ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

AN ELABORATIVE MULTIPLEX PRIMER DESIGN ALGORITHM TO DEVELOP COMPREHENSIVE MULTIGENE PANELS

ALPER AKKUŞ

M.Sc. THESIS

DEPARTMENT OF GENOME STUDIES

SUPERVISOR
Prof. Özden Hatırnaz Ng

SECONDARY SUPERVISOR
Asst. Prof. Özkan Özdemir

ISTANBUL-2024

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

AN ELABORATIVE MULTIPLEX PRIMER DESIGN ALGORITHM TO DEVELOP COMPREHENSIVE MULTIGENE PANELS

ALPER AKKUŞ
M.Sc. THESIS

DEPARTMENT OF GENOME STUDIES

SUPERVISOR
Prof. Özden Hatırnaz Ng

SECONDARY SUPERVISOR
Asst. Prof. Özkan Özdemir

ISTANBUL-2024

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethical rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

12.09.2024
Alper Akkuş

PREFACE AND ACKNOWLEDGEMENT

First and foremost, I extend my deepest gratitude to my advisors, Professor Özden Hatırnaz Ng and Asst. Professor Özkan Özdemir, for their invaluable support throughout the preparation of this thesis. Their guidance is crucial in shaping both the content and direction of all my work.

I am deeply honored to be part of ACURARE – the Rare Diseases and Orphan Drugs Application and Research Center. My sincere appreciation goes to our former director and founder, Professor Uğur Özbek, whose vision and commitment laid the foundation for our groundbreaking research. I wish to express my sincere thanks to Dr. Yuk Yin (Peter) Ng from Istanbul Bilgi University, who will always be remembered as my invaluable undergrad advisor. Also, I would like to thank all the instructors from the Genome Studies Department for their guidance and support.

Furthermore, I am profoundly grateful to all my colleagues at ACURARE for their unwavering support. I would also like to extend my heartfelt thanks to my dear friends—İlayda Şahin, Gizem Önder, Ufuk Amanvermez, T. Taygun Turan, Eylül Aydın and Aybike Ş. Bulut—for their constant encouragement. I owe a deep debt of gratitude to Berk Ergun whose guidance in bioinformatics helped shape my research. Also, I am thankful to Assoc. Prof. Nihat Buğra Ağaoğlu for all feedback. Additionally, I must acknowledge the exceptional efforts of our fellow, Yunus Eken, and our intern, Hüma Günay, who contributed to the experimental part of this thesis.

I would like to acknowledge Professor Naci Çine for his invaluable support by providing essential patient data that greatly enriched this research. Additionally, I extend my warmest regards and heartfelt appreciation to all the individuals battling cancer who agreed to contribute their data to genomic research.

Finally, I must express my profound gratitude to my parents. Their unwavering support, both in times of health and illness, has been the cornerstone of my achievements, allowing me to complete my education and this thesis.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

AUC	Area Under the Curve
AP	Average Precision
APC	Adenomatous Polyposis Coli
AS_FilterStatus	Allele-Specific Filter Status
AS_SB_TABLE	Allele-Specific Strand Bias Table
BAM	Binary Alignment Map
BRAF	B-Raf Proto-Oncogene
BRCA1	Breast Cancer Type 1 Susceptibility Protein
BRCA2	Breast Cancer Type 2 Susceptibility Protein
BRIP1	BRCA1 Interacting Protein C-terminal Helicase 1
BQSR	Base Quality Score Recalibration
BWA-MEM2	Burrows-Wheeler Aligner MEM version 2
CDH1	Cadherin 1
CDx	Companion Diagnostics
CHEK2	Checkpoint Kinase 2
cfDNA	Cell-Free DNA
CTC	Circulating Tumor Cells
ctDNA	Circulating Tumor DNA
DET	Detectability Condition from vLoD; Determines if detectability threshold is passed (“Yes”) or not (“No”)
DETS	Detectability Score from vLoD; Gives the odds of a true positive (correct detection of a variant) against the probability of errors (both sequencing errors and false positives)
DP	Depth; total number of reads covering the position
ECNT	Event Count
EGFR	Epidermal Growth Factor Receptor
EPCAM	Epithelial Cellular Adhesion Molecule
FDA	Food and Drug Administration
FP	Forward Primer
GATK	Genome Analysis Tool Kit

GERMQ	Genotype Quality for Error Model
HER2	Human Epidermal Growth Factor Receptor 2
HRD	Homologous Recombination Deficiency
LoD	Limit of Detection
MBQ	Median Base Quality
MFRL	Median Fragment Length
MGPT	Multi-Gene Panel Testing
MLH1	MutL Homolog 1
MSH6	MutS Homolog 6
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction
POPAF	Population Allele Frequency Prior
RET	RET Proto-Oncogene
ROC	Receiver Operating Characteristic
RU	Repeat Unit
SAM	Sequence Alignment Map
SMRT	Single Molecule Real-Time (sequencing)
STR	Short Tandem Repeat
STRQ	Short Tandem Repeat Quality
TAD	Topologically Associating Domains
TCEP	Tris(2-carboxyethyl)phosphine
TIRF	Total Internal Reflection Fluorescence
TLOD	Tumor LOD Score; Log odds ratio of the data supporting the variant in tumor versus normal (From Mutect2)
TMB	Tumor Mutational Burden
TP53	Tumor Protein P53
TPPTS	Trisodium tris(3-sulfonatophenyl)phosphine
VAF	Variant Allele Fraction
VCF	Variant Call Format
vLoD	Variant Limit of Detection

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

An Elaborative Multiplex Primer Design Algorithm to Develop Comprehensive Multigene Panels

This thesis presents the development of a comprehensive multiplex primer design algorithm to facilitate the creation of multi-gene panels (MGPT) for the detection of somatic variants in various cancer types. Cancer, being a predominantly genetic disease driven by accumulated genomic alterations, requires advanced diagnostic tools that can accurately identify these variations to inform treatment decisions. The study leverages next-generation sequencing (NGS) technologies to design and validate primers targeting 18 key genes frequently mutated in cancers such as breast, ovarian, colorectal, and pancreatic cancers. The primers designed in this work cover over 94% of the target loci, ensuring broad and accurate variant detection with minimal off-target effects. A critical feature of this research is the integration of open-source primer design tools with custom algorithms, which allow for the optimization of primer specificity and the reduction of dimerization and secondary structure formation, such as hairpins. The development of a bioinformatics pipeline that includes a novel model for calculating the detectability of genetic variants further strengthens the diagnostic accuracy, especially for low-fraction variants. The outcomes of this study demonstrate the potential for improving precision oncology by providing a cost-effective diagnostic tool that enables the detection of somatic variants. The research also addresses the need for accessible cancer diagnostics, particularly in regions such as Türkiye and the MENA region, where the high cost of existing commercial multi-gene panels limits their widespread use. By designing a multiplex panel that can detect multiple gene variants simultaneously, the study will contribute to the growing demand for personalized cancer therapies, allowing for more targeted treatment approaches and improved patient outcomes.

Keywords: Genetics, Bioinformatics, Cancer, Multi-gene panels, Precision oncology

ÖZET

Kapsamlı Çok-Gen Panelleri Geliştirmek Üzere Detaylı Çoklu Primer Tasarım Tekniği

Bu tez, çeşitli kanser türlerinde somatik varyantların tespitine yönelik kapsamlı bir Çok-Gen Paneli (MGPT) geliştirilmesi için yenilikçi bir çoklu primer tasarım algoritmasının geliştirilmesini ele almaktadır. Kanser, genomdaki değişiklikler sonucu gelişen bir hastalık olduğundan, bu değişiklikleri doğru bir şekilde tespit edebilen tanı araçlarına duyulan ihtiyaç artmaktadır. Çalışmada, özellikle meme, yumurtalık, kolon ve pankreas kanserleri gibi birçok kanser türünde sıkça mutasyona uğrayan 18 önemli geni hedef alan primerlerin tasarımı için yeni nesil dizileme (NGS) teknolojileri kullanılmıştır. Bu çalışmada tasarlanan primerler, hedeflediğimiz lokusların %94'ünden fazlasını kapsamaktadır ve bu sayede geniş kapsamlı ve doğru varyant tespiti sağlanmaktadır. Araştırmanın en önemli çıktılarından biri, açık kaynaklı primer tasarım araçları ile özel algoritmaların entegre edilmesi sonucunda primer özgüllüğünün artırılması ve dimerizasyon ile ikincil yapı oluşumlarının (*hairpin* gibi) en aza indirilmesidir. Ayrıca, düşük fraksiyonlu varyantların tespit edilebilirliğini hesaplamak için yenilikçi bir biyoenformatik analiz modeli geliştirilmiş olup, bu model genetik varyantların tespit hassasiyetini artırmaktadır. Geliştirilen tanı aracı, çoklu gen varyantlarının aynı anda tespit edilmesine olanak tanıyarak, somatik varyantların doğru bir şekilde tespit edilmesini mümkün kılmaktadır. Bu yenilikçi yaklaşım, yüksek maliyetli tanı kitlerine erişimin sınırlı olduğu Türkiye ve MENA bölgesi gibi bölgelerde kanser tanı ve tedavi süreçlerini iyileştirmek için büyük bir ihtiyaç olan uygun maliyetli tanı araçlarını sağlamaktadır. Aynı zamanda, bu panel hassas (*precision*) kanser tedavilerine duyulan artan talebi karşılayarak, hedefe yönelik tedavi yaklaşımlarının yaygınlaşmasına ve hastaların tedavi sonuçlarının iyileştirilmesine katkıda bulunmaktadır.

Anahtar Kelimeler: Genetik, Biyoinformatik, Kanser, Çok-gen panelleri, Hassas onkoloji

1 INTRODUCTION AND AIM

Cancer remains a leading cause of mortality worldwide, accounting for millions of new cases and deaths each year (1). As a predominantly genetic disease, cancer is primarily driven by somatic variants, which accumulate in cells over time, leading to disruptions in cell growth, division, and apoptosis (2). Although hereditary predispositions contribute to cancer risk, most cancers are sporadic and result from somatic variants acquired over a lifetime (3). Detecting these variants is crucial for enabling personalized treatments that target the specific molecular abnormalities in a patient's tumor (4).

Advances in next-generation sequencing (NGS) technologies have revolutionized cancer research by allowing high-throughput analysis of tumor genomes, enabling the identification of genetic alterations (5). NGS has paved the way for precision oncology, where therapies are designed to target the unique genetic variants within a patient's tumor (6). This shift toward personalized treatment has improved patient outcomes by facilitating earlier diagnosis and more effective therapeutic interventions (7).

One of the most impactful applications of NGS in clinical oncology is multi-gene panel testing (MGPT), which allows for the simultaneous analysis of multiple cancer-related genes (8). These panels can identify key somatic variants that inform treatment decisions and provide insights into tumor biology (9). Despite the advantages of MGPT, high costs and limited accessibility, particularly in regions like Türkiye and the Middle East and North Africa (MENA), present significant barriers to widespread adoption (10). As a result, there is an urgent need for affordable diagnostic tools that can provide accurate and comprehensive genetic analysis in resource-limited settings.

The primary objective of this thesis is to design and develop a cost-effective MGPT prototype for detecting somatic variants in both solid and liquid biopsy samples. The test targets 18 key cancer-related genes, which collectively cover over 94% of the loci frequently mutated in cancers such as breast, ovarian, colorectal, and

pancreatic cancers (11). By focusing on these frequently mutated genes, this research aims to develop a robust diagnostic tool that enables accurate detection of variants, thus supporting personalized cancer therapies.

A key innovation of this research is the development of a custom algorithm for multiplex primer design. This algorithm optimizes primer specificity while minimizing issues such as dimerization and secondary structure formation, ensuring efficient coverage of target loci (12). Additionally, a bioinformatics pipeline has been developed to improve the detectability of low-frequent variants, which is particularly relevant in liquid biopsy applications where circulating tumor DNA (ctDNA) concentrations are often low (13).

By providing a cost-effective and accessible MGPT platform, this research addresses a critical gap in cancer diagnostics. It offers a scalable solution that can be deployed in regions with limited access to high-cost diagnostics, potentially improving cancer detection and treatment outcomes, particularly in Türkiye and the MENA region (14).

2 BACKGROUND

2.1 Human Genome Project and Its Impact

The Human Genome Project (HGP), completed in 2003, was a historic milestone in genetics, resulting in the sequencing of the entire human genome. This project revealed that humans share 99.9% of their DNA, with the remaining 0.1% accounting for individual differences, including susceptibility to diseases such as cancer (15). The HGP paved the way for significant advances in cancer research by enabling the identification of genetic variations that drive tumor development and progression (16).

The most significant impact of the HGP on cancer research has been the development of next-generation sequencing (NGS) technologies. These technologies allow researchers to rapidly sequence entire genomes, identifying somatic variants—variants that acquired during a person's lifetime that occur in specific cells and are not passed to offspring (17). The introduction of NGS has revolutionized our understanding of the genetic alterations involved in cancer, leading to the discovery of numerous cancer-related variants and facilitating the development of targeted therapies (18).

Additionally, the HGP laid the groundwork for large-scale cancer genomics initiatives such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). These projects have cataloged the genetic variants present in thousands of tumors, providing researchers with invaluable resources for understanding the molecular underpinnings of cancer (19). The data generated by these initiatives have shown that most cancers are driven by combinations of variants in multiple genes, rather than by a single mutation, which has led to a paradigm shift in cancer treatment strategies (20).

2.2 Variant Types: Somatic and Germline Variants

Cancer is a genetic disease driven by variants that alter the function of critical genes involved in cell growth, division, and death. These variants can be classified into two main types: somatic and germline variants. Somatic variants are variants that occur in non-reproductive cells during an individual's lifetime. They are typically caused by environmental factors, such as exposure to ultraviolet radiation or carcinogens, and are restricted to the tumor cells (21). Somatic variants are responsible for the majority of cancers, as they accumulate over time and disrupt normal cellular regulation, leading to uncontrolled cell growth (22). For instance, variants in oncogenes such as *KRAS* and tumor suppressor genes like *TP53* are commonly observed in various cancers (23).

In contrast, germline variants are inherited variants that are present in every cell of the body and can be passed on to offspring. These variants predispose individuals to an increased risk of developing certain cancers. A well-known example of germline variants is found in the *BRCA1* and *BRCA2* genes, where variants significantly increase the risk of breast and ovarian cancers (24). Individuals carrying these variants may benefit from enhanced screening protocols or preventive measures, such as prophylactic surgery (25).

The identification of both somatic and germline variants is crucial for developing targeted cancer therapies. Advances in NGS technologies have made it possible to detect these variants simultaneously, providing clinicians with a comprehensive genetic profile of the tumor and allowing for more personalized treatment approaches (26).

Detecting genetic variants in cancers not only facilitates the association with treatments but also enables the identification of variant carrier individuals, thereby allowing the implementation of risk-reducing methods. As an instance, since *BRCA* variants influence treatment choices in patients with epithelial ovarian cancer, testing for these variants has become a routine part of healthcare services worldwide. New

findings for targeted therapies in patients with these variants are emerging (23-26). Somatic variants play a crucial role in understanding tumor development (Figure 1).

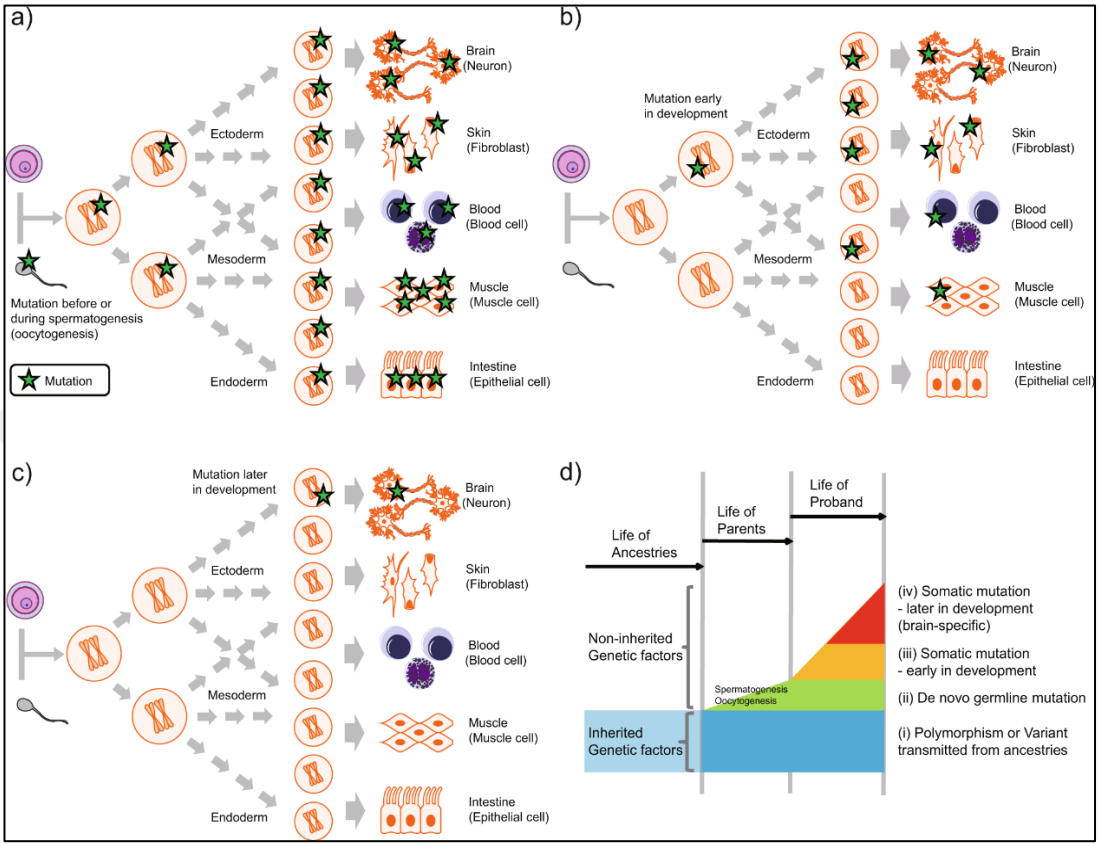

Figure 1. Somatic variogenesis

The distribution of somatic variants highlights the stages at which these variants can occur and their possible effects on cancer formation. Basically, it can be observed that during different development stages, a variant can exist in all, some, or a single tissue throughout the body. “Figure 1a” represents disease-causing variant exists across all tissues because of a *de novo* variant before or during the spermatogenesis. “Figure 1b” and “Figure 1c” represent somatic variant presence in different tissues due to the appearance stage (27, 28).

De novo variants can arise during the production of sperm or eggs, leading to changes that are present from the very beginning of an organism's development. These variants are transmitted from the gametes to the embryo, affecting all body tissues of the individual that develops. The probability that these *de novo* germline variations

will be inherited by the next generation is 50%. In contrast, somatic variants that occur early in development, before tissues have differentiated, can be present in multiple but not all cells or tissues of the individual (27, 28).

Variations that occur later in development, post-differentiation of tissues, are confined to specific parts of specific tissues, such as the brain. These later-stage somatic variants typically have a lower allele frequency and do not pass to offspring. Furthermore, an individual's genetic composition encompasses a complex mix of genetic variations, as stated in Figure 1: (i) inherited polymorphisms and variants from ancestors, (ii) *de novo* germline variants, (iii) early developmental somatic variants, and (iv) later developmental tissue-specific somatic variants. While the inherited variations are transmitted from previous generations, the other types are acquired and contribute collectively to the individual's phenotype. This analysis particularly highlights the influence of somatic variants, occurring at various developmental stages, on the traits of an individual (27, 28).

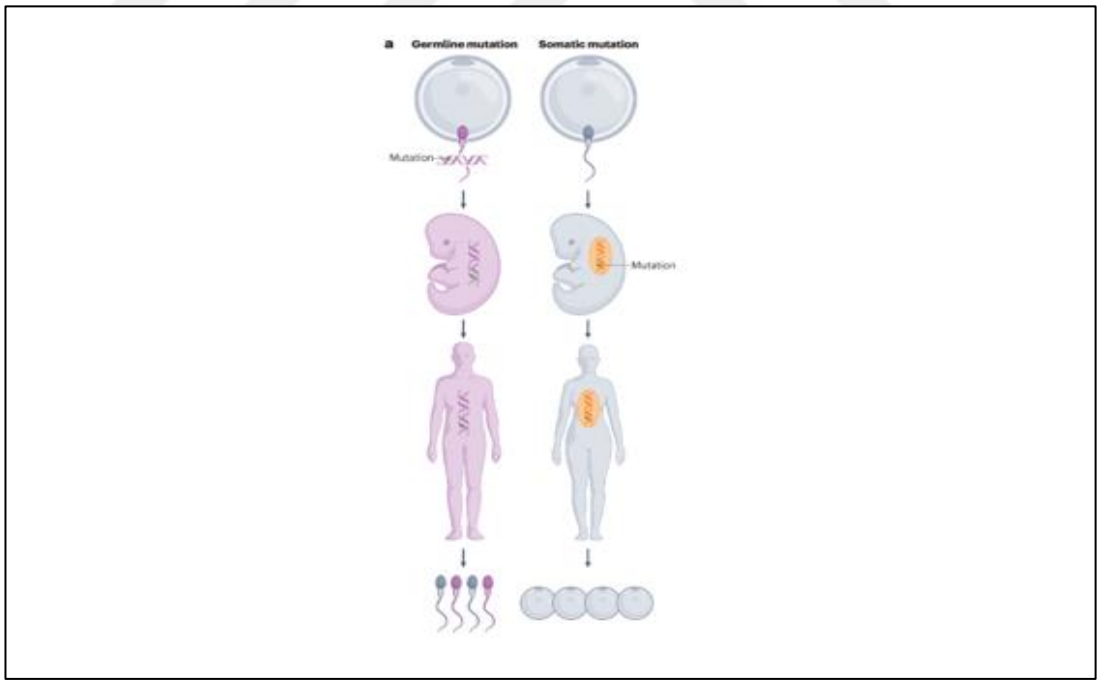

Figure 2. Somatic vs. germline variants

Cancers caused by somatic and germline variants exhibit different patterns of inheritance and manifestation (Figure 2) (29). A comparison between these two types of genetic alterations, illustrating how germline variants are inherited and present in every cell, while somatic variants arise spontaneously in a subset of cells post-conception. This distinction is crucial for understanding the different pathways through which cancer can develop and the implications for diagnosis and treatment.

Sequencing data often shows distinct patterns for germline and somatic variants. Germline variants typically display allelic fractions around 50% for heterozygous variants and near 100% for homozygous variants, indicating their presence in all cells. Somatic variants, however, are characterized by much lower allelic fractions, indicating their presence in only a subset of cells (29).

The research into germline variants focuses on understanding how variants are passed through generations, tracing ancestry, exploring genetic associations, and studying the selection of these variants in broader human populations. Research on somatic variants centers on mutation rates over an individual's lifespan, cell lineage tracking, characterizing variants specific to somatic cells, and selection processes within groups of cells (29, 30).

2.3 Generations of Sequencing

DNA sequencing technologies have evolved significantly over the past several decades, facilitating groundbreaking discoveries in cancer research. The ability to sequence entire genomes with increasing speed and accuracy has been critical for understanding the genetic basis of diseases like cancer. These sequencing technologies have progressed through three generations: first-generation sequencing (Sanger sequencing), second-generation sequencing (next-generation sequencing, NGS), and third-generation sequencing (single-molecule sequencing).

2.3.1 First generation: Sanger Sequencing

First-generation sequencing, also known as Sanger sequencing, was developed by Frederick Sanger in the 1970s. This method was the first to allow the determination of the nucleotide sequence of DNA through the selective incorporation of chain-terminating dideoxynucleotides during DNA replication. Despite its pivotal role in early genomic research, Sanger sequencing had several limitations. It was labor-intensive, expensive, and could only sequence short DNA fragments—typically around 700 base pairs per reaction (31). These limitations made it unsuitable for large-scale cancer genomics projects that require the analysis of complex genetic alterations, such as large structural variants (32).

2.3.2 Second generation: Next-Generation Sequencing (NGS)

The development of second-generation sequencing technologies—commonly referred to as NGS—revolutionized the field of genomics by dramatically increasing throughput. NGS platforms, such as Illumina's sequencing by synthesis (SBS), allow for the simultaneous sequencing of millions of DNA fragments. This parallel processing of DNA fragments significantly reduces both time and costs, making large-scale genomic studies feasible (33). In SBS, DNA fragments are randomly broken and attached to adapters, which are then amplified in clusters. Fluorescently labeled nucleotides are incorporated during synthesis, emitting signals that are detected in real time, allowing researchers to rapidly determine the sequence of bases (34). The high throughput of NGS platforms has made it possible to sequence entire genomes or large gene panels, enabling the identification of genetic variants associated with cancer (35). In Illumina's SBS, the process begins with the preparation and immobilization of DNA templates onto a solid surface within the flow cell, where DNA molecules are amplified to form dense clusters of identical fragments. This step of cluster generation is essential because it ensures a sufficient fluorescent signal for accurate base calling. The subsequent enrichment of these DNA templates is achieved through bridge amplification, where the fragments bind to oligonucleotides attached to the flow cell surface, forming clusters through repeated rounds of amplification (Figure 3) (36-39).

Once DNA templates are prepared, sequencing begins by introducing a mixture containing primers, DNA polymerase, and four types of chemically modified nucleotides. These nucleotides are blocked at the 3′-O position with an azidomethyl group, which prevents the incorporation of additional nucleotides after one is added. Additionally, each nucleotide is labeled with a distinct fluorescent dye—adenine, cytosine, guanine, and thymine are tagged with different colors, allowing for precise identification during sequencing (40). The azidomethyl group ensures that only a single nucleotide is incorporated during each cycle, maintaining tight control over the sequencing process and reducing errors from multiple nucleotide additions (41).

After nucleotide incorporation, the flow cell is washed to remove any unincorporated reagents, leaving only the extended DNA strand. This strand is then imaged using total internal reflection fluorescence (TIRF) microscopy, which excites the fluorescent dyes attached to the nucleotides. Platforms such as Illumina's HiSeq and MiSeq use two or four laser channels to detect the specific fluorescent signal emitted by each nucleotide (42). These signals are used to determine which base has been added at each cluster location.

Following imaging, the fluorescent tag is chemically cleaved, and the blocking group is removed by tris(2-carboxyethyl) phosphine (TCEP), restoring the 3′-OH group to allow for the next nucleotide to be incorporated. This cyclic process—nucleotide incorporation, imaging, and chemical regeneration—continues for multiple cycles, allowing the system to sequence millions of DNA clusters simultaneously. This method provides the deep coverage and accuracy required for genomic applications such as whole-genome sequencing and targeted sequencing (43).

On the other hand, NGS technologies have proven especially valuable for studying tumor heterogeneity—the genetic diversity within a single tumor (Figure 3). Tumor heterogeneity plays a significant role in cancer treatment outcomes, as different subclones of a tumor may have distinct variants and exhibit different responses to therapy (36, 43). The ability to sequence DNA at high depth with NGS has made it

possible to identify these subclones, offering insights into tumor evolution and the development of drug resistance.

