Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN
Scienze Biochimiche e Biotecnologiche

Ciclo: XXIX

Settore Concorsuale di Afferenza: 05/E1
Settore Scientifico Disciplinare: BIO/10

New approaches for the molecular profiling of human cancers through omics data analysis

Presentata da: Ítalo Faria do Valle

Coordinatore Dottorato
Prof. Santi Mario Spampinato

Relatore
Prof. Gastone Castellani
Co-relatore
Prof. Daniel Remondini

Esame finale anno 2017

"Só o que eu quis, todo o tempo, o que eu pelejei para achar, era uma só coisa - a inteira - cujo significado e vislumbrado dela eu vejo que eu sempre tive."

Grande Sertão: Veredas

João Guimarães Rosa

Acknowledgements

This PhD was one of the most amazing experiences in my life! I have learned so much, and I must thank all the people that were somehow involved on the process and that helped me in so many different ways throughout these 3 years. I would like to thank:

- the CAPES Foundation, Ministry of Education of Brazil, for giving me the opportunity and all the support I needed for this PhD;
- the professor Gastone Castellani, for believing in me from the very beginning (already in Brazil!), for having trust in me, for providing me the opportunities to learn and to participate in many different projects, and for listening and encouraging my ideas. I see him as a mentor that taught me the many different aspects necessary for becoming a successful scientist;
- the professor Daniel Remondini, for all the scientific discussions, for introducing me to the world of networks, and for the trust he put in me these three years. I am very thankful for having this opportunity of learning from him not only theory and techniques, but the very process of scientific thinking. He is an inspiration of the kind of scientist and professor I want to be one day;
- Enrico Giampieri and Giulia Menichetti, for all the help and teaching;
- Claudia Sala and Silvia Vitali, for all the friendship and support during these three years;
- the profressor Gerardo Manfreda (Department of Agricultural and Food Sciences, University of Bologna) and his group: Alessandra DeCesare, Frederique Pasquali, and Federica Palma; for all the opportunities of learning and participating in their studies;
- the professor Giovanni Martinelli (Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna) and his students Giorgia Simonetti, Antonella Padella, Anna Ferrari, and Samantha Bruno; for introducing me to the world of cancer research, and for helping us to experimentally validate our results;
- the professor Alessia Ciarrochi (Laboratory of Translational Research, Arcispedale S. Maria Nuova-IRCCS, Reggio Emilia), for giving me the opportunity to participate in the exciting research of her group and for inspiring me once again to the world of molecular biology and functional genomics;

- Finally, for my wife, Danielle, my family, and all my friends, without their support nothing would be possible.

Thank you!

Abstract

Cancer is a disease caused by alterations in the cell genomes. Different types of genomic alterations can cause uncontrolled cell proliferation, cell invasiveness, and resistance to therapy. The advances in high-throughput technologies has permitted a comprehensive characterization of cancer cell genomes, transcriptomes, and proteomes. Therefore, a big challenge in cancer research relies on new computational methods able to integrate large amount of "omics" data, reduce its complexity, and translate it in biological and clinical knowledge. In this thesis, we present three studies in which we applied *ad hoc* computational methods, combined with standard approaches, for the molecular profiling of human cancers using omics data.

The first study is entitled *"Detection of somatic single nucleotide polymorphisms in whole-exome sequencing data"*. Several studies have proposed software packages, filters and parametrizations for the detection of somatic polymorphisms in sequencing data, but many research groups have been reporting low concordance among different methods. Our main goal in this study was to develop a pipeline of analysis able to detect a wide range of single nucleotide mutations with high validation rates. We combined two standard tools - Genome Analysis Toolkit (GATK) and MuTect - to create the GATK-LOD_N method. As proof of principle, we applied our pipeline to exome sequencing data of hematological (Acute Myeloid and Acute Lymphoblastic Leukemias) and solid (Gastrointestinal Stromal Tumor and Lung Adenocarcinoma) tumors. We created simulated datasets and performed experimental validation (Sanger sequencing) to test the pipeline sensitivity and specificity.

The second study is entitled *"Network integration of multi-tumor omics data for the discovery of novel targeting strategies"*. We characterized the gene expression profiles of 11 tumor types, retrieved from The Cancer Genome Atlas data portal, aiming the discovery of multi-tumor drug targets and new strategies of drug combination and repurposing. First, we clustered tumors based on their transcriptomical correlation profiles. Then, we applied a network-based analysis, integrating gene expression profiles and protein interactions of cancer-related proteins. This allowed us to define three multi-tumor gene signatures, with genes belonging to the following biological categories: NF-κB signaling, chromosomal instability, ubiquitin-proteasome system, DNA metabolism, and apoptosis. We demonstrated the validity of our selection by evaluating the gene signatures based on mutational, pharmacological and clinical evidences. Moreover, we defined new pharmacological strategies validated by *in vitro* experiments that showed inhibition of cell growth in two tumor cell lines, with significant synergistic effect.

The third study is entitled *"Searching for the molecular mechanisms of tumor progression in thyroid cancer by gene expression data analysis"*. As the incidence of thyroid cancer continue to increase over the past years, we still need to understand the molecular mechanisms that cause the progression from less aggressive to highly

invasive and incurable forms of this tumor. We evaluated thyroid gene expression profiles of normal, Papillary Thyroid Carcinoma (PTC) and Anaplastic Thyroid Carcinoma (ATC) tissue samples (n=279). We observed that samples grouped in a progressional trend according to tissue type. The main biological processes affected in the normal to PTC transition were related to extracellular matrix and cell morphology; and those affected in the PTC to ATC transition were related to the control of cell cycle. We separated genes according to trends of up and down regulation, and then defined signatures related to each step of tumor progression. By mapping the gene signatures onto protein-protein interaction and transcriptomical regulatory networks, we could prioritize gene signatures for following experimental validation (ongoing experiments).

Contents

Acknowledgements **2**

Abstract **4**

1 The genomic basis of cancer **9**

Part I
Detection of somatic single nucleotide polymorphisms in whole-exome sequencing data **15**

2 Introduction **15**
2.1 Somatic Mutations in Cancer 15
2.2 Detecting Somatic Mutations in Sequencing Data 16
2.3 Low concordance of variant calling pipelines 18
2.4 Objectives 18

3 Material and Methods **19**
3.1 Sequencing Data 19
3.2 Pipeline for the discovery of Somatic Single Nucleotide Variants (sSNV) 19
3.2.1 Quality Control Check 19
3.2.2 Alignment and alignment post processing 20
3.2.3 Base Quality Score Recalibration 21
3.2.4 Variant detection 21
3.2.5 Variant Annotation 24
3.2.6 Availability of data and material 24
3.3 Pipeline Testing 24
3.3.1 Testing MuTect thresholds 24
3.3.2 Simulated datasets 24
3.3.3 Experimental validation 25

4 Results **26**
4.1 Comparison of methods 26
4.2 Pipeline Testing 28
4.2.1 Testing MuTect thresholds 28
4.2.2 Simulated datasets 29
4.2.3 Experimental Validation 31

5 Discussion **32**

Part II
Network integration of multi-tumor omics data for the discovery of novel targeting strategies 35

6 Introduction 35
6.1 Oncogenic alterations across human cancers . . . 35
6.2 Data integration in cancer genomics . . . 36
6.3 Objectives . . . 36

7 Material and Methods 38
7.1 Gene Expression Datasets . . . 38
7.2 Tumor clustering . . . 39
7.3 Multi-tumor gene signatures . . . 40
7.4 Validation of the multi-tumor gene signatures . . . 40
7.4.1 Gene signatures and mutational data . . . 40
7.4.2 Gene signatures and pharmacological data . . . 42
7.4.3 Gene signatures and prognosis . . . 42
7.4.4 Gene signatures and *in vitro* inhibition . . . 43

8 Results 44
8.1 Tumor clustering . . . 44
8.2 Multi-tumor gene signatures . . . 44
8.3 Validation of the multi-tumor gene signatures . . . 48

9 Discussion 54

Part III
Searching for the molecular mechanisms of tumor progression in thyroid cancer by gene expression data analysis 58

10 Introduction 58

11 Material and Methods 60
11.1 Data and Processing . . . 60
11.2 Differential Expression Analysis . . . 60
11.3 Signatures of Tumor Progression . . . 61
11.4 Analysis of gene signatures in biological networks . . . 61

12 Results 63
12.1 Clustering Gene Expression Profiles . . . 63
12.2 Differential Expression Analysis . . . 63
12.3 Trends . . . 65

13 Discussion 72

List of Publications **75**

List of Figures **75**

List of Tables **77**

Bibliography **79**

Appendices **90**
A Part I 90

Chapter 1

The genomic basis of cancer

Cancer comprises a group of diseases characterized by uncontrolled proliferation of cells that can invade normal tissues and metastasize to distant organs. It is a major cause of morbidity and mortality, representing 15% of the world deaths in 2012 [1]. The incidence of cancer has increased from 12.7 million in 2008 to 14.1 million in 2012, indicating that the number of new cases may rise in 75%, bringing the number of cancer patients close to 25 million over the next two decades [2].

The first insights about the connection between genome abnormalities and cancer development emerged in the late nineteenth and early twentieth centuries, when David von Hansemann [3] and Theodor Boveri [4] examined cancer cells under the microscope and observed the presence of bizarre chromosomal aberrations.

After the discovery of the DNA as the genetic material and the determination of its structure, increasingly refined analyses of cancer cell chromosomes demonstrated that specific and recurrent genomic abnormalities were associated with particular cancer types. In the beginning of the 1980s, it was demonstrated that normal cells could be transformed into cancer cells after receiving the genomic DNA from human cancers. This study permitted the first detection of an oncogenic mutation in a human gene: the point mutation G>T that causes a glycine to valine substitution in the HRAS gene [5, 6]. With this landmark finding, the genomic basis of cancer became firmly established.

Nowadays, as result of the advances in the molecular biology techniques, we have an extensive list of genomic alterations related to cancer development. Nucleotide substitutions, insertions and deletions (indels) of bases may modify proteins, causing their activation, as occurs in many oncogenes, or the loss of their function, typical of many tumor suppressor genes (Figure 1.1a). Chromosomal rearrangements damage normal genes or generate chimeric ones (gene fusions) that affect the development and maintenance of malignancy states (Figure 1.1b). Copy Number Variation (CNV) events cause gain and loss of gene copies through the duplication or deletion of chromosome segments (Figure 1.1c). Epigenetic processes, such as DNA methylation and histone modifications, can alter the chromatin structure and, consequently, the regulation of mechanisms as transcription, DNA repair, and DNA replication (Figure 1.1d).

The mechanisms causing the genomic alterations have both internal and external origins. For instance, environmental and life-style factors as tobacco-smoke and ultraviolet (UV) radiation exposure are associated with high mutation rates in lung and melanoma cancers, respectively (Reviewed in [8]). Other processes as aging,

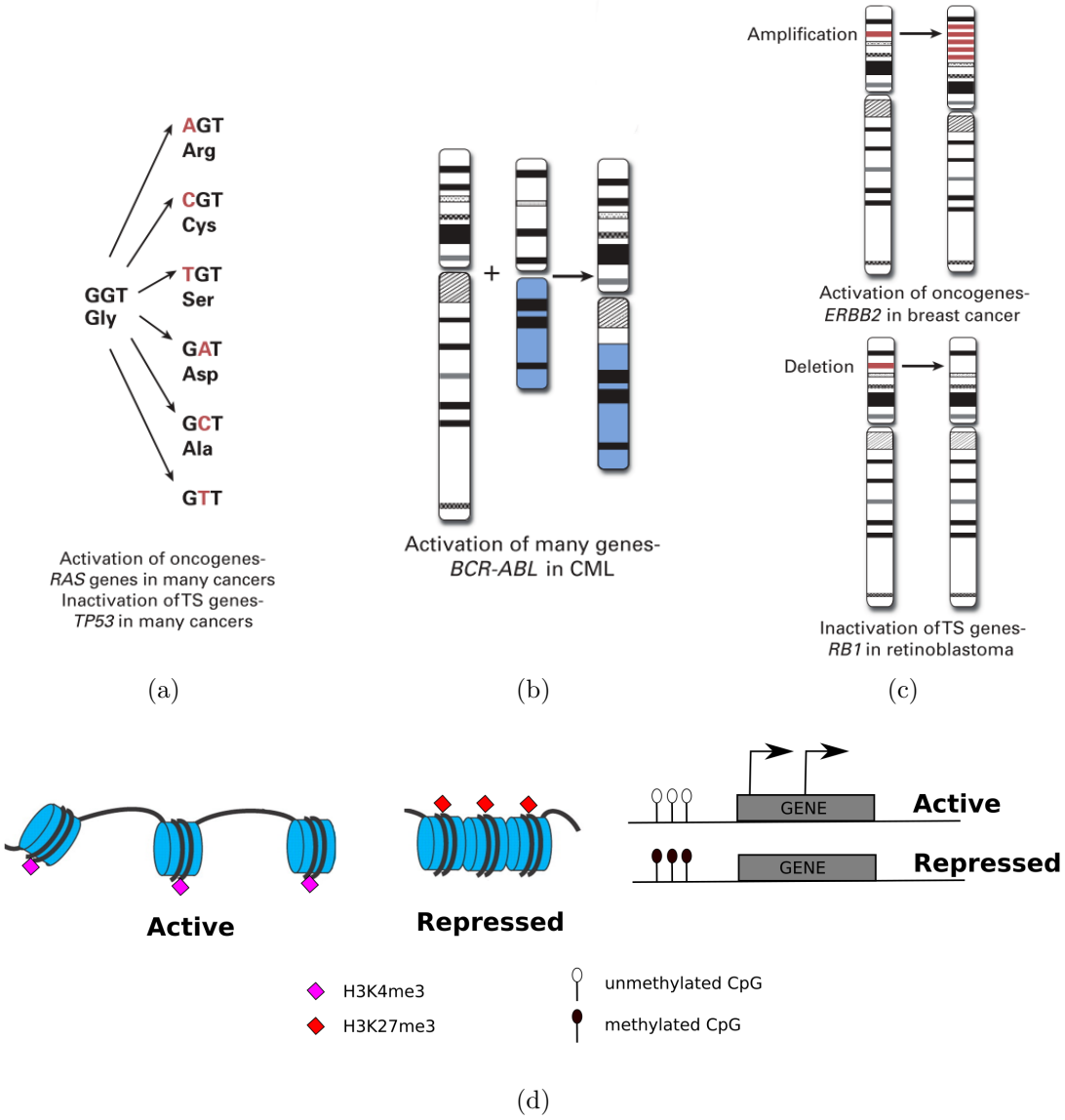


Figure 1.1: Examples of genomic alterations in cancer: (a) Nucleotide Substitutions, (b) Chromosomal Rearrangements, (c) Copy Number Variation, and (d) Epigenetic Modifications. TS: tumor suppressor; CML: Chronic Myeloid Leukemia; H3K4me3: histone modification associated with transcription activation; H3K27me3: histone modification associated with transcription repression. (Figures a, b, and c adapted from MacConaill& Garraway, 2010 [7]).

chronic inflammation, and carcinogens exposure may cause oncogenic epigenetic alterations (Reviewed in [9]). The malfunctioning of DNA damage-sensor and DNA repair mechanisms can increase mutation rates and cause chromosomal abnormalities, while erroneous chromosomal segregation during cell replication may result in large chromosomal rearrangements.

The proof of concept that genomics could bring new strategies for personalized therapy came in 2002, with the discovery of the BRAF V600E mutation in more than 50% of melanomas and the subsequent development of the inhibitor to treat patients having this mutation [10]. Since then, the research for cancer genes through the characterization of tumor genomic alterations has provided many insights for the

Table 1.1: Categories of Genomic Alteration and Exemplary Cancer Genes Exemplary

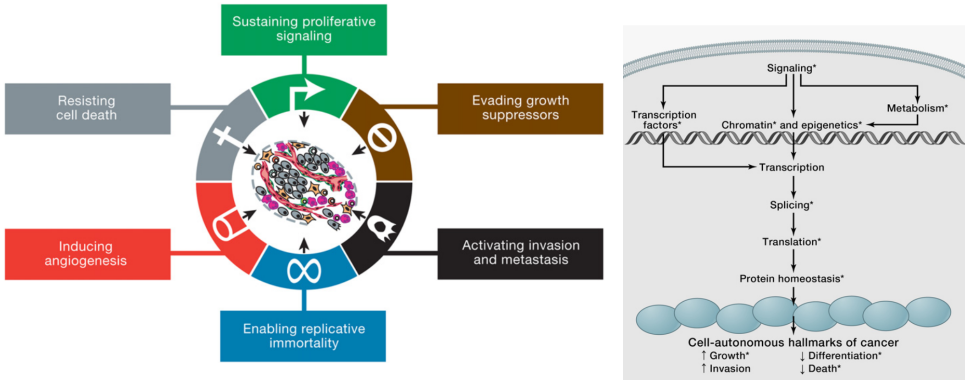
Genomic Alteration	Exemplary Cancer Gene	Type of Cancer	Targeted Therapeutic Agent
Translocation	BCR-ABL	Chronic myelogenous leukemia	Imatinib
	PML-RAR	Acute promyelocytic leukemia	All-trans-retinoic acid
	EML4-ALK	Breast, colorectal, lung	ALK inhibitor
	ETS gene fusions	Prostate	—
	Other	Leukemias, lymphomas, sarcomas	—
Amplification	EGFR	Lung, colorectal, glioblastoma, pancreatic	Cetuximab, gefitinib, erlotinib, panitumumab, lapatinib
	ERBB2	Breast, ovarian	Trastuzumab, lapatinib
	KIT, PDGFR	GISTs, glioma, HCC, RCC, CML	Imatinib, nilotinib, sunitinib, sorafenib
	MYC	Brain, colon, leukemia, lung	—
	SRC	Sarcoma, CML, ALL	Dasatinib
	PIK3CA	Breast, ovarian, colorectal, endometrial	PI3-kinase inhibitors
Point Mutation	EGFR	Lung, glioblastoma	Cetuximab, gefitinib, erlotinib, panitumumab, lapatinib
	KIT, PDGFR	GISTs, glioma, HCC, RCC, CML	Imatinib, nilotinib, sunitinib, sorafenib
	PIK3CA	Breast, ovarian, colorectal, endometrial	PI3-kinase inhibitors
	BRAF	Melanoma, pediatric astrocytoma	RAF inhibitor
	KRAS	Colorectal, pancreatic, GI tract, lung	Resistance to erlotinib, cetuximab (colorectal)

ALK: anaplastic lymphoma kinase; GIST: Gastrointestinal stromal tumor; HCC: hepatocellular carcinoma; RCC: renal cell carcinoma; CML: chronic myelogenous leukemia; ALL: acute lymphoblastic leukemia; PI3: phosphatidylinositol-3. (Table adapted from MacConaill & Garraway, 2010 [7])

understanding, classification and treatment of cancer types (Table 1.1). Usually, the so-called cancer genes encode proteins belonging to a wide range of biological categories: signal transduction pathways, metabolism, histone modification, nucleosome remodeling, DNA methylation, RNA splicing, protein homeostasis, and others.

However, we still need to connect the cancer genes into cancer processes in order to truly understand and treat cancer. Hanahan and Weiberg [11] have proposed "halmark" processes that generally become deregulated in tumorigenesis and metastasis: i) sustaining proliferative signaling, ii) evading growth supressors, iii) activating invasion and metastasis, iv) enabling replicative immortality, iv) inducing angiogenesis, and v) resisting cell death (Figure 1.2a). For instance, mutations in tyrosine kinase receptors and cell cycle inhibitors lead to effects that can be understood as "jamming the accelerator pedal" or "eliminating the breaks" on cell growth, respectively. However, these connections remain obscure for several cancer genes, since many of them affect multiple coregulated targets that act in several processes (Figure 1.2b) [12]. Therefore, identifying the full range of target genes for designing therapeutic strategies requires global genomic investigations at the DNA, RNA and protein levels.

Initial cancer genome projects had to be carried out with what today seem like primitive technologies, but the advances in microarray techonology, like comparative genomic hybridization and high-density single nucleotide polymorphism arrays, inaugurated a new phase for high-throughput and high resolution in cancer genomics research. The emergence of the Next-Generation Sequencing (NGS) technologies [13–15] revolutionized the research by lowering the costs and propelling an explosion of sequencing data. In parallel, methods were developed to capture specific portions of the genome as the 2% of genomic DNA containing known exons (the "exome"). The NGS technologies made possible the use of a single platform for



(a) (b)

Figure 1.2: (a) The hallmarks of cancer (Adapted from Hanahan & Weiberg, 2011 [11]). (b) Alterations in a range of cellular processes presumably contribute to cancer through their action on one or more target genes, mRNAs, or proteins, although the precise targets remain unknown in many cases (illustrated by shaded ovals) (Adapted from Garraway & Lander, 2013 [12]).

accessing all categories of genome alterations: point mutations, copy number variations, chromosomal rearrangements, gene expression measurement, identification of alternative splicing, detection of DNA methylation, chromatin structure mapping, and others. The application of NGS platforms to characterize and quantify biological molecules inaugurated the "omics" revolution, allowing scientists to explore proteins (proteomics), metabolites (metabolomics), RNA molecules (transcriptomics), epigenetic markers (epigenomics), and several molecules and processes.

Armed with this new technologies, large-scale efforts organized many groups across the world to collaboratively characterize tumor genomes and report their findings from both tissue site-specific and pan-cancer perspectives. The most prominent examples are: The Cancer Genome Atlas (TCGA), The International Cancer Genome Consortium (ICGC), and the Pediatric Cancer Genome Project (PCGP). The TCGA (`https://cancergenome.nih.gov/`) is jointly supported and managed by the National Cancer Institute and the National Human Genome Research Institute of the US National Institute of Health. The project has generated multidimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset (2.5 pentabytes of data describing tumor and matched normal tissues from more than 11,000 patients) has been used widely by the research community, contributing to more than a thousand of cancer studies by independent researchers. The ICGC (`http://icgc.org/`) was launched to coordinate large-scale cancer genome studies in tumors from 50 different cancer types and subtypes of clinical and societal importance. The PCGP (`http://explore.pediatriccancergenomeproject.org/`) was organized by the St. Jude Children's Research Hospital and The Genome Institute at the Washington University (US), with the stated goal of sequencing the whole genome of 600 tumors (and matched non-tumor germline samples) to define the landscape of somatic mutations underlying major subtypes of pediatric cancer.

The huge amount of genomic information generated by these projects brings a set of new challenges: handling, processing and analyzing these massive datasets. The detection of different genomic alterations requires specialized algorithms and statistical methods able to deal with false negatives produced by technical problems (sequencing errors, alignment artifacts) and biological factors as: admixture of non-

cancer cells (tumor purity), copy number variations inherent in cancer genomes (ploidy), and the presence of variant subclones within the cancer cell population (heterogeneity).

Since the non-functional genomic alterations outnumber the functional oncogenic events, a harder challenge is their classification according to the consequence in cancer development. The "driver" events grant growth advantage to cells and have been positively selected during cancer progression. The remaining genomic alterations are considered as "passengers": random events that have simply accumulated over the course of development and cell growth.

In face of all these challenges, the advance in cancer research requires better methods for integrate large amounts of molecular data, reduce its complexity and translate it in biological and clinical meaning. We attempted these goals in the three studies presented in this thesis: first, by developing a method for better detection of somatic mutations in cancer sequencing data; secondly, by defining multi-tumor targets and new drug repurposing strategies through the study of genes expression profiles from 11 tumor types; and finally, by evaluating gene expression profiles to investigate the molecular mechanisms involved in thyroid cancer progression.

Part I
Detection of somatic single nucleotide polymorphisms in whole-exome sequencing data

Chapter 2

Introduction

2.1 Somatic Mutations in Cancer

The term somatic mutation refers to changes in the DNA that occurs in a developing somatic tissue, while the germline mutations are those inherited from the parents and transfered to the offspring. These mutations include single nucleotide variants (SNVs), insertions and deletions (indels) of bases, DNA rearrangements and copy number alterations. Our current understanding of cancer genetics is grounded on the principle that cancer arises from a clone that has accumulated a set of somatic aberrations that leads to malignant transformation (Figure 2.1). Consequently, somatic mutations play a crucial role in cancer development, progression and chemotherapy resistance.

The somatic landscape of cancer genomes has been characterized by the sequencing of all human protein-coding exons (WES, Whole Exome Sequencing), which detects approximately 20,000 SNVs, or by the sequencing of the entire genome (WGS, Whole Genome Sequencing), which detects 3-4 million SNVs. The analysis of several samples from different tumors revealed large variability in the frequency of mutational profiles across cancer types (Figure 2.2) [17]. For example, pediatric and hematological tumors have the lowest mutation rates while lung cancer and melanoma present the highest rates. The mutational profiles are related with signatures of carcinogenesis mechanisms, such as the high proportion of G →C transver-

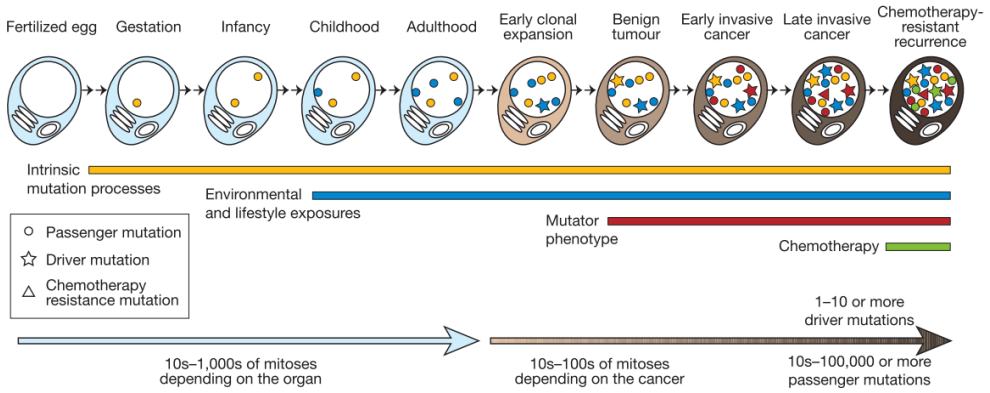


Figure 2.1: The lineage of mitotic cell divisions from the fertilized egg to a single cell within a cancer. The figure shows the acquisition of somatic mutations and the processes that contribute to them (Adapted from Stratton, Campbell & Futreal, 2009 [16]).

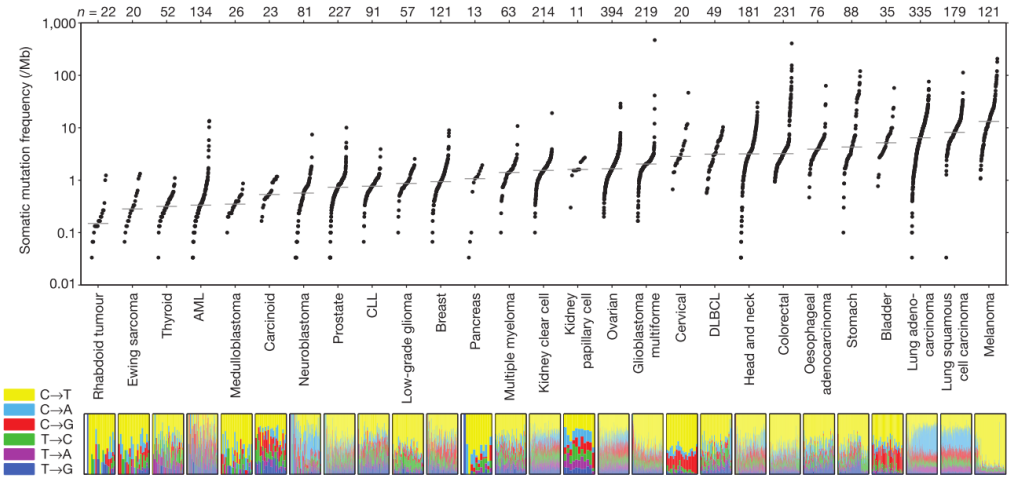


Figure 2.2: Somatic mutational frequency and mutation profiles observed in 3,083 tumor–normal paired exomes. Each dot represents a tumor–normal pair and their vertical position indicates the their frequency of somatic mutations. The bottom panel shows the relative proportions of the six possible base-pair substitutions (Adapted from Lawerence et al, 2013 [17]).

sions in lung cancers, attributed to tobacco smoke exposure; and the high proportion of C $\rightarrow$T transitions in melanomas, caused by ultraviolet radiation-induced DNA damage and repair .

However, the detection of somatic mutations in normal-cancer paired samples presents several challenges. The first one relies on the fact that we need to distinguish between germline and somatic mutations, and, at the same time, deal with many sources of biological variation: tumor heterogeneity, subclonality, copy number variation events, and tumor samples contaminated with normal cells or vice-versa. The other problems arise from technical problems as: sequencing artifacts, mapping errors, and differential sequencing depth between normal and tumor samples.

2.2 Detecting Somatic Mutations in Sequencing Data

The most important steps in the detection of SNVs in sequencing data are: quality control, alignment, alignment post-processing, quality score recalibration, variant and genotype calling, and variant filtering.

Quality Control The sequencing data may present base calling errors, poor quality reads, adapter contamination, and a wide range of chemistry and instrument failures [18]. For that reason, the first step must be the data quality control by removing, trimming or correcting reads that do not meet a defined standard.

Alignment, Alignment Post-Processing and Base Recalibration The next step is the alignment of reads to a reference genome. The most widely used references are provided by the University of Santa Cruz (versions hg18, hg19 and hg38) and by the Genome Reference Consortium (versions GRCh36-38). The choice of alignment tool and the corresponding settings significantly affect the analysis outcome

as wrongly aligned reads may result in artificial deviations from the reference. The alignment tools are based on either hash table indexing procedures or data compression algorithms, as the "Burrows-Wheller transform" (BWT). BWT-based aligners are fast, memory-efficient and particularly useful for aligning repetitive reads; however, they tend to be less sensitive than the state-of-the-art hash-based algorithms [19].

The first step of the alignment post-processing is the sort of reads according to their chromosomal position. Then, a common practice is to remove (or mark) the PCR duplicates, which occur when two copies of the same original DNA fragment is sequenced. They can be detected by selecting the sequencing reads that start and end at exactly the same position. Finally, one can realign reads around small indels, since differences in resolving the indels may cause artificial variant positions in the downstream analysis.

Previous works demonstrated that the per-base quality scores issued by the sequencing platforms may often deviate from the true error rate [20]. Since the variant and genotype calling depend on the these scores, it is very important to obtain well-calibrated quality scores, which is provided by softwares as GATK and SOAPsnp.

Variant and Genotype Calling The term variant calling refers to determining, in the sequencing data, positions that differ from a reference sequence; and genotype calling refers to the process of determining the the individual's genotype for the positions having a variant. The early variant calling approaches relied on counting the abundance of high-quality nucleotides at a single site, but probabilist frameworks has been developed to provide ways of quantifying uncertainty about the variant calls. In brief, it is assumed that one can compute a genotype likelihood $p(E|G)$ for a genotype G using the E evidence, i.e, all the read data for a particular individual at a particular site. By using a genotype prior $p(G)$, we can calculate the posterior probability $p(G|E)$ of a genotype G with the Bayes' formula:

$$p(G|E) = \frac{p(E|G)p(G)}{p(E)} \tag{2.1}$$

We can compute the most plausible genotype $\hat{G}$ by:

$$\hat{G} = \text{argmax}_G\{p(E|G) * p(G)\} \tag{2.2}$$

The prior for the evidence $p(E)$ remains constant in the maximization and can therefore be omitted. Thus, it is sufficient to find the genotype that maximizes the posterior probability. In this case, the evidence E simply consists of the bases quality values in each read i. Thus, $p(E|G)$ can be calculated directly from the data by taking the product of $p(E_i|G)$ over all i. More precisely, it will be expressed as:

$$p(E|G) = \prod_i p(E_i|G) \tag{2.3}$$

Variant Filtering Filtering is an essential step in reducing the number of false-positive SNVs. Typically, the filtering approaches check for deviations from the Hardy-Weinberg equilibrium, low-quality scores, systematic differences in quality

scores for major and minor alleles, extreme read depths, strand bias (when a disproportional number of plus and minus strands are observed), spanning deletions, adjacent indels, within-read position, and presence of particular surrounding sequence motifs. [19, 21].

2.3 Low concordance of variant calling pipelines

Since the variant detection methods present different error-modeling approaches and prior assumptions, several studies demonstrated low concordance between variant sets called by different pipelines, softwares and parametrizations [21–25]. Therefore, the current challenge in clinical genomics relies on providing detection approaches that present the lowest level of false positives, in order to ensure the correct clinical interpretation and chemotherapy selection, while, at the same time, reduce the level of false negatives, since even a single variant missed can mean the difference between discovering a disease-contributing mutation or not.

The false positives originate from factors as: i) batch effect and sample preparation [26], ii) read depth [23], iii) sequencing and alignment artifacts [27], iv) and variant calling biases that algorithms may have towards specific types of SNVs [25]. The available tools either detect too many false positives in order to get all true positives or lose too many true positives in order to reduce the number of false positives. In the first case, the researcher spends much time and resource validating the set of candidate variants to select the true ones. In the second case, important mutations that explain the biological characteristics of the cancer cells, may be missed. One could reduce the number of false positives by increasing stringency filters or intersecting different variant calls, but it usually results in a consistently increase of false negative rates [28, 29]. As different tools and filtering approaches usually present variability in performance according to studies and tumor types, the research community faces a big challenge choosing the right pipeline among all available options.

2.4 Objectives

In this study, we aimed to develop a pipeline that detects single nucleotide variants in sequencing data of cancer samples with high specificity and sensitivity rates. To accomplish that, we combined the benefits of using two standard tools: Genome Analysis Toolkit (GATK) and MuTect. GATK independently calls variants in the normal and tumor samples, while MuTect performs the analysis simultaneously. In order to ensure the somatic classification of the GATK results and reduce its false positive calls, we created the GATK-LOD_N method. Briefly, it is part of the MuTect algorithm that is applied downstream to the GATK analysis in order to ensure the correct somatic classification and reduce its false positive calls.

We applied our pipeline to hematological (Acute Myeloid and Acute Lymphoblastic Leukemias) and solid (Gastrointestinal Stromal Tumor and Lung Adenocarcinoma) tumors. We also created artificial tumor samples to test the sensitivity and specificity of our pipeline. Our results show that the pipeline performed well and we believe that it can be helpful in discovery studies aimed to profile the somatic mutational landscape of cancer genomes.

Chapter 3

Material and Methods

3.1 Sequencing Data

The whole exome sequencing data from Acute Myeloid Leukemia (n=37) and Acute Lymphoblastic Leukemia patients (n=41) were kindly provided by the professor Giovanni Martinelli, from the *Dipartimento di Medicina Specialistica, Diagnostica e Sperimentale* of the Università di Bologna. The targeted region comprised 62 Mb of 201,121 exonic regions sequenced by the Illumina HiSeq2000 platform, which produced a per sample average of 55.2 and 63 million 100 bp paired-end reads for the AML and ALL cohorts, respectively. The AML and ALL data sets are available upon request to the Next Generation Sequencing for Targeted Personalized Therapy of Leukemia consortium. We also selected two public data sets of Illumina HiSeq 2000 whole exome sequencing from NCBI Sequence Read Archive: 1) seven Gastrointestinal Stromal Tumors (GIST) samples, and their matching peripheral blood samples, with an average of 35.5 million 100 bp paired-end reads per sample [SRA: SRR1299130-141 and SRR1299144-147] [30] [U+2060]; and 2) two Lung Adenocarcinoma samples, and their normal counterparts, with an average of 56.5 million 100 pb paired-end reads per sample [SRA: ERR160124, ERR160136, ERR166338, and ERR166339] [31] [U+2060]. After the quality control check, the average of final coverages in the tumor cohorts were: 72X (± 29X) for AML, 119X (± 28X) for ALL, 76X (± 7X) for GIST, and 133X (± 64X) for Lung Adenocarcinomma (See appendix tables A1, A2 ,A3, A4).

3.2 Pipeline for the discovery of Somatic Single Nucleotide Variants (sSNV)

The Figure 3.1 summarizes steps of the pipeline that we created for the discovery of somatic single nucleotide variants in paired normal-cancer sequencing data.

3.2.1 Quality Control Check

Initially, the sequencing reads were submitted to a quality control check. It was based on the per-base error estimation emitted by the sequencing machines, which provides the probability (P) of a DNA base calling error. A common approach to present this probability is by the Phred quality score (Q):

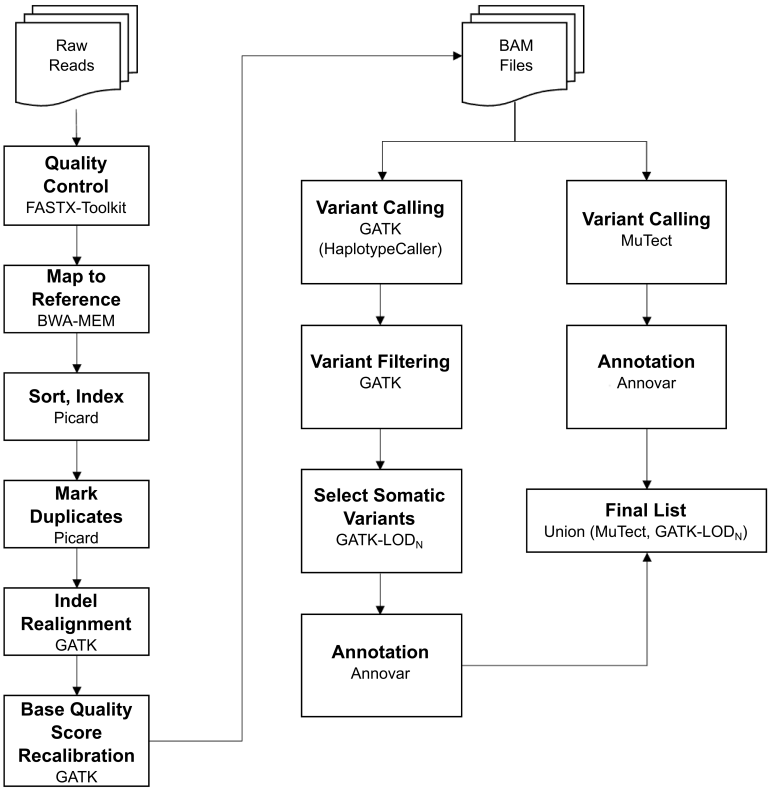


Figure 3.1: The steps and tools of the pipeline for the discovery of sSNVs in paired normal-cancer sequencing data.

$$Q = -10 \log_{10} P \tag{3.1}$$

We determined $Q \geq 20$ as our threshold for base quality, in order to have an accuracy of 99%. We applied the fastq_quality_filter.pl script to remove reads having more than 80% of low quality bases. Then, we applied the fastq_quality_trimmer.pl script for trimming the remaining reads with low quality bases in their 3' extremities. Both scripts are from the FASTX-Toolkit (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) .

3.2.2 Alignment and alignment post processing

After the quality control check we aligned the reads to the human reference genome version hg19/GRCh37 using BWA-MEM from the Burrows-Wheeler Alignment tool (BWA) package [32]. The BWA is based on backward search with Burrows-Wheeler Transform (BWT) to efficiently align short sequencing reads against a large reference sequence. The alignment was sorted according to the chromosome coordinates and indexed to permit an efficiently access to the reads in the file. Both sorting and indexing steps were performed by applying SortSam and BuildBamIndex tools, from Picard (`https://broadinstitute.github.io/picard/`).

In the next step, we marked the duplicated reads. Read duplication originates from the DNA sample preparation. In the first step of the sample preparation the DNA is shattered in random fragments and adapters are ligated to the extremities of all fragments. Then, all fragments are amplified by PCR (Polymerase chain reaction) to create multiple copies of each original DNA molecule. The DNA is

diluted in the flow cells (glass slides where the sequencing chemistry occurs) in order to hybridize with short oligonucleotides complementary to the adapter sequences. The duplication occurs when two copies of the same original molecule hybridize with the oligonucleotides in a flow cell. These duplicates are identified in the sequencing data by selecting reads with identical start and end positions. We marked the duplicated reads with MarkDuplicates, from Picard.

The alignment of indels (insertion/deletions) often produce mapping artifacts. They result from many mismatching bases near to the indel event, which can be miscalled as true variants. Indel errors also originate from platform-specific artifacts. For example, sequencing data from Hi-Seq and Mi-Seq (Illumina) usually present increased indel error rate around inverted repeats sequences and after long homopolymer stretches in GC-rich regions [33–35]. We applied the IndelRealigner from the Genome Analysis Toolkit (GATK) package (version 3.0) [36] to reduce the number of false positives originating from indel artifacts. First, its algorithm identify reads spanning an indel; then it performs the Smith-Waterman algorithm to provide a consistent alignment that minimizes mismatching bases across all reads.

3.2.3 Base Quality Score Recalibration

The variant calling algorithms rely heavily on the quality scores assigned to the individual bases in each sequenced read. However, due to various sources of systematic technical errors, the sequencing machines may over- or underestimate the base qualities [20]. The GATK package provides a base quality score recalibration (BQRS) tool (BaseRecalibrator) that applies machine learning methods to model the base errors empirically and adjust the quality scores accordingly. This process is accomplished by analyzing the covariation among several features of each base as: reported quality scores, the position within the read, and dinucleotide context. We applied the base quality score recalibration to improve the accuracy for subsequent variant calling.

3.2.4 Variant detection

In order to obtain a large set of somatic variant candidates, we combined the results from three different variant detection strategies: we applied the standard tools HaplotypeCaller (GATK) and MuTect [37] and we created the GATK-LOD$_{\mathrm{N}}$ method.

HaplotypeCaller, VariantRecalibrator and VariantFiltration (GATK)

For the first variant detection strategy, we applied the HaplotypeCaller, VariantRecalibrator, and VariantFiltration tools from the GATK package. The HaplotypeCaller performs the variant calling by applying a Bayesian algorithm that estimates the probability for the homozygous (AA or BB) and heterozygous (AB) genotypes.

The VariantRecalibrator was applied for filtering variants in the largest datasets: AML and ALL. Its algorithm assigns scores to the candidate variants through the variant quality score recalibration (VQSR) process. The VQSR is based on the idea that variants with similar characteristics as previously known variants are likely to be real, whereas those with unusual characteristics are more likely to be machine or

Table 3.1: List of parameters and thresholds used for SNV hard filtering

QualByDepth (QD)	The variant confidence divided by the unfiltered depth	< 2
FisherStrand (FS)	Phred-scaled p-value using Fisher's Exact Test to detect strand bias	> 60
RMSMappingQuality (MQ)	The Root Mean Square of the reads mapping qualities	< 40
MappingQualityRankSumTest (MQRankSum)	The u-based z-approximation from the Mann-Whitney Rank Sum Test for mapping qualities (reads with reference bases vs those with the alternate allele) (Only applied to heterozygous calls)	< -12.5
ReadPosRankSumTest (ReadPosRankSum)	The u-based z-approximation from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele. If the alternate allele is only seen near the ends of reads, this is indicative of error (Only applied to heterozygous calls)	< -8

data processing artifacts. It employs a variational Bayes Gaussian mixture model (GMM) to estimate the probability that each variant is a true polymorphism rather than a sequencing, alignment or data processing artifact. The set of variants v_i are treated as an n-dimensional point cloud, each variant v_i positioned by its covariate annotation vector, $\vec{v}$. A mixture of Gaussians is fit to the set of likely true variants, here approximated by the variants already present in the HapMap3 database. Following training, this mixture model is used to estimate the probability of each variant call being true.

Since the number of samples was too small to apply the variant quality score recalibration for the GIST and Lung Adenocarcinoma datasets, we followed the GATK best practices instructions and applied a "hard" filtering approach by using the VariantFiltration tool. The table 3.1 shows the list of parameters and their thresholds used for SNV hard filtering.

MuTect

In our second variant detection strategy we applied the MuTect software [37]. It analyzes the tumor and normal samples at the same time, performing the variant detection, filtering, and their classification in germ-line or somatic.

This variant detection algorithm was designed to detect somatic mutations with very low allele fractions. It analyzes the data at each site considering two alternate models: (i) a reference model, M_0, which assumes there is no variant at the site and any observed non-reference bases are sequencing errors, and (ii) a variant model, M_f^m, which assumes that the site contain a true variant allele m at the allele fraction f. The variant m is declared as a somatic candidate if the log-likelihood ratio of the data under the variant and reference models (i.e., the log-odds (LOD) score) exceeds a predefined decision threshold.

Mutect applies six filters for variant filtration and they are listed in the table 3.2.

GATK-LOD$_N$

The GATK software analyzes the tumor and control samples separately. The somatic mutations are classified if they are observed in the tumor samples but not in the normal samples. However, many variants are artifacts derived from: coverage

Table 3.2: Description of filters and default thresholds applied by MuTect

Filter name	Description and default thresholds
Proximal gap	Remove sites with nearby misaligned small insertion and deletion events. Reject candidate site if there are ≥ 3 reads with indels in an 11-base-pair window centered on the candidate mutation
Poor mapping	Two tests are used to identify such sites: (i) candidates are rejected if $\geq 50\%$ of the reads in the tumor and normal samples have a mapping quality of zero; and (ii) candidates are rejected if they do not have at least a single observation of the mutant allele with mapping quality score ≥ 20
Triallelic site	Reject sites where MuTect is considering an alternate allele C in the tumor sample and the normal samples is heterozygous with alleles A and B. Although this is biologically possible, calling at these sites generates many false positives.
Strand bias	Reject sites that have the majority of the alternate alleles observed in a single read direction. This test is performed by stratifying the reads by direction and then applying the core detection statistic on the two data sets. It also calculates the sensitivity to have passed the threshold given the data. Candidates are rejected when the strand-specific LOD is < 2.0 in directions where the sensitivity to have passed that threshold is $\geq 90\%$
Clustered position	Reject sites hallmarked by the alternate alleles being clustered at a consistent distance from the start or end of the read alignment. It is performed by calculating the median and median absolute deviation of the distance from both the start and end of the read and reject sites that have a median ≤ 10 (near the start/end of the alignment) and a median absolute deviation ≤ 3 (clustered)
Observed in control	Reject sites in the tumor data by looking the matched normal sample for evidence of the alternate allele beyond what is expected from random sequencing error. A candidate is rejected if, in the normal sample, there are (i) ≥ 2 observations of the alternate allele or they represent $\geq 3\%$ of the reads; and (ii) their sum of quality scores is > 20

differences, and sequencing and alignment errors. To ensure the somatic classification of each SNV candidate from the GATK output, we developed a method called GATK-LOD$_{\mathrm{N}}$, which was adapted from the MuTect algorithm and is based on the LOD_N score. If the LOD_N was less than a predefined decision threshold, the variant was classified as somatic. In the following lines we explain the calculation of the LOD_N score, as defined in the MuTect original publication [37]:

$$LOD_N = \log_{10}\left(\frac{L(M_0)P(m,f)}{L(M_{0.5}^m)P(\text{germ line})}\right) \geq \log_{10}\theta_N \tag{3.2}$$

The terms M_0 and $M_{0.5}^m$ represent the variant model M_f^m when $f = 0$ and $f = 0.5$, respectively. The $f = 0.5$ represents a germline heterozygous variant. The likelihood of the model M_f^m is given by:

$$L(M_f^m) = P(b_i|e_i, r, m, f) = \prod_{i=1}^{d} P(b_i|e_i, r, m, f) \tag{3.3}$$

The reference alleles are denoted as r, the called base at the read i as b_i and e_i as the probability of base miscalling. The term $P(m, f)$ was determined by assuming that $P(m)$ and $P(f)$ are statistically independent and that $P(f)$ is uniformly distributed (that is, $P(f) = 1$). The $P(m)$ was set as a typical mutation frequency of $3 * 10^{-6}$. The P(germline) was distinguished according (i) sites known to be variant in the population (i.e., present in dbSNP database) and (ii) all other sites. There are $30 * 10^{-6}$ sites known to be variant in the human population according to dbSNP release 134, which is 1000 variants/Mb. A given individual typically has $3 * 10^{-6}$ variants in their genome, 95% of which are in dbSNP sites. Therefore, we expect 50 variants/Mb not at dbSNP, that is, P(germline|non-dbSNP site) $= 5 * 10^{-5}$. At dbSNP sites we expect 95% of variants in the $3 * 10^{-6}$ sites in the dbSNP database, yielding P(germline|dbSNP site) $= 0.095$.

The LOD_N score was calculated from the results of the GATK variant filtration.

3.2.5 Variant Annotation

The process of adding biological information to each variant is called annotation, and it was performed by the Annovar software [38]. The gene-based annotation identified whether the variants caused protein coding changes and which aminoacids were affected. It was performed by using the information available in the Ensembl Gene annotation database for the build 37 of the human genome (`www.ensembl.org/`). Variants were removed if reported in the dbSNP138 and 1000 genomes databases with minor allele frequency (MAF) greater than 0.05.

3.2.6 Availability of data and material

The scripts and instructions for running the main pipeline steps are available for the community in the link: `https://bitbucket.org/BBDA-UNIBO/wes-pipeline`.

3.3 Pipeline Testing

3.3.1 Testing MuTect thresholds

Part of the tumor samples were previously profiled by Sanger Sequencing and it permitted us to evaluate the number of False Negatives in the MuTect output. We asked if we could could lower the MuTect decision thresholds in order to reduce its selectivity and increase the number of True Positives without increasing too much the number of False Positives. We created an adapted version of the MuTect algorithm in which we lowered the threshold that determines the mutation detection ($\theta_T \geq 6.5$) and the threshold that determines if the mutation is a somatic event ($\theta_{N|\text{dbSNP site}} \geq 5.5$). This adapted-MuTect was applied to the SNVs in the GATK output (prior to the variant filtering). We set as new thresholds to the minimum values that would permit the correct classification of 10 false negative variants.

3.3.2 Simulated datasets

We simulated datasets to evaluate the specificity and sensitivity of the methods MuTect, GATK and GATK-LOD_N.

As each read is independently sequenced, datasets can be simulated by splitting the sequencing data of the same sample in many subsets. We selected the alignment (80X) of the saliva sample a1025 from the AML cohort and randomized its reads. Then, we split it in two subsets by applying the bamutils tool from the NGSUtils package [39]. By considering one of the halves as a normal sample and the other as a tumor sample, we applied the methods and evaluated the number of called sSNVs. Since these two simulated samples originated from the same saliva sample, all called variants were considered as false positives.

The sensitivity was calculated by creating artificial tumor samples. We adapted the mutate_sample.py script from the Shimmer package [40] to create mutations in the alignment of the a1025 saliva sample. Three artificial tumor samples were created with 22, 25 and 25 sSNVs, which had variant allelic frequencies (VAF) in the

Table 3.3: Artifical Tumor Samples

Chromosome	Position	REF>ALT	Artificial Tumors Variant allelic frequencies			Normal Variant Allelic Frequencies
			0.02-0.26	0.5-0.86	0.97-1	
11	19854088	G>A	0.03	0.69	1.00	0.00
11	36484167	C>T	0.08	0.62	1.00	0.03
11	4608116	T>C	0.13	0.71	1.00	0.02
11	4661826	T>C	0.11	0.60	0.97	0.03
11	4673788	G>A	0.26	0.64	1.00	0.02
11	4928841	T>C	0.13	0.61	1.00	0.00
11	5372856	A>G	0.24	0.69	1.00	0.02
11	5373562	C>A	0.09	0.68	1.00	0.03
11	5443887	T>C	0.10	0.86	1.00	0.00
11	5443893	G>A	0.10	0.86	1.00	0.00
11	5462255	C>G	0.16	0.56	1.00	0.00
11	5906203	T>G	0.19	0.70	1.00	0.00
11	6519642	G>A	0.08	0.61	1.00	0.00
11	824789	T>C	0.11	0.63	1.00	0.03
12	25398281	C>T	0.12	0.63	1.00	0.00
12	75715330	C>A	0.13	0.60	1.00	0.00
22	24891418	A>C	0.21	0.70	1.00	0.03
22	44083442	T>C	NA	0.78	1.00	0.00
13	101289801	C>A	0.13	0.65	1.00	0.00
20	61537337	G>T	0.13	0.65	1.00	0.00
17	48557299	G>T	0.11	0.74	1.00	0.00
5	45262378	G>T	0.08	0.50	1.00	0.00
1	94476902	T>C	0.15	0.65	1.00	0.00
2	110372199	G>T	NA	0.57	1.00	0.00
5	64907465	C>A	0.10	0.57	1.00	0.00

Table 3.4: Number of SNVs submitted to experimental validation

	Mutation Detection	Mutation Classification
GATK	48	14
GATK-LOD$_N$	9	4
MuTect	22	8

range of 0.02 to 0.25, 0.5 to 0.86, and 0.97 to 1.0, respectively (Table 3.3). For each artificial tumor sample, we created subsets by randomly excluding reads in order to simulate sequencing coverages in the range of 5X to 80X, with intervals of 5X. The creation of the subsets was performed by the DownsampleBam tool of Picard. We then evaluated the performance of each variant calling method at the different coverage levels.

3.3.3 Experimental validation

We selected a subset of SNV candidates for experimental validation. They were selected from the output of each method (MuTect, GATK, and GATK-LOD$_N$) and we tested if these variants were confirmed in the tumor sample (Mutation Detection) and if they were truly classified as somatic events (Mutation Classification) (Table 3.4). Variants with allelic frequency higher than 0.2 were validated by Sanger Sequencing and those with allelic frequency lower than 0.2 were validated by using the Illumina TrueSight Myeloid Sequencing Panel and Illumina MiSeq sequencing. The results were analyzed by the VariantStudio software (Illumina), according to manufacturer's instruction.

Chapter 4

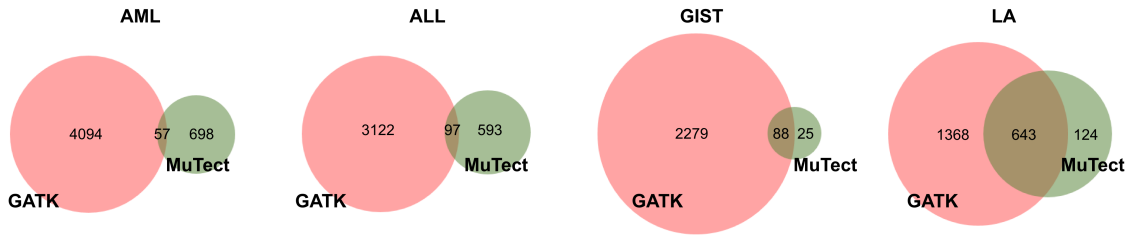
Results

4.1 Comparison of methods

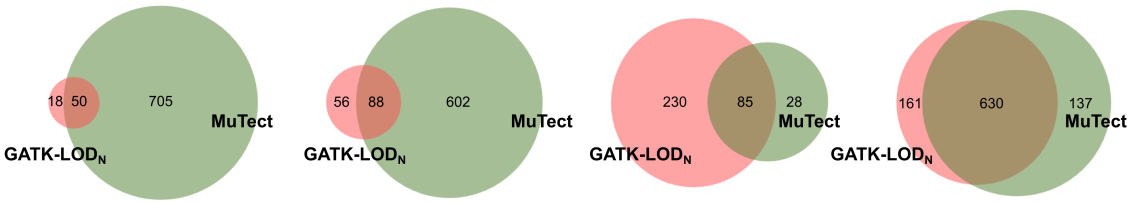
We built a pipeline for discovery of somatic single nucleotide variants (sSNVs) in whole exome sequencing data and applied it to Acute Myeloid Leukemia (AML), Acute Lymphoid Leukemia (ALL), Gastrointestinal Stromal Tumor (GIST), and Lung Adenocarcinoma (LA) samples.

First, we compared the results of the variant detection procedures MuTect and GATK. GATK detected more sSNVs than MuTect in all datasets: 5.5, 4.6, 20.9 and 2.6 times more than MuTect in the datasets AML, ALL, GIST, and LA, respectively. The results also showed low concordance between GATK and MuTect results: 1.1%, 2.54%, 3.67%, 30.11%, for the AML, ALL, GIST and ALL datasets, respectively (Figure 4.1a). This low concordance indicated that there were several method-specific sSNVs that could be considered as final candidates by merging the results from both methods. However, as GATK presented larger numbers of candidates in comparison with MuTect, we hypothesized that the GATK results also presented proportionally larger numbers of false positives.

In order to merge both results without increasing the number of false positives, we created the GATK-LOD_N variant detection procedure to filter the false positives from the GATK candidates. By comparing GATK and GATK-LOD_N, we observed that the latter filtered 98.36%, 95.52%, 86.69%, and 60.66% of the GATK candidates in the AML, ALL, GIST, and LA datasets, respectively (Figure 4.1b). As we can see, the filter strongly reduced the GATK candidates in the hematological tumors, but approximately 10% of the GATK specific sSNVs remained after the filtering in the solid tumors. Interestingly, after the filtering in the GIST dataset, GATK-LOD_N final candidates still represented three times more candidates than the MuTect results.



(a)



(b)

Figure 4.1: Comparison of the number of sSNVs between GATK and MuTect before (4.1a) and after (4.1b) applying the GATK-LOD$_N$ filter for each whole exome sequencing dataset. AML: Acute Myeloid Leukemia, ALL: Acute Lymphoblastic Leukemia, GIST: Gastrointestinal Stromal Tumor, LA: Lung Adenocarcinoma.

Table 4.1: MuTect False Negatives

Patient	Gene	Chr	Pos	Ref	Alt	θ_N	θ_T	N_{alt}	N_{ref}	T_{alt}	T_{ref}
b2042	GBP4	1	89652071	T	C	3.31	4.72	0	11	2	23
b2042	GBP4	1	89652072	C	T	3.01	4.52	0	10	2	23
b2002	YLMP1	14	75245170	T	A	NA	14.79	0	3	5	12
b2002	MADCAM1	19	501786	C	A	7.12	5.45	6	29	9	31
b2002	MADCAM1	19	501801	A	C	7.22	5.22	0	24	2	29
b2002	MADCAM1	19	501802	G	C	7.22	5.50	0	24	2	26
a1025	ATXN1	6	16327915	A	C	19.25	6.14	0	64	3	108
a1024	CENPO	2	25022705	C	T	4.51	26.56	0	15	9	25
b2035	GBP4	1	89652071	T	C	4.81	5.19	0	16	2	13
b2035	GBP4	1	89652072	C	T	4.81	4.87	0	16	2	15

* All listed positions are reported in the dbSNP database
$N_{alt,ref}$: Number of reads having the alternate or reference allele in the normal sample
$T_{alt,ref}$: Number of reads having the alternate or reference allele in the tumor sample
NA: MuTect did not calculate the LOD_N when the normal samples had a total read depth coverage < 8.

Table 4.2: Number of variants found by MuTect, before and after relaxing the θ_T and θ_N parameters for six Acute Myeloid Leukemia (AML) normal-cancer sample pairs

Patients	MuTect	MuTect Adapted[1]
a1024	11	39
a1025	31	41
b1014	22	54
b2002	10	25
b2035	43	419
b2042	58	338

[1] Applying the computation of θ_T and θ_N, from the MuTect algorithm, with lowered threshold values (4.5 and 3, respectively) downstream to the GATK analysis

4.2 Pipeline Testing

4.2.1 Testing MuTect thresholds

We observed that MuTect miscalled variants that were previously confirmed by Sanger sequencing (Table 4.1) and we tested if we could lower the algorithm stringency and reduce the number of false negatives. For each one of the reported false negatives, we calculated the MuTect parameters that are used for sSNV detection and classification (θ_T and θ_N, respectively). We observed that the minimum threshold values for their correct classification would be $\theta_T \geq 4.5$ and $\theta_{N|\text{dbSNP site}} \geq 3$.

We observed that applying these decision thresholds increased the number of final candidates approximately 1.3 to 19 times in comparison with the original MuTect output (Table 4.2). This result means that reducing the MuTect stringency may increase the number of true positives, but with the cost of many false positives.

Table 4.3: Performance of MuTect and GATK-LOD$_N$ for artificial tumor samples having variants with diverse allelic frequencies

		Artificial Tumor Samples		
		Low Frequency VAF: 0.02 – 0.26	Intermediate Frequency VAF: 0.5 – 0.86	High Frequency VAF: 0.97 – 1
MuTect	Somatic Candidates	22	25	25
	TP	19	25	25
	FN	0	0	0
	FP	3	0	0
	PPV	19/22	25/25	25/25
	FDR	3/22	0/25	0/25
GATK-LOD$_N$	Somatic Candidates	27	32	33
	TP	17	23	23
	FN	5	5	2
	FP	5	7	8
	PPV	17/22	23/30	23/31
	FDR	5/22	7/30	8/31

TP: True positives, FN: False negatives, FP: False positives, PPV: Positive Predictive Value (#TP / (#FP + #TP)), FDR: False Discovery Rate (#FP / (#FP + #TP)), VAF: Variant Allelic Frequency. GATK results were not reported in the table since it detected more than 2200 candidates out of 22 or 25 TPs.

4.2.2 Simulated datasets

Simulated data permitted the evaluation of sensitivity and specificity of the three variant detection approaches (MuTect, GATK and GATK-LOD$_N$). We created the first simulated dataset by splitting a saliva sample alignment (80X) in two. As both samples originated from the same alignment, all sequencing and mapping artifacts were the same, but the splitting would create small deviations of the allelic frequencies in the two halves. By counting the number of resulted sSNVs, we observed how each method deals with small coverage deviations in normal and tumor samples. MuTect detected the lowest (8), GATK-LOD$_N$ an intermediate (35), and GATK the highest (76) number of false positives.

Then, we applied technical replicates of the same saliva sample to the pipeline. In this case, each replicate had its own set of sequencing and mapping artifacts and the number of sSNVs resulted from each method reflected how they deal with sequencing/mapping errors. Again, MuTect detected the lowest (7), GATK-LOD$_N$ an intermediate (33), and GATK the highest (84) number of false positives.

We measured the methods sensitivity by creating three artificial tumors from the same saliva sample alignment: we inserted high-frequency SNVs in the first (n=25, VAF: 0.97 to 1.0), intermediate-frequency in the second (n=25, VAF: 0.5 to 0.86), and low-frequency in third (n=22, VAF: 0.02 to 0.25).

GATK presented the worst performance, detecting 2,206 candidates out of 22 or 25 true positive variants. MuTect presented a Positive Predictive Value (PPV) of 0.86 (19/22) for low VAF mutations and its false negatives either presented VAF under 0.1 or low sequencing depths (Table 4.3). GATK-LOD$_N$ presented a PPV of 0.77 (17/22) for the low allelic frequency variants, but it also missed variants with very low VAFs (< 0.095) (Table 4.3). MuTect detected all intermediate and high allelic frequency variants, while GATK-LOD$_N$ presented PPVs of 0.76 (23/30) and 0.74 (23/31), respectively (Table 4.3).

In order to evaluate the detection methods performance at different coverage levels, we simulated sequencing coverages, in the range of 5X to 80X, for each ar-

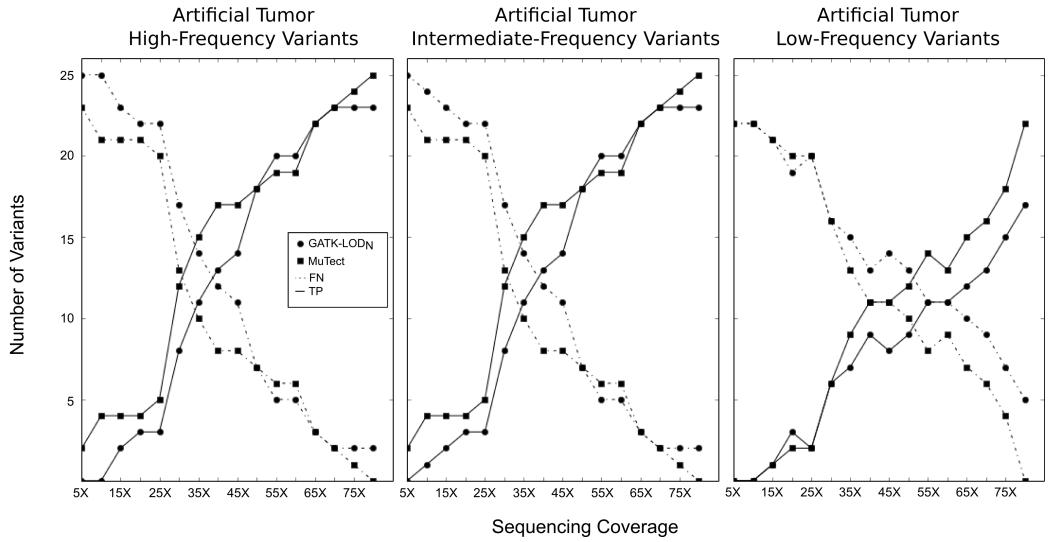


Figure 4.2: Number of False Negatives and True Positives at different coverage levels. Three artificial tumors were created with 22, 25 and 25 SNVs, which had variant allelic fractions range of 0.02 to 0.25, 0.5 to 0.86, and 0.97 to 1.0, respectively. We counted the number of False Negatives (FN) and True positives (TP) for different levels of simulated sequencing coverage.

tificial tumor sample. We observed that, at different coverage levels, GATK-LOD$_N$ and MuTect presented almost identical performance for the artificial tumors with high and intermediate variant frequency SNVs, except in the number of false negatives detected by GATK-LOD$_N$ at the coverage interval of 5 to 20X. GATK-LOD$_N$ presented increased number of detected true positives than MuTect in the coverage interval of 50 to 55X for high and intermediate-frequency variants, and in the coverage 20X for low-frequency variants (Figure 4.2).

4.2.3 Experimental Validation

We evaluated the performance of each variant detection method by performing two rounds of experimental validation (Sanger sequencing): 1) in the first we tested the sSNVs presence in the tumor samples (SNV detection), and 2) in the second we tested both tumor and normal samples, in order to evaluate the correct classification as a somatic event (SNV classification).

We observed that 37.5% (18/48) and 27.7% (5/18) of GATK variants were correctly detected and classified, respectively. MuTect presented the highest performance in both rounds (85.7% (6/7) and 66.6% (2/3), respectively). The GATK-LOD_N presented intermediate performance, resulting in 66.6% (6/9) and 75% (3/4) validated variants, for mutation detection and classification, respectively (Table 4.4).

Table 4.4: The GATK-LOD_N method increases the GATK performance for both mutation detection and classification

	SNV Detection [a]		SNV Classification [b]	
	Tested	Validated	Tested	Validated
GATK-LOD_N (specific)	4	1	2	2
GATK-LOD_N (all variants)	9	6	4	3
GATK (without GATK-LOD_N) (specific)	37	11	9	2
GATK (without GATK-LOD_N) (all variants)	48	18	14	5
MuTect (specific)	22	21	8	8
MuTect (all variants)	29	27	11	10
MuTect & GATK	7	6	3	2

The Sanger sequencing validation was performed in two rounds: in the first round we tested whether the methods correctly detected the mutation and in the second one we assessed whether the methods correctly classified the mutations as somatic events. The variant subsets tested (AML datatset) presented variants method specific and variants detected by one or more methods.

[a] variants tested for correct mutation detection

[b] variants tested for correct classification as somatic events

Chapter 5

Discussion

Cataloguing somatic single nucleotide variants (sSNVs) is essential to understand the genetic bases of cancer and extensive efforts have been made towards accurate variant calling approaches. The main challenge relies on removing errors derived from multiple sources, detecting rare variants in highly heterogeneous tumor samples, and detecting rare variants at small sequencing coverage levels. Different approaches may be more successful dealing with some of these challenges and less so with others. Therefore, we did not considered the option of considering just the results obtained from one tool, since it would risk the selection of errors for which the algorithm is vulnerable [21]. Another option would be taking the intersection of multiple variant callers, but it would result in high false negative rates, since the concordance between tools is very small and each one might uniquely identify true variants [28]. Our data show that the combination of standard tools - Genome Analysis Toolkit (GATK) and MuTect – improves the range of detected SNVs in whole exome sequencing data of cancer samples. We also developed the GATK-LOD$_N$ method, which reduced the number of GATK false positive calls.

The GATK is one of the variant callers used by the 1000 Genomes Consortium [41]. It uses a Bayesian model to estimate the likelihood of a genotype given the sequenced reads that cover the locus and it independently calls genotypes in tumor and normal samples, being the somatic candidates reported as those only present in the tumor sample. A previous study demonstrated that the SNVs found only by GATK had relatively high validation rates [28], but, in our dataset, its results presented several false positives (Table 4.4). These false calls are likely germline variants that are not called in normal samples because of low sequencing coverage or low allelic frequency.

MuTect jointly analyzes tumor and normal samples by also applying a Bayesian classifier to detect SNVs with very low allelic fractions, requiring only a few supporting reads, followed by tuned filters to ensure specificity. When applied to a cohort of 11 tumor types, it presented high sensitivity and specificity, with validation rates superior to 90% [37]. MuTect was applied to the Acute Myeloid Leukemia dataset of the TCGA project, but it was latter demonstrated that its results contained a consistent presence of false negatives, leading to undervaluation of somatic variants occurrence in this tumor type [29]. In our dataset, we observed at least 10 MuTect false negatives (Table 4.1), but we discarded the option of relaxing the algorithm decision thresholds, because, even if it detected variants previously miscalled, the final results included many false positives (Table 4.2).

The combination of different approaches has been suggested in literature. For instance, Kim *et. al*, 2014 [42] combined the output of different variant callers by forming linear combinations of different calls to predict the true somatic status of a given variant. The authors also demonstrated that, by including genomic features in the model through a Feature-weighted linear stacking approach, their method could learn in which type of locus each algorithm is typically right. However, our pipeline resulted as a straightforward and simple way of combining the advantage of different algorithms: GATK presented high amounts of false positive calls (type I error), MuTect presented high amounts of false negative calls (type II error), and so the GATK-LOD_N method is an option of increasing the range of detected SNVs without severely compromising sensitivity and specificity.

Our results show that GATK-LOD_N reduced the number of GATK false positives and detected variants that were missed by MuTect (Figure 4.1). The experiments in the simulated artificial tumor samples and the sequencing validation showed that GATK-LOD_N increased the GATK performance (Figure 4.2 and Table 4.4, respectively). However, we performed the validation experiments just for variants from the hematological tumors (available in our laboratories) and the validation rate might change for solid tumors. We observed that the GATK-LOD_N also outperformed MuTect in some simulated sequencing coverages (Figure 4.2). As sequencing datasets usually present large variability in coverage and quality, the different error modeling approaches and prior assumptions associated with these two methods should permit good performances in a wide range of scenario possibilities.

The results show that GATK-LOD_N filtered more variants in the hematological tumors than in the solid tumors and we hypothesized that a possible cause might be that normal samples from hematological tumors are more prone to contamination by cancer cells. Although GATK-LOD_N provided a small number of variants in the hematological datasets, even a single variant can give insights into the mechanisms of malignant transformation and help design personalized therapeutic approaches [43]. We observed that the Lung Adenocarcinoma presented the biggest concordance between methods, maybe because patients with this type of cancer usually presents high mutation frequencies and harbors more somatic mutations compared with other cancer types [17, 44].

The GATK-LOD_N is suitable for application together with other post-calling filtering strategies proposed in the literature: strand bias, nearby polymorphisms and technology specific sequencing errors removal [33, 35, 45]. For instance, Carson et al [26] suggested new thresholds for genotype and variant filters to be used in conjunction with the GATK pipeline analysis that could increase the GATK-LOD_N performance in population-based studies. Altogether, the GATK-LOD_N allows enough flexibility to deal with different study designs and requirements about how stringent the analysis must be.

Finally, we believe that the GATK-LOD_N can be of service in large-cohort discovery studies, helping in the understanding of cancer biology through the discovery of somatic single nucleotide variants in cancer sequencing data.

Part II
Network integration of multi-tumor omics data for the discovery of novel targeting strategies

Chapter 6

Introduction

6.1 Oncogenic alterations across human cancers

For decades, anatomical localization and histological features have guided the identification of cancer subtypes, but the genomic profiling of tumor samples has revealed differences and similarities that go beyond the histopathological classification.

The diversity in genomic alteration patterns often stratifies tumors from the same organ or tissue, while tumors in different tissues may present similar patterns. For example, TP53 mutations drive high-grade serous ovarian, serous endometrial and basal-like breast carcinomas, all of which share a global transcriptomical signature involving the activation of similar oncogenic pathways [46, 47]. Other commonalities across tumor types include inherited and somatic inactivation of the BRCA1-BRCA2 pathway in both serous ovarian and basal-like breast cancers, microsatellite instability in colorectal and endometrial tumors, and the POLE-mediated ultramutator phenotype characterized by extremely high mutation rates, common to both colon and endometrial cancers [46, 48]. Hoadley *et. al*, 2013[U+2060] [49] suggests that lung squamous, head and neck, and a subset of bladder cancers form a unique cancer category typified by specific alterations, while copy number, protein expression, somatic mutations and activated pathways divide bladder cancer into different subtypes. The analysis of cancer transcriptomes revealed that the same tumor type may originate from several cell types and different biological processes may lead to the same malignant transformation. Moreover, the activation of similar pathways may occur across different cancers, as exemplified by high-grade serous ovarian, serous endometrial and basal-like breast carcinomas [46, 47, 50].

However, we still need to translate this increasing amount of knowledge into practical applications for cancer treatment and classification. A recent study divided tumor samples into a hierarchical system of classification based on two major classes: tumors primarily affected by mutations, and tumors primarily affected by copy number alteration events. For each class, the authors found a list of oncogenic signatures shared by tumor samples, and, based on the cellular processes that these genomic alterations reflected, they proposed therapeutic strategies for the different tumor classes [51]. A systematic pharmacogenomic profiling in cancer cell lines revealed several associations between drugs and genomic features (mutational, transcriptional and CNV profiles) that correlate with drug sensitivity [52]. For example, plasma cell lineage correlate with sensitivity to IGF1 receptor inhibitors;

AHR expression associates with MEK-inhibitor efficacy in NRAS-mutant cell lines, and SLFN11 expression predicts sensitivity to topoisomerase inhibitors [53].

The development of a molecular and functional perspective across tumors will result in the description of similar genomic profiles that will enable us to repurpose therapies from one cancer to another.

6.2 Data integration in cancer genomics

The huge amount of heterogeneous types of data for a large number of tumors requires novel approaches capable to integrate such information into a unified framework. To reach this goal, data integration methodologies have to meet many computational challenges, which arise arise owing to different sizes, formats and dimensionalities of the data being integrated, as well to their complexity, noisiness, information content and mutual concordance (i.e. the level of agreement between datasets) [54].

One integration approach is to describe biological data from different aspects of cellular information level as omic layers (Figure 6.1). In the cancer research, high-throughput technologies has permitted an extensive description of these layers: 1) the genome layer through the sequencing of cancer cell genomes; 2) the transcriptome layer, by microarray and RNA-sequencing technologies; and 3) the proteome layer, by the description of protein physical interactions through yeast two-hybrid assays and affinity purification with mass spectometry.

Each omic layer can be represented by networks (or graphs), with nodes representing entities (genes, proteins, etc) and links representing the relationships between nodes. In biological networks, links can represent physical, functional or chemical relationships between pairs of nodes. Network approaches have been the most widely used method for modeling and analyzing omics data, but we still need improved methods of network integration to detect and study gene-gene associations.

6.3 Objectives

In this study, we aimed to combine gene expression and mutational data from several cancer types in order to find multi-tumor drug targets, prognostic markers, and a molecular taxonomy for effective cancer categorization. Our approach relied on reducing the complexity of thousand of genes to a curated subset of cancer-related genes described by the Ontocancro database. Based on tumor expression profiles extracted from The Cancer Genome Atlas (TCGA) data portal, we performed a tumor-wise clustering approach to define clusters of tumors. Then, to find gene signatures for the tumor clusters, we applied a network analysis approach that combined: the curated set of cancer-related pathways described in the Ontocancro database, gene expression profiles, and the BioPlex protein-protein interaction network.

The relevance of the gene signatures was assessed by: considering the mutational and clinical data available for the cancer types considered in this study, drug-gene associations according to the DrugBank database, and by evaluating the existence of ongoing clinical trials that investigate the inhibition of signature genes. Finally

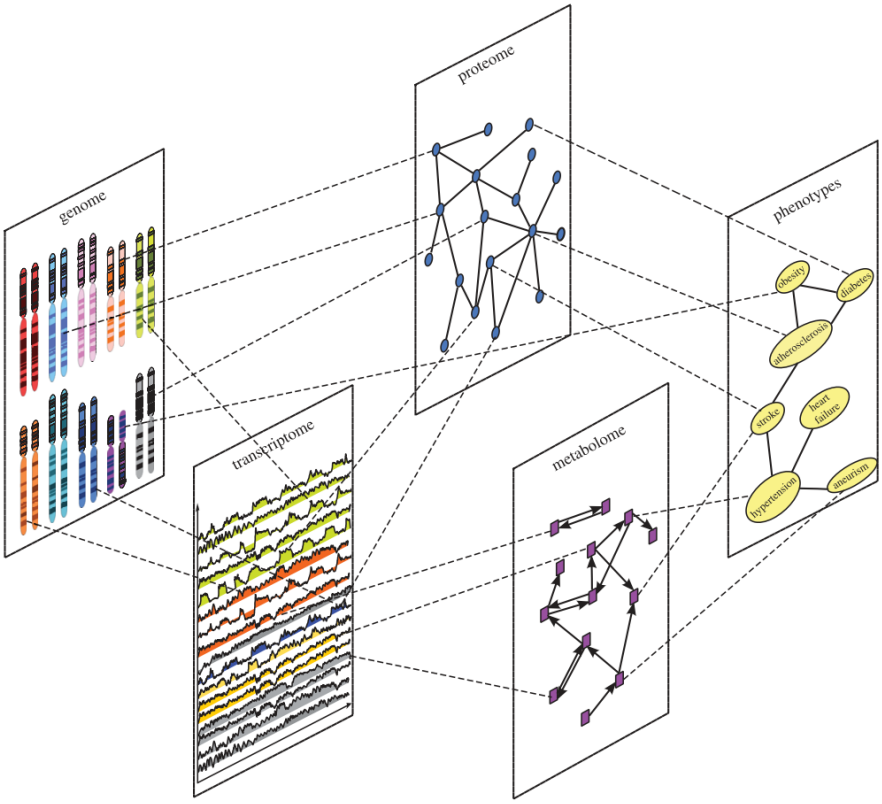


Figure 6.1: A schematic illustration of the molecular information layers of a cell (Adapted from Gligorijevic & Przulj, 2015 [54]).

we treated two cancer cell line models with three existing drugs that target genes in the signatures and evaluated their potential to inhibit cell growth.

We believe that our study will help both clinical and research communities, providing novel targets for multi-drug approaches and for the repurposing of existing drugs.

Chapter 7

Material and Methods

7.1 Gene Expression Datasets

The gene expression datasets used in this study were retrieved from The Cancer Genome Atlas (TCGA) Data Portal. The entire dataset included Agilent expression arrays of 2,378 samples from 11 tumor types (Table 7.1).

In order to reduce the data high dimensionality (17,814 genes), we performed a knowledge-based selection of cancer-related genes having protein protein interaction information. For the proteomic information we selected the BioPlex protein-protein interaction (PPI) network [55]. This network represent the results from an experiment of affinity purification of epitope-tagged proteins followed by mass spectrometry (AP-MS), which resulted in a human interaction map of 56553 interactions among 10961 proteins. The list of cancer-related genes was retrieved from the Ontocancro database (`http://ontocancro.inf.ufsm.br/`), which provides curated annotations for cancer-associated genes related to specific biological functions as: cell cycle, DNA damage response, and inflammation. We selected for this study, the genes present in both BioPlex network and Ontocancro database, resulting in a list of 760 genes.

Table 7.1: List of tumors and the respective number of gene expression arrays analyzed

Abbreviation	Cancer	Number of patients
BRCA	Breast invasive carcinoma	593
COAD	Colon adenocarcinoma	172
GBM	Glioblastoma multiforme	595
KIRC	Kidney renal clear cell carcinoma	72
KIRP	Kidney renal papillary carcinoma	16
LGG	Brain lower grade glioma	27
LUAD	Lung adenocarcinoma	32
LUSC	Lung squamous cell carcinoma	155
OV	Ovarian serous cystadenocarcinoma	590
READ	Rectum adenocarcinoma	72
UCEC	Uterine corpus endometrial carcinoma	54
	Total	2,378

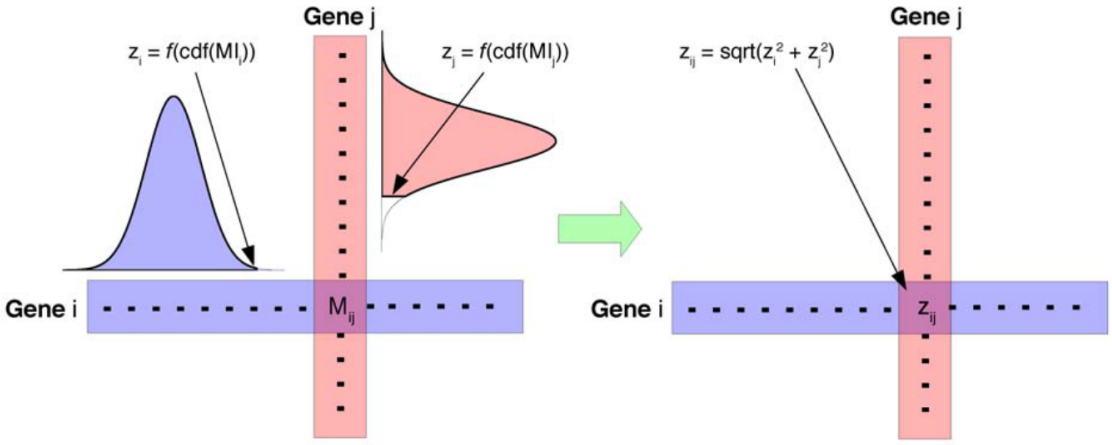


Figure 7.1: The Context Likelihood of Relatedness (CLR) algorithm. The CLR algorithm computes a z-score for each gene-gene correlation and it depends on the distribution of the Pearson's coefficients for all possible interactions of a given pair of genes i and j (Figure adapted from [57]).

7.2 Tumor clustering

In order to find clusters of tumors, we grouped the tumor types according to their transcriptomical profiles. We studied the relation between genes in each tumor dataset by calculating a correlation matrix containing pairwise Pearson r_{ij} coefficients across all samples.

In order to filter false correlations and indirect influences, the absolute correlation values ($|r_{ij}|$) were adjusted with the Context Likelihood of Relatedness (CLR) algorithm [56, 57] (Figure 7.1). In this study, the CLR estimated the correlation likelihood for a particular pair of genes i and j by comparing their correlation coefficient to a background distribution (the null model). The background distribution was considered to be two set of values: $\{r_i\}$ and $\{r_j\}$, the set of all correlation coefficients for gene i and j, respectively. The CLR algorithm approximates this background distribution as a joint normal distribution with $\{r_i\}$ and $\{r_j\}$ as independent variables. Thus, the final form of the likelihood estimation was:

$$f(Z_i, Z_j) = \sqrt{Z_i^2 + Z_j^2} \tag{7.1}$$

where Z_i and Z_j are the z-scores of $r_{i,j}$ marginal distributions, being the $f(Z_i, Z_j)$ the joint likelihood measure. In this study, we used the CLR function implemented in the R/Bioconductor package "minet" [58] [U+2060].

The matrices containing the z-scores computed by the CLR algorithm were clustered using the hierarchical clustering procedure. The clustering was based on the element-wise Euclidean distance between each pair of tumor matrices A and B, calculated as follows:

$$d(A, B) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{i,j} - b_{i,j})^2} \tag{7.2}$$

.

We applied the Ward linkage method, which minimizes the total within-cluster variance.

7.3 Multi-tumor gene signatures

A network approach was applied to find gene signatures that characterized the tumor clusters. First, we created a backbone network (BioPlex-Ontocancro) by selecting genes present in the BioPlex PPI network that were also annotated in the Ontocancro database (Figure 7.2).

Then, for each tumor cluster, the gene-gene correlation coefficients were computed and their absolute values were adjusted with the CLR algorithm. Each correlation matrix was superimposed to the BioPlex-Ontocancro network, producing networks (one for each cluster) in which genes were linked only if they correlated in the expression profiles and if they presented a physical interaction in the PPI network. For the cluster 1 and 3 networks, we selected the giant components (245 and 244 nodes, respectively), and for the cluster 2 we selected the two biggest components (149 and 118 nodes). The networks were analyzed and visualized by Networkx Python package and Cytoscape [59]`[U+2060]`.

We hypothesized that the most central genes in each cluster network would be those most functionally important for the tumors of each cluster. We selected the Spectral Centrality (SC) [60] measure to select the most central nodes, which calculates the effect of a node removal on the network diffusivity based on the spectral properties of the Laplacian graph. Thus, the gene signatures of each cluster were defined as the genes (nodes) having the SC measure above the 90^{th} percentile.

7.4 Validation of the multi-tumor gene signatures

7.4.1 Gene signatures and mutational data

In order to verify the relation between genes in the signatures and genes commonly mutated in cancer, we retrieved the somatic mutational data from TCGA data portal for the considered tumors. To avoid cancer unrelated mutations, we considered only mutations reported in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (`http://cancer.sanger.ac.uk`). The COSMIC database is the world's largest resource for exploring somatic mutational data in human cancers. In its latest release (August, 2014), it describes more than 2 million coding point mutations in over one million tumor samples, and all the mutational information is manually curated from the scientific literature [61].

We asked if the mutated genes were closer to the signature genes in the PPI network in comparison with all other genes. To quantify it, for each signature gene, we calculated the minimum distance (in terms of shortest path in the network) required to reach a mutated gene, and represented each signature as the average of minimum distances ($\bar{d}^{\,\text{real}}_{\text{min}}$). Then, we performed a permutation test by creating 10^6 random gene signatures, having the same size as the originals, and recalculated the average minimum distance to nearest mutated gene ($\bar{d}^{\,\text{random}}_{\text{min}}$). A p-value was calculated as the proportion of random signatures presenting an average minimum distance smaller than the real one ($\bar{d}^{\,\text{random}}_{\text{min}} < \bar{d}^{\,\text{real}}_{\text{min}}$).

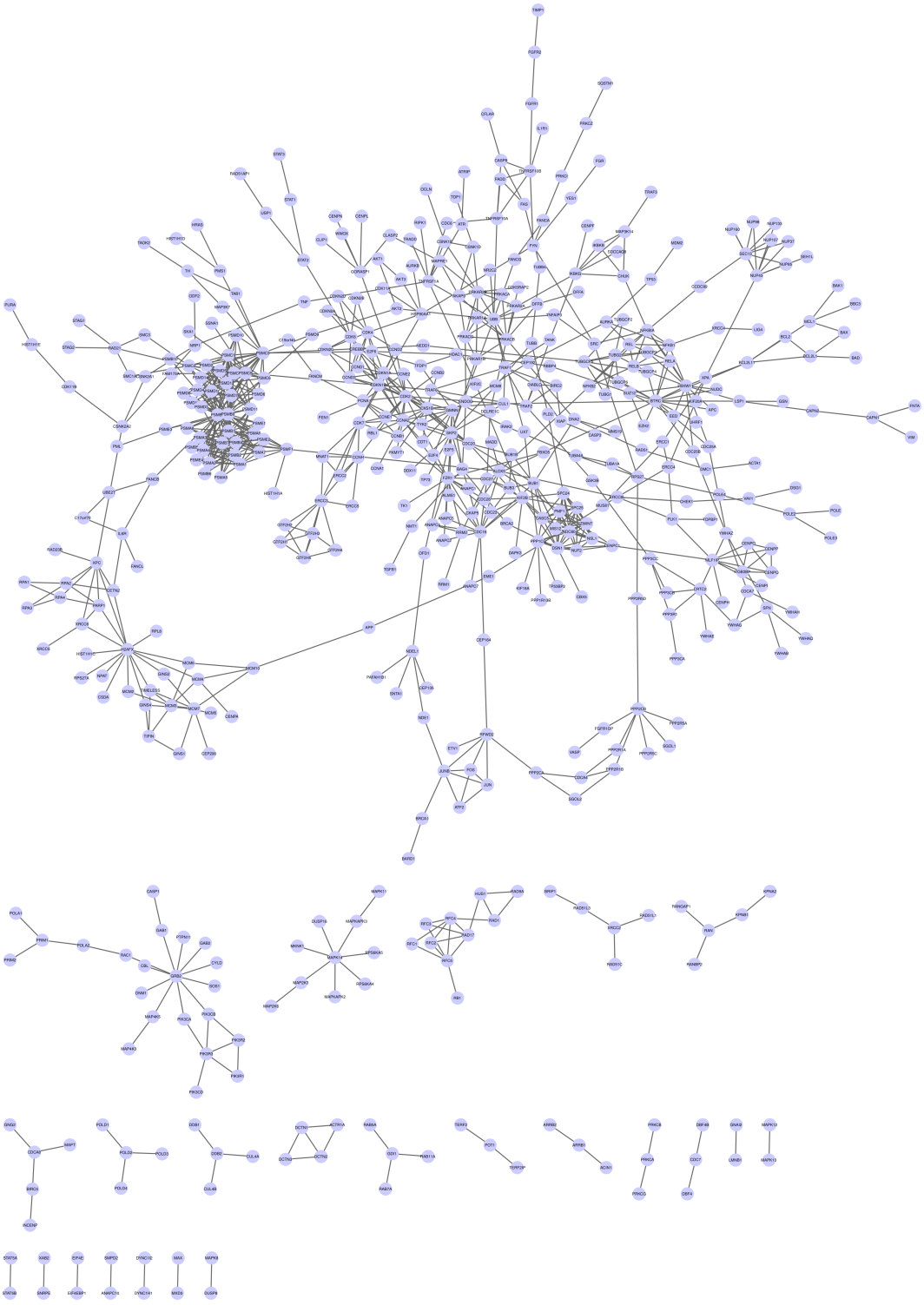


Figure 7.2: BioPlex Ontocancro Network. Network built from the genes present in both BioPlex network and Ontocancro database.

7.4.2 Gene signatures and pharmacological data

We retrieved drug-gene annotations from the DrugBank (`www.drugbank.ca`) and Drug Gene Interaction (DGIdb, `www.dgidb.org`). The DrugBank database is a resource that combines drug data (chemical, pharmacological and pharmaceutical) with drug target information (sequence, structure and pathway) [62]. The DGIdb is a web resource that combines twenty seven data sources describing drug-gene interactions and gene drugability [63]. By using the data from these databases, we evaluated possible drug target annotations for the genes in the signatures.

We also asked if the genes in the signatures have been evaluated in ongoing clinical trials as inhibition targets. To answer this question, we retrieve data from the the Aggregate Analysis of ClinicalTrials.gov (AACT) database (`www.ctti-clinicaltrials.org/aact-database`). The ClinicalTrials.gov is a web-based resource that provides access to information on publicly and privately supported clinical studies.

7.4.3 Gene signatures and prognosis

The prognostic potential of each gene signature was evaluated by considering the clinical data (days to death) available in the TCGA data portal. First, we clustered the patients having clinical information based on the expression levels of the gene signatures. We performed the k-means clustering procedure, which is one of the simplest unsupervised clustering algorithms. The algorithm classify a given dataset into a certain number of clusters (k) fixed *a priori*. The algorithm starts with two random cluster centers and iteratively moves the centers to minimize the total within cluster variance (until convergence) [64]. We fixed the number of clusters as 2 ($k = 2$), assuming the existence of two patient groups: one with good and another with bad survival outcome. We applied the k-means algorithm implemented in the Python package 'scikit'.

Then, we calculated survival curves for each one of the two patient groups defined by the k-means. The survival curves are the estimated probability of patient survival for a given time and we applied the Kaplan-Meier method. In this method, the survival probability at any particular time is calculated as:

$$S_t = \frac{n - d}{n} \tag{7.3}$$

being n the number of subjects living at the start; and d the number of died subjects. Survival to any point time is calculated as the product of all survival probabilities of preceding time intervals [65]. To compare the two survival curves we applied the log-rank test, which evaluates the null hypothesis that there is no difference regarding the groups in the probability of an event (here death) at any time point. Thus, each time an event occur, the test calculate the observed number of deaths and the number expected if there were no difference between groups. The χ^2 is used to test the null hypothesis, using the number of groups minus one (i.e., 2-1=1) degrees of freedom [66]. The survival curves and the log-rank test were calculated using the Python package "lifelines".

7.4.4 Gene signatures and *in vitro* inhibition

We asked if the defined gene signatures were good candidates for drug targeting. We selected two drugs that inhibit genes from the cluster 2 signature: Bortezomib and BI6727, which inhibit the genes PSMB3 and PLK1, respectively. We also selected the drug PF-00477736 (Selleckchem) for inhibiting the genes CHK1/2, which are not present in the gene cluster 2 signature but are strictly related to the genes in the signature.

We tested the effect of the inhibition of these genes in two cancer cell models: the glioblastoma T98G and the breast adenocarcinoma MCF-7 cell lines, obtained from ATCC and DSMZ, respectively. Cells were cultured at a density of 10^5 cells/ml in RPMI medium plus 10% FBS (plus 5% Sodium orthovanadate for T98G) for 72 hours with increasing drug concentrations, testing the single drugs or their combination. One hour and 30 minutes before the end of treatment, WST-1 reagent was added to the cell medium and cell viability was measured according to manufacturer's instruction (Roche). The dose-effect response and the IC50 of each drug were calculated using GraphPad Prism 6 (GraphPad Software).

We evaluated if the results indicated a synergistic effect, which means that the drug combinations provided better results in comparison with the single agent treatments. We computed the Combination Index (CI) using the CompuSyn software (ComboSyn Inc), in which values < 1, $= 1$, and > 1 indