Discovering Subclones and Their Driver Genes in Tumors Sequenced at Standard Depths

by

Navid Ahmadinejad

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2019 by the
Graduate Supervisory Committee:

Li Liu, Chair
Carlo Maley
Valentin Dinu

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Understanding intratumor heterogeneity and their driver genes is critical to designing personalized treatments and improving clinical outcomes of cancers. Such investigations require accurate delineation of the subclonal composition of a tumor, which to date can only be reliably inferred from deep-sequencing data (>300x depth). The resulting algorithm from the work presented here, incorporates an adaptive error model into statistical decomposition of mixed populations, which corrects the mean-variance dependency of sequencing data at the subclonal level and enables accurate subclonal discovery in tumors sequenced at standard depths (30-50x). Tested on extensive computer simulations and real-world data, this new method, named model-based adaptive grouping of subclones (MAGOS), consistently outperforms existing methods on minimum sequencing depth, decomposition accuracy and computation efficiency. MAGOS supports subclone analysis using single nucleotide variants and copy number variants from one or more samples of an individual tumor. GUST algorithm, on the other hand is a novel method in detecting the cancer type specific driver genes. Combination of MAGOS and GUST results can provide insights into cancer progression. Applications of MAGOS and GUST to whole-exome sequencing data of 33 different cancer types' samples discovered a significant association between subclonal diversity and their drivers and patient overall survival.

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor Dr. Li Liu for all her support and encouragement through the years. This PhD would not have been possible without her guidance. I would like to thank her for pushing me to be a more rounded researcher. She has helped me and supported me every step of these process and I could not have asked for a better adviser. Additionally, I would like to thank Dr. Carlo Maley for all of his great help and knowledge and for being generous with his time and advice. Also, I would like to extend my gratitude to the Dr.Valentine Dinu, for his understanding, his continued support and his help.

Special thanks to ASU's Biomedical Informatics department for their incredible support and guidance, in particular, Lauren Madjidi, Maria Hanlin, Dr. Dongwen Wang and Dr.George Runger. I will always be grateful.

In addition, I would like to thank my lab mate Pramod Chandrashekar and my friends and colleagues Arjun Magge, Verah Nyarige, Meredith Abrams, Lu Zheng, among other students in my department for all their support.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

PUBLISHED WORK

The majority of the work presented in this dissertation has not been published yet. The only section that is in press and will be published soon, is the GUST algorithm. GUST in introduced and discussed in details in chapter 3. This work is an algorithm designed with collaboration between Drs. Li Liu, Carlo Maley, Sudhir Kumar, Pramod Chandrashekar and myself, Navid Ahmadinejad. I am the second author of this paper. The study was designed by Li Liu, Carlo Maley and Sudhir Kumar. I contributed by applying the GUST algorithm to the TCGA data and analyzing the results as discussed in the publication. Pramod Chandrashekar developed the online database. All co-authors have granted their permissions for this work to be included in my dissertation.

CHAPTER 1

INTRODUCTION

Cancer is a major public health problem worldwide and after heart disease is the second most common cause of death in the United States. In 2019, it is projected that 1,762,450 new cancer cases and 606,880 cancer deaths will happen in the United States.(Siegel, Miller, and Jemal 2019)

In spite of the extraordinary amount of effort and money spent on cancer research and treatment, it remains difficult to achieve a cure in most cases. Nevertheless, all that research has been contributing to the greater understanding of cancer biology, how cancer evolves and its complexities. Unraveling the secrets of how cancer evolves and progresses is crucial to designing treatment approaches. Like most problems, to find a solution, it is vital to understand all aspects of the problem. In cancer, also, in order to decide on a treatment approach, it is essential to understand the nature of the disease, how it progresses, how it evolves, and how it will react to environmental stresses.

1.1 Clonal Evolution

The clonal evolution model proposed by Peter Nowell (Nowell 1976), established an evolution model for cancer that a was a foundation for further studies in understanding cancer (Greaves and Maley 2012). Nowell suggested that cancer is an evolutionary process that is driven by cells acquiring somatic mutations. The accumulation of these mutations in a cell affects the fitness of the cell. These mutations, depending on which gene they occur in, can give the cell a selective advantage so that the cell proliferates faster or survives better than its peers. When this proliferation gets out of control, it can result in the death of the host. Not all of these mutations observed in cancers give the cell

a selective advantage. Some may increase the mutation rate in the cell (Sottoriva and Graham 2015), and some may be passenger mutations and have no effect on the fitness of the cell (Sottoriva and Graham 2015).

When a cell is hit with a mutation, all its descendants will carry that mutation. As mentioned, only some mutations will give the cell a selective advantage, and most mutations are passenger mutations, meaning they do not affect the reproduction or survival of the cell. The mutations that give the cell a growth or survival advantage are called driver mutations (Stratton, Campbell, and Futreal 2009; Christopher Greenman et al. 2007). These mutations happen in cancer genes. The majority of mutations in these cells are passenger mutations in comparison to the number of driver mutations that provide the cell with selective advantage (Christopher Greenman et al. 2007). The mutations in cancer cells can be caused by an error in the DNA repair process or from mutagenic exposures, such as ultraviolet light, cigarette carcinogens, and from DNA damaging therapies such as chemotherapy (Stratton 2011). Although, mutations can also occur just from the normal error rate of DNA polymerases when cells copy their genomes during the cell cycle. There are many cancer cases in which causes of the mutations have remained unexplained (Stephens et al. 2005).

Cancers also evolve through clonal expansions. This means that when a driver mutation hits a cell, the cell proliferates faster than its peers. All the descendants of that mutant cell, called a clone, will share the same genomic profile; hence, they will also carry the same mutation. In addition, new cells in that clone may accumulate their own new mutations that they maybe passenger or driver mutations. Since the majority of mutations are passengers, it is likely that the cells will accumulate more passenger

mutations. After some cell generations, the resulting group of cells will share a similar genomic profile with passenger and driver mutations. This group of cells, which were driven from the initial driver mutation, form a clone that is genetically and may be functionally different from the other cells in the neoplasm. In a similar fashion, other driver mutations hitting different cells can create different subclones from their background genomic profile (Nowell 1976; Greaves and Maley 2012). This will cause intratumor heterogeneity, which is discussed in more detail in a later section. It is important to note that the term "clone," as mentioned in Nowell (Nowell 1976), implies a population of cells descendant from a single cell of origin.

1.2 Intratumor Heterogeneity

Cancer development is an evolutionary process that typically originates from a single cell or clone and grows into a diverse population of cells via incessant somatic mutations and natural selection (Nowell 1976; Greaves and Maley 2012). In this dynamic process, different cell populations (i.e., subclones) emerge, expand and diminish over time and space, leading to a heterogeneous malignancy with multifarious clinical presentations. This dynamic often results in two types of heterogeneity; intertumor heterogeneity and intratumor heterogeneity (J. Liu, Dang, and Wang 2018).

Intertumor heterogeneity refers to the differences between tumors, even if they are the same clinical type. There are differences in the genomic profiles of the tumor samples from different patients. Intertumor heterogeneity in part is because of the altered genotype and phenotype of the patients from etiological and environmental factors (Davidson et al. 2016; Llovet, Burroughs, and Bruix 2003), and also the fact that acquiring mutations in cells is not deterministic and mutations may hit different genes in

different patients. Intratumor heterogeneity, on the other hand, refers to the within tumor heterogeneity, meaning the tumor is composed of various populations of cells with different genomic alterations, which is the result of the mutations and clonal expansion within a cancer. The driver mutations generate these subpopulations, and the passenger mutations are carried in the clonal expansions.

Understanding intratumor heterogeneity is critical to understanding the dynamics of the disease. Most cancer therapies eliminate some clonal populations but not all (Aktipis et al. 2011). Better understanding the composition of the tumor will have significant clinical implications in designing treatments. Understanding this dynamic system provides valuable knowledge to facilitate early diagnosis, effective treatment, and outcome monitoring of cancers (Aktipis et al. 2011; Andor et al. 2014; Nik-Zainal et al. 2012; Gerlinger and Swanton 2010; Fisher, Pusztai, and Swanton 2013; Ma, Ennis, and Aparicio 2012). Before explaining the therapeutic effects and importance of this clonal structure and intratumor heterogeneity, it is essential to talk about the different types of mutations, the genes they hit, and their effects on cancer progression. In the next section, we explain different types of mutation, what sort of hallmarks are essential to the development of cancer and the role of mutations in the cell acquiring these hallmarks.

1.3 Cancer Hallmarks and Driver Mutations

Gene alterations in the body are called germline mutations if they are in the sperm or egg that fused to generate the first cell of the body (the zygote) and are inherited in the DNA of every cell in the body. Germline mutations are passed on from parents to their children, and they exist in all of the cells, whereas the somatic mutations are alterations that can happen in the non-germ cells (the somatic cells) during the lifetime of an

organism. They are only passed to the decedent of that specific cell and not the other cells in the body. Tumorigenesis, which means the process of formation of a tumor, is believed to be a consequence of somatic mutation accumulation, including base substitutions, insertions and deletions of bases, inversions, translocations, and changes in the copy number of DNA segments, in the genome of cancer cells (Christopher Greenman et al. 2007; Tao et al. 2011; Harrington 2016). Germline mutations can predispose a person to heritable cancer, either by providing one of the sufficient mutations to cause a cancer, and thereby reducing the number of mutations that must be accumulated to complete the process of tumorigenesis, or by increasing the mutation rate in somatic cells (Croce 2008; Stratton 2011).

As previously discussed, mutations can be divided into passenger and driver mutations (Segal et al. 2008). Driver mutations are the ones that confer a fitness advantage in some way and contribute directly to tumorigenesis. The contribution to tumorigenesis can also be explained in the concept of hallmarks of cancer (Fortunato et al. 2017).

Cancer hallmarks are biological capabilities that cancer cells acquire during tumorigenesis. These capabilities are vital in the progression of cancer, and with all of them at work, cancer cells can survive, and tumors can grow. There are eight important hallmarks proposed by Hanahan et al., including sustaining proliferative signaling, evading growth suppressors, ignoring apoptosis signals, enabling replicative immortality, inducing angiogenesis, altering cell metabolism, evading the immune system, and activation of metastasis and invasion (Hanahan and Weinberg 2011). These complementary traits enable tumorigenesis. Understanding how these capabilities are

activated is key to understanding the cancer progression. These hallmarks are activated and induced by somatic alterations in different genes. Some of the hallmarks are the result of a gene being activated and gain function, and some are the result of a gene losing its function and being deactivated. The genes that are activated by acquiring mutations are called proto-oncogenes, and the genes that lose their functions are called tumor suppressor genes. Once a proto-oncogene acquires a mutation that increases the chance of cancer, the mutated gene is called an oncogene. However, the distinction between proto-oncogenes and oncogenes is somewhat esoteric, and in practice is often ignored. Here, for simplicity, I will just refer to them as oncogenes. Oncogenes are genes that are cancer causing when activated, typically causing increased proliferation of the clone. Their activation causes the oncogenesis, which is the process in which a healthy cell transforms into a cancer cell, which leads to the cell to proliferate in an uncontrolled manner (Krump and You 2018). The tumor suppressor genes, on the other hand, are genes that increase the chance of developing cancer when they are deactivated. These genes are responsible for sensing and responding to DNA damage, repairing DNA, sending apoptosis signals or preventing the cell from proliferating uncontrollably (Morris and Chan 2015).

Since the driver genes are the main reason for oncogenesis, being able to identify these genes can reveal the mechanism of the disease progression. It becomes very important to be able to identify driver mutations and driver genes. These identifications need to be supported by independent observation of these events occurring more frequently in multiple neoplasms than the expected in the normal sample (Greaves and Maley 2012; Maley et al. 2004; Llovet, Burroughs, and Bruix 2003). The type of mutation (silent, missense or nonsense)

is also important in classifying it as driver or passenger (Bignell et al. 2010; Youn and Simon 2011). In the next section, we discuss the clinical implications of the tumor heterogeneity, the current approaches and their shortcomings.

1.4 Therapeutic Importance and Current Approaches to Study Tumor Heterogeneity

Understanding the dynamic system behind the progression of cancer provides valuable knowledge to facilitate early diagnosis, effective treatment, and outcome monitoring of cancers (Aktipis et al. 2011; Andor et al. 2014; Nik-Zainal et al. 2012; Gerlinger and Swanton 2010; Fisher, Pusztai, and Swanton 2013; Ma, Ennis, and Aparicio 2012). The heterogeneity in the tumors results in a non-uniform and genetically diverse population of cells both across the tumor at a single time point (spatial heterogeneity) or across different time points (temporal heterogeneity).

The heterogeneity in the tumor is the main cause of resistance to therapy in cancers; therefore, controlling the disease and designing the best therapy approach depends on understanding this subclonal structure (Dagogo-Jack and Shaw 2018; Hiley et al. 2014).

Because mutations in driver genes cause cancer, an important approach to therapy has been the development of drugs that target those driver genes (Higgins and Baselga 2011; Morris and Chan 2015; Sawyers 2004). This is one reason that identifying the oncogenes and tumor suppressor genes is important. Detecting the driver genes can also help with early detection of the disease, identifying risk factors, and generating animal models of cancer. Ideally, we would like to therapeutically reactivate tumor suppressor

genes and deactivate oncogenes, though at present we only have methods to deactivate oncogenes, and even in that case, we only have been successful at inhibiting oncogenes that are tyrosine kinases.

The clonal structure of a tumor can provide valuable information. It is essential to identify the driver genes and then study whether they are clonal mutations or subclonal mutations. Because in most cancers, it is the clonal mutations that are driving the progression of cancer, but it is the rare mutations that are subclonal that cause relapse and often resistance to different therapies (Hinohara and Polyak 2019; Schmitt, Loeb, and Salk 2016).

Whole-exome or whole-genome sequencing is a common approach to studying intratumor heterogeneity (Schwartz and Schäffer 2017; Egan et al. 2012; Landau et al. 2013). By tracking relative abundances of genomic variants in a collection of cancerous cells, scientists aim to quantify the genetic diversity of a tumor and to reconstruct the phylogeny of subclones. The underlying principle is that cells of similar genetic compositions belong to the same clone. While single-cell sequencing is on the rise, bulk sequencing remains the dominant technology used to interrogate an amalgam of heterogeneous cells collectively. Researchers rely on *in silico* analysis to de-convolute the mixed populations of clones within the sample. Single-cell sequencing is a promising technology to examine the genetic compositions of individual cells, but it currently suffers from low and uneven genome coverage, low accuracy of variant calls, and prohibitive cost which all limit its usage for subclonal investigations (Gawad, Koh, and Quake 2016). The majority of current studies perform bulk sequencing of tens of thousands of heterogeneous cells from a tumor followed by clonal deconvolution. In

some studies, generic clustering tools (Fraley and Raftery 1999) and packages have been used to analyze the bulk sequencing data (Ding et al. 2012), but the problem with this approach is that some manual curation is required. In addition, these methods are not tuned to use sequencing information such as sequencing depth and Variant allele frequency (VAF). Several methods have been developed to address this weakness, such as SciClone (C. A. Miller et al. 2014), PyClone (Roth et al. 2014), and Expands (Andor et al. 2014). Despite algorithmic differences, these methods make a common assumption that variant allele frequency (VAFs, i.e., the fraction of reads containing the mutant allele among total reads) is indicative of the relative abundance of cells carrying these mutations. Thus, subpopulation discovery is translated into grouping mutations with similar VAFs in a tumor. If multiple samples of a tumor are sequenced, mutations defining a subpopulation are expected to have concordant changes of VAFs across these samples. The discovered subclones can be further analyzed to infer phylogenetic relationships, test for selection, and other analysis regarding tumor heterogeneity and clonal evolution of the disease. SciClone and PyClone are able to analyze multiple samples but EXPANDS can only do single sample analysis.

1.5 Limitations of Existing Methods

Next-gen sequencing technologies have opened a vast range of genomic analysis capabilities to researchers (Shendure and Ji 2008; Meldrum, Doyle, and Tothill 2011). Whole-genome and whole-exome sequencing are becoming the standard approach to study the clonal evolution in cancer as well. These sequencing technologies produce the counts of reference and alternate reads on each sequenced position. The total number of sequenced DNA segments that cover a position is referred to as the coverage or

sequencing depth. The variant allele frequency (VAF) is defined as the number of alternate reads divided by the sequencing depth, which is the sum of the reference reads and the alternate reads. The VAF is the feature used in the existing algorithms for detecting the clonal structure in tumors. Because of the reliance on VAFs, existing methods require the sequencing depth of a tumor sample to be at least 100x coverage, or suggest a minimum number of mutations (Miller et al. 2014; Andor et al. 2014). Some of them were even developed to work on only deeply sequenced data (depth>1000x) (Roth et al. 2014). They are not suited for samples sequenced at a standard 30-50x depth (Sims et al. 2014). This precludes their use on the overwhelming majority of samples sequenced to date, including those generated by collaborative consortia, such as the TCGA Pan-Cancer Atlas (average depth = 68x). Unfortunately, clinical samples are often sequenced at a much lower depth, since 30x-50x is considered sufficient to call germline mutations with high confidence (Griffith et al. 2015).

An independent evaluation of SciClone showed that consistent clonal identifications can only be produced when the sequencing coverage is 300x or higher (Griffith et al. 2015). Given the constraint on the sequencing depth of existing methods, valuable information embedded in tens of thousands of tumor genomes is unexploited. New methods to characterize subclonal structure accurately from sequencing data at a shallow-to-medium coverage are urgently needed.

At the algorithmic level, the constraints of sequencing depths in current methods are at least partially due to unexplained variances in bulk sequencing data. A key assumption taken by these methods is the correspondence between variant allele frequencies (VAFs, i.e., the fraction of reads containing a specific mutant allele among

total reads) and cellular prevalence (i.e., the fraction of cells carrying this particular mutant among all cells). Subclone discovery is then translated into a task of clustering similar VAFs (Miller et al. 2014). However, VAFs are also influenced by technical factors, such as sequencing depth. Due to randomness in sequencing procedures, the same subclone may give rise to a dispersed cluster of VAFs when sequenced at a low depth, but a tight cluster of VAFs when sequenced at a high depth (Griffith et al. 2015). The spread of VAFs also correlates with cellular prevalence. VAFs of variants in a common subclone are expected to scatter more broadly than those in a rare subclone (Griffith et al. 2015). Without considering these confounders, subclones reported by current methods are inevitably adulterated, especially when the sequencing depth is not high enough to create strong contrasts between subclones with similar cellular prevalence.

To detangle these technical variabilities from biological variabilities, we have developed a new method, named model-based adaptive grouping of subclones (MAGOS) that explicitly models the impact of sequencing depth and cellular prevalence on the variance of VAFs in subclone decomposition. Cellular prevalence refers to the proportion of cancer cells in the sample that share the same group of mutations. Through extensive tests using computer simulations and real-world data, we show that MAGOS can accurately delineate subclonal structures of tumor samples sequenced at depths as low as 30x. MAGOS is also the fastest program when compared to SciClone and PyClone, showing an acceleration of 3-20 fold. We implemented MAGOS as an R package that is freely available at GitHub (https://github.com/liliulab/magos).

In the next chapter, we discuss the approach we took and explain the principles and the methods that MAGOS uses in contrast to the commonly known algorithms, then we present the results of the simulations we ran for the evaluation of MAGOS and the results of executing MAGOS on the 33 cancer projects publicly available on TCGA. In the third chapter, we introduce GUST, which is an algorithm that identifies cancer specific driver genes. We evaluate its performance and present the results of executing GUST on the TCGA data. In the fourth chapter, combine the results of MAGOS with GUST to find the (sub)clonal distribution of driver genes and report if there are any features with prognosis value.

CHAPTER 2

MODEL-BASED ADAPTIVE GROUPING OF SUBCLONES: MAGOS

The purpose of MAGOS, as well as the other existing algorithms, is to group variants that emerge and evolve together into a cluster based on similarities of VAFs. Each cluster thus corresponds to a subclonal expansion. In this context, we use the VAF cluster and subclone interchangeably.

Each method has a different approach to finding the mutational clones. Some of them were designed to work on a specific type of data, but they have been used in various studies before. PyClone, SciClone, and EXPANDS are the algorithms that are used in the most recent studies. Each has its limitations and assumptions. In here, we will discuss their approached and limitations.

2.1 Existing Methods and Their Limitations

2.1.1 PyClone

PyClone is a Bayesian clustering method for grouping sets of deeply sequenced mutations into clonal and subclonal clusters. PyClone requires the mutations to be deeply sequenced at coverage > 1000x. It can also be used to cluster sequenced mutations across multiple samples. The primary assumption in the PyClone approach is that no site can be mutated more than once, and the mutations do not disappear (they do not revert to wildtype). The allelic prevalence of mutations has information from several sources in it. It contains normal contamination, meaning the proportion of contaminating normal cells in the sample (purity), the proportion of cells carrying the mutation, and technical noise. The main advantage of the analysis of multiple sequencing samples (temporal or spatial)

is the possibility of sets of mutations that their VAFs shift together. This is a great way to identify and separate the clones and mutations that are proliferating in different rates.

PyClone is a hierarchical Bayesian statistical model. The inputs of the model are a set of deeply sequenced mutations from one or more samples from the same cancer. PyClone output is the posterior densities for the model parameters and the clonal structure of the mutations. PyClone uses Beta-Binomial densities over Binomial models because it has more efficiency in modeling the datasets with more variance in allelic prevalence measurements. It also utilizes Bayesian non-parametric clustering to discover groupings of mutations and the number of groups concurrently. This eliminates the need for fixing the number of clusters *a priori* and allows cellular prevalence estimates to reflect uncertainty in this parameter.

The PyClone package provides tools for performing Dirichlet Process clustering of mutations. The model outputs a posterior density for each mutation's VAF and a matrix containing the probability that any two mutations occur in the same cluster. Then the model merges two mutations to the same cluster if they have similar VAF in the sample(s). To obtain a flat clustering of the mutations from the matrix of pairwise probabilities, we construct a dendrogram and find the cut point that optimizes the MPEAR criterion (Fritsch and Ickstadt 2009). More detailed information on the PyClone method is available in Roth et al (Roth et al. 2014).

The traditional view of clonal heterogeneity and tumor phylogenies is to cluster cells by their mutational composition — however, PyClone clusters mutations that appear at similar cellular frequencies (similar VAFs). In simple cases, where there is only one clone, the clustering derived may be correct; however, if multiple subclones exist at

similar cellular frequencies, the model will cluster the associated mutations together and fails to identify distinct clusters. The mutations get separated, and we start to see a contrast between the VAFs of mutations from different clusters as we increase the sequencing depth. Sequencing multiple samples also helps to identify the correct clusters in these situations because it is expected that the mutations from the same cluster shift together and have similar VAF across all samples.

PyClone is specifically designed for the problem of inferring the clonal structure of single nucleotide mutations in deeply sequenced tumor samples. However, the model is quite generic in the sense that it only assumes the sequenced sample is a heterogeneous mixture of cells. Because of this fact, PyClone is not recommended to be used on low coverage data. The results from PyClone on low coverage input are not consistent. Another limitation of PyClone, specifically on single samples with low coverage sequencing, is the fact that it tends to over-estimate the number of clusters; it finds additional clusters with only a single or a few mutations assigned to them. Because of these limitations, using PyClone on samples similar to TCGA data is not recommended.

2.1.2 SciClone

SciClone is a method for identifying the subclones across one or many samples and estimate the number of existing subclones. It mainly focuses on the copy-number neutral(diploid) and loss of heterozygosity free regions of the genome, which leads to higher confidence variant allele frequencies and inference on the clonality. However, it limits the number of mutations that can be analyzed since many cancers are highly aneuploid. The regions of copy number alterations and the loss of heterozygosity are provided as an input from whole-genome sequencing or whole-exome sequencing. The

single nucleotide variants should be sequenced at a sufficient depth from whole-exome sequencing or deeply sequenced from targeted sequencing. The approach that SciClone takes can be used on any input that can be described as a frequency.

In SciClone, the clustering of the VAFs is performed using a variational Bayesian mixture model (VBMM) (Bishop 2006). Several approaches, before SciClone, have taken Dirichlet process models to perform clustering(Roth et al. 2014; Shah et al. 2012) since they could automatically infer the number of clusters. VBMMs, in addition to automatically inferring the number of clusters and providing probabilistic interpretations of the clusters, can be scaled to high dimensions and have an efficiency advantage over MCMC techniques that are used in the algorithms before SciClone.

One fundamental principle in identifying the subclones is mixture modeling. The mutations from different subclones are mixed in the bulk sequencing data, and the observed data is a representation of that mixture. SciClone's fundamental objective is to identify the subclones from the mixture. The VAF is defined as the number of alternate reads over the total number of reads. The distribution of the VAF from each cluster is assumed to have a beta distribution. The probability of each mutation coming from each cluster is then determined by using the probability distribution function of each cluster.

The SciClone model is initialized by performing k-means clustering on the data with ten initial clusters. This adds randomness to the clustering, and local optima may be selected. This may lead to inconsistent results in some samples. SciClone also removes any clusters having less than the largest of three variants or 0.5% of "N," the total number of mutations. So it is biased against finding clones with few mutations. If clusters are removed, the algorithm is again executed until it reaches convergence.

Based on the analysis on Miller et al.(Miller et al. 2014), SciClone can be useful on samples with as few as 29 SNVs, and in more complex cases, identifying the subclones may require two hundred or more variants and that subclones be separated by VAFs of ~7% or more.

The accuracy of the detected subclones is increased with the number of mutations in the data. In exome-sequencing, many passenger mutations are likely to be missed; thus, whole-genome sequencing is more likely to capture the full spectrum of mutations and produce a high confidence clonal structure. Also, SciClone is also able to incorporate additional temporal or spatial samples into its analysis to provide a more accurate view of the subclonal structure of the tumor. The suggested minimum coverage of SciClone is 100x, which limits the use of this approach on low coverage data. This is shown in the results chapter.

2.1.3 EXPANDS

EXPANDS is another algorithm designed to identify these subpopulations in a tumor using allele frequencies (VAFs) from whole-exome or whole-genome sequencing and copy number alterations. EXPANDS estimates the tumor purity as well.
The EXPANDS model (Andor et al. 2014), similar to previous approaches, makes a few assumptions. First, it assumes that each locus is mutated only once. Meaning cells are mutated in that locus, or they are normal. It means the mutation occurs at a single time point, and all the cells carry that mutation are descendants of the original cell. The same mutation is not likely to occur in multiple distinct cells independently. Another assumption is that multiple passenger mutations accumulate in a cell before a driver

mutation hits the cell and causes clonal expansion. SciClone and PyClone also have this assumption.

The EXPANDS model uses these assumptions in four steps. Cell frequency estimation, clustering, filtering, and assignment of mutations to clusters. In the first step, using the single nucleotide variations and the copy number information, EXPAND combines both information inputs and estimates the fraction of cells carrying each mutation. Thus, it does not require mutations to be in diploid regions of the genome. If all mutations are from diploid regions, then the measure directly translates into VAF. If the mutation occurs in an aneuploid region of the genome, the VAF must be scaled by the number of copies in which it occurs before the clone's frequency can be estimated. Since the equations used at this step are not deterministic, the results for the cell frequency estimation are in a probability form as a function of f in the range [0,1] ($p(f)$= probability that the mutation has VAF=f). In the clustering step, mutations are clustered based on their cell frequency probability distributions. The clustering is performed in two steps. The aim of the clustering step in EXPANDS is to merge mutations that have common peaks in their p(f) distribution. The mutations are grouped together by hierarchical clustering of the p(f)s using Kullback-Leibler divergence as a distance measure (Joyce 2011). Then, the cell frequency at each cluster maxima is considered as the size of the cluster. In the second step, each cluster is extended by mutations with similar distributions in an interval around the cluster maxima.

In the filtering step, the clusters that all the mutations assigned to them all have the same cell frequency as each other. EXPANDS calculates the tumor purity is the size of the

largest subclone. Then each locus is assigned to the cluster that its cell frequency estimation is closest to.

EXPANDS is sensitive to the mutation count as well as the coverage of the mutations. The prediction accuracy of EXPANDS increases with the number of mutations. EXPANDS also tends to over-estimate the number of clusters and in lower coverage data fails to detect the correct number of clusters. This makes EXPANDS a poor candidate for analyzing samples with a small number of mutations and low coverage.

2.2 MAGOS Algorithm

We designed MAGOS in a way so that it has better performance on low coverage data, as well as high coverage. In the rest of this chapter, we discuss how MAGOS works. In the Results chapter, we discuss evaluations and comparisons we ran on real and simulated data.

MAGOS discovers subclones from bulk sequencing data based on the concept of mixture distributions. It represents VAFs of somatic mutations that occur and evolve together as samples drawn from a beta distribution. When multiple subclones are present, the observed VAFs are a mixture of samples from multiple beta distributions. Therefore, the identification of subclones is equivalent to decomposing mixed beta distributions. MAGOS supports subclone analysis of a tumor containing single nucleotide variants (SNVs) and copy number variants (CNVs) obtained from one or more samples. To illustrate the algorithms of MAGOS, we start with simple scenarios and gradually introduce complexities into the data model.

2.2.1 Algorithm for SNVs from Single Sample

The simplest scenario involves a single tumor sample that has only SNVs, no CNVs and no contamination of normal cells. The goal is to group the mutations that belong to the same (sub)clone and estimates its frequency. For this, we need to find clusters of SNVs with similar VAFs. Given a variant i, we denote the total number of reads aligned to this position as its sequencing depth e_i, and denote the fraction of reads containing the mutant allele among all reads as its VAF $v_i \in (0, 1)$. For a set of m variants belonging to the same subclone, we model their VAFs as random samples from a beta distribution $Beta(\alpha, \beta)$ and require the two shape parameters (α and β) to satisfy

$$\begin{cases} \alpha + \beta = \bar{e} \\ \frac{\alpha}{\alpha+\beta} = \bar{v} \end{cases} \quad (1)$$

where $\bar{e}$ is the mean sequencing depth and $\bar{v}$ is the mean VAF of these variants. This configuration has the desired property that the variance of the beta distribution is positively correlated to the mean VAF and negatively correlated to the mean sequencing depth. The proof of this concept is in Appendix A. Through this setup, we link the variance of VAFs to cellular prevalence and sequencing depth.

When multiple subclones are present, the observed VAFs are a mixture of samples from multiple beta distributions, each defined by a set of shape parameters. Therefore, identification of subclones is equivalent to decomposing mixed beta distributions (Fig. 2.1). We solve this problem with a two-phase algorithm that performs agglomerative hierarchical clustering and adaptive partitioning.

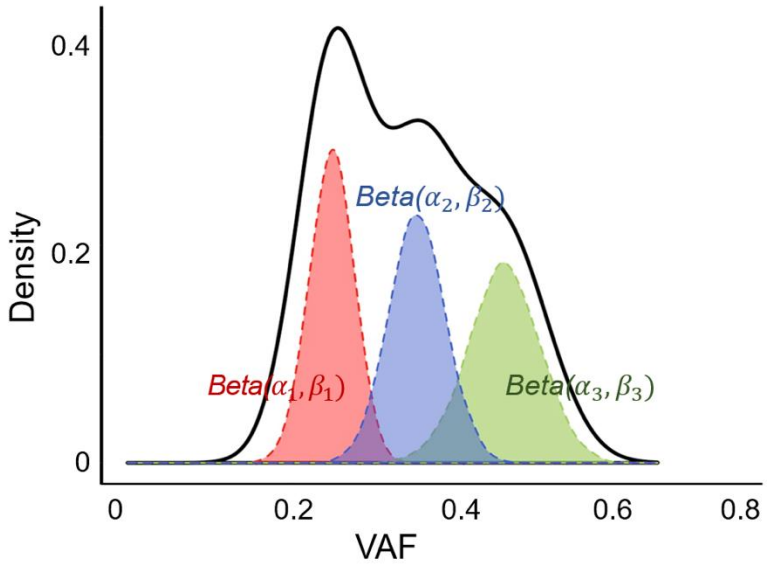

Figure 2.1. Beta mixture distribution. The observed distribution (black curve) of VAFs is a combination of multiple hidden groups of VAFs, each forming a beta distribution (shaded curves) defined by different parameters.

In the first phase, we organize variants into a hierarchical tree structure by progressively grouping variants with similar VAFs into a cluster. Starting with leaf nodes each consisting of an individual variant, we iteratively merge a pair of nodes with the shortest distance (defined below) among all pairs to create a new cluster till all variants are merged into one root cluster (Fig. 2). Given two nodes (i.e., clusters), C_1 and C_2 consisting of m_1 and m_2 variants, respectively, we define their distance d as a weighted sum of negative log likelihood that VAFs of all variants in C_1 and C_2 are drawn from the same beta distribution,

$$d(C_1,\ C_2) = w\ \sum_{i\in\{C_1,\ C_2\}} -\log(\mathrm{P}(v_i\ ;\ Beta(\alpha,\beta))) \qquad (2)$$

where α and β are calculated by solving equation (1) and the weight $w = 1/(m_1 + m_2) \cdot var(v) \cdot range(v)$. Because the distance is down-weighted by the variance and range of VAFs, given two pairs of clusters with similar values of log

likelihood, MAGOS will choose the pair with a smaller variance and a narrower range to merge at an earlier step.

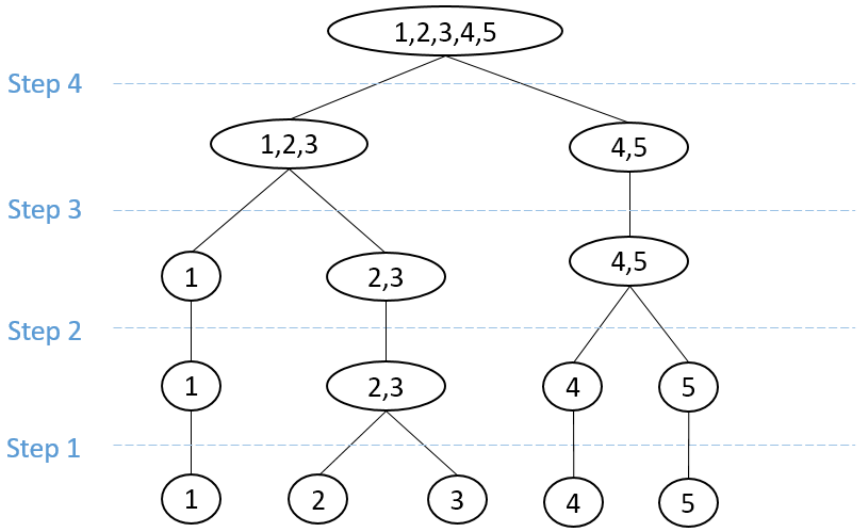

Figure 2.2. We start with 5 clusters each containing a single mutation, i.e., {1}, {2}, {3}, {4}, {5}. We then fit a beta distribution on each possible pairwise combinations of the 5 mutations, which produces 10 corresponding d values. Because the pair {2} and {3} has the minimum distance, we put them together to create a new cluster {2, 3}. Next, we compute the pairwise distance between the remaining 4 clusters and merge {1} and {2, 3} because their distance is the smallest. We iterate this process until all mutations are merged into a single cluster, i.e., the root.

In the second phase, we identify boundaries of distinct beta distributions by traversing and partitioning the tree into clades (i.e., aggregation of clusters below a branching point). Unlike traditional approaches that cut the tree at a fixed branch level, we perform an adaptive splitting (Fig. 2.3). Along the root-to-leaves path, we examine the clade at each branching point and test the null hypothesis that VAFs in this clade are drawn from the same beta distribution. This is done by comparing the observed variance of VAFs with the expected variance of VAFs. Specifically, given a clade containing m variants, we assume they belong to the same subclone and compute α and β by solving equation (1). We then draw m random samples $x_{1:m} \sim Beta(\alpha, \beta)$ and calculate $var(x)$. By

repeating this process 1,000 times, we derive 1,000 $var(x)$ values representing the null distribution. We then use one-sample one-sided t-test to evaluate if $var(v) \leq \overline{var(x)}$. We reject the null hypothesis if the p-value < 0.01, which indicates VAFs of this clade are from heterogeneous beta distributions and need to be partitioned further. Otherwise, we consider this clade as homogeneous and stop traversing below this branching point. We repeat this process till we find homogeneous clades along all branches or we reach the leaf nodes. Each of the resulted homogeneous clade represents a unique cluster.

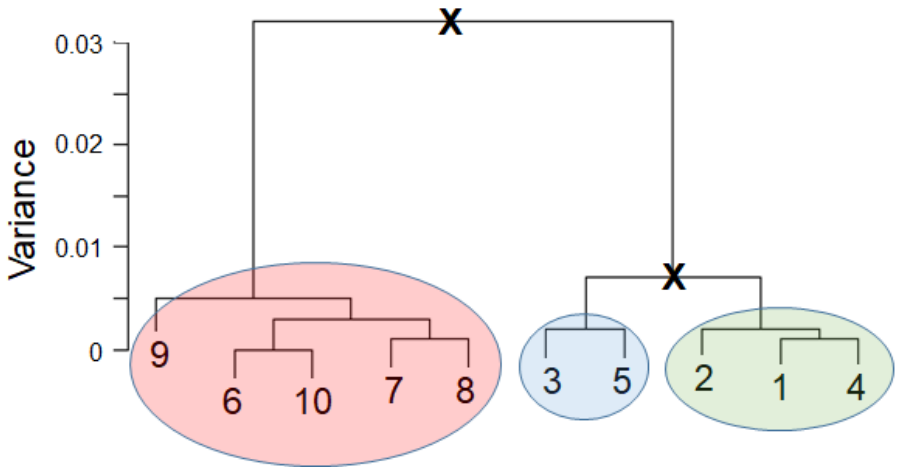

Figure 2.3. Hierarchical clustering and adaptive partitioning. In this example, ten variants at the leaf nodes are progressively grouped into clusters based on VAF similarities to form a tree structure. To partition the tree, we follow the root-to-leave paths. At each branching point, the variance of the VAFs of the clade is compared with the expected variance. A cluster is accepted if the variance is lower than the expected value. Otherwise, it is rejected and partitioning continues (marked by black crosses). In this example, the red, blue and green cluster are accepted.

2.2.1.1 Algorithm for CNVs in Single Sample

For variants located in CNV regions, we assume they do not form new clusters but instead belong to clusters identified from the diploid SNV analysis. Given a variant i, the expected VAF v_i' reflects the cellular prevalence ρ, average ploidy φ of the genomic region it resides and the number k of copies carrying the mutant allele,

$$v_i' = \frac{k\rho}{\varphi} \tag{3}$$

Note that φ is the average ploidy of the focus region in the entire sample and takes a continuous value. Finding the cluster assignment g from among existing SNV clusters $\{1, \dots, C\}$ is to solve

$$\arg\min_{k,g} |v_i - v_i'| = \arg\min_{k,g} \left| v_i - \frac{k\rho_g}{\varphi} \right| \tag{4}$$

where v_i is the observed VAF. We limit the search space of k to integers between 1 and $10 \times \varphi$.

2.2.2 Algorithm for SNVs from Multiple Samples

When multiple samples of a tumor are analyzed, we expect that VAFs representing the same subclone to change concordantly across all samples. However, because the sequencing depth and cellular prevalence of a subclone vary across samples, we need to estimate the beta distribution of this subclone in each sample separately.

MAGOS takes two matrices (R and C) as inputs. The matrix R contains the number of reads mapped to the reference allele ($e_{r,i}^s$) and the number of reads mapped to the alternative allele ($e_{a,i}^s$) for each SNV i in each sample s. The matrix C contains the

average ploidy (φ_j^s) of each CNV j in each sample s. We categorize SNVs into two groups based on if they are located in regions affected by CNVs.

For a SNV i not affected by CNVs in any sample, its sequencing depth in sample s is $e_i^s = e_{r,i}^s + e_{a,i}^s$ and VAF is $v_i^s = e_{a,i}^s / e_i^s$. Given a collection of such SNVs, our task is to organize them into groups so that SNVs in the same group have similar VAFs that are significantly different from variants in a different group. We achieve this task with a two-phase hierarchical clustering and adaptive partitioning algorithm.

In the hierarchical clustering phase, we start with leaf nodes each consisting of an individual SNV, and iteratively merge a pair of nodes with the minimum distance among all pairs to create a new cluster till all SNVs are merged into one root cluster. We define the distance between two nodes (i.e., clusters) C_1 and C_2 as

$$d(C_1, C_2) = \max_s \left(\frac{1}{(|C_1|+|C_1|) \cdot var(v^s) \cdot range(v^s)} \sum_{i \in \{C_1, C_2\}} -\log(\mathrm{P}(v_i^s ; Beta(\alpha^s, \beta^s))) \right) \quad (5)$$

where $v^s = \{v_i^s\}$ for $i \in \{C_1, C_2\}$, and α^s and β^s are the two shape parameters of a beta distribution computed by solving

$$\begin{cases} \alpha^s + \beta^s = \overline{e^s} \\ \frac{\alpha^s}{\alpha^s + \beta^s} = \overline{v^s} \end{cases} \quad (6)$$

where $e^s = \{e_i^s\}$ for $i \in \{C_1, C_2\}$.

In the adaptive partitioning phase, we traverse the tree along the root-to-leaf path. At each branching point, we examine the clade that includes all SNVs below this point. If a clade contains m variants that belong to the same subclone, VAFs of these variants in each sample must come from the same beta distribution, $v^s \sim Beta(\alpha^s, \beta^s)$. To test this hypothesis, we first calculate values of α^s and β^s by solving equation [9] for each sample s. We then draw m random samples $x_{1:m}^s \sim Beta(\alpha^s, \beta^s)$ and calculate $var(x^s) =$

$\sum(x^s - \overline{x^s})^2 / m$. By repeating this sampling process 1,000 times, we derive 1,000 $var(x^s)$ values representing the null distribution. Using the one-sample one-sided t-test, we assess if $var(v^s) \leq \overline{var(x^s)}$. A p-value >0.01 indicates that VAFs of this clade in sample s are indeed from the same beta distribution (i.e., homogeneous). We perform this test for all samples and accept the null hypothesis if all samples produce p-values >0.01. Then, we stop traversing below this branching point and consider variants in this clade constitute a subclone. Otherwise, we will partition this node by following the path and examine the clades at the next lower level. We repeat this process until we find homogeneous clades along all branches or we reach the leaf nodes. Each of the resulted homogeneous clade then represents a unique subclone.

2.2.2.1 Algorithm for CNVs in Multiple Sample

Next, we match SNVs located in CNV regions to subclones identified from the above analysis. Given a SNV j located in a CNV region that has an unknown copy number k^s carrying the mutant allele in sample s, we find the best match subclone g by solving

$$\arg\min_{k^s,g} \sum_s \left| v_j^s - \frac{2k^s \overline{v_g^s}}{\varphi_j^s} \right| \quad (7)$$

where φ_j^s is the average ploidy of the region harboring this variant, and $\overline{v_g^s}$ is the mean VAF of subclone g identified from the previous analysis using only SNVs not affected by CNVs. We limit the search space of k^s to integers between 1 and $10 \times \varphi_j^s$.

2.2.3 Optimization to improve computational efficiency

The standard hierarchical agglomerative clustering procedure requires calculations of all pairwise distances at each step, which leads to an exponential increase of computational complexity as the number of variants grows. However, given the narrow range of VAFs between 0 and 1, not all pairwise comparisons are necessary, especially for variants with highly similar VAFs across all samples. Eliminating unnecessary comparisons is particularly important at the initial clustering steps because the computational complexity of a bifurcating tree is roughly determined by the number of leaf nodes. Based on this principle, we have identified several scenarios in which clusters can be formed without exhaustive search.

- For a single tumor sample
    - If the differences of VAFs of some mutations are too small (< 0.01) to be meaningful, these mutations are collapsed into one cluster immediately above the leaf nodes.
    - By sorting VAFs in an ascending order, it is only necessary to compare adjacent VAFs of leaf nodes.
- For multiple tumor samples
    - Given a set of mutations, if the differences of their VAFs in all samples are < 0.01, we collapse them into one cluster.
    - If we are performing clustering on a large number of mutations, we break them into smaller groups and analyze them in smaller sets. This approach is discussed in the next section.

2.2.3.1 Single Sample Optimization

In single sample clustering, we only have one set of VAFs. If we have n mutations to cluster, there will be $\binom{n}{2}$ possible pairwise combinations. We use an upper triangular matrix to store the distances. In this matrix, the element in row i and column j corresponds to the distance between mutation i and j. After finding the minimum distance, the corresponding row and column are removed from the matrix, but a new column and row are added which will store the distances between the new cluster and the other unchanged clusters. We update this matrix until the last step that the matrix is reduced to a 2 by 2 matrix, which corresponds to the last step of merging where there are only two clusters left to be merged. For large a number of mutations, the pairwise calculations will slow down the process and many computed distances are never used. For example, in the first step, we do not need to calculate distances between mutations that are far from each other. In order to make the process more efficient, first, we sort the VAFs, and then we calculate the distances between the mutations that are next to each other. Because we know that for three mutations, we will merge the closest one first and there is no need to calculate the distance between the points with higher distance (Fig.4).

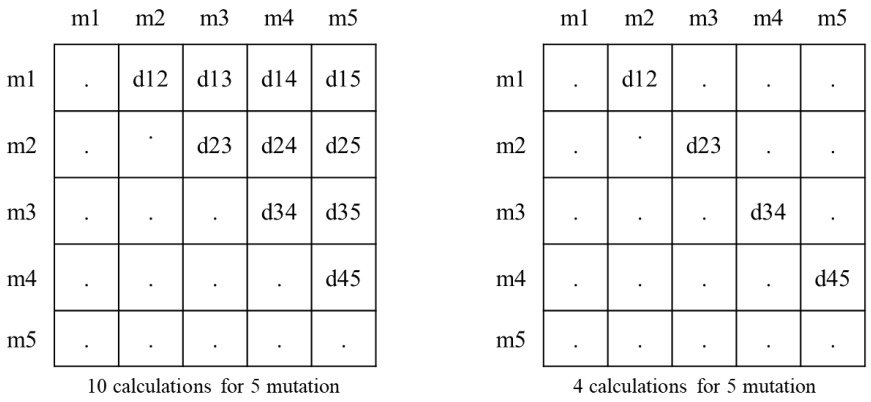

Figure 2.1: For 5 mutations $\binom{5}{2} = 10$ need to be calculated then the pair with the lowest distance are merged. As for the optimized version (on the right) only 4 calculations are needed. As number of mutations increase this difference, contribute largely to computation speed. For n mutations $\binom{n}{2}$ calculation is needed but for optimized version, only $n - 1$.

Another approach that makes the process faster in single sample clustering is that in the preprocessing step, we round the VAFs and collapse the identical ones into one cluster. This will significantly reduce the number of irrelevant computations. For a large number of mutations, the mutations are rounded to two decimal points. The mutations with the exact VAF are merged. The total possible number of starting clusters will reduce to 99 (clusters with mean VAF of 0.01 to 0.99). The hierarchical clustering is then performed on the resulting clusters, and a lot of lengthy initial steps are eliminated.

2.2.3.2 Multiple Sample Optimization

Although the optimization steps in the single sample approach significantly speed up the process, not all of them can be extended to multiple sample clustering because we have more than one set of mutations, and sorting VAFs based on one sample will not help

much. The computational issue for the multiple sample approach is that we add another dimension to our computations. This means that for every pair of mutations, we need to compute the distance between the two across all samples. The pairwise comparison is the nature of the hierarchical clustering. In single sample optimization, we managed to avoid doing all the pairwise calculations because we could sort the mutations and only look at the adjacent groups. However, we do not have that privilege in multiple samples. The idea behind the optimization in the single sample is to merge the mutations that have the same VAF and avoid calculations. So to implement the same idea in multiple samples, we need to merge the mutations that have the same VAF across all samples and collapse them into one group before performing the hierarchical clustering.

In order to do that, we first compute the Euclidean distance between each pair of variants based on their VAFs in all samples. We then construct an incidence matrix and set the entry in row x and column y to 1 if the Euclidean distance between variants x and y is less than a threshold across all samples (the default value for the threshold in the algorithm is 0.01 but can be modified in the code). Using an undirected graph created from this incidence matrix, we search for the largest complete subgraphs and collapse variants belonging to each complete graph into a leaf-node in our initial hierarchical clustering step. To build the graph we take each mutation as a vertex, and we put an edge between two vertices if the value of element corresponding to the two mutations is 1. Each "complete" subgraph represents a group of mutations in which all are inside a sphere with a diameter of the threshold value across all samples, meaning all of these specific mutations are closer to each other than the threshold value (Fig 5).

The benefit of this approach is that in a single step, we can construct multiple clusters and reduce number of leaf nodes significantly, and we avoid calculating the unnecessary distances between the mutations among these clusters and other clusters. Because the number of leaf-nodes determines the computational complexity of a bifurcating tree, this graph-based reduction step accelerates the speed of MAGOS significantly.

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11
m1	.	1	1	1	.	.	.	.	.	.	.
m2	.	.	1	1	.	.	.	.	.	.	1
m3	.	.	.	1	1	.	.	.	.	.	.
m4	.	.	.	.	.	.	.	.	.	.	.
m5	.	.	.	.	.	1	1	1			
m6	.	.	.	.	.	.	1	.	.	.	.
m7	.	.	.	.	.	.	.	.	.	.	.
m8	.	.	.	.	.	.	.	.	.	.	.
m9	.	.	.	.	.	.	.	.	.	1	
m10	.	.	.	.	.	.	.	.	.	.	.
m11	.	.	.	.	.	.	.	.	.	.	.

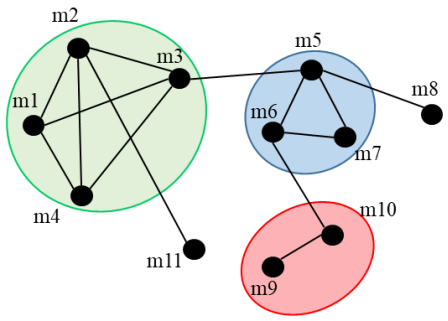

Figure 2.5: The incidence matrix and the corresponding undirected graph created from the distance matrix. In creating the graph, we merge the mutations from the biggest complete graph. For example, m5 also creates a complete graph with m8, but since the subgraph created with m6 and m7 has three mutations, we select that and keep m8 as a separate mutation, m5 and m8 are not merged in this step.

2.3 MAGOS Performance

Discovering the clonal structure of cancer have always been a big challenge. Lack of accurate data, insignificant depth, and lack of ground truth are among the reasons that make it hard to design and tune algorithms and secondly evaluate any developed algorithm. Most of the previous studies have relied heavily on simulations. Either, clones are simulated from real germline mutations or completely simulating the reads with proposed models. Simulations are used to evaluate models and tune the parameters of the models. Although the simulations cannot precisely replicate real data, it is the best that can be done in the absence of real data with ground truth. Because of the levels of heterogeneity in tumor samples, it is costly to be able to capture all of the diversity in the sample. Ideally, in order to get the most accurate view of the tumor, we need to perform single-cell sequencing on a large number of cells from the single tumor, which has its constraints and practicality issues. Bulk sequencing is still the preferred approach. By using bulk sequencing, we need to sequence multiple samples from the tumor with high depth to capture an accurate view of the tumor structure.

Recently, a few studies have tried to sequence a few ultra-deep sequencing samples to establish a high-quality data set to be used in genome sequencing analysis (Griffith et al. 2015). In Griffith et al., they have sequenced two samples from an AML patient up to 10000x coverage to get the most accurate mutation data. The problem with these kinds of studies is that the number of samples is limited, and it is not practical to have all the sequencing samples sequenced at that coverage. Of course, this study has provided us with a valuable evaluation dataset. As mentioned in chapter one, as the sequencing coverage increases, the variance of the clusters decreases, and clusters

become apparent, and clustering becomes trivial when the coverage is high enough. In this study, the data with the highest sequencing depth can be considered as the truth. Using the labels form the high coverage data, and map them back to the low coverage set, we can have the ground truth on the low coverage set.

In this chapter, we present the performance of MAGOS on simulated single samples. We extensively tested MAGOS on simulated data with different depth and different clonal structure and tried to cover almost every clonal possibility. Next, we tested MAGOS on multiple simulated samples. We also ran SciClone and PyClone on each dataset to compare their result with MAGOS. Finally, we used the data published in Griffith et al. to compare the performance of MAGOS with SciClone and PyClone on three different coverages and showed that MAGOS performance is superior to the other methods.

2.3.1 Simulating Subclones in Single Tumor Sample

In here, we explain the down sampling approach that we used in order to generate the simulation data for the single sample evaluation. In each simulation, we create an artificial tumor sample containing two subpopulations sequenced at an average depth of e. Variants in each subpopulation are drawn from a pool that is the founding clone in the primary acute myeloid leukemia sample sequenced at >10,000x depth by Griffith et. Al (Griffith et al. 2015). Because this sample has an estimated purity of 90.3% and the founding clone contains only heterozygous somatic mutations, the mean VAF of variants in this pool is $u = 0.451$. To create a subclone containing m variants with a mean VAF v, we first draw m random variants from the pool. For a given variant, there are e_r^0 number of reads mapped to the reference allele and e_a^0 number of reads mapped to the

alternative allele in the pool. We down-sample these reads to the lower sequencing depth e according to Poisson distributions

$$\begin{cases} e_r = Pois\left(e_r^0 \eta \frac{1-v}{1-u}\right) \\ e_a = Pois\left(e_a^0 \eta \frac{v}{u}\right) \end{cases}$$

where $\eta = e/(e_r^0 + e_a^0)$, and e_r and e_a are the number of reads mapped to the reference allele and to the alternative allele in the simulated tumor sample, respectively. Using this strategy, we generate two subclones each containing $m = 100$ somatic variants and combine them to create an admixture, representing a single tumor sample with a two-subclone structure. We vary the mean VAF v of each subclone between 0.05 and 0.45 and the overall sequencing depth e at 30x, 60x, 100x, 200x, 300x, 500x.

2.3.2 *J* Score

The *J* score is a modification of the $\%G_E$ value used by Miura et. al. to evaluate subclone discovery methods (Miura et al. 2018). However, the *J* score accommodates subclone sizes (i.e. number of variants in each subclone) when quantifying the similarities between two sets of clusters. Specifically, we denote T as a set of clusters representing the ground truth, and D as a set of clusters representing predictions. For each cluster $t \in T$, we find its most similar cluster $d \in D$ based on the Jaccard Index

$$I = \frac{|A \cap B|}{|A \cup B|}$$

where A is the set of variants defining cluster t, and B is the set of variants defining cluster d. After all truth clusters are matched, if there remain unmatched clusters in D, we use the Jaccard Index to find their most similar clusters in T. After finding the best matched pair for all truth clusters and all predicted clusters, we compute the J score as

$$J = \frac{n_p I_p}{\sum_p n_p} \times 100$$

where I_p is Jaccard Index of the p^{th} pair, and n_p is the number of variants in the truth cluster involved in this pair. J score takes a value between 0 and 100, in which 0 means no overlap between any truth clusters and any predicted clusters, and 100 means perfect matches between truth and predictions.

2.3.3 Performance on Simulated Datasets

2.3.3.1 Performance on simulated single tumor samples

In evaluating a model, specifically when there is no ground truth, using simulations is the most common approach. In evaluating MAGOS, we needed to determine how sensitive the algorithm is to depth, the difference between subclones, and number of mutations. In order to do that, we simulated a huge number of data sets that could give us the necessary insight into how good the algorithm can perform and how much it could be trusted. This could be shown in comparison to other popular packages, e.g. SciClone and PyClone. We ran both of these methods on each dataset and compared their results. At the first part, the goal is to estimate the lower bound of sequencing depth and difference of mean VAFs between the clusters that MAGOS can detect.

To estimate the lower bound of sequencing depth and difference of mean VAFs ($\Delta\bar{v}$) between subclones that can be detected by MAGOS, PyClone, and SciClone, we simulated tumor samples with a simple two-population structure. The Griffith et al. study has sequenced a primary leukemia sample at various depths from 60x to over 10,000x (Griffith et al. 2015). This sample had an estimated purity of 90.3% and consisted of 1,343 somatic variants in the founding clone. The distributions of VAFs of these variants

confirmed that the variance of VAFs was negatively correlated with the sequencing depth (Pearson correlation coefficient= –0.54, correlation test p-value=0.02). Using these variants as a pool, we randomly drew two populations each containing 100 variants. We then adjusted the $\bar{v}$ of each population between 0.05 and 0.45 with an interval of 0.05, combined these two populations and simulated read counts at an average sequencing depth of 30x, 50x, 100x, 200x, 300x and 500x via a Poisson-based down-sampling procedure which is discussed later in this chapter. For each combination of $\bar{v}$ values and sequencing depth, we created 10 artificial admixtures. To quantify decomposition accuracies, we computed a weighted Jaccard index (*J*) that considered both the number of clusters identified and the assignment of variants. A *J* score takes a value between 0 and 100, with 100 indicating a perfect match between true compositions and inferred compositions. The *J* score is discussed in more details later in this chapter.

We first examined admixtures in which the $\bar{v}$ of one population was 0.45 representing a founding clone and the $\bar{v}$ of the other population was lower than 0.45 representing a derived clone. We presented three examples to illustrate different decomposition results of these methods. In an admixture with a $\Delta\bar{v} = 0.2$ sequenced at a 30x depth, the distributions of VAFs of the two populations showed substantial overlaps covering a continuous spectrum of VAFs from 0.09 to 0.85 (Fig. 6). While MAGOS found the correct number of clusters, PyClone and SciClone reported excessive clusters, mistaking the large technical variance as variance caused by mixture of multiple populations.

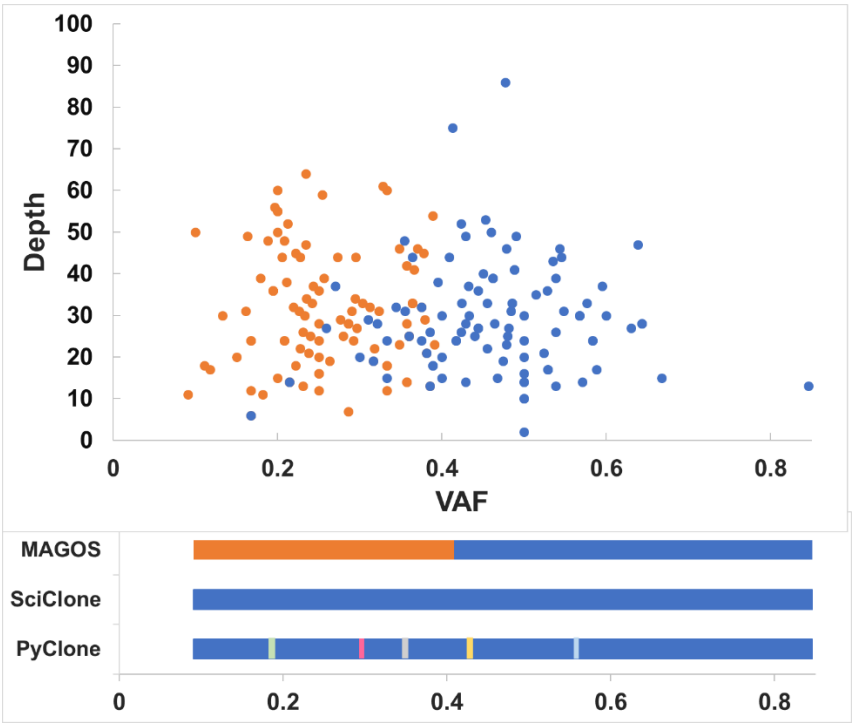

Figure 2.6: Performance of MAGOS, PyClone and SciClone on simulated single tumor samples, consisting of two subclones. The scatter plots show two simulated clusters of variants in a tumor sample. The mean VAFs of the two clusters are 0.45 and 0.20 and the average depth is 30x. The scatter plots is the truth assignment of the mutations. The blue being the clone and the orange cluster being the subclone. MAGOS was able to detect two clusters whereas SciClone only detected one cluster. PyClone as well, found one main cluster, but it detected a few clusters with only one mutation assigned to them.

At the 300x depth, the same admixture produced well-separated distributions of VAFs (Fig. 7). Both MAGOS and SciClone then found two clusters correctly. PyClone however still reported more than two clusters with many clusters containing one or two variants. As we see as the depth increases, the clustering task becomes easier as the distance between the clusters becomes apparent.

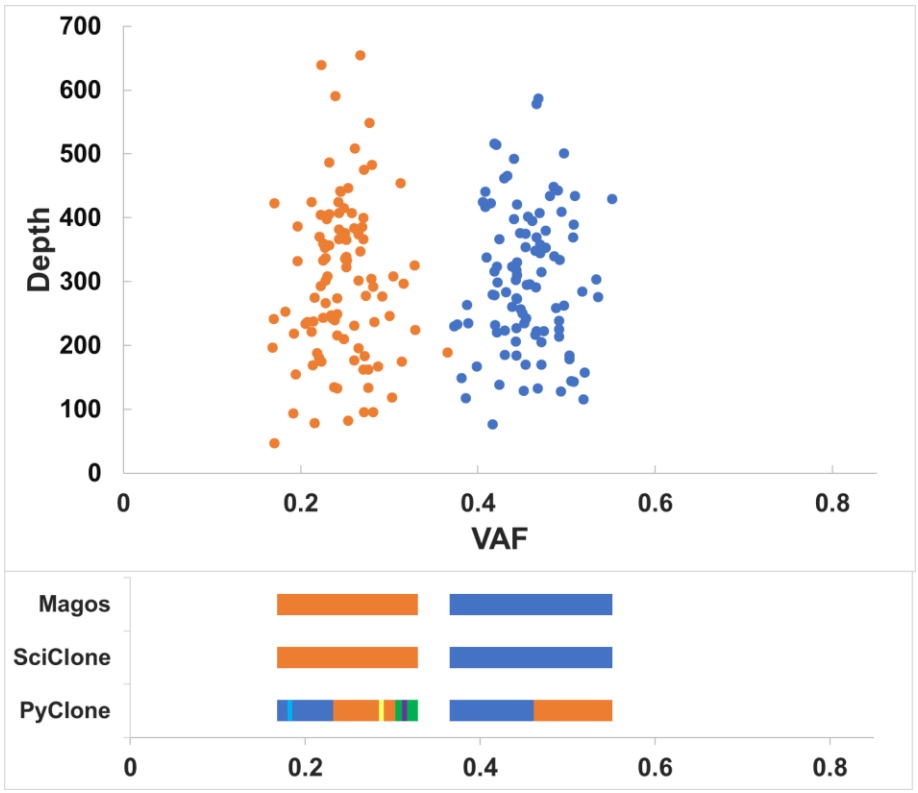

Figure 2.7: Performance of MAGOS, PyClone and SciClone on simulated single tumor samples, consisting of two subclones. The scatter plots show two simulated clusters of variants in a tumor sample. The mean VAFs of the two clusters are 0.45 and 0.20 and the average depth is 300x.The blue being the clone and the orange cluster being the subclone. MAGOS and SciClone were able to detect two clusters. PyClone as well, found two main cluster, but it detected a few clusters with only one mutation assigned to them. PyClone also failed to perform the correct assignment of the mutations belonging to the main clusters.

When we reduced the difference between the subclones frequencies ($\Delta\bar{v}$) to 0.05, it was extremely challenging to divide the two populations even at the 300x coverage (Fig. 8). In this case, SciClone reported one cluster and PyClone reported three main clusters, respectively. Although MAGOS successfully recognized the existence of two clusters, it assigned only 75% of the variants to the correct cluster.

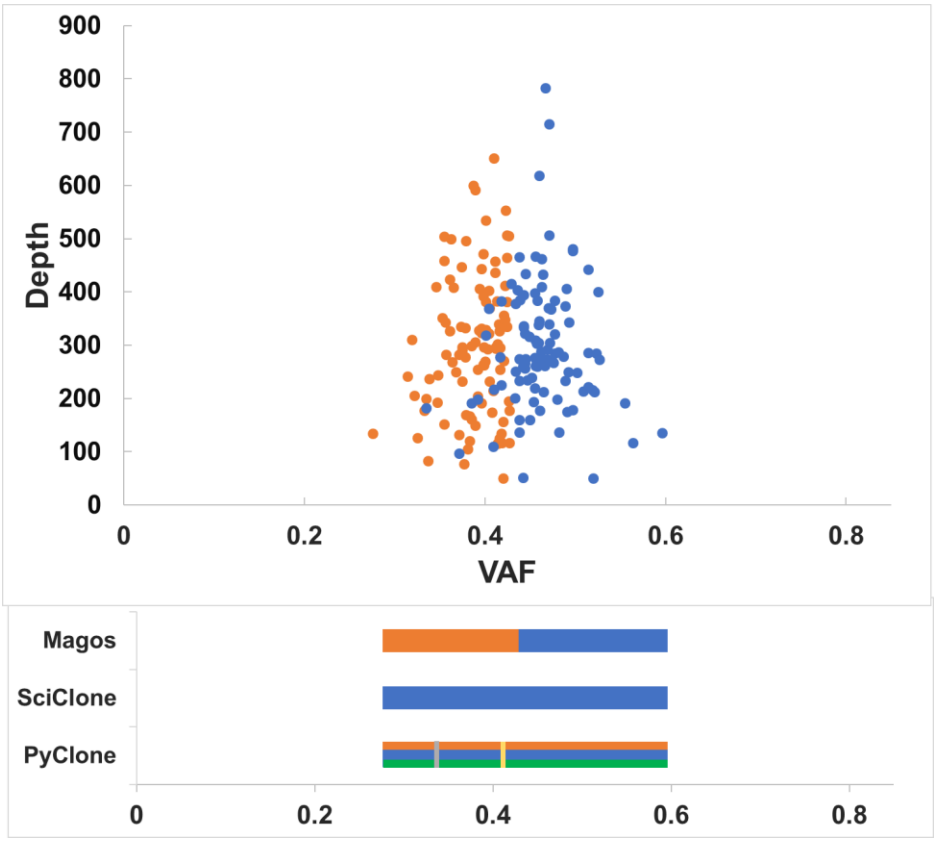

Figure 2.8: Performance of MAGOS, PyClone and SciClone on simulated single tumor samples, consisting of two subclones. The scatter plots show two simulated clusters of variants in a tumor sample. The mean VAFs of the two clusters are 0.45 and 0.4 and the average depth is 300x.The blue being the clone and the orange cluster being the subclone. MAGOS was able to detect two clusters with 75% correct assignment. SciClone found only one cluster. PyClone, found three main clusters, but it detected a few clusters with only one mutation assigned to them.

In the next step, we wanted to determine what is the minimum VAF difference between the clusters that each algorithm can achieve 80% minimum accuracy in different depths. Using J >80 as the accuracy threshold, we recorded the minimum $\Delta\bar{v}$ value between the two populations at a given sequencing depth for each method. The advantage of MAGOS was the most prominent at the depths of 30x – 50x (Fig. 9). In these simulations, MAGOS could produce accurate decompositions with $\Delta\bar{v}$ as low as 0.25.

PyClone required a $\Delta\bar{v}$ of at least 0.35. SciClone could not achieve J >80 at any level of $\Delta\bar{v}$. MAGOS retained the leading position till the sequencing depth increased to 200x, beyond which both MAGOS and SciClone could decompose the admixtures equally well. Interestingly, the minimum $\Delta\bar{v}$ for PyClone remained at 0.35 across all sequencing depths.

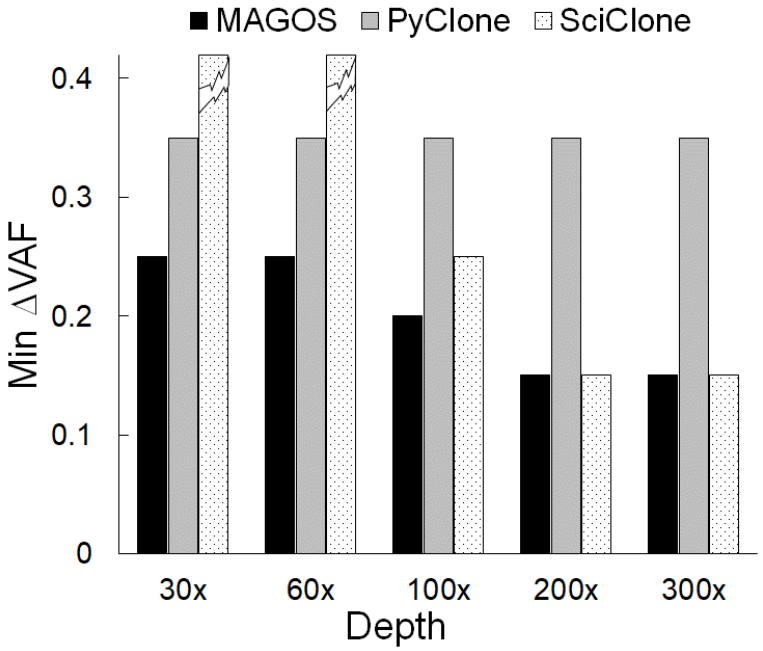

Figure 2.9: Minimum ΔVAF of two subclones that can be decomposed with an accuracy J score >80 by each method. Broken tops on bars indicate J scores >80 cannot be achieved.

Next, we examined the decomposition accuracies of all admixtures. The J score of all three methods was positively correlated with the $\Delta\bar{v}$ value (linear regression coefficients for MAGOS, PyClone and SciClone are 1.30, 1.03 and 0.87, respectively, all p-values $<10^{-12}$). The J score was positively correlated with the sequencing depth for MAGOS and SciClone (coefficients are 0.08, 0.15, respectively, p-value $<10^{-16}$), but not

for PyClone (coefficient=0.006, p-value=0.51). At the 30x depth, MAGOS could achieve an average *J* score ≥ 80 when $\Delta\bar{v} \geq 0.25$ (Fig. 10). In a total of 100 such admixtures, the average *J* score of MAGOS was 86.6, which was significantly better than that of PyClone (73.7, t test p-value=0.008) and SciClone (54.4, p-value=3.5×10^{-8}). As the depth increased to 300x, MAGOS could achieve an average *J* score ≥ 80 when $\Delta\bar{v} \geq 0.15$ (Fig. 11). In a total of 210 such admixtures, the average *J* score of MAGOS was 97.2, which was significantly better than that of PyClone (64.9, t test p-value=2×10^{-8}) but slightly worse than SciClone (98.7, p-value=0.02).

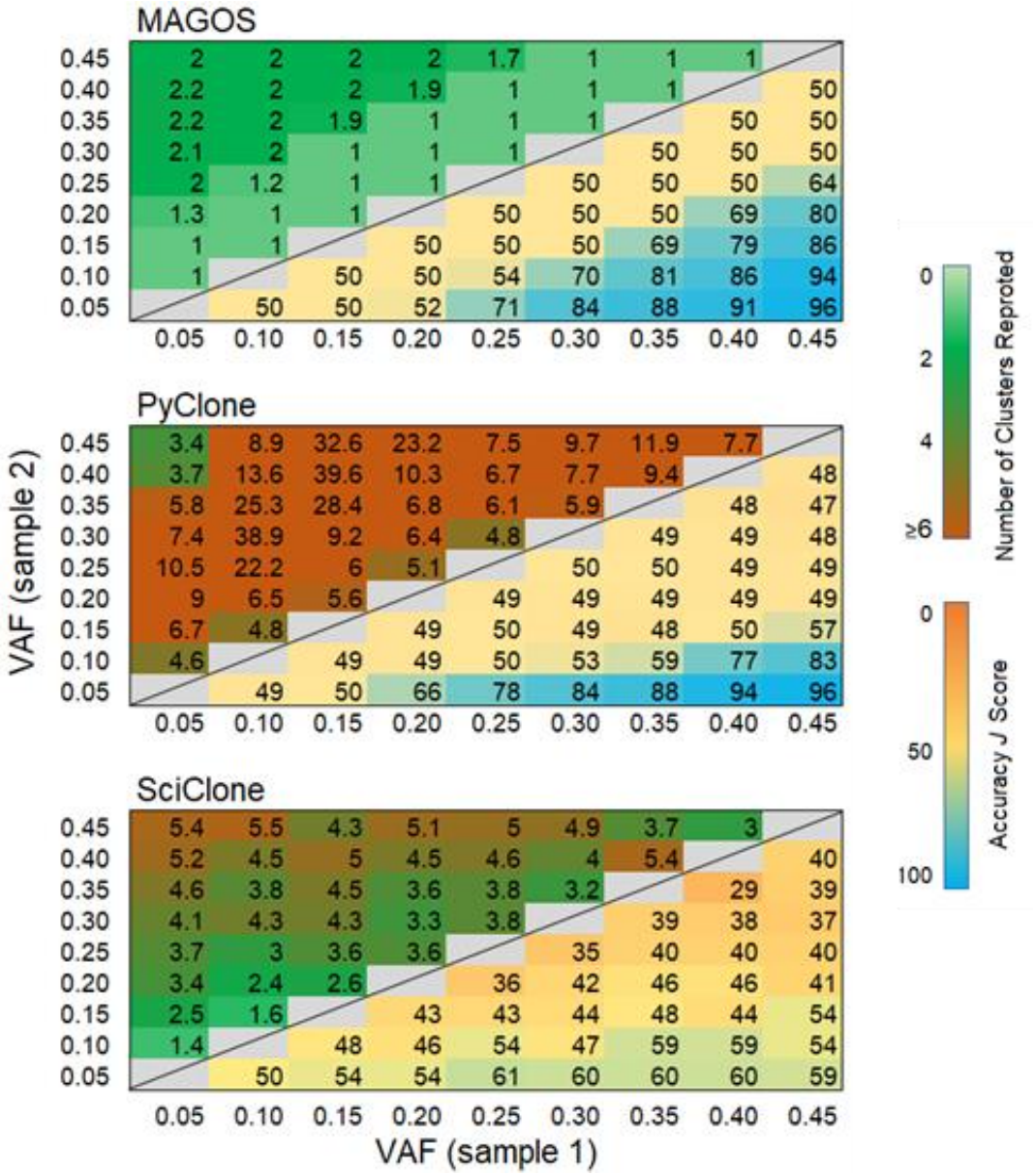

Figure 2.10: Number of reported clusters (upper triangle) and *J* scores (lower triangle) at sequencing depths of 30x. Displayed values are averages of 10 simulations. Perfect decompositions shall report 2 clusters and a *J* score value of 100

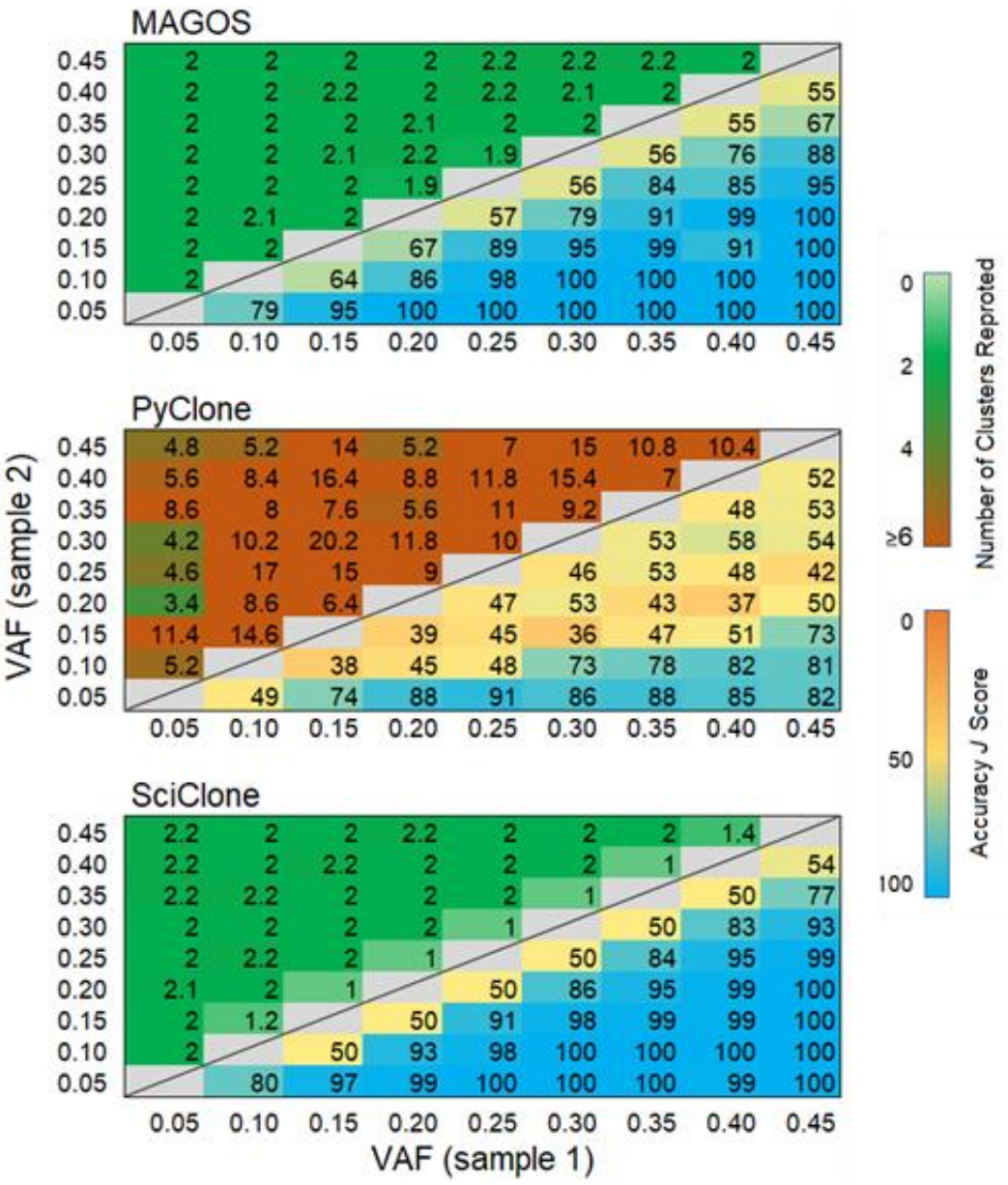

Figure 2.11: Number of reported clusters (upper triangle) and *J* scores (lower triangle) at sequencing depths of 300x. Displayed values are averages of 10 simulations. Perfect decompositions shall report two clusters and a *J* score value of 100

2.3.3.2 Performance on simulated multiple tumor samples

To evaluate the performance of MAGOS on delineating complicated subclonal structure embedded in multiple samples from an individual tumor, we cannot use the simulation we used in the previous section. Because we need to construct more complicated structures. For this matter, we used an established method (El-Kebir et al. 2015) to simulate the admixtures.

The number of sequenced samples from a patient is very important in detecting the clonal structure and as the number of samples increases the accuracy of the detected structure increases significantly. In our multiple sample simulations, we generated set of simulations for two, three and four samples separately.

For the two samples simulations, each simulation set contains 200 variants distributed among 3 subclones, and 10 replicates were generated at sequencing depth of 30x, 50x, 100x and 300x each. Across all depths, MAGOS consistently outperformed the other two methods (Fig 12). The largest improvement was at 30x depth where SciClone and PyClone had a *J* score of 0.66 and 0.78, respectively and MAGOS had a *J* score of 0.82 (paired t-test p-value$<10^{-4}$). MAGOS remained at the leading performance at the sequencing depths increased to 50x and 100x (p-value<0.05). As the sequencing depth reached 300x, SciClone a performed equally well as MAGOS, achieving similar mean *J* scores of 0.95. Interestingly, although all three methods showed better performances at higher sequencing depth, PyClone was the least affected.

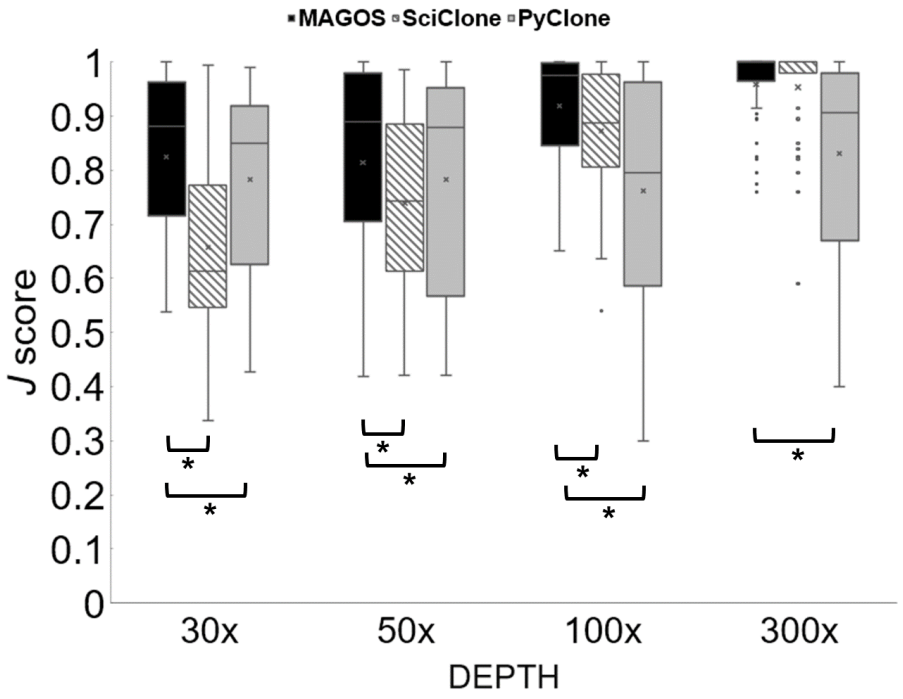

Figure 2.12: Performance on simulated two tumor samples, each consisting of three subclones. Accuracies of different methods tested on tumors sequenced at depth from 30x to 300x. Asterisks indicate a significant better performance of MAGOS as compared to the other two methods.

Similar to the two sample case, for three samples and four samples cases and for each sequencing depths of 30x, 50x, 100x and 300x, we simulated 10 different set of simulations, each set contains 200 variants distributed among 3 subclone. Therefore, for three sample and four sample case, we simulated 40 simulations each.

In both three and four samples, on low coverage, MAGOS consistently performs better than both PyClone and SciClone. As the depth is increased, SciClone's performance improves more than PyClone, but does not get better than MAGOS. MAGOS consistently performs better than all the other two.

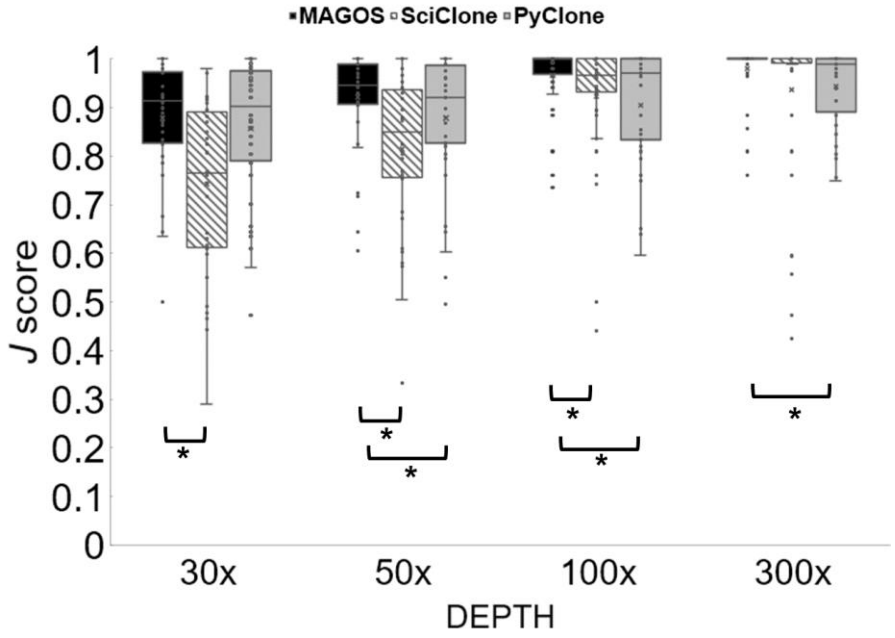

Figure 2.13: Performance on three tumor samples, each consisting of three subclones. Accuracies of different methods tested on tumors sequenced at depth from 30x to 300x. Asterisks indicate a significant better performance of MAGOS as compared to the other two methods.

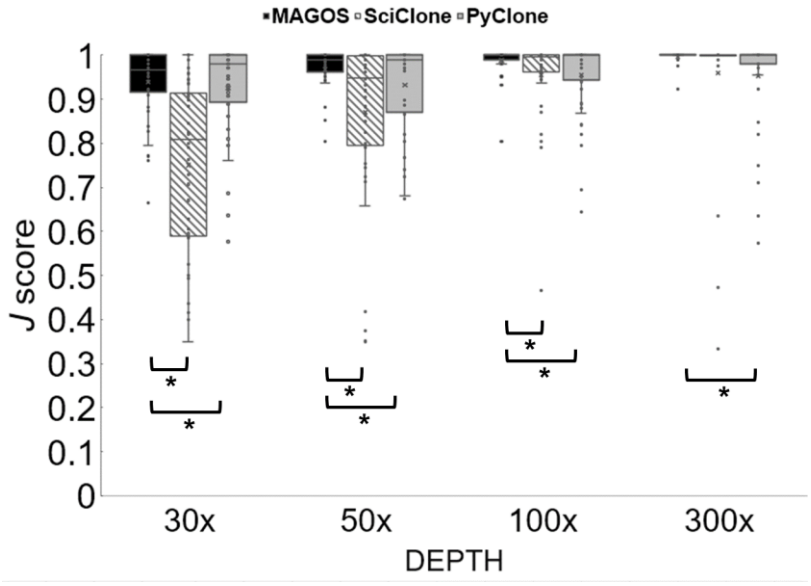

Figure 2.14: Performance on four tumor samples, each consisting of three subclones. Accuracies of different methods tested on tumors sequenced at depth from 30x to 300x. Asterisks indicate a significant better performance of MAGOS as compared to the other two methods.

2.3.4. Performance on real-world sequencing data

In the Griffith et al. study (Griffith et al. 2015), they sequenced the primary tumor sample and the relapse tumor sample of an individual with AML. The goal of the study was to optimize cancer genome sequencing analysis. They claimed that current sequencing strategies are inadequate on complex tumors. More importantly, they presented a comprehensively sequenced and validated dataset to be used as a resource for the community. They performed whole-genome sequencing of the tumor to depths greater than 300x. They also performed targeted sequencing up to 10000x coverage. They performed extensive filtering and manual review on the mutations to produce a set of 1337 high quality mutations. On these set of mutations, they published the sequencing experiment results on them. The resulting set had read counts for this selected set from different range of depths, from low coverage to deeply sequenced. For our use, we used this set of mutations, compared the MAGOS's performance on this set on the different coverages, and compared it to SciClone and PyClone. We showed that MAGOS is able to detect acceptable clonal structure on this set, even on the low coverage reads.

We used the ultra-deep sequencing data published by Griffith et.al to assess the accuracy and reproducibility of MAGOS and the other two methods. This dataset contains 1,337 high-quality somatic SNVs detected in a primary sample and a relapsed sample from a patient with acute myeloid leukemia. Each sample was sequenced at up to 10,000x depth for validation that represents the most comprehensively sequenced tumor. Although the true subclonal structures of these samples are unknown, we followed the authors' suggestion and used clusters detected at highest depths as the benchmark to evaluate clusters found at lower depths (data at 30x, 60x and 300x were available). The

"best truth" consisted of five clusters with high confidences and two clusters with low confidences.

At the depth of 300x, MAGOS, PyClone and SciClone performed equally well, each reporting five to seven tight clusters (Fig. 2.15).

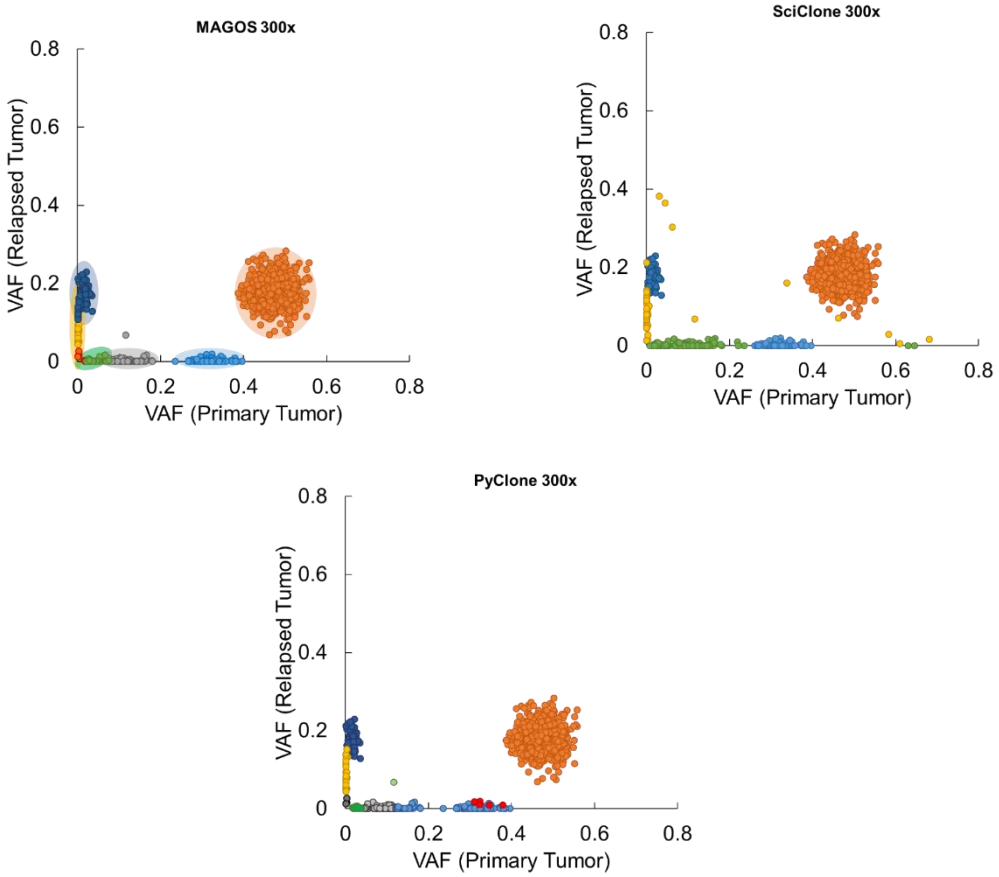

Figure 2.15: Performance on empirical data of two samples (primary tumor and relapsed tumor) from the same patient. Scatter plots show VAFs of variants sequenced at average depths of 300x. Dots of the same color are variants assigned to the same cluster by each method. Shaded ellipses represent "true" clusters inferred from data at ~10,000x sequencing depth. Radiuses of an ellipse correspond to 2 standard deviations of VAFs of variants belonging to a true cluster.

When the depth drops to lower than 60x, VAFs of these subclones showed large overlaps. However, MAGOS was still able to decompose the structure correctly, reporting six clusters. SciClone had great difficulties in separating overlapping clusters and reported only 3 subclones. Results from PyClone was similar to MAGOS but included small-interspersed clusters (Fig. 2.16).

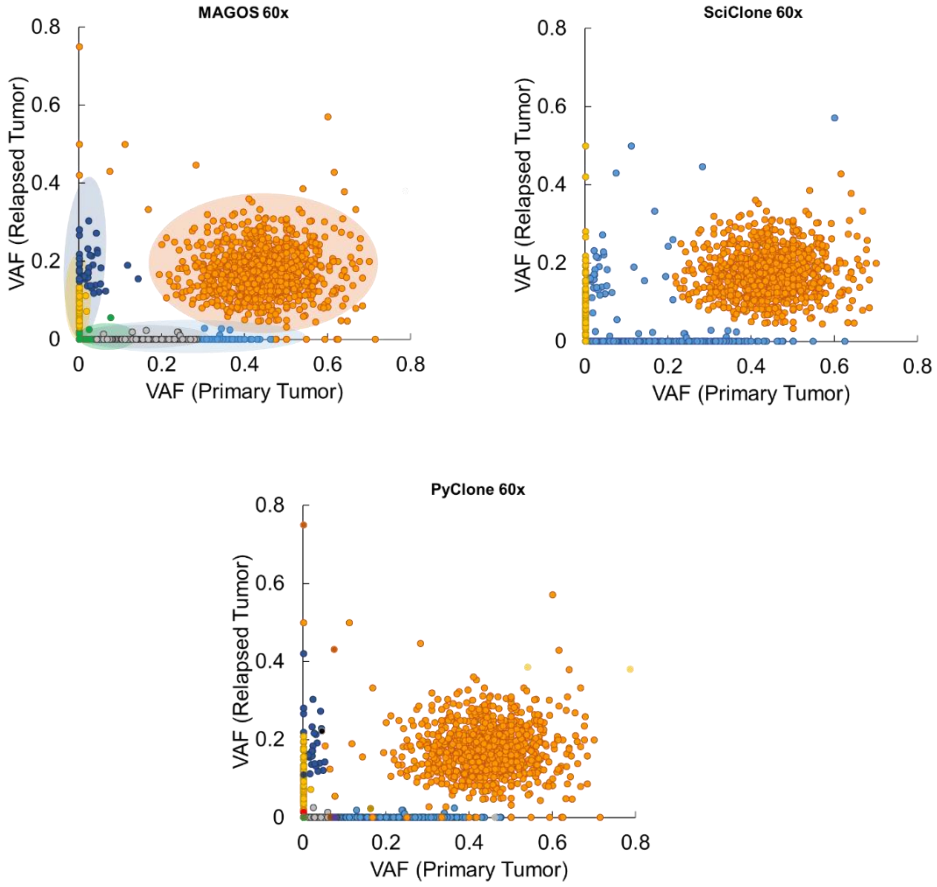

Figure 2.16: Performance on empirical data of two samples (primary tumor and relapsed tumor) from the same patient. Scatter plots show VAFs of variants sequenced at average depths of 60x. Dots of the same color are variants assigned to the same cluster by each method. Shaded ellipses represent "true" clusters inferred from data at ~10,000x sequencing depth. Radiuses of an ellipse correspond to 2 standard deviations of VAFs of variants belonging to a true cluster.

At the depth of 30x, MAGOS was the only method reporting the correct number of clusters and assigning variants to the correct cluster with high accuracies. SciClone reported results similar to 60x. PyClone added more than 10 small-interspersed clusters in its result (Fig. 2.17).

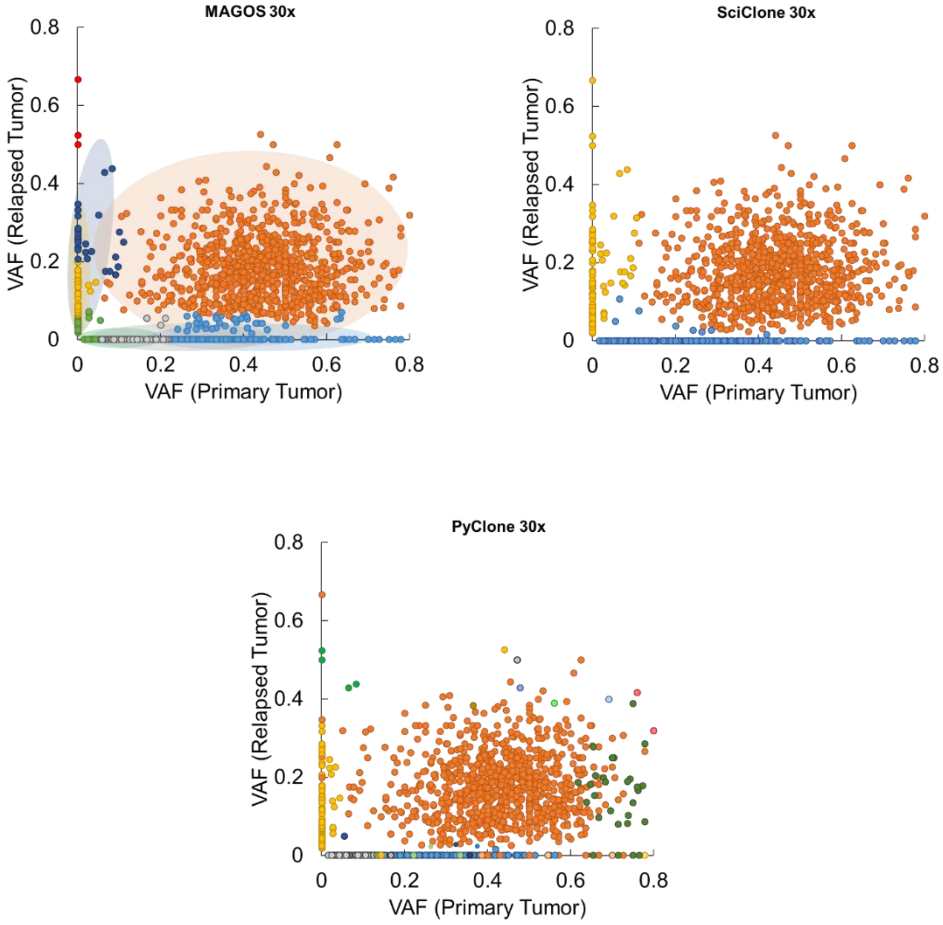

Figure 2.17: Performance on empirical data of two samples (primary tumor and relapsed tumor) from the same patient. Scatter plots show VAFs of variants sequenced at average depths of 30x. Dots of the same color are variants assigned to the same cluster by each method. Shaded ellipses represent "true" clusters inferred from data at ~10,000x sequencing depth. Radiuses of an ellipse correspond to 2 standard deviations of VAFs of variants belonging to a true cluster.

2.3.5. Computational Efficiency of MAGOS

Computational speed and efficiency becomes important when analyzing large number of samples and the combination of accuracy and efficiency is critical in clinical settings. Although, these algorithms are used in research mostly but in designing pipelines, it is critical to be computationally efficient. In this section, we compared computational speed of MAGOS to SciClone and PyClone in three different settings. We evaluated the speed sensitivity of MAGOS to number of mutations, depth of coverage and number of samples.

At first, we tested all three algorithms on simulated datasets with 2 samples, fixed depth of 100x and varying number of mutations (50, 100, 300, 500 and 1000 mutations). For each setting, we simulated 10 simulation sets and recorded the computational speed of the three (Fig. 2.18). As MAGOS is significantly faster than SciClone and PyClone across all number of mutations.

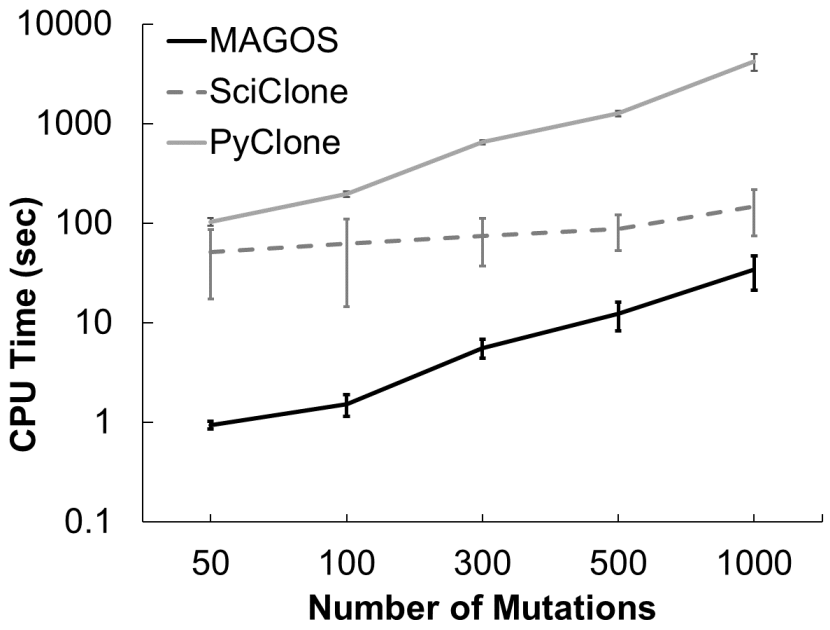

Figure 2.18: Computational efficiency comparisons tested on simulated tumors. (A) Each tumor has two samples and each sample contains 50 to 1,000 mutations sequenced at 100x depth.

Then we tested the speed against the sequencing depth of the data. We fixed number of samples and the number of mutations and varied the depth from 30x to 1000x. In this situation as well, as the depth got larger, MAGOS and SciClone performed faster. The reason for this is that with increasing depth the clusters get denser, and the clustering task becomes easier for MAGOS and SciClone. However, PyClone is not sensitive over depth, and its computation speed is independent of sequencing depth (Fig. 2.19).

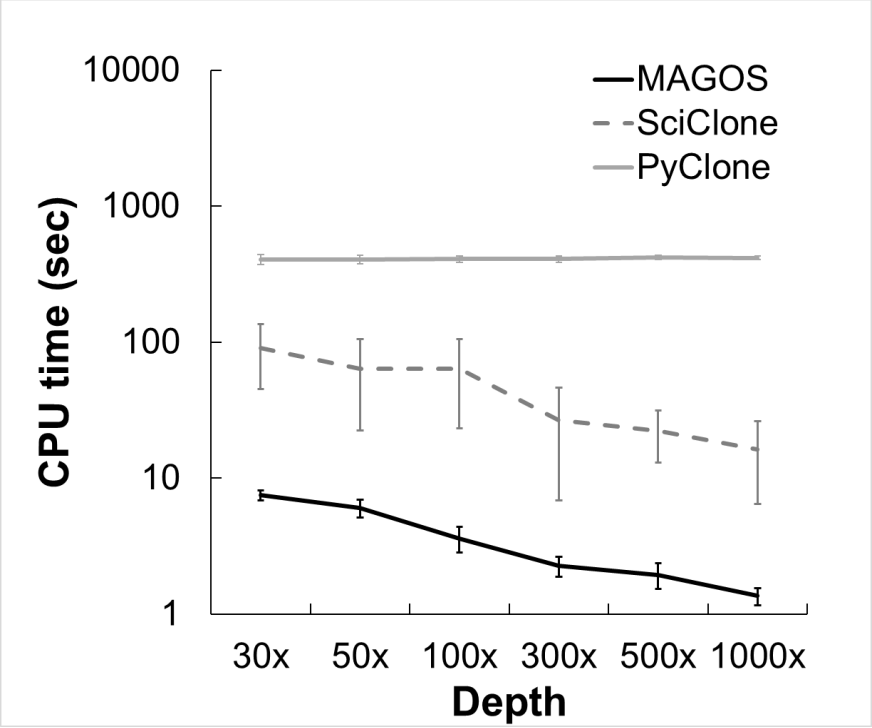

Figure 2.19: Each tumor has two samples and each sample contains 100 mutations sequenced at depth from 30x to 1,000x.

At last, we tested the speed over different number of samples. We fixed the depth (=100x) and number of mutations (=200) and simulated data for 1 to 4 samples. As expected, for all three algorithms, speed decreased by adding additional samples, because number of required computations increased (Fig. 2.20).

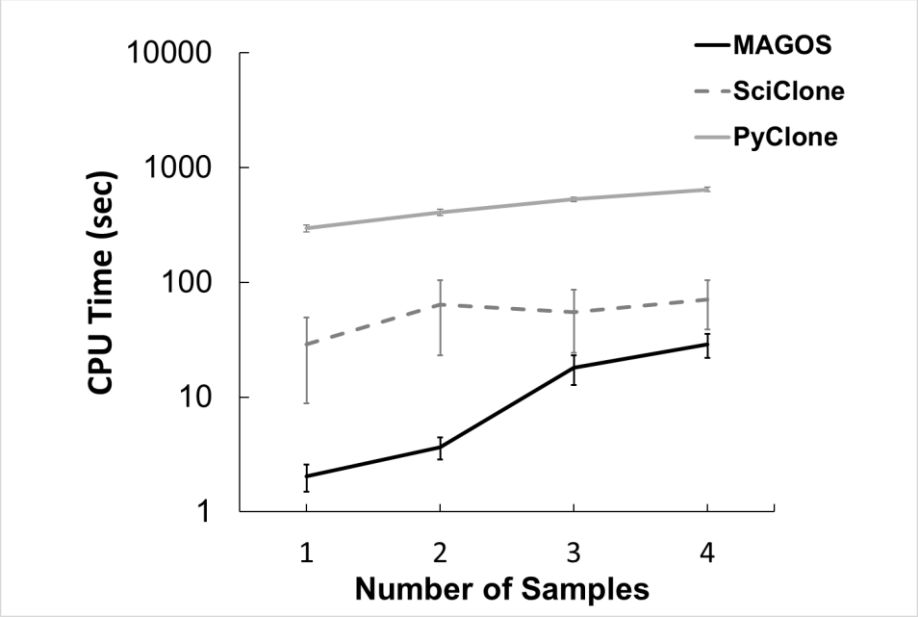

Figure 2.20: Each tumor has 1 to 4 samples and each sample contains 100 mutations sequenced at 100x depth.

2.4 Application to TCGA Data

The TCGA data is valuable source of data from different cancer types as well as clinical data. By executing MAGOS on each sample for 33 different cancer types, we detected the number of clones/subclones in each sample. From the clinical data and clonal data, we analyzed the association between the number of subclones and the survival of the patient. In some of the cancer types, we were able to find strong association between number of clones and the survival of the patient, meaning the number of subclones identified by MAGOS has prognostic power on these cancer types. On the other hand, none of the reported features had any prognostic power on the survival of the patient. In the coming sections, we discuss out significant findings.

2.4.1. Liver Hepatocellular Carcinoma (LIHC)

We applied the MAGOS method to whole-exome sequencing data of 331 liver hepatocellular carcinoma samples from the TCGA project. The majority (79.2%) of these tumors contained 3 or 4 subclones (Table 2.1). Using Cox proportional hazard regressions, we tested if the number of subclones in a tumor was significantly associated with patient overall survival via age at diagnosis, sex and tumor stages as covariates. We found a significant association among tumors of stage III (p-value=0.01, HR=1.67, Fig. 2.21). For comparisons, the number of mutations in a tumor is not a significant prognostic factor among these tumors (p-value=0.44). Therefore, the subclone number is a novel prognostic factor for stage-3 liver cancers that is independent of age at diagnosis, sex and total number of mutations. In Fig. 2.21 it is observed that the number of subclones are negatively correlated with the survival. The patients with fewer subclones, tend to survive longer.

Table 2.1: the distribution of cluster counts across tumor stages in LIHC

		NUMBER OF SUBCLONES IN A TUMOR							
		1	**2**	**3**	**4**	**5**	**6**	**subtotal**	
TUMOR STAGE	**I**	0	25	66	62	13	0	166	50.2%
	II	1	7	38	28	6	0	80	24.2%
	III	0	9	35	30	6	1	81	24.5%
	IV	0	1	1	2	0	0	4	1.2%
	subtotal	1	42	140	122	25	1		
		0.3%	12.7%	42.3%	36.9%	7.6%	0.3%		

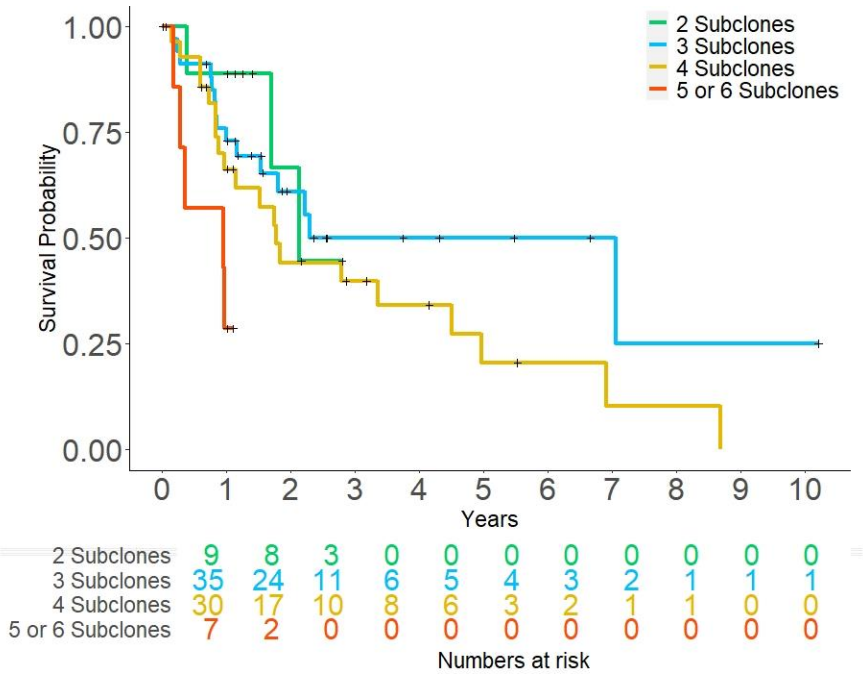

Figure 2.21: Kaplan-Meier plot of liver cancers at stage III. Tumors were stratified into groups based on the number of subclones.

2.4.2. Adenoid cystic carcinoma (ACC)

In Adenoid cystic carcinoma, we discovered some interesting relationships between the number of subclones and the survival. By performing the same analysis, we discovered that the tumor stage and the number of subclones are correlated (p-value= 0.0005, Table 2.2). In the survival plots, we can see that the number of subclones are not strongly associated with survival although by looking at the events after 6 months, the effect of number of subclones become significant. Although in the Cox proportional hazard regression analysis, tumor stage is the strongest prognosis factor and not the cluster count.

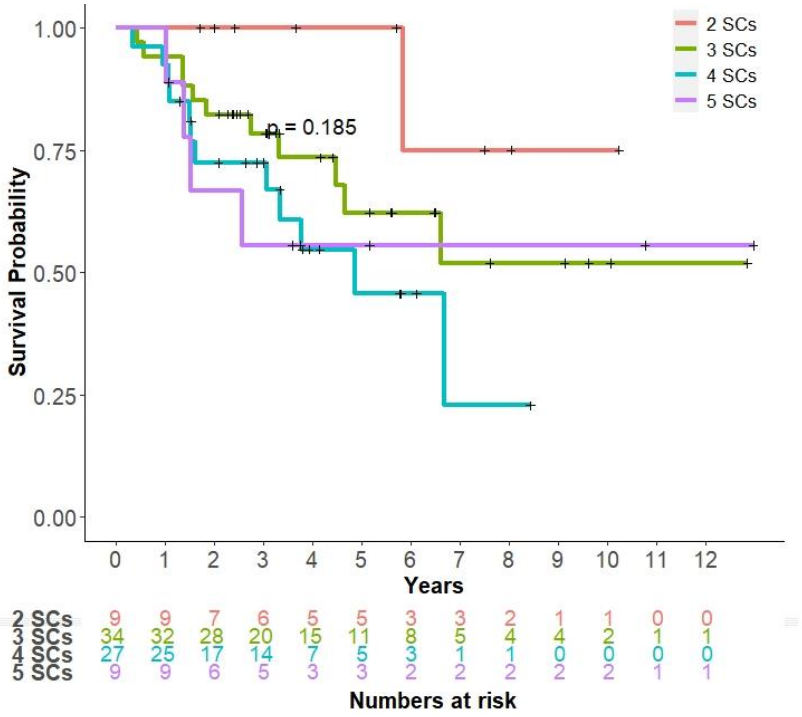

Figure 2.22. Kaplan-Meier plot of ACC. Tumors were stratified into groups based on the number of subclones.

Table 2.2. Counts of subclones across different tumor stages in ACC

		NUMBER OF SUBCLONES IN A TUMOR				
		2	**3**	**4**	**5**	**subtotal**
TUMOR STAGE	**I**	5	3	1	0	9 11.4%
	II	4	19	11	3	37 46.8%
	III	0	6	7	3	16 20.3%
	IV	0	6	7	2	15 19.0%
	subtotal	9	34	26	8	
		11.4%	43.0%	32.9%	10.1%	

2.4.3. Ovarian Cancer (OV)

In the analysis of 272 ovarian cancer primary tumor samples from TCGA, we discovered that the age at diagnosis has the most prognostic power among the available features (p-value=0.0017). Although, by using the Cox proportional hazard model, number of detected subclones by MAGOS also has significant effect on survival of the patient (p-value=0.046). Since tumor stages is not reported, the only important features are age at diagnosis and number of subclones whereas number of mutations is not significant (p-value= 0.32). In ovarian cancer, MAGOS detected 1, 2, 3, 4, 5 and 6 clones. In Fig. 2.23, we look at the survival plots for subclones greater than 4 and samples with fewer than 4 subclones. Interestingly, in OV, if the number of subclones are more than 3, the patients tend to survive longer.

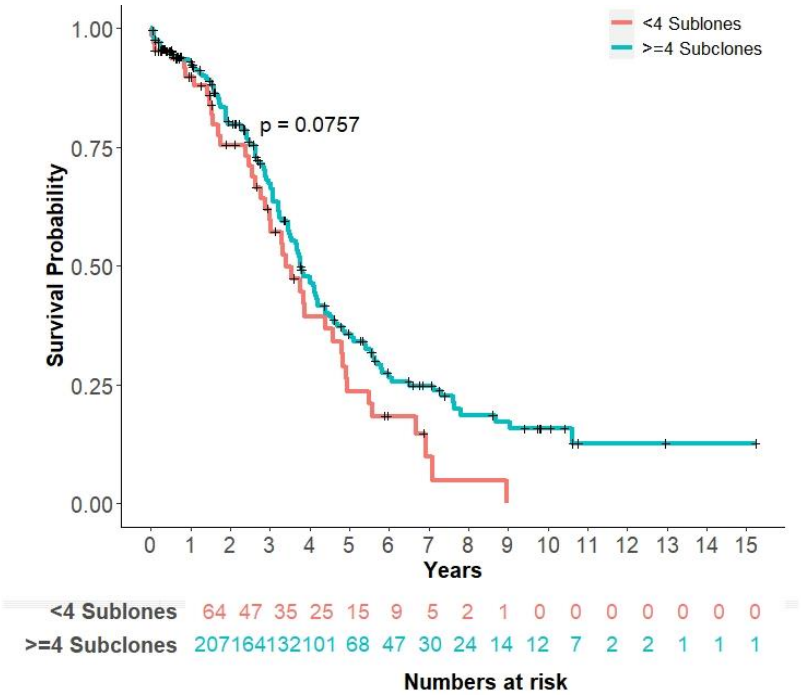

Figure 2.23. Kaplan-Meier plot of OV. Tumors were stratified into groups based on the number of subclones.

2.4.4 Thymoma (THYM)

The analysis of THYM, showed that the number of subclones detected by MAGOS is the most promising feature in predicting survival of the patient. In the Cox proportional hazard model, number of subclones, is the most significant feature (p-value=0.045). Interestingly, age at diagnosis, gender, and number of mutations do not have any significant prognosis (p-values= 0.46, 0.54, 0.1 respectively). The survival plot also indicates this prognosis power. Patients with fewer number of subclones tend to live longer.

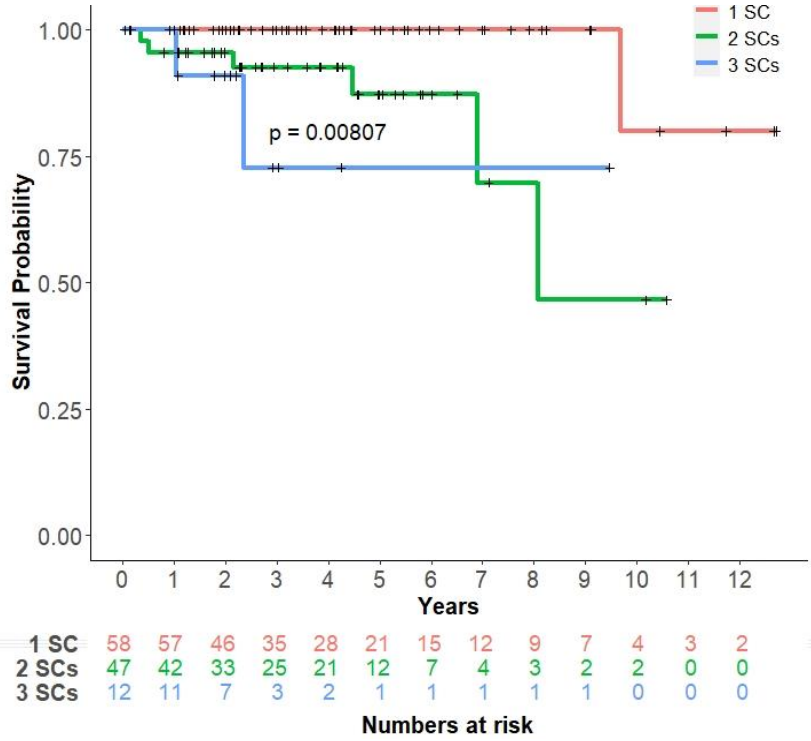

Figure 2.24. Kaplan-Meier plot of THYM. Tumors were stratified into groups based on the number of subclones.

2.4.5 Other Cancers

By analyzing the results from MAGOS on the other cancer types, we were not able to find significant association between survival and the number of subclones as strongly as the LIHC, THYM, ACC and OV. Although, in other cancer types we were able to detect the association if we looked at a specific subset of the data. In this approach, we were able to see the difference in survival curves at the later years, but the effect was not significant overall. We were able to see significant effect when we filtered the patients based on the minimum survival time. For example, in rectum adenocarcinoma, there is a significant difference between the survival rates of the patients that have greater than or equal to four subclones vs the patients with fewer than four subclones if we only look at the patients that remained in the study after 6 months (Fig. 2.25).

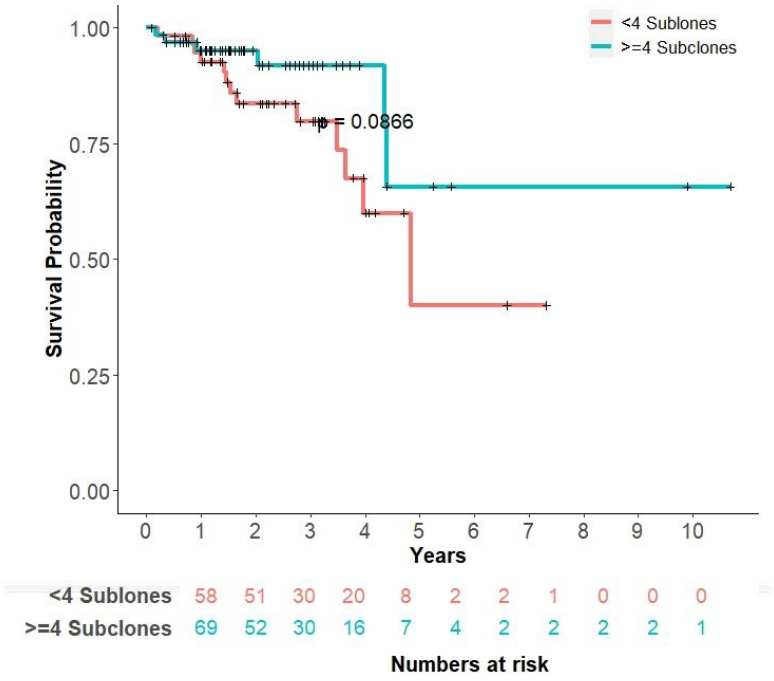

Figure 2.25. Kaplan-Meier plot of READ. Tumors were stratified into groups based on the number of subclones.

We observed a similar trend in GBM, ESCA, CESC, HNSC, KIRP and LGG. We were not able to further derive any solid conclusions on the effect of cluster count on survival. In other cancer types, even the survival curves did not show any meaningful information.

2.5 Discussion

Cancer, as an evolutionary process, is born with a heterogeneous and dynamic nature (Nowell 1976; Greaves and Maley 2012). Precision identification and intervention of cancers shall consider the past, present and future of each tumor. With NextGen sequencing technologies, we can now catch snapshots of this process and potentially reconstruct the evolutionary history and trajectory of a tumor (Griffith et al. 2015; Ding et al. 2012; Landau et al. 2013; Gerlinger and Swanton 2010; Fisher, Pusztai, and Swanton 2013). While single-cell sequencing is a promising technology to examine the genetic compositions of individual cells, uneven genome coverage, low accuracy of variant calls and prohibitive cost limit its usage in subclonal investigations (Ma, Ennis, and Aparicio 2012; Sims et al. 2014; Andor et al. 2014; Navin et al. 2010). The majority of current studies and likely many others in the near future rely on bulk sequencing of mixed tumor cells and computational decomposition to identify variants that occur and evolve together. Several challenges emerge in these analyses.

First, sequencing depth is a key factor affecting the accuracy of identified subclones (Sims et al. 2014). Shown in both the simulated and real-world data, as the sequencing depth drops, the centroid of each cluster remains unchanged while each cluster becomes more scattered, eventually leading to overlaps. Explicit modeling of this

correlation enables MAGOS to accommodate large variances at a lower depth. However, variants in overlapping areas are impossible to separate. Instead of using VAF cutoffs to create artificial borders, a more informative measure is the probability of each variant belonging to a specific clone, which MAGOS reports.

Second, the difference of mean VAFs between subclones limits the power of distinguishing them. Increasing sequencing depth does not change the centroid of each cluster, thus helps little on detecting subclones with similar VAFs. Contrarily, additional samples from the same tumor helps segregate these clusters that are otherwise undiscernible. As MAGOS enables subclonal identifications from genomes sequenced at standard depths, the saved cost can be better invested on analyzing more samples. The benefit of sequencing additional samples is more evident at low sequencing depth, as shown in our test of PyClone. PyClone performs significantly worse than MAGOS when only a single sample is analyzed at low depth. However, when analyzing 2 samples at depth 60x, PyClone results are similar to that from MAGOS.

Third, as multiple samples from a tumor help identify segregating subclones and whole genome sequencing reveals noncoding variants, these additional data also increase the computational complexity, which in turn requests efficient algorithms. We optimized the efficiency of MAGOS that tested its CPU time using simulated tumors. We varied the number of samples for each tumor, the number of mutations in each sample and the mean sequencing depth. Across all configurations, MAGOS showed 3-20x acceleration as compared to SciClone and PyClone, making it a fast and reliable method for subclone decompositions. The identified clusters can be used for further analyses, such as tumor phylogenetic inferences.

CHAPTER 3

CANCER-TYPE SPECIFIC DRIVERS & PROGNOSTIC VALUES

3.1 Introduction

In tumor development, oncogenes (OGs) and tumor-suppressor genes (TSGs) work complementarily to promote and maintain abnormal cell growth (Morris and Chan 2015; Weinberg 1993). OGs cause cancers through gain-of-function variants, whereas TSGs operate by loss of function. While there are a few well known OGs (e.g., RAS) and TSGs (e.g., TP53), it is fast becoming clear that the tumor-enabling activities of a gene is not the same for all types of cancers. Activities of driver genes depend strongly on their cellular contexts because of tissue specific organizations of cancer pathways(Schaefer and Serrano 2016; Schneider et al. 2017; Visvader 2011). Prediction of functional status of genes in different cancer types and cellular contexts is critical for not only understanding tumor biology, but also informing targeted therapies and drug repurposing (Morris and Chan 2015; Schneider et al. 2017; Sleire et al. 2017).

Interestingly, only one computational method (20/20+) is available to predict OGs and TSGs (Tokheim et al. 2016). 20/20+ is an extension of the 20/20 rule in which OGs have >20% mutations causing missense changes at recurrent positions and TSGs have >20% mutations causing inactivating changes (Vogelstein et al. 2013). However, recurrent missense mutations are not a deterministic feature of OGs because these events can cluster at functionally neutral positions due to high mutational rates (Schaub et al. 2018), and many TSGs harbor hotspots of inactivating missense mutations (Iacobuzio-Donahue et al. 2004; M. L. Miller et al. 2015). Meanwhile, random mutational processes may introduce protein-truncating mutations (i.e., nonsense and frameshifting mutations)

into OGs, which increase in frequency via genetic drift with no significant impact on tumor fitness and mislead annotations (Lipinski et al. 2016; Mort et al. 2008; Schaub et al. 2018). Therefore, conventional ratiometric measures are inadequate to distinguish these two groups of genes.

Because tumor development is an evolutionary process, cells carrying somatic mutations are under natural selection within tumors. The positive selection promotes advantageous genotypes that confer higher fitness to a tumor. The negative selection eliminates genotypes with adverse effects. Neutral evolution lets insignificant genotypes to drift up or down in frequency. In OGs, gain of functions may be achieved via missense mutations, which are expected to be positively selected. In contrast, protein-truncating mutations (e.g., nonsense mutations and frame-shifting mutations) often inactivate an OG and are detrimental to tumor fitness, resulting in negative selection. In TSGs, both protein-truncating mutations and missense mutations can be positively selected when they result in the loss of functions. Otherwise, they may drift neutrally or be even under negative selection if they disrupt essential biological functions. For passenger genes (PGs) that do not have significant impact on tumor fitness, we expect that all mutations be under neutral selection (Sun et al. 2017; Williams et al. 2016).

We tested whether the difference in evolutionary dynamics of missense and truncating mutations has sufficient signal and power to improve the detection of OGs and TSGs beyond that of ratiometric measures. Such contrast is essential to distinguish TSGs deactivated by missense mutations from OGs activated by missense mutations, which is a challenging task for conventional ratiometric measures because hotspots of missense mutations are present in both cases. Furthermore, when activities of a gene vary across

cancer types, the direction and magnitude of somatic selection will change accordingly, enabling contextual classification of driver genes.

Our analysis of 10,172 tumor exomes from the cancer genome atlas (The Cancer Genome Atlas 2013) project revealed significant differences in selective patterns of OGs, TSGs and PGs. Based on these patterns, we developed a computational method, named genes under selection in tumors (GUST) that integrates somatic selection of genes in tumor development, molecular conservation during species evolution, and conventional ratiometric measures to classify cancer genes in different tissues and organs.

3.2 Existing Method: 20/20+

Next-generation DNA sequencing of the exome has detected hundreds of thousands of small somatic variants (SSV) in cancer. Distinguishing genes containing driving mutations rather than simply passenger SSVs from a cohort sequenced cancer samples requires sophisticated computational approaches. 20/20+ integrates many features indicative of positive selection to predict oncogenes and tumor suppressor genes from small somatic variants. The features capture mutational clustering, conservation, mutation in silico pathogenicity scores, mutation consequence types, protein interaction network connectivity, and other covariates (e.g. replication timing). 20/20+ uses ratiometric features of mutations by normalizing for the total number of mutations in a gene. 20/20+ uses Random Forest and used the set of OGs and TSGs identified by the original 20/20 rule as a training set. Each gene was scored as the fraction of trees that voted for OG, TSG, or passenger gene. A driver score for each was calculated as the sum of the OG and TSG scores. Even though, 20/20+ extends the 20/20 rule and implements a more sophisticated model and predictive variables, but at its core, it still uses the ratiometric

measures. In our analysis, we show that integrating somatic selection of genes in tumor development improves the driver gene detection. We explain the GUST method, and compare the performance of GUST with 20/20+.

3.3 GUST Method

3.3.1 Curation of Cancer-Type Specific Functions of Driver Genes

To train a random forest model, we needed cancer-type specific functional annotations of cancer genes. Because these annotations are not currently available, we conducted manual curations using two lists of genes with complementary information. The first list consisted of 36 OGs, 48 TSGs and 21 genes with dual OG/TSG roles annotated in the cancer gene consensus (CGC, version 87, (Sondka et al. 2018)). The tumor-activating or -suppressing roles of these genes have been confirmed with cancer hallmarks in experimental assays and are attributable to coding substitutions or indels (Hanahan and Weinberg 2011). The second list consisted of 235 computationally predicted driver genes assigned to specific cancer types (Bailey et al. 2018). These predictions were based on a meta-analysis of the TCGA samples with multiple computational programs. These two lists shared 70 genes. We then retrieved somatic mutations of these 70 genes from the TCGA project (The Cancer Genome Atlas 2013). For a gene to qualify as an OG in a specific cancer type, it needs to be annotated as an OG or a dual-role gene in the CGC, predicted as a driver in the meta-analysis of the matching cancer type, and display mutational hotspots in the corresponding TCGA tumor samples. For a gene to qualify as a TSG in a specific cancer type, it needs to be annotated as a TSG or a dual-role gene in the CGC, predicted as a driver in the meta-analysis, and have an overabundance of truncating mutations or missense mutations in the

corresponding TCGA tumor samples. For a gene to qualify as a PG in a specific cancer type, it needs to be predicted as a PG in the meta-analysis and shows no mutational hotspots nor overabundance of truncating mutations in corresponding TCGA tumor samples. Genes that did not meet these requirements were removed. The final collection consisted of 55 OG annotations, 174 TSG annotations and 304 PG annotations that involved a total of 50 known driver genes and 33 cancer types.

3.3.2 Somatic Selection Features

Given a gene with somatic mutations reported in a collection of tumor samples, we denote the selection coefficient of missense mutations as ω, and the selection coefficient of protein-truncating (nonsense and frameshifting) mutations as φ. To account for differences in mutational rates, we consider seven substitution types (1: A→C or T→G, 2: A→G or T→C, 3: A→T or T→A, 4: C→A or G→T, 5: C→G or G→C, 6: C→T or G→A at non-CpG sites, and 7: C→T or G→A at CpG sites), one insertion type and one deletion type. Based on the statistical framework proposed by Greenman et al. (Chris Greenman et al. 2006), the probability of observing these mutations is a product of multinomial distributions

$$L(\{s_k, m_k, n_k, i_k, f_k\}_k) = \prod_k \frac{t_k!}{s_k!\, m_k!\, n_k!\, i_k!\, f_k!} \frac{(S_k)^{s_k}(\omega M_k)^{m_k}(\varphi N_k)^{n_k}(I_k)^{i_k}(\varphi F_k)^{f_k}}{(S_k + \omega M_k + \varphi N_k + I_k + \varphi F_k)^{t_k}} \qquad (3.1)$$

where s_k, m_k, n_k , i_k and f_k are the observed numbers of synonymous, missense, nonsense, in-frame indel and frameshifting indel mutations in the k^{th} rate category, respectively; S_k, M_k, N_k , I_k and F_k are the corresponding expected numbers of changes

computed by saturated mutations, in which we introduced each possible single nucleotide mutation one at a time; and $t_k=s_k+m_k+n_k+i_k+f_k$ is the total number of observed mutations. The values of log(ω) and log(φ) are determined by maximizing the log likelihood L and constrained within the range of [-5, 5]. The sign and absolute value of log(ω) and log(φ) indicate the direction and magnitude of somatic selection. Values around 0 indicate neutral somatic evolution.

3.3.3 GUST Algorithm

GUST is a random forest model that predicts the class label (OG, TSG or PG) of a gene based on 10 features (Table 3.1, Fig. 3.1). In addition to the log(ω) and log(φ) values, we also compute ratiometric measures to detect mutational hotspots and conservational measures to estimate substitutional rate across species. Specifically, given a gene and a set of somatic missense mutations detected in tumor samples, we applied density estimates with a rectangular kernel and a bandwidth of 5 protein positions to aggregate closely-spaced mutations into peaks and denoted the highest peak as the summit. To estimate evolutionary conservation of a gene, we downloaded multiple sequence alignments of 100 vertebrate species from the UCSC Genome Browser (James Kent et al. 2002), and computed the substitution rate of each protein position (Kumar et al. 2012; L. Liu and Kumar 2013). The average substitution rate over all positions measures the gene-level conservation. The average substitution rate over positions in a summit measures the conservation of a mutational hotspot. For a given gene/cancer-type pair in the curated annotations, we retrieved somatic mutations from the corresponding TCGA tumor samples and computed values of the 10 features. Using these training data, we constructed a random forest classifier with 200 trees. For each gene, this model

produces three probability scores of it being an OG, a TSG or a PG, respectively. It assigns the class label based on the highest probability score. For all predictions, GUST reports random forests probability score, sensitivity and specificity. For OG or TSG predictions, GUST also reports false discovery rate (FDR).

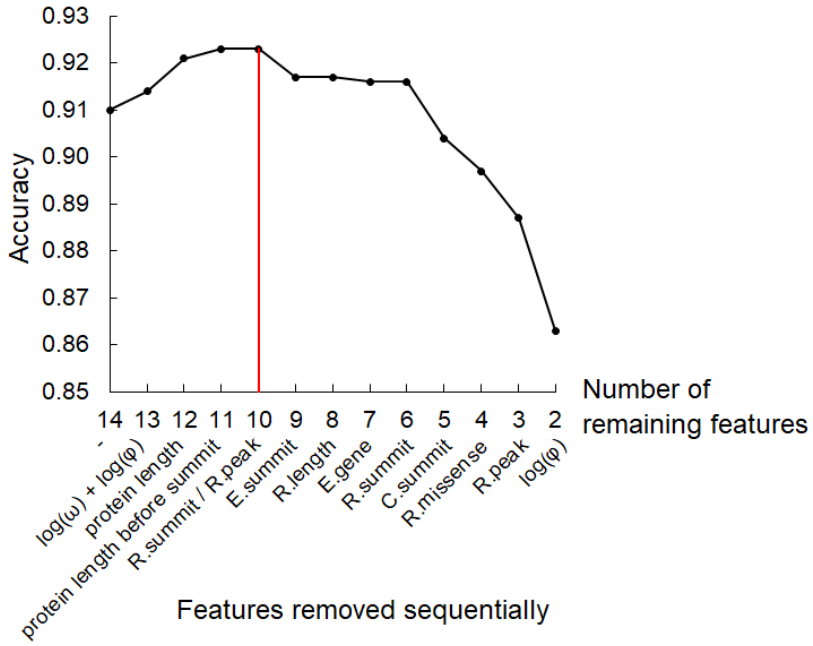

Figure 3.1. Prediction accuracies with decreasing numbers of predictors.

Table 3.1: Features tested, removed and included in GUST random forest classifier

Category		Predictor	Description
The final 10 features included in GUST	Somatic selection	log(ω)	Selection coefficient of missense mutations
		log(φ)	Selection coefficient of truncating mutations
	Ratiometrics	R.missense	Fraction of missense mutations over all mutations
		R.truncating	Fraction of truncating mutations over all mutations
		R.peak	Fraction of missense mutations under peaks
		R.summit	Fraction of missense mutations under the highest peak
		C.summit	Count of missense mutations under the highest peak
		R.length	Mean fraction of lengths of truncated proteins
	Conservation	E.gene	Mean substitution rate of all positions of the protein
		E.summit	Mean substitution rate of positions under the highest peak
Features tested, but not included in GUST		log(ω) + log(φ)	Sum of selection coefficient of missense and truncating mutations
		protein length	Length of protein
		protein length before summit	Position of summit divided by the length of protein
		R.summit / R.peak	Ratio of R.summit and R.peak

3.3.3.1 Feature Selection and Random Forest Classifier

Each record in the training data involved one gene and one cancer type. For a given gene/cancer-type pair, we retrieved somatic mutations from the corresponding TCGA tumor samples and computed values of the 14 features. We then employed a backward selection procedure to identify the most informative features. We started with a baseline model by building a random forest classifier with 200 trees, in which the predictors included all 14 features and the response was class labels. For each gene, this

model produced three probability scores of it being an OG, a TSG or a PG, respectively. It assigned the class label based on the highest probability score. We computed the overall prediction accuracy via 10-fold gene-holdout cross-validations. We then removed one feature, built an alternative random forest model with the remaining 13 features and computed the cross-validation accuracy. Among these 13-predictor alternative models, we identified the one with the highest accuracy and set it as the new baseline. We repeated the above procedure to sequentially remove one feature at a time, and recorded the highest accuracy at each iteration (Fig 3.1). The highest accuracy was 0.923 that corresponded to two models, one using 10 predictors and the other using 11 predictors. We chose the 10-predictor model for GUST because adding the 11th feature did not improve the accuracy. For the 10-predictor random forests classifier, we built ROC curves for one-vs-rest predictions, extrapolated the values corresponding to the probability score and estimated the sensitivity and specificity associated with a prediction.

To compute false discovery rates (FDR), we simulated passenger genes by randomly moving the observed mutations across genes. Given all mutations observed for a cancer type, we first stratify them by seven substitution types (1: A→C or T→G, 2: A→G or T→C, 3: A→T or T→A, 4: C→A or G→T, 5: C→G or G→C, 6: C→T or G→A at non-CpG sites, and 7: C→T or G→A at CpG sites), one insertion type and one deletion type. Within each stratum, we randomly moved the mutations across genes following a uniform distribution, which generated 18,810 simulated passenger genes. We then applied GUST to classify these genes. GUST misclassified 271 genes as OGs and 432 genes as TSGs. We then used the scores of the misclassified OGs to build an

empirical null distribution of OGs. For a test gene predicted as an OG with a score s, we calculated the FDR as the fraction of simulated genes with a score ≥s. Similarly, we built an empirical null distribution of TSGs, against which FDR can be calculated for a test gene predicted as a TSG.

3.4 GUST Results

3.4.1 Different Selection Patterns of Cancer Genes

For each gene/cancer-type pair in our manual annotations, we retrieved somatic mutations in the matching tumor samples from the TCGA project, and computed the somatic selection coefficients. We found that missense mutations in OGs were under stronger positive selection than in TSGs, as the mean log(ω) was 4.18 and 1.68, respectively ($P < 10\text{-}10$, Fig. 3.2). In contrast, protein-truncating mutations showed positive selection in TSGs (mean log(φ) =4.08), but negative selection in OGs (mean log(φ) = –3.25, respectively). The effect size of the differences observed is very large, and the P values highly significant ($P < 10\text{-}8$). The selection measures observed on PGs were close to zero (mean log(ω) =0.60 for missense and mean log(φ)= –0.28 for nonsense mutations). Therefore, TSGs, OGs, and PGs show significant evolutionary differences. The distribution of log(φ) values of PGs had two peaks. The largest peak located close to 0, consistent with the expected neutral selection of PGs. The second peak located close to –5, indicating that loss of function of these PGs is detrimental to tumor growth. Interestingly, many genes in the second peak are established TSGs in other cancer types where loss of their functions is beneficial to tumors. For example, the BCOR gene regulates apoptosis in stomach cancer and had overabundant truncating mutations (The Cancer Genoma Atlas 2013). However, this gene was depleted of truncating mutations in

melanoma (Fig. 3.3). Such contrast suggested that although disabled TSGs promote tumor growth in certain cellular contexts, maintaining their activities may be essential for tumor development in other contexts. We then examined the joint distributions of log(ω) and log(φ) values and found that somatic selection patterns reflected the contextual activities of a gene (Fig. 3.4). For example, the PIK3CA gene had high log(ω) values and low log(φ) values in bladder cancers, breast cancers and colorectal cancers, consistent with its well-known OG role. The log(ω) and log(φ) values of this gene were close to zero in melanoma, indicating lack of a role resulting in neutral patterns. Recently, the passenger role of PIK3CA in melanoma has been proposed in a study that shows PIK3CA-mutated melanoma cells rely on cooperative signaling to promote cell proliferation and PI3K inhibitors do not repress tumor growth in the absence of other activating driver genes in melanoma (Silva et al. 2017).

For TSGs, such as TP53, their high log(φ) values occupied spaces distant from OGs in the distribution plot (Fig. 3.4). As discussed earlier, TSGs with hotspots of missense mutations, such as the FBXW7 gene in uterine carcinosarcoma (Fig. 3.5) are challenging to distinguish from OGs using ratiometric methods. Based selection measures (log(ω)=5.0, log(φ) =3.9), this gene is unambiguously separated from OGs (log(φ)<<0), consistent with our expectations.

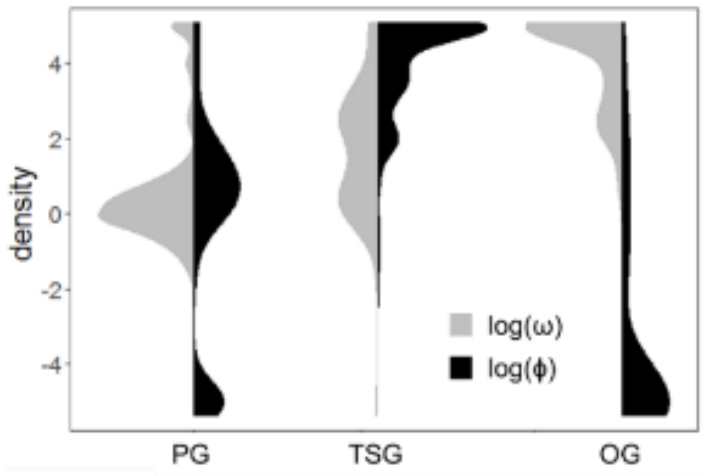

Figure 3.2 Split violin plot showing densities of log(ω) and log(φ) values for PGs, TSGs and OGs

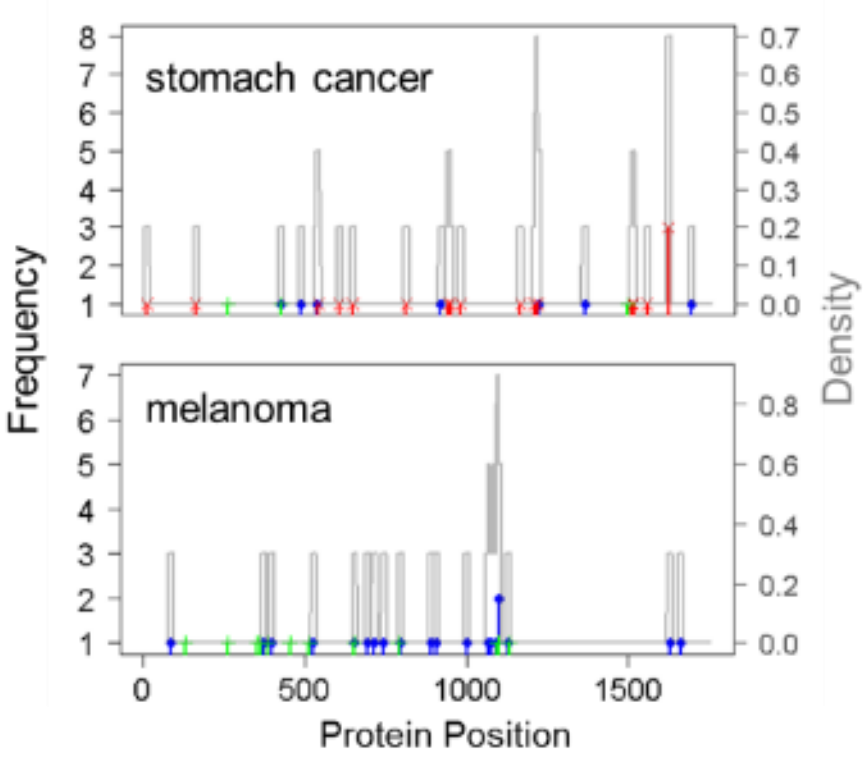

Figure 3.3. Positional distribution of somatic mutations of the BCOR gene in stomach cancer and in melanoma. Vertical lines represent frequencies of various type of mutations at a given position. Synonymous, missense and truncating mutations are represented by green, blue and red lines, respectively. Gray lines are density curves.

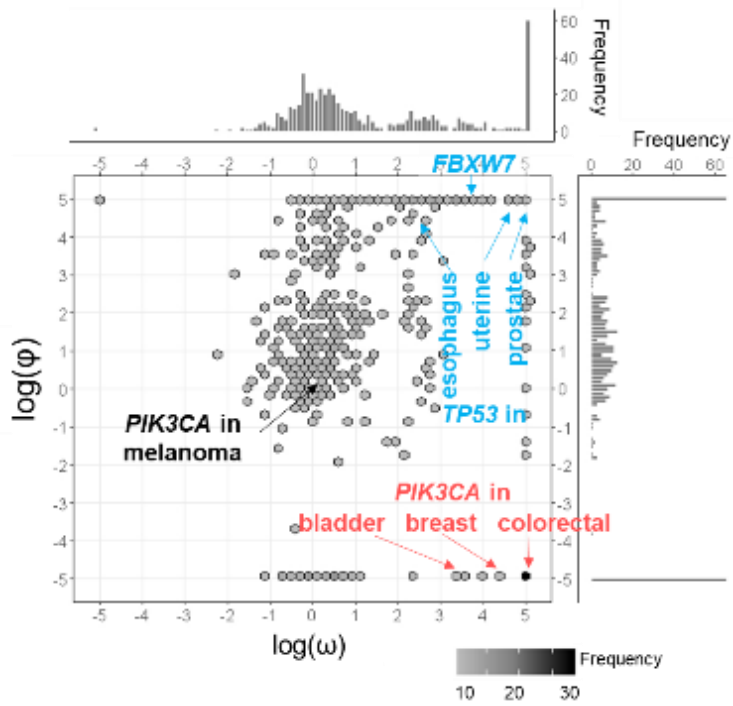

Figure 3.4. Scatter plot of log(ω) and log(φ) values. Shades of hexagon bins represent the number of observations.

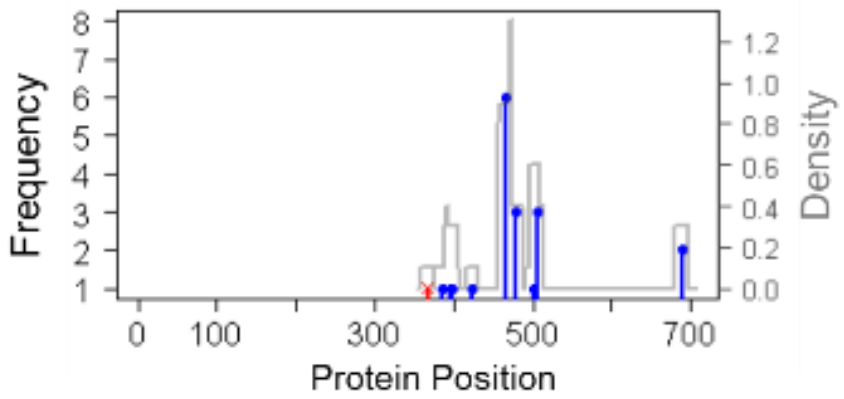

Figure 3.5. Positional distribution of somatic mutations of the FBXW7 gene in uterine carcinosarcoma

3.4.2 Performance of the GUST Method

We trained a random forest classifier (GUST) using the 10 features of the curated genes. Via 10-fold gene-holdout cross-validations, the testing accuracy of GUST was 0.92. As a comparison, the accuracy of 20/20+ on the entire training dataset was 0.86. To calculate traditional performance metrics, we converted three-class predictions to binary predictions by contrasting one class with the other two classes combined, i.e., one-vs-rest predictions. In all categories, GUST showed better or comparable performance than 20/20+. The largest improvements were on the precision of identifying OGs and TSGs, which increased from 0.78 – 0.82 in 20/20+ to 0.85 – 0.92 in GUST (Table 3.2). The receiver operating characteristic (ROC) curves reconfirmed the superior performance of GUST (Fig. 3.6). Compared to 20/20+, GUST had a significantly higher area under the curve (AUC) value of the PG-vs-rest ROC curve (0.97 vs. 0.94, DeLang's test P=0.0008), and a significantly higher AUC value of the TSG-vs-rest ROC curve (0.97 vs. 0.93, P=0.001). However, the AUC values of the OG-vs-rest ROC curves were not significantly different between these two methods (0.99 vs. 0.97, P=0.21).

To evaluate the concordance of GUST classifications with other methods that predict cancer drivers but do not distinguish OGs and TSGs, we computed a driver score by adding the OG and TSG scores of each gene. The TCGA PancanAtlas consortium reported a collection of putative driver genes based on consensus predictions from 12 computational methods (Bailey et al. 2018). We first examined the 510 gene/cancer-type pairs (204 unique genes) predicted as drivers by ≥2 methods. In this permissive list, GUST predicted 373 pairs (73.1%, 145 unique genes) as drivers. We then examined the 283 gene/cancer-type pairs (109 unique genes) predicted as drivers by ≥3 methods. In this

more stringent list, GUST predicted 254 pairs (89.8%, 96 unique genes) as drivers. These results showed that drivers predicted by GUST had a high agreement with existing methods while providing additional OG/TSG classifications.

Table 3.2. Performance of GUST and 20/20+

		Binary Classes			***Three Classes***
	Positive	OG, TSG	OG	TSG	
	Negative	PG	PG, TSG	PG, OG	
GUST	*TPR*	0.93	0.84	0.93	-
	TNR	0.94	0.98	0.95	-
	PPV	0.92	0.85	0.9	-
	NPV	0.95	0.98	0.96	-
	ACC	0.94	0.97	0.94	0.92
	AUC	0.97	0.99	0.97	0.98*
20/20+	*TPR*	0.90	0.95	0.86	-
	TNR	0.85	0.97	0.90	-
	PPV	0.82	0.78	0.81	-
	NPV	0.92	0.99	0.93	-
	ACC	0.88	0.97	0.89	0.86
	AUC	0.94	0.97	0.93	0.95*

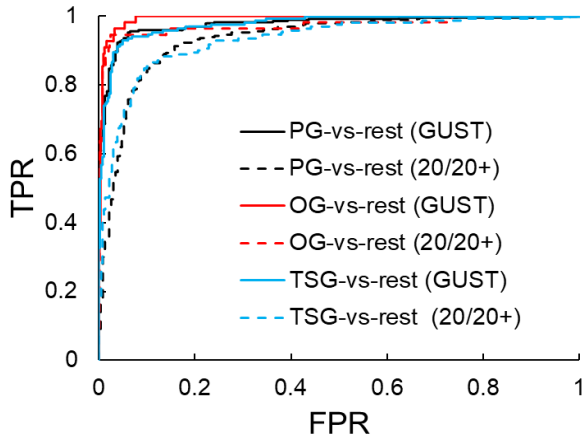

Figure 3.6. ROC curves of one-vs-rest predictions for GUST and for 20/20+

To measure the importance of each predictor in the random forest model, we computed the mean decreased Gini index by per-muting out-of-bag samples (Louppe et al. 2013). The most informative predictors are the selection coefficients and fraction of truncating mutations, followed by the selection coefficient and fraction of missense mutations (Fig. 3.7). Interestingly, evolutionary conservation was not very informative, which may be because a vast majority of drivers are known to occur at highly conserved positions (Dudley et al. 2012), providing a limited power to dis-criminate OGs and TSG.

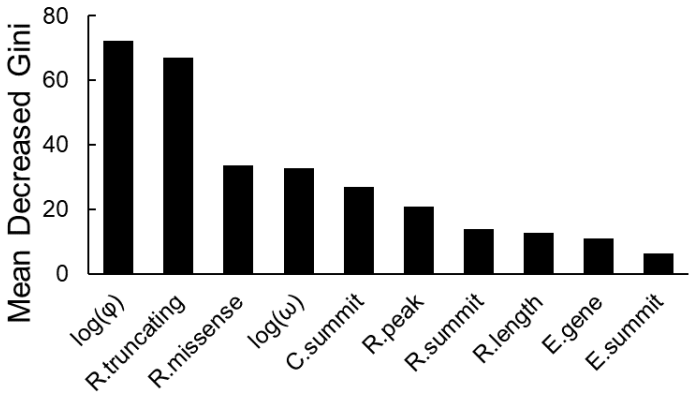

Figure 3.7. Variable importance of each feature in the random forest model.

Although recurrence among patients has been taken as a surrogate of mutations under functional selection, recent investigations have shown that passenger hotspot mutations are common (Buisson et al. 2019; Hess et al. 2019). For example, multiple samples of various cancer types harbored a C->T or C->G mutation at position 931 of the MB21D2 gene (Fig. 3.8). Buisson et al. discovered that this mutational hotspot is due to its location in a hairpin loop susceptible to mutagenesis and functions as a passenger (Buisson et al. 2019). GUST analysis confirmed that the selection pattern of this gene

was consistent with neutral evolution in individual cancer types and in the combined samples (Fig. 3.9). Thus, GUST predicted the MB21D2 gene as a PG correctly. This demonstrated the effectiveness of quantifying the contribution of genetic alterations to tumor fitness in cancer gene classifications.

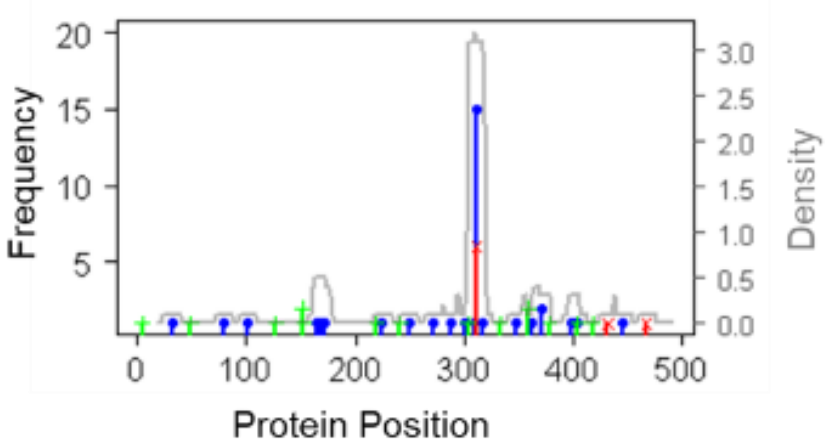

Figure 3.8. Positional distribution of somatic mutations of the MB21D2 gene. Mutations were combined from tumor samples of bladder cancer, cervical cancer, head and neck cancer, lung adenocarcinoma, and lung squamous cell carcinoma. A mutation hotspot is located at coding position 931 that corresponds to protein position 311.

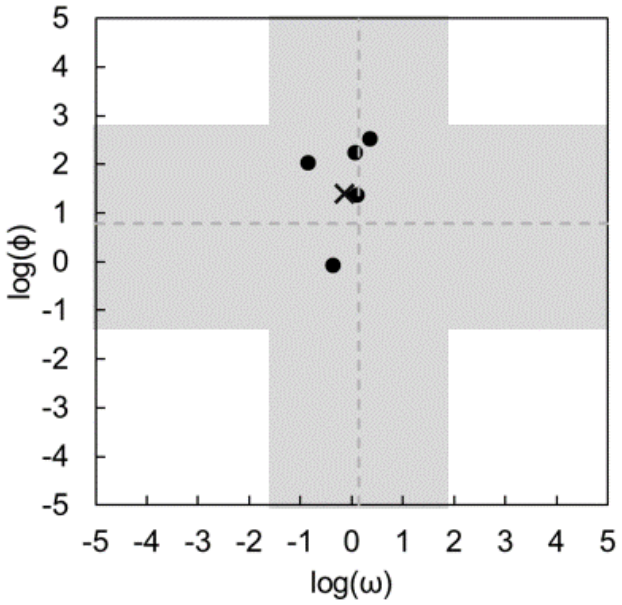

Figure 3.9. Selection coefficients estimated for the MB21D2 gene in individual cancer types (dots) and for combined samples (cross). Broken lines are the mean selection coefficient of all genes analyzed using all TCGA samples. Shaded areas are the 95% confidence intervals of the mean selection coefficients.

3.4.3 Application to TCGA Data

We retrieved somatic mutations from whole-exome sequencing data of 10,172 TCGA tumor samples spanning 33 cancer types. We then removed low-quality mutations, hyper-mutated or hypo-mutated samples, genes with fewer than 4 protein altering mutations, and genes mutated in <2% of tumors which we will discuss below. We applied GUST to the remaining 9,663 samples. We predicted 161 OGs of which 98 were unique genes in 29 cancer types. We also predicted 331 TSGs of which 179 were unique genes in 33 cancer types (Fig. 3.10).

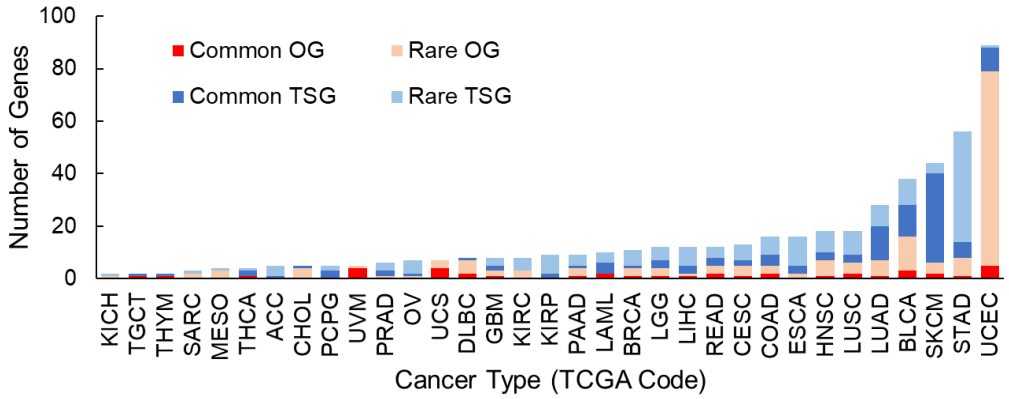

Figure 3.10. Number of common and rare OGs and TSGs found in each cancer type.

3.4.3.1 Preprocessing TCGA data

The TCGA project provides whole-exome sequencing data of 10,172 tumor samples representing 33 cancer types (The Cancer Genoma Atlas 2013). We retrieved somatic mutations called by the GATK/MuTect pipeline (Cibulskis et al. 2013) against the hg38 human reference genome from the Genomic Data Commons (GDC) data portal (Grossman et al. 2016). We removed low-quality mutations and kept single nucleotide substitutions causing synonymous, missense or nonsense changes, and indels causing in-frame or frame-shifting changes of the encoded proteins. Due to the rare occurrences (<1%) and low confidence (Jian, Boerwinkle, and Liu 2014) of predicted splice site mutations, we did not include these mutations in the analysis. We computed the mutational load of a tumor as the number of mutations it contained. For each cancer type, we removed samples with mutational loads outside the 1.5 interquartile range below the first quartile or above the third quartile, respectively. A total of 194 hypo-mutated and 315 hyper-mutated were removed. For each cancer type, we removed less frequently mutated genes that had fewer than 4 protein-altering mutations or were mutated in less than 2% of tumor samples to avoid non-convergence problems during maximum

likelihood estimations of selection coefficients. Because uterine corpus endometrial carcinoma had a significantly higher mutational load than the other cancer types (z test p-value=0.017), we increased the threshold for this cancer type to remove genes mutated in less than 5% of tumor samples.

3.4.3.2 Novel Driver Genes

The GUST-predicted drivers consisted of 55 putative OGs and 97 putative TSGs that were classified as PGs in the CGC database (Sondka et al. 2018). Most (81.7%) of these new putative drivers were annotated in only one cancer type and had low probability scores. To estimate the confidence of each prediction, we computed the sensitivity and specificity of each one-vs-rest prediction based on the ROC curves. We then derived a list of high-confidence drivers consisting of 22 OGs with OG-vs-rest specificity ≥0.99 and 74 TSGs with TSG-vs-rest specificities ≥0.99, all of which had a PG-vs-rest sensitivity ≥0.99. This short list of high-confidence drivers included two novel OGs and 28 novel TSGs not annotated in the CGC. The two novel OGs (CNOT9 in melanoma and GTF2I in thymoma) had single mutational hotspots disrupting highly conserved protein positions (Fig. 3.11, Fig 3. 12). The GTF2I mutant stimulates cell proliferation in vitro and has been associated with favorable prognosis of thymoma (Roy 2017).

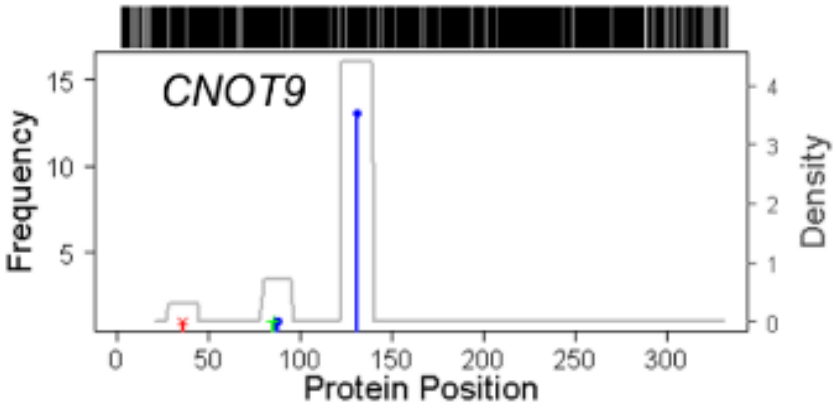

Figure 3.11. Positional distributions of somatic mutations of CNOT9 in melanoma. Evolutionary conservation of each position, measured as number of substitutions per billion years is displayed above each plot.

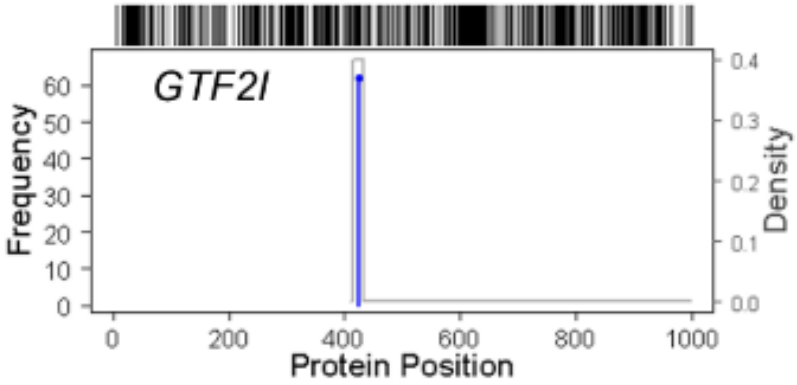

Figure 3.12. Positional distributions of somatic mutations in GTF2I in thymoma. Evolutionary conservation of each position, measured as number of substitutions per billion years is displayed above each plot.

All of the novel TSGs had an overabundance of truncating mutations (FIG 3.15). For example, frameshifting mutations in SOX9 were observed in 40 colon cancers (Fig. 3.13). As an atypical tumor suppressor, SOX9 has been shown to interact with nuclear β-catenin. Inactivation of SOX9 causes loss of inhibition of the oncogenic Wnt/β-catenin signaling pathway and is associated with patient survivals (Prévostel et al. 2016). Some novel TSGs harbor mutational hotspots. For instance, the N583fs frameshifting mutation in BMPR2 introduced premature stops of protein synthesis and was observed in nine stomach adenocarcinomas (Fig. 3.14). We searched the literature and found supporting evidence of the tumor suppressing functions of 22 (78.6%) novel TSGs. Many of these novel TSGs were also annotated as putative drivers by other computational methods (Bailey et al. 2018).

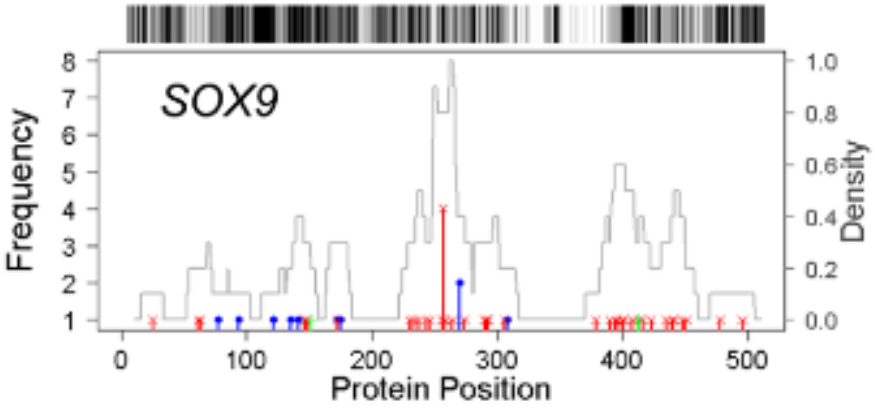

Figure 3.13. Positional distributions of somatic mutations in SOX9 in colon cancer. Evolutionary conservation of each position, measured as number of substitutions per billion years is displayed above each plot.

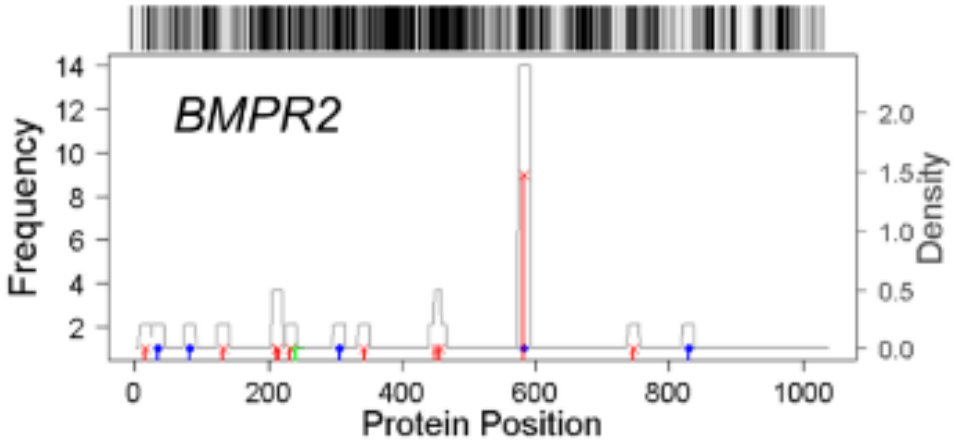

Figure 3.14. Positional distributions of somatic mutations in BMPR2 in stomach adenocarcinomas. Evolutionary conservation of each position, measured as number of substitutions per billion years is displayed above each plot.

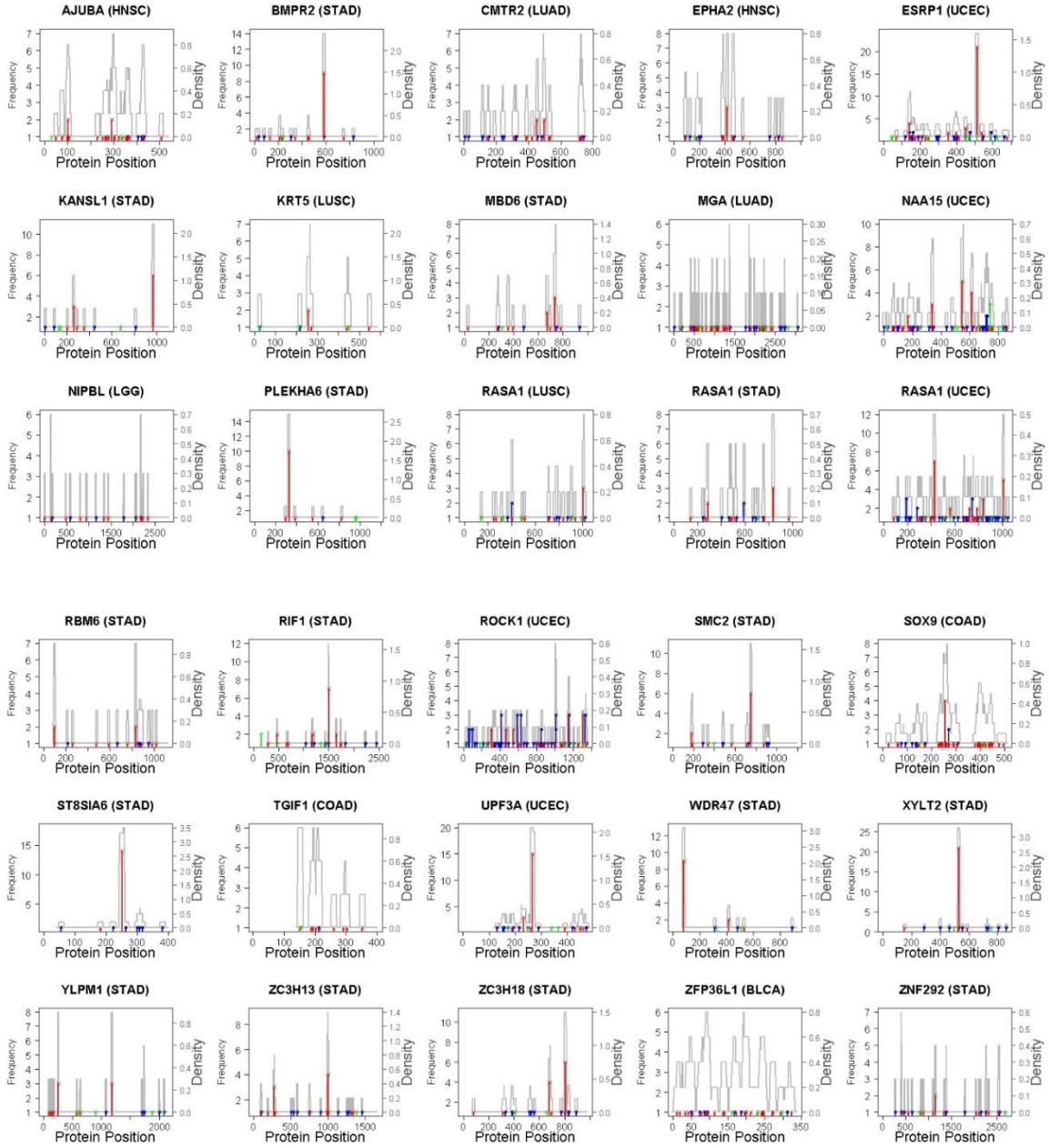

Figure 3.15. Mutational distribution of 28 novel TSGs.

3.4.3.3 Spectrum of Tissue Specificity

Even after removing low-confidence predictions, most of the drivers annotated by GUST were engaged in only one cancer type, showing high tissue specificities. Only 13 (59.1%) OGs and 25 (33.8%) TSGs in this high-confidence set are broad-spectrum drivers, promoting tumorigenesis in two or more cancer types (Fig. 3.16). The most prevalent OG was the PIK3CA gene found in 15 cancer types with high confidence, followed by the KRAS/NRAS/HRAS genes found in 13 cancer types. The most prevalent TSG was the TP53 gene found in 18 cancer types, followed by the ARID1A gene found in 10 cancer types.

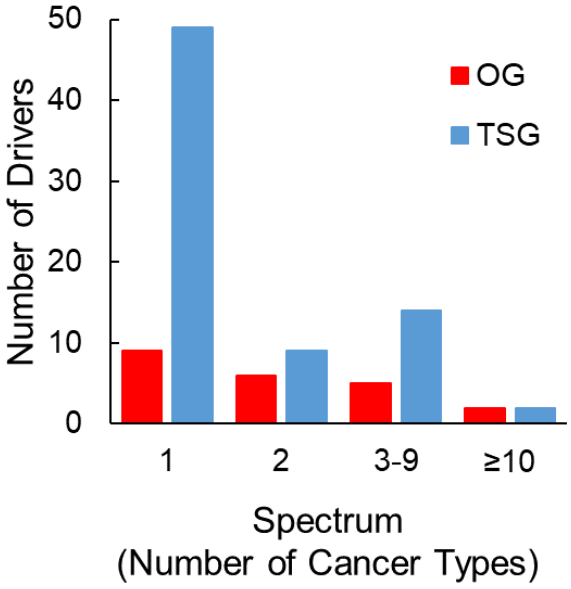

Figure 3.16. Distribution of driver genes with different spectrum of tissue specificity.

Furthermore, 11 out of the 13 broad-spectrum OGs possessed multiple hotspots (one-sided proportional test P <0.05 after Bonferroni corrections, Supplementary Figure 5, Supplementary Methods). For each significant hotspot, we examined the affected

functional domains as annotated in the NCBI Gene database. A representative example is the EGFR gene. In lung adenocarcinoma, 48% of missense mutations clustered at a single mutational hotspot affecting the tyrosine kinase activation loop (Fig. 3.17). In glioma, only one mutation hit this loop (chi-square test P < 10-18), and 69.3% of all missense mutations clustered at two hotspots affecting the extracellular domains independent of kinase activities. The contextual selection of mutations averting the kinase catalytic domain in glioma suggests an alternate path of activating EGFR signaling. In fact, several studies have reported the associations of these hotspot mutations with different levels of EGFR activities (Kamburov et al. 2015; Niu et al. 2016; Porta-Pardo et al. 2017).

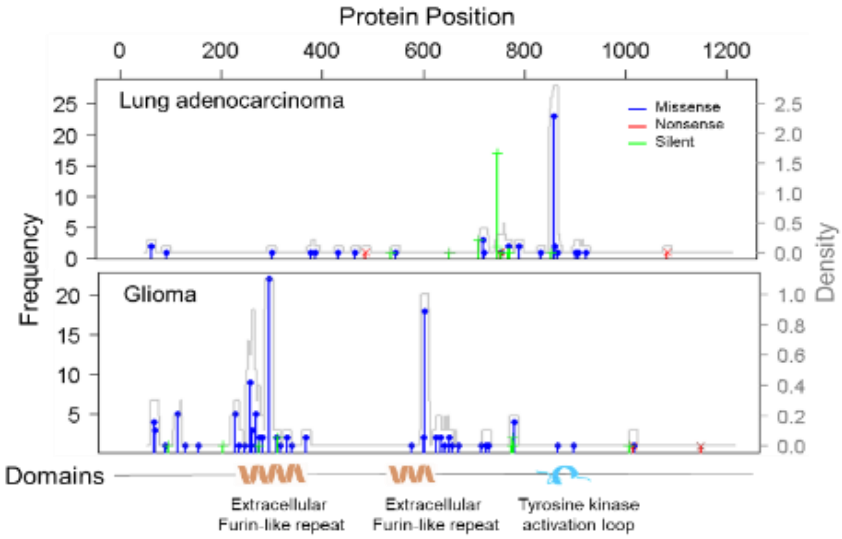

Figure 3.17. Positional distribution of mutations in the EGFR gene in lung adenocarcinoma and glioma (low-grade glioma and glioblastoma combined)

Interestingly, each of the 33 cancer types engaged at least one broad-spectrum driver and multiple tissue-specific drivers, implicating the synchrony of convergent and

divergent disease pathways. Clustering of cancers based on broad-spectrum driver genes grouped cancer types largely matching their tissue and cellular origins (Fig. 3.18).

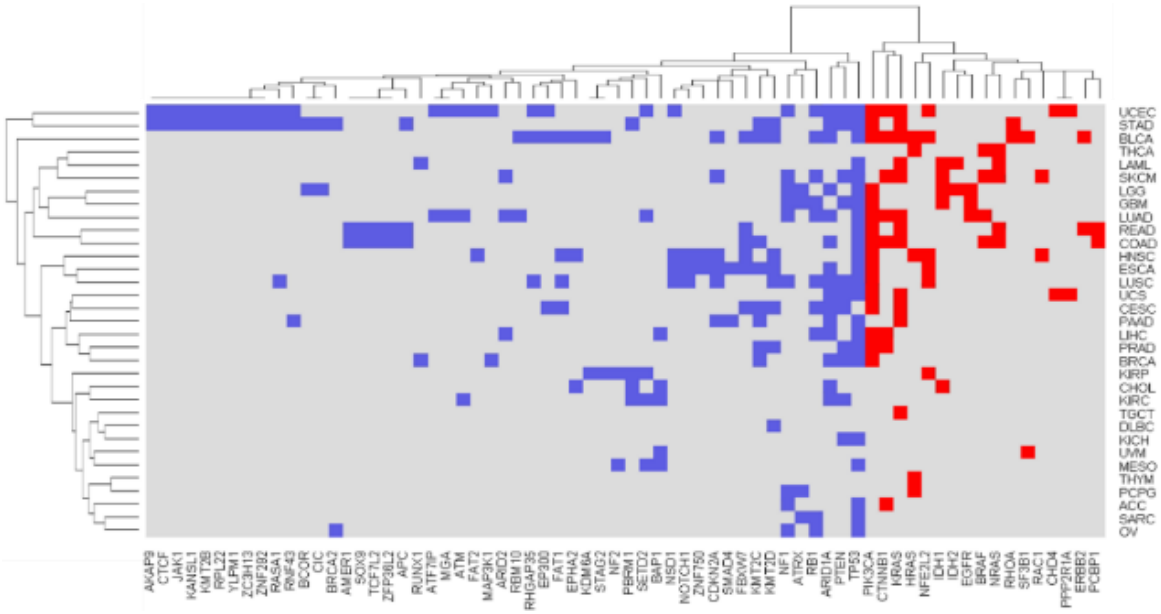

Figure 3.18. Two-way clustering of driver genes and cancer types. Driver genes found in more than one cancer type are used (OGs in red and TSGs in blue).

3.5 Discussion

Distinguishing OGs and TSGs in individual cancer types is critical to understanding cancer etiology and pinpointing clinically actionable targets. In this chapter, we proved that protein-coding mutations in OGs and TSGs are under different somatic selection, and subsequently discussed the GUST method to discover cancer-type specific functions of cancer driver genes. We compared GUST with the 20/20+ method, which is the only available method to classify OGs and TSGs. Both GUST and 20/20+ employ a random forest model to integrate features extracted from tumor exomes. Although GUST uses only 10 features compared to 24 features in 20/20+, the accuracy of

GUST is consistently higher. In the GUST model, selection measures contribute the most information content. In 20/20+, the p-value of enrichment of inactivating mutations is the most informative feature. Interestingly, this feature is also related to selection, although it is not a strict evolutionary measure (Kryazhimskiy and Plotkin 2008; Temko et al. 2018). These results suggest that using a small number of features engineered on evolutionary mechanisms is more powerful than feeding a large number of raw features to machine learning models. Furthermore, given the scarcity of known drivers for specific cancer types, reducing the number of features in predictive models helps mitigate overfitting problems.

We acknowledge that a driverMAPS (Zhao et al. 2019) method has been recently developed that estimates selection coefficients of a gene under three competing models (i.e., a PG, an OG and a TSG model). However, this method later combines the OG model and the TSG model into a driver model and contrasts it with the PG model to predict driver genes. Consequently, the reported posterior likelihood and false discovery rate are for the purpose of distinguishing drivers and passenger, but not OGs versus TSGs. Via personal communications with the authors of driverMAPS, we confirmed that this method does not provide statistical significance of OG and TSG classifications. Therefore, we did not compare GUST with driverMAPS.

While we discovered many known and novel cancer driver genes, none of them showed dual OG/TSG roles with high confidence in our analysis. A straightforward explanation is that GUST makes predictions based on protein altering substitutions and indels, thus it is unable to capture genes acting through other mechanisms, such as noncoding regulatory variants, copy number variants, translocations, fusions, differential

expressions, post-translational modifications and epigenetic regulations. Further investigations will shed light on key switches that divert paths of dual-role drivers. We also note that genes with only a small number of mutations may cause non-convergence problems during maximum likelihood estimations of selection coefficients, which limits the application of GUST to discovering rare drivers.

For practical use, we have built an online database (https://liliulab.shinyapps.io/gust) with precomputed results of analyzing TCGA samples. Users can query the database and visually inspect somatic selection patterns and conservational patterns of selected genes. Combined with information indicating whether or not a gene has been annotated by CGC as a driver or a drug target, users can make informed decisions on prioritizing candidate genes for further investigations. The R implementation of the GUST algorithm is available on Github (https://github.com/liliulab/gust).

CHAPTER 4

SUBCLONAL DISTRIBUTIONS OF CANCER DRIVERS & PROGNOSTIC VALUES

4.1 Introduction

In chapter 2, we discussed the clinical implication and importance of discovering clonal structure of the tumors. We developed MAGOS to discover this structure and applied it to data from 33 cancer types from TCGA. We found interesting associations between the number of subclones and the survival of patients, in LIHC and ACC and THYM. In chapter 3, we introduced GUST, a random forest method to predict cancer-specific driver genes. We argued that genes' roles in cancers are not universal and depends on the cancer type. GUST is able to predict each gene's role, specific to each cancer type. We applied GUST to TCGA data, and detected the driver genes for each sample (OG, TSG, PG). We considered the cluster with the highest VAF as the major clone and the remaining clusters as sub-clones.

In this chapter, we are focusing on combining the results of chapter 2 and chapter 3 by overlaying clonal information and the roles of the genes. In doing so, we will investigate whether subclonal drivers are effective for patient survival. The results of this analysis will help us understand the progression of the disease, may point to potential candidates for targeted therapies, and help with early disease detections by discovering potential drivers.

4.2 Combining MAGOS and GUST

MAGOS works with mutations from a tumor sample from an individual patient, whereas GUST utilizes all of the mutations from a specific cancer type as input and predicts the class of each gene for that cancer. In order to overlay the results, the clonality information for each specific mutation comes from MAGOS, and the gene type comes from the GUST output. For each patient, we will have the counts of clonal OGs, clonal TSGs, subclonal OGs and subclonal TSGs. Using the clinical data for each patient we can study the survival time of the patients and the clonal/driver features.

4.3 MAGOS + GUST Results

This section highlights the results found by analyzing the association between the survival and the clonal distribution of the driver genes on the 33 cancer types from TCGA. We discovered an association between the clonal distribution of the drivers and the survival of the patients, in ACC, READ, LIHC, OV and HNSC. Although, in the majority of the cancer types, we were not able to detect any significant association.

4.3.1 Adenoid Cystic Carcinoma (ACC)

Subclonal Tumor suppressor genes in Adenoid cystic carcinoma showed an association with increased survival of the patients. We saw that if a patient does not have any TSG mutations, they tend to survive longer (p-value= 0.05, Table 4.1, Fig. 4.1). However, most of this TSGs occurs in the subclone and there are only 4 samples that have TSG mutations in their clone (Fig 4.2). Patients with no TSG in their subclone tend to survive longer as well (p-value= 0.04, Fig 4.3).

Table 4.1. Clonal distribution of the driver genes detected in ACC. The gray cells correspond to the p-value of a test with: h0: having a *Driver (All/OG/TSG)* in *Clone(Either/Clone/Subclone)* is not associated with survival. The white cells correspond to the number of samples without any driver vs the number of samples with the driver.

	ALL	**Clonal**	**Subclonal**
OG+TSG	0.02	1	0.02
	30 vs 32	51 vs 11	40 vs 22
OG	0.2	0.9	0.01
	50 vs 12	54 vs 8	58 vs 4
TSG	0.05	0.9	0.04
	38 vs 24	58 vs 4	42 vs 20

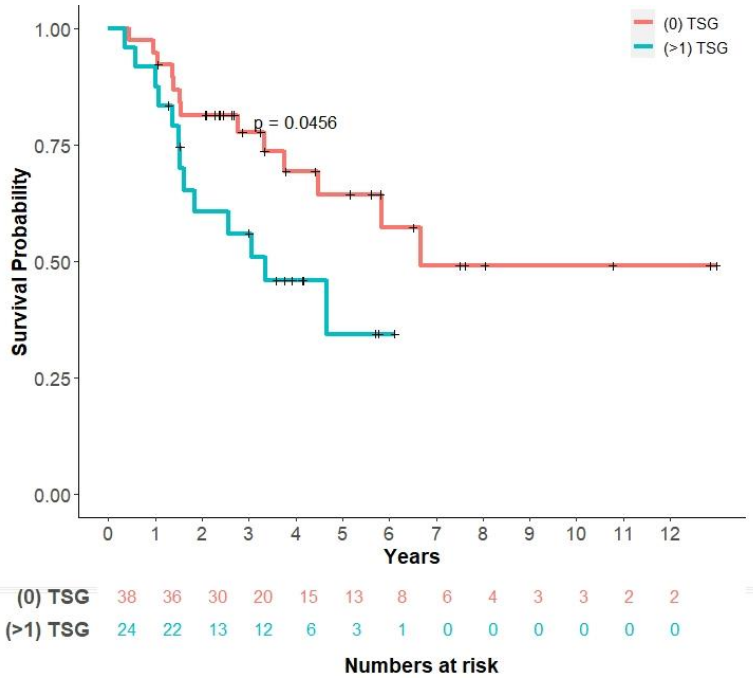

Figure 4.1. Kaplan Meier curve of the patients with no TSG in neither clone nor subclone vs patients with at least one TSG.

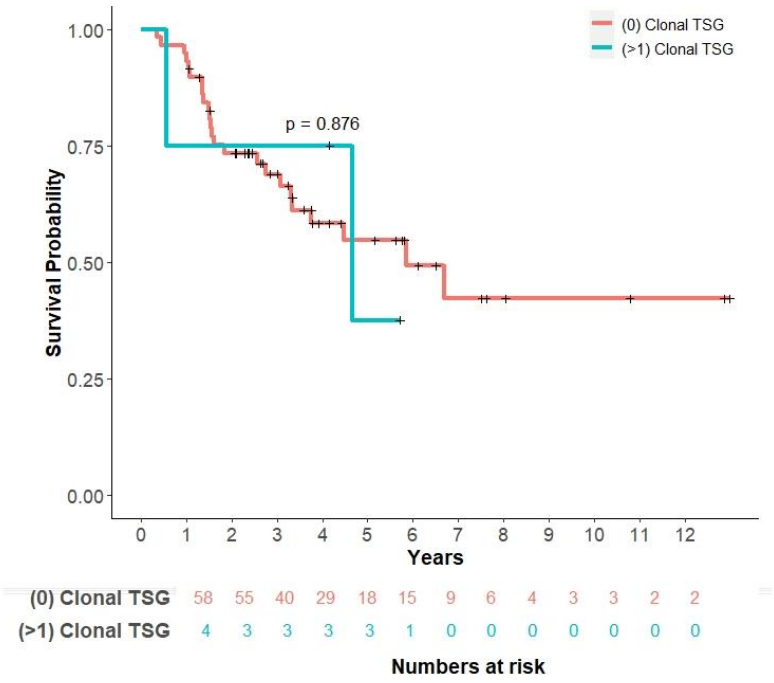

Figure 4.2. Kaplan Meier curve of the patients with no TSG in their clone vs patients with at least one clonal TSG.

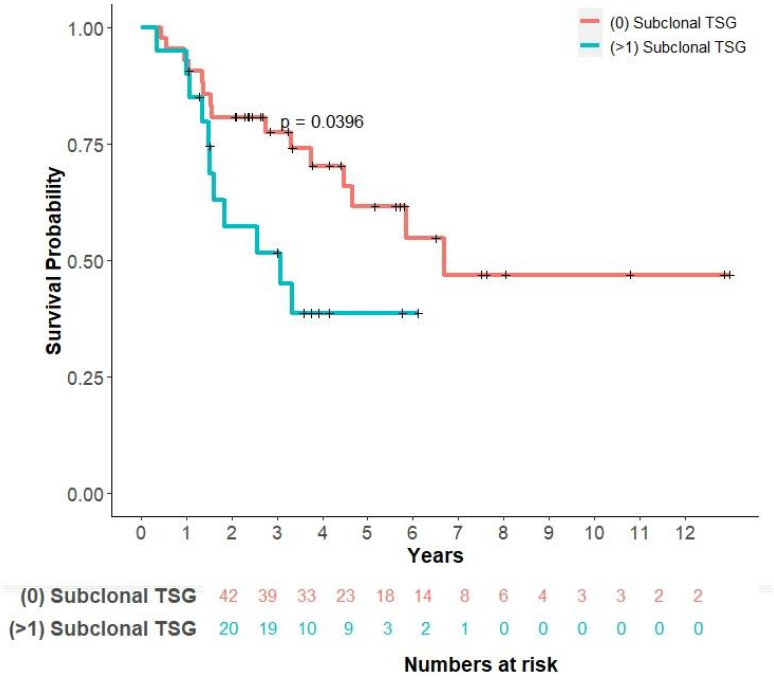

Figure 4.3. Kaplan Meier curve of the patients with no TSG in subclone vs patients with at least one subclonal TSG.

4.3.2 Liver Hepatocellular Carcinoma (LIHC)

Results for LIHC have already been reported in the MAGOS section. In patients with stage III of LIHC, we found significant association between survival of the patient and the detected number of subclones. In adding the GUST analysis results, we also found an association between the survival of the patient and the number of clonal/subclonal drivers across all of the samples. This was also observed in the data from stage III LIHC.

Across all of the LIHC samples, we discovered that the existence of subclonal driver mutations is associated with the survival (p-value=0.05), whereas clonal drivers do not have the same effect on survival (p-value=0.6) (Table 4.2, Fig. 4.4, Fig. 4.5 and Fig. 4.6). Interestingly, not having a driver gene in either the clone or subclone will not increase the survival rate of patients.

Table 4.2. Clonal distribution of the driver genes detected in LIHC. The gray cells correspond to the p-value of a test with: h0: having a *Driver (All/OG/TSG)* in *Clone (Either/Clone/Subclone)* is not associated with survival. The white cells correspond to the number of samples without any driver vs number of samples with the driver.

	ALL	**Clonal**	**Subclonal**
OG+TSG	0.3	**0.6**	**0.05**
	36 vs 324	135 vs 225	107 vs 253
OG	0.6	**0.3**	**0.06**
	235 vs 125	297 vs 63	289 vs 71
TSG	0.3	0.5	0.2
	52 vs 309	159 vs 202	122 vs 238

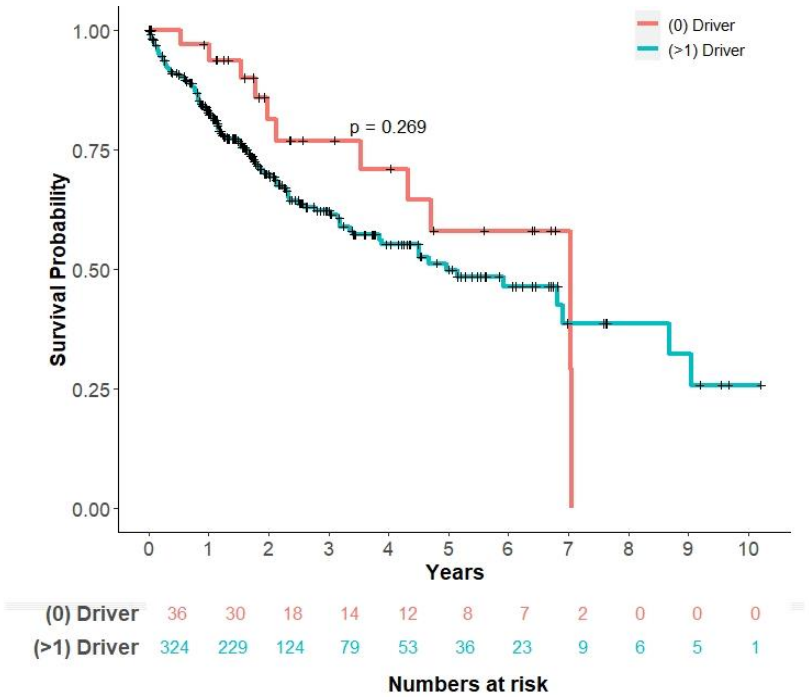

Figure 4.4. Kaplan Meier curve of the patients with no Driver vs patients with at least one Driver in either clone or subclone.

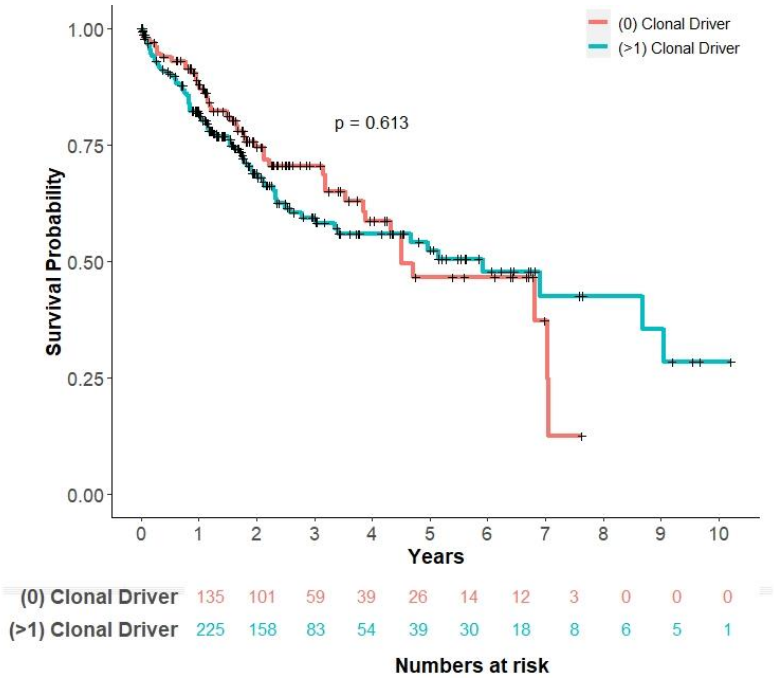

Figure 4.5. Kaplan Meier curve of the patients with no driver in clone vs patients with at least one clonal driver.

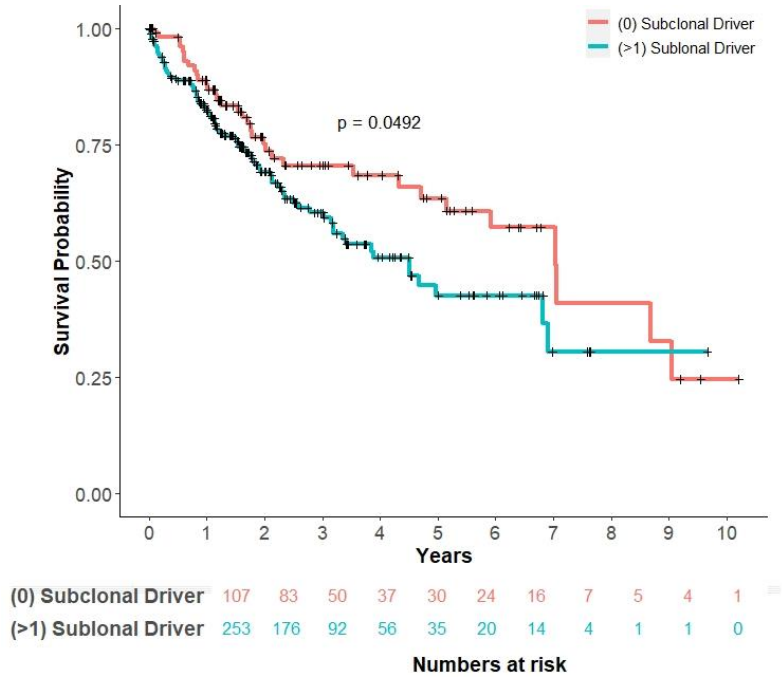

Figure 4.6. Kaplan Meier curve of the patients with no subclonal driver vs patients with at least one subclonal driver.

We used Cox proportional hazard regression to test if the number of clonal or subclonal drivers in a tumor is associated with patient overall survival. We included age at diagnosis, tumor stages and number of detected clusters as covariates. We found significant association between number of subclonal oncogenes and the survival (p-value=0.039, HR=1.38). For comparison, age at diagnosis is not a significant prognostic factor among these tumors (p-value=0.31). Among all LIHC tumors, the number of subclonal oncogenes is a novel prognostic factor for liver cancers that is independent of age at diagnosis, and total number of clusters.

We also analyzed Stage III LIHC and found interesting results. We discovered that in stage III liver cancer, subclonal OGs have the most significant prognosis power

(p-value=0.01). Existence of Subclonal OG are the most significant feature in predicting survival, even more than age at diagnosis (Table 4.3, Fig. 4.7, Fig. 4.8, Fig. 4.9).

Table 4.3. Clonal distribution of the driver genes detected in LIHC Stage III. The gray cells correspond to the p-value of a test with: h0: having a *Driver (All/OG/TSG)* in *Clone (Either/Clone/Subclone)* is not associated with survival. The white cells correspond to the number of samples without any driver vs number of samples with the driver.

	ALL	**Clonal**	**Subclonal**
OG+TSG	0.3	0.9	0.2
	7 vs 76	32 vs 51	22 vs 61
OG	**0.07**	0.7	**0.01**
	51 vs 32	65 vs 18	68 vs 15
TSG	0.6	0.8	0.5
	12 vs 71	41 vs 42	25 vs 58

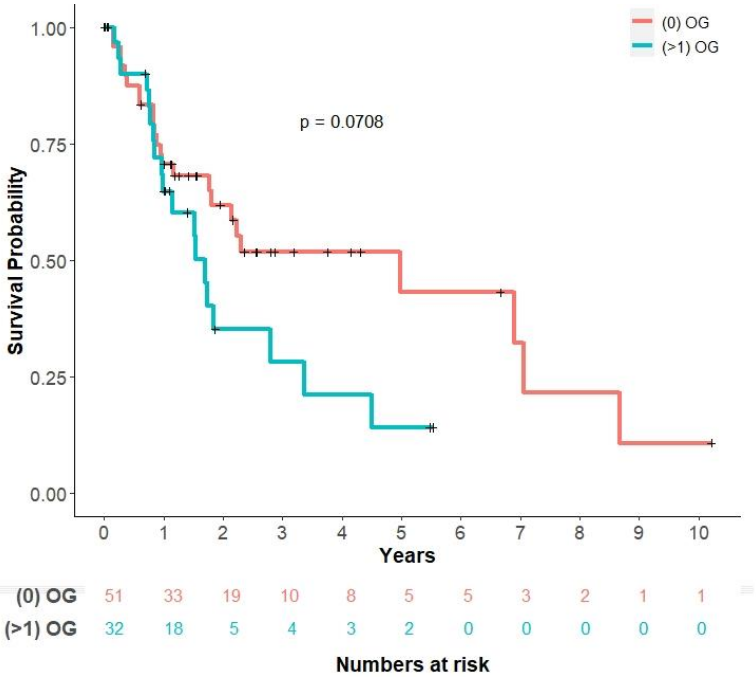

Figure 4.7. Kaplan Meier curve of the patients with no OG vs patients with at least one OG.

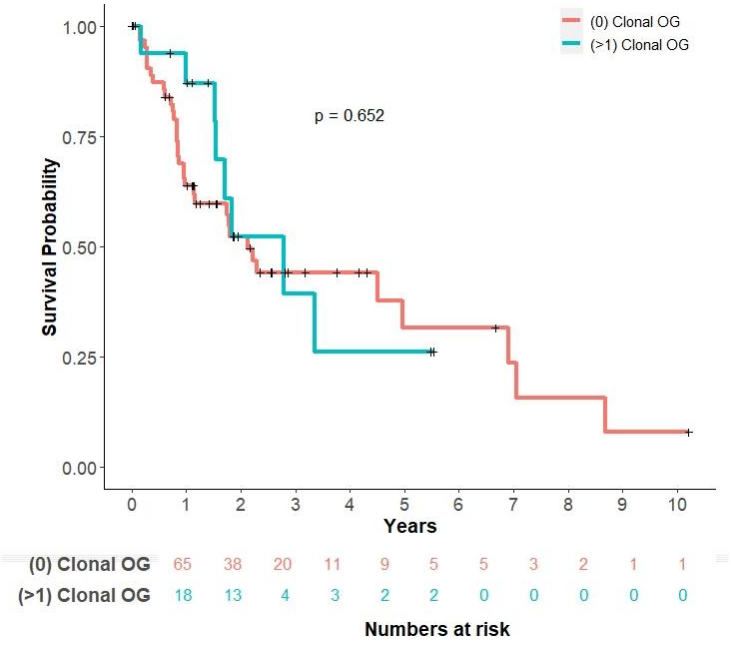

Figure 4.8. Kaplan Meier curve of the patients with no clonal OG vs patients with at least one clonal OG.

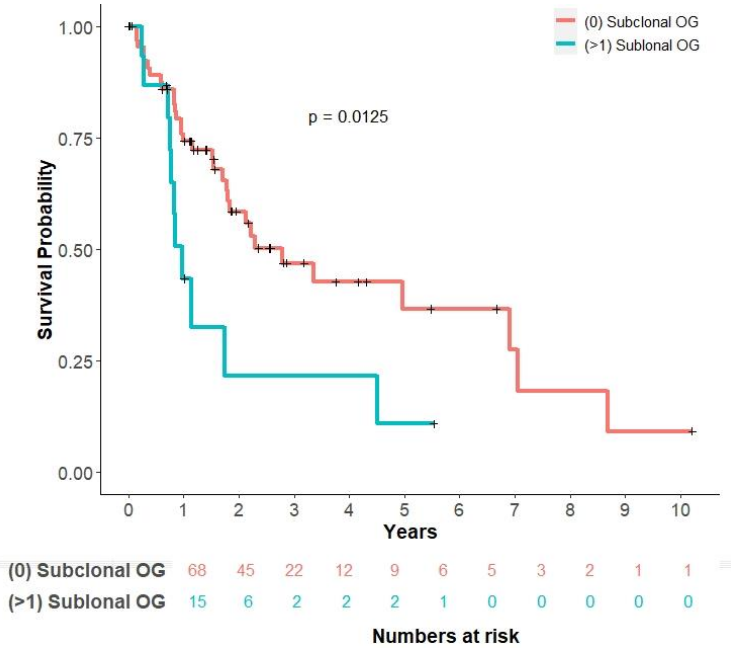

Figure 4.9. Kaplan Meier curve of the patients with no subclonal OG vs patients with at least one subclonal OG.

Using Cox proportional hazard regression, we tested if the number of clonal or subclonal drivers in a tumor is associated with patient overall survival in stage III LIHC, We included age at diagnosis, and number of detected clusters as covariates. In analyzing all of the LIHC tumors across all tumor stages, number of clusters, do not have any associations. Number of subclonal oncogenes also is not associated with survival as opposed to all LIHC data.

4.3.3 Head and Neck Squamous Cell Carcinoma (HNSC)

In head and neck squamous cell carcinoma, we detected associations between the lack of subclonal drivers and the survival of the patients. By looking closely, we discovered that this association is mainly from subclonal TSGs (p-value=0.03) rather than the OGs. Because subclonal OGs do not have association with survival (p-value=0.4, Table 4.4, Fig.4.10, Fig. 4.11, Fig.4.12, Fig.4.13).

Table 4.4. Clonal distribution of the driver genes detected in HNSC. The gray cells correspond to the p-value of a test with: h0: having a *Driver (All/OG/TSG)* in *Clone (Either/Clone/Subclone)* is not associated with survival. The white cells correspond to the number of samples without any driver vs number of samples with the driver.

	ALL	**Clonal**	**Subclonal**
OG+TSG	1	**0.5**	**0.05**
	6 vs 488	88 vs 405	46 v 447
OG	0.6	0.1	0.4
	321 vs 173	435 vs 59	369 vs 125
TSG	0.5	**0.1**	**0.03**
	8 v 485	99 v 394	51 vs 442

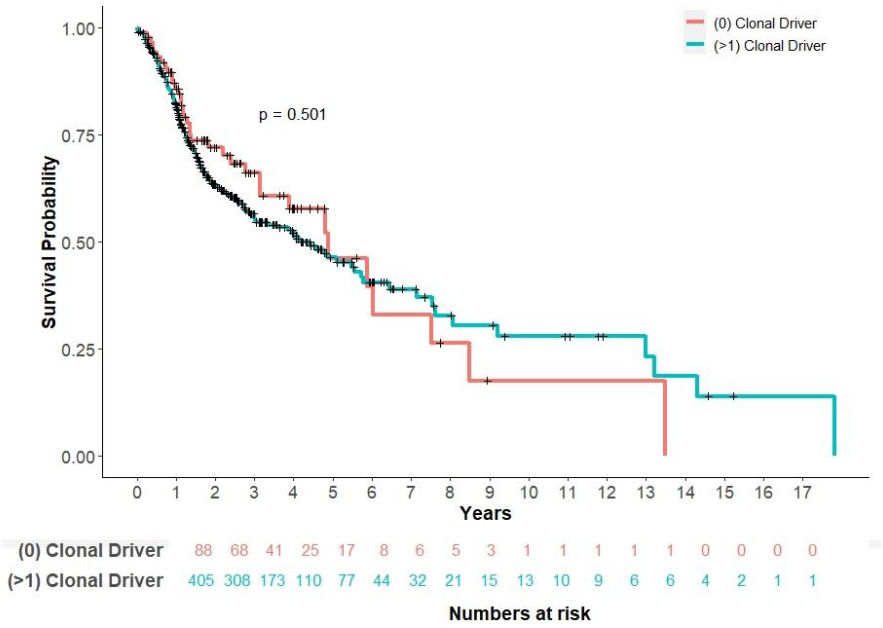

Figure 4.10. Kaplan Meier curve of the patients with no clonal driver vs patients with at least one clonal driver.

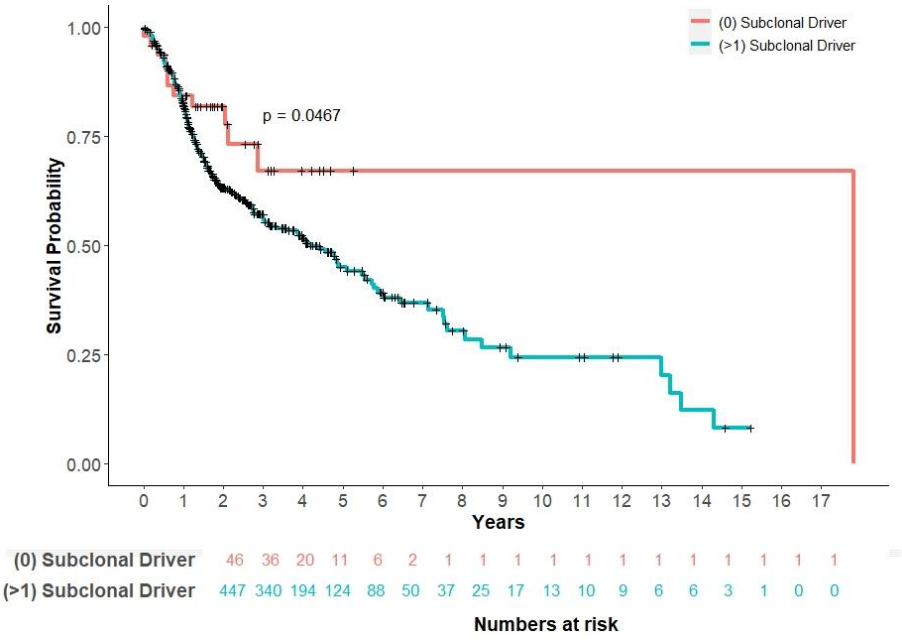

Figure 4.11. Kaplan Meier curve of the patients with no subclonal driver vs patients with at least one subclonal driver.

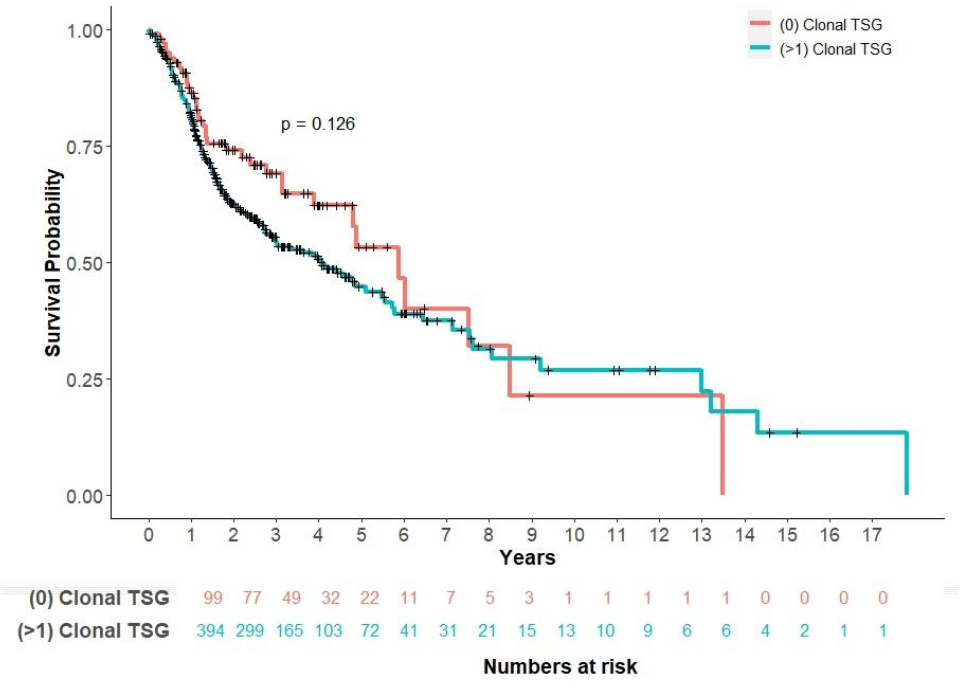

Figure 4.12. Kaplan Meier curve of the patients with no clonal TSG vs patients with at least one clonal TSG.

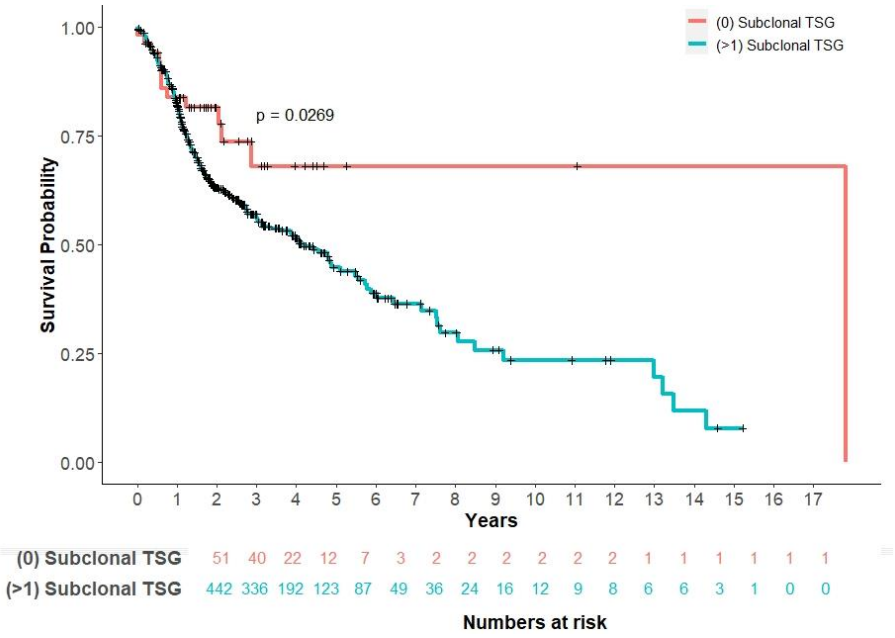

Figure 4.13. Kaplan Meier curve of the patients with no subclonal TSG vs patients with at least one subclonal TSG.

Using Cox proportional hazard regression, we did not find any associations between the clonal/subclonal drivers and the survival of the patients.

4.3.4 Low Grade Glioma (LGG)

In Low Grade Glioma, OGs seem to have prognostic value. First, we discovered that having a clonal driver (p-value=0.0002, table 4.5) effects the rate of patient survival. Not having any drivers is significant in predicting survival. However, looking further at the clonal distribution of the drivers, the clonal drivers are more important than the subclonal drivers (table 4.5). In figures 4.14-4.19 we can see the effect of clonal drivers and the OG, specifically.

Table 4.5. Clonal distribution of the driver genes detected in LGG. The gray cells correspond to the p-value of a test with: h0: having a *Driver (All/OG/TSG)* in *Clone (Either/Clone/Subclone)* is not associated with survival. The white cells correspond to the number of samples without any driver vs number of samples with the driver.

	ALL	**Clonal**	**Subclonal**
OG+TSG	0.01	**0.0002**	**0.1**
	14 vs 494	85 vs 423	147 vs 361
OG	**0.00001**	**0.00001**	0.07
	54 vs 454	178 vs 330	365 vs 143
TSG	0.5	0.6	0.6
	63 vs 447	206 vs 304	191 vs 319

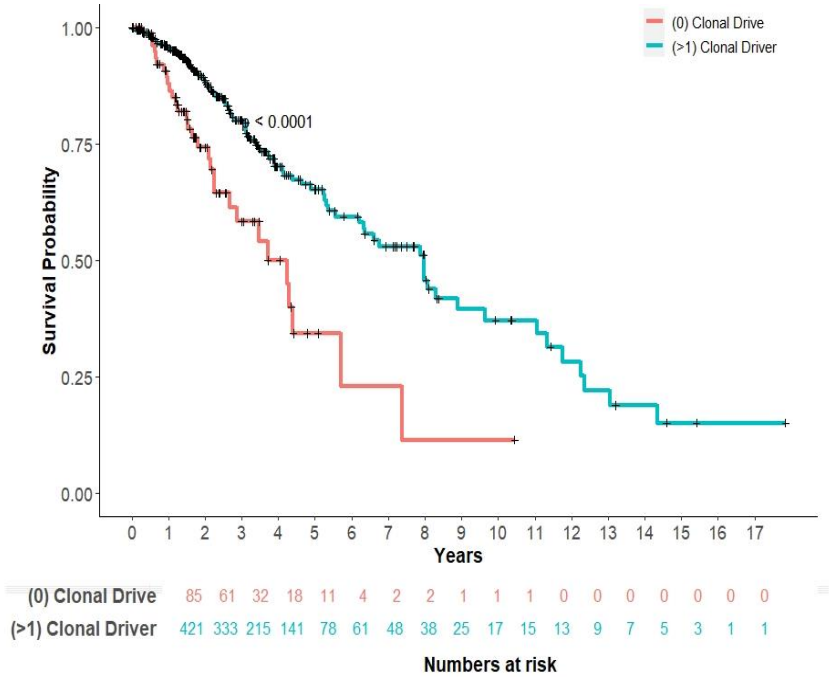

Figure 4.14. Kaplan Meier curve of the patients with no clonal driver vs patients with at least one clonal driver.

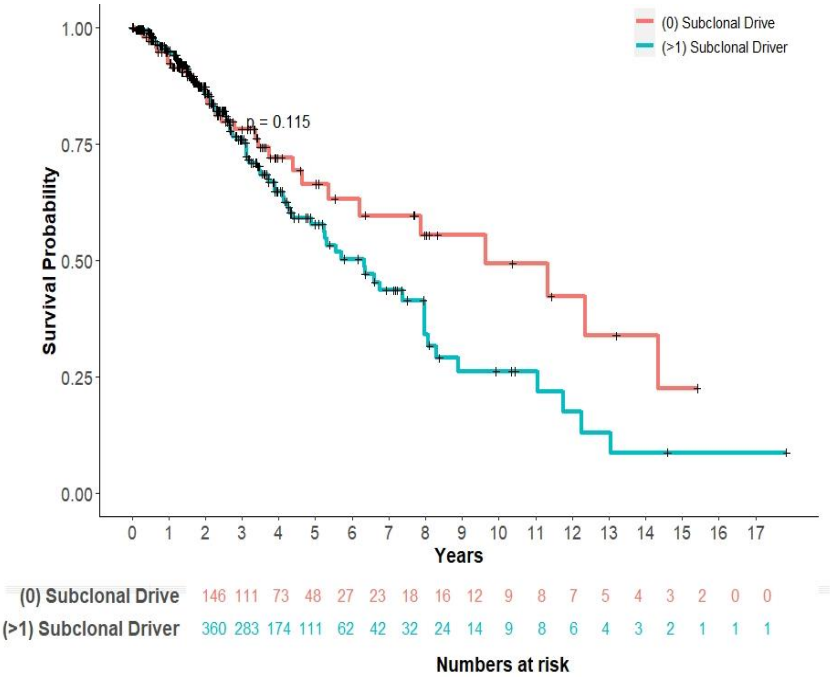

Figure 4.15. Kaplan Meier curve of the patients with no subclonal driver vs patients with at least one subclonal driver.

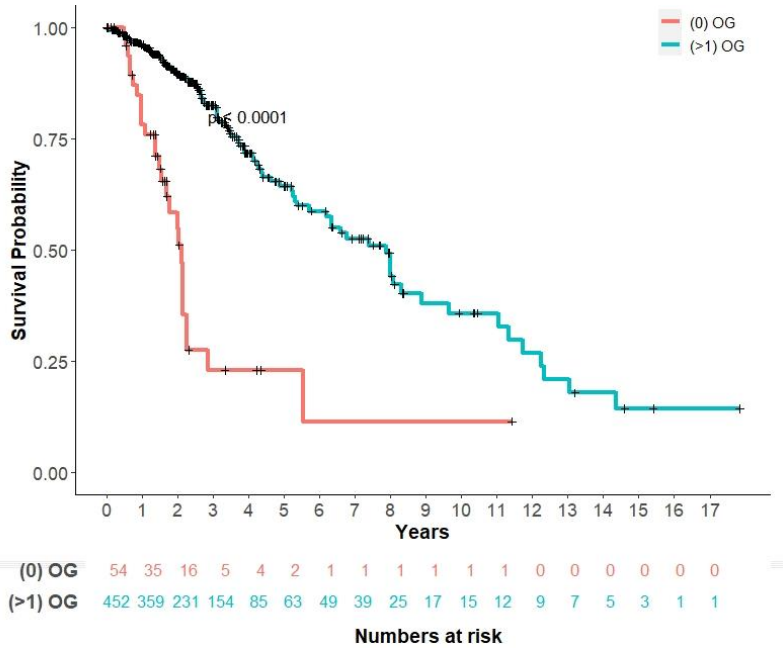

Figure 4.16. Kaplan Meier curve of the patients with no OG vs patients with at least one OG.

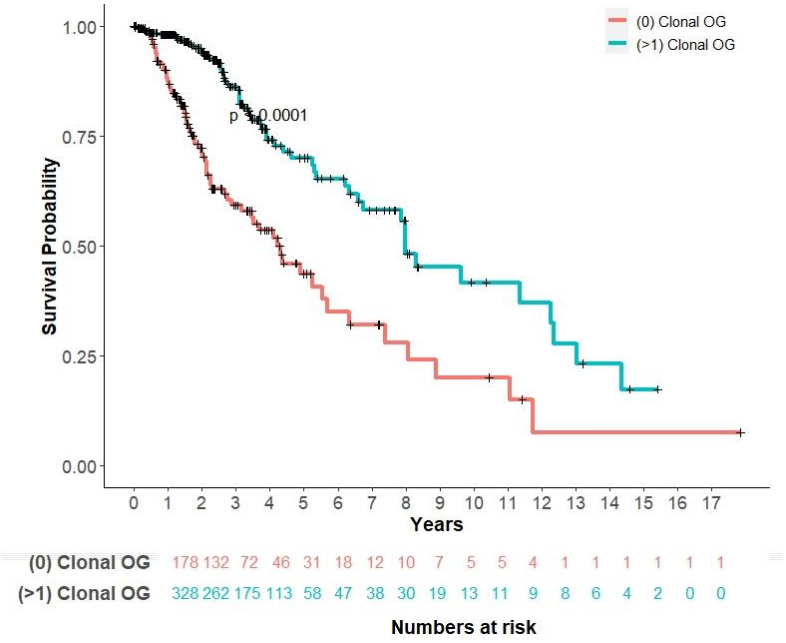

Figure 4.17. Kaplan Meier curve of the patients with no clonal OG vs patients with at least one clonal OG.

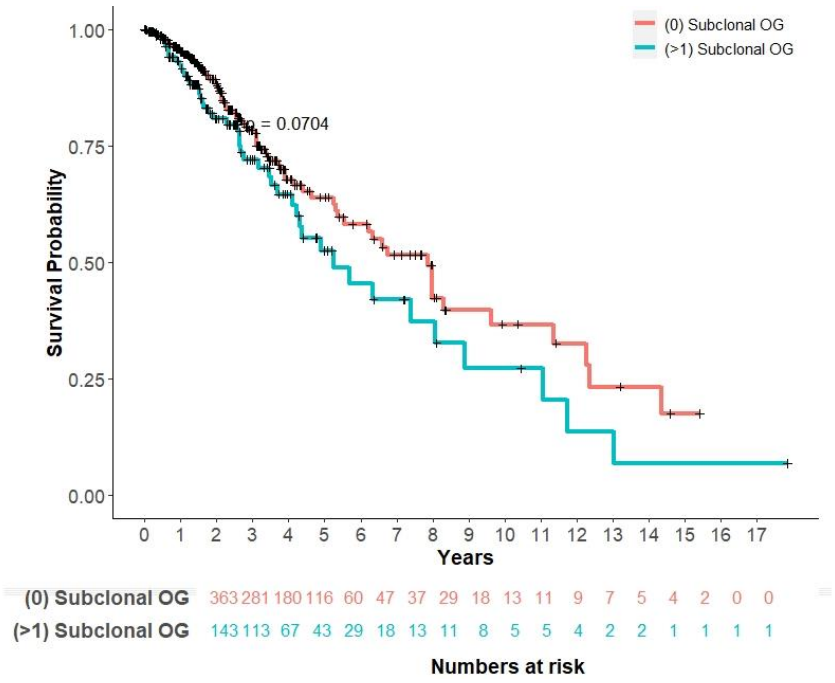

Figure 4.18. Kaplan Meier curve of the patients with no subclonal OG vs patients with at least one subclonal OG.

We did not have tumor stage data reported for the LGG patients. We used Cox proportional hazard regression to test if there is any association between the clonal/subclonal drives, number of clusters, and other covariates and the survival of the patients. We found that using age at diagnosis, clonal/subclonal drivers, and the number of clusters as covariates, age at diagnosis (p-value<0.0001, HR=1), number of clonal oncogenes (p-value<0.0001, HR=0.319) and number of subclonal oncogenes (p-value<0.028, HR=0.59) have significant association with survival. Although, the association between the number of clonal and subclonal oncogenes and the survival is negative. In other words, the patients with more clonal/subclonal oncogenes live longer than the patients with fewer oncogenes.

4.3.5 Rectum Adenocarcinoma (READ)

In the rectum adenocarcinoma, we observed an association between the distribution of oncogenes and the survival rate of patients. We observed a strong association between subclonal drivers and survival, though, there were not many samples that did not have any subclonal drivers (table 4.6). Interestingly, the clonal distribution of OGs have a strong association with survival. In figures 4.19-4.21 we can see that the samples with subclonal OG have higher a survival chance vs patients with subclonal OGs. This example is one case that demonstrates the power of the results that can be obtained from analyzing the results of MAGOS and GUST. We can see that the existence of clonal OGs are not associated with survival whereas subclonal OG is strongly associated with the survival of the patient.

Table 4.6. Clonal distribution of the driver genes detected in READ. The gray cells correspond to the p-value of a test with: h0: having a *Driver (All/OG/TSG)* in *Clone (Either/Clone/Subclone)* is not associated with survival. The white cells correspond to the number of samples without any driver vs number of samples with the driver.

	ALL	**Clonal**	**Subclonal**
OG+TSG	0.9	0.4	0.005
	1 vs 132	26 vs 107	8 vs. 125
OG	0.6	**0.3**	**0.05**
	52 vs 81	90 v 43	87 vs 46
TSG	0.9	1	0.003
	1 vs 132	39 vs 94	14 vs 119

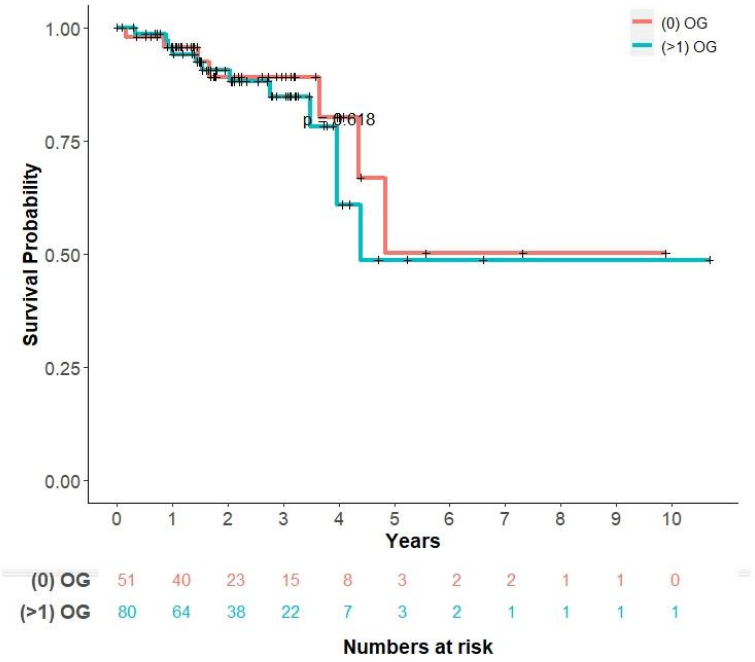

Figure 4.19. Kaplan Meier curve of the patients with no OG vs patients with at least one OG.

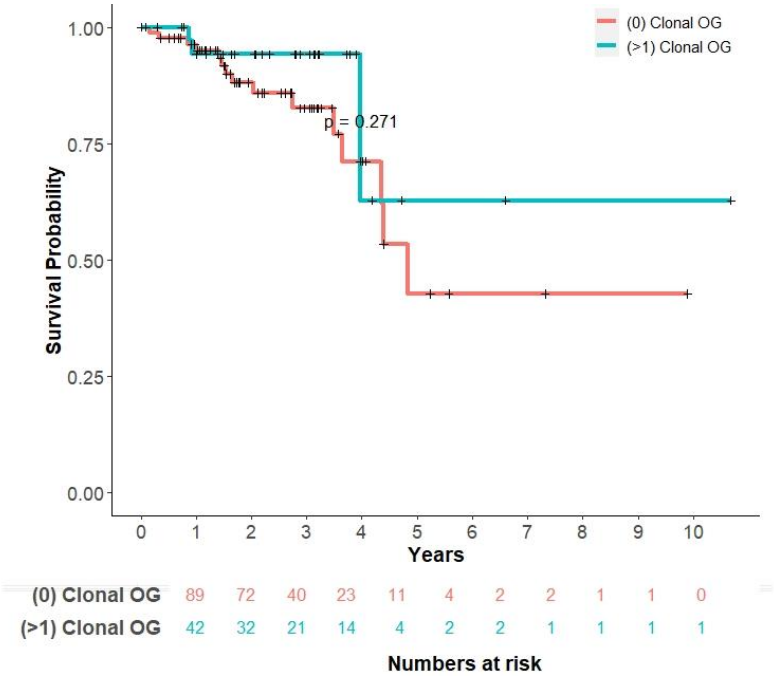

Figure 4.20. Kaplan Meier curve of the patients with no clonal OG vs patients with at least one clonal OG.

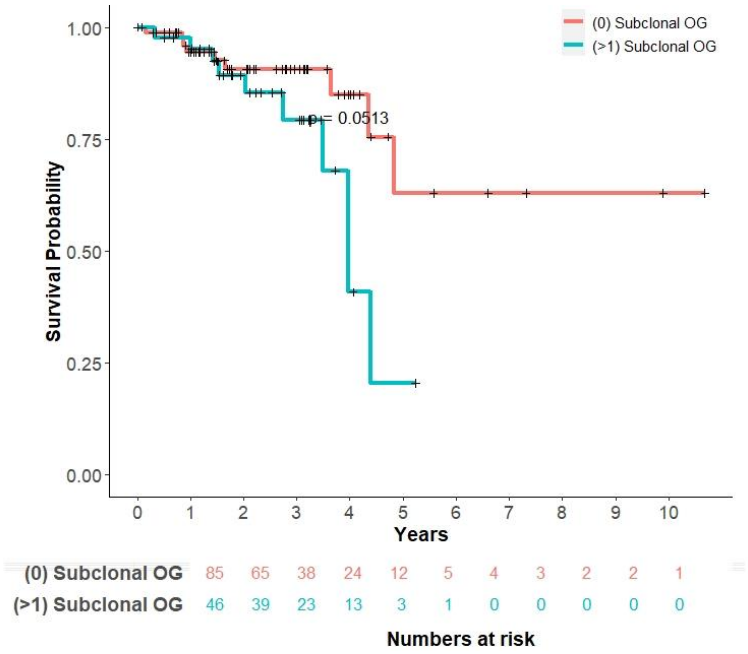

Figure 4.21. Kaplan Meier curve of the patients with no subclonal OG vs patients with at least one subclonal OG.

Using Cox proportional hazard regression, we tested if the number of clonal or subclonal drivers in a tumor is associated with patient overall survival in READ, We included age at diagnosis, and number of detected clusters and tumor stage as covariates. Number of subclonal tumor suppressor genes (p-value=0.032, HR=0.45) and age at diagnosis (p-value=0.03, HR=1) were found to be significantly associated with the survival. The number of subclonal tumor suppressor genes have a negative association with the survival. The patients with more subclonal TSGs tend to live longer.

4.3.6 Ovarian Cancer (OV)

In Ovarian cancer samples, subclonal drivers have the most impact on survival. However, interestingly in OV, the association has a negative effect. This means that if the patient has subclonal driver, they tend to live longer than if they do not have subclonal drivers (table 4.7, Fig.4.22, Fig 4.23, and Fig.4.24).

Table 4.7. Clonal distribution of the driver genes detected in OV. The gray cells correspond to the p-value of a test with: h0: having a *Driver (All/OG/TSG)* in *Clone (Either/Clone/Subclone)* is not associated with survival. The white cells correspond to the number of samples without any driver vs number of samples with the driver.

	ALL	**Clonal**	**Subclonal**
OG+TSG	0.5	**0.3**	**0.03**
	6 vs 270	115 v 160	23 v 252
OG	0.07	0.4	0.03
	234 v 41	270 v 5	239 v 36
TSG	0.4	**0.3**	**0.02**
	7 v268	117 v 158	24 v 251

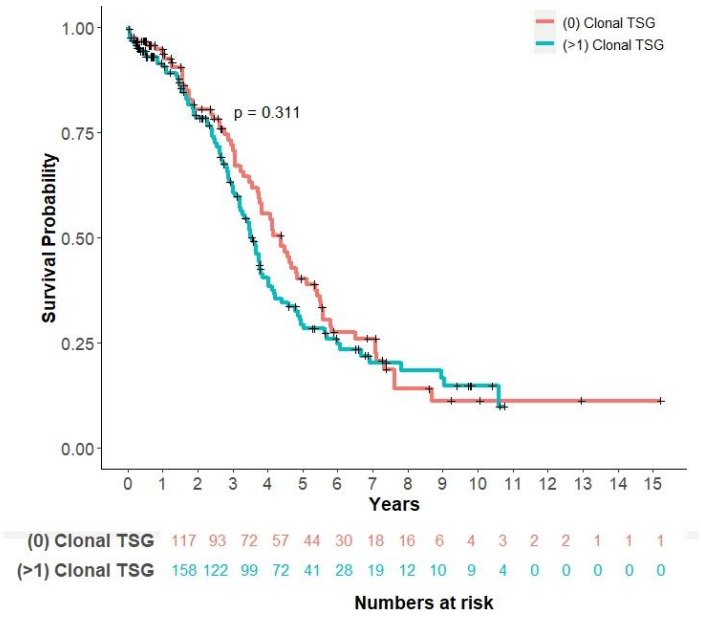

Figure 4.22. Kaplan Meier curve of the patients with no clonal TSG vs patients with at least one clonal TSG.

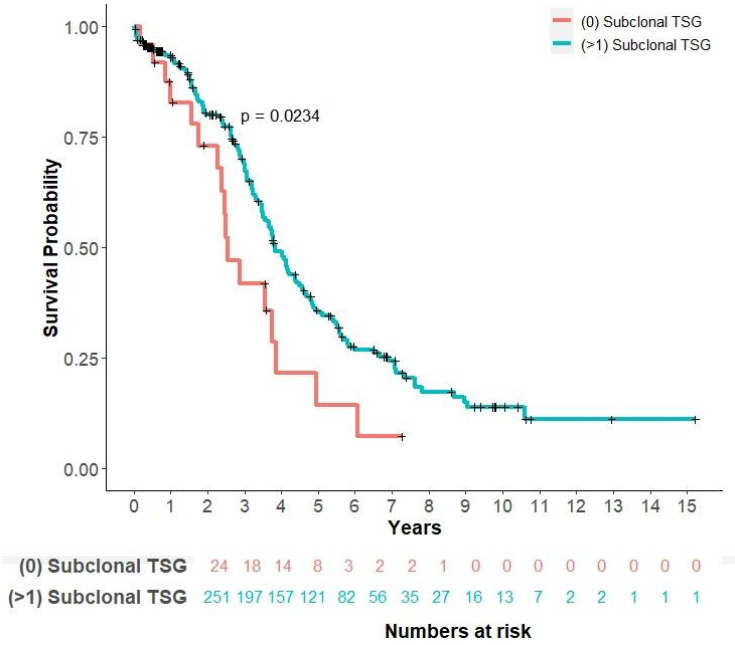

Figure 4.23. Kaplan Meier curve of the patients with no subclonal TSG vs patients with at least one subclonal TSG.

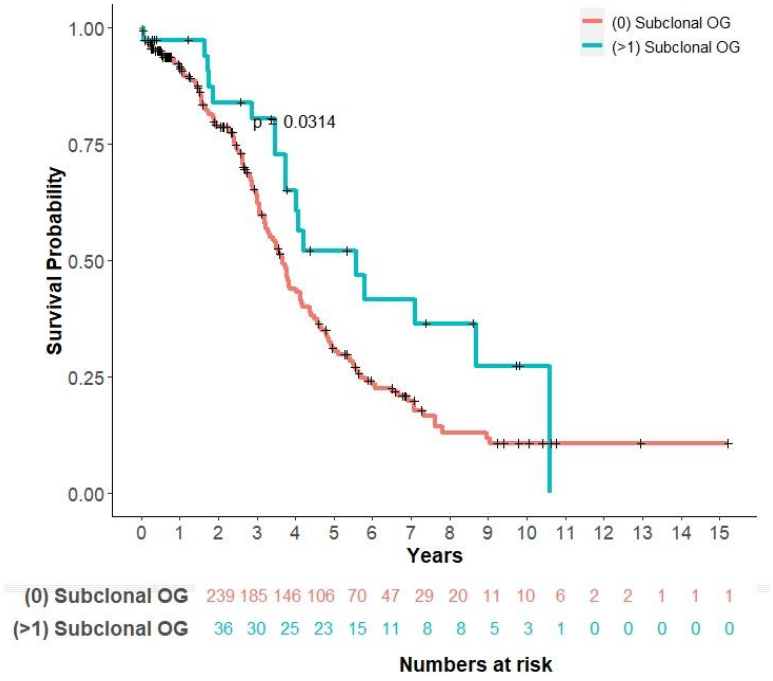

Figure 4.24. Kaplan Meier curve of the patients with no subclonal OG vs patients with at least one clonal OG.

We did not have tumor stage data reported for the OV patients. We used Cox proportional hazard regression to test if there is any association between the clonal/subclonal drives, number of clusters, and other covariates and the survival of the patients. We found that using age at diagnosis, clonal/subclonal drivers, and the number of clusters as covariates, age at diagnosis (p-value=0.0012, HR=1) and number of clusters (p-value=0.03, HR=0.83) have significant association with survival.

4.3.7 Pan Cancer Analysis

After looking at each cancer separately, we conducted a pan cancer analysis. We tested whether across all of the cancers, if there are any features with prognosis value. We performed Cox proportional regression. Including number of clonal and subclonal drivers, age at diagnosis, tumor stage, and number of clusters and the expected survival of each cancer type as covariates in the model. We found that age at diagnosis, number of clonal oncogenes, number of clonal tumor suppressor genes, number of detected clones, number of mutations and tumor stage are significantly associated with the survival of the patient (table 4.8).

Table 4.8. Cox regression results.

	Coefficient	exp(Coef)	P-Value
Age at Diagnosis	**<0.001**	1	**<0.001**
Number of Clonal OGs	**-0.107**	0.898	**0.025**
Number of Subclonal OGs	0.024	1.024	0.502
Number of Clonal TSGs	**-0.037**	0.964	**0.009**
Number of Subclonal TSGs	-0.01	0.99	0.117
Number of Clusters	**0.084**	1.088	**0.004**
Number of Mutations	<0.001	0.999	0.08
Expected Survival	**-0.032**	0.969	**<0.001**
Number of Clonal Passengers	**0.003**	1.003	**<0.001**
Number of Subclonal Passengers	0.001	1.001	0.14
Tumor Stage II	**0.594**	1.811	**<0.001**
Tumor Stage III	**0.952**	2.59	**<0.001**
Tumor Stage IV	**1.347**	3.848	**<0.001**

The number of oncogenes and tumor suppressor genes found to be significantly associated with survival, but interestingly this association is reversed. In other words, more oncogenes or tumor suppressor genes in the clone results in longer survival. But for number of clusters this association is direct; more subclones, result in shorter survival time.

4.4 Discussion

Previous sections in this chapter discussed the clinical implication and importance of discovering the clonal structure of the tumor and discussed how MAGOS can be used to discover this structure. Using MAGOS, we grouped the mutations from 33 cancer types' samples from TCGA. We found significant associations between the number of subclones and the survival of the patients in LIHC, ACC and THYM. Next, we used GUST to predict cancer-specific driver genes. We applied GUST to TCGA data, and classified each mutated gene in each sample (OG, TSG, and PG). We considered the cluster with the highest VAF as the major clone and the remaining clusters as sub-clones. In this chapter, we combined the results of MAGOS and GUST and discovered associations between the clonal distribution of drives and the survival of each patient. By overlaying clonal information and the roles of the genes, we were able to identify important features with prognosis power that were significant in predicting the survival of the patient. In the pan cancer analysis, we discovered that the clonal drivers show a protective effect on survival. The patients with more clonal oncogenes or tumor suppressor genes survive longer vs the patients with fewer clonal drivers. On the other hand the number of subclones detected by MAGOS, has a negative effect on the survival. The patients with more subclones have shorter life span.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this study, we attempted to better understand the progression of cancer and tried to find meaningful information from the available data. The aim was to study intratumor heterogeniety in cancer because it would provide us with valuable insights on how we can design treatment approaches to control the progression of the disease. We developed MAGOS to help us deconvolute the observed frequency of mutations in sequenced tumor samples. By applying MAGOS to TCGA data, we found that the number of subclones that could be detected by MAGOS have significant prognosis power in some cancer types (LIHC stage III).

In the next phase, we used GUST to predict the role of each gene in each cancer. Using TCGA data, we detected the number of driver genes in each sample of each cancer. The results can help us make hypotheses about the ways that each cancer may behave and progress. The next level of analysis was to merge the results of both studies. We merged our findings from both studies in order to look at the clonal distribution of drivers across different cancer types. Interestingly, the results showed that the total number of drivers in a sample may not be associated with the survival of the patient, but the clonal distribution of that specific driver may have strong associations. Our study showed that, in regards to LIHC, the lack of drivers is not associated with survival, but the lack of subclonal drivers can extend the survival of the patient.

The results we found are promising, but there is room to improve. There are several possible areas for development. The next step would be to develop methods to construct the clonal phylogeny, which by itself can provide valuable insights. Other

possible directions for development may include other covariates to improve prognostic values of our approach, and studying the disease mechanisms that we observed in order to discover the reasons that some tumors have positive correlations between clone counts with patient survival.

REFERENCES

Aktipis, C. Athena, Virginia S.Y. Kwan, Kathryn A. Johnson, Steven L. Neuberg, and Carlo C. Maley. 2011. "Overlooking Evolution: A Systematic Analysis of Cancer Relapse and Therapeutic Resistance Research." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0026100.

Andor, Noemi, Julie V. Harness, Sabine Müller, Hans W. Mewes, and Claudia Petritsch. 2014. "Expands: Expanding Ploidy and Allele Frequency on Nested Subpopulations." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt622.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. "Comprehensive Characterization of Cancer Driver Genes and Mutations." *Cell*. https://doi.org/10.1016/j.cell.2018.02.060.

Bignell, Graham R., Chris D. Greenman, Helen Davies, Adam P. Butler, Sarah Edkins, Jenny M. Andrews, Gemma Buck, et al. 2010. "Signatures of Mutation and Selection in the Cancer Genome." *Nature*. https://doi.org/10.1038/nature08768.

Bishop, Christopher M. 2006. "Machine Learning and Pattern Recogniton." In *Information Science and Statistics*.

Buisson, Rémi, Adam Langenbucher, Danae Bowen, Eugene E. Kwan, Cyril H. Benes, Lee Zou, and Michael S. Lawrence. 2019. "Passenger Hotspot Mutations in Cancer Driven by APOBEC3A and Mesoscale Genomic Features." *Science (New York, N.Y.)*. https://doi.org/10.1126/science.aaw2872.

Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." *Nature Biotechnology*. https://doi.org/10.1038/nbt.2514.

Croce, Carlo M. 2008. "Oncogenes and Cancer." *New England Journal of Medicine*. https://doi.org/10.1056/NEJMra072367.

Dagogo-Jack, Ibiayi, and Alice T. Shaw. 2018. "Tumour Heterogeneity and Resistance to Cancer Therapies." *Nature Reviews Clinical Oncology*. https://doi.org/10.1038/nrclinonc.2017.166.

Davidson, Nancy E., Scott A. Armstrong, Lisa M. Coussens, Marcia R. Cruz-Correa, Ralph J. DeBerardinis, James H. Doroshow, Margaret Foti, et al. 2016. "AACR Cancer Progress Report 2016." *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*.

Ding, Li, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S. Welch, Julie K. Ritchey, et al. 2012. “Clonal Evolution in Relapsed Acute Myeloid Leukaemia Revealed by Whole-Genome Sequencing.” *Nature*. https://doi.org/10.1038/nature10738.

Dudley, Joel T., Yuseob Kim, Li Liu, Glenn J. Markov, Kristyn Gerold, Rong Chen, Atul J. Butte, and Sudhir Kumar. 2012. “Human Genomic Disease Variants: A Neutral Evolutionary Explanation.” *Genome Research*. https://doi.org/10.1101/gr.133702.111.

Egan, Jan B., Chang Xin Shi, Waibhav Tembe, Alexis Christoforides, Ahmet Kurdoglu, Shripad Sinari, Sumit Middha, et al. 2012. “Whole-Genome Sequencing of Multiple Myeloma from Diagnosis to Plasma Cell Leukemia Reveals Genomic Initiating Events, Evolution, and Clonal Tides.” *Blood*. https://doi.org/10.1182/blood-2012-01-405977.

El-Kebir, Mohammed, Layla Oesper, Hannah Acheson-Field, and Benjamin J. Raphael. 2015. “Reconstruction of Clonal Trees and Tumor Composition from Multi-Sample Sequencing Data.” In *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv261.

Fisher, R., L. Pusztai, and C. Swanton. 2013. “Cancer Heterogeneity: Implications for Targeted Therapeutics.” *British Journal of Cancer*. https://doi.org/10.1038/bjc.2012.581.

Fortunato, Angelo, Amy Boddy, Diego Mallo, Athena Aktipis, Carlo C. Maley, and John W. Pepper. 2017. “Natural Selection in Cancer Biology: From Molecular Snowflakes to Trait Hallmarks.” *Cold Spring Harbor Perspectives in Medicine*. https://doi.org/10.1101/cshperspect.a029652.

Fraley, Chris, and Adrian E. Raftery. 1999. “MCLUST: Software for Model-Based Cluster Analysis.” *Journal of Classification*. https://doi.org/10.1007/s003579900058.

Fritsch, Arno, and Katja Ickstadt. 2009. “Improved Criteria for Clustering Based on the Posterior Similarity Matrix.” *Bayesian Analysis*. https://doi.org/10.1214/09-BA414.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. “Single-Cell Genome Sequencing: Current State of the Science.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2015.16.

Gerlinger, M., and C. Swanton. 2010. “How Darwinian Models Inform Therapeutic Failure Initiated by Clonal Heterogeneity in Cancer Medicine.” *British Journal of Cancer*. https://doi.org/10.1038/sj.bjc.6605912.

Greaves, Mel, and Carlo C. Maley. 2012. “Clonal Evolution in Cancer.” *Nature*. https://doi.org/10.1038/nature10762.

Greenman, Chris, Richard Wooster, P. Andrew Futreal, Michael R. Stratton, and Douglas F. Easton. 2006. “Statistical Analysis of Pathogenicity of Somatic Mutations in Cancer.” *Genetics*. https://doi.org/10.1534/genetics.105.044677.

Greenman, Christopher, Philip Stephens, Raffaella Smith, Gillian L. Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, et al. 2007. “Patterns of Somatic Mutation in Human Cancer Genomes.” *Nature*. https://doi.org/10.1038/nature05610.

Griffith, Malachi, Christopher A. Miller, Obi L. Griffith, Kilannin Krysiak, Zachary L. Skidmore, Avinash Ramu, Jason R. Walker, et al. 2015. “Optimizing Cancer Genome Sequencing and Analysis.” *Cell Systems*. https://doi.org/10.1016/j.cels.2015.08.015.

Grossman, Robert L., Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. 2016. “Toward a Shared Vision for Cancer Genomic Data.” *New England Journal of Medicine*. https://doi.org/10.1056/NEJMp1607591.

Hanahan, Douglas, and Robert A. Weinberg. 2011. “Hallmarks of Cancer: The next Generation.” *Cell*. https://doi.org/10.1016/j.cell.2011.02.013.

Harrington, Kevin J. 2016. “The Biology of Cancer.” *Medicine (United Kingdom)*. https://doi.org/10.1016/j.mpmed.2015.10.005.

Hess, Julian M., Andre Bernards, Jaegil Kim, Mendy Miller, Amaro Taylor-Weiner, Nicholas J. Haradhvala, Michael S. Lawrence, and Gad Getz. 2019. “Passenger Hotspot Mutations in Cancer.” *Cancer Cell*. https://doi.org/10.1016/j.ccell.2019.08.002.

Higgins, Michaela J., and José Baselga. 2011. “Targeted Therapies for Breast Cancer.” *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI57152.

Hiley, Crispin, Elza C. de Bruin, Nicholas McGranahan, and Charles Swanton. 2014. “Deciphering Intratumor Heterogeneity and Temporal Acquisition of Driver Events to Refine Precision Medicine.” *Genome Biology*. https://doi.org/10.1186/s13059-014-0453-8.

Hinohara, Kunihiko, and Kornelia Polyak. 2019. “Intratumoral Heterogeneity: More Than Just Mutations.” *Trends in Cell Biology*. https://doi.org/10.1016/j.tcb.2019.03.003.

Iacobuzio-Donahue, Christine A., Jason Song, Giovanni Parmiagiani, Charles J. Yeo, Ralph H. Hruban, and Scott E. Kern. 2004. “Missense Mutations of MADH4: Characterization of the Mutational Hot Spot and Functional Consequences in Human Tumors.” *Clinical Cancer Research*. https://doi.org/10.1158/1078-0432.CCR-1121-3.

James Kent, W., Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research*. https://doi.org/10.1101/gr.229102. Article published online before print in May 2002.

Jian, Xueqiu, Eric Boerwinkle, and Xiaoming Liu. 2014. “In Silico Tools for Splicing Defect Prediction: A Survey from the Viewpoint of End Users.” *Genetics in Medicine*. https://doi.org/10.1038/gim.2013.176.

Joyce, James M. 2011. “Kullback-Leibler Divergence.” In *International Encyclopedia of Statistical Science*. https://doi.org/10.1007/978-3-642-04898-2_327.

Kamburov, Atanas, Michael S. Lawrence, Paz Polak, Ignaty Leshchiner, Kasper Lage, Todd R. Golub, Eric S. Lander, and Gad Getz. 2015. “Comprehensive Assessment of Cancer Missense Mutation Clustering in Protein Structures.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1516373112.

Krump, Nathan A., and Jianxin You. 2018. “Molecular Mechanisms of Viral Oncogenesis in Humans.” *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-018-0064-6.

Kryazhimskiy, Sergey, and Joshua B. Plotkin. 2008. “The Population Genetics of DN/DS.” *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1000304.

Kumar, Sudhir, Maxwell Sanderford, Vanessa E. Gray, Jieping Ye, and Li Liu. 2012. “Evolutionary Diagnosis Method for Variants in Personal Exomes.” *Nature Methods*. https://doi.org/10.1038/nmeth.2147.

Landau, Dan A., Scott L. Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S. Lawrence, Carrie Sougnez, et al. 2013. “Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia.” *Cell*. https://doi.org/10.1016/j.cell.2013.01.019.

Lipinski, Kamil A., Louise J. Barber, Matthew N. Davies, Matthew Ashenden, Andrea Sottoriva, and Marco Gerlinger. 2016. “Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine.” *Trends in Cancer*. https://doi.org/10.1016/j.trecan.2015.11.003.

Liu, Jinping, Hien Dang, and Xin Wei Wang. 2018. “The Significance of Intertumor and Intratumor Heterogeneity in Liver Cancer.” *Experimental and Molecular Medicine*. https://doi.org/10.1038/emm.2017.165.

Liu, Li, and Sudhir Kumar. 2013. “Evolutionary Balancing Is Critical for Correctly Forecasting Disease-Associated Amino Acid Variants.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/mst037.

Llovet, Josep M., Andrew Burroughs, and Jordi Bruix. 2003. “Hepatocellular Carcinoma.” In *Lancet*. https://doi.org/10.1016/S0140-6736(03)14964-1.

Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. “Understanding Variable Importances in Forests of Randomized Trees.” In *Advances in Neural Information Processing Systems*.

Ma, Qianli C., Catherine A. Ennis, and Samuel Aparicio. 2012. “Opening Pandora’s Box-the New Biology of Driver Mutations and Clonal Evolution in Cancer as Revealed by next Generation Sequencing.” *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2012.01.008.

Maley, Carlo C., Patricia C. Galipeau, Xiaohong Li, Carissa A. Sanchez, Thomas G. Paulson, and Brian J. Reid. 2004. “Selectively Advantageous Mutations and Hitchhikers in Neoplasms: P16 Lesions Are Selected in Barrett’s Esophagus.” *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-03-3249.

Meldrum, Cliff, Maria A. Doyle, and Richard W. Tothill. 2011. “Next-Generation Sequencing for Cancer Diagnostics: A Practical Perspective.” *Clinical Biochemist Reviews*.

Merlo, Lauren M.F., and Carlo C. Maley. 2010. “The Role of Genetic Diversity in Cancer.” *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI42088.

Miller, Christopher A., Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, et al. 2014. “SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003665.

Miller, Martin L., Ed Reznik, Nicholas P. Gauthier, Bülent Arman Aksoy, Anil Korkut, Jianjiong Gao, Giovanni Ciriello, Nikolaus Schultz, and Chris Sander. 2015. “Pan-Cancer Analysis of Mutation Hotspots in Protein Domains.” *Cell Systems*. https://doi.org/10.1016/j.cels.2015.08.014.

Miura, Sayaka, Karen Gomez, Oscar Murillo, Louise A. Huuki, Tracy Vu, Tiffany Buturla, and Sudhir Kumar. 2018. “Predicting Clone Genotypes from Tumor Bulk Sequencing of Multiple Samples.” *Bioinformatics (Oxford, England)*. https://doi.org/10.1093/bioinformatics/bty469.

Morris, Luc G.T., and Timothy A. Chan. 2015. “Therapeutic Targeting of Tumor Suppressor Genes.” *Cancer*. https://doi.org/10.1002/cncr.29140.

Mort, Matthew, Dobril Ivanov, David N. Cooper, and Nadia A. Chuzhanova. 2008. “A Meta-Analysis of Nonsense Mutations Causing Human Genetic Disease.” *Human Mutation*. https://doi.org/10.1002/humu.20763.

Navin, Nicholas, Alexander Krasnitz, Linda Rodgers, Kerry Cook, Jennifer Meth, Jude Kendall, Michael Riggs, et al. 2010. “Inferring Tumor Progression from Genomic Heterogeneity.” *Genome Research*. https://doi.org/10.1101/gr.099622.109.

Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, et al. 2012. “The Life History of 21 Breast Cancers.” *Cell*. https://doi.org/10.1016/j.cell.2012.04.023.

Niu, Beifang, Adam D. Scott, Sohini Sengupta, Matthew H. Bailey, Prag Batra, Jie Ning, Matthew A. Wyczalkowski, et al. 2016. “Protein-Structure-Guided Discovery of Functional Mutations across 19 Cancer Types.” *Nature Genetics*. https://doi.org/10.1038/ng.3586.

Nowell, Peter C. 1976. “The Clonal Evolution of Tumor Cell Populations.” *Science*. https://doi.org/10.1126/science.959840.

Porta-Pardo, Eduard, Atanas Kamburov, David Tamborero, Tirso Pons, Daniela Grases, Alfonso Valencia, Nuria Lopez-Bigas, Gad Getz, and Adam Godzik. 2017. “Comparison of Algorithms for the Detection of Cancer Drivers at Subgene Resolution.” *Nature Methods*. https://doi.org/10.1038/nmeth.4364.

Prévostel, Corinne, Cyrine Rammah-Bouazza, Hélène Trauchessec, Lucile Canterel-Thouennon, Muriel Busson, Marc Ychou, and Philippe Blache. 2016. “SOX9 Is an Atypical Intestinal Tumor Suppressor Controlling the Oncogenic Wnt/β-Catenin Signaling.” *Oncotarget*. https://doi.org/10.18632/oncotarget.10573.

Roth, Andrew, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. 2014. “PyClone: Statistical Inference of Clonal Population Structure in Cancer.” *Nature Methods*. https://doi.org/10.1038/nmeth.2883.

Roy, Ananda L. 2017. “Pathophysiology of TFII-I: Old Guard Wearing New Hats.” *Trends in Molecular Medicine*. https://doi.org/10.1016/j.molmed.2017.04.002.

Sawyers, Charles. 2004. “Targeted Cancer Therapy.” *Nature*. https://doi.org/10.1038/nature03095.

Schaefer, Martin H., and Luis Serrano. 2016. “Cell Type-Specific Properties and Environment Shape Tissue Specificity of Cancer Genes.” *Scientific Reports*. https://doi.org/10.1038/srep20707.

Schaub, Franz X., Varsha Dhankani, Ashton C. Berger, Mihir Trivedi, Anne B. Richardson, Reid Shaw, Wei Zhao, et al. 2018. “Pan-Cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas.” *Cell Systems*. https://doi.org/10.1016/j.cels.2018.03.003.

Schmitt, Michael W., Lawrence A. Loeb, and Jesse J. Salk. 2016. “The Influence of Subclonal Resistance Mutations on Targeted Cancer Therapy.” *Nature Reviews Clinical Oncology*. https://doi.org/10.1038/nrclinonc.2015.175.

Schneider, Günter, Marc Schmidt-Supprian, Roland Rad, and Dieter Saur. 2017. “Tissue-Specific Tumorigenesis: Context Matters.” *Nature Reviews Cancer*. https://doi.org/10.1038/nrc.2017.5.

Schwartz, Russell, and Alejandro A. Schäffer. 2017. “The Evolution of Tumour Phylogenetics: Principles and Practice.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2016.170.

Segal, Neil H., D. Williams Parsons, Karl S. Peggs, Victor Velculescu, Ken W. Kinzler, Bert Vogelstein, and James P. Allison. 2008. “Epitope Landscape in Breast and Colorectal Cancer.” *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-07-3095.

Shah, Sohrab P., Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, et al. 2012. “The Clonal and Mutational Evolution Spectrum of Primary Triple-Negative Breast Cancers.” *Nature*. https://doi.org/10.1038/nature10933.

Shendure, Jay, and Hanlee Ji. 2008. “Next-Generation DNA Sequencing.” *Nature Biotechnology*. https://doi.org/10.1038/nbt1486.

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal. 2019. “Cancer Statistics, 2019: {Cancer} {Statistics}, 2019.” *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21551.

Silva, Jillian M., Marian M. Deuker, Bruce C. Baguley, and Martin McMahon. 2017. “PIK3CA-Mutated Melanoma Cells Rely on Cooperative Signaling through MTORC1/2 for Sustained Proliferation.” *Pigment Cell and Melanoma Research*. https://doi.org/10.1111/pcmr.12586.

Sims, David, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. 2014. “Sequencing Depth and Coverage: Key Considerations in Genomic Analyses.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3642.

Sleire, Linda, Hilde Elisabeth Førde-Tislevoll, Inger Anne Netland, Lina Leiss, Bente Sandvei Skeie, and Per Øyvind Enger. 2017. “Drug Repurposing in Cancer.” *Pharmacological Research*. https://doi.org/10.1016/j.phrs.2017.07.013.

Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. 2018. “The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers.” *Nature Reviews Cancer*. https://doi.org/10.1038/s41568-018-0060-1.

Sottoriva, Andrea, and Trevor Graham. 2015. “A Pan-Cancer Signature of Neutral Tumor Evolution.” *BioRxiv*. https://doi.org/10.1101/014894.

Stephens, Philip, Sarah Edkins, Helen Davies, Chris Greenman, Charles Cox, Chris Hunter, Graham Bignell, et al. 2005. “A Screen of the Complete Protein Kinase Gene Family Identifies Diverse Patterns of Somatic Mutations in Human Breast Cancer.” *Nature Genetics*. https://doi.org/10.1038/ng1571.

Stratton, Michael R. 2011. “Exploring the Genomes of Cancer Cells: Progress and Promise.” *Science*. https://doi.org/10.1126/science.1204040.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. “The Cancer Genome.” *Nature*. https://doi.org/10.1038/nature07943.

Sun, Ruping, Zheng Hu, Andrea Sottoriva, Trevor A. Graham, Arbel Harpak, Zhicheng Ma, Jared M. Fischer, Darryl Shibata, and Christina Curtis. 2017. “Between-Region Genetic Divergence Reflects the Mode and Tempo of Tumor Evolution.” *Nature Genetics*. https://doi.org/10.1038/ng.3891.

Tao, Yong, Jue Ruan, Shiou Hwei Yeh, Xuemei Lu, Yu Wang, Weiwei Zhai, Jun Cai, et al. 2011. “Rapid Growth of a Hepatocellular Carcinoma and the Driving Mutations Revealed by Cell-Population Genetic Analysis of Whole-Genome Data.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1108715108.

Temko, Daniel, Ian P.M. Tomlinson, Simone Severini, Benjamin Schuster-Böckler, and Trevor A. Graham. 2018. “The Effects of Mutational Processes and Selection on Driver Mutations across Cancer Types.” *Nature Communications*. https://doi.org/10.1038/s41467-018-04208-6.

The Cancer Genoma Atlas. 2013. “TCGA.” *National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI)*.

Tokheim, Collin J., Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin. 2016. “Evaluating the Evaluation of Cancer Driver Genes.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1616440113.

Visvader, Jane E. 2011. “Cells of Origin in Cancer.” *Nature*. https://doi.org/10.1038/nature09781.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. 2013. “Cancer Genome Landscapes.” *Science*. https://doi.org/10.1126/science.1235122.

Weinberg, Robert A. 1993. “ Oncogenes and Tumor Suppressor Genes. F. Macdonald , C. H. J. Ford .” *The Quarterly Review of Biology*. https://doi.org/10.1086/418032.

Williams, Marc J., Benjamin Werner, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva. 2016. “Identification of Neutral Tumor Evolution across Cancer Types.” *Nature Genetics*. https://doi.org/10.1038/ng.3489.

Youn, Ahrim, and Richard Simon. 2011. “Identifying Cancer Driver Genes in Tumor Genome Sequencing Studies.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq630.

Zhao, Siming, Jun Liu, Pranav Nanga, Yuwen Liu, A. Ercument Cicek, Nicholas Knoblauch, Chuan He, Matthew Stephens, and Xin He. 2019. “Detailed Modeling of Positive Selection Improves Detection of Cancer Driver Genes.” *Nature Communications*. https://doi.org/10.1038/s41467-019-11284-9.

APPENDIX A

PROOF

We model each cluster as a Beta distribution. We show that the variance of a beta distribution is positively correlated to the mean VAF and negatively correlated to the mean sequencing depth. This is important in clustering mutations in low coverage. Because if the model does not consider the coverage, the model will over estimate the number of clusters. In here, we show why the variance is higher in low coverage and also why the clusters with bigger mean VAF have larger variance.

The variance of the beta distribution is: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ replacing the depth: $\frac{\alpha\beta}{(\bar{e})^2(\bar{e}+1)}$

To look at the effect of VAF on variance, we keep the depth constant. With constant depth, higher VAF means bigger α. Since $\bar{e} = \alpha + \beta$ is constant, therefore β gets smaller. Therefore αβ is bigger. This results in higher variance. Since the denominator of the variance is constant and the numerator which is αβ increases. It is concluded that with constant depth, with higher VAF, the variance of the subclone increases.

To study the effects of depth of sequencing for a subclone at a fixed VAF, we expect that the variance gets smaller if the depth of the sequencing is increased. If the new reference and alternate reads are $\alpha'$and β' the increased depth is $\alpha' + \beta'$and we have $\alpha' + \beta' > \alpha + \beta$. By keeping the VAF fixed, we have: $\frac{\alpha}{\alpha+\beta} = \frac{\alpha'}{\alpha'+\beta'}$

It can also be shown that 1-VAF is also fixed. $1 - \text{VAF} = \frac{\beta}{\alpha+\beta} = \frac{\beta'}{\alpha'+\beta'}$

Now, we can rewrite the new variance as: $\frac{\alpha'\beta'}{(\alpha'+\beta')^2(\alpha'+\beta'+1)} = \frac{\alpha'}{(\alpha'+\beta')} \times \frac{\beta'}{(\alpha'+\beta')} \times \frac{1}{(\alpha'+\beta'+1)}$

We showed that the first and second part of the variance are unchanged but since $\alpha' + \beta'$has increased, the variance decreases.

APPENDIX B

NOTE ON UNPUBLISHED WORK

The GUST algorithm, discussed in chapter 3 of this work is in press and expected to be published shortly after my dissertation defense. GUST is an algorithm designed with collaboration between Drs. Li Liu, Carlo Maley, Sudhir Kumar, Pramod Chandrashekar and myself, Navid Ahmadinejad. I am the second author of this paper. I contributed by applying the GUST algorithm to the TCGA data and analyzing the results as discussed in the publication. Pramod Chandrashekar developed the online database. All co-authors have granted their permissions for this work to be included in my dissertation.