Integrative Analysis of Genomic Aberrations in Cancer and Xenograft Models

by

Sen Peng

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

Approved October 2015 by the
Graduate Supervisory Committee:

Valentin Dinu, Chair
Matthew Scotch
Garrick Wallstrom

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

No two cancers are alike. Cancer is a dynamic and heterogeneous disease, such heterogeneity arise among patients with the same cancer type, among cancer cells within the same individual's tumor and even among cells within the same sub-clone over time. The recent application of next-generation sequencing and precision medicine techniques is the driving force to uncover the complexity of cancer and the best clinical practice. The core concept of precision medicine is to move away from crowd-based, best-for-most treatment and take individual variability into account when optimizing the prevention and treatment strategies. Next-generation sequencing is the method to sift through the entire 3 billion letters of each patient's DNA genetic code in a massively parallel fashion.

The deluge of next-generation sequencing data nowadays has shifted the bottleneck of cancer research from multiple "-omics" data collection to integrative analysis and data interpretation. In this dissertation, I attempt to address two distinct, but dependent, challenges. The first is to design specific computational algorithms and tools that can process and extract useful information from the raw data in an efficient, robust, and reproducible manner. The second challenge is to develop high-level computational methods and data frameworks for integrating and interpreting these data. Specifically, Chapter 2 presents a tool called Snipea (SNv Integration, Prioritization, Ensemble, and Annotation) to further identify, prioritize and annotate somatic SNVs (Single Nucleotide Variant) called from multiple variant callers. Chapter 3 describes a novel alignment-based algorithm to accurately and losslessly classify sequencing reads from xenograft models. Chapter 4 describes a direct and biologically motivated framework and associated methods for identification of putative aberrations causing survival difference in GBM patients by integrating whole-genome sequencing, exome sequencing, RNA-Sequencing, methylation array and clinical data. Lastly, chapter 5 explores longitudinal

and intratumor heterogeneity studies to reveal the temporal and spatial context of tumor evolution. The long-term goal is to help patients with cancer, particularly those who are in front of us today. Genome-based analysis of the patient tumor can identify genomic alterations unique to each patient's tumor that are candidate therapeutic targets to decrease therapy resistance and improve clinical outcome.

DEDICATION

To my father (Bingcheng Peng) and mother (Zhongying Wang) who are always in support of me and my Fiancée (Xiaoxiao Wang) without whom this dissertation might be finished sooner.

ACKNOWLEDGMENTS

First and foremost, I'd like to express my sincere gratitude to my ASU advisor Dr. Valentin Dinu and my TGen mentors Drs. Nhan Tran and Michael Berens. They not only provided unsurpassed and continuous guidance, but also challenged me for self-motivated research and critical thinking. Their advice and genuine encouragement have been invaluable to me throughout my graduate career. Besides all the guidance and training, the financial support from The Ben & Catherine Ivy foundation and TGen enabled me to pursue my degree.

I would also thank Drs. Matthew Scotch and Garrick Wallstrom for serving on my PhD committee and Dr. David Kaufman for serving on my PhD comprehensive exam committee. They are so responsive and provided helpful feedback and timely suggestions to my questions about research and draft of this thesis.

A special thanks to Brock Armstrong, who worked closely with me and helped me through hard times. This work has greatly benefited from many useful discussions with Brock and his advices are on not only research but also personal life. My deep thanks also go to Drs. Harshil Dhruv, Timothy Whitsett, Jonathan Keats, Jeff Kiefer, Seungchan Kim, Bodour Salhia, Leland Hu, Jann Sarkaria, Michael Prados and Joanna Phillips, who guided me through troubleshooting approaches and discussion with genomics and programming.

Fortunately I had the opportunity to collaborate with and learn from all my friends and colleagues: Christophe Legendre, Megan Russell, Ahmet Kurdoglu, Jessica Aldrich, Tyler Izatt, Winnie Liang, Mark Teng, Venkata Yellapantula and Barrie Bradley.

Last but not the least, I am grateful to my parents and family for their constant love, unconditional support and confidence in me.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

ABBREVIATIONS

aCGH - Array Comparative Genomic Hybridization

AF - Allele Fraction

BAM - Binary Alignment Map

BBB - Blood Brain Barrier

BWA - Burrows-Wheeler Aligner

CBS - Circular Binary Segmentation

CCDS - Consensus Coding DNA Sequence

CLIA - Clinical Laboratory Improvement Amendments

CNV - Copy Number Variation

COSMIC - the CatalOgue of Somatic Mutations in Cancer

CRISPER - Clustered Regularly Interspaced Short Palindromic Repeats

CV - Coefficient of Variation

DNA - Deoxyribonucleic Acid

dbSNP - The Single Nucleotide Polymorphism Database

dbNSFP - Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations

ENCODE - Encyclopedia of DNA Elements

ESP - Exome Sequencing Project

FDR - False Discovery Rate

FPKM - Fragments per Kilo Base of Exon per Million Fragments Mapped

GATK - Genome Analysis Tool Kit

GFF - Gene Feature Format

GTF - Gene Transfer Format

HPC - High Performance Computing

HD - Huntington's Disease

ICGC - International Cancer Genome Consortium

INDEL - Insertions and Deletions

ITH - Intratumor Heterogeneity

mRNA - Messenger Ribonucleic Acid

NGS - Next generation Sequencing

NHLBI - National Heart, lung and Blood Institute

NSCL - Non-small Cell Lung Cancer

PCR - Polymerase Chain Reaction

q-PCR - Quantitative Polymerase Chain Reaction

PDX - Patient Derived Tumor Xenograft

RNA - Ribonucleic Acid

SAM - Sequence Alignment Map

SNP - Single Nucleotide Polymorphism

SNV - Single Nucleotide Variant

TGen - Translational Genomics Research Institute

VAF - Variant Allele Frequency

VCF - Variant Call Format

WS - Weighted Score

WES - Whole Exome Sequencing

WGS - Whole Genome Sequencing

CHAPTER 1

INTRODUCTION

1.1. Cancer is a heterogeneous and complex disease

No two cancers are alike. Cancer could be regarded as a complex, heterogeneous and evolutionary process. Spontaneous cancer commonly results from a series of mutations within a single cell (Marusyk and Polyak 2010). However, it has been previously reported that distinct subpopulations exist in various human cancers, including acute myeloid leukemia (AML), breast cancer, ovarian cancer, colorectal cancer, glioblastoma and pancreatic cancer (Marusyk and Polyak 2010, Meacham and Morrison 2013).

Genetic changes (such as SNV), environmental differences (such as radiation) and reversible changes (such DNA repair during replication and epigenetic changes) are believed to mainly contribute to morphological, physiological and functional heterogeneity in cancer (Marusyk, Almendro et al. 2012).

Therapeutic resistance has been proven to be one of the foremost obstacles limiting the clinical efficacy of cancer drug treatments across many tumor types. Recent evidence indicated that intrinsic heterogeneity of tumors is one of the main mechanisms of acquired drug resistance (Yap, Gerlinger et al. 2012). The response and thus killing of “sensitive” population of the tumor hardly could avoid the flourish of “insensitive” part and therefore cause the inevitable recurrence of the cancer.

Tumor heterogeneity describes the observation that cancer cells display substantially distinct genetic and phenotypic features, such as cellular morphology, gene expression, metabolism, metastatic potential and invasiveness. In this case, although it remains unclear what inheritable mechanism and specific model the cancer cells follow, tumor heterogeneity definitely have profound implications both for tumor progression and

therapeutic response. In despite of the substantial clinical significance of tumor clonal heterogeneity, the issue still remains relatively poorly explored. To address this, a more systematic technique is desired to characterize the extent and molecular signatures of clonal heterogeneity of different cancer types through longitudinal stages of tumor progression.

1.2. Precision medicine and Massively parallel sequencing

In the field of cancer treatment, “one size” doesn’t fit all. Knowing the existence of heterogeneity within most cancer types, the idea of precision medicine is becoming increasingly popular and it is currently accepted that molecular information will improve the precision with which patients are categorized and treated.

In this manner, the concept of treatment has now been extended to “5P” - Preemptive, predictive, personalized, participatory and precise (Bradley, Golding et al. 2011). The core concept of precision medicine is to take individual variability into account when optimizing the prevention and treatment strategies (Jameson and Longo 2015). Naturally the first the foremost task would be to accurately identify such individual variability in an effective fashion cost wise and time wise.

Massively parallel sequencing, with increasing throughput and reducing cost, has enabled us to simultaneously screen for potential biomarkers -- genetic and epigenetic variants: copy number variants (CNV), single nucleotide variant (SNV), gene fusions, translocations, methylation and over-expression or under-expression of genes. Those individual characteristics of each patient could then be used to design tailored medical treatment, which may show higher susceptibility to the patient’s tumor. Achieving the goals of precision medicine thus will require the proper handling, integration and interpretation of multi-parametric “-omics” data. On the other hand, the knowledge

gathered during the precision medicine studies would in turn enable better understanding of disease mechanisms and improvement of large-scale biological database (such as “Genomics of Drug Sensitivity in Cancer”) in the long term (Yang, Soares et al. 2013).

1.3. Sequencing technology for personalized cancer treatment : Application and Challenges

In order to fulfill the bi-directional communication gap of “Bench side and Bedside”, a key component of translational research is the management, integration and analysis of large amount of both genomic and clinical data.

Some initial studies followed the “common disease – common variant” hypothesis and successfully identified potential cause of genetics disease (Klein, Zeiss et al. 2005). One example would be the identification of cause of Huntington’s disease (HD) to be excessive CAG trinucleotide repeats in the HD gene (Myers 2004). However, due to the complexity of tumor and rareness of some genetic disorders, such hypothesis is challenged. In this case, more comprehensive characterization is imperative for identifying the key changes in the DNA as well as gaining the biological insights for diagnosis and therapies.

The increasing affordability of high throughput assays (Exome sequencing, RNA sequencing, Whole-genome sequencing, and methylation assay) make it possible for the simultaneous measurement of several genomic features in the same biological samples.

However, such emerging genomics technology and bioinformatics make the cancer field rapidly realize that the bottleneck in discovery is no longer data generation, but data analysis. As genetic features play a significant role in the metabolism and the function of the cells, the integration of genetic information (genomics – proteomics –metabolomics

- phenotype) to cancer research is now perceived by scientists not as a future trend but rather as a demanding need.

The field of computational biology and bioinformatics are dealing with a growing range of genomic data types, including RNA transcriptional expression, SNV and INDEL (small insertion and deletion), DNA copy number variation, translocation, gene fusions and epigenetic markers. Moreover, integration of findings from multiple experimental approaches will be necessary to distinguish hidden interactions between various layers of data.

Therefore, genomic data integration - the process of statistically combining diverse sources of information from these experiments - is becoming increasingly prevalent and challenging.

1.4. Xenograft model as a key tool for cancer study and drug development

Numerous murine models have been developed to study human cancer. One of the most widely used models is the human tumor xenograft. In this model, human tumor cells are transplanted, either subcutaneously or orthotopically into the organ type in which the tumor originated, using immune-compromised mice that do not reject human cells such as severely compromised immune-deficient (SCID) mice or other immune compromised mice.

In general, human cancer xenografts represent the gold standard method used to investigate the factors involved in drug discovery, cancer stem cell biology, metastasis prediction, as well as response to therapy. Compared to in vitro cell culture models, xenografts usually show a higher validity across most assays. These models would be very helpful for testing in vivo the effect of novel/repurposing drugs which is individualized and was decided by genetic makeup of specific patient (DeRose, Wang et

al. 2011). A study led by Spanish National Cancer Research Center has demonstrated successful implementation of xenograft models as investigational platform for therapeutic decision-making. This study showed that known actionable alterations (such as NF1, PI3KA, and DDR2) accurately predicted the ineffectiveness of treatment of the patients based on the failure of the same drug tested on the xenograft models (Garralda, Paz et al. 2014).

With the decrease in sequencing cost and increase in the sequencing capacity over the past decades, more and more studies have begun to apply NGS technology to xenograft models. One study (Bradford, Farren et al. 2013) established that RNA-seq could be applied to xenograft model to gain better understanding of drug mechanism of action and identify both tumor and host biomarkers. Specifically, they found increased expression of genes related to inflammatory response in mouse and induction of hypoxia genes in human when xenograft mouse were treated with cediranib (a potent vascular endothelial growth factor receptor kinase inhibitor).

Due to unclear boundary of human and mouse stromal cells, samples that are extracted for parallel sequencing will inevitably contain various amounts of mouse contamination. Since there exists high degree of homology between human and mouse genome, short sequencing reads originating from mouse genome could potentially have an impact on the accuracy of downstream analysis (such as copy number analysis, gene expression profiling). Recent evidence points to the need for an appropriate method that could separate graft reads from host reads and thus ensure precise genomic aberration identification (Leong, Marini et al. 2014). For example, researchers showed that mRNA level change of stem/serrated/mesenchymal (SSM) genes were mostly due to mouse stromal expression instead of human cancer cells (Isella, Terrasi et al. 2015). Thus, it is important to apply bioinformatic techniques to ensure species specific gene expression.

1.5. Biology of Glioblastoma Multiforme (GBM)

GBM is a common and extremely malignant form of brain cancer. The disease most commonly affects adults in their sixth decade of life; gliomas also affect children with an incidence of 2 to 3 out of 100,000, with 14% of these diagnosed as malignant glioma. Of the 12,000 or so patients diagnosed with GBM each year, about half die within the first year of diagnosis, with most of the rest succumbing to their disease within five years. Current treatment options for GBM patients are limited and largely ineffective. The mechanisms driving the development and recurrence of GBM are still unknown. This fact greatly limits the successful treatment of this disease. Currently, standard treatment includes surgical resection followed by radiation along with concurrent and adjuvant chemotherapy using temozolomide (TMZ), which only extends the current median survival to 14.6 months.

Unfortunately, possibly due to the aforementioned heterogeneous and highly invasive characteristics, GBM often exhibits a high resistance to these standard therapies and recurrence is nearly assured. However, there is no established second-line regimen. In order to address the dismal prognosis & management of patients with GBM, it is essential to transform traditional clinical trial paradigms to allow for rapid and efficient therapeutic development. Thus, the development of new combinational therapies, together with an increase in the selectivity of the treatments based on a detailed molecular characterization of these tumors has significant potential to enhance the survival of patients suffering from GBM (Verhaak, Hoadley et al. 2010).

Genomic characterization will constitute an ever-increasing fundamental role in the delivery of individualized care for oncologic patients. Emerging genomics technology and bioinformatics are now resulting in the molecular sub-classification of cancers with applications for more accurate characterization of disease, prognosis, and therapeutic

selection (Hanahan and Weinberg 2011, Brennan, Verhaak et al. 2013). Under this paradigm, therapy selection is guided by the molecular profile of targetable mutations and gene pathways that vary among patients. This phenomenon is well represented in GBM, which is among the most genetically heterogeneous and lethal of all human cancers (Marusyk, Almendro et al. 2012).

1.6. Integration framework is needed to analyze multi-parametric "-omics" data

There is an increasing trend toward acquiring a number of types of data from the same patient in both clinical and research field. Once we get patient tumor aberrations from various types of data (E.g., Copy number, SNV, expression information), the remaining challenge is how to effectively integrate them and try to identify potential cancer causing variants ("drivers") and their corresponding drug/treatment.

In recent large-scale cancer genome studies (Leary, Lin et al. 2008, Parsons, Jones et al. 2008, Verhaak, Hoadley et al. 2010) preliminary integration approaches have been successfully applied; however, these approaches have been tailored to very specific niches and studies. In addition, previous studies are often at most of two or three data types; while in our GBM survival outlier project only, we have more than five types of data. A systematic framework shall be developed to integrate them. Moreover, Current frameworks often lack the ability to predict key "drivers" in the disease, let alone the possible individualized treatment.

Genomics, transcriptomics, proteomics, epigenomics and metabolomics data each of course enables us to get a specific and insightful view of genome functions, but those views are often limited to one-dimension. Just like complex biological processes, data describing those processes are usually complementary and shall not be treated totally

independently. To maximize the utilization of all available information, we should consider each assay as part of "big picture" with unified, global view.

Few previous attempts have been made to integrate various "-omics" data in a systematic manner. Integration of exome sequencing data, RNA-Seq, whole genome and epigenetic data in a coherent fashion is critical to comprehensive understanding of molecular interactions in complex genetic diseases. For instance, integrating highly informative yet individual datasets offer the potential to answer many long-standing research questions: what impact does variants in genetic code have on the gene expression variation? To what extent does the methylation and other regulatory elements contribute the disease phenotypes and gene expression? Is there always a corresponding structural rearrangement at DNA level for each gene fusion event at RNA level? Therefore, a more effective interpretation of accumulated and interacting information in the data analysis is in imperative demand.

In this thesis, I present a computational and biology-motivated framework for integrating and interpreting multiple genomics data. My goal is to develop a general, scalable and rigorous statistical framework as well as algorithms designed for solving specific contexts.

1.7. Temporal and spatial tumor evolution

Solid tumors - especially the highly invasive types such as GBM and lung cancer - are often diverse in three dimensions and their oncogenesis processes are dynamic. For example, a study (de Bruin, McGranahan et al. 2014) focus on lung cancer sequenced 25 spatially distinct regions from human non-small cell lung cancer (NSCLC) and found evidence of intratumor heterogeneity (ITH) and branched evolution. Another study (Zhang, Fujimoto et al. 2014) also indicated that all lung tumor regions they sequenced

showed clear evidence of intratumor heterogeneity but suggested that single region sequencing might be enough to identify the majority of known gene mutations.

Such ITH may contribute to the clinical impact on drug resistance, surgery planning and actionable targets strategy (Yap, Gerlinger et al. 2012). Therefore, understanding the initiation, maintenance and evolution of tumor could shed insight into potential therapeutic interventions. Traditional tumor evolution model focused on linear cumulative genomic alterations over time, however, the authors in a leukemia study (Egan, Shi et al. 2012) reported the presence of a "clonal tides" model in tumor evolution. They found that the dominate clones at diagnosis then followed this "tide" model under the conceivable selective pressure of therapeutic drugs and could only be detected at alternating time points after treatment. These findings represent a novel paradigm in evolutionary biology and emphasize Darwinian mutational progression and shifting dominance of different clones over time.

Although aforementioned several studies showed that subpopulations have been identified in lung cancer samples, the extent of genomic diversity or effect of such evolution in other tumor types still remains rudimentary. Moreover, few attempts have been made to explore temporal dynamics of tumor progression and the contribution of somatic aberrations to driver tumor growth overtime.

To resolve spatial and temporal tumor evolution in GBM, we performed multiregional and longitudinal whole-exome, whole-genome and RNA sequencing on samples collected from patients who had primary, relapse and post-relapse GBM. Approaches to delineate the spatial heterogeneity to identify subclones will be described in Chapter 5. Our goal is to investigate the extent of genomic diversity in GBM and to infer ancestral relations between tumor regions and tumor samples over time.

1.8. Summary

In summary, harnessing aforementioned flood of data requires the development of computational tools that can address two distinct, but dependent, challenges. The first is developing computational algorithms and tools that can process the data in an efficient, robust, and reproducible manner and thus extract useful information as output. For example, some of those tools will identify and catalog the genetic alterations present in cancerous cells. The second challenge is to develop computational methods and data frameworks for integrating and interpreting these data. This will augment our understanding of cancer at a systems level and generate testable hypotheses. Ultimately, it is hoped that this knowledge will be translated into better therapeutics and better diagnostics that can pair cancer patients with the appropriate treatment. In this thesis, chapter 2 and chapter 3 will mainly focus on the first challenge; chapter 4 and 5 will mostly address the second challenge. Chapter 2 presents a tool called Snipea (SNv Integration, Prioritization, Ensemble, and Annotation) to further identify, prioritize and annotate somatic SNVs called from multiple variant callers. Chapter 3 describes a novel alignment-based algorithm to accurately and efficiently classify sequencing reads from xenograft models. Chapter 4 describes a direct and biologically motivated framework and associated methods for identification of putative aberrations causing survival difference in GBM patients by integrating whole-genome sequencing, exome sequencing, RNA-sequencing, methylation array and clinical data. Lastly, chapter 5 explored longitudinal and intratumor heterogeneity studies to reveal the temporal and spatial context of tumor evolution.

CHAPTER 2

SNIPEA: SNV INTEGRATION, PRIORITIZATION, ENSEMBLE, AND ANNOTATION

2.1. Introduction

2.1.1. SNV is essential for insightful cancer genome analysis

Cancer is a complex disease caused by genetic mutations and generally requires several mutations to circumvent cellular defenses against carcinogenesis. These mutations and epigenetic modifications alter the expression or activity of genes responsible for maintaining the balance between cell proliferation and cell death.

Two types of mutations are present as essential drivers in many human cancer types: germline and somatic mutations. Somatic mutations are limited to descendants of the original mutated cells (such as tumor) and thus not passable to progeny. But germinal mutations exist in parents' germ cells and may be transmitted to some or all progeny. Compared to the relatively low contributions of germline alleles to the carcinogenesis, we usually focus on somatic mutations for determination of potential tumorigenic mechanisms.

These somatic mutations further fall broadly into two categories: 1. oncogenes which trigger growth and differentiation and are often over-expressed or contain activating mutations. 2. Tumor suppressor genes which act as safeguards that induce apoptosis or retard the cell cycle, thereby preventing unrestrained growth (Goya, Sun et al. 2010, Wu, Li et al. 2014).

In addition, exome sequencing has been widely used in detecting pathogenic non-synonymous single nucleotide variants including SNVs and INDELs (small insertion and deletion). By sequencing to high depth and comparing the results to matched normal

tissue, it is possible to identify cancer-specific somatic mutations that may be contributing to carcinogenesis.

2.1.2. Accuracy is critical for downstream analysis but remains an unsolved issue

Recent evidence points to considerable attention to roles of SNV in cancer genome studies: Jiao et al. used the allele frequency of SNV was used to explore the subclonal lineage in patient tumor samples (Jiao, Vembu et al. 2014), Salari et al. applied SNVs as linkage markers for phylogenetic tree reconstruction to compare and track longitudinal tumor progression (Salari, Saleh et al. 2013), SNV is regarded as key components to translation of pharmacogenomics to clinical applications, somatic molecular alterations-Gene-Drug association (e.g. Drug Bank) and has made individualized medicine possible (Crews, Hicks et al. 2012, Law, Knox et al. 2014). In addition, SNV could be used to identify and monitor tumor burden and progression from circulating tumor DNA in plasma (Forshew, Murtaza et al. 2012).

All those studies relied on a limited number of filtered SNVs; if any inaccurate SNV calling occurred, it could easily result in false conclusions. In light of this, the essential pre-requisite for accurate downstream analysis is to identify a list of high quality somatic variants from each available assay (Whole-genome or Exome). However, distinguishing genuine somatic mutations from artifacts (including germline variation, alignment mismatch, inherent machine error, tumor heterogeneity, normal contamination and genomic complexity) in NGS data is a particularly immense challenge.

Several factors make detection of these mutations non-trivial. Error rates from current generation sequencing machines are still significantly higher than that seen in capillary-based sequencing. This can be overcome by increasing the sequence depth, using

mapping algorithms that can tolerate mismatches, and using some sort of consensus calling at each base pair. Even we assume the difficulty of technical issues would be largely improved (such as alignment inaccuracy and sequencer error) by rapid development of bioinformatics and sequencing technology – let alone bioinformatics field is still far from mature – the fundamental biological problem will still prevent SNV calling from an easy task.

One issue is tumor purity. Tumor samples are derived from a heterogeneous population of cells, and also may have significant admixture with non-cancerous cells. There may also be variations in copy-number that lead to unconventional allelic ratios.

Most of the times, it is almost impossible to obtain a 100% pure tumor sample and thus it will be sequenced "with contamination". Succinctly, we expect heterozygous somatic SNVs to be present in a sample at a frequency around 50% (and homozygous changes to be present at 100%). If the tumor is impure, these fractions will drop. For instance, assuming our sample is 10% normal cells, the actual frequency would change to 45% and 90%. Regions with loss of heterozygosity (LOH) or single copy number loss provide larger numbers of somatic mutations that should occur at nearly 100%, making this estimation easier.

Another complicating problem is that tumors are usually heterozygous and has abnormal variations (e.g. Copy number variation). We have to deal with mutations present along the entire tumor, but also mutations present in subpopulations of this tumor. It is much more difficult to determine these mutation genotypes because their frequencies are conditioned with the size of its population within the tumor. For example, considering a pure sample, a mutation with a frequency of 15% will be considered heterozygous, but it could also be homozygous if the mutation belongs to a tumor subpopulation that resides in 15% of the total tumor area. The combinations of those two problems, along with the

complexity of genome, make the variant allele frequency (VAF) range across almost the entire percentage spectrum, from 0% to 100% (Spencer, Tyagi et al. 2014). To address this, several tools and algorithms have been developed and widely used in academic and industry community.

2.1.3. Single tool has its pros and cons

Since accuracy of SNV calling has a huge impact on downstream analysis and even clinical decision making, the choice of somatic mutation detection tool/algorithm thus may have the most substantial influence on the output. The research community is still defining a best practice for cancer somatic variant calling; thus far no single tool has dominating performance. Multiple previous studies have done comparison between different somatic variant callers and suggested significant room for improvement in regards to the accuracy and discrimination of SNV and SNPs (Wang, Jia et al. 2013, Spencer, Tyagi et al. 2014, Xu, DiCarlo et al. 2014).

Although each of those tools has been compared with some previous softwares and validated using certain datasets, their relative merits and robustness are largely unknown. Some tool will perform better with low coverage or low allelic frequency SNV and some tools will have better accuracy with alternate alleles in normal samples. Moreover, different strategies can perform better in different conditions based on varying parameters such read-depth, allele frequencies in tumor and control samples and tumor purity. After all, there might be no single approach that is in some sense best. Different approaches may work better with different data sets and with different disease. Most variant-calling algorithms are usually highly parameterized. Although we could tune the parameters to generate high quality output for each tool, we usually don't have the resources to validate every finding and thus we are unable to know the optimal

selection threshold. In this fashion, an ensemble idea (similar to data mining technique random forest) could be applied to produce better SNV calling. However, in the literature, there are limited previous works that discuss how to merge results from different variant callers.

2.1.4. Not every SNV has the same impact on clinical outcome

"Driver" genes & pathways are those that are positively selected in the tumor microenvironment and provide growth advantage during cancer development. "Passenger" genes & pathways are viewed as byproducts of oncogenesis and make little or no contributions to tumor progression (Stratton, Campbell et al. 2009). Non-synonymous variants are viewed as intrinsically more reasonable phonotype candidates compared to common, synonymous and non-coding variants. In this manner, not every SNV shall be further explored with equal significance especially when we have limited time and resources.

The idea of "Common disease, common variant" which hypothesizes common variants in the population result in the fraction of susceptibility to common disease has outdated. But discovering driver cancer genes and pathways are still a vital part in cancer genomics and is traditionally identified by ranking the genes & pathways simply based on their alteration recurrence (Tamborero, Gonzalez-Perez et al. 2013).

However, due to the limitation of the traditional approaches such as potential neglecting of low frequently mutated driver genes, the driver mutation should not only rely on the recurrence of the SNV event. The functional impact as well as pathway/network changes shall also be taken into consideration.

Ranking algorithm that takes into account both accuracy and biological functions shall be introduced to the process of prioritizing somatic alterations. Snipea is thus developed

to rank the SNVs by the authenticity as well as their significance in potential disease causing functions. Top-ranking SNVs are believed to be more likely to be involved in oncogenesis, hence could be used to guide specialized cancer genomics researcher for further validation and mechanisms studies.

2.1.5. Annotation is the key bridge between informatics and biology

Once a genome is sequenced and analyzed, it needs to be annotated to make sense of it. Annotation is important to bridge the gap among bioinformatics, biological significances and clinical application. The complete genome annotation spans a diverse range of layers: nucleotide-level annotation, protein-level annotation and process-level annotation. The process of annotation -– integrating layers of information to the raw DNA sequence – has enabled us to extract & interpret the biological insights, and then place such findings into the context of prior knowledge of biological networks (Stein 2001).

Nevertheless, genome annotation is still considered an unreliable and inaccurate procedure due the following potential reasons: 1) Error and inconsistency. The incompleteness of human genome and inconsistency of previous genomic studies, resulted in a considerable amount of errors. For instance, more variants would be annotated using ENSEMBL transcript set compared to Refseq set. 2) Automation. Manually input of each variant into the database search window is a nuisance and a major bottleneck of analysis. 3) Benchmarking. Benching marking is hard and not always straightforward to decide whether the results will improve on the original annotation since ground-truth is often unavailable (Yandell and Ence 2012).

Several mutation prediction tools have been previously developed to meet the strong need of functional annotation for genetic variants (Wang, Li et al. 2010, Cingolani, Platts et al. 2012). However, those current annotation tools such as SNPEFF and ANNOVAR do

not perform any interactions with large-scale genetic variation database (E.g. dbSNP, dbNSFP and COSMIC). In this case, those tools focus more on the nucleotide and protein level annotation; more comprehensive annotation (pathway and system level) needs to be included in the output as well.

It is obviously tedious and time-consuming to manually search for information in large database. Therefore, an efficient, cross-platform algorithm shall be applied to incorporate update-to-date annotation information to cancer genomic variants report.

Snipea will incorporate a comprehensive annotation engine that integrate industry standard cancer database (e.g. COSMIC), known short genetic variations (e.g. dbSNP, NCBI), pathway databases (e.g. KEGG), drug-gene interaction databases (e.g. Drug Bank), pharmacogenomics annotations (e.g. PharmGKB) and hand-curated cancer gene annotations from the literature to generate annotated cancer variant reports for somatic variants. Integration with these public archives will not only enable us to know whether the SNV has been identified before but also provide us with a more comprehensive annotation.

In addition, integrating data retrospectively with the above databases might help us to carry out statistical analysis to identify frequent mutations in a certain cancer type in the population. It might also help us to identify alternative treatments for specific patients if the same mutation in previous study has been reported and targeted in a clinical trial.

Moreover, by incorporating the information of public database such as 1000 genomic project and dbSNP, we could post-filter our variants list to leave only relatively rare and novel mutations. Generally speaking, those drivers for cancer are less likely to be present in those common SNP databases with high allele frequencies. On the other hand, we could utilize the crosstalk with COSMIC (catalogue of somatic mutations relating to human cancers) to rescue any ignored variants for downstream analysis.

2.2. Methods and Materials

2.2.1 Snipea implementation and usage

Snipea is a software application designed to systematically integrate, prioritize and annotate somatic SNV and small INDELs called from multiple variant callers.

Snipea takes options from multiple variant callers instead of one. The idea is somewhat similar to the "random forests" which is a data mining ensemble prediction method. We run each selected variant callers with "best practice" parameter setting and then use majority voting to combine outputs of the each callers.

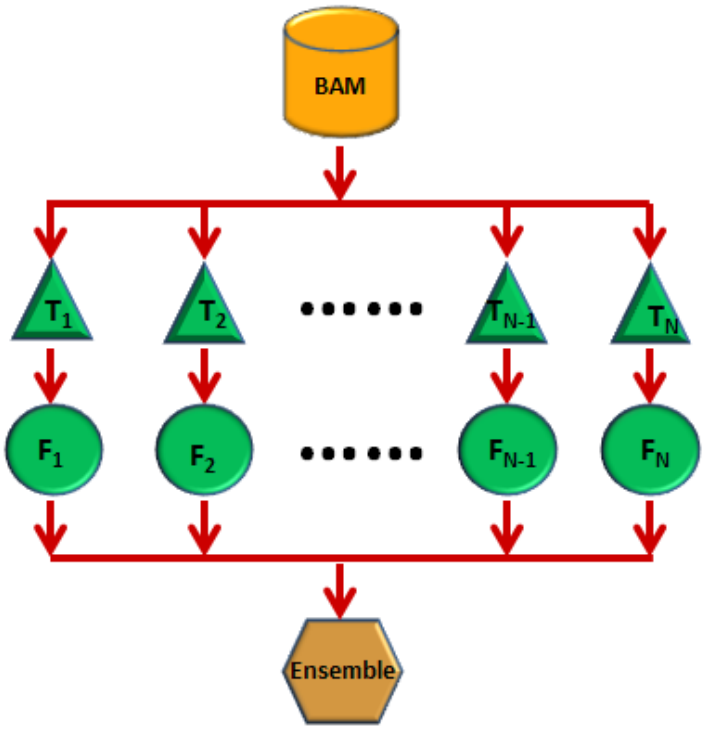


Figure 2.1. Schema of Snipea ensemble and integration

Since different variant callers generate different output file format, we used the Variant Call Format (VCF) 4.2 format (Danecek, Auton et al. 2011) as the universal standard and integrate those results based on the "primary key" field of chromosome number and coordinate positions of SNVs. We created index for each Snipea run for potential large quantity of SNVs called in a cancer sample. Snipea not only annotates the functional

impact of SNV using mainstream annotation tools such as Snpsift, SNPEFF and ANNOVAR but also crosstalk with public large-scale genetic variation databases (E.g. dbSNP, dbNSFP and COSMIC). The annotation was performed in batch using PBS (Portable Batch System) script jobs and fully automated and in the end all those information was incorporated into the final comprehensive output file.

Snipea calculates the authenticity and statistical significance that an SNV is causative for a query disease and hence provides a means of prioritizing candidate SNVs. The ranking of SNVs is based on the integration of numerous sources of information (including the consensus of callers, functional impact of SNV, quality of the data and public database record) using a weighted sum approach.

Based on previous literatures (Roberts, Kortschak et al. 2013, Wang, Jia et al. 2013, Spencer, Tyagi et al. 2014, Xu, DiCarlo et al. 2014), discussion in bioinformatics forum (http://seqanswers.com/ and https://www.biostars.org/) and our in-house testing at TGen, we selected Seurat, Strelka and Mutect (Saunders, Wong et al. 2012, Christoforides, Carpten et al. 2013, Cibulskis, Lawrence et al. 2013) as our input variant callers to Snipea. All three callers are widely used in research community and publicly available software tools, but no tool seems have completely satisfactory performance.

A brief workflow is shown in figure 2.2.

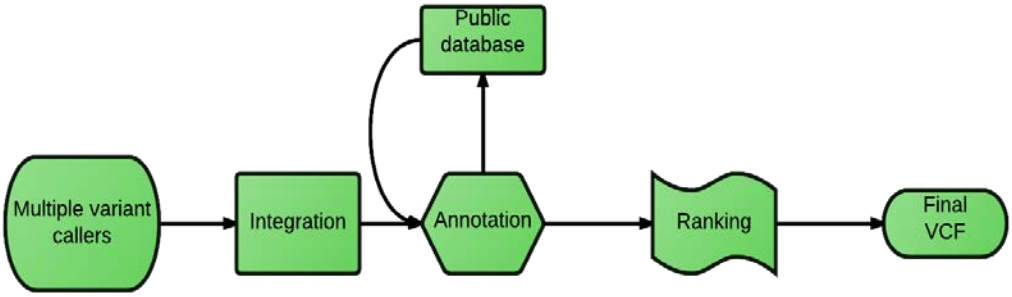


Figure 2.2. Overview of the Snipea workflow. The process begins with sorted output of somatic SNV and INDEL calls from selected tools and ultimately produces a prioritized and annotated comprehensive file in variant call format

2.2.2 Ensemble and Integration

The first step in Snipea is to integrate results with various output formats into one standard universal format. Almost all SNV callers have the common columns "Chromosome" and "Position", in this case, those two fields are used as primary key to create index for each record in output files. The integration is processed in the order of users' input and the count of callers for each position is tracked and stored in a new column named "callers_count". A special case would be the heterogeneous SNV changes at one position where two alternative nucleotide changes may then be called by different tools. The reference allele and the alternative allele are further used to create two distinct records. Therefore, we will have two or more lines for a position with more than one nucleotide alteration.

I processed SNV and small INDELs with two separate functions due to the different length of altered nucleotide length. As Strelka seems to have a better performance of calling the INDELs, the prevalence for the INDELs has been assigned to Strelka. For the SNVs, the prevalence is in the order of Seurat, Strelka and Mutect.

Each caller measures the quality of reported variants in a different way. The consensus quality score is calculated using the joint probability of a somatic variant and a specific genotype in the normal sample as well as the mapping/base quality in the position. The mapping and base quality was used to filter out data with low confidence calls and the mean average of joint probability was taken and then converted back to Phred scale score using formula $Q = -10 \times \log_{10}(\text{Prob.})$.

The final file would be in the VCF format which is a text file format for storing genotype data and gene sequence variation. Each VCF file generally has three parts: it starts with Meta-information lines (lines beginning with "##"); then it is followed with one header line which contains the basic column names (beginning with "#CHROM"); the rest would be the actual data records contain genome variants. An example VCF file is shown in Appendix A. Finally, an R program is run to automatically generate Venn diagrams which depict the overlap and unique SNV and INDEL calls for those input variants callers.

2.2.3 Prioritization

The ranking of SNVs is based on the integration of numerous sources of information (including the consensus of callers, functional impact of SNV, quality of the data and public database record) using a weighted sum approach. Specifically, the ranking algorithm takes into account the callers' agreement, clinical impact (protein-altering aberrations such as non-synonymous SNV, Stop code gained/loss or Frame shift mutation), quality score from each tool and public database status (listed in COSMIC, etc.).

Snipea's ranking Algorithm is based on final weighted score (WS) and implemented using pseudo codes as follows:

1. Measure the Callers agreement. WS=100* Callers_count

2. Add Protein altering (Non-synonymous, stop code gain/loss, frame shift) index. If SNV has any functional impact on protein change, for callers_count = 1, WS = WS+50; for callers_count > 1, WS = WS+100.

3. Add Consensus Quality scores. Consensus quality score is calculated based on quality score from each tool as elucidated above and the range of consensus quality score is from 0 to 60. WS = WS+ Consensus quality score.

4. Add Public database status score. If SNV is stored in large-scale public cancer database, WS = WS+40.

5. Normalize to 100 scale and sort. Final WS = WS/5. Sort the output file based on the final weighted score in the descending order.

2.2.4. Annotation

Snipea still applies current annotation tools such as Snpsift, SNPEFF. It first sorts the VCF to makes it compatible with Genome Analysis Tool Kit (GATK) (McKenna, Hanna et al. 2010) format. Then it adds annotation from SNPEFF (version 3.6c) using ensemble Gene Transfer Format (GTF) reference file (version 37.74).

Lastly it seeks information from various public databases for each SNV record: dbSNP (The Single Nucleotide Polymorphism database) database (version 137 corresponding to Hg19/GRch37 Assembly); National Heart, lung and Blood Institute (NHLBI) Exome sequencing project (ESP) SNP (ESP6500SI-V2_snps_indel.vcf); 1000 genome project (1000G, phase 1 high confidence snps); COSMIC (Catalogue of somatic mutations in cancer) database (Cosmic coding mutation version 66) and dbNSFP (database of human non-synonymous SNVs and their functional predictions and annotations) database (version 2.4). All the information is appended to the “INFO” column in VCF file and could be easily used to query or filtering during downstream analysis.

2.2.5. Prerequisite and implementation

The assumption for Snipea is that the inputs files contain filtered variants, which will minimize the number of variants to process. If users would like to change to their own filtering parameters, they could modify them in the "Snipea.SNV_filtering.sh". Users also could use “Snipea.main.sh –-help” for examples of expected files.

The prefix name for the output Snipea file is by default based on the Seurat's input filename but could also be defined by users. The prefix is any string of character up to the first dot encounter into the filename. For instance: if the input filename is << my_fileName_preFix.passed.filterd.snvs.vcf >>, the prefix will be: <<my_fileName_preFix>>.

To better call genotype from variants data, we define thresholds for genotypes as follows: Homozygote reference if allele fraction (AF) <= 0.200, heterozygote if 0.200 < AF < 0.900, and Homozygote alteration if AF >=0.900. There are two tiers of information in Strelka, we only use the more stringent tier1 values to calculate allele frequency and genotype. During the implementation, we also created several customized flags and tags. All of those extra fields are compatible with VCF format and have been added to the VCF Header.

Snipea requires the following tools/software to be installed: basic Linux commands including cat, case, cd , cut, sed, grep, awk, sort, uniq, getopt, wc, mv, cp, head , pwd, echo, for while, if, bash, read, mkdir, wait, do, rm, local; Portable Batch system (PBS) module; bedtools version 2.17 or up; R version 2.15.2 with features TIFF, PNG enabled; R package Venn Diagram 1.6.5 or up, Vennerable 2.2 or up, gplots 2.11.0 or up.

To implement Snipea, refer to the following example with the mandatory parameters:

```
bash DIR_INSTALLATION/Snipea.main.sh \
--output DIR_OUTPUT \
--seu "Input.seurat.vcf" \
```

--slksnv "Input.strelka.passed.somatic.snvs.vcf" \

--slkindel "Input.strelka.passed.somatic.indels.vcf" \

--mtcsnv “Input.MuTect.filt.vcf"

2.2.6. Validation data and Pre-processing

The establishing of SNV gold standards from actual patient samples is challenging for multiple reasons. For example, benchmarking is generally resource intensive and methods used to estimate 'ground truth' from validation data may still exhibit sources of error (such as machine error, inappropriate threshold) in despite of independent technology (such as q-PCR) or higher depth of coverage (such as deep sequencing). The performance of the proposed Snipea approach is thus determined using a synthetic validation dataset from International Cancer Genome Consortium (ICGC) Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge.

ICGC DREAM somatic mutation calling challenge utilized BAMSurgeon (https://github.com/adamewing/bamsurgeon) tool for adding synthesized mutations to BAM files (Ewing, Houlahan et al. 2015). The data synthesized is originally from real tumor and normal samples and consists of millions of short reads with around 83 nt. Briefly, for each synthetic sample, I randomly sampled a deeply sequenced (60x-80x genome coverage) BAM file into two mutually exclusive subsets of about equal size (30-40x). I then applied BAMSurgeon to generate a non-overlapping spectrum of mutations, randomly select mutation in known cancer-associated genes and add to one of the sub-BAMs. The BAM to which the synthetic mutations were added turns into the “tumor” and the other sub-BAM becomes “normal”. In this manner, complete ground-truth is known for each dataset and we could evaluate the performance of various tools using those gold standard variants.

I downloaded data through GeneTorrent client, an open-source software developed by Annai Systems from the website (https://dream.annailabs.com/cghub/data/analysis/download/*). SNV and INDELs were called using three aforementioned variant callers using manual recommended parameter setting. All those computationally intensive algorithms were run in a cluster high performance computing (HPC) environment.

2.3. Results

2.3.1. Comparison of accuracy among Snipea and individual variant callers

Our first goal was to generate an ensemble and integration format of SNVs from all input variant callers and evaluate the performance of Snipea using a validation dataset. For this purpose, we presented the snapshot of output from each individual caller and their integrated format. As shown in figure 2.3 through 2.5, each caller has different output and all that information was integrated into a universal format in figure 2.6. Some notable differences reside in the "Info" column which exhibit detailed statistics for this specific mutation in tumor and normal samples.

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  Sample_NTLH_0001_1_P
B_Whole_C5_A1S5U_T00027     Sample_NTLH_0001_1_BR_Whole_T4_A1S5U_T00026
1       13302   rs180734498     C       T       .       REJECT  DB      GT:AD:BQ:DP:FA  0:12
,1:.:11:0.077       0/1:18,2:32:20:0.100
```

Figure 2.3. Snapshot of Mutect output

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NORMAL  TUMOR
1       2121071 .       G       C       .       PASS    NT=ref;QSS=57;QSS_NT=57;SGT=GG->CG;S
OMATIC;TQSS=2;TQSS_NT=2     DP:FDP:SDP:SUBDP:AU:CU:GU:TU    54:1:0:0:0,0:0,1:53,54:0,0
92:3:0:0:0,0:26,28:63,66:0,0
```

Figure 2.4. Snapshot of Strelka output

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
1       254680  .       A       G       40.20   PASS    AR1=0.000;AR2=0.625;DNA_ALT_ALLELE_F
ORWARD=2;DNA_ALT_ALLELE_FORWARD_FRACTION=0.667;DNA_ALT_ALLELE_REVERSE=3;DNA_ALT_ALLELE_REVER
SE_FRACTION=0.600;DNA_ALT_ALLELE_TOTAL=5;DNA_ALT_ALLELE_TOTAL_FRACTION=0.625;DNA_REF_ALLELE_
FORWARD=1;DNA_REF_ALLELE_REVERSE=2;DNA_REF_ALLELE_TOTAL=3;DP1=10;DP2=8;LN=1;MVBQ1=30.5;MVBQ2
=30.5;MVC1=20.5;MVC2=24.0;MVMQ1=22.0;MVMQ2=22.0;PILEUP1=AAAAaaaaaa;PILEUP2=GaGAgagg;TYPE=som
atic_SNV
```

Figure 2.5. Snapshot of Seurat output

The final output VCF for Snipea includes various information for each caller and they are separated by a semicolon ";". Allele Pileup are also extracted and incorporated into Seurat field and this could be viewed as IGV visualization for this position in text format.

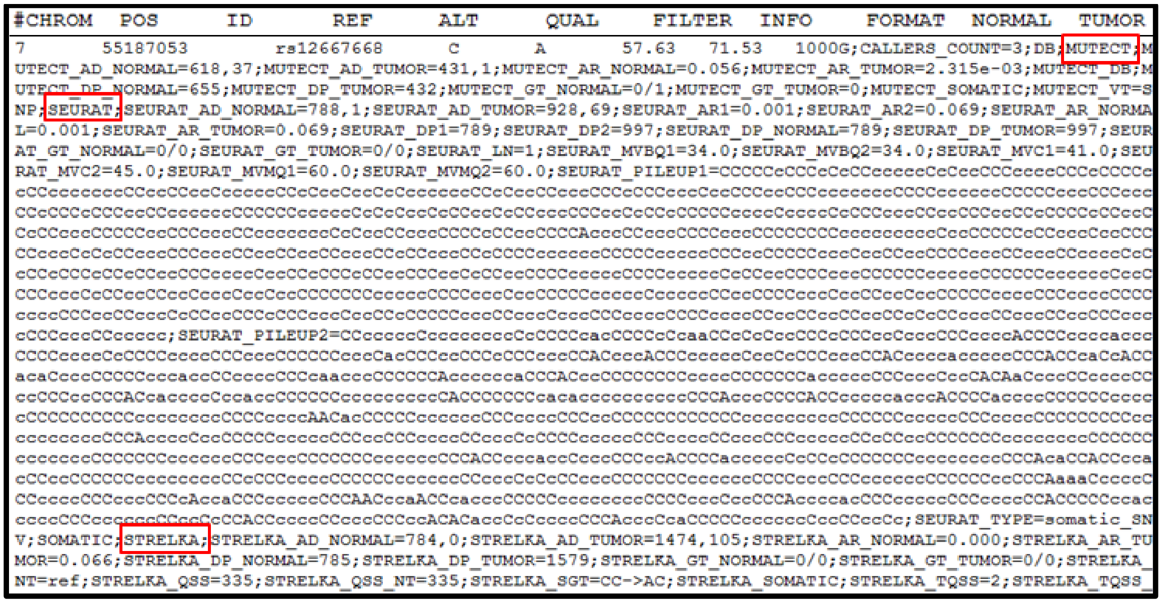


Figure 2.6. Snapshot of Snipea output. The red boxes highlighted the separated tools; additional information is followed by each tool name.

Next we further examined the performance of Snipea and explored the trade-offs among sensitivity, the false positive rate and F-measure for those methods. We assessed the sensitivity and false positive rate of each variant calling algorithm on each synthetic set using the gold standard variants provided by ICGC DREAM SNV challenge. Sensitivity was calculated as follows:

$$\text{Sensitivity (Recall) (\%)} = \frac{\text{Gold starndard variants detected}}{\text{Total variants}} \times 100$$

False positive rate per Mb (Xu, DiCarlo et al. 2014) was calculated as follows:

$$\text{FPR } (\text{Mb}^{-1}) \text{ (False Positive Rate per Mb)} = \frac{\text{FP}}{(\text{FP} + \text{TN})} \times 10^6$$

F-measure was calculated using harmonic mean of precision and recall as follows:

$$\mathrm{F-measure} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{(\mathrm{Precision} + \mathrm{Recall})}$$

The summary is listed in Table 2.1. As we can see, for both set1 and set2, the highest F-measure was achieved by Snipea. Mutect achieved higher sensitivity in both sets, but at the expense of a much higher false positive rate.

	Set1						
	TP	**FP**	**FN**	**FPR(Mb^{-1})**	**Precision**	**Recall (Sens.)**	**F-measure**
Mutect	3440	4040	224	134.6485	0.4599	0.9389	0.6174
Strelka	3168	1248	496	41.5983	0.7174	0.8646	0.7842
Seurat	3287	3875	377	129.1500	0.4590	0.8971	0.6072
Snipea	3396	1251	268	41.6983	0.7308	0.9269	0.8172
	Set2						
Mutect	4213	5312	444	177.0353	0.4423	0.9047	0.5941
Strelka	3643	501	1014	16.6997	0.8791	0.7823	0.8279
Seurat	3928	4427	729	147.5449	0.4701	0.8435	0.6038
Snipea	4025	1849	632	61.6295	0.6852	0.8643	0.8336
	Combined						
Mutect	7653	9352	668	311.6362	0.4500	0.9197	0.6044
Strelka	6811	1749	1510	58.2966	0.7957	0.8185	0.8069
Seurat	7215	8302	1106	276.6568	0.4650	0.8671	0.6053
Snipea	7421	3100	900	42.5315	0.8406	0.8089	0.8245

Table 2.1. The precision and recall of three variant callers and Snipea.

We then examined the performance trade-offs by varying cutoffs for consensus quality score, which reflect the joint probability estimate of a somatic variant and a normal reference by three callers. The ROC-like curves summarizing sensitivity and specificity were listed the figure 2.7 & 2.8. In dataset 1, Snipea achieved significantly better false positive rate than Mutect and Seurat, and also better sensitivity than Strelka. On the other hand, in dataset 2, although Snipea has a slightly higher false positive rate than Strelka, the sensitivity is still much better than Strelka. This result further confirmed that variant-calling is still immature field and there is unlikely a single best caller. We show

that Snipea could integrate outputs from several callers, learn from their discrepancies and have a robust balance of sensitivity and specificity to some extent.

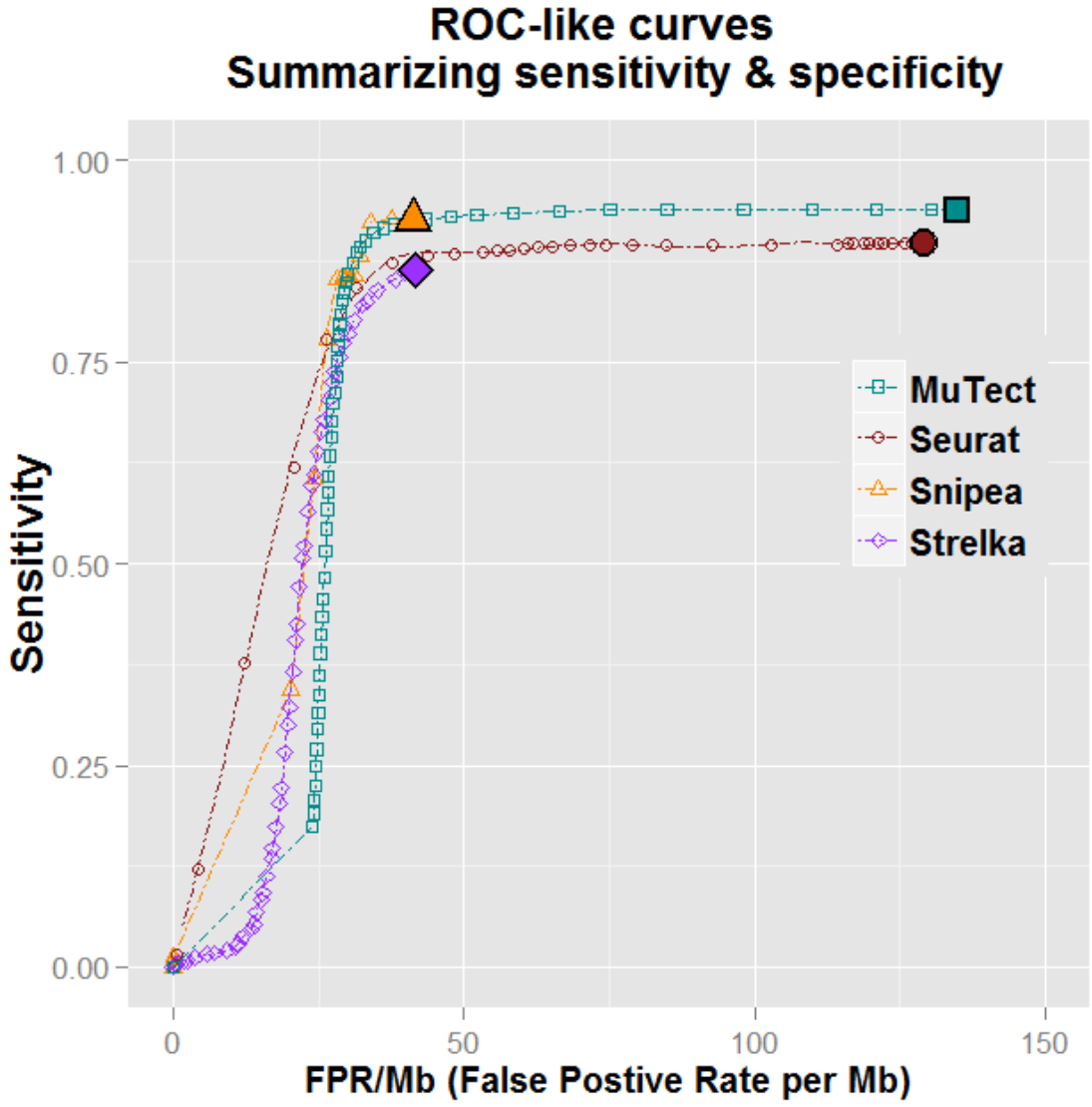


Figure 2.7. ROC-like curves summarizing sensitivity and specificity of Mutect, Seurat, Snipea and Strelka in dataset 1 of the ICGC DREAM challenge.

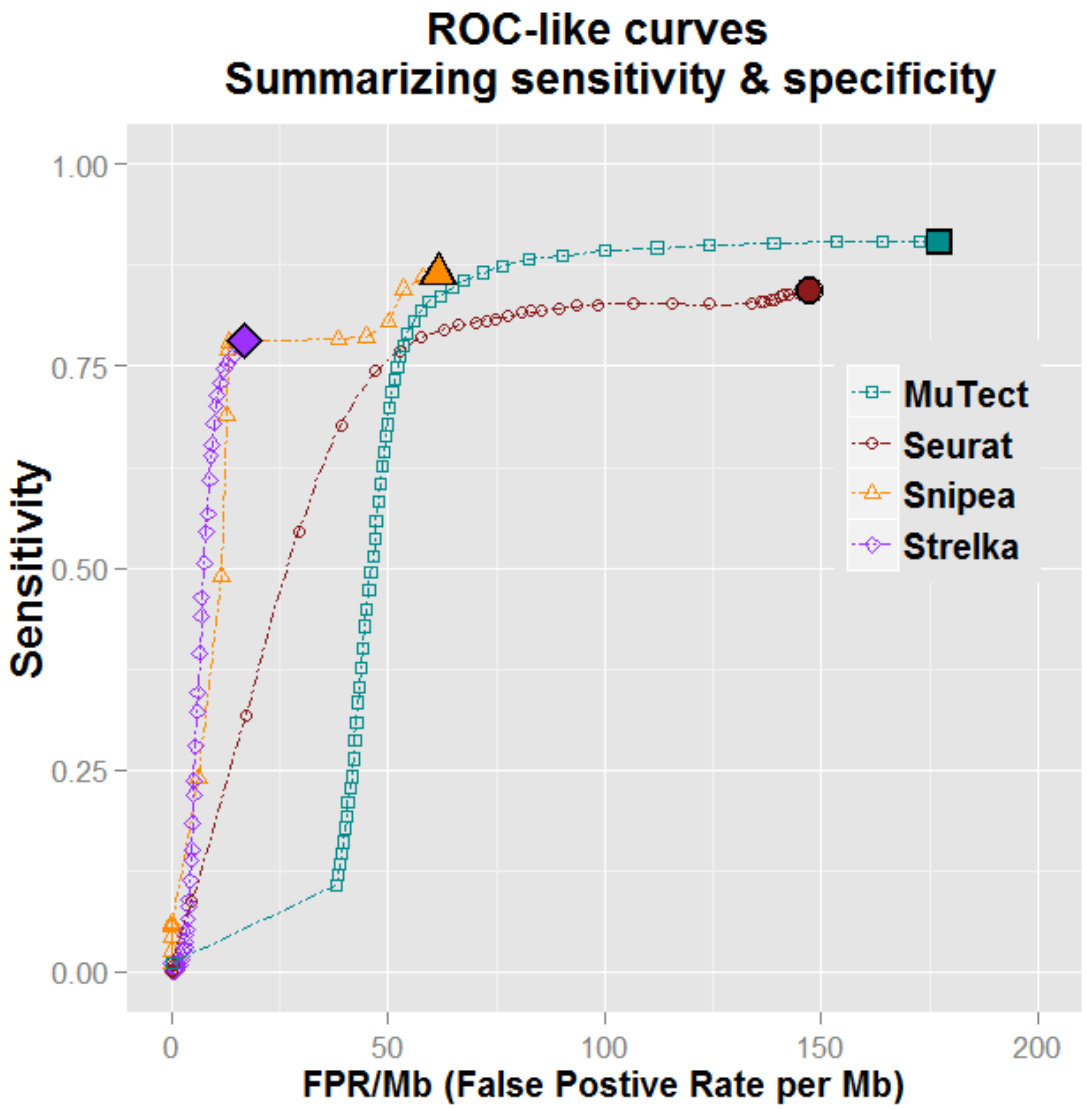


Figure 2.8. ROC-like curves summarizing sensitivity and specificity of Mutect, Seurat, Snipea and Strelka in dataset 2 of the ICGC DREAM challenge.

A Venn diagram of the mutations will be automatically produced for each Snipea run. Figure 2.9 shows an example of such Venn diagram. The counts of mutations detected by three callers were also listed in the diagram.

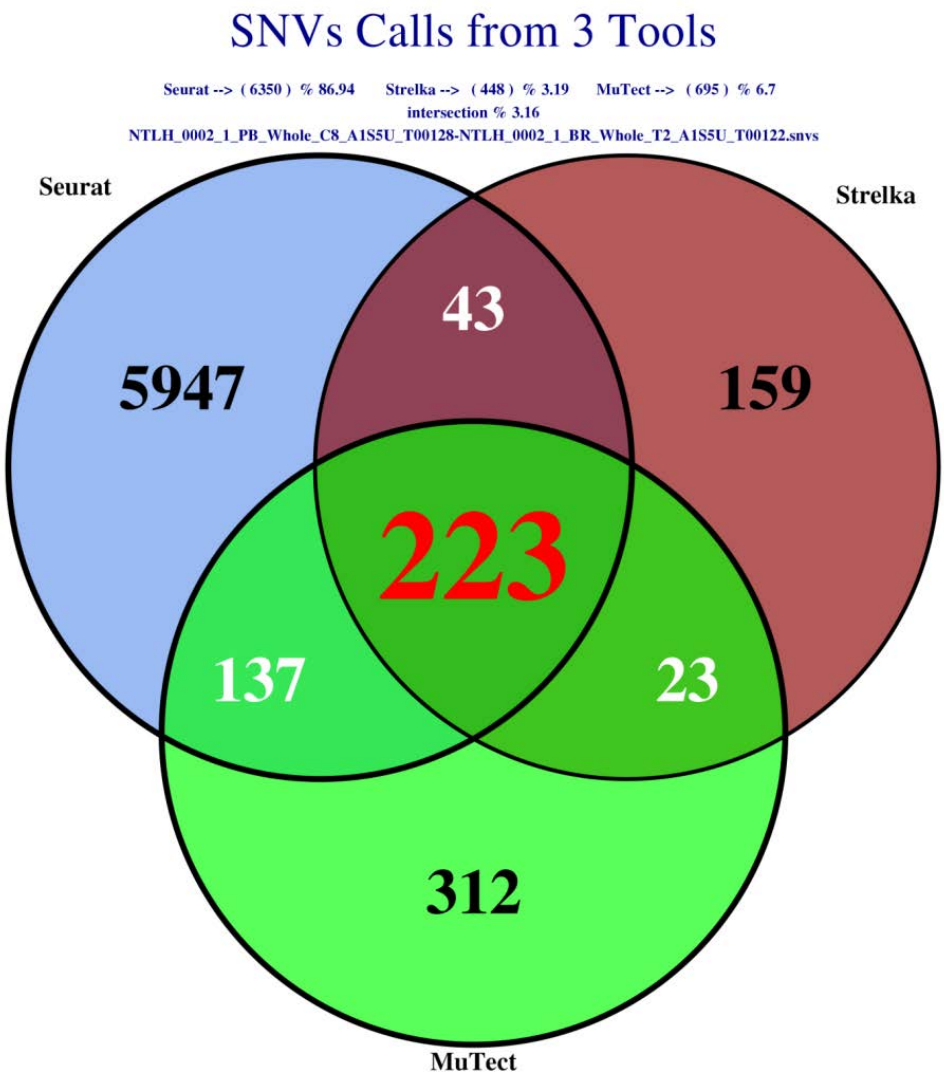


Figure 2.9. Venn diagram showing the overlap of SNV calls from three callers.

Overall, our analysis demonstrated that the algorithms differ in sensitivity and specificity but Snipea acts like the buffer solution and achieves robust results. Users therefore could select their algorithm depending on which performance characteristics is the highest priority. In research studies, users could rely on Snipea to get the balanced performance characteristics. In clinical setting, tools with lower false positive rate might be preferable than others.

2.3.2. Ranking and Annotation

Snipea generates high ranks for many known important genes in GBM (TP53, PTEN, NF1, PIK3R1, ERBB2, EGFR and RB1) (Verhaak, Hoadley et al. 2010), we regard this observation as a proof of concept for our algorithm.

Figure 2.10 indicates an example of PTEN small deletion in chromosome 10. Since we did filtering during the Snipea processing, the status of the "FILTER" column in the VCF file will always be "PASS". We decided to replace the column with the final weighted score. Specifically PTEN gene was called by all three callers and predicted to have a frame-shift protein change function; it also has a high consensus quality score from all three callers and has been reported in COSMIC before. Therefore, it ended up with a 99.61 high weighted score and ranked high among all SNVs.

Snipea still applies current annotation tools such as Snpsift, SNPEFF and ANNOVAR, which will list the gene symbol, splicing region (exonic/intronic) and predict the SNV effects and functions. Moreover, Integration with public archives will not only enable us to know whether the SNV has been identified before but also provide us with a more comprehensive annotation. For example, dbSNP would include information from other sources of information at NCBI such as GenBank, PubMed, LocusLink and the Human Genome Project data; COSMIC will combine curation of the scientific literature with tumor sequencing data from the Cancer Genome Project; dbNSFP will compile prediction scores from four popular algorithms (SIFT, Polyphen2, LRT, and Mutation Taster), along with a conservation score (PhyloP). Those scores would be obtained in a high-throughput fashion to evaluate SNVs function in a short time.

Moreover, the annotation and ranking algorithm is independent of variant calling tools and implemented as separate functions. In this case, we could easily and continually update the various versions/release of biological database. However, it is thus vital to

record the version information in our final output so that if needed, previous results could be checked and validated retrospectively.

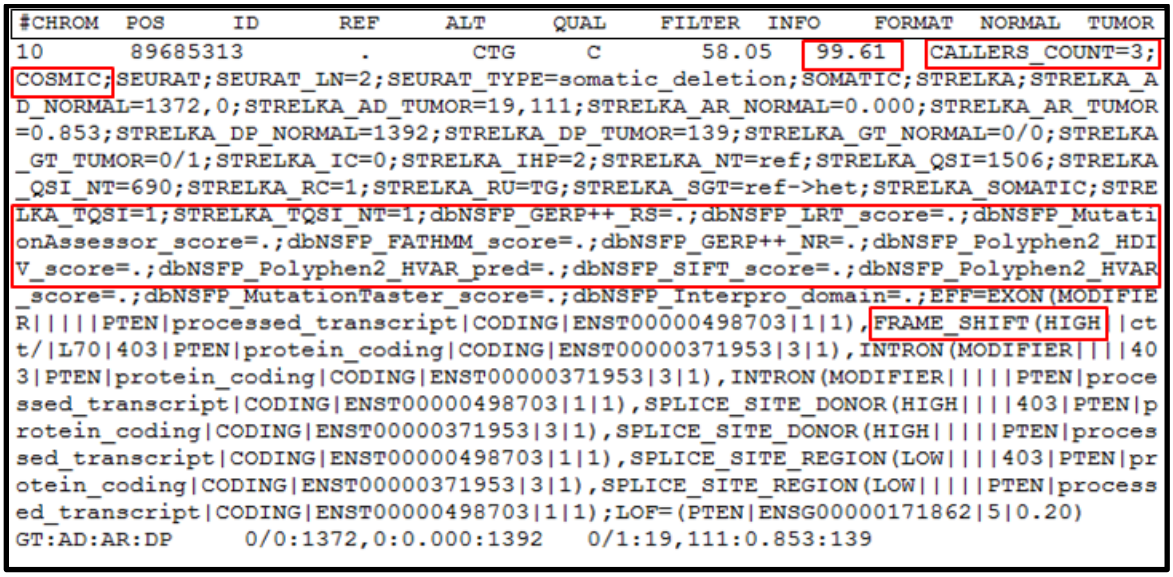


Figure 2.10. An example SNV call as PTEN deletion, content in red boxes shows the weighted score, callers counts, information extracted from large-scale public database and functional prediction.

Snipea is implemented in UNIX shell and R and is supported to run on UNIX/Linux based platforms. The algorithm, UNIX shell scripts and R code are available for all non-commercial users via https://github.com/spengInformatics/Snipea.

2.4. Conclusion and Discussion

As the key upstream step in cancer sequencing data analysis, identifying SNVs and INDELs with balanced sensitivity and specificity is critical for downstream analysis and treatment planning. One conclusion that can be drawn from Snipea tool is that the "one model fits all" approach is likely not optimal for cancer samples due to their complexity and heterogeneity. We demonstrated that Snipea has superior accuracy and specificity than the result of each single tool. Moreover, the ranking and annotation functions shed

light into the functional impact and could be used to pinpoint potential “drivers” of specific tumor types.

The limitations of Snipea include: 1. Currently it is only designed for cancer in general, it does not take specific tumor types into account. 2. Snipea is mainly designed for SNVs with known biological knowledge and thus might be biased for discovery purpose. 3. Snipea will usually run robustly if all input results from various callers have a decent performance but it may still be vulnerable to extreme outlier tool output. However, those limitations could be addressed with further development and version update.

Although Snipea is currently written for integration of three variant callers, it could be easily extended to more or different callers based on users’ preference. In addition, more databases or updated version could also be incorporated into the Snipea output without difficulty. The functions and methods in Snipea were deliberately divided into independent function classes which are implemented in a "cascade" type fashion. This ensures the easy update of different variant tools or annotation database version since the modified functions have little dependency on others. Furthermore, the ensemble and integration method could also be extended to other analysis other than SNV detection. Data from different regional parts of tumor, different stages (such as primary V.S. relapse) of tumor or different patients could also be integrated to garner biological insights.

Snipea is mainly designed for SNVs involved in clinical relevance or decision making; therefore it is more suitable for known biological knowledge than discovery purpose. However, users shall choose the strategy based on priority of their performance characteristics. For example, CLIA certified labs would probably pay more attention to the accuracy - especially specificity - for clinical samples. They would like to discriminate true somatic SNV from alternate alleles from normal samples and discard SNV with

insufficient evidence. While on the other hand, during drug discovery process, false negative (FN) is less tolerated compared to false positive (FP). FP could also be eliminated later in drug efficacy step but FN would never be rescued and a potential new medicine would be missed. One solution to such trade-off dilemma is to produce output by two models, one is standard default and the other is high Confidence (HC). The final somatic variant score is calculated from the probability of allele frequency that is unequal in tumor & normal samples given the observed sequence data. In this case, we may define P(somatic) > 0.95 as standard and P(somatic) > 0.999 as HC. The Phred-scale Quality score that converts P(call in Alt is wrong) becomes an intuitive criteria. Higher Phred QUAL score indicates higher confidence call. In the end, based on their needs, users are able to choose list of variants based on their stringency and confidence.

Through many ongoing projects such as The cancer genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), we start to know that same SNV might play different role in different tissue/tumor types and thus have distinct regulatory mechanism or signaling pathway (Alexandrov, Nik-Zainal et al. 2013, Weinhold, Jacobsen et al. 2014). With the help of such knowledge, we might also to able to adjust the ranking algorithm to account for tissue types or tumor types. We even can take a “crowd-sourcing” approach to adjust ranking algorithm and assign credibility to each tool based on user’s feedback from research community.

Snipea will usually run robustly under the assumption that all input results from various callers have a decent performance. But this might be not always true. Besides all the difficulties we discussed above for somatic SNV calling in cancer samples, the poor consensus between the predictions made by each tool may be explained in part by difference in the stringency settings (even by default). Each tool is sensitive to its own filtering threshold and initial requirement of mapping quality. Based on our manual

check and several previous studies (Cibulskis, Lawrence et al. 2013, Koboldt, Larson et al. 2013), the sources of discrepancies in those callers likely come from the germline variants. A small dissimilarity in stringency settings could cause the “butterfly effect” and potentially lead to large difference in false calling germline variants as somatic ones. To address this, a pilot run might be used to tune parameters for each tool. Specifically, 5 million reads could be randomly extracted from target bam files as pilot bam file, then 20 different parameter combinations for each tool were tried and finally Snipea was applied to get the consensus SNVs set with 10 or more overlap. In this case, using this consensus SNV list as standard, we could determine the parameter setting for each variant caller for every sample. Furthermore, if a candidate mutation found in dbSNP but NOT in COSMIC, it has a higher chance to be a germline variant instead of somatic one. Ranking algorithm perhaps could take advantage of this and give less weight to SNV only reported in dbSNP.

CHAPTER 3

NOVEL APPROACH FOR ACCURATE AND LOSSLESS CLASSIFICATION OF XENOGRAFT SEQUENCING READS

3.1. Introduction

3.1.1. Sequencing in xenograft models

The xenograft model is a well-established investigational platform for studies involved in drug discovery, cancer and stem cell biology, metastasis prediction, as well as response to therapy. Recent growth in the use of patient-derived tumor xenografts (PDX) points to the key advantage of xenograft models: 1. They are biologically stable. When engrafted in to host species, it is usually stable for mutational status, metastatic potential, gene-expression patterns, drug susceptibility and even genetic heterogeneity. 2. They have a short cycle. Generally results can be obtained in a matter of a few weeks in terms of response to therapy. 3. They reproduce the actual environment. Orthotopic xenografts can be appropriately placed to mimic the human organ environment (Joo, Kim et al. 2013). This is especially important for brain tissue in which the blood brain barrier (BBB) plays a vital role in therapeutic decision making.

The advent and prosperity of next generation techniques fuelled an acceleration and expansion of decoding genetic makeup in xenograft models. Global gene expression profiles from RNA sequencing data could be used to examine the similarities between originating tumors and xenografts. Mutation changes (before treatment V.S. after treatment) from exome sequencing data could be used to define drug-sensitivity patterns and explore selection process. Methylation profiling from bisulfite sequencing can be applied to measure modification of DNA methylome and identify upstream transcriptional regulators (Aparicio, Hidalgo et al. 2015).

In a word, biological and molecular characteristics (including genetic, epigenetic or signaling) in a patient tumor could be recapitulated and tested in a massively parallel fashion using xenograft models. Those results in turn will allow for detailed understanding of tumorigenic aberrations and individualized clinical decision making.

3.1.2. Homology between mouse and human

A large variety of animal species including dogs, sheep, pigs, birds, monkeys, zebrafish and of course mice have been widely used to better understand genetic and physiological processes involved in human disease (Agca 2012). However, there is no doubt that mice are the most commonly used mammalian xenograft model in biomedical community, due to their exceedingly well-characterized genetics and low cost. Mice have many advantages over other model organisms: their genetic, biological and behavior characteristics closely reproduce those of humans; a comprehensive collection of genetic and molecular databases are already available and the mice are small and relatively inexpensive which made them easily housed and maintained.

65 million years divergence during evolution does not stray us significantly from the mouse. Over 95% of the mouse (*Mus musculus)* genome is similar to human (Homo Sapiens) genome. A previous mammalian genome project utilized Whole-Genome Sequence and Assembly (WGSA) method (Church, Goodstadt et al. 2009) to show that 95% of the mouse genome could be lined up with a region on the human genome and 99% of mouse genes turn out to have direct counterparts in humans. To date only 300 or so genes appear to be unique to one species or the other.

However, this homology serves as a double-edged sword. Indeed the mouse model can better mimic the biological mechanism in human but it also raises the classification problem – it's harder for us to tell them apart when needed. As shown in figure 3.1, when

a mixture of human and mouse tissue were sequenced, it becomes a challenge for us to separate sequencing reads for each species bioinformatically.

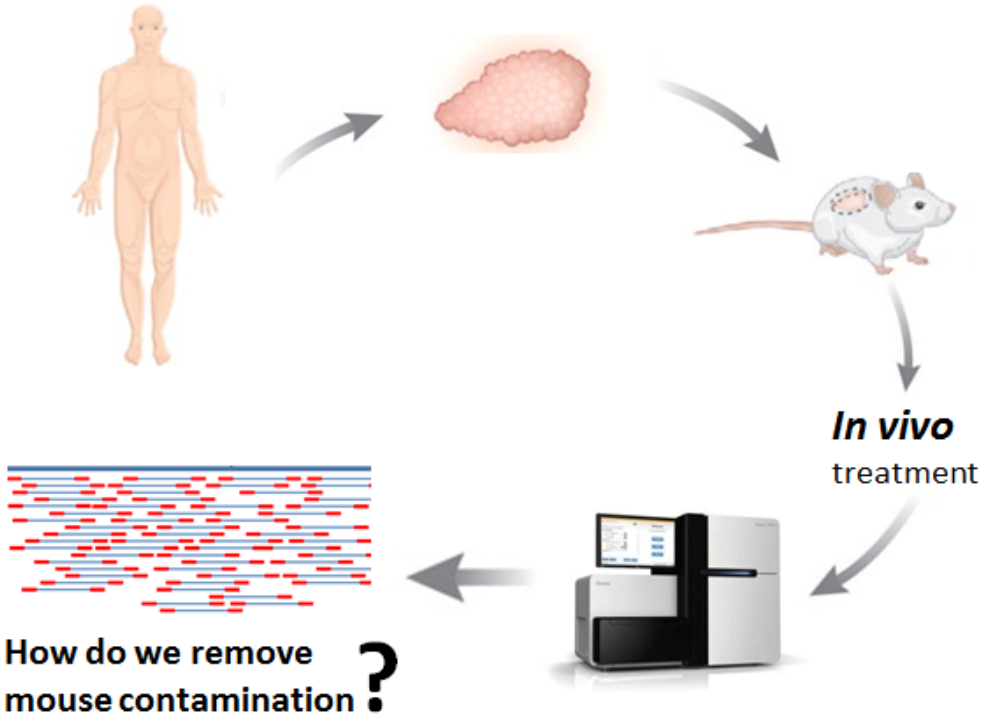


Figure 3.1. Schematic overview of potential sequencing classification difficulty in xenograft models.

3.1.3. K-mer method and its problem

In order to have an accurate representation of human biological events, a K-mer based tool called xenome was proposed to partition the sequencing reads in xenograft models (Conway, Wazny et al. 2012).

K-mer (also known as n-mer/k-tuples/n-grams) is originally used in cryptography or pattern matching for computational sequence analysis. In the context of Bioinformatics, k-mers refer to all the possible substrings of length k from DNA sequencing reads produced by the sequencer. The number of possible k-mers given n possibilities (4

in the case of DNA e.g. ACTG) is n^k, and counting the occurrences of all those subsequences is an important step in many sequence alignment and sequence assembly tools.

The authors described xenome as k-mer decompositions of host and graft reference (Conway, Wazny et al. 2012). The first step is to construct an index structure for both reference genomes; the k could be defined by users but has a recommended value of 25. Then classification was done by taking each coming read and map it to the pre-computed classes (graft, host, both, ambiguous and neither).

Xenome is a simple and straightforward algorithm with decent performance. However, it does have several potential aspects for improvement: First, xenome does not allow any mismatches in the sequence. We know if DNA sequences were generated totally randomly, there is a roughly 25% error rate. To date, next generation sequencing platforms have greatly decreased this percentage, among them Illumina Hiseq and Miseq show the lowest error rate up to around 0.5% (Quail, Smith et al. 2012). But if the error does not occur at the position which is different between host and graft reference genome, such error shall still be regarded as machine error and kept in the downstream analysis. Moreover, the “ambiguous” class in xenome was not further explained by the author and thus usually discarded at the end of classification. The second potential problem with xenome is the management of the “both” class. We already know that the human and mouse share a great amount of DNA sequence, therefore it is expected a significant portion of reads would fall into the “both” category for each run. One option to deal with them would be to simply ignore them, but this would run into the problem of under-estimating the expression of some genes, especially ones with high homology between mouse and human. On the other hand, if we were to keep them as-is, we would obviously run into the opposite problem, of over-estimating the expression of

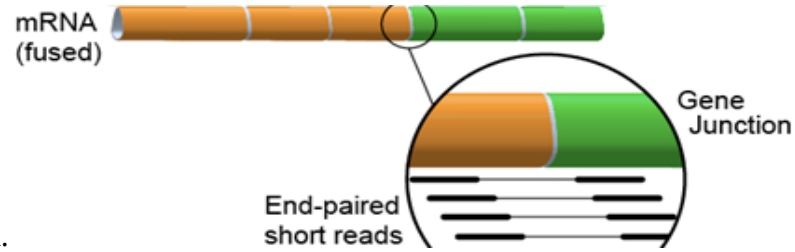
homologous genes. Last, xenome is likely to miss any influential genomic aberrations that do not previously exist in the reference genome. For example, two separate genes could form novel hybrid transcripts called gene fusion. Often the case, such genomic abnormality is novel and the information is not included in the reference genome. In this case, the spanning reads across the “break point” of genes would conceivably classified by xenome into the “neither” category and we would lose the capability to call the fusion genes in the downstream analysis.

3.1.4. Gene fusion detection using RNA-seq

Gene fusions are novel hybrid transcripts that are formed by two previously isolated genes. Gene fusion is a crucial genomic event in human cancer because abnormal proteins made by them appear to be more active than normal versions. Therefore, this abnormal genetic rearrangement can drive the development of cancer and provide an opportunity for potential prognostic tools and druggable targets in anti-cancer treatment. Gene fusion could be caused by several chromosome abnormalities including translocation, chromosome inversion and interstitial deletion (Wang, Xia et al. 2013).

Next-generation sequencing technologies enable the possibility of systematic characterization of cancer cell transcriptome, including the accurate gene expression profiling and the detection of expressed fusion gene products. Since fusions genes were formed by the breakage and re-joining of two isolated genes, RNA sequencing has the potential to discover gene fusion events based on aligned reads matching on both sides of fusion and critical reads spanning the entire break point. Figure 3.2 presents a schematic view of gene fusion, and key RNA-seq reads that span the break point (in blue box).

A.



B.

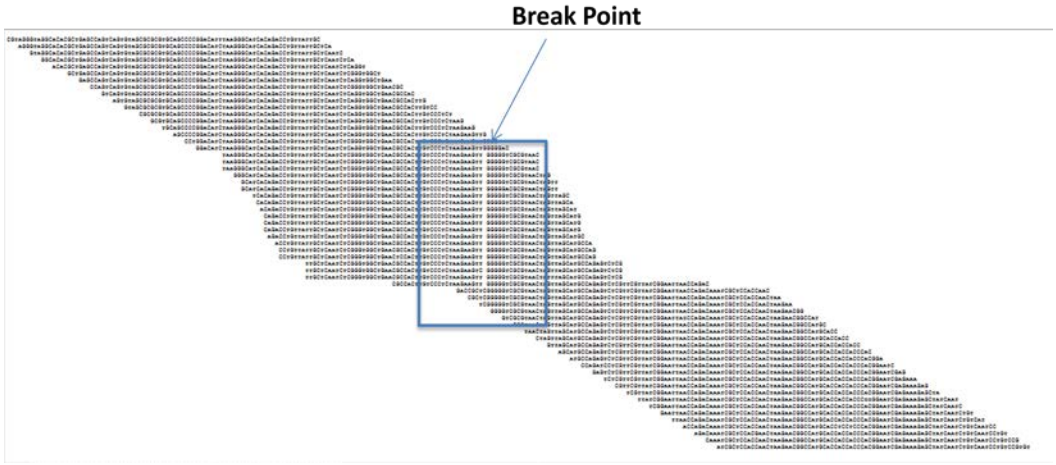


Figure 3.2. A) shows a schematic view of fusion gene. B) shows an actual gene from a sequence alignment point of view. Blue box highlights a fusion break point and spanning reads that support the fusion.

Until recently, fusion genes have been an underappreciated class of genomic anomaly in cancer biology. Since then the field of fusion gene discovery is constantly evolving to garner more biological knowledge.

Gene fusion could be used as a signature to identify new tumor subtypes. Pierron et al. (Pierron, Tirode et al. 2012) observed a new gene fusion between BCOR (encoding the BCL6 co-repressor) and CCNB3 (encoding the testis-specific cyclin B3) and defined a novel subtype of bone sarcoma. The authors confirmed that BCOR-CCNB3-positive cases are biologically distinct from other sarcomas and may be caused by a newly identified gene fusion mechanism. This subtype-specific gene fusion product has revolutionized the diagnostics of sarcoma and has provided new insight into oncogenesis. Gene fusion might also be the driver for pushing cells towards cancer and thus be regarded as novel

drug targets for therapeutic intervention. Columbia University Medical Center reported that a small subset (around 3%) of GBMs acquire tumorigenic chromosomal translocations that fuse the tyrosine kinase coding domains of fibroblast growth factor receptor (FGFR) genes (FGFR1 or FGFR3) to the transforming acidic coiled-coil (TACC) coding domains of TACC1 or TACC3, respectively. They also found that oral usage of an FGFR kinase inhibitor prolongs survival of mice with intracranial FGFR3-TACC3-initiated glioma. In this case, FGFR-TACC gene fusions could help us to potentially identify GBM patients who would benefit from targeted FGFR kinase inhibitor treatment (Singh, Chan et al. 2012).

The presence of contaminating mouse DNA or RNA affects the accuracy of downstream NGS analysis. In addition, currently most NSG techniques still use short-read methodology and physical and biochemical removal of mouse can introduce a significant source of technical bias and usually require large amounts of resources. This leaves opportunity for an in-silico bioinformatic solution to classify reads to their species-of-origin. Previous efforts such as the xenome tool have been made to tackle this issue but remain imperfect. Here, we describe and benchmark another alignment-based strategy to aim for more accurate and lossless classification of xenograft sequencing reads.

3.2. Material and Methods

To this end, we developed a novel alignment-based classification strategy and an algorithm to redistribute the “both” reads that takes into account the relative expression levels of mouse and human genes, gene boundaries, and differing mapping quality.

3.2.1. An Alignment-based method

This proposed approach does not try to reinvent the wheel but takes advantage of maturely developed algorithms specially designed for alignment.

The very first step in analysis of RNA-seq data is the proper alignment (mapping) of the reads to the reference genome, which is complicated by the presence of introns.

Normally the mapping of RNA-seq reads to genes is a straightforward process that involves a sequence aligner (like STAR or TopHat) (Trapnell, Pachter et al. 2009, Dobin, Davis et al. 2013), a reference genome, and a corresponding GTF file that contains the location and structure of known genes. In the case of xenograft samples, the mapping step is complicated by the fact that the RNA is extracted from a mixture of mouse and human cells. When the sequenced reads are mapped to the mouse or human genomes, the reads can either map exclusively to the human genome (human reads), map exclusively to the mouse genome (mouse reads), or map to both genomes (multi-mapping reads).

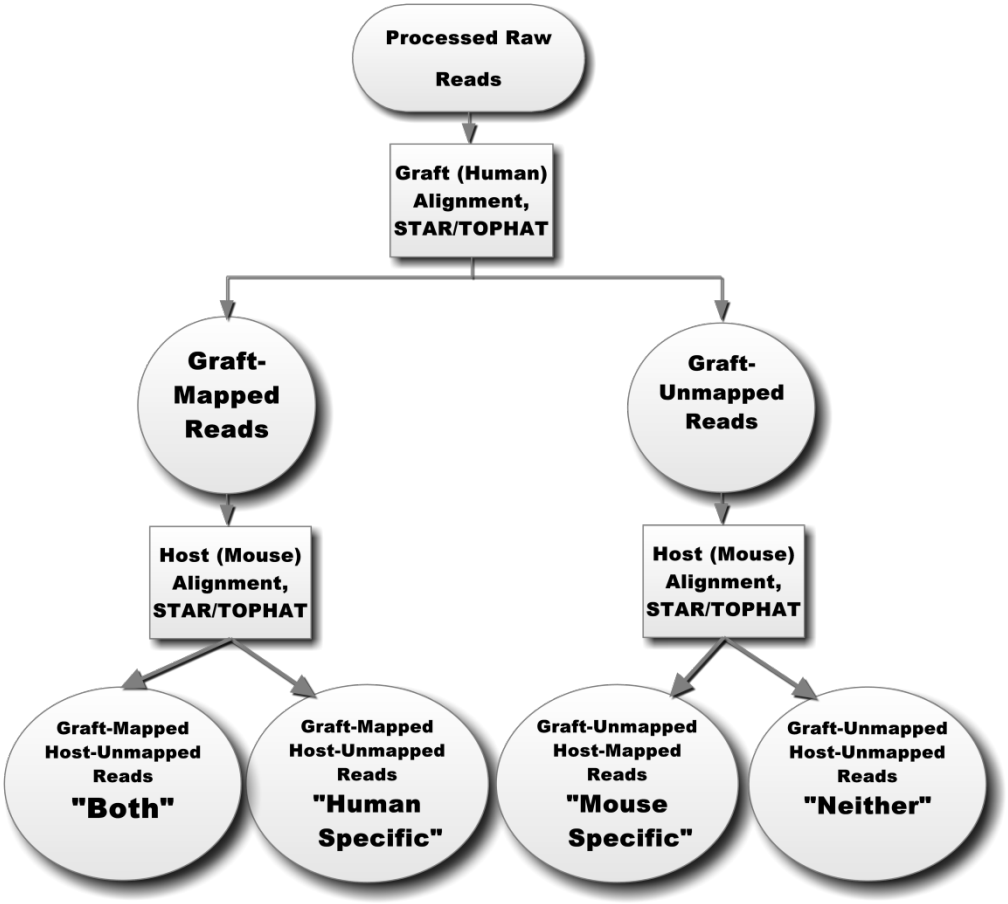


Figure 3.3. Workflow of alignment-based strategy.

Briefly, as shown in figure 3.3, the pre-processed reads were aligned to graft genome (in our case it is human genome) where reads are classified into graft-mapped and graft-unmapped classes. Then those classes were realigned to the host genome (in our case it is the mouse genome) where reads were further divided into host related classes. Specifically, graft-mapped and host-unmapped reads shall fall into the category of "Both", graft-mapped and host-mapped reads shall be regards as "Human specific"; graft-unmapped and host-mapped reads are "Mouse specific", graft-unmapped and host-unmapped reads shall be viewed as "Neither".

$$L(y_j, f_{class}(x_{.j})) = \begin{cases} 1 & \text{if } y_j = f_{class}(x_{.j}) \\ 0 & \text{if } y_j \neq f_{class}(x_{.j}) \end{cases}$$

More formally, we compute the above function for each class as classification score.

3.2.2. Read extraction

Standard formats are needed for storage, processing and communication in NGS analysis. There are a lot of NGS related file formats, the most common ones are: FASTQ and FASTA files for reference sequence and pre-processed sequence data; BAM and SAM files for alignment output; and GTF, GFF and BED files for annotation and feature description; VCF for variants storage and management. FASTQ is a text-based format for storing both DNA sequence and its corresponding quality score. SAM format is a generic format for storing large nucleotide sequence alignments. Usually for each alignment, we use FASTQ files as input and BAM/SAM files as output, mapped and unmapped reads are stored in BAM/SAM formatted files (Li, Handsaker et al. 2009). A table of mandatory fields of FASTQ and SAM files are listed in Appendix B.

The first technical problem for alignment strategy we met is how to extract mapped or unmapped reads from SAM files and locate the corresponding reads in raw FASTQ files. We used bitwise FLAG (as listed in Table 3.1) status in BAM files to separate and extract mapped/unmapped reads. In this case, 0x104 (104=100+4) could be used to separate mapped and unmapped reads in BAM files. Since we need to go back to FASTQ file to extract corresponding reads and make a new FASTQ as further input, we treated reads that mapped to multiple locations as one read. Once the multi-mapping reads are identified, these are temporarily excluded from the analysis and the uniquely mapping reads are counted first. Reads that map exclusively to the mouse or human genomes are mapped to the GTF file using htseq-count (Anders, Pyl et al. 2015).

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

Table 3.1. Bitwise flag of SAM files.

Next we used the field of “sequence ID” to extract relevant reads in FASTQ based on information of BAM files. To make things more complicated, the FASTQ reads are paired which indicates both Read1 and Read2 need to be addressed. Paired-end sequencing is largely used in today’s research since it is more likely to align to a reference and could facilitate detection of novel transcripts, gene fusions as well as repetitive sequence

elements. Since the experiment was done using paired-end sequencing, we also filter reads that are not properly paired, or have an unmapped mate. A Perl script was created to perform aforementioned functions. Since FASTQ and SAM are large files, all jobs were completed in TGen PNAP HPC cluster and Saguaro High Performance Computing cluster computer at Arizona State University (http://a2c2.asu.edu/resources/saguaro/).

3.2.3. Take further care of "Both" Category Reads

Reads that map exclusively to either genome can be handled using normal means, but things become complicated with the multi-mapping reads. One option would be to simply ignore them, but you would run into the problem of under-estimating the expression of some genes, especially ones with high homology between mouse and human. If we were to keep them as-is, we would run into the opposite problem, of over-estimating the expression of homologous genes. The third option would be split the multi-mapping reads between the mouse and human genomes, so we can get a better estimate of their "true" expression levels. In this case, the method takes into account the number of uniquely mapping reads that already map to the genes. The multi-mapping reads are then split between the genes according to the proportion of unique reads in each gene and randomly assigned to each group based on their ratio. For example, if 5 multi-mapping read were perfectly aligned to a mouse gene that had 15 mouse specific reads, and the analogous human gene with 10 human specific reads, then 15/(15+10)*5=3 of the reads would go to the mouse gene, and 2 of the reads would go to the human gene. The new total count would be 18 reads for the mouse gene, and 12 reads for the human gene. Since many methods that use RNA-seq data (like DEseq and EdgeR) require integer counts for the genes, final count was rounded to the smallest following integer using the ceiling function.

3.2.4. Data sets used for performance comparison

40 PDX samples (named from JC001 to JC040) were generated at the Mayo Clinic by transferring primary GBM patient tumor directly into an immune-deficient mouse. Passage of brain tissue was sent to TGen for H&E staining. RNA extraction was done at TGen using QIAGEN AllPrep Kit.

RNA samples were then sent to Department of Systems Biology at Columbia University where pair-end RNA sequencing was completed. Raw data was copied back to TGen for quality check downstream analysis. The transcriptome was sequenced at high depth coverage with average 110 million paired reads in each sample. At TGen, all samples were assessed for overall quality using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and low quality reads were filtered and if needed, hard trimming was performed using in-house script.

Figure 3.4 shows initial check using xenome run output. It summarizes the number of reads that map to either the mouse or human genome in each sample. The reads that uniquely map to the mouse and human genomes are shown in dark blue and light blue, respectively.

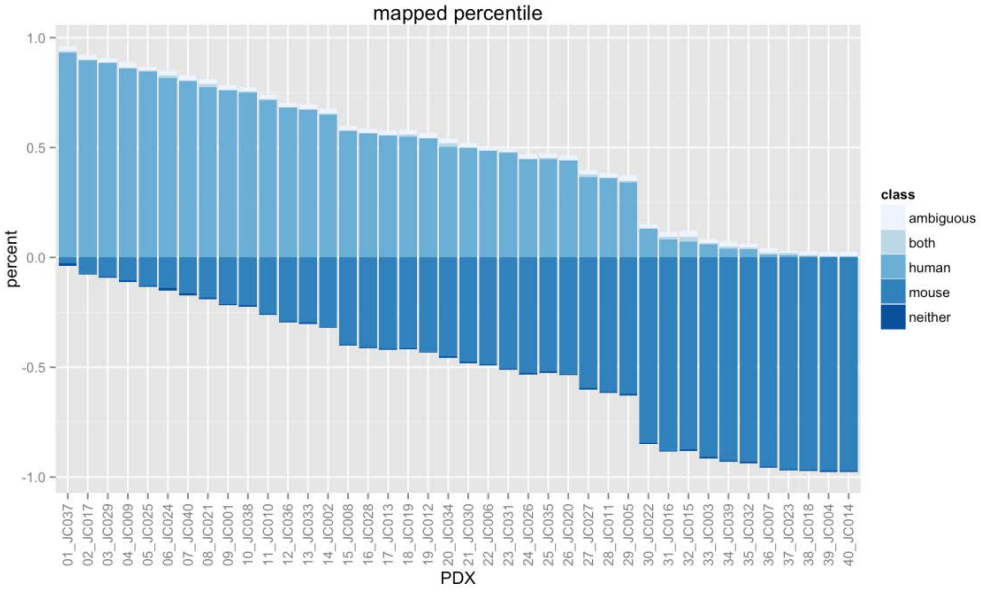


Figure 3.4. The number of reads that uniquely map to the mouse (in light blue) or human genomes (in dark blue).

For most samples, a significantly greater number of reads are assigned to the human genome. This is to be expected for the human xenograft samples. However, for some samples there is a bias toward reads that map to the mouse genome. This might be due to variability in the collection of the samples, and it might be a good idea to analyze the samples separately. But this in turn serves as an excellent test dataset for the comparison between our alignment-based method and K-mer approach. The wide range of human mouse ratio would reduce the algorithm bias towards “pure” samples and achieve a balanced evaluation.

3.3. Results

3.3.1. Accuracy of classification

The alignment strategy generated on average 19.13 GB output in BAM format which is around 110 million reads per sample. First, we checked the correlation of human percentage output between our alignment-based method and current K-mer method. As shown in figure 3.5, it indicates high agreement of both approaches, with Pearson correlation coefficient of 0.99937 and Spearman correlation coefficient of 0.99812. This result suggests that both methods did a decent job to roughly classify xenograft sequencing reads.

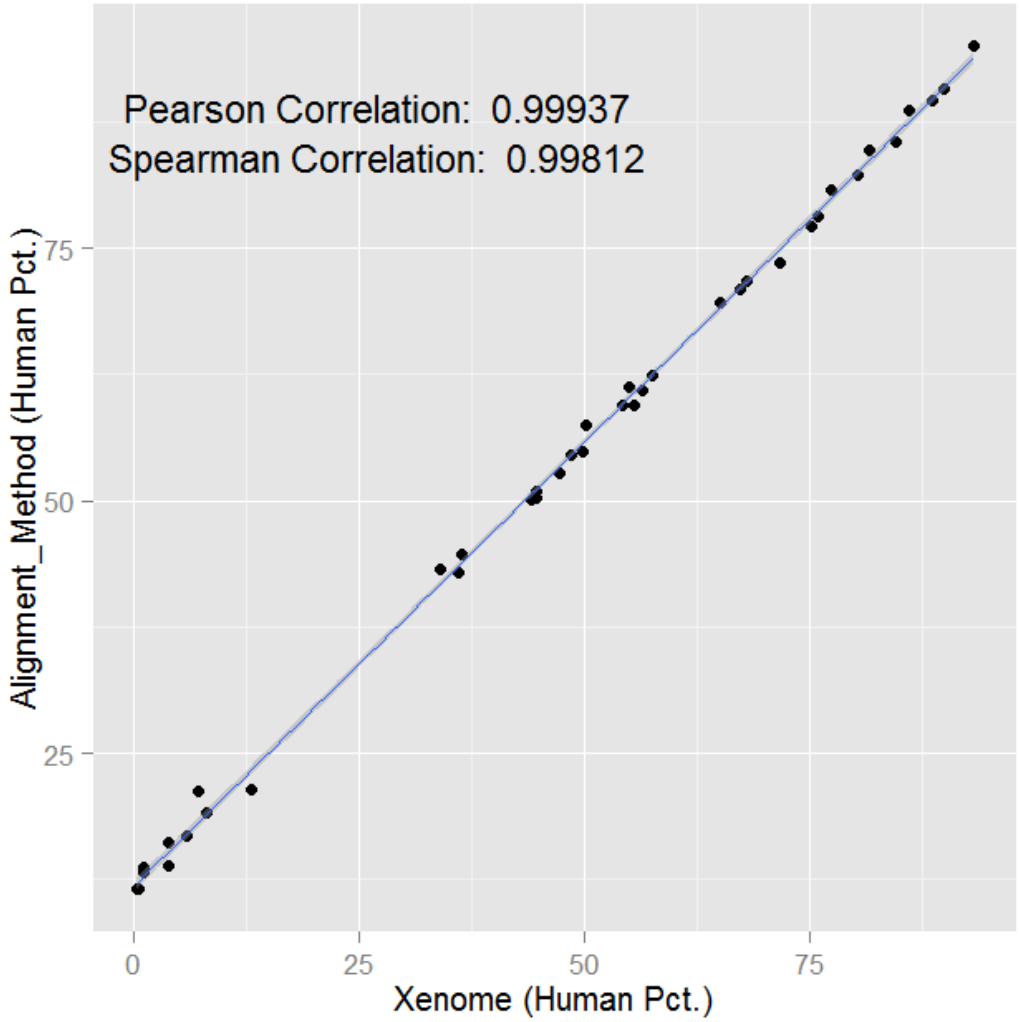


Figure 3.5. Correlation of human reads percentage using xenome and alignment-based methods.

To evaluate the performance of our strategy and compare it to xenome, we used the following evaluation criteria: number of human-related reads, final percentage of aligned reads, pipeline run time, accuracy of read classification and the capability to keep necessary information for downstream analysis. In terms of final alignment percentage and run time (didn't take into account for alignment, since xenome will perform alignment in later step), those two methods achieved very similar performance and we will focus on the other three standards.

Our alignment based method, presented in figure 3.6, assigned on average 6.3 million more pairs to human-related (human only, both classes) reads compared to xenome approach.

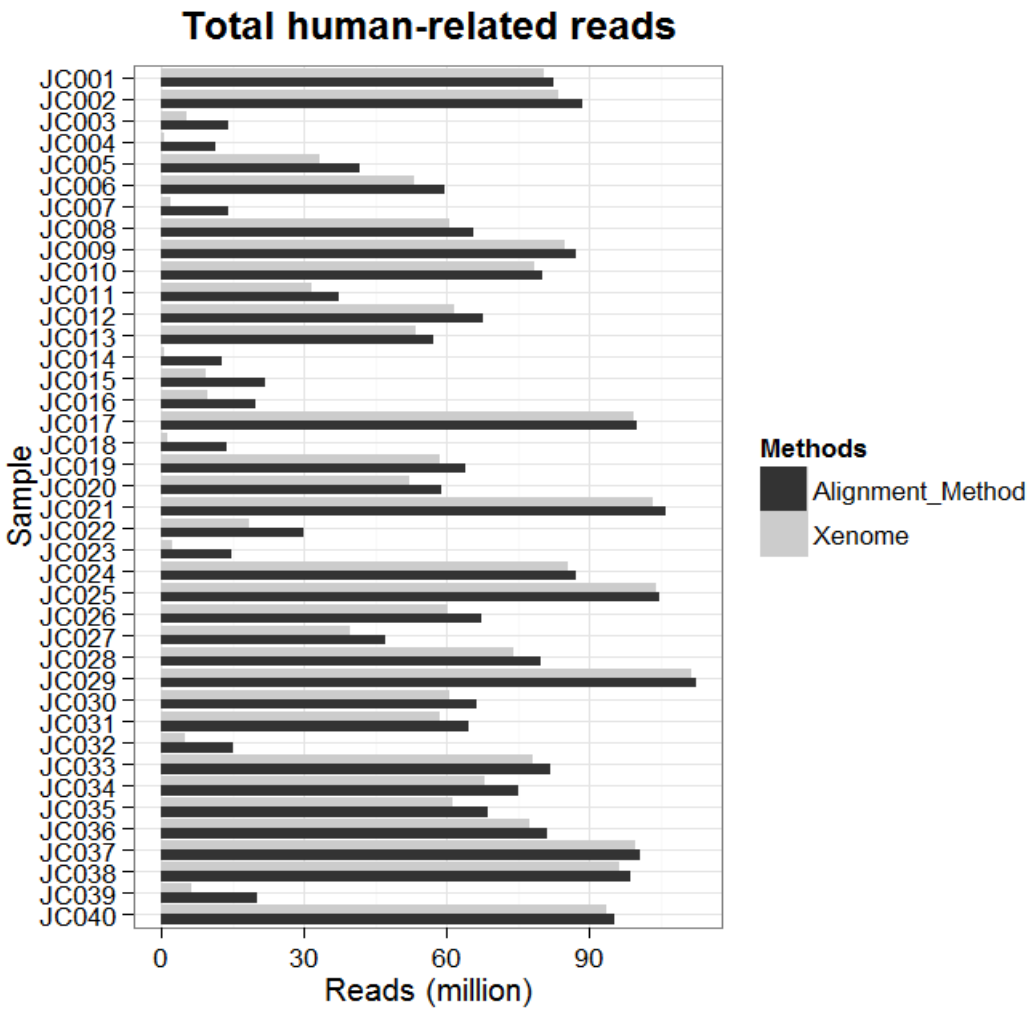


Figure 3.6. Number of aligned reads (in million pairs) in human-related class for Alignment-based and xenome methods

Next, we compared and looked into details of the difference between those two methods. We can tell from figure 3.7 that Alignment method assigns significant more reads (~10%) to “Both” class compared to xenome tool (~0.6%).

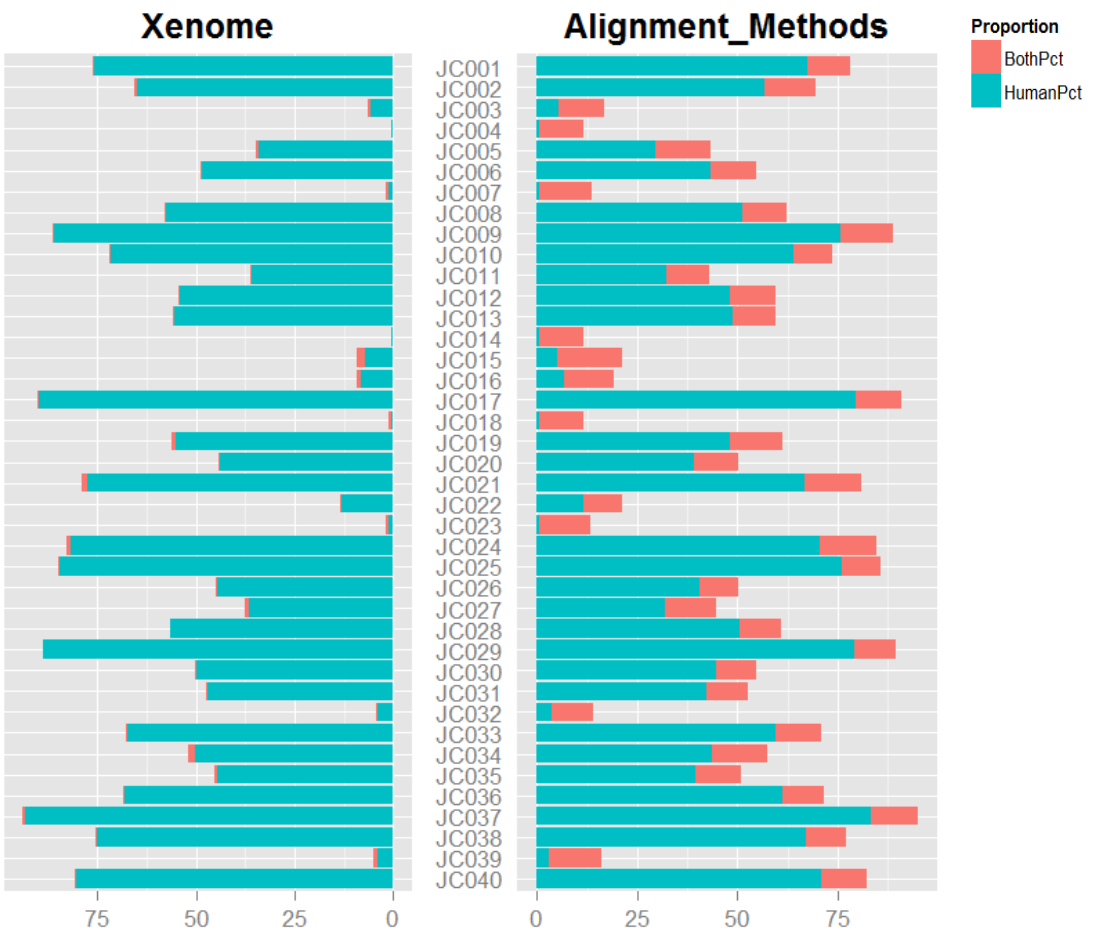


Figure 3.7. Read distribution of two methods. The x-axis is algined the human-related reads (human only in cyan and both in red) percentage by two methods.

In a typical sample processed by alignment-based method, there are 7,000,000–9,000,000 reads that align to both the mouse and human genomes which represent around 10% of the total aligned reads. Of these reads, approximately 20% are filtered out because they do not map to regions of the genome that are within known gene boundaries. After filtering, there are 5,000,000 – 7,000,000 reads per sample that map within known gene boundaries in the mouse and human genomes. As one would expect with human xenograft samples, the greatest number of multi-mapping reads map to a known gene in the human genome. This suggests that the reads mapped to non-transcribed regions of the mouse genome that share homology with genes in the human

genome. We are aware that more read alignment does not necessary mean more accurate classification. Therefore, we then align again only the “both” class read to the human reference and compared with the “Complete List of Human and Mouse Homologs (downloaded from http://www.informatics.jax.org/homology.shtml) in Mouse Genome Database (MGD) of the Jackson Laboratory. For genes with an expression FPKM (Fragments Per Kilo base of transcript per Million mapped reads) more than 1, 83.67 % of them were found in the human mouse homology list and for genes of FPKM value more than 0.1, 77.5% of genes were found in the homology list. This result implies the majority of reads in “both” class belongs to the homology genes between human and mouse. It is concordant with the fact that protein-coding regions of the mouse and human genomes are 95 percent identical due to their requirement for similar evolutionarily conserved functions.

Finally, reads in “both” group were split fairly and randomly between the mouse and human genes based on their ratio of exclusively aligned reads. Housekeeping gene expression (Eisenberg and Levanon 2013) was used to assess the accuracy of read classification. Housekeeping genes are involved and required for the maintenance of basic cellular function and thus are expected to preserve constant expression level in all conditions and human cell types. Based on large amount of public available dataset, human housekeeping genes are defined using the following criteria: (1).They have to express in all tissues. (2). Low variance across all tissue types, specifically, they have to show a log2 (FPKM) standard deviation less than 1. (3). No exceptional expression in any single tissues (Eisenberg and Levanon 2013). Those criteria resulted in a list of 3808 human housekeeping genes and of which 11 genes are selected as highly uniform and strongly expressed genes (Appendix C).

All 40 PDX samples were processed under different procedures (including K-mer xenome method, Alignment-based method but keep only human specific reads, Alignment-based method but keep all reads in “both” category, Alignment-based method and assign “both” class reads based on one ratio for all genes and Alignment-based methods but with our proposed strategy to assign reads randomly based on the human mouse ratio for each gene). Cufflinks (Trapnell, Williams et al. 2010) was used to quantify gene expression in the unit of PFKM. Coefficient of variation (CV), a dimensionless number shows the extent of variability in relation to the mean, was used to measure the degree of variation. A low CV is suggestive of a constant gene expression across all samples. Together with the definition of human housekeeping genes, we would view samples with low CV indicates more accurate and precise classification of reads.

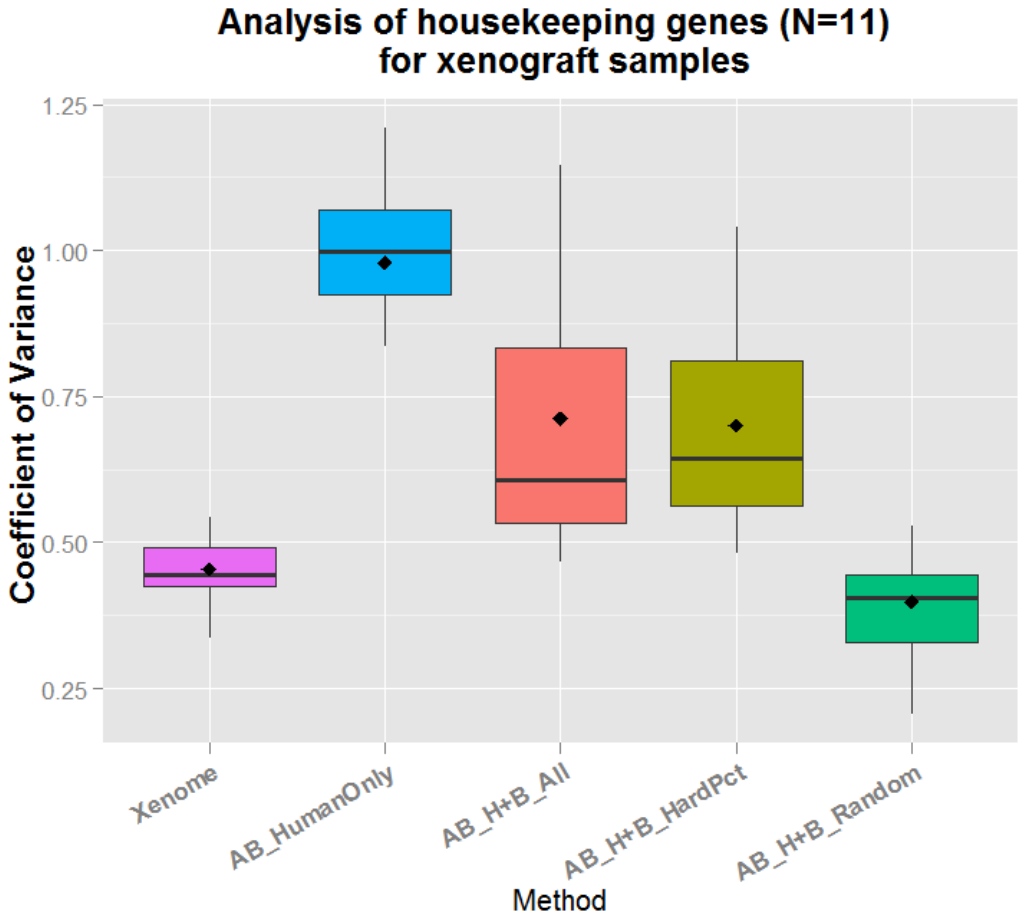


Figure 3.8. Coefficient of variation analysis of expression of 11 highly uniform and strongly expressed genes in xenograft models using various approaches.

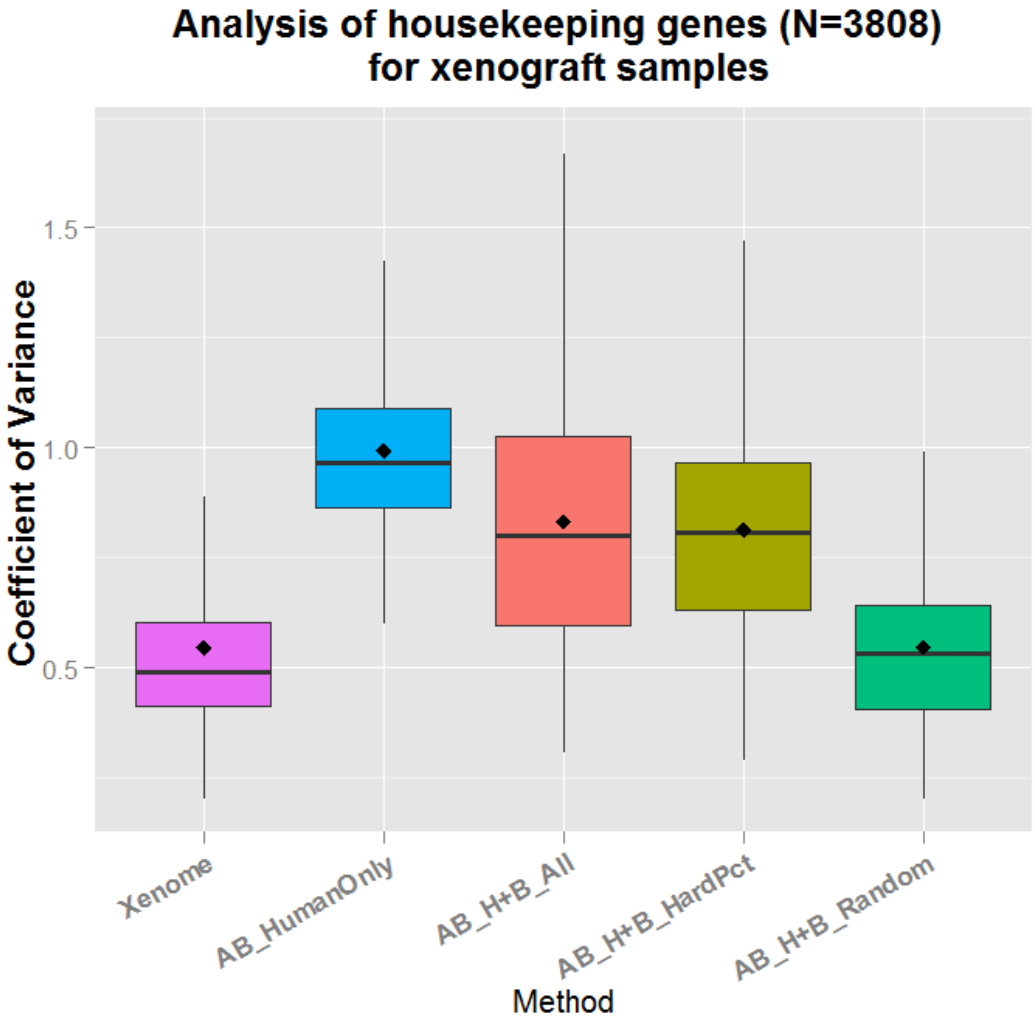


Figure 3.9. Coefficient of variation analysis of expression of 3808 housekeeping genes in xenograft models using various approaches.

We can see from figure 3.8 and 3.9 that both the highly uniform short list and full list show the similar trend and our alignment-based method with proper handling with "Both" category reads displays lowest CV. Therefore, our approach achieved better accuracy with more reads aligned.

3.3.2. Gene fusions

We also explored the calling of gene fusion for both methods. To avoid internal bias, we chose an totally independent and widely accepted gene fusion detector, Tophat-fusion (Kim and Salzberg 2011) instead of STAR's fusion to call potential gene fusions. Output of gene fusions and their related reads from tophat-fusion were used as "standard" and evaluate the capability of keeping fusion related reads after classification using xenome and alignment-based methods.

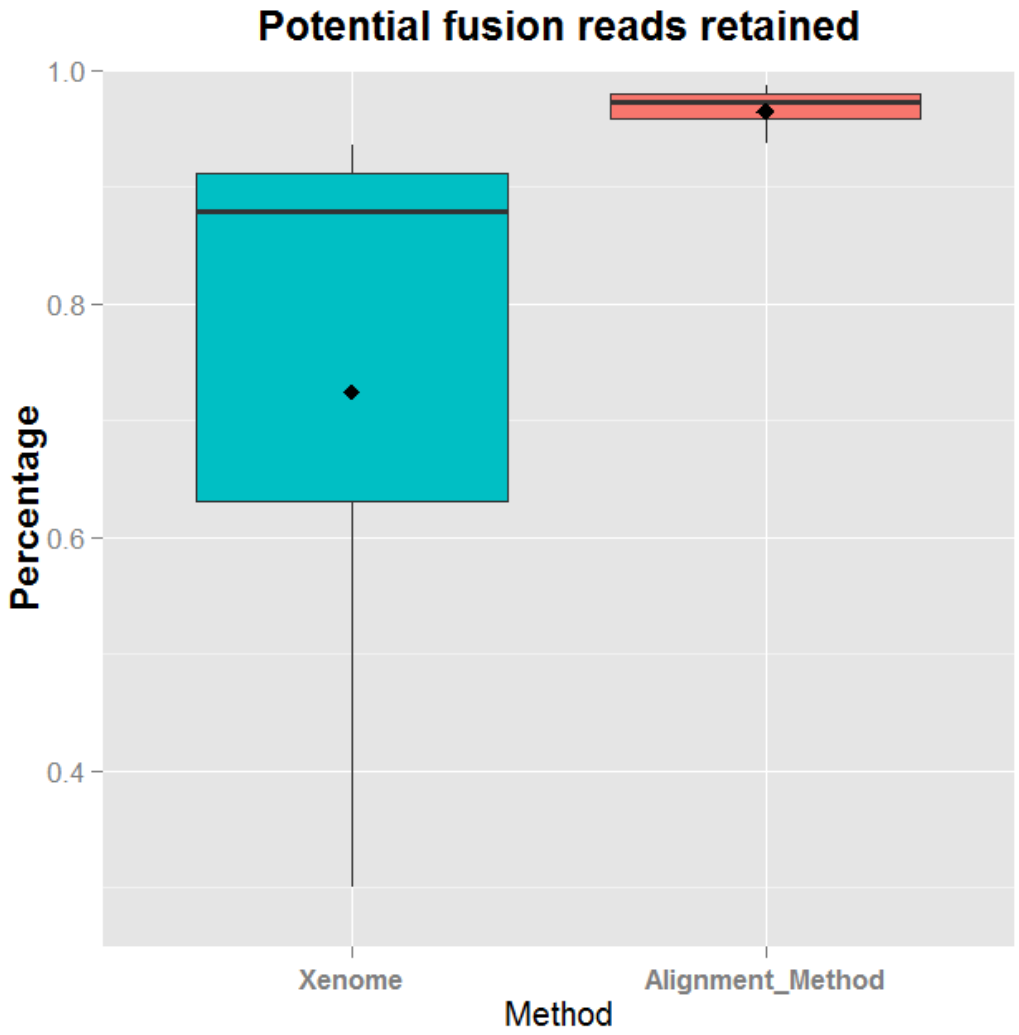


Figure 3.10. Percentage of potential gene fusion products supporting reads that were retained by two methods.

As shown in figure 3.10, our strategy was able to retain constantly on average 96.4% fusion reads while the performance of xenome varies largely across samples and could only keep on average 72.4% related reads. More importantly, the stringent characteristics of xenome would have much lower chance to preserve reads that spans cross the break point of two fused genes. Unfortunately, those critical spanning reads happen to be the key evidence of gene fusion calling.

3.3.3. Double-edged sword

Sometimes the stringent characteristic of xenome k-mer methods will become double-edged sword and show its advantage side. Usually we applied matched tumor-normal whole exome DNA sequencing to detect somatic single nucleotide variant (SNV), but sometimes RNA-seq is also used for variant calling to validate our findings in exome or to see what fraction of variants is expressed during the biosynthesis of mRNA. To explore this, we utilized Varscan 2 (Koboldt, Larson et al. 2013) with identical parameters to call variants from BAM files generated by xenome and our approach respectively. Figure 3.11 displays the number of variants for both approaches. We can see that alignment-based method generally called significant more variants when compared to k-mer xenome tool. The only exception is the JC37 sample, the number variants called by xenome is more than that by our method.

This result is expected to some extent due to the stringency of xenome during the classification step. However, xenome also tends to miss a lot of variants that are filtered before variant calling.

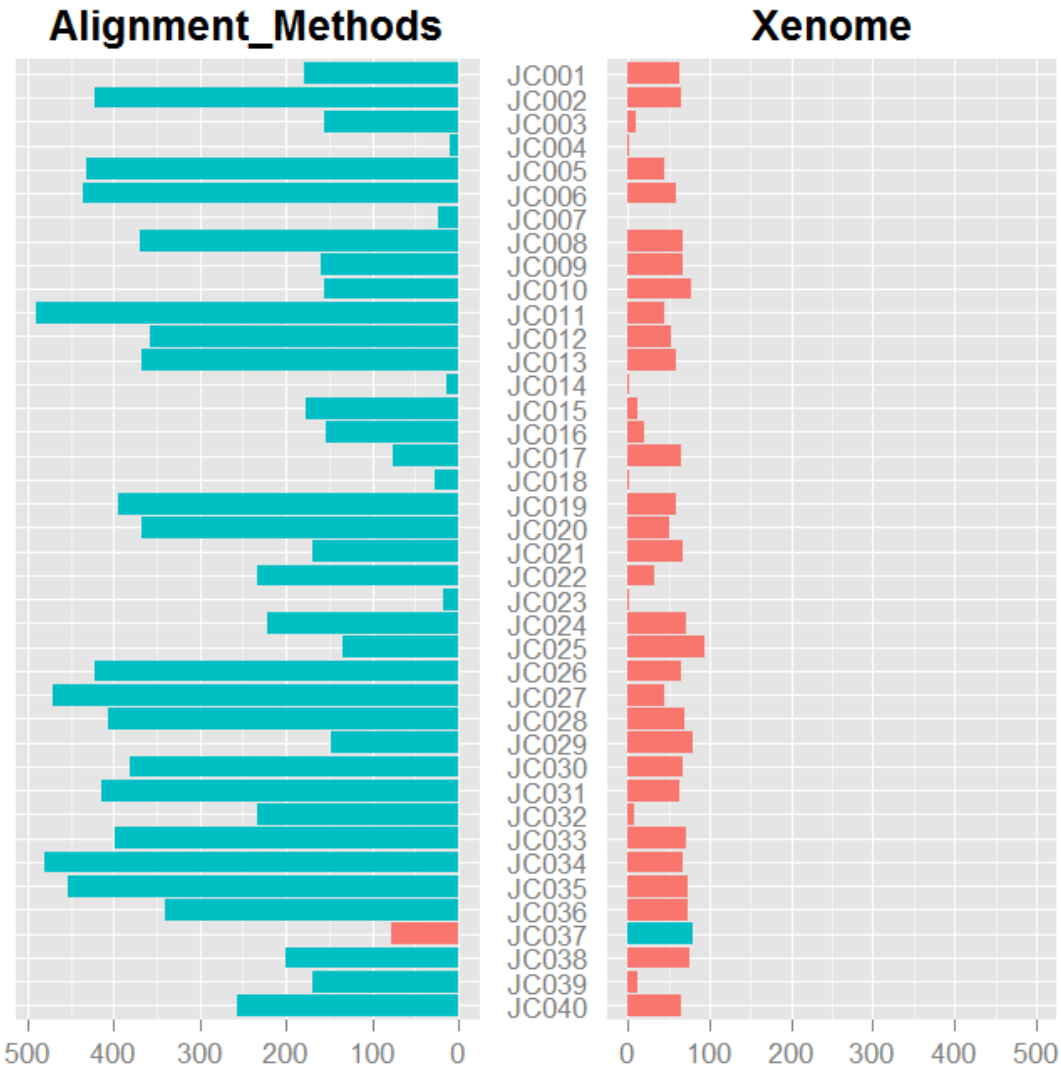


Figure 3.11. Number of Variants called by VARSCAN with same parameter, the x axis is the unit of thousand (k). Cyan color indicates more variants called in the current method and cyan shows vice versa.

We then manually checked for random variant output for the purpose of validating the calls. As shown in figure 3.12, the snapshot of IGV in PTEN region reveals an example of discordance of two methods in regards to variant calling. Both methods produce the small insertion call but only alignment method called another potential SNV. We further checked the UCSC genome browser for sequence in this region for multiple species (e.g. Rhesus, mouse, dog, zebrafish). In this specific case, the mutation “T” happens to be the

sequence difference between mouse and other species including human. Therefore, we probably shall ignore the variant at this position and view it as false positive calling. Nevertheless, we also found cases where the mutation is correctly called but filtered out by xenome at a lower occurrence rate.

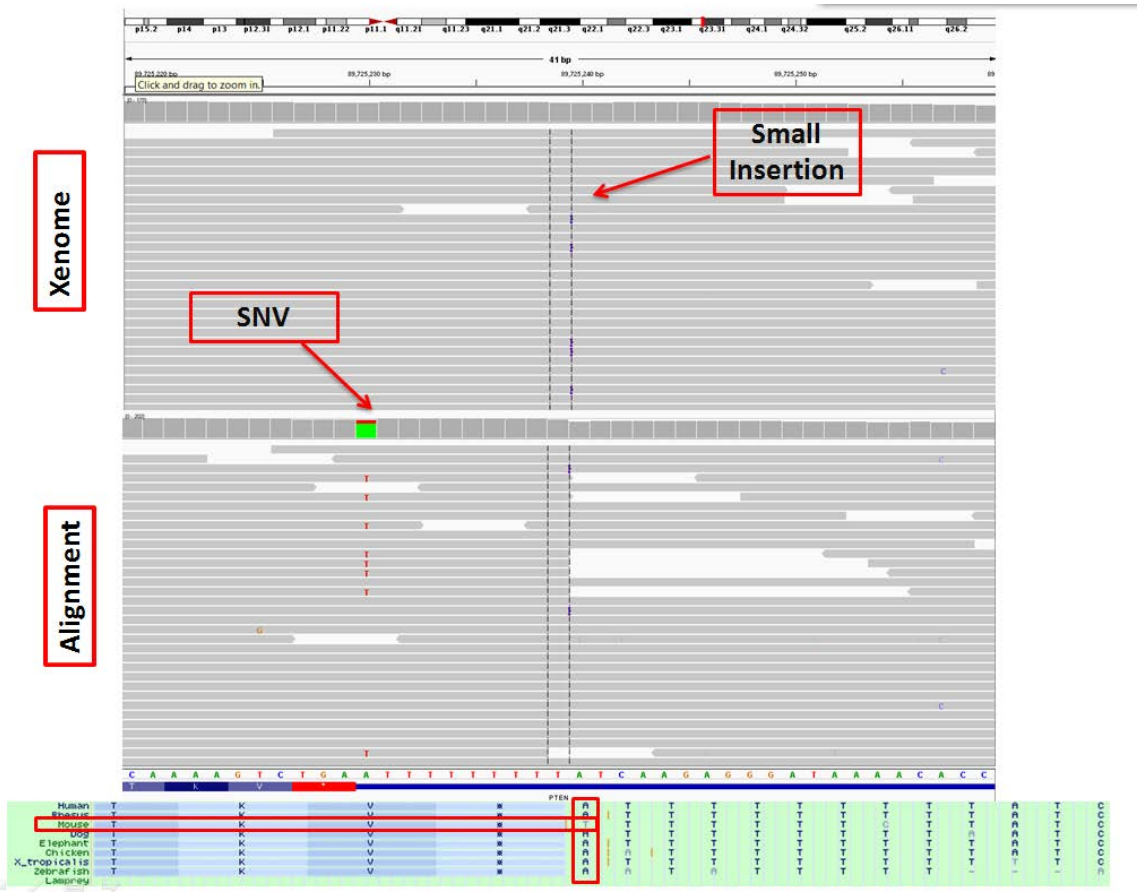


Figure 3.12. Snapshot of IGV visualization of a SNV and a small insertion in PTEN gene. Multiple reference genomes are listed in the bottom.

The alignment-based approach is implemented in UNIX shell, Perl (version 5.10.1) and portable batch system (PBS) and is supported to run on UNIX/Linux based platforms. The algorithm, UNIX shell scripts and Perl code are available for all non-commercial users via https://github.com/spengInformatics/xenograft.

3.4. Conclusion and Discussion

Using an in-silico approach relied on developed and widely used aligners, such as STAR, Tophat or Bowtie, we have shown that xenograft sequencing reads can be accurately and losslessly classified and the data was internally consistent when compared with current K-mer method. In spite of being a more liberal method than xenome (higher number of reads fall into the “both” class), our approach is more robust and could achieve higher accuracy in regards to gene expression quantification. Moreover, our strategy retains the capability to discover one of the key genomic events - gene fusions which can only be efficiently detected in large scale using RNA -seq.

The limitations of our alignment-based approach include: 1. Since our method is more lenient in regards to classification, more false positive calls are expected and should be eliminated in further validation. 2. The performance of our approach might be associated with different aligners or distinct tool versions. 3. More sophisticated algorithm is needed for better differentiation of machine error and true variants between the human and mouse.

In addition, customizable and flexible stringency could be applied in our methodology by tuning the parameter for mapping tools. Customized modifications in specific parameters including number of mismatches allowed and minimum mapping quality score enable user to have elastic control of the stringency and therefore provide robustness of our approach. Furthermore, K-mer method can be viewed as a “pre-processing” of the data but our alignment-based classification process and filters the sequencing reads along with the aligning step which avoids extra computing time and storage resources. Rather than being a disadvantage, our approach will align mouse related reads into mouse genome at the same time as byproduct. Intriguingly, this would provide us with informative knowledge about genomic features in stromal cells. Although the alignment–based approach is designed for RNA-seq read classification for

xenograft model, it is extensible for other NSG techniques for DNA samples including Exome-capture target sequencing and low-coverage whole genome sequencing. For example, we can change the aligner to DNA mapping tools such as BWA or Bowtie (Li and Durbin 2010) and go through similar workflow. The script for filtering and extracting reads could be directly applied in the case DNA sequencing as long as the sequencer uses the similar type of synthesis chemistry. Additionally, we can foresee the advantage of alignment-based method in calling chromosomal translocations from DNA sequencing compared to xenome approach for similar aforementioned reasons.

Again, in regards to the variant calling using RNA-seq, we would argue that users shall choose the strategy based on priority of their performance characteristics. For example, CLIA certified labs would like to initiate a clinical trial and want to validate the genotype before further testing. Instead of putting extensive amount of money and resources into the validation, they probably would prefer xenome processed files for lower chance of false positive calls. Generally speaking, we recommend calling variants based on xenome processed reads due to its relatively low false positive rate. In this case, we could avoid being overwhelmed by tons of data and draw meaningful biological conclusions.

Our strategy is not limited to vertebrate model animals. Previous findings already show that the viruses play an important role in human cancer and disease and some of them could even integrate into human genome. Infection of hepatitis B virus (HBV) or hepatitis C virus (HCV) greatly promotes the chance of getting hepatocellular carcinoma and cervical carcinoma is almost exclusively caused by human papillomaviruses (HPV). Authors in one study observed recurrent fusion events of human papillomavirus insertions in RAD51B and ERBB2 genes (Tang, Alaei-Mahabadi et al. 2013). In light of this, even with “pure” human samples, we would apply our approaches to classify sequencing reads into human and non-human ones and then align non-human reads to

currently known virus genomes. Based on the analysis of virus patterns, we would be able to explore many biological events including co-adaptation between virus and host mRNA expression and detection of viral integration such as host–virus fusions. At the very least, revealing the type of virus in the human sample is the low-hanging fruit by above analysis but would provide us great benefit and help for early detection of virus-associated cancer.

CHAPTER 4

INTEGRATION FRAMEWORK FOR GENOMIC CHARACTERIZATION OF SURVIVAL OUTLIERS IN GLIOBLASTOMA MULTIFORME

4.1. Introduction

4.1.1. Glioblastoma multiforme is an extremely malignant form of brain cancer

Glioblastoma multiforme (GBM) is a common and extremely malignant form of brain cancer. The disease most commonly affects adults in their sixth decade of life. Of the 12,000 or so patients diagnosed with GBM each year, about half die within the first year of diagnosis, with most of the rest succumbing to their disease within five years. The mechanisms driving the development and recurrence of GBM are still unknown. This fact greatly limits the successful treatment of this disease.

Unfortunately, sometimes GBM exhibits a high resistance to these standard therapies and recurrence is nearly assured. However, there is no established second-line regimen. In order to address the dismal prognosis & management of patients with GBM, it is essential to transform traditional clinical trial paradigms to allow for rapid and efficient therapeutic development. Thus, the development of new combinational therapies, together with an increase in the selectivity of the treatments based on a detailed molecular characterization of these tumors has significant potential to enhance the survival of patients suffering from GBM (Verhaak, Hoadley et al. 2010).

4.1.2. Outlier survivors

Standard treatment includes radiation and chemotherapy with the DNA alkylating agent temozolomide (TMZ), which only extends the median survival to approximately 15 months. Unfortunately, GB exhibits a high resistance to the standard therapy regimen

and recurrence is virtually assured due largely to highly invasive cells that aggressively disperse into the surrounding normal brain.

However, a small percentage of GB patients respond to standard treatment and benefited with an average survival time greater than two years, with some patients living longer than ten years. In this case, “GBM outliers” can be identified: patients who responded (long-term survivor, LTS, average OS is 30 months) versus those who failed rapidly (short-term survivor, STS, average OS is 7 months). To date, it is unclear why some of these some of individuals with the same diagnosis of GB die quickly, while others have extended survival. Thus, studying the genomics of these “outlier” GB patients could inform ways to better treat GB patients.

Several factors besides tumor size and location determine a patient’s survival changes. These include age at diagnosis (where younger patients often receive more aggressive treatment that is multimodal), functional status (which has a significant negative correlation with age), and histologic and genetic markers. Among them, genetic markers could effectively provide prognostic prediction of survival (Mutation and Pathway Analysis working group of the International Cancer Genome 2015) and thus are essential for transforming traditional clinical trial paradigms to allow for rapid and efficient therapeutic development. Since all patients with primary GB tumor received the same standard of care, what genomic features lead to such different outcome? The rationale is that genome-based analysis of the primary patient tumor can identify genomic alterations unique to each patient’s tumor that are candidate therapeutic targets to decrease therapy resistance and improve clinical outcome (Agarwal, Sane et al. 2011, Weiss, Liang et al. 2013). Moreover, by integrating and comparing data of patient groups (STS vs. LTS), we may further identify frequently altered events that distinguish patients with different survival outcome. We believe that targeting these genetic alterations

within the patient tumor in combination with standard of care therapy directly addresses the unique features of each tumor and will improve survival relative to standard of care alone (Stupp, Hegi et al. 2009, Barajas, Hodgson et al. 2010, Nowsheen, Whitley et al. 2012).

4.1.3. Integrative analysis for comprehensive understanding of the disease

Genomic characterization will constitute an ever-increasing fundamental role in the delivery of individualized care for oncologic patients. Emerging genomics technology and bioinformatics are now resulting in the molecular sub-classification of cancers with applications for more accurate characterization of disease, prognosis, and therapeutic selection (Hanahan and Weinberg 2011, Brennan, Verhaak et al. 2013). Under this paradigm, therapy selection is guided by the molecular profile of targetable mutations and gene pathways that vary among patients. This phenomenon is well represented in GBM, which is among the most genetically heterogeneous and lethal of all human cancers (Marusyk, Almendro et al. 2012).

Based on previous studies, germline and somatic genomic alterations have been directly linked to tumorigenesis, malignant progression, and drug resistance. Specific genomic alterations have utility to inform targeted therapeutic approaches. For example, patients with activated EGFR mutation positive non-small cell lung cancer have improved progression-free survival and overall survival when treated with gefitinib (Douillard, Shepherd et al. 2010). Similarly, a molecular profiling approach to select treatments resulted in longer progression-free survival in a significant percentage of patients with refractory cancers (Von Hoff, Stephenson et al. 2010).

Recent technological advances in Next Generation Sequencing (NGS) platforms enable the determination of entire human genomes rapidly and at reasonable cost. There is also

an increasing trend toward acquiring a number of types of data from the same patient in both clinical and research field. Moreover, publicly accessible collections of data for various cancer types (such as The Cancer Genome Atlas (TCGA)) are cataloged and stored. However, our cohort is unique and superior in interrogating aforementioned questions compared to those data repositories. For instance, in TCGA, most studies were carried out independently and thus most patients only have one or two array/sequencing data available and very few patients have complete genomic profiling. Only 6 patients are available in the TCGA glioblastoma dataset who have comprehensive and complete "-omics" data including copy number profiles, mRNA expression profiles, mutation sequencing data, methylation (HM450) data and clinical data (treatment naïve patients with standard therapy and survival days available). Unfortunately, all of those 6 patients are short-term survivors with average survival of 6 months (range 3.3 to 9.3 months), thus making the comparison of LTS and STS impossible.

To Identify and characterize the genomic signatures indicative of tumor vulnerability, we performed comprehensive genomic analysis of various "-omics" data from GBM outliers. Genomics, transcriptomics and epigenomics data each of course enables us to get a specific and insightful view of genome functions, but those views are often limited to one-dimension. Just like complex biological processes, data describing those processes are usually complementary and shall not be treated totally independently. To maximize the utilization of all available information, we considered each assay as part of "big picture" with unified, global view. Integration of Exome sequencing data, RNA-Seq, whole genome and epi-genetics data in a coherent and systematic fashion is critical to comprehensive understanding of molecular interactions in complex genetic diseases. For example, integrating highly informative yet individual datasets offer the potential to answer many long-standing research questions: what impact does variants in genetic

code have on the gene expression variation? To what extent the methylation and other regulatory elements contribute the disease phenotypes and gene expression? Is there always a corresponding structural rearrangement at DNA level for each gene fusion event at RNA level?

In a word, even after comprehensive enumeration of these genetic alterations becomes routine, understanding how they work together to cause the cancer phenotype is still a daunting challenge. As described above, primary analyses can offer some insight, especially when applied to large cohorts of samples. But our integrative methods promise to reveal a much more accurate view of what is going on in the cell by combining two or more disparate sources of information. Therefore, the development of new combinational therapies, together with an increase in the selectivity of the treatments based on a detailed molecular characterization of these tumors is likely to provide unique insights into specific molecular mechanisms of survival and has significant potential to enhance the survival of patients suffering from GBM (Verhaak, Hoadley et al. 2010, Brennan, Verhaak et al. 2013).

4.2. Materials and Methods

4.2.1. Ethics Statement and Sample Collection

Clinical information was assimilated from patient records (figure 4.1) from the Case Western Reserve University. Informed consent was obtained for each patient on protocols approved by Case Western Reserve University Institutional Review Board. Clinical data elements collected comprise of gender, age at diagnosis/surgery, pathology (i.e. pre-treatment/recurrence/secondary tumor), MGMT status, G-CIMP status, IDH1 status, therapy class, vital status, overall and progression-free survival. Tissue specimens and matched blood samples were collected fresh frozen and maintained below −80°C

until nucleic acid extraction.

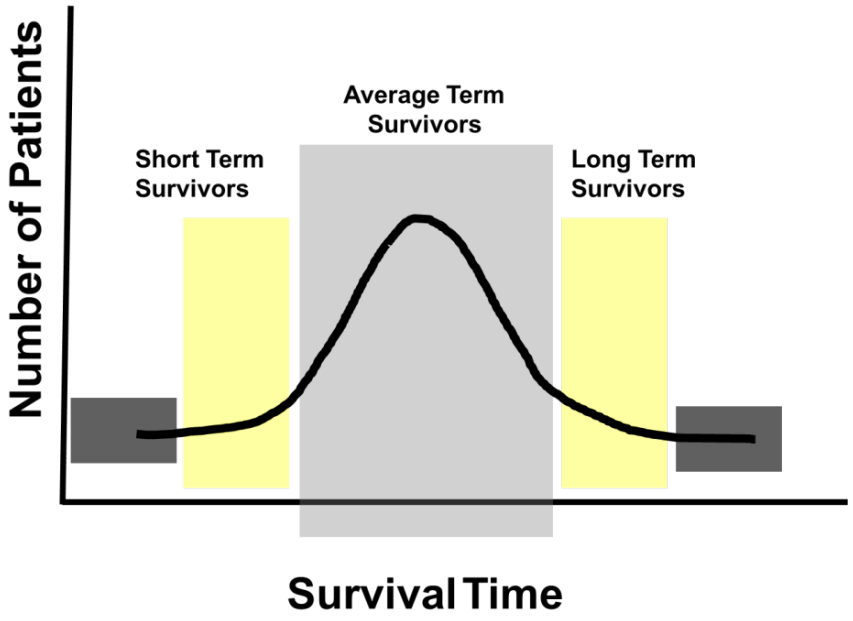


Figure 4.1. Schema of patient selection for Outlier study

4.2.2. Sample Selection and DNA/RNA Isolation

For this work, samples from 23 patients with known clinical data elements as listed above were screened and 23 samples, collected from treatment-naïve primary GBM patients who subsequently received surgery and standard of care treatment, representing two distinct survival groups were selected for the study. These included all long-term survivors who had an overall survival (OS) of > 18 months but < 60 months (LTS) and all short-term survivors who had an OS of > 3 months but < 12 months (STS).

Samples from TCGA dataset were selected, by the criteria of deceased patient who received the same standard of care treatment (surgical debulking followed by radiation and chemotherapy) and whose survival days fall within one standard deviation of mean survival days of the LTS and STS groups, and used as a validation dataset.

DNA and total RNA from fresh frozen tissue specimen were isolated using DNAeasy

Blood and Tissue Kit (Qiagen #69504) and RNAeasy Mini Kit (Qiagen # 74104). Matched normal patient DNA was extracted and purified from the blood using a QIAamp DNA Blood Midi Kit/QIAamp Mini Kit from QIAGEN as previously described (TCGA, 2008) by Case Western Reserve University and supplied to TGen.

4.2.3. Next Generation Sequencing (NGS)

All NGS data acquisition and analysis was carried out using previously described methods (Borad, Champion et al. 2014). Methods for whole genome sequencing, exome sequencing, RNA sequencing and their analysis are described briefly here.

4.2.3.1. Whole genome sequencing

1.1 µg genomic tumor and normal DNA was used to generate separate long insert whole genome libraries for each sample using Illumina's (San Diego, CA) TruSeq DNA Sample Prep Kit (catalog# FC-121-2001). In summary, genomic DNAs are fragmented to a target size of 900–1000 bp on the Covaris E210. 100 ng of the sample was run on a 1% TAE gel to verify fragmentation. Samples were end repaired and purified with Ampure XP beads using a 1:1 bead volume to sample volume ratio, and ligated with indexed adapters. Samples are size selected at approximately 1000 bp by running samples on a 1.5% TAE gel and purified using Bio-Rad Freeze 'N Squeeze columns and Ampure XP beads. Size selected products are then amplified using PCR and products were cleaned using Ampure XP beads. Whole genome libraries were prepared using Illumina's TruSeq DNA Sample Prep Kit.

4.2.3.2. Exome sequencing

1.1 µg genomic tumor and normal DNA for each sample was fragmented to a target size

of 150–200 bp on the Covaris E210. 100 ng of fragmented product was run on TAE gel to verify fragmentation. The remaining 1 µg of fragmented DNA was prepared using Agilent's SureSelect^{XT} and SureSelect^{XT} Human All Exon 50 Mb kit (catalog# G7544C). Exome libraries were prepared with Agilent's (Santa Clara, CA) SureSelect^{XT} Human All Exon V4 library preparation kit (catalog# 5190-4632) and SureSelect^{XT} Human All Exon V4+UTRs (catalog# 5190-4637) following the manufacturer's protocols.

4.2.3.3. RNA sequencing

1 µg of total RNA for each sample was used to generate RNA sequencing libraries using Illumina's TruSeq RNA Sample Prep Kit V2 (catalog# RS-122-2001) following the manufacturer's protocol.

4.2.3.4. Paired end sequencing

Libraries with a 1% phiX spike-in were used to generate clusters on HiSeq Paired End v3 flowcells on the Illumina cBot using Illumina's TruSeq PE Cluster Kit v3 (catalog# PE-401-3001). Clustered flowcells were sequenced by synthesis on the Illumina HiSeq 2000 using paired-end technology and Illumina's TruSeq SBS Kit.

4.2.4. Alignment and Variant Calling

4.2.4.1. Whole genome and whole exome

For whole genome and exome sequencing, we aligned FASTQ files with BWA 0.6.2 to GRCh37.62 and the SAM output were converted to a sorted BAM file using SAMtools 0.1.18. We then processed BAM files through INDEL realignment, mark duplicates, and recalibration steps in this order with GATK 1.5 where dpsnp135 was used for known SNPs and 1000 Genomes' ALL.wgs.low_coverage_vqsr.20101123 was used for known

INDELs. Lane level sample BAMs were then merged with Picard 1.65 if they were sequenced across multiple lanes. Comparative variant calling for exome data was conducted with Seurat.

We applied previously described copy number and translocation detection to the whole genome long insert sequencing data (Craig, O'Shaughnessy et al. 2013) (these are available through https://github.com/davcraig75/tgen_somaticSV). Copy number detection was based on a log2 comparison of normalized physical coverage (or clonal coverage) across tumor and normal whole genome long-insert sequencing data, where physical coverage was calculated by considering the entire region a paired-end fragment spans on the genome, then the coverage at 100 bp intervals was kept. We normalized normal and tumor physical coverage, smoothed and filtered for highly repetitive regions prior to calculating the log2 comparison. We used Genomic Identification of Significant Targets in Cancer (GISTIC) to identify regions of the genome that were significantly amplified or deleted across the LTS and STS groups (Mermel, Schumacher et al. 2011). GISTIC calculated a statistic (G-score) for the frequency of occurrence and the amplitude of the aberration. We computed the statistical significance of each aberration by comparing the observed G-score to the results expected by chance. Regions with false-discovery rate (FDR) q-values less than 0.25 were considered statistically significant.

Translocation detection was based on discordant read evidence in the tumor whole genome sequencing data compared to its corresponding normal data. In order for the structural variant to be called, their needs to be greater than 7 read pairs mapping to both sides of the breakpoint. The unique feature of the long-insert whole-genome sequencing was the long overall fragment size (~1 kb), where by two 100 bp reads flank a region of ~800 bp. The separation of forward and reverse reads increases the overall probability that the read pairs do not cross the breakpoint and confound mapping.

4.2.4.2. RNA

For RNA sequencing, lane level FASTQ files were appended together if they were across multiple lanes. These FASTQ files were then aligned with STAR 2.3.1 and Tophat 2.0.8 to GRCh37.62 using ensembl.63.genes.gtf as GTF file. Changes in transcript expression were calculated with Cuffdiff 2.1.1 in FPKM format using upper-quartile normalization. Genes with mean FPKM less than 0.1 were filtered out and surrogate variable analysis (SVA) was applied to remove batch effect (Detecting and correcting systematic variation in large-scale RNA sequencing data). We used Student's t test to call differentially expressed genes (DEG) between LTS and STS groups using a p-value 0.05 as cutoff. We aligned novel fusion discovery reads with Tophat-Fusion 2.0.8 (Kim and Salzberg 2011). Clustering was performed using R heatmap.2 package with Euclidean Distance method and McQuitty clustering method.

We performed unsupervised hierarchical clustering using expression of genes known to be related to the genome instability and are included in the CIN 70 gene list (Carter, Eklund et al. 2006). Additionally, to identify specific molecular programming that might be driving outcome to standard of care treatment, ontology and pathway enrichment analysis was carried out using genes differentially expressed between LTS and STS groups.

4.2.5. DNA Methylation Profiling

Genomic DNA from patient samples was profiled for DNA methylation using IlluminaInfiniumHumanMethylation450 (HM450) platform, which interrogates 482,421 CpG sites as described previously (Hjelm, Salhia et al. 2013). Briefly, 1 µg genomic DNA per sample was used for the Illumina Infinium HumanMethylation450 Bead Chip, and the chip was prepared and ran according to the manufacturer's instructions (Illumina) (Bibikova, Barnes et al. 2011). Differential methylation was

defined as a site that had a beta-value difference of at least 20% (i.e. ≥0.2). Analysis of differential methylation was performed for all methylation sites, as well as those specific to CpG Islands, shore or shelves.

4.2.6. Integration framework

The genomic sequencing coverage was more than 100X for exome and 10X for whole genome for tumor and germline genomes (Supplementary Table). Somatic mutations including SNVs, INDELs, translocations, intra-chromosomal rearrangements (inversion, etc.), and copy number alterations were determined from sequencing of tumor and germline pairs. Once we get patient tumor aberrations from various types of data (E.g. Copy number, SNV), the remaining challenge is how to effectively integrate them and try to identify potential cancer causing variants ("drivers") & their corresponding drug/treatment. A bunch of previous studies have proposed methods to integrate data, however, their approaches often lack the ability to predict key "drivers" in the disease, let alone the possible individualized treatment (Beroukhim, Getz et al. 2007, Parsons, Jones et al. 2008, Verhaak, Hoadley et al. 2010, Salhia, Kiefer et al. 2014).

We applied a framework (figure 4.2) which is a direct and biologically motivated approach to analyze these data types in an integrated way. For instance, it has been documented that SNVs in certain gene promoters may strengthen or weaken the binding affinity of the transcription machinery, but it is unclear exactly which factors modulate the strength of the effect. By integrating genomic assays with expression data, it is possible to assess how a particular gene's expression is regulated by these mutations. Such framework provides a comprehensive synthesized set of guideline for the systematic data analysis and integration.

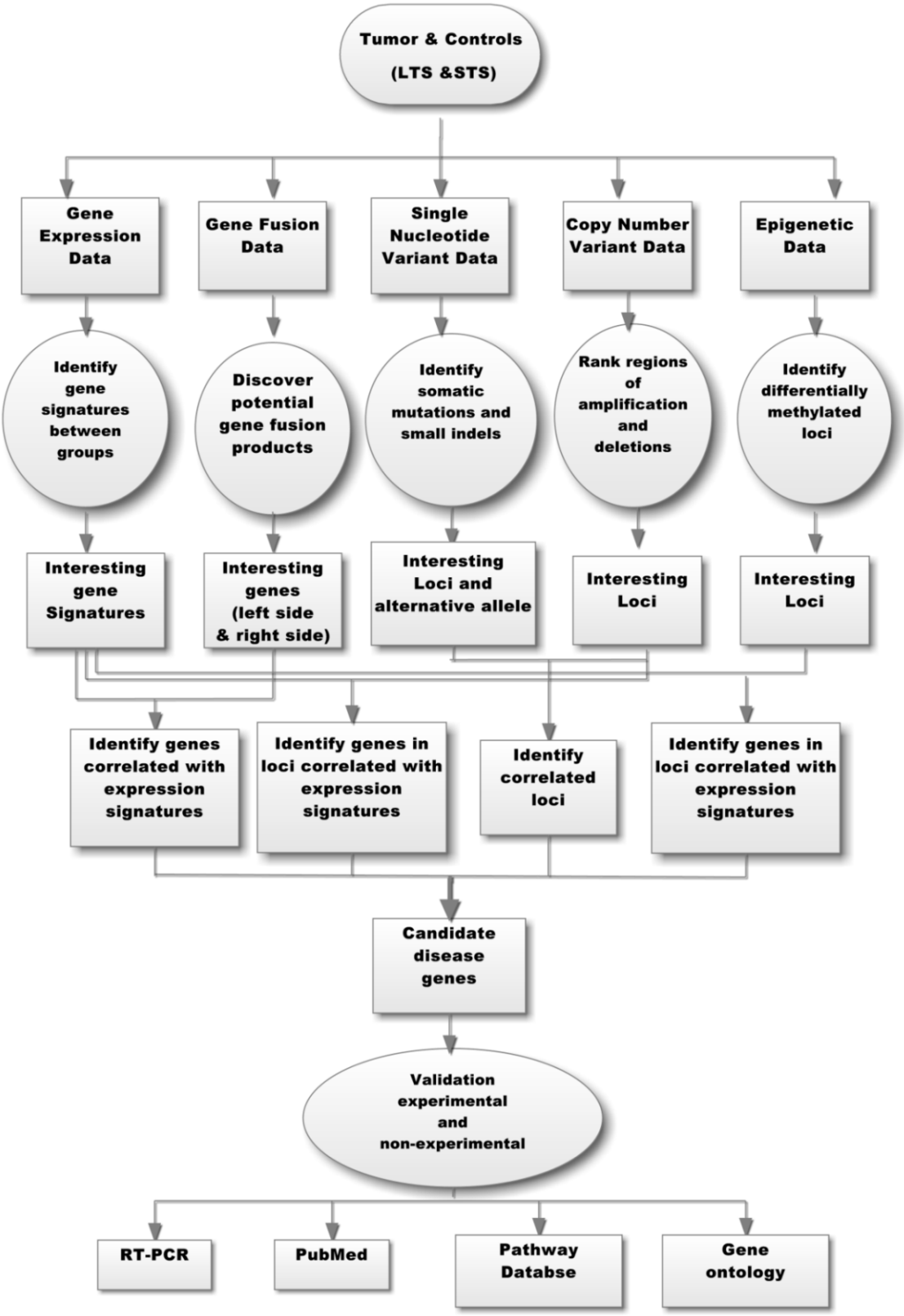


Figure 4.2. Tentative workflow of the biologically motivated framework.

4.3. Results

4.3.1. Clinical characteristics of patients

Our long-term and short-term survivor cohorts consisted of patients diagnosed with primary GBM. All tumor specimens were treatment naïve and contained an average of 75% tumor cellularity (range 50%-95%). Long-term survivors are defined as patients with GBM with an average overall survival (OS) of 33 months (range 18-57 months), and short-term survivors are patients with an average OS of 6 months (range 3-11months; Table I).

Patient #	Case Western Sample ID	Path Diagnosis	Age	Gender	Race	Estimate of Resection	Post-Resection Therapy (Rad. + TMZ)	Overall Survival (days)	Survival Cohort
1	252	GB	69	F	White	GTR	Y	102	**STS**
2	118	GB	51	M	White	GTR	Y	148	**STS**
3	274	GB	70	M	White	GTR	Y	151	**STS**
4	215	GB	56	M	White	STR	Y	157	**STS**
5	535	GB	71	M	White	GTR	Y	184	**STS**
6	795	GB	71	M	White	STR	Y	229	**STS**
7	58	GB	51	M	White	GTR	Y	282	**STS**
8	509	GB	50	M	White	STR	Y	304	**STS**
9	87	GB	70	M	White	STR	Y	307	**STS**
10	33	GB	70	M	White	GTR	Y	331	**STS**
11	45	GB	83	F	White	STR	Y	596	**LTS**
12	288	GB	55	F	Black	GTR	Y	748	**LTS**
13	422	GB	68	F	White	STR	Y	749	**LTS**
14	7	GB	63	F	White	GTR	Y	772	**LTS**
15	22	GB	71	M	Other	GTR	Y	1042	**LTS**
16	317	GB	55	F	White	GTR	Y	1208	**LTS**
17	2	GB	56	F	White	GTR	Y	1267	**LTS**
18	88	GB	61	M	White	STR	Y	1713	**LTS**

GTR = Gross Total Resection (> 95% by volume), STR = Sub Total Resection (≤ 95% by volume), Y = Yes, N = No, F = Female, M = Male, STS = Short-term survivor, and LTS = Long-term survivor

Table 4.1. Clinical characteristics of the primary GB patients in the study.

4.3.2. Genomic landscape

Compared to STS, we identified a significantly greater number of genomic alterations in LTS. Similar number of somatic coding mutations were identified in our cohorts, LTS showed a total of 425 somatic coding mutations, with an average of 53 mutations/tumor (range 34-80) whereas, STS harbor a total of 347 somatic coding mutations, with an

average of 35 mutation/Tumor (range 2-56) in STS (Figure 4.3). In addition, there is no significant difference of translocation events observed between LTS and STS where the average is 93 and 73, respectively. However, LTS cohort displayed a two-fold increased in copy number variants (CNV) loss with an average of 155 CNV loss as compared to STS with an average of 74 CNV loss; similar average CNV gains are observed in both LTS (18) and STS (12).

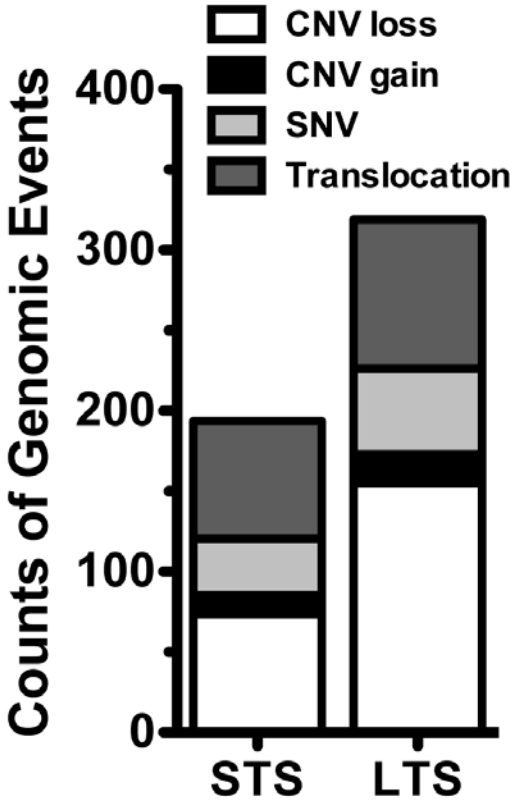


Figure 4.3. Number of genomic alteration events between STS and LTS group.

We used Snipea to detect the list of somatic single nucleotide variants (SNV) and small INDELs. Overall, we note the trend of more mutations in LTS compared to STS. If we overlap our findings with the previously identified, frequently altered genes of primary GBM, the spectrum showed similar frequency of known drivers variants in EGFR, CDKN2A and PTEN but more alterations in LTS for other genes.

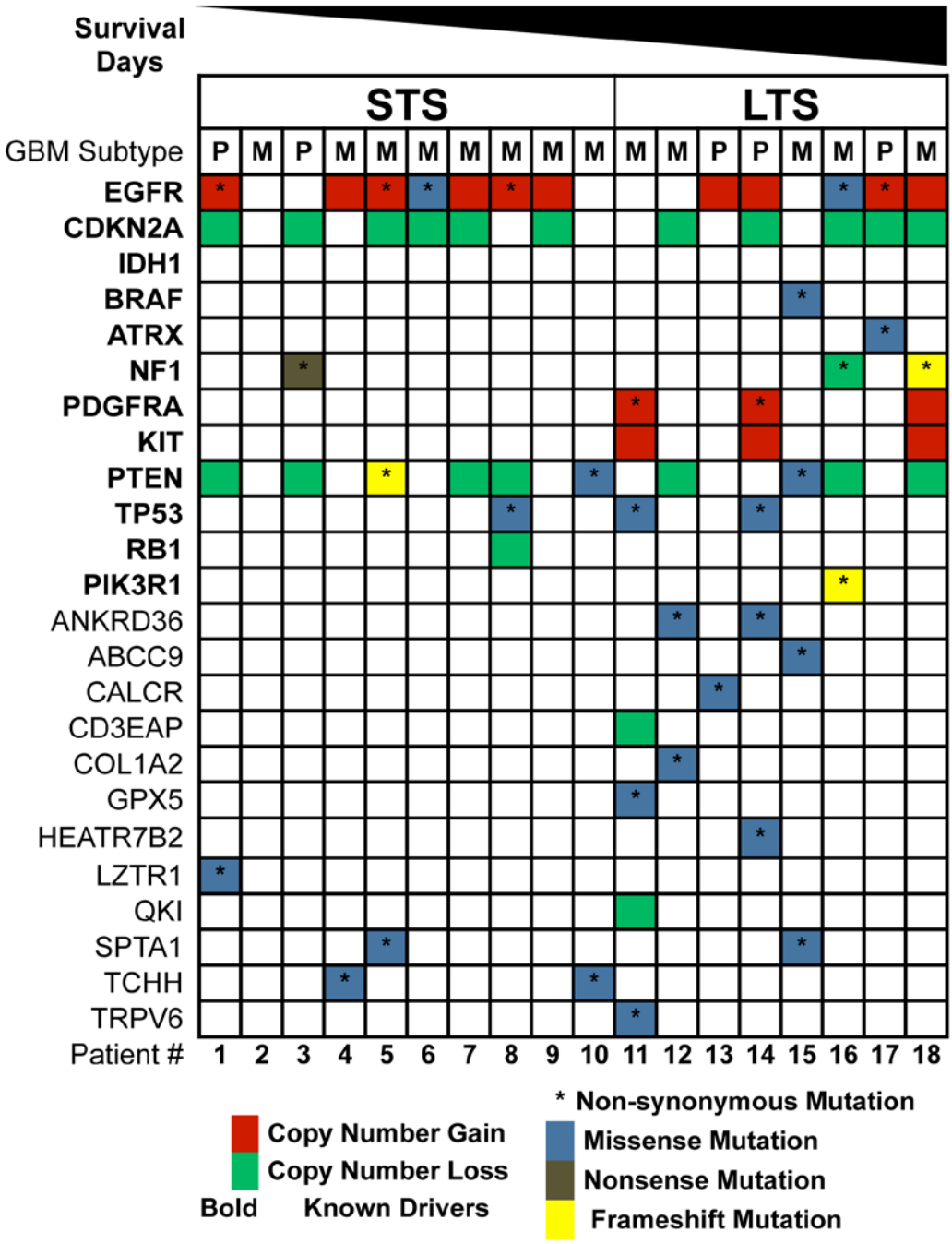


Figure 4.4. Genomic Alterations Identified in Outlier Cohort. The spectrum of alterations identified within the two patient groups was mapped against the subset of frequently altered genes previously identified in primary GBM (TCGA, Cell 2013). Red indicates copy number gain, green is copy number loss, an asterisk (*) Indicates non-synonymous mutation, where missense mutations are colored blue, nonsense mutations

colored grey, and frame-shift mutations colored yellow. The order of patients is ranked by their survival days in a descending order. We added a few known drivers in primary GB and no IDH1 mutation was found in our cohort.

This co-mutations figure provide a comprehensive analysis profile within outlier exome data and enables us to rapidly infer the relationships between co-occurring results across patients and between survival groups.

4.3.3. Copy number analysis

A schematic of the total copy number changes in LTS and STS cohorts is shown in figure 4.5. Of particular interest, among the focal amplification unique to LTS that occurred in more than 1 tumor includes PDGFR and KIT at (chromosome 4q12). In addition, focal amplification was also detected LTS cohort at 12q14.1 and 16p11.2. Moreover, the focal deletions observed in LTS as compared to STS, including 19q13.33, 17q11.2, 17q21.2 and 2p22.1. Both LTS and STS showed similar frequency of focal amplification of EGFR at 17p12.1, focal deletion of CDKN2A/B at 9p21.3, and PTEN deletion Chr. 10. Those observations were further confirmed by GISTIC analysis, which is shown in figure 4.6. The sum of CNV events is much greater in the LTS samples with the greatest difference being the high number of deletions in LTS samples. The most frequently observed CNV's were the classic GBM events such as EGFR amplification and CDKN2A deletion.

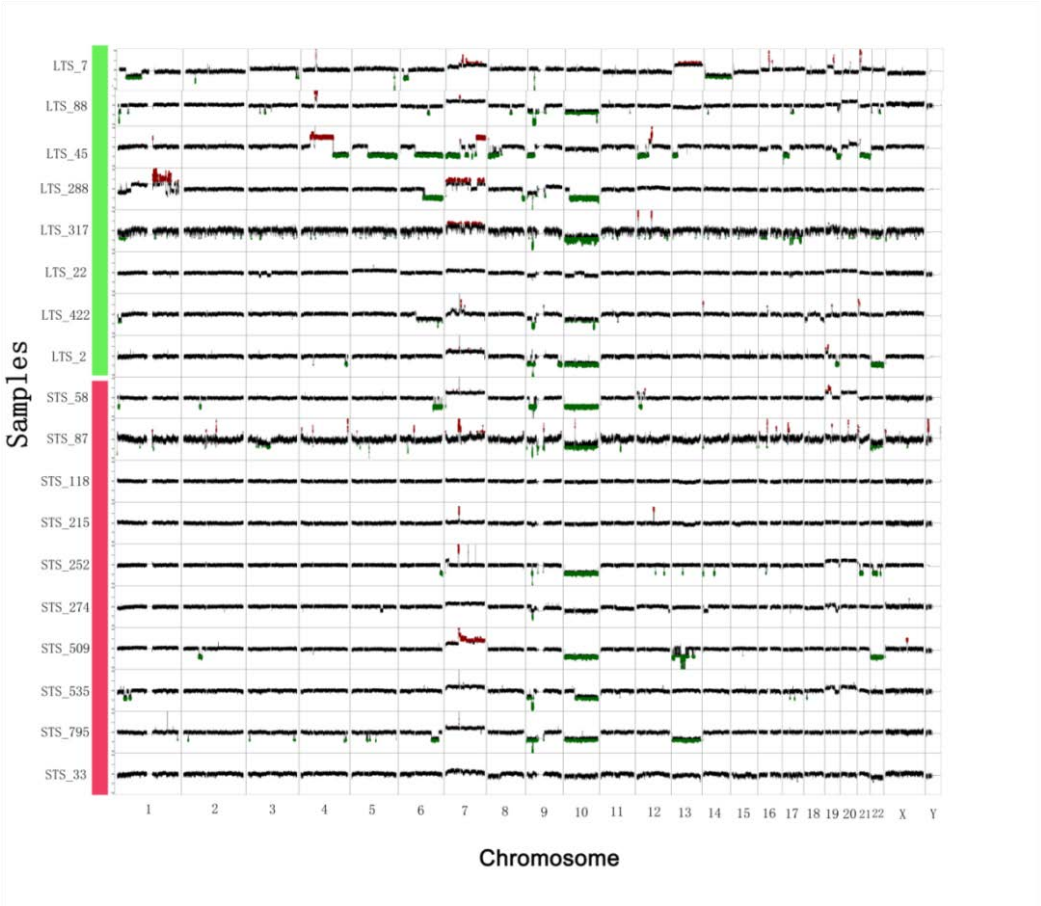


Figure 4.5. CNV compilation plot: Long Term Survivors are highlighted in Green on the y-axis. Short Term Survivors are highlighted in Red on the y-axis. X-axis defines the chromosomes. Each sample line represents DNA copy number data as the log2 ratio of fragment counts in the tumor specimen relative to the matched normal DNA. Copy Number Variations (CNV's) can be identified in spikes where the log2 ratio is greater the +1 or less than -1. Amplifications are marked in red and deletions marked in green. While both cohorts have classic GBM CNV's (Chr 7 gain, EGFR amplification, Chr 10 loss) the long term samples have many more CNV events.

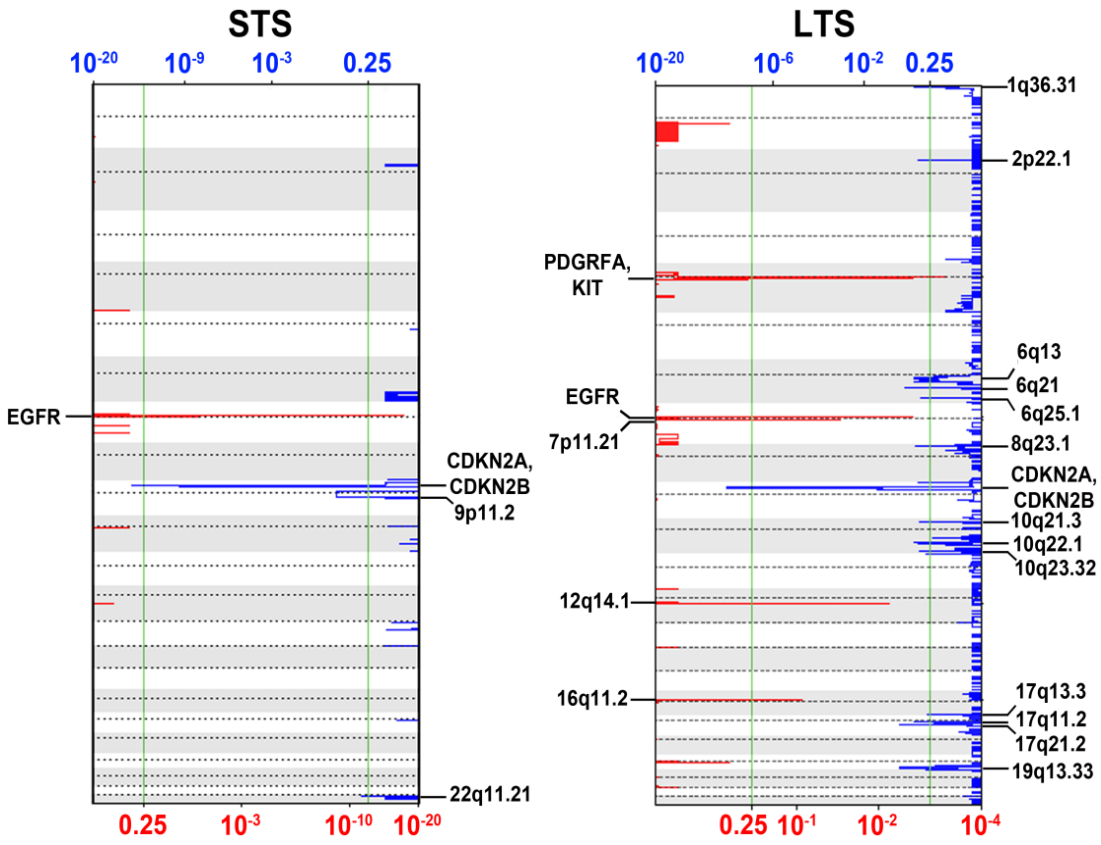


Figure 4.6. GISTIC Analysis plot: Each plot shows the frequency of CNV's for each cohort (Left: STS Right: LTS). In each plot the bottom axis defines the frequency of CNV amplifications in red. The top axis of each plot defines the frequency of CNV deletions. The x-axis of each graph represents the genome with chromosomes marked with alternating white and gray regions. The vertical green bars mark a threshold of significance. Annotations of a variety of significant regions are highlighted in boldface. Q-values for amplified (red) and deleted (blue) regions are displayed along the x-axis on bottom and top of figures, respectively.

To validate our observation of increased genomic alteration in LTS cohort, we examined the CNV changes in GBM samples in the TCGA database. To ensure the sample cohorts are similar, we select the TCGA specimens with available genomic data based on the following criteria as our outlier cohorts which includes deceased patients, patients

receiving same standard of care (surgery with TMZ and IR therapies), and patients survival days. Based on these criteria, we identified 44 LTS and 28 STS in the TCGA dataset. Examination of the CNV alteration showed that the LTS cohort displayed increased genomic alterations as compared to STS (Figure 4.7; p = 0.034), thus corroborating with our GBM outlier dataset (p = 0.016).

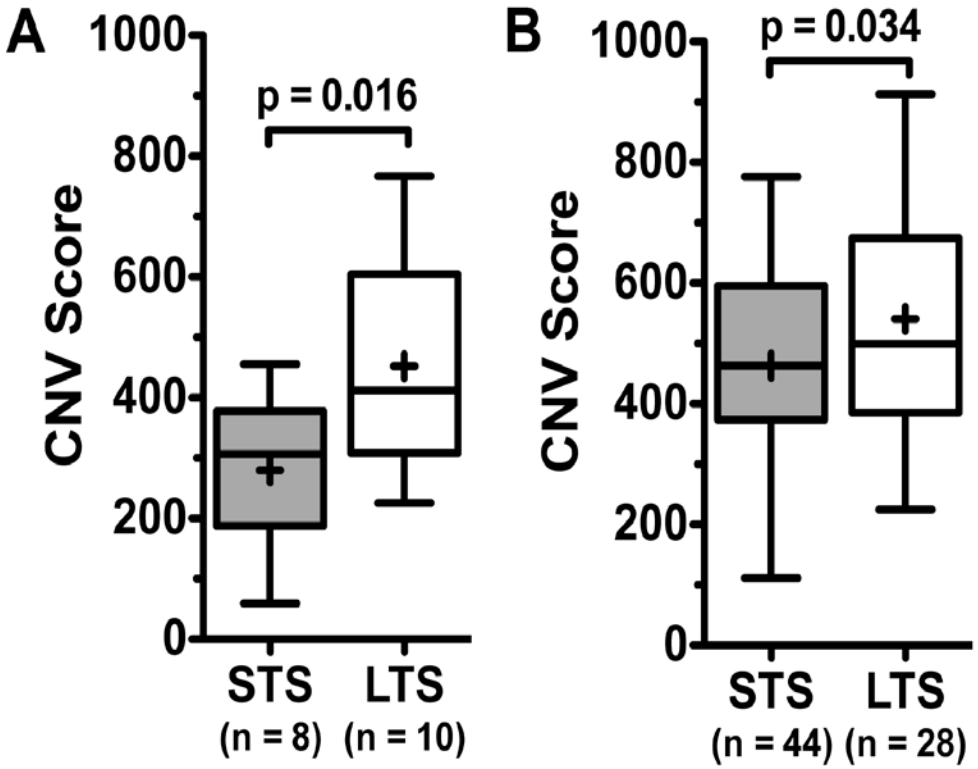


Figure 4.7. Validation of Copy number alterations in TCGA dataset. A) Box-plot showing the CNV score of STS (n=8) and LTS (n=10) in outlier dataset B) Box-plot illustrating the difference of CNV score in STS (n=44) and LTS (n=28) patients who met the same criteria. P-value was calculated using one-tail T-test.

4.3.4. Methylation analysis

We next assessed global DNA methylation patterns in the outlier GBM cohorts. We used the 450K-methylation platform to compare LTS and STS cohorts. We performed a logit

transformation on each sample, where logit transformation converts otherwise heteroscedastic beta values (bounded by 0 and 1) to M values following a Gaussian distribution. Overall, the analysis revealed 89 (263 regions) differentially methylated CpG loci (DML) encompassing 69 unique genes. We used Beta values for generating box plots to represent overall methylation levels across DML for LTS and STS. Median overall methylation was lower in STS (β =0.374) than in LTS (β =0.472) (Figure 4.8), indicating hypomethylation in STS. Examination of the overall methylation levels of the GBM Outliers in TCGA also showed an overall significant hypomethylation status in STS, corroborating with our data ($p < 0.0001$). Next we examined the distribution of DML across chromosomes, plotting the distribution of hypo- and hyper-DML after normalization to chromosome length (Figure 4.9). According to the analysis in STS cohort, chromosomes 10, 19, and 20 had the most hypomethylated loci and chromosomes 5, 8, 16, and 22 had the most hypermethylated loci, with chromosomes 10, 19, and 20 having the most overall DML.

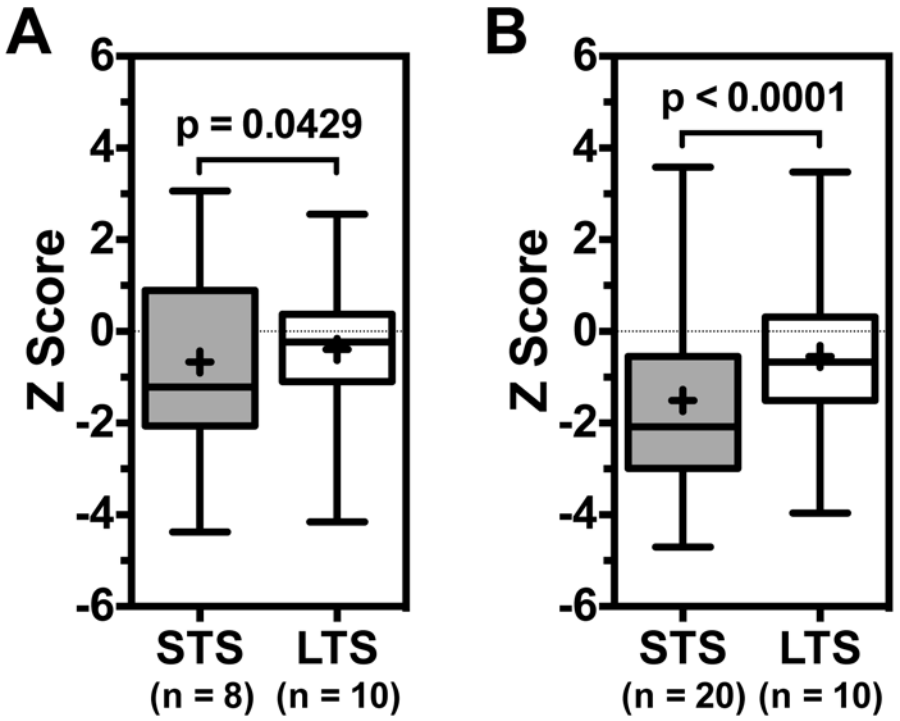


Figure 4.8. Differential methylation analysis in GBM. A) Boxplot showing the higher beta value of differentially methylated loci in LTS compared to STS in GB outlier dataset. B) Boxplot illustrating the similar observations in TCGA dataset.

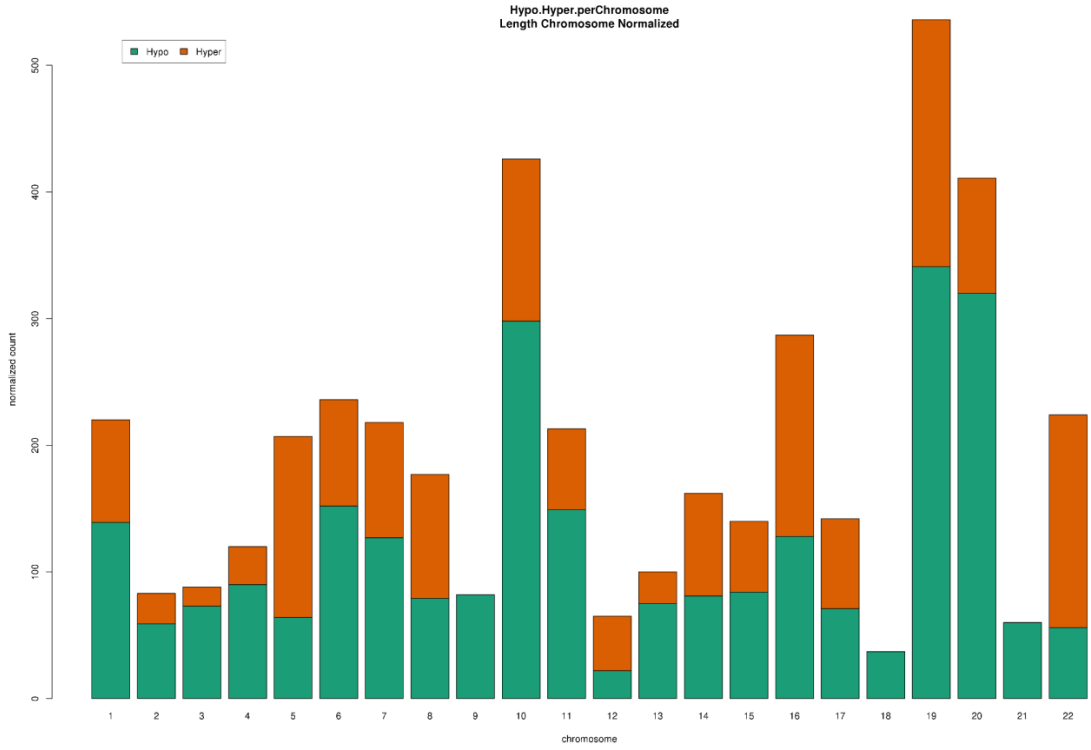


Figure 4.9. Bar plot comparing the hypomethylated and hypermethylated probes on each chromosome between STS and LTS survival cohorts. On the y-axis the number of differentially methylated probes (hypo in green, hyper in red) are shown. The x-axis defines chromosomes. This figure suggests that chromosomes 10, 19, 20 are frequently hypomethylated in STS samples relative to LTS samples. Chromosomes 10, 16, 19, and 22 are frequently hypermethylated in STS samples relative to LTS samples.

Next we wanted to examine the regional and functional CpG distribution of DML in the Outlier GBM cohorts. Functional distribution relates CpG position to transcription start sites (TSS –200 to –1500 bp, 5′ untranslated region (UTR), and exons 1 for coding genes as well as gene bodies. Overall, majority of probes (>40%) were situated in gene bodies,

followed by ~20% of probes situated -1500 bp of TSS. STS cohort appears to harbor more hypothemethylation in probes situated in -200 bp of TSS and both exon 1 and 5'UTR region (Figure 4.10A).

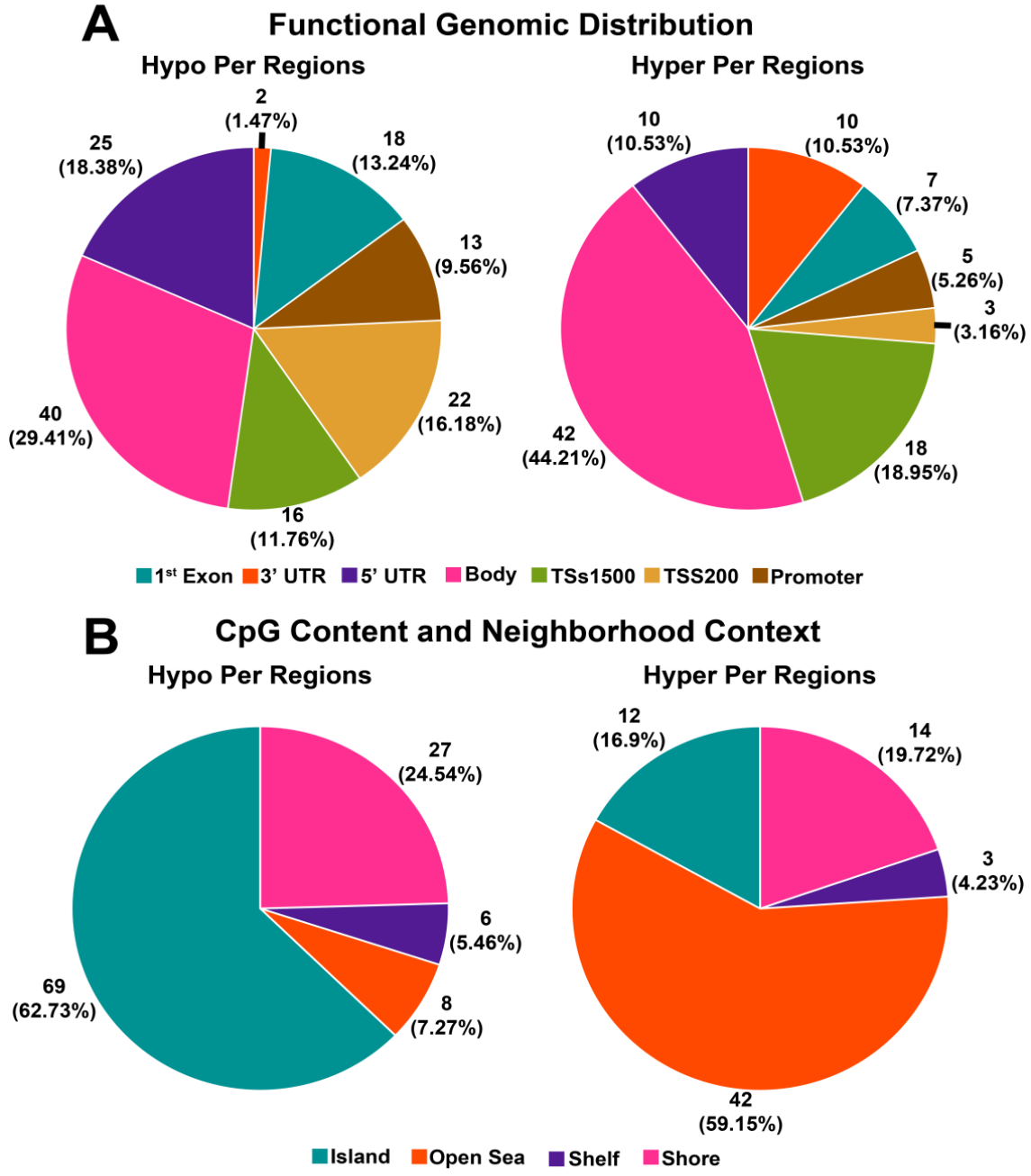


Figure 4.10. Functional distribution analysis of differentially methylated loci in GB outlier.

Regional distribution of DML was assessed based on their proximity to the closest CpGisland. In addition to island cores, shores are 0-2 kb from CpG islands, shelves are 2-4 kb away and open sea regions are isolated loci without a designation. When comparing the STS to the LTS cohorts, we identified the majority of DML in STS that were hypomethylated were in the islands (62.73%) and the shores (27%) (Figure 4.10B). In addition we note that the majority of hypermethylated loci (59.15%) were located in the open sea as compared to the majority of hypomethylated loci being situated in CpG island (62.73%) (Figure 4.10B).

Interestingly, unsupervised clustering analysis of DML demonstrated a distinct separation of LTS and STS samples, consisting of 89 probes (Figure 4.11).

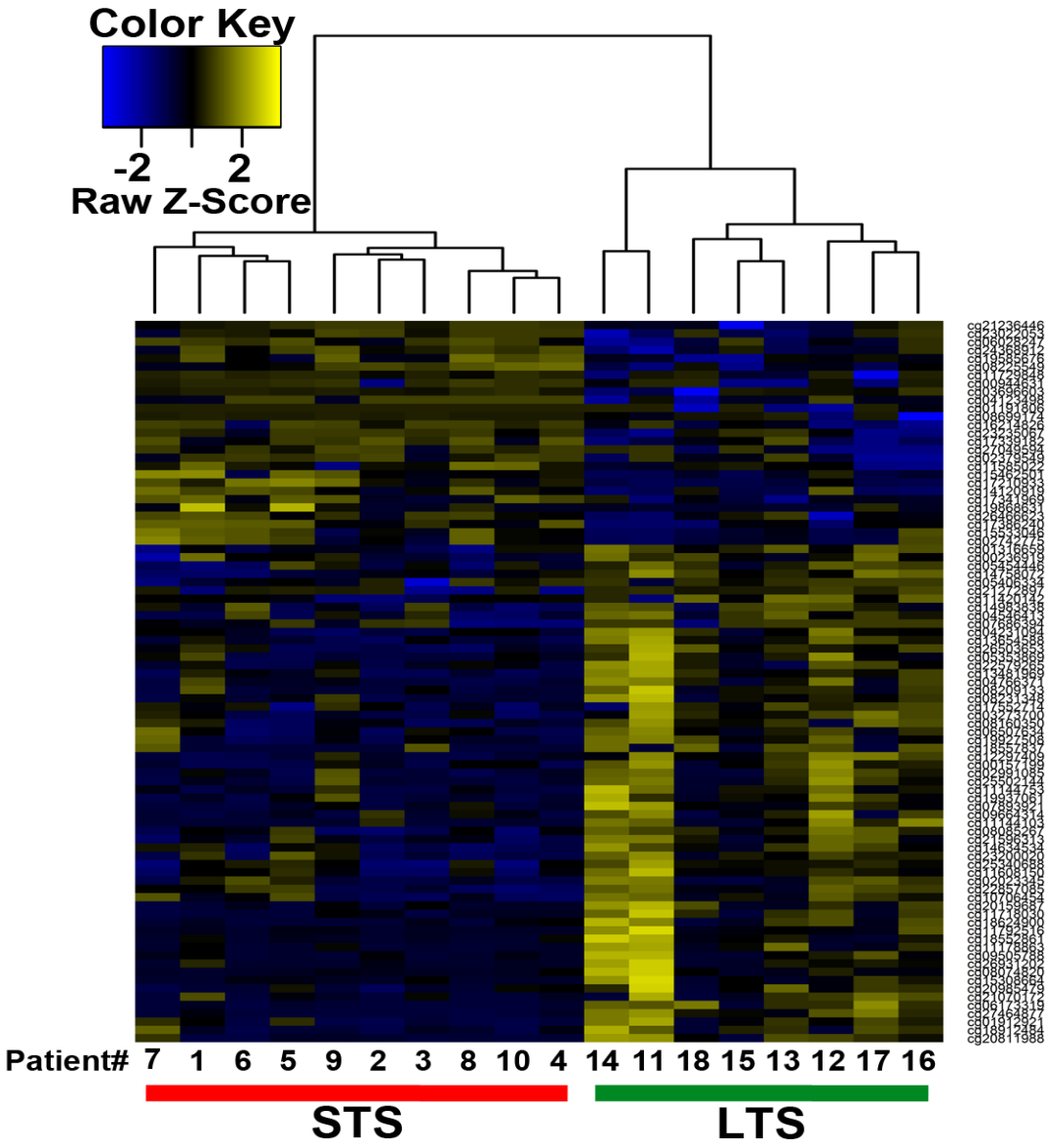


Figure 4.11. Heat map of differentially methylated probes found with an absolute Delta Beta value of greater than 0.2 in LTS and STS samples. Y-axis lists the differentially methylated probes and the x-axis shows the samples with LTS samples highlighted in green and STS samples highlighted in red.

4.3.5. mRNA expression analysis

Gene expression profiling was performed using Cufflinks to identify differentially expressed genes (DEG) in Outlier cohort. The comparison identified 615 differentially

expressed genes. A heatmap was generated to show the clear separation of LTS and STS (Figure 4.12). To reveal patterns of expression, we also performed unsupervised hierarchical clustering with genes known to be related genome instability included in the CIN 70 gene list (Carter, Eklund et al. 2006) using Euclidean distance. The trend observed was that LTS patient cluster generally has lower expression of CIN70 genes compared to STS patient.

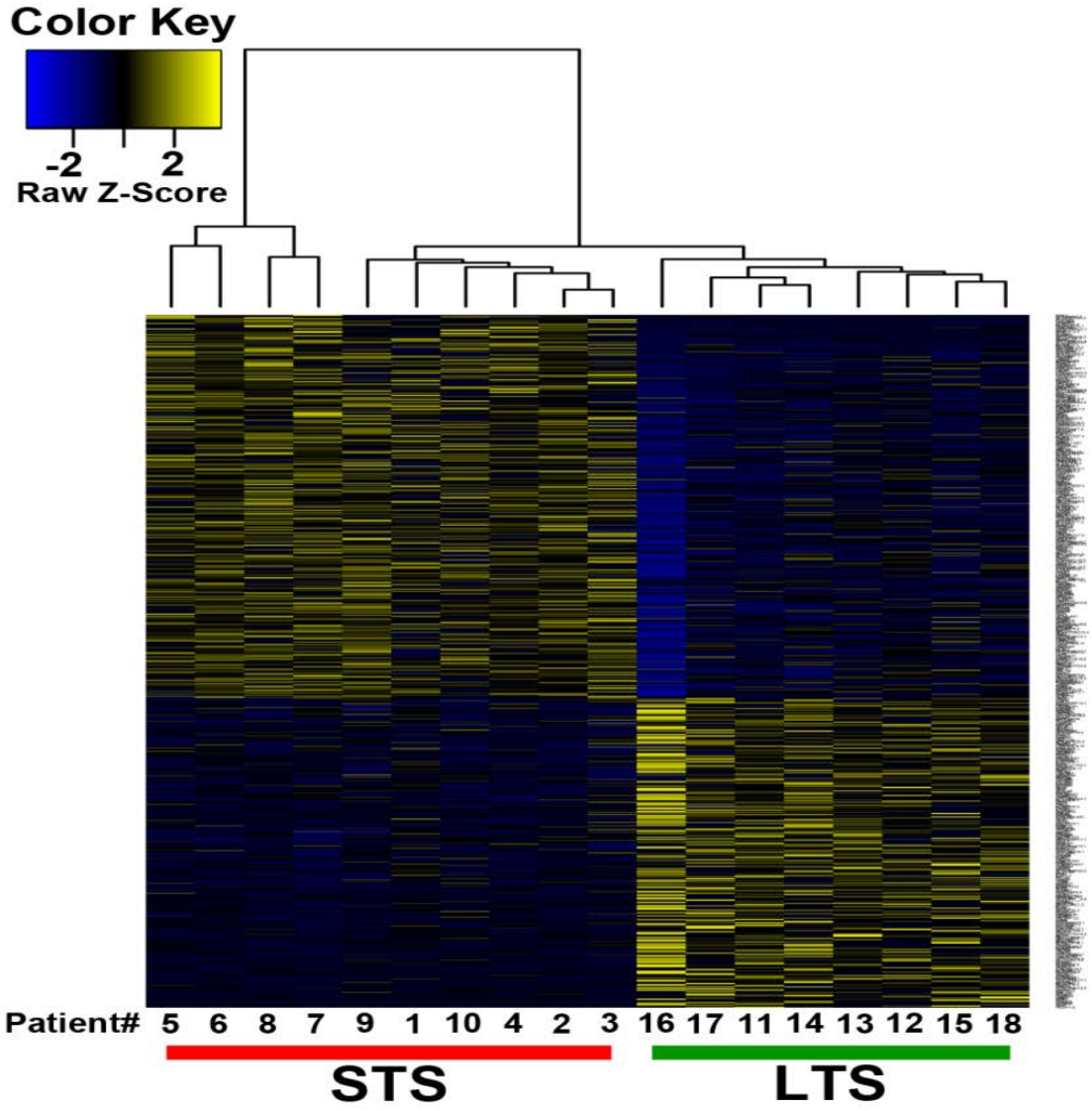


Figure 4.12. Analysis of differentially expressed genes in Outlier cohort. Hierarchal clustering of 615 genes distinguishing LTS from STS group.

In order to gain a better understanding of altered biological pathways and distinct cancer network patterns, we performed IPA (Ingenuity Systems) core analysis and biological concept enrichment analysis. All analysis was performed in IPA core analysis module using default settings. The most significant IPA canonical pathways for LTS-STS comparison seem to be involved in immune response and lipid metabolism.

The Disease and functions analysis portion of IPA enriched biological functions that are significantly altered among our differentially expressed genes (DEG). Significant categories were sorted based on their p-values and with a minimum of 10 supporting molecules. 59 out of 65 categories are mapped to “Cancer” and “Neurological Disease” functional annotations. We then applied the output of upstream regulator analysis to discover potential functional regulator in our DEG list. Six upstream regulators were identified as activated and three were identified as inhibited using z-score of 1 as filtering threshold. One of the activated regulator in STS was STAT5 a/b, NFKB and IFNG. Inhibited regulators in STS (thus activated in LTS) were MAPK1 and ERK1/2 and ESR1.

Two interesting highly scored and activated regulators NFKB and IFNG have two commonly regulated genes which is illustrated in a combined network (Appendix D)

In order to determine representative biology associated with the DEG between LTS and STS samples we performed biological concept enrichment analysis using ClueGO software. DEGs for LTS and STS were analyzed as two separate gene lists using GO Biological Process, GO Cellular Component, KEGG, Reactome, and Wiki Pathways.

The network-based modeling takes knowledge from prebuilt canonical pathways as well as potential network rules from each sample. Those analyses revealed a number of interesting biological concepts associated with LTS and STS gene expression changes. The LTS enriched biological concepts include those associated with development and morphogenesis, and also mTOR signaling pathway. The STS concepts are centered

around metabolic processes, APC degradation, and immune processes associated with MHC class I antigen presentation (figure 4.13).

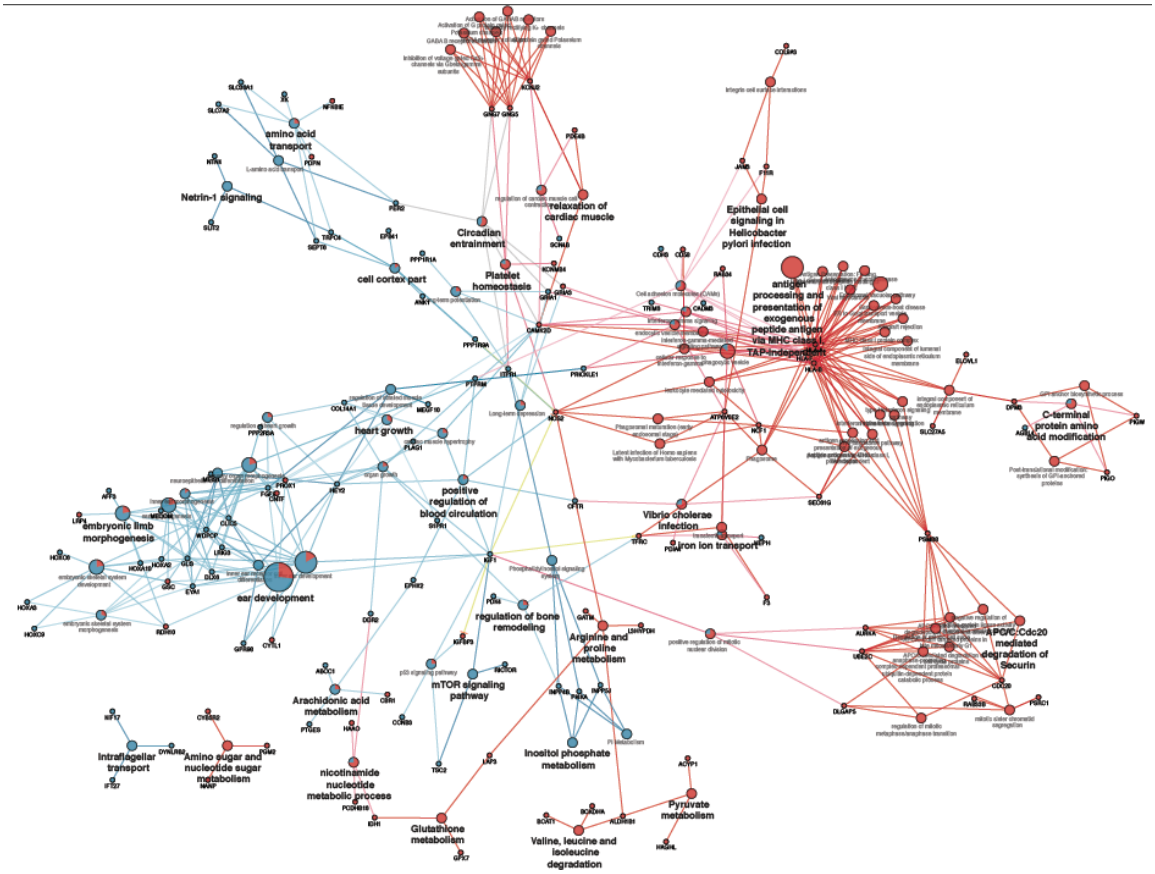


Figure 4.13. Biological concept analysis of differentially expressed genes between LTS and STS. Enrichment analysis was performed and results visualized with ClueGO v2.1.5 + CluePedia v1.1.5. The nodes in the diagram represent an enriched ontology category, pathway, or gene. Blue nodes are enriched for LTS samples and red are enriched in STS samples. Nodes with enrichment for both are represented as pie charts. Node size corresponds to how many genes in that ontology/pathway concept. Edges represent the statistical association between the nodes based on gene membership and location with gene ontology.

4.3.6. Combined methylation and mRNA expression analysis

We combined gene expression and methylation analysis in order to identify several interesting genes that were hypomethylated and over-expressed in STS. Of note, we identified SLC10A4 and FAM24B. This finding highlights the possible epigenetic activation of those genes in the STS tumor progression. Furthermore, by overlapping outlier methylation analysis with TCGA validation data, we note two genes, DOCK2 and MIR886, were consistently hypomethylated in STS compared to LTS. They were previously found to be critical in regulating cell proliferation and migration.

If we used a more lenient cutoff of beta value to call differentially methylated loci, 888 probes were found to be differentially methylated between survival groups and 598 were annotated with gene symbols. The overlap is illustrated in figure 4.14. DNA methylation is one of the most vital epigenetic mechanisms that could regulate gene expression. Methylation and gene expression are often, but not always, correlated in the direction of negative for promoter CpG islands and positive for gene-body regions.

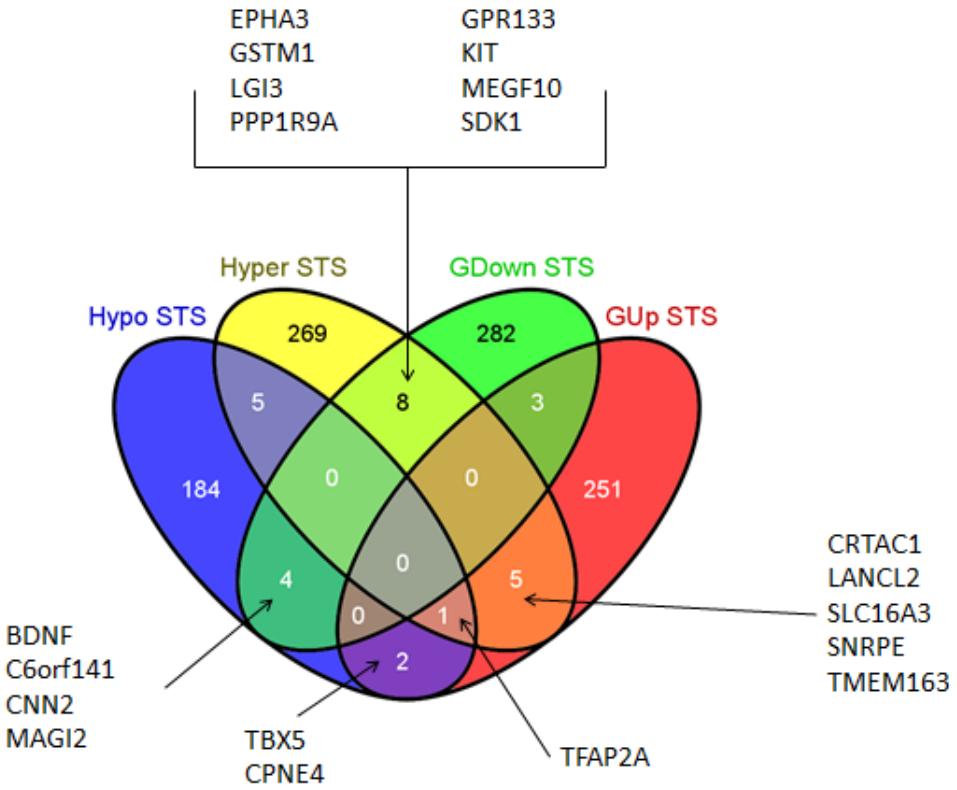


Figure 4.14. Four-way Venn diagrams were used to identify genes that were overlapped between methylation and expression analysis. The Blue oval represents genes annotated with hypomethylation in the STS group, the Yellow oval represent genes with hypermethylation in the STS group, the green oval represents genes with decrease expression in the STS group, and the Red oval represents genes with increased expression in the STS group. Genes that overlap in multiple lists are listed.

4.3.7. Data visualization

As seen from various methods and analysis outlined above, there has not been an comprehensive visualization approach to integrate exome, RNA, whole-genome and methylation data in a coherent fashion and abstract higher-level information such as pathway and network analysis. In light of this, we applied several data representation tools to incorporate those data into one figure for better identification of correlations among biological processes. Mutational “lollipop” plot (figure 4.14) could be used to visualize the distribution of functional protein changes across cohorts. Circos plot (figure 4.15) can be applied to integrate all key alterations from multiple assays into one united diagram.

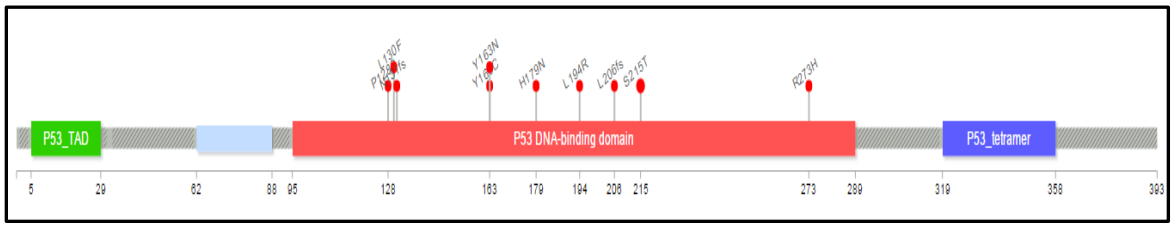


Figure 4.15. A mutational "lollipop" plot, which illustrates the discovered amino acid changes for TP53 in our Outlier dataset in the context of previously reported protein mutations (red dot, if not reported, color is blue) and known domains.

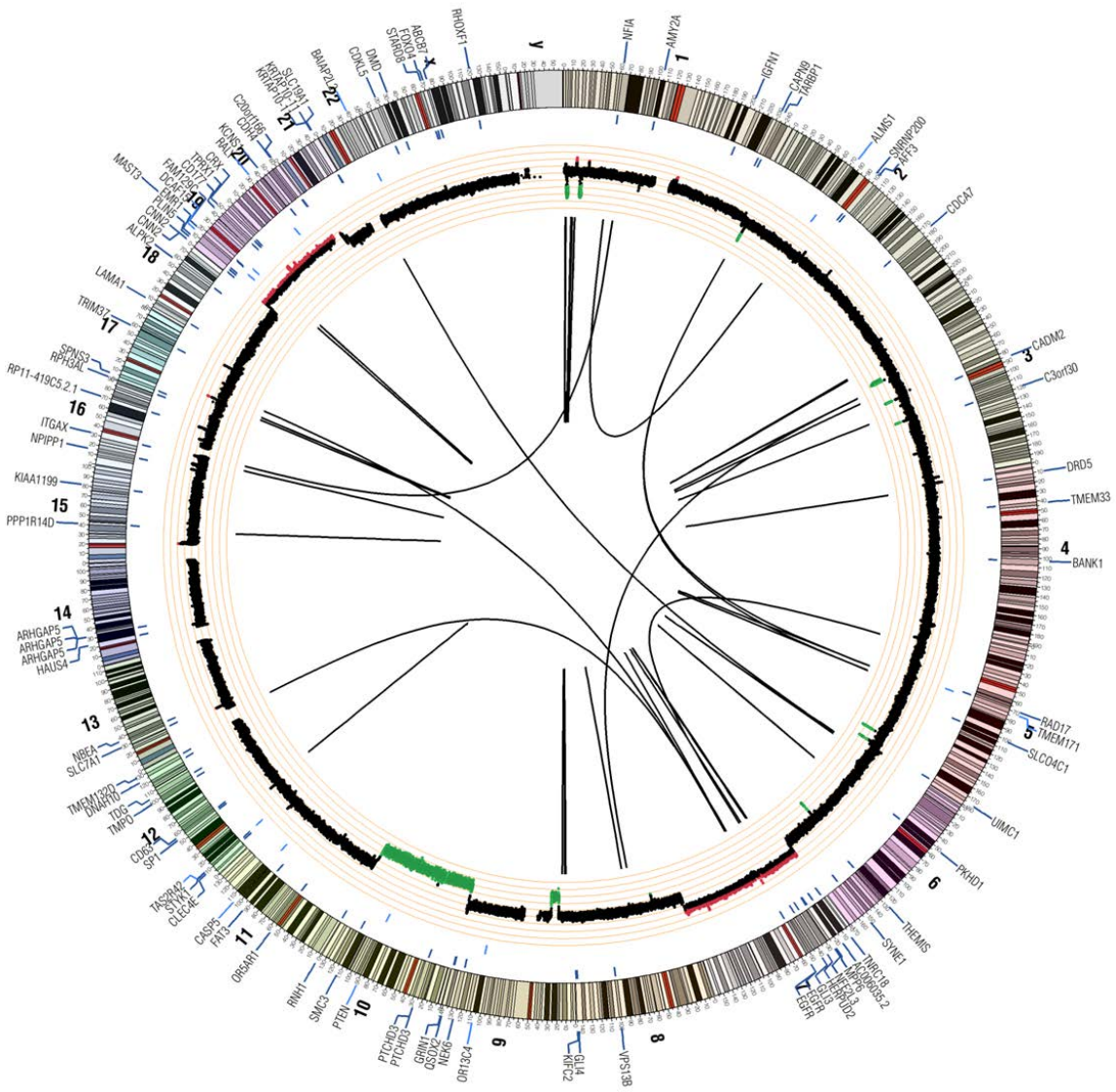


Figure 4.16. Circos plot summarizes all significant genomic events that were identified in one GBM patient. Copy number changes are shown in the inner circle plot with red marking amplifications and green marking deletions. SNVs are indicated with dark blue tick marks and INDELs are indicated with light blue tick marks. Translocations are linked using arc inside the circle.

4.4. Conclusions and Discussion

Massively parallel sequencing, with increasing throughput and reducing cost, make it

possible for the simultaneous measurement of several genomic features in the same biological samples. Those individual characteristics of each patient could then be used to design tailored medical treatment which may show higher susceptibility to this patient's tumor. Since glioblastoma multiforme is still a common and extremely malignant form of brain cancer, there is urgency to retrospectively explore the genomic difference of short and long term survivors as well as prospectively influence the treatment planning. Despite the general poor prognosis for patients with Glioblastoma multiforme (GBM), a proportion survives well beyond the median survival of 14 months following diagnosis. To elucidate molecular features associated with disproportionately protracted survival, we conducted deep genomic comparative analysis of a cohort of patients receiving standard therapy (surgery plus concurrent radiation and temozolomide) wherein "GBM outliers" were identified: patients who responded (long-term survivor, LTS) versus those who failed rapidly (short-term survivor, STS). Those genomic alterations were assessed using next generation sequencing technology at the level of DNA copy number, DNA methylation, DNA somatic mutation and mRNA expression. We demonstrated that we were able to identify the genetic variations of GB patients with outlying survival. Copy number analysis found out that both STS and LTS have similar frequency of common gains and losses of GB such as EGFR (chromosome 7), CDKN2A (chromosome 9) and PTEN (chromosome 10) but LTS showed a significantly "nosier" genome in other regions. Whole-exome sequencing detected frequently mutated genes from previous studies (EGFR, NF1, TP53, PTEN, etc) for both groups but displayed the trend of higher number of somatic mutation in LTS. Methylation analysis also presented distinct epigenetic modifier and functional patterns between STS and LTS which may affect key regulatory functions. Differential expression and network analysis reveals enriched biological processes associated with development in LTS and metabolic processes in STS.

Those genomic signatures may serve as classification factors and predicts vulnerability of GBM to standard therapy.

Those high level integration and interpretation require accurate and trustworthy prior steps (e.g. SNV calling, removal of mouse contamination) as prerequisite. Just like groundwork foundation is important for building construction, the complete weight of the analysis “building” relies on the prerequisite “foundation”. For example, without accurate SNV calling and CNV analysis, the association generated from co-mutation analysis and circos investigation would easily lead to wrong conclusions.

The limitations of the outlier work include: 1. We only have a medium-sized cohort and patient recruitment was based on the patient visit and thus not totally randomized. 2. The outlier study did not take into account the tumor heterogeneity which is common in GBM tumor types and therefore could affect variations within each survival group. 3. The outlier cohort lacks proteomics data. As a result, we were unable to fully investigate those genomic variants that are not expressed in protein form.

Previous studies, where large scale genomic characterization was implemented, have associated altered retinoic acid signaling (PMID: 21346226), enhanced immune-related gene expression (PMID: 22802421), distinct DNA methylation profiles (PMID:23291739), and MGMT methylation and IDH1/2 mutation status (PMID: 24615357) with long-term survival in GB. To date, there is no genomic study that comprehensively examines the outliers of primary GB patients from both ends of survival spectrum. Although the Cancer Genome Atlas (TCGA) GBM database provides genomic analyses of primary GB tumors, samples with “multi-omics” data including copy number variants, exome sequencing, mRNA expression profiles and global methylation data for the GB outliers are only available for six GB specimens. Thus, we identified glioma patients and employed Next Generation Sequencing (Exome, RNA, and Whole

Genome Sequencing) and methylation profiling to identify genetic, epigenetic, and transcriptomic differences between long-term survivors (LTS, OS > 18 months) and short-term survivors (STS, OS < 12 months).

The observation of more genomic alterations in treatment naïve primary GB tumors in LTS compared to STS may be due to the following mechanism. It is possible that those patients are actually responders to current standard of care for GBM treatment. The tumor cells that try to gain advantages of fitness and growth with more genetic abnormalities perhaps be vulnerable to standard therapeutic approaches. Such trade-off between growth and drug resistance is inevitable owing to limited resources in organisms. Tumors usually have stronger competency than other cells to adapt now and mutate later. But when their adaption capability mismatch with novel and rapidly changing environments (treatment in our case), those cells might be pushed over the edge of the cliff and go through apoptosis. Specifically, high expression of CIN 70 signatures and fewer genomic aberrant events together make the tumor in STS more immune (resistant) to the alkylating chemo and radiation therapy and therefore associated with poor prognostics.

When an aberration has been identified and characterized, prior knowledge from previous literature could be applied to predict associated therapeutic implications and Contraindication. The range of knowledge base could be defined with different purposes for levels of FDA-approved drug, repurposing drug and investigational drug (Prados, Byron et al. 2015). Interestingly, the drug prediction lists 2.4 fold more chemotherapy agents for LTS patients (on average 3.13 per patient) compared to STS patients (on average 1.3 per patient) based on their genomic alterations. This match is consistent with our hypothesis that LTS patients are suitable for standard therapy and therefore become potential responders with improved survivals.

In our outlier cohort, we also noticed a potential gender inequality. The majority of female patients were long term survivors and on the contrary, most male patients were short term survivors. In order to investigate the role of age in prediction of survival, we performed statistical testing to see if such gender effect also holds in TCGA with larger number of samples. There is no significant difference (P = 0.332) in survival between male and female TCGA samples that have been filtered for treatment and survival to match the samples in the Outliers project. Moreover, even within the STS and LTS, TCGA dataset did not show a significant divergence between male and female patients (STS p-value: 0.8581; LTS p-value: 0.9628).

Our long-term goal is to help patients with GBM, particularly the patient who is in front of us today. We posit that genomic instability predicts vulnerability of GBM to standard therapy and coupled with genetic and epigenetic signatures may identify patients where front-line entry into alternative, targeted regimens would be a preferred, more-efficacious management. Those patients who are categorized to be potential LTS patient could just stick to standard therapy and for patients with STS signature, their treatment selection shall be guided by the molecular profile of targetable mutations and gene pathways that vary among patients. In light of this, molecular/genomic signatures in patient tumors may direct optimal or effective therapy selection, thereby enabling personalized treatment planning. The net result of this approach will be to have more effective therapy directed to identify features in profiled patient cancer specimens as opposed to the current paradigm of indiscriminately exposing patients to chemotherapeutic toxins and hoping for a response. Our studies have highlighted a number of genetic and epigenetic alterations occurring in STS and LTS which indicate targetable mutations and hold promise for better clinical outcomes.

CHAPTER 5

LONGITUDINAL AND INTRATUMOR HETEROGENEITY STUDIES TO REVEAL THE TEMPORAL AND SPATIAL CONTEXT OF TUMOR EVOLUTION

5.1. Introduction

5.1.1. Tumor adapts and evolves over time

Evolution is considered the ultimate oncogenesis experiment but even within a short period of time, tumor constantly adapts and evolves instead of staying static. Cancer-promoting mutations such as oncogenes and tumor suppressor gene variations do not occur at once. Over time, accumulated mistakes and selective pressures that cancer cell acquires drive tumor evolution. For example, cells in the "core" of a proliferating tumor face hypoxia and shortages of nutrients. Invasive "rim" cells possess the metastatic potential but still must adapt to micro-environment in a foreign tissue. Such selective pressures are not only limited to natural characteristics of tissues but also include the human intervention. For instance, surgery, chemotherapy, radiation therapy and drug treatment all could influence the evolutionary context for the remaining tumor cells.

From the perspective of long-term natural selection and evolution, development involving "hallmarks" of cancer may reveal the genetic progression that leads to cancer. Previous attempts have been made to identify the six types of evolutionary explanations for our vulnerability to cancer: (1) Adaption capability mismatch with novel and rapidly changing environments (e.g., Cigarette smoking→ lung cancer) (2) Co-evolution with faster-evolving pathogens (e.g., HBV,HCV → hepatocellular carcinoma and HPV→ cervical cancer) (3) Limitations on what evolution can do (e.g.,. mutations → cancer and the ability of cancer cells to evade immune system's detection) (4) Trade-offs

between traits with opposite functions (e.g.,. capacity for fast tissue repair versus risk of cancer) (5) Reproductive success at the expense of longevity (e.g., Competitiveness promoting allele enrichment in males may increase susceptibility to prostate cancer) (6) Defenses with costs as well as benefits (e.g., inflammation is crucial for defending against harmful stimuli, but it also damages tissues and makes them more vulnerable to cancer) (Aktipis and Nesse 2013).

5.1.2. Tumors are not spatially uniform

Recent NGS technologic advances have enabled more comprehensive, rapid and deeper analysis of individual cancer genomes at the single-nucleotide level. Such development contributes a lot to the studies of the longitudinal tumor progression where vital information was gathered to reconstruct evolutionary history. Unfortunately, cancer is complex and we have another important issue to address - tumor morphologic heterogeneity. Tumors are often heterogeneous with mixed populations of cells, even at a single time point. Knowledge about cancer branched evolution would shed light on the consequences of such heterogeneity for clinical treatment planning, drug discovery, mechanisms of drug resistance and biomarker validation. Despite increasing observations of intratumor heterogeneity at the histopathologic, genetic and epigenetic level, our understanding of the extent of such subclonal diversity, as well as its underlying causes, has remained relatively unclear.

Researchers have taken the strategy of sequencing spatially distinct regions in tumors and showed clear evidence of intratumor heterogeneity at genetic, transcriptomic, and functional levels. They were also able to identify the majority of known gene mutations as signatures for each region and use such information to infer phylogenetic tree and

subclonal evolution (de Bruin, McGranahan et al. 2014, Xu, DiCarlo et al. 2014, Zhang, Fujimoto et al. 2014).

In recent years, researchers have improved the resolution of ITH studies even to single cell level. Nicholas Navin (Navin, Kendall et al. 2011) and his colleagues performed breast cancer single-cell sequencing by isolating nuclei with flow-sorting and then amplifying DNA using whole genome amplification for massively parallel sequencing. They explored the population substructure and relationship between subpopulations using neighbor joining and phylogenetic lineage approaches. They found out that robust high-resolution genetic profiles could be obtained by sequencing a single cell and we can make meaningful inferences about the clonal evolution and metastasis of cancer by examining multiple cells from the same cancer (Navin, Kendall et al. 2011).

5.1.3. Integrating and monitoring tumor heterogeneity in space and time may have profound clinical influence

Heterogeneity in tumors may help them to evade detection through traditional biomarkers methodology and therefore influence clinical outcome (Murugaesu, Chew et al. 2013). The changing spatial and temporal nature of cancer during its progression implies the need for longitudinal prospective monitoring of cancer evolution and integration of clinically actionable biomarkers identified from genetic diverse intratumor regions.

The first issue such strategy could help to solve is tumor sampling bias. Tumor sampling bias may arise due to constant changes in the subclonal architecture of the tumor. Biopsies in one region of a heterogeneous primary tumor will discover genomic aberrant events for that specific region but may also miss more heterogeneous events not shared by all regions/subclones of the tumor. To make things worse, heterogeneity itself is also

dynamic and evolves over time. Current therapeutic decision making in clinical setting is often made with reference to the one biopsy from a random position in patient's tumor which is probably obtained months or years previously. Such approaches suppose to work well and guide treatment planning if and only if these aberrant genomic events occur ubiquitously throughout all subclones and continue to maintain their presence. However, we already know from previous evidence that such assumptions are not likely to be always true. In this case, comprehensive sequencing on multiple regions or even on single cell would help us to detect a complete list of variants. Following treatment based on the complete list altogether could further prevent clonal competition with alternating dominance (clonal tide model) and achieve better clinical outcome (Egan, Shi et al. 2012). Another important application is the identification of key drivers of heterogeneity, either within the tumor or between primary tumors and their metastatic sites. In contrast to linear models of tumor evolution with sequentially acquired somatic mutations, up-to-date finding is consistent with the fact that branched evolutionary variants in driver genes result in homogeneous tumor cell expansion. By integrative analysis of multiple heterogeneous regions, we could detect and infer the driver events for genomic heterogeneity (as the consensus mutation in the "trunk" of evolution tree) and may provide novel treatment to limit tumor diversity and adaptation.

Similarly, Comparison of paired primary and metastatic biopsies may raise our ability to pinpoint "trunk" thus vital genomic events for therapeutic targeting.

The third impact is the increase of capability to monitor and track the subclonal changes. Development of noninvasive diagnostic techniques would enable us to monitor and track the subclonal dynamics of tumor subpopulation. It may not only enhance our understanding of resistance mechanisms (branches are aimed and "pruned" at the expense of outgrowth of other branches with distinct resistance aberrations), but also

aims to pursue earlier, safer, more accurate and faster ways of diagnosing cancer. Monitoring the biomarkers would provide safer medical testing procedures and potentially more accurate diagnoses, leading to treatment that are more precise, according to each patient's condition. Moreover, knowing the progression of tumor over time, we may use such information to help us detect cancer in its infancy. Noninvasive diagnostics using liquid biopsy such as circulating DNA and extracellular RNA holds promises for early detection of cancer signatures and thus earlier diagnoses, treatment and even cures.

5.2. Materials and Methods

5.2.1. Patient clinical characteristics

5.2.1.1. Patient information of longitudinal study

Two patients were selected for this longitudinal study for their availability of recurrent and post-recurrent GBM tumor. Tumor DNA and RNA with matched normal blood DNA were extracted for Exome and RNA sequencing respectively. First patient is a 61 year old female with the primary tumor located at the right posterior frontal lobe. First surgery was performed to remove the tumor and obtain "primary" biopsy, followed by radiation and temozolomide treatment. Second surgery was done to remove the recurrent tumor, followed by Avastin monotherapy and then Avastin with CPT-11 combination therapy. Third surgery was taken to remove the post-recurrent tumor and patient was placed with Gliadel wafers. First patient passed away with 22 month survival. Second patient is a 55 year old female with primary GBM tumor located at left temporal lobe. Similarly, three surgeries were performed to remove "primary", "recurrent" and "post-recurrent" tumor respectively. However, the treatment for patient two is different. The treatment for primary and recurrent tumor is radiation with combination drugs of temozolomide and

erlotinib which is an EGFR inhibitor. The patient refused standard treatment for the post-recurrent tumor in the beginning and then took Avastin and CPT-11 at the same time.

5.2.1.2. Patient information for spatial study

9 patients were included with the study, SNV and copy number profiling were accessed for multiple regions of two patients' tumors using Exome sequencing and CNV of other 7 patients were determined with Array CGH (comparative genomic hybridization) (aCGH). To ensure the maximum probability of covering all subpopulation, several core regions (indicated as Enhanced region (Enh) in MRI image) and rim regions (indicated as Brain around Tumor (BAT)) were taken as biopsies.

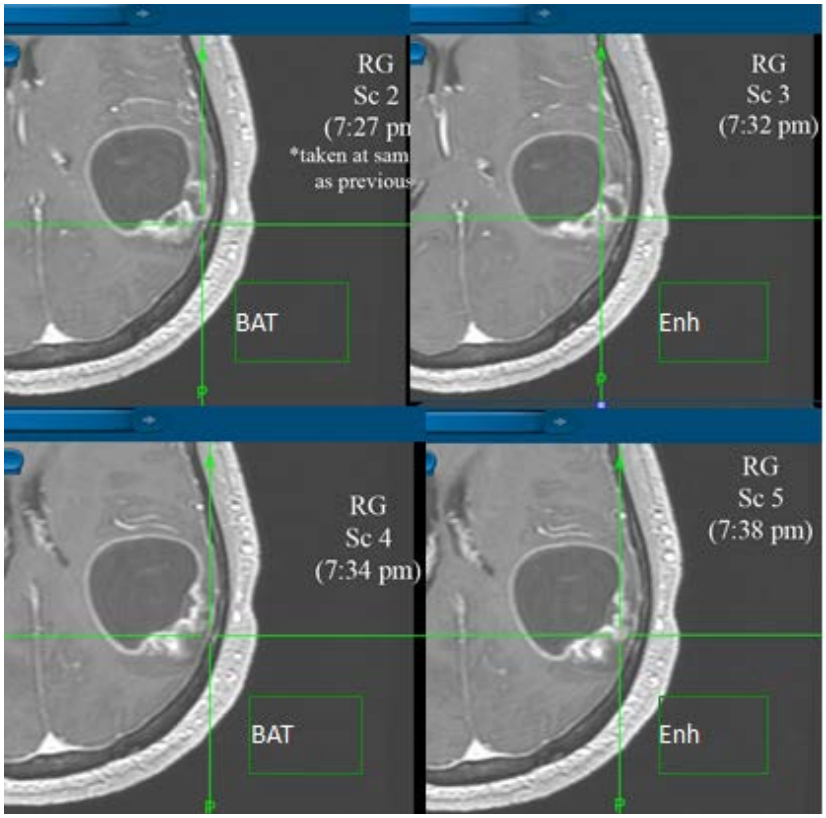


Figure 5.1. MRI images for patient 5 and present 4 different sampling regions (Courtesy of Dr. Leland Hu from Mayo Clinic)

5.2.2. DNA and RNA extraction

DNA and total RNA from fresh frozen tissue specimen were isolated using DNAeasy Blood and Tissue Kit (Qiagen #69504) and RNAeasy Mini Kit (Qiagen # 74104). Matched normal patient DNA was extracted and purified from the blood using a QIAamp DNA Blood Midi Kit/QIAamp Mini Kit from QIAGEN as previously described.

5.2.3. RNA library construction and sequencing

We constructed paired-end mRNA libraries from 2 ug of total RNA following the TruSeq RNA Sample Preparation v2 Guide, Revision A (Illumina). Libraries were amplified for 10 PCR cycles in order to reduce potential PCR duplicates. Final libraries were quantified by Bioanalyzer Finallibrary size (mode) fell between 260 and 280bp.Libraries were pooled and sequenced on the Illumina2000 HiSeq platform, usingv3 chemistry (one TruSeq PE Cluster Kit, and three TruSeq SBS 50-cycle HS kits per flow cell) (Illumina) to produce 83 x 83 paired end reads.

5.2.4. Whole exome library construction and sequencing

Whole genome libraries for exome capture were constructed from 3ug genomic DNA following the SureSelect^{XT} for Illumina PE v1.2 Kit (Agilent), with SureSelect^{XT} 70 Mb capture using V4+UTR capture oligos (Agilent). Libraries were pooled following capture, and amplified for 12 PCR cycles. Final libraries were quantified by High Sensitivity DNA chip (Agilent), and Qubit DNA High Sensitivity kit (Life Technologies). Pools averaged between 3000 and 8000 pM, with an average library size (mode) around 360 bp.

5.2.5. Alignment and Variant Calling

For whole exome sequencing, FASTQ files were aligned with BWA 0.6.2 to GRCh37.62 and the SAM output were converted to a sorted BAM file using SAMtools 0.1.18. BAM files were then processed through INDEL realignment, mark duplicates, and recalibration steps in this order with GATK 1.5 where dpsnp135 was used for known SNPs and 1000 Genomes' ALL.wgs.low_coverage_vqsr.20101123 was used for known INDELs. Lane level sample BAMs were then merged with Picard 1.65 if they were sequenced across multiple lanes. Comparative variant calling for exome data was conducted with Seurat, Strelka, Mutect and processed with Snipea.

For RNA sequencing, lane level FASTQ files were appended together if they were across multiple lanes. These FASTQ files were then aligned with STAR 2.3.1 and TopHat 2.0.8 to GRCh37.62 using ensembl.63.genes.gtf as GTF file. We calculated changes in transcript expression with Cuffdiff 2.1.1 in FPKM format using upper-quartile normalization. We then used DEseq and Edger to call differentially expressed genes (DEG) between LTS and STS groups using p-value 0.05 as cutoff. For novel fusion discovery reads were aligned with TopHat-Fusion 2.0.8 (Kim and Salzberg 2011).

5.2.6. Array CGH preparation and analysis

For array CGH, DNA from frozen tumor was extracted using the Allprep DNA/RNA mini kit (Qiagen) following the manufacturer's protocol. With FFPE samples DNA was extracted using Allprep FFPE DNA/RNA mini kit (Qiagen). Pooled male or female DNA from a commercial source (Promega) was used as a reference. Tumor samples and references (1 μg each) were digested with DNaseI and labeled with Cy-5 dUTP and Cy-3 dUTP, using the BioPrime labeling kit (Life Technologies). All labeling reactions were assessed using a Nanodrop assay before mixing and hybridized to 244,000 feature human genome CGH arrays (Agilent Technologies) according to manufacturer's

instructions (CGH enzymatic protocol v6.2; Ref # G4410-90010). Microarray slides were scanned using an Agilent 2565C DNA scanner, and the images were analyzed with Agilent Feature Extraction version 10.5, using default settings. Data were assessed with a series of quality control metrics and analyzed using an aberration detection algorithm (ADM2) (Lipson, Aumann et al. 2006) implemented in the Genomic Workbench software package (Agilent). ADM2 identifies all aberrant intervals in a given sample with consistently high or low log ratios based on the statistical score derived from the average normalized log ratios of all probes in the genomic interval multiplied by the square root of the number of these probes. This score represents the deviation of the average of the normalized log ratios from its expected value of zero and is proportional to the height, h (absolute average log ratio), of the genomic interval and to the square root of the number of probes in the interval.

5.3. Results

5.3.1. Longitudinal study

Our first purpose of this longitudinal investigation was to explore how GBM tumor evolve, adapt and change over time. Presented in figure 5.2 and 5.3, somatic SNVs was listed and overlap among primary, relapse and post-relapse was shown for patient 1 and 2 respectively. Key mutations were marked in bold. For patient one, tumors at each stage share a list of common mutations throughout the entire time, but each tumor also acquired roughly 9 times more unique SNVs. In regards to patient 2, similar result was observed but post-relapse tumor showed fewer unique SNVs compared to primary and relapse. Based on that information, a phylogenetic tree that depicts relatedness among those tumors could be generated.

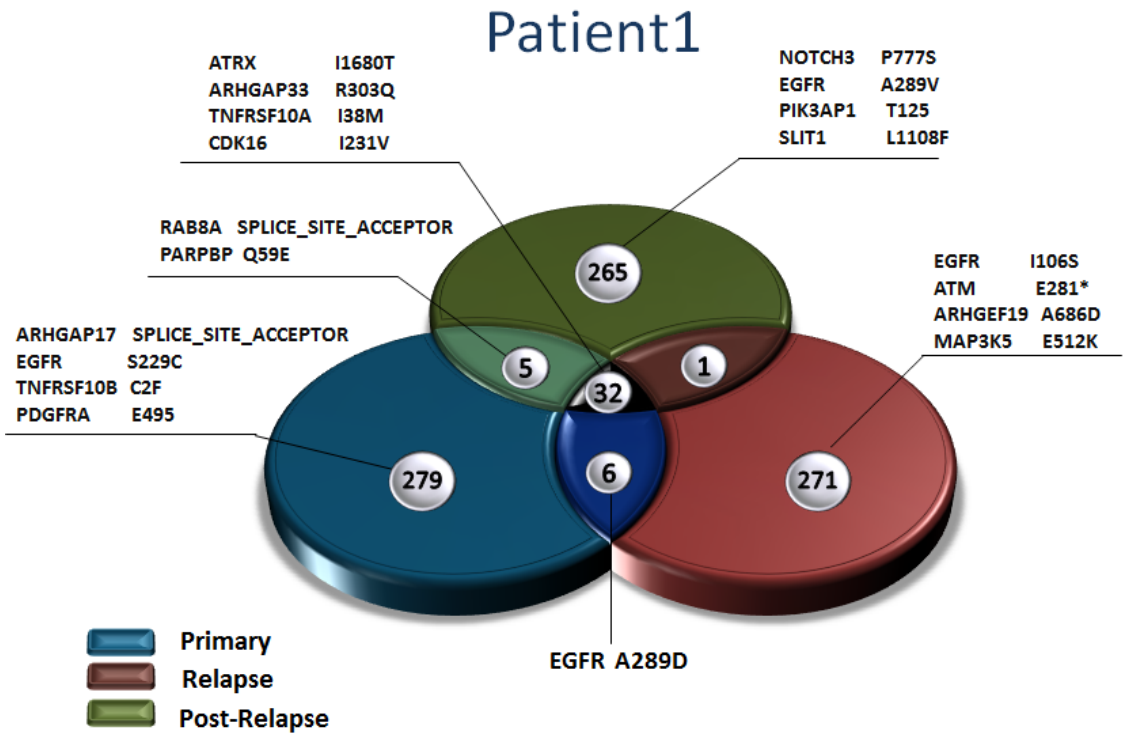


Figure 5.2. 3D Venn diagram showing unique and overlapping SNVs in primary, relapse and post-relapse tumor of GBM patient one.

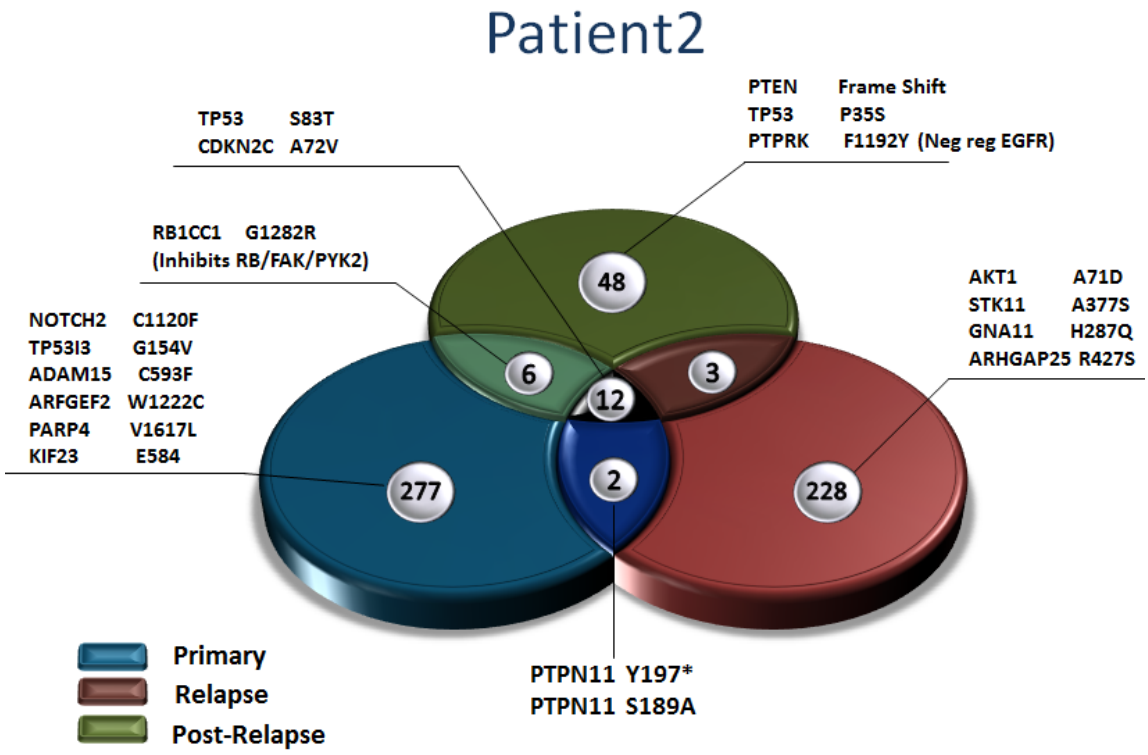


Figure 5.3. 3D Venn diagram showing unique and overlapping SNVs in primary, replapse and post-relapse tumor of GBM patient two.

We further looked into the sources of SNVs. From table 5.1, Post-relapse tumor in patient one exhibited huge numbers of somatic TMZ-associated mutations. TMZ-associated mutations were defined as C>T/G>A transition mutations that are not detected in treatment naïve primary tumor (Johnson, Mazor et al. 2014).

Patient 01		**TMZ-associated SNV**	total SNV	Percentage
	Relapse	32	297	10.77%
	Post-Relapse	203	302	67.22%
Patient 02				
	Relapse	14	245	5.71%
	Post-Relapse	10	69	14.49%

Table 5.1. Table shows us the numbers and percentage of TMZ-associated SNVs in recurrent tumors

Temozolomide, (TMZ) is an oral chemotherapy drug used in GBM standard therapy. Since it is an alkylating agent, TMZ has ability to alkylate/methylate DNA, which damages the DNA and triggers the death of tumor cells but it will also harms normal cells and potentially introduces mutations. Previous studies have showed that the majority of TMZ treated patients with hypermuations appear to be caused by TMZ-associated mutagenesis.

Our data also suggests the high number of unique mutations in recurrent tumor of patient one is attributable to TMZ exposure. This result is consistent with the finding of prior work. Therefore, optimizing glioblastoma temozolomide chemotherapy not only could increase TMZ efficiency to avoid resistance but also may keep off acquiring additional harmful mutations.

Copy number profiles were then examined for patient 1 and 2. As shown in figure 5.4, copy number variants between patients and among different stages are distinct.

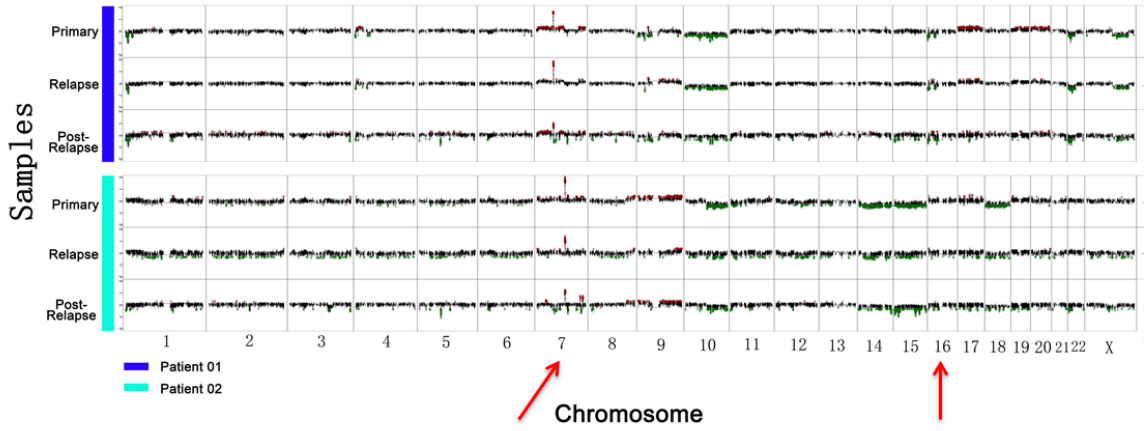


Figure 5.4. Copy number Variants compilation plot for patient 1 and 2. Red arrows indicate the dissimlar region of chromosome 7 and 16.

Spefically, in figure 5.5, EGFR amplification was contantly detected for all three stages of tumor in patient 1. However, there's no EGFR treatment applied to that patient,

instead, Avastin (a Vascular endothelial growth factor inhibitor) was tried for second line therapy.

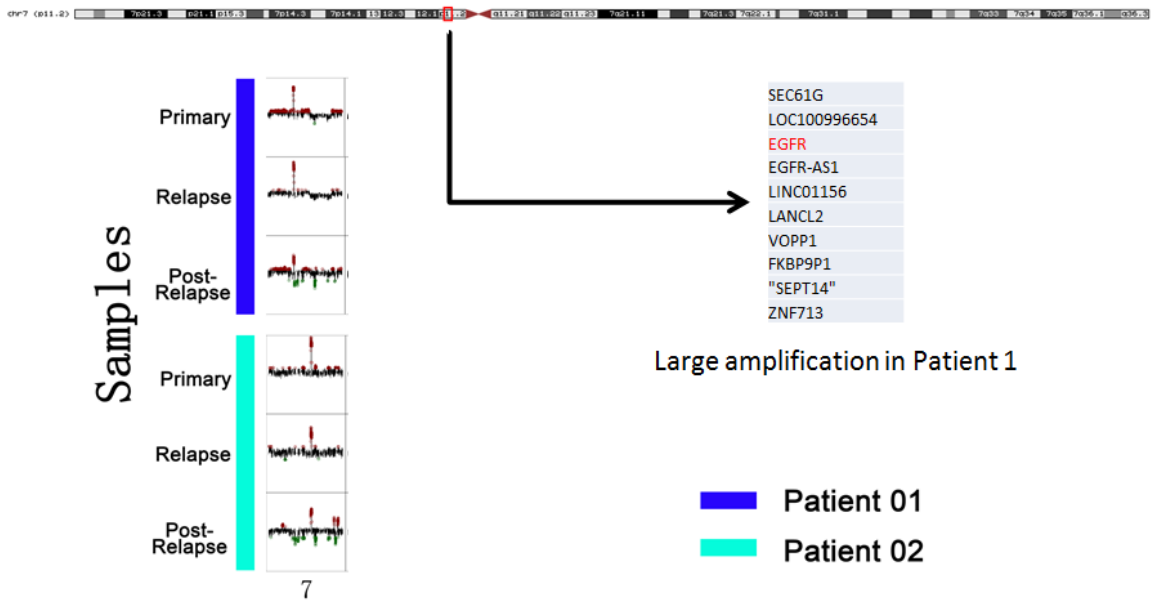


Figure 5.5. Enlarged copy number alteration at chromosome 7 with detailed genes for patient 1.

On the contrary, as shown in figure 5.6, patient 2 had CDK6 amplication in all three phases and gained additional amplication at BRAF and EZH2 region in post-relapse tumor. Ironically, Patient 2 was precribed with erlotinib (an EGFR inhibitor) but no corrosponding treatment for CDK6, BRAF or EZH2.

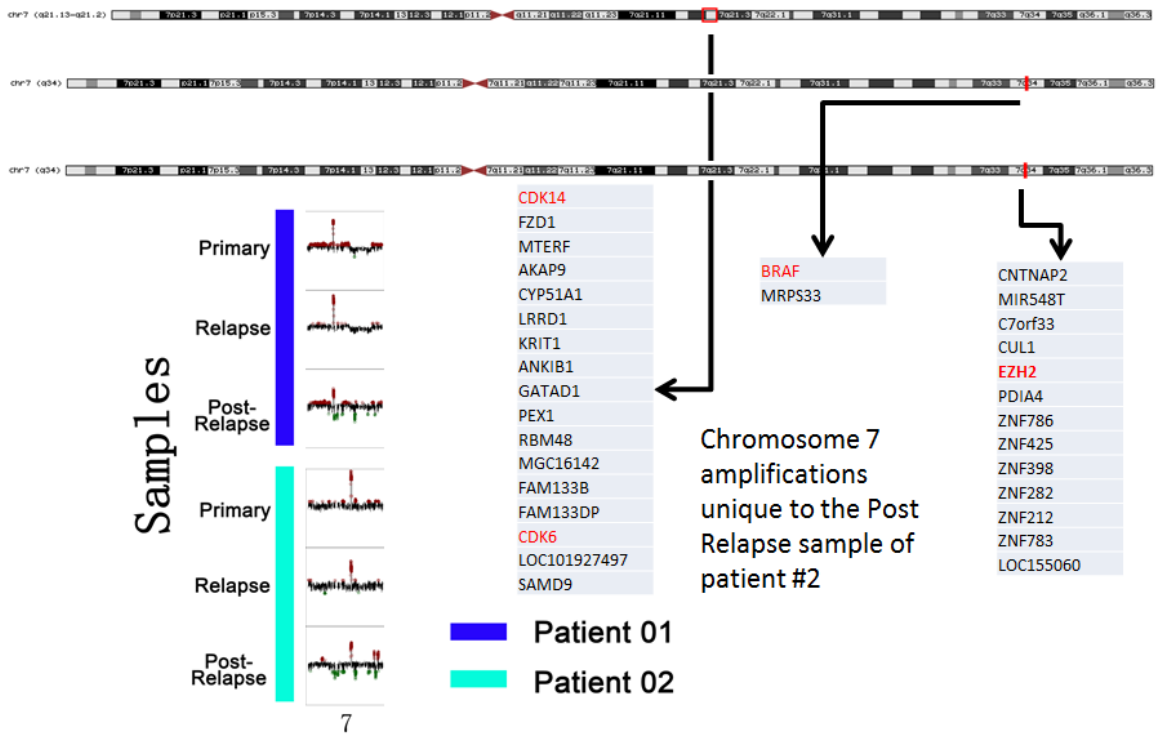


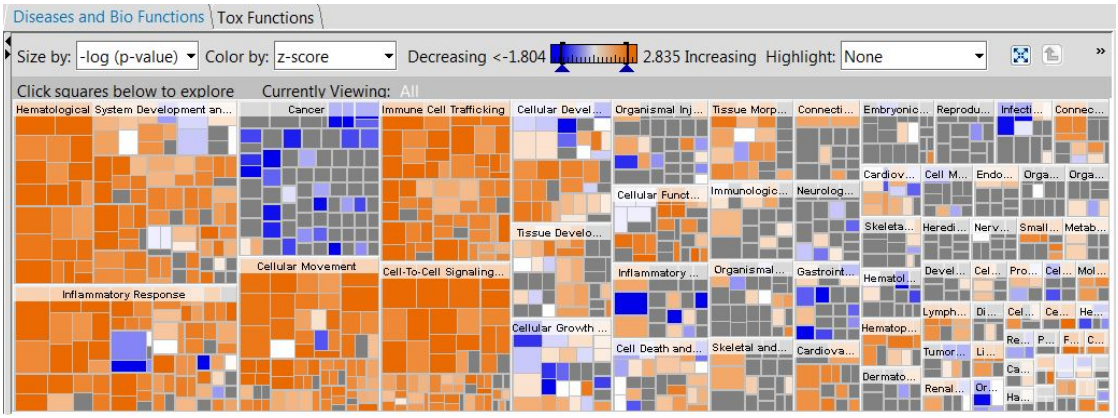
Figure 5.6. Enlarged copy number alteration at chromosome 7 with detailed genes for patient 2.

Had we known that patient two didn't have EGFR amplication, we could avoid treating her with Erlotinib. We already know that most drugs have possible side effect and unexpected interactions. The rational is why should we put a patient on a drug if we already know that drug will not do any benefit to the patient? Although there is no way to know for sure if a drug will cause side effects for patients, this "try and see" strategy at least put patients at huge risk for adverse drug reaction and mutagenesis while provide little or none benefit to treat target disease.

Differentially gene expression, canonical pathway and network analysis were then performed to assess the higher level transcriptomic changes during tumor progession based on RNA sequencing data. Ingenuity Pathway Analysis (IPA) was introduced to further identify unique pathway involved in multiple tumor recurrence (Jimenez-Marin, Collado-Romero et al. 2009). IPA (Core) Analysis is the process of mapping data to the

IPA Knowledge Base (KB) to create molecular networks and to divide data into diseases, biological functions and signaling & metabolic canonical pathways that are over represented. All analysis was performed in IPA core analysis module using default settings. Figure 5.7 shows a snapshot of bio function analysis panel in relapse and post-relapse tumors of patient 2 (other comparisons are not shown and could be found in Appendix E).

A.



B.

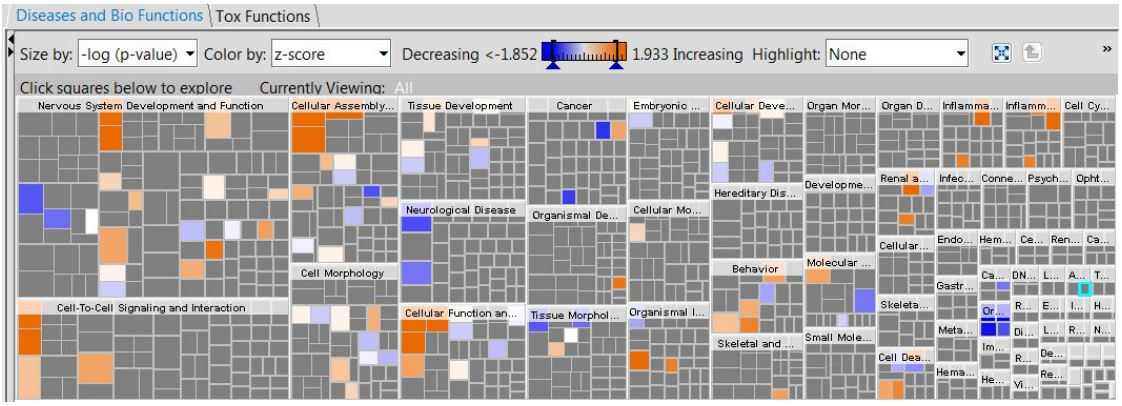


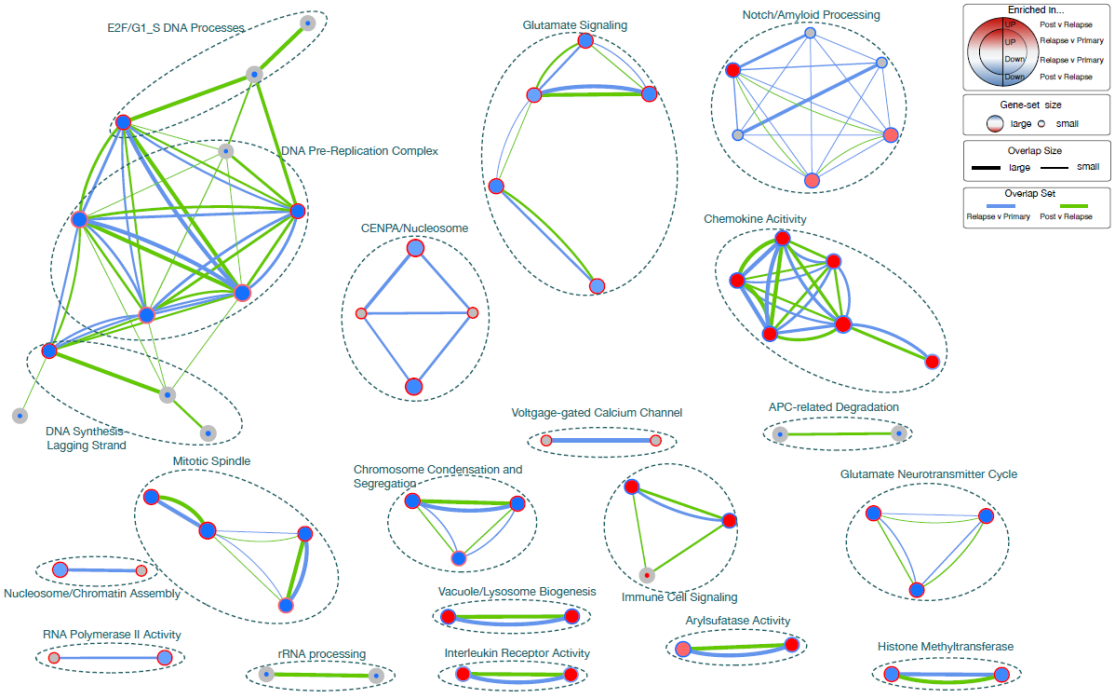
Figure 5.7. Snapshot of bio function analysis panel using IPA for relapse and post-relapse tumors in patient 2.

As expected, the enriched differentially expressed genes were mainly related to cancer (Relapse FDR P-value: 3.03 E−14; Post-relapse FDR P-value: 2.75 E−05), and Cell-to cell signaling and interaction (Relapse FDR P-value: 1.27 E−15; Post-relapse FDR P-value: 1.47 E−10). Interestingly, in relapse tumor, inflammatory response (FDR P-value: 3.33

E–14) and immune cell trafficking (FDR P-value: 8.22 E–15) pathways were highly ranked while Nervous system development and function (FDR P-value: 2.39 E-09) ranked high in post-relapse tumor. Those findings suggests immune response and inflammation may play a role in first recurrence of paitent 2's GBM and specific neuroligcal related pathways were more affected in second recurrence of the tumor.

To gain a better understanding of distinct cancer network patterns, enrichment maps technique was applied using Cytoscape (Shannon, Markiel et al. 2003) for both patients. The network-based modeling take knowledge from prebulit canonical pathways as well as potential network rules from each sample (e.g. differentially expressed, amplified, deleted and mutated genes).

A.



B.

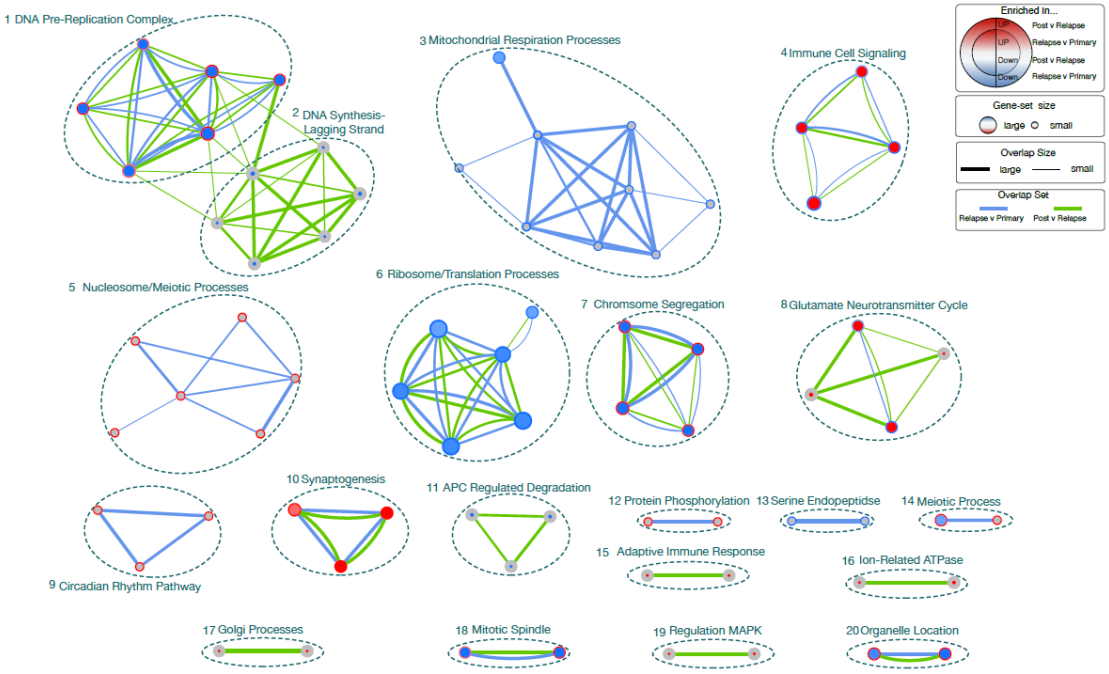


Figure 5.8. Enrichment map analysis for two patients. The size of circle indicates the size of the involved gene set. The thickness of lines shows the overlap degree of two gene sets. The color of outer circle stands for the comparison between post-relapse and relapse tumors and the color of inside circle stands for the comparison between relapse and primary tumors (red indicates up-regulation and blue indicates down-regulation).

We could see from figure 5.8 that DNA pre-replication complex and DNA synthesis lagging strand networks (down-regulated in Relapse tumor but up-regulated in post-relapse tumor) were enriched in both patients. This result suggests that relapse tumor might develop the invasiveness at the expense of DNA replication proliferation and post-relapse tumor regain such ability to grow tumor. With the help of such knowledge, more targeted and precise treatment could have been designed for both patients.

5.3.2. Spatial study

To determine the intratumor heterogeneity of copy number aberrations, we also performed multi-region whole-exome and aCGH on a total of 36 tumor regions from 9

patients. Spatial intratumor heterogeneity was identified in all nine patients.

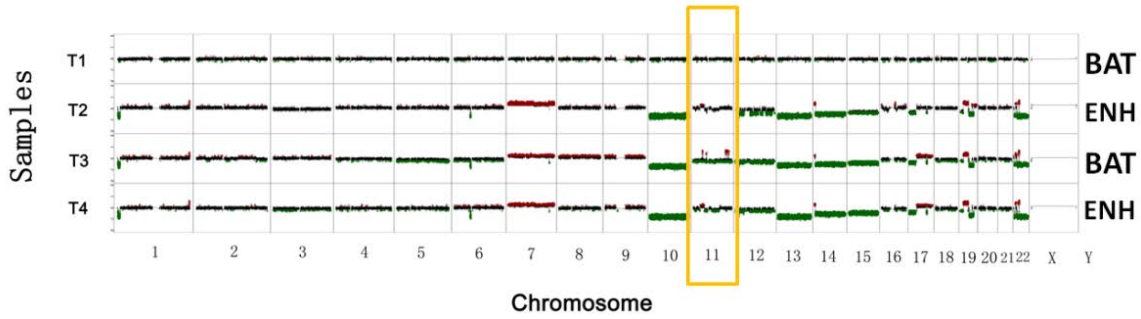


Figure 5.9. CNV compilation plot for four tumor regions of patient 5. Clear evidence of heterogeneity in chromosome 11 was highlighted in yellow box.

Specifically, as shown in figure 5.9, patients 5 displayed evidence for heterogeneity among several regions. A large fraction of genome had undergone copy number alterations expect for one BAT region; another BAT (Brain Around Tumor, Rim) shows very similar copy number profiling to other two ENH (Enhanced tumor, Core) sites except for the chromosome 11 region. Three additional copy number gains were detected in that rim region. Those novel genomic aberrations may later serve as drivers for constant tumor progression.

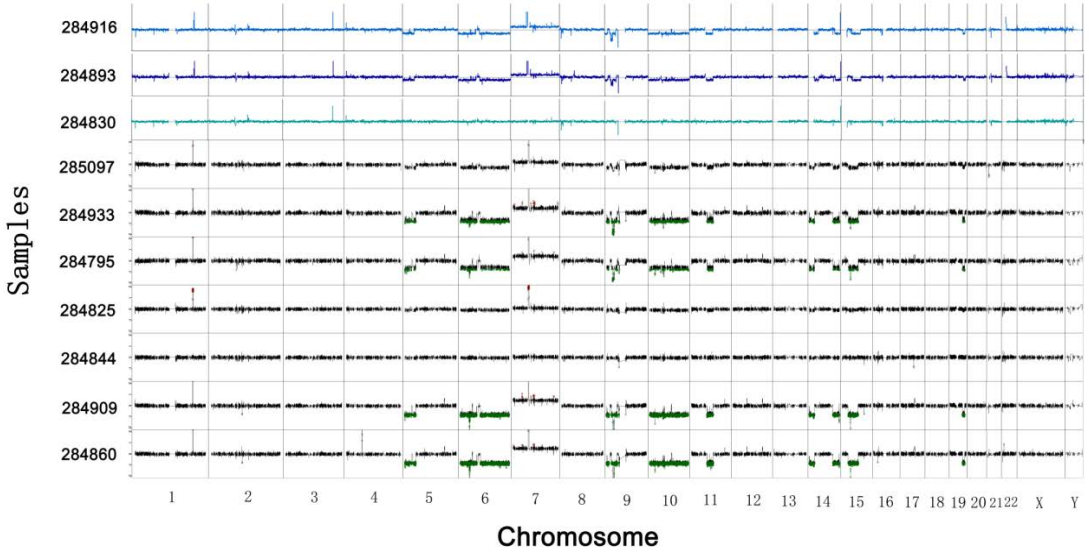


Figure 5.10. CNV compilation plot for 10 tumor regions of patient 8. Copy number variants of first three samples were assessed using aCGH and last 7 regions were evaluated by whole exome sequencing.

From figure 5.10, we could see that aCGH and sequencing data agree well on most significant CNVs, but exome sequencing data provides better coverage and resolution. Most tumor locations displayed comparable patterns of copy number profiles, while sample “284844” had few copy number alterations and sample “284825” was only present with local major amplifications but lack of global copy number events.

To better investigate GBM cancer evolution, we performed a follow-up aCGH study for two recurrent tumors of the same patient. We collected ENH and BAT regions from primary, relapse and post-relapse tumors from the same patient 1676612. This unique regionally separated spatial study, coupled with longitudinal follow-up, enables us to decipher drivers of subclonal expansion, mechanisms of tumor recurrence and resolve the evolutionary principles of GBM tumor progression.

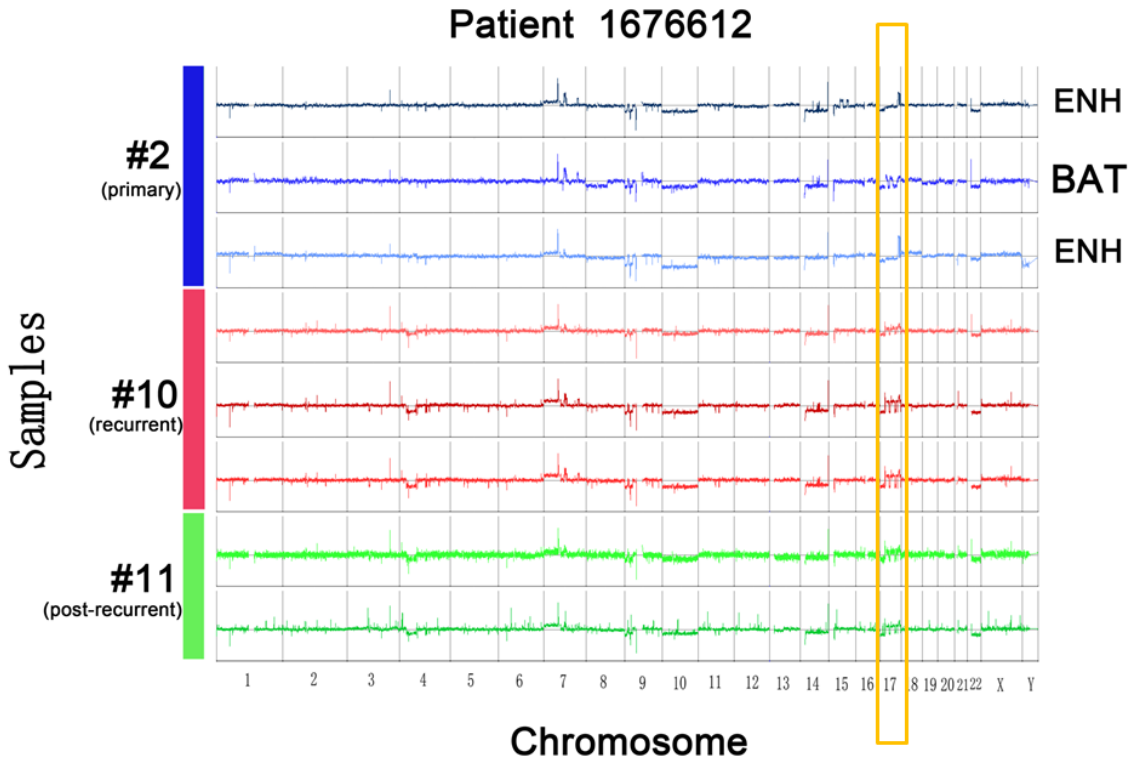


Figure 5.11. aCGH compilation plot for patient 1676612. Key events in chromosome 17 are highlighted in yellow box.

As shown in figure 5.11, relapse and post-relapse tumors had the exact same pattern of copy number aberrations in chromosome 17 as BAT (rim) region in primary tumor. This finding suggests that though surgical procedure removed the majority tumor chunk (enhanced core region in MRI image), remaining rim tumor cells would form another clone and begin to develop and eventually cause the recurrence of tumor.

5.4. Conclusion and Discussion

Through longitudinal study of primary and recurrent tumor of same patients, we showed that we could identify “trunk” mutations which exist in all staged tumor as well as unique “branch” variants for each tumor. Copy number and mutational analysis based on sequencing data greatly helped us to find potential druggable driver genes and genomic events. In addition, the higher level pathway and network analysis would enable us to group genetic aberrations using prior knowledge and thus avoid overwhelming data by reducing thousands of altered mutations and expressed gene to a limited number but more interpretable set of biological processes. Network analysis would also assist us to discover less-frequently mutated but functionally important targets. Through analyzing genomes of multiple surgically resected tumor regions, we were able to confirm the existence of intratumor heterogeneity in glioblastoma.

The limitations of the longitudinal and spatial studies include: 1. The sample size for longitudinal study is small due to restrained number of patients who underwent multiple recurrence of tumor. In order to expand the sample size, we may perform the comparisons only between primary and relapse tumors. 2. More dedicated algorithms are required to deal with the exponential growth in complexity. For each tumor sample,

we often already have multi-parametric “-omics” data. Adding time and space as confounding factors would greatly increase the integration complexity and hence needs appropriate handling. 3. The efficient delivery of molecular therapies is hampered by insufficient exploration into the mechanism of heterogeneity in cancer. Further investigation is needed and should be integrated within the process of segregation of genetic changes in tumor cells during the clonal expansion and tumor progression.

The longitudinal and spatial studies taught us a lesson and reinforced our belief and need of precision medicine. If we were able to know that patient 1 had EGFR amplification instead of VGFR alteration, and patient 1 had CDK6 instead of EGFR aberrations, we could have avoided giving patients the unnecessary drugs. If we were able to know the additional BRAF and EZH2 copy number gain in post-relapse tumor of patient 2 and uniquely acquired mutations for both patients, we could have adjusted treatment plan, quickly and efficiently, for the patients before optimal therapeutic window elapses. If we were able to realize the relapse tumor in patient 1676612 grew from rim BAT tumor cells which have distinct features from core cells, we could have designed treatment strategy based on the right information instead of sticking to data from surgically removed tumor. Those inappropriate or even wrong clinical decisions not only miss the precious treating time to relieve or control the patient’s illness, but also may possibly cause side effects and therefore worsen the current situation. And we shall not blame this on doctors, without the help and evidence from genomic profiling; they don’t have much to do with options and patient’s specificity. In this case, we shall utilize genomic biomarker classifiers – those are validated in clinical trials - to guide doctors and design therapeutic plans tailored to each patient. This could tremendously improve patient benefit compared to treating the patient with a standard of care treatment or simply based on guesses. Moreover, adaptive treatment - monitoring the genetic shift

during treatment and adjusting therapies accordingly - can considerably improve the therapeutic benefit to adverse effects ratio. Those strategies are cost-effective for patients as well as society.

Those studies also could be coupled with xenograft models. For instance, in one of our recent studies, figure 5.12 illustrates the xenograft passages could represent the patient tumor very well and provide a useful tool for clonality investigation.

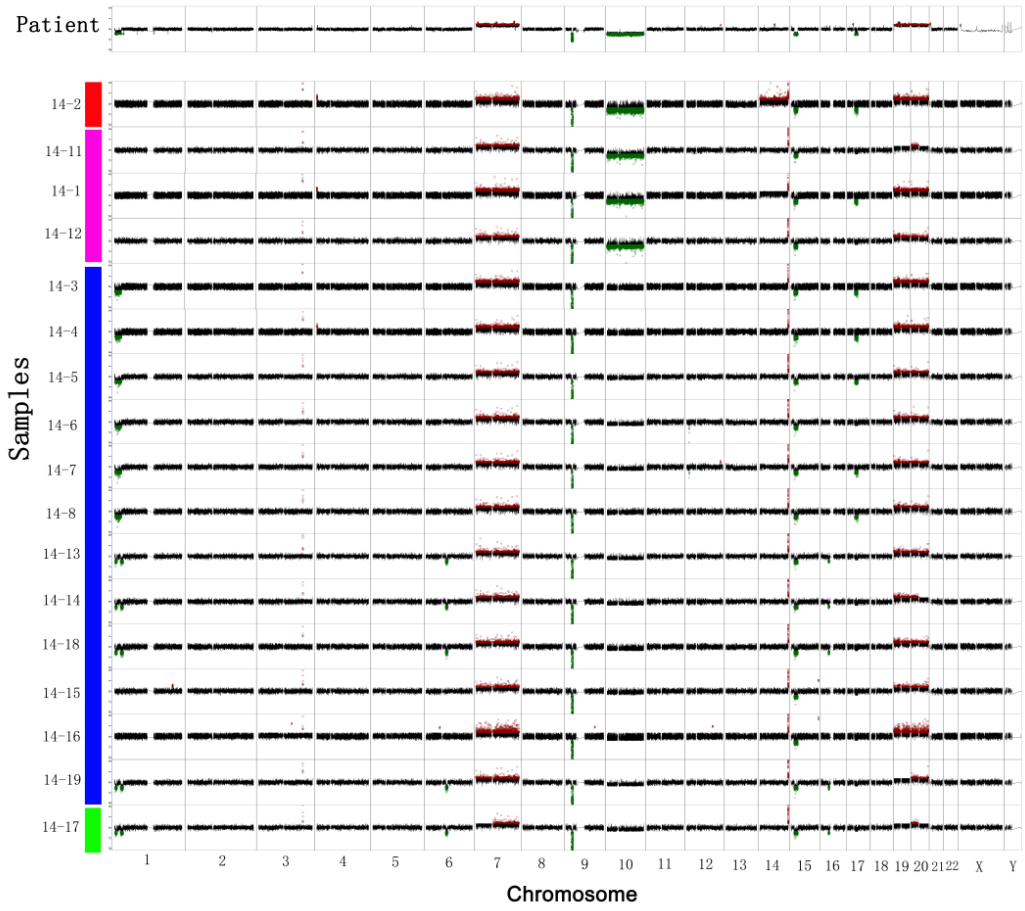


Figure 5.12. CNV compilation plot for GBM patient and xenograft passages. X-axis is the chromosome and Y-axis is different xenograft samples, colors indicate multiple clusters.

Another point we would like to stress is that spatial and temporal evolution is not independent; they are happening at the same time and may interact with each other. Natural selection, like species striving for survival, or therapy selection, like selective killing of sensitive cancer cells, facilitate the progression towards favoring remaining clones with resistance to anticancer drugs and high proliferation rate. Those spatially resolved clones evolve over time and may gain advantage and replace other regions just like the expansion of an invasive species (Korolev, Xavier et al. 2014). Potentially the intratumor heterogeneity may undermine the clinical decision making to control tumor growth since single biopsy might not represent all clones in the tumor. However, we could also turn this evolution against cancer by guiding our therapeutic interventions to exploit the evolutionary response of tumors. For example, a drug cocktail that targets the altered genetic signatures and also suppresses the evolution of resistance could be developed to control tumor. Furthermore, lineage tracing techniques could be applied to uncover and monitor the population dynamics within the tumor. Briefly, a reported gene that can be switched on in a subset of cells is used a marker for a specific clone (Alcolea and Jones 2013). The dynamics of multicolor clones can then be monitored spatially and over time, thereby providing insights into the evolution of tumors.

CHAPTER 6

CONCLUSION, DISCUSSION AND NEXT STEPS

Complexity is the very nature of cancer. Mutations, epigenetic modification, heterogeneity and rapid evolution already make cancer one of the most complex biological systems of all. Nevertheless, as the Greek philosopher Aristole said：”The whole is greater than the sum of its parts”. Complex interaction and signaling networks between the players only boost the complexity to higher order. Fortunately, the wide application of massively parallel sequencing and the idea of precision medicine, provides a bird's-eye view of cancer genome with high-resolution. In today’s genomics-driven world, revolution of cancer research, diagnosis and therapy with genomic testing is simply a must. To this end, the studies described in this thesis give examples of tools to uncover information about the altered functional modules and ways to integrate individual data to higher level from a biological organization perspective.

In Chapter 2, we demonstrated that Snipea could achieve better SNV calls and provide biological insights by prioritizing and annotating them. One logical next step would be to try to incorporate pathway and network into the ranking and annotation of Snipea. Grouping genetic aberrations using prior knowledge about network interactions allows investigation of cellular mechanism and determination of altered oncogenic pathway (Mutation and Pathway Analysis working group of the International Cancer Genome 2015). Another SNV related expansion is to call somatic mutation without matched normal samples. The availability of the normal matched samples is obviously desired but sometimes unobtainable, especially for samples in retrospective studies. In this case, we could develop an algorithm to make the distinction between germline and somatic variants based on tumor alone. The quickest and simplest way is to find a “pooled” or “standard” sample as reference. For Copy number and other analysis, we perhaps shall

use pool samples to reduce the variance. By the central limit theorem, pooling independent samples helps reduce variance in depth-of-coverage and increases precision of the methods. However, for variant caller, sticking to one commercial control sample might be beneficial. Since the chance of low allele frequency variant in this patient tends to be filtered out, but with more people, more will be filtered out. Another strategy would be “filtering and rescue” approach. We could first conduct germline/or somatic calling using one reference genome with best practice setting, then filter out the potential germline variants based on large scale databases such as dbSNP, NHLBI and 1000Genome. We then might want to rescue known cancer-related somatic variants in COSMIC for further validation. In addition, if the tumor sample is impure, which is often the case for solid tumors, we can take advantage of the fact that the allele fraction of variants would be shifted away from 50% (heterozygous mutation) or 100% (homozygous mutation).

In chapter 3, we presented an alignment-based approach could accurately and losslessly classify sequencing reads from xenograft models. Especially this method takes good care of “both” category reads which could be aligned to both genomes due to homology. In addition to k-mer and alignment-based methodologies, a “combined genome” strategy was mentioned by Yip and his colleagues (Tso, Lee et al. 2014). This approach tries to combine the host and graft reference genomes into an artificial genome, and align all sequencing reads to it. They claimed that such method handles with ambiguous reads which have DNA sequence similarity to both genomes and is expected to yield a lower false negative alignment rate and intermediate false positive rate. This method might be superior in some specific situation and shall be included in our benchmarking in the future studies. Another obvious follow up study would be to test the variant-associated drugs in vivo using xenograft models. By comparing patient tumor genome and

xenograft sequencing data, we could easily identify those druggable variants that are retained in xenograft. In light of this, investigational drugs can be tested in those model animals to assess drug safety and efficacy before going directly to expensive clinical trials. In chapter 4 and 5, we showed how to integrate and interpret multi-parametric "-omics" data using biologically motivated framework and to investigate tumor evolution over time and space. The importance of integrative analysis cannot be overemphasized; the integrated knowledge of datasets allows confirmation as well as novel hypothesis generation. For instance, if a variant is detected at DNA (Genome-sequencing) and RNA level (RNA-seq), the confidence of this call would dramatically increase. On the other hand, if any correlations were established, it may guide novel biological hypothesis. An example would be that we probably can identify new regulation factor if we discover a high correlation between a SNV and consistent altered expression and methylation status of the gene where SNV resides in. The comprehensive examination of an individual's unique genetic and biochemical make-up raises the promises of "N of 1" study for everyone. This is trending quickly from benchside (research) to bedside (CLIA). Moreover, how to detect cancer in its infancy and eliminate cancer before they form become our next daunting challenge. To overcome such challenge, it requires developing biomarkers with high specificity, better ways to select genomic subsets of interest, monitoring individuals for very early signs and incorporating genomic insights into an individual's daily health behavior and health care.

With the advance of data collection techniques and computational data processing, several association studies have successfully identified cancer-susceptibility genetic variants and original driving event of cancer. We could apply inhibitor drugs to suppress the downstream signaling of such variants. However, we are only dealing with the "consequence" of the alteration. Since we already detected the very specific variants, we

might be able to use the direct approach and to reverse that variant. It was not plausible until the CRISPER (Clustered Regularly Interspaced Short Palindromic Repeats) technology was invented. The CRISPR/Cas system takes advantage of adaptive immunity in select bacteria and archaea to confer resistance and elimination to foreign genetic elements. Genome engineering by these RNA-programmable Cas9 nucleases has been widely used for gene editing (adding, disrupting or changing the sequence of specific genes) at almost any desired location (Shalem, Sanjana et al. 2014). CRISPR aims to directly repair the defect and actually correct the DNA itself. This exciting technology offers the promise not only of better understanding of the genetic machinery but also the potential to fix that machinery with much more efficient and precise gene modification.

Lastly, while those approaches were mostly implemented on dataset in GBM, they could be easily extended to other types of tumor. Together, this dissertation aims to develop an innovative, cost-effective and time-sensitive methodology to identify clinically actionable vulnerabilities in cancer patients. The molecular/genomic signatures in patient tumors may direct optimal or effective therapy selection, thereby enabling personalized treatment planning. We would have more effective therapy directed to identify features in profiled patient cancer specimen as opposed to the current paradigm of indiscriminately exposing patients to chemotherapeutic toxins and hoping for a response.

REFERENCE

Agarwal, S., R. Sane, R. Oberoi, J. R. Ohlfest and W. F. Elmquist (2011). "Delivery of molecularly targeted therapy to malignant glioma, a disease of the whole brain." Expert Rev Mol Med **13**: e17.

Agca, Y. (2012). "Genome resource banking of biomedically important laboratory animals." Theriogenology **78**(8): 1653-1665.

Aktipis, C. A. and R. M. Nesse (2013). "Evolutionary foundations for cancer biology." Evol Appl **6**(1): 144-159.

Alcolea, M. P. and P. H. Jones (2013). "Tracking cells in their native habitat: lineage tracing in epithelial neoplasia." Nat Rev Cancer **13**(3): 161-171.

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, I. Australian Pancreatic Cancer Genome, I. B. C. Consortium, I. M.-S. Consortium, I. PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell and M. R. Stratton (2013). "Signatures of mutational processes in human cancer." Nature **500**(7463): 415-421.

Anders, S., P. T. Pyl and W. Huber (2015). "HTSeq--a Python framework to work with high-throughput sequencing data." Bioinformatics **31**(2): 166-169.

Aparicio, S., M. Hidalgo and A. L. Kung (2015). "Examining the utility of patient-derived xenograft mouse models." Nat Rev Cancer **15**(5): 311-316.

Barajas, R. F., Jr., J. G. Hodgson, J. S. Chang, S. R. Vandenberg, R. F. Yeh, A. T. Parsa, M. W. McDermott, M. S. Berger, W. P. Dillon and S. Cha (2010). "Glioblastoma multiforme regional genetic and cellular expression patterns: influence on anatomic and physiologic MR imaging." Radiology **254**(2): 564-576.

Beroukhim, R., G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S. Lander, I. K. Mellinghoff and W. R. Sellers (2007). "Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma." Proc Natl Acad Sci U S A **104**(50): 20007-20012.

Bibikova, M., B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J. B. Fan and R. Shen (2011). "High density DNA methylation array with single CpG site resolution." Genomics **98**(4): 288-295.

Borad, M. J., M. D. Champion, J. B. Egan, W. S. Liang, R. Fonseca, A. H. Bryce, A. E. McCullough, M. T. Barrett, K. Hunt, M. D. Patel, S. W. Young, J. M. Collins, A. C. Silva, R. M. Condjella, M. Block, R. R. McWilliams, K. N. Lazaridis, E. W. Klee, K. C. Bible, P. Harris, G. R. Oliver, J. D. Bhavsar, A. A. Nair, S. Middha, Y. Asmann, J. P. Kocher, K. Schahl, B. R. Kipp, E. G. Barr Fritcher, A. Baker, J. Aldrich, A. Kurdoglu, T. Izatt, A. Christoforides, I. Cherni, S. Nasser, R. Reiman, L. Phillips, J. McDonald, J. Adkins, S. D. Mastrian, P. Placek, A. T. Watanabe, J. Lobello, H. Han, D. Von Hoff, D. W. Craig, A. K. Stewart and J. D. Carpten (2014). "Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma." PLoS Genet **10**(2): e1004135.

Bradford, J. R., M. Farren, S. J. Powell, S. Runswick, S. L. Weston, H. Brown, O. Delpuech, M. Wappett, N. R. Smith, T. H. Carr, J. R. Dry, N. J. Gibson and S. T. Barry (2013). "RNA-Seq Differentiates Tumour and Host mRNA Expression Changes Induced by Treatment of Human Tumour Xenografts with the VEGFR Tyrosine Kinase Inhibitor Cediranib." PLoS One **8**(6): e66003.

Bradley, W. G., S. G. Golding, C. J. Herold, H. Hricak, G. P. Krestin, J. S. Lewin, J. C. Miller, H. G. Ringertz and J. H. Thrall (2011). "Globalization of P4 medicine: predictive, personalized, preemptive, and participatory--summary of the proceedings of the Eighth International Symposium of the International Society for Strategic Studies in Radiology, August 27-29, 2009." Radiology **258**(2): 571-582.

Brennan, C. W., R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, R. Beroukhim, B. Bernard, C. J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S. A. Shukla, G. Ciriello, W. K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, D. D. Bigner, E. G. Van Meir, M. Prados, A. Sloan, K. L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, D. W. Andrews, A. Guha, M. Iacocca, B. P. O'Neill, G. Foltz, J. Myers, D. J. Weisenberger, R. Penny, R. Kucherlapati, C. M. Perou, D. N. Hayes, R. Gibbs, M. Marra, G. B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, P. W. Laird, D. Haussler, G. Getz, L. Chin and T. R. Network (2013). "The somatic genomic landscape of glioblastoma." Cell **155**(2): 462-477.

Cancer Genome Atlas Research, N. (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." Nature **455**(7216): 1061-1068.

Carter, S. L., A. C. Eklund, I. S. Kohane, L. N. Harris and Z. Szallasi (2006). "A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers." Nat Genet **38**(9): 1043-1048.

Christoforides, A., J. D. Carpten, G. J. Weiss, M. J. Demeure, D. D. Von Hoff and D. W. Craig (2013). "Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs." BMC Genomics **14**: 302.

Church, D. M., L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, W. Hlavina, Y. Kapustin, P. Meric, D. Maglott, Z.

Birtle, A. C. Marques, T. Graves, S. Zhou, B. Teague, K. Potamousis, C. Churas, M. Place, J. Herschleb, R. Runnheim, D. Forrest, J. Amos-Landgraf, D. C. Schwartz, Z. Cheng, K. Lindblad-Toh, E. E. Eichler, C. P. Ponting and C. Mouse Genome Sequencing (2009). "Lineage-specific biology revealed by a finished genome assembly of the mouse." PLoS Biol **7**(5): e1000112.

Cibulskis, K., M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander and G. Getz (2013). "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples." Nat Biotechnol **31**(3): 213-219.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly (Austin) **6**(2): 80-92.

Conway, T., J. Wazny, A. Bromage, M. Tymms, D. Sooraj, E. D. Williams and B. Beresford-Smith (2012). "Xenome--a tool for classifying reads from xenograft samples." Bioinformatics **28**(12): i172-178.

Craig, D. W., J. A. O'Shaughnessy, J. A. Kiefer, J. Aldrich, S. Sinari, T. M. Moses, S. Wong, J. Dinh, A. Christoforides, J. L. Blum, C. L. Aitelli, C. R. Osborne, T. Izatt, A. Kurdoglu, A. Baker, J. Koeman, C. Barbacioru, O. Sakarya, F. M. De La Vega, A. Siddiqui, L. Hoang, P. R. Billings, B. Salhia, A. W. Tolcher, J. M. Trent, S. Mousses, D. Von Hoff and J. D. Carpten (2013). "Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities." Mol Cancer Ther **12**(1): 104-116.

Crews, K. R., J. K. Hicks, C. H. Pui, M. V. Relling and W. E. Evans (2012). "Pharmacogenomics and individualized medicine: translating science into practice." Clin Pharmacol Ther **92**(4): 467-475.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin and G. Genomes Project Analysis (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

de Bruin, E. C., N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, E. Gronroos, M. A. Muhammad, S. Horswell, M. Gerlinger, I. Varela, D. Jones, J. Marshall, T. Voet, P. Van Loo, D. M. Rassl, R. C. Rintoul, S. M. Janes, S. M. Lee, M. Forster, T. Ahmad, D. Lawrence, M. Falzon, A. Capitanio, T. T. Harkins, C. C. Lee, W. Tom, E. Teefe, S. C. Chen, S. Begum, A. Rabinowitz, B. Phillimore, B. Spencer-Dene, G. Stamp, Z. Szallasi, N. Matthews, A. Stewart, P. Campbell and C. Swanton (2014). "Spatial and temporal diversity in genomic instability processes defines lung cancer evolution." Science **346**(6206): 251-256.

DeRose, Y. S., G. Wang, Y. C. Lin, P. S. Bernard, S. S. Buys, M. T. Ebbert, R. Factor, C. Matsen, B. A. Milash, E. Nelson, L. Neumayer, R. L. Randall, I. J. Stijleman, B. E. Welm and A. L. Welm (2011). "Tumor grafts derived from women with breast cancer

authentically reflect tumor pathology, growth, metastasis and disease outcomes." Nat Med **17**(11): 1514-1520.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Douillard, J. Y., F. A. Shepherd, V. Hirsh, T. Mok, M. A. Socinski, R. Gervais, M. L. Liao, H. Bischoff, M. Reck, M. V. Sellers, C. L. Watkins, G. Speake, A. A. Armour and E. S. Kim (2010). "Molecular predictors of outcome with gefitinib and docetaxel in previously treated non-small-cell lung cancer: data from the randomized phase III INTEREST trial." J Clin Oncol **28**(5): 744-752.

Egan, J. B., C. X. Shi, W. Tembe, A. Christoforides, A. Kurdoglu, S. Sinari, S. Middha, Y. Asmann, J. Schmidt, E. Braggio, J. J. Keats, R. Fonseca, P. L. Bergsagel, D. W. Craig, J. D. Carpten and A. K. Stewart (2012). "Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides." Blood **120**(5): 1060-1066.

Eisenberg, E. and E. Y. Levanon (2013). "Human housekeeping genes, revisited." Trends Genet **29**(10): 569-574.

Ewing, A. D., K. E. Houlahan, Y. Hu, K. Ellrott, C. Caloian, T. N. Yamaguchi, J. C. Bare, C. P'ng, D. Waggott, V. Y. Sabelnykova, I.-T. D. S. M. C. C. participants, M. R. Kellen, T. C. Norman, D. Haussler, S. H. Friend, G. Stolovitzky, A. A. Margolin, J. M. Stuart and P. C. Boutros (2015). "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection." Nat Methods.

Forshew, T., M. Murtaza, C. Parkinson, D. Gale, D. W. Tsui, F. Kaper, S. J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton and N. Rosenfeld (2012). "Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA." Sci Transl Med **4**(136): 136ra168.

Garralda, E., K. Paz, P. P. Lopez-Casas, S. Jones, A. Katz, L. M. Kann, F. Lopez-Rios, F. Sarno, F. Al-Shahrour, D. Vasquez, E. Bruckheimer, S. V. Angiuoli, A. Calles, L. A. Diaz, V. E. Velculescu, A. Valencia, D. Sidransky and M. Hidalgo (2014). "Integrated next-generation sequencing and avatar mouse models for personalized cancer treatment." Clin Cancer Res **20**(9): 2476-2484.

Goya, R., M. G. Sun, R. D. Morin, G. Leung, G. Ha, K. C. Wiegand, J. Senz, A. Crisan, M. A. Marra, M. Hirst, D. Huntsman, K. P. Murphy, S. Aparicio and S. P. Shah (2010). "SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors." Bioinformatics **26**(6): 730-736.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hjelm, B. E., B. Salhia, A. Kurdoglu, S. Szelinger, R. A. Reiman, L. I. Sue, T. G. Beach, M. J. Huentelman and D. W. Craig (2013). "In vitro-differentiated neural cell cultures progress towards donor-identical brain tissue." Hum Mol Genet **22**(17): 3534-3546.

Isella, C., A. Terrasi, S. E. Bellomo, C. Petti, G. Galatola, A. Muratore, A. Mellano, R. Senetta, A. Cassenti, C. Sonetto, G. Inghirami, L. Trusolino, Z. Fekete, M. De Ridder, P. Cassoni, G. Storme, A. Bertotti and E. Medico (2015). "Stromal contribution to the colorectal cancer transcriptome." Nat Genet **47**(4): 312-319.

Jameson, J. L. and D. L. Longo (2015). "Precision Medicine - Personalized, Problematic, and Promising." N Engl J Med.

Jiao, W., S. Vembu, A. G. Deshwar, L. Stein and Q. Morris (2014). "Inferring clonal evolution of tumors from single nucleotide somatic mutations." BMC Bioinformatics **15**: 35.

Jimenez-Marin, A., M. Collado-Romero, M. Ramirez-Boo, C. Arce and J. J. Garrido (2009). "Biological pathway analysis by ArrayUnlock and Ingenuity Pathway Analysis." BMC Proc **3 Suppl 4**: S6.

Johnson, B. E., T. Mazor, C. Hong, M. Barnes, K. Aihara, C. Y. McLean, S. D. Fouse, S. Yamamoto, H. Ueda, K. Tatsuno, S. Asthana, L. E. Jalbert, S. J. Nelson, A. W. Bollen, W. C. Gustafson, E. Charron, W. A. Weiss, I. V. Smirnov, J. S. Song, A. B. Olshen, S. Cha, Y. Zhao, R. A. Moore, A. J. Mungall, S. J. Jones, M. Hirst, M. A. Marra, N. Saito, H. Aburatani, A. Mukasa, M. S. Berger, S. M. Chang, B. S. Taylor and J. F. Costello (2014). "Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma." Science **343**(6167): 189-193.

Joo, K. M., J. Kim, J. Jin, M. Kim, H. J. Seol, J. Muradov, H. Yang, Y. L. Choi, W. Y. Park, D. S. Kong, J. I. Lee, Y. H. Ko, H. G. Woo, J. Lee, S. Kim and D. H. Nam (2013). "Patient-specific orthotopic glioblastoma xenograft models recapitulate the histopathology and biology of human glioblastomas in situ." Cell Rep **3**(1): 260-273.
Kim, D. and S. L. Salzberg (2011). "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts." Genome Biol **12**(8): R72.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable and J. Hoh (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-389.

Koboldt, D. C., D. E. Larson and R. K. Wilson (2013). "Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection." Curr Protoc Bioinformatics **44**: 15 14 11-15 14 17.

Korolev, K. S., J. B. Xavier and J. Gore (2014). "Turning ecology and evolution against cancer." Nat Rev Cancer **14**(5): 371-380.

Law, V., C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou and D. S. Wishart (2014). "DrugBank 4.0: shedding new light on drug metabolism." Nucleic Acids Res **42**(Database issue): D1091-1097.

Leary, R. J., J. C. Lin, J. Cummins, S. Boca, L. D. Wood, D. W. Parsons, S. Jones, T. Sjoblom, B. H. Park, R. Parsons, J. Willis, D. Dawson, J. K. Willson, T. Nikolskaya, Y. Nikolsky, L. Kopelovich, N. Papadopoulos, L. A. Pennacchio, T. L. Wang, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, B. Vogelstein and V. E. Velculescu (2008). "Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers." Proc Natl Acad Sci U S A **105**(42): 16224-16229.

Leong, T. L., K. D. Marini, F. J. Rossello, S. N. Jayasekara, P. A. Russell, Z. Prodanovic, B. Kumar, V. Ganju, M. Alamgeer, L. B. Irving, D. P. Steinfort, C. D. Peacock, J. E. Cain, A. Szczepny and D. N. Watkins (2014). "Genomic characterisation of small cell lung cancer patient-derived xenografts generated from endobronchial ultrasound-guided transbronchial needle aspiration specimens." PLoS One **9**(9): e106862.

Li, H. and R. Durbin (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform." Bioinformatics **26**(5): 589-595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Lipson, D., Y. Aumann, A. Ben-Dor, N. Linial and Z. Yakhini (2006). "Efficient calculation of interval scores for DNA copy number data analysis." J Comput Biol **13**(2): 215-228.

Louie, B., P. Mork, F. Martin-Sanchez, A. Halevy and P. Tarczy-Hornoch (2007). "Data integration and genomic medicine." J Biomed Inform **40**(1): 5-16.

Marusyk, A., V. Almendro and K. Polyak (2012). "Intra-tumour heterogeneity: a looking glass for cancer?" Nat Rev Cancer **12**(5): 323-334.

Marusyk, A. and K. Polyak (2010). "Tumor heterogeneity: causes and consequences." Biochim Biophys Acta **1805**(1): 105-117.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.

Meacham, C. E. and S. J. Morrison (2013). "Tumour heterogeneity and cancer cell plasticity." Nature **501**(7467): 328-337.

Mermel, C. H., S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim and G. Getz (2011). "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers." Genome Biol **12**(4): R41.

Murugaesu, N., S. K. Chew and C. Swanton (2013). "Adapting clinical paradigms to the challenges of cancer clonal evolution." Am J Pathol **182**(6): 1962-1971.

Mutation, C. and C. Pathway Analysis working group of the International Cancer Genome (2015). "Pathway and network analysis of cancer genomes." Nat Methods **12**(7): 615-621.

Myers, R. H. (2004). "Huntington's disease genetics." NeuroRx **1**(2): 255-262.

Navin, N., J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks and M. Wigler (2011). "Tumour evolution inferred by single-cell sequencing." Nature **472**(7341): 90-94.

Nowsheen, S., A. C. Whitley and E. S. Yang (2012). "Biomarkers to assess the targeting of DNA repair pathways to augment tumor response to therapy." Curr Mol Med **12**(6): 788-803.

Parsons, D. W., S. Jones, X. Zhang, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, L. A. Diaz, Jr., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. Marie, S. M. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu and K. W. Kinzler (2008). "An integrated genomic analysis of human glioblastoma multiforme." Science **321**(5897): 1807-1812.

Pierron, G., F. Tirode, C. Lucchesi, S. Reynaud, S. Ballet, S. Cohen-Gogo, V. Perrin, J. M. Coindre and O. Delattre (2012). "A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion." Nat Genet **44**(4): 461-466.

Prados, M. D., S. A. Byron, N. L. Tran, J. J. Phillips, A. M. Molinaro, K. L. Ligon, P. Y. Wen, J. G. Kuhn, I. K. Mellinghoff, J. F. de Groot, H. Colman, T. F. Cloughesy, S. M. Chang, T. C. Ryken, W. D. Tembe, J. A. Kiefer, M. E. Berens, D. W. Craig, J. D. Carpten and J. M. Trent (2015). "Toward precision medicine in glioblastoma: the promise and the challenges." Neuro Oncol **17**(8): 1051-1063.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow and Y. Gu (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." BMC Genomics **13**: 341.

Roberts, N. D., R. D. Kortschak, W. T. Parker, A. W. Schreiber, S. Branford, H. S. Scott, G. Glonek and D. L. Adelson (2013). "A comparative analysis of algorithms for somatic SNV detection in cancer." Bioinformatics **29**(18): 2223-2230.

Salari, R., S. S. Saleh, D. Kashef-Haghighi, D. Khavari, D. E. Newburger, R. B. West, A. Sidow and S. Batzoglou (2013). "Inference of tumor phylogenies with improved somatic mutation discovery." J Comput Biol **20**(11): 933-944.

Salhia, B., J. Kiefer, J. T. Ross, R. Metapally, R. A. Martinez, K. N. Johnson, D. M. DiPerna, K. M. Paquette, S. Jung, S. Nasser, G. Wallstrom, W. Tembe, A. Baker, J. Carpten, J. Resau, T. Ryken, Z. Sibenaller, E. F. Petricoin, L. A. Liotta, R. K.

Ramanathan, M. E. Berens and N. L. Tran (2014). "Integrated genomic and epigenomic analysis of breast cancer brain metastasis." PLoS One **9**(1): e85448.

Saunders, C. T., W. S. Wong, S. Swamy, J. Becq, L. J. Murray and R. K. Cheetham (2012). "Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs." Bioinformatics **28**(14): 1811-1817.

Shalem, O., N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench and F. Zhang (2014). "Genome-scale CRISPR-Cas9 knockout screening in human cells." Science **343**(6166): 84-87.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-2504.

Singh, D., J. M. Chan, P. Zoppoli, F. Niola, R. Sullivan, A. Castano, E. M. Liu, J. Reichel, P. Porrati, S. Pellegatta, K. Qiu, Z. Gao, M. Ceccarelli, R. Riccardi, D. J. Brat, A. Guha, K. Aldape, J. G. Golfinos, D. Zagzag, T. Mikkelsen, G. Finocchiaro, A. Lasorella, R. Rabadan and A. Iavarone (2012). "Transforming fusions of FGFR and TACC genes in human glioblastoma." Science **337**(6099): 1231-1235.

Spencer, D. H., M. Tyagi, F. Vallania, A. J. Bredemeyer, J. D. Pfeifer, R. D. Mitra and E. J. Duncavage (2014). "Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data." J Mol Diagn **16**(1): 75-88.

Stein, L. (2001). "Genome annotation: from sequence to biology." Nat Rev Genet **2**(7): 493-503.

Stratton, M. R., P. J. Campbell and P. A. Futreal (2009). "The cancer genome." Nature **458**(7239): 719-724.

Stupp, R., M. E. Hegi, W. P. Mason, M. J. van den Bent, M. J. Taphoorn, R. C. Janzer, S. K. Ludwin, A. Allgeier, B. Fisher, K. Belanger, P. Hau, A. A. Brandes, J. Gijtenbeek, C. Marosi, C. J. Vecht, K. Mokhtari, P. Wesseling, S. Villa, E. Eisenhauer, T. Gorlia, M. Weller, D. Lacombe, J. G. Cairncross, R. O. Mirimanoff, R. European Organisation for, T. Treatment of Cancer Brain, G. Radiation Oncology and G. National Cancer Institute of Canada Clinical Trials (2009). "Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial." Lancet Oncol **10**(5): 459-466.

Tamborero, D., A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, C. Kandoth, J. Reimand, M. S. Lawrence, G. Getz, G. D. Bader, L. Ding and N. Lopez-Bigas (2013). "Comprehensive identification of mutational cancer driver genes across 12 tumor types." Sci Rep **3**: 2650.

Tang, K. W., B. Alaei-Mahabadi, T. Samuelsson, M. Lindh and E. Larsson (2013). "The landscape of viral expression and host gene fusion and adaptation in human cancer." Nat Commun **4**: 2513.

Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.

Tso, K. Y., S. D. Lee, K. W. Lo and K. Y. Yip (2014). "Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts?" BMC Genomics **15**: 1172.

Verhaak, R. G., K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes and N. Cancer Genome Atlas Research (2010). "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1." Cancer Cell **17**(1): 98-110.

Von Hoff, D. D., J. J. Stephenson, Jr., P. Rosen, D. M. Loesch, M. J. Borad, S. Anthony, G. Jameson, S. Brown, N. Cantafio, D. A. Richards, T. R. Fitch, E. Wasserman, C. Fernandez, S. Green, W. Sutherland, M. Bittner, A. Alarcon, D. Mallery and R. Penny (2010). "Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers." J Clin Oncol **28**(33): 4877-4883.

Wang, K., M. Li and H. Hakonarson (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Res **38**(16): e164.

Wang, Q., P. Jia, F. Li, H. Chen, H. Ji, D. Hucks, K. B. Dahlman, W. Pao and Z. Zhao (2013). "Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers." Genome Med **5**(10): 91.

Wang, Q., J. Xia, P. Jia, W. Pao and Z. Zhao (2013). "Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives." Brief Bioinform **14**(4): 506-519.

Weinhold, N., A. Jacobsen, N. Schultz, C. Sander and W. Lee (2014). "Genome-wide analysis of noncoding regulatory mutations in cancer." Nat Genet **46**(11): 1160-1165.

Weiss, G. J., W. S. Liang, M. J. Demeure, J. A. Kiefer, G. Hostetter, T. Izatt, S. Sinari, A. Christoforides, J. Aldrich, A. Kurdoglu, L. Phillips, H. Benson, R. Reiman, A. Baker, V. Marsh, D. D. Von Hoff, J. D. Carpten and D. W. Craig (2013). "A pilot study using next-generation sequencing in advanced cancers: feasibility and challenges." PLoS One **8**(10): e76438.

Wu, J., Y. Li and R. Jiang (2014). "Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies." PLoS Genet **10**(3): e1004237.

Xu, H., J. DiCarlo, R. V. Satya, Q. Peng and Y. Wang (2014). "Comparison of somatic mutation calling methods in amplicon and whole exome sequence data." BMC Genomics **15**: 244.

Yandell, M. and D. Ence (2012). "A beginner's guide to eukaryotic genome annotation." Nat Rev Genet **13**(5): 329-342.

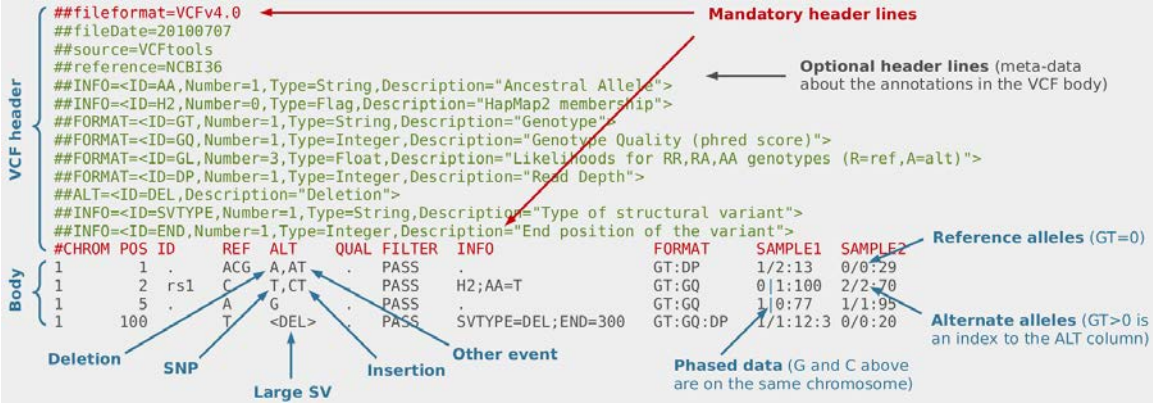
Yang, W., J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott and M. J. Garnett (2013). "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells." Nucleic Acids Res **41**(Database issue): D955-961.

Yap, T. A., M. Gerlinger, P. A. Futreal, L. Pusztai and C. Swanton (2012). "Intratumor heterogeneity: seeing the wood for the trees." Sci Transl Med **4**(127): 127ps110.

Zhang, J., J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C. W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, Wistuba, II and P. A. Futreal (2014). "Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing." Science **346**(6206): 256-259.

APPENDIX A

EXAMPLE OF VCF 4.2 FILE FORMAT



Cited from http://vcftools.sourceforge.net/VCF-poster.pdf

APPENDIX B

MANDTORY FIELDS OF FASTQ FILE AND SAM FILE

FASTQ file:

Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a-z, A-Z, 0-9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a-z, A-Z, 0-9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	X coordinate of cluster
<y_pos>	Numerical	Y coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered, N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. See below.
<index sequence>	ACTG	Index sequence

SAM file:

Column	Description
QNAME	Query template/pair NAME
FLAG	bitwise FLAG
RNAME	Reference sequence NAME
POS	1-based leftmost Position/coordinate of clipped sequence
MAPQ	Mapping Quality (Phred-scaled)
CIGAR	extended CIGAR string
MRNM	Mate Reference sequence Name ('=' if same as RNAME)
MPOS	1-based Mate Position
LEN	inferred Template Length (insert size)
SEQ	query Sequence on the same strand as the reference
QUAL	query Quality (ASCII-33 gives the Phred base quality)
OPT	variable Optional fields in the format TAG:VTYPE:VALUE

APPENDIX C

LIST OF HOUSE KEEPING GENES (EISENBERG AND LEVANON 2013)

11 highly uniform and strongly expressed genes:

Gene Name	RefSeq accession number	Gene description	Genomic coordinates (hg19) of strongly and uniformly expressed exons		
C1orf43	NM_015449	chromosome 1 open reading frame 43	chr1	154192817	154192883
			chr1	154186932	154187050
			chr1	154186368	154186422
			chr1	154184933	154185100
			chr1	154184795	154184854
CHMP2A	NM_014453	charged multivesicular body protein 2A	chr19	59065411	59065579
			chr19	59063625	59063805
			chr19	59063421	59063552
EMC7	NM_020154	ER membrane protein complex subunit 7	chr15	34382517	34382656
			chr15	34380253	34380334
			chr15	34376537	34376687
GPI	NM_000175	glucose-6-phosphate isomerase	chr19	34857687	34857756
			chr19	34859487	34859607
			chr19	34868639	34868786
			chr19	34869838	34869910
			chr19	34872370	34872424
			chr19	34884152	34884213
			chr19	34884818	34884971
			chr19	34887205	34887335
			chr19	34887485	34887562
			chr19	34890111	34890240
			chr19	34890460	34890536
			chr19	34890623	34890690
PSMB2	NM_002794	proteasome subunit, beta type, 2	chr1	36101910	36102033
			chr1	36096874	36096945
			chr1	36070833	36070883
PSMB4	NM_002796	proteasome subunit , beta type, 4	chr1	151372456	151372663
			chr1	151372917	151373064
			chr1	151373239	151373321
			chr1	151373714	151373831
RAB7A	NM_004637	member RAS oncogene family	chr3	128525214	128525433
			chr3	128526385	128526514
			chr3	128532169	128532262
REEP5	NM_005669	receptor accessory protein 5	chr5	112256859	112256953
			chr5	112238076	112238215
			chr5	112222711	112222880
SNRPD3	NM_004175	small nuclear ribonucleoprotein D3	chr22	24953642	24953768
			chr22	24963951	24964144
VCP	NM_007126	Valosin containing protein	chr9	35067887	35068060
			chr9	35066671	35066814
			chr9	35064150	35064282
			chr9	35062213	35062347

			chr9	35061999	35062135
			chr9	35061573	35061686
			chr9	35061011	35061176
			chr9	35060797	35060920
			chr9	35060309	35060522
			chr9	35059489	35059798
			chr9	35059060	35059216
			chr9	35057372	35057527
			chr9	35057116	35057219
VPS29	NM_016226	vacuolar protein sorting 29 homolog	chr12	110930800	110931036
			chr12	110929812	110929927

Full list of housekeeping genes:

AAAS	COMMD1	GSTM4	NAA15	RANGRF	TJAP1
AAGAB	COMMD10	GSTO1	NAA20	RAP1A	TLE1
AAMP	COMMD3	GTDC2	NAA38	RAPGEF1	TLK1
AAR2	COMMD3-BMI1	GTF2A1	NAA50	RAPGEF2	TM2D1
AARS	COMMD5	GTF2B	NAA60	RARS	TM2D2
AARS2	COMMD6	GTF2F1	NABP2	RARS2	TM2D3
AARSD1	COMMD7	GTF2F2	NACA	RB1CC1	TM9SF1
AASDHPPT	COMMD9	GTF2H1	NACA2	RBAK	TM9SF2
AATF	COMT	GTF2H4	NACC1	RBBP4	TM9SF3
ABCB10	COPA	GTF2H5	NACC2	RBBP7	TM9SF4
ABCB7	COPB1	GTF2I	NAE1	RBCK1	TMBIM1
ABCD3	COPB2	GTF3A	NAMPT	RBFA	TMBIM4
ABCE1	COPE	GTF3C1	NANS	RBM10	TMBIM6
ABCF1	COPG1	GTF3C2	NAP1L4	RBM12	TMCC1
ABCF2	COPS2	GTF3C3	NAPA	RBM12B	TMCO1
ABCF3	COPS3	GTF3C5	NARF	RBM14	TMCO3
ABHD10	COPS4	GTF3C6	NARFL	RBM14-RBM4	TMED1
ABHD12	COPS5	GTPBP10	NARG2	RBM15	TMED10
ABHD13	COPS6	GTPBP4	NARS	RBM15B	TMED2
ABHD14A	COPS7A	GTPBP5	NARS2	RBM17	TMED4
ABHD16A	COPS7B	GTPBP8	NAT10	RBM18	TMED5
ABHD4	COPS8	GUK1	NBN	RBM19	TMED7
ABHD8	COPZ1	GZF1	NBR1	RBM23	TMED7-TICAM2
ABI1	COQ10B	H1FX	NCAPH2	RBM27	TMED9
ABT1	COQ2	H2AFV	NCBP2	RBM28	TMEM101
ACAD9	COQ4	H2AFX	NCK1	RBM33	TMEM106B
ACADVL	COQ5	H2AFY	NCKIPSD	RBM34	TMEM106C
ACAP3	COQ6	H2AFZ	NCL	RBM39	TMEM115
ACBD3	CORO1C	HADH	NCLN	RBM4	TMEM120A

ACBD5	COX11	HADHA	NCOA1	RBM41	TMEM126A
ACBD6	COX14	HAGH	NCOA6	RBM42	TMEM127
ACIN1	COX15	HARS	NCOR1	RBM5	TMEM128
ACLY	COX16	HARS2	NCSTN	RBM6	TMEM129
ACOT13	COX19	HAT1	NDEL1	RBM7	TMEM131
ACOT8	COX20	HAUS3	NDFIP1	RBM8A	TMEM134
ACOT9	COX4I1	HAUS4	NDNL2	RBMX	TMEM141
ACOX1	COX5B	HAUS7	NDST1	RBMXL1	TMEM147
ACOX3	COX6B1	HAX1	NDUFA10	RBX1	TMEM14B
ACP1	COX6C	HBP1	NDUFA11	RC3H2	TMEM14C
ACSF3	COX7A2	HBS1L	NDUFA12	RCAN1	TMEM161A
ACSL3	COX7A2L	HCCS	NDUFA13	RCHY1	TMEM167B
ACSS2	COX7C	HCFC1	NDUFA2	RCN2	TMEM168
ACTR10	COX8A	HDAC2	NDUFA3	RDH14	TMEM177
ACTR1A	CPD	HDAC3	NDUFA4	RDX	TMEM179B
ACTR1B	CPNE1	HDAC6	NDUFA5	REEP3	TMEM18
ACTR5	CPNE2	HDAC8	NDUFA6	REEP5	TMEM184C
ACTR8	CPNE3	HDDC3	NDUFA7	RELA	TMEM185B
ACVR1	CPOX	HDGF	NDUFA8	REPIN1	TMEM186
ACVR1B	CPSF2	HDHD3	NDUFA9	REPS1	TMEM187
ADCK2	CPSF3L	HDLBP	NDUFAF2	RER1	TMEM189
ADCK4	CPSF4	HEATR2	NDUFAF3	REST	TMEM189-UBE2V1
ADH5	CPSF6	HEATR5A	NDUFAF4	REXO1	TMEM19
ADI1	CPSF7	HEBP1	NDUFB10	RFC1	TMEM192
ADIPOR1	CRADD	HECTD3	NDUFB11	RFC2	TMEM199
ADIPOR2	CRBN	HELZ	NDUFB2	RFC5	TMEM203
ADK	CRCP	HEMK1	NDUFB3	RFK	TMEM205
ADNP	CREB3	HERC4	NDUFB4	RFNG	TMEM214
ADO	CREBZF	HERPUD1	NDUFB5	RFT1	TMEM219
ADPRH	CREG1	HERPUD2	NDUFB6	RFWD2	TMEM222
ADPRHL2	CRELD1	HEXA	NDUFB7	RFXANK	TMEM223
ADPRM	CRIPAK	HEXDC	NDUFB8	RGP1	TMEM230
ADSL	CRIPT	HEXIM1	NDUFB9	RHBDD1	TMEM242
AES	CRK	HGS	NDUFC1	RHBDD3	TMEM248
AFF4	CRKL	HIAT1	NDUFC2	RHOA	TMEM251
AFTPH	CRLS1	HIATL1	NDUFC2-KCTD14	RHOB	TMEM256
AGFG1	CRNKL1	HIBADH	NDUFS2	RHOT1	TMEM258
AGGF1	CRTC2	HIGD1A	NDUFS3	RHOT2	TMEM259
AGPAT1	CRY2	HIGD2A	NDUFS4	RIC8A	TMEM30A
AGPAT3	CSGALNACT2	HINFP	NDUFS5	RIN2	TMEM33
AGPAT6	CSNK1A1	HINT1	NDUFS6	RING1	TMEM39A
AGPS	CSNK1A1L	HINT2	NDUFS7	RINT1	TMEM41A
AHCY	CSNK1D	HIST1H2BC	NDUFS8	RIOK1	TMEM41B
AHSA1	CSNK1G3	HIVEP1	NDUFV1	RIOK2	TMEM42
AIMP1	CSNK2A3	HMBS	NDUFV2	RIOK3	TMEM5

AIP	CSNK2B	HMG20A	NECAP1	RIPK1	TMEM50A
AK2	CSRP2BP	HMG20B	NEDD8	RMDN1	TMEM50B
AK3	CST3	HMGB1	NEDD8-MDP1	RMDN3	TMEM55B
AKAP8	CSTB	HMGN3	NEIL2	RMI1	TMEM57
AKAP9	CSTF1	HMGXB3	NEK4	RMND1	TMEM59
AKIP1	CSTF2T	HMGXB4	NEK9	RMND5A	TMEM60
AKIRIN1	CTAGE5	HMOX2	NELFB	RMND5B	TMEM62
AKIRIN2	CTBP1	HN1L	NELFCD	RNASEH1	TMEM63B
AKR1A1	CTCF	HNRNPA0	NELFE	RNASEH2C	TMEM64
AKR7A2	CTDSP2	HNRNPA2B1	NENF	RNASEK	TMEM66
AKT1	CTNNA1	HNRNPAB	NEU1	RNF10	TMEM69
AKT1S1	CTNNB1	HNRNPC	NF2	RNF103	TMEM70
AKTIP	CTNNBIP1	HNRNPD	NFATC2IP	RNF11	TMEM81
ALAD	CTNNBL1	HNRNPF	NFE2L2	RNF111	TMEM87A
ALDH3A2	CTNND1	HNRNPH1	NFIL3	RNF113A	TMEM9
ALDH9A1	CTSA	HNRNPH2	NFKBIB	RNF115	TMEM9B
ALG11	CTSD	HNRNPK	NFKBIL1	RNF121	TMF1
ALG5	CTTN	HNRNPL	NFU1	RNF126	TMLHE
ALG8	CTU2	HNRNPM	NFX1	RNF13	TMPO
ALG9	CUEDC2	HNRNPR	NFYB	RNF14	TMUB1
ALKBH1	CUL1	HNRNPU	NFYC	RNF141	TMUB2
ALKBH2	CUL2	HNRNPUL1	NGDN	RNF146	TMX1
ALKBH3	CUL4A	HNRNPUL2	NGLY1	RNF167	TMX2
ALKBH5	CUL4B	HNRPDL	NGRN	RNF181	TMX4
ALS2	CUL5	HNRPLL	NHP2	RNF185	TNFAIP1
ALYREF	CUTA	HP1BP3	NHP2L1	RNF187	TNFAIP8L2
AMBRA1	CUX1	HPS1	NIF3L1	RNF216	TNIP1
AMD1	CWC15	HPS6	NINJ1	RNF220	TNKS2
ANAPC10	CWC22	HS1BP3	NIP7	RNF25	TNPO1
ANAPC11	CWC25	HS2ST1	NIPA2	RNF26	TNPO3
ANAPC13	CXXC1	HS6ST1	NIPBL	RNF31	TNRC6A
ANAPC15	CXXC5	HSBP1	NISCH	RNF34	TOB1
ANAPC16	CXorf40A	HSCB	NIT1	RNF4	TOLLIP
ANAPC2	CXorf40B	HSD17B10	NIT2	RNF40	TOMM20
ANAPC5	CXorf56	HSD17B12	NKAP	RNF5	TOMM22
ANAPC7	CYB5B	HSD17B4	NKIRAS2	RNF6	TOMM40
ANKFY1	CYB5D2	HSPA14	NMD3	RNF7	TOMM5
ANKH	CYB5R3	HSPA4	NME1-NME2	RNH1	TOMM6
ANKHD1	CYC1	HSPA5	NME2	RNMTL1	TOMM7
ANKHD1-EIF4EBP3	CYFIP1	HSPA8	NME3	RNPEP	TOMM70A
ANKRD10	CYHR1	HSPA9	NME6	ROMO1	TOP1
ANKRD17	CYP2U1	HSPBP1	NMRK1	RP9	TOP2B
ANKRD28	D2HGDH	HSPE1-MOB4	NMT1	RPA2	TOPORS

ANKRD39	DAD1	HTATIP2	NOA1	RPA3	TOR1A
ANKRD46	DAG1	HTRA2	NOB1	RPAIN	TOR1AIP2
ANO6	DAGLB	HTT	NOC2L	RPAP3	TOR1B
ANP32A	DALRD3	HUS1	NOL10	RPF1	TOR3A
ANP32B	DAP3	HUWE1	NOL11	RPF2	TOX4
ANP32C	DARS	HYOU1	NOL12	RPL10A	TP53RK
ANP32E	DARS2	HYPK	NOL6	RPL11	TPCN1
ANXA6	DAXX	IAH1	NOL7	RPL14	TPD52L2
ANXA7	DAZAP1	IARS	NOL8	RPL26L1	TPGS1
AP1B1	DBT	IARS2	NOLC1	RPL27	TPI1
AP1G1	DCAF10	IBA57	NOM1	RPL30	TPP2
AP1M1	DCAF11	IBTK	NONO	RPL31	TPRA1
AP2A1	DCAF12	ICK	NOP10	RPL32	TPRG1L
AP2A2	DCAF13	ICMT	NOP14	RPL34	TPRKB
AP2M1	DCAF5	ICT1	NOP16	RPL35	TPRN
AP2S1	DCAF7	IDE	NOP2	RPL35A	TPST2
AP3B1	DCAF8	IDH3A	NOP56	RPL36AL	TRA2A
AP3D1	DCAKD	IDH3B	NOP58	RPL4	TRA2B
AP3M1	DCTD	IDH3G	NOP9	RPL6	TRAF6
AP3S1	DCTN2	IDI1	NPC1	RPL7L1	TRAF7
AP3S2	DCTN3	IER3IP1	NPC2	RPL8	TRAP1
AP4B1	DCTN4	IFNAR1	NPLOC4	RPN1	TRAPPC1
AP5M1	DCTN5	IFNGR1	NPRL2	RPN2	TRAPPC10
APEH	DCTN6	IFRD1	NPRL3	RPP14	TRAPPC11
APEX1	DCTPP1	IFT27	NQO2	RPP25L	TRAPPC12
APEX2	DCUN1D3	IKZF5	NR1H2	RPP30	TRAPPC13
APH1A	DCUN1D4	IL13RA1	NR2C1	RPP38	TRAPPC2L
API5	DCUN1D5	IL6ST	NR2C2AP	RPRD1B	TRAPPC3
APIP	DDA1	ILF2	NR3C2	RPS13	TRAPPC4
APOA1BP	DDB1	ILKAP	NRBP1	RPS19BP1	TRAPPC5
APOL2	DDB2	ILVBL	NRDE2	RPS23	TRAPPC6B
APOOL	DDOST	IMMT	NRIP1	RPS24	TRAPPC8
APOPT1	DDRGK1	IMP3	NSA2	RPS27L	TRAPPC9
APPL2	DDX1	IMP4	NSD1	RPS5	TRIAP1
APTX	DDX10	IMPAD1	NSDHL	RPS6	TRIM26
ARAF	DDX17	INF2	NSFL1C	RPS6KA3	TRIM27
ARCN1	DDX18	ING1	NSMCE1	RPS6KB1	TRIM28
ARF1	DDX19A	INO80B	NSMCE2	RPS6KB2	TRIM3
ARF5	DDX19B	INO80E	NSMCE4A	RPUSD3	TRIM39
ARF6	DDX21	INPP5A	NSRP1	RQCD1	TRIM39-RPP21
ARFGAP2	DDX23	INPP5K	NSUN2	RRAGA	TRIM41
ARFGAP3	DDX24	INSIG2	NSUN5	RRM1	TRIM44
ARFGEF2	DDX27	INTS1	NSUN6	RRN3	TRIM56
ARFIP1	DDX39B	INTS10	NT5C	RRNAD1	TRIM65
ARFIP2	DDX3X	INTS12	NT5C3	RRP1	TRIM8
ARFRP1	DDX41	INTS3	NT5DC1	RRP36	TRIP12
ARHGAP35	DDX42	INTS4	NTAN1	RRP7A	TRIP4

ARHGAP5	DDX46	INVS	NTMT1	RRP8	TRMT1
ARHGDIA	DDX47	IP6K1	NTPCR	RRS1	TRMT10C
ARHGEF10L	DDX49	IP6K2	NUB1	RSAD1	TRMT112
ARHGEF11	DDX54	IPO7	NUBP1	RSBN1L	TRMT12
ARHGEF40	DDX56	IPO8	NUBP2	RSC1A1	TRMT1L
ARIH1	DDX59	IPO9	NUCB1	RSL1D1	TRMT2A
ARIH2	DEDD	IRAK1	NUCKS1	RSPRY1	TRNAU1AP
ARIH2OS	DEF8	IREB2	NUDC	RSRC1	TRNT1
ARL1	DEGS1	IRF2BP1	NUDCD1	RSRC2	TRPC4AP
ARL14EP	DEK	IRF2BP2	NUDCD2	RTCA	TRPT1
ARL5A	DENND1A	IRF2BPL	NUDT14	RTFDC1	TRUB2
ARL6IP4	DENND4A	IRGQ	NUDT15	RTN4	TSC2
ARL8A	DENR	ISCU	NUDT2	RUFY1	TSEN15
ARL8B	DERA	ISOC2	NUDT21	RUVBL1	TSEN34
ARMC1	DERL1	IST1	NUDT22	RWDD1	TSFM
ARMC10	DERL2	ISY1	NUDT3	RWDD3	TSG101
ARMC5	DESI1	ISY1-RAB43	NUDT9	RXRA	TSN
ARMC6	DEXI	ITCH	NUFIP2	RXRB	TSNAX
ARMC7	DFFA	ITFG1	NUP107	SAE1	TSPAN17
ARMC8	DGCR14	ITFG3	NUP133	SAMD1	TSPAN31
ARMCX3	DGCR2	ITGB1	NUP153	SAMD4B	TSPYL1
ARMCX5	DGCR6L	ITGB1BP1	NUP54	SAMD8	TSR1
ARNT	DHPS	ITM2B	NUP62	SAMM50	TSR2
ARPC1A	DHRS12	ITPA	NUP85	SAP18	TSR3
ARPC2	DHRS7B	ITPK1	NUPL2	SAP30	TSSC4
ARPC5L	DHX15	ITPKC	NUTF2	SAP30BP	TSTA3
ARV1	DHX16	ITPRIPL2	NXF1	SAP30L	TSTD2
ASB1	DHX29	IVNS1ABP	NXT1	SAR1A	TTC1
ASB6	DHX30	IWS1	OAT	SARNP	TTC17
ASB7	DHX32	JAGN1	OAZ1	SARS	TTC19
ASB8	DHX33	JAK1	OAZ2	SART1	TTC32
ASCC1	DHX36	JKAMP	OBFC1	SART3	TTC33
ASCC3	DHX38	JMJD4	OCEL1	SAT2	TTC37
ASF1A	DHX8	JMJD6	OCIAD1	SAV1	TTC4
ASH2L	DHX9	JMJD7	ODC1	SBDS	TTC7B
ASNA1	DIABLO	JMJD8	OGFOD1	SCAF1	TTC9C
ASNSD1	DIDO1	JOSD2	OGFOD3	SCAF11	TTI1
ASPSCR1	DIEXF	JTB	OGFR	SCAF4	TTI2
ASUN	DIMT1	JUND	OGG1	SCAF8	TUBA1B
ASXL1	DIRC2	KANSL2	OGT	SCAMP2	TUBA1C
ATAD1	DIS3	KANSL3	OLA1	SCAMP3	TUBB
ATAD3A	DIS3L2	KARS	OPA1	SCAND1	TUBD1
ATE1	DKC1	KAT2B	OPA3	SCAP	TUBGCP2
ATF1	DLD	KAT5	ORC4	SCARB2	TUBGCP4
ATF2	DLG1	KAT8	ORMDL1	SCFD1	TUFM
ATF4	DLGAP4	KBTBD2	ORMDL2	SCFD2	TUSC2
ATF6	DLST	KBTBD4	ORMDL3	SCNM1	TUT1

ATF7	DMAP1	KBTBD7	OS9	SCO1	TVP23B
ATF7IP	DNAAF2	KCMF1	OSBP	SCO2	TXLNA
ATG12	DNAJA2	KCTD20	OSBPL2	SCOC	TXLNG
ATG13	DNAJA3	KCTD21	OSBPL9	SCP2	TXN2
ATG16L1	DNAJB11	KCTD6	OSGEP	SCRIB	TXNDC11
ATG2A	DNAJB12	KDM2A	OSGIN2	SCRN3	TXNDC12
ATG2B	DNAJB9	KDM4A	OSTM1	SCYL1	TXNDC15
ATG3	DNAJC10	KDM5C	OTUB1	SCYL2	TXNDC17
ATG4B	DNAJC11	KDSR	OTUD5	SCYL3	TXNDC9
ATG4D	DNAJC14	KHDRBS1	OVCA2	SDAD1	TXNL1
ATG5	DNAJC17	KHNYN	OXA1L	SDCBP	TXNL4A
ATG7	DNAJC19	KHSRP	OXNAD1	SDCCAG3	TXNL4B
ATIC	DNAJC2	KIAA0100	P4HTM	SDCCAG8	TXNRD1
ATL2	DNAJC21	KIAA0141	PA2G4	SDE2	TYK2
ATMIN	DNAJC3	KIAA0195	PABPN1	SDF2	TYW1
ATOX1	DNAJC4	KIAA0196	PACSIN2	SDF4	U2AF1
ATP2C1	DNAJC5	KIAA0232	PAF1	SDHA	U2AF1L4
ATP5A1	DNAJC7	KIAA0319L	PAFAH1B1	SDHAF2	U2AF2
ATP5B	DNAJC8	KIAA0391	PAGR1	SDHB	UAP1
ATP5C1	DNAJC9	KIAA0754	PAICS	SDHC	UBA1
ATP5D	DNASE2	KIAA0947	PAIP1	SDHD	UBA2
ATP5F1	DNLZ	KIAA1143	PAIP2	SDR39U1	UBA3
ATP5G2	DNM1L	KIAA1191	PAK1IP1	SEC11A	UBA5
ATP5G3	DNM2	KIAA1429	PAK2	SEC13	UBA52
ATP5H	DNTTIP1	KIAA1430	PAM16	SEC16A	UBAC2
ATP5J	DNTTIP2	KIAA1586	PANK2	SEC22B	UBALD1
ATP5J2	DOHH	KIAA1704	PANK3	SEC22C	UBAP1
ATP5J2-PTCD1	DOLK	KIAA1715	PANK4	SEC23A	UBAP2L
ATP5L	DPAGT1	KIAA1919	PANX1	SEC23IP	UBB
ATP5O	DPH1	KIAA1967	PAPD4	SEC24A	UBC
ATP5S	DPH2	KIAA2013	PAPD7	SEC24B	UBE2A
ATP5SL	DPH3	KLC4	PAPOLA	SEC24C	UBE2B
ATP6AP1	DPH5	KLF3	PARK7	SEC31A	UBE2D2
ATP6V0A2	DPM1	KLF9	PARL	SEC61A1	UBE2D3
ATP6V0B	DPP7	KLHDC2	PARN	SEC61B	UBE2D4
ATP6V0C	DPY30	KLHDC3	PARP1	SEC61G	UBE2E1
ATP6V0D1	DR1	KLHL20	PARP3	SEC62	UBE2E2
ATP6V0E1	DRAM2	KLHL25	PARP9	SEC63	UBE2E3
ATP6V1C1	DRAP1	KLHL36	PATL1	SECISBP2	UBE2F
ATP6V1D	DRG2	KLHL5	PATZ1	SEH1L	UBE2G2
ATP6V1E1	DROSHA	KLHL8	PAXBP1	SEL1L	UBE2H
ATP6V1F	DSCR3	KPNA1	PBDC1	SELK	UBE2I
ATP6V1G1	DTWD1	KPNB1	PBX2	SELO	UBE2J1
ATP6V1H	DUSP11	KRCC1	PCBP1	SELRC1	UBE2J2
ATPAF2	DUSP14	KRR1	PCBP2	SELT	UBE2K
ATPIF1	DUSP16	KTI12	PCDHGB5	SENP2	UBE2L3
ATRAID	DUSP22	KTN1	PCF11	SENP3	UBE2M

ATRN	DUT	KXD1	PCGF1	SENP5	UBE2N
ATXN10	DVL3	L3MBTL2	PCGF5	SENP6	UBE2NL
ATXN1L	DYM	LACTB	PCID2	SEPHS1	UBE2Q1
ATXN2	DYNC1LI1	LAGE3	PCIF1	SERBP1	UBE2R2
ATXN2L	DYNLL2	LAMP1	PCM1	SERF2	UBE2V1
ATXN7L3	DYNLRB1	LAMP2	PCMT1	SERGEF	UBE2V2
ATXN7L3B	DYNLT1	LAMTOR1	PCNA	SERINC1	UBE2W
AUH	E2F4	LAMTOR2	PCNX	SERINC3	UBE2Z
AUP1	E4F1	LAMTOR3	PCNXL4	SERPINB6	UBE3A
AURKAIP1	EAF1	LAMTOR4	PCSK7	SERTAD2	UBE3B
AXIN1	EAPP	LAMTOR5	PCYOX1	SET	UBE3C
AZI2	EARS2	LAP3	PCYT1A	SETD2	UBE4A
AZIN1	EBAG9	LAPTM4A	PDAP1	SETD3	UBE4B
B3GALT6	EBNA1BP2	LARP1	PDCD2	SETD5	UBFD1
B4GALT3	ECD	LARP4	PDCD5	SETD6	UBIAD1
B4GALT5	ECH1	LARP7	PDCD6	SETD7	UBL3
B4GALT7	ECHDC1	LARS2	PDCD6IP	SETD8	UBL4A
BABAM1	ECHS1	LCOR	PDE12	SETDB1	UBL5
BAD	ECI1	LDHA	PDE6D	SF1	UBL7
BAG1	ECI2	LEMD2	PDGFC	SF3A1	UBOX5
BAG4	ECSIT	LENG1	PDHB	SF3A3	UBP1
BAG6	EDC3	LEPROT	PDHX	SF3B1	UBQLN1
BAHD1	EDC4	LETM1	PDK2	SF3B14	UBQLN2
BANF1	EDEM3	LETMD1	PDLIM5	SF3B2	UBQLN4
BAP1	EDF1	LGALSL	PDP2	SF3B3	UBR2
BAZ1B	EED	LHPP	PDS5A	SF3B4	UBR7
BBS4	EEF1B2	LIAS	PDZD11	SF3B5	UBTD1
BCAP29	EEF1E1	LIG3	PDZD8	SFSWAP	UBTF
BCAP31	EEF2	LIG4	PEBP1	SGK196	UBXN2A
BCAS2	EEFSEC	LIN37	PEF1	SGMS1	UBXN4
BCAT2	EFCAB14	LIN54	PELO	SGPL1	UBXN6
BCCIP	EFHA1	LIN7C	PELP1	SGSM3	UCHL3
BCKDHA	EFR3A	LINS	PEPD	SGTA	UCHL5
BCKDK	EFTUD1	LIPT1	PES1	SH3BP5L	UCK1
BCL2L1	EFTUD2	LMAN1	PET100	SH3GLB1	UCK2
BCL2L13	EGLN2	LMBRD1	PET117	SHARPIN	UCKL1
BCL2L2-PABPN1	EHMT1	LMF2	PEX1	SHOC2	UEVLD
BCL7B	EI24	LMO4	PEX11A	SIAH1	UFC1
BCLAF1	EID2	LNX2	PEX11B	SIAH2	UFD1L
BCS1L	EIF1	LOC100129361	PEX12	SIGMAR1	UFL1
BECN1	EIF1AD	LOC100289561	PEX13	SIKE1	UFSP2
BFAR	EIF1B	LOC441155	PEX14	SIL1	UGP2
BIRC2	EIF2A	LOC729020	PEX16	SIRT2	UHRF1BP1L
BIVM-ERCC5	EIF2AK1	LONP1	PEX19	SIRT3	ULK1

BLMH	EIF2AK3	LONP2	PEX2	SIRT5	ULK3
BLOC1S1	EIF2AK4	LPCAT3	PEX26	SIRT6	UNC50
BLOC1S2	EIF2B2	LPIN1	PEX5	SIVA1	UNG
BLOC1S3	EIF2B3	LPPR2	PEX6	SKIL	UPF1
BLOC1S4	EIF2B4	LRFN3	PFDN2	SKIV2L	UPF2
BLOC1S6	EIF2B5	LRPAP1	PFDN4	SKIV2L2	UPF3B
BLZF1	EIF2D	LRPPRC	PFDN5	SKP1	UPRT
BMI1	EIF2S1	LRRC14	PFDN6	SLC15A4	UQCC
BMS1	EIF2S2	LRRC24	PFN1	SLC20A1	UQCR10
BNIP1	EIF3A	LRRC28	PGAM5	SLC25A11	UQCR11
BNIP2	EIF3B	LRRC40	PGBD3	SLC25A26	UQCRB
BOD1	EIF3D	LRRC41	PGK1	SLC25A28	UQCRC1
BOLA1	EIF3E	LRRC42	PGLS	SLC25A3	UQCRC2
BOLA3	EIF3G	LRRC47	PGP	SLC25A32	UQCRHL
BPGM	EIF3H	LRRC57	PGPEP1	SLC25A38	UQCRQ
BPNT1	EIF3I	LRRC59	PGRMC2	SLC25A39	URGCP
BPTF	EIF3J	LRRC8A	PHACTR4	SLC25A44	URI1
BRAT1	EIF3K	LRRFIP2	PHAX	SLC25A46	URM1
BRD2	EIF3L	LRSAM1	PHB	SLC25A5	UROD
BRD4	EIF3M	LSG1	PHB2	SLC27A4	UROS
BRD7	EIF4A1	LSM1	PHC2	SLC30A1	USB1
BRD9	EIF4A3	LSM10	PHF10	SLC30A5	USE1
BRE	EIF4E2	LSM14A	PHF12	SLC30A9	USF1
BRF1	EIF4G1	LSM14B	PHF20L1	SLC35A2	USF2
BRF2	EIF4G2	LSM2	PHF23	SLC35A4	USP10
BRIX1	EIF4G3	LSM3	PHF5A	SLC35B1	USP14
BRK1	EIF4H	LSM4	PHKB	SLC35B2	USP16
BRMS1	EIF5	LSM5	PHPT1	SLC35C2	USP19
BRPF1	EIF5A	LSM6	PHRF1	SLC35E1	USP22
BRPF3	EIF5AL1	LSM7	PI4K2A	SLC35E3	USP25
BSDC1	EIF5B	LSMD1	PI4KA	SLC35F5	USP27X
BSG	EIF6	LSS	PI4KB	SLC38A2	USP33
BTBD2	ELAC2	LTV1	PIAS1	SLC39A1	USP38
BTD	ELAVL1	LUC7L2	PICALM	SLC39A3	USP39
BTF3	ELF2	LUC7L3	PICK1	SLC39A7	USP4
BUB3	ELK1	LUZP6	PIGC	SLC41A3	USP47
BZW1	ELK4	LYRM1	PIGF	SLC46A3	USP5
C10orf12	ELL2	LYRM4	PIGG	SLC48A1	USP7
C10orf2	ELMOD3	LYRM5	PIGH	SLIRP	USP8
C10orf76	ELOVL1	LYSMD1	PIGK	SLMO2	USP9X
C10orf88	ELP2	LYSMD3	PIGP	SLTM	UTP11L
C11orf1	ELP3	LYSMD4	PIGS	SMAD2	UTP14A
C11orf24	ELP4	LZTR1	PIGT	SMAD4	UTP14C
C11orf31	ELP6	M6PR	PIGU	SMAD5	UTP15
C11orf57	EMC1	MAD2L1BP	PIGW	SMAP1	UTP23
C11orf58	EMC10	MAD2L2	PIGX	SMARCA2	UTP3
C11orf73	EMC2	MAEA	PIGY	SMARCA4	UTP6
C11orf83	EMC3	MAGED1	PIH1D1	SMARCAL1	UXS1

C12orf10	EMC4	MAGEF1	PIK3C3	SMARCB1	UXT
C12orf23	EMC6	MAGOH	PIK3CB	SMARCE1	VAC14
C12orf29	EMC7	MAGT1	PIK3R1	SMC1A	VAMP3
C12orf44	EMC8	MAK16	PIK3R4	SMC5	VAMP5
C12orf45	EMC9	MALSU1	PIN1	SMCR7L	VAPA
C12orf5	EMD	MAN1A2	PINK1	SMEK1	VAPB
C12orf52	EMG1	MAN1B1	PINX1	SMEK2	VARS2
C12orf57	ENDOG	MAN2A2	PIP5K1A	SMG5	VBP1
C12orf65	ENOPH1	MAN2B2	PITHD1	SMG7	VCP
C12orf66	ENSA	MAN2C1	PITPNA	SMG8	VDAC3
C14orf1	ENTPD4	MAP1LC3B2	PITPNB	SMIM11	VEZT
C14orf119	ENTPD6	MAP2K1	PITRM1	SMIM12	VIMP
C14orf142	ENY2	MAP2K2	PLA2G12A	SMIM8	VMA21
C14orf166	EPC1	MAP2K5	PLAA	SMNDC1	VPS16
C14orf2	EPM2AIP1	MAP3K7	PLBD2	SMPD1	VPS18
C14orf28	EPN1	MAP4K4	PLD3	SMPD4	VPS25
C15orf38-AP3S2	EPRS	MAPK1	PLEKHA1	SMU1	VPS26A
C15orf57	ERAL1	MAPK1IP1L	PLEKHJ1	SMUG1	VPS26B
C16orf13	ERAP1	MAPK6	PLEKHM1	SNAP23	VPS28
C16orf62	ERCC1	MAPK8	PLGRKT	SNAP29	VPS29
C16orf72	ERCC2	MAPK9	PLIN3	SNAP47	VPS33A
C16orf91	ERCC3	MAPKAP1	PLOD1	SNAPC3	VPS36
C17orf49	ERCC5	MAPKAPK2	PLOD3	SNAPC5	VPS37A
C17orf51	ERGIC2	MAPKAPK5	PLRG1	SNAPIN	VPS4A
C17orf58	ERGIC3	MAPRE2	PMF1	SND1	VPS51
C17orf59	ERH	MARCH2	PMF1-BGLAP	SNF8	VPS52
C17orf70	ERI3	MARCH5	PMPCA	SNRNP200	VPS53
C17orf85	ERICH1	MARCH6	PMPCB	SNRNP25	VPS72
C18orf21	ERLEC1	MARCH7	PMS1	SNRNP27	VRK2
C18orf25	ERO1L	MARK3	PMVK	SNRNP35	VRK3
C18orf32	ERP44	MARK4	PNISR	SNRNP40	VTA1
C18orf8	ESD	MARS	PNKD	SNRNP48	VTI1A
C19orf43	ESF1	MARS2	PNKP	SNRNP70	VTI1B
C19orf53	ETF1	MAT2B	PNN	SNRPA	WAC
C19orf60	ETFA	MAVS	PNO1	SNRPB	WAPAL
C19orf70	ETFB	MAX	PNPLA6	SNRPB2	WARS2
C1GALT1	ETV6	MAZ	PNPLA8	SNRPC	WBP11
C1QBP	EWSR1	MBD1	PNPO	SNRPD1	WBP1L
C1orf109	EXD2	MBD2	PNPT1	SNRPD2	WBP2
C1orf122	EXOC1	MBD3	PNRC2	SNRPD3	WBP4
C1orf123	EXOC2	MBD4	POFUT1	SNRPG	WBSCR22
C1orf174	EXOC3	MBLAC1	POLD2	SNUPN	WDR1
C1orf43	EXOC4	MBNL2	POLDIP2	SNW1	WDR12
C1orf50	EXOC7	MBTPS1	POLDIP3	SNX12	WDR13
C1orf52	EXOC8	MBTPS2	POLE3	SNX13	WDR18

C20orf111	EXOSC1	MCAT	POLE4	SNX17	WDR20
C20orf24	EXOSC10	MCCC1	POLG	SNX18	WDR24
C21orf2	EXOSC2	MCEE	POLH	SNX19	WDR25
C21orf33	EXOSC4	MCFD2	POLK	SNX2	WDR26
C21orf59	EXOSC7	MCM3AP	POLL	SNX25	WDR3
C22orf28	EXOSC8	MCM7	POLM	SNX3	WDR33
C22orf29	EXT2	MCMBP	POLR1C	SNX4	WDR36
C22orf32	EXTL3	MCOLN1	POLR1D	SNX5	WDR41
C2orf47	FADD	MCPH1	POLR1E	SNX6	WDR43
C2orf49	FAF1	MCRS1	POLR2A	SNX9	WDR44
C2orf69	FAF2	MCTS1	POLR2B	SOCS4	WDR45
C2orf74	FAHD1	MCU	POLR2C	SOCS6	WDR45B
C2orf76	FAM104B	MDC1	POLR2D	SOD1	WDR46
C3orf17	FAM108A1	MDP1	POLR2E	SON	WDR55
C3orf37	FAM108B1	ME2	POLR2F	SPAG7	WDR59
C3orf38	FAM114A2	MEAF6	POLR2G	SPAG9	WDR5B
C3orf58	FAM118B	MECP2	POLR2H	SPATA2	WDR6
C4orf27	FAM120A	MED10	POLR2I	SPATA5L1	WDR61
C4orf3	FAM120AOS	MED11	POLR2J	SPCS1	WDR70
C4orf52	FAM120B	MED13	POLR2K	SPCS3	WDR73
C5orf15	FAM122A	MED14	POLR2L	SPECC1L	WDR74
C5orf24	FAM127B	MED16	POLR3C	SPEN	WDR75
C6orf1	FAM134A	MED19	POLR3E	SPG11	WDR77
C6orf106	FAM134C	MED20	POLR3GL	SPG21	WDR81
C6orf120	FAM136A	MED21	POLR3K	SPG7	WDR83OS
C6orf136	FAM149B1	MED24	POM121	SPHAR	WDR85
C6orf226	FAM160A2	MED29	POM121C	SPNS1	WDR89
C6orf47	FAM160B1	MED31	POMGNT1	SPOP	WDTC1
C6orf57	FAM160B2	MED4	POMP	SPPL2B	WIBG
C6orf62	FAM162A	MED6	POMT1	SPPL3	WIPI2
C6orf89	FAM168B	MED7	POP4	SPRYD3	WIZ
C7orf25	FAM173A	MED8	POP5	SPRYD7	WRAP53
C7orf26	FAM173B	MEF2A	POP7	SPSB3	WRB
C7orf49	FAM174A	MEF2BNB	PPA1	SPTSSA	WRNIP1
C7orf50	FAM175B	MEMO1	PPA2	SPTY2D1	WSB2
C7orf55	FAM177A1	MEN1	PPAN	SRA1	WTAP
C7orf55-LUC7L2	FAM178A	MEPCE	PPAN-P2RY11	SRD5A3	WTH3DI
C7orf73	FAM192A	METAP1	PPARA	SREBF2	WWP1
C8orf33	FAM199X	METAP2	PPARD	SREK1IP1	WWP2
C8orf40	FAM200A	METRN	PPCS	SRM	XIAP
C8orf59	FAM204A	METTL13	PPFIA1	SRP14	XPA
C8orf76	FAM206A	METTL14	PPHLN1	SRP19	XPC
C8orf82	FAM208B	METTL16	PPID	SRP54	XPNPEP1
C9orf123	FAM20B	METTL17	PPIE	SRP68	XPO1
C9orf16	FAM210B	METTL18	PPIF	SRP72	XPO7
C9orf37	FAM32A	METTL20	PPIG	SRP9	XPOT

C9orf64	FAM35A	METTL21A	PPIH	SRPR	XRCC5
C9orf69	FAM3A	METTL23	PPIL4	SRPRB	XRCC6
C9orf78	FAM50A	METTL2A	PPM1A	SRR	XYLT2
C9orf89	FAM50B	METTL2B	PPM1B	SRRD	YAF2
CAB39	FAM58A	METTL3	PPP1CA	SRRM1	YARS
CALCOCO2	FAM63A	METTL5	PPP1CC	SRSF1	YARS2
CALM1	FAM73B	MFAP1	PPP1R10	SRSF10	YIF1A
CALR	FAM8A1	MFAP3	PPP1R11	SRSF11	YIF1B
CALU	FAM96A	MFF	PPP1R15B	SRSF2	YIPF1
CAMTA1	FAM96B	MFN1	PPP1R37	SRSF3	YIPF3
CAMTA2	FAM98A	MFSD11	PPP1R7	SRSF4	YIPF4
CANT1	FARS2	MFSD12	PPP1R8	SRSF7	YIPF5
CANX	FARSA	MFSD3	PPP2CA	SRSF8	YIPF6
CAPN1	FARSB	MFSD5	PPP2CB	SS18L2	YKT6
CAPN7	FASTK	MGAT2	PPP2R1A	SSB	YME1L1
CAPNS1	FASTKD2	MGAT4B	PPP2R2A	SSBP1	YPEL2
CAPRIN1	FASTKD5	MGME1	PPP2R2D	SSNA1	YRDC
CAPZA2	FBRSL1	MGMT	PPP2R3C	SSR1	YTHDC1
CAPZB	FBXL15	MGRN1	PPP2R4	SSR2	YTHDF1
CARKD	FBXL17	MGST3	PPP2R5A	SSR3	YTHDF2
CARS	FBXL3	MIA3	PPP2R5B	SSRP1	YTHDF3
CARS2	FBXL4	MIB1	PPP2R5C	SSSCA1	YWHAB
CASC3	FBXL5	MICALL1	PPP2R5D	SSU72	YWHAE
CASC4	FBXL6	MICU1	PPP2R5E	ST3GAL2	YY1
CASP3	FBXO11	MID1IP1	PPP4C	ST6GALNAC6	YY1AP1
CASP7	FBXO18	MIDN	PPP4R1	ST7	ZADH2
CASP9	FBXO22	MIEN1	PPP4R2	STAM	ZBED4
CBR4	FBXO28	MIER1	PPP5C	STAM2	ZBED6
CBX3	FBXO3	MIF	PPP6C	STAMBP	ZBTB1
CBX5	FBXO38	MIF4GD	PPP6R2	STARD3	ZBTB10
CC2D1A	FBXO42	MIIP	PPP6R3	STARD7	ZBTB11
CC2D1B	FBXO45	MINOS1	PPWD1	STAT3	ZBTB14
CCAR1	FBXO6	MIS12	PQBP1	STAU1	ZBTB17
CCBL1	FBXO7	MITD1	PQLC1	STAU2	ZBTB18
CCDC12	FBXW11	MKI67IP	PQLC2	STIM1	ZBTB21
CCDC124	FBXW2	MKKS	PRADC1	STIP1	ZBTB25
CCDC127	FBXW4	MKLN1	PRCC	STK11	ZBTB33
CCDC130	FBXW5	MKNK1	PRDM4	STK16	ZBTB39
CCDC137	FBXW7	MKRN2	PRDX1	STOM	ZBTB44
CCDC149	FCF1	MLEC	PRDX2	STOML1	ZBTB45
CCDC174	FDFT1	MLF2	PRDX3	STOML2	ZBTB5
CCDC22	FDPS	MLH1	PRDX5	STRAP	ZBTB6
CCDC23	FDX1	MLLT1	PRDX6	STRIP1	ZBTB7A
CCDC25	FECH	MLLT10	PREB	STRN3	ZBTB8OS
CCDC47	FEM1C	MLST8	PREP	STT3A	ZC3H10
CCDC50	FEN1	MLX	PRKAA1	STT3B	ZC3H11A
CCDC51	FEZ2	MMAA	PRKAB1	STUB1	ZC3H13

CCDC59	FGFR1OP2	MMADHC	PRKACA	STX10	ZC3H15
CCDC71	FH	MMS19	PRKAG1	STX17	ZC3H18
CCDC86	FIBP	MNAT1	PRKAR1A	STX4	ZC3H3
CCDC90A	FICD	MNF1	PRKRIP1	STX5	ZC3H7A
CCDC92	FIP1L1	MOB4	PRMT1	STX8	ZC3H7B
CCDC94	FIS1	MOGS	PRMT5	STXBP3	ZCCHC10
CCM2	FIZ1	MON1A	PRMT7	STYXL1	ZCCHC11
CCNB1IP1	FKBP3	MON2	PROSC	SUB1	ZCCHC3
CCNDBP1	FKBP8	MORC2	PRPF18	SUCLA2	ZCCHC7
CCNG1	FKBPL	MORF4L2	PRPF19	SUCLG1	ZCCHC9
CCNH	FKRP	MOSPD1	PRPF3	SUCLG2	ZCRB1
CCNK	FLAD1	MPC2	PRPF31	SUGP1	ZDHHC14
CCNL1	FLCN	MPDU1	PRPF4	SUGT1	ZDHHC16
CCNL2	FLOT1	MPG	PRPF40A	SUMO1	ZDHHC2
CCNY	FLOT2	MPHOSPH10	PRPF4B	SUMO3	ZDHHC3
CCPG1	FNDC3A	MPI	PRPF6	SUN2	ZDHHC4
CCT3	FNTA	MPLKIP	PRPF8	SUPT4H1	ZDHHC5
CCT4	FNTB	MPND	PRPS1	SUPT5H	ZDHHC8
CCT5	FOPNL	MPPE1	PRPSAP1	SUPT6H	ZFAND1
CCT6A	FOXK2	MPV17L2	PRR14	SUPT7L	ZFAND2B
CCT7	FOXP4	MRFAP1	PRRC1	SUPV3L1	ZFAND3
CCT8	FOXRED1	MRFAP1L1	PRRC2A	SURF1	ZFAND5
CD164	FPGS	MRI1	PRRC2B	SURF4	ZFAND6
CD320	FPGT	MRM1	PRUNE	SURF6	ZFP91
CD46	FRA10AC1	MRP63	PSEN1	SUV420H1	ZFPL1
CD63	FTO	MRPL1	PSEN2	SUZ12	ZFR
CD81	FTSJ1	MRPL10	PSENEN	SYAP1	ZFYVE1
CD82	FTSJ2	MRPL11	PSKH1	SYF2	ZFYVE19
CD99L2	FTSJ3	MRPL12	PSMA1	SYMPK	ZFYVE27
CDC123	FTSJD1	MRPL13	PSMA2	SYNCRIP	ZGPAT
CDC16	FTSJD2	MRPL14	PSMA3	SYNJ2BP	ZHX1
CDC23	FUBP1	MRPL15	PSMA4	SYNJ2BP-COX16	ZHX1-C8ORF76
CDC27	FUK	MRPL16	PSMA5	SYPL1	ZHX2
CDC37	FUNDC2	MRPL17	PSMA6	SYS1	ZHX3
CDC37L1	FXN	MRPL18	PSMA7	SYVN1	ZKSCAN1
CDC40	FYTTD1	MRPL19	PSMB1	SZRD1	ZMAT2
CDC42	FZR1	MRPL2	PSMB2	TAB1	ZMAT3
CDC5L	G3BP1	MRPL20	PSMB3	TAB2	ZMAT5
CDIP1	GAA	MRPL21	PSMB4	TACO1	ZMPSTE24
CDIPT	GABARAP	MRPL22	PSMB5	TADA1	ZMYM2
CDK12	GABARAPL2	MRPL23	PSMB6	TADA3	ZMYND11
CDK13	GABPB1	MRPL24	PSMB7	TAF10	ZNF121
CDK16	GADD45GIP1	MRPL27	PSMC2	TAF11	ZNF131
CDK2AP1	GALK2	MRPL28	PSMC3	TAF12	ZNF134
CDK4	GALNS	MRPL3	PSMC4	TAF13	ZNF138

CDK5RAP1	GALNT1	MRPL30	PSMC5	TAF15	ZNF142
CDK8	GALNT2	MRPL32	PSMC6	TAF1D	ZNF143
CDK9	GALT	MRPL33	PSMD1	TAF4	ZNF146
CDS2	GANAB	MRPL35	PSMD10	TAF5L	ZNF174
CDV3	GAPVD1	MRPL36	PSMD11	TAF8	ZNF181
CDYL	GARS	MRPL37	PSMD12	TAF9	ZNF189
CEBPG	GART	MRPL38	PSMD13	TALDO1	ZNF195
CEBPZ	GATAD2A	MRPL4	PSMD14	TAMM41	ZNF197
CECR5	GATAD2B	MRPL40	PSMD2	TANGO2	ZNF207
CELF1	GATC	MRPL41	PSMD3	TANGO6	ZNF22
CENPB	GBA	MRPL42	PSMD4	TANK	ZNF226
CENPT	GBA2	MRPL43	PSMD5	TAOK2	ZNF232
CEP104	GBF1	MRPL44	PSMD6	TAPBP	ZNF24
CEP57	GCC1	MRPL45	PSMD7	TAPT1	ZNF259
CEP63	GCDH	MRPL46	PSMD8	TARDBP	ZNF274
CERK	GCLC	MRPL47	PSMD9	TARS	ZNF277
CERS2	GCLM	MRPL48	PSME1	TATDN1	ZNF280D
CGGBP1	GDE1	MRPL49	PSME3	TATDN2	ZNF281
CHAMP1	GDI2	MRPL50	PSMF1	TAX1BP1	ZNF3
CHCHD1	GDPGP1	MRPL51	PSMG2	TAZ	ZNF32
CHCHD2	GEMIN7	MRPL52	PSMG3	TBC1D1	ZNF322
CHCHD3	GEMIN8	MRPL53	PSMG4	TBC1D14	ZNF326
CHCHD4	GET4	MRPL54	PSPC1	TBC1D15	ZNF330
CHCHD5	GFER	MRPL55	PTCD1	TBC1D20	ZNF335
CHCHD7	GFM1	MRPL9	PTCD3	TBC1D22A	ZNF33A
CHD1L	GFOD2	MRPS10	PTDSS1	TBC1D23	ZNF343
CHD4	GGCT	MRPS11	PTEN	TBC1D7	ZNF347
CHD8	GGNBP2	MRPS12	PTGES2	TBC1D9B	ZNF37A
CHERP	GGT7	MRPS14	PTGES3	TBCA	ZNF384
CHID1	GHDC	MRPS15	PTOV1	TBCB	ZNF394
CHKB	GHITM	MRPS16	PTP4A2	TBCC	ZNF397
CHMP1A	GID8	MRPS17	PTPMT1	TBCCD1	ZNF398
CHMP2A	GINM1	MRPS18A	PTPN1	TBCD	ZNF408
CHMP2B	GIPC1	MRPS18B	PTPN11	TBCE	ZNF41
CHMP4A	GLCE	MRPS18C	PTPN23	TBK1	ZNF410
CHMP4B	GLE1	MRPS2	PTRH1	TBRG1	ZNF414
CHMP5	GLG1	MRPS21	PTRH2	TBRG4	ZNF419
CHMP6	GLI4	MRPS22	PTRHD1	TCAIM	ZNF438
CHP1	GLO1	MRPS23	PUF60	TCEANC2	ZNF444
CHPT1	GLRX2	MRPS24	PUM1	TCEB1	ZNF446
CHRAC1	GLRX3	MRPS25	PUM2	TCEB2	ZNF48
CHST12	GLRX5	MRPS26	PURA	TCEB3	ZNF480
CHST7	GLT8D1	MRPS27	PURB	TCERG1	ZNF491
CHTOP	GLTP	MRPS28	PUS3	TCF12	ZNF506
CHUK	GLTPD1	MRPS30	PUS7	TCF20	ZNF507
CHURC1	GLYR1	MRPS31	PUSL1	TCF25	ZNF513
CHURC1-FNTB	GMPPA	MRPS33	PWP1	TCP1	ZNF518A

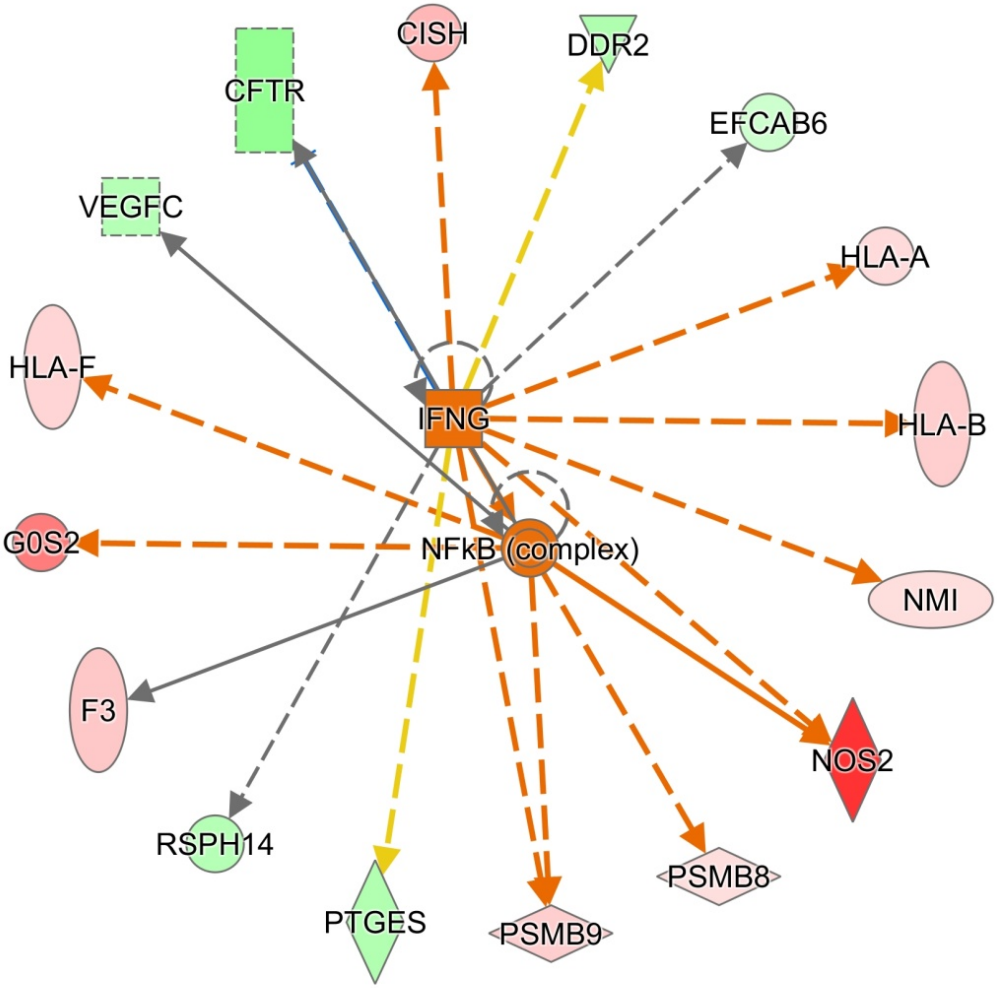
CIAO1	GMPR2	MRPS34	PWP2	TCTN3	ZNF526
CIB1	GNB1	MRPS35	PWWP2A	TDP2	ZNF561
CIC	GNB2	MRPS5	PXMP4	TDRD3	ZNF574
CINP	GNE	MRPS6	PYCR2	TECR	ZNF576
CIR1	GNL2	MRPS7	PYGO2	TEF	ZNF579
CIRH1A	GNL3	MRPS9	PYURF	TEFM	ZNF580
CISD1	GNPAT	MRRF	QARS	TELO2	ZNF592
CISD2	GNPDA1	MRS2	QRICH1	TERF2	ZNF593
CISD3	GNPNAT1	MRTO4	QRSL1	TERF2IP	ZNF598
CKAP4	GNPTG	MSANTD3	QSOX1	TEX2	ZNF620
CLCC1	GNS	MSH3	QTRT1	TEX261	ZNF622
CLCN3	GOLGA1	MSH6	R3HCC1	TEX264	ZNF623
CLCN7	GOLGA2	MSL3	R3HDM2	TFAM	ZNF638
CLINT1	GOLGA3	MSMP	RAB10	TFB1M	ZNF639
CLK3	GOLGA5	MSRA	RAB11A	TFB2M	ZNF641
CLNS1A	GOLGA7	MSRB2	RAB11B	TFCP2	ZNF644
CLOCK	GOLGB1	MTA2	RAB14	TFDP1	ZNF649
CLP1	GOLPH3	MTCH1	RAB18	TFE3	ZNF654
CLPP	GOLT1B	MTCH2	RAB1A	TFG	ZNF655
CLPTM1	GOPC	MTDH	RAB1B	TFIP11	ZNF664
CLPTM1L	GORASP1	MTERFD1	RAB21	TFPT	ZNF668
CLPX	GORASP2	MTERFD2	RAB22A	TGIF2-C20orf24	ZNF672
CLTA	GOSR1	MTERFD3	RAB2A	TGOLN2	ZNF687
CLTB	GOSR2	MTFMT	RAB2B	THADA	ZNF688
CLTC	GPAA1	MTFR1	RAB3GAP1	THAP3	ZNF691
CMAS	GPANK1	MTFR1L	RAB3GAP2	THAP4	ZNF7
CMC1	GPATCH4	MTIF3	RAB40C	THAP5	ZNF706
CMC2	GPBP1	MTM1	RAB4A	THAP7	ZNF721
CMC4	GPBP1L1	MTMR1	RAB5A	THOC5	ZNF740
CMPK1	GPHN	MTMR3	RAB5B	THOC7	ZNF76
CNBP	GPI	MTMR6	RAB5C	THOP1	ZNF764
CNIH	GPKOW	MTO1	RAB6A	THRAP3	ZNF770
CNIH4	GPN1	MTPAP	RAB7A	THTPA	ZNF777
CNNM2	GPN2	MTRR	RAB9A	THUMPD3	ZNF787
CNNM3	GPN3	MTSS1	RABEP1	THYN1	ZNF805
CNOT1	GPR107	MTX2	RABEPK	TIA1	ZNF814
CNOT11	GPR108	MUL1	RABGEF1	TIAL1	ZNF830
CNOT2	GPS1	MUS81	RABGGTA	TICAM1	ZNF865
CNOT3	GPS2	MUT	RABGGTB	TIGD5	ZNF91
CNOT4	GPX4	MVD	RAD1	TIGD6	ZNHIT1
CNOT7	GRAMD4	MXD4	RAD17	TIMM10	ZNHIT3
CNST	GRHPR	MXI1	RAD23B	TIMM10B	ZNRD1
COA1	GRINA	MYBBP1A	RAD50	TIMM13	ZRANB1
COA3	GRIPAP1	MYEOV2	RAD51C	TIMM17A	ZRANB2
COA4	GRPEL1	MYL12B	RAF1	TIMM17B	ZSCAN21
COA5	GRSF1	MYNN	RALA	TIMM21	ZSCAN29
COA6	GRWD1	MYO1E	RALBP1	TIMM22	ZSCAN32

COASY	GSK3A	MYPOP	RALY	TIMM44	ZSWIM1
COG1	GSK3B	MZF1	RAN	TIMM50	ZSWIM7
COG2	GSPT1	MZT2A	RANBP1	TIMM8B	ZSWIM8
COG3	GSPT2	MZT2B	RANBP2	TIMM9	ZW10
COG4	GSR	N4BP1	RANBP3	TIMMDC1	ZXDA
COG7	GSS	N4BP2L2	RANBP6	TINF2	ZXDB
COG8	GSTK1	NAA10	RANGAP1	TIPRL	ZZZ3

APPENDIX D

NFKB AND IFNG NETWORK

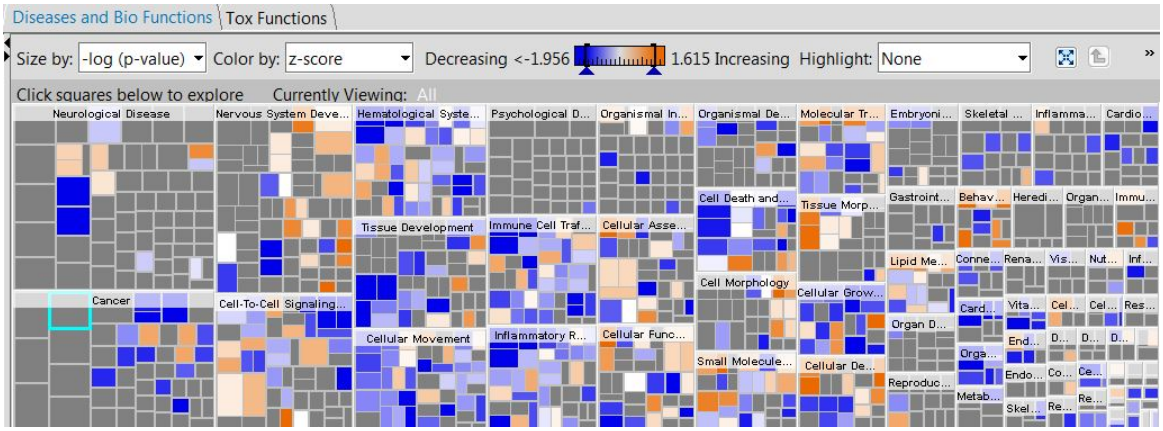
IFNG,NFkB (complex) 4



APPENDIX E

BIO FUNCTION ANALYSIS PANEL IN RELAPSE AND POST-RELAPSE TUMORS OF PATIENT ONE

Bio function analysis between primary and relapse tumor of patient one:



Bio function analysis between relapse and post-relapse tumor of patient one: