IDENTIFICATION OF GENE MUTATIONS INVOLVED IN DRUG RESISTANCE IN LIVER CANCER USING RNA-SEQ DATA ANALYSIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
MIDDLE EAST TECHNICAL UNIVERSITY
BY

MONA SHOJAEI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

SEPTEMBER 2016

IDENTIFICATION OF GENE MUTATIONS INVOLVED IN DRUG RESISTANCE IN LIVER CANCER USING RNA-SEQ DATA ANALYSIS

Submitted by MONA SHOJAEI in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son _____
Head of Department, **Health Informatics**

Assoc. Prof. Dr. Rengül Çetin Atalay _____
Supervisor, **Health Informatics**

Assist. Prof. Dr. Aybar Can Acar _____
Co-Supervisor, **Health Informatics**

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son _____
Health Informatics, Middle East Technical University

Assoc. Prof. Dr. Rengül Çetin Atalay _____
Health Informatics, Middle East Technical University

Assist. Prof. Dr. Aybar Can Acar _____
Health Informatics, Middle East Technical University

Assist. Prof. Dr. Can Alkan _____
Computer Engineering, Bilkent University

Assoc. Prof. Dr. A.Elif Erson Bensan _____
Molecular Biology, Middle East Technical University

Date: **05.09.2016**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this wok.

Name, Last name : MONA SHOJAEI

Signature : _____

ABSTRACT

IDENTIFICATION OF GENE MUTATIONS INVOLVED IN DRUG RESISTANCE IN LIVER CANCER USING RNA-SEQ DATA ANALYSIS

Shojaei, Mona

MSc., Department of Bioinformatics

Supervisor: Assoc. Prof. Dr. Rengül Çetin Atalay

Co-Supervisor: Assist. Prof. Dr. Aybar Can Acar

September 2016, 68 pages

A significant concern in cancer research is the detection of cancer associated somatic mutations. Liver cancer is the 5th most common and 2nd deadliest cancer in the world. Several somatic mutations were previously reported in liver cancer but their relations to chemotherapeutic response was not studied in detail. In this study, the relationship between mutation status and drug treatment response of well-differentiated Huh7 and poorly-differentiated Mahlavu liver cancer cells were analyzed. The RNA-Seq data of each cancer cell line (as control) was compared to "sorafenib" and "PI3K/Akt Pathway inhibitors" treated data. Somatic mutations associated with drug resistance were comparatively identified with MuTect tool. The results were then filtered to distinguish the missense mutations. The common genes among drug-resistant sets were found to be associated with liver cancer perseverance and aggressiveness. SLC39A5, FRG1, PPHLN1 and SRP9 gene mutations were found to be the most significant, shared among three drug treated sets. The sets were further investigated in detail to discover the liver cancer associated survival genes. Using our results, appropriate targets can be defined that play critical roles in cancerous cell growth for drug development purposes.

Keywords: mutation detection, liver cancer, Mutect, drug resistance, RNA-Seq

ÖZ

KARACİĞER KANSERİNDE İLACA DİRENÇLİLİK MÜTASYONLARININ RNA-DİZİ ANALİZİ İLE BULUNMASI

Shojaei, Mona

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Doç. Dr. Rengül Çetin Atalay

Ortak Tez Yöneticisi: Yrd. Doç. Dr. Aybar Can Acar

September 2016, 68 sayfa

Kanser araştırmasında önem kazanmış konulardan biri, kansere neden olan somatik mutasyonların tespitidir. Günümüzde besinci en yaygın görülen ve ikinci en ölümcül kanser, karaciğer kanseridir. Karaciğer kanseri ile ilişkili pek çok somatik mutasyon belirlenmiştir; ancak bunların kemoterapatik ilaçlara tepkileri detaylı bir şekilde çalışılmamıştır. Bu çalışmada iyi huylu Huh7 ve habis Mahlavu karaciğer kanseri hücrelerinin mutasyon durumlarının ilaca verdikleri tepkiyle ilişkisi incelenmiştir. Her iki kanser hücre hattının muamele görmemiş kontrol RNA-sekans verisi, "sorafenib" ve "PI3K/Akt Yolak inhibitörü" ile muamele görmüş RNA sekans verileriyle karşılaştırılmıştır. MuTect aracı kullanılarak ilaca dirençliliğe neden olan somatik mutasyonlar karşılaştırmalı olarak belirlenmiştir. Elde edilen sonuçlar, yanlış anlam mutasyonlarını ayırt etmek için filtrelenmiştir. İlaca dirençlilik sağlayan gen setleri arasında ortak olan genlerin karaciğer kanserinin direnç ve agresifliğiyle ilişkili olduğu bulunmuştur. Bu bağlamda en anlamli mutasyonlarin SLC39A5, FRG1, PPHLN1 ve SRP9 gen mutasyonlari olduğu gözlenmiştir. Bu mutasyonlar üç farkli ilaçla muamele edilmiş setlerde ortak olarak belirlenmiştir. Karaciğer kanseri ile ilişkili sağkalım genlerinin keşfedilmesi amacıyla bu setler daha detaylı incelenmiştir. Bu çalışmanın sonuçları kullanılarak kanser hücre gelişimi için önemli olan hedefler belirlenebilir ve bu bilgi karaciğer kanseri için ilaç geliştirme alanında kullanılabilir

Anahtar Sözcükler: mutasyon tespiti, Karaciğer kanseri, Mutect, ilaca dirençlilik, RNA

To My Family and Those Who Believed in Me

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Prof. Rengül Atalay for her continuous support of my Master study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, my sincere thanks also goes to my co-advisor Prof. Aybar C. Acar for his encouragement, insightful comments, and hard questions, which led me to success.

I would also like to thank my committee members, professor Yeşim Aydın Son, professor Can Alkan, professor A.Elif Erson Bensan for serving as my committee members. I also want to thank them for letting my defense be an enjoyable moment, and for their brilliant comments and suggestions.

My particular thanks goes to Dr. Tunca Doğan because of his insightful comments in writing this thesis.

My deepest gratitude goes to CanSyl lab for providing me the data and useful analysis. I thank my fellow group mates Kübra Narcı and Damla Gözen who helped me in editing and translation and criticized me in writing.

In particular I would like to show my gratitude to my roommate Ezgi Yavuzyılmaz who gave me her unconditional support during these years.

I wish to present my huge thanks and love to Navid Mohammadvand who helped me a lot in different branches during this thesis. Thank you for your encouragement and all your supports.

Special thanks to my family, Words cannot express how grateful I am to my mother, and father for all of the sacrifices that they have made on my behalf. Your prayer for me was what sustained me thus far.

This work was supported by TÜBİTAK grant #110S388.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HCC	**Hepatocellular Carcinoma**
HHCM	**Mahlavu Hepatocellular Carcinoma**
HDMSO	**Huh7's Control Group**
MDMSO	**Mahlavu's Control Group**
PI3K	**Phosphoinositide 3-kinase**
PIP2	**Phosphatidylinositol biphosphate**
PIP3	**Phosphatidylinositol triphosphate**
SOR	**Sorafenib**
PI3Kα	**PI3K's α isoform**
PI3Kβ	**PI3K's β isoform**
H_ALPHA	**Huh7 treatment by inhibiting PI3K's α isoform**
H_BETA	**Huh7 treatment by inhibiting PI3K's β isoform**
H_S_ALPHA	**Huh7 treatment by inhibiting PI3K's α isoform+ Sorafenib**
H_S_BETA	**Huh7 treatment by inhibiting PI3K's β isoform+ Sorafenib**
HSOR	**Huh7 treatment by Sorafenib**
M_ALPHA	**Mahlavu treatment by inhibiting PI3K's α isoform**
M_BETA	**Mahlavu treatment by inhibiting PI3K's β isoform**
M_S_ALPHA	**Mahlavu treatment by inhibiting PI3K's α isoform+ Sorafenib**
M_S_BETA	**Mahlavu treatment by inhibiting PI3K's β isoform+ Sorafenib**

MSOR	**Mahlavu treatment by Sorafenib**
SNV	**Single Nucleotide Variant**
dbSNP	**Single Nucleotide Polymorphism Database**
COSMIC	**Catalogue of Somatic Mutations in Cancer**
LOD	**Log Odds**
GATK	**Genome Analysis Toolkit**

CHAPTER 1

INTRODUCTION

1.1 Motivation

Cancer term refers to the normal cell transformation into a tumor one. Having abnormal growth, the tumor cells can proliferate beyond the boundaries and affect nearby tissues. Cancerous cells can also spread through bloodstream or lymph to other parts of the body. Cancer is the leading cause of death worldwide and the risk increases with aging. Although the biological, chemical and physical factors trigger the disease, the exact reason of abnormality is still unknown. Liver cancer, known to be the sixth most common cancer, causes a high number of deaths each year. The most common type of liver cancer is Hepatocellular Carcinoma and some factors like hepatitis C and hepatitis B viruses increase the risk of the disease (Vale, 2013).

Appropriate treatment of hepatocellular carcinoma (HCC, primary liver cancer) depends on the disease stage, patient's age and overall health and individual priorities. There are many different treatment strategies with high cure rate in early detections. Surgery can be the first solution in early stages in which the tumor size is reduced by using chemotherapy drugs before the operation. However, the treatment is ineffective in advanced cases and the liver tumor resists to chemotherapy drugs. Radiation is the other method, which can be recommended if the liver is devoid from any disease like liver cirrhosis. However, deficiencies and tumor distinction in progressive stages causes fail in treatments. In addition to chemotherapy, different targeted therapies are used for treatment of liver cancer. Targeted therapy is a newer method, which blocks cancer associated proteins or prevents new blood vessels' formation. Unlike chemotherapy, which affects the whole normal and cancerous fast-growing cells, targeted drugs attack to specific molecules in cancer cells and have much less impact on healthy tissues. Sorafenib (GM & Santoro, 2009) is one of those targeted drugs, which is considered as an efficient treatment in liver cancers. Targeted cancer therapies can be used alone or in combination with other treatments such as chemotherapy, radiotherapy and surgery.

The validity of targeted therapy depends on finding appropriate targets, playing a critical role in cancerous cells growth and survival. In order to identify the optimal targets, different approaches focus on proteins with different expression levels in normal and cancer cells, search for chromosomes abnormalities and investigate the mutations that change the type of produced amino acids in tumors. As a result, the good target identification leads to development of new therapies, which dedicatedly act on specific targets and provide better treatment for HCC patients.

1.2 Scope and Goal

Recently, significant achievements have been acquired in liver cancer treatment and many somatic mutations have been detected relating to this fatal cancer. In spite of that, the effect of mutations on chemotherapeutic responses and the HCC resistance mechanism is left off. In addition, most of the related studies have been performed with DNA data. Having look from a different aspect, our aim is using RNA-Seq data to investigate the relation between somatic mutations and treatment compounds. For this purpose, the effective genes in HCC progression are identified using the chemical knockdown. Indeed, the resistant genes alter in response to different inhibitors and continue the activity to provide the vital substances for cancer growth. The investigations in mutations' type and repetition lead to distinguish the most significant genes, which involve in cancer. The missense mutations are considerable for protein coding different from the unaffected genes. In addition, the genes with missense mutations, which are present in most of the HCC treated samples, possess higher resistance to applied treatments. Passing through all considered filters, the remained genes are eligible to be defined as biomarkers for future treatments.

The main objective of this study is to analyze the mutated regions, their location on chromosomes and related gene expression in drug treated HCC samples with the concern of finding resistant genes, which assist tumor tissue in abnormal cell growth to guarantee the cancer survival.

1.3 Contribution

In order to make a precise selection of significant mutations, we preferred Mutect (Cibulskis et al., 2013), which is known to be a sensitive method in mutation detection. The exonic regions of two liver cancer cell lines, Huh7 as well-differentiated and Mahlavu as poorly-differentiated are selected for further analysis (Yuzugullu et al., 2009). The cell lines are then treated with "PI3K/Akt Pathway inhibitors", "sorafenib" and their combination to examine the behavior of cancer cells under different treatments. Mutect comparison results for control (untreated) and treated HCC samples reveal

various alterations in response to the applied treatments. Detected regions are indeed representing mutated sites, which alter to escape from drugs' therapeutic effects. Continually, we restricted the analysis domain to missense mutations, applying filters to ineffective mutations. The obtained missense mutations are significant variations, involved in cancer growth by altering and producing new proteins.

Investigating the response of cancerous cells to different chemical compounds clarifies the genes, which involve in cancer perseverance. The remarkable genes with higher survival rank in HCC treatments can be further studied for therapeutic purposes. It would be helpful to consider different perspectives to get vital genes of cancer and filter out the rest. A key point here is to separate driver genes from passengers. Drivers contribute to cancer progression, where passengers are generally known to be neutral as they change from patient to patient. However, it is verified that the sufficient accumulation of passenger genes can damage the tumor and play a role in cancer analysis. The change in gene expression level of cancerous cells is the other issue, worth investigating. The genes activation or silencing may occur in tumor tissue. Tumor suppressor genes, which control the abnormal cell growth, can be inactivated. Subsequently, their affected genes can express in high amount to trigger cancer. The expression level evaluation leads to quantify the target genes effect in disease process. The last important thing is referring to previous studies about detected genes to verify their importance.

The ability to identify the effective genes in cancerous cell growth and continually reach the new cancer targets is a great achievement for new targeted-drug design. Drugs with specific targets can find the appointed place and turn the target gene off to disturb the cancer cells' proliferation. Accordingly, the mutated genes activities are stopped and the disease progression is prevented.

1.4 Outline

This thesis, which defines new probable targets for novel targeted treatments, comprises 5 chapters. The chapters are entitled "Introduction", "Background and Literature Review", "Materials and Methods", "Results" and "Discussion", respectively.

The first chapter gives a brief description of the issues in the aimed research field and the goal of study, which can be a possible solution for the relating problems. The second chapter includes information about HCC and its mechanism, mutations and the possible treatments of the disease. Further, various methods that are available for detecting mutations in cancer tissue are described in detail. It also goes over important studies and conclusions relating to each topic. In the third part, the selected method for mutation detection (Mutect) is illustrated extensively. Besides comprehensive definition of required datasets for analysis, prerequisites for program executing are given in detail.

The additional steps to provide the optimal list of mutations are presented in the last part of the related chapter. The fourth chapter comes next to represent the examination results. The efficacy of different treatments is evaluated and the survived genes in each group are compared to specify the ones in common. Finally, in the fifth chapter a lot of studies are reviewed to verify the significance of detected mutations and to find out the other important cancers that are in deal with outstanding altered genes.

CHAPTER 2

BACKGROUND AND RELATED WORKS

2.1 Liver Cancer

Liver cancer is the aggregation of cancerous cells that primarily form in the liver or spread to this organ after growing in other parts of the body. Different liver cancer types arise according to their generating cell types.

Hepatocellular carcinoma (HCC), the one rising from hepatocellular cells of the liver, is the most prevalent type of mentioned cancer group, accounting for approximately 75 percent of all liver cancers and being the third cause of cancer deaths worldwide (Altekruse, McGlynn, & Reichman, 2009). There are various factors, which trigger normal liver cells to turn into cancerous ones. Although the existence of all factors doesn't lead to an obligatory illness, they certainly increase the risk of getting the disease. Infection with hepatitis B and hepatitis C viruses, high alcohol consumption and liver cirrhosis are some of the critical elements among these causing factors. Liver can tolerate low alcoholic drinks amount by wiping out the poison and regenerating the injured cells. However, regular drinking diminishes and ruins the capability of regeneration by terminating in liver cirrhosis, which widely puts the liver at the risk of a cancer. In addition, blood-transmitted infection by hepatitis C increases the harmful effects of regular drinking on liver tissue. Same as hepatitis C, hepatitis B can spread by blood or either by the means of other fluids and cause many destructive effects. Furthermore the fat mass accumulation in the liver, which normally does not carry any special harm for the tissue, can lead to liver inflammation and severe scarring in high amount of fat depositing conditions (Figure 1.1) (Dragani, 2010).

HCC has no effective treatment, but striving is for slowing down the tumor growth, decreasing the pain of the disease, and in the best situation disease progression disturbed by stopping cancerous cells activity and getting full recovery of the patient.

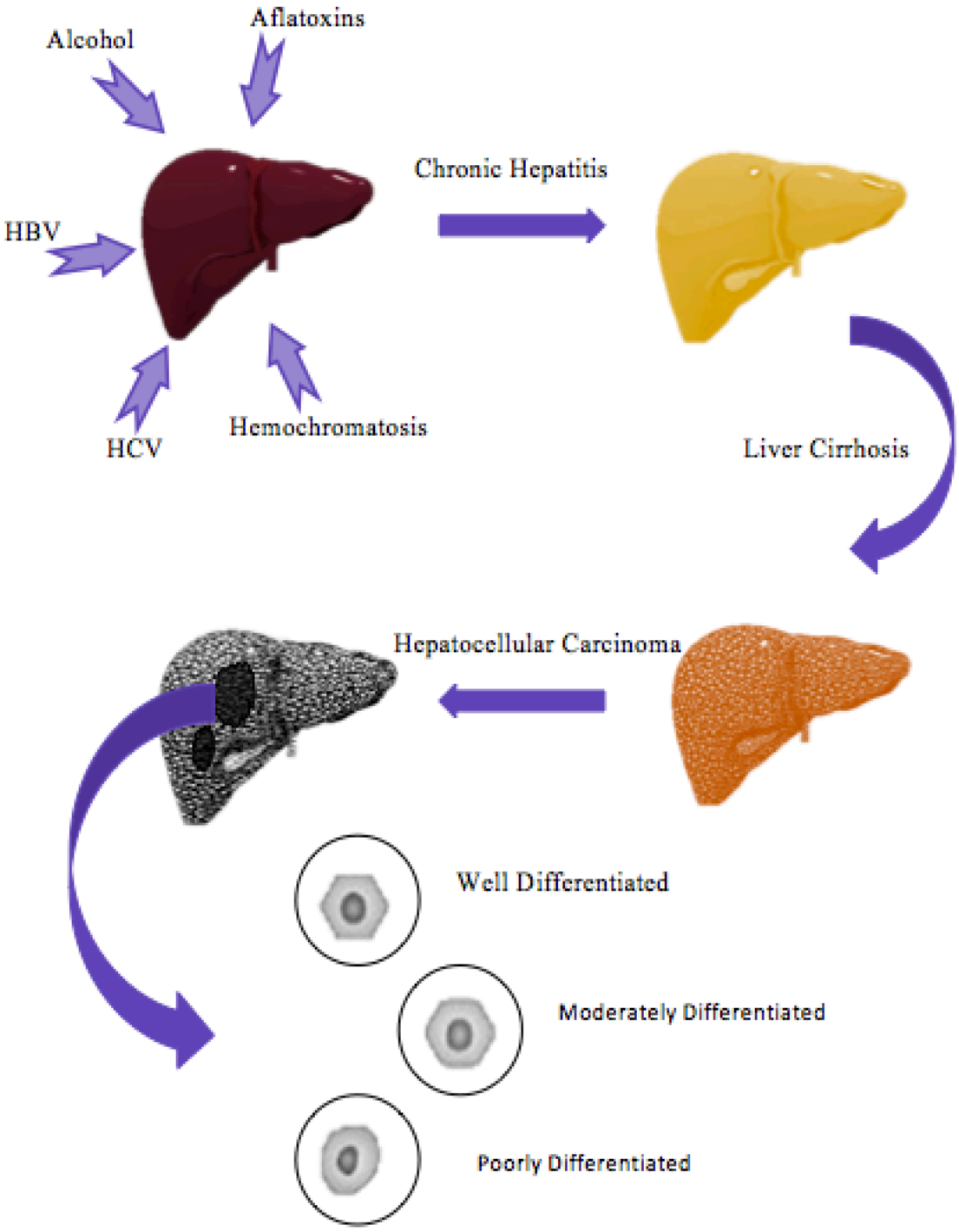


Figure 2-1: Liver canser stages. A) Healthy liver under risk factors, B) Chronic hepatitis, causes the scar tissue by continuous inflammation of liver, C) Cirrhotic liver, the blood flow blocked form as a result of frequent scarring, D) Liver cancer

2.2 Pathways and liver cancer

Pathways such as VEGF, FGF, MAPK and PI3K are a biochemical torrent of regulated reactions inside the cell, which play important roles in liver cancer. Out of all PI3K, for being a hyperactive pathway, holds a significant figure in this disease.

PI3K or Phosphoinositide-3 kinases are the family of lipid kinases with the ability to phosphorylate PIP2 (phosphatidylinositol biphosphate) and produce PIP3 by taking a role in the PI3K pathway. PI3K enzymes are categorized in three classes where the class one includes PI3Kα, PI3Kβ, and PI3Kδ isoforms. Although α isoform is more predominant and PI3Kβ is the latter in terms of getting a role in signaling processes, they both are ubiquitously expressed forms. Furthermore the two isoforms have different activating signals and paths that tyrosine kinase receptors and G protein-coupled receptors lead to activation of one isoform based upon the related signals type and conduction (Wu & Li, 2012).

The PI3K/mTOR pathway is a central oncogenic pathway deregulated in cancer (Fritsch et al., 2014) and affected by some upstream factors that alter primarily. PTEN, the tumor suppressor gene is one of the mutated factors, resulting in disruption of pathway regulation. In fact PTEN regulates PI3K enzyme activities by dephosphorylating their products and keeping them in balance. In case of antagonistic factors elimination, the enzymes switch into their hyperactive form to continue phosphorylating and assisting cancer development (Figure 2.2)

According to Schwartz et al., preventing the PI3K hyperactivity requires inhibition of both isoforms simultaneously since, suppression of one isoform leads to another one's activation. So, disturbing PI3Kα and PI3Kβ function across with drug injection is useful for obtaining more efficient results in drug treatment (Schwartz et al., 2015).

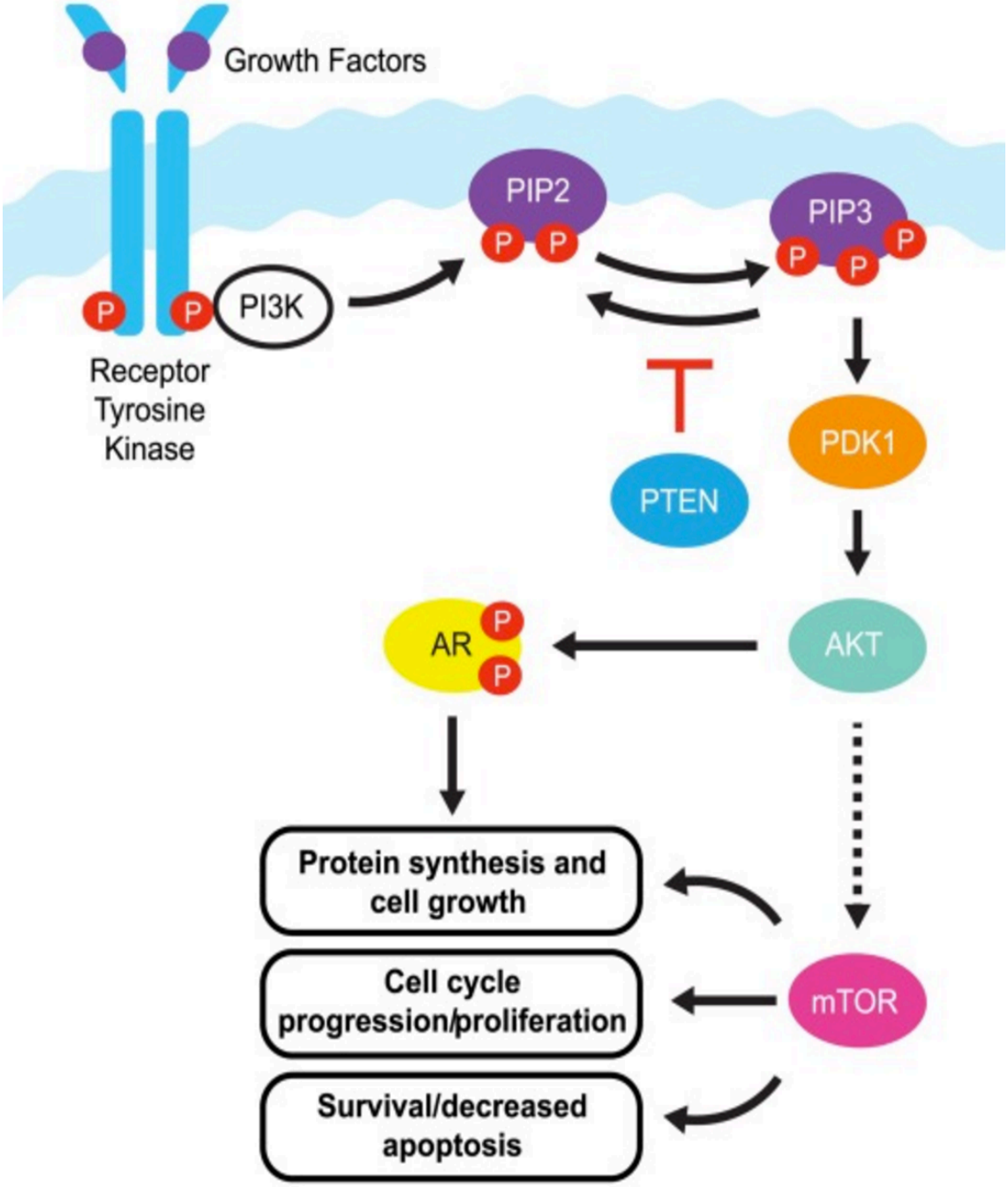


Figure 2-2: The PI3K/PTEN/Akt pathway. PI3K is activated by growth factors and activates mTOR in return to provide the cell survival. PTEN disturbs the process by decomposing essential substance of mTOR activation (PIP3) that the lack of it guarantees the cell growth. (Phin, Moore, & Cotter, 2013 with copyright permission).

2.3 Liver cancer therapy

The way of HCC therapy changes according to its growth rate, phase of proliferation, tumor type and location. In early diagnoses, the first step of treatment process is surgery, to remove tumor tissue from liver or new liver transplant in long-term cirrhosis verifications. Radiofrequency ablation and Microwave ablation are other ways of destroying cancer cells by heating them up, using radio waves and microwaves mechanisms, respectively. Alcohol (ethanol) injection is another option for patients with low number of tiny tumors. Direct injection of ethanol into cancer cells leads to tumor tissues extermination, by dehydrating the component cells. Chemoembolization method with many side effects comes next as a solution in hard liver cirrhosis to ruin blood vessels and control targeted tissue. Finally, in advanced cases, biological therapy is suggested where cancer supporter pathways are disturbed to prevent tumor activity. Sorafenib is the most common and effective drug of the mentioned therapy group. It inhibits tyrosine kinase, the contributory factor of cancer, by stopping its signal conduction into the cancerous cells and preventing tumor blood feeding sources formation. In addition, as previously explained about PI3K's hyperactivity, it is a good target for these drug treatments (Ahmed & Lobo, 2009).

2.4 Mutations and Liver cancer

HCC is a process, which starts with well-differentiated stage, continued with moderately and poorly differentiated ones and ends with undifferentiated condition. Huh7 and Mahlavu are HCC's different cell line types, which are categorized in well and poorly differentiated stages, respectively. In each case many different mutations occur and consequently alter the genome sequencing order. These alterations have an important role in disturbing pathway activities and are the fundamentals of liver cancer.

Mutation term refers to genome alterations, which is created as a result of irreparable damages to DNA. These alterations are significant for being detectable in both genes and non-genic regions and participating in essential processes inside the body. The damage reason and initiation point create many different types of mutations. Spontaneous mutations occur by being affected from changes in hydrogen bonding pattern, hydrolysis, a purine base missing and denaturation of the replicated strand. Error-prone replication bypass is another reason that most probably stimulates spontaneous alterations by transmitting harmful effects to a new strand. The repairment operations of inevitable errors, like double strand breaking, may also lead to mutation in some cases. Induced mutation is another type, which arises from chemicals and radiation damage.

Another classification of mutations is defined depending on the effects of these alterations. They may influence the function, fitness, structure or the protein sequence. Impact on the function can cause fatal mutations or leads to gain or loss of the function. Likewise Fitness impressing can either have harmful, beneficial or neutral results.

Structural mutations have two types as large-scale in chromosomal structure and small-scale in a small gene, where the latter is divided into three additional parts named point mutations, insertions and deletions. Point mutation is the result of exchanging or substitution of a single nucleotide to another. Insertion and deletion, the other mutation types, are the result of adding and losing some amino acids in DNA, respectively. Affecting the protein sequence leads to alterations and these alterations are subsets of point mutations. Single nucleotide conversion (point mutation) may result in missense mutation by reproducing a new codon with a new amino acid or may create the nonsense mutation by changing into a stop codon without any efficient protein production. The third situation is coding for the same protein and being the silent mutation. Furthermore, frameshift mutations can appear as a result of insertion and deletion, since they make changes in codon reading frame and leads to interruption of their produced massages. Inheritance way of the mutations is also a base of classifying the alteration as germline and somatic subgroups. Germline mutations arise in reproductive cells and can transmit to the next generation while the somatic mutations involve non-reproductive cells and can't transfer to the children from their parents.

2.4.1 Drugs and Mutations

Drugs are the chemical compounds that are widely produced with the aim of impressing and treating the specific targets. Some of the targets are highly sensitive to drugs where the others are very resistant. Resistant mutations are a kind of point mutations that are remained constant and rescued from applied treatments by becoming resistant to the drugs therapeutic effects. (Usually drugs are single targeted and are specialized for their target disruption that any changes in the target will lead to their inefficacious.) The organisms may implement different methods to earn the mentioned ability. They can deactivate drugs, alter the target binding site or metabolic pathway, and reduce drug dosage by decreasing its penetration into the cell or increasing its pumping out rate from cell surface. After getting the resistance property to chemical treatments they become stronger by reproducing the resistant genes and transmitting their trait to subsequent descendants (Luqmani, 2005; Zahreddine & Borden, 2013)

The Huh7 cell line was originally taken from a liver tumor of a 57-year-old Japanese male in 1982 (Nakabayashi, Taketa, Miyano, Yamane, & Sato, 1982). Different cell line propagation in the medium of chemical compound has been done with the purpose of developing knowledge and facilitating treatment in the field of HCC disease. According to establishing results of mentioned cell lines, Huh7, with the capability of releasing essential cell growth factors and independency from medium serum, was distinguished as a cell line with distinct characteristics. These properties made it suitable for investigation about the regulatory mechanisms of gene expression across with, becoming the elected cell line for HCV (hepatitis c virus) (Bréchot, 1996) related cancer type studies.

Mahlavu (HHCM) cell line was originally taken from the human genome, but unexpectedly, it mostly made up of L1 repeat elements (HHCM NCBI, 2016). However, there is a research, which verifies oncogenic activity of Mahlavu after translation and transfection in to rat liver, and also describes the target cell transformation after being infected (Yang et al., 1990). Another study had been conducted for examining hepatitis B virus antigen expression level in HHCM. The medium investigation results had not shown any antigen secretion from Mahlavu cell lines, whereas the cytoplasmic region also had a very low amount of antigen (Oefinger, Bronson, & Dreesman, 1981).

Some firm work has been done by Haluk Yuzugullu and his group mates, regarding the disclosure of differences between HCC's well and poorly differentiated cell lines and definition of their specific characteristics. For this reason, they focused on β-catenin mutations that relate with chromosomal consistency, high patient persistence and less tumor attack. These mutations are detectable in a group of HCC related cell lines, and activate a set of pathways known as canonical Wnt pathways (Yuzugullu et al., 2009) with responsibility of transferring the exterior signals to the cell inside for regulating gene transcription. Wnt pathways, which are categorized into three main groups named canonical, noncanonical and noncanonical planar cell polarity, are activated by receptor binding ligands to start their specialized task. Based on considering the activation of defined pathways in well-differentiated group and deactivation of them in the other, they specified the cell lines properties across with expression levels of canonical pathway genes and found their signaling activity.

The results support the hypothesis of canonical signaling pathway activation in well-differentiated subset, as their ligands expression is high for this group, whereas the noncanonical ligands are highly expressed in poorly-differentiated subsets and Wnt5a like ligands from noncanonical set act as an antagonist, preventing the canonical pathway activation. Also, according to comprehensive implementation of various markers they observed the loss of lineage and epithelial markers and overexpression of mesenchymal in poorly differentiated groups. Furthermore, the observations showed high movement and attack for this group, while the vice versa is true for huh7 and its group members. Researchers represented the canonical and noncanonical pathways complementarity as they relate to cancer start up and development processes, respectively.

In addition, they verified that the "well differentiated cell line type" expression for Huh7 represents its relation with cancer's early stages as it can rarely be seen in advanced levels. Morphological traits of these cells are very similar to that of normal liver cells, hence their size is smaller and structural organization is totally different from normal ones. On the other hand, the poorly differentiated definition of Mahlavu declares its relation to advanced proliferation stages of cancer and implies multiform structure with low cytoplasm (Yuzugullu et al., 2009).

2.5 Next Generation Sequencing

DNA sequencing, the procedure of shaping the exact order of nucleotides, is remarkable in discovering disease treatment and developing research in the medical realm. Earlier methods of DNA sequencing were known for having low throughput and high cost that brought the idea of creating new techniques. Next generation sequencing by providing various methods like Ion Torrent, Sanger and Illumina, is used in different sequencing projects to satisfies the throughput problem and aims to reduce the costs (Buermans & den Dunnen, 2014).

Exome, the DNA protein coding part includes the whole RNAs of genomes and is sequenced in order to detect alterations involved in diseases. Although allocating a small portion of genome to itself, it consists of the highest number of variations leading to disruptions (Meynert, Ansari, FitzPatrick, & Taylor, 2014).

In contrast with Exome, Transcriptome (Frith, Pheasant, & Mattick, 2005) is a collection of RNAs in an appointed cell group. RNA sequencing, a subset of Next Generation Sequencing (NGS), is a term used for transcriptome sequencing. The higher sample throughput of NGS with augmented base coverage of DNA sequencing, is suitable for dynamic nature of transcriptome and simplifies sequencing of the transcribed RNAs in a cell, to make available the potency, to look at post transcriptional modifications, the boundaries of exon/intron and changes in gene expression.

2.6 Mutect

Somatic point mutations detection is a critical phase of cancer recognition where different methods are designed for achieving success in the case. Methods varying in specificity and sensitivity make possible the selection of most appropriate way for this purpose. High sensitive methods are able to find more mutated regions, while the high specificity leads to reduction in error rate of counting healthy regions as the mutated ones. The simultaneous ascendency of these properties provides the chance for precise and accurate detection of SNVs. Allelic fraction; another involved factor in detection method's evaluation represents the frequency of a variant in the sample population. Infer from above definition very high sensitive tools are required in order to detect the mutations with low allelic fractions. On the other hand the sequence reading depth and coverage have a significant role in discovering the mutated regions. The first expression refers to each nucleotide's reading time in a sequence. For this reason deep sequencing or high reading depths leads to read numbers larger than the sequence length. The second expression represents the average read number of a nucleotide in a given sequence. As a result, deep sequencing and high coverage can support more confident and correct determining of variations (Cibulskis et al., 2013).

2.6.1 Methods evaluation

The downsampling and virtual tumor benchmarking methods are used in order to assess the efficiency of various mutation finding tools. Downsampling is a way for measuring the method's performances, utilizing the read-sets driven from a data set, with primarily validated somatic mutations, until catching the desired coverage depth. The procedure's read-set selection is random and valuable from the aspect of saving the original data's allelic fraction, since, it eliminates the reads without attending to discover whether it carries the mutant allele or not. However, there are some deficiencies with the method like: 'limitation in the read set numbers', 'overestimating the sensitivity because of excluding the variants that are not originally distinguished' and 'disability in specificity assessment', which brings the idea of applying the virtual normal-tumor method to solve the problem.

This method's process starts with creating the virtual normal and tumor samples from two separated sequenced data, belonging to a single normal sample. Each detected mutation of comparison result is a false positive since the prepared normal samples are not different and does not carry any mutation. After that, embedding a number of known mutations in primary virtual tumor sample creates a new tumor sample. As the embedded segments of tumor are precisely known, matching the results with normal sample assists in estimating the sensitivity of each detection method (Figure 2.3).

Comparing somatic mutation detection methods, Mutect with extremely high sensitivity even in low allelic fractions and approximately high specificity, was distinguished from others and became an outstanding tool for investigations in the field. These observations represent Strelka as a second reliable method for variant detection from the perspective of having high sensitivity in both high and low allelic fractions. JointSNVMix comes next by showing more sensitivity than Strelka in high fractions, but placing in the third place as a result of sensitivity descending in samples with low allelic fraction. SomaticSniper allocates the last rank to itself by showing the lowest sensitivity in both samples. Mutect is developed at Broad Institute using GenomeAnalysisToolkit, the software package for high throughput data analysis, to improve the sSNVs identification in next generation sequencing data of cancer genomes (Cibulskis et al., 2013).

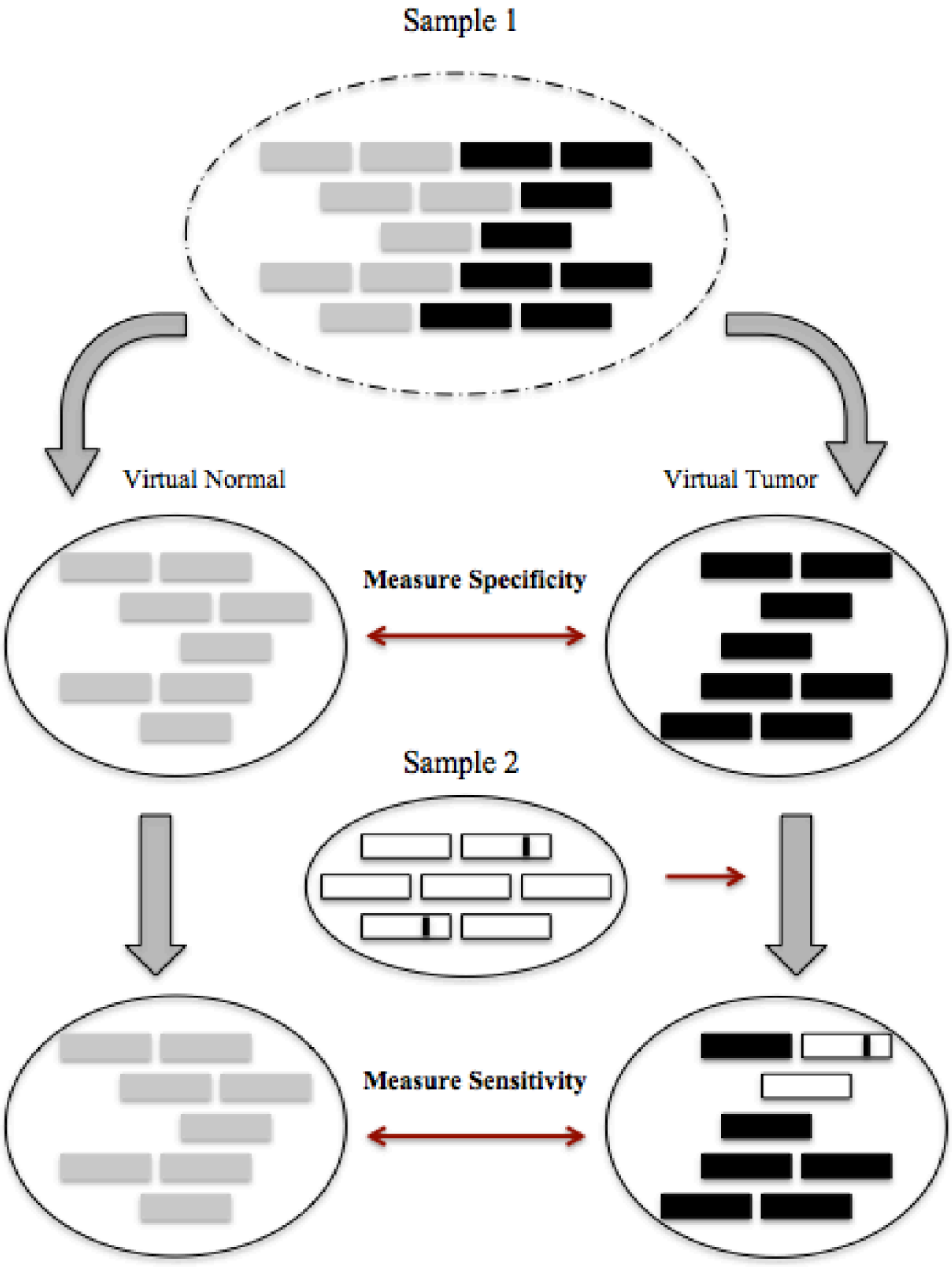


Figure 2-3: Overview of virtual tumor approach.

2.6.2 Mutect processing steps and components

There are some essential elements need to be supplied as inputs of the system, to profit from Mutect analysis. The inputs consist of the reference genome, COSMIC (Forbes et al., 2009) dbSNP (Sherry et al., 2001) data, and two samples as normal and tumor. The reference is an organism genome's fasta file, which is selected according to the origin of prepared normal and tumor samples with the aim of comparison. Catalog Of Somatic Mutations In Cancer (COSMIC) is a database that contains all previously discovered variations relating with cancer genome processes. Either, the other input named dbSNP, provides the database for single nucleotide polymorphism to enable the separation of SNVs from the single nucleotide variation of a population with a high frequency. Finally the tumor and its matched normal data are used to find differences between altered and normal genome and recognize the somatic mutations from that of germline.

Mutect starts with preprocessing stage, continued with statistical analysis and ends with post processing function. The preprocessing involves the elimination of reads with low quality score representing the high error probability or the highly mismatched sites to prevent from noise in the sample. Statistical analysis begins with the purpose of determining sSNVs after primary operations. In this regard, first the tumor data is compared in mutated and reference site to find the altered regions and variation ratio of the sample. Then it checks the results by comparing the normal data with reference to verify that the predicted mutations are not in normal. To do so it uses Bayesian classifiers across with defining a threshold of each case. The cutoffs guarantee that the false positive rate is less than half of the somatic mutations rate.

$$LOD_T = log_{10}\left(\frac{P(observed\ data\ in\ tumur|site\ is\ mutated)}{P(observed\ data\ in\ tumor|site\ is\ reference)}\right)$$

$$LOD_T > log_{10}(0.5\times10^{-6}) \approx 6.3$$

$$LOD_N = log_{10}\left(\frac{P(observed\ data\ in\ normal|site\ is\ reference)}{P(observed\ data\ in\ normal|site\ is\ mutated)}\right)$$

$$LOD_N > log_{10}(0.5\times10^{-2}) \approx 2.3$$

In order to remove artifact mutations, post-processing is applied and wiped out suspected candidates using a number of filters. There are some common filters in all processes. LOD score (log odds) is the one that is previously illustrated. Next is patient contamination possibility that leads to elimination. Presence of candidates in dbSNP and not in COSMIC, causes the LOD_N cutoff rise to that of dbSNP for prohibiting the germline variation's count-in as somatic mutations by mistake. Further, passing more than one predicted mutation through the whole filters stimulates a filter named "Triallelic Site" as the mentioned case is almost impossible. In post-processing, additional filters are applied into detected mutations to get more confidant results. Proximal Gap, Poor Mapping and Strand Bias are some of the involved filters in the mechanism which provide the confident outcomes (Cibulskis et al., 2013).

2.6.3 Mutect output

Mutect produces three files as Call-stats, containing the candidates, the wiggle, which shows the read coverage of the experimental data and a vcf in demanded cases. Coverage should be at least fourteen in tumor and eight in normal, to be in the acceptable range of sensitivity for variant calling. Call-stats is the main file with the list of mutations that gives very detailed information for each participant and can be selected partially according to the specific purpose of the users. If required, the prepared vcf file results show predicted mutations, which are placed in keep or rejected group.

CHAPTER 3

MATERIALS AND METHODS

3.1 Materials

This section includes the processing steps of Mutect and the required inputs for mutation detection analysis. The inputs' preparation and related softwares are illustrated in detail.

3.1.1 Data Discription

RNA-Seq data was obtained from treatment of liver cancer cells with Sorafenib or PI3K inhibitors (Figure 3.1).

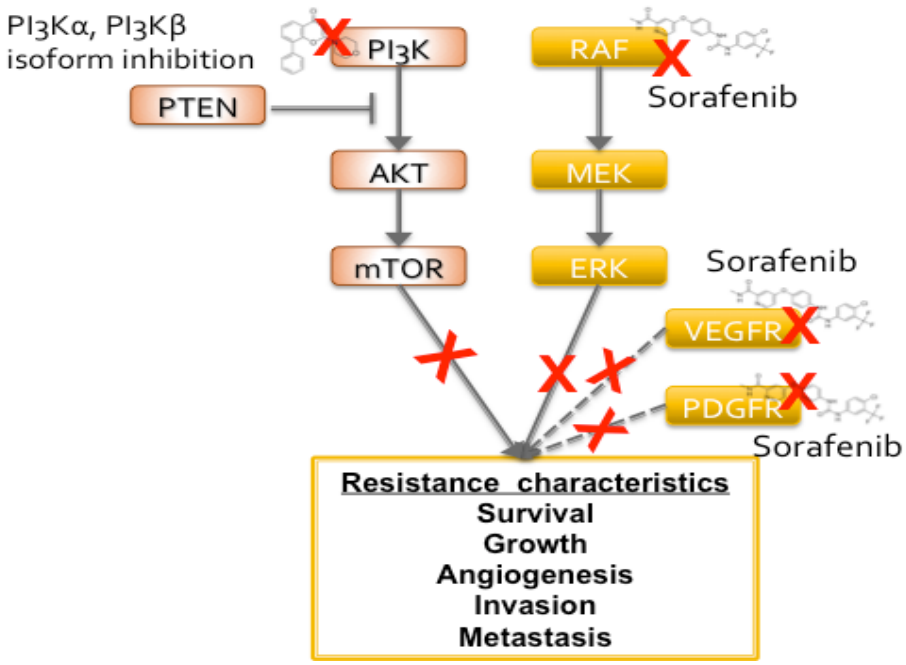


Figure 3-1: Liver cancer treatment with Sorafenib or PI3K inhibitors.

In this study PI3Kα and β isoform inhibitors were used. The RNA-Seq data was provided by CANSYL laboratory ODTU. Huh7 and Mahlavu cell lines were treated in IC_{50} concentration for 48 hours with the following condition.

- Control (DMSO)
- Sorafenib (SOR)
- PI3Kα inhibitor (ALPHA)
- PI3Kβ inhibitor (BETA)
- Sorafenib with PI3Kα inhibitor (S_ALPHA)
- Sorafenib with PI3Kβ inhibitor (S_BETA)

3.1.2 Mutect

Mutect, a reliable method in detecting somatic mutations, is known to be an outstanding technique in the field because of special properties (such as high sensitivity even in low allelic fractions). Based on Bayesian classifier, Mutect acts very well in SNVs detection where, it requires a low number of supporting reads. Additionally, the tool's applied filters guaranty the high specificity like other methods along with a remarkable sensitivity. To start with the process some prior procedures need to be done and downloading the .jar file from broad institute is what we started with. We used the MuTect-1.1.7 jar file for our work, which was accessible from GATK download page at **https://www.broadinstitute.org/gatk/download/auth?package=MuTect**.
The required inputs were prepared continually after downloading the .jar file.

3.1.2.1 Normal vs. Tumor Data

One of the most important inputs of Mutect is tumor/normal data, which is supplied for comparison. The normal and tumor samples comparison is a considerable step in cancer studies that reveals the significant alterations and resistant mutations in disease cases. However, they are generally from the same individual to allow the separation of germline mutations and prevent error in somatic mutation detection. We have two Huh7 (HDMSO) and Mahlavu (MDMSO) cancer cell lines' data in normal and their various drug treated forms in tumor group to detect the stable mutations and resistant genes after comparison (Table 3.1).

Table 3-1: Normal vs. Tumor in mutation detection. The normal group has HDMSO (Huh7) and MDMSO (Mahlavu) as untreated control samples. The tumor group includes the two samples' treatment with different drug combinations where "S" and "SOR" stands for Sorafenib, ALPHA stands for PI3Kα inhibition and BETA stands for PI3Kβ inhibition.

Normal	Tumor
HDMSO	H_ALPHA
HDMSO	H_BETA
HDMSO	H_S_ALPHA
HDMSO	H_S_BETA
HDMSO	HSOR
MDMSO	M_ALPHA
MDMSO	M_BETA
MDMSO	M_S_ALPHA
MDMSO	M_S_BETA
MDMSO	MSOR

3.1.2.2 HG19/GRCH37

Human genome is a complete set of nucleic acid sequence in Homo sapiens, which is made of two twisting, paired strands inhabit in 23 pairs of chromosomes within the nucleus. The complete sequencing of human genome is available in different releases and out of those we have chosen hg19, which is equal to grch37 from genome reference consortium (Church et al., 2011). Mutect requires reference (hg19) in fasta format, which displays the data in single-letter form with a brief description prior to the sequences. Hg19 fasta file is accessible online at different browsers such as **http://genome.ucsc.edu/** and **http://www.1000genomes.org/**.

3.1.2.3 dbSNP

dbSNP represents the database of single nucleotide polymorphism and differs from SNV in frequency of occurrence. Although, there are some somatically mutated sites locating in this group. The database submission provides the easy access of Mutect to this collection and assists in correct detection of SNVs. dbsnp_132_b37.leftAligned.vcf.gz for hg19 is accessible online at **http://gatkforums.broadinstitute.org/**.

3.1.2.4 COSMIC

COSMIC is the collection of somatic mutations, which are found previously. Investigating the database content during Mutect analysis prevents from wrong elimination of variants, existing in both dbSNP and COSMIC databases. The b37_cosmic_v54_120711.vcf, compatible with hg19, is accessible online at **http://gatkforums.broadinstitute.org/**.

3.1.3 RefSeq

The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation of medical, functional, and diversity studies (Pruitt, Brown, Tatusova, & Maglott, 2002). The data are available for many organisms of different clades. Various assemblies of organized data provide the facility to use the genome sequence in required version and serve a standard reference for intended analysis. In this survey, the assembly was settled on hg19 and the RefSeq data was obtained from "RefGene" table through the UCSC table browser. The output gives a detailed description of genes, including the chromosomal location, start-codon, range, strand and exon-frame. We implemented the produced information for classifying the candidate mutations as missense, nonsense or silent ones, writing a python code.

3.1.4 TopHat

Mutect requires the normal/tumor inputs in .bam file format. However, the raw data from laboratory was in fastq form, that we used a consistent method to perform the required conversion. The implemented tool should map the data to reference (hg19) to produce an acceptable result. Tophat (Trapnell, Pachter, & Salzberg, 2009), satisfying the needs, is an eligible tool for the work, which creates the bam files in an accurate way to feed into the Mutect. The implemented tool and its prerequisite softwares (boost, bowtie) were downloaded and installed from **https://ccb.jhu.edu/software/tophat/index.shtml**.

3.1.5 FastQC

FastQC (Andrews S., 2010) is a tool to check the quality of reads from high throughput data, acquired in fastq format. The analysis results-in various statistical plots to expose low qualities and deficiencies by the read sets before starting the mutation detection analysis. From different created plots, the “per base sequence quality” informs us of the need for trimming or masking the reads, specifying the number and position of the bad reads. This software can be installed on the machine and run using simple codes. In addition, it is accessible online at galaxy (https://usegalaxy.org/), which provides the easy access for many analysis tools.

3.2 Method

Out of different techniques in the field, Mutect is known to be precise in low allelic fractions and has high sensitivity. Thus we decided to use Mutect as our leading program to determine somatic mutations. To use Mutect effectively, we need to prepare the default inputs for the program. Our data consist of two cell line types of liver cancer; one, HDMSO (Huh7) as well differentiated and second, MDMSO (Mahlavu) as poorly differentiated and their drug treated forms, including the treatment by Sorafenib, inhibiting α isoform, inhibiting β isoform and their mixture as sorafenib+PI3Kα inhibition, sorafenib+PI3Kβ inhibition which we got in .bam file format. The original data was in fastq format, which converted to bam, the only acceptable format of Mutect, using Tophat tool. Tophat creates bam files by mapping the fastq data to reference genome of the related organism - hg19 in our experiment - and requires some subsidiary softwares like Boost (Duffy & Bv, n.d.), Bowtie (Langmead & Salzberg, 2012) and SAMtools (Li et al., 2009) to be able to process the files. We started with Tophat by downloading Boost, which was required for the first compiling. Subsequently bowtie was installed and indexed for the reference genome. SAMtools, the last complementary software with Tophat, should also be installed on the system in appropriate version suited for our purpose. The output of Tophat program is accepted_hits.bam from which we had 10, each for one of the Huh7, Mahlavu or their drug treated forms. After creating bam files we need to prepare the other input files; Mutect jar, hg19, COSMIC, dbSNP, which are obtainable from gatk forum. Hg19, our reference genome, should be in fasta format as required in Mutect. The related COSMIC and dbSNP files named b37_cosmic_v54_120711.vcf and dbsnp_132_b37.leftAligned.vcf.gz, respectively, are also downloaded from the mentioned forum. After preparing all inputs we started the program successfully and got the somatic mutations (Figure 3.2) (Figure 3.3).

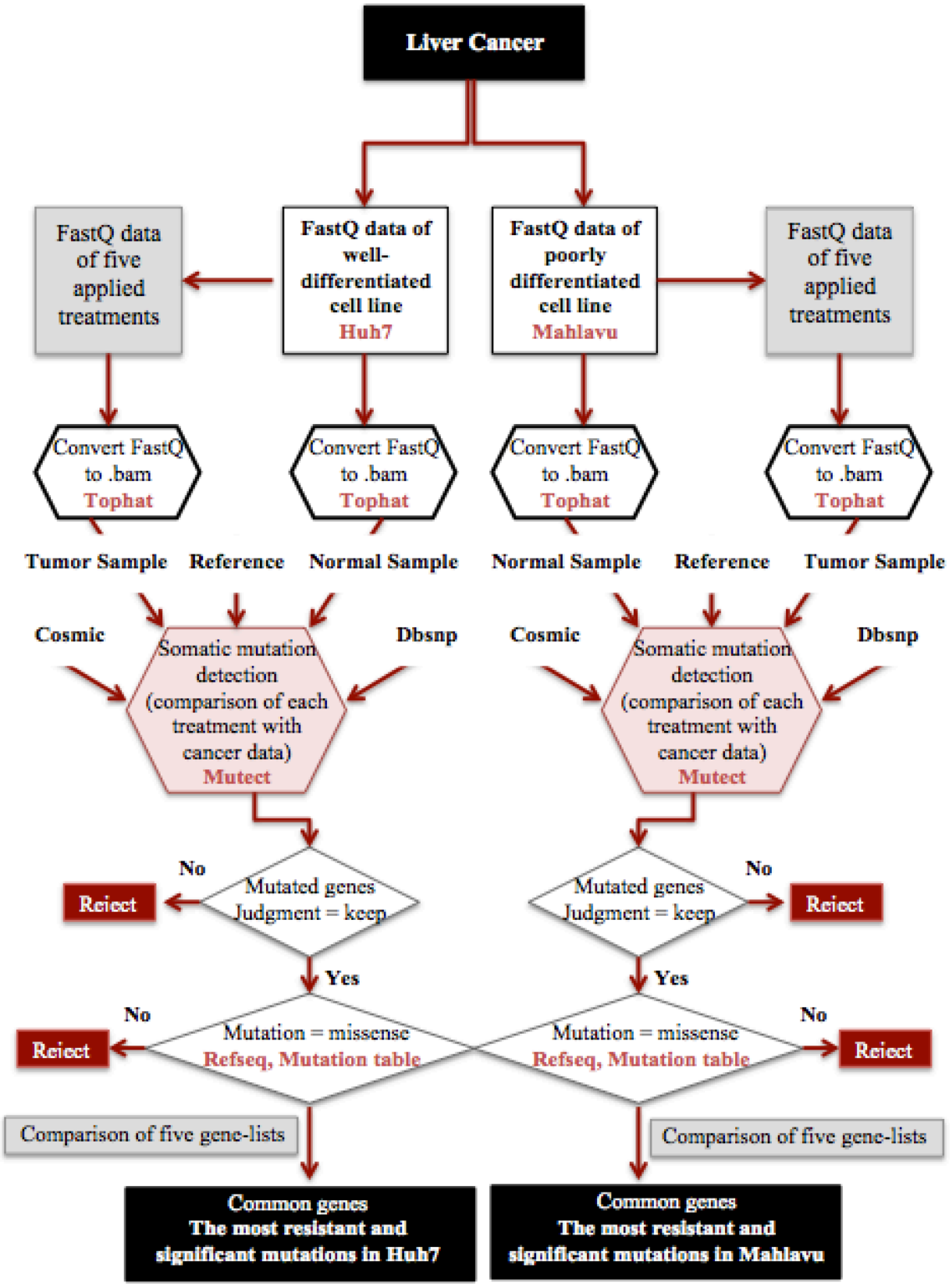


Figure 3-2: Workflow of mutation detection.

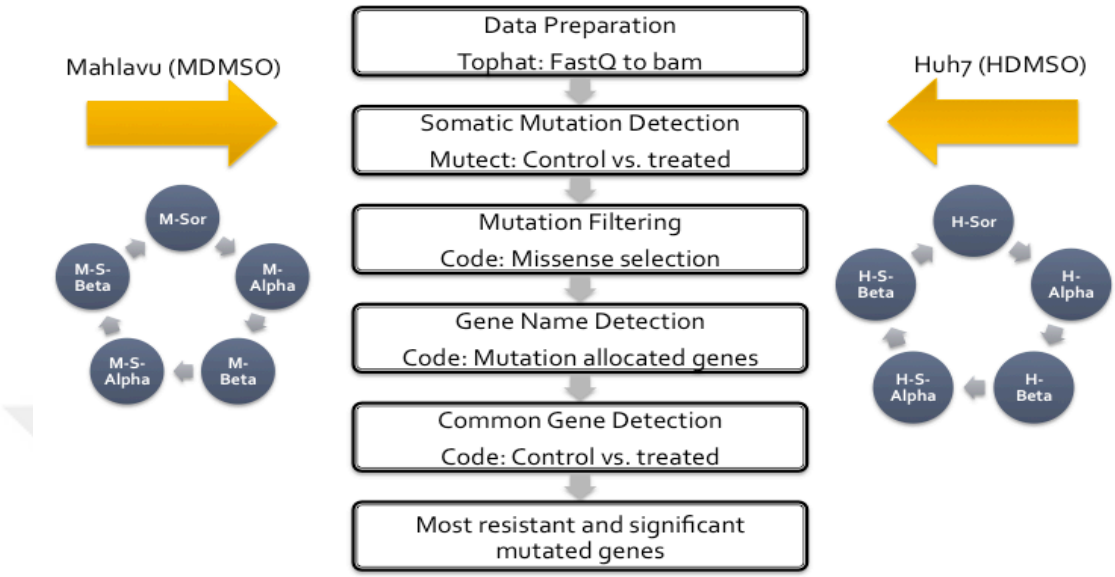


Figure 3-3: Data and processing steps

3.2.1 Quality Control

The quality control analysis of Huh7 and Mahlavu data sets are primarily done to evaluate the read quality, GC content and adapter contamination. The analysis results were acceptable and the sequences were highly scored (from 28-40) in the green area (Figure 3.4) (for more details see Appendix A). Furthermore, the GC content comparison to a normal distribution plot passed the quality thresholds by revealing a smooth distribution over the length of all sequences. Getting check marks for all of the desired analysis, we crossed to mutation detection steps without any filtering changes.

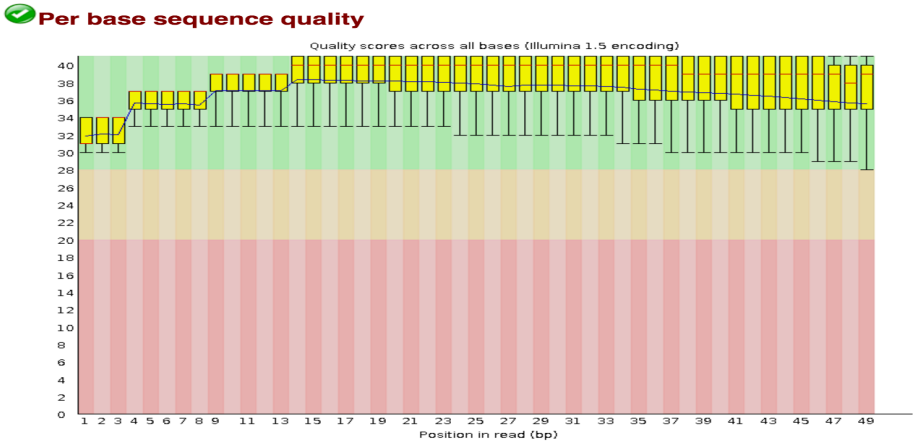


Figure 3-4: Quality control analysis of H_ALPHA

3.2.2 Data Mapping using Tophat

TopHat is a spliced read mapper for RNA-Seq that uses Bowtie as a super speed short read aligner (Trapnell et al., 2009). TopHat identifies splice junctions without the demand for annotation from reference. TopHat forms a database of plausible splice junctions by applying primary mapping data to plot reads versus mentioned junctions to validate them.

Boost and Bowtie are the prerequisite programs to compile TopHat (for more details see Appendix B). It is also possible to use the compiled version on Linux. After finishing the compilation steps we can start the analysis with Tophat using the command below.

```
$tophat –G .gtf_file $bowtieindex/hg19 reads1.fq,reads2.fq
```

The gtf file used in the command line is the reference's release data that we got from Gencode (Harrow et al., 2012). Gencode, the encyclopedia of genes and gene variants provides the facility of obtaining required information from accumulated reference gene-sets for human or mouse genomes. We used the current released data collection to get the related data for hg19 (GRCh37). Fastq1 and Fastq2, the paired_end reads of Illumina NGS, were acquired as primary data in laboratory. This data is used to get the precise results by sequencing from both left and right sides and aligning to the reference genome. Finally, Tophat output named "accepted_hits.bam" was saved in Tophat_out by default.

3.2.3 Data Preparation

Output bam files require further preparation to be acceptable by Mutect. One of the essential terms of input samples is denoting a read group for each read and specifying the read groups in the header. Read groups are defined with @RG sign, including critical information like applied platform, library type and sample name for processing. The read groups are assigned to specify the samples of related data and compensate variability from one sequencing run to the next. The read groups' absence causes an error, which should be solved using an appropriate tool. We utilized Picard, the java-based command line tools' pack, to deal with AddReadGroup error and correct deficiencies. In this work, Illumina was the applied platform, where the library in use was bar, the platform unit was 1 and the sample name was unique for each, appointed by considering the analyzed data. It is notable that we created new files for the outputs instead of overwriting the same bam files.

```
java –jar $picard.jar AddOrReplaceReadGroups I=$accepted_hits.bam
O=$accepted_hits01.bam RGID=1 RGPL=illumina RGPU=1 RGSM=[“sample-
name”] CREATE_INDEX=True
```

Another prerequisite of the system is to satisfy mapping conditions for karyotypically ordered reference. The lexicographical sorted inputs, led to inconveniences by submitting chromosomes in a contrary order with hg19. They require chromosome rearrangement from 1 to 22, followed by X, Y and M, respectively (though M can be the leading of the list instead). Because of getting an error with implication of “unmatched input with reference”, we implemented the ReorderSam tool from Picard to resort the bam files.

```
java –jar $ReorderSam.jar I=$accepted_hits01.bam
O=$accepted_hits02.bam REFERENCE=$hg19.fasta
```

Settling down the bam file’s ordering problem, we got the incompatible contig error, relating to dbSNP and COSMIC, in the next step. In order to sort the vcf files according to the reference, the reordering process was done using vcfsorter.pl, which was downloaded and executed using hg19 dict file for mapping. Because the index of newly reordered vcfs files are needed for Mutect analysis, their indices had been made using *tabix* command, continually.

```
Perl $vcfsorter.pl $hg19.dict $dbsnp.vcf  > dbsnp.reordered.vcf

Perl $vcfsorter.pl $hg19.dict $cosmic.vcf  > cosmic.reordered.vcf

tabix –p vcf dbsnp.reordered.vcf

tabix –p vcf cosmic.reordered.vcf
```

N_CIGAR_READS, another error of RNA-Seq data processing, can handle by allowing or filtering the reads. We allowed N cigars, which implies introns to analyze full RNA data. The -U phrase at the beginning of the command shows that it is an unsafe option. Although the existence of this sign is currently a supported option for RNA-Seq processing, the matter is still under development.

```
-U ALLOW_N_CIGAR_READS
```

Mutect requires a standard quality score (ASCII 33), which is defined considering the Sam/Bam specification. However, some methods result-in datasets with higher scores by encoding from ASCII 64. The argument for bug fix, exchanges the read scores with acceptable ones by subtracting 31 from each. Getting the related error, we had to use the command below as the program refuses to process without receiving the assumed quality scores.

```
-fixMisencodedQuals
```

3.2.4 Mutect

We run Mutect code after solving all problems and preparing input files correspond to the instructions. Including valuable information, "call-stats" and "wiggle" are the most significant output files among several results produced by Mutect. Wiggle contents represent the coverage depth of each read to verify their sufficient coverage in the samples. Call-stats (out.txt), the main output of the Mutect, transfers a vast amount of information by categorizing them under different fields. However, the results are very comprehensive and confusing that make the users to select the specific subsets, essential for their work.

```
java -Xmx2g -jar $mutect-1.1.7.jar --analysis_type Mutect --
reference_sequence $hg19.fa --cosmic
$b37_cosmic_v54_120711.reordered.vcf --dbsnp
$dbsnp_138.hg19.reordered.vcf --input_file:normal <DMSO_BAMFILE> --
input_file:tumor <TREATED_BAMFILE> --out out.txt --coverage_file
out.wig.txt -U ALLOW_N_CIGAR_READS -fixMisencodedQuals
```

Judgment is one of the fields in the result file, which informs us of keeping or ignoring the calls by specifying them with reject or keep labels. It was the first filter that we applied to remove the artifact mutations. After rejection process of nonessential calls, we kept those with "keep" label to find somatic mutation host genes and types. To do this, columns including the contig, position, context, ref_allele, alt_allele and dbSNP_site information were imported to a new excel file and utilized in codes named mutation type and gene name detection. “Contig” results refer to chromosomes, carrying mutations and “position” shows the exact location of candidates on the related chromosomes. Furthermore, “context” reveals the prior and following bases to the mutated regions from reference. Continually, the columns labeled with “ref_allele” and “alt_allele” represent the candidates before (reference) and after (mutant) alteration. Finally,

"dbSNP_site", placed in the last column of our new file, classifies the candidates as dbSNP or novel mutations.

3.2.5 Gene Name Detection

In order to detect the origin of mutations, we wrote a python code and linked the candidates to their belonging genes. The Package named "PyEnsembl" was used in python to get the Ensembl reference genome information. PyEnsembl installation was done using "pip", and after that the reference data were downloaded and installed in proper release (75) corresponding to hg19.

```
pip install pyensembl

pyensembl install --release=["release 75"] --species=["human"]
```

PyEnsembl acts based on mapping the locus of candidates to the reference. After installation of required packages, the excel files were imported into python and essential columns for gene symbol detection, contig and position, were appended into a list. In the next step we created *for* loops to define separate lists for contigs and positions without labels. In order to prepare the data for gene analysis, position contents were defined as integer and "chr" signs were removed from contig lists. Finally, importing the PyEnsembl and using related reference data, we detected the genes responsible for mutations.

```
import pyensembl

ensembl = pyensembl.EnsemblRelease(release=75)

genes = ensembl.genes_at_locus(contig= ["contig-number"], position=
["position-number"])
```

3.2.6 Mutations' Type Detection

There are some methods like "SNPef" (http://snpeff.sourceforge.net/), "Gemini" (http://gemini.readthedocs.io/en/latest/content/installation.html#automated-installation) and "Varapp" (https://pypi.python.org/pypi/varapp-backend-py/1.0.1.2), which are available online for variant filtering. However, we used our own methodology.

Specifying the types of variants, resulted from Mutect analysis, is a major step toward detecting the significant mutations. For this purpose, primarily, the two different

combinations of codons, each for one of the reference and altered states, are gained by substitution of "x" character of "context" column with "ref_allele" and "alt_allele" columns' content one by one. Since, "x" represents affected nucleotide by alterations, the two distinct sequence order belonging to mutated and reference sites, were formed after the replacement. However, it was essential to consider the reading frame of sequences to extract the right codons for comparison. The presence of three reading frames (0, 1, 2) for each of the positive and negative DNA strands, influences the nucleotides' order and position in codons creation. Along with exchanging the reading direction in opposite strands, we appointed the first, second and third positions of codons for replaced characters in x by reading frames of 0, 1 and 2, respectively. RefSeq database, by containing the strand and reading frame data of sequences, specifies the reading direction and assists with defining the nucleotides position in codons (Figure 3.5). Additionally, we switched all of the A characters with 'U', 'T' with 'A', 'C' with 'G' and 'G' with 'C' to gain the RNA form of sequences (Figure 3.4).

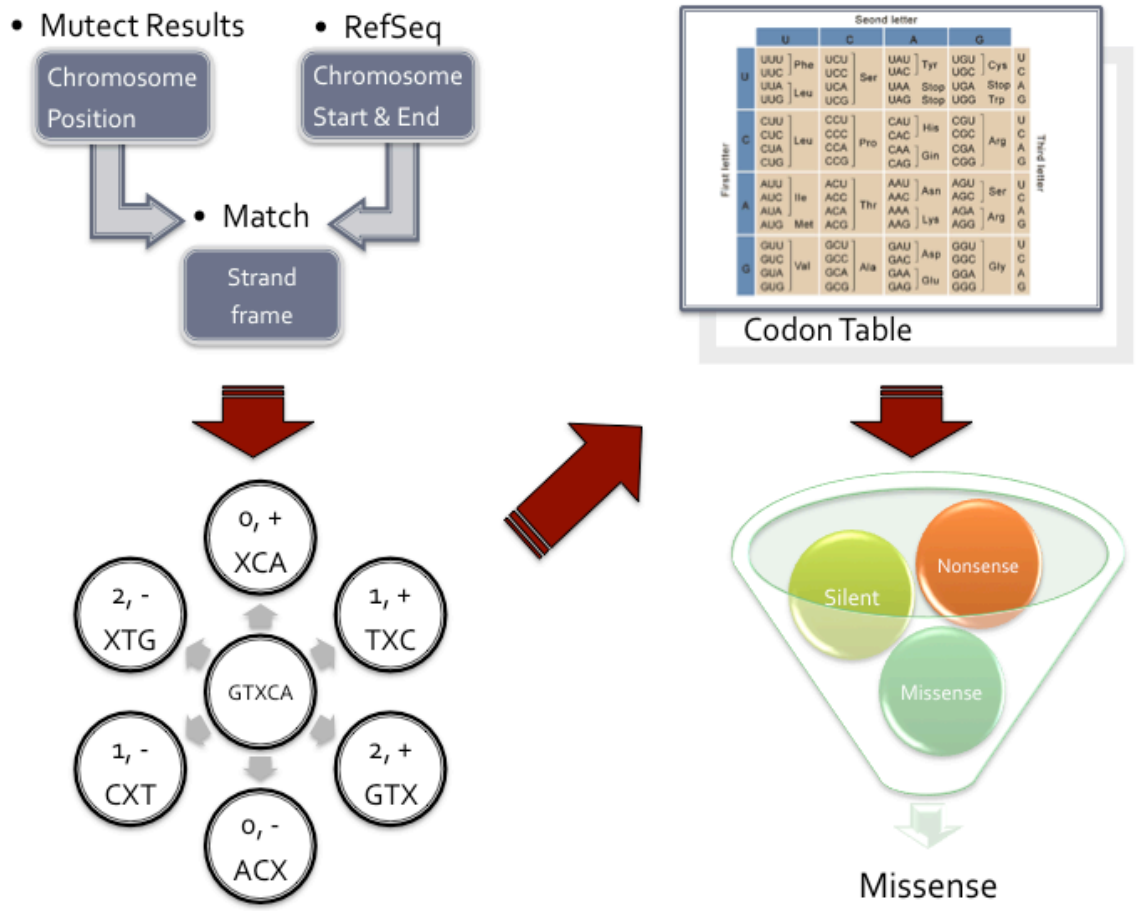


Figure 3-5: Workflow of mutation type detection

```
For each mutations in each dataset

    Take chromosome names and position columns and check the reference
for match

    If it matches then check the matches from reference for any
possible pair sets of related strand {+,-} and reading frame {0, 1, 2}

    Print out a list for each pair{{+0},{+1},{+2},{-0},{-1},{-2}}

Go back to one

If all of the mutations are checked end the program
```

The produced codons comparison was done to distinguish mutations, which result-in different amino acids production. We categorized the codons' comparison results to nonsense, missense and silent groups by referring to DNA nucleotides table, which shows the coded amino acids.

```
if amino_ref == 'Stop' or amino_alt == 'Stop':

    print ('Nonsense')

elif amino_ref == '_' and amino_alt == '_':

    print ('_')

elif amino_ref == amino_alt and amino_ref != '_' and amino_alt != '_':

    print('Silent')

else :

    print ('Missense')
```

3.2.7 Common Genes Detection

In order to detect the resistant mutations for applied drugs we compared missense subsets and found the common mutated genes in different treatments of Huh7 and Mahlavu cancer cell lines. The comparison of two cell line treatments was distinctly done, among five members of treatment groups. The genes high mutuality in the treatment lists refers to their high resistance. We investigated the existence of each gene and counted the exact number of their repetition in all treatments to find the high resistant genes in each group.

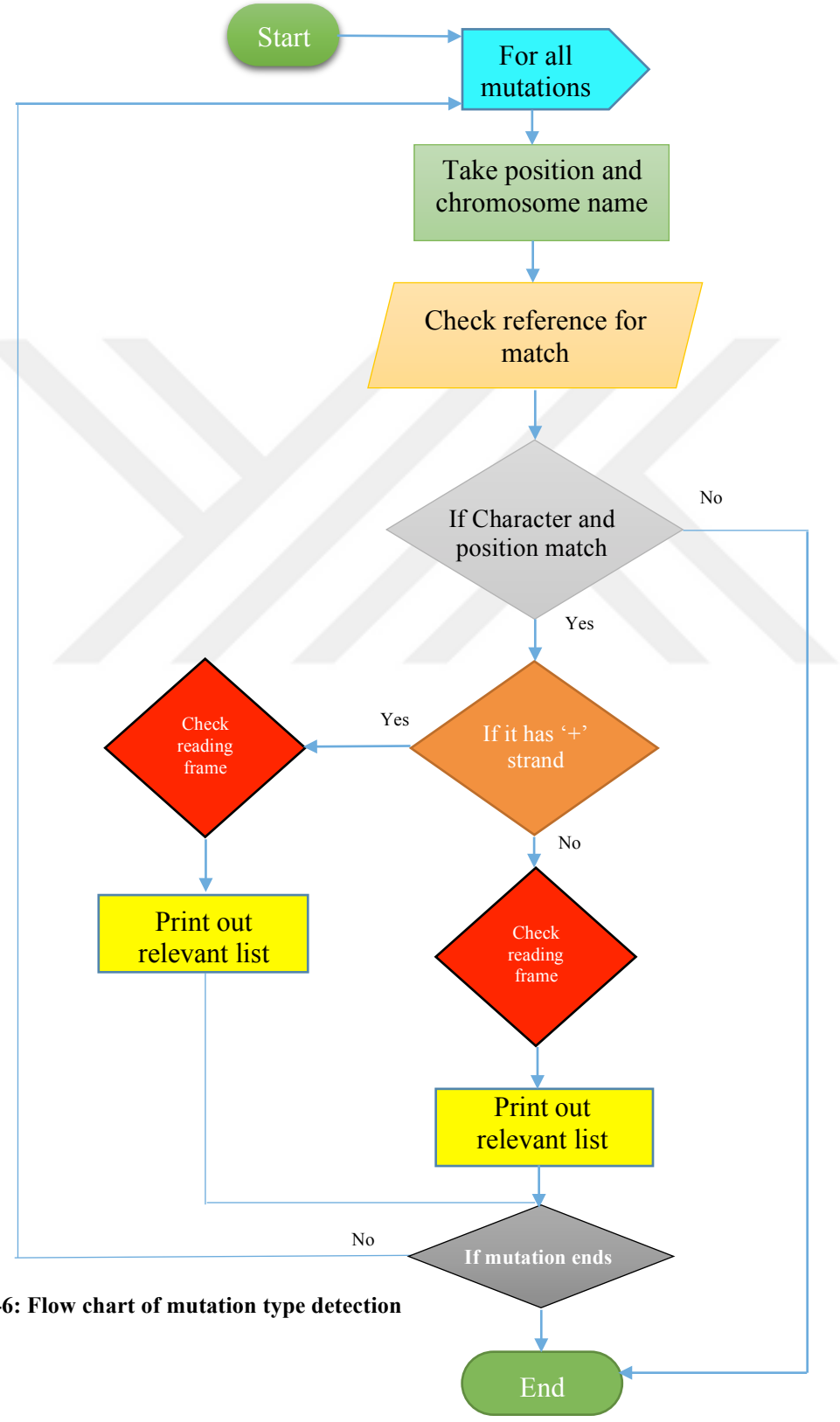


Figure 3-6: Flow chart of mutation type detection

CHAPTER 4

RESULTS

Each cancerous cell line's comparison with its drug treated forms, using Mutect, revealed various mutations. Every compared set has a series of unique mutations, hence there are a number of shared ones between certain groups. Variations arise from treatment of Huh7 and Mahlavu cell lines with different drugs and refer to resistant genes, which alter to fight against the drugs' fatal impacts. The genes that are common in most of the sets are very important, as they are the probable reasons of liver tumors' perseverance and distribution. We tested different combinations of drug treatments to analyze their interactions and subsequently find the mutated and resistant genes in each treatment.

Table 4-1: Mutect results of Huh7 and Mahlavu compared to drug treated forms

Normal & Tumor	# Mutation	# Missense
HDMSO vs. H_ALPHA	610	79
HDMSO vs. H_BETA	508	90
HDMSO vs. H_S_ALPHA	930	77
HDMSO vs. H_S_BETA	594	58
HDMSO vs. HSOR	546	54
MDMSO vs. M_ALPHA	488	92
MDMSO vs. M_BETA	377	78
MDMSO vs. M_S_ALPHA	882	116
MDMSO vs. M_S_BETA	379	72
MDMSO vs. MSOR	633	65

DMSO (control)

SOR: Sorafenib ALPHA: PI3Kα inhibitor BETA: PI3Kβ inhibitor

S_ALPHA: Sorafenib with PI3Kα inhibitor S_BETA: Sorafenib with PI3Kβ inhibitor

4.1 Huh7 treatment

Mutations related to applied drugs on Huh7 cell line and their interactions are illustrated below.

4.1.1 PI3K-α and PI3K-β inhibition

Huh7 cancer cell line was treated by PI3K-α and PI3K-β's activities inhibition, where α isoform's inhibition led to 79 mutations and β isoform's inhibition caused 90 gene alterations. The two sets had 3 genes in common which was *PLEKHO1*, *LIMK2*, *RBM14*, resisting to α and β isoforms disruption, and mutate whether with PI3K-α or PI3K-β inhibition. *PLEKHO1* may involve in actin capping protein (CP) phosphorylation and thereby regulate the actin cytoskeleton. Moreover, it behaves as a tumor suppressor to inhibit tumor growth by representing negative regulation effect on cell proliferation. The other common genes, *LIMK2* participates in phosphorylation process and *RBM14*, depending on isoform type, acts as transcriptional repressor or nuclear receptor co-activator (Figure 4.1) (Table 4.2).

4.1.2 Sorafenib, PI3K-α inhibition and Sorafenib+PI3K-α inhibition

We treated Huh7 with Sorafenib, PI3K-α inhibition and their mixture in order to compare the targeted drugs' effect on this cancer cell line and find out how they interact with each other in combined treatments. Where Sorafenib usage by itself leads to 54 and PI3K-α inhibition results in 79 mutations, their combination sorafenib+PI3K-α inhibition exhibits 77 mutations in Huh7 cell line treatment. PI3K-α inhibition and Sorafenib has 5 and 6 genes in common with their combined form, respectively. However, there is not any mutual mutated gene in therapies with pure drugs, and subsequently there is no shared mutation among these three therapeutic methods. Presence of common genes between drugs and their combined forms (but not single treatments) represents the creation of a totally different substance, having mutual genes with its constitutive parts. However, the interaction does not provide an improved treatment, that even we could not get the Sorafenib treatment results, though the outcome is a bit better than treatment with PI3K-α inhibition, exclusively (Figure 4.2) (Table 4.3).

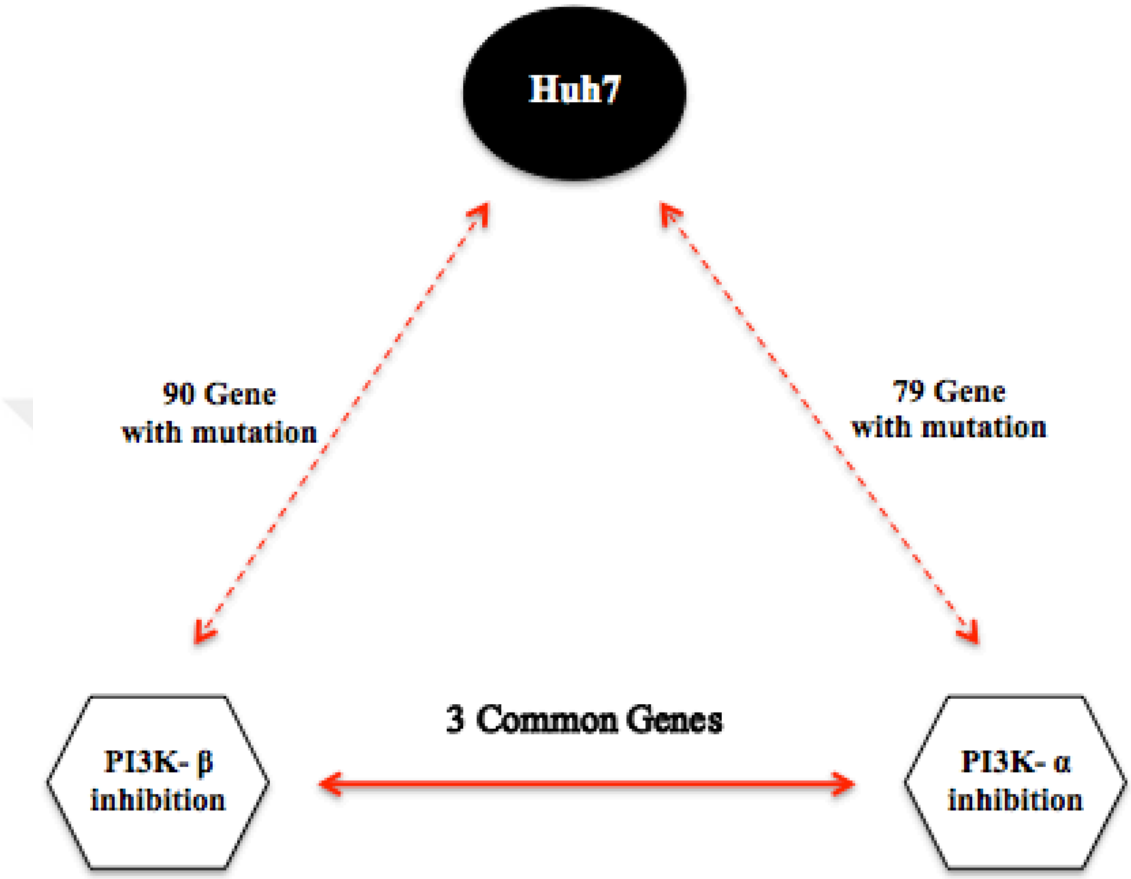


Figure 4-1: Interaction related to Huh7; PI3K-α inhibition and PI3K-β inhibition

Table 4-2: Common survival mutations in Huh7 treatment with PI3K-α and PI3K-β inhibition

Common Genes	UniProtKB	Protein	Organism
PLEKHO1	Q53GL0	Pleckstrin homology domain-containing family O member1	Homo sapiens
LIMK2	P53671	LIM domain kinase 2	Homo sapiens
RBM14	Q96PK6	RNA-binding protein 14	Homo sapiens

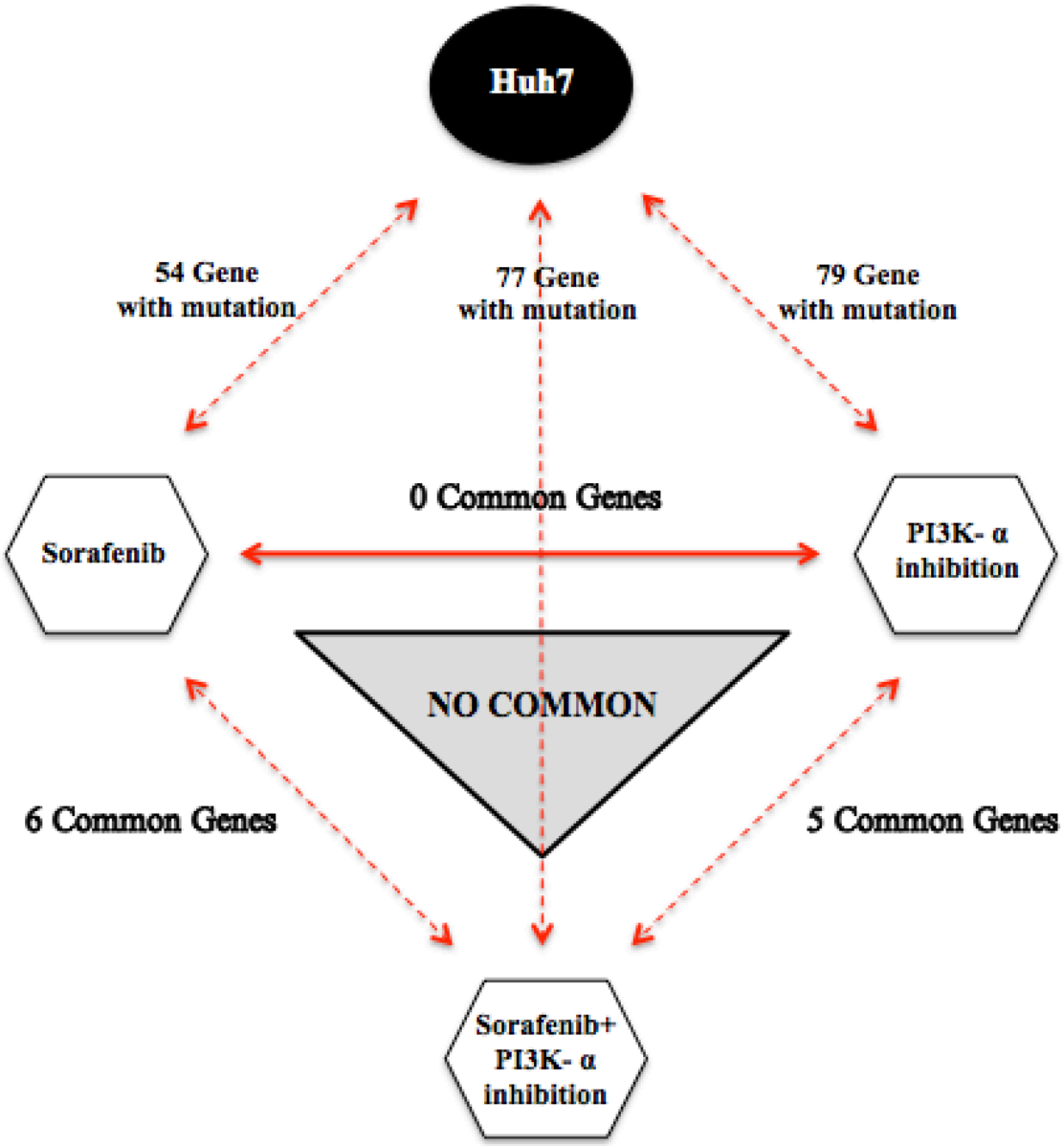


Figure 4-2: Interaction related to Huh7; Sorafenib, PI3K-α inhibition and Sorafenib+PI3K-α inhibition.

4.1.3 Sorafenib, PI3K-β inhibition and Sorafenib+PI3K-β inhibition

The other drug combination applied for Huh7 was sorafenib+PI3K-β inhibition, which encountered with 58 resistant genes, where the individual treatment with Sorafenib had 54 and PI3K-β inhibition had 90 survived genes, unaffected by that specific drug. Like previously illustrated Sorafenib combination with α isoform disruption, this new combination of Sorafenib with PI3K-β inhibition did not lead to descent in survived genes in comparison with solely treatment with Sorafenib. And hence, it was not a very beneficial method as one PI3K isoform's inhibition can activate the other isoform for signal conduction. PI3K-β inhibition, Sorafenib and their mixture have one survived gene in common named *SLC39A5*, which is resistant to all of the three mentioned therapies. *SLC39A5* is a member of the ZIP family (zinc transporters) that conducts zinc (the vital cofactor of transcription enzymes and other functions) from outside into the cell. Variations in this gene lead to critical defects. Additionally, it participates in eye development and plays a role in TGF signaling pathway regulation (Figure 4.3) (Table 4.3).

4.2 Mahlavu Treatment

Mutations related to applied drugs on Mahlavu cell line and their interactions are illustrated below.

4.2.1 PI3K-α and PI3K-β inhibition

We treated Mahlavu, the poorly differentiated liver cancer cell line by inhibiting PI3K-α and PI3K-β isoforms one by one and obtained 92 resistant genes for the first and 78 resistant genes for the latter treatment. Out of these *CTC-512J12.4* and *FRG1* are mutual in survived genes' list of both methods. CTC-512J12.4 is the zinc finger protein *ZNF229* (*ZNF229*) mRNA that originally found in human. *ZNF229* is probably a participant of transcriptional regulators and functions in DNA binding. *FRG1* gene is coding for a protein that regulates exons splicing in pre-mRNA, transports mRNAs and acts in rRNA processing. Furthermore, it attends as a regulator in muscle development and related alterations in severe muscle diseases (Figure 4.4) (Table 4.4).

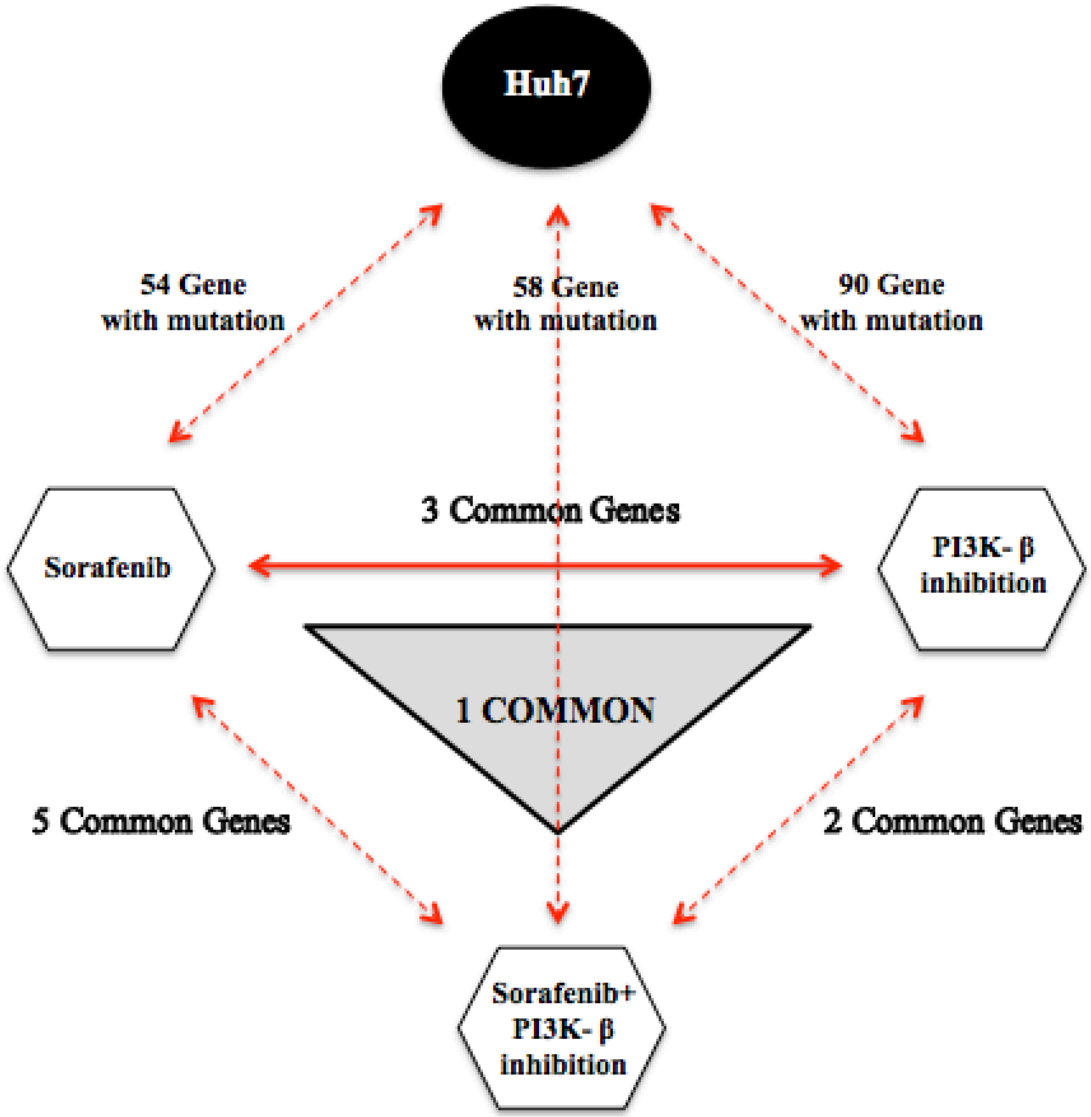


Figure 4-3: Interaction related to Huh7; Sorafenib, PI3K-β inhibition and Sorafenib+PI3K-β inhibition

Table 4-3: Common survival mutations in Huh7 treatment with Sorafenib, PI3K-β inhibition and Sorafenib+PI3K-β inhibition

Common Genes	UniProtKB	Protein	Organism
SLC39A5	Q6ZMH5	Zinc transporter ZIP5	Homo sapiens

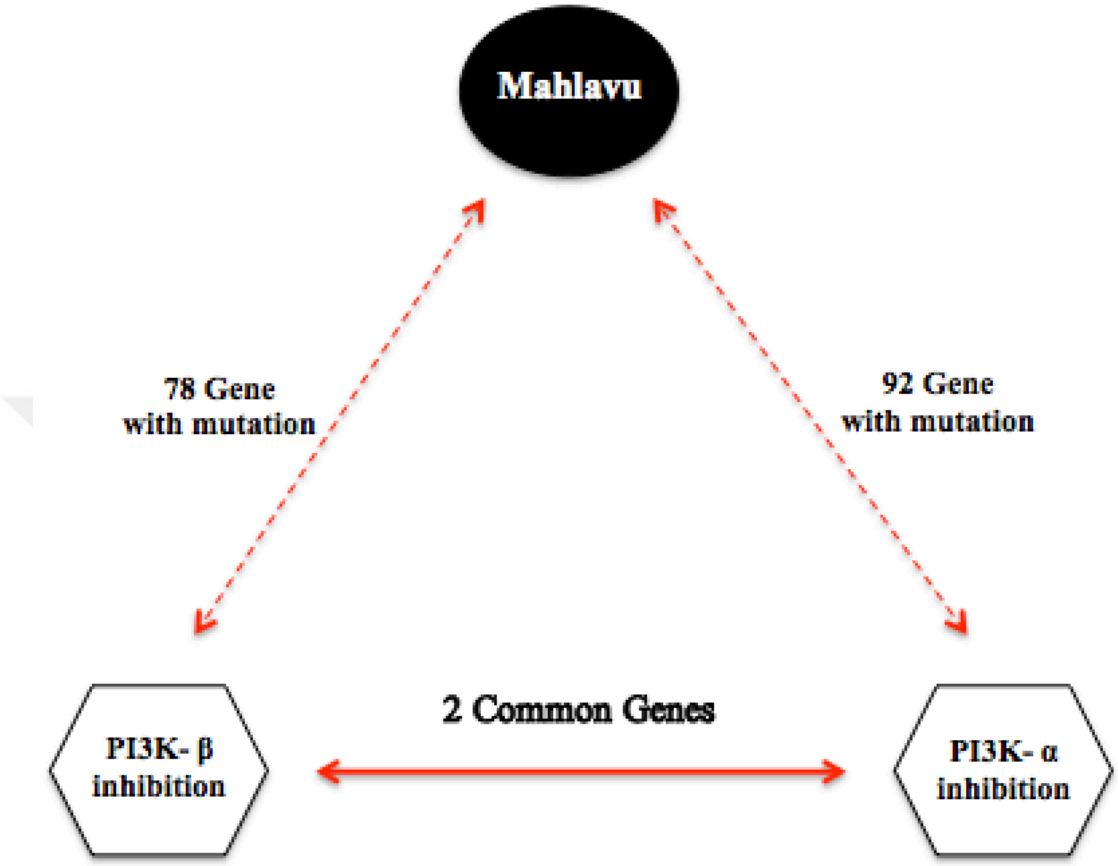


Figure 4-4: Interaction related to Mahlavu; PI3K-α inhibition and PI3K-β inhibition

Table 4-4: Common survival mutations in Mahlavu treatment with PI3K-α and PI3K-β inhibition

Common Genes	UniProtKB	Protein	Organism
FRG1	Q14331	Protein FRG1	Homo sapiens
CTC-512J12.4	-	-	Homo sapiens

4.2.2 Sorafenib, PI3K-α inhibition and Sorafenib+PI3K-α inhibition

Mahlavu cell line treatment with PI3K-α inhibition, Sorafenib and sorafenib+PI3K-α inhibition had 92, 65 and 116 survived genes, respectively. Higher number of mutated genes in Mahlavu combined treatment displayed the negative interaction between drugs, which increases resistant genes in the targeted cancer cell line. PI3K-α inhibition has 4 genes in common with each of the two other methods where, Sorafenib and sorafenib+PI3K-α inhibition have just three mutual genes. In total they have two survived genes in common; 1) *PPHLN1*, one of the shared genes among all methods, is a component of a multi-protein complex, and acts in some cells keratinization process to replace the cytoplasm with keratin. It also presents as co-repressor in transcription regulating processes of the cells. 2) *FRG1*, the second shared gene of whole comparison is explained previously, which also existed in common genes list of PI3K's α and β isoforms inhibition (Figure 4.5) (Table 4.5).

4.2.3 Sorafenib, PI3K-β inhibition and Sorafenib+PI3K-β inhibition

Last treatment comparison of Mahlavu was among Sorafenib, PI3K-β inhibition and their combination. The first treatment had 78 unaffected genes where the combined cure decreased this number to 72. Sorafenib with the least number of mutations (65 genes) obtained the highest effectiveness rate of treatment in this comparison. Each therapy method has some shared genes with others but they totally have 3 genes known as *CTC-512J12.4*, *PPHLN1* and *SRP9* in common that are preserved in all treatments. *CTC-512J12.4* and *PPHLN1* are repetitive from previous treatments. *SRP9* produce an RNA binding protein, which interacts with signal recognition particles and participates in protein secretion into the membrane to regulate the membrane function (Figure 4.6) (Table 4.6).

4.3 Outcome

As a summation, it is notable that *FRG1* exists in all of the Mahlavu treatments and two Huh7 treatments (Sorafenib and its combination with PI3K-β's inhibition). *CTC-512J12.4* and *PPHLN1* come next by being resistant to all treatments except one. And in the end *SRP9* stays in the third rank showing resistance to three of the mentioned therapies.

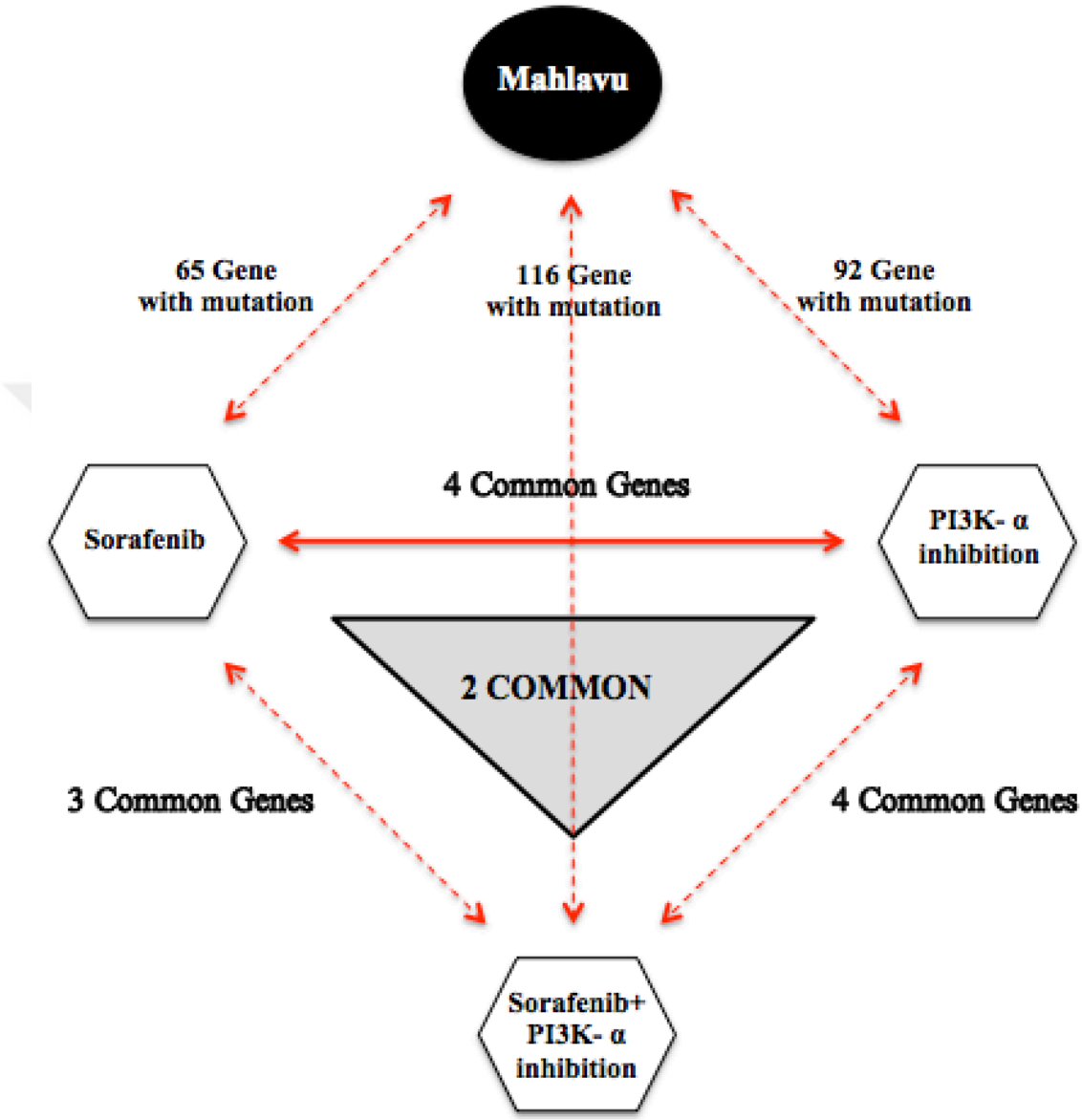


Figure 4-5: Interaction related to Mahlavu; Sorafenib, PI3K-α inhibition and Sorafenib+PI3K- α inhibition

Table 4-5: Common survival mutations in Mahlavu treatment with Sorafenib, PI3K-α inhibition and Sorafenib+PI3K-α inhibition

Common Genes	UniProtKB	Protein	Organism
FRG1	Q14331	Protein FRG1	Homo sapiens
PPHLN1	Q8NEY8	Periphilin-1	Homo sapiens

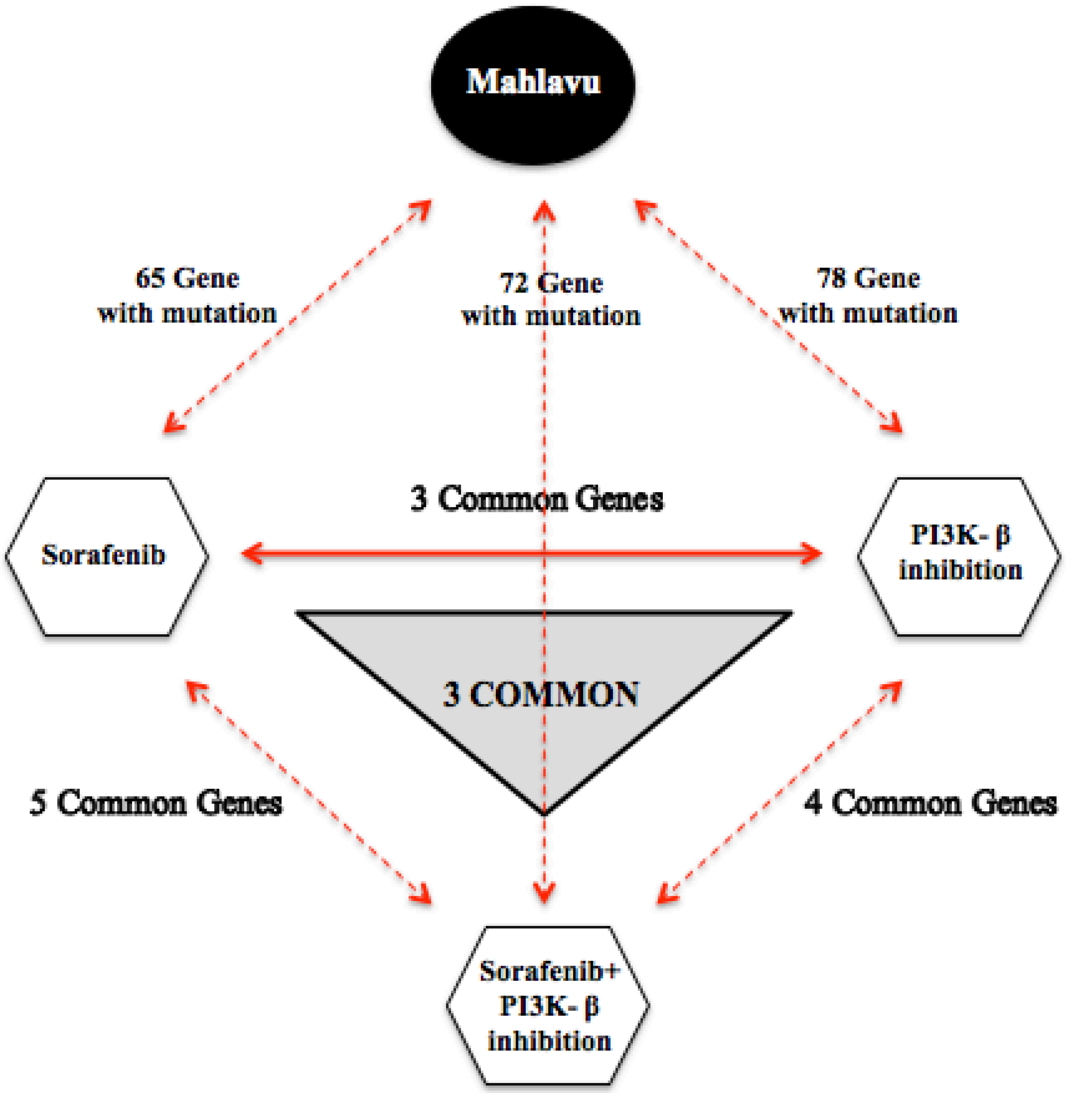


Figure 4-6: Interaction related to Mahlavu; Sorafenib, PI3K-β inhibition and Sorafenib+PI3K-β inhibition

Table 4-6: Common survival mutations in Mahlavu treatment with Sorafenib, PI3K-β inhibition and Sorafenib+PI3K-β inhibition

Common Genes	UniProtKB	Protein	Organism
FRG1	Q14331	Protein FRG1	Homo sapiens
CTC-512J12.4	-	-	Homo sapiens
SRP9	P49458	Signal recognition particle 9 kDa	Homo sapiens

CHAPTER 5

DISCUSSION

5.1 Summary

In this project, we aimed to find and investigate the involved genes in Hepatocellular Carcinoma that assist the tumor tissue in surviving, progressing and spreading to other parts of the body, acquiring some special traits. In this regard, frequently mutated genes were detected by comparing the applied treatments' on HCC related cancerous samples. The relation of target cell line with common mutations of all treatments was interpreted, considering the survived genes' roles and characteristics in the body. Normal cell's function is under the supervision of different regulators to preserve the biological system discipline and stability. In contrast, the activity of regulating factors is suppressed or circumvented in cancerous cells to get the capability of proliferating uncontrollably, independent from all limitative factors. As a whole, a number of general principles fulfillment, considered as cancer fundamentals, is required for categorizing the abnormal tissues as cancerous ones. Mentioned principles are actually cancer's essential traits, which were originally classified in six groups as sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis (Hanahan & Weinberg, 2011). In addition, newly established two enabling characteristics called "inflammation" and "genomic instability" along with emerging characteristics as "avoiding immune destruction" and "deregulating cellular energetics" have been developed the domain of cancer hallmarks.

Cancer cells maintain proliferating capability by producing growth signals and increasing related receptors to become independent from regulating growth factors. Additional to limited growth signals, the antigrowth factors, which prevent the cells growth in unusual conditions, are also disrupted in tumor tissues to get the chance for irregular dividing. The other barrier for cell division is apoptosis, which causes cells to die as a result of morphological changes, whereas the cancerous cells are able to rescue and stay alive by getting resistance to this habitual process. In contrast to normal cells, cancerous cells are also capable of infinitive replication as their chromosome telomeres, the division supporters, are maintained and not truncated with cell division. The next

trait facilitating cancer activity is the formation of new feeding vessels for nutrients provision in tumor's required amount. The defined function, known as angiogenesis, is conducted through specific growth factors and assists the spread of tumor domains by carrying the pioneer seeds to other parts of the body (metastasis). Later, two stated features of cancer reveal the mutated cells' effect on rearrangement of metabolism corresponding to their desires. They escape from clutches of immune system suppressors by abusing their inflammatory responses in favor of progression.

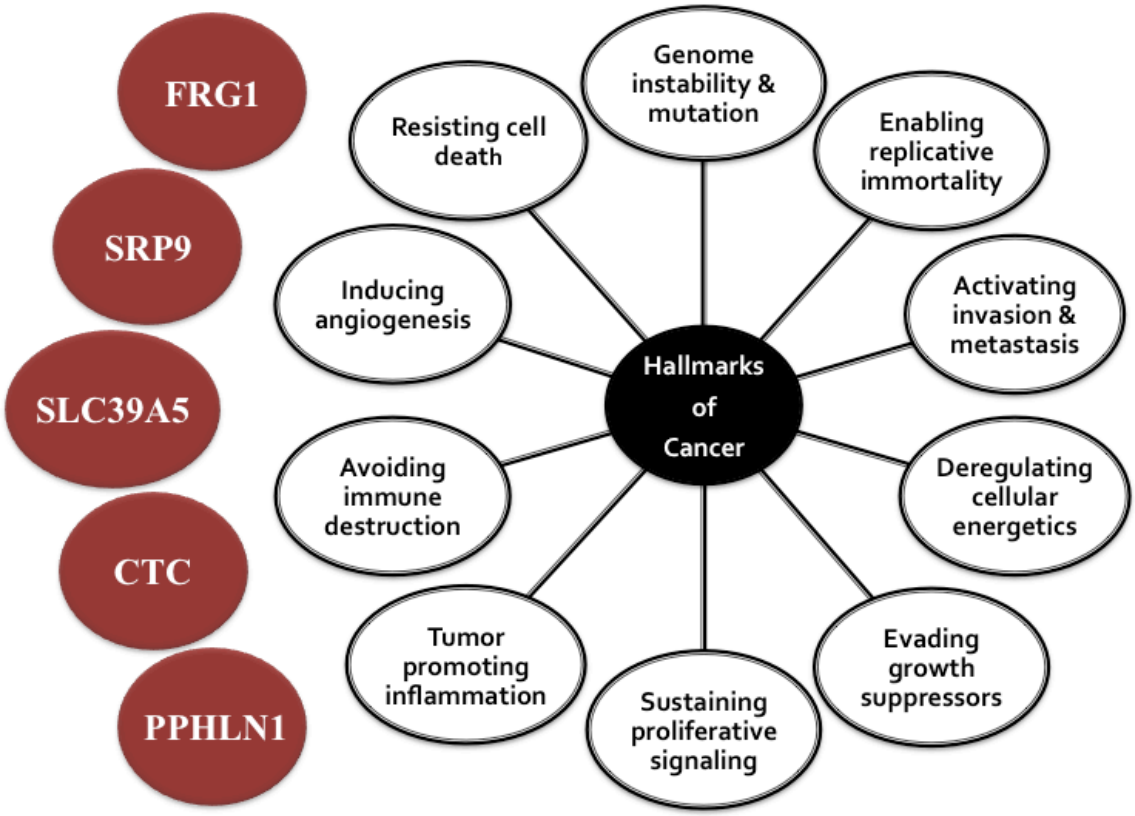


Figure 5-1: Hallmarks of cancer. The six original cancer hallmarks and established enabling and emerging characteristics in 2011(Adapted from Hanahan & Weinberg, 2011).

We investigated each detected gene in detail and inferred from remarkably exhibited cancerous characteristics (Table 5.1), we verified their interference in Hepatocellular Carcinoma.

Supported by Ryan D. Wuebbles et al (2009) FSHD region gene (FRG1), the explicitly resistant gene of all Mahlavu treatments, is crucial for angiogenesis and linked to vasculature development. This muscular dystrophy related gene is mainly observable in

patients with vascular disordering that the gene knockdown leads to decrease in angiogenesis. As previously illustrated, angiogenesis is one of the essential hallmarks of cancer. In cancer condition, the FRG1 gene, containing this specificity, is overexpressed to increase angiogenesis for oxygen supply and maintenance. Additionally, at recent international liver cancer association conference, FRG1 is reported as one of the novel targets for therapeutic affairs of intrahepatic cholangiocarcinoma (iCCA) ("IMISCOE 9 th Annual Conference," 2015; Wuebbles, Hanel, & Jones, 2009).

PPHLN1, common among four Mahlavu treatment results, is the other survived gene, which displays its biological function by applying regularity effects on transcription. Periphilin-1 (PPHLN1) is also known as CDC7 expression repressor (CR) and Gastric cancer antigen Ga50. CDC7 is a participant in DNA replication, which is repressed by PPHLN1 overexpression. Subsequently, CDC7 down-regulation leads to S phase interruption and captures its progression. This mechanism is induced by some anti-cancer drugs like Hydroxyurea to inhibit DNA replication (Kurita et al., 2004). Further, Jie Wang et al investigation about autoantibodies' relation with basal-like breast cancer resulted-in 13 anti-bodies secretion including PPHLN1. These remarkable antibodies are observable in cancer condition and in comparison to control samples they seem like valuable targets for therapeutic purposes. As a conclusion, these declaration reveal the relation of PPHLN1 with breast and gastric cancers (Wang et al., 2015).

CTC is the ZNf229 partial coding sequence, which appears as a resistant gene in four Mahlavu treatments. ZNF229, the Zinc finger protein 229, is a DNA binding protein and biologically acts as transcription regulator. In Haijun Gong et al's survey "the detection of effective pathways and genes in pancreatic cancer", four signaling pathways and 15 genes' direct relation with the disease has revealed. DNA-dependent transcription pathway is one of the participants, including six various genes like ZNF229, ensures the development of pancreatic related tumors. As a result of this experiment most of the zinc finger family members deal with disordering caused by kidney and pancreas cancer (Gong, Wu, & Clarke, 2014).

The signal recognition particle 9 kDa protein (SRP9) is the other discussable survived gene, being shared among three Mahlavu treatments. SRP9 conducts the process of protein targeting to membrane during translation. Jung-hyun Rho et al survey in 2008 with the purpose of detecting effective biomarkers in cancer therapy demonstrates the up-regulation of SRP9 in colorectal adenocarcinoma (Rho, Qin, Wang, & Roehrl, 2008). They also refer to Yuefang Liu et al survey in 2007, which verifies the SRP9 mRNA up-regulation in hepatocellular carcinomas to involve in different processes like transcription. They interpret the particular increase of membrane protein synthesis as an inevitable action correspond to higher metabolic activities in neoplastic cells (Liu et al., 2007).

Similar as FRG1 in Mahlavu treatment, SLC39A5 possesses the highest rate of resistance to Huh7 applied drugs, shared among three treatment results. SLC39A5, the zing transporter ZIP5, with regulatory effect on TGF beta signaling pathway supports its

intense anti-proliferating activities and assists pathway's tumor suppressing duty. In progressive cancer stages the growth suppressing activity of TGF beta pathway can switch to tumor cells' proliferation activator (Guo et al., 2014; Hanahan & Weinberg, 2011). Jin J et al survey (2015) is another experiment, which verifies the importance of SLC39A5 in cancer treatment, indicating the significant inhibitory effect of ZIP5 knockdown on esophageal squamous cell carcinoma (ESCC). Gene alteration leads to disturbance in proliferation that various gene-expression determining methods state the highest level of ZIP5 expression in ESCC and lowest amount of it in normal tissues. It also mentions the 28 and 38% reduction in proliferation rate, according to the MTT and CCK-8 assays, respectively (Jin J et al., 2015). In addition Benjamin P. Weaver et al reported the participation of ZIP4, another member of zinc transporter family, in HCC by inhibiting the cell death, enhancing cell cycle and finally increasing the invasion ability of cancerous cells. As a result the gene knockdown leads to apoptosis and returns the stability back to the cells (Weaver et al., 2010).

FRG1 is the most remarkable gene, which is shared in seven data sets out of ten. This gene is also known as one of the driver genes in Prostate adenocarcinoma. Besides FRG1, CTTN and CDC27, which are common between two data sets, act as drivers in Lung squamous cell carcinoma and Prostate adenocarcinoma, respectively (Chang et al., 2013).

Table 5-1: Significant genes' relation with different cancers

FRG1	• Resistant gene of all Mahlavu treatments and two Huh7 treated samples • Crucial for angiogenesis and linked to vasculature development (Wuebbles, 2009) • A novel target for intrahepatic cholangiocarcinoma (iCCA) treatment ("IMISCOE, 2015) • Driver gene in Prostate adenocarcinoma
PPHLN1	• Common among four Mahlavu treatment results • Regulate transcription, CDC7 expression repressor, Gastric cancer antigen (Kurita, 2004) • A secreted antibody in basal-like breast cancer and a therapeutic target (Wang, 2015)
CTC	• ZNf229 partial coding sequence, resistant gene in four Mahlavu treatments • Effective in pancreatic cancer as a component of DNA-dependent transcription pathway • ZNF family members deal with disordering by kidney and pancreas cancer (Gong, 2014)
SRP9	• Shared among three Mahlavu treatments, protein targeting to membrane during translation • Effective biomarker in cancer therapy, up-regulation in colorectal adenocarcinoma • Up-regulation in hepatocellular carcinomas to involve in different processes (Rho, 2008)
SLC39A5	• Highest rate of resistance to Huh7 applied drugs, shared among three treatment results • Zing transporter,Tumor suppressor, switch to proliferation activator in cancer (Guo, 2014) • Inhibitory effect of ZIP5 knockdown on esophageal squamous cell carcinoma (Jin, 2015) • ZIP4 participates in HCC by inhibiting the cell death (Weaver, 2010)

More detailed investigation reveals the critical alterations, which transform the type of amino acids in significant mutations (Figure 5.2). FRG1, the common survived gene of whole Mahlavu and some of the Huh7 applied treatments, includes the amino acid changes from Leucine, Glycine, Arginine and Phenylalanine to Glutamine, Valine, Histidine and Serine, respectively. Amino acids are categorized in different groups according to their side-chains and charges. The main alterations result-in coding new amino acids, where their types are different from the initial ones. It is observable that mutations in FRG1 generally switched the hydrophobic side-chain to polar amino acids. However, the FRG1 results contain another type of mutation, which alters Glycine, a special case with just hydrogen in its R group position, to an amino acid with hydrophobic side chain (Valine). The mutation in PPHLN1 leads to change from Leucine to Isoleucine, which is observable in four Mahlavu treatments. The affected amino acid in CTC mutation is Serine with polar uncharged side-chain. Serine switches to Glycine (a special case) as a result of all Mahlavu applied treatments except one. SRP9 mutations represent the change from a polar amino acid, Serine, to a one from the special cases group, Proline, as an effect of three Mahlavu treatment compounds. In addition the change from Phenylalanine to Leucine is detected in one of the Mahlavu's treated cases. The SLC39A5 mutation displays the change from Serine to Proline in three huh7 treatments like SRP9 in Mahlavu treatments.

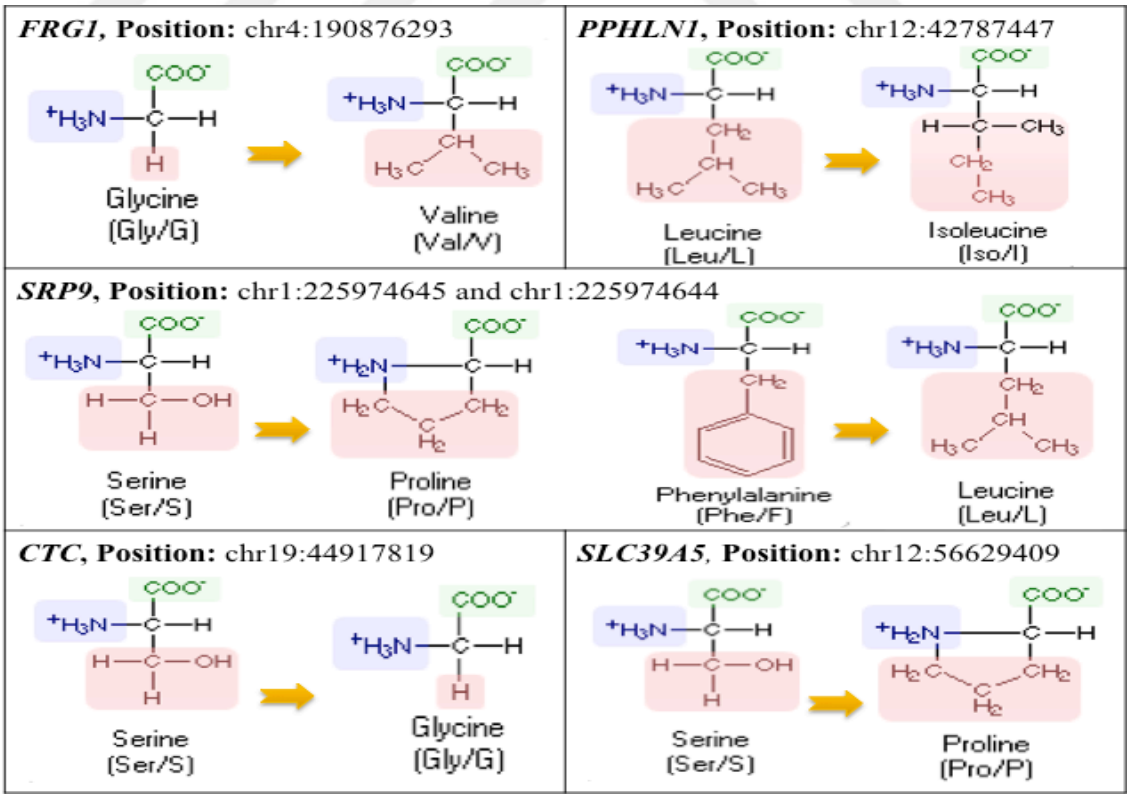
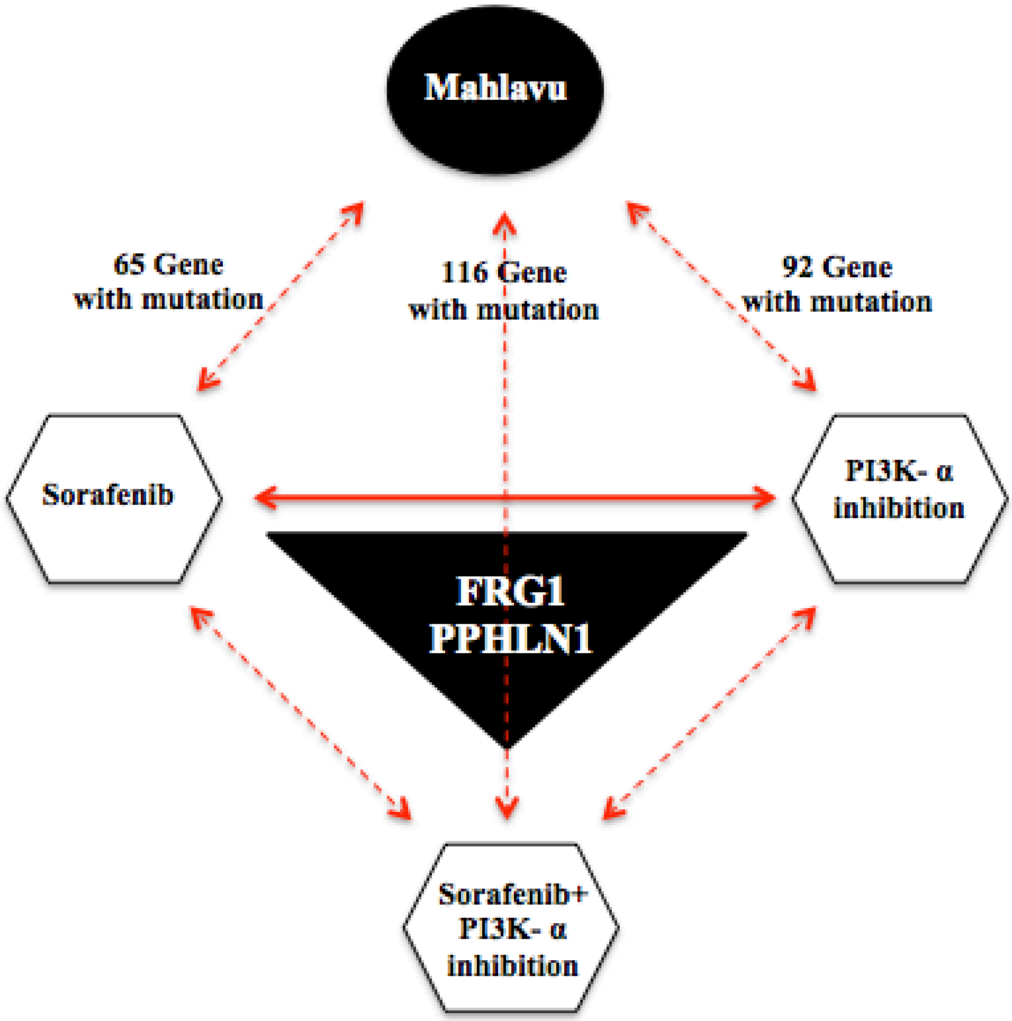


Figure 5-2: Mutations effect on amino acid changes

5.2 Future Direction

There are many genes in HCC that are involved in resistance to various treatments. Differential gene expression is a way to evaluate the impact of these genes on disease progression. In this study, we have used an approach, which involves chemical knockdown by various Akt Pathway inhibitors. The mutated genes that we identified during the chemical knockdown studies can be further studied in gene expression vs. patient survival data. As shown in Figure5.3, FRG1 gene expression can be exploited more in detail in clinics for a patient follow-up, because we found the frg1 gene mutations are correlated with liver cancer under drug treatments.



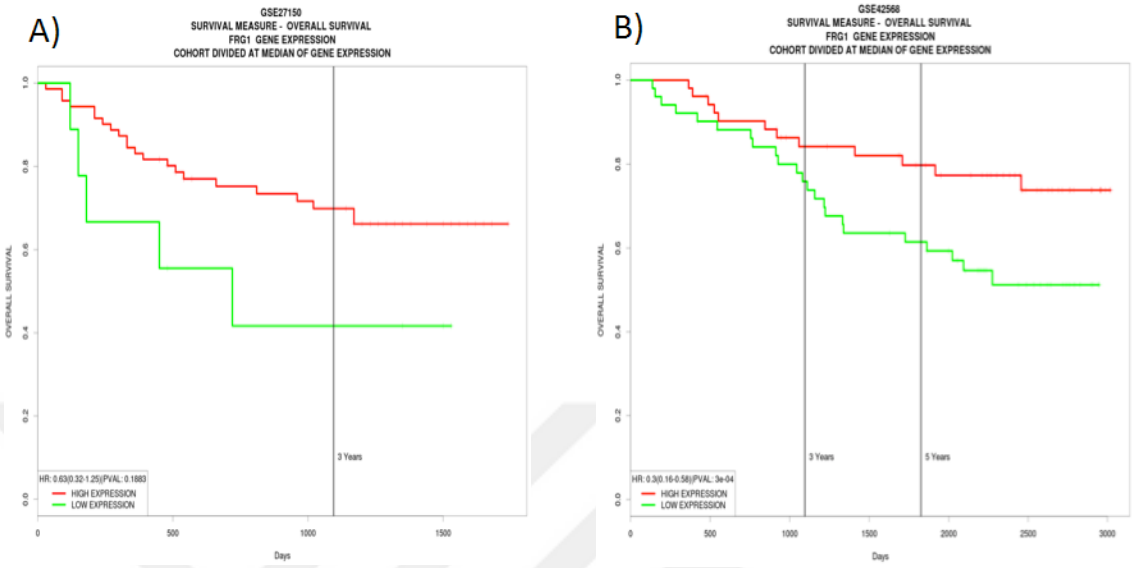


Figure 5-3: FRG1 gene expression level and survival correlation analysis in A) Hepatocellular Carcinoma vs. matched noncancerous liver tissue, B) Breast cancer

In efficacy validation, the genes with highest survival rank (resistant to most of the applied treatments) are the most remarkable genes and have the priority for investigation. The candidate genes that are found in this study can be exploited in therapeutic response in patients. The significant genes with their specific survival mutations can then be used as biomarkers for the planning and the follow-up of treatment in liver cancer as well as other cancers.

Furthermore online databases such as "TCGA" (http://cancergenome.nih.gov/), "METABRIC" (https://www.synapse.org/#!Synapse:syn1688369/wiki/27311) and "CCLE" (http://software.broadinstitute.org/software/cprg/?q=node/11), can be used for identifying genes related with survival from the list of genes detected in table 4.1 in addition to the selected genes in table 5.1.

REFERENCES

Ahmed, I., & Lobo, D. N. (2009). Malignant tumours of the liver. *Surgery*, *27*(1), 30–37. http://doi.org/10.1016/j.mpsur.2008.12.005

Altekruse, S. F., McGlynn, K. A., & Reichman, M. E. (2009). Hepatocellular carcinoma incidence, mortality, and survival trends in the United States from 1975 to 2005. *Journal of Clinical Oncology*, *27*(9), 1485–1491. http://doi.org/10.1200/JCO.2008.20.7753

Bréchot, C. (1996). Hepatitis C virus: molecular biology and genetic variability. *Digestive Diseases and Sciences*, *41*(12 Suppl), 6S–21S. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9011478

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1842*(10), 1932–1941. http://doi.org/10.1016/j.bbadis.2014.06.015

Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., … Shaw, K. R. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. http://doi.org/10.1038/ng.2764

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*(7), 1–5. http://doi.org/10.1371/journal.pbio.1001091

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219. http://doi.org/10.1038/nbt.2514

Duffy, D. J., & Bv, D. E. (n.d.). The Boost C ++ Libraries : Part II, 80–87.

Forbes, S. a, Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., & Menzies, A. (2009). *Europe PMC Funders Group The Catalogue of Somatic Mutations in Cancer (COSMIC)*. http://doi.org/10.1002/0471142905.hg1011s57.

Frith, M. C., Pheasant, M., & Mattick, J. S. (2005). The amazing complexity of the human transcriptome. *European Journal of Human Genetics : EJHG*, *13*(8), 894–7. http://doi.org/10.1038/sj.ejhg.5201459

Fritsch, C., Huang, A., Chatenay-Rivauday, C., Schnell, C., Reddy, A., Liu, M., … Sellers, W. R. (2014). Characterization of the Novel and Specific PI3Kα Inhibitor NVP-BYL719 and Development of the Patient Stratification Strategy for Clinical Trials. *Molecular Cancer Therapeutics*, *13*(5), 1117–29. http://doi.org/10.1158/1535-7163.MCT-13-0865

Gong, H., Wu, T. T., & Clarke, E. M. (2014). Pathway-gene identification for pancreatic cancer survival via doubly regularized Cox regression. *BMC Systems Biology*, *8*(Suppl 1), S3. http://doi.org/10.1186/1752-0509-8-S1-S3

Guo, H., Jin, X., Zhu, T., Wang, T., Tong, P., Tian, L., … Xia, K. (2014). SLC39A5 mutations interfering with the BMP/TGF-β pathway in non-syndromic high myopia. *Journal of Medical Genetics*, *51*, 518–25. http://doi.org/10.1136/jmedgenet-2014-102351

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*(5), 646–674. http://doi.org/10.1016/j.cell.2011.02.013

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, *22*(9), 1760–1774. http://doi.org/10.1101/gr.135350.111

IMISCOE 9 th Annual Conference. (2015), (September), 9.

Kurita, M., Suzuki, H., Masai, H., Mizumoto, K., Ogata, E., Nishimoto, I., … Matsuoka, M. (2004). Overexpression of CR/periphilin downregulates Cdc7 expression and induces S-phase arrest. *Biochemical and Biophysical Research Communications*, *324*(2), 554–561. http://doi.org/10.1016/j.bbrc.2004.09.083

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Meth*, *9*(4), 357–359. Retrieved from http://dx.doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Liu, Y., Zhu, X., Zhu, J., Liao, S., Tang, Q., Liu, K., … Feng, Z. (2007). Identification of differential expression of genes in hepatocellular carcinoma by suppression subtractive hybridization combined cDNA microarray. *Oncology Reports*, *18*(4), 943–951. http://doi.org/10.3892/or.18.4.943

Luqmani, Y. A. (2005). Mechanisms of drug resistance in cancer chemotherapy. *Medical Principles and Practice*, *14*(SUPPL. 1), 35–48. http://doi.org/10.1159/000081921

Meynert, A. M., Ansari, M., FitzPatrick, D. R., & Taylor, M. S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics*, *15*(1), 247. http://doi.org/10.1186/1471-2105-15-247

Nakabayashi, H., Taketa, K., Miyano, K., Yamane, T., & Sato, J. (1982). Growth of Human Hepatoma Cell Lines with Differentiated Chemically Defined Medium, *42*(September), 3858–3863.

Oefinger, P. E., Bronson, D. L., & Dreesman, G. R. (1981). Induction of hepatitis B surface antigen in human hepatoma-derived cell lines. *The Journal of General Virology*, *53*(Pt 1), 105–13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6268737

Phin, S., Moore, M. W., & Cotter, P. D. (2013). Genomic Rearrangements of PTEN in Prostate Cancer. *Frontiers in Oncology*, *3*(September), 240. http://doi.org/10.3389/fonc.2013.00240

Pruitt, K., Brown, G., Tatusova, T., & Maglott, D. (2002). The Reference Sequence (RefSeq) Database. *The NCBI Handbook*, 1–24.

Rho, J. H., Qin, S., Wang, J., & Roehrl, M. H. A. (2008). Proteomic expression analysis of surgical human colorectal cancer tissues: Up-regulation of PSB7, PRDX1, and SRP9 and hypoxic adaptation in cancer. *Journal of Proteome Research*, *7*(7), 2959–2972. http://doi.org/10.1021/pr8000892

Schwartz, S., Wongvipat, J., Trigwell, C. B., Hancox, U., Carver, B. S., Rodrik-Outmezguine, V., … Rosen, N. (2015). Feedback suppression of PI3K?? signaling in PTEN-mutated tumors is relieved by selective inhibition of PI3K?? *Cancer Cell*, *27*(1), 109–122. http://doi.org/10.1016/j.ccell.2014.11.008

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–11. http://doi.org/10.1093/nar/29.1.308

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. http://doi.org/10.1093/bioinformatics/btp120

Vale, P. (2013). Multidisciplinary Treatment of Hepatocellular Carcinoma. *Recent Results in Cancer Research*, *190*, 21–33. http://doi.org/10.1007/978-3-642-16037-0

Wang, J., Figueroa, J. D., Wallstrom, G., Barker, K., Park, J. G., Demirkan, G., …

LaBaer, J. (2015). Plasma autoantibodies associated with basal-like breast cancers. *Cancer Epidemiology Biomarkers and Prevention*, *24*(9), 1332–1340. http://doi.org/10.1158/1055-9965.EPI-15-0047

Weaver, B. P., Zhang, Y., Hiscox, S., Guo, G. L., Apte, U., Taylor, K. M., … Andrews, G. K. (2010). Zip4 (Slc39a4) expression is activated in hepatocellular carcinomas and functions to repress apoptosis, enhance cell cycle and increase migration. *PLoS ONE*, *5*(10). http://doi.org/10.1371/journal.pone.0013158

Wu, X., & Li, Y. (2012). Signaling Pathways in Liver Cancer. *Liver Tumors*, 37–58. Retrieved from http://www.intechopen.com/source/pdfs/27576/InTech-Signaling_pathways_in_liver_cancer.pdf

Wuebbles, R. D., Hanel, M. L., & Jones, P. L. (2009). FSHD region gene 1 (FRG1) is crucial for angiogenesis linking FRG1 to facioscapulohumeral muscular dystrophy-associated vasculopathy. *Disease Models & Mechanisms*, *2*(5-6), 267–74. http://doi.org/10.1242/dmm.002261

Yang, S. S., Zhang, K., Vieira, W., Taub, J. V, Zeilstra-ryalls, J. H., & Somerville, R. L. (1990). A Human Hepatocellular Carcinoma 3.0-Kilobase DNA Sequence Transforms Both Rat Liver Cells and NIH3T3 Fibroblasts and Encodes a 52-Kilodalton Protein1'2, (Cm 22131).

Yuzugullu, H., Benhaj, K., Ozturk, N., Senturk, S., Celik, E., Toylu, A., … Ozturk, M. (2009). Canonical Wnt signaling is antagonized by noncanonical Wnt5a in hepatocellular carcinoma cells. *Molecular Cancer*, *8*, 90. http://doi.org/10.1186/1476-4598-8-90

Zahreddine, H., & Borden, K. L. B. (2013). Mechanisms and insights into drug resistance in cancer. *Frontiers in Pharmacology*, *4 MAR*(March), 1–8. http://doi.org/10.3389/fphar.2013.00028

APPENDICES

APPENDIX A

QUALITY CONTROL RESULTS

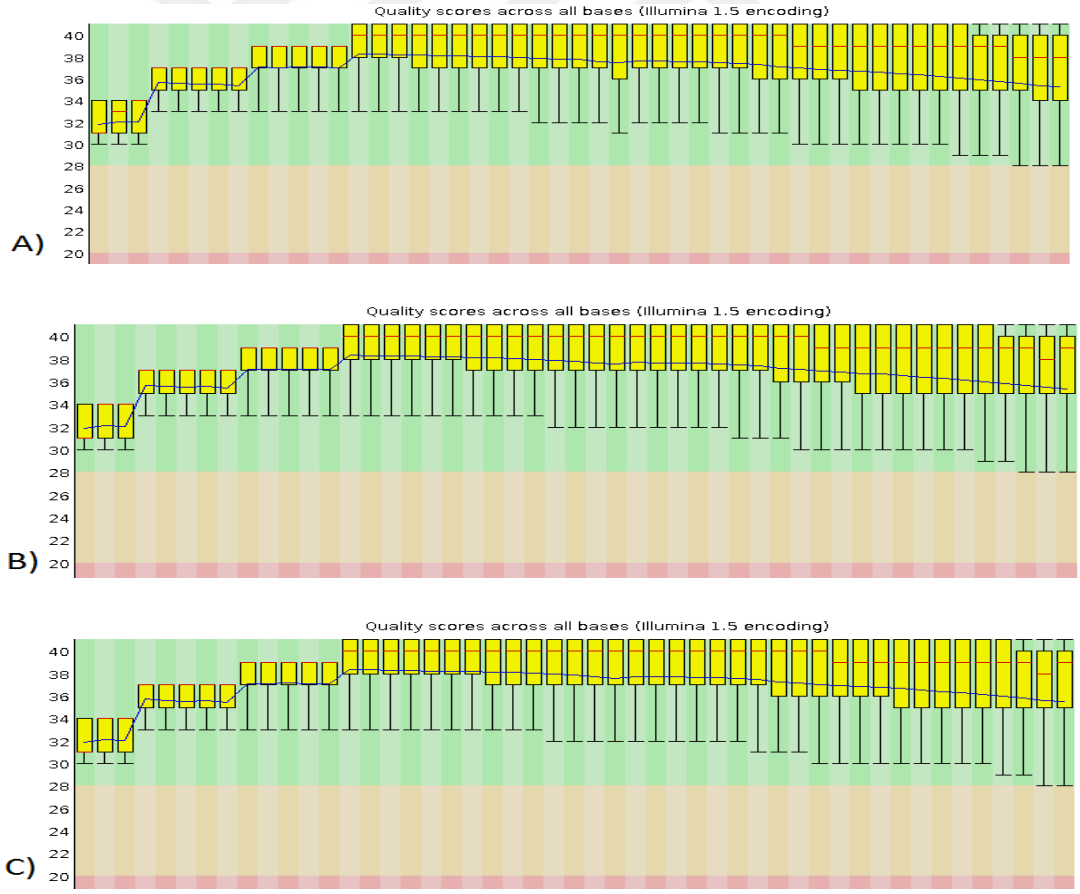


Figure A.1: Per base sequence quality in A) H-BETA, B) H-S-ALPHA, C) H-S-BETA

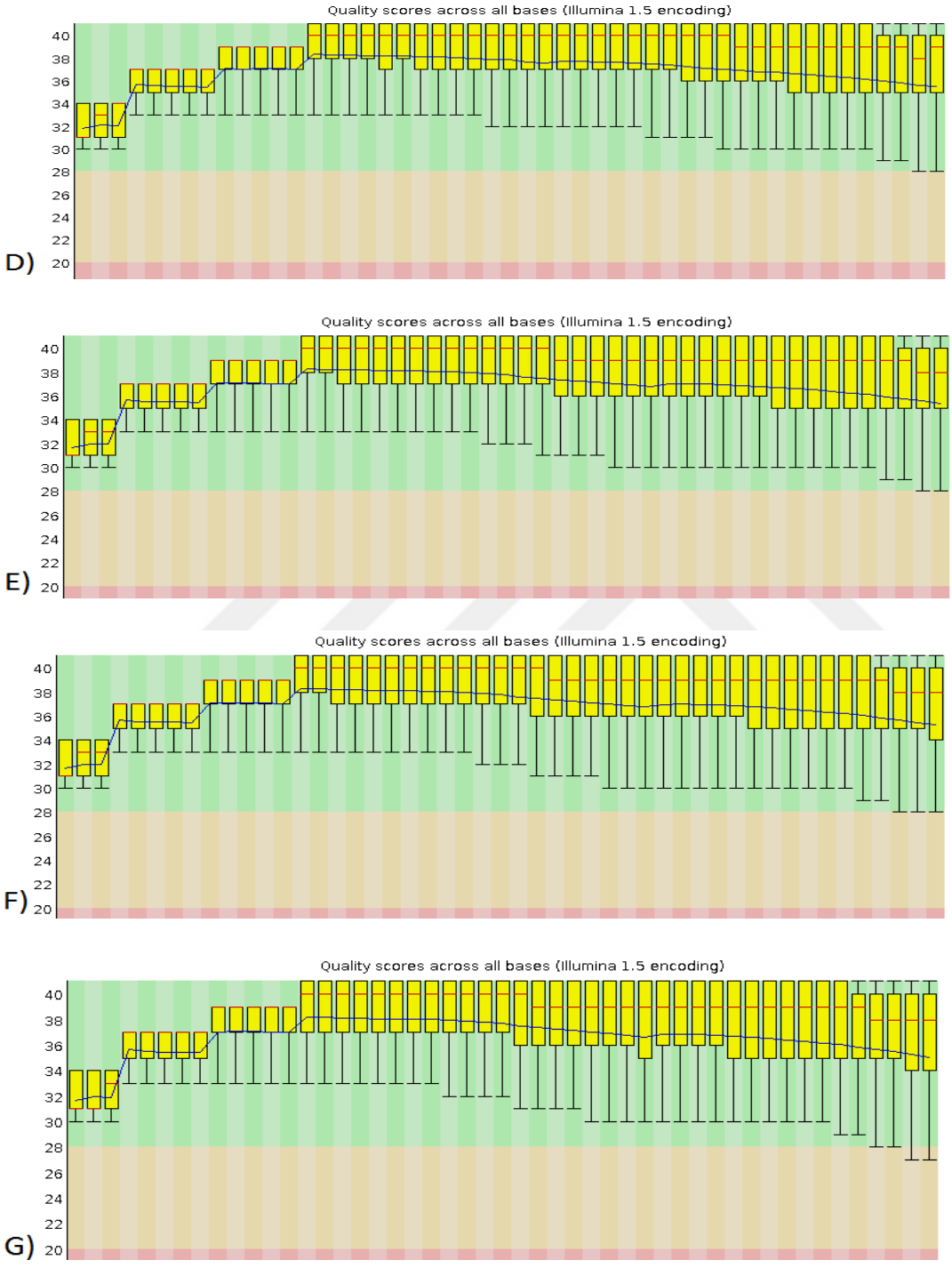


Figure A.2: Per base sequence quality in D) H-SOR, E) M-ALPHA, F) M-BETA, G) M-S-ALPHA

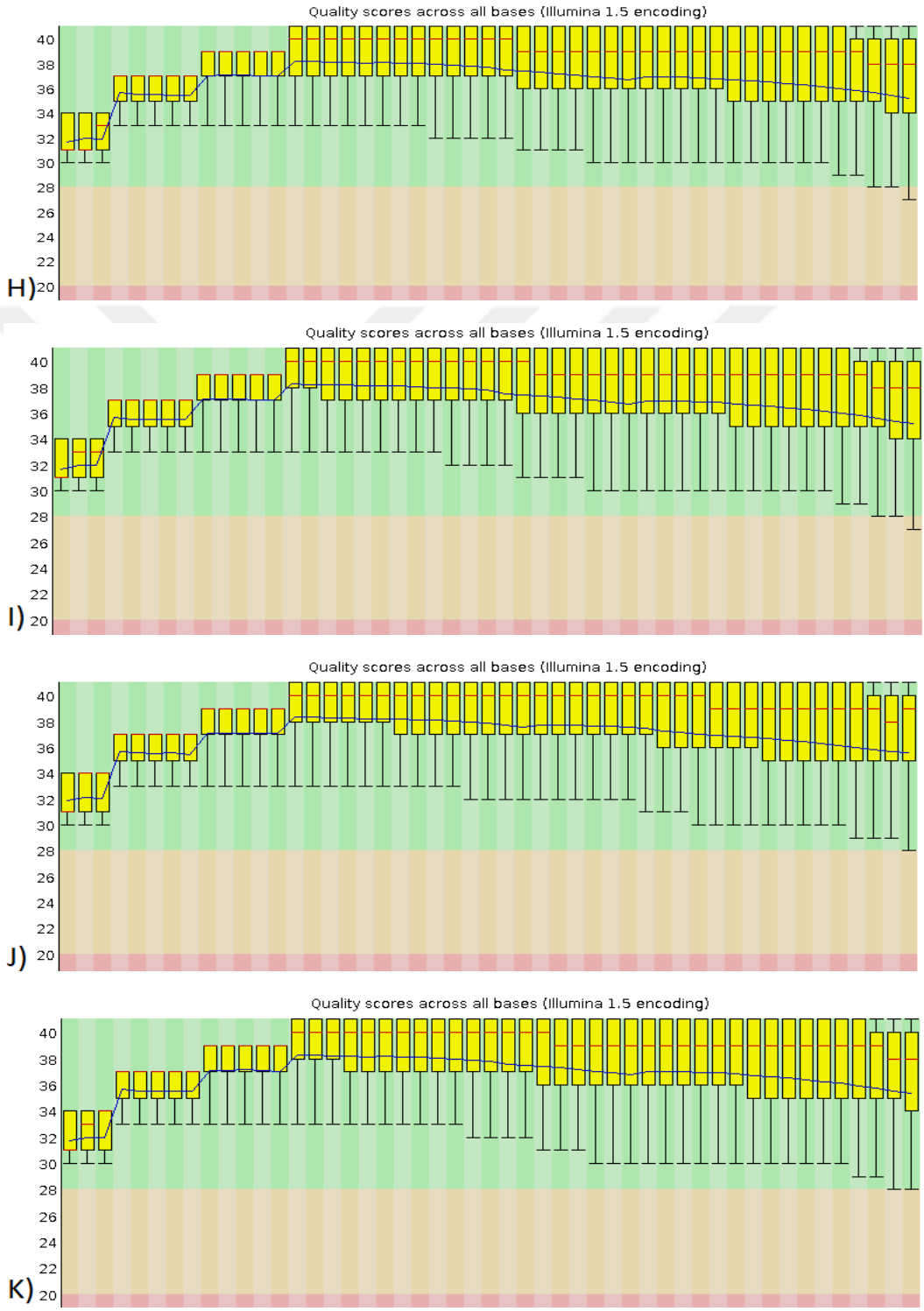


Figure A.3: Per base sequence quality in H) M-S-BETA, I) M-SOR, J) HDMSO, K) MDMSO

APPENDIX B

TOPHAT COMPILATION

We need to download and install boost libraries to initiate the tophat procedure. Version 1.59 is selected and installed according to requested instructions. First, in order to install and build the boost we downloaded and made a specific directory for it. Running the codes below gave us the Boost.Build to develop our system.

```
./bootstrap.sh

./b2 install

b2      --build-dir=["The-directory-where-boost-builds"]      toolset=
["toolset-name"]   stage
```

In addition, using *touch* command and defining the path in the shell we saved the installation path on the system to prevent any probable disruption in finding boost built file by Tophat. In the next step we downloaded and installed Tophat in the same directory with boost. The default directories are set to be local/bin directory.

```
./configure --with-boost=/path/to/boost_prefix_dir
```

With *configure* command we specify the needed directory for installation and with *"with-boost"* we specify the path needed for our installed boost. Now we run Tophat installation with following commands.

```
make

make install
```

The other required program, Bowtie is also downloaded and installed like pervious steps in the same directory. All the paths consisting boost, Tophat and Bowtie should be defined to the computer so that the program work in a harmony. In addition Bowtie must be indexed to our reference genome, which is downloadable from iGenomes.

APPENDIX C

SOURCE CODES

C.1 Gene Name Detection

```
output = []
import pyensembl
ensembl = pyensembl.EnsemblRelease()

f = open('mutect driven data') #open the file in read universal
mode
for line in f:
    cells = line.split( "," ) #comma separated file
    output.append( ( cells[ 1 ], cells[ 2 ] ) ) #since we want
the first and the second columns(contig & position)
f.close()
list1 = []
list2 = []

for i in range (1, len(output)): #for first column(contig row
number)
    con = output[i][0]
    list1.append(con)

for j in range (1, len(output)): #for second column(position row
number)
    pos = output[j][1]
    list2.append(pos)

list2 = [int(i) for i in list2] # list2 items as integer

list1 = ' '.join(list1).replace('chr','').split() #remove chr
phrase from list1 items
#print list1

list3 = []
import pyensembl
ensembl = pyensembl.EnsemblRelease(release=75) #use
hg19(release=75)
for i in range (0, len(list1)): #dont have titles anymore
    genes = ensembl.genes_at_locus(contig= list1[i], position=
list2[i]) #give contig & position to ensembl
    print genes
```

C.2 Mutation Type Detection

```
di = []
h = open( 'RefSeq data' ) #refseq data from ucsc, keep CDS rows
only(remove exon rows)
for line in h:
    cells = line.split( "," )
    di.append( ( cells[ 0 ], cells[ 1 ], cells[ 2 ], cells[ 3 ],
cells[ 4 ]) )
h.close()

d = []
for i in range (1, RefSeq Data Length): #row number of the file
    d.append([ di[i][ 0 ], di[i][ 1 ], di[i][ 2], di[i][ 3 ],
di[i][ 4 ]] )

output = []

f = open( 'mutect data' ) #open csv file of mutect result for
mv_mvalpha
for line in f:
    cells = line.split( "," )
    output.append( ( cells[ 0 ], cells[ 1 ], cells[ 2 ], cells[
3 ], cells[ 4 ], cells[ 5 ] ) )
f.close()

list1 = []
n = 1
j = 3
for i in range(1,len(output)):
    text = (output[i][j]) #get the file "context" labeled column
which includes 7 characters
        text2 = (list (text))
        if n in range(1,len(output)):
             text2 [3] = (output[n][4]) #put the content of
"ref" labeled column instead of "context" x character
             X1 = text2
             list1.append(X1)

list2 = []
n = 1
j = 3
for i in range(1,len(output)):
        text3 = (output[i][j]) #get the file "context" column
which includes 7 characters
        text4 = (list (text3))
        if n in range(1,len(output)):
             text4 [3] = (output[n][5]) #put the content of
```

```
"alt" column in "context" x character
            X2 = text4
            list2.append(X2)


result1 =[]
for i in range (0, len(list1)):#we dont have labels anymore
    result1.append( [ list1[i][ 1 ], list1[i][ 2 ], list1[i][
3], list1[i][ 4 ], list1[i][ 5 ]] ) #remove the first and last
characters from "context" column since we dont need them for
this operation
    i += 1

result2 =[]
for i in range (0, len(list2)):
    result2.append( [ list2[i][ 1 ], list2[i][ 2 ], list2[i][
3], list2[i][ 4 ], list2[i][ 5 ]] ) #remove the first and last
characters from "context" column since we dont need them for
this operation
    i += 1


for i in range (0, len(result1)): # convert T to A, C to G, G to
C, A to U as we have RNA
    if result1[i][0] == 'T':
        result1[i][0] = 'A'
    elif result1[i][0] == 'C':
        result1[i][0] = 'G'
    elif result1[i][0] == 'G':
        result1[i][0] = 'C'
    elif result1[i][0] == 'A':
        result1[i][0] = 'U'
    if result1[i][1] == 'T':
        result1[i][1] = 'A'
    elif result1[i][1] == 'C':
        result1[i][1] = 'G'
    elif result1[i][1] == 'G':
        result1[i][1] = 'C'
    elif result1[i][1] == 'A':
        result1[i][1] = 'U'
    if result1[i][2] == 'T':
        result1[i][2] = 'A'
    elif result1[i][2] == 'C':
        result1[i][2] = 'G'
    elif result1[i][2] == 'G':
        result1[i][2] = 'C'
    elif result1[i][2] == 'A':
        result1[i][2] = 'U'
```

```
    if result1[i][3] == 'T':
        result1[i][3] = 'A'
    elif result1[i][3] == 'C':
        result1[i][3] = 'G'
    elif result1[i][3] == 'G':
        result1[i][3] = 'C'
    elif result1[i][3] == 'A':
        result1[i][3] = 'U'
    if result1[i][4] == 'T':
        result1[i][4] = 'A'
    elif result1[i][4] == 'C':
        result1[i][4] = 'G'
    elif result1[i][4] == 'G':
        result1[i][4] = 'C'
    elif result1[i][4] == 'A':
        result1[i][4] = 'U'
    i += 1
#print(result1, '\n')
#The same thing has done for result2

ch_name = []
pos_num = []
list3 = []
list4 = []
listi = []


for i in range (1, len(output)):
    con = output[i][1] #put contig column index in the con
variable
    ch_name.append(con)

for j in range (1, len(output)):
    pos = output[j][2] #put position column index in the pos
variable
    pos_num.append(pos)


c = 1
for i in range (0,len(ch_name)):
    ch = ch_name[i]
    po = pos_num[i]
    for j in range (0, RefSeq Data Length): #compare each item's
contig and position with refseq to find it's location and read
order
        c += 1
        if d[j][0] == ch and d[j][1] <= po and d[j][2] >= po:
            if d[j][3] == '+':
```

```
                if d[j][4] == '0\n':
                    list3.append([result1[i][2], result1[i][3],
result1[i][4]])
                    list4.append([result2[i][2], result2[i][3],
result2[i][4]])
                    listi.append(i)
                    break
                elif d[j][4] == '1\n':
                    list3.append([result1[i][1], result1[i][2],
result1[i][3]])
                    list4.append([result2[i][1], result2[i][2],
result2[i][3]])
                    listi.append(i)
                    break
                else:
                    list3.append([result1[i][0], result1[i][1],
result1[i][2]])
                    list4.append([result2[i][0], result2[i][1],
result2[i][2]])
                    listi.append(i)
                    break
            else :
                if d[j][4] == '0\n':
                    list3.append([result1[i][4], result1[i][3],
result1[i][2]])
                    list4.append([result2[i][4], result2[i][3],
result2[i][2]])
                    listi.append(i)
                    break
                elif d[j][4] == '1\n':
                    list3.append([result1[i][3], result1[i][2],
result1[i][1]])
                    list4.append([result2[i][3], result2[i][2],
result2[i][1]])
                    listi.append(i)
                    break
                else:
                    list3.append([result1[i][2], result1[i][1],
result1[i][0]])
                    list4.append([result2[i][2], result2[i][1],
result2[i][0]])
                    listi.append(i)
                    break

print (len(listi)) #we find its length to know the number of
values we got from refseq

for i in range (0, len(listi)): #put listi length number
```

```
# do this operation after finding the len((listi))
    print (listi[i])

list_ref1 = []
for i in range (0, len(listi)):
    ref_res = (''.join(list3[i])) #join each 3 chr in ref list
    i += 1
    list_ref1.append(ref_res)
for i in range (0, len(listi)):
    print (list_ref1[i])


list_alt1 = []
for i in range (0, len(listi)):
    alt_res = (''.join(list4[i])) #join each 3 chr in alt list
    i += 1
    list_alt1.append(alt_res)
for i in range (0, len(listi)):
    print (list_alt1[i])



listi_new = []           #run this part after executing the
previous step
k = 0
for i in range (1,length of listi):  #from first value to last
value of listi
    if listi[k] != i: #if a number is in list1 append it to
listi_new, otherwise append _
        listi_new.append('_')
    else:
        listi_new.append(listi[k])
        k += 1

for i in range (1,length of listi): #print listi_new
    print (listi_new[i])



listi = []
list_ref = []
list_alt = []

m = open('Codon List')
for line in m:
    cells = line.split( "," )
    listi.append( ( cells[ 4 ]) ) #row number which include
mutation
```

```
    list_ref.append( ( cells[ 2 ]) ) #ref codon list
    list_alt.append( ( cells[ 3 ]) ) #alt codon list
m.close()

n = 0
for i in listi:
    listi[n] = i.replace('\n','')
    n += 1

list_ref_new = []
j = 0
for i in range (0, len(listi)):
    if listi[i] == '_':
        list_ref_new.append('_')
    else:
        list_ref_new.append(list_ref[j])
        j += 1

for i in range (0, Codon List Length):
    print (list_ref_new[i])

list_alt_new = []
j = 0
for i in range (0, Codon List Length):
    if listi[i] == '_':
        list_alt_new.append('_')
    else:
        list_alt_new.append(list_alt[j])
        j += 1

for i in range (0, Codon List Length):
    print (list_alt_new[i])

list3 = []                  #start this part after finishing the
previous part
list4 = []
m = open('Ordered Codon List')
for line in m:
    cells = line.split( "," )
    list3.append( ( cells[ 5 ]) ) #ref codon list
m.close()

m = open('Ordered Codon List')
for line in m:
    cells = line.split( "," )
    list4.append( ( cells[ 6 ]) ) #alt codon list
m.close()
```

```
list3_new = []
list4_new = []
for i in range (0, len(list3)):
    j = list3 [i]
    list3_new.append(j)
    l = list4 [i]
    list4_new.append(l)
n = 0
for i in list4_new:
    list4_new[n] = i.replace('\n','')
    n += 1

codon_lookup = {
    ('AUU'): 'Ile',
    ('AUC'): 'Ile',
    ('AUA'): 'Ile',
    ('CUU'): 'Leu',
    ('CUC'): 'Leu',
    ('CUA'): 'Leu',
    ('CUG'): 'Leu',
    ('UUA'): 'Leu',
    ('UUG'): 'Leu',
    ('GUU'): 'Val',
    ('GUC'): 'Val',
    ('GUA'): 'Val',
    ('GUG'): 'Val',
    ('UUU'): 'Phe',
    ('UUC'): 'Phe',
    ('AUG'): 'Met',
    ('UGU'): 'Cys',
    ('UGC'): 'Cys',
    ('GCU'): 'Ala',
    ('GCC'): 'Ala',
    ('GCA'): 'Ala',
    ('GCG'): 'Ala',
    ('GGU'): 'Gly',
    ('GGC'): 'Gly',
    ('GGA'): 'Gly',
    ('GGG'): 'Gly',
    ('CCU'): 'Pro',
    ('CCC'): 'Pro',
    ('CCA'): 'Pro',
    ('CCG'): 'Pro',
    ('ACU'): 'Thr',
    ('ACC'): 'Thr',
    ('ACA'): 'Thr',
    ('ACG'): 'Thr',
```

```
    ('UCU'): 'Ser',
    ('UCC'): 'Ser',
    ('UCA'): 'Ser',
    ('UCG'): 'Ser',
    ('AGU'): 'Ser',
    ('AGC'): 'Ser',
    ('UAU'): 'Tyr',
    ('UAC'): 'Tyr',
    ('UGG'): 'Trp',
    ('CAA'): 'Gin',
    ('CAG'): 'Gin',
    ('AAU'): 'Asn',
    ('AAC'): 'Asn',
    ('CAU'): 'His',
    ('CAC'): 'His',
    ('GAA'): 'Glu',
    ('GAG'): 'Glu',
    ('GAU'): 'Asp',
    ('GAC'): 'Asp',
    ('AAA'): 'Lys',
    ('AAG'): 'Lys',
    ('CGU'): 'Arg',
    ('CGC'): 'Arg',
    ('CGA'): 'Arg',
    ('CGG'): 'Arg',
    ('AGA'): 'Arg',
    ('AGG'): 'Arg',
    ('UAA'): 'Stop',
    ('UAG'): 'Stop',
    ('UGA'): 'Stop',
    ('_'): '_',
}

for i in range(0, len(list3)):
    ref = list3_new[i]
    alt = list4_new[i]
    amino_ref = (codon_lookup.get(ref))
    amino_alt = (codon_lookup.get(alt))
    if amino_ref == 'Stop' or amino_alt == 'Stop':
        print ('Nonsense')
    elif amino_ref == '_' and amino_alt == '_':
        print ('_')
    elif amino_ref == amino_alt and amino_ref != '_' and
amino_alt != '_':
        print('Silent')
    else :
        print ('Missense')
    i += 1
```

C.3 Common Genes Detection

```
hal = []
hbe = []
hsal = []
hsbe = []
hsor = []


m = open('Missense Results of H-ALPHA')
for line in m:
    cells = line.split( "," )
    hal.append( ( cells[ 0 ]) )
m.close()

m = open('Missense Results of H-BETA')
for line in m:
    cells = line.split( "," )
    hbe.append( ( cells[ 0 ]) )
m.close()

m = open('Missense Results of H-S-ALPHA')
for line in m:
    cells = line.split( "," )
    hsal.append( ( cells[ 0 ]) )
m.close()

m = open('Missense Results of H-S-BETA')
for line in m:
    cells = line.split( "," )
    hsbe.append( ( cells[ 0 ]) )
m.close()

m = open('Missense Results of H-SOR')
for line in m:
    cells = line.split( "," )
    hsor.append( ( cells[ 0 ]) )
m.close()
n = 0
for i in hal:
    hal[n] = i.replace('\n','')
    n += 1
n = 0
for i in hbe:
    hbe[n] = i.replace('\n','')
    n += 1
n = 0
for i in hsal:
```

```
    hsal[n] = i.replace('\n','')
    n += 1
n = 0
for i in hsbe:
    hsbe[n] = i.replace('\n','')
    n += 1
n = 0
for i in hsor:
    hsor[n] = i.replace('\n','')
    n += 1

new_list = []
for item in hal:
    if item in hbe:
        if item in hsal:
            if item in hsbe:
                if item in hsor:
                    new_list.append('C')
                elif item == '':
                    new_list.append('_')
                else:
                    new_list.append('D')
print (new_list)


name_list = []
for item in hal:
    if item in hbe:
        if item in hsal:
            if item in hsbe:
                if item in hsor:
                    name_list.append(item)
                    print (item)
```