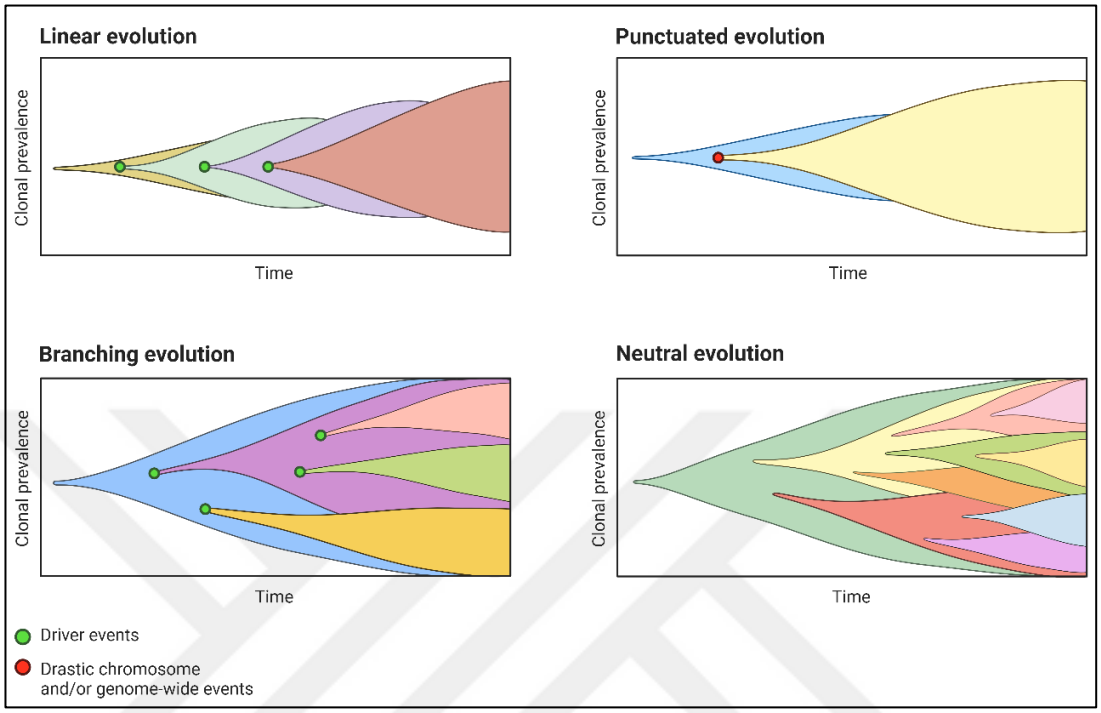

Figure 3. Clonal evolutions in cancers (Generated from a Biorender template)

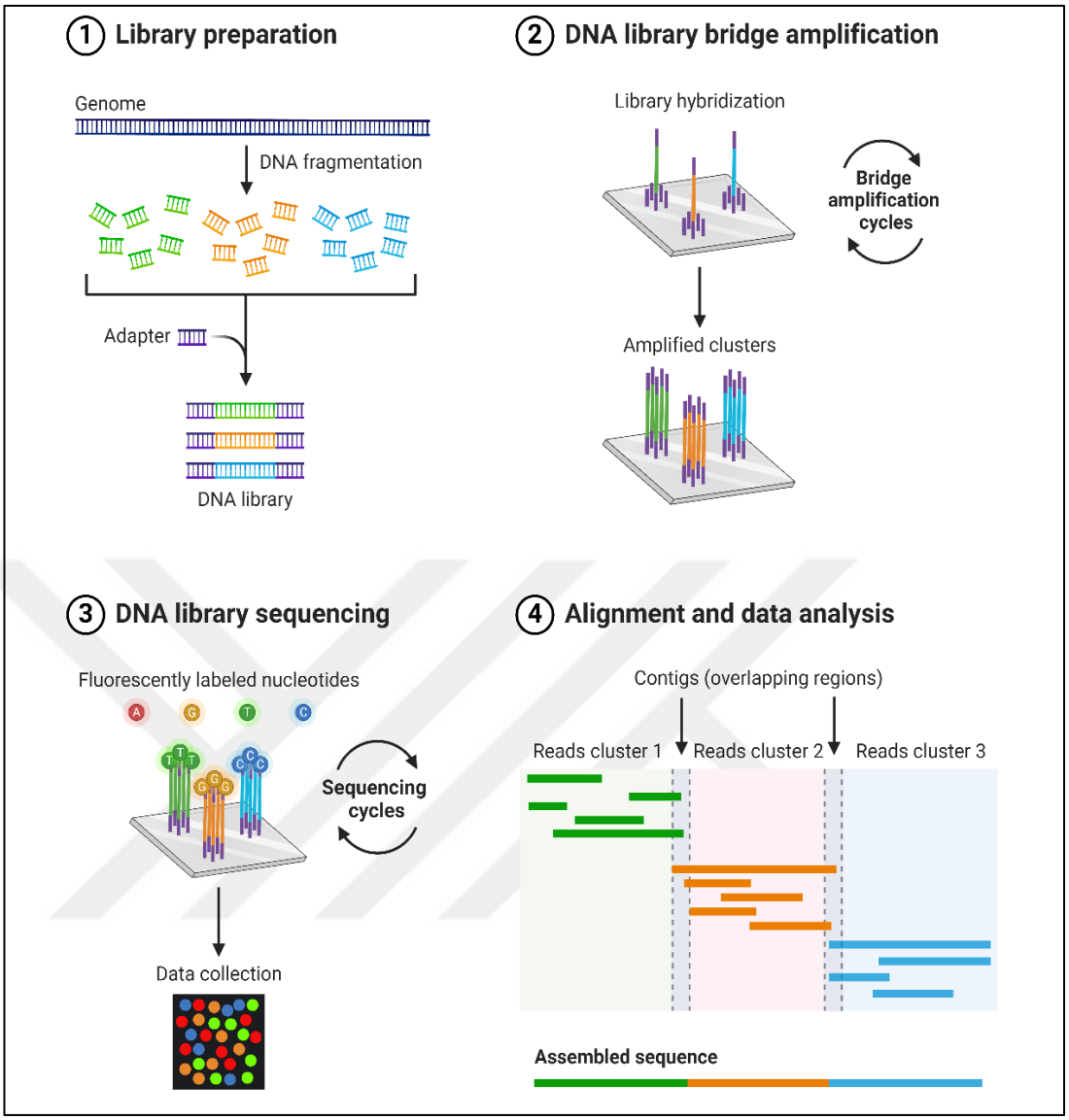

Figure 4. Cyclic reversable termination demonstration (Generated from a Biorender template)

Using the same techniques but with different target enrichment strategies, the range of NGS platforms can be compared to visualize the trade-offs between read lengths and throughput levels that researchers must consider when choosing a sequencing method (Figure 5).

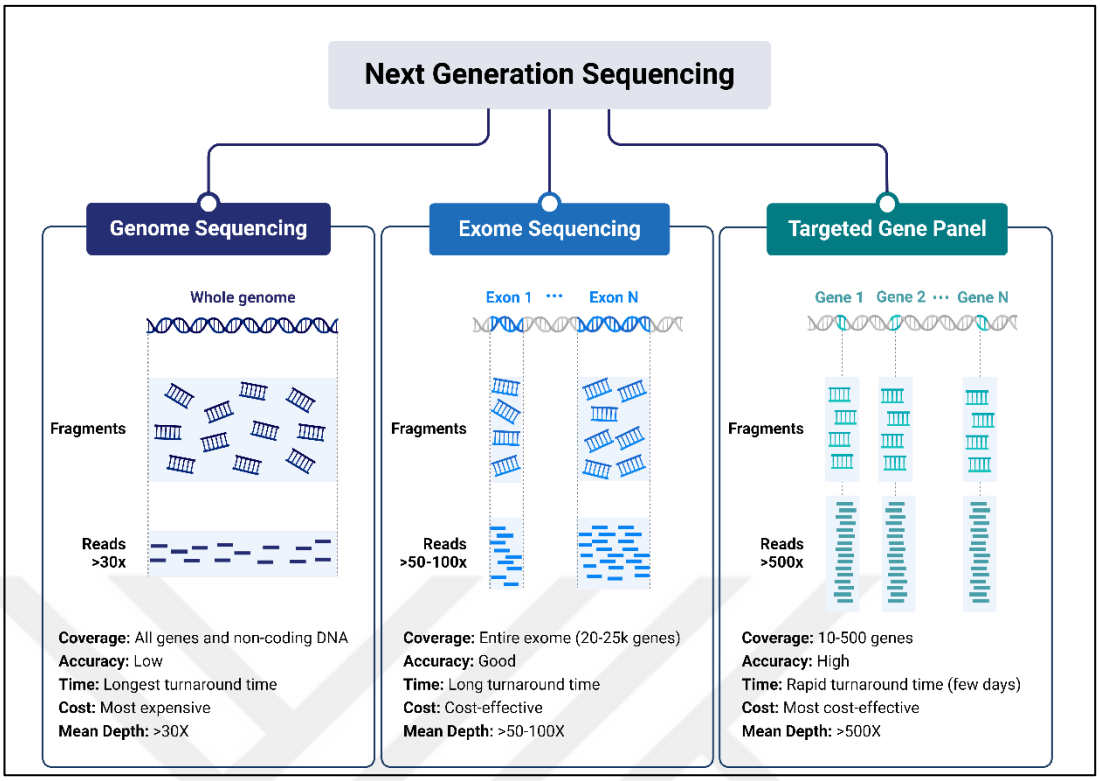

Figure 5. NGS comparison in terms of target span (Generated from a Biorender Template)

2.3.3 Third generation: single molecule sequencing

Third-generation sequencing represents the latest advancements in DNA sequencing technologies. These platforms, such as Pacific Biosciences' Single Molecule Real-Time (SMRT) sequencing and Oxford Nanopore sequencing, allow for the direct sequencing of single DNA molecules without the need for amplification (44).

This technology provides much longer read lengths compared to NGS, which is essential for detecting large structural variants and rearrangements commonly found in cancer genomes. In addition, third-generation sequencing is capable of detecting epigenetic modifications—chemical changes to DNA that affect gene expression without altering the underlying nucleotide sequence. Epigenetic changes play a crucial role in cancer progression, as they can regulate oncogene activation and tumor suppressor gene silencing (45).

Oxford Nanopore Technologies (ONT) has introduced a radically different approach to sequencing that focuses on the direct detection of single-stranded DNA or RNA molecules passing through a nanopore embedded in a biological or synthetic membrane. Nanopore sequencing involves threading the nucleic acid strand through a nanoscale pore—typically a protein pore embedded in an electrically insulating membrane—under an applied voltage. As each nucleotide passes through the pore, it disrupts the ionic current flowing through the nanopore in a sequence-specific manner, with each nucleotide (adenine, cytosine, guanine, thymine, or uracil) causing a distinct change in current. These changes are detected in real-time, allowing the sequence of the nucleic acid to be determined by interpreting the pattern of current disruptions (46).

In summary, third-generation sequencing technologies like SMRT and nanopore sequencing offer a profound leap forward in our ability to interrogate genomes, transcriptomes, and epigenomes with unprecedented precision. Their ability to generate long reads, detect epigenetic modifications, and sequence in real-time without amplification bias makes them indispensable tools in modern genomics, enabling deeper insights into genetic diversity, disease mechanisms, and potential therapeutic targets.

2.4 Cancer Research in the NGS Era

The development and adoption of next-generation sequencing (NGS) have revolutionized cancer research by enabling the comprehensive analysis of tumor genomes. Prior to NGS, cancer research was largely confined to the study of individual genes or small regions of the genome. However, with the advent of NGS technologies, researchers can now sequence entire tumor genomes, uncovering the full spectrum of somatic variants, structural variants, and copy number alterations that drive cancer (47).

One of the most significant outcomes of NGS technology is its role in large-scale cancer genomics projects, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). These initiatives have cataloged

the genetic variants present in thousands of tumor samples across a wide range of cancer types (48). These comprehensive datasets have provided invaluable insights into the molecular mechanisms of cancer, revealing that most cancers are driven by combinations of variants in multiple genes, rather than by single, isolated variants.

NGS-based studies have identified driver variants, which are genetic alterations that directly contribute to cancer development and progression. Identifying these variants has been instrumental in the development of targeted therapies—drugs designed to inhibit the activity of specific oncogenes or restore the function of tumor suppressor genes. For example, *EGFR* variants in non-small cell lung cancer (NSCLC) and *BRAF* variants in melanoma can be targeted with specific inhibitors, leading to improved outcomes for patients with these cancers (49). The ability to match patients with targeted therapies based on the genetic alterations in their tumors represents one of the major successes of precision oncology.

NGS technologies have also facilitated the development of liquid biopsies, which allow for the detection of circulating tumor DNA (ctDNA) in the bloodstream. Liquid biopsies provide a non-invasive alternative to traditional tissue biopsies and enable the real-time monitoring of tumor progression and treatment response (50). Liquid biopsies are particularly useful for tracking tumor evolution, as they allow for the detection of new variants that may arise during therapy, providing a way to adjust treatment plans accordingly (51).

In addition to its applications in basic research, NGS has become a crucial tool in clinical oncology. Gene panels, which are designed to sequence a specific set of cancer-related genes, are widely used in clinical practice to identify actionable variants. These multi-gene panels enable clinicians to assess the genetic profile of a tumor and make informed decisions about treatment. Table 2, in the Methods part, lists the genes included in the multi-gene panel test (MGPT) developed for this research, which targets 18 key cancer-related genes, including *BRCA1, TP53,* and MLH1 (52). These genes are frequently mutated in breast, colorectal, and ovarian cancers, making them important targets for diagnostic testing and treatment planning.

NGS has not only expanded our understanding of the genetic basis of cancer but has also reshaped clinical practice by enabling more personalized treatment approaches. As NGS technologies continue to improve, the hope is that they will provide even deeper insights into the genetic underpinnings of cancer, leading to more effective therapies and better outcomes for patients.

2.4.1 Gene panel testing in cancer diagnostics

The use of multi-gene panels in cancer diagnostics has become an essential tool for identifying genetic variants associated with cancer risk, prognosis, and treatment responses. Gene panels enable the simultaneous analysis of multiple genes that are known to be frequently mutated in cancers, providing a comprehensive view of the tumor's genetic landscape. The ability to screen for variants in several cancer-related genes is critical for developing precision treatment plans (53).

In recent years, gene panels have been designed to target specific genes that play a role in cancer development. These include genes like *TP53*, *BRCA1*, *BRCA2*, and *MLH1*, which are associated with different cancer types such as breast, ovarian, colorectal, and lung cancers. For example, *BRCA1* and *BRCA2* variants significantly increase the risk of developing breast and ovarian cancers, and testing for these variants is now standard practice in oncology (53).

Gene panel testing is particularly valuable for targeted therapy, as it helps clinicians identify specific variants that can be treated with targeted drugs. For instance, *EGFR* variants in non-small cell lung cancer (NSCLC) are treated with *EGFR* inhibitors, while *BRAF* variants in melanoma are targeted by *BRAF* inhibitors (54). The data generated by gene panels allow for a personalized approach to cancer treatment, where therapies are tailored based on the specific genetic alterations in a patient's tumor, improving the efficacy of the treatment while minimizing side effects.

Additionally, the use of liquid biopsies in gene panel testing is gaining traction. Liquid biopsies, which analyze circulating tumor DNA (ctDNA) from blood samples,

provide a non-invasive alternative to traditional tissue biopsies. This method allows for the real-time monitoring of tumor evolution and treatment response, making it possible to detect emerging variants that may lead to drug resistance.

2.4.2 Companion Diagnostics (CDx)

Companion Diagnostics (CDx) is an *in vitro* technique that initiates certain features related to a disease and provides relevant conclusions about the diagnosis, prognosis, and therapeutics. In the genetics field, CDx is usually applied as a high throughput targeted sequencing to visualize variants and associate these with FDA (United States of America Food and Drug Administration) approved therapeutic approaches. Knowledge of the molecular and genetic changes underlying cancers enables the development of individualized approaches to diagnosis, disease prediction, prognosis and treatment strategies (FDA; The United States of America Food and Drug Administration, 2018) (8, 9, 52) (Table 1).

Furthermore, somatic testing with Next Generation Sequencing (NGS) allows the identification of patients with HRD (Homologous Recombination Deficiency) or MRD (Minimal Residual Disease) for whom targeted therapies can be used. Therefore, the search for somatic variants in tumor tissue is significant. The homologous recombination deficiency (HRD) test is performed in peripheral blood using established techniques. Since the prevalence of variants in some genes is beyond normal, there may be differences in the interpretation of such tests. Moreover, loss of heterozygosity frequently occurs with the progression of cancer.

Initially, a heterozygous state is required, which shows that the locus of interest lacks a functioning tumor suppressor gene pair. Despite this loss, many people continue to be in good health since there is still one intact gene on the opposite pair of chromosomes. A point mutation or other process that results in a loss of heterozygosity event can suppress the tumor suppressor gene's surviving copy, rendering it useless for the body's defense. Small amounts of cancer cells that persist in an individual when they are in remission are referred to as minimal residual disease (MRD) (no symptoms

or signs of disease). It is the main factor in recurrent cancer. MRD can be predicted through NGS. The burden (a heavy amount) of somatic variants in tumor tissue have resolved different characterizations of cancer disease (15, 16, 53-56). For example, the development of immune checkpoint inhibitors (ICI) is a significant achievement in cancer treatment. Immune checkpoints are known to regulate immune response and crucial to inhibit random attacks. Cancer cells are known to have higher levels of immune checkpoint molecule expression.

Therefore, the higher expression of immune checkpoint molecules results in a lower detection of cancer cells by the immune system. However, non-inherited (somatic) variants in tumor tissue can accumulate as a biomarker of cancer disease. The Tumor Mutational Burden (TMB) is the mathematical expression to provide relevant therapeutic knowledge calculated by the ratio of somatic variants per mega base. TMB, as a biomarker, can be used as an indicator to decide in using ICI therapy for different cancers. Usually, more than 20 somatic variants per mega base is a cutoff for TMB and ICI therapy. Some proteins are not visible as a whole gene product but as multiple polypeptides. These fragmented polypeptides can form major histocompatibility complexes and be present as antigens in the cell membrane (57-59).

Hence, if these antigens have comparably higher TMB, they get caught by the immune system. By illuminating biomarkers for patient classification and resistance mechanisms for therapeutic targeting, an awareness of the genetic aspects of response and resistance to checkpoint inhibition may improve the advantages for patients. Consequently, ICI therapy is used to induce immune response for individuals that have high TMB. The detection of TMB is derived from the NGS Data (60, 61).

CDx in Precision Oncogenomic is a collective methodology that arises from the uniqueness of tumors (62-66). The loss of genome instability results from the deficient DNA repair mechanism and unpredictably increases over time. Genome instability causes different genomic variants in the same tumor tissue. Tumors are known to have instable genomes, caused by the alterations of the cell division processes. A high variation rate in any cellular linage prior to carcinogenesis results in an instable

genome. Since tumor will abolish countless variations due to the loss of DNA repair mechanism, different tumor sites are very likely to have different cellular origins. These differentiations of cellular origins are susceptible to change and generate new somatic variants (67, 68). Figure 6 at the end of this section provides an overview of precision and Oslerian (general) medicine approaches.

Therefore, liquid biopsy samples are prospective to yield more information on constantly evolving heterogeneity on a broader scale (62). Plasma contains cell-free genomic DNA content along with circulating tumor DNA (ctDNA). Cell-free DNA (cfDNA) is all DNA molecules that are available in the bloodstream. On the other hand, ctDNA is a cfDNA that originates from tumor tissue. Following the apoptotic DNA fragmentation, DNA content is released into the plasma due to an unknown pathway. NGS-based methods are available for Liquid biopsy samples to enable companion diagnostics and interpret somatic variants in cancer. The FDA has published or approved a variety of personalized treatment options for cancer patients. Since everyone has a unique genome, their genomic factors in a cancerous state, as well as therapeutic approaches, are also unique. Taken together, one of the main purposes of the study is to reach these driver somatic variants and associate these with approved therapeutics or prognostics. (U.S. Food and Drug Administration, 2018) (62, 63, 69)

Nonetheless, the count of somatic variants needs to be discreet. The NGS may have impurities that yield false positive variant calls. The intratumor heterogeneity sets a limitation for the detection of somatic variants since they are likely to be lower in depth. The problem of when to know which read is correct in high throughput sequencing data is achieved with the Limit of Detection (LoD). The limit of Detection is the minimal read count for a variant to be distinguished from the overall depth. Statistically, the LoD is the minimum concentration that can be confidently detected and distinguished from the blank buffer/liquid.

The binomial distribution-based calculations are available to standardize the somatic variant detection in a given sequencing depth (70-72). The design of the Multi-

Gene Panel Testing (MGPT) is closely linked with the Limit of Detection (LoD). A well-designed MGPT must be sensitive enough to detect low-frequency variants, which directly influence the LoD. By carefully selecting target loci and optimizing primer sequences, we ensure that the MGPT can identify variants even at low concentrations, thereby improving the overall accuracy and reliability of the test.

2.5 Challenges and Biological Questions

Despite the rapid advancements in next-generation sequencing (NGS) technologies, several challenges persist in the field of cancer genomics, particularly when it comes to accurately detecting and interpreting genetic variants. One of the primary challenges is achieving sufficient depth of coverage to detect low-frequency variants, especially in heterogeneous tumor samples. Tumor heterogeneity, wherein different regions of a tumor or even different cells within a tumor harbor distinct variants, complicates the detection of driver variants. In addition, technical limitations, such as sequencing errors or polymerase chain reaction (PCR) amplification biases, can introduce false positives or negatives during variant calling (6, 42).

Another major challenge lies in the interpretation of variants with low fractions. While NGS can identify numerous genetic alterations, not all of them have well-established in data mining. Low variant allele fractions leave clinicians unsure of how to incorporate these findings into treatment plans. This highlights the need for more comprehensive databases and improved bioinformatics tools to aid in the detection of such variants (43, 50).

In light of these challenges, this research seeks to address the question of how best to design an affordable and efficient multi-gene panel test (MGPT) that can accurately detect somatic variants in solid tumors.

Specifically, the focus is on creating a panel that minimizes the risk of false positives and maximizes sensitivity for low-frequency variants, which are often critical for understanding cancer progression and treatment resistance. The development of a

custom primer design algorithm is central to this research, as it ensures that the primers used in the MGPT can cover the target loci efficiently while reducing issues such as dimerization and the formation of secondary structures (9, 55, 101).

Primer design is a crucial aspect of targeted sequencing approaches, such as multi-gene panel testing (MGPT), which are designed to identify genetic variants in specific regions of interest. The efficiency and accuracy of sequencing depend heavily on the quality of the primers used in the amplification of target DNA regions. In this study, a custom primer design algorithm was developed to address the common challenges associated with multiplex PCR-based sequencing (9, 100, 101).

One of the primary challenges in primer design is ensuring primer specificity. Primers must bind specifically to the target DNA regions without hybridizing to non-target sequences. Non-specific binding can result in off-target amplification, leading to inaccurate variant detection and the generation of false positives. The custom algorithm employed in this study uses multiple criteria to ensure high specificity, including thermodynamic stability, melting temperature (Tm), and the avoidance of primer-dimer formation (55, 101).

Another critical issue in primer design is the formation of secondary structures, such as hairpins or primer dimers. These structures can significantly reduce the efficiency of PCR amplification by competing with the target DNA for primer binding. The primer design algorithm implemented in this research includes features to minimize secondary structure formation, ensuring that the primers maintain high amplification efficiency even when used in a multiplex setting (56, 101).

Additionally, the algorithm was designed to cover more than 94% of the targeted loci, ensuring that even small or low-frequency variants can be detected with high sensitivity. This is particularly important for liquid biopsy applications, where circulating tumor DNA (ctDNA) is present at low levels in the bloodstream. The ability to accurately detect rare variants in ctDNA is essential for monitoring tumor progression and treatment response in real-time (50, 57).

The development of high-quality primers is also vital for minimizing amplification biases, which can skew the representation of different alleles during sequencing. Amplification bias occurs when certain DNA fragments are preferentially amplified over others, leading to an over- or under-representation of specific variants (58, 101).

In the context of the thesis, target enrichment using PCR plays a crucial role in the detection of somatic variants in multi-gene panel testing (MGPT). Target enrichment refers to the selective amplification of specific regions of the genome that are known to be associated with cancer. By focusing on these regions, PCR enables the efficient identification of clinically relevant variants while minimizing the sequencing of non-target regions.

In this thesis, a multiplex PCR-based approach was utilized, allowing for the simultaneous amplification of multiple target regions in a single reaction. The biological question driving this approach is: “How can we optimize the detection of low-frequency somatic variants in heterogeneous tumor environments and liquid biopsy samples using a cost-effective and scalable enrichment strategy?” This question addresses the core challenge of capturing critical variants that may be present in only a small fraction of the sample, while ensuring that the method remains feasible for broad clinical application (50, 101).

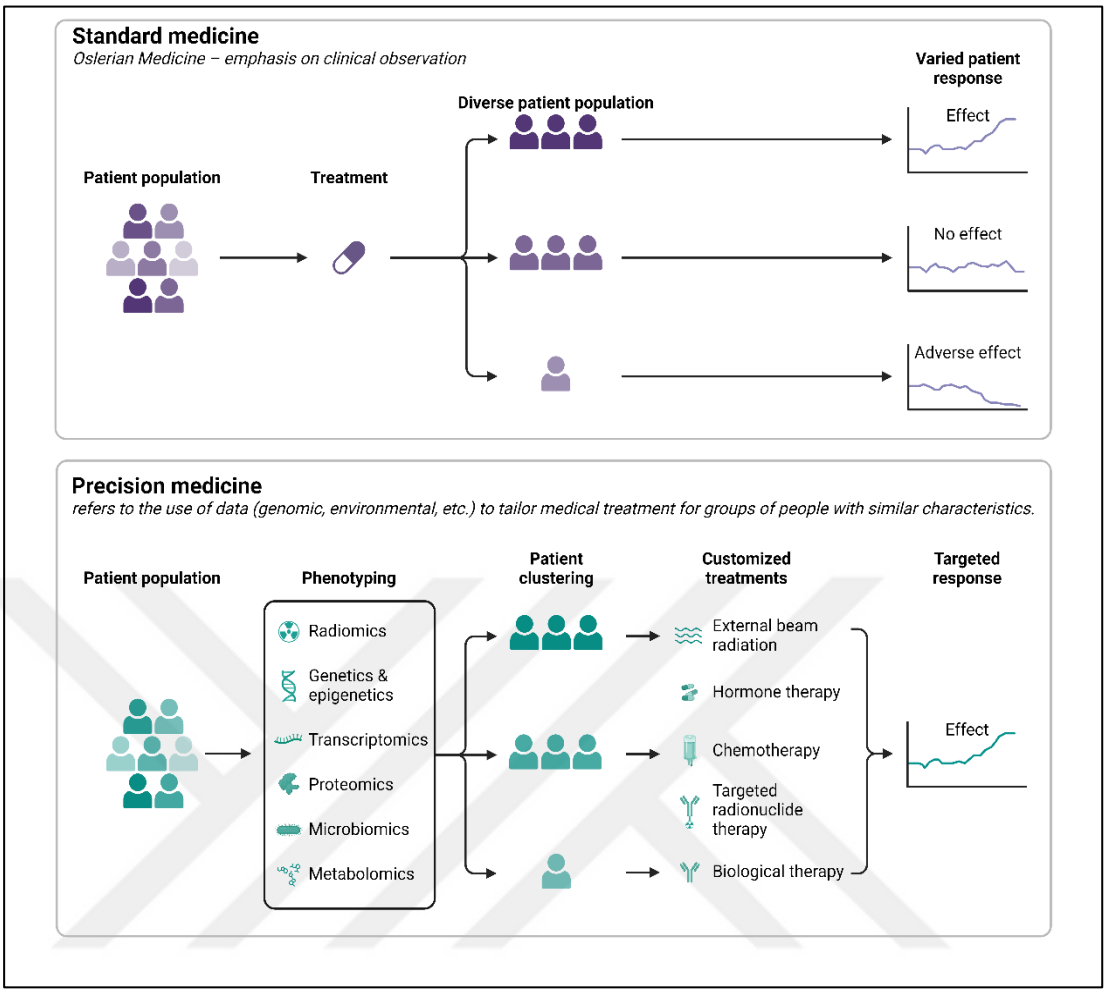

Figure 6. Precision Medicine vs. Oslerian Medicine Overview (Generated from a Biorender Template)

Here we provided FDA approval status for diagnostic kits which are well-known world-wide. The lack of FDA approval for certain CDx does not imply that they are ineffective. As mentioned, Companion Diagnostics (CDx) is an *in vitro* device that enables the determination of possible treatments along with the side effects of each treatment for an individual. (These tests might currently be under investigation or might have received approval from health authorities in other countries.)

Table 1. List of Companion Diagnostics (CDx) worldwide along with assessed cancer types, target biomarkers, FDA approval status, and manufacturer name

CDx Test	Cancer Type	Biomarker Targeted	FDA Approved	Manufacturer
Vysis ALK Break Apart FISH Probe Kit	Non-small cell lung cancer	ALK	Yes	Abbott Molecular
Cobas EGFR Mutation Test v2	Non-small cell lung cancer	EGFR	Yes	Roche
BRACAnalysis CDx	Breast and ovarian cancer	BRCA1/2	Yes	Myriad Genetics
myChoice CDx	Ovarian cancer	Homologous recombination deficiency	Yes	Myriad Genetics
MammaPrint	Breast cancer	70-gene signature	Yes	Agendia
Prosigna Breast Cancer Prognostic Gene Signature Assay	Breast cancer	PAM50 gene signature	Yes	NanoString Technologies
Ventana HER2/neu (4B5) Rabbit Monoclonal Primary Antibody	Breast and gastric cancer	HER2	Yes	Roche
KRAS Mutation Test Kit	Colorectal cancer	KRAS	Yes	Qiagen
BRAF V600E Mutation Test	Melanoma	BRAF V600E	Yes	bioMérieux
Therascreen EGFR RGQ PCR Kit	Non-small cell lung cancer	EGFR	Yes	QIAGEN
PD-L1 IHC 22C3 pharmDx	Multiple cancers	PD-L1	Yes	Agilent Technologies
Guardant360 CDx	Multiple cancers	Multiple biomarkers	Yes	Guardant Health
FoundationOne CDx	Multiple cancers	Multiple biomarkers	Yes	Foundation Medicine
Oncomine Dx Target Test	Non-small cell lung cancer	EGFR, BRAF, ROS1, etc.	Yes	Thermo Fisher Scientific
MSK-IMPACT	Multiple cancers	Multiple biomarkers	Yes	Memorial Sloan Kettering
Oncomine Comprehensive Assay	Multiple cancers	Multiple biomarkers	No	Thermo Fisher Scientific
Tempus xT	Multiple cancers	Multiple biomarkers	No	Tempus

Table 1. List of Companion Diagnostics (CDx) worldwide along with assessed cancer types, target biomarkers, FDA approval status, and manufacturer name (continue)

CDx Test	Cancer Type	Biomarker Targeted	FDA Approved	Manufacturer
GuardantOMNI	Multiple cancers	Multiple biomarkers	No	Guardant Health
ArcherMET	Non-small cell lung cancer	MET exon 14 skipping	No	ArcherDX
InVisionFirst-Lung	Non-small cell lung cancer	EGFR, ALK, ROS1, etc.	No	Inivata
Loxo Oncology RET Fusion CDx	Multiple cancers	RET fusions	No	Loxo Oncology
Epic Sciences Liquid Biopsy	Multiple cancers	Circulating tumor cells (CTCs)	No	Epic Sciences
Caris MI Profile	Multiple cancers	Multiple biomarkers	No	Caris Life Sciences
NeoGenomics MultiOmyx	Multiple cancers	Multiple biomarkers	No	NeoGenomics
Biocartis Idylla™ EGFR Mutation Test	Non-small cell lung cancer	EGFR	No	Biocartis

3 MATERIALS AND METHODS

3.1 Primer Design

In the context of targeted sequencing, selecting optimal primer sequences for target enrichment is pivotal to ensure both efficiency and accuracy. After deciding on the desired loci for the MGPT, we began our methodology with a literature-based three-step comparison to outline their primer design efficiencies. Consequently, we selected the most useful tool for generating primer sets (Figure 7).

Besides, we developed a scoring algorithm for *in-silico* downstream primer sequence analysis using Python programming language version 3.12.2 (2024) (73). This algorithmic design is accomplished in Visual Studio Code, augmented with WSL (Windows Subsystem for Linux), providing a versatile and integrated development environment (74, 75). For this part of the dissertation, the computational work is completed on an *ASUS Vivobook Pro* Laptop, which is equipped with an Intel Core i5 processor, 16GigaByte of memory, and a 512GigaByte SSD.

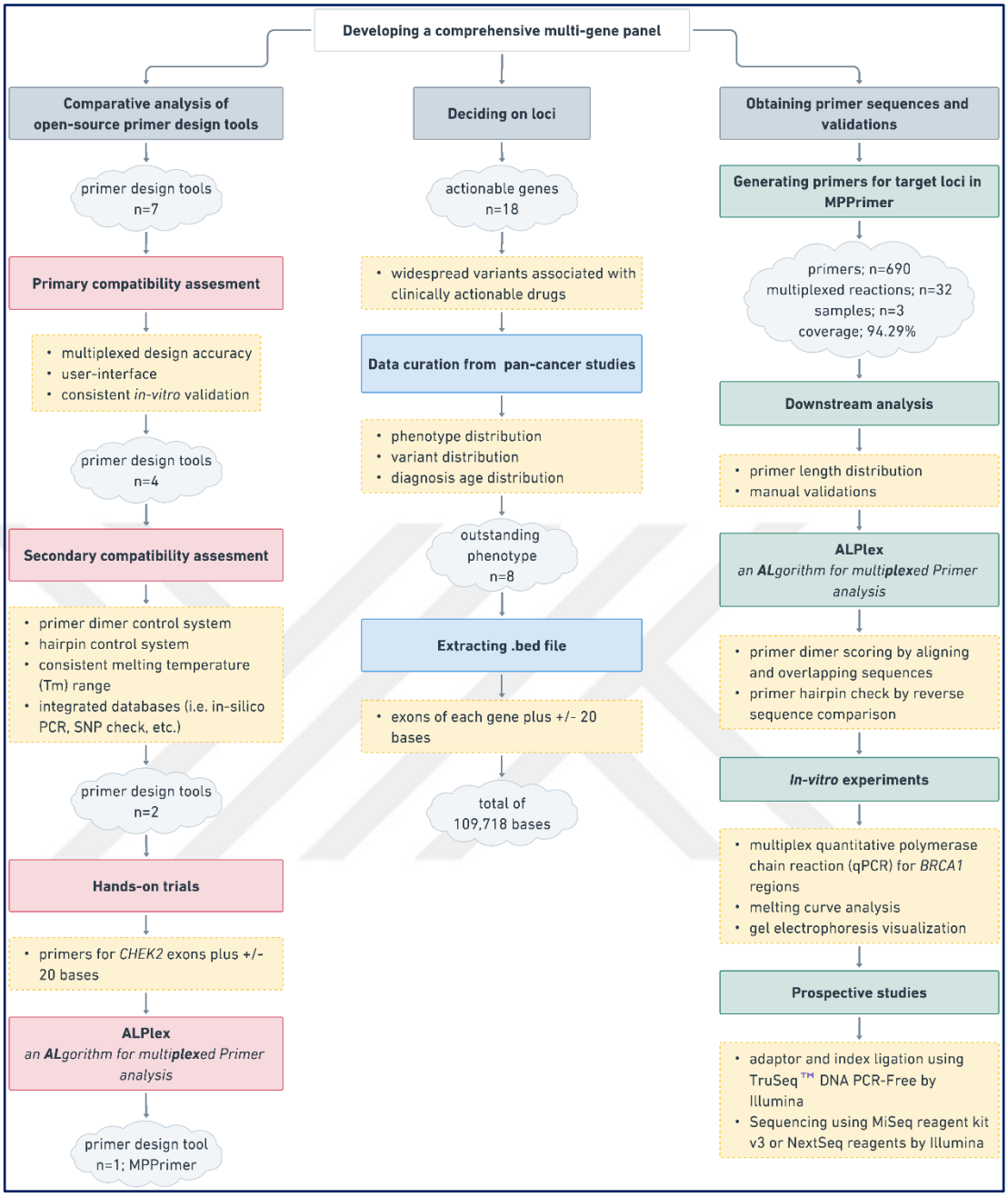

Figure 7. Summary of the MGPT development

3.1.1 Cancer-related gene list for the initial MGPT

We began to work on the loci selection for the MGPT design by using the COSMIC database version 99 (Catalogue of Somatic Variants in Cancer; November 18, 2023) (76, 77). We anchored our selection on two main criteria: the occurrence of somatic variants across different cancer types and actionability.

A list of 18 genes, known for their pivotal roles in tumorigenesis, is prepared (Table 2) (19, 78). For instance, the *TP53* gene, renowned for its widespread somatic variants in various cancers, was chosen for its vital role in maintaining genomic stability and tumor suppression (19). Similarly, genes like *MLH1* and *MSH6* were selected due to their association with mismatch repair deficiency, a common characteristic in sporadic colorectal and endometrial cancers (78). Likewise, the inclusion of genes such as *EPCAM* and *CDH1* was justified by their roles in tumor progression and metastasis, as evidenced in colorectal and gastric cancers (52, 79). This gene list aims to enhance the ability to detect prevalent somatic variants in a small-scale initial panel design.

Thereafter, the gene list is curated in a total of 10 pan-cancer study's patient data from the cBioPortal software (80-82) (release 2024q1) to ensure suitable loci selection. Studies included in the gene set curation are shown in a pie chart (Figure 8) and listed in Appendix 1. Publicly available clinical data is downloaded from the Portal and extracted into a pandas Data Frame. This is followed by converting the "*Mutation Count"* column to a numeric type to facilitate calculations and removing invalid entries to maintain reliability.

We applied a logarithmic transformation to the mutation counts to address data skewness and large-range variability in the violin plot. We finally generated plots visually representing the phenotypic distribution and variants across different cancers to provide a perspective on the desired gene list. Python script that extracts the data into different plots for the selected genes is available in Appendix 2.

The pie chart in Figure 8 displays fractions of all available pan-cancer studies on the cBioPortal for Cancer Genomics website. These studies are specifically selected for analysis based on a list of key genes associated with sporadic cancers. The chart shows the proportions of samples in each study who have at least one variant in any of these 18 genes. The figure is adapted from the cBioPortal on April 2, 2024.

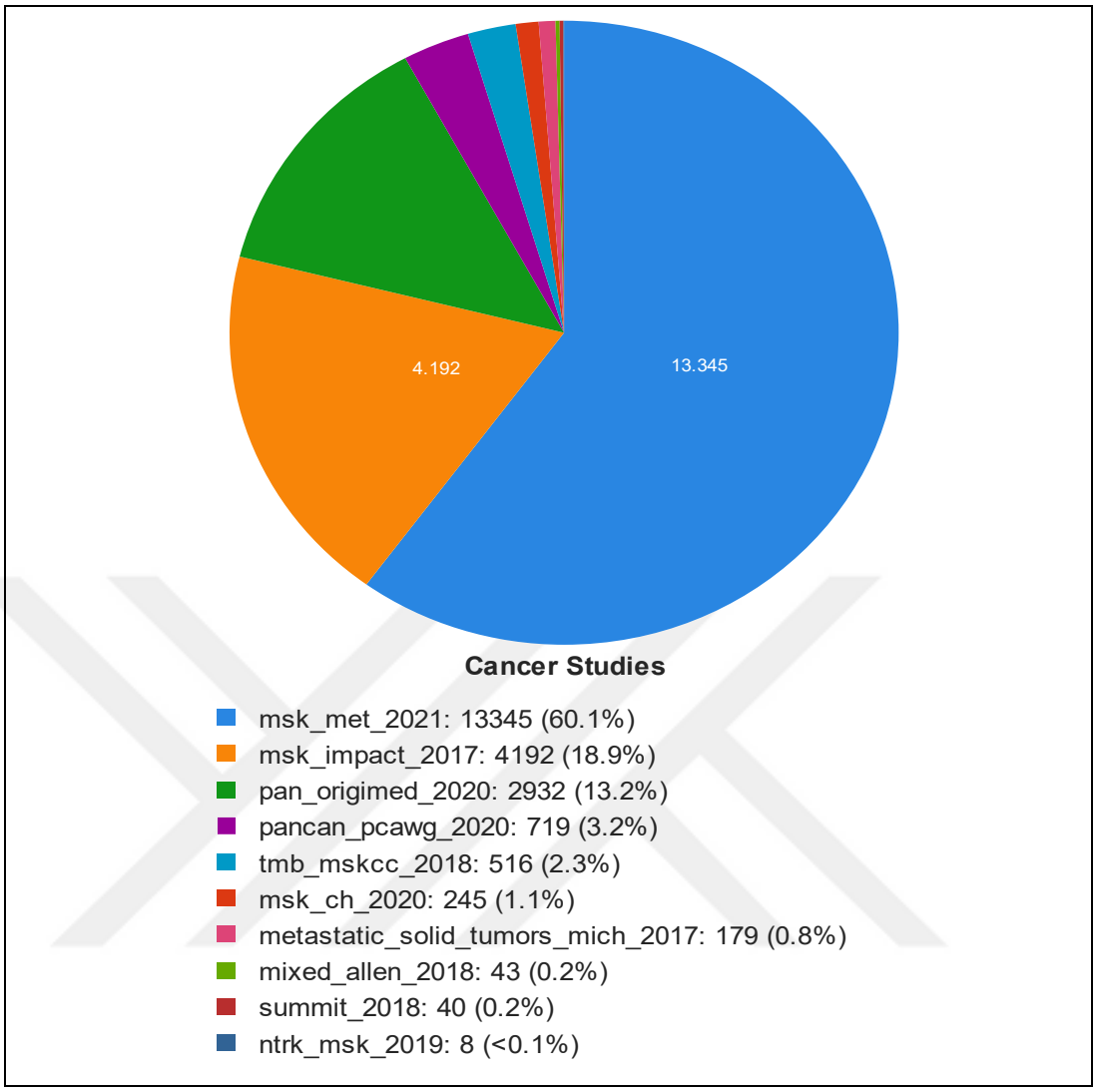

Figure 8. Sample distribution, after filtering the desired genes (Table2) among the pan-cancer studies

Table 2. Desired gene list for the MGPT in alphabetical order (76)

Gene	Full Name	Description
APC	Adenomatous Polyposis Coli	A tumor suppressor gene, with somatic variants often leading to colorectal cancer progression
BRCA1	Breast Cancer Type 1 Susceptibility Protein	Frequently mutated somatically in breast and ovarian cancers, crucial for DNA repair
BRCA2	Breast Cancer Type 2 Susceptibility Protein	Like BRCA1, it plays a key role in DNA repair; variants are common in breast and ovarian cancers
BRIP1	BRCA1 Interacting Protein C-terminal Helicase 1	Involved in DNA repair; somatic variants contribute to ovarian and breast cancer risk
CDH1	Cadherin 1	Encodes E-cadherin, a cell adhesion protein; somatic variants are implicated in gastric and breast cancers
CHEK2	Checkpoint Kinase 2	Involved in cell cycle control, DNA repair, and apoptosis
EPCAM	Epithelial Cellular Adhesion Molecule	Overexpression due to somatic variants can lead to tumorigenesis in colorectal and other cancers
MLH1	MutL Homolog 1	A mismatch repair gene, with somatic variants often found in various sporadic cancers
MSH6	MutS Homolog 6	Part of the mismatch repair system; somatic variants are linked to colorectal and endometrial cancers

Table 2. Desired gene list for the MGPT in alphabetical order (76) (continue)

Gene	Full Name	Description
MUTYH	MUTY Homolog	Involved in DNA repair; somatic variants are associated with an increased risk of colorectal cancer
PALB2	Partner and Localizer of BRCA2	Works with BRCA2 in DNA repair; somatic variants can lead to breast, ovarian, and pancreatic cancers
PMS2	Postmeiotic Segregation Increased 2	A Mismatch repair gene: somatic variants are associated with colorectal and endometrial cancers
PTEN	Phosphatase and Tensin Homolog	A tumor suppressor gene: somatic variants are common in cancers like glioblastoma, prostate, and endometrial
RAD51C	RAD51 Paralog C	Involved in DNA repair and homologous recombination; somatic variants increase ovarian and breast cancer risk
RAD51D	RAD51 Paralog D	Also involved in DNA repair; somatic variants are implicated in ovarian cancer
SMAD4	SMAD Family Member 4	Tumor suppressor in the TGF-beta signaling pathway; somatic variants found in pancreatic and colorectal cancers
STK11	Serine/Threonine Kinase 11	A tumor suppressor gene: somatic variants are linked to increased cancer risk, notably in lung adenocarcinoma
TP53	Tumor protein p53	It is a critical tumor suppressor gene involved in various cancers

3.1.2 Comparative analysis of open-source primer design tools

Here, we performed a three-step evaluation for seven open-source primer design tools, focusing on their compatibility in target enrichment for a comprehensive MGPT. The first evaluation is centered around four principal aspects: algorithmic efficiency, design accuracy, user interface, and consistent *in-vitro* validations. In the primary comparison, we pointed out four primer design tools given their proven efficacy in terms of our principles. A secondary comparison for the remaining tools was conducted using published studies as a source of information to assess their efficiency in addressing primer dimerization, hairpin formation, melting temperature (Tm) consistency, and target specificity. Here, two of the tools were not chosen to work with.

Subsequently, we observed the performance through *hands-on* as a third step of primer design scenarios. We used the remaining two algorithms to provide us primer sequences for the canonical *CHEK2* exons plus twenty bases from each end. This preliminary analysis was meant to mimic the conditions and prevent the challenges present in multiplex PCR, providing insights into the effectiveness of each tool in a real-world context. Hereby, we came up with an algorithm that will print out primer sequences while taking our interests into consideration. Contrarily, we eventually used the other algorithm to seek primers for our uncovered and *hard-to-call* regions. However, it could not generate feasible primers as well (i.e: heptad repeats in *PMS2* gene).

3.1.3 Generating primer sequences

We first identified our target genes as a material after deciding the source algorithm for primer sequences. We then extracted locus information for canonical sequences from the UCSC (The University of California Santa Cruz) database, using the RefSeq (Reference Sequence Database) numbers based on the GRCh38.p14 (38th version and 14th patch of the Genome Reference Consortium Human). We choose *"Exons plus 20 bases on each end"* option in the Table Browser of the UCSC

to include donor and acceptor splice sites (83-86). Thus, a total of 109,718 bases of genomic regions are deployed into BED files -- browser extensible data format (Ensembl, 2024) (87).

Furthermore, we obtained primer sequences from the MPPrimer web server (version 2) (88). In the beginning, we made several adjustments to the primer design to enhance the performance of the PCR. These adjustments include setting the GC content of the primers to be at least 30%, ideally around 50%, and not exceeding 70%. We aimed for an optimal primer melting temperature range of 59.0°C to 63.0°C to prevent polymerase competition. To avoid hairpins and dimers we chose to keep the minimum delta G to -3.5. We designed the primers to avoid any SNPs within nine bases of their 3' ends to reduce the likelihood of incorporating errors from single nucleotide polymorphisms. Additionally, we ensured that the primers would not produce nonspecific amplicons, as an additional purification process is not feasible in multiplex PCR, where amplicon lengths are typically between 200 and 250 bases. Lastly, we displayed coverage as a bar chart to demonstrate the coverage rate throughout the target loci using Python (Appendix 7).

3.1.4 ALPlex: an algorithm for multiplexed primer analysis

Suppose we have two primers, Primer A (5'-AGCT-3') and Primer B (5'-AGCT-3'). At first glance, they are identical and not complementary. However, in PCR, Primer B will incorporate in its reverse complement form to bind to the target DNA, which would be 5'-AGCT-3' for Primer A and 3'-TCGA-5' for Primer B's reverse complement. In this case, the reverse complement of Primer B (5'-AGCT-3') can bind to Primer A, leading to dimerization. Thus, we wanted to analyze each set of primers that will be together in a reaction targeting different loci. A Python-based algorithm is designed to evaluate potentials for primer dimers and hairpins. *ALPlex* stands out as a critical innovation in this study, offering a robust and automated solution that enhances the accuracy and efficiency of multiplex PCR primer design. By automating these assessments within an environment using unique computer functions, the algorithm aims to enhance the primer design efficacy. By automating these assessments within

an environment using unique computer functions, the algorithm aims to enhance the primer design efficacy.

The “reverse_complement” function creates the reverse complement of a DNA sequence, essential for evaluating potential primer-primer interactions and dimer formation. The “check_structure*”* function predicts the likelihood of hairpin structures which can inhibit primer functionality. Another sophisticated component, the *“advanced_dimer_score”* function, quantifies the potential for dimerization by aligning a primer with the reverse complement of another, focusing on the complementarity length and GC content to assess stability. The main function, *“analyze_primers”*, orchestrates the workflow by loading primer sequences from an Excel file, applying the aforementioned calculations, and outputting analyzed data to a new Excel file. This function processes each primer to determine its potential hairpin structures, and calculates dimer scores for all combinations, documenting these findings in a table for easy evaluation.

Once the reverse complement is obtained, the algorithm uses the “*advanced_dimer_score”* function to compute a dimerization score based on possible matches between a primer and the reverse complement of every other primer in the dataset. This function iterates through each base of the first primer and attempts to align it with every base of the reverse complement of the second primer. For each potential complementarity point, the function extends the match as long as the bases are complementary. The “matching” process involves:

- Incrementing a length counter for each matching base pair.
- Increasing a GC count if the matching bases are either G or C, due to the stronger hydrogen bonding of GC pairs compared to AT pairs.

The score for each matching is then calculated as the sum of the complementarity length and the GC count. The rationale behind this scoring method is that longer matches and matches with higher GC content are more likely to result in stable and significant dimer formation. After calculating the dimer scores for all possible matches

between primer pairs, the algorithm records these scores and identifies the maximum score associated with each primer. A high dimer score suggests a high potential for stable dimer formation and is a critical metric for evaluating primer suitability. Primers with high dimer scores may be flagged or excluded from the final selection to avoid potential PCR efficiency problems.

The hairpin check in the PCR primer analysis algorithm is designed to identify potential secondary structures within a primer sequence that might impede its ability to effectively bind to the target DNA sequence. The first step in checking for hairpin structures is to reverse the primer sequence. This is simply flipping the order of nucleotides in the sequence to simulate the folding back mechanism of the strand.

After reversing, the algorithm checks for complementary base pairing within the sequence itself. This involves identifying segments of the reversed sequence that can align and hybridize with other parts of the same sequence. The algorithm essentially looks for matches between bases in the reversed sequence and the original sequence that are characteristic of base pairing (A with T, and G with C). A match means there is the potential to fold back on itself and form a stable hairpin structure. The hairpin check implemented in the function "*check_structure"* investigates whether the reversed sequence is found within the original sequence or vice versa.

3.2 *In-vitro* Validation

3.2.1 DNA sample

- A DNA sample from a healthy donor was provided by the Biobank Unit of Acıbadem MAA University.

3.2.2 Primers

We purchased services from MSM Products and Services Company (www.msm.ist) to have our primers synthesized by Macrogen

(https://www.macrogen.com/). The Macrogen Oligonucleotide Purification Cartridge (MOPC™) is a specialized custom technique used for the purification of oligonucleotides (short DNA or RNA sequences) is applied for all primers.

3.2.3 Chemicals

- Agarose (Merck, Germany)
- Loading dye (Merck, Germany)
- Nuclease-free water (Thermo Fisher, USA)
- 50X TAE Buffer (Sigma, Germany)

3.2.4 Consumables and Laboratory Kits

- Microcentrifuge tube (Isolab, Germany)
- Micropipette (Gilson, USA)
- Parafilm M Bemis Sealing Film
- Pipettes (Thermo Fisher, USA)
- 15 mL and 50 mL falcon tubes (Isolab, Germany)

3.2.5 Devices

- CFX96 Real-Time System, C1000 Touch Thermal Cycler (Biorad, USA)
- ChemiDoc MP Imaging System with universal Hood III (BioRad, USA)
- Electrophoresis instrument (Biorad, USA)
- Refrigerator (+4°C, -20°C, -80°C) (Nuve, Turkiye)

3.2.6 Kits

- A key component of the validation study will be the use of BlasTaq 2X qPCR MasterMix (abmgood, Canada), a reagent designed for accurate and sensitive real-time quantitative PCR (qPCR) analysis.

3.2.7 Quantitative polymerase chain reaction (qPCR)

The efficiency of the primers was evaluated by the qPCR (Figure 9). Primers for PCR were initially synthesized at a concentration of 100 µM. For use in the reactions, these were diluted to 10 µM by combining 10 µL of each primer with 90 µL of nuclease-free water. Subsequently, a 10X primer mixture was prepared by combining 10 µL of each 10 µM primer and diluting to a final volume of 500 µL with nuclease-free water, achieving a concentration of 2 µM for each primer in the mix. The final PCR reaction volume was set to 10 µL, containing 1 µL of 10X primer mix, 5 µL of 2X Master Mix, 1 µL of DNA template (at 100 ng/µL), and 3 µL of nuclease-free water. Since we wanted to keep background noise and costs minimized, we merged only five primer pairs as a multiplex set up.

The PCR reactions were subjected to the following thermal cycling conditions: initial enzyme activation at 95°C for 3 minutes, followed by 40 cycles of denaturation at 95°C for 15 seconds, and annealing/extension at 60°C for 1 minute. A melting curve analysis was performed to assess the specificity of the amplification products, with a gradient from 65°C to 95°C, increasing by 0.2°C every 5 seconds (Figure 9).

We initially tested a subset of five primer pairs from a single primer set specific to the *BRCA1* regions to evaluate the efficiency of these primers in a multiplex PCR reaction. After amplifying the DNA with these five primer pairs, we used the PCR product as a template for further amplification, diluting it at a 1:500 ratio. If distinct amplification bands are observed through gel electrophoresis from this diluted PCR product, it would confirm the presence of our target amplicons in the prior multiplexed reaction.

Moving forward, we plan to increase the number of primers and conduct multiplex amplification on both *BRCA1* and *BRCA2* regions. This will allow us to complete a localized *BRCA* test, serving as a preliminary product to demonstrate our capability for performing comprehensive Multi-Gene Panel Testing (MGPT) in future projects.

3.2.8 Gel electrophoresis

Agarose gel electrophoresis was performed to visualize the PCR products. A 2% agarose gel was prepared by dissolving 2 g of agarose in 100 mL of TAE (Tris-acetate-EDTA) buffer. The mixture was heated in a microwave until fully dissolved and then cooled to 50-70°C before adding ethidium bromide (EtBr) at the manufacturer's recommended concentration. The solution was poured into a gel casting tray with a comb, and the gel was allowed to solidify at room temperature for 20-30 minutes.

The solidified gel was placed in an electrophoresis chamber, which was filled with TAE buffer until the gel was submerged. Samples, prepared by mixing 5 µL of PCR product with 1 µL of 6X loading dye, were loaded into the wells alongside a DNA ladder for molecular weight reference. Electrophoresis was conducted at 100-110 volts for 60-70 minutes. Following the run, the gel was visualized using a UV transilluminator or gel documentation system to capture images of the DNA bands.

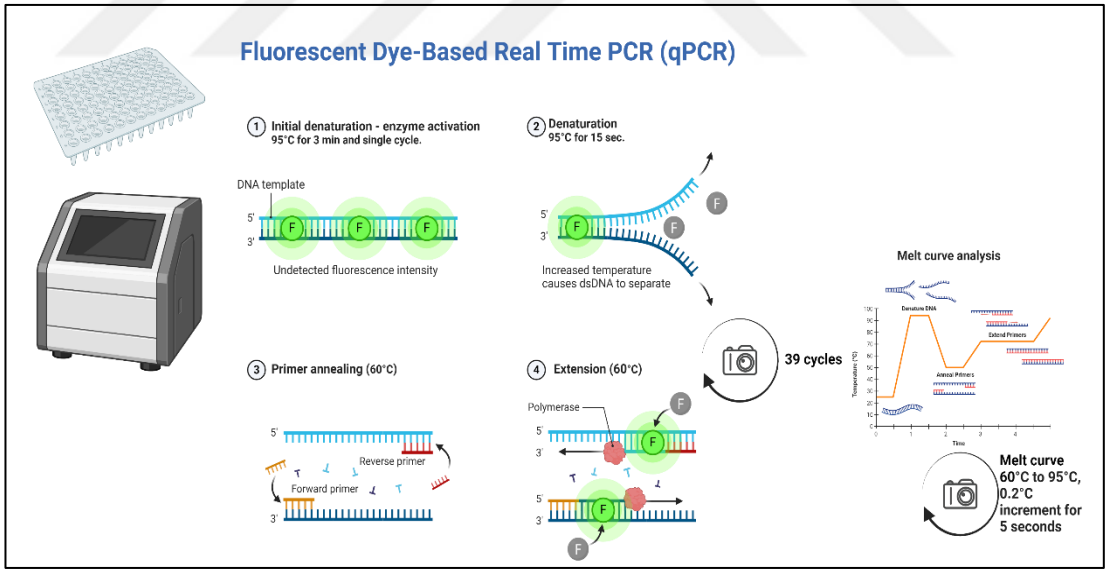

Figure 9. Real time PCR phases (Generated from a Biorender Template)

3.2.9 Non-template control preparation

A non-template control from the same DNA template is prepared through conventional polymerase chain reaction specifically for 156 bp amplicon of housekeeping *GAPDH* gene (Figure 10).

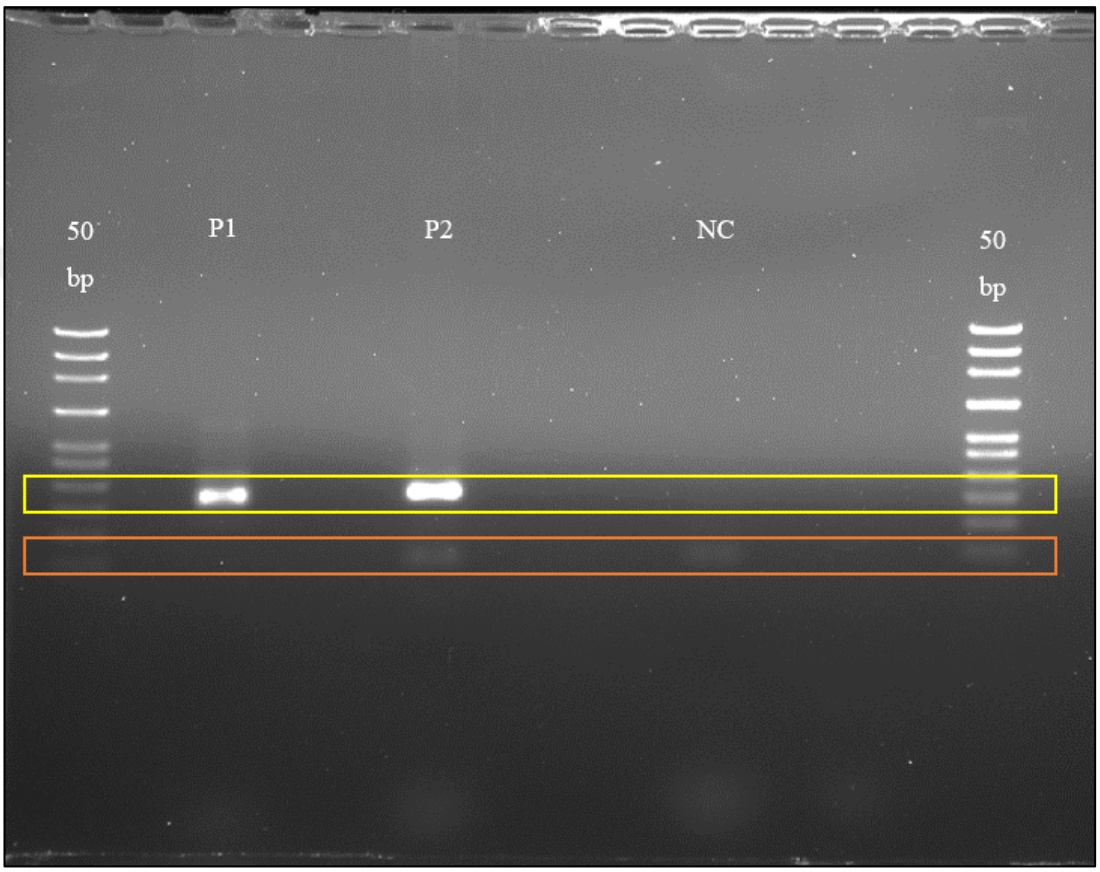

Figure 10. Gel electrophoresis visualization for NTC sample preparation

The first PCR product (P1) has a higher primer concentration, though shows tiny non-specific amplifications. Primer dimers and expected amplicon length are highlighted in orange and yellow, respectively. Therefore, we continued using product 2 by diluting 20 times.

3.3 Variant Calling and Detection Analysis

An MGPT is beneficial if and only if it gives reliable results and understandable reports. Hereby, any patient would benefit from the personalized treatment strategies, as mentioned before. Therefore, we built an automatic computational pipeline to call somatic variants properly from the raw biological data, based on the Genome Analysis Tool Kit (GATK) Best Practices (github.com/akkusalper/mgpt_pipeline). The design of the Multi-Gene Panel Testing (MGPT) is critical for ensuring the precise detection of genetic variants, especially those present at low frequencies. The effectiveness of MGPT is inherently connected to the analysis pipeline and limit of detection (LoD), as the capacity to identify low frequent variants hinges on the sensitivity of the panel. By integrating LoD considerations into the MGPT design, we ensure that the panel can robustly detect clinically significant variants, thereby improving the accuracy and reliability of diagnostic results.

Additionally, a variant LoD algorithm is designed and implemented to the pipeline to give an allele specific mathematical understanding of variant calls (github.com/akkusalper/vLoD). In this part, we employed all computational tasks within the configuration of 64 cores and 256 GB of memory at the Acıbadem Mehmet Ali Aydınlar University Biobank Unit.

3.3.1 Patients

In this part of the study, we benchmarked the variant limit of detection (vLoD) algorithm using 652 MGPT data derived from liquid biopsies of patients who applied to Kocaeli University, School of Medicine, Medical Genetics Department. Patient data is utilized with respect to the allowance from ATADEK-Acıbadem Mehmet Ali Aydınlar University Medical Research Evaluation Board (The project was discussed in the *ATADEK* meeting on May 2, 2024, with reference number 2024/7. It was concluded that the project, with decision number 2024-7/266, is ethically appropriate from a medical perspective).

3.3.2 Analysis pipeline

In this study, we employed a genomic analysis workflow to ensure accuracy and reliability in variant detection from sequencing data (Figure 11). Initially, quality control of raw sequencing reads was performed using FastQC (89) which provides a quick and informative overview of potential issues in the data. Following quality assessment, reads were aligned to the reference genome utilizing BWA-MEM2 (90), a fast and accurate alignment tool that handles sequencing reads up to 1Mb. Aligned reads were then converted from SAM (sequence alignment map) to BAM (binary alignment map) format and sorted using SAMtools (91), which also facilitated subsequent analyses. Duplicate reads, which can arise during sequencing due to PCR amplification artifacts, were identified and marked using GATK's MarkDuplicatesSpark (92). enhancing the accuracy of variant calling by minimizing false positives. To further refine the quality of the sequencing data, base quality score recalibration (BQSR) was conducted using GATK's BaseRecalibrator and ApplyBQSRSpark tools (93).

Variant calling was performed with GATK's Mutect2 (94), specifically designed for detecting somatic point variants in heterogeneous cancer samples. This was followed by stringent variant filtering with GATK's FilterMutectCalls to retain only high-confidence variants, critical for accurate downstream analyses. To predict the detectability of these variants, we processed the filtered variant call format (VCF) files using a custom vLoD script (Figure 11).

Finally, we integrated the detectability status back into the VCF (Appendix 9), providing a dataset that facilitates robust downstream genetic analyses and interpretation, essential for personalized medicine applications and genetic research. This methodological approach ensures data quality and reliability in identifying clinically relevant genetic variants. Tumor Mutational Burden (TMB) can be calculated by counting the number of non-synonymous variants present in any VCF file (github.com/akkusalper/tmb).

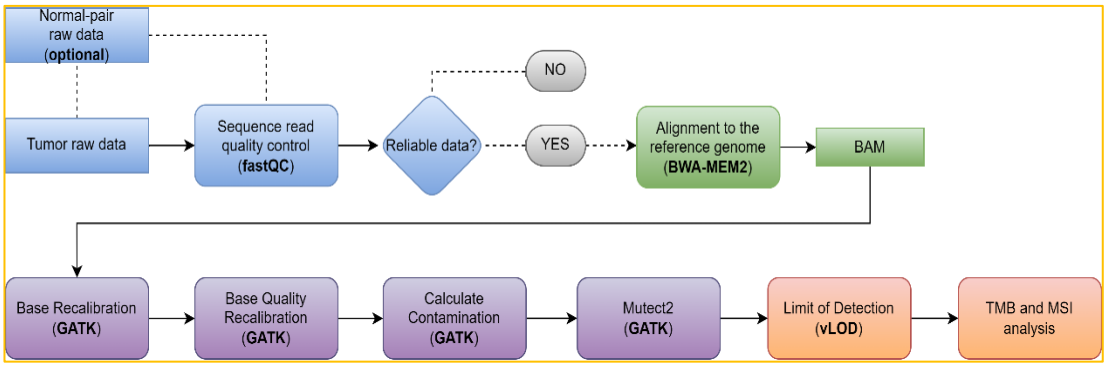

Figure 11. Analysis pipeline flow chart

Flow chart shows computations in the Analysis Pipeline. It can work both with and without a healthy data (so-called normal) pair. Steps are adapted from the Genome Analysis Tool Kit's Best Practices Workflows for calling somatic variants (https://shorturl.at/b9Mi2).

3.3.3 Variant limit of detection tool (vLoD)

In the context of somatic variant detection, traditional methods for determining the limit of detection (LoD) often fall short due to the unique challenges posed by tumor heterogeneity, low-frequency variants, and sequencing errors. Somatic variants, which occur in a subset of cells within a tissue, often exist at very low variant allele fractions (VAFs), making them much harder to distinguish from background noise or sequencing artifacts. Conventional approaches are typically optimized for germline variants, which are present in nearly 50% or 100% of the reads and are therefore easier to detect. However, these methods may lack the sensitivity required to accurately call low-frequency somatic variants, especially when the variant exists in a small proportion of the sample. As a result, false negatives can occur, meaning potentially critical variants are overlooked.

Somatic variant callers, including often struggle to distinguish true low-frequency somatic variants from sequencing errors. Sequencing technologies, particularly those based on next-generation sequencing (NGS), inherently produce errors, with error rates typically ranging from 0.1% to 1% per base, depending on the platform. This

presents a challenge for detecting variants at low variant allele frequencies (VAFs), where the signal of a true variant may be indistinguishable from noise. In tumor samples where the variant frequency might be lower than the sequencing error rate, leading to false positives. This issue is compounded when analyzing low-input or degraded samples (like formalin-fixed paraffin-embedded tissue).

Many somatic variant callers, including rely on a matched normal sample to distinguish somatic variants from germline variants. The absence of a matched normal sample can significantly reduce the accuracy of variant calls, as the caller must rely on population frequency databases to filter germline variants. Without a matched normal, somatic variant callers may incorrectly classify germline variants as somatic, especially if the variant is rare in the population. Conversely, they may also fail to detect somatic variants that are indistinguishable from germline polymorphisms based solely on frequency data. This limitation is especially problematic for rare or private variants not well represented in public databases (90-96).

To address these limitations, more advanced bioinformatics tools are needed ones that can account for sequencing depth, error rates, and the low VAFs typical of somatic variants. This is where our algorithm becomes essential, offering a more robust and sensitive approach to determine the minimum detectable variant allele frequency, providing improved accuracy in identifying true somatic variants.

In this step, our bioinformatics approach incorporates an algorithm designed to calculate the minimum variant allele frequency detectable in a VCF file. Upon executing a Python script that calculates the detection limit using the provided approach, a mathematical expression of detectability for each allele is derived through different functions.

The logarithm of odds (LOD) score, accounting for sequencing error, is calculated based on the probability of a variant being a true positive or a false positive. Initially, the algorithm extracts variant details (chromosome, position, reference, and alternate alleles) from the VCF file, then determines the coverage area and read count through

the BAM file. These values are crucial in calculating the Variant Allele Fraction (VAF), which significantly influences the LOD score. Users can adjust the code parameters to optimize detection sensitivity and specificity. The resulting file includes the LOD score, coverage, and variant read count for each detectable variant.

The LOD score scales the probability of a variant's presence. With default parameters, the likelihood of a false positive call is one in a thousand at 3 LOD points. This range can serve as a detectability paradigm for variant calls. Essentially, this score represents the absence or presence of a variant as a binary event, calculated using the formula in Equation 1 (github.com/akkusalper/vLoD). A LOD score of 2.5 means that the odds of the data being observed if the variant is present are $10^{2.5}$ times higher than if the variant is absent. This translates to odds of about 316:1 in favor of the variant being present (Figure 12).

By implementing the algorithm within a parallel processing framework, we significantly reduce processing time and enhance the scalability of our approach. Future work will involve testing the script's performance and exploring further optimization potential, ultimately leading to the final version. Our computational technique will provide reporting capabilities for variations based solely on tumor data without the comparison to a *"healthy data"* counterpart (Figure 13).

- VAF: The fraction or percentage of sequencing reads that support the alternate allele at a given genomic position.
- True positive rate (p_{tp}): The probability that the variant is genuinely present (i.e., it's not due to random sequencing errors).
- False positive rate (p_{fp}): The probability that the alternate allele is incorrectly identified as a variant when it's not genuinely present.
- Sequencing error probability (p_{se}): The likelihood of observing an alternate allele due to inherent sequencing errors.
- The log-odds score equation in the next page (LOD score) then combines these parameters to gauge the confidence in the presence of an alternate allele.

$$\text{LOD score} = log_{10} \frac{(p_{tp}\ x\ VAF)}{(1 - VAF)\ x\ p_{se} + VAF\ x\ p_{fp}}$$

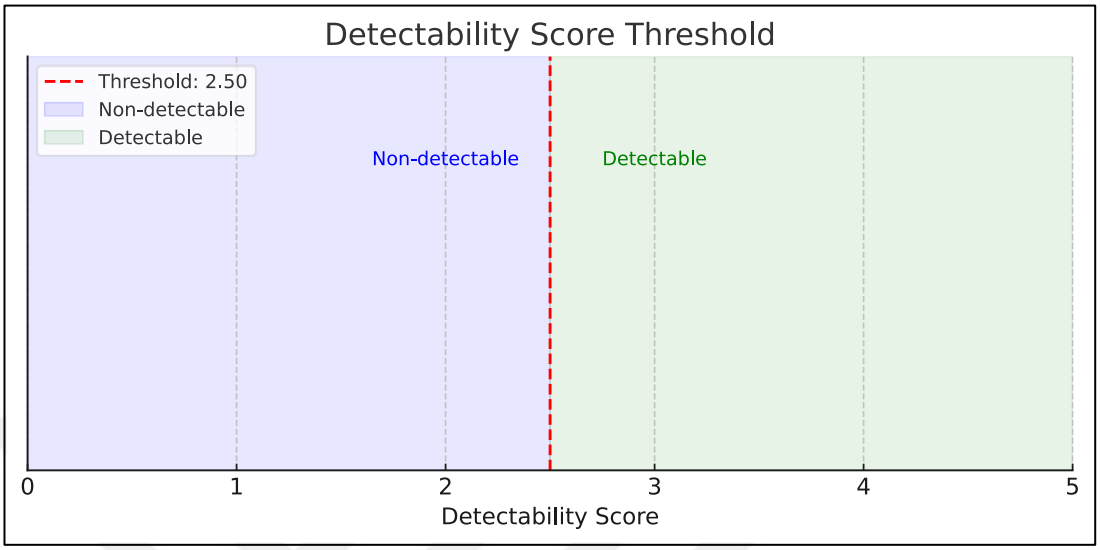

Figure 12. Detectability score threshold.

This LOD scoring helps in distinguishing true genetic variants from sequencing errors or random noise. A higher LOD score indicates stronger evidence that a variant is truly present. In clinical genetics, an LOD score of +3 (or equivalently, a likelihood ratio of 1000:1) is often considered strong evidence for the presence of a linkage, while an LOD score of -2 (or a likelihood ratio of 1:100) suggests strong evidence against it. The exact value of the odds ratio when the LOD score is 2.5 is approximately 316.23:1. This means that the odds of the data being observed under the hypothesis that the genetic variant is present are about 316 times greater than the odds of the data being observed under the hypothesis that the variant is absent.

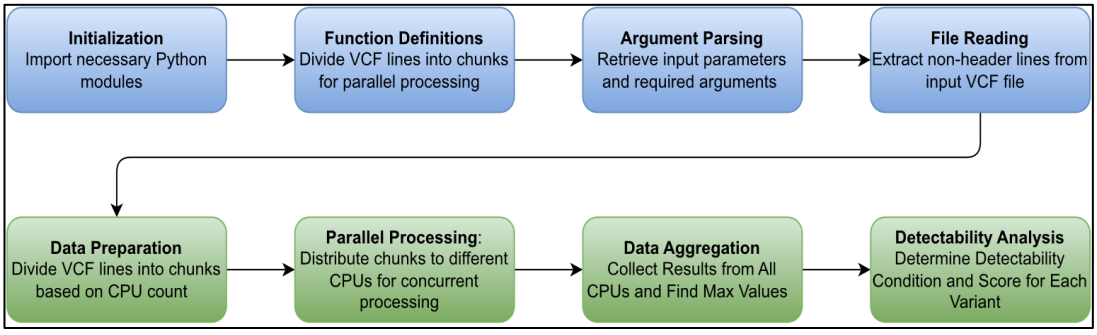

Figure 13. vLoD processing flow and algorithm design

This tool utilizes a sophisticated algorithm to analyze variant call format (VCF) and BAM files. It calculates log-odds scores by employing binomial probabilities to distinguish between true positives and false positives, incorporating the variant allele fraction (VAF). The algorithm sets a minimum VAF threshold, ensuring variants are detected with confidence above a predefined error rate (default set at 0.0001).

3.3.3.1 Benchmark of vLoD

The vLoD tool was tested for benchmarking purposes on 652 liquid biopsy data, encompassing a MGPT of 20 genes. Statistical validation was achieved through Fisher's exact test, applied at both variant and gene levels, with a significance threshold of $p<0.05$. This testing confirmed the correlation between the predicted and actual detections of variants, affirming the tool's accuracy (95) (Figure 14).

We gathered genomic data, preprocessed it, and assessed the tool's performance using key parameters. These variables assessed sensitivity, specificity, precision, to define the algorithm's reliability through false and true positives. We also conducted correlation analysis and visualized performance using Receiver Operating Characteristic (ROC).

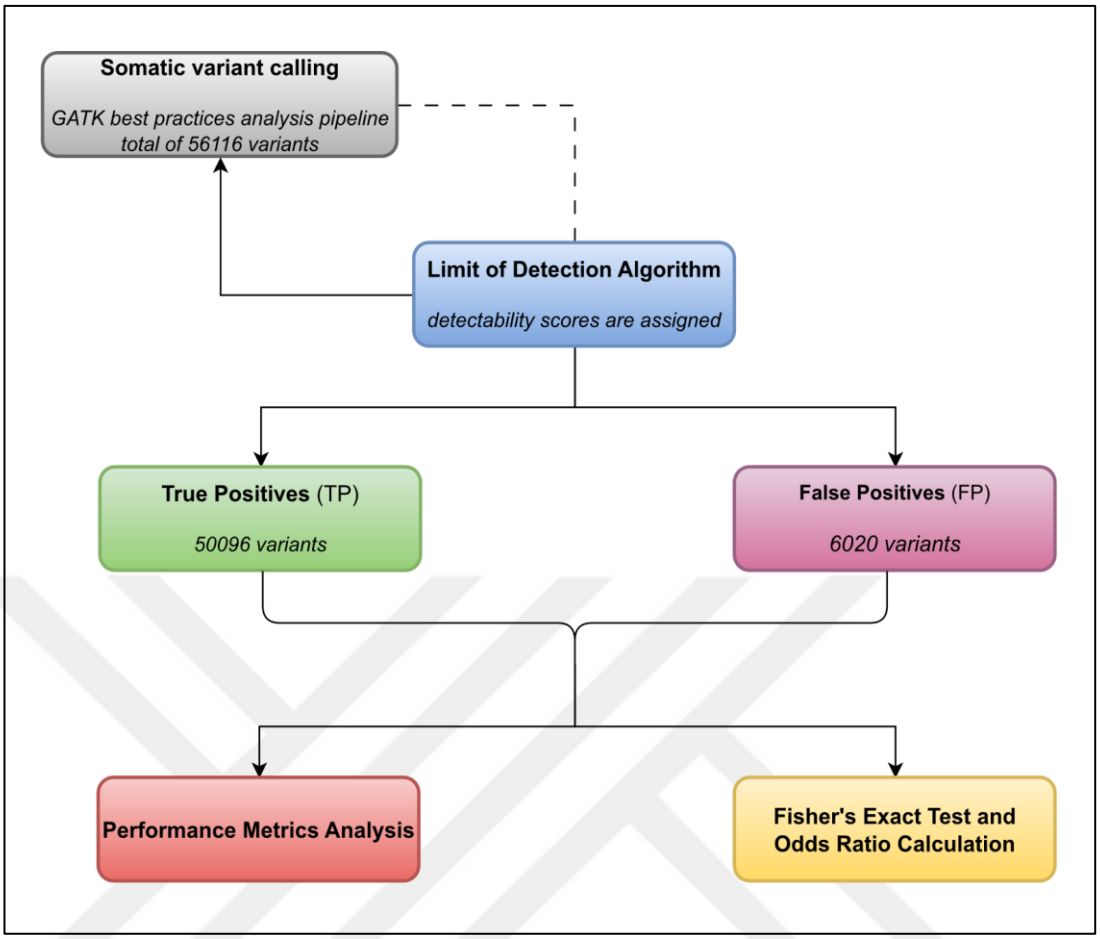

Figure 14. Benchmarking methodology

4 RESULTS

4.1 Primer Design

Here, we designed a total of 690 primers that cover over 94% of the 109,718 bases of the target loci. This part exhibits a virtual, and small-scale MGPT prototype that is theoretically effective based on *in-silico* evaluations. Hypothetically, we will be able to work with 3 samples in a single 96-wellplate multiplex PCR run.

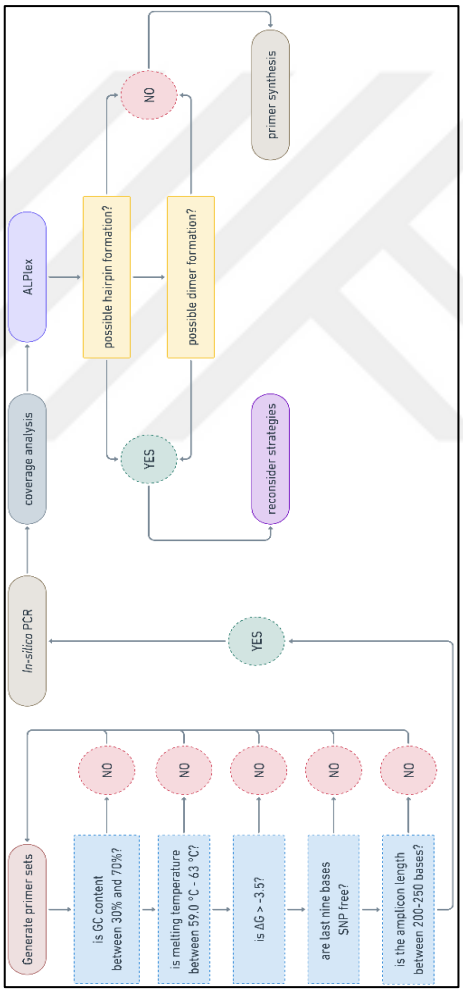

Figure 15. Our *in-silico* workflow for designing and validating primers

4.1.1 Cancer-related gene list for the initial MGPT

First, we observed that our gene list outlines different cancers including breast colorectal, and pancreatic carcinomas (Figure 16). After the curation, numbers of variants seen versus the numbers of volunteers among the curated patient data is displayed at Figure 17A. Furthermore, the age distribution in the pan-cancer studies indicates late onset diagnosis (Figure 17B). Additionally, the plot for the numbers of variants observed in our gene list shows an uneven distribution (Figure 18).

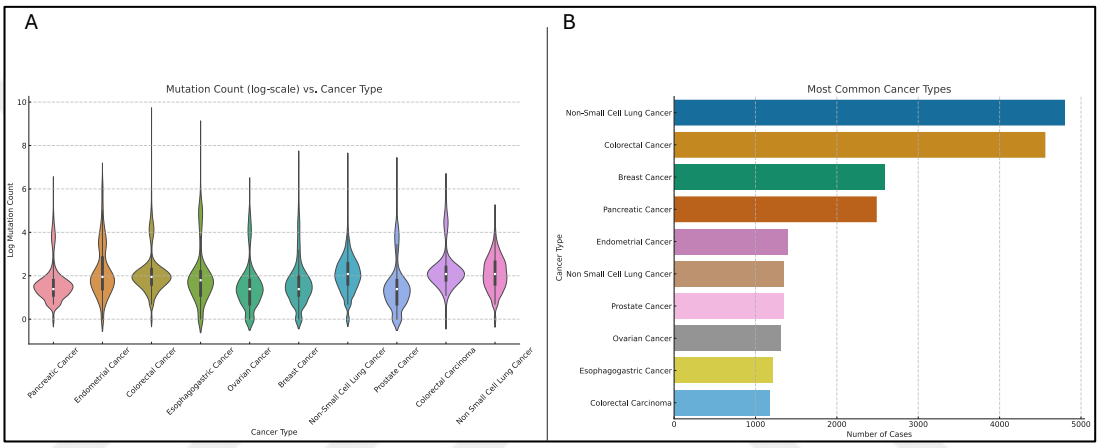

Figure 16. Descriptive statistics for curated data

A) Number of cases are demonstrated in the X-axis through bars, while cancer types are sorted from the highest occurrence to the lowest in the Y-axis.
B) Mutation counts in logarithmic scale shown as a Violin plot.
Cancers with more than 100 occurrences are extracted. These figures are prepared in Python.

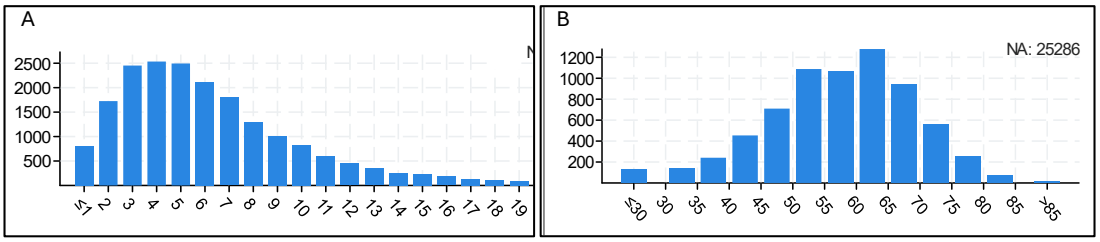

Figure 17. Numbers of variants and diagnosis age distribution among samples

Plot A indicates numbers of variants seen versus the numbers of volunteers among the curated patient data. According to plot B, the chosen genes are highly mutated through late onset diagnosis. Plots are adapted from the cBioPortal in April 2024. (NA: not available samples).

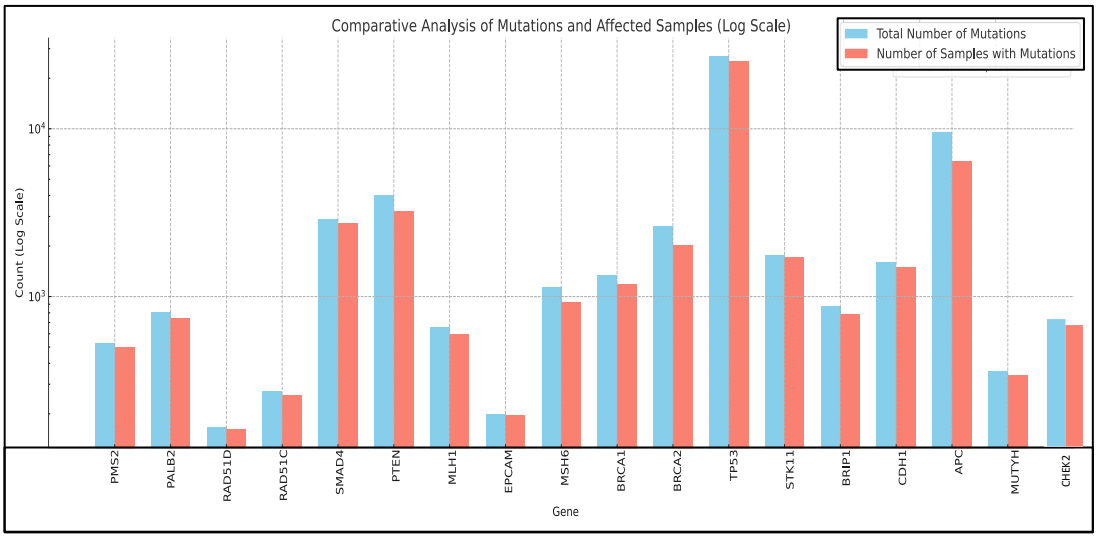

Figure 18. Comparative plot for the numbers of variants observed in our gene list

This figure generated using Python displays the mutation accumulation throughout each gene to project the potential patients that would benefit from this MGPT design. The chart presents a comparative analysis of variants and affected volunteers for several genes, commonly associated with cancer or hereditary diseases, with the y-axis on a log scale. It shows the total number of variants (light blue) and the number of volunteers with variants (pink) in each gene, allowing for insights into the prevalence of variants and their impact on patient populations. Key genes like *TP53*, *PTEN*, and *BRCA1/2* display high mutation frequencies and many affected samples, aligning with their known roles in cancer development. In contrast, other genes like *MSH6* and *EPCAM* show fewer variants and affected samples, suggesting they are less frequently mutated in the curated data. This data highlights important genes for further study and potential therapeutic targeting, particularly in the context of precision medicine where treatments can be tailored to a patient's genetic profile.

4.1.2 Comparative analysis of open-source primer design tools

We applied a three-step evaluation to a total of seven open-source primer design tools to get the optimum primer sequences in the process of multiplex PCR primer design (Table 3). As a disclaimer, a final decision was made for the most favorable

tool for given compatibility requirements, including the complexity of the target sequences and span.

Table 3. The list of studies included in the comparative analysis of open-source primer design tools

Primer Design Tool	Description
Multiplex Primer Design – MPD	Multiplex primer design for next-generation targeted sequencing (96)
PrimerJinn	A tool for rationally designing multiplex PCR primer sets for amplicon sequencing and performing *in silico* PCR (97)
PMP Primer Design	A tool to automatically design multiplex PCR primer pairs for specific targets using diverse templates (98)
HiPlex2	A simple and robust approach to targeted sequencing-based genetic screening (99)
openPrimeR	An R-based tool for evaluating and designing multiplex PCR primers (100)
MPPrimer	A program for reliable multiplex PCR primer design (88)
NGS PrimerPlex	High-throughput primer design for multiplex polymerase chain reactions (101)

Tools eliminated in the primary compatibility assessment, Multiplex Primer Design – MPD, PrimerJinn, and PMP Primer Design, are highlighted in yellow; tools eliminated in the secondary compatibility assessment, HiPlex2 and openPrimeR, are highlighted in blue; tools evaluated by hands-on-trial, MPPrimer and NGS PrimerPlex, are highlighted in green.

4.1.2.1 Primary compatibility assessment

We assessed seven primer design tools based on various parameters to determine their compatibility with our objectives for multiplex PCR in targeted loci. This evaluation was based on a superficial review of the existing literature. Out of the seven tools, four (openPrimeR, MPprimer, NGS PrimerPlex, and HiPlex2) were critically assessed for their suitability. The remaining three tools were eliminated due to

significant limitations in their application or validation. (Table 4). Column titles have been abbreviated as follows: 3-D Complementarity (3-D Comp.), Customization Options (Customization), Database Usage (DB Usage), Documentation & Support (Docs & Support), G-C content (GC%), Hard to Sequence Regions (Hard to Seq. Regions), In Vitro Validation Span (In Vitro), Melting Temperature (Melting Temp), Primer Specificity (Primer Spec.), Repetitions (Reps), and User-Friendly Interface (User Interface).

4.1.2.2 Secondary compatibility assessment

Although openPrimeR was initially considered a strong candidate for our study, its absence of integrated sequence or variant databases and a limited number of *in-vitro* studies ultimately led to its exclusion. Similarly, HiPlex was not selected due to its inefficiencies in primer design relative to the specific needs of our study. The comparative assessment of various primer design tools is detailed in Table 5, highlighting the distinct strengths and weaknesses of each. While all these tools may perform adequately under different conditions or for other types of studies, only two (MPPrimer and NGS Primer Plex) out of the seven tools evaluated were able to meet our criteria and provide suitable primer sequences for our project.

4.1.2.3 Hands-on-trial compatibility assessment

While both tools utilize algorithms that incorporate primer3 libraries, MPprimer (101) was selected over the NGS PrimerPlex (88) tool due to its lower potential for dimerization and greater ease of use. Despite the primer sequences from the NGS PrimerPlex tool appearing suitable, there was a significant presence of repetitive sequences within each primer, as detailed in Appendix 3. This issue influenced the decision to opt for MPprimer, which better met our requirements for efficient primer design. We assessed various parameters to determine each designer's compatibility with our objectives, drawing from a superficial review of the existing literature.

Table 4. Primary compatibility assessment results

Tool	3-D Comp.	Customization	DB Usage	Docs & Support	GC%	Hard to Seq. Regions	*In Vitro*	Melting Temp.	Primer Spec.	Reps	User Interface
openPrimeR	+	-	+	-	+	+	-	+	+	+	+
MPprimer	+	-	+	+	+	-	+	+	+	-	+
NGS PrimerPlex	-	+	+	+	-	+	+	-	+	-	-
HiPlex2	-	-	+	-	+	-	+	-	+	-	-
MPD	+	+	-	+	-	-	-	-	-	-	-
PrimerJinn	+	+	-	-	+	+	-	+	+	-	-
PMP	+	-	+	-	+	-	-	+	+	-	-

Table 5. Secondary compatibility assessment results

Criteria	openPrimeR	MPprimer	NGS PrimerPlex	HiPlex2
Primer Specificity	Non-specific binding potential is not present, with smaller inputs	Little to no potential of non-specific amplification	Little to no potential of non-specific amplification	Neutralizes by re-grouping primer sets
User-Friendly Interface	User-friendly command-line usage	User-friendly interface	Complex, command-line usage	It has no server to work with
Documentation & Support	Limited documentation and support	Extensive documentation and support	Extensive documentation and support	Extensive documentation and support
Customization Options	Limited customization options	Limited customization options	Various customization options	Limited customization options
***In Vitro* Validation Span**	Limited: used only for immune receptors and pathogens	Extensive: specificity, primer dimerization	Moderate: tested on bacterial and human samples	Moderate: requiring extensive optimization of reaction conditions
G-C content	Can be optimized	Can be optimized	Lots of high-GC primers	Suboptimal GC content
Repetitions in the Primer Sequence	Primers are well evaluated with smaller inputs.	It is inadequate, but tries to eliminate the repetitive sequences	It is inadequate, but tries to eliminate the repetitive sequences	It is inadequate, but tries to eliminate the repetitive sequences
Hard to sequence regions	Designed to eliminate repetitive elements	It is inadequate, but tries to cover the target completely	Excellent, but causes potential dimerization	Suitable for small-to-medium target range.
3-D complementarity	Seeks to eliminate potential secondary structures	Seeks to eliminate potential secondary structures	Moderate: It is inadequate for our study	Potential primer dimers
Melting Temperature	Ensure that the melting temperature is within an appropriate range	Ensures that the melting temperature is within an appropriate range	The difference in melting temperatures is greater than expected	It is inadequate for our study
Database Usage	Integration to different databases	NCBI and UCSC are integrated	BLAST and dbSNP (GRCh 37)	BLAST

4.1.3 Adapting primer sequences

Stated evaluations resulted that only feasible primers would be designed through MPPrimer in our panel design. Primer sequences are adapted from the MPPrimer (version 2) (88). Currently, 94.29% of the target loci are covered and prepared for amplification, as demonstrated in Figure 19. All the loci, except *PMS2*, exhibit a desirable expected target enrichment rate of 85%. Even though the demonstration is made by gene names, target loci exist of their exons along with 20 bases from each end. These results lay the foundation for implementing a local Multi-gene Panel Test (MGPT) as outlined in this study. We also displayed the lengths of primers in nucleotide bases on the y-axis and the number of primers corresponding to each length (See Figure 20).

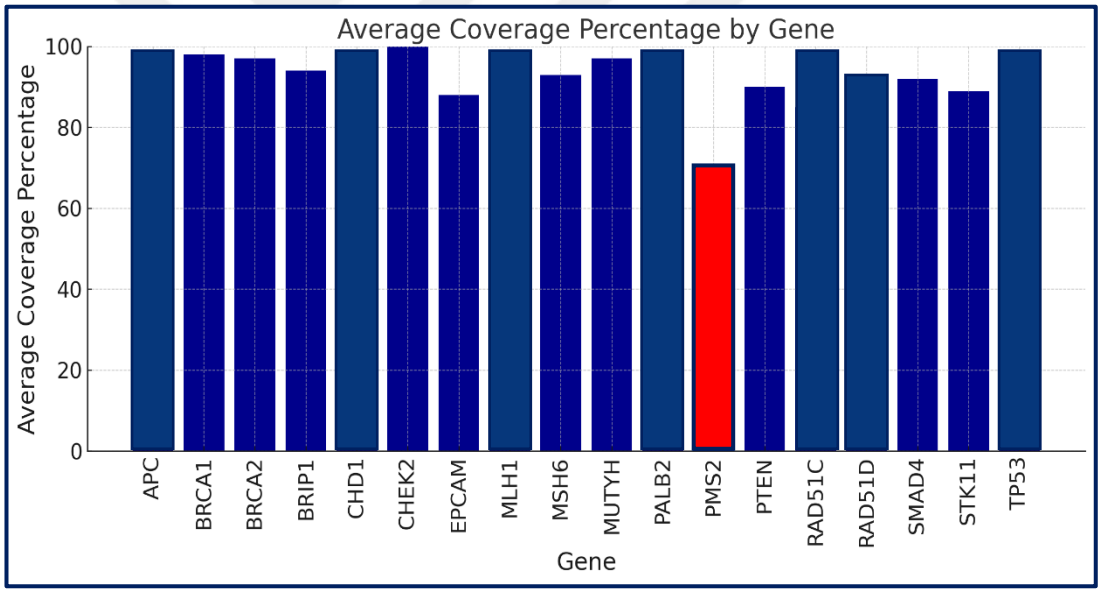

Figure 19. Average theoretical expected coverage percentage by gene regions

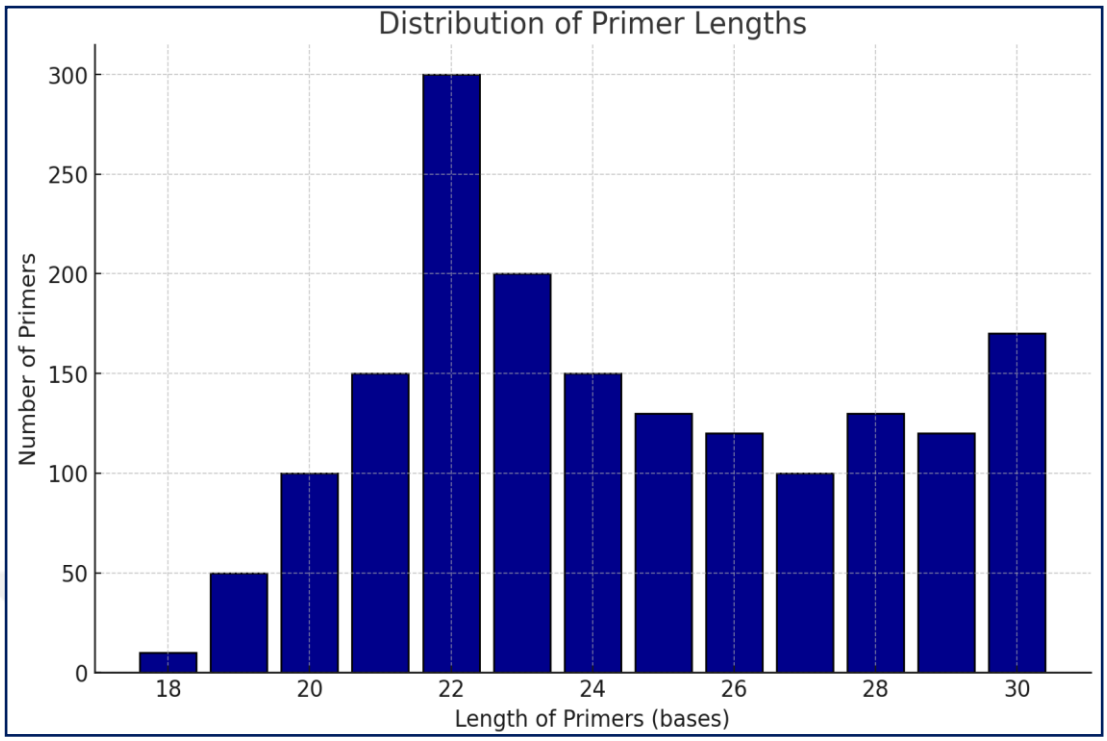

Figure 20. Distribution of primer lengths by bases

4.1.4 ALPlex: an algorithm for multiplexed primer analysis

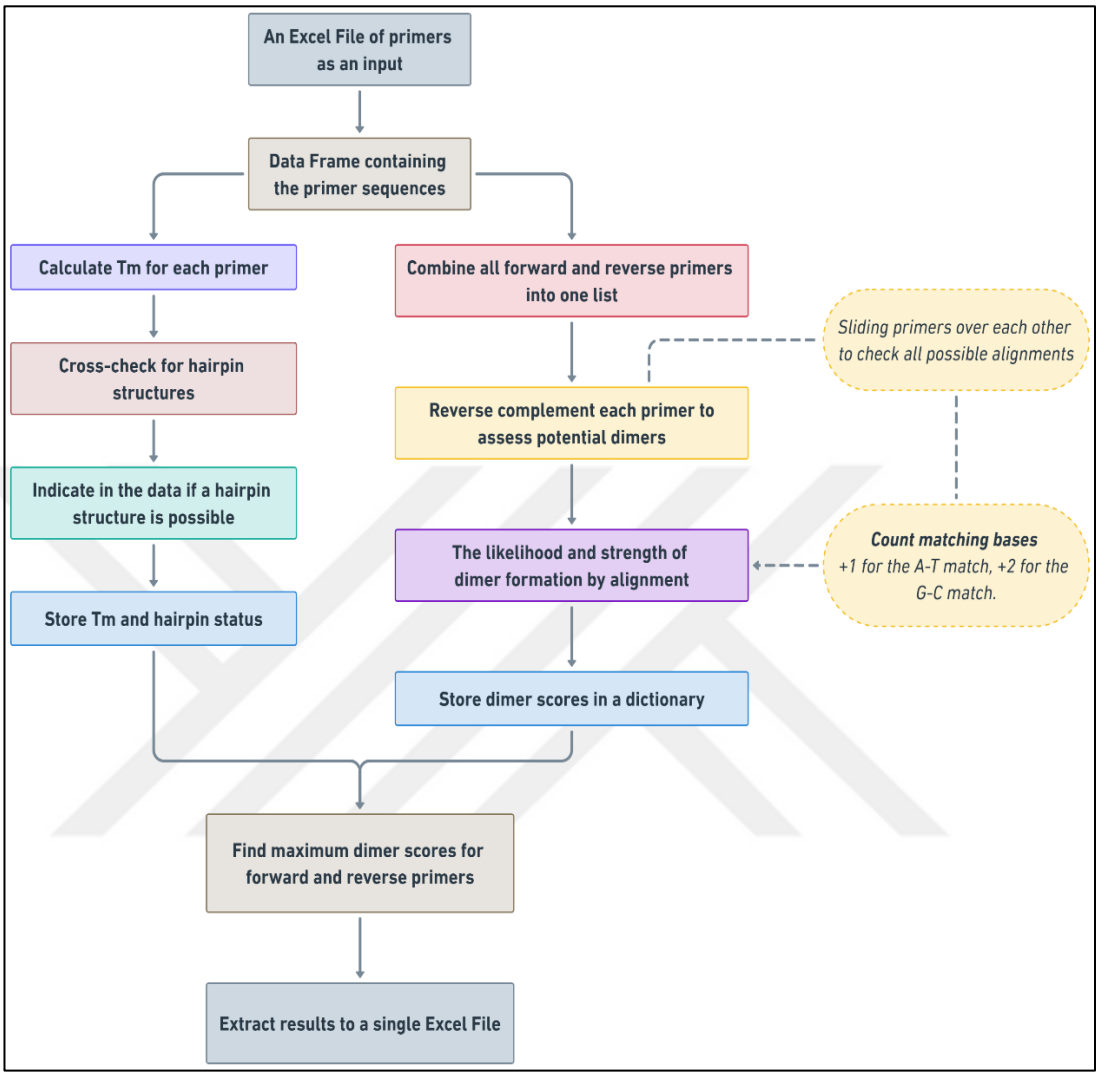

Figure 21. ALPlex (an algorithm for multiplexed primer analysis) set up

In this section, we present ALPlex, a custom-built algorithm specifically designed to perform comprehensive primer analysis for multiplex PCR applications (Figure 21). ALPlex was developed to automate the detection and evaluation of potential primer dimerization and hairpin formations, which are common issues in multiplex PCR that can compromise the specificity and efficiency of amplification. This novel algorithm integrates advanced methodologies, utilizing primer sequences adapted from MPPrimer (version 2) (88), and introduces a new scoring system that ranks primers based on their structural integrity and potential for interaction. By prioritizing primers with minimized risks of non-specific interactions, ALPlex ensures that selected primer

sets are both highly efficient and robust, reducing amplification errors and optimizing multiplex PCR performance.

4.1.4.1 Primer dimer analysis

ALPlex includes an innovative method for assessing primer dimerization potential, addressing a significant challenge in PCR primer design. The algorithm begins by generating the reverse complement of the second primer in a primer pair, enabling the evaluation of complementarity between the two sequences. ALPlex focuses on the degree of base complementarity between primers, reflecting their biochemical interaction potential.

Using nested loops, the algorithm compares each base in the first primer with each base in the reverse-complemented second primer. For each pair of complementary bases, the algorithm extends the comparison to identify longer regions of complementarity. The scoring system incorporates both the length of these complementary regions and the biochemical stability of the base pairs.

Specifically, the scoring function assigns one point for each complementary base pair, with an additional point awarded for guanine-cytosine (GC) pairs due to their higher stability, as GC pairs form three hydrogen bonds compared to the two formed by adenine-thymine (AT) pairs. This allows ALPlex to assess not only the extent of complementarity but also the thermodynamic strength of the interaction.

The algorithm retains the highest observed score for each primer pair, denoted as the "*maximum dimer score*," which represents the strongest potential for dimer formation. By considering both the number of complementary bases and their chemical stability, the scoring system provides a nuanced and robust evaluation of primer dimerization.

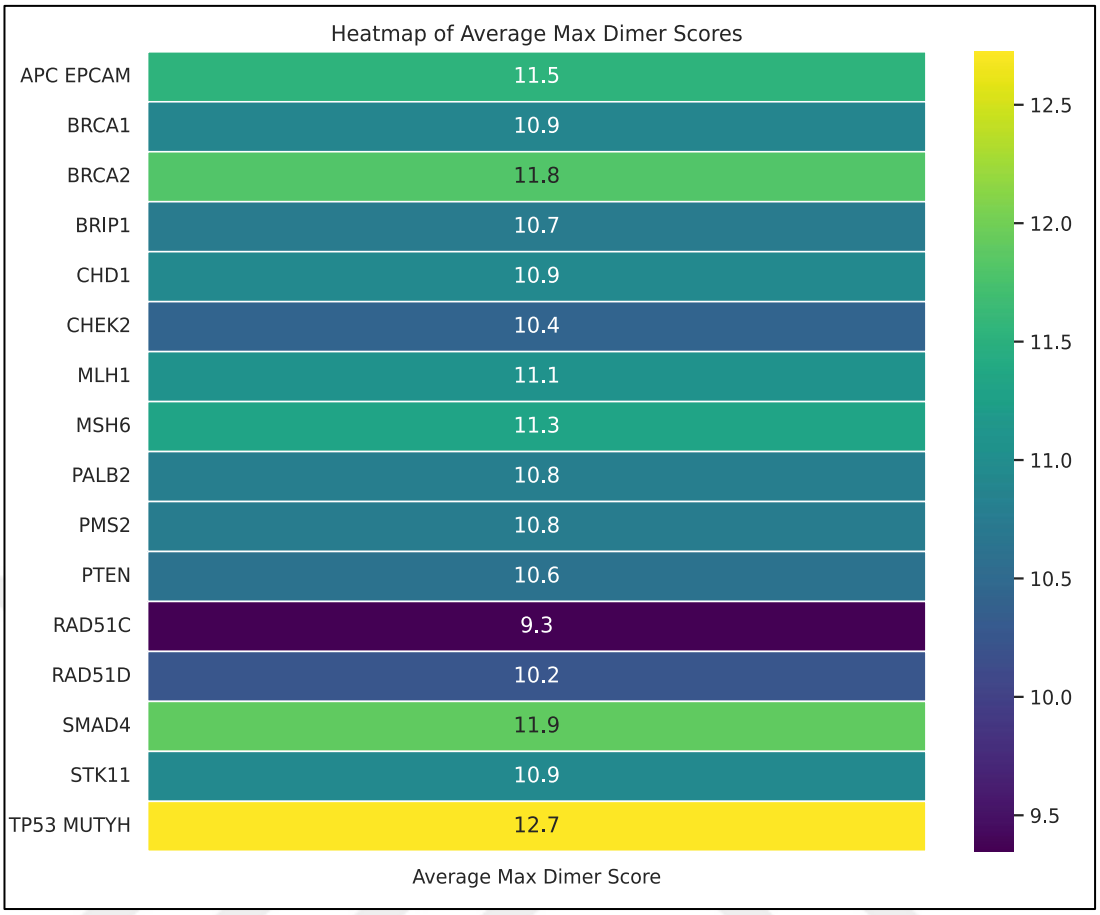

Figure 22. The average maximum dimer scores throughout the primer sets.

Primer sets are denoted on the left as the gene regions that they bind. Primer sets amplifying gene regions are stated as gene names, primer sets which run for multiple gene regions at the same time are stated together. i.e.: “TP53 MUTYH” is for *TP53* and *MUTYH* gene regions (exons +/-20 bases from each end, explained in section 4.1.3).

In summary, ALPlex provides a powerful and novel approach to primer analysis for multiplex PCR. By automating the detection of primer dimers and hairpins and introducing a complementarity-based scoring system, the algorithm ensures that selected primers are both specific and robust, significantly reducing the risk of non-specific amplifications. The results of the analysis, as visualized in the heatmap (Figure 22), provide an intuitive and detailed overview of the dimerization potential across the primer sets, allowing researchers to make informed decisions when designing primers for complex genomic applications (Appendix 5-8).

4.1.4.2 Hairpin analysis

For hairpin estimates, our algorithm plays a critical role in evaluating the potential for hairpin formation in PCR primers. The function initiates by reversing the primer sequence, creating a scenario where the natural tendency of the sequence to fold back onto itself can be evaluated. It then checks for the presence of the reversed sequence within the original sequence or vice versa. The function does not specify a particular number of bases that need to align to qualify as a potential hairpin. Instead, it checks for any instance where the reversed sequence can be found within the original sequence.

If the function returns "No," it suggests that there is no significant portion of the sequence that is complementary to another segment in the reverse orientation. Thus, this implies a low risk of hairpin formation under typical PCR conditions. A return value of "Yes" indicates that a portion of the primer is complementary to another segment, suggesting potential for hairpin structure formation. This method did not yield any potential hairpins for our primer sets. We deliberately used a made-up sequence to get a *"Yes"* on Table 6. Palindromic sequences are straightforward examples where the sequence and its reverse complement are identical, leading to potential hairpin structures. In PCR primer design, such sequences are generally avoided because they can easily fold back on themselves, forming stable hairpin structures that can interfere with proper annealing to the target DNA.

Table 6. An example output from the *in-silico* downstream analysis for primer sequences

ForwardPrimer (Fp)	ReversePrimer (Rp)	Fp Hairpin	Rp Hairpin	Max Dimer Score (Fp)	Max Dimer Score (Rp)
AGCTCGA	ATTG***CGA	Yes	No	6	6
ACTGG***ACC	GAGT***AGG	No	No	9	8

A made-up primer sequence (5' – 3') is used to obtain a "yes" for the hairpin evaluation; highlighted in green.

4.1.5 MGPT design

The current Multiple Gene Panel Test (MGPT) prototype, which includes 32 distinct primer sets, has been employed to facilitate the simultaneous amplification of multiple samples in a single PCR run. As a target enrichment strategy, this setup (Table 7) is particularly might be useful for high-throughput genetic analysis, allowing for rapid processing of samples Excluding negative controls, it is technically designed for 3 samples simultaneously as highlighted in different colors out casting 32 amplicons in Figure 23, as a prototype demonstration.

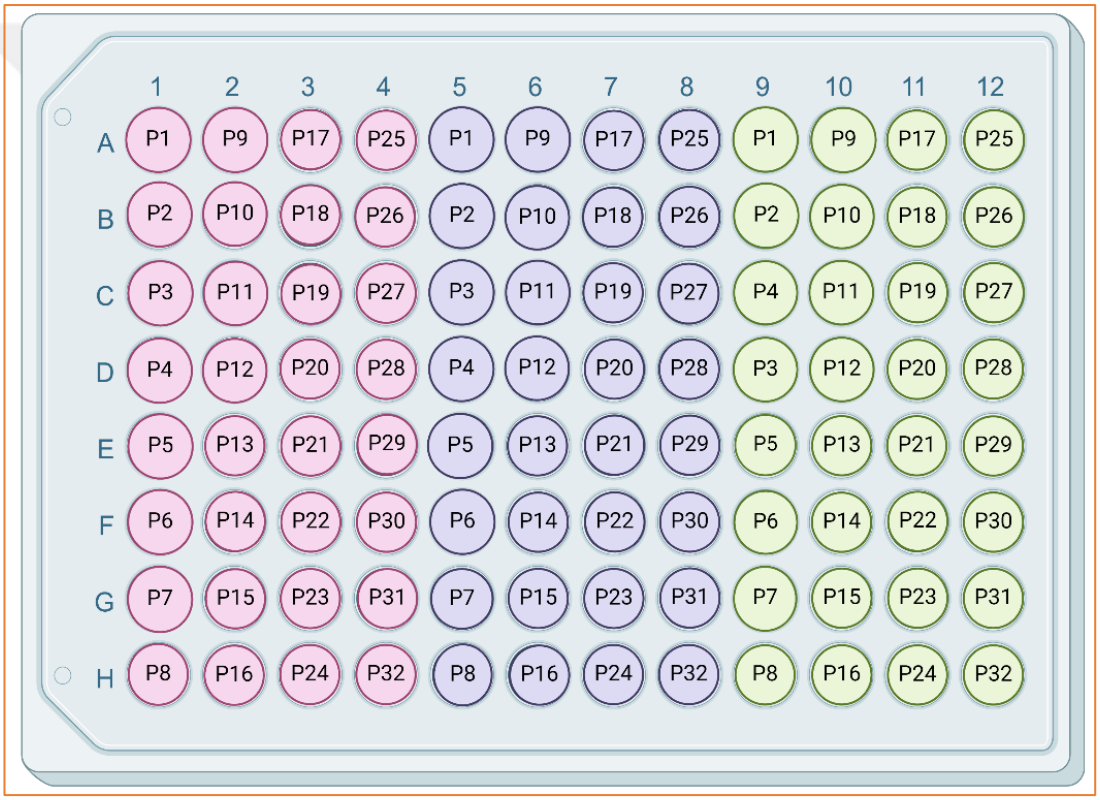

Figure 23. Initial MGPT design demonstration on a 96-well plate.

A 96-well plate is modeled. For practical implementation, the MGPT is designed to accommodate the amplification of three samples in a single PCR run, utilizing all 32 primer sets. Each sample is designated to specific wells in the PCR plate, with wells reserved for each patient differentiated by separate colors. Each well labeled as "P#" corresponds to a specific primer set, tailored to amplify a particular locus of interest.

4.2 *In-vitro* Validation

We observed amplification(s) according to Cq values for the DNA sample and the higher Cq among non-template and negative controls. Each of the listed primers in Table 9, are used in the next reaction to check PCR product as a so-called DNA template. Electrophoresis shows the expected lengths of amplicons for each of the primers (Figure 24).

4.2.1 Quantitative polymerase chain reaction and gel electrophoresis

Table 7. Multiplexed qPCR Cq values.

Well	Fluor	Content	Cq
A01	SYBR	DNA (100ng)	17,02
A02	SYBR	DNA (150ng)	14,91
A11	SYBR	NTC	37,56
A12	SYBR	Neg Ctrl	27,84

Table 8. Multiplexed qPCR melting curve values, temperatures in ^{o}C units.

Well	Fluor	Content	Melt Temperature	Peak Height	Begin Temperature	End Temperature
A01	SYBR	DNA **P1** (100ng)	83,00	235,80	76,60	86,60
A02	SYBR	DNA **P2** (150ng)	82,60	191,17	77,40	85,60
A11	SYBR	NTC	None	None	None	None
A12	SYBR	Neg Ctrl	76,60	108,68	74,00	80,00

Table 9. First five primer info from the first primer set (*BRCA1* regions).

ID	Amplicon Position	ForwardP	ReverseP	FpTm	RpTm	FpSize	RpSize	FpGC(%)	RpGC(%)
T1P1	chr17:43091362-43091612	CC***CA	AC***GA	61.13	61.23	26	23	38.46	43.48
T1P2	chr17:43091710-43091957	CC***GC	CC***GC	60.67	60.90	24	23	45.83	47.83
T1P3	chr17:43092060-43092307	CC***GT	AC***GC	60.44	60.56	24	22	45.83	45.45
T1P4	chr17:43092408-43092652	GA***TC	GC***TC	60.87	60.67	28	24	35.71	45.83
T1P5	chr17:43092741-43092990	GG***TG	GG***AG	60.76	60.87	25	25	44.00	44.00

T1P1: Tube 1 Pimer Pair 1, Fp/Rp: Forward/Reverse Primer, Tm: Melting Temperature in ºC, GC: Guanine and Cytosine Percentage, NTC: Non-template control, Neg Ctrl: Negative Control

Table 10. Pair by pair amplification Cq values.

Well	Fluor	Content	Cq	Cq Mean
A01	SYBR	T1P1	5,16	5,16
A02	SYBR	T1P2	2,27	2,27
A03	SYBR	T1P3	4,81	4,81
A04	SYBR	T1P4	7,18	7,18
A05	SYBR	T1P5	6,36	6,36
B01	SYBR	Neg Ctrl		0,00
B02	SYBR	Neg Ctrl		0,00
B03	SYBR	Neg Ctrl		0,00
B04	SYBR	Neg Ctrl		0,00
B05	SYBR	Neg Ctrl		0,00
C01	SYBR	NTC		0,00
C02	SYBR	NTC		0,00
C03	SYBR	NTC		0,00
C04	SYBR	NTC		0,00
C05	SYBR	NTC		0,00

Table 11. Pair by pair amplification melting curve results, Temperatures in ºC units.

Well	Fluor	Content	Melt Temperature	Peak Height	Begin Temperature	End Temperature
A01	SYBR	T1P1	83,00	338,53	80,80	86,20
A02	SYBR	T1P2	81,20	294,30	78,40	84,40
A03	SYBR	T1P3	81,80	340,96	78,80	85,80
A04	SYBR	T1P4	80,20	349,27	77,60	83,00
A05	SYBR	T1P5	81,20	270,40	78,00	84,20
B01	SYBR	Neg Ctrl	None	None	None	None
B02	SYBR	Neg Ctrl	None	None	None	None
B03	SYBR	Neg Ctrl	None	None	None	None
B04	SYBR	Neg Ctrl	None	None	None	None
B05	SYBR	Neg Ctrl	73,40	100,57	68,40	80,60
C01	SYBR	NTC	None	None	None	None
C02	SYBR	NTC	None	None	None	None
C03	SYBR	NTC	None	None	None	None
C04	SYBR	NTC	None	None	None	None

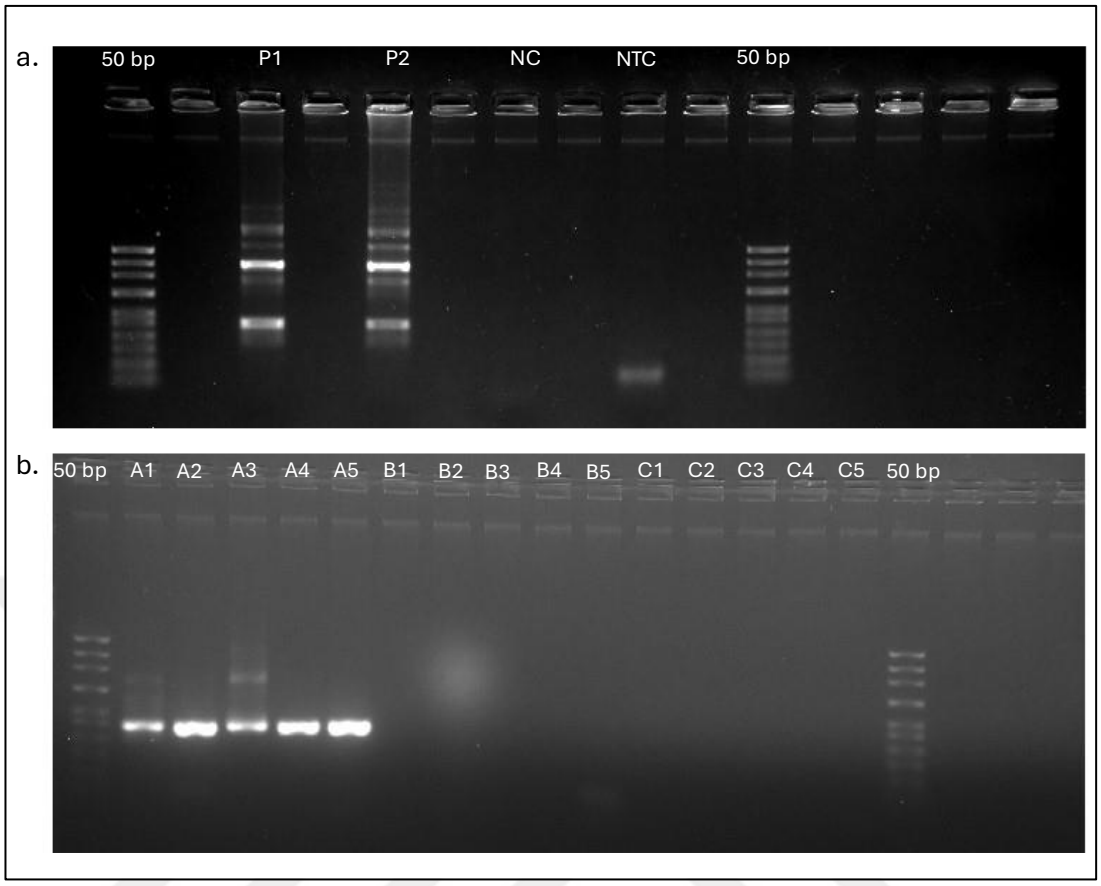

Figure 24. Gel electrophoresis results of multiplexed PCR products.

First, we observed multiplex PCR products in gel through different DNA template concentrations along with Negative Control (NC) and Non-template control (NTC). Using this PCR product, an additional PCR is performed using primer pairs individually to see amplifications clearly.

Gel electrophoresis of multiplexed PCR products. Lanes P1 and P2 correspond to the reactions using DNA concentrations of 100 ng and 150 ng, respectively. The NC (Negative Control) and NTC (Non-Template Control) lanes show the results of control reactions. The 50 bp DNA ladder is included in the first and last lanes for size reference. The NTC lane shows no visible bands, confirming the absence of contamination. The faint band in the NC lane may suggest minor primer-dimer formation, but no significant contamination is observed.

Gel electrophoresis of products from the multiplexed PCR (lanes A1-A5), negative controls (lanes B1-B5), and non-template controls (lanes C1-C5). Lanes A1 to A5 correspond to the five different primer pairs used in the prior multiplex PCR. The B1 to B5 lanes serve as negative controls for each corresponding primer pair, while the C1 to C5 lanes represent non-template controls. The 50 bp DNA ladder is again present in the first and last lanes for size reference. The presence of clear bands in lanes A1 to A5 indicates successful amplification of the target sequences in the multiplex reaction. However, faint or diffuse bands in some lanes may suggest the presence of non-specific amplification or primer-dimer formation. Notably, the absence of significant bands in the negative controls (B1-B5) and non-template controls (C1-C5) suggests that the multiplex PCR was mostly specific, with minimal contamination or non-specific amplification.

4.3 Variant Calling and Detection Analysis

In this section, we present the findings from the evaluation of our newly developed variant limit of detection (vLoD) tool designed for precision oncogenomic. This tool was engineered to address the challenges associated with detecting variants in different next generation sequencing applications from various samples. By employing a sophisticated algorithm that integrates genomic data processing with parallel computational techniques, we aimed to enhance the accuracy and efficiency of variant detection.

4.3.1 Benchmark of vLoD

This study presents the results of our newly developed computational framework designed to optimize the detection of somatic variants in multigene panel testing (MGPT) from liquid biopsy samples, leveraging next-generation sequencing (NGS) data. The main objective of our methodology was to refine the accuracy and efficiency of variant detection processes, which are critical in personalized cancer therapies and precision medicine. Accordingly, our approach turned into an automated system which can operate with not only somatic, but also germline variants as well.

The analytical process began with the rigorous preprocessing of raw sequencing reads to ensure optimal data quality. This initial step involved using industry-standard tools for quality control checks, alignment of sequences to a reference genome, and the identification and removal of duplicate reads that could potentially skew variant calling accuracy. Following data preparation, we employed a custom-developed variant limit of detection prediction (vLoD) algorithm designed to detect somatic point and small insertions/deletions variants with high specificity and sensitivity. This algorithm processes sequence alignment files to extract variant information and then applies a binomial probability model to compute a log-odds score (See the Equation in Figure 12). This score helps determine the likelihood of each variant being a true calling versus a sequencing artifact. Crucially, our algorithm integrates variant allele frequency (VAF) in its calculations, adjusting the detection threshold to minimize false positives while maximizing true positive detections.

For the practical application and validation of our tool, we utilized a comprehensive dataset of 652 liquid biopsy samples, encompassing 20 genes implicated in various cancers. This dataset was processed through our parallelized algorithmic structure to calculate the LoD for each variant reduced computational time.

Specifically, the use of a 64-core configuration allowed us to process data at a speed of 44.32 kilobases per minute (kb/min). This setup illustrates the principle that as the number of cores increases, the ability to handle simultaneous operations expands, effectively decreasing the time required to analyze large datasets. Statistical validation of our tool's performance was rigorously conducted using Fisher’s exact test (Table 12) to compare the correlation between predicted and actual variant detections. Overall, our results indicate a high level of accuracy, as evidenced by an 89.92% positive predictive value (PPV), with a statistical significance of $p<0.005$. Further, gene-based analyses underscored the tool’s reliability, showing significant detection across all evaluated genes ($p<0.003$). This establishes the vLoD tool's potential as a crucial resource for enhancing precision in oncogenomic diagnostics.

Table 12. Contingency table for the Fisher's exact test.

Classification	True Positive	True Negative
Predicted Positive	Group A → 1654	Group B → 5947
Predicted Negative	Group C → 73	Group D → 48442

Variant occurrences are categorized into four groups. Each variant is evaluated by their status in VCF files. The low p-value of 0.0001 from the Fisher's exact test indicates a statistically significant association between the predicted and actual results. This suggests that the observations are not just due to random chance.

An Odds Ratio of approximately 184.56 means that the odds of predicting a positive result (given it's actually positive) are roughly 184.56 times the odds of predicting a positive result (given it's actually negative).

Here we have a total of 73 variants that Mutect2 included as a somatic variant that we statistically filtered out. On the other hand, we have a total of 5947 variants that Mutect2 filters as either germline or non-qual to count in, but we statistically included.

Group A: If an alternate allele is called a somatic variant (by Mutect2; GATK) and is detectable (by the vLoD),

Group B: If an alternate allele is not called a somatic variant (by Mutect2; GATK) but is detectable (by the vLoD),

Group C: If an alternate allele is called a somatic variant (by Mutect2; GATK) but is not detectable (by the vLoD),

Group D: If an alternate allele is not called a somatic variant (by Mutect2; GATK) and is not detectable (by the vLoD).

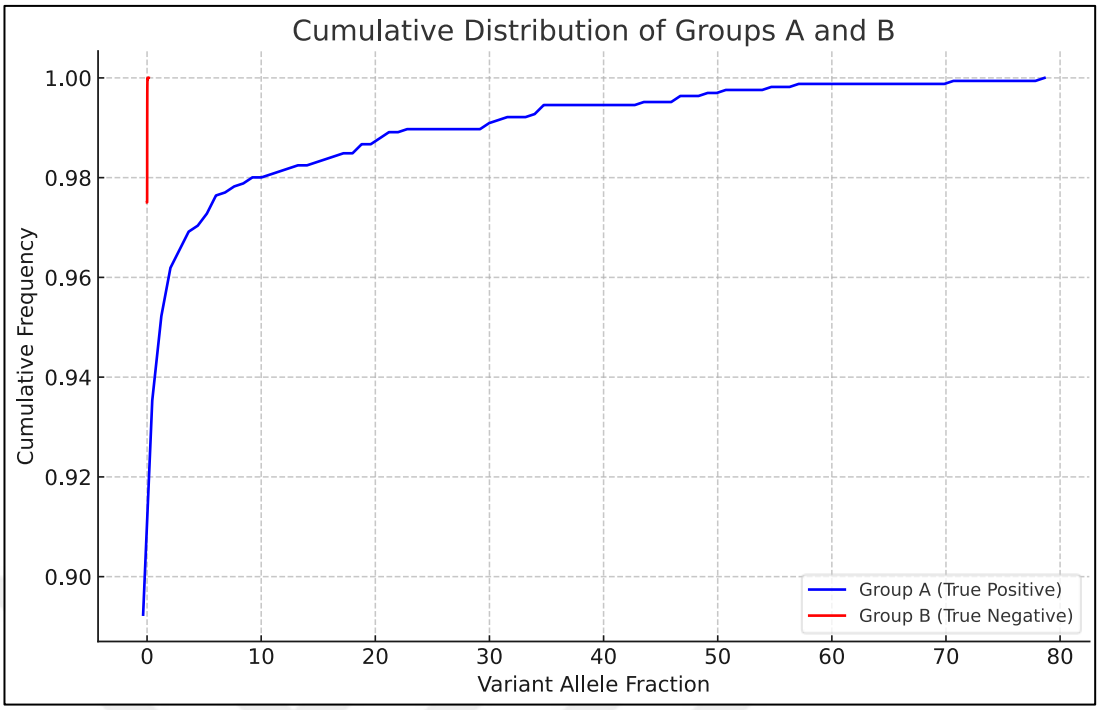

Figure 25. Cumulative distribution of groups A and B.

Here, we can identify at what point the variant allele fraction (VAF) overlaps with the true and false positives in terms of detectability status. Cumulative correct result frequencies across various VAFs provides an insight into the reliability of the vLoD prediction (Figure 25).

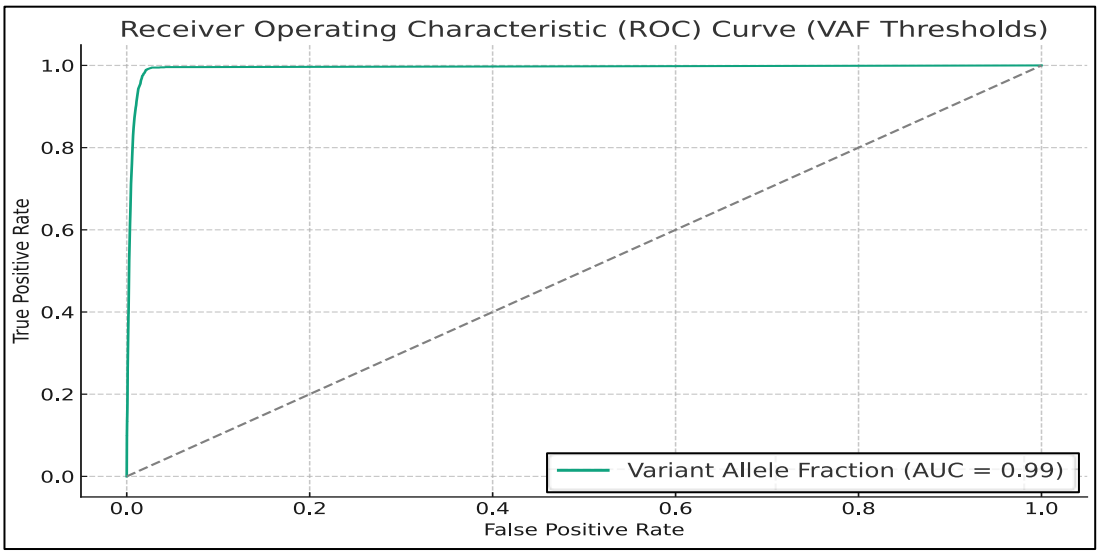

Figure 26. Receiver operating characteristic (ROC) curve with VAF thresholds.

This Plot charts precision (the proportion of true positives among all positive predictions) on the y-axis and recall (the proportion of true positives detected relative to their actual occurrences) on the x-axis. The area under the curve (AUC) measures the model's overall capacity to distinguish between classes throughout the range of VAF thresholds (Figure 26).

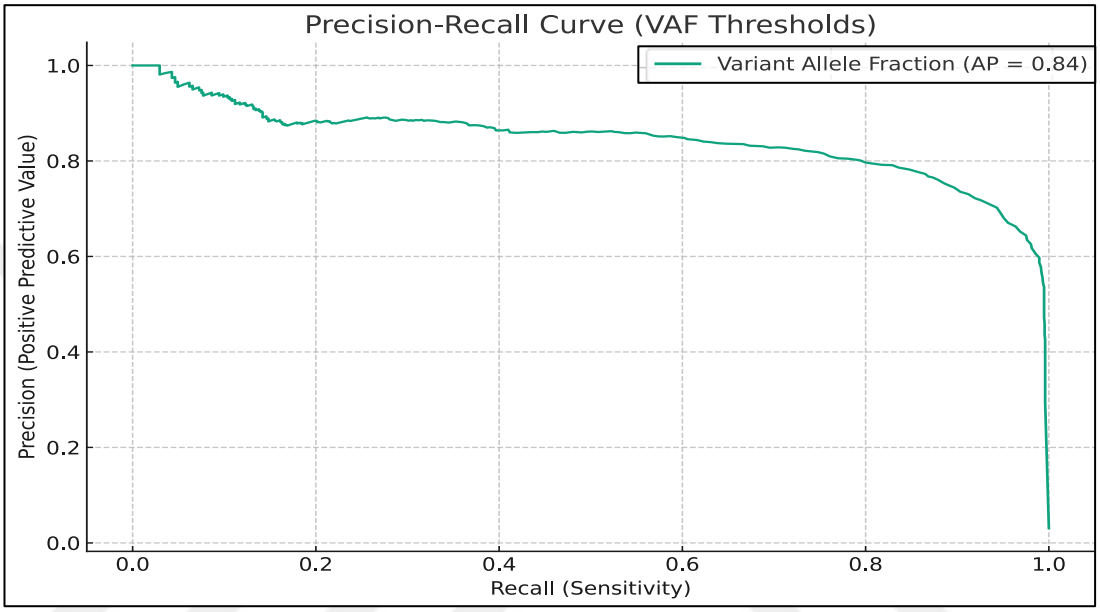

Figure 27. Precision-recall curve.

Figure 27 illustrates the precision-recall curve with a high AUC, indicating effective detection of the positive class (high recall) and precise predictive accuracy (high precision). The term "AP" refers to “Average Precision”, a metric that evaluates the overall performance of classification models. Here, Average Precision condenses the data from the precision-recall curve into a single comprehensive performance value across VAF thresholds.

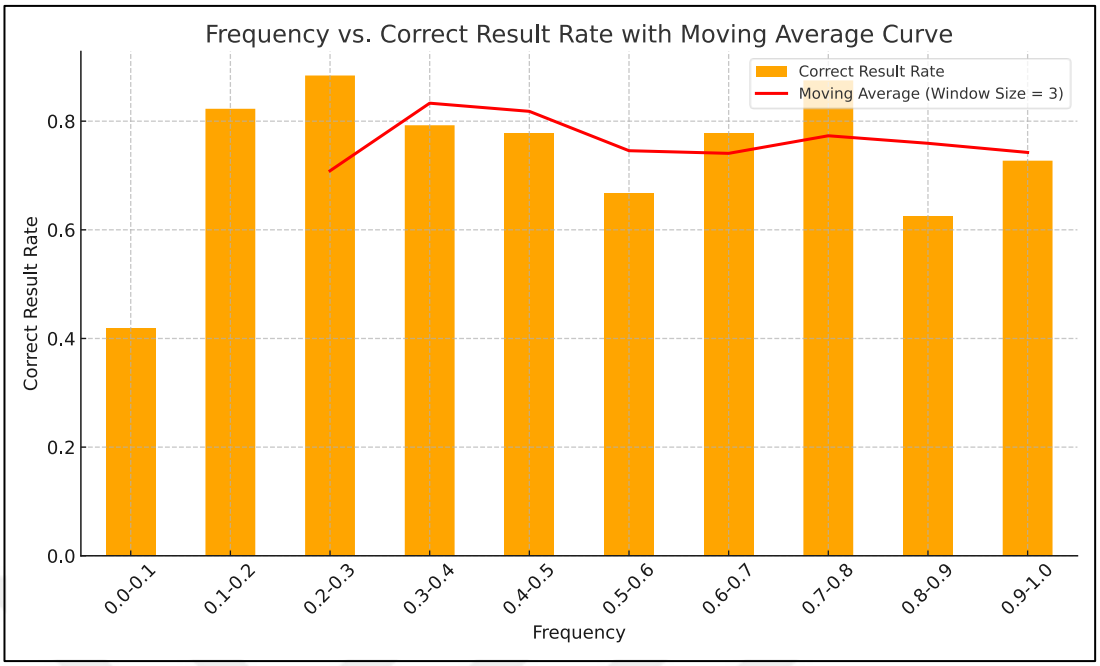

Figure 28. VAF vs. correct result rate.
This figure displays the average rate of correct results across various VAF.

5 DISCUSSION

The thesis emphasizes the pivotal role of advanced molecular diagnostics in the management of cancer, a globally significant health issue.

The adoption of next-generation sequencing (NGS) technologies has substantially enhanced the ability to delineate the molecular complexities of various cancers, thereby improving early detection, prognosis, and the customization of therapeutic strategies. The proposed development of a multi-gene diagnostic kit seeks to utilize these technological advancements to furnish reliable and precise assessments of somatic variants critical for understanding cancer progression and resistance to treatments.

In recent years, the integration of NGS with companion diagnostics has provided a deeper understanding of tumor biology, particularly in revealing the complexity of genetic and epigenetic alterations that drive cancer development and behavior (102).

Furthermore, the thesis acknowledges the need for sustainable and cost-effective approaches in medical diagnostics. It advocates for a targeted enrichment strategy that focuses exclusively on disease-associated loci, thereby curtailing the generation of superfluous data and reducing the labor intensity typically involved. This approach is expected not only to lessen unnecessary sequencing and analysis but also to decrease the financial and environmental costs.

Unfortunately, the absence of a locally based companion diagnostic multi-gene panel test (MGPT) in Türkiye has rendered diagnostic processes both costly and time intensive. This thesis proposes the initial design of a small-scale, comprehensive MGPT as a preliminary step towards developing a more accessible diagnostic kit. This endeavor aims to bridge the gap in local diagnostics infrastructure, potentially leading to more efficient and economically viable cancer management solutions within the region.

5.1 Primer Design

We made a strategic and cost-effective development of a multi-gene panel test (MGPT), using open-source resources, starting from primer design to gene selection and bioinformatics analysis. The selection of optimal primer sequences is crucial for the success of targeted sequencing, as it directly affects the efficiency and accuracy of the entire diagnostic process. The methodical approach adopted in the thesis, involving a comparative analysis of open-source primer design tools, is commendable for its depth and rigor. This evaluation ensures that the chosen tools are effective in terms of algorithmic efficiency and design accuracy and compatible with the specific requirements of cancer-related genetic studies.

However, the practical implications of primer design choices extend beyond the scope of computational assessments. The reliability of these tools in a clinical setting, particularly their capacity to consistently produce valid results under variable experimental conditions, remains a crucial consideration. Although the thesis outlines a thorough evaluation process, the application of these tools in clinical diagnostics would benefit from a more extensive *in-vitro* validation studies, including a broader range of samples and conditions to mimic clinical scenarios more closely. Given the lack of funds, we were unable to apply several laboratory procedures, like adapting primer sequences from several open-source tools and evaluating them not only *in-silico* but also *in-vitro.*

The Multiplex Primer Design (MPD) tool is a notably efficient tool in processing multiple targets at once, a feature indispensable for next-generation sequencing applications. Even though it considers uncovering single nucleotide polymorphisms (SNP), it presents limitations in processing highly complex or AT/GC-rich sequences, given its lower coverage. A significant concern is the inconsistent *in-vitro* validations, which raises doubts about the reliability of the primer design under varied experimental conditions. Overall, we chose not to obtain primer sequences from the MPD tool (96).

