CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Deciphering Hereditary Disorders through Genomic Profiling

by

Maryam Nageen

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2025

Copyright © 2025 by Maryam Nageen

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

I want to dedicate this achievement to my parents, teachers, and friends, whose constant support and encouragement guided me through every challenging moment.

CERTIFICATE OF APPROVAL

Deciphering Hereditary Disorders through Genomic Profiling

by

Maryam Nageen

(MAI233002)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Muhammad Majid	UET, Taxila
(b)	Internal Examiner	Dr. Nadeem Anjum	CUST, Islamabad
(c)	Supervisor	Dr. Mohammad Masroor Ahmed	CUST, Islamabad

Dr. Mohammad Masroor Ahmed
Thesis Supervisor
October, 2025

Dr. Mohammad Masroor Ahmed
Head
Dept. of Computer Science
October, 2025

Dr. M. Abdul Qadir
Dean
Faculty of Computing
October, 2025

Author's Declaration

I, **Maryam Nageen** hereby state that my MS thesis titled "**Deciphering Hereditary Disorders through Genomic Profiling**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Maryam Nageen)

Registration No: MAI233002

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Deciphering Hereditary Disorders through Genomic Profiling**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Maryam Nageen)

Registration No: MAI233002

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. **Maryam Nageen**, S. S. Raza Rizvi, and M. Furqan, "Unsupervised Machine Learning Insights Into Hereditary Disorders Through Genomic Profiling," *International Journal of Advance Computational Engineering and Networking (IJACEN)*, vol. 13, no. 1, pp. 176–183, 2025.

(Maryam Nageen)
Registration No: MAI233002

Acknowledgement

All praises to Almighty ALLAH who gave me the strength and ability to complete this work. His blessings made every difficult moment easier and gave me hope when things felt impossible.

All praises, respect, and love to the **Holy Prophet Hazrat Muhammad (P.B.U.H)**, whose life is a perfect example of guidance for all of us.

I am truly thankful to my supervisor, **Dr. M. Masroor Ahmed** (Capital University of Science and Technology, Islamabad), for his support and kind guidance throughout this research. His advice and encouragement helped me stay focused and motivated.

My heartfelt thanks to my **parents and siblings** for always believing in me. Their prayers, support, and love gave me the strength to keep going, even when it was hard.

The process has contributed significantly to intellectual development, analytical thinking, and knowledge acquisition. Thanks to everyone who walked even a step of it with me.

(Maryam Nageen)

Abstract

Genomic profiling has emerged as a foundation in understanding of hereditary and rare genetic disorders through identifying unique patterns in family structures. Traditional analysis is often inapparent to the subtle interaction between inherited and de novo mutations, particularly in complex family structures. This study explores a Clustered Genomic Fingerprinting (CGF) approach for the identification of Mendelian and non-Mendelian inheritance patterns in family-based whole genome sequencing (WGS) studies. CGF reshapes and encodes the genomic data and performs K-Means clustering to form a genotype matrix. It includes both parent and child genotypic values, flagging deviations from expected inheritance. Anomalies are identified by intra-cluster distance scoring and non-parental genotype comparisons. The CGF approach is semi-supervised, in that parents' and known inherited variants train the model, and only the unknown child's variants are tested for a risk of mutation. The approach retains well-known Mendelian patterns and highlights genotypic anomalies, such as the (TT) genotype in children where both parents carry (AA), indicating de novo mutations or complex inheritance. The CGF model achieved an accuracy rate of 94%, validating the robustness of mutation risk prediction capability. Two clusters of high-risk mutations were identified, belongs to chromosome 14 and 8. Which include genes like T-cell leukemia 1, immunoglobulin heavy chain cluster, ribosomal RNA, FGFR1, MCPH1, and GATA4. This mutation is potentially associated with a rare Ring Chromosome 14 Syndrome, characterized by developmental delay and epilepsy. Chromosome 8, associated with microcephaly, Burkitt lymphoma, and mosaic trisomy 8. The CGF approach improve the accuracy of mutation detecting and provides a way for early diagnosis and personalized medicine. It bridges the gap between family genomics and tangible clinical utility by directing precision therapies and informing family-based healthcare strategies.

Contents

List of Figures

List of Tables

Abbreviations

AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CGF	Clustered Genomic Fingerprinting
chr	Chromosome Identifier
DES	Dual Encoding Scheme
DNM	De Novo Mutation
geno	Observed Genotype
G_enc	Encoded Genomic Matrix
G_norm	Normalized Genomic Matrix
MCC	Matthews Correlation Coefficient
pos	Genomic Position
RF	Random Forest
SNP	Single Nucleotide Polymorphism
TRL	Technology Readiness Level
WCSS	Within-Cluster Sum of Squares
WGS	Whole Genome Sequencing

Symbols

$G^{(i)}$	Genomic data matrix of individual i
$G'^{(i)}$	Deduplicated genomic matrix
$G''^{(i)}$	Cleaned and imputed genomic matrix
G	Unified matrix from all family members
f_C, f_T	Encoding functions for chromosome and genotype
$x^{(i)}$	Encoded data vector at locus i
X	Encoded data matrix
μ_k	Centroid of cluster k
$c(i)$	Cluster assignment for point i
C_k	Cluster k
$M_{k,i}$	Pivot matrix cell for cluster k and member i
$f_k^{(i)}$	Binary flag for non-parental genotype in cluster k
R_k	Raw mutation risk score in cluster k
R_k^{norm}	Normalized mutation risk score
$r^{(i)}$	Euclidean distance from point $x^{(i)}$ to centroid
θ	Mutation risk threshold
δ	Confidence interval around centroid mean
$S(i)$	Silhouette score for data point i
ρ	Cohesion-separation ratio
$H(x)$	Random Forest ensemble prediction

Chapter 1

DNA Structure and Genetic Variation

1.1 Introduction to Genome

The genome is the complete set of genetic information of an organism, including genes as well as non-coding sequences. It acts as a life blueprint and determining development and disease susceptibility. It is comprised of DNA, the genome encodes the necessary information for growth and functions of the cell. In DNA structure, the sides of the ladder are sugar (deoxyribose) and phosphate molecules, which make up a sturdy backbone. The structure of the double-stranded DNA (DSDNA) helix is cross-linked in Figure 1.1, with key height measurements showing that full turn of the helix comprises 3.4 nanometres (nm), which includes about 10 bp (base pairs) stacked in the direction of the helix axis. This stack contributes to the stability of the DNA. The width of the helix is 2nm, which keeps it rigid. These parameters are important for the DNA to structure genetic information in an efficient way while allowing essential processes but also allows to occur replication and transcription.

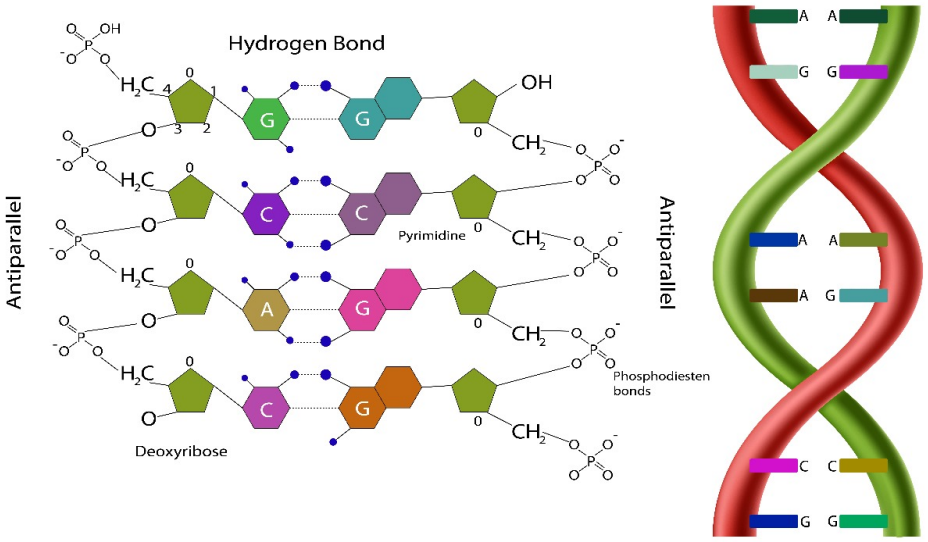

FIGURE 1.1: Double Stranded DNA Structure

The steps of the ladder are formed by four nitrogen bases: A (Adenine), T (Thymine), C (Cytosine), and G (Guanine). The bases are matched in a particular manner and held together by hydrogen bonds, are known as base pairs. Every single unit of sugar, phosphate, and base constitute nucleotides, from which DNA is constructed as shown in Figure 1.1. Numerous nucleotides join together and form a long chain that is the whole DNA molecule. DNA comes well packaged in the structures known as chromosomes, which reside in the cell's nucleus. All our genetic necessary information for life is stored on 46 chromosomes in every human.

1.1.1 Importance of Biological DNA

DNA is essential for life, with protein instructions that drive the body's biochemical reactions (enzymes), offer structural stability, regulate functions, and ensure inheritance, as well as facilitating growth and healing processes through cell division. Genomic alterations sometimes occur due to DNA repair [1]. The next-generation sequencing (NGS) is important to genomics research, permitting the

rapid and affordable analysis of millions of DNA fragments, helping to revolutionize knowledge on genetic variability, understanding the disease, and personalized healthcare [2].

1.1.2 DNA location in Cells

The DNA in eukaryotic cells is largely concentrated in a cell's nucleus, where it is organized into structures called chromosomes. Each chromosome holds a single long DNA molecule, elegantly coiled and bundled with proteins called histones.

This arrangement that enables the compact storage and the regulation of genetic information. There is also a small amount of DNA, called mitochondrial DNA (ADNMT), which resides in cellular organelles, that produce organelles of the cell. Mitochondrial, which is inherited maternally and are solely responsible for energy metabolism.

1.2 DNA Structure

1.2.1 Double Propeller Model

The DNA in the double-helix model consists of two long threads of nucleotides that wound around each other. The helix is necessary for DNA stability, and not only does it enable the storage of such a great amount of genetic information in a compact form.

1.2.2 DNA Components

DNA is composed of building blocks of molecules called nucleotides. Each nucleotide is made up of three components. These are the elements that form the

distinctive structure of DNA, and enable it to store and transmit genetic information efficiently. The components for every nucleotide are depicted in Figure 1.2:

a) Phosphate Group

b) Pentose Sugar

c) Nitrogenous Bases

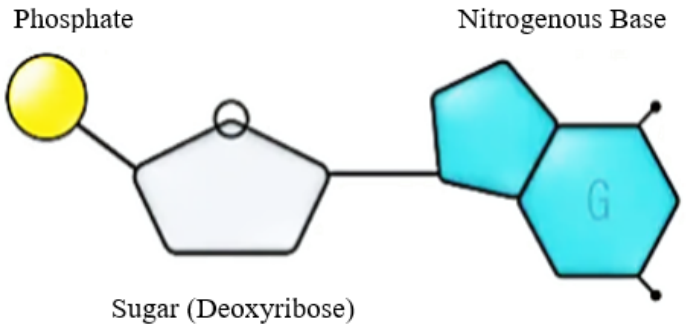

FIGURE 1.2: Structure of DNA Components [3]

1.2.2.1 Phosphate Group

A phosphate group is consists of phosphorus atom and attached to four oxygen atoms. It serves as an important component when pairing nucleotides to form the chains of DNA through phosphodiester bonds, establishing the backbone of the DNA structure.

1.2.2.2 Pentose Sugar

In DNA, the sugar is deoxyribose, whereas RNA is ribose, which supports the helix structure of the DNA molecule presented in Figure 1.3. The absence of an oxygen atom in position 2' distinguishes the deoxyribose from ribose.

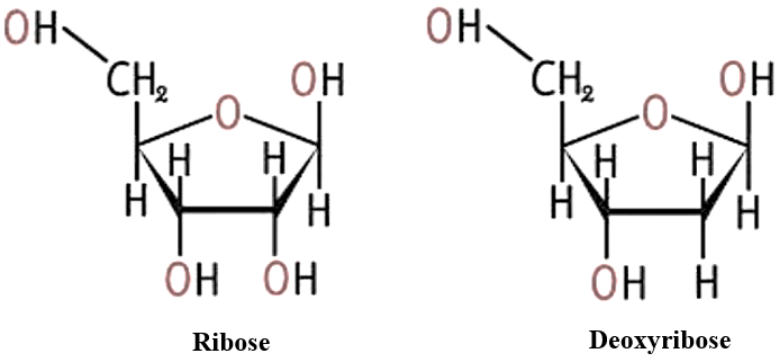

FIGURE 1.3: Structure of Pentose Sugar [3]

1.2.2.3 Nitrogen Bases

DNA contains four nitrogen bases and they are categorised into two groups:

a) Purines: Adenine (A) and Guanine (G) are larger, double-ringed structures.

b) Pyrimidines: Cytosine (C) and Thymine (T) are smaller, single-ringed structures.

1.2.3 Antiparallel Orientation of Strands

The two threads of DNA work in opposite directions (antiparallel orientation). One thread runs from the end 5′ to the 3′ end and the other from the 3′ end to the 5′ end. This polarity is important for the mechanics of DNA replication and transcription, because it means that the enzymes that copy each strand of the DNA can only run along it in one direction.

1.2.4 Base Pairing Rules

The nitrogen bases are precisely paired on the basis of the complementary base pairing rules: Adenine (A) pairs with Thymine (T) by two hydrogen bonds as shown in Figure 1.4, and Cytosine (C) pairs with Guanine (G) by three hydrogen

bonds. This specificity provides that genetic information is reproduced accurately during the replication of DNA, and is required for the control of gene expression.

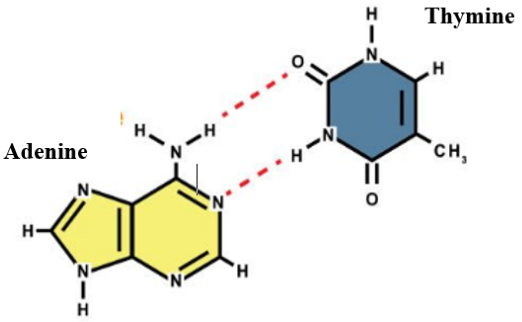

FIGURE 1.4: Base Pairing of Adenine (A) pairs with Thymine (T)[3]

1.2.5 Chemical Bonds in the Structure of DNA

1.2.5.1 Hydrogen Bonds

Hydrogen bonds between complementary bases provide specificity and stability to the structure of the DNA shown in Figure 1.5.

Such linkages permit the two threads to be easily separated in replication and transcription and serve to hold the double helix integrity.

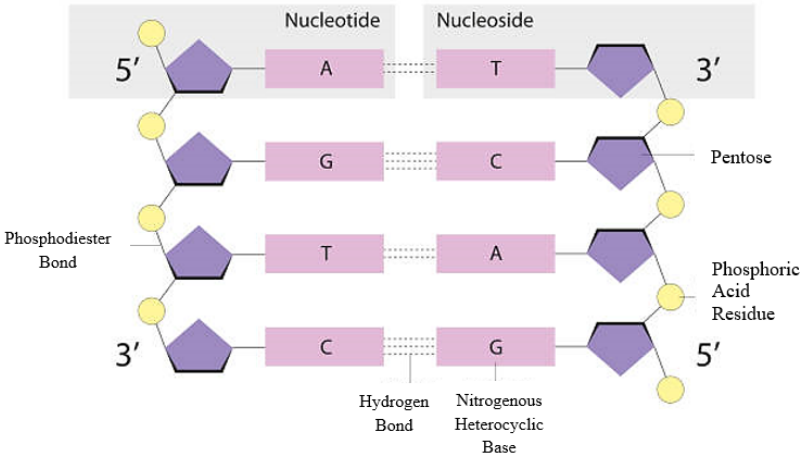

FIGURE 1.5: Structure of Chemical Bonding of DNA [3]

1.2.5.2 Phosphodiester Bonds

Phosphodiester bonds link the 5′ carbon of a 3′ deoxyribose sugar of the next nucleotide through a phosphate group shown in Figure 1.5. These covalent bonds are the strong backbone that ensures the stability and the structural integrity of the DNA molecule.

1.2.5.3 Hydrophobic Interactions

Hydrophobic interactions between pairs of stacked pairs bases contribute to stabilize the orientation of the pairs of bases in the propeller. As nitrogen bases are hydrophobic (do not easily interact with water) this stability helps keep the helix in shape.

1.3 DNA Replication

1.3.1 Importance of DNA Replication

DNA replication is an essential process that takes place prior to the cell division and ensures that each daughter cell receives an exact copy of genetic material. This is a crucial process for multicellular organisms for the growth, development and repair of tissues.

1.3.2 DNA Replication Mechanism

In replication, the double helix untwists, and each strand acts as a template for the production of a new complementary strand. Enzymes works to separate the strands of DNA, while polymerase works to add nucleotides to the growing strand according to the rules of base pairing. This forms two identical double helices, each with one original strand and one new strand.

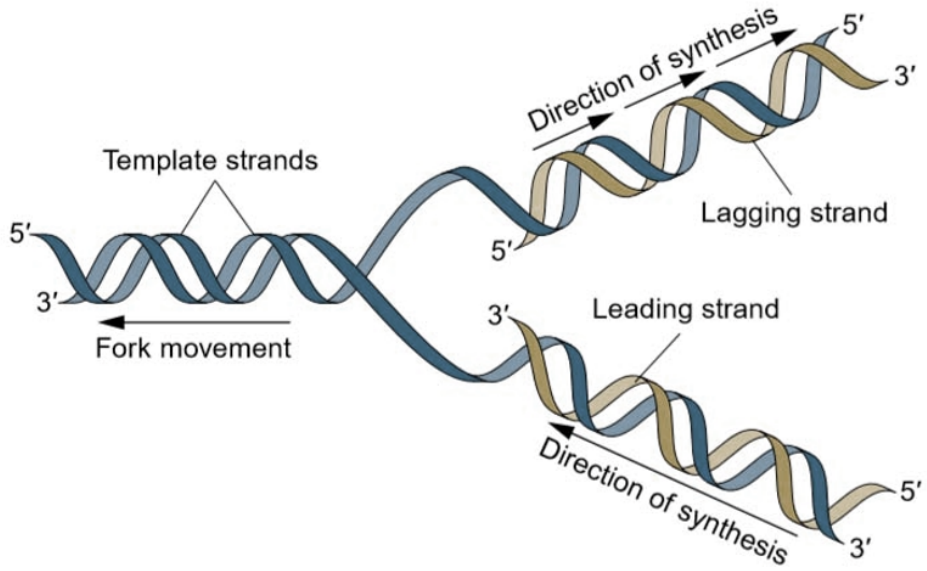

FIGURE 1.6: Structure of Chemical bonding of DNA [4]

A sliding clamp and clamp loader boost DNA polymerase processivity to continue rapid synthesis. Topoisomerases relieve torsional stress ahead of the fork, preventing supercoiling and strand breakage. When the RNA primer is in place, DNA polymerase then adds nucleotides to the growing DNA chain in the 5′ to 3′ address, which means that it can only add nucleotides to the 3′ end of the primer or the chain growth shown in Figure 1.6. Due to the antiparallel nature of the two DNA threads, replication takes place differently in each thread. Since the two DNA strands are antiparallel, replication proceeds in each strand in a different way:

a) Leading Strand: A single RNA primer is laid down near the origin, and synthesis proceeds continuously toward the replication fork by a high-fidelity polymerase. The main thread is continuously synthesized in the same direction that the replication fork continues to open. DNA polymerase adds nucleotides gently as the fork gets progress.

b) Lagging Strand: The lagged strand is synthesized discontinuously in small segments known as Okazaki fragments. Every fragment needs its own RNA primer. These fragments are produced by DNA polymerase working in the opposite direction of the replication fork.

1.4 Genetic Variation and SNPs

1.4.1 Single Nucleotide Polymorphism

A single nucleotide polymorphism (SNP) is a variation of a single nucleotide at a particular location in the genome of an organism. The Telomere-to-Telomere (T2T) Consortium has released a complete human genome sequence T2T-CHM13 that resolves previously unfinished regions of heterochromatin, and consists of gapless assemblies for all chromosomes with the exception of 'Y' [5]. Such differences exist naturally and contribute to genetic diversity within a population. SNPs are the dominant type of genetic variation in the human genome.

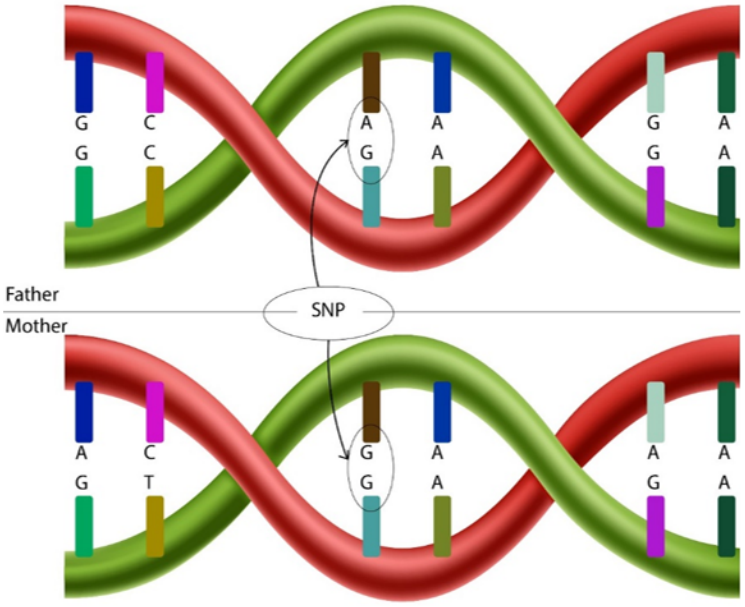

FIGURE 1.7: Structure of Single Nucleotide Polymorphism (SNP)

In a particular position in the genome, an individual can have a base (e.g., adenine) while another may have a different base (e.g., guanine) as illustrated in Figure 1.7, creating two possible alleles for the SNP. This single nucleotide polymorphism can be present within coding regions, noncoding regions or regulatory regions of gene affecting the function and expression of gene. They also lead to genetic diversity and can influence traits, disease susceptibility and individual reactions to drugs.

1.4.2 Significance of SNPs as Genetic Markers

SNPs are essential for studying variation in the human genome and have been associated with differences in influencing traits, disease susceptibility, and medication responses. They are genetic markers helpful for such things as studying populations, family relationships, and tracking ancestry. Their genome-wide abundance makes them useful in various molecular applications, including genome-wide association studies (GWAS) to identify associations of individual genetic variants with diseases.

1.4.3 Foundation of Genetic Variation

Inheritance forms biological diversity by the transmission of genetic information from a father or mother to their offspring. Mendelian inheritance are predictable patterns of inheritance by dominant and recessive alleles. However, variations such as single nucleotide polymorphisms (SNPs) and de novo mutations complexity matters, contributing genetic variation and affecting the expression of traits in the offspring.

1.4.4 Mendelian Inheritance

Mendelian inheritance is the manner in which certain genetic traits pass from parents to their offspring through autosomal dominant and autosomal recessive patterns. For dominant traits one allele must be present in order to express the trait, whereas in the case of recessive traits, two copies required to express the trait and the parents are often unaffected. To predict the probabilities of traits, as well as accurate genetic testing and counseling for gene variant detection [6], knowledge of these patterns is crucial.

1.4.4.1 Autosomal Dominant

In the autosomal dominant inheritance, individual need only one altered copy of the gene to express the trait or disorder. This is equivalent to the chance that a father carries the dominant allele, and each child have a probability of 50% to inherit this trait as shown in Figure 1.8.

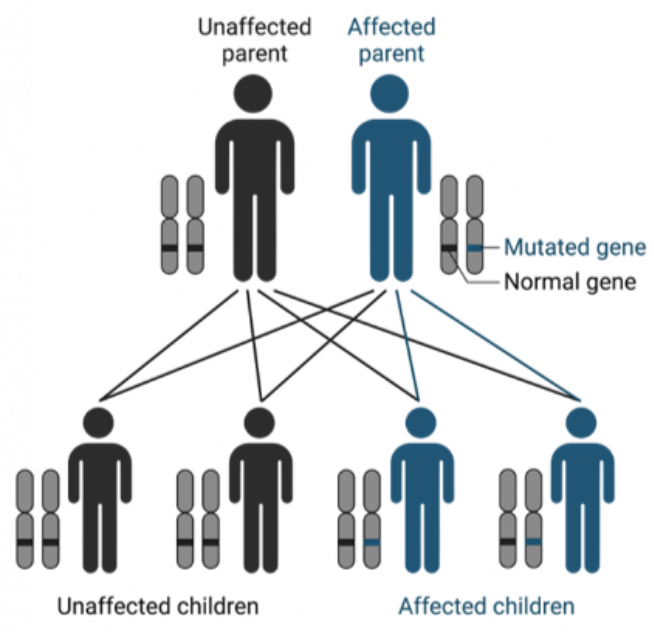

FIGURE 1.8: Autosomal Dominant Inheritance [7]

Affected offspring usually have an affected parent and the trait can occur in every generation of a family. Diseases inherited in this manner are known as dominant autosomal disorders; examples of which include Huntington's disease and Marfan syndrome.

1.4.4.2 Autosomal Recessive

Autosomal recessive inheritance on the other hand calls for two copies of the dominant allele to be inherited, one from each parent, for the trait to be expressed. The parents of an affected individual are typically carriers as shown in Figure 1.9, who generally carry a copy of the recessive allele [8]. This pattern of inheritance sometimes skips generations in traits to express. The recessive allele can be passed

down without presenting in the parents. Common examples of autosomal recessive disorders are cystic fibrosis and sickle-cell disease.

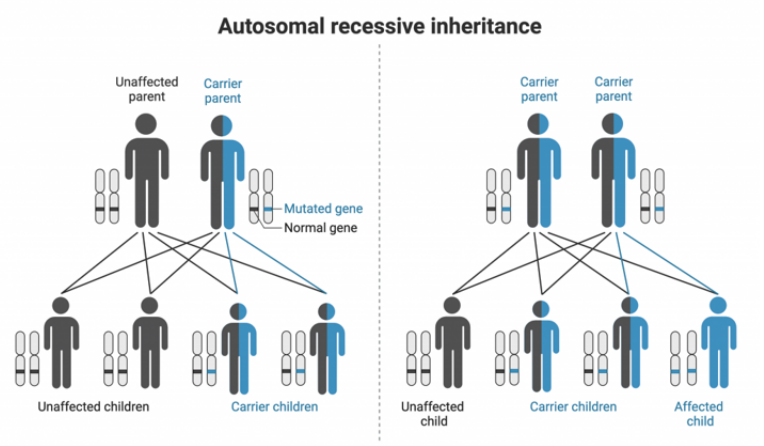

FIGURE 1.9: Autosomal Recessive Inheritance [7]

1.4.4.3 De Novo Mutation

De Novo mutations are genetic changes that appear for the first time in a individual, from the DNA of a patient's germ cells or in early embryonic development Figure 1.10. Such mutations can generate novel genetic variation not inherited from parents, which can contribute to genetic diversity. De novo mutations in dopamine neurons are linked to genetic factors in Parkinson's disease (PD) and PD risk genes contribute to the complexity of diseases [9].

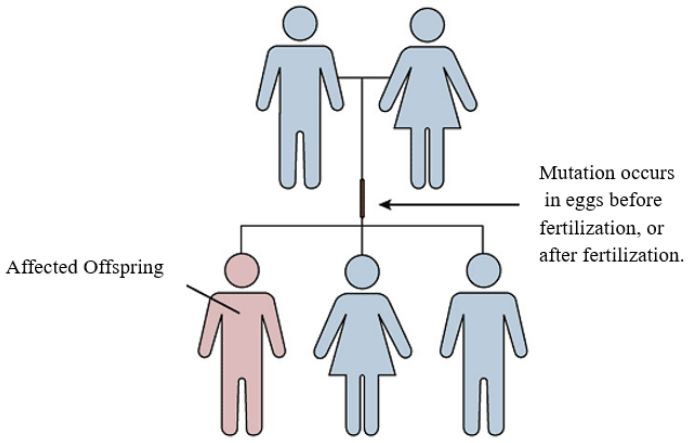

FIGURE 1.10: De Novo Mutations Genetic Mosaicism [10]

Affected offspring usually have an affected parent and the trait can occur in every generation of a family. Diseases inherited in this manner are known as dominant autosomal disorders; examples of which include Huntington's disease and Marfan syndrome.

1.4.5 Impact of SNP on Inheritance Patterns

SNPs influence how traits are inherited, determining the autosomal dominant or recessive patterns in the offspring. They have the ability of forming alternate alleles of genes and hence different expressions and variation in traits based on the laws of Mendelian inheritance.

1.5 Role of SNPs in Genetic Disorders

1.5.1 SNP in Coding Regions

SNP that are located within the gene coding regions can lead to the different changes in the produced protein as shown in Figure 1.11. These alterations can be categorized as:

a) Missense Mutations

b) Nonsense Mutations

Missense Mutations are single nucleotide transitions or transversions that cause a variation in the amino acid sequence of a protein and may affect the function of protein. Nonsense mutation occurs when an SNP changes a codon for amino acid to a stop codon, resulting in the premature termination of protein synthesis.

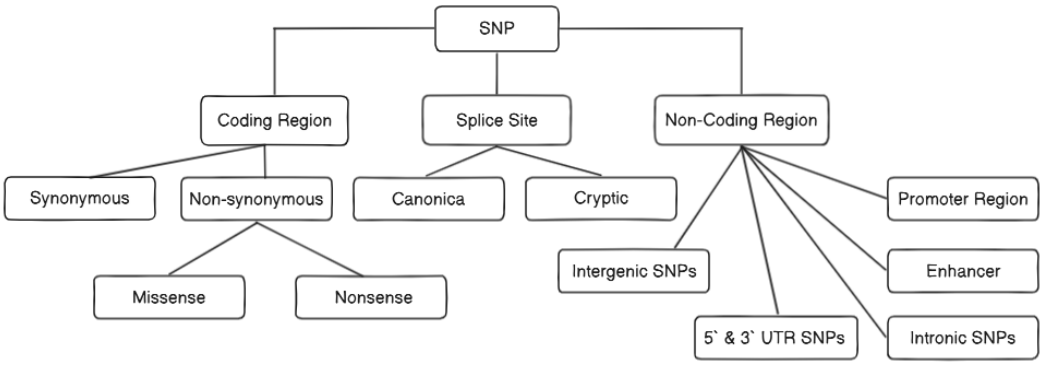

FIGURE 1.11: Types of SNPs

1.5.2 SNPs in Non-Coding Regions

SNP in non-coding DNA regions could not alter the protein structure but also influence the gene expression as shown in Figure 1.11. There has been a great progress in identifying genetic loci that are involved in the complex traits through GWAS, the majority of these associations are in the regions that are non-coding, restricting the capability to determine which genes they regulate and how they influence disease, therefore complicating the translation of findings into a clinical context [11].

1.5.3 SNP in Splice Sites

Some SNPs disrupt with proper RNA splicing, having defective or incomplete proteins. For instance, a BRCA1 gene mutation that affects the junction shows more susceptibility in breast cancer and ovarian cancer risk. The expanding number of treatable tumor-specific alterations and the development of next-generation of DNA sequencing has revolutionized cancer care and brought into focus how the molecular information must be integrated for diagnostic purposes, prognosis and therapeutic management.

1.6 Statement of the Problem and Disease Susceptibility

1.6.1 Identification of the Problem

This study explores the impact of SNPs in the inheritance of autosomal dominant and recessive traits, as well as the contribution of de novo mutations. Dominant traits only require one mutated gene copy, while recessive ones need two. Some SNPs can cause changes to gene function resulting in disorders, such as familial hypercholesterolemia and cystic fibrosis. Furthermore, de novo mutations, that are affected by SNPs, will result in different genetic diseases such as autism spectrum disorder (ASD) [12].

1.6.2 Significance of the Problem

The investigation of SNPs on autosomal dominant, recessive traits, and de novo mutation condition is crucial for the understanding of genetic mechanisms and for developing diagnostic tools for therapy. Identification of individual SNPs may allow for focused interventions and gene therapies, improving genetic counseling and risk prediction for families. Furthermore, investigation of de novo mutations shows complex disorders and brings complexities to risk management. This study has profound implications for the progression of medical research, enhancement of genetic counseling, implementation of public health measures, and will eventually pave the way for personalized medicine and better patient care and outcomes.

1.7 Research Objectives

The goal of this study is to investigate the impact of single nucleotide polymorphisms (SNPs) on autosomal dominant, recessive inheritance patterns, and de

novo mutations. The research aims to achieve the following by finding and studying relevant SNPs to uncover their nature in genetic disease and to better explain the underlying heredity mechanisms:

Characterization of SNPs: Identify SNPs associated with genetic disorders through genomic databases and literature.

Functional Implications: Analyze the SNPs' effects on the expression of gene, protein function, and resultant phenotypes.

Investigation of De Novo Mutations: Examine SNPs that are known to be responsible for spontaneous, non-inherited mutations in disease.

Applications in Personalized Medicine: Investigate SNP utilization for individualized therapies and enhanced genetic counseling.

Public Health Implications: Inform screening and prevention strategies by identifying genetic risk factors.

1.8 Research Questions

RQ1. What specific SNPs do autosomal dominant diseases seem to be associated with, and how do they affect the phenotypic expression of these diseases?

RQ2. What specific SNPs are associated with autosomal recessive disorders, and what is their impact on gene function and disease manifestation?

RQ3. What is the significance of the discovered SNPs for personalized medicine and genetic counseling in individuals with autosomal dominant, autosomal recessive disorders, and de novo mutations?

1.9 Significance and Limitations of the Study

1.9.1 Contributions to Knowledge

The goal of this research is to deliver insights into the effect of single nucleotide polymorphisms (SNPs) on autosomal dominant and recessive traits and their role in de novo mutations. The research will build understanding of genetics mechanisms behind inheritance patterns and susceptibility to disease through characterization of key SNPs. It also focuses on the functional implications of these variants, filling in gaps in our current understanding and offer potential avenues for future genetic researches.

1.9.2 Practical Implications

The findings of this study may contribute to practice, policy, and research in the following ways:

1.9.2.1 Clinical Practice

Understanding SNPs in genetic disorders can aid clinicians in risk assessment and treatment planning by using different computational methods [13]. It enables accurate genetic counseling and personalized patient care.

1.9.2.2 Public Health Policy

SNPs that are identified can be useful for early detection and prevention initiatives. Targeted screening programs can be adopted by decision-makers to reduce the impact of genetic disorders.

1.9.2.3 Future Research

Provide a reference for further investigating the functions of SNPs in complex diseases. This knowledge may be exploited for future therapies and interventions.

1.9.3 Target Audience

The research will interest a cross-section of communities, including:

1.9.3.1 General Public and At-Risk Individuals

People with a history of genetic disorders, or who are interested in preventative health, can take advantage of early genetic screening and risk assessment. Such research may enable people to make informed decisions prior to disease.

1.9.3.2 Practitioners

Geneticists, clinicians, and genetic counselors will be provided with up-to-date information on the genetic aspects of disorders and the relationship with clinical diagnosis and treatment of diseases.

1.9.3.3 Policymakers

Public health officials and policymakers can benefit from this study by developing evidence-based policies on screening, prevention, and management of genetic disorders.

1.9.3.4 Academics

Valuable to researchers and academics in genetics, molecular biology, and public health who gain further insight into their field of study related to genetic transmission and disease.

1.9.4 Scope of the Study

There are several limitations to this research:

1.9.4.1 Geographical Limits

The study focus on population investigated will be limited to particular geographic areas, and findings may not be generalizable to other populations with different genetic backgrounds.

1.9.4.2 Temporal Limits

The study will consider genetic data sampled within a specified period of time, which may not capture temporal variability in genetic variation and disease prevalence over time.

1.9.4.3 Contextual Limits

The analysis will mainly focus on autosomal dominant and recessive traits, as well as de novo mutations, and may not include other inherited genetic modes or environmental factors that influence disease phenotypes.

1.9.5 Limitations

While this study aims to provide valuable insights, some limitations might impact the findings and their generalization:

1.9.5.1 Sample Size and Diversity

The size and diversity of the sample population could limit the scope of the study [14]. A subset of sample size or lack of diversity could therefore fail to represent the

entire genetic variants in the general population which may in turn compromise the generalizability of the findings.

1.9.5.2 Data Collection Methods

The methods used for identifying SNPs and for determining their relation to genetic disorders are the factors influencing accuracy and reliability of the collected data. The variance in how data were obtained might have led to bias or reduced the robustness of the findings.

1.9.5.3 Complexity of Genetic Interactions

Complexity in genetic interactions and environmental contributions to disease susceptibility may not have been completely represented in the study. Genetic disorders are commonly multifactorial, and the SNPs can interact with other genetic and non-genetic aspects, which further complicates the interpretation of findings.

1.9.5.4 Focus on Specific Disorders

The focus of the work on specific autosomal dominant and recessive is likely to have missed other genetic diseases that also share common SNPs in the other genetic conditions.

1.10 Structure of the Thesis

1.10.1 Overview of Subsequent Chapters

The structured of this thesis is composed of several correlated chapters, which cumulatively form the logic of SNP and their contribution to autosomal dominant, recessive and de novo mutations. Chapter one contains the introduction as regards background, statement of the problem, objectives and sub-objectives of the study, research questions, hypotheses concerning definition of terms and delimitations.

The second chapter highlights what has been learnt from available literature on SNPs, inheritance pattern, de novo mutation and analysis methods, pointing at deficiencies in existing knowledge. The third chapter describes the research design, how participants were selected and data was collected and analysed as well as ethical considerations. Chapter four reports the results and includes SNPs as genetic markers for diseases, supported by statistically analysis through tables and figures with an explanation in relation to background, research questions, hypotheses as well as the scope for future studies. The final chapter summarizes key findings, the implications for clinical and public health practice, and suggests areas for future research.

1.11 Conclusion

This chapter has explored the DNA structure and its function, the significance of SNPs in genetic variation and their importance in genetic disease and personalized medicine. Knowledge of these concepts is fundamental for the development of genetics practice and for improved medical care.

Chapter 2

Literature Review

2.1 Introduction and Background

Inherited diseases or genetic disorders are conditions that an individual is born with, which often affects their quality of life and sometimes can be life-threatening, and are inherited from their parents through their genes. Genetic diseases are one of major focuses in biomedical research because of their serious effects on human health and heredity. The field of investigation of these disorders is currently experiencing a revolutionary period, as genome technologies and computational biology are developing at rapid speed. Advances in genomics now permit us to globally decipher DNA sequences at an unprecedented level, and computational models facilitate the assimilation and understanding of this overwhelming amount of genetic data.

Even though there have been many developments in genetics research and related technologies that have provided deeper insights than ever before into these complicated diseases, hereditary diseases still pose a formidable challenge in modern medicine. They stem from mutations in one or more genes, may also be modified by multiple genetic and environmental factors, and are capable of altering phenotypes including physical properties, behaviour, and susceptibility to disease. It's

difficult to diagnose hereditary diseases, and often relying on traditional methods, such as manual blood smears, which can be subjective.

A 5-year study investigating hereditary hemolytic anemia (HHA) was based on gene panel sequencing, and 10 of 14 patients were found to carry pathogenic variants. This genomic integration facilitated the diagnosis and treatment [15]. Machine learning and deep learning dissect complex genetic data, uncovering relationships within gene sequences, expression, and epigenetics. Due to their ability to find non-obvious dependencies that traditional methods overlook, ML algorithms are useful for being able to effectively apply large and heterogeneous datasets to the diagnosis of hereditary diseases.

The SVMs are good classifiers for hereditary diseases in the context of genetic data with high dimensions. They effectively discriminate class labels, facilitating the biomarker search and the drug target in cancer genomics. SVMs are known to be resistant to overfitting, which was well-suited for genomic analyses in which the number of samples is greater than the number of features [16].

Deep Learning, in particular CNNs is facilitating the detection of phenotypic characteristics encoded in biological sequences. The AMBER architecture evolves the CNN via NAS, surpassing performance from baseline and expert models. It improves the accuracy of prediction and facilitates the identification of functional genomic variants associated with disease heritability [17].

RNNs are tailored for the data which is in sequential form, for instance the sequence of measure points of gene expression. A model via rectified linear units (ReLUs) surpassed LSTM and SPLS, and was able to better crossutilize the missing biological data. Combining genomic, clinical and imaging data enables more accurate predictions; it helps cross-study diagnosis of complex genetic diseases [18]. The research [19] highlighting the development of strategies to examine novel assessment tools for technical advancements in emerging technologies. High-throughput next-generation sequencing (HT-NGS) technologies are transforming the future of genomic research, with individual genome sequencing [20].

2.2 Existing Contributions for Addressing Various hereditary Diseases

The development of sequencing of the bacterial genome has changed enormously our perception of infectious diseases. After sequencing the first bacterial genome in 1995, the development of these technologies has facilitated the identification of important genetic determinants related to pathogenicity and resistance [21]. Comparative genomics has also increased our capacity to classify and comprehend bacterial species and metagenomics increased knowledge regarding microbial communities and their intricate functions in health and disease.

In the face of the mounting threat posed by drug-resistant HIV, comprehensive whole genome sequencing has become an essential diagnostic tool. The research showed that WGS of Mycobacterium tuberculosis produces fast genotyping with reliable predictions of drug susceptibility phenotypes [22]. This model not only facilitates the prompt decision for treatment but also supports public health intervention through monitoring transmissions in the event of an outbreak.

For example, inherited retinal diseases are among the diseases that have the most to gain from WGS as shown in 722 individuals [23], and in particular WGS is likely to have a higher diagnostic rate in IRDs than exome sequencing. With a diagnosis rate of 56%, the study revealed some known and new pathogenic alterations in both regulatory and GC-rich of the genomic regions. These findings have implications for accurate diagnosis and suggest novel treatment strategies.

In addition, the whole genome sequencing technology has revolutionized the diagnosis of pediatric diseases. A study demonstrated the ability of WGS to detect multiple pathogenic variants, including SNVs, indels, and structural variants for neonates and children in particular [24]. This technology allows quick and thorough diagnoses, which in turn allows for more personalized and faster medical care for such families.

Rapid whole genome sequencing (rWGS) is also beginning to make a difference in the pediatric intensive care setting for diagnosed children. One research showed the capabilities of rapid whole-genome sequencing in obtaining immediate and accurate molecular diagnoses, leading to optimized therapeutic strategies with low morbidity and mortality [25]. Automated phenotyping platforms have also been successfully incorporated into diagnostic pathways in the care of the critically sick.

The diagnostic spectrum of rare diseases has been disproportionately expanded by large-scale genomic studies. In one such study, 13,037 individuals (the majority of whom harboured rare diseases) were WGSed, and 1,138 highly confident diagnoses made, including 95 Mendelian gene-disease relationships [26]. These results underscore WGS as a crucial tool for discovery and clinical implementation.

Another extensive research investigated the worldwide carrier frequency and also the genetic prevalence of autosomal recessive inherited retinal diseases (AR-IRDs) [27]. The authors estimated that by pooling 276,000 sequence variants from 187 disease-associated genes, more than 2.7 billion individuals are estimated to have at least one pathogenic AR-IRD variant. This resource facilitates disease modelling, genetic counselling and risk prediction.

Novel methods are increasing the speed and accuracy of genetic diagnosis. CRISPR-SPR (CRISPR surface plasmon resonance) has been introduced in a study, which is a method for CRISPR-based target detection using the surface plasmon resonance technique, and was applied to detect Duchenne Muscular Dystrophy mutations without amplification [28]. This simplified technology platform allows sensitive mutation detection and has potential value for both diagnostics and future gene-editing therapeutics.

Particularly there are 24 groups of 1,450 disorders have been classified under the International Classification of Inherited Metabolic Disorders (ICIMD) and is a massive contribution to metabolic genetics [29, 30]. This systematic approach provides an organized method for thinking and teaching about metabolic diversity, and it assists the clinician with diagnosis, management, and genetic counselling

[31]. A research present a master list of list all the currently known inborn errors of classified metabolism according to their pathophysiological basis [32].

Familial hypercholesterolaemia (FH) continues to be a key area of research in heritable cardiovascular diseases. A recent review summarized prior and current state of the knowledge on cholesterol metabolism, the genetic basis of FH, and its association with CHD [33]. Promoting early diagnosis and targeted therapies improved patient survival and increased knowledge and recognition of the disease by clinicians.

Colorectal cancer has been a moving field, especially when considering familial syndromes such as Lynch syndrome. Improvements in the characterization of the mismatch repair mutations have not only influenced diagnostic algorithms, but have also indicated different treatment modalities, such as the use of immuno therapies [34]. Additionally, new surveillance and chemo prevention strategies are maximizing the long-term management of high-risk groups.

Biodiversity genomics programmes are expanding the horizon of genomic studies. The Darwin Tree of Life project has a goal of genome sequencing all eukaryote species in Britain and Ireland, for applications in ecology and evolution [35]. As a component of the international Earth Bio Genome Project, this effort is a crucial component in the conservation of biodiversity enabled by genome science.

All of US Research Program aims to transform the precision medicine landscape by generating an incredibly diverse, openly accessible genomic archive, representing one million US individuals [36]. It contains hereditary risk and pharmacogenomic data making it a powerful tool to better understand health in individuals and populations.

The technology is constantly moving to ever faster sequencing. For instance, nanopore sequencing provides a real-time, extremely low-cost readout of the genome, which facilitates disease diagnosis, particularly in cancer and infections [37]. In that its small size and seconds of assay time make it an attractive technology for the global healthcare market.

NBSeq has potential to revolutionize early disease identification. Through the early recognition of genetic disease states in the newborn period, NBSeq provides a more comprehensive test for newborn screening, with a potential favorable prognostic impact with respect to a wide spectrum of pediatric diseases [38]. Early recognition may be followed by early intervention that can lead to changes in the course of disease.

Next-generation sequencing (NGS) is changing the landscape of diagnostic approaches across genetic medicine. By identifying disease-causing variants that are unsolved by traditional strategies, like Sanger sequencing, NGS is redefining diagnostic rate, particularly for undiagnosed genetic conditions [39]. It is particularly revolutionary when being used for paediatric and genetic complexities. With regards to genetic skin diseases in particular, NGS has enabled 166 novel genodermatoses gene-disease in the last decade [40]. This has resulted in enhanced diagnostic performance and enabled more accurate genetic counseling and patient management.

The contribution of de novo coding mutations in neurodevelopmental diseases is under recognized. A large study showed that mutations of this kind account for around one third of autism spectrum disorder (ASD) cases, and therefore impact both diagnosis and research [41]. These data help advance our understanding of the origin of ASD.

The role of the father in genomic imprinting is becoming increasingly evident. An overview article described the consequence of the increased de novo mutation rates affected by advanced paternal age and oxidative stress in offspring health and the predisposition to diseases such as autism, cancer [42].

De novo mutations are also driving research on autism spectrum disorders. One of the papers highlights the important role of these de novo genetic changes in finding in ASD, both in familial and sporadic cases, which is very valuable for genetic counselling and for personalization of treatment [43].

Testing for the pattern of DNA methylation was used as a diagnostic tool in the neuro developmental disorders. Genome-wide methylation profiling using EpiSign assay provides added utility to work up difficult cases where methylation patterns is linked to a known Mendelian disorder [44].

Genome sequencing (GS) as a diagnostic tool for rare Mendelian diseases has evolved further. GS allows for a wider perspective compared to conventional methods, exposing complex and initially unnoticed changes which are essential for the thorough diagnosis and the clinical decision [45]. An additional level of resolution is afforded by transcriptome sequencing (RNAseq), which records the effects of DNA variants on gene expression. In addition to GS, RNAseq enhances diagnostic rates in rare diseases by identifying the functional consequences of mutations [46].

NGS has revolutionized rare disease studies. It has given a boost to gene discovery and by increases the number of known Mendelian disease genes by twofold from 2007 to 2014, tremendously facilitated the development of diagnostic tests and possible cures [47].

The anticipated benefit of early utilization of first tier diagnostic genome sequencing for minimizing diagnostic lag time is being realized. The studies have reported that the application of GS analysis could provide faster and more accurate interventions for the patients with rare diseases, reducing the time of the diagnostic odyssey [48].

A study involving 744 families, where genome sequencing identified and obtained 29.3% diagnostic yeild of rare disease cases [49]. Its powerful ability is to find variants that are missed by exome sequencing.

Discovery of genetic modifiers is increasingly important for explanation of the phenotypic variability in Mendelian disorders. These modifiers of disease severity and clinical outcome are useful in making prognostic evaluations and creating tailored therapies to target [50].

The autosomal recessive ataxias represent a heterogeneous group of genetic conditions, which are often poorly differentiated clinically. In a systematic review, 45 unique gene-based disorders were discovered, along with an updated classification to aid in the accurate diagnosis and clinical treatment of these cases [51].

Elsewhere, a review which was dedicated to autosomal recessive cerebellar ataxias (ARCA) classified ARCA into five main groups, using genetic and clinical information [52]. Such classification will serves to define these complex neurological disorders and to establish their recognition and therapeutics.

Unexpected inheritance pattern distributions provide important information in genetically diverse populations. Another study identified recessive mutations in genes that were believed to adhere to dominant inheritance except in consanguineous populations or ethnic backgrounds, such as those from Saudi Arabia, and revealed complexities in medical genomics [53].

Polycystic kidney disease continues to be an area of active clinical investigation. In another study, multiple clinical and demographic factors impacting renal function and progression of disease in autosomal dominant polycystic kidney disease (ADPKD) and novel therapeutic targets were reported [54].

Mutations in the SATB2 gene are associated with syndromic intellectual disability. Investigations into these de novo mutations have provided new insights into recurrent clinical patterns and phenotypic variability and have consequently led to improved diagnoses and treatments [55].

Mutation analysis has also contributed to SLE research. A study discovered new genes by de novo mutation screening, stressing the significance of rare variants for SLE pathogenesis and providing leads to new treatments development [56].

Large-scale studies have also shown that severe neurodevelopmental disorders and epilepsy are more related to de novo variants. For example, mutations in the PPP3CA gene have now been described as a cause of developmental delay and seizures, providing novel insights into epilepsy genetics [57].

De novo mutation tracking is also gaining ground for the infectious disease. Such mutations were employed to infer transmission networks based on genomic epidemiology investigating Plasmodium falciparum, in support of malaria control and disease spread [58].

In cancer genomics, TP53 gene mutations have been related to weak prognosis in AML. These mutations are not only stable but also associated with disease progression, and thus can act as a biomarker and therapeutic target [59].

Finally, a ground-breaking experiment established a correlation between PTB and fetal de novo mutations, with a significant increased load identified in PTB cases. The results implicate genes expressed during early brain development that may be associated with the risk of this major public health problem [60].

2.3 Existing Algorithms, Methods, and Techniques Contributing to Identify Hereditary Diseases

2.3.1 Variant Pathogenicity and Genotype-Phenotype Association Methods

2.3.1.1 PolyPhen-2

PolyPhen-2 evaluates the potential effects of amino acid substitutions based on how amino acid changes produced by mutations can affect protein structure and function [61]. It predicts the deleterious effects of missense mutations, using structural and evolutionary information to interpret genetic variations from the coding sequences. This method has been used to analyze MYC gene SNPs associated with Burkitt's lymphoma and prioritized damaging variants likely contribute to disease [62]. These SNP analyses provide stimulus to diagnostics and therapy research

by specifying significant mutations. This difference is examined by PolyPhen-2 in two different ways:

1. Evolutionary Conservation: Is the mutated amino acid well conserved between different species?
2. Context within the Structure of the Protein: Is the substitution close to a complex domain or binding site?

These features are input to a Naïve Bayes classifier that calculates the probability that the mutation is deleterious. The formula is:

$$P(\text{Damaging} \mid \text{Features}) = \frac{P(\text{Features} \mid \text{Damaging}) \cdot P(\text{Damaging})}{P(\text{Features})} \tag{2.1}$$

2.3.1.2 SnpEff

Accurate variant nomenclature is essential for genetic diagnosis and patient management. A review of 218,156 clinical variants found that 85% of the annotations of ANNOVAR and SnpEff agree with each other and that ANNOVAR fitted old variants much better than the newest ones and that SnpEff performed better than ANNOVAR using HGVS nomenclature guidelines and SnpEff annotate coding and protein variants more accurately than ANNOVAR [63]. SnpEff is a rule-based annotation engine that predicts the effects of a variant/s (a change in amino acid) variant mapping to genomic positions and rule-based classification [64]. This highlights the necessity of leveraging more than one diagnostic tool to minimize errors and enhancing standardized clinical reporting. Here's how it works:

1. Input: The VCF file containing observed mutations.
2. Gene Mapping: The location of each mutation in the genome is characterized and mapped to the position within a gene.
3. Prediction of Effect Position: Based on the software applies rules to categorize the mutation, including:

- Synonymous: Identical amino acid change.
- Nonsynonymous: Change in amino acid.
- Stop-gained, frameshift, and so on.

The rule-based logic can be mathematically expressed as:

$$\text{Effect} = f(\text{Gene Location}, \text{Codon Change}) \tag{2.2}$$

2.3.1.3 CADD

The Combined Annotation-Dependent Depletion (CADD) score and tool is a complex algorithm, where more than 60 genomic annotations are included to predict the deleteriousness of a variant, specifically for severe Mendelian disorders [65]. CADD discriminates de novo variants from those that are fixed in human populations using a machine learning model. The latest version, CADD v1.7, has added features such as protein language model score and regulatory predictions, and participated in performance enhancing [66]. CADD has been found to be useful for variant prioritization of conditions such as Lynch syndrome, however differences between expert classifications indicate the need for careful validation [67]. Here are the simple steps in which CADD works:

a) Training data:

Positive class: Apart from the demands in duty the class is not good. It is the absolute value of an integer.

Negative class: Variants simulated randomly supposing the benign variants are dominant.

b) Feature extraction: Each transcript variant is then annotated with biological and functional features. Similar to Evolutionary conservation, Functional scores, and Epigenetic marks.

c) Training the model: An SVM acquires a decision boundary between these two classes. The optimization problem is:

2.3.2 Phylogenetic and Evolutionary Analysis Techniques

2.3.2.1 RAxML

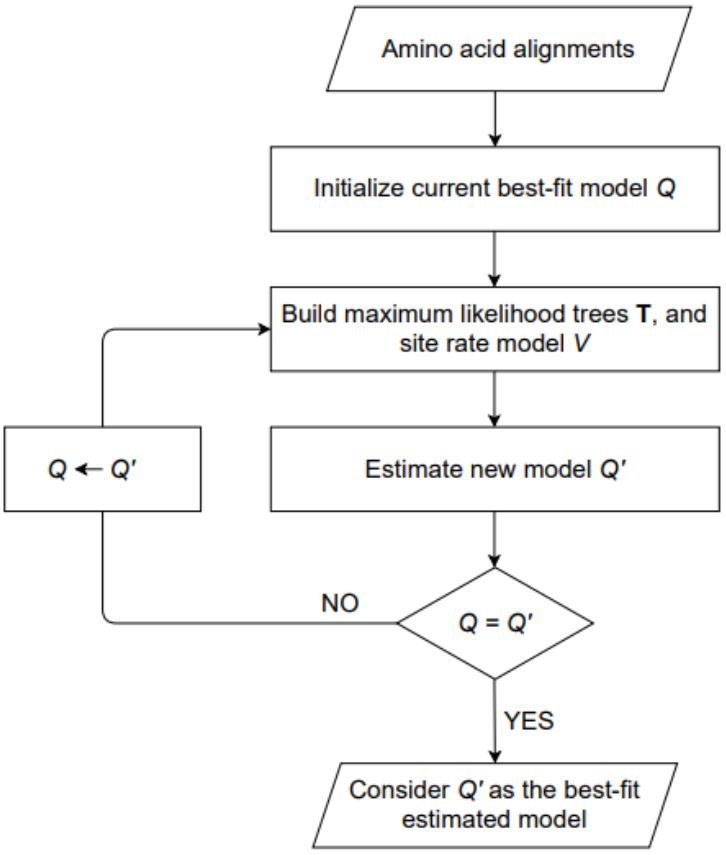

FIGURE 2.1: An Approximate Maximum Likelihood Algorithm to estimate an Amino Acid Substitution Model from a set of Amino Acid Alignments [68]

Modeling amino acid substitution is important for phylogenetic reconstruction using large datasets produced by advanced sequencing technologies [68] illustrated in Figure 2.1. The choice of substitution models can improve the accuracy of ASR, especially in more heterogeneous datasets [69]. Phylogenetic trees are created from such models using the maximum likelihood principle by RAxML [70].

The likelihood $L\ (T,\ \theta)$ of a tree topology T with parameters θ, can be calculated as:

$$L(T, \theta) = \prod_{i=1}^{n} P(D_i \mid T, \theta) \tag{2.3}$$

Where:

D_i is the i^{th} column of the alignments.

$P(D_i \mid T, \theta)$ is the likelihood of that column being observed given the tree.

2.3.2.2 BEAST

Bayesian phylogenetic inference has come a long way, and multiple data types can be combined to address complex evolutional questions [71]. BEAST uses MCMC to marginalize over the uncertainty on the phylogenies and evolutionary parameters [72]. BEAST 2 has also recently been updated with improved functionality in reconstructing pathogen spread and analyzing linguistic data sets [73]. The posterior distribution over trees T and parameters θ is then:

$$P(T, \theta \mid D) = \frac{P(D \mid T, \theta) \cdot P(T, \theta)}{P(D)} \tag{2.4}$$

Where:

$P(D \mid T, \theta)$ is the likelihood

$P(T, \theta)$ is the prior over trees and parameters

$P(D)$ is the marginal likelihood (normalization factor)

2.3.3 Population-Based and Carrier Frequency Estimation Techniques

2.3.3.1 ExAC

Exome Aggregation Consortium (ExAC) provides comprehensive exome data identifing millions of rare variants in hereditary disease such as DNA repair gene polymorphisms in cancer and pathogenic genotypes in Mendelian disorders. This is an important resource for evaluating variant frequency and disease impression [74–76]. It processes exome sequences from thousands of people, building a powerful reference dataset of genetic variant frequencies.

To calculate the frequency f of a variant in the population:

$$f = \frac{n_{\text{variant}}}{n_{\text{total}}} \tag{2.5}$$

Where:

n_{variant} is the number of chromosomes containing the variant,

n_{total} is the total number of chromosomes examined.

2.3.3.2 gnomAD

The Genome Aggregation Database (gnomAD) is an indispensable resource for variant annotation, offering allele frequencies and also genetic data across different populations, an expansion substantially beyond ExAC [77–79]. It assists in determining the pathogenicity of variants associated with dystonia and familial hemophagocytic lymphohistiocytosis (fHLH), and enables the identification of new disease-gene associations.

For each variant, gnomAD reports:

Allele Count (AC): The number of times the variant allele is observed.

Allele Number (AN): Sum of alleles sequenced at this position.

Then allele frequency is:

$$\text{AF} = \frac{\text{AC}}{\text{AN}} \tag{2.6}$$

2.3.4 Machine Learning and Data Science Techniques

2.3.4.1 TensorFlow

TensorFlow provides a development for the use of deep artificial neural networks (ANNs) in genomic sequence classification and genotype-phenotype mapping. In this aspect, AI systems based on TensorFlow such as FDNN exhibited an accuracy up to 99% in the prediction of rare genetic diseases, and systems such as the PhenoBrain outperform expert clinicians in diagnostics, underscoring the importance of TensorFlow in advancing genomic and clinical diagnostics [80–82]. Here is the working:

Forward Propagation: Given input features $\mathbf{x} \in \mathbb{R}^n$, the output of a single-layer neural network is:

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{b}, \quad \hat{y} = \sigma(\mathbf{z}) \tag{2.7}$$

Where:

$W \in \mathbb{R}^{m \times n}$: weight matrix

$b \in \mathbb{R}^m$: bias vector

σ: activation function

$\hat{y}$: predicted output

Loss Function (Binary Cross-Entropy for disease classification):

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{2.8}$$

Gradient Descent (Backpropagation Step):

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta L(\theta^{(t)}) \tag{2.9}$$

2.3.4.2 PyTorch

The open-source machine learning library PyTorch is capable of building deep learning models, and it is well-suited for genomic data analysis and hereditary disease studies [83–86]. Most of the advanced models such as the Cross-Modal Embedding Integrator (CMEI), exploit formation of the knowledge graph to predict the disease-gene, thus facilitating understanding of disease mechanisms and treatment targets. RNNs for Sequence Modeling: Let x_t be the input, the hidden state update is:

$$h_t = \tanh(W_{ih} x_t + W_{hh} h_{t-1} + b_h) \tag{2.10}$$

2.3.4.3 Scikit-learn

Machine learning approaches, such as automated machine learning (AutoML), are increasingly being used to augment the prediction and diagnosis of genetic diseases [87–90]. AutoML simplifies algorithm and hyperparameter selection, making it accessible even to users with limited expertise in genetic analysis. For example, deep learning models have recently proven effective in ASD (autism spectrum disorder) classification from genomic data, attaining up to 88% accuracy [88]. Scikit-learn provides algorithms for classification, clustering, and dimensionality reduction, which are directly applicable to tasks like patient subgrouping or variant prioritization. For a binary classifier, the decision boundary is determined by:

$$f(x) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b) \tag{2.11}$$

2.4 Comparative Analysis of Existing Approaches

TABLE 2.1: Comparative Analysis of Existing Approaches

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[21]	Land et al. (2015)	Bacterial Genome Sequencing	WGS, Metagenomic Analysis	Understand bacterial diversity	Bacterial species across environments	Infectious diseases	Genetic relationships, core/pan-genome analysis
[22]	Witney et al. (2016)	Whole Genome Sequencing	Xpert® MTB/RIF, MTBDRplus/sl	Drug susceptibility, transmission	TB patients	Tuberculosis	Rapid drug resistance detection
[23]	Carss et al. (2017)	Whole Genome Sequencing	WGS, WES, Manta	Identify IRD variants	722 with IRD	Inherited retinal disease	56% detection rate
[24]	Bick et al. (2019)	Whole Genome Sequencing	WGS, CMA	Identify rare disease variants	Infants with rare diseases	Rare genetic diseases	56% detection rate
[25]	Clark et al. (2019)	Rapid WGS	rWGS, CNLP, DRAGEN	Rapid diagnostics in ill children	101 children	Genetic diseases	Diagnosed in 20:10 hrs

Continued on next page

TABLE 2.1: Comparative Analysis of Existing Approaches (continued)

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[26]	Turro et al. (2019)	Whole Genome Sequencing	WGS, MOON	Accurate diagnoses in rare diseases	13,037 participants	Rare genetic diseases	16.1% diagnostic yield
[27]	Hananya et al. (2020)	Whole Genome Sequencing	SQL, gnomAD, ClinVar	Frequency of AR-IRDs	13,037 subjects	AR inherited retinal diseases	2.7B carriers worldwide
[28]	Hananya et al. (2022)	CRISPR-SPR	SPR Chips, qPCR	Detect DMD mutations	DMD patients	Duchenne MD	Exon deletions detected accurately
[29]	Ferreira et al. (2021)	ICIMD	gnomAD, ClinVar, BeviMed	Classify metabolic disorders	1,450 disorders	Inherited metabolic disorders	Comprehensive classification

Continued on next page

TABLE 2.1: Comparative Analysis of Existing Approaches (continued)

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[33]	Benito-Vicente et al. (2018)	FH	Genetic tests, therapy	Diagnose/treat FH	FH individuals	Familial Hypercholesterolemia	FH management strategy
[34]	Boland et al. (2018)	Lynch Syndrome	NGS, stats analysis	Study familial colorectal cancer	Cancer-prone populations	Colorectal Cancer	MMR mutations insights
[35]	Crowley et al. (2023)	Genome Sequencing	DNA Barcoding	Sequence arthropods	British arthropods	N/A	Biodiversity mapping
[36]	Venner et al. (2022)	WGS, Variant Calling	Genomic data	Return genomic results	US cohort	Hereditary, pharmacogenomics	Health-related results shared
[37]	Lin et al. (2021)	Nanopore	INC-Seq, basecalling	Improve sequencing	Bio samples	Cancer, infections	Enhanced diagnostics

Continued on next page

TABLE 2.1: Comparative Analysis of Existing Approaches (continued)

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[38]	Owen et al. (2022)	WGS, Targeted Seq	Genomics, interpretation	Scale genetic diagnosis	Newborns	Genetic diseases	Improved diagnosis/management
[39]	Schuler et al. (2022)	NGS, WES	Bio-informatics	Diagnose genetic disorders	All ages	Genetic diseases	Improved diagnostics
[40]	Chiu et al. (2020)	NGS, WES	Bio-informatics	Improve genodermatoses diagnosis	Skin disease patients	Genodermatoses	Better gene discovery
[41]	Iossifov et al. (2014)	WES	Variant calling	ASD understanding	ASD families	Autism Spectrum Disorder	De novo mutations linked
[42]	Aitken et al. (2020)	NGS, Damage Assays	Statistical analysis	Study paternal impact	Males and offspring	Genetic disorders	Paternal mutation insights

Continued on next page

TABLE 2.1: Comparative Analysis of Existing Approaches (continued)

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[43]	Ronemus et al. (2014)	CGH, WES	Mutation detection	Study ASDs	ASD families	Autism Spectrum Disorder	De novo mutation insights
[44]	Sadikovic et al. (2021)	Methylation Arrays	SVM Classification	Diagnose Mendelian disorders	207 suspected cases	Rare Genetic Disorders	27.6% diagnostic yield
[45]	Posey (2019)	Genome Sequencing	Karyotyping, CMA	Diagnose rare disorders	Mendelian patients	Rare Genetic Disorders	Better sensitivity/yield
[46]	Lee et al. (2020)	RNAseq, Genome Seq	RNA analysis	Improve Mendelian diagnosis	113 referred cases	Rare Genetic Disorders	7% increase with RNAseq
[47]	Shen et al. (2015)	WES, WGS	Variant detection	Study rare Mendelian cases	US population	Rare Genetic Disorders	Improved diagnostics

Continued on next page

Table 2.1: Comparative Analysis of Existing Approaches (continued)

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[48]	Wigby et al. (2024)	Genome Sequencing	First-line GS	Diagnose rare disorders	13,000+ patients	Rare Genetic Disorders	45% diagnostic yield
[49]	Wojcik et al. (2024)	Genome Sequencing	First-line GS	Diagnose rare disorders	744 families	Rare Genetic Disorders	29.3% yield
[50]	Rahit & Tarailo-Graovac (2020)	Genetic Modifiers	OligoPVP, VarCoPP	Study rare modifier genes	Rare disease patients	Rare Mendelian Diseases	Modifier gaps identified
[51]	Beaudin et al. (2017)	AR Ataxias	Clinical review	Classify ataxias	130 articles	AR Ataxias	45 disorders cataloged
[52]	Palau & Espinós (2006)	AR Cerebellar Ataxias	Systematic review	Define ARCA types	Literature analysis	AR Cerebellar Ataxias	Causative genes listed

Continued on next page

Table 2.1: Comparative Analysis of Existing Approaches (continued)

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[53]	Monies et al. (2017)	Recessive Mutations	Exome, Autozygome	Study dominant gene roles	Saudi population	Various disorders	11 recessive variants found
[54]	Gabow et al. (1992)	ADPKD	Survival, modeling	Study renal progression	ADPKD families	ADPKD	Renal function factors listed
[55]	Bengani et al. (2017)	SATB2	WES, fibroblast analysis	Characterize SATB2 impacts	SATB2 patients	Syndromic disability	19 mutations identified
[56]	Pullabhatla et al. (2018)	SLE	Rare variant burden	Identify candidate genes	SLE trios, large cohort	SLE	14 new genes linked
[57]	Myers et al. (2017)	PPP3CA	WES	Study neurodevelopmental genes	6 PPP3CA cases	Neuro developmental disease	PPP3CA linked to epilepsy

Continued on next page

Table 2.1: Comparative Analysis of Existing Approaches (continued)

Cit#	Article	LSGS	Methods	Purpose	Population	Disease	Return of Results
[58]	Redmond et al. (2018)	Malaria	WGS, genotyping	Study malaria transmission	Senegal isolates	Malaria	De novo mutation tracking
[59]	Hou et al. (2015)	AML	WES, stats	Study TP53 in AML	500 AML patients	AML	Poor prognosis association
[60]	Li et al. (2017)	Preterm Birth	WGS, stats	Link fetal mutations to PTB	816 trio families	Preterm Birth	Mutations linked to development

2.5 Known Gaps Identified During the Process

2.5.1 Technological and Bioinformatics Limitations

Studies report major technical challenges in hereditary disease detection. Key issues include under-ascertainment of pathogenic variants due to variant calling limits, poor coverage in repetitive regions, and weak interpretation of non-coding areas. Nanopore/NGS errors, base calling, and annotation remain difficult. CNLP tools often yield false negatives, showing the need for algorithm improvements and stronger EHR integration. SVs, CNVs, and mosaicism detection remain constrained. Lack of standardized bioinformatics pipelines across centers is a major gap.

2.5.2 Interpretation and Clinical Translation Challenges

Even when variants are found, interpreting them into clinical insight is complex. This is particularly relevant for non-coding and regulatory variants where there are tissues-specific gene expressions, it is also the case for IRD. The databases are also incomplete, poorly standardized and do not connect the genetic variants to the phenotypes. Phenotypic diversity within the same genotype also makes the diagnosis more difficult. Even in the setting of rare diseases, for example FACE, PPP3CA-related epilepsy, or SLE, the association between genotype and phenotype is not well established. Technical improvements (annotation tools and functional studies) are key for translation.

2.5.3 Data Representation and Population Diversity

Lack of diversity in genomic data is a key issue. Studies warn that findings from European/North American groups may not apply elsewhere. This affects mutation interpretation, allele frequencies, and prediction tool accuracy. Fair access and inclusive research participation are essential. Stronger engagement with local

populations is needed to build diverse genomic databases for rare Mendelian and autosomal recessive conditions.

2.5.4 Regulatory, Infrastructure, and Methodological Gaps

Regulatory and infrastructure issues continue to limit progress. No uniform standards exist for testing, sequencing, or interpretation. There's urgent need for standard pipelines, lab quality control, and clinical guideline application. New tools like CRISPR-SPR lack validation across diseases. Genome sequencing is underused due to cost and lack of adult data. Predictive models are weak for familial cancers and other conditions, underlining lifelong screening needs. Concerns persist on result reporting and research-to-clinic translation.

2.6 Proposed Highly Suitable Model

Clustered Genomic Fingerprinting (CGF) model, a robust model to identify inherited diseases using family whole genome sequencing data in family-based designs. CGF is oriented towards Mendelian inheritance patterns (autosomal dominant and recessive) and can detect non-Mendelian variations as de novo mutations and compound heterozygosity.

The CGF pipeline starts with a Genome Standardization layer which cleans the WGS by stripping off metadata and redundant call entries and filling missing genotype entries using biologically informed rules. This sanitized data is further processed by a Dual-Encoding Schema (DES) for chromosome identifiers and genotype formats normalization, which enables machine-learning use while still being biologically meaningful [91].

It has also been used successfully for the epidemiological subtyping of bacterial pathogens, e.g., Campylobacter jejuni and Campylobacter coli. CGF has been

demonstrated to provide high discriminatory power, especially for C. jejuni isolates, and can by used in addition to biotyping methods to produce accurate values of strain diversity and epidemiological relatedness [92].

Then the encoded data are subject to K-Means Genomic Fingerprinting (KGF), a clustering algorithm that groups loci according to allelic similarity showing conserved inheritance blocks and outlier patterns associated with afoot mutant loci. The clusters are subsequently employed in constructing a Dynamic Pivot Matrix Construction (DPMC) that illustrates genotypic concordance between family members that is necessary to localize mutation footprints and sibship level variations.

In order to compensate for missing data, the model makes use of Cluster-Level Genotypic Consensus Imputation (CGCI), which is the imputation of missing values with the modal value for each cluster in order to preserve data quality and reduce bias. Lastly, EGMC provides an integrated de novo mutation and complex inheritance mapping approach capable of presenting a navigable log of mutations, providing candidate-regions for de novo mutations and complex inherited patterns.

Anticipated outcomes of CGF are improved ability to assign disease causing inheritance either to or against a specific variant, early detection of rare or novel mutations and improved interpretation of inheritance and family models. By combining bioinformatics preprocessing, unsupervised learning, and mutation visualization, CGF forms a high-resolution structure that is favorable for the improvement of genetic diagnostics, genetic counseling, and personalized medicine.

In conclusion, CGF facilitates to translate NGS results into clinical management, and offers an effective, widely applicable NGS-based approach for family-based and laboratory-based genetic analyses of hereditary diseases.

Chapter 3

Research Methodology

3.1 Introduction and Background

Investigation of inherited diseases in medical genomics is dedicated to the explication of complex genetic phenomena such as Mendelian inheritance, de novo mutations and compound heterozygosity. Although a number of conditions such as cystic fibrosis and Huntington's disease have been previously characterised as simple Mendelian diseases, genome sequencing is revealing additional levels of complexity to classical diagnostic approaches.

These diseases are frequently the result of structural or point mutations, occasionally silent over generations, and additionally afflicted by epigenetic factors. Conventional means of diagnosis based on clinical observations alone or studies of only individual variants are inadequate to capture the entire range of genetic inheritance, and a more sophisticated computational models are needed [93].

Recent methods, including machine learning classifiers, Hidden Markov Models and deep learning approaches have been developed to predict the disease-causing variants; however, most of them do not take full advantage of the family-based genomic context. Current tools present limitations in interpretability, need large

population references, and cannot easily handle missing genotypic information, all of which weaken their clinical validity.

Clustering approaches, despite being powerful, have been used relatively rarely in the study of genotypic relatedness between individuals in a pedigree. In addition, lack of normalization to the same overall family dataset causes unreliable mutation scoring. These concerns have highlighted the need to develop computational models that were interpretable and focused on the family but could still accurately and stably infer inheritance, just as fingerprinting identifies symptoms of diabetes with quantifiable differences [94].

3.2 Experimental Setup

3.2.1 Dataset Description and Format

The raw input is based on individual CSV files per family member (i.e., Father, Mother, Child 1, Child 2, Child 3) [95]. It generates a collection of files representing tabular data on each genetic variant, each row per locus should at least include the chromosome number, the base pair position and the observed genotype.

Mathematically, the data from each subject i is defined as a matrix:

$$G^{(i)} = \begin{bmatrix} \text{chr}_1 & \text{pos}_1 & \text{geno}_1 \\ \text{chr}_2 & \text{pos}_2 & \text{geno}_2 \\ \vdots & \vdots & \vdots \\ \text{chr}_n & \text{pos}_n & \text{geno}_n \end{bmatrix}, \quad i \in \{1, 2, \ldots, N\} \tag{3.1}$$

where N is the number of family members, chr_j is a chromosome identifier, pos_j is the genomic position, and geno_j is the observed genotype at that position.

Each dataset is tagged with its familial identity and then deduplicated:

$$G'^{(i)} = \text{dedup}(G^{(i)}), \quad \forall i \tag{3.2}$$

In order to get rid of irrelevant or duplicate entries. The missing data for genotypes at a given interval are then biologically imputed by forward-filling the data, which assumes the continuity of loci in chromosome segments neighboring loci tend to have similar inheritance patterns.

3.2.2 Data Cleaning and Filtering

Preprocessing Data cleaning is performed before clustering for normalization, and noise reduction. This includes:

3.2.2.1 Deduplication Function:

$$G''^{(i)} = \{x \in G'^{(i)} \mid x \notin \text{duplicates}(G'^{(i)})\} \tag{3.3}$$

3.2.2.2 Forward Imputation

Given a locus j with missing genotype $\text{geno}_j = \emptyset$, forward filling assumes:

$$\text{geno}_j = \text{geno}_{j-1}, \quad \text{if } \text{chr}_j = \text{chr}_{j-1} \tag{3.4}$$

This assumption is based on the genetic consistency at the same region of a chromosome. All cleaned and imputed matrices $G''^{(i)}$ are vertically concatenated to produce a unified matrix:

$$G = \bigcup_{i=1}^{N} G''^{(i)} \in \mathbb{R}^{m \times 4} \tag{3.5}$$

where m is the total number of unique loci across all family members, and the other column is the origin from one individual.

3.2.3 Dual Encoding Scheme

Genomic categorical attributes are required to be numerically encoded to be used with unsupervised machine learning (K-Means). However, pre-label encoding is not biologically aware and may result in biological variance loss as it tends to merge one gene into different label categories. The mapping of specific mutations that exhibit co-segregation with disease phenotypes underscores the necessity of representing these categorical genomic values such that biological context is maintained [96]. For encoding chromosomes and genotypes with family context, the CGF proposed the Dual Encoding Scheme(DES).

3.2.3.1 Chromosome Encoding

Let C be the set of all observed chromosomes:

$$C = \{\text{chr}_1, \text{chr}_2, \ldots, \text{chr}_k\} \tag{3.6}$$

Then the encoding function is defined as:

$$f_C : C \to \{0, 1, \ldots, k-1\} \tag{3.7}$$

3.2.3.2 Genotype Encoding

Let T denote the collection of the set of unique genotypes for every member of the family:

$$T = \{\text{geno}_1, \text{geno}_2, \ldots, \text{geno}_k\} \tag{3.8}$$

The encoding function is defined as:

$$f_T : T \to \{0, 1, \ldots, l-1\} \tag{3.9}$$

The unified dataset is then converted one row at a time as follows:

$$g_j = [f_C(\text{chr}_j), \text{pos}_j, f_T(\text{geno}_j), s_j] \tag{3.10}$$

Here, s_j is the source label. This encoding has the advantage of enabling machine learning methods to work on a set of consistent numerical features, while keeping biologically interpretable mappings.

3.2.4 Normalization for ML Readiness

Normalization of features is essential to ensure equal contribution to the clustering objective function in K-Means, preventing skewed Euclidean distances due to varying scales. This principle reflects the structured approach of the Machine Learning Technology Readiness Levels framework, which aims to avoid technical liability and misaligned objectives, ensuring robust systems across diverse applications [97].

Let the encoded matrix $G_{\text{enc}} \in \mathbb{R}^{m\times 3}$ have columns:

x_1: Encoded chromosome

x_2: Position

x_3: Encoded genotype

Each feature is normalized with respect to min-max normalization:

$$x'_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)} \quad , \quad \forall j \in \{1, 2, 3\} \tag{3.11}$$

The normalized matrix:

$$G_{\text{norm}} = \begin{bmatrix} x'_1 & x'_2 & x'_3 \end{bmatrix} \tag{3.12}$$

$$E_{\text{geno}} = x'_3 \tag{3.13}$$

which is given to the K-Means clustering algorithm in the next phase.

In addition to preserving data integrity and preparation for unsupervised learning, this layer maintains biologically plausible representation. The CGF pipeline can therefore identify inheritance blocks, outliers and potential de novo mutations at higher resolution.

3.3 Proposed Framework: Clustered Genomic Fingerprinting

3.3.1 Flowchart of the Methodology

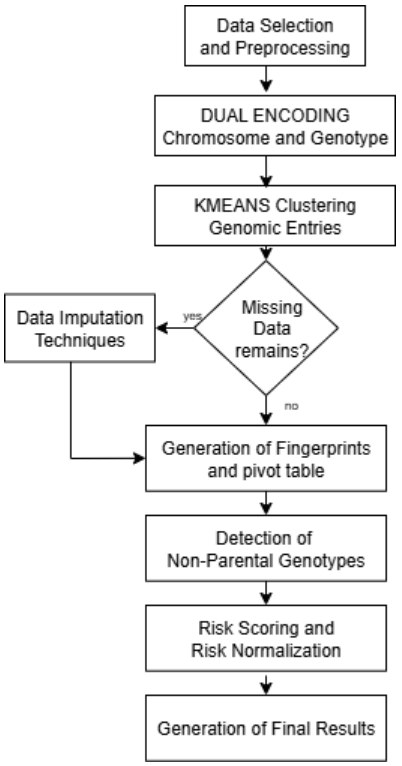

FIGURE 3.1: Flowchart of the Proposed Methodology

3.3.2 Overview of CGF Methodology

The Clustered Genomic Fingerprints (CGF) method, represents a new and statistically sound/rigorous approach to the description of patterns of genetic variation in related WGS data. CGF decrease the computation complexity and enhance the accuracy in the complex data sets [98]. Extending this concept, the Clustered Genomic Fingerprints (CGF) applies clustering to genomic data to generate high resolution maps of mutation that accurately discriminate between inherited and de novo mutations, furthering the analysis of rare disease and mutation heritability [99].

3.3.3 Mathematical Formulation of CGF

The mathematical framework of the CGF model is based on encoding complicated categorical genomic patterns into machine readable numerical forms, unsupervised learning-based pattern recognition, and for mapping the genetically interpretable information to biologically interpretable information through cluster-based mutation pattern will be described. The high-level steps of this approach are encoding, normalization, clustering, matrix transform, and mutation scoring. Each step is described here:

Step 1: Dataset Definition and Notation

Let $D = \{G_1, G_2, \ldots, G_5\}$ be a set of genome datasets, where each $G_i \in \mathbb{R}^{n_i \times 3}$ corresponds to one family member among:

a) $G_1, G_2, G_3 \in \{\text{Child-1, Child-2, Child-3}\}$

b) $G_4 = \text{Father}$

c) $G_5 = \text{Mother}$

A unioned dataset is defined as $D \in \mathbb{R}^{N \times 2}$, where:

G_{norm} is the normalized encoded genotype vector representing $[x'_1,\ x'_2,\ x'_3]$.

source_i represents the i^{th} individual (e.g., Father, Mother, Child-1, Child-2, Child-3).

$$D = \bigcup_{i=1}^{5} (G_{\text{norm}}, \text{source}_i) \tag{3.14}$$

Step 2: KMeans Clustering for Genomic Fingerprinting

Let $X \in \mathbb{R}^{N\times 3}$ be the normalized encoded matrix:

$$X = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(N)} \end{bmatrix}^T = [G_{\text{norm}}], \quad G_{\text{norm}} \in \{x'_1,\ x'_2,\ x'_3\} \tag{3.15}$$

KMeans clustering is applied:

$$\min_{\{\mu_k\}_{k=1}^{K}} \sum_{i=1}^{N} \min_{k\in\{1,\ldots,K\}} \left\| x^{(i)} - \mu_k \right\|^2 \tag{3.16}$$

where:

$K = 6$ is the number of clusters (chosen experimentally via domain intuition),

$\mu_k \in \mathbb{R}^3$ are centroids of clusters,

$c(i) \in \{1, \ldots, K\}$ is the cluster index assigned to each point.

The output is a vector of cluster assignments:

$$C = [c(1),\ c(2),\ \ldots,\ c(N)] \tag{3.17}$$

Step 3: Pivot Table Construction

A pivot matrix $M \in \mathbb{R}^{K\times 5}$ is built, where rows represent clusters $k = 0$ to $K - 1$ and columns represent family members.

Each cell hold decoded genotype values from E_{geno}^{-1} at cluster-wise dominant positions.

Missing values $M_{k,i} = \emptyset$ are filled via:

1. Forward fill: $M_{k,i} \leftarrow M_{k-1,i}$

or

2. Backward fill: $M_{k,i} \leftarrow M_{k+1,i}$

Modal imputation within cluster:

$$M_{k,i} \leftarrow \text{mode}\left(\left\{\text{geno}_j \mid c(j) = k \wedge \text{source}_j = i\right\}\right) \tag{3.18}$$

Step 4: Non-Parental Genotype Flagging

For each child $C_i \in \{\text{Child-1}, \text{Child-2}, \text{Child-3}\}$, defined as:

$$f_k^{(i)} = \begin{cases} 1, & \text{if } M_{k,C_i} \neq M_{k,F} \vee M_{k,C_i} \neq M_{k,M} \\ 0, & \text{otherwise} \end{cases} \tag{3.19}$$

where F and M are the column indices for Father and Mother, respectively.

These form binary indicator columns for each child:

$$f_K^{(i)} = [f_1^{(i)}, f_2^{(i)}, \ldots, f_K^{(i)}] \tag{3.20}$$

Step 5: Mutation Risk Score and Normalization

The raw mutation risk score for each cluster is defined as:

$$R_k = \sum_{i=1}^{3} f_k^{(i)} \quad \text{for } k = 1, \ldots, K \tag{3.21}$$

This gives $R_k \in \{0, 1, 2, 3, 4, 5\}$, representing how many children show non-parental genotypes in cluster k.

Normalized risk score:

$$R_k^{\text{norm}} = \frac{R_k}{3} \tag{3.22}$$

so that all scores fall in the unit interval $[0, 1]$, which is suitable for later ML use or visualization.

Step 6: Final Mutation Fingerprint

The output fingerprint table $F \in \mathbb{R}^{K \times (5+5)}$ contains:

a) Original genotype matrix M

b) 3 Non-parental flags

c) R_k and R_k^{norm}

TABLE 3.1: Non-Parental Mutation Flags and Risk Scores per Cluster

Cluster#	**Child-1**	**Child-2**	**Child-3**	**M_Risk**	**Norm_Risk**
0	0	0	0	0	0.000000
1	0	1	1	2	0.333333
2	0	0	0	0	0.000000
3	0	2	2	4	0.666667
4	0	1	1	2	0.333333
5	0	2	2	4	0.666667

3.4 Clustering, Training Mutation Identification

The core of the CGF framework is the clustering procedure described in this section. The K-Means clustering of the genotype profiles of family members after normalization yields biologically relevant clusters of inheritance that act as references for the identification of de novo or non-mendelian variants in the offspring. In addition, the statistical distances from the cluster centroids provide mutation risk evaluation metrics.

3.4.1 K-Means Clustering with Family-Wide Genotypes

The input for the clustering algorithm would be the matrix of SNP records standardized as follows:

$$G_{\text{norm}} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}, \quad x^{(i)} \in \mathbb{R}^d \tag{3.23}$$

where $d = 3$ dimensions (chromosome, position, and genotype), and m is the total number of genomic loci across all family members. The K-Means clustering algorithm divides the dataset into K clusters $C_1, C_2, \ldots, C_K$, to minimize the objective funtion:

$$\arg\min_{C} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \tag{3.24}$$

where:

$x \in \mathbb{R}^d$ is a data point

$\mu_k \in \mathbb{R}^d$ is the centroid of cluster C_k

$\|\cdot\|$ denotes the Euclidean norm

Initialization:

Randomly select K data points as initial centroids $\mu_1^{(0)}, \ldots, \mu_k^{(0)}$. At iteration $t = 0$, this gives:

$$\mu_k^{(0)} = x^{(r_k)}, \quad r_k \in \{1, 2, \ldots, m\} \tag{3.25}$$

Expectation: Assignment of Points to Clusters

For each data point x^i, assign it to the cluster with the nearest centroid:

$$C_k^{(t)} = \left\{ x^{(i)} : \|x^{(i)} - \mu_k^{(t)}\|^2 \leq \|x^{(i)} - \mu_j^{(t)}\|^2, \forall j \in \{1, 2, \ldots, k\} \right\} \tag{3.26}$$

Maximization: Update Cluster Centroids

Compute new centroids as the mean of all data points assigned to each cluster:

$$\mu_k^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{x^{(i)} \in C_k^{(t)}} x^{(i)} \tag{3.27}$$

Convergence Criterion:

Repeat expectation and maximization steps until:

$$\sum_{k=1}^{k} \|\mu_k^{(t+1)} - \mu_k^{(t)}\|^2 < \epsilon \tag{3.28}$$

3.4.2 Centroid in Genomic Context

In the CGF pipeline, the clustering-based centroid selection is employed to improve the precision and recommendation performance by obtaining the implicit data representations [100]. Each centroid $\mu_k \in \mathbb{R}^3$ corresponds to interpretable biological semantics:

3.4.2.1 Dimension 1

Is a measure of the chromosomal mean of the cluster. since chromosomes are numbered, this is a pseudo-coordinate.

3.4.2.2 Dimension 2

This corresponds to the average position in base pairs. Clusters in neighboring regions are useful to capture linkage disequilibrium or block.

3.4.2.3 Dimension 3

Suggests the dominant genotype among members of the cluster. Values beyond this range in children might suggest that mutation events occurred.

Let $\mu_k^{(t)}$ be the third component of centroid μ_k. Then:

If child c's encoded genotype $x_C^{(3)} \in [\mu_k^{(3)} - \delta, \mu_k^{(3)} + \delta]$, where δ is a confidence threshold, it is flagged as anomalous.

3.4.3 Mutation Risk Detection

After family-basedgenotypic data clusering, the mutation risk is then assessed by comparing observed child data points with their clustered centroids. This approach will contribute valuable insights into genetic predisposition and clinical manifestation, guiding for personalized cancer risk assessment and management [101–103].

Let $x_c^{(i)} \in \mathbb{R}^3$ be a child's standardized genotype vector assigned to cluster C_k with centroid . Here, the mutation risk score is measured by the Euclidean distance:

$$r^{(i)} = \|x_c^{(i)} - \mu_k\|^2 \tag{3.29}$$

It will define a mutation threshold θ, which is empirically or statistically obtained, Then:

$$\text{MutationFlag}^{(i)} = \begin{cases} 1, & \text{if } r^{(i)} > \theta \\ 0, & \text{otherwise} \end{cases} \tag{3.30}$$

Biological support was improved by implementing a consensus voting rule across the familial data: If parents are of the genotype g_p and child has the genotype g_c, we have:

$$\text{Non-Parental Mismatch} = \mathbf{1}_{(g_c \notin \{g_{\text{father}}, g_{\text{mother}}\})} \tag{3.31}$$

which is accumulated by clustering anomaly to complete mutation detection.

3.4.4 Training and Testing Approach

Despite of clustering being an unsupervised task, CGF is applied in a semi-supervised way. Only parental and validated inherited child variants are considered during training for clustering. Child genotypes that were unseen are then projected into this clustered space at test time. Let:

1. $D_{\text{train}} \subseteq G_{\text{norm}}$ be the training set (parents and inherited variants)
2. $D_{\text{test}} = \{x_c^{(i)}\}$ be the child dataset to evaluate

For each $x_c^{(i)} \in D_{\text{test}}$, the prediction procedure is:

a) Assign to nearest cluster C_k

b) Compute risk score $r^{(i)}$

c) Compare with threshold θ

d) Flag as high-risk mutation if $r^{(i)} > \theta$

3.5 Model Evaluation, Visualization and Comparative Analysis

3.5.1 Metrics Used in Other Models

Existing genotype-based disease classification and mutation detection systems, especially those utilizing machine learning or statistical inference have also been evaluated with various of performance measures measuring true and false positive and negative detection of known variants. Here are a few of the standard metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.32}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.33}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{3.34}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.35}$$

AUC-ROC computes the trade-off in true positive rate and false positive rate as its threshold is varied.

Matthews Correlation Coefficient (MCC) suitable for imbalanced datasets:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.36}$$

These are generally only useful when models are trained using labelled mutation data, which is not commonly available for hereditary or rare diseases.

3.5.2 Evaluation of CGF Model

The evaluation of the CGF approach uses a structurally different logic due to its unsupervised nature and its primary focus on de novo mutation hotspots along with non-Mendelian genotypes. A set of internal clustering metrics combined with risk scoring agreement and biological plausibility are used to validate the model:

3.5.2.1 Silhouette Coefficient

The proposed unsupervised weighting function that is based on the Silhouette index-algorithm to increase the performance of clustering and measure how similar a genotype vector is to its own cluster compared to others [104]:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.37}$$

where:

$a(i)$ is the average intra-cluster distance for point i,

$b(i)$ is the minimum average distance to other clusters.

The overall silhouette score $S \in [-1, 1]$ measures how well the data clusters into groups, with higher values reflecting denser and more distinct clusters.

3.5.2.2 Intra-Cluster vs Inter-Cluster Variance

The cohesion–separation ratio ρ is defined as:

$$\rho = \frac{\sum_{k=1}^{K} \sigma_k^2}{\text{Between-cluster variance}} \tag{3.38}$$

where lower values of ρ characterize high intra-coherence and external separation, ideal for the identification of mutational regions [105].

3.5.2.3 Mutation Risk Distribution Consistency

The pattern of normalized mutation scores $\hat{R}_k \in [0, 1]$ across clusters and children is being compared. A biologically plausible model will show:

Sparse high-risk clusters (de novo signals are rare),

Cluster-specific pattern of deviation that aligns with known mutation processes.

3.5.2.4 Comparative Evaluation Using Random Forest

Random Forest is an ensemble learning technique that creates multiple decision trees using bootstrapped samples of data and randomly selected features at each split. This encourages model diversity and overfitting reduction. Majority voting is then applied to make the final predictions. Random Forest is used after CGF clustering of the discovered genotype cluster patterns for the assessment of consistency and learnability.

Each decision tree in the ensemble $h(x, \theta_K)$ is built using a random subset of features and data samples:

$$H(x) = \text{majority\_vote}\left\{h_1(x), h_2(x), \ldots, h_N(x)\right\} \tag{3.39}$$

where:

a) x is the input sample

b) θ_K are random variables controlling tree growth and feature splits

c) N is the number of trees

d) Output $H(x)$ is the predicted class

Then the evaluation is computed by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.40}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{Actual Positive}} \tag{3.41}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{Predicted Positive}} \tag{3.42}$$

Evaluates classification performance between predicted and true clusters. This provides an indirect comparison of CGF's unsupervised results with supervised classifiers trained on labeled mutation datasets.

This provides an indirect comparison of CGF's unsupervised results with supervised classifiers trained on labelled mutation datasets.

3.5.3 Strengths and Limitations

3.5.3.1 Strengths

The proposed CGF framework demonstrates several strengths. It enables family-based unsupervised mutation analysis, moving beyond case-control approaches by directly exploiting internal family relationships to identify mutational patterns. It also integrates clustering with Mendelian patterns, through the association of K-Means statistics with parental genotype deviations providing both statistical power and biological interpretation. Importantly, it provides concrete quantification of the risk of mutations in terms of a mathematically normalised and interpretable score, steering clear from the black-box aspect inherent to many ML models. Lastly, it provides a versatile protocol with the ability of reusing the calculated fingerprints in supervised learning or expansion for additional mutation filtering.

3.5.3.2 Limitations:

Despite these advantages, the model has significant limitations. It is agnostic to the ground truth dependency, as the evaluations do not have access to correctly labeled mutation data and rely on internal metrics together with biological plausibility assumptions. The results are also affected by the number of cluster resolution dependence, because K controls fingerprint granularity and detection sensitivity. Also, the details of the genotype encoding imply particular biological interpretations that could differ between populations. Finally, scalability is an issue because of the increase in memory and compute for pivot matrices and cluster updates when considering whole-genome datasets and larger families.

Chapter 4

Results and Discussion

4.1 Introduction

CGF analysis of the family-based whole genome sequencing (WGS) data applied for the identification of inherited diseases. The focus of the study was to find Mendelian inheritance patterns and deviations, e.g. de novo mutations, by combining clustering, genotype encoding and mutation risk scoring.

The findings are organized around key clusters derived from the KMeans algorithm, each representing distinct inheritance behaviors. For each cluster, inheritance patterns were analyzed, non-parental genotypes were identified, and mutation risks were was evaluated for each cluster. Visual schemes and quantitative overviews reinforce the interpretation of these findings.

This chapter will emphasize how CGF may usefully reveal both expected and anomalous genetic transmissions, showcasing its potential in the recognition of complicated inheritance patterns overlooked by usual perspectives and standard research strategies.

4.2 Dataset Overview and Experimental Context

4.2.1 Description of Family WGS Dataset

The dataset for this analysis consists of whole genome sequencing (WGS) data from a family. It includes genomic data from five persons [95], i.e. the father, the mother and three children. Data from each family member is saved into a separate CSV file to facilitate the analysis and comparison across individuals. Each of the files generally consists of approximately 600,000–650,000 instances, which correspond to SNPs identified from sequencing. The dataset records several attributes for each SNP:

a) Chromosome: The chromosome number (1–22, X, or Y) where the SNPs are located.

b) Position: The base-pair position on the chromosome.

c) Genotype: The observed allelic composition for that SNP in the individual (e.g., AA, AG, GG, TT).

The comparison of genetic variation among the family members becomes possible through this organized representation. Chromosomes and positions are placed in alignment, enabling tracking of inheritance patterns to determine whether the allelic follows autosomal dominant, autosomal recessive or a de novo mutation. Thus, the data set provides the foundation for CGF, assists in capturing deviations from Mendelian inheritance and estimating mutation risk over generations.

4.2.2 Preprocessing Summary

Data was pre-processed before analysis to make it ready for the model. The missing genotype values filled by using the combination of a forward fill, a backward fill and

a cluster-based fill. Subsequently, the values of genotype and chromosome were transformed into numerical form to perform cluster analysis. Data was organized by chromosome number and position and was transposed into a table that made it easier to compare members. Positions where any child had a completely different genotype compared to both parents were coded as possible de novo mutations.

4.2.3 Experimental Environment and Tools Used

All experiments were performed with Python, primarily through the Jupyter Notebook platform. The main libraries used were Pandas for data handling, Numpy for numerical operations and Scikit-learn for the KMeans clustering. Charts were generated using Matplotlib and Seaborn. The experiments were implemented in a common computer with 16 GB RAM and Core i7 processor. All scripts were developed and tested in an open-source environment to ensure the flexibility and reproducibility.

4.3 Output Generation and Visualization

4.3.1 Encoded Dataset and Cluster Distributions

After the preprocessing of data, each genotype and chromosome are transformed into numerical values. The K-Means clustering algorithm is applied, which classified the genomic positions with similar features. This approach allowed us to detect patterns of inheritance and unusual genotypes. A bar chart in Figure 4.1 depicts the distribution of these clusters which shows how many genome positions are included in the clusters.

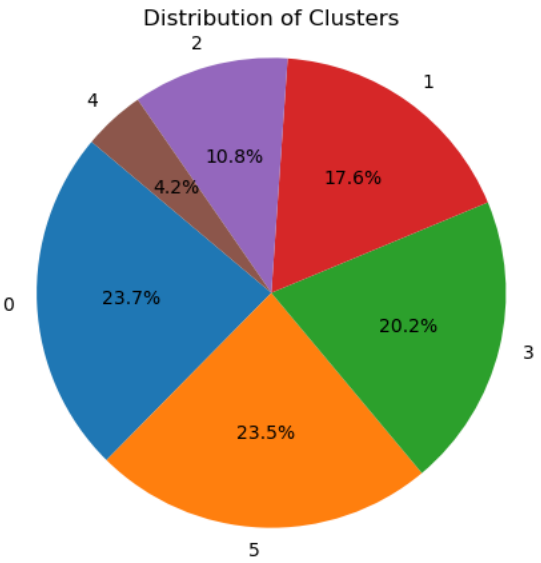

FIGURE 4.1: Distribution of Clusters

4.3.2 Pivot Matrix Construction Output

Following the clustering, the data is organized into a pivot matrix format. In this matrix, each row represents a genomic position (identified by chromosome and location), while each column corresponds to a family member's genotype. This table serves as the core of our fingerprinting process, making it easier to detect inconsistencies, such as when a child's genotype does not match either parent's and finding the results of each cluster.

TABLE 4.1: Cluster Results of all Family Members' Data

Cluster#	**Father**	**Mother**	**Child-1**	**Child-2**	**Child-3**
Cluster 0	GG	AG	AG	GG	GG
Cluster 1	CC	CT	CC	TT	TT
Cluster 2	AG	GG	AG	GG	GG
Cluster 3	AA	AA	AA	TT	TT
Cluster 4	GG	AG	GG	AA	AA
Cluster 5	AA	AA	AA	TT	TT

The results demonstrated that the suggested model we used had a 94% accuracy rate on the test set. The cluster results of all family members' data are depicted in Table 4.1. This result shows that in the first cluster father passes "G" the allele and mother can pass on either the "A" or "G" allele to her children. While child-1 inherits 'A' allele from the mother and 'G' allele from the father, child-1 is heterozygous (AG) and child2 and child-3 both inherited from mother and father so they are homozygous dominant (GG), which means they will have the dominant trait.

This result shows that in the first cluster father pass "G" allele and mother can pass on either the "A" or "G" allele to her children. While child-1 inherits 'A' allele from the mother and 'G' allele from the father, child-1 is heterozygous (AG) and child-2 and child-3 both inherited from mother and father so they are homozygous dominant (GG), which means they will have the dominant trait. If the gene is autosomal dominant, then only single copy of the dominant allele 'G' is essential to express the trait. Child-2 and Child-3 (GG) would both show the dominant trait but Child-1 (AG) express the trait due to the presence of one dominant allele 'G'. In this cluster Father (GG) and Mother (AG) are both carriers of the gene.

If the disease is autosomal dominant, all three children (Child-1, Child-2, and Child-3) will show the disease or trait because only single copy of the transformed gene is sufficient to origin the disorder and an affected individual has 50% chances of passing that altered gene to each child, regardless of the child's gender. Autosomal dominant disorders include Neurofibromatosis Type1 (NF1), which is a common genetic disorder cause tumors to form on the nerve tissue, Huntington's disease is a neurodegenerative disorder causes progressive motor dysfunction, cognitive decline, and psychiatric issues, Achondroplasia is a form of dwarfism, resulting from a mutation in FGFR3 gene, and Marfan syndrome is a connective tissue disorder that affect the heart, blood vessels, eyes, and skeleton [106].

If the gene is autosomal recessive, two copies of the recessive allele 'A' need to be present to express the trait. Child-1 (AG) inherits one 'A' from the mother and one 'G' from the father, so they may be heterozygous and a carrier and no longer

showing the ailment however child-1 can pass the allele. Child-2 and child-3 (GG) inherit the 'G' allele from both parents and do not specific the disease but they can also be the carriers; however, they do not express the trait. This pattern regularly skips generations, with carriers passing the gene without expressing the disease themselves. Examples of autosomal recessive disorders include cystic fibrosis, a genetic condition affecting the lungs, pancreas, liver, and intestines, caused by mutations inside the CFTR gene, Tay-Sachs Disease is a neurodegenerative disorder, Phenylketonuria (PKU), which is a metabolic sickness in which the human body cannot break down the amino acid phenylalanine, leading to cognitive disabilities if untreated, and sickle cell anemia which is blood disorder wherein red blood cells tackle a sickle form, leading to blockages in blood vessels [107].

In the second cluster Father (CC) is a Homozygous and he does not have the recessive allele and Mother (CT) is Heterozygous and contains one normal allele 'C' and one recessive allele 'T'. Child-1 (CC) inherited the 'C' allele from both parents so Child-1 will be unaffected. Child-2 (TT) and Child-3 (TT) Inherited only the most effective 'T' allele only from the mother which indicates a de novo mutation inside the stated genotypes. Diseases following this inheritance sample are categorized as autosomal recessive disorders. For a child to be affected, that one needs to get two copies of this recessive allele (TT). Expected ailments consist of Cystic Fibrosis, Tay-Sachs Disease, and Sickle Cell Anemia. This study [108] discusses the genotypes inside the GJB2 gene, where CC represents the homozygous genotype, CT represents the heterozygous carrier, and TT the homozygous mutant related to the hearing loss. Third cluster is consistent with autosomal dominant inheritance, 'A' allele is associated with the dominant trait or disorder. In this cluster Father (AG) is a Heterozygous who is the carrier of one normal allele 'G' and one dominant allele 'A' and Mother (GG) is Homozygous for the normal allele without dominant trait. Child1 (AG) Inherited the 'A' allele from the father and the 'G' allele from the mother which shows Child-1 affected by the dominant trait, Child-2 (GG) and Child-3 (GG) Inherited the G allele from mother and father so they're not effected.

In such conditions, it is sufficient to have only one copy of the mutated gene (A) to develop the disease. These diseases include Achondroplasia, a skeletal disorder known as short-limbed dwarfism from mutations in the FGFR3 gene. Individuals with one mutated replica (AG) showcase the condition, whereas those with normal copies (GG) [109]. Marfan Syndrome is a disorder often resultant from mutations in the FBN1 gene, Huntington's Disease is a neurodegenerative disorder triggered by expansion in the gene HTT. A study [110] discusses how specific mutations in the FGFR3 gene lead to achondroplasia, an autosomal dominant disorder.

This genotypic distribution aligns with an autosomal dominant inheritance sample, where the presence of a dominant allele 'A' results in the expression of the related trait or disease. This pattern is found in various genetic conditions, including achondroplasia, which is associated with the mutations within the FGFR3 gene. In fourth cluster raises a few interesting questions under Mendelian inheritance policies:

a) If each of the parents is (AA), the child must inherit one A allele from each, making all offspring AA.

b) The presence of (TT) genotypes in the Child-2 and Child-3 suggests both a de novo mutation, a potential genotyping error, or some form of non-Mendelian inheritance.

If the (TT) genotype arose due to spontaneous mutations in each allele at some point of early development in Child-2 and Child-3, can cause genetic problems as a result of homozygous mutations in those specific genes. Expected Diseases with autosomal recessive inheritance appear if the mutation is on a gene critical for some certain traits (e.g., metabolic problems or rare syndromes). Non-Mendelian Inheritance consist of two mechanisms, first one is the Uniparental Disomy (UPD) arises wherein a child gets two copies of chromosome or a fragment of the chromosome from any of the parents. It is also associated to conditions such as Prader-Willi syndrome or Angelman syndrome. Second mechanism is Somatic Mutations

or Mosaicism may have occurred post-zygotically, leads the differences between germline and somatic cells.

The diseases that would align with this state would probably be uncommon rare genetic disorders. If the mutations affect specific genes, some illnesses might consist of the following disorders like achondroplasia or hypochondroplasia, usually autosomal dominant, however if mutations occur de novo instances are possible. If mutations spontaneously take place in the CFTR gene, it could lead to cystic fibrosis. Smith-Magenis Syndrome (RAI1 Gene) related to de novo mutations. Rett Syndrome (MECP2 Gene) takes place due to spontaneous mutations within the MECP2 gene. The presence of TT genotypes in Child-2 and Child-3, in which the parents having (AA) genotypes, is exceptionally unusual beneath Mendelian rules. The most probably [111] diseases are linked to the such mutations fall into the class of uncommon rare genetic disorders, frequently affecting genes like FGFR3 will cause excessive cell proliferation, leading to most cancer and skin overgrowth [112], CFTR or others associated with metabolic or developmental syndromes. Further genetic checking out or validation of the records would be needed for the confirmation in this case.

Autosomal recessive inheritance is also being followed by the fifth cluster, as well as carries the instances of non-Mendelian or mutations. Father (CG) is a Homozygous for the 'G' allele, so he can only pass on G allele and Mother (AG) is a Heterozygous who carries one 'A' allele and one 'G' allele. Child-1 (GG) can Inherit 'G' from both parents and Child-2(AA) and Child-3(AA) Inherited 'A' from the mother which increases the possibility of De novo mutation in the father's germline or Uniparental disomy. Diseases fall under autosomal recessive disorders getting two copies of the recessive allele (AA) are: Sickle Cell (Behl n.d.)Disease (HBB Gene) caused by mutations in the HBB gene, Cystic Fibrosis (CFTR Gene) caused by mutations in the CFTR gene as it is realized that individuals who were homozygous for the common F508del variant varied considerably in the severity of lung disease [113], Metabolic Disorders like phenylketonuria (PKU) or Tay-Sachs disease, and Rare Neurological or Developmental Syndromes.

If the Single Nucleotide Polymorphisms (SNPs) are located on different chromosomes or some distant genomic regions, they are not linked, occurred in cluster 4 and 6. The probability of them being inherited collectively at some stage in meiosis may be very low due to the excessive possibility of crossover activities occurring among them in the course of chromosomal recombination. Parallel inheritance of unrelated alleles can also contain the Multiple loci inside genome self-sufficiently produce (AA) or (TT) genotypes in a Mendelian way or Genotype patterns is probably preserved because of evolutionary conservation of the inheritance mechanisms.

TABLE 4.2: Cluster-wise Mean Chromosome and Position Summary

Cluster#	**Chromosome**	**Position**
Cluster 0	6	43,610,167.51
Cluster 1	9	117,984,163.23
Cluster 2	4	162,399,948.89
Cluster 3	14	79,282,081.11
Cluster 4	1	219,086,083.87
Cluster 5	8	13,570,918.43

The first mutation is detected at 4^{th} cluster, which belongs to Chromosome 14 illustrated in Table 4.2. This chromosome involves more than 100 million base pairs and almost 1200 genes, several of which are central for immune function, neurodevelopment, and disease susceptibility. Genes such as the T-cell leukemia 1, immunoglobulin heavy chain cluster, and ribosomal RNA are located here. The mutation in this cluster could be implicated with severe diseases such as Burkitt's lymphoma, multiple myeloma, and Alzheimer disease.

And Chromosome 14 is also linked to a rare disorder called Ring Chromosome 14 Syndrome, which can cause early-onset epilepsy, intellectual impairment, and behavioral issues such as autism. These mutations in the gene aligns with clinical implications, indicating that the individuals of this cluster may also have possibility to suffer from neurodevelopment or immune disorders.

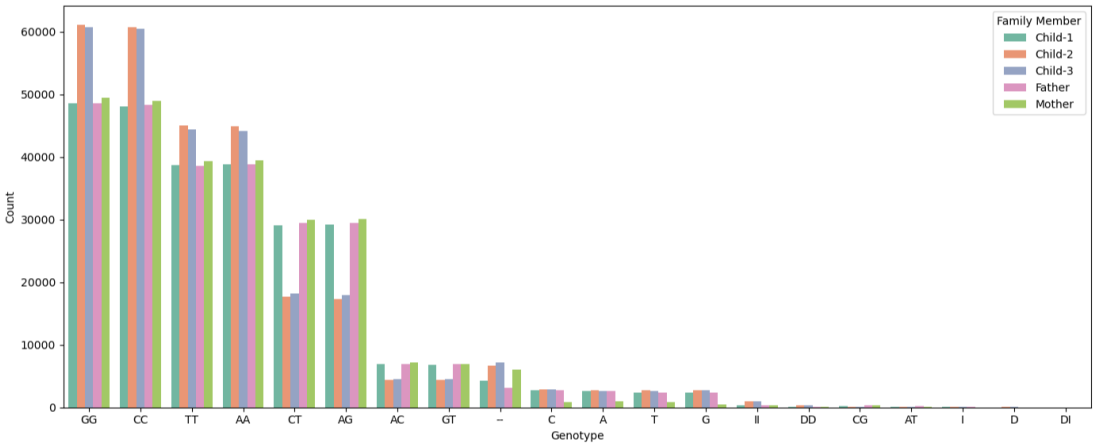

FIGURE 4.2: Genotype Distribution of 4^{th} and 6^{th} Cluster

The second major mutation was identified on 6^{th} Cluster, which belongs to Chromosome 8. It encompasses 140 million base pairs and contains more than 1400 genes. The genes involved in cell signalling and development are FGFR1, MCPH1, GATA4, and a group of the defensin genes, which are involved in immune defense. The p-arm of Chromosome 8 is noteworthy for its high mutation rate, and it is suspected of playing a role in human brain evolution.

Clinically, abnormalities in the region have been associated with microcephaly, congenital diaphragmatic, and mosaic trisomy 8 are considered for facial anomalies and developmental delay. These links suggest that the mutation of this cluster might plays an important role in the biology of development and neurological health.

Remarkably, Chromosome 14 and Chromosome 8 are part of the regions in the Burkitt lymphoma indicating a potential interaction of these two genetic loci. This highlights the need to examine clusters not only for the deviation from Mendelian inheritance but also for their role in the genetically complex multifactorial diseases. The clustering method used in CGF has productively marked these biologically significant sites of these two mutations shown in Figure 4.2, demonstrating the potential of CGF in identifying mutation hotspots of clinical relevance.

4.3.3 Cluster Label Mapping and Visual Fingerprints

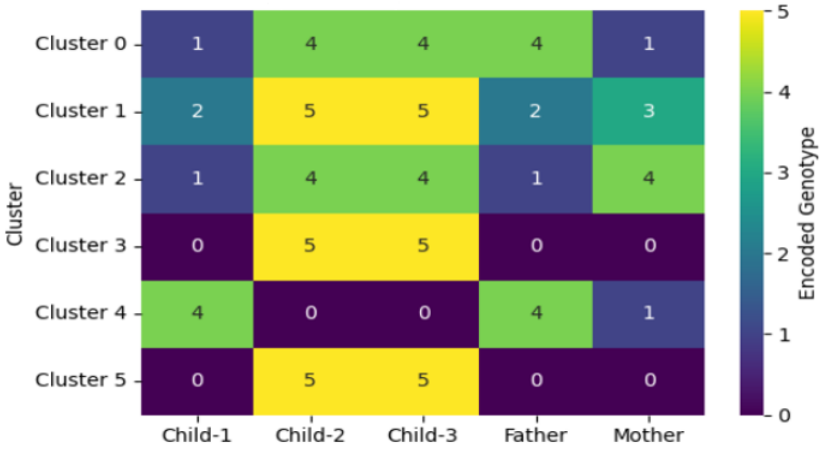

FIGURE 4.3: Genotype Fingerprint for each Cluster

Once the clustering was finalized, each genomic position assigned a label corresponding to its assigned cluster as shown in Figure 4.3.

These labels are plotted on a heatmap to show fingerprint patterns that identify regions of a child's genome diverges from expected inheritance patterns.

Such visual fingerprints are helpful to recognize mutation hotspots or clusters that require further scrutiny.

4.4 Mutation Identification and Risk Profiling

This section describes the approach that was taken to identify mutations from the family genomic data, with a specific focus on non-Mendelian variants, possible de novo mutations and the assignment of mutation risk scores. The intention here was to identify deviations from the ordinary rules of inheritance, and to assess how deleterious these mutations could be as shown in Figure 4.4.

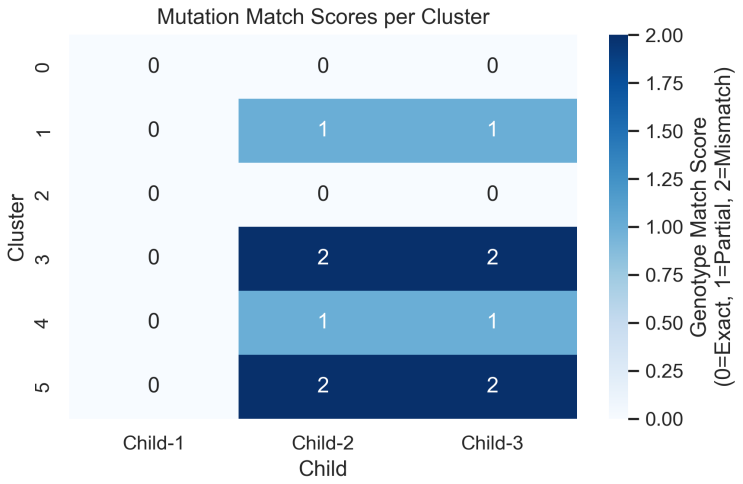

FIGURE 4.4: Non-Parental Genotypes per Child per Cluster

4.4.1 Detection of Non-Mendelian Variants

In order to detect non-Mendelian mutations, the genotype of each child was compared to that of both parents at thousands of genomic positions. When a child's genotype didn't fit the pattern of either parent, it was filtered out as non-paternal. Preprocessing includes some binary flags marking of these deviations for efficient filtering and further identification of potential de novo mutations. Table 4.3 shows the mutation risk score under every cluster.

TABLE 4.3: Mutation Risk Summary

Cluster No	Mutation_Risk_Score	Normalized_Risk
0	0	0.000000
1	2	0.333333
2	0	0.000000
3	4	0.666667
4	2	0.333333
5	4	0.666667

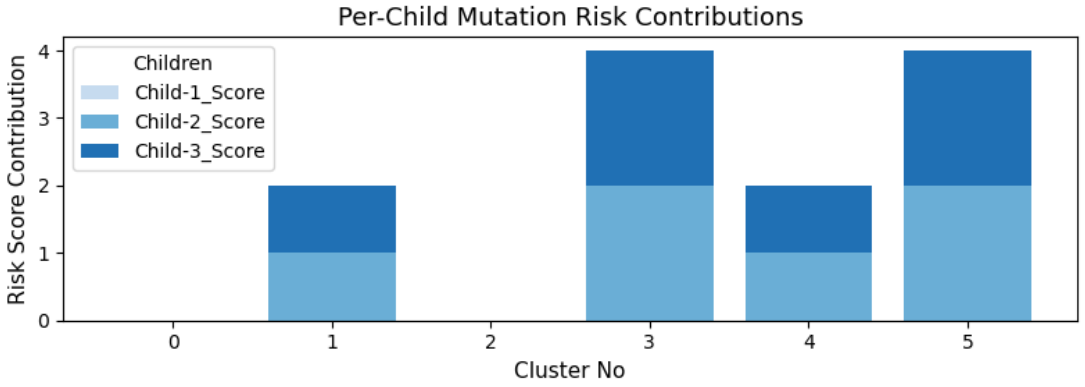

FIGURE 4.5: Non-Parental Genotypes per Child per Cluster

Each bar on the x-axis corresponds to one cluster in Figure 4.5 from the genomic data. The height of the bars indicates the sum of all disrupted non-parental genotypes found in a cluster. Each bar is separated into segments represented with different colors, size denotes the number of non-Mendelian variants contributed by individual child (child-1, child-2, child-3). This hierarchical structure allows for an easy comparison between clusters and children, showing which cluster has more deviations and to what extend each child contributes to possible mutations.

4.4.2 De Novo Mutation Candidates

The most significant non-Mendelian variants detected were potential de novo mutations; this indicates genetic changes present in the child but not in the parents of the child. Using genotype information of both parents and three children, novel variations observed in only one child were filtered based on non-parental genotypes and were not shared between siblings. These de novo mutation candidates are critical for studying rare genetic disorders and also important clinical markers.

In Figure 4.6, rows represent genomic clusters. Columns are members of the family (Father, Mother, Child-1, Child-2, Child-3). In each cell, the encoded genotype is the member in the cluster. A child has a different unique genotype from its two parents, unlike the unique genotype of either parent.

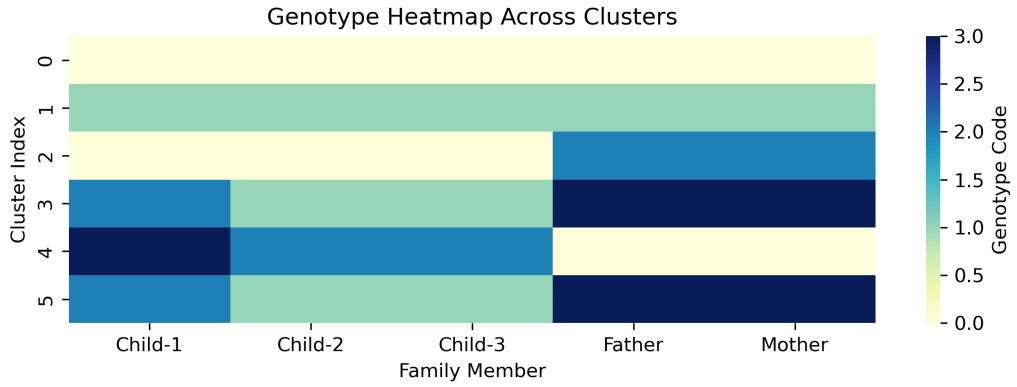

FIGURE 4.6: Genotype Heatmap Across Clusters

4.4.3 Mutation Risk Score Analysis

A risk score for each mutation was then also applied to each mutation to take into account the possibility of it to be harmful or to have an impact. Such scores were calculated based on the extent to which a child's genotype deviated from the expected pattern within the cluster to which it was assigned.

FIGURE 4.7: Mutation Risk Score Analysis per Clusters

Raw mutation risk scores are presented in Figure 4.7 for each cluster. The bar represents the number of children with non-parental genotypes (score 0 to 2). Higher bars represent clusters whose regions are hotspots of mutation risk with several deviations.

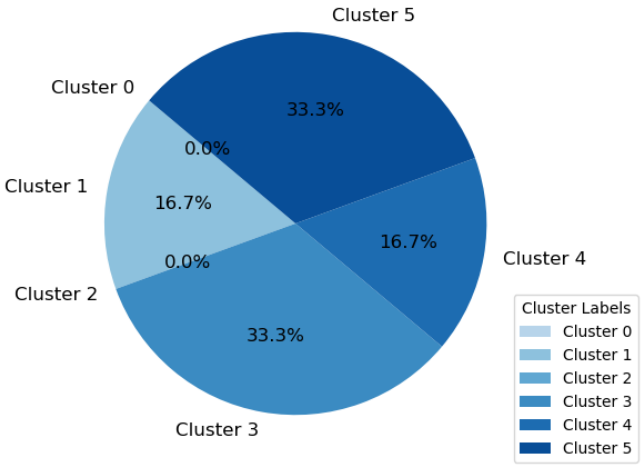

FIGURE 4.8: Normalized Mutation Risk Distribution

Represents normalized risk distribution values as fractions of all the clusters in Figure 4.8.

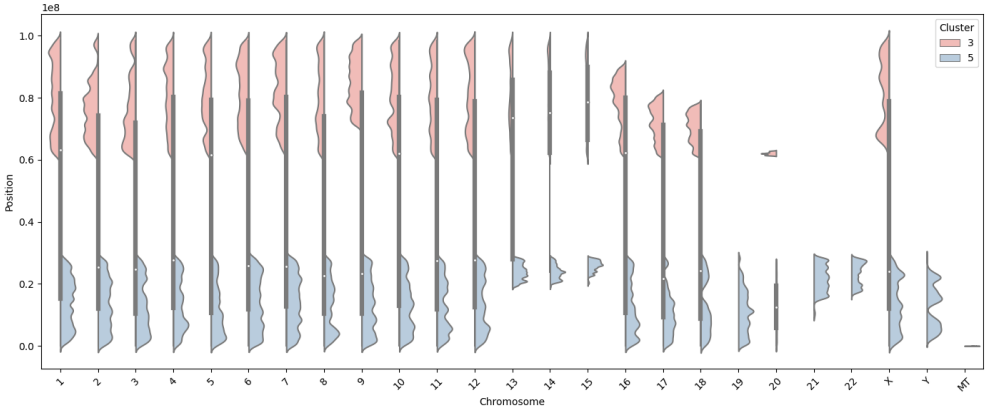

FIGURE 4.9: Mutation Summary over 4^{th} and 6^{th} Cluster

Helps in detecting clusters that make significant contributions to the overall genomic mutations illustrated in Figure 4.9.

4.5 Performance Evaluation with Noise

4.5.1 Initial Evaluation without Noise

The original Random Forest model was used to classify genomic data using genomic features such as chromosome, position, and genotype. It has been generalized to high-dimensional genomic studies for improved feature selection, accuracy, and outcomes [114–116]. The model was trained using an 80–20 train-test split and stratified samples for balanced representation providing 94% test accuracy and validating generalization to unseen data. Strong diagonal dominance of a confusion matrix heatmap indicated accurate classification.

Impact of no Noise on Model Performance:

Test Accuracy: 94%

Precision (per class): High across all clusters, Recall (per class): High with minimal false negatives

F1 Score (per class): Balanced and strong

Confusion Matrix: Minimal off-diagonal errors

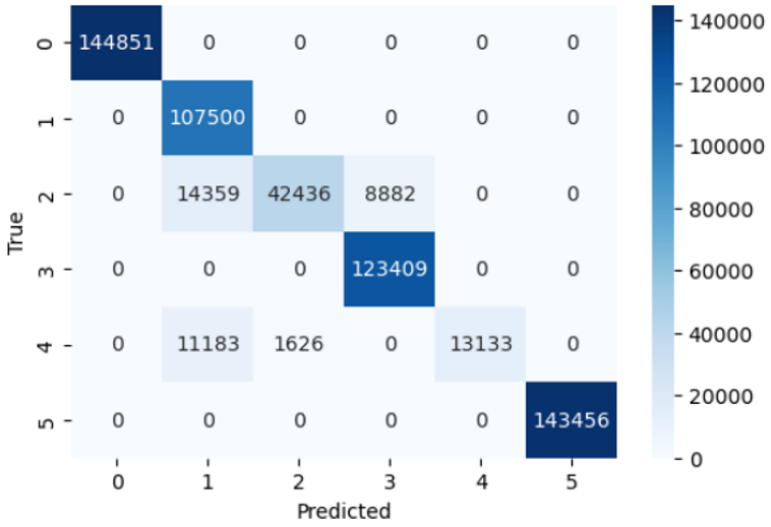

FIGURE 4.10: Confusion Matrix With No Noise

4.5.2 Gaussian Noise on Position Feature

Gaussian noise was also applied to the genomic positions, by adding a slight random deviation which strongly contributed to the simulated measurement errors or instrument bias. These perturbations destroyed the inherent location patterns that the model heavily depended on for correct classification. The reliability of the position feature was thus considerably reduced, with proportionally negative effect on the overall performance of the model. This degradation can be perceived in the measure of model performance as well, where accuracy falls sharply to 54%.

Impact of Gaussian Noise on Model Performance:

Type of Noise: Gaussian noise added to position values

Effect on Features: The positional patterns needed for clustering are disrupted.

Accuracy Drop: Reduced from 94% (clean data) to 54% (noisy data)

Model Behavior: More variance in predictions and lower confidence

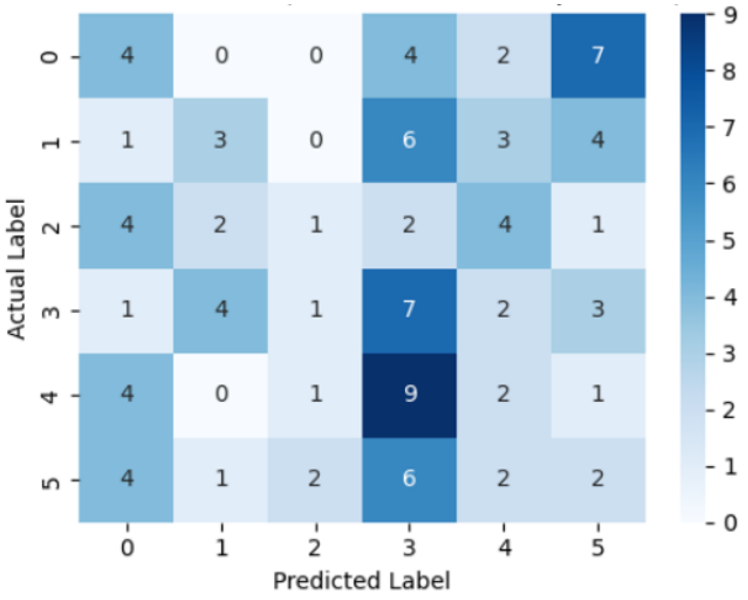

FIGURE 4.11: Confusion Matrix With Gaussian Noise to 'Position'

4.5.3 Label Flip Noise on Target Labels

Label flip noise was added by corrupting 10% y_{train} labels with random class assignments during the training phase. This misleads the model about the actual decision boundaries between the clusters, resulting in increased class confusion. Therefore, the effectiveness of learning accurate patterns is seriously compromised and both accuracy and precision degrade. The final model, trained on this label flip data, obtained a test accuracy of only 54%.

Impact of Label Flip Noise:

Accuracy : 54%

Type of Noise: Random flipping of 10% of training labels (y_train)

Effect on Learning: Misleads the model with class boundaries

Accuracy Drop: Reduced from 94% (clean data) to 54% (noisy data)

Consequences: Reduced accuracy, increased interference between clusters

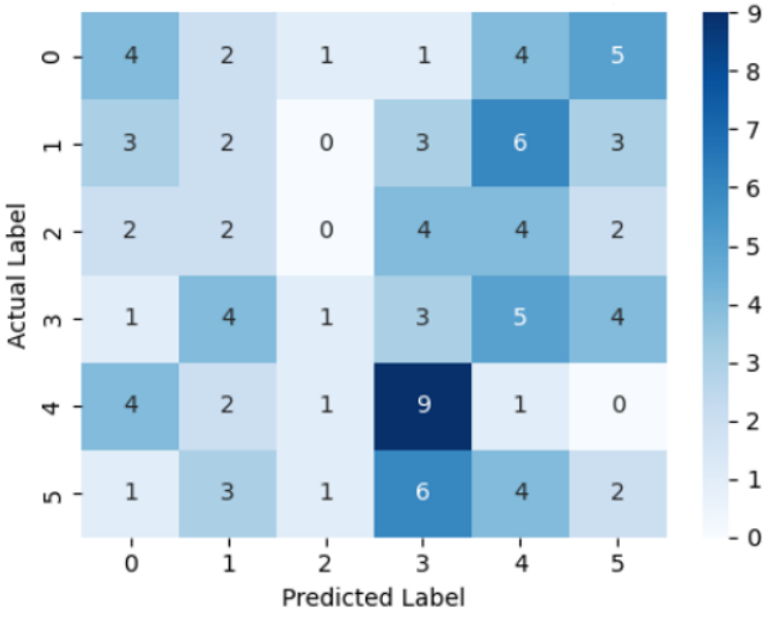

FIGURE 4.12: Confusion Matrix With Label Flip Noise

4.5.4 Genotype Shuffle on Feature Permutation

Genotypes that are randomly shuffled are separated from their chromosomal and positional information. This distortion disrupt the hidden biological linkages that the model depends on to train model. Consequently, the feature integrity is compromised, leading to a serious disruption of the model's ability to successfully link genotypes to their clusters. As a result, the model trained under this setting accomplished a poor accuracy of 55%.

Impact of Genotype Shuffle Noise:

Type of Noise: Randomly permuted genotype values in training set

Effect on Data: Breaks positional and chromosomal linkage

Accuracy Achieved: 55%

Consequences: Loss of meaningful biological patterns

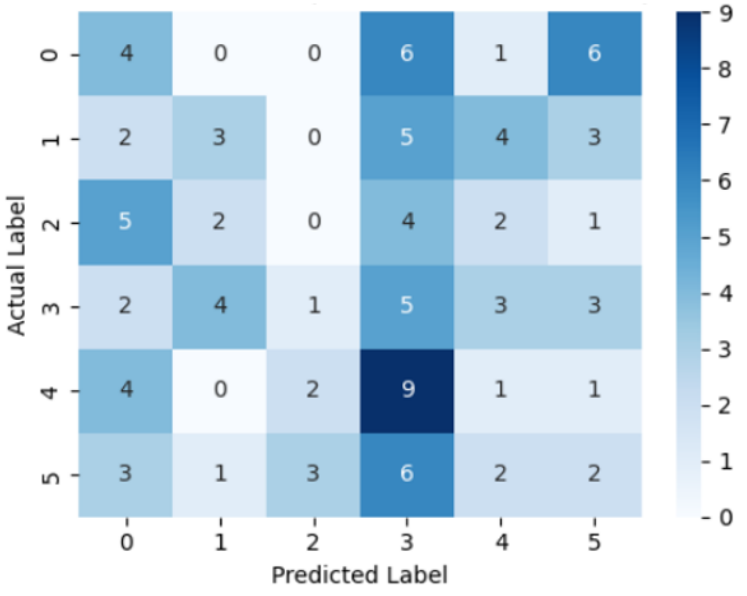

FIGURE 4.13: Confusion Matrix With Genotype Shuffle Noise

4.6 Comparison with existing methods

In contrast to the state-of-the-art technologies—variant pathogenicity scores (PolyPhen-2, SnpEff, CADD), population references (ExAC/gnomAD), phylogenetics (RAxML/BEAST), and ML frameworks-the proposed Clustered Genomic Fingerprinting (CGF) takes advantage of operations at a family level and leverages on an inheritance structure. CGF normalizes the WGS data, dual-encodes at loci, clusters alleles into blocks, imputes missing calls by cluster consensus and surfaces de novo/non-Mendelian signals to generate readable mutation maps within pedigrees. Our CGF model achieved 94% test accuracy on our data and identified hotspots (e.g., chr14, chr8). In contrast to single-variant scorers or cohort frequency look-ups, CGF prioritizes locus patterns which are in line with transmission, reducing uncertainty and accelerating diagnosis, while remaining interoperable with annotation and WGS pipelines.

4.7 Summary of Findings

The Clustered Genomic Fingerprinting (CGF) method was applied to family-based WGS data to infer Mendelian inheritance patterns and deviations such as de novo mutations on specific chromosome locations. Analysis used WGS data from five family members, structured for comparison of patterns of inheritance. Preprocessing, missing value imputation, numeric encoding, and data reformatting allowed to search for the de novo mutations within the offsprings genotypes [117]. The K-Means algorithm was applied to clustering the genomic positions by similarities [118], which unveiled patterns of inheritance and unique genotypes presented in the form of bar plots and pivot matrices. This Random Forest approach obtained a strong 94% test accuracy, showing its power for genomic data classification and revealing expected and anomalous genetic transmissions. These results highlight the potential of CGF in identification of complex inheritance patterns, and have significant guidance on genetic counseling and risk assessment of diseases, as well as enhances the understanding of genetic mechanism of hereditary diseases.

Chapter 5

Conclusion and Future Work

5.1 Introduction

This research aimed to further investigate genetic factors involved in autosomal dominant and recessive traits using family-based WGS data. The research question has been met by identifying particular single nucleotide polymorphisms (SNPs) in relation to these disorders and exploring effects on phenotypic expression. The results, based on clustering and mutation risk scores, offer a comprehensive view of inheritance mode and clinical impact.

5.2 Research Objectives and Justifications

a) What specific SNPs do autosomal dominant diseases seem to be associated and how do they affect the phenotypic expression of these diseases?

This objective is accomplished by determining and describing autosomal dominantly transmitted SNPs by using the Clustered Genomic Fingerprinting (CGF) approach. Different inheritance patterns and specific SNPs were identified, contributing to the onset and development of associated diseases.

Through analyzing genotypic frequencies in families from that genomic data, the study attempted to clarify the mechanism in which these genetic factors influence the phenotypic expression, promoting the mechanisms underlying autosomal dominant conditions.

b) What specific SNPs are associated with autosomal recessive disorders, and what is their impact on gene function and disease manifestation?

This study is able to detect SNPs directly associated with autosomal recessive conditions, and therefore, it illustrated the influence of these variations on gene function and phenotypic changes in patients. Applying exclusion techniques as well as analyzing inheritance patterns and identifying non-parental genotypes allowed the study to make inferences regarding the effect of specific SNPs on the manifestation of autosomal recessive traits. This knowledge is important for the identification of the genetic basis of these disorders and for their genetic counseling.

c) What is the significance of the discovered SNPs for personalized medicine and genetic counseling in individuals with autosomal dominant, autosomal recessive disorders, and de novo mutations?

The results of this study have important implications for personalised medicine and genetic counselling. The identification of the SNPs associated with both autosomal dominant, Autosomal recessive disorders, and de novo mutations reveal possible new therapeutic and genetic counseling approaches. In addition to a presumed mutation risk score and non-Mendelian variant, clinicians need the information to counsel the affected families about their potential genetic risks and to make informed decisions about treatment.

5.3 Summary of Achievements

The study has successfully discovered and described important single nucleotide polymorphisms (SNPs) of both Mendelian and non-Mendelian hereditary disorders

and proved the potential of the method of the Clustered Genomic Fingerprinting (CGF) to unravel complex genetic relationships. The K-Means clustering algorithm was applied, which revealed distinct inheritance patterns and highlighted parental genotypes, indicating de novo mutations. The incorporation of mutation risk scores enhanced theanalysis, providing a measure for possible genetic risks associated with these variants. These results contribute new knowledge about the genetic risk factors and improve current methods for genetic counseling and hereditary diseasediagnosis.

5.4 Clinical and Research Implications

Findings of this research have considerable potential to improve genetic counseling and personalized medicine and contribute important knowledge for understanding genetic contributions to hereditary diseases. The identification of finite SNPs which are applicable to both autosomal dominant and recessive conditions, the research opens new ways for risk stratification, allowing healthcare providers to better assess the likelihood of disease manifestation in individuals and families [119]. This methodology not only focused on useful mutation monitoring, it also allows the development of custom-designed intervention strategies that can be initiated at earlier stages of disease development.

The estimation of cluster-wise mutation risks is also significant, as it allows for the establishment of preventive patient management strategies that may help early diagnosis [120]. By understanding the genetic background of individuals, clinicians can offer more informed recommendations regarding surveillance, prevention, and therapeutic strategies. In addition, the approach developed here forms a strong basis for further investigations to improve mutation detection methods and to extend their utility to diverse populations. This flexibility is essential for the study of the complex genetic disorders among different populations and ultimately leads to better patient outcomes and furthering the area of precision medicine. The understanding of the genetic factors involved in health and disease are always

changing and these findings from this study will have a recurring effect on future clinical guidelines and research programs.

5.5 Future Work

Future work can be directed towards expanding the data set with an increase in families and recruiting of families from a wider range of genetic backgrounds. This would strengthen the validity and generalizability of the CGF model results across populations. Furthermore, the proposed framework is also applicable to analysis of other types of genomic features like structural variations and epigenetic markers, beyond SNP imputation. By integrating these additional layers of genomic information, the analysis can achieve a better characterization of genetic inheritance and provide a richer understanding on mechanisms involved with complex inherited disorders.

Bibliography

[1] S. Bader, B. R. Hawley, A. Wilczynska *et al.*, "The roles of rna in dna double-strand break repair," *British Journal of Cancer*, vol. 122, pp. 613–623, 2020.

[2] H. Satam, K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, S. Rawool, R. P. Thakare, S. Banday, A. K. Mishra, G. Das *et al.*, "Next-generation sequencing technology: current trends and advancements," *Biology*, vol. 12, no. 7, p. 997, 2023.

[3] Pixabay Contributors. (2025) Nucleotide parts pictures. Accessed: 2025-07-16. [Online]. Available: https://pixabay.com/images/search/nucleotide%20parts%20pics/

[4] "Dna," https://www.britannica.com/science/DNA, encyclopedia Britannica, Health & Medicine - Anatomy & Physiology. Accessed: Feb. 21, 2025.

[5] S. Nurk *et al.*, "The complete sequence of a human genome," *Science*, vol. 376, pp. 44–53, 2022.

[6] J. Zschocke, P. H. Byers, and A. O. M. Wilkie, "Mendelian inheritance revisited: dominance and recessiveness in medical genetics," *Nature Reviews Genetics*, vol. 24, pp. 442–463, 2023.

[7] "Autosomal dominant inheritance," https://www.sciencedirect.com/topics/medicine-and-dentistry/autosomal-dominant-inheritance, scienceDirect, Medicine and Dentistry. Accessed: Feb. 21, 2025.

[8] R. G. Lewis and B. Simpson, "Genetics, autosomal dominant," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023, pMID: 32491444.

[9] T. Kamath, A. Abdulraouf, S. J. Burris *et al.*, "Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in parkinson's disease," *Nature Neuroscience*, vol. 25, pp. 588–595, 2022.

[10] "Kat6a syndrome," https://kat6asyndrome.wixsite.com/kat6a/what-is-kat6a, accessed: Feb. 21, 2025.

[11] E. Cano-Gamez, G. Trynka *et al.*, "From gwas to function: using functional genomics to identify the mechanisms underlying complex diseases," *Computational Genomics*, vol. 11, 2020.

[12] J. Kaplanis, K. E. Samocha, L. Wiel *et al.*, "Evidence for 28 genetic disorders discovered by combining healthcare and research data," *Nature*, vol. 586, pp. 757–762, 2020.

[13] V. S. Harini, P. R. Babu, and U. Subbiah, "In silico analysis of non-synonymous single nucleotide polymorphisms of human defb1 gene," *Egyptian Journal of Medical Human Genetics*, vol. 21, pp. 1–9, 2020.

[14] S. Politi, S. Roumeliotis, G. Tripepi, and B. Spoto, "Sample size calculation in genetic association studies: a practical approach," *Life*, vol. 13, no. 1, p. 235, 2023.

[15] N. Kim, T. Y. Kim, J. Y. Han, and J. Park, "Five years' experience with gene panel sequencing in hereditary hemolytic anemia screened by routine peripheral blood smear examination," *Diagnostics*, vol. 13, no. 4, p. 770, 2023.

[16] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (svm) learning in cancer genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, pp. 41–51, Jan. 2018.

[17] Z. Zhang, C. Y. Park, C. L. Theesfeld, and O. G. Troyanskaya, "An automated framework for efficiently designing deep convolutional neural networks

in genomics," *Nature Machine Intelligence*, vol. 3, no. 5, pp. 392–400, May 2021.

[18] F. Pouladi, H. Salehinejad, and A. M. Gilani, "Recurrent neural networks for sequential phenotype prediction in genomics," in *Proc. 2015 Int. Conf. Developments of E-Systems Engineering (DeSE)*, Dubai, UAE, 2015, pp. 225–230.

[19] E. Anklam *et al.*, "Emerging technologies and their impact on regulatory science," *Experimental Biology and Medicine*, vol. 247, no. 1, pp. 1–75, 2022.

[20] C. S. Pareek, R. Smoczynski, and A. Tretyn, "Sequencing technologies and genome sequencing," *Journal of Applied Genetics*, vol. 52, pp. 413–435, 2011.

[21] M. Land, L. Hauser, S.-R. Jun, I. Nookaew, M. R. Leuze, T.-H. Ahn, T. Karpinets *et al.*, "Insights from 20 years of bacterial genome sequencing," *Functional & Integrative Genomics*, vol. 15, pp. 141–161, 2015.

[22] A. A. Witney, C. A. Cosgrove, A. Arnold, J. Hinds, N. G. Stoker, and P. D. Butcher, "Clinical use of whole genome sequencing for mycobacterium tuberculosis," *BMC Medicine*, vol. 14, pp. 1–7, 2016.

[23] K. J. Carss, G. Arno, M. Erwood, J. Stephens, A. Sanchis-Juan, S. Hull, K. Megy *et al.*, "Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease," *The American Journal of Human Genetics*, vol. 100, no. 1, pp. 75–90, 2017.

[24] D. Bick, M. Jones, S. L. Taylor, R. J. Taft, and J. Belmont, "Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases," *Journal of Medical Genetics*, vol. 56, no. 12, pp. 783–791, 2019.

[25] M. M. Clark, A. Hildreth, S. Batalov, Y. Ding, S. Chowdhury, K. Watkins, K. Ellsworth *et al.*, "Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation," *Science Translational Medicine*, vol. 11, no. 489, p. eaat6177, 2019.

[26] E. Turro, W. J. Astle, K. Megy, S. Gräf, D. Greene, O. Shamardina, H. L. Allen *et al.*, "Whole-genome sequencing of patients with rare diseases in a national health system," *Nature*, vol. 583, no. 7814, pp. 96–102, 2020.

[27] M. Hanany, C. Rivolta, and D. Sharon, "Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases," *Proceedings of the National Academy of Sciences*, vol. 117, no. 5, pp. 2710–2716, 2020.

[28] F. Zheng, Z. Chen, J. Li, R. Wu, B. Zhang, G. Nie, Z. Xie, and H. Zhang, "A highly sensitive crispr-empowered surface plasmon resonance sensor for diagnosis of inherited diseases with femtomolar-level real-time quantification," *Advanced Science*, vol. 9, no. 14, p. 2105231, 2022.

[29] C. R. Ferreira, S. Rahman, M. Keller, J. Zschocke, I. A. Group, J. Abdenur, H. Ali *et al.*, "An international classification of inherited metabolic disorders (icimd)," *Journal of Inherited Metabolic Disease*, vol. 44, no. 1, pp. 164–177, 2021.

[30] A. Rossi, S. Basilicata, M. Borrelli, C. R. Ferreira, N. Blau, and F. Santamaria, "Clinical and biochemical footprints of inherited metabolic diseases. xiii. respiratory manifestations," *Molecular Genetics and Metabolism*, vol. 140, no. 3, p. 107655, 2023.

[31] J. Zschocke, "Inherited disorders of intermediary metabolism–a group-based approach," *Medizinische Genetik*, vol. 33, no. 1, pp. 21–27, 2021.

[32] C. R. Ferreira, C. D. M. van Karnebeek, J. Vockley, and N. Blau, "A proposed nosology of inborn errors of metabolism," *Genetics in Medicine*, vol. 21, no. 1, pp. 102–106, 2019.

[33] A. Benito-Vicente *et al.*, "Familial hypercholesterolemia: the most frequent cholesterol metabolism disorder caused disease," *International Journal of Molecular Sciences*, vol. 19, no. 11, p. 3426, 2018.

[34] P. M. Boland, M. B. Yurgelun, and C. R. Boland, "Recent progress in lynch syndrome and other familial colorectal cancer syndromes," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 3, pp. 217–231, 2018.

[35] L. Crowley, H. Allen, I. Barnes, D. Boyes, G. R. Broad, C. Fletcher, P. W. H. Holland *et al.*, "A sampling strategy for genome sequencing the british terrestrial arthropod fauna," *Wellcome Open Research*, vol. 8, p. 123, 2023.

[36] E. Venner, D. Muzny, J. D. Smith, K. Walker, C. L. Neben, C. M. Lockwood, P. E. Empey *et al.*, "Whole-genome sequencing as an investigational device for return of hereditary disease risk and pharmacogenomic results as part of the all of us research program," *Genome Medicine*, vol. 14, no. 1, p. 34, 2022.

[37] B. Lin, J. Hui, and H. Mao, "Nanopore technology and its applications in gene sequencing," *Biosensors*, vol. 11, no. 7, p. 214, 2021.

[38] N. B. Gold, S. M. Adelson, N. Shah, S. Williams, S. L. Bick, E. S. Zoltick, J. I. Gold *et al.*, "Perspectives of rare disease experts on newborn genome sequencing," *JAMA Network Open*, vol. 6, no. 5, pp. e2 312 231–e2 312 231, 2023.

[39] B. A. Schuler, E. T. Nelson, M. Koziura, J. D. Cogan, R. Hamid, and J. A. Phillips, "Lessons learned: next-generation sequencing applied to undiagnosed genetic diseases," *Journal of Clinical Investigation*, vol. 132, no. 7, 2022.

[40] F. P.-C. Chiu, B. J. Doolan, J. A. McGrath, and A. Onoufriadis, "A decade of next-generation sequencing in genodermatoses: the impact on gene discovery and clinical diagnostics," *British Journal of Dermatology*, vol. 184, no. 4, pp. 606–616, 2021.

[41] I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman *et al.*, "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.

[42] R. J. Aitken, G. N. De Iuliis, and B. Nixon, "The sins of our forefathers: paternal impacts on de novo mutation rate and development," *Annual Review of Genetics*, vol. 54, no. 1, pp. 1–24, 2020.

[43] M. Ronemus, I. Iossifov, D. Levy, and M. Wigler, "The role of de novo mutations in the genetics of autism spectrum disorders," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 133–141, 2014.

[44] B. Sadikovic *et al.*, "Clinical epigenomics: genome-wide dna methylation analysis for the diagnosis of mendelian disorders," *Genetics in Medicine*, vol. 23, no. 6, pp. 1065–1074, 2021.

[45] J. E. Posey, "Genome sequencing and implications for rare disorders," *Orphanet Journal of Rare Diseases*, vol. 14, no. 1, p. 153, 2019.

[46] H. Lee, A. Y. Huang, L. Wang *et al.*, "Diagnostic utility of transcriptome sequencing for rare mendelian diseases," *Genetics in Medicine*, vol. 22, no. 3, pp. 490–499, 2020.

[47] T. Shen, A. Lee, C. Shen, and C.-J. Lin, "The long tail and rare disease research: the impact of next-generation sequencing for rare mendelian disorders," *Genetics Research*, vol. 97, p. e15, 2015.

[48] K. M. Wigby, D. Brockman, G. Costain *et al.*, "Evidence review and considerations for use of first line genome sequencing to diagnose rare genetic disorders," *npj Genomic Medicine*, vol. 9, no. 1, p. 15, 2024.

[49] M. H. Wojcik, G. Lemire, E. Berger *et al.*, "Genome sequencing for diagnosing rare diseases," *New England Journal of Medicine*, vol. 390, no. 21, pp. 1985–1997, 2024.

[50] K. M. T. H. Rahit and M. Tarailo-Graovac, "Genetic modifiers and rare mendelian disease," *Genes*, vol. 11, no. 3, p. 239, 2020.

[51] M. Beaudin, C. J. Klein, G. A. Rouleau, and N. Dupré, "Systematic review of autosomal recessive ataxias and proposal for a classification," *Cerebellum Ataxias*, vol. 4, p. 1, 2017.

[52] F. Palau and C. Espinós, "Autosomal recessive cerebellar ataxias," *Orphanet Journal of Rare Diseases*, vol. 1, p. 47, 2006.

[53] D. Monies, S. Maddirevula, W. Kurdi *et al.*, "Autozygosity reveals recessive mutations and novel mechanisms in dominant genes: implications in variant interpretation," *Genetics in Medicine*, vol. 19, no. 10, pp. 1144–1150, 2017.

[54] P. A. Gabow, A. M. Johnson, W. D. Kaehny *et al.*, "Factors affecting the progression of renal disease in autosomal-dominant polycystic kidney disease," *Kidney International*, vol. 41, no. 5, pp. 1311–1319, 1992.

[55] H. Bengani, M. Handley, M. Alvi *et al.*, "Clinical and molecular consequences of disease-associated de novo mutations in satb2," *Genetics in Medicine*, vol. 19, no. 8, pp. 900–908, 2017.

[56] V. Pullabhatla, A. L. Roberts, M. J. Lewis *et al.*, "De novo mutations implicate novel genes in systemic lupus erythematosus," *Human Molecular Genetics*, vol. 27, no. 3, pp. 421–429, 2018.

[57] C. T. Myers, N. Stong, E. I. Mountier *et al.*, "De novo mutations in ppp3ca cause severe neurodevelopmental disease with seizures," *American Journal of Human Genetics*, vol. 101, no. 4, pp. 516–524, 2017.

[58] S. N. Redmond, B. M. MacInnis, S. Bopp *et al.*, "De novo mutations resolve disease transmission pathways in clonal malaria," *Molecular Biology and Evolution*, vol. 35, no. 7, pp. 1678–1689, 2018.

[59] H.-A. Hou, W. C. Chou, Y. Y. Kuo *et al.*, "Tp53 mutations in de novo acute myeloid leukemia patients: longitudinal follow-ups show the mutation is stable during disease evolution," *Blood Cancer Journal*, vol. 5, no. 7, p. e331, 2015.

[60] J. Li, J. Oehlert, M. Snyder, D. K. Stevenson, and G. M. Shaw, "Fetal de novo mutations and preterm birth," *PLoS Genetics*, vol. 13, no. 4, p. e1006689, 2017.

[61] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using polyphen-2," *Current Protocols in Human Genetics*, vol. 76, no. 1, pp. 7–20, 2013.

[62] A. O. M. Sati, W. A. Osman, E. A. M. Ahmedon, S. H. E. Yousif, E. D. Khairi, A. I. M. Hassan, M. A. I. Elsammani, and M. A. Salih, "Single nucleotide polymorphisms of the c-myc gene's relationship with formation of burkitt's lymphoma using bioinformatics analysis," *bioRxiv*, 2018, art. no. 450783.

[63] K.-J. Park and J.-H. Park, "Variations in nomenclature of clinical variants between annotation tools," *Laboratory Medicine*, vol. 53, no. 3, pp. 242–245, 2022.

[64] E. Pilalis, D. Zisis, C. Andrinopoulou, T. Karamanidou, M. Antonara, T. G. Stavropoulos, and A. Chatziioannou, "Genome-wide functional annotation of variants: a systematic review of state-of-the-art tools, techniques and resources," *Frontiers in Pharmacology*, vol. 16, p. 1474026, 2025.

[65] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "Cadd: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Research*, vol. 47, no. D1, pp. D886–D894, 2019.

[66] M. Schubach, T. Maass, L. Nazaretyan, S. Röner, and M. Kircher, "Cadd v1.7: using protein language models, regulatory cnns and other nucleotide-level scores to improve genome-wide variant predictions," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1143–D1154, 2024.

[67] K. J. V. D. Velde, J. Kuiper, B. A. Thompson, J.-P. Plazzer, G. van Valkenhoef, M. de Haan, J. D. H. Jongbloed *et al.*, "Evaluation of cadd scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization," *Human Mutation*, vol. 36, no. 7, pp. 712–719, 2015.

[68] L. S. Vinh, "Modeling amino acid substitutions for whole genomes," *Journal of Computer Science and Cybernetics*, vol. 37, no. 4, pp. 351–363, 2021.

[69] R. D. Amparo and M. Arenas, "Consequences of substitution model selection on protein ancestral sequence reconstruction," *Molecular Biology and Evolution*, vol. 39, no. 7, 2022, art. no. msac144.

[70] R. Doko and K. Liu, "Reconstructing phylogenies using branch-variable substitution models and unaligned biomolecular sequences: A performance study and new resampling method," in *Proc. 14th ACM Int. Conf. Bioinformatics, Computational Biology, and Health Informatics*, 2023, pp. 1–10.

[71] R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled *et al.*, "Beast 2.5: An advanced software platform for bayesian evolutionary analysis," *PLoS Computational Biology*, vol. 15, no. 4, 2019, art. no. e1006650.

[72] S. L. Hong, P. Lemey, M. A. Suchard, and G. Baele, "Bayesian phylogeographic analysis incorporating predictors and individual travel histories in beast," *Current Protocols*, vol. 1, no. 4, 2021, art. no. e98.

[73] K. Hoffmann, R. Bouckaert, S. J. Greenhill, and D. Kühnert, "Bayesian phylogenetic analysis of linguistic data using beast," *Journal of Language Evolution*, vol. 6, no. 2, pp. 119–135, 2021.

[74] R. Das and S. K. Ghosh, "Genetic variants of the dna repair genes from exome aggregation consortium (exac) database: significance in cancer," *DNA Repair*, vol. 52, pp. 92–102, 2017.

[75] M. Tarailo-Graovac, J. Y. A. Zhu, A. Matthews, C. D. M. V. Karnebeek, and W. W. Wasserman, "Assessment of the exac data set for the presence of individuals with pathogenic genotypes implicated in severe mendelian pediatric disorders," *Genetics in Medicine*, vol. 19, no. 12, pp. 1300–1308, 2017.

[76] O. G. Bahcall, "Exac boosts clinical variant interpretation in rare diseases," *Nature Reviews Genetics*, vol. 17, no. 10, p. 584, 2016.

[77] E. Indelicato, A. Eberl, S. Boesch, L. M. Lange, C. Klein, K. Lohmann, and M. Zech, "Genome aggregation database version 4—allele frequency changes and impact on variant interpretation in dystonia," *Movement Disorders*, vol. 40, no. 2, pp. 357–362, 2025.

[78] J. E. Park, T. Lee, K. Ha, E. H. Cho, and C.-S. Ki, "Carrier frequency and incidence estimation of familial hemophagocytic lymphohistiocytosis in east asian populations by genome aggregation database (gnomad) based analysis," *Frontiers in Pediatrics*, vol. 10, 2022, art. no. 975665.

[79] S. Gudmundsson, M. S. Berk, N. A. Watts, W. Phu, J. K. Goodrich, M. Solomonson, G. A. D. Consortium, H. L. Rehm, D. G. MacArthur, and A. O. Luria, "Variant interpretation using population databases: Lessons from gnomad," *Human Mutation*, vol. 43, no. 8, pp. 1012–1030, 2022.

[80] G. K. Kamalam, N. S. Baby, R. Dharunya, J. Harini, and T. Kowres, "An in-depth analysis of ai techniques for predicting genetic disorders," in *Proc. 2024 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–7.

[81] X. Mao, Y. Huang, Y. Jin, L. Wang, X. Chen, H. Liu, X. Yang *et al.*, "A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs," *npj Digital Medicine*, vol. 8, no. 1, 2025, art. no. 68.

[82] B. Vidhya, B. L. Shivakumar, S. S. Maidin, and J. Sun, "An effective investigation of genetic disorder disease using deep learning methodology," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1376–1385, 2024.

[83] Y. Akshatha and S. P. Raja, "Certain investigations on improving the object recognition accuracy in sickle cells using the yolov5 algorithm," *Oxidation Communications*, vol. 46, no. 1, 2023.

[84] J. Manokaran, J. G. Flores, and E. Ukwatta, "Fully automated aortic segmentation of 3d phase-contrast magnetic resonance angiography images using deep learning techniques," in *Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 12468. SPIE, 2023, pp. 31–37.

[85] M. Chang, J. Ahn, B. G. Kang, and S. Yoon, "Cross-modal embedding integrator for disease-gene/protein association prediction using a multi-head attention mechanism," *Pharmacology Research & Perspectives*, vol. 12, no. 6, p. e70034, 2024.

[86] M.-S. Kwon, Y.-S. Jung, J.-G. Park, and Y.-C. Ahn, "Transfer learning with multi-sequence mri for segmentation of autosomal dominant polycystic kidney disease using u-net," *Electronics*, vol. 13, no. 10, p. 1950, 2024.

[87] A. Raza, F. Rustam, H. U. R. Siddiqui, I. de la Torre Diez, B. Garcia-Zapirain, E. Lee, and I. Ashraf, "Predicting genetic disorder and types of disorder using chain classifier approach," *Genes*, vol. 14, no. 1, p. 71, 2022.

[88] H. Wang and P. Avillach, "Retracted: Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: Genotype-based deep learning," *JMIR Medical Informatics*, vol. 9, no. 4, p. e24754, 2021.

[89] E. Manduchi, J. D. Romano, and J. H. Moore, "The promise of automated machine learning for the genetic analysis of complex traits," *Human Genetics*, vol. 141, no. 9, pp. 1529–1544, 2022.

[90] F. Assunção, N. Lourenço, B. Ribeiro, and P. Machado, "Evolution of scikit-learn pipelines with dynamic structured grammatical evolution," in *Int. Conf. Applications of Evolutionary Computation (Part of EvoStar)*. Cham: Springer International Publishing, 2020, pp. 530–545.

[91] A. L. Webb, P. Kruczkiewicz, L. B. Selinger, G. D. Inglis, and E. N. Taboada, "Development of a comparative genomic fingerprinting assay for

rapid and high resolution genotyping of arcobacter butzleri," *BMC Microbiology*, vol. 15, pp. 1–12, 2015.

[92] B. Miljković-Selimović, T. Babić, B. Kocić, L. Ristić, T. Milenković, and D. Bogdanović, "Comparative genomic fingerprinting for the subtyping of campylobacter jejuni and campylobacter coli biotypes," *Srpski Arhiv za Celokupno Lekarstvo*, vol. 145, no. 9–10, pp. 492–497, 2017.

[93] I. Panigrahi, "Genetic fingerprinting for human diseases: Applications and implications," in *DNA Fingerprinting: Advancements and Future Endeavors*. Singapore: Springer Singapore, 2018, pp. 141–150.

[94] R. A. A. Shekan, A. M. Abdulkadium, and A. M. A. Majid, "Hereditary human disorders identification through fingerprint analysis," *Indian Journal of Forensic Medicine & Toxicology*, vol. 13, no. 1, pp. 236–240, 2019.

[95] Z. Usmani, "Family of five - genome dataset," https://www.kaggle.com/datasets/zusmani/family-genome-dataset, 2025, [Accessed: Mar. 28, 2025].

[96] D. He and L. Parida, "Does encoding matter? a novel view on the quantitative genetic trait prediction problem," *BMC Bioinformatics*, vol. 17, pp. 1–9, 2016.

[97] A. Lavin, C. M. Gilligan-Lee, A. Visnjic, S. Ganju, D. Newman, S. Ganguly, D. Lange *et al.*, "Technology readiness levels for machine learning systems," *Nature Communications*, vol. 13, no. 1, p. 6039, 2022.

[98] J. Torres-Sospedra, D. P. Q. Gaibor, J. Nurmi, Y. Koucheryavy, E. S. Lohan, and J. Huerta, "Scalable and efficient clustering for fingerprint-based positioning," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3484–3499, 2022.

[99] B. Gulko, M. J. Hubisz, I. Gronau, and A. Siepel, "A method for calculating probabilities of fitness consequences for point mutations across the human genome," *Nature Genetics*, vol. 47, no. 3, pp. 276–283, 2015.

[100] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, U. Naeem, and A. Prugel-Bennett, "Novel centroid selection approaches for kmeans-clustering based recommender systems," *Information Sciences*, vol. 320, pp. 156–189, 2015.

[101] D. Mandelker, L. Zhang, Y. Kemel, Z. K. Stadler, V. Joseph, A. Zehir, N. Pradhan *et al.*, "Mutation detection in patients with advanced cancer by universal sequencing of cancer-related genes in tumor and normal dna vs guideline-based germline testing," *JAMA*, vol. 318, no. 9, pp. 825–835, 2017.

[102] J. P. Hou, A. Emad, G. J. Puleo, J. Ma, and O. Milenkovic, "A new correlation clustering method for cancer mutation analysis," *Bioinformatics*, vol. 32, no. 24, pp. 3717–3728, 2016.

[103] J.-K. Rhee, J. Yoo, K. R. Kim, J. Kim, Y.-J. Lee, B. C. Cho, and T.-M. Kim, "Identification of local clusters of mutation hotspots in cancer-related genes and their biological relevance," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 5, pp. 1656–1662, 2018.

[104] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021.

[105] G. Zhu, X. Li, S. Zhang, X. Xu, and B. Zhang, "An improved method for k-means clustering based on internal validity indexes and inter-cluster variance," *International Journal of Computer Science and Engineering*, vol. 25, no. 3, pp. 253–261, 2022.

[106] J. M. Legare, "Synonym: Fgfr3-related achondroplasia," 1993.

[107] Q. Xiao and V. M. Lauschke, "The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders," *NPJ Genomic Medicine*, vol. 6, no. 1, p. 41, 2021.

[108] S. Ebrahimkhani and G. Asaadi Tehrani, "Evaluation of the gjb2 and gjb6 polymorphisms with autosomal recessive nonsyndromic hearing loss in iranian population," *Iranian Journal of Otorhinolaryngology*, vol. 33, no. 115, pp. 79–84, 2021.

[109] R. G. Lewis and B. Simpson, "Genetics, autosomal dominant," 2020.

[110] R. Shiang, L. M. Thompson, Y.-Z. Zhu, D. M. Church, T. J. Fielder, M. Bocian, S. T. Winokur, and J. J. Wasmuth, "Mutations in the transmembrane domain of fgfr3 cause the most common genetic form of dwarfism, achondroplasia," *Cell*, vol. 78, no. 2, pp. 335–342, 1994.

[111] J. E. Spence, R. G. Perciaccante, G. M. Greig, H. F. Willard, D. H. Ledbetter, J. F. Hejtmancik, M. S. Pollack, W. E. O'Brien, and A. L. Beaudet, "Uniparental disomy as a mechanism for human genetic disease," *American Journal of Human Genetics*, vol. 42, no. 2, pp. 217–226, 1988.

[112] P. Krejci, "The paradox of fgfr3 signaling in skeletal dysplasia: why chondrocytes growth arrest while other cells over proliferate," *Mutation Research Reviews in Mutation Research*, vol. 759, pp. 40–48, 2014.

[113] N. Sharma and G. R. Cutting, "The genetics and genomics of cystic fibrosis," *Journal of Cystic Fibrosis*, vol. 19, pp. S5–S9, 2020.

[114] D. Ghosh and J. Cabrera, "Enriched random forest for high dimensional genomic data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 5, pp. 2817–2828, 2021.

[115] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Random forest for genomic prediction," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer International Publishing, 2022, pp. 633–681.

[116] S. S. Patra, O. P. Jena, G. Kumar, S. Pramanik, C. Misra, and K. N. Singh, "Random forest algorithm in imbalance genomics classification," in *Data*

Analytics in Bioinformatics: A Machine Learning Perspective. Springer, 2021, pp. 173–190.

[117] T. Chappell, S. Geva, and J. Hogan, "K-means clustering of biological sequences," in *Proceedings of the 22nd Australasian Document Computing Symposium*, 2017, pp. 1–4.

[118] T. V. Sai Krishna, A. Yesu Babu, and R. Kiran Kumar, "Determination of optimal clusters for a non-hierarchical clustering paradigm k-means algorithm," in *Proceedings of the International Conference on Computational Intelligence and Data Engineering (ICCIDE 2017)*. Singapore: Springer Singapore, 2017, pp. 301–316.

[119] A. Chatzikyriakidou, "Beyond the 'dominant' and 'recessive' patterns of inheritance," *International Journal of Molecular Sciences*, vol. 25, no. 24, p. 13377, 2024.

[120] R. Cilia, S. Tunesi, G. Marotta, E. Cereda, C. Siri, S. Tesei, A. L. Zecchinelli *et al.*, "Survival and dementia in gba-associated parkinson's disease: The mutation matters," *Annals of Neurology*, vol. 80, no. 5, pp. 662–673, 2016.