Predicting potential cancer driver genes using hybrid approach

P.A.A Iloshini

2019

Predicting potential cancer driver genes using hybrid approach

A dissertation submitted for the Degree of Master of Science in Computer Science

P.A.A.Iloshini
University of Colombo School of Computing
2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: P.A.A Iloshini
Registration Number: 2016/MCS/042
Index Number: 16440424

Signature: Date: 06/05/2019

This is to certify that this thesis is based on the work of Ms. P.A.A Iloshini
under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:
Supervisor Name: Ms. M W A C R Wijesinghe

Signature: Date:

Abstract

Identifying cancer driver genes remains great significance since it assists in increasing the survival rate by defining cohesive treatments at early stages. Not only single algorithms but also hybrid approaches to identify driver genes do exist, but systematic ways to combine and optimize the existing algorithms on large datasets are few. By identifying the drawbacks of existing cancer driver genes identification methods, this approach formulates an effective hybrid method (Dots Witer) to identify potential cancer driver genes in cancer. The Dots Witer pipeline summarizes somatic mutations, genes involved in tumorigenesis. The input pancancer dataset consists of 2397 small somatic variants of Breast Invasive Carcinoma and 1017 small somatic variants of Lung Adenocarcinoma. Dots Witer pipeline can be applied to genes that are targeted by single nucleotide variants (SNVs) and small insertions and/or deletions (indels). The Dots Witer integrates the tools, DOTS Finder and WITER in order to identify the driver genes efficiently and effectively. This pipeline identifies 656 cancer progression genes out of 1438 genes in Breast carcinoma and 42 cancer progression genes out of 102 in Lung Adenocarcinoma. Since existing tools shows compatibility issues due to technological stack of each tool, the Dots Witer provide a consistent and common platform to execute the given exome/genome sequence dataset. Moreover due to the limitation of the processing power and the storage of the workstation, Dots Witer provides a distributed solution to scatter the ensemble approach. Compare to the existing cancer driver gene detection algorithms, this pipeline gives a higher fraction of predicted driver genes by integrating Fisher's method.

Acknowledgement

I have taken much effort for creating this research project. However, it would not have been possible without the kind support and help of many people. I would like to extend my sincere thanks to all of them.

I am highly indebted to Ms. M.W.A.C.R Wijesinghe, main research supervisor, for her guidance and constant supervision as well as for providing necessary information regarding the research project. In addition I would like to express my sincere gratitude towards Mr.T.Kartheeswaran, co-supervisor for his guidance to make the research successful. Moreover I would like to be thanked for Project Coordinator (Individual Research Project) Dr. L N C De Silva, all academic and non-academic staff members of University of Colombo School of Computing

Also I would like to thank my family members for their kind co-operation and encouragement which help me in completion of this research project.

Thanks and appreciations also go to my colleagues in developing the project and people who have willingly helped me out with their abilities.

Contents

List of Figures

List of Tables

List of Abbreviations

DNA	deoxyribonucleic acid
mRNA	messenger Ribonucleic Acid
miRNA	micro Ribonucleic Acid
SNP	single nucleotide polymorphisms
CNV	copy number variations
CGAP	cancer genome anatomy project
CNA	copy number alteration
SNV	single-nucleotide variant
CCF	fraction of cancer cells
VAF	variant allele frequency
OG	oncogenes
TSG	tumor suppressor genes
OC-SVM	one-class support vector machine
PPI	Protein–protein interaction
MIF	mutational impact function
LoF	loss of function
sSNV	synonymous single-nucleotide variant
nsSNV	non-synonymous single nucleotide variants
BMR	background mutation rate
MAF	mutation annotation format
HTTP	hypertext transfer protocol

CHAPTER 1

Introduction

1.1 Background

Cancer is an involuted genetic disease that is caused by certain changes to genes and it has become a leading genetic disease across the world that result from both inherited and acquired changes in DNA. There are about 100 types of cancer which can affect any part of the human body. According to the statistics of 2016 by World Health Organization (figure1.1),[37] the combination of Trachea, bronchus and lung cancers has become one of top ten causes of deaths globally.

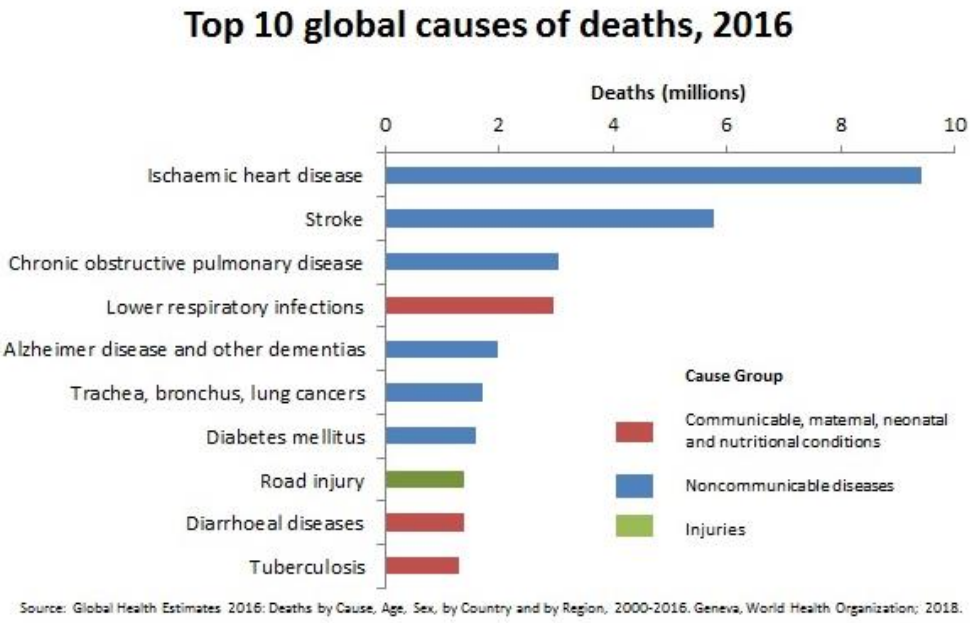

Figure 1.1: Top 10 global causes of deaths, 2016 by WHO

Most human cancers arise from an accumulation of genetic deviations in somatic cells[10]. Tumor genomes of different tissues may contain several somatic mutations. Only a few, critical deviations are caused in tumorigenesis, while the rest are relatively not harmful and make little or no contribution at all[33]. The difference between these two types of deviations in a cancer genome is commonly referred to as 'driver' versus 'passenger' mutations. A driver mutation is referred to as the main cause of tumorigenesis. Genes that bear driver mutations are called cancer driver genes and the remaining genes are identified as passengers. Predicting cancer driver genes remains a major challenge because it assists in increasing the survival rate by defining cohesive treatments at early stages. Predictive algorithms play a

major role as potential methods in filtering driver genes from passenger genes with the help of genomic data generated from Next Generation Sequencing.

1.2 Motivation

Even though there are multiple advanced methods are available to predict driver genes, only few of the methods are effective on large data sets. Moreover the efficiency and accuracy of the individual methods are not promising. According to the evaluation of existing cancer driver gene prediction algorithms by [30], amount of driver genes identified through one algorithm is comparatively low. Moreover the researches [26] have proven that the accuracy of predicting cancer driver genes is increased by using hybrid algorithm rather than using a single algorithm (Table 1.1)

Table 1.1: Percentage of accuracy in cancer driver genes prediction algorithms when uses individually and combined way

	DrGap	**MutSigCV**	**Intogen**
DrGap	4% (individually)	Unrevealed	56%
MutSigCV	Unrevealed	8% (individually)	72%
Intogen	56%	72%	43% (individually)

This study attempts to discover an approach that can predict a complete list of potential cancer driver genes in large datasets. That will also serve as a blueprint for future biological and clinical endeavors in cancer genes prediction.

1.3 Goals and objectives

By addressing the drawbacks of existing cancer driver genes prediction methods, this research aims to formulate an effective method to predict probable driver genes in cancer. In addition the key objectives of the research are as follows:

1. To identify Driver mutations and passenger mutations from the given clinical records
2. To identify computational techniques and algorithms, available for predicting cancer driver genes from the driver mutations

3. To identify relevant molecular profiling platforms that should be used(Platforms will be included exome sequencing, mRNA sequencing, SNP arrays and reverse phase protein arrays) in driver gene prediction.

1.4 Scope

This study focuses two(02) common cancer types, Breast Invasive Carcinoma and Lung Adenocarcinoma. This uses multiple and complementary methods based on mutation rate based algorithms and function prediction based algorithms for a more accurate prioritization of cancer driver candidates.

1.5 Research contribution

Extensive study of currently available methods for cancer driver gene identification will be done through this research. After identifying the drawbacks related to performance, accuracy and reliability of existing cancer driver gene identification methods, an optimized hybrid approach will be proposed. Ultimately this research focuses on identifying a complete list of cancer driver genes.

1.6 Outline of the Thesis

Chapter 02 - Background/Literature Review

This chapter explains a comprehensive summary of previous research related to the current research topic and review of the area being researched.

Chapter 03 - Research Methodology and Design

It includes the process used to collect information, data and other research techniques that have been used to detect the driver genes of Cancer.

Chapter 04 - Evaluation and Results

It shows the ultimate outcome of the research (Cancer driver gene set) and the techniques that used to evaluate the outcome with the result of the other existing algorithms.

Chapter 05 - Conclusion and Future Work

It shows the summary of the research work and the development works that supposed to be implemented in future.

CHAPTER 2

Background/Literature Review

2.1 Background

2.1.1 DNA and Mutations

Deoxyribonucleic acid (DNA) contains the biological directions for life, stored inside living beings. Coiled tightly around proteins called histones, the DNA is packaged within 23 chromosome pairs in cell nuclei. Our DNA comprises of lengthy strings of molecules called nucleotides. These nucleotides are bonded together containing of a group of phosphate, a group of sugar and one of four types of nitrogen bases: adenine (A), thymine (T), guanine (G) and cytosine (C). The most stable form of DNA is structured using hydrogen bonds between base pairs, binding adenine with thymine, and guanine with cytosine. This is how the DNA "ladder" form is built up. Though this form is the most common, DNA also looks as single stranded. The following figure (figure 2.1) [17] illustrated the DNA Structure

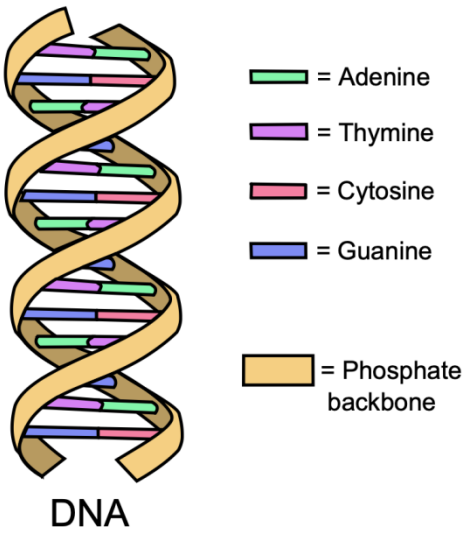

Figure 2.1: DNA Structure

A change in DNA, the genetic sequence, is called a mutation. There are 2 basic types of genetic mutations: Somatic mutation and Germline mutations. Somatic mutation is a modification in DNA that occurs after conception. Somatic mutations can take place in any of the cells of the body apart from the germ cells (sperm and egg) and therefore the alterations are not passed on to children.

A germline mutation happens in a sperm cell or egg cell. It moved within the time period of conception from a parent to a child. While the multicellular diploid eukaryotic organism grows right into an infant, the mutation from the initial sperm or egg cell is copied into each cellular in the body. Since the mutation affects propagative cells, it may be passed from generation to generation.

Most human cancers arise from an accumulation of genetic deviations in somatic cells [10]. Tumor genomes of different tissues may contain several somatic mutations. Only a few, critical deviations are caused in tumorigenesis, while the rest are relatively not harmful and make little or no contribution at all [34]. The difference between these two types of deviations in a cancer genome is commonly referred to as 'driver' versus 'passenger' mutations. A driver mutation is referred as the main cause of tumorigenesis. Driver mutations are carried via genes and they are known as cancer driver genes and the remaining genes are identified as passengers.

2.1.2 Types of somatic mutation

Real patients' data obtained from cancer genomic repositories are been used for analysis. The next key step is to prioritize the list of somatic mutations. The cancer can be happened due to several types of somatic mutations, which comprise single nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations (CNAs) and chromosomal/structural rearrangements. SNVs are sequence variations that relates to a single nucleotide. Synonymous SNVs (sSNVs) are not changed the protein series and non-synonymous single nucleotide variants (nsSNVs) alter the protein series. Indels are the additions or losses of short nucleotide series in a genetic material of a cell. CNAs denote the additions or removals of DNA sectors. Several ways can be identified to categorize copy number alterations, according to the sizes and types of variations. Chromosomal rearrangements are the variations in the organization of a chromosome. It is occured due to inversion (turn around of a chromosomal section), deletion, duplication, translocation (parts of the cromosomes are combined each other) and transpositions (short DNA sections moves to a different position). Recently many algorithms have been developed to discover not only nsSNVs, but also other cancer caused mutated genes as well.

2.1.3 Types of genes linked to cancer

Genes that support to cancer progression can be divided into different categories: Tumor suppressor genes and oncogenes.

Tumor suppressor genes (TSG) are defending genes. Naturally, they limit cell growth by observing the cell division rate into novel cells, fixing mismatched DNA, controlling when a cell dies. When a tumor suppressor gene alters, cells grow uncontrollably and they may ultimately produce a tumor.

Oncogene (OG) is a genetic substantial that carries the ability to prompt cancer. When proto-oncogenes have been changed, it gives the result of altered sequence of deoxyribonucleic acid (DNA). The proto-oncogene helps propagation of normal cells. A range of proto-oncogenes are involved in different key steps of cell growth. An alteration in the proto-oncogene's sequence or in the amount of protein it produces can affect with its normal role in cellular regulation. Uncontrolled cell growth can be resulted in the formation of a cancerous tumor ultimately.

2.2 Data Portals

Yang [38] describes web-based cancer genomics facts repositories, alongside with tools and assets to manipulate and analyze these data. The large genomic database called Catalogue of Somatic Mutations in Cancer (COSMIC) is described here. The database is up to date each 2 months and has consequently far built-in 15,047 complete most cancers genomes from 1,058,292 samples. Data is accessed by means of key words and registered users can download it. The SNP500Cancer database is used to store the data corresponding to sequence of single nucleotide polymorphisms (SNPs) in cancer and different diseases. The cBioPortal for Cancer Genomics is a convenient portal for researchers to explore, visualize, and analyze multidimensional cancer genomics data. cBioPortal methods authentic molecular profiling data from most cancers tissues and cell lines into smaller datasets.

2.3 Tools and techniques for cancer driver gene prediction

Much research has been undertaken so far for the clinical management of cancer. However, due to the large explosion of the quantity of cancer data, there is an increased requirement for the intervention of computational techniques to make sense of the data.

Researchers [32] have introduced SomInaClust, a method that accurately recognizes driver genes and categorizes them into oncogenes or tumour suppressor genes. SomInaClust was proven to perceive candidate driver genes with excessive accuracy. This is proven with the aid of the evaluation of the breast cancer dataset, from which the well-known but hardly ever mutated most cancers genes CDKN1B, KRAS and MEN1 had been recognized. On the opposite hand, the approach was used to disorganize frequently mutated genes like TTN and MUC4 that have been defined as “artefacts” with the aid of others. The consequences obtained with SomInaClust have been as compared with those obtained via the previously posted driver gene prioritization techniques MutSigCV [19],OncodriveFM [11] and OncodriveClust [27] at the identical dataset.

Another study [39] proposed a new approach for figuring out most cancers driver genes, which gives progressed accuracy. The new technique gives the functional impact of mutations on proteins, versions in background mutation rate among tumors and the redundancy of the genetic code. To locate driving genes, each gene is examined for whether or not its mutation rate is drastically higher than the background (or passenger) mutation rate. According [16] there are numerous techniques and algorithms for investigating breast cancer driving force mutation genes. IntOGen [12] that could discover alterations at transcriptomics degree.By using the cancer driver gene identification approach of MESA [15] predicts cancer driver genes based totally on patterns of mutation hotspot.

Researchers [23] offered a new computational method for identifying genomic alterations that arise at low frequencies. Driver–passenger discrimination method is examined based totally on time of the mutation in sizeable simulation studies and applied it to cross-sectional copy number alteration (CNA) information from ovarian cancers, CNA and single-nucleotide variation (SNV) data from breast tumors and SNV information from colorectal cancer. The mutation timing approach will assist identifying from cancer genome records the alterations that manage tumor development. When figuring out the pan-genomic classification of adrenocortical carcinoma, researches [43] qualified the tumor pattern facts on as a minimum one molecular profiling platform. mRNA sequencing, miRNA sequencing and SNP arrays

are considered as platforms.Mutation calling was finished by way of five impartial callers, and a voting mechanism has applied to generate the final mutation set. MutSigCV used to determine extensively mutated genes. GISTIC2.Zero [20] was used to discover recurrent deletion and amplification peaks. Consensus clustering turned into derives miRNA, mRNA and methylation.

If the mutation changes the activity of proteins at some phases of tumor development and if the mutation is functional, it is considered as a driver. A driver gene ought to incorporate as a minimum one driver mutation[8]. Approaches for identifying driver genes can be classified into three categories named mutation rate based approaches, function prediction based approaches, and hybrid approaches. Mutational Significance in Cancer (MuSiC) [7], Mutation Significance (MutSig, MutSigCV, MutSigFN) [19] ActiveDriver [22], ContrastRank [30] are classified into mutation rate based approaches. OncodriveFM, OncodriveCLUST, Oncodrive-CIS [29], Oncodrive-ROLE [24], InVeX [13] are identified as function prediction based approaches and hybrid approaches are the combination of mutation rate based approaches and function prediction based approaches.Researchers [27] have proven that the hybrid techniques permit identifying a comprehensive and reliable list of cancers driver genes. The five strategies including MuSiC-SMG [7], MutSigCV, OncodriveFM, OncodriveCLUST and ActiveDriver were used to predict different cancer driver genes. Lists of 291 cancer driver genes are accommodated and investigated 3,205 tumors from 12 different cancer types. Among those genes, some have no longer formerly recognized as cancer drivers and sixteen have clear bias for a specific tumor type.

Another study [3] has delivered a new database known as DriverDB. It employed eight computational techniques to become aware of driver genes of most cancer types. Four methods, which include MutsigCV, Simon [39], OncodriverFM and ActiveDriver, are based on mutation frequencies. MEMo [5], Dendrix [33], MDPFinder [42] and NetBox [2] have been used as subnetwork based algorithms. Three levels of biological mechanisms are used to (Gene Oncology, Pathways and Protein/Genetic Interaction) to assist researchers to understand the connection among driver genes.

Researches [41] have proposed a different technique to become aware of the cancer driver genes. CDriver, a new approach that integrates signatures of somatic point mutations (SNVs and short indels) at three stages. Population stage , cellular stage and molecular stage are those three stages. Population stage is the proportion of affected individuals (recurrence), cellular stage suggests the fraction of most cancers cells harboring a somatic mutation (CCF),

and molecular stage, is the functional effect of the variant allele. Existing solutions for identifying driver genes rely on the recurrent mutation of genes throughout a huge number of cancers victims [7] and techniques based totally on molecular selection signatures, together with functional impact and mutation clustering. CCF is computed by means of the variant allele frequency (VAF) multiplied by two. Ultimately apply the values received from those three levels to the bayesian inference models and predict the driver genes for 12 types of cancer.

Wei, P.J and the researchers [36] presented a gene length-based network method, named DriverFinder, to identify driver genes by integrating somatic mutations, copy number variations and gene-gene interaction network. Since exceptional mutated drivers are willing to be left out via frequency-based strategies, they've proposed a novel approach to discover driver genes by combining gene-gene interaction network. The gene-gene interplay network is built via combining preceding gene-gene interplay network and Pearson correlated coefficient network. With the aid of analyzing interindividual variant in tumor and normal expression, the outline matrix is constructed.Secondly, in order to rank the mutated genes which are based on the coverage, greedy algorithm is used. In each repetition of the greedy algorithm, the mutated gene of the bipartite graph which pertains to the most outlying expression genes is selected. Until all of the outlying expression genes are investigated via the least mutated genes, repetition is clogged. Genes with the maximum outlying expression are considered as candidate driver genes. Finally, the statistical significance test on null distribution is implemented to these putative cancer driver genes. Moreover researches estimated the performance of DriverFinder with frequency-based method [1] and MUFFINN [4] and acquired a high performance compare to other existing methods.

According to Porta and Godzik, [21] most cancer driver genes can stumble on using the distribution of somatic missense mutations among the protein's functional areas. E-Driver has proposed to perceive driver genes according to the missense mutation. Initially all missense mutations in a protein are examined by the E-Driver. It then detects its protein functional areas.E-Driver iterates through each functional area, calculating the p values of the mutation distribution. Once the p values of all of the areas of all mutated proteins in the cohort are grabbed, the Benjamini-Hochberg false discovery rate set of rules is implemented to correct multiple testing. Those areas with a q value < 0.05 are taken into consideration as positive. In order to evaluate the validity of the technique,reanalyzed the pan-most cancers dataset of the TCGA. The dataset has been formerly analyzed using four distinctive techniques to detect cancer drivers from mutation records (MuSiC, OncodriveFM, OncoCLUST and

ActiveDriver) and the novel approach will be able to locate new potential cancer driver genes as well.

Researchers [35] have proposed a different approach to identify potential novel cancer drivers as those somatic mutations that overlap with known pathogenic mutations in Mendelian diseases. The underlying rationale is if a gene is mutated at significantly greater rate than the background mutation rate, it is more likely to be oncogenic. In this study, it first identified overlapping mutations between pathogenic variants in HGMD [25] and cancer somatic mutations from the COSMIC database. Those overlapping mutations with high recurrence in cancers were subjected to mutual exclusivity analysis with known oncogenes in each tumor type in order to identify novel oncogenic drivers.Researchers [6] proposed a new approach LOTUS to prognosticate cancer driver genes. LOTUS is a machine learning-based approach that estimates a scoring feature to rank candidate genes by means of decreasing probability that they're oncogenes (OG) or tumor suppressor genes (TSG), given set of known OGs and TSGs. The score of a candidate gene is a weighted sum of similarities among the candidate gene and the recognized cancer genes, where the weights are optimized with the aid of a one class support vector machine (OC-SVM) set of rules. Another significant function of LOTUS is to predict driver genes precise to individual cancers sorts. Later, it makes use of a multitask gaining knowledge of method to jointly examine scoring functions for all most all the cancers sorts via sharing information about investigated driver genes in various cancers types.

Researchers [14] recognized a unique technique, MaxMIF to distinguish the driver genes from the passenger genes by means of effective integration of somatic mutation records and molecular interplay records the use of a maximal mutational impact function. Three stages can be identified in MaxMIF. The first one is it computes a mutation rating for each candidate driver gene based on somatic mutation data. Second, it calculates a mutational impact function (MIF) value for each pairs of candidate genes, determining their mutational influences in step with their bond with PPI networks .Two genes should have a strong mutational impact if they each have an excessive mutation rate. Finally, it computes a singular maximal mutational impact function value for every candidate gene by considering about all its acquaintances inside the PPI networks to rank the candidate genes consistent with their maximal mutational impact function values. Tested the MaxMIF on six mutation datasets of Pan-Cancer and 19 datasets of individual cancer sorts from TCGA and earned a higher output with compare to the alternative existing driver genes prediction techniques.

According to the Schroeder M.P and his research team [24] cancer driver genes can be classified according to their role. Researchers presented an automated approach, OncodriveROLE. It is a machine learning-based pipeline that classifies cancer driver genes in step with their role, the use of numerous properties related to the pattern of modifications throughout tumors. Approaches for detecting loss of characteristic (LoF) and Act driver genes appearing throughout tumor samples exist are foremost theories behind on this method. The first approach includes in at once detecting genes that show off recognized alterations patterns corresponding to the tumor suppressors and oncogenes from mutations and CNA statistics. In the second method, first driver genes performing in tumor samples are detected by means of combining the signals of positive selection. Then, in a third step, those drivers are categorized into the two aforementioned lessons exploiting comparable alteration styles as within the first technique.

Predicting driver genes in cancer genomic data is a major role of future biological and clinical endeavors in cancer genes prediction. Several existing algorithms which embed different kind of approaches (eg: Mutation rate based approaches, function prediction based approaches etc) can be identified to predict the cancer driver genes by reviewing the literature related to cancer driver genes prediction. Due to increasing the amount of cancer data and types of cancer, there should be an optimized model to predict the driver genes in cancer.

CHAPTER 3

Research Methodology and Design

In order to identify the cancer driver genes more accurately and efficiently from the given dataset, ensemble approach can be used. Here it describes the overall methodology behind the new approach (Dots Witer) and the unique feature of the Dots Witer algorithm. Figure 3.1 is shown the methodology of the cancer driver gene prediction hybrid approach

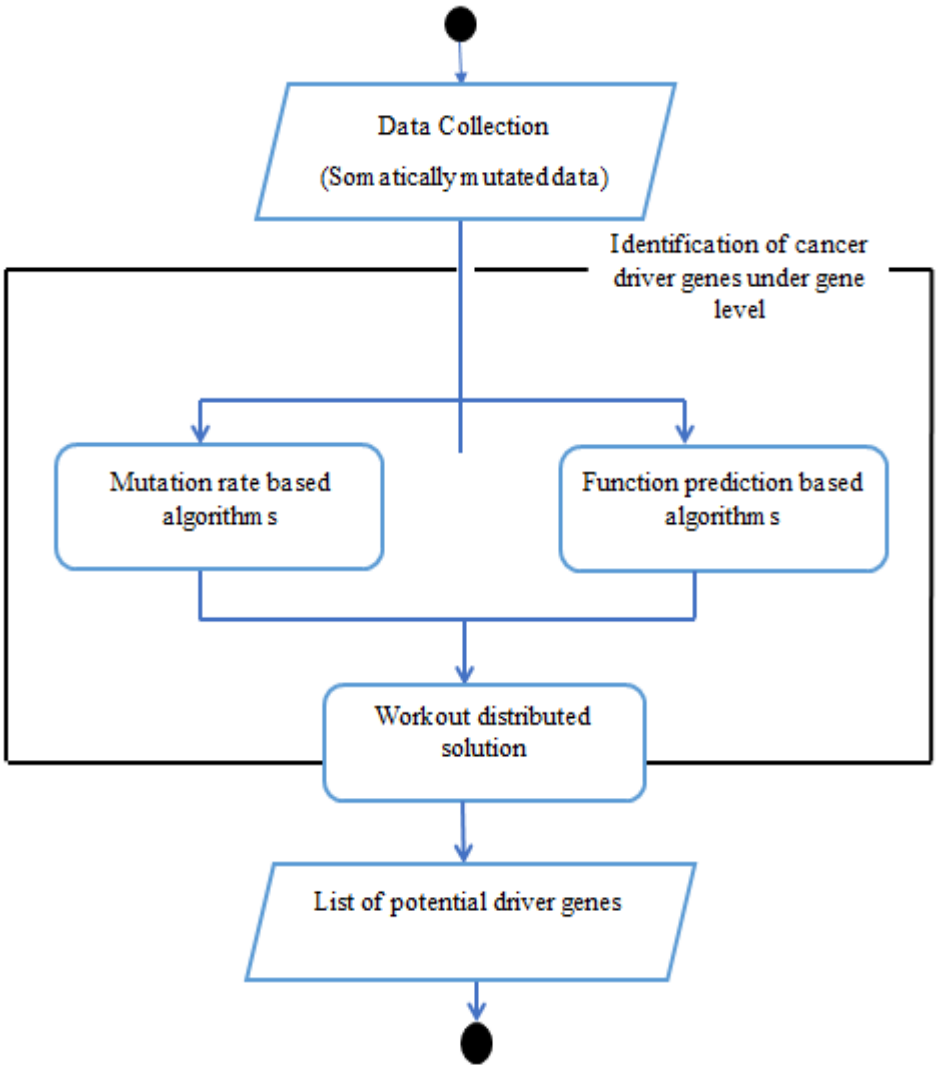

Figure 3.1: Cancer driver gene prediction methodology

Methodology is described under three (03) steps as 3.1 Data Collection (Somatically mutated data), 3.2 Identification of Cancer driver genes under gene level with workout distributed solution and 3.3 List of potential driver genes.

Step 3.1 – Data Collection (Somatically mutated data)

Real patients' data obtained from cancer genomic repositories are been used for this analysis. A key phase is to prioritize the series of somatic mutations out of the obtained cancer genomic data. The cancer is determined by diverse types of somatic mutations, which contain single nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations (CNAs), fusion genes, chromosomal/structural rearrangements.

Real patients' data obtained from Cancer Genome Atlas are been used via cBioPortal software for this analysis. The tool, Dots Witer accepts the input dataset as Mutation Annotation Format (MAF) file for a set of cancer patients that can be categorized by different conditions. Input pancancer dataset consists of 2397 small somatic variants related to Breast Invasive Carcinoma and 1017 data sample consists of Lung Adenocarcinoma.

Step 3.2 – Identification of Cancer driver genes under gene level with workout distributed solution

To be a driver, a mutation should be functional and change the activity of proteins at some stages of tumor growth. A driver gene needs to include as a minimum one driver mutation. In order to identify driver gene, there are three approaches under gene level analysis (driver gene identification) including mutation rate based approaches, function prediction based approaches, and ensemble approaches. Mutation rate based approach deals with the Background Mutation Rate(BMR).Function prediction based approach has a comparable idea but avoids the difficulties of estimating BMR and predicts the functional impact of a particular mutation within the coding pattern of a protein. Ensemble approaches use both mutation rate based approach and function prediction based approach to detect driver gene and it helps to increase accuracy significantly.

Dots Witer is a newly introduced pipeline that used to identify cancer drivers among tumor types and to visualize the results of the analysis. Mainly it builds upon the concept of small somatic variants (SSV) such as single nucleotide variants (SNVs) and small insertions and/or deletions (indels). The Dots Witer pipeline integrates a result set of tumor genomes which is analyzed with various mutation-calling workflows. It currently includes DOTS-Finder [9],a tool that integrates the approach of assessing the type of mutations (for example, missense/truncating/silent) with a protein function prediction based approach (functional step) and a mutation rate based approach (frequentist step) to identify tumor suppressor genes (TSGs) and oncogenes (OGs) genes separately and the tool WITER [18] that works with

synonymous and non-synonymous mutations with a frequentist step. Dots Witer pipeline also considers the somatic variants and integrate functional step and a frequentist method in order to identify the cancer driver genes.

Since the DOTS-Finder and WITER tools arise compatibility issues due to technological stack of each tool, the Dots Witer provide a consistent and common platform to execute the given exome/genome sequence dataset. Dots Witer pipeline gives more accurate result set by integrating the different result set of each tool.

Since the limitation of the processing power and the storage of the particular workstation, this is used a distributed solution to scatter the ensemble approach. The Dots Witer pipeline flow is illustrated in Figure 3.2.

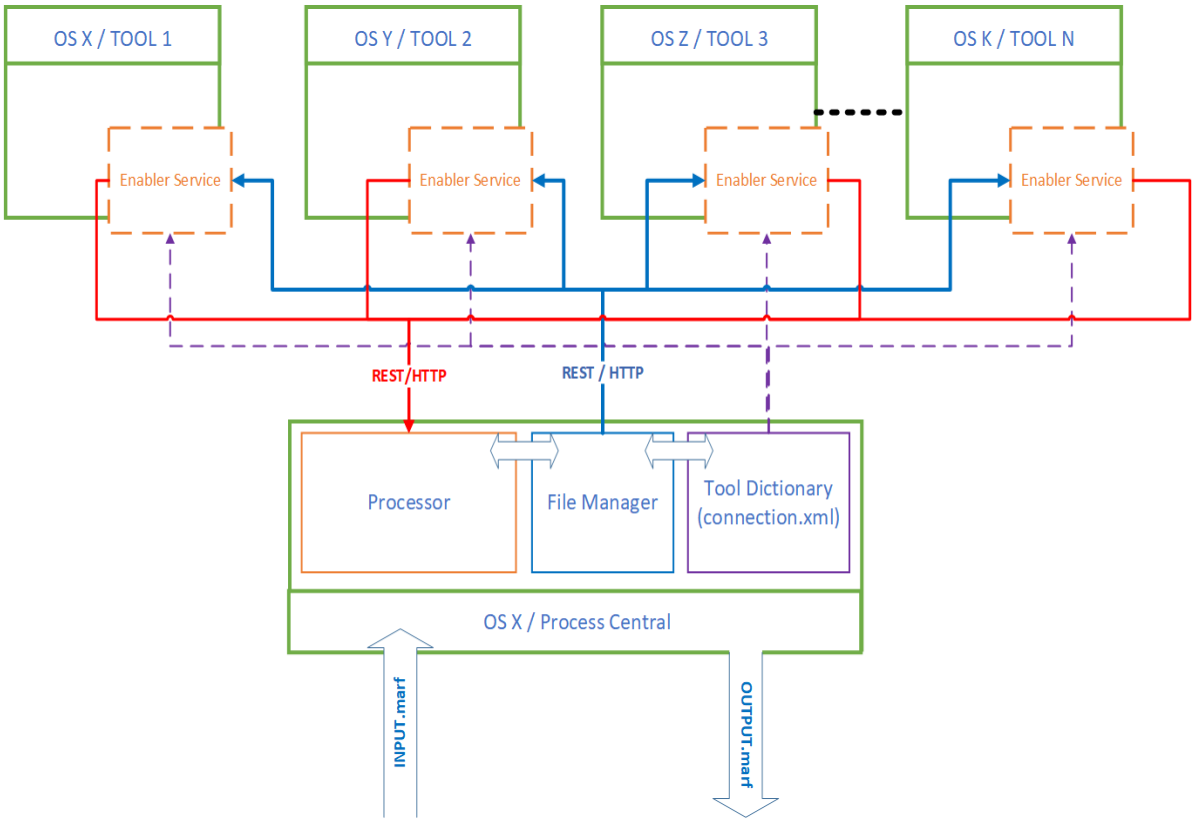

Figure 3.2: Dots Witer Pipeline flow

This application of this pipeline mainly focusses on integrating MultiTech Driver Gene finding tools in to Dots Witer algorithm. Application consists of two main parts.

1. Service Application
2. Utilizer application

3.2.1 Service application

This application is created for managing Driver gene prediction tools by allowing those tools via HTTP/REST. This part of application consist of cherrypy web server, metadata xml and a process engine (Figure 3.3) which is written in Python 2.7.

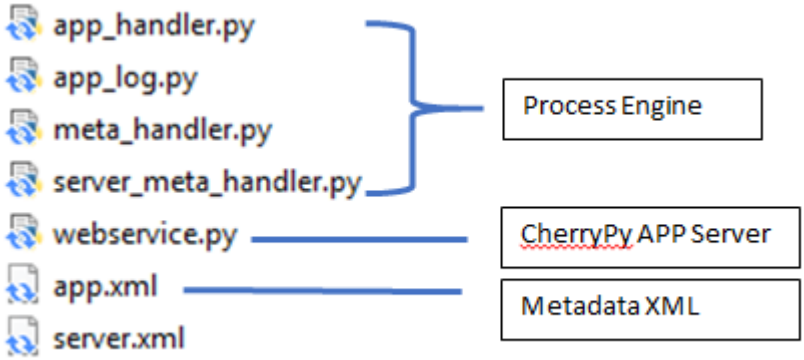

Figure 3.3: Service Application of Dots Witer pipeline

Once it installed in the computer which consist relevant Driver gene, it allows user to access relevant tool via HTTP calls.

Metadata xml can be used to control integration (app.xml) *(see the appendix A.)* Each instance of the service application contains its own copy of metadata xml, which has tagged with a unique key. Each key and operational port and URL will be saved in a common XML file *(see the appendix B.)* call server.xml. Later, utilizer is referring to server xml for identifying possible tools.

As the typical input/output files contain huge amount of cancer data special algorithm has been implemented to handle data transferring through HTTP. This algorithm allows users to pass file as chucks with the size of their desire. As implemented algorithm, checks for special tag which has been prefixed in each incoming text (prefix: data). On availability, process engine pushes relevant data in to a file which will later use as input file. On unavailability, process engine executes the tool with newly created input data and send back the output via HTTP (Hypertext transfer protocol). This mechanism reduces the complexity of Service application by avoiding the usage of FTP (File transfer protocol).

3.2.2 Utilizer application

Utilizer integrate all service applications together in to Dot Witer. This mainly refers server.xml to identify possible tools for driver gene prediction. An algorithm has been

implemented to transfer input data in to each tool and collecting their outputs in to one place. Afterwards, it runs Fisher's theory to have a combined P value as output. Utilizer use both python 2.7 and R as its main programing languages.

This setup allow user to use any Driver gene prediction tool without considering technical complexities.

Step 3.3 – List of potential driver genes

The tool, DOTS Finder accepts the input dataset as Mutation Annotation Format (MAF) (*see the appendix C*) file for a set of patients that can be categorized by various criteria. After analyzing the details of the MAF file, p-value for each genes are calculated and p-value ≤ 0.1 are identified as candidate driver OGs or TSGs. WITER also accepts the input dataset as Mutation Annotation Format (MAF) (*see the appendix D*) with slight difference compare to the input file content of DOTS Finder. It also listed out the existing genes with relevant p-value and the genes which have p- value < 0.1 considered as statistically significant and identified those as driver genes. Dots Witer pipeline execute the input data set through DOTS-Finder and WITER algorithms parallelly and identify likely drivers across the tumor samples. The pipeline combines the P values computed with the aid of either technique for each gene into a single P value using Fisher's method. It produces one integrated P value for each gene. The following algorithm of Dots Witer is illustrated in figure 3.4.

In order to avoid possible dependence between the two P values included in the combination, the Dots Witer considers as significant those with a false discovery rate (FDR) below 0.1

