, 1000 Genomes)
Gene-disesase association databases (HGMD, OMIM, ClinVar, etc.)
Population allele frequency? Disease-associated gene?
VARIANT-LEVEL FILTER: Remove variants that are unlikely to alter gene function (e.g., high allele frequency)
GENE-LEVEL FILTER: Remove variants in genes that are not associated with patient phenotype
MODE-OF-INHERITANCE FILTER: Remove variants that are inconsistent with (suspected) mode of inheritance
Filtered variants
Prioritized variants
Causal variant(s) not identified
Minor allele frequency
Variant type (missense, loss of function, etc.)
Computational predictions
1. What is the likelihood of disrupting gene and/or protein function?
Data mining (PubMed, ClinVar, HGMD, etc.)
2. Does disruption of gene and/or protein lead to disease?
CLINICAL GRADE ASSESSMENTS
Variant assessment
Additional gene assessments
Variant classification
Case discussions (ordering physician, other experts, colleagues)
3. Does the variant cause or contribute to the patient's phenotype?
CLINICAL RESULT REPORTING
Integration of results with patient phenotype
Functional studies where assays are available or critical

Fig 5: Flow chart of analysis

The varminer is a user interface software which enables efficient analysis of the samples based on predefined parameters (defined filters), which can also be modified

Content: Variant annotation grid displays the data columns from the VariMat file such as chromosome number, variant start position, base change, variant type, zygosity, read depth of the variant, gene in which variant is found, cDNA change, protein change, exon number, Ensembl transcript ID, RefSeq transcript ID, RefSeqProtein ID, EnsemblProtein ID, ClinVar ID, SwissVar, 1000G MAF, ExAC MAF, SIFT, PolyPhen2 HDiv, PolyPhen2_HVar, LRT, VariantQuality, Reference base depth, Altered base depth, disease database, MedVar database Samples (variant in internal samples database), MedVarDb MAF (internal samples database) etc.

Hyperlinks: Some of the columns contents are hyperlinked to external databases to enable better analysis and to make informed decision

- **Gene name** is hyperlinked to OMIM database (Online Mendelian Inheritance in Man, An Online Catalog of Human Genes and Genetic Disorders) which provides information on the associated disease with variants in that gene. OMIM focuses on the relationship between phenotype and genotype.
- **Variant position** is hyperlinked to University of California, Santa Cruz (UCSC) Genome Browser, which contains the reference sequence and working draft assemblies for a large collection of genomes. It is an interactive website offering access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms, integrated with a large collection of aligned annotations. The UCSC filters can be modified to access data with relevant to the variant
- **SwissVar, Clinvar, rsIDs etc** is hyperlinked to their respective sources. Thus the significance of the variants and their literature reference can be confirmed. Additional details with respect to variant can also be obtained from these sites.

Filters: The results obtained depends solely on the basis of the filters and their parameters used. Thus it is essential to understand the biology of the disease, mode of inheritance, commonly observed mutations and prevalence for analysis. Most of the column content in the varminer can be filtered based on the applied parameters. Find below an example of a predefined basic filter.

Panel filter: Entering the panel name in the “Disease Filters” tab populates the gene list corresponding to the panel and can be used for filtering. These gene list corresponds to that offered for a particular test e.g. muscular dystrophy and congenital myopathy, deafness panel, MODY panel, Charcot-Marie-Tooth disease panel etc.

Included genes: Specific genes to be checked can be manually added to this tab

Excluded genes: During the process of analysis, gene names can be excluded by entering the list in this tab (for e.g. genes that have been checked and excluded)

OMIM filters: Entering the disease phenotype and diagnosis in this tab populates the genes reported in OMIM for the phenotype entered.

Synopsis Filters: Entering the disease phenotype and diagnosis in this tab populates all the genes based on OMIM clinical synopsis.

The above filters can be used in combination of two or more based on the requirement (For e.g. basic filter and synopsis filter). In addition to the above filters, column filters are also available, which can be used to exclude or include specific parameters. For example only “Homozygous” variants can be filtered or only “missense” variants can be filtered. The understanding of the contents of the varimat file and filters of the Varminer is essential in performing efficient analysis.

Criteria for filtering and prioritization of the variants for reporting:

The ultimate goal of filtering and prioritization is to weigh the available evidence in favor of a likely role of the genetic variants detected in a patient.

1. **Phenotype / clinical indications / diagnosis:** The variants prioritized for reporting is mainly based on clinical indications and diagnosis. The clinical indications given in the TRF is are duly read and understood with respect to the genetic basis of the disorder and gene list is selected accordingly.
2. **Assumed mode of inheritance based on pedigree or family history or phenotype:** Based on given clinical synopsis or diagnosis or any specific phenotype the mode of inheritance of the disease is studied and included. For e.g. some phenotypes are very well known mode of inheritance. Cystic fibrosis is inherited in an autosomal recessive mode.

   If pedigree and details of other affected and unaffected members are given, these information can also be used to infer the inheritance of the disease for that particular family
3. **Minor allele frequency:** The variants detected in the clinically relevant genes are systematically prioritized to distinguish baseline polymorphisms from clinically significant variants based on minor allele frequencies in the population databases (1000 genomes, EXAC, internal database) whose threshold value depends on the prevalence of the disease condition under study. In principle, when a variant is more common than the incidence of a disease (disease penetrance is taken into consideration) the variant is unlikely to be disease causing. Conversely, when a variant rare in the general population, the variant is likely to cause disease. However, rarity alone is not sufficient to indicate pathogenicity.
4. **Reported status:** The clinically relevant variant(s) are checked whether is reported in literature for given phenotype, by using the following databases.

   Clinvar - http://www.ncbi.nlm.nih.gov/clinvar/

SwissVar - http://swissvar.expasy.org/

HGMD - http://www.hgmd.cf.ac.uk/ac/index.php

If the variant is reported in any of these databases with the suspected or similar phenotype, the variant is prioritised for reporting.

5. **Prediction based on *in-silico* tools:** The current ACMG guidelines suggests that multiple lines of computational evidence support deleterious effect on the gene. Thus in our analysis we use five different predictive tools SIFT, PolyPhen2, Mutation Taster2 and LRT in assigning significance and clinical correlation for reporting. The in silico function prediction algorithms assesses the effect of single nucleotide variations on the protein and their prediction may vary based on the criteria and parameters used. For example SIFT uses median conservation score to measure protein conservation whereas PolyPhen2 also uses a normalized cross-species conservation score and combines this with a variety of protein structural features. Thus the predictive accuracy, sensitivity, specificity, true positive and true negative call percentages varies among these tools. The results from these tools may also vary in accordance with the particular edition/version used. Thus we have included the version details and source of the in silico tools in the "Test Methodology" section of the report ("The in silico predictions are based on Variant Effect Predictor tool, Ensembl release 87 (SIFT version - 5.2.2; PolyPhen - 2.2.2); LRT version - November, 2009 release from dbNSFPv3.1 and Mutation Taster2 based on build NCBI 37 / Ensembl 69"), which can be sometimes different from the online/real time based predictions as they are updated more frequently. (Any updates in the version or any additional computational tools used will be updated in the "Test Methodology" section of the clinical report. It is to be noted that the *in silico* predictions of the online versions of the tools may vary due to the difference in the version as they may be more frequently updated.)
6. **Genotype-Phenotype correlation:** The prioritised variants based on all the above criteria is checked for genotype-phenotype correlation to ascertain the significance of the variant (s).

PINDEL analysis: The size of the indels called in GATK tool is only around 10bp. Thus to identify large indels PINDEL tool is used. The variants are annotated similar to that of the GATK VCF file using VEP and other annotations as in Varimat file. This annotated PINDEL_varimat file can be uploaded onto the varminer for analysis and prioritization of variants as described above.

Integrative Genomics Viewer (IGV): The variants thus prioritized for reporting are visualized in the IGV, a high-performance visualization tool for interactive exploration of variants viewer. This aligns the BAM file of the variant against the reference sequence of NCBI. This tool is freely available and is installed in the laptop. The final BAM file used for variant calling is uploaded on to the tool for visualizing the variant. It helps us to identify skewed strand ratio, bad quality reads, allele change and visualize depth. A screen shot a variant is depicted in the picture below.

Variant Interpretation:

Variant selected after filtering based on the above criteria in a sample is evaluated and classified according to ACMG guidelines. The variants sequence change variants are weighed as follows

Very Strong:

- Variant reported in previous studies with substantial evidence that the variant causes the disease in question
- Variants in a gene where loss of function is a known mechanism of disease, including: nonsense or frameshift changes including interruption of the normal start or stop codon, alteration of a splice donor or acceptor site.

Strong:

- Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
- De novo (both maternity and paternity confirmed) in a patient with the disease and no family history
- Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product
- The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls

Moderate:

- Located in a mutational hot spot and/or critical and well-established functional domain
- Absent from controls population
- Detected in trans with a pathogenic variant for recessive disorders
- Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants

- Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before
- Assumed de novo, but without confirmation of paternity and maternity

Supporting:

- Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease
- Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease
- Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)
- Patient's phenotype or family history is highly specific for a disease with a single genetic etiology
- The variant detected is checked for its presence in HGMD database, and also other disease databases like LOVD and UMD. The publication (pubmed indexed) for the reported variant is checked for its clinical relevance, functional studies, presence in control/ unaffected subjects and segregation studies.

Variant Classification – ACMG guidelines:

After sequence change has been reviewed it is classified into these categories:

Pathogenic: A disease causing variation in a gene which can explain the patients' symptoms has been detected. This usually means that a suspected disorder for which testing had been requested has been confirmed.

Likely Pathogenic: A novel or previously reported variants, with some evidence suggesting that this is very likely to contribute to the development of disease however, the scientific evidence is currently insufficient to prove this conclusively.

Benign: A variant that are present at a higher frequency in the general population than expected for that disease and which is known not to be responsible for disease has been detected. Generally no further action is warranted on such variants when detected.

Likely Benign: This category includes variants in intronic region that are unlikely to affect splicing, variants in highly variable region of a gene, variants for which multiple lines of computational evidence suggest no impact on gene/gene product, well-established functional studies show no deleterious effect and variants not segregating with the disease in a family

Variant of uncertain significance: A variant has been detected, such as missense variant, inframe deletions/insertions, but it is difficult to classify it as either pathogenic (disease causing) or benign (non-disease causing) based on current available scientific evidence. It is probable that their significance may change, subject to availability of scientific evidence.

Report Generation:

A clinical report is generated for a given patient based on the variant classification and other demographical information.

A report format consists of mainly following heads:

Table 1: The contents in the first table include details of the subject; referring clinician; test requested and sample details.

Clinical Diagnosis / Symptoms / History: Details given in TRF are added in this section of the report, which includes patient details, family history, clinical symptoms/indications shown by the patient, lab test results and any differential diagnosis or suspected disease.

Results: The variant details are entered in a tabular form as shown below

This section consists of three parts; annotation as per HGVS annotation; corresponding disease information from OMIM database and significance of the observed variant derived based on ACMG guidelines. The HGVS notation is obtained from the varminer annotated file and variant class is confirmed by using the web based internal tool and the values are saved in the databas. The web based tool is based on ACMG guidelines 2015. Transcript used for reporting the observed variant is according the Transcript SOP.

Variant Interpretation and Clinical Correlation:

The variant observed in the individual is correlated with the disease condition. The location and annotation of the variant is detailed which includes the chromosomal position, depth of the observed variant; amino acid change in the HGVS notation.

The second section describes the phenotype and inheritance pattern reported in OMIM for the corresponding gene mutations and from literature as applicable.

The third section of this content explains the supporting evidence for reporting a particular variant. This includes

- minor allele frequency of the variant in the population databases (1000Genome database and ExAC database)
- If the observed variant has been previously reported to literature to be associated with the phenotype.
- Multiple *in silico* tool predictions of the variant in affecting the protein (PolyPhen-2, Mutation Taster, SIFT, LRT)
- Location of the variant in the protein/ domain based on Uniprot (http://www.uniprot.org/)
- Conservation of the base/amino acid across species

The fourth section describes the final concludes by attributing significance to the detected variant.

The description/ templates of various variant class are available to maintain uniformity in reporting

Recommendations:

The following recommendations are mandatory to be included in the report as applicable

1. Sanger sequencing is recommended to confirm the variant(s)

- We do not do Sanger sequencing as a part of NGS testing; Sanger is recommended by ACMG guidelines to confirm the variants detected in NGS; thus this is explicitly mentioned in the report
- when a gene has pseudogenes in the human genome
- When there is stand bias observed
- Low coverage of the variant less than 10x depth

2. It is recommended to sequence the variant in the reported variation in the parents and the other affected and unaffected members of the family to ascertain the significance of the variant.
3. The maternal cell contamination (MCC) in this prenatal sample cannot be ruled out. MCC, if present, can lead to discrepancies with the reported results.

7.7 Limitations of the NGS test:

- Intronic variants are not assessed using this method.
- Large deletions of more than 10 bp or copy number variations /chromosomal rearrangements cannot be assessed using this method.

- Certain genes may not be covered completely and few mutations could be missed.
- The mutations have not been validated by Sanger sequencing.
- Incidental or secondary findings (if any) that meet the ACMG guidelines [22] can also be given upon request.
- The maternal cell contamination (MCC) in this prenatal sample cannot be ruled out. MCC, if present, can lead to discrepancies with the reported results.
- For all hematological disorders for which blood transfusion is done, genetic testing should only done after 3 months post transfusion
- The classification of variants of unknown significance can change over time based on additional evidences may become available. Please contact MedGenome at a later date to inquire about any changes.

References: The list of references used in the report are listed in this section of the report. In addition to those references listed below; additional sources of literature / database evidences are added in this section.

Coverage of the genes analyzed: As per the test requested the coverage of the genes listed in the panel is given as appendix. The physical coverage of the genes listed with >1x depth of coverage. An example is given below:

Checklist for reporting: Those who analyze the samples are to strictly follow the checklist. It is mandatory to share the prescribed checklist along with the report. A check list was followed while preparing a report.

7.8 Future directions

Although disease-targeted testing may remain useful in the short term, as our knowledge improves, addition of newly identified genes to the panel (which has to be synthesized according to customized order) may take time, may be laborious and may not be cost-effective for the laboratory. Many laboratories have now shifted to performing WES and limiting the analysis to genes associated with phenotype and filling up the gaps with Sanger sequencing. Although the cost may be 2–5 times higher than targeted panel sequencing (which can go as high as USD 3000 depending on the number of genes in the panel), it allows for reanalysis of the data when new gene associations are made. In addition, some commercial companies have developed kits that capture only medically relevant regions of the genomes. These kits eliminate the risk of finding variants in genes whose function is unknown.

Currently, clinicians order NGS testing based on detailed phenotyping and after excluding single candidate genes—also referred to as a “phenotype first” approach. As NGS testing becomes more easily available, there will be a tendency for clinicians to perform the testing first and then assess the patient’s phenotype to match the genotype—referred to as “genotype first” approach.

It is also possible that in the not so distant future, a relatively healthy individual may perform WES first and then consult a clinician for interpretation of the findings and subsequent evaluations. It may not be long before newborns are screened for inherited disorders using WES. However, there are many ethical, medical, and logistical challenges that need to be addressed before this becomes common practice.

Finally, there have been recent reports on the role of somatic mutations in Mendelian diseases. This has been advanced by the ability to perform deep-targeted NGS to detect low level mutations in patients with Mendelian disorders, majority of which would have been missed by conventional testing. As our understanding of the role of somatic mutations in other human diseases increases, deep-targeted NGS may be the test of choice in these disorders.

7.9. Conclusion

NGS is a useful diagnostic test for the majority of Mendelian disorders and is gaining acceptance in the medical community. However, there remain certain key challenges that need to be addressed before NGS testing becomes part of routine clinical practice.

8. How to complement in corporate?

Ambitious employees always find opportunities to climb up the corporate ladder. There is no doubt that the individuals who are eager to climb the professional ladder of success, but the thing matters is whether they have the discipline and the positive interest to reach the desired goal. Successful individuals always keep track of great principles, which would help them to reach the highest level.

□ **Taking Initiative-** One should be more involved in any difficulty his colleague or boss has encounter. One should always be ready for any sort of work which has nothing to do with his work, and that’s what makes the difference between a leader and a common employee. One should always be ready to help.

- **Being always prepared**- Preparation which is good always helps an individual to stay focused on your agenda. It also shows the managers that he is serious in his work, and he takes company meetings seriously. Avoiding being careless at the meeting, with no preparation whatsoever. Good preparations may lead to promotion.

- **Being Responsible**- A good employee is a responsible employee. Adding up responsibilities should be one's forte. He should also be ready to undertake more responsibilities if necessary. When he fails at certain task assigned to him, he should be the one standing up to admit his fault and ask for the apology. Managers are always on the lookout for such a person who is not only ready to shoulder any responsibilities, but willing to bear the consequences of his failure or irresponsibility.

- **Never Postpone**- He should finish his task on time. He should ensure that the tasks must be completed before anything. Postponing of work is looked upon as less dedicated, and not focused.

- **Effective Communication**- Proper communication is an essential tool to successful career. If someone is able to put his thoughts and opinion across in an effective manner, he will get the most attention. His communication skills will lead him to sit among bosses and managers. He will be asked to share his views in company's meetings. He should ask politely, state clearly, say friendly, share his opinions carefully. When he is angry, he should try to say slowly.

8.1 REFERENCES

[1] Baird, P. A, Genetic disorders in children and young adults: a population study. Am J Hum Genet. 1988;42(5):677–93

[2] Veltman, J. A, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012;13(8):565–75.

[3] McKusick-Nathans Institute of Genetic Medicine JHUB, MD. Online Mendelian Inheritance in Man, OMIM®. http://www.omim.org (2015). Accessed 15 Feb 2015.

[4] Bamshad, M. J, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745–55.

[5] Shashi V, et al. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genet Med. 2014;16(2):176–82.

[6] Margulies, M. "Genome sequencing in microfabricated high-density picolitre reactors." *Nature.* 437.7057 (2005): 376.

[7] Harris, S. R, *et al.*, Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis. 2013;13:130–6.

[8] Antonarakis, S. E., *et al.*, Mendelian disorders deserve more attention. *Nature Reviews Genetics* **7**, 277–282 (2006)

[9] Chial, H. "Mendelian genetics: patterns of inheritance and single-gene disorders." *Nature Education* 1.1 (2008): 63.

[10] Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet. 2013;14(4):295–300.

[11] Board ACMG. of Directors. Points to consider for informed consent for genome/exome sequencing. Genet Med. 2013;15(9):748–9.

[12] Worthey EA, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 2011;13(3):255–62.

[13] Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. N Engl J Med. 2014;370(25):2418–25.

[14] MacArthur DG., et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469–76.

[15] Need AC, et al. Clinical application of exome sequencing in undiagnosed genetic conditions. J Med Genet. 2012;49(6):353–61.

[16] Oz-Levi D, et al. Exome sequencing as a differential diagnosis tool: resolving mild trichohepatoenteric syndrome. Clin Genet 2014