PrimerJinn, a relatively new tool in literature, incorporates algorithms that include *in-silico* PCR capabilities to enhance specificity. This results in high design accuracy, particularly advantageous for applications targeting specific sequences, such as drug resistance genes in pathogens. The primarily command-line interface, however, may deter technical problems. The tool is proficient in minimizing issues like primer dimerization and hairpin structures; however, achieving consistent Tm without empirical adjustments remains challenging. While some *in-vitro* validations have been performed, the lack of human sample validation remains a considerable drawback for its broader application (97).

The PMP Primer Design Tool demonstrates good algorithmic efficiency and is capable of handling diverse primer design tasks effectively. It excels in design accuracy, although its performance may depend on the target sequence. The interface is more complex than its counterparts because it requires manual database construction, which could restrict usage capacity. The tool manages primer dimerization and hairpin structures, with a reasonable Tm consistency and specificity, although these factors can vary depending on the genome's complexity or the assay's multiplexing. Unfortunately, this tool has been eliminated due to limited *in-vitro* studies for human samples (98).

The selection of genes based on their prevalence in various cancers and their roles in tumorigenesis is strategically secure. Using databases such as COSMIC to anchor this selection ensures that the chosen genes are relevant and their variants actionable. This approach not only enhances the clinical relevance of the MGPT but also increases its potential utility in identifying targetable variants that can guide therapy decisions.

Nevertheless, the selection process could be further enriched by incorporating emerging data on genetic variants that may have been underrepresented in previous studies. As cancer genomics is rapidly evolving, continual updates and validation of gene lists against the latest research and clinical data are essential to maintain the relevance and effectiveness of the MGPT. Conversely, this was an initial small-scale and comprehensive kit that made us overcome some significant challenges during the

design. Yet we are aware that we could make the target loci size bigger, it was important to take budget and time into account for the first prototype. Also, we kept the design revisable for potential new MGPT designs in our future studies.

The integration of custom and known computational tools for data analysis is a key strength of this thesis, facilitating the handling of complex genomic data and enhancing the interpretability of results. The use of Python and its famous libraries for data manipulation and visualization is particularly effective in providing easiness while performing different tasks without explicitly creating new complex functions. We hope that the thesis will expand on the specific challenges, such as dealing with high variability in mutation rates across samples or the integration of multi-omics data. Addressing these challenges would provide a more comprehensive understanding of the limitations and capabilities of the current bioinformatics tools and strategies.

The development and refinement of primer sequences for targeted sequencing are central to enhancing the accuracy and efficiency of molecular diagnostics in cancer. Our research utilizes state-of-the-art genomic databases and bioinformatics tools, detailing a methodology that enhances the specificity and efficiency of PCR amplifications. Each step is critical for ensuring the reliability of the PCR amplifications in clinical diagnostics.

The utilization of the UCSC genome database for extracting locus information using RefSeq numbers establishes a robust foundation for primer design. This approach ensures adherence to a standardized genetic reference, which is crucial for the reproducibility of the results across different research settings. The selection of "*Exons plus 20 bases on each end*" from the Table Browser is a strategic decision that enhances the likelihood of capturing significant regulatory and splicing-related genetic variations. This methodological choice is pivotal for targeting clinically relevant variants that might influence cancer phenotype and patient response to a treatment.

The optimization of primer design parameters such as GC content and melting temperature addresses the technical challenges associated with PCR. Setting these

parameters to optimize the primer-template annealing conditions minimizes the occurrence of non-specific bindings and maximizes the efficiency of the amplification process. We are currently working further to explore the impact of these parameters on the overall yield and specificity of the multiplex PCR, potentially through a series of empirical tests that could validate the *in-silico* findings with actual experimental data.

The development of a custom Python algorithm for primer dimer and hairpin structure analysis is a significant advancement in primer design technology. The algorithm's functions provide a detailed assessment of potential complications in PCR reactions. This proactive computational evaluation helps in refining the primer design before laboratory implementation, thus saving resources, and enhancing the reliability of the diagnostic tests. The dimerization score calculation, which considers both alignment length and GC content, is commendable. However, the discussion of these results in the thesis could be enhanced by comparing them with established validations or through simulations that illustrate the potential impact of identified dimers on PCR amplification efficacy. Such an analysis would not only validate the computational predictions but also strengthen the case for the algorithm's utility in clinical settings.

ALPlex presents a novel approach to primer analysis in multiplex PCR, offering both strengths and limitations. Its main advantage lies in the automation of the detection of potential primer dimers and hairpin structures, reducing the likelihood of non-specific amplification and ensuring more reliable PCR outcomes. By incorporating a complementarity-based scoring system that accounts for the biochemical stability of base pairs, particularly guanine-cytosine (GC) pairs, ALPlex provides a more nuanced evaluation of primer interactions compared to traditional alignment-based methods. This can improve primer selection, especially in multiplex setups where multiple primers interact in the same reaction.

However, as with any bioinformatics tool, ALPlex is not without its limitations. Its scoring system, while innovative, may not capture all the complexities of primer behavior in actual biological systems, where factors like reaction conditions and

template quality can affect PCR outcomes. Additionally, the algorithm's reliance on base complementarity may overlook other factors that contribute to primer efficiency, such as secondary structures or the influence of thermodynamic properties beyond GC content.

Furthermore, while ALPlex automates and accelerates the process of primer analysis, it is important to recognize that no algorithm can fully replace experimental validation. Primers deemed ideal by ALPlex may still need to be tested in laboratory conditions to ensure their performance in real-world applications. As such, ALPlex can serve as a valuable tool for the initial stages of primer design but should be used in conjunction with empirical testing and other validation methods to optimize PCR results. Overall, ALPlex offers a useful addition to the toolbox for multiplex PCR primer analysis, though it should be considered one part of a larger, more comprehensive workflow.

Additionally, as a future route, the way ALPlex works in its sliding window approach will be integrated into well-known alignment algorithms to improve its accuracy and scoring system. Also, the scoring system will be incorporated through biochemical properties to mimic *in-vitro* conditions prior to experiments.

The coverage rate achieved signifies the effectiveness of the adapted primer sequences in capturing the genomic regions of interest. Nevertheless, the challenges associated with complex genomic regions, such as the *PMS2* gene heptad repeats, underscore the limitations of current primer design strategies. The thesis could benefit from a deeper exploration into these limitations, perhaps the use of alternative technologies like long-read sequencing, which might overcome these hurdles. Still, our focus was using a next generation sequencing approach with a lower cost.

The field of molecular diagnostics could greatly benefit from integrating newer genomic technologies and more sophisticated bioinformatics analyses. Continuous updates to genomic databases and refinements to algorithms for primer design should be pursued to keep pace with the rapidly evolving landscape of cancer genomics.

Additionally, incorporating machine learning models to predict primer efficiency and specificity based on historical data could further enhance the precision of primer design. Moreover, extending the primer design framework to accommodate RNA-based targets could open new avenues for understanding post-transcriptional modifications and their role in cancer. This expansion would be particularly pertinent for cancers where RNA splicing variants are crucial in disease progression.

As proposed in this thesis, the development of a local MGPT has significant implications for improving cancer diagnostics in Türkiye. By reducing reliance on international diagnostic resources, which can be costly and time-consuming, a localized approach could enhance the accessibility and timeliness of cancer diagnostics. Additionally, this initiative could pave the way for the development of more such kits tailored to the specific genetic backgrounds including cancer profiles common in the region. Although we believe that the current loci coverage is insufficient to define a critical parameter, tumor mutational burden, expanding the loci in the future using the same design techniques would enable us to achieve this goal.

Upon examining the *in-vitro* validation results, the observed Cq values in the negative control and non-template control (*GAPDH 156 bp PCR product*) differ a lot in the multiplexed qPCR. The gel electrophoresis results suggest the presence of a dimer or hairpin structure, likely due to the absence of a template sample. This is inferred from the appearance of a 50-100 bp band in the negative control, which is absent in the non-template control sample. Although there is a possibility of slight cross-contamination, this seems less likely, as the band does not appear when individual primer pairs are used to amplify the prior multiplexed product. Both the Cq values and gel electrophoresis data indicate that the desired amplicons are present. However, without sequencing the amplicons and aligning them to the human reference genome, we cannot confirm this with absolute certainty. Unfortunately, due to time and budget constraints, we were unable to prepare libraries and sequence the products.

Multiplexed PCR reactions are prone to producing non-specific amplicons due to the increased complexity of the reaction. However, these unwanted products can be

minimized by optimizing various reaction conditions to improve primer specificity and reduce interactions that lead to non-specific amplification.

The results clearly demonstrate that the most prominent bands correspond to the expected sizes of our target amplicons, even though the primers used were not HPLC (High-Performance Liquid Chromatography) purified. This lack of purification might contribute to the presence of minor non-specific products, as well. By using HPLC-purified primers, only the full-length, correctly synthesized oligonucleotides are used in the PCR reaction, which significantly reduces the likelihood of non-specific binding and amplification. Though, due to time and budged constraints we were only dedicated to seeing if the primers would work when multiplexed as soon as possible and as cheap as possible. Therefore, in future experiments, it would be beneficial to use HPLC-purified primers to enhance the specificity of the amplification and reduce the occurrence of non-specific bands. Additionally, verifying the amplicons through sequencing would be essential for confirming the accuracy of the multiplexed PCR results.

In conclusion, while the thesis presents a robust framework for the development of a cancer-specific MGPT, future work should focus on extensive *in-vitro* clinical validation, adaptation to emerging genomic data, and integration with other diagnostic modalities. This would not only enhance the diagnostic accuracy but also ensure the clinical efficacy of the MGPT in personalized cancer treatment strategies.

5.2 Variant Calling and Detection Analysis

The cornerstone of precision oncology lies in the accurate and timely detection of genetic variants from clinical samples. This thesis details the development and validation of a multi-gene panel testing (MGPT) approach using advanced bioinformatics tools to enhance the reliability of somatic variant detection in liquid biopsies. Central to this endeavor is the implementation of a robust computational pipeline integrated with a variant limit of detection (vLoD) algorithm, facilitating precise and individualized cancer treatment strategies.

The design of the bioinformatic pipeline, based on the Genome Analysis Tool Kit (GATK) Best Practices, ensures a robust framework for variant calling. The incorporation of GATK's tools, such as Mutect2 and FilterMutectCalls, provides a refined approach for detecting somatic variants, which are critical for tailoring cancer therapies. The vLoD algorithm enhances this process by introducing a quantitative threshold for variant detection, thereby minimizing the risk of false positives and false negatives, which are common challenges in high-throughput sequencing analyses.

While the methodology demonstrates high scientific rigor, certain aspects warrant further scrutiny. For instance, the reliance only on liquid biopsy samples, while minimally invasive and clinically convenient, may introduce biases due to the variable quality and quantity of circulating tumor DNA (ctDNA). Such factors could potentially affect the sensitivity and specificity of the vLoD algorithm and, by extension, the overall reliability. Thus, the benchmarking samples should be expanded.

The successful deployment of the MGPT and its proper analysis holds significant implications for clinical practice, particularly in the realm of personalized medicine. By enabling more accurate detection of somatic variants, this approach facilitates the development of tailored therapies that are more effective and less toxic than traditional treatment methods. Moreover, the rapid processing capabilities of the bioinformatics pipeline ensure timely insights into patient genomics, which is essential for critical care decisions. From a research perspective, the MGPT platform provides a valuable tool for exploring the genetic underpinnings of cancer. The extensive data generated could be used to uncover novel biomarkers for early detection and prognosis, thereby advancing the field of cancer genomics.

Looking ahead, further benchmark for the vLoD algorithm is necessary to check its utility across different types of diseases and stages of disease. Expanding the methodology to include germline variants and integrating it with other genomic data types could provide a more comprehensive understanding of accuracy and precision. Additionally, future work could explore the application of artificial intelligence and machine learning techniques to predict the existence and the phenotypic effects of a

variant more accurately. Such advancements would not only improve the diagnostic power of the MGPT platform but also reduce the labor load, making the system more efficient and scalable.

In short, the incorporation of a local-based MGPT and the novel vLoD algorithm integrated custom-made Analysis Pipeline represents a significant step forward in precision oncology, offering promising avenues for both clinical and research applications. Continued innovation and critical usage of these technologies will be essential to determine the full potential for cancer diagnosis and treatment in Türkiye given its advancing global position in health tourism.

6 CONCLUSION

In conclusion, we developed and validated a MGPT through *in-silico* techniques. The creation of a prototype MGPT that covers over 94% of targeted loci represents a significant advancement in the field of genomic diagnostics. The use of open-source primer design tools and a custom algorithm (ALPlex) has ensured high specificity in primer design, crucial for reducing primer-primer interactions and unwanted secondary structures, thus enhancing the overall reliability of PCR amplifications. Furthermore, the implementation of a novel analysis pipeline capable of calculating allele detectability (vLoD) exemplifies a breakthrough in the application of mathematical models to improve the accuracy and personalization of cancer therapies. The results of this research contribute a robust methodological framework that not only advances targeted sequencing technologies but also holds great promise for the enhancement of companion diagnostic tools in Türkiye. Even though we could not provide laboratory evidence for the MGPT, we took the first steps of a local diagnostic kit. This foundation is a critical step toward the regional development of clinical practices that can lead to better patient-specific therapeutic strategies and improved healthcare outcomes.

7 REFERENCES

1. World Health Organization. Cancer. 2022. Available at: https://www.who.int/news- room/fact-sheets/detail/cancer
2. Weinberg, R.A. The Biology of Cancer. 2nd ed., Garland Science, 2013.
3. Stratton MR, Campbell PJ, Futreal PA. "The cancer genome." Nature, vol. 458, no. 7239, 2009, pp. 719–724.
4. Vogelstein B, Papadopoulos N, Velculescu VE, et al. "Cancer genome landscapes." Science, vol. 339, no. 6127, 2013, pp. 1546-1558.
5. Morash M, Mitchell H, Beltran H, et al. "The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology." Journal of Personalized Medicine, vol. 8, no. 3, 2018, p. 30.
6. Koboldt DC, Steinberg KM, Larson DE, et al. "The next-generation sequencing revolution and its impact on genomics." Cell, vol. 155, no. 1, 2013, pp. 27-38.
7. Dienstmann R, Rodon J, Barretina J, et al. "Genomic medicine frontier in human solid tumors: Prospects and challenges." Journal of Clinical Oncology, vol. 31, no. 15, 2013, pp. 1874-1884.
8. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. "K-ras variants and benefit from cetuximab in advanced colorectal cancer." New England Journal of Medicine, vol. 359, no. 17, 2008, pp. 1757-1765.
9. Kechin A, Borobova V, Boyarskikh U, et al. "NGS-PrimerPlex: High-throughput Primer Design for Multiplex PCR." PLOS Computational Biology, vol. 16, no. 12, 2020, p. e1008468.
10. Geynisman DM, Whelan CM, Meeker CR, et al. "Economic model comparing universal tumor testing and sequential testing for Lynch syndrome based on a large integrated health system." Journal of Clinical Oncology, vol. 36, no. 31, 2018, pp. 3178-3185.
11. Aaltonen LA, Abascal F, Abeshouse A, et al. "Pan-cancer Analysis of Whole Genomes." Nature, vol. 578, no. 7793, 2020, pp. 82–93.
12. Muller PAJ, Vousden KH. "p53 Variants in Cancer." Nature Cell Biology, vol. 15, no. 1, 2013, pp. 1-12.
13. Maron JL, Johnson KL, Slonim DK, et al. "Circulating fetal cell analysis: A review of the past and future applications." Journal of Clinical Medicine, vol. 5, no. 2, 2016, p. 34.
14. Wu L, Yao H, Chen H, et al. "Landscape of somatic alterations in large-scale solid tumors from an Asian population." Nature Communications, vol. 13, no. 1, 2022.
15. Lander ES, Linton LM, Birren B, et al. "Initial sequencing and analysis of the human genome." Nature, vol. 409, 2001, pp. 860-921.
16. Weinberg RA. The Biology of Cancer. 2nd ed., Garland Science, 2013.
17. Koboldt DC, Steinberg KM, Larson DE, et al. "The next-generation sequencing revolution and its impact on genomics." Cell, vol. 155, no. 1, 2013, pp. 27-38.

18. Morash M, Mitchell H, Beltran H, et al. "The Role of Next-Generation Sequencing in Precision Medicine." Journal of Personalized Medicine, vol. 8, no. 3, 2018, p. 30.
19. The Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of human colon and rectal cancer." Nature, vol. 487, no. 7407, 2012, pp. 330-337.
20. International Cancer Genome Consortium. "International network of cancer genome projects." Nature, vol. 464, no. 7291, 2010, pp. 993-998.
21. Stratton MR, Campbell PJ, Futreal PA. "The cancer genome." Nature, vol. 458, no. 7239, 2009, pp. 719-724.
22. Vogelstein B, Papadopoulos N, Velculescu VE, et al. "Cancer genome landscapes." Science, vol. 339, no. 6127, 2013, pp. 1546-1558.
23. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. "K-ras variants and benefit from cetuximab in advanced colorectal cancer." New England Journal of Medicine, vol. 359, no. 17, 2008, pp. 1757-1765.
24. Ford D, Easton DF, Stratton M, et al. "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families." American Journal of Human Genetics, vol. 62, no. 3, 1998, pp. 676-689.
25. Domchek SM, Friebel TM, Singer CF, et al. "Association of risk-reducing surgery in BRCA1 or BRCA2 mutation carriers with cancer risk and mortality." Journal of the American Medical Association, vol. 304, no. 9, 2010, pp. 967-975.
26. Shendure J, Ji H. "Next-generation DNA sequencing." Nature Biotechnology, vol. 26, no. 10, 2008, pp. 1135-1145.
27. George A, Kaye S, Banerjee S, George A, Kaye S, Banerjee S. Delivering widespread BRCA testing and PARP inhibition to patients with ovarian cancer. Nature Reviews Clinical Oncology 2016 14:5. 2016-12-13;14(5).
28. Nishioka M, Bundo M, Iwamoto K, Kato T, Nishioka M, Bundo M, et al. Somatic variants in the human brain: implications for psychiatric research. Molecular Psychiatry 2018 24:6. 2018-08-07;24(6).
29. Yu Z, Coorens THH, Uddin MM, Ardlie KG, Lennon N, Natarajan P, et al. Genetic variation across and within individuals. Nature Reviews Genetics 2024. 2024-03-28.
30. Yu Z, Jin J, Tin A, Köttgen A, Yu B, Chen J, et al. Polygenic Risk Scores for Kidney Function and Their Associations with Circulating Proteome, and Incident Kidney Diseases. Journal of the American Society of Nephrology. December 2021;32(12).
31. Heather JM, Chain B. "The sequence of sequencers: The history of sequencing DNA." Genomics, vol. 107, no. 1, 2016, pp. 1-8.
32. Lander ES, Linton LM, Birren B, et al. "Initial sequencing and analysis of the human genome." Nature, vol. 409, 2001, pp. 860-921.
33. Metzker ML. "Sequencing technologies — the next generation." Nature Reviews Genetics, vol. 11, no. 1, 2010, pp. 31-46.

34. Koboldt DC, Steinberg KM, Larson DE, et al. "The next-generation sequencing revolution and its impact on genomics." Cell, vol. 155, no. 1, 2013, pp. 27-38.
35. Goodwin S, McPherson JD, McCombie WR. "Coming of age: ten years of next-generation sequencing technologies." Nature Reviews Genetics, vol. 17, no. 6, 2016, pp. 333-351.
36. Swanton C. "Intratumor heterogeneity: Evolution through space and time." Cancer Research, vol. 72, no. 19, 2012, pp. 4875-4882.
37. Shendure J, Ji H, Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology 2008 26:10. 2008-10-09;26(10).
38. Metzker ML, Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics 2010 11:1. 2009-12-08;11(1).
39. Goodwin S, McPherson JD, McCombie WR, Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 2016 17:6. 2016-05-17;17(6).
40. Metzker ML. "Sequencing technologies — the next generation." Nature Reviews Genetics, vol. 11, no. 1, 2010, pp. 31-46.
41. Mardis ER. "Next-generation sequencing platforms." Annual Review of Analytical Chemistry, vol. 6, 2013, pp. 287-303.
42. Goodwin S, McPherson JD, McCombie WR. "Coming of age: ten years of next-generation sequencing technologies." Nature Reviews Genetics, vol. 17, no. 6, 2016, pp. 333-351.
43. Bentley DR, Balasubramanian S, Swerdlow HP, et al. "Accurate whole human genome sequencing using reversible terminator chemistry." Nature, vol. 456, no. 7218, 2008, pp. 53-59.
44. Eid J, Fehr A, Gray J, et al. "Real-time DNA sequencing from single polymerase molecules." Science, vol. 323, no. 5910, 2009, pp. 133-138.
45. Marusyk A, Almendro V, Polyak K. "Intra-tumour heterogeneity: A looking glass for cancer?" Nature Reviews Cancer, vol. 12, no. 5, 2012, pp. 323-334.
46. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community - PubMed. Genome biology. 11/25/2016;17(1).
47. The Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of human colon and rectal cancer." Nature, vol. 487, no. 7407, 2012, pp. 330-337.
48. International Cancer Genome Consortium. "International network of cancer genome projects." Nature, vol. 464, no. 7291, 2010, pp. 993-998.
49. Chapman PB, Hauschild A, Robert C, et al. "Improved survival with vemurafenib in melanoma with BRAF V600E mutation." New England Journal of Medicine, vol. 364, no. 26, 2011, pp. 2507-2516.
50. Diaz LA, Bardelli A. "Liquid biopsies: Genotyping circulating tumor DNA." Journal of Clinical Oncology, vol. 32, no. 6, 2014, pp. 579-586.
51. Zhang L, Wang W, Wang S. "Gene expression profiling-based immune signature for tumor classification." Oncotarget, vol. 8, no. 12, 2017, pp. 19114-19123.

52. Ford D, Easton DF, Stratton M, et al. "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families." American Journal of Human Genetics, vol. 62, no. 3, 1998, pp. 676-689.
53. Chapman PB, Hauschild A, Robert C, et al. "Improved survival with vemurafenib in melanoma with BRAF V600E mutation." New England Journal of Medicine, vol. 364, no. 26, 2011, pp. 2507-2516.
54. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies - PubMed. Blood. 04/08/2010;115(14).
55. Minimal Residual Disease Detection in Acute Lymphoblastic Leukemia - PubMed. International journal of molecular sciences. 02/05/2020;21(3).
56. Clinical implications of sequential MRD monitoring by NGS at 2 time points after chemotherapy in patients with AML - PubMed. Blood advances. 05/25/2021;5(10).
57. Establishing guidelines to harmonize tumor mutational burden (TMB): in silico assessment of variation in TMB quantification across diagnostic platforms: phase I of the Friends of Cancer Research TMB Harmonization Project - PubMed. Journal for immunotherapy of cancer. 2020 Mar;8(1).
58. The repertoire of mutational signatures in human cancer - PubMed. Nature. 2020 Feb;578(7793).
59. Tumor Mutation Burden-From Hopes to Doubts - PubMed. JAMA oncology. 07/01/2019;5(7).
60. Tumor Mutational Burden and Efficacy of Immune Checkpoint Inhibitors: A Systematic Review and Meta-Analysis - PubMed. Cancers. 11/15/2019;11(11).
61. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic - PubMed. Annals of oncology : official journal of the European Society for Medical Oncology. 01/01/2019;30(1).
62. Liquid biopsies come of age: towards implementation of circulating tumour DNA - PubMed. Nature reviews Cancer. 2017 Apr;17(4).
63. Quantitation of genomic DNA in plasma and serum samples: higher concentrations of genomic DNA found in serum than in plasma - PubMed. Transfusion. 2001 Feb;41(2).
64. Utility of serum DNA and pyrosequencing for the detection of EGFR variants in non-small cell lung cancer - PubMed. Cancer genetics. 2013 Mar;206(3).
65. Companion biomarkers: paving the pathway to personalized treatment for cancer - PubMed. Clinical chemistry. 2013 Oct;59(10).
66. Tumor heterogeneity: causes and consequences - PubMed. Biochimica et biophysica acta. 2010 Jan;1805(1).
67. From polyploidy to aneuploidy, genome instability and cancer - PubMed. Nature reviews Molecular cell biology. 2004 Jan;5(1).
68. DNA damage response--a double-edged sword in cancer prevention and cancer therapy - PubMed. Cancer letters. 03/01/2015;358(1).

69. Aging and the rise of somatic cancer-associated variants in normal tissues - PubMed. PLoS genetics. 01/04/2018;14(1).
70. Standardization of Sequencing Coverage Depth in NGS: Recommendation for Detection of Clonal and Subclonal Variants in Cancer Diagnostics - PubMed. Frontiers in oncology. 09/04/2019;9.
71. Oncogene Concatenated Enriched Amplicon Nanopore Sequencing for rapid, accurate, and affordable somatic mutation detection - PubMed. Genome biology. 09/06/2021;22(1).
72. Assessing Limit of Detection in Clinical Sequencing - PubMed. The Journal of molecular diagnostics : JMD. 2021 Apr;23(4).
73. Foundation PS. Python Language Reference. v.3.10 ed2021.
74. Corporation M. Visual Studio Code. version 1.75 ed2023.
75. Ltd. C. Ubuntu. 22.04 ed2024.
76. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research. 2017;45(D1):D777-D83.
77. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nature Reviews Cancer. 2004;4(3):177-83.
78. Boland CR, Goel A. Microsatellite Instability in Colorectal Cancer. Gastroenterology. 2010;138(6):2073-87.e3.
79. Beggs AD, Latchford AR, Vasen HFA, Moslein G, Alonso A, Aretz S, et al. Peutz-Jeghers syndrome: a systematic review and recommendations for management. Gut. 2010;59(7):975-86.
80. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal - PubMed. Science signaling. 04/02/2013;6(269).
81. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data - PubMed. Cancer discovery. 2012 May;2(5).
82. Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal - PubMed. Cancer research. 12/01/2023;83(23).
83. Cole CG, Mccann OT, Collins JE, Oliver K, Willey D, Gribble SM, et al. Finishing the finished human chromosome 22 sequence. Genome Biology. 2008;9(5):R78.
84. Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Research. 2004;32(90001):493D-6.
85. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, Mcveigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016;44(D1):D733-D45.
86. Sayers EW, Bolton EE, Brister R, J, Canese K, Chan J, Comeau C, Donald, et al. Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Research. 2023;51(D1):D29-D38.
87. Ensembl. BED File Format - Definition and supported options: EMBL-EBI; 2024 Available from: https://www.ensembl.org/info/website/upload/bed.html.

88. Shen Z, Qu W, Wang W, Lu Y, Wu Y, Li Z, et al. MPprimer: a program for reliable multiplex PCR primer design. BMC Bioinformatics. 2010;11(1):143.
89. S. A. FastQC: a quality control tool for high throughput sequence data. 2010.
90. Heng L. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. v2 ed. Cornell University2013.
91. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
92. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297-303.
93. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43(5):491-8.
94. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point variants in impure and heterogeneous cancer samples. Nature Biotechnology. 2013;31(3):213-9.
95. Akkuş A, Ergun B, Gökbayrak M, Çine N, Özdemir Ö, Ng ÖH. 13P A novel algorithm for predicting variant detectability in oncogenomic analysis. ESMO Open. 2023;8(1):101659.
96. Wingo TS, Kotlar A, Cutler DJ. MPD: multiplex primer design for next-generation targeted sequencing. BMC Bioinformatics. 2017;18(1).
97. Limberis JD, Metcalfe JZ. primerJinn: a tool for rationally designing multiplex PCR primer sets for amplicon sequencing and performing in silico PCR. BMC Bioinformatics. 2023;24(1).
98. Yang L, Ding F, Lin Q, Xie J, Fan W, Dai F, et al. A tool to automatically design multiplex PCR primer pairs for specific targets using diverse templates. Scientific Reports. 2023;13(1).
99. Hammet F, Mahmood K, Green TR, Nguyen-Dumont T, Southey MC, Buchanan DD, et al. Hi-Plex2: a simple and robust approach to targeted sequencing-based genetic screening. BioTechniques. 2019;67(3):118-22.
100. Kreer C, Döring M, Lehnen N, Ercanoglu MS, Gieselmann L, Luca D, et al. openPrimeR for multiplex amplification of highly diverse templates. Journal of Immunological Methods. 2020;480.
101. Kechin A, Borobova V, Boyarskikh U, Khrapov E, Subbotin S, Filipenko M. NGS-PrimerPlex: High-throughput primer design for multiplex polymerase chain reactions. PLOS Computational Biology. 2020;16(12):e1008468.
102. Morash M, Mitchell H, Beltran H, Elemento O, Pathak J. The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. Journal of Personalized Medicine. 2018;8(3):30.
103. Nguyen B, Fong C, Luthra A, Smith SA, Dinatale RG, Nandakumar S, et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell. 2022;185(3):563-75.e11.

104. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature Medicine. 2017;23(6):703-13.
105. Wu L, Yao H, Chen H, Wang A, Guo K, Gou W, et al. Landscape of somatic alterations in large-scale solid tumors from an Asian population. Nature Communications. 2022;13(1).
106. Aaltonen LA, Abascal F, Abeshouse A, Aburatani H, Adams DJ, Agrawal N, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82-93.
107. Rosen EY, Goldman DA, Hechtman JF, Benayed R, Schram AM, Cocco E, et al. TRK Fusions Are Enriched in Cancers with Uncommon Histologies and the Absence of Canonical Driver Variants. Clinical Cancer Research. 2020;26(7):1624-32.
108. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nature Genetics. 2020;52(11):1219-26.
109. Robinson DR, Wu Y-M, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548(7667):297-303.
110. Hyman DM, Piha-Paul SA, Won H, Rodon J, Saura C, Shapiro GI, et al. HER kinase inhibition in patients with HER2- and HER3-mutant cancers. Nature. 2018;554(7691):189-94.
111. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nature Genetics. 2019;51(2):202-6.
112. BioRender.com [Internet]. Toronto (ON): BioRender; 2024. Available from: https://app.biorender.com/

8 APPENDIX

APPENDIX 1. List of Studies Used in The Gene Set Curation from the cBioPortal

Study abbreviation and title	Number of samples*	Fraction by percent*
msk_met_2021: Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients **(103)**.	13.345	60.1%
msk_impact_2017: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients **(104)**.	4.192	18.9%
pan_origimed_2020: Landscape of somatic alterations in large-scale solid tumors from an Asian population **(105)**.	2.932	13.2%
pancan_pcawg_2020: Pan-cancer analysis of whole genomes **(106)**.	719	3.2%
tmb_mskcc_2018: TRK Fusions Are Enriched in Cancers with Uncommon Histologies and the Absence of Canonical Driver Variants **(107)**.	516	2.3%
msk_ch_2020: Cancer therapy shapes the fitness landscape of clonal hematopoiesis **(108)**.	245	1.1%
metastatic_solid_tumors_mich_2017: Integrative clinical genomics of metastatic cancer **(109)**.	179	0.8%
mixed_allen_2018: Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors **(50)**.	43	0.2%
summit_2018: HER kinase inhibition in patients with HER2- and HER3-mutant cancers **(110)**.	40	0.2%
ntrk_msk_2019: Tumor mutational load predicts survival after immunotherapy across multiple cancer types **(111)**.	8	<0.1%

* We present the remaining samples after the curation with the gene list, not all the patients in each cohort.

APPENDIX 2. Python Script used in the Data Curation from the cBioPortal

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data
file_path = '/path/to/data.tsv'
data = pd.read_csv(file_path, sep='\t')

# Preprocessing for mutation count plots
data['Mutation Count'] = pd.to_numeric(data['Mutation Count'], errors='coerce')
filtered_data = data.dropna(subset=['Mutation Count'])
filtered_data = filtered_data[filtered_data['Mutation Count'] > 0]

# Plot 1: Mutation Count (log-scale) vs. Cancer Type
plt.figure(figsize=(12, 8))
sns.violinplot(x='Cancer Type', y=np.log(filtered_data['Mutation Count']), data=filtered_data)
plt.xticks(rotation=45)
plt.title('Mutation Count (log-scale) vs. Cancer Type')
plt.tight_layout()
plt.savefig('mutation_count_log_scale_vs_cancer_type.svg', format='svg')
plt.close()

# Plot 2: Total Number of Variants vs. Cancer Type (without log-scale)
plt.figure(figsize=(12, 8))
sns.violinplot(x='Cancer Type', y='Mutation Count', data=filtered_data)
plt.xticks(rotation=45)
plt.title('Total Number of Variants vs. Cancer Type')
plt.tight_layout()
plt.savefig('total_mutation_vs_cancer_type.svg', format='svg')
plt.close()

# Gene-based analysis
# Mutation Count per Gene
gene_mutation_counts = data['Gene'].value_counts()
gene_mutation_counts_df = gene_mutation_counts.reset_index()
gene_mutation_counts_df.columns = ['Gene', 'Mutation Count']

# Plot 3: Mutation Count vs. Gene
plt.figure(figsize=(12, 6))
sns.barplot(x='Gene', y='Mutation Count', data=gene_mutation_counts_df, palette='colorblind')
plt.title('Mutation Count vs. Gene')
plt.xticks(rotation=45)
plt.savefig('mutation_count_vs_gene.svg', format='svg')
plt.close()
```

APPENDIX 3. NGS PrimerPlex Primer Sequences for *CHEK2*

Left Primer Sequence	Right Primer Sequence	Amplicon Start – Hg37 (chr 22)	Amplicon End – Hg37 (chr 22)	Left Primer	Right Primer	Left Primer	Right Primer	Left GC (%)	Right GC (%)
TTCTGTCCTCTGTCTCATGTCTCT	AGAAGAAGCCTTAAGACACCCG	29089941	29090050	64	64	24	22	46	50
ATGCTCAGAGAAAGGTGGATAC	AGAAATCTTTATATCTGAATGCCACTGA	29091200	29091304	61	62	22	28	45	32
CTGAACAAGAATCTACAGGAATAGCC	CGTGACTTAAAGCCAGAGAATGTTTTA	29092839	29092948	63	64	26	27	42	37
CTTTATAAGACAGTCCTCTTCTTGA	GTCATGCCTGCCTTTCTGTGT	29092889	29092998	60	65	25	21	36	52
CTGTCAAAAGAATTGAGGGCTTC	TTTGACAAAGTGGTGGGGAATAAAC	29095797	29095906	62	64	23	25	43	40
CAGGTAGCTTCTTTCAGGC	CCCAGGAATGAACCCCTT	29095863	29095963	60	60	19	18	53	56
GCCTACATTAGATTCTTTGGTGGCT	TGATGCAGAAGATTATTATATTGTTTTGG	29099417	29099525	65	61	25	29	44	28
AATTCCAAAACAATATAATAATCTTCTGC	CTTGGGAGTTTCTCACTACTTTCCCTTTT	29099493	29099601	59	67	29	29	24	41
CTGAAAGGCTTTATACTCTTCTCATAT	TGAAACAGAAATAGAAATTTTGAAAAAGC	29105922	29106030	60	61	27	29	33	24
ATTTCTGTTTCAACATTGAGAGC	CAACTGAGTTTAACTGTAAATGTTTTT	29106019	29106128	60	59	23	27	35	26
GTGATCAGCCTTTTATTGGTACT	TCGAGAGGAAAACATGTAAGAAAGT	29107871	29107976	60	62	23	25	39	36
CCTTTTGCTGATGATCTTTATGGCT	GACAATCACTATCTTTGTTTTTCCCT	29107927	29108035	64	62	25	26	40	35
AAACCACCAATCACAAATGTATAGTG	CCTAAGGCATTAAGAGATGAATACA	29115320	29115429	62	60	26	25	35	36
TAATTTACCTTCCAAGAGTTTTTTGACAT	TTCTTCCTTAGTTTTTGTCTTTTTTGAT	29115375	29115484	62	61	28	28	29	25
CTCTTAATGCCTTAGGATAAACTGACT	CCTCTTGTGAATAAATTAATGAAACCC	29115414	29115523	62	61	27	27	37	33
TGTAAGGACAGGACAAATTTTCC	TGAAATTGCACTGTCACTAAGC	29120887	29120996	61	61	23	22	39	41
ATAATATTACCTTTATTTCTGCTTAGTGA	GGCAATGGAACCTTTGTAAATACA	29120955	29121064	59	62	29	24	24	38
GCGTTTTCCTTTCCCTACAAG	GTTCTCTATTTTAGGAAGTGGGTC	29121017	29121126	61	60	21	24	48	42
TTCCATTGCCACTGTGATCT	CTATAAATCTCTGCTATTCAAAGTCTGA	29121055	29121162	61	61	20	28	45	32
ACAATGACCAAATTACCAGCTCT	TGCTGAAAAGAACAGATAAATACCGA	29121180	29121289	62	63	23	26	39	35
CCGAAAGTGTTTCTTGCTGTATG	TGAATGACAACTACTGGTTTGGG	29121240	29121349	62	63	23	23	43	43
TCAGCAGTGGTTCATCAAAGC	GCCCTCTGATGCATGCTTTTA	29121284	29121384	63	63	21	21	48	48
ACGTGATACTATACAACAAAGGGT	CTGAGGACCAAGAACCTGAG	29130362	29130471	61	61	24	20	38	55
GGGGCAGGGGTAGGCTC	CTCTGGGACACTGAGCTCCTTAGA	29130435	29130542	66	67	17	24	76	54
GGAATAGAATAGAGTTCCTGAGTG	CCAGCTCCTCTACCAGCACG	29130486	29130594	60	66	24	20	42	65
GAGGACTGGCTGGAGTTT	CTGTTCACAGCCCCATGG	29130552	29130659	60	62	18	18	56	61
GGAGCCTTGGGACTGG	TTCTTTTTGAGGTCGTGATGTCTC	29130617	29130726	60	63	16	24	69	42
CACTGCTGCCATGAGACTGCT	TTTTAAGTCTTGTGCCTTGAAACTCAC	29130661	29130770	67	64	21	27	57	37

APPENDIX 4. MPPrimer Primer Sequences for *CHEK2*

ForwardPrimer(Fp)	ReversePrimer(Rp)	FpTm (°C)	RpTm (°C)	Fp Length	Rp Length	FpGC (%)	RpGC (%)	Start (chr22) – Hg38	End (chr22) – Hg38
tccacctggtaataCAACTTTCTGT	TCCAGCCAGTCCTCTCACTC	60.48	61.13	25	20	40	60	28734353	28734557
ACTGGGTAACGCTGCCATG	TTGGAGAGCATGTTTGCCCT	60.88	60.41	19	20	57	50	28734657	28734839
GCACAAAGCCCAGGTTCCAT	CTACTAGTCGAAAGCGGCCC	61.71	60.75	20	20	55	60	28687752	28687962
gaggcagaggctacagttagc	ACCTACTTGGCGCCTGAAGT	60.72	62.01	21	20	57	55	28695602	28695784
GACAGGACAAATTTTCCTCCTATGAGA	GTCTGAAACAAAATGTTCTCTATTTTAGGA	61.36	59.58	27	30	40	30	28724930	28725122
CAATGACCAAATTACCAGCTCTCCTA	CGTTTGATACATGAAATTCAACAGCC	61.41	60.50	26	26	42	38	28725217	28725393
GGGAAGTTATGAAGACGTGTTAATAAAAGG	AGGAGTGGTAGGTCTCATAATTAAAAACAT	61.58	61.32	30	30	36	33	28711882	28712050
CAGAAATGAGAAACCACCAATCACA	TCAGTGATCGCCTCTTGTGAATAA	60.53	60.62	25	24	40	41	28719345	28719521
GTTTCTGTCCTCTGTCTCATGTCTC	CTGTTAATTCTGGCATACTCTTACTGA	61.39	59.56	25	27	48	37	28693974	28694164
GGCCCCACTACTACATACATACG	CTTGTGGTTTTCCTCTTGGGAGT	60.58	61.23	23	23	52	47	28703399	28703604
TCCCCCTTCAGCTCAAAACTAC	GGTTTTGATTGGCTGAGGGTG	60.49	60.24	22	21	50	52	28741671	28741846
TTCCTGAGTGGACACTGTCTCT	ACTCACCTTTGTTGTTGGACAC	60.96	59.76	22	22	50	45	28734532	28734739
ttcattcaggctgggcaacag	TCCGTGGTTTGAACACGAAAGA	61.69	61.21	21	22	52	45	28687664	28687872
ACAGCAGCACACACAGCT	CTGTGGTGAGGACTCAGTTGTC	60.30	61.07	18	22	55	54	28687920	28688129
CTCCTAAACTCCAGCAGTCCAC	ACTGATCTAGCCTACGTGTCTTC	60.88	59.90	22	23	54	47	28695746	28695948
GCGTTTTCCTTTCCCTACAAGC	TACCCATGTATCTAGGAGAGCTGG	60.87	61.05	22	24	50	50	28725049	28725206
GAAGAAACTCCCACCACAGCA	ATGCCACTGAGAATGCCACTT	61.00	60.80	21	21	52	47	28695076	28695277
GTTTTAATCCACGGTCCCTCGA	CGTTTTCCCCCAGGAATGAAC	60.88	59.97	22	21	50	52	28699764	28699962
GTTTCTGAACAAGAATCTACAGGAATAGC	GTGAGATGTGTGTGTTGGTAACTTTG	61.19	61.49	29	26	37	42	28696874	28697044
CTTATCAGCTCCTTAAGCCCAGA	TTGGACGGACATTTTTCCTCCC	60.15	61.35	23	22	47	50	28689101	28689302
GAAAGGCTTTATACTCTTCTCATATTTTGA	GTAGAGCTGGGTTTGGAACTCAG	59.09	61.40	30	23	30	52	28709964	28710152

APPENDIX 5. Analysis for the *CHEK 2* Primers from MPPrimer

ForwardPrimer(Fp)	ReversePrimer(Rp)	Dimer Score (Fp)	Dimer Score (Rp)
tccacctggtaataCAACTTTCTGT	TCCAGCCAGTCCTCTCACTC	8	11
ACTGGGTAACGCTGCCATG	TTGGAGAGCATGTTTGCCCT	9	12
GCACAAAGCCCAGGTTCCAT	CTACTAGTCGAAAGCGGCCC	11	8
gaggcagaggctacagttagc	ACCTACTTGGCGCCTGAAGT	9	9
GACAGGACAAATTTTCCTCCTATGAGA	GTCTGAAACAAAATGTTCTCTATTTTAGGA	9	9
CAATGACCAAATTACCAGCTCTCCTA	CGTTTGATACATGAAATTCAACAGCC	19	11
GGGAAGTTATGAAGACGTGTTAATAAAAGG	AGGAGTGGTAGGTCTCATAATTAAAAACAT	9	9
CAGAAATGAGAAACCACCAATCACA	TCAGTGATCGCCTCTTGTGAATAA	10	9
GTTTCTGTCCTCTGTCTCATGTCTC	CTGTTAATTCTGGCATACTCTTACTGA	10	9
GGCCCCACTACTACATACATACG	CTTGTGGTTTTCCTCTTGGGAGT	8	11
TCCCCCTTCAGCTCAAAACTAC	GGTTTTGATTGGCTGAGGGTG	10	10
TTCCTGAGTGGACACTGTCTCT	ACTCACCTTTGTTGTTGGACAC	10	10
ttcattcaggctgggcaacag	TCCGTGGTTTGAACACGAAAGA	11	11
ACAGCAGCACACACAGCT	CTGTGGTGAGGACTCAGTTGTC	12	13
CTCCTAAACTCCAGCAGTCCAC	ACTGATCTAGCCTACGTGTCTTC	10	9
GCGTTTTCCTTTCCCTACAAGC	TACCCATGTATCTAGGAGAGCTGG	9	19
GAAGAAACTCCCACCACAGCA	ATGCCACTGAGAATGCCACTT	13	9
GTTTTAATCCACGGTCCCTCGA	CGTTTTCCCCCAGGAATGAAC	11	9
GTTTCTGAACAAGAATCTACAGGAATAGC	GTGAGATGTGTGTGTTGGTAACTTTG	9	12
CTTATCAGCTCCTTAAGCCCAGA	TTGGACGGACATTTTTCCTCCC	10	9
GAAAGGCTTTATACTCTTCTCATATTTTGA	GTAGAGCTGGGTTTGGAACTCAG	9	13

APPENDIX 6. Analysis for the *CHEK2* Primers from NGS PrimerPlex

ForwardPrimer(Fp)	ReversePrimer(Rp)	Max Dimer Score (Fp)	Max Dimer Score (Rp)
TTCTGTCCTCTGTCTCATGTCTCT	AGAAGAAGCCTTAAGACACCCG	12	10
ATGCTCAGAGAAAGGTGGATAC	AGAAATCTTTATATCTGAATGCCACTGA	10	10
CTGAACAAGAATCTACAGGAATAGCC	CGTGACTTAAAGCCAGAGAATGTTTTA	11	10
CTTTATAAGACAGTCCTCTTCTTGA	GTCATGCCTGCCTTTCTGTGT	13	11
CTGTCAAAAGAATTGAGGGCTTC	TTTGACAAAGTGGTGGGGAATAAAC	11	10
CAGGTAGCTTCTTTCAGGC	CCCAGGAATGAACCCCTT	10	10
GCCTACATTAGATTCTTTGGTGGCT	TGATGCAGAAGATTATTATATTGTTTTGG	11	32
AATTCCAAAACAATATAATAATCTTCTGC	CTTGGGAGTTTCTCACTACTTTCCCTTTT	32	10
CTGAAAGGCTTTATACTCTTCTCATAT	TGAAACAGAAATAGAAATTTTGAAAAAGC	11	15
ATTTCTGTTTCAACATTGAGAGC	CAACTGAGTTTAACTGTAAATGTTTTT	15	9
GTGATCAGCCTTTTATTGGTACT	TCGAGAGGAAAACATGTAAGAAAGT	10	11
CCTTTTGCTGATGATCTTTATGGCT	GACAATCACTATCTTTGTTTTTCCCT	9	9
AAACCACCAATCACAAATGTATAGTG	CCTAAGGCATTAAGAGATGAATACA	11	23
TAATTTACCTTCCAAGAGTTTTTGACAT	TTCTTCCTTAGTTTTTGTCTTTTTTGAT	10	9
CTCTTAATGCCTTAGGATAAACTGACT	CCTCTTGTGAATAAATTAATGAAACCC	23	9
TGTAAGGACAGGACAAATTTTCC	TGAAATTGCACTGTCACTAAGC	11	13
ATAATATTACCTTTATTTCTGCTTAGTGA	GGCAATGGAACCTTTGTAAATACA	13	15
GCGTTTTCCTTTCCCTACAAG	GTTCTCTATTTTAGGAAGTGGGTC	11	9
TTCCATTGCCACTGTGATCT	CTATAAATCTCTGCTATTCAAAGTCTGA	15	10
ACAATGACCAAATTACCAGCTCT	TGCTGAAAAGAACAGATAAATACCGA	9	9
CCGAAAGTGTTTCTTGCTGTATG	TGAATGACAACTACTGGTTTGGG	8	8
TCAGCAGTGGTTCATCAAAGC	GCCCTCTGATGCATGCTTTTA	13	11
ACGTGATACTATACAACAAAGGGT	CTGAGGACCAAGAACCTGAG	10	10
GGGGCAGGGGTAGGCTC	CTCTGGGACACTGAGCTCCTTAGA	11	10
GGAATAGAATAGAGTTCCTGAGTG	CCAGCTCCTCTACCAGCACG	9	9
GAGGACTGGCTGGAGTTT	CTGTTCACAGCCCCATGG	13	10
GGAGCCTTGGGACTGG	TTCTTTTTGAGGTCGTGATGTCTC	10	8
CACTGCTGCCATGAGACTGCT	TTTTAAGTCTTGTGCCTTGAAACTCAC	13	10

APPENDIX 7. Python Script for the Coverage Analysis of Primers

```
# Importing necessary libraries
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns

# Extracting Information
def calculate_overall_coverage(directory):
    total_coverage_percentages = []
    for file_name in os.listdir(directory):
        file_path = os.path.join(directory, file_name)
        if file_path.endswith(".txt"):  # Ensure processing only text files
            try:
                df = pd.read_csv(file_path, delimiter='\t')
                df['CovRate'] = df['CovRate'].str.rstrip('%').astype(float)
                # Calculate average coverage rate for the file and add to list
                average_coverage_rate = df['CovRate'].mean()
                total_coverage_percentages.append(average_coverage_rate)
            except Exception as e:
                print(f"Error processing {file_name}: {e}")
    if total_coverage_percentages:
        overall_average_coverage = sum(total_coverage_percentages) / len(total_coverage_percentages)
        return overall_average_coverage
    else:
        return 0

# Usage
directory = 'path_to_extracted_files'  # directory path
overall_coverage_percentage = calculate_overall_coverage(directory)
print(f"Overall Coverage Percentage: {overall_coverage_percentage:.2f}%")

# Preparing the plot
plt.figure(figsize=(14, 8))
bar_plot = sns.barplot(x=genes, y=average_covs, palette='colorblind')
plt.xticks(rotation=45, ha='right', fontsize=18, fontname='Times New Roman', fontstyle='italic')
plt.yticks([0, 0.2, 0.4, 0.6, 0.8, 1], ['0%', '20%', '40%', '60%', '80%', '100%'], fontsize=14)
plt.xlabel('Gene', fontsize=16)
plt.ylabel('Average Coverage Percentage', fontsize=16)
plt.title('Average Coverage Percentage by Gene', fontsize=20, fontname='Times New Roman')
plt.ylim(0, 1)

# Applying font settings manually to each xtick label
for label in bar_plot.get_xticklabels():
    label.set_fontname('Times New Roman')
    label.set_fontstyle('italic')
    label.set_fontsize(18)

# Saving as SVG
overall_svg_path_percent_with_text_updated =
'/mnt/data/overall_coverage_percent_with_text_updated.svg'
plt.savefig(overall_svg_path_percent_with_text_updated, format='svg')
plt.close()
```

APPENDIX 8. Python Script for the *In-silico* Downstream Primer Analysis

```
import pandas as pd
from itertools import combinations

def reverse_complement(seq):
    complement = {'A': 'T', 'T': 'A', 'G': 'C', 'C': 'G'}
    return ''.join([complement[base] for base in seq[::-1].upper()])

def check_structure(sequence):
    reversed_seq = sequence[::-1]
    hairpin = sequence in reversed_seq or reversed_seq in sequence
    return "Yes" if hairpin else "No"

def advanced_dimer_score(primer1, primer2):
    primer2_rc = reverse_complement(primer2)
    max_score = 0
    for i in range(len(primer1)):
        for j in range(len(primer2_rc)):
            length = 0
            gc_count = 0
            while (i + length < len(primer1)) and (j + length < len(primer2_rc)) and (primer1[i + length]
== primer2_rc[j + length]):
                if primer1[i + length] in 'GC':
                    gc_count += 1
                length += 1
            score = length + gc_count
            max_score = max(max_score, score)
    return max_score

def analyze_primers(excel_file_path, output_file_path):
    primers_df = pd.read_excel(excel_file_path)
    primers_df['Fp Hairpin'] = primers_df['ForwardPrimer(Fp)'].apply(check_structure)
    primers_df['Rp Hairpin'] = primers_df['ReversePrimer(Rp)'].apply(check_structure)
    primer_sequences = list(primers_df['ForwardPrimer(Fp)']) + list(primers_df['ReversePrimer(Rp)'])
    dimer_scores = {}
    for primer1, primer2 in combinations(primer_sequences, 2):
        score = advanced_dimer_score(primer1, primer2)
        if score > 0:
            dimer_scores[(primer1, primer2)] = score

    for index, row in primers_df.iterrows():
        fp = row['ForwardPrimer(Fp)']
        rp = row['ReversePrimer(Rp)']
        fp_scores = [score for (p1, p2), score in dimer_scores.items() if p1 == fp or p2 == fp]
        rp_scores = [score for (p1, p2), score in dimer_scores.items() if p1 == rp or p2 == rp]
        primers_df.at[index, 'Max Dimer Score (Fp)'] = max(fp_scores) if fp_scores else 0
        primers_df.at[index, 'Max Dimer Score (Rp)'] = max(rp_scores) if rp_scores else 0

    primers_df.to_excel(output_file_path, index=False)
    return primers_df

if __name__ == "__main__":
    input_file_path = '/file/path/the_input.xlsx'  # file path
    output_file_path = '/file/path/the_output.xlsx'  # output file path
    result_df = analyze_primers(input_file_path, output_file_path)
    print(f"Results exported to {output_file_path}")
```

APPENDIX 9. Example VCF Info Column from Analysis Pipeline

INFO
AS_FilterStatus=SITE;AS_SB_TABLE=83,77|27,21;DP=236;ECNT=2;GERMQ=93;MBQ=20,20;MFRL=152,186;MMQ=60,60;MPOS=38;POPAF=7.30;RPA=8,7;RU=T;STR;STRQ=14;TLOD=38.19;DET=Yes;DETS=2.837792926593422
AS_FilterStatus=SITE;AS_SB_TABLE=75,75|26,28;DP=209;ECNT=1;GERMQ=93;MBQ=20,20;MFRL=130,149;MMQ=60,60;MPOS=45;POPAF=7.30;TLOD=114.02;DET=Yes;DETS=2.8904845167880007
AS_FilterStatus=SITE;AS_SB_TABLE=114,125|2,2;DP=254;ECNT=1;GERMQ=93;MBQ=20,26;MFRL=152,104;MMQ=60,60;MPOS=12;POPAF=7.30;TLOD=4.77;DET=No;DETS=1.7323937598229686
AS_FilterStatus=SITE;AS_SB_TABLE=77,77|8,8;DP=205;ECNT=1;GERMQ=93;MBQ=20,20;MFRL=142,158;MMQ=60,60;MPOS=31;POPAF=7.30;TLOD=25.01;DET=Yes;DETS=2.688775655272845
AS_FilterStatus=SITE;AS_SB_TABLE=81,92|29,33;DP=260;ECNT=1;GERMQ=93;MBQ=20,20;MFRL=143,132;MMQ=60,60;MPOS=41;POPAF=7.30;RPA=6,5;RU=G;STR;STRQ=93;TLOD=48.02;DET=Yes;DETS=2.88779283759875

This table presents an example of the VCF (Variant Call Format) info column resulting from the Analysis Pipeline, with vLoD (variant Level of Detection) results incorporated. Abbreviations are listed below.

AS_FilterStatus: Allele-Specific Filter Status
AS_SB_TABLE: Allele-Specific Strand Bias Table
DP: Depth; total number of reads covering the position.
ECNT: Event Count
GERMQ: Genotype Quality for Error Model
MBQ: Median Base Quality
MFRL: Median Fragment Length
MMQ: Median Mapping Quality
MPOS: Median Position
POPAF: Population Allele Frequency Prior
RPA: Repeat Primitive Allele
RU: Repeat Unit
STR: Short Tandem Repeat
STRQ: Short Tandem Repeat Quality
TLOD: Tumor LOD Score; Log odds ratio of the data supporting the variant in tumor versus normal (From Mutect2)
DET: Detectability Condition from vLoD; Determines if detectability threshold is passed (“Yes”) or not (“No”)
DETS: Detectability Score from vLoD; Gives the odds of a true positive (correct detection of a variant) against the probability of errors (both sequencing errors and false positives)

Results coming from the vLoD algorithm are highlighted in light green.

9 CURRICULUM VITAE