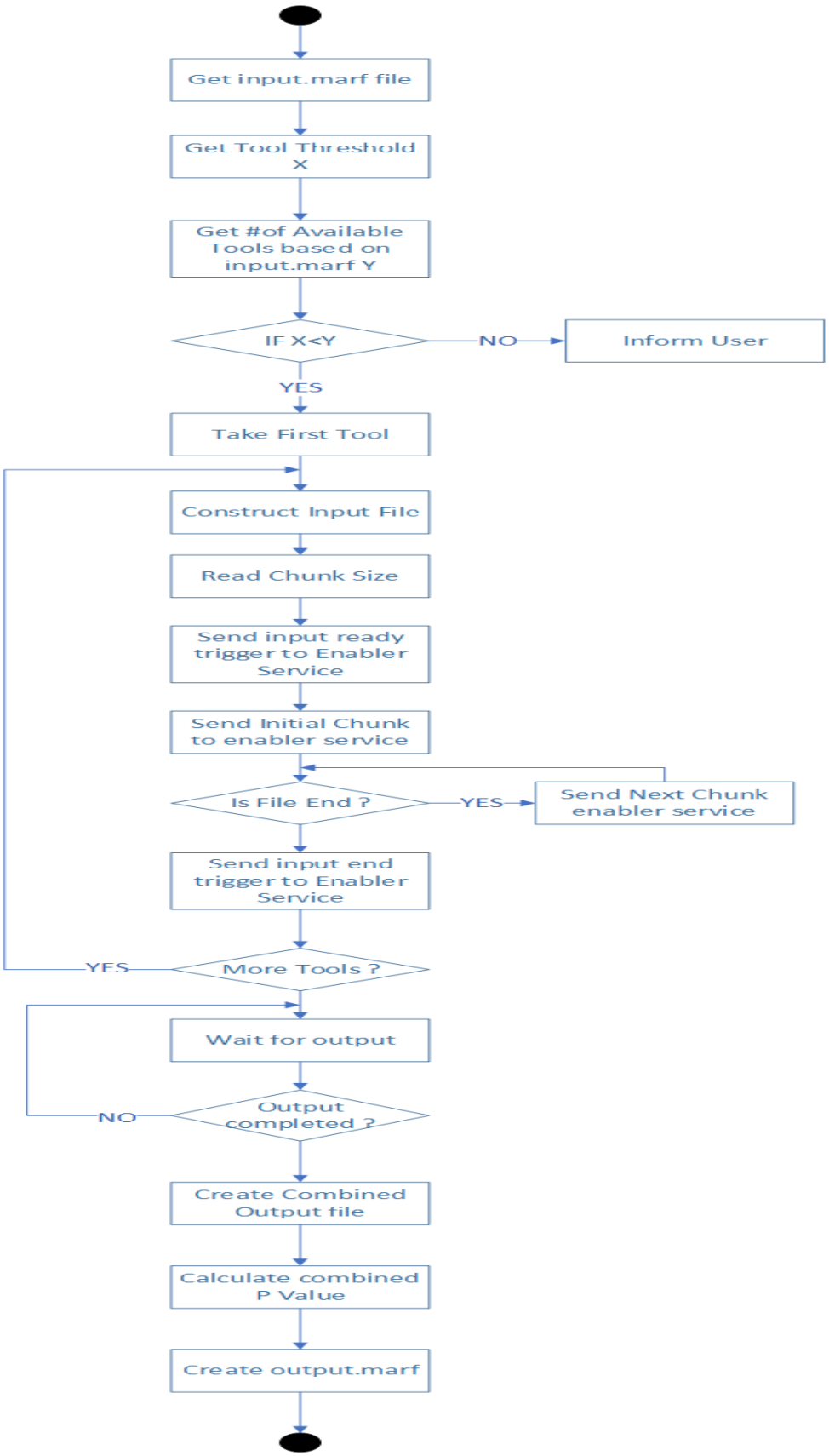

Figure 3.4 : The main algorithm of Dots Witer pipeline

CHAPTER 4

Evaluation and Results

Cancer is a critical disease which caused by somatic mutations on genetic materials of an affected cell in the human body. Identifying the driver genes for the cancer types is a major task of cancer genomics in patient care. Predictive algorithms became the potential method to filter the driver genes from passenger genes with the help of genomic information from Next Generation Sequencing.

4.1 Cancer driver gene prediction tools used in evaluation

Following most common and widely used tools for cancer driver gene prediction are evaluated against known sample set.

1. MutsigCV
2. OncodriveClust
3. OncodriveRole
4. OncodriveCIS
5. OncodriveFml
6. 20/20 Rule
7. Dots Finder
8. WITER

4.2 Tools evaluation procedure

Following tools are been evaluated against following criteria

1. Compatibility of the tool with multiple operating systems
2. Tool dependencies
3. License availability
4. Integration compatibility of algorithm

Depending on the results, most convenient two algorithms named Dots Finder and WITER are selected for the proposed hybrid approach. Results are as follows (Table 5.1)

Table 4.1: Evaluation of the existing cancer driver gene prediction algorithms

Tool	Compatible OS	Primary Language	Required Tools and platforms	Input file formt
MutsigCV	Linux/Unix	Matlab	Matlab and Matlab runtime	.maf, .txt
OncodriveCLUST	Linux/Unix/Windows	Python	Python 2.5 or above	.txt, .mcv
Oncodrive-ROLE	Linux/Unix/Windows	Python	Python 2.5 or above	.txt, .mcv
Oncodrive-CIS	Linux/Unix/Windows	Python	Python 2.5 or above	.txt, .mcv
OncodriveFML	Linux/Unix	Python	Python 2.5 or above	.txt, .mcv
20/20 Rule	Linux/Unix	Python	Python 2.5 or above	.maf, .txt
Dots Finder	Linux/Unix	Python	Python 2.5 or above	.maf
WITER	MS Windows / Mac OS X / Linux	Java	Java version 1.8.0	.maf

During the evaluation MutsigCV has been identified as an incompatible tool for integration, due to the requirement of commercial license of MATLAB platform and limited number of compatible operating systems. The tools which are coming under Oncodrive family (Functional prediction based approach) require different types of input files and it's problematic to convert the original input files to those required file contents. Though 20/20 rule is an accurate tool for identify cancer driver genes, it spends more time even to execute a small dataset. Since both Dots Finder and Witer algorithms use same input formats(.maf), easiness of input file type conversion from real data format, average time execution, compatibility with Operating Systems and capability of smooth installation, those two algorithms are chosen for hybrid approach called Dots Witer.

Dots Witer pipeline is used Python 2.7 as the compiler. Since required pycurl package cannot be installed as mentioned in setup.py script, it was installed manually using pip installer.

4.3 Results related to Dots Witer Pipeline

Dots Witer pipeline used to identify cancer drivers among tumor types and to visualize the results of the analysis of most currently available large data sets of tumor somatic mutations. The pipeline integrates a result set of tumor genomes which is analyzed with various mutation-calling workflows. It currently includes DOTS-Finder,a tool that integrates a protein function approach (functional step) and a frequentist method (frequentist step) to identify tumor suppressor genes (TSGs) and oncogenes (OGs) genes separately and the tool

Witer that works with synonymous and non-synonymous mutations with a frequentist method and ratiometric approach. For the evaluation purpose, the DOT Finder and WITER algorithms are executed individually.

4.3.1 Breast Invasive Carcinoma

DOTS Finder executes 2397 small somatic variant data set related to Breast Invasive Carcinoma and it generated tumor suppressor genes (TSGs) file and oncogenes (OGs) genes file separately with the corresponding p-values. Two of the main types of genes that play a role in cancer are OGs and TSGs. OGs must be activated to cause cancer and when tumor suppressor genes don't work properly, cells can grow out of control, which can lead to cancer. After executing the Breast Invasive Carcinoma sample data set, DOTS Finder identifies TP53 and TNS3 as oncogenes and other 65 unique tumor suppressor genes such as ADAR, AP3B2, BRCA2 etc. The content of output files are as follows (Table 5.2 and Table 5.3)

Table 4.2: TSGs list by DOTS Finder - Breast Invasive Carcinoma

Gene	OncoGene_Entropy_Score	TSG_Score	MissenseType	TruncatingType	p_FI_Total	p_FI_Onco	Global_P_Value
TP53	4.090949666	2.083936323	23	12	1.91E-07	2.15E-05	0
TNS3	1.694666761	4.521465883	4	10	0.5	1	0.014859144
RB1	0	1.908064623	1	3	0.0625	1	0.121511238
ADAR	0	1.397940009	0	1	1	1	0.999992587
AP3B2	0	1.146128036	0	1	1	1	0.999992587
ARHGAP21	0	1.185636577	0	1	1	1	0.999992587
ASB2	0	1.028028724	0	1	1	1	0.999992587
B3GALTL	0	1.185636577	0	1	1	1	0.999992587
BRCA2	0	1.738388821	1	2	0.25	1	0.999992587
C16orf54	0	1.230448921	0	1	1	1	0.999992587
CEACAM1	0	1.006803708	0	1	1	1	0.999992587
CHMP4C	0	1.858837851	0	2	0.5	1	0.999992587
CHRM3	0	1.16879202	0	1	1	1	0.999992587
COL22A1	0	1.204119983	0	1	1	1	0.999992587
CPVL	0	1.007178585	0	1	1	1	0.999992587
DCLRE1B	0	1.139879086	0	1	1	1	0.999992587
DDX31	0	1.021189299	0	1	1	1	0.999992587
EIF4G2	0	1.007178585	0	1	1	1	0.999992587
EML1	0	1.021189299	0	1	1	1	0.999992587
FPGT	0	1.006803708	0	1	1	1	0.999992587

Table 4.3: OGs list by DOTS Finder - Breast Invasive Carcinoma

Gene	OncoGene_Entropy_Score	TSG_Score	MissenseType	TruncatingType	p_FI_Total	p_FI_Onco	Global_P_Value
TP53	4.090949666	2.083936323	23	12	1.91E-07	2.15E-05	0
TNS3	1.694666761	4.521465883	4	10	0.5	1	0.031355657

WITER also executes 2397 small somatic variant data set related to Breast Invasive Carcinoma individually and generates a list of cancer associated genes with p-value. Here It identifies 655 cancer progression genes such as TP53, TNS3, ADAR, BRCA2, ERBB2 etc. The following table (Table 5.4) shows the WITER output.

Table 4.4: WITER output - Breast Invasive Carcinoma

GeneSymbol	ResponseVar	ResponseVarScore	ExplanatoryVar	AvgCodingLen	expr	reptime	hic	constraint_score	Residual	P
TP53	341	341	3	1.448	14.54284996	213	34	1.378924009	27.48757067	1.236E-166
PIK3CA	340	340	6	3.307	12.90393121	613	11	5.42012467	18.96634074	1.61834E-80
GATA3	106	106	2	1.36	12.11686978	675	-2	2.949085091	13.52528856	5.54538E-42
MAP3K1	104	104	3	4.639	12.20904323	412	38	1.52670372	9.843353435	3.66142E-23
CDH1	63	63	3	2.729	13.73657124	240	30	0.809569415	9.675275777	1.92027E-22
CBFB	24	24	1	0.625	13.88191234	125	48	2.473490631	7.021966277	1.09384E-12
AKT1	28	28	1	1.554	13.47551291	247	36	4.027570836	6.797973135	5.30506E-12
PTEN	28	28	0	1.781	12.46719768	300	34	3.714996503	6.602826774	2.01695E-11
MAP2K4	30	30	0	1.293	10.73278145	583	-52	3.491928477	6.399752339	7.78146E-11
TNS3	23	23	3	2.234	13.88701787	147	45	4.86477391	5.315484425	5.3187E-08
RUNX1	24	24	4	1.639	12.01318547	429	45	2.484099017	5.303919692	5.66711E-08
PIK3R1	25	25	1	2.406	10.79955517	619	32	2.416685771	5.137905132	1.38909E-07
TBX3	23	23	1	2.272	10.38495645	548	-30	2.997800057	4.63297388	1.80225E-06
ADAR	19	19	1	2.922	12.76548841	450	41	1.983363117	4.416787702	5.00893E-06
FOXA1	17	17	1	1.429	11.11960136	487	-58	0.017679788	4.389843476	5.67161E-06
CASP8	14	14	1	1.763	13.25642213	386	37	0.817747403	4.234582437	1.14488E-05
ERBB2	19	19	1	3.94	14.06626305	226	38	3.376613895	4.077858006	2.27263E-05
ZFP36L1	12	12	0	1.875	13.71012671	260	32	0.614822974	3.961154031	3.72942E-05
BRCA2	21	21	1	4.115	13.50969574	276	30	7.892268258	3.866418954	5.52226E-05

Dots-Witer pipeline identifies 65 cancer progression genes from DOTS Finder and WITER algorithms such as ADAR, BRCA2, TP53, PIK3CA etc. The combined p-value for those cancer progression genes are calculated using Fisher's approach. 0.05 is considered as level of significance ($p<0.05$) when identifying cancer driver genes. Because R. A. Fisher's argument that one in twenty chance represents an unusual sampling occurrence. The corresponding output file is shown in table 5.5

Table 4.5: Output of Dots Witer pipeline – Breast Invasive Carcinoma

Gene	DOTS Finder p value(X)	WRITER p value(Y)	ln(X)	ln(y)	Sum = ln(X) + ln(y)	Statistic = -2 (sum)	Dots Writer p value (Combined P value)
TP53	0	1.236E-166		-382.0172624	-382.0172624	764.0345247	1.236E-166
PIK3CA		1.61834E-80		-183.7254057	-183.7254057	367.4508115	1.61834E-80
GATA3		5.54538E-42		-94.99560872	-94.99560872	189.9912174	5.54538E-42
MAP3K1		3.66142E-23		-51.66160556	-51.66160556	103.3232111	3.66142E-23
CDH1		1.92027E-22		-50.00440417	-50.00440417	100.0088083	1.92027E-22
CBFB		1.09384E-12		-27.54132966	-27.54132966	55.08265933	1.09384E-12
AKT1		5.30506E-12		-25.96235971	-25.96235971	51.92471943	5.30506E-12
PTEN		2.01695E-11		-24.62684755	-24.62684755	49.2536951	2.01695E-11
MAP2K4		7.78146E-11		-23.27669218	-23.27669218	46.55338436	7.78146E-11
CTCF		5.3187E-08		-16.74945231	-16.74945231	33.49890463	5.3187E-08
RUNX1		5.66711E-08		-16.68600154	-16.68600154	33.37200308	5.66711E-08
PIK3R1		1.38909E-07		-15.78944657	-15.78944657	31.57889314	1.38909E-07
RB1	0.121511238	5.00893E-06	-2.10774853	-12.2042889	-14.31203743	28.62407485	6.08641E-07
TBX3		1.80225E-06		-13.22647459	-13.22647459	26.45294917	1.80225E-06
FOXA1		5.67161E-06		-12.08003683	-12.08003683	24.16007367	5.67161E-06
CASP8		1.14488E-05		-11.37762224	-11.37762224	22.75524448	1.14488E-05
ERBB2		2.27263E-05		-10.69198969	-10.69198969	21.38397939	2.27263E-05
ADAR		3.72942E-05		-10.1966728	-10.1966728	20.39334559	3.72942E-05

The outcome is presented as the format A | B CGC*, where A is the number of genes predicted to be drivers by the pipeline and B is the number of genes in the list A included in the Cancer Gene Census [40]

Number of genes predicted as drivers by DOTS-Finder: 67 | 28 CGC*
Number of genes predicted as drivers by WITER: 655 | 86 CGC*
Number of genes predicted as drivers by Dots-Witer : 65 | 32 CGC*

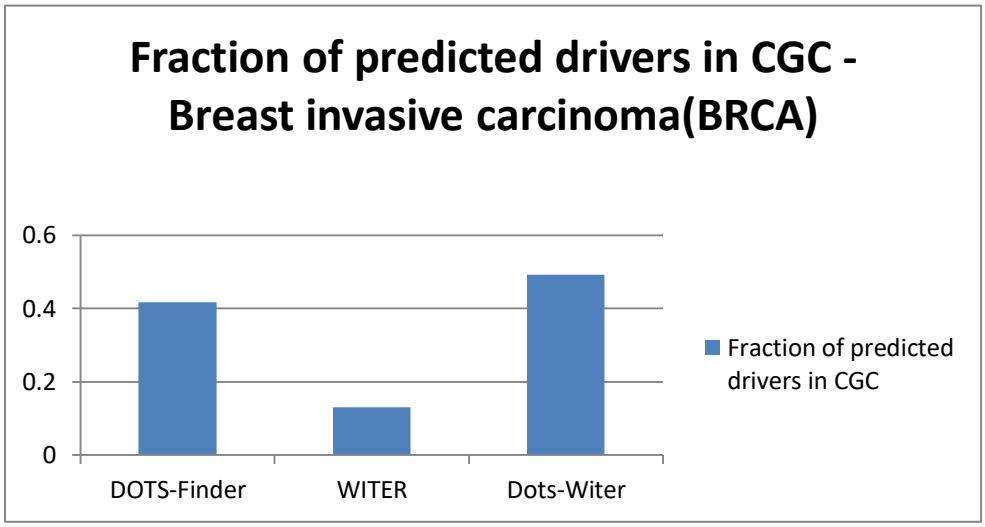

Figure 4.1: Fraction of predicted driver genes in CGC- Breast Invasive Carcinoma

4.3.2 Lung Adenocarcinoma

In order to identify cancer driver genes, DOTS Finder executes 1017 small somatic variant data set related to Lung Adenocarcinoma and it generated tumor suppressor genes (TSGs) file and oncogenes (OGs) genes file separately with the corresponding p-values. After executing the Lung Adenocarcinoma sample data set, DOTS Finder identifies EGFR, KRAS and TP53 as oncogenes and other 05 unique tumor suppressor genes such as STK11, NF1 RB1,LTK and FYN. The content of output files are as follows (Table 5.6 and Table 5.7)

Table 4.6: TSGs list by DOTS Finder -Lung Adenocarcinoma

Gene	OncoGene_Entropy_Score	TSG_Score	MissenseType	TruncatingType	p_FI_Total	p_FI_Onco	Global_P_Value
STK11	-0.517115327	3.409051294	9	26	2.53E-06	0.04002896	0
TP53	2.884648059	1.202938451	43	24	8.11E-07	4.96E-06	0
NF1	2.015813574	1.43964897	6	10	0.018816442	0.578125	1.24E-09
RB1	0	5.198868649	0	7	0.0078125	1	3.91E-05
LTK	0.798729307	1.315641131	3	3	0.15625	0.875	0.00027616
FYN	0	1.174421741	0	2	0.25	1	0.379186503

Table 4.7: OGs list by DOTS Finder -Lung Adenocarcinoma

Gene	OncoGene_Entropy_Score	TSG_Score	MissenseType	TruncatingType	p_FI_Total	p_FI_Onco	Global_P_Value
EGFR	4.150015528	0	34	0	0.993409085	0.422777148	0
KRAS	22.21766777	0	61	0	0.000530302	2.21E-08	0
TP53	2.884648059	1.202938451	43	24	8.11E-07	4.96E-06	0

WITER also executes 1017 small somatic variant data set related to Lung Adenocarcinoma individually and generates a list of cancer associated genes with p-value. Here it identifies 80 cancer progression genes such as EGFR, KRAS STK11, NF1, RB1, EPHB1, DDR1 etc, including oncogene and Tumor suppressor genes together. Table 5.8 illustrates the result set of WITER.

Table 4.8: WITER output- Lung Adenocarcinoma

Chromosome	StartPositionHg19	ReferenceAlternativeAllele	rsID	MostImportantFeatureGene	MostImportantGeneFeature	RefGeneFeatures	GENCODEFeatures	SIFT_score	p
3	89468500	C/A	.	EPHA3	missense	:c.2034C>A:p.D678E:(17Exons	n11:missense;EPHA3:ENSG0	0.039000001	0
20	54945233	G/A	.	AURKA	missense	:c.1193C>T:p.S398L:(10Exons	se;AURKA:ENSG00000087586.	0.135000005	0
3	119582320	C/T	.	GSK3B	missense	:c.1081G>A:p.V361I:(12Exons	SK3B:ENSG00000082701.15_5	0.136999995	0
X	105159737	C/T	.	NRK	missense	c.2365C>T:p.P789S:(29Exons)	T00000243300:c.2365C>T:p.P7	0.214000002	0
2	37516509	G/A	.	PRKD3	missense	3:c.707C>T:p.P236L:(18Exons	00000115825.9_2:ENST000002	0.254000008	0
10	26457765	A/T	.	MYO3A	missense	:c.3236A>T:p.Q1079L:(35Exon	:ENSG00000095777.15_4:ENST	0.254000008	0
11	17156525	G/A	.	PIK3C2A	missense	0:c.809C>T:p.A270V:(32Exons	PIK3C2A:ENSG00000011405.13	0.280999988	0
12	52306967	C/T	.	ACVRL1	missense	exon2:missense;ACVRL1:NM_	000550683:c.188C>T:p.A63V:	0.395999998	0
4	96046157	C/G	.	BMPR1B	missense	exon6:missense;BMPR1B:NM	on6:missense;BMPR1B:ENSG	0.430999994	0
1	32745298	C/T	.	LCK	missense	:c.991C>T:p.P331S:(13Exons)	:c.991C>T:p.P331S:(13Exons)	0.666000009	0
12	18443908	T/C	.	PIK3C2G	missense	70:c.881T>C:p.I294T:(32Exons	79:c.881T>C:p.I294T:(32Exons	0.745999992	0
3	138117376	G/A	.	MRAS	missense	exon4:missense;MRAS:NM_	on3:missense;MRAS:ENSG00	0.995000005	0
4	55139779	C/G	.	PDGFRA	missense	7828:c.1515C>G:p.D505E:(24E	D507166:c.1018-1313C>G:(24E	1	0
22	21097011	C/T	.	PI4KA	missense	62:c.3405G>A:p.M1135I:(54Ex	4KA:ENSG00000241973.10_3:	.	0
12	53876420	C/T	.	MAP3K12	missense	12:missense;MAP3K12:NM_0	A:p.G723R:(14Exons):exon11	0.057	0.001
17	19285705	G/A	.	MAPK7	missense	on5:missense;MAPK7:NM_1	T00000395604:c.2089G>A:p.G	0.059999999	0.001
3	39452296	C/T	.	RPSA	missense	C>T:p.R102C:(7Exons):exon4:	48C>T:(1Exons):upstream;RP	0.067000002	0.001
X	3533911	T/A	.	PRKX	missense	4:c.896A>T:p.H299L:(9Exons)	;PRKX:ENSG00000183943.5_2	0.092	0.001
22	20843446	G/A	.	KLHL22	missense	A;KLHL22:NM_032775:c.53C>	:ENST00000494929:(3Exons):r	0.167999998	0.001

Dots-Witer pipeline identifies 43 cancer progression genes from DOTS Finder and WITER algorithms such as LTK, EGFR , MYO3B, CDK15 etc. The combined p-value for those cancer progression genes are calculated using Fisher's approach. 0.05 is considered as level of significance ($p<0.05$) when identifying cancer driver genes. The result set of Lung Adenocarcinoma is shown in table 5.9.

Table 4.9: Output of Dots Witer pipeline - Lung Adenocarcinoma

A	B	C	D	E	F	G	H
Gene	DOTS Finder p value(X)	WRITER p value(Y)	ln(X)	ln(y)	Sum = ln(X) + ln(y)	Statistic = -2 (sum)	Dots Writer p value (Combined P value)
NF1	1.24067E-09	0.006	-20.50761212	-5.115995801	-25.62360793	51.24721585	7.44404E-12
RB1	3.91024E-05		-10.14932621		-10.14932621	20.29865241	3.91024E-05
LTK	0.00027616		-8.194530199		-8.194530199	16.3890604	0.00027616
STK11	0	0.006		-5.115995801	-5.115995801	10.2319916	0.006
MAP3K12		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
MAPK7		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
RPSA		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
PRKX		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
KLHL22		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
ERBB4		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
SEMA3F		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
EGFR		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
MYO3B		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
EPHA1		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
JUNB		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
MAP3K3		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
KRAS		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
CDK15		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
LMTK2		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
FOXO3		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
KMT2A		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756
RBL1		0.001		-6.907755231	-6.907755231	13.81551046	0.007907756

Number of genes predicted as drivers by DOTS-Finder: 8 | 5 CGC*

Number of genes predicted as drivers by WITER: 80 | 24 CGC*

Number of genes predicted as drivers by Dots-Witer : 43 | 29 CGC*

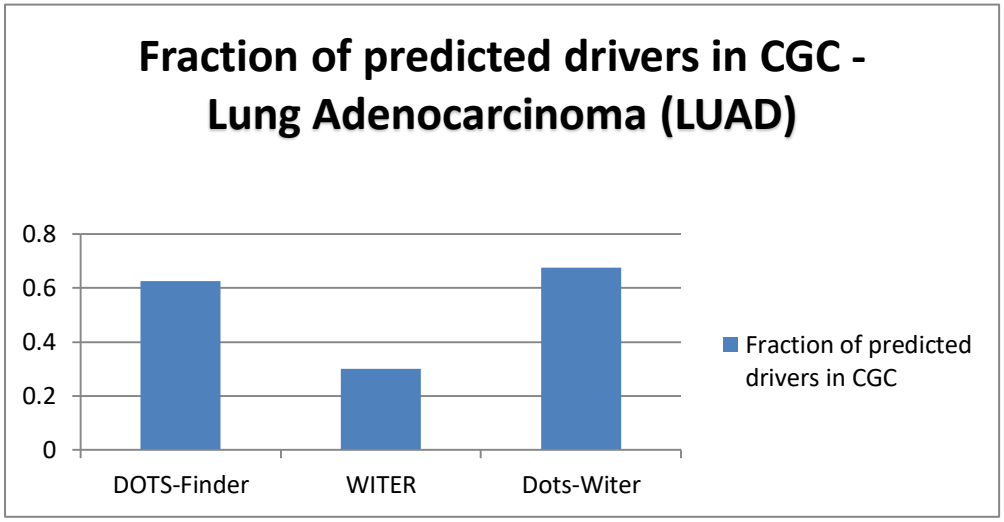

Figure 4.2: Fraction of predicted driver genes in CGC -Lung Adenocarcinoma

CHAPTER 5

Conclusion and Future Work

Dots Witer is a pipeline used to identify cancer drivers among tumor types and to visualize the results of the analysis of most currently available large data sets of tumor somatic mutations. Mainly it builds upon the concept of small somatic variants (SSV) such as single nucleotide variants (SNVs) and small insertions and/or deletions (indels). The Dots Witer pipeline integrates a result set of tumor genomes which is analyzed with various mutation-calling workflows. Dots Witer pipeline integrate functional prediction based step and mutation rate based method in order to identify the cancer driver genes. It is a more reliable pipeline and gives higher fraction of predicted driver genes. Dots Witer pipeline works as a distribution solution and it works as a common platform for other individual driver detection algorithms.

As the diversity of input files has been identified as a bottleneck, new mechanism to manage such complexities needs to be introduced to the Dots Witer algorithm. Since the common main idea of each of these tools is assisting relevant responsible bodies by foreseen potential driver genes, having a common standard for input output files will be an advantage. Dots Witer has a potential to promote that requirement. Therefore, such a standard will be introduced as a future improvement to the algorithm.

Intervention of powerful developer community can make this application grow faster with new ideas and refinements to the algorithm. To encourage such a community, this tool will be documented and published as a free and open source tool in GIT Hub Most of the tools which were evaluated for integration did not have such suitability and lack of documentation of those tools together with long response time creates some additional overhead for tool users. In order to address this, tool builders' community will be maintained attached to the Dots Witer.

Appendices

1. Appendix A - app.xml in service application of Dots Witer pipeline

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!-- This xml contains metadata and description of commad which we uses within the application -->
<!-- sample XML content -->
<py_script key="writer_hjhjtkjdnnxxdfd">
    <script_location>\writer\application\</script_location>
    <script_name>run.sh</script_name>
    <command>java -Xmx6g -jar witer.jar --maf-file |input_file| --out |output_file|
    --excel --db-gene refgene,gencode --gene-feature-in 0,1,2,3,4,5,6,7 --iter-gene</command>
    <options>
        <option name="input_file">examples/hg19_BreastAdenocarcinoma.maf</option>
        <option name="output_file">output/hg19_BAC</option>
        <!-- <option name=""></option>
        <option name=""></option> -->
    </options>
    <url_extention>writer</url_extention>-->
</py_script>
```

2. Appendix B - server.xml in service application of Dots Witer pipeline

```xml
<?xml version="1.0" encoding="UTF-8"?>
<installations>
    <!-- Sample installation -->
    <installation key="writer_hjhjtkjdnnxxdfd">
        <type>Writer</type>
        <nick_name>Writer APP</nick_name>
        <description>This is for test</description>
        <app_host>127.0.0.1</app_host>
        <app_port>3000</app_port>
    </installation>
    <installation key="DOTS_oaskjhkjhsdahkjsa">
        <type>Dots_Finder</type>
        <nick_name>Dots Finder</nick_name>
        <description>This is for test</description>
        <app_host>127.0.0.1</app_host>
        <app_port>3001</app_port>
    </installation>
</installations>
```

3. Appendix C - input file format of DOTS Finder approach

Breast_Invasive_Carcinoma.marf

Hugo_Symbol	Entrez_Gene_Id	NCBI_Build	Chromosome	Start_Position	End_Position	Variant_Classification	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2	dbSNP_RS	Tumor_Sample_Barcode	protein_change
ABCC9	0	GRCh37	12	22078995	22078995	Missense_Mutation	C	C	T		SA018	p.R96Q
C8B	0	GRCh37	1	57397545	57397545	Missense_Mutation	C	C	A		SA018	p.R520L
CD300E	0	GRCh37	17	72608853	72608853	Missense_Mutation	A	A	G		SA018	p.L186P
CDC42BPA	0	GRCh37	1	227182571	227182571	Missense_Mutation	C	C	T		SA018	p.G1580S
CPXM2	0	GRCh37	10	125528080	125528080	Missense_Mutation	A	A	C		SA018	p.S421A
DST	0	GRCh37	6	56481248	56481248	Intron	A	A	C		SA018	p.I2339M
DUSP3	0	GRCh37	17	41852113	41852113	Missense_Mutation	C	C	T		SA018	p.D107N
GRIK2	0	GRCh37	6	102337572	102337572	Missense_Mutation	G	G	A		SA018	p.D528N
HEATR5B	0	GRCh37	2	37268436	37268436	Splice_Site	C	C	A		SA018	p.?
HIST1H1T	0	GRCh37	6	26108162	26108162	Missense_Mutation	C	C	T		SA018	p.V54M
HNRNPH1	0	GRCh37	5	179047973	179047973	Missense_Mutation	T	T	C		SA018	p.D106G
HRNR	0	GRCh37	1	152191472	152191472	Missense_Mutation	G	G	A		SA018	p.S878L
KIAA0556	0	GRCh37	16	27789901	27789901	Missense_Mutation	G	G	A		SA018	p.R1603H
KLHL13	0	GRCh37	X	117033283	117033283	Missense_Mutation	G	G	A		SA018	p.A519V
KRT34	0	GRCh37	17	39535257	39535257	Missense_Mutation	C	C	G		SA018	p.E392Q
PPP1R42	0	GRCh37	8	67929884	67929885	Frame_Shift_Del	TA	TA	-		SA018	
MYO3A	0	GRCh37	10	26385582	26385582	Missense_Mutation	T	T	C		SA018	p.C583R
NBEAL1	0	GRCh37	2	204013799	204013799	Missense_Mutation	A	A	G		SA018	p.Q478R
NCL	0	GRCh37	2	232326413	232326413	Missense_Mutation	C	C	T		SA018	p.D151N
NCL	0	GRCh37	2	232326437	232326437	Missense_Mutation	C	C	T		SA018	p.E143K
NOS1	0	GRCh37	12	117691478	117691478	Missense_Mutation	G	G	T		SA018	p.D871E
NPR2	0	GRCh37	9	35805626	35805626	Missense_Mutation	G	G	A		SA018	p.R669Q
PCDH10	0	GRCh37	4	134073673	134073673	Missense_Mutation	C	C	G		SA018	p.S793C

4. Appendix D - input file format of WITER approach

hg19_BreastAdenocarcinoma_new.marf

Gene	Tumor_Sample_UUID	Tumor_Type	Chromosome	Start_Position	End_Position	Variant_Classification	Reference_Allele	Tumor_Allele1	Tumor_Allele2	Protein_Change
A1BG	TCGA-A8-A06P	Breast Adenocarcinoma	chr19	58864307	58864307	Missense_Mutation	C	A	C	p.E109D
A1BG	TCGA-A8-A06P	Breast Adenocarcinoma	chr19	58864307	58864307	Missense_Mutation	C	A	C	p.E109D
A1BG	TCGA-E9-A1NH	Breast Adenocarcinoma	chr19	58864366	58864366	Missense_Mutation	G	A	G	p.R90C
A1BG	TCGA-E9-A22B	Breast Adenocarcinoma	chr19	58862784	58862784	Missense_Mutation	C	T	C	p.A295T
A1CF	BR-MEX-015	Breast Adenocarcinoma	chr10	52566490	52566490	Splice_Site	C	T	C	p.*603*
A1CF	TCGA-BH-A0HP	Breast Adenocarcinoma	chr10	52595854	52595854	Missense_Mutation	G	A	G	p.A203V
A1CF	TCGA-BH-A18P	Breast Adenocarcinoma	chr10	52595937	52595937	Silent	G	A	G	p.I175I
A2M	TCGA-A2-A0EY	Breast Adenocarcinoma	chr12	9246090	9246090	Silent	C	T	C	p.E737E
A2M	TCGA-A8-A08G	Breast Adenocarcinoma	chr12	9251298	9251298	Nonsense_Mutation	G	A	G	p.R586*
A2M	TCGA-B6-A0IC	Breast Adenocarcinoma	chr12	9220358	9220358	Silent	-	T	-	p.K167fs*
A2M	TCGA-B6-A0IQ	Breast Adenocarcinoma	chr12	9256962	9256962	Missense_Mutation	G	T	G	p.P380Q
A2M	TCGA-BH-A18H	Breast Adenocarcinoma	chr12	9230409	9230409	Missense_Mutation	T	C	T	p.Y1055C
A2M	TCGA-BH-A1FN	Breast Adenocarcinoma	chr12	9254262	9254262	Nonsense_Mutation	G	T	G	p.Y425*
A2M	TCGA-C8-A138	Breast Adenocarcinoma	chr12	9242995	9242995	Silent	G	A	G	p.N851N
A2M	TCGA-D8-A1JK	Breast Adenocarcinoma	chr12	9221429	9221429	Nonsense_Mutation	G	A	G	p.Q1425*
A2M	TCGA-E9-A1ND	Breast Adenocarcinoma	chr12	9242989	9242989	Silent	C	T	C	p.R853R
A2M	PD5936a	Breast Adenocarcinoma	chr12	9248202	9248202	Missense_Mutation	T	C	T	p.Y649C
A2M	PD5934a	Breast Adenocarcinoma	chr12	9243024	9243024	Nonsense_Mutation	C	A	C	p.E842*
A2ML1	TCGA-A1-A0SO	Breast Adenocarcinoma	chr12	8994108	8994108	Missense_Mutation	G	C	G	p.W408C
A2ML1	TCGA-A8-A08P	Breast Adenocarcinoma	chr12	8995779	8995779	Missense_Mutation	G	A	G	p.R433H
A2ML1	TCGA-AN-A0FT	Breast Adenocarcinoma	chr12	9000231	9000231	Silent	G	A	G	p.A590A
A2ML1	TCGA-AR-A251	Breast Adenocarcinoma	chr12	8988187	8988187	Missense_Mutation	G	A	G	p.E190K
A2ML1	TCGA-B6-A0WZ	Breast Adenocarcinoma	chr12	9020914	9020914	Missense_Mutation	C	T	C	p.P1341L
A2ML1	TCGA-BH-A0AV	Breast Adenocarcinoma	chr12	8998791	8998791	Silent	C	T	C	p.F552F
A2ML1	TCGA-BH-A0HP	Breast Adenocarcinoma	chr12	9001389	9001389	Missense_Mutation	C	G	C	p.S636C

References

[1] Bashashati, A. et al., 2012. DriverNet : uncovering the impact of somatic driver mutations on transcriptional networks in cancer.

[2] Cerami, E. et al., 2010. Automated network analysis identifies core pathways in glioblastoma. PLoS ONE.

[3] Cheng, W.C. et al., 2014. DriverDB: An exome sequencing database for cancer driver gene identification. Nucleic Acids Research.

[4] Cho, A. et al., 2016. MUFFINN : cancer gene discovery via network analysis of somatic mutation data. Genome Biology, pp.1–16. Available at: http://dx.doi.org/10.1186/s13059-016-0989-x.

[5] Ciriello, G. et al., 2012. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research.

[6] Collier, O., Stoven, V. & Vert, J.-P., 2018. LOTUS: a Single- and Multitask Machine Learning Algorithm for the Prediction of Cancer Driver Genes. , pp.1–35. Available at: http://dx.doi.org/10.1101/398537.

[7] Dees, N.D. et al., 2012. MuSiC: Identifying mutational significance in cancer genomes. Genome Research.

[8] Djotsa Nono, A.B., Chen, K. & Liu, X., 2016. Computational Prediction of Genetic Drivers in Cancer. eLS, (February), pp.1–16.

[9] G. E. M. Melloni et al., "DOTS-Finder: A comprehensive tool for assessing driver genes in cancer genomes," Genome Med., vol. 6, no. 6, pp. 1–13, 2014.

[10] Garraway, L.A. & Lander, E.S., 2013. Review Lessons from the Cancer Genome. Cell, 153(1), pp.17–37. Available at: http://dx.doi.org/10.1016/j.cell.2013.03.002.

[11] Gonzalez-Perez, A. & Lopez-Bigas, N., 2012. Functional impact bias reveals cancer drivers. Nucleic Acids Research, 40(21).

[12] Gonzalez-Perez, A. et al., 2013. IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods.

[13] Hodis, E. et al., 2012. A landscape of driver mutations in melanoma. Cell.

[14] Hou, Y. et al., 2018. MaxMIF: A New Method for Identifying Cancer Driver Genes through Effective Data Integration. Advanced Science.

[15] Jia, P. et al., 2014. MSEA : detection and quantification of mutation hotspots through mutation set enrichment analysis. , pp.1–16.

[16] Kumar Rajendran, B. & Deng, C.-X., 2017. Characterization of potential driver mutations involved in human breast cancer by computational approaches. Oncotarget.

[17] Kumar, S. (2019). A Proper Approach on DNA Based Computer. [online] Pubs.sciepub.com. Available at: http://pubs.sciepub.com/ajn/3/1/1/ [Accessed 3 Jun. 2019]. - dna

[18] L. Jiang et al., Genomic characterization of additional cancer-driver genes using a weighted iterative regression accurately modelling background mutation rate. 2018.

[19] Lawrence, M.S. et al., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499(7457), pp.214–218.

[20] Mermel, C.H. et al., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology.

[21] Porta-Pardo, E. & Godzik, A., 2014. E-Driver: A novel method to identify protein regions driving cancer. Bioinformatics.

[22] Reimand, J., Wagih, O. & Bader, G.D., 2013. The mutational landscape of phosphorylation signaling in cancer. Scientific Reports.

[23] Sakoparnig, T., Fried, P. & Beerenwinkel, N., 2015. Identification of Constrained Cancer Driver Genes Based on Mutation Timing. PLoS Computational Biology.

[24] Schroeder, M.P. et al., 2014. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. In Bioinformatics.

[25] Stenson, P.D. et al., 2014. The Human Gene Mutation Database : building a comprehensive mutation repository for clinical and molecular genetics , diagnostic testing and personalized genomic medicine. , pp.1–9.

[26] Strekerud & Kristoffer., 2014. A Comparison of Computational Tools for Prediction of Cancer Driver Genes. Institutt for informatikk [2799], pp.1–63.

[27] Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N., 2013. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics.

[28] Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., et al., 2013. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific Reports, 3.

[29] Tamborero, D., Lopez-Bigas, N. & Gonzalez-Perez, A., 2013. Oncodrive-CIS: A Method to Reveal Likely Driver Genes Based on the Impact of Their Copy Number Changes on Expression. PLoS ONE.

[30] Tian, R., Basu, M.K. & Capriotti, E., 2014. ContrastRank: A new method for ranking putative cancer driver genes and classification of tumor samples. In Bioinformatics.

[31] Tokheim, C.J. et al., 2016. Evaluating the evaluation of cancer driver genes. , 113(50).

[32] Van den Eynden, J. et al., 2015. SomInaClust: Detection of cancer genes based on somatic mutation patterns of inactivation and clustering. BMC Bioinformatics, 16(1).

[33] Vandin, F., Upfal, E. & Raphael, B.J., 2012. De novo discovery of mutated driver pathways in cancer. Genome Research

[34] Vogelstein, B. et al., 2013. NIH Public Access. , 339(6127), pp.1546–1558.

[35] Wang, C. et al., 2017. IDENTIFY CANCER DRIVER GENES THROUGH SHARED MENDELIAN DISEASE. , pp.473–484.

[36] Wei, P.J. et al., 2017. Driver finder: A gene length-based network method to identify cancer driver genes. Complexity.

[37] Who.int. ,2019. The top 10 causes of death. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [Accessed 28 May 2019]

[38] Yang, Y. et al., 2015. Databases and Web Tools for Cancer Genomics Study. Genomics, Proteomics & Bioinformatics, 13(1), pp.46–50. Available at: http://dx.doi.org/10.1016/j.gpb.2015.01.005.

[39] Youn, A. & Simon, R., 2011. Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics, 27(2), pp.175–181.

[40] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer, pp. 1.

[41] Zapata, L. et al., 2017. Bayesian inference of cancer driver genes using signatures of positive selection. bioRxiv.

[42] Zhao, J. et al., 2012. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics, 28(22), pp.2940–2947.

[43] Zheng, S. et al., 2016. Comprehensive pan-genomic characterization of adrenocortical carcinoma. Cancer Cell, 29(5), pp.723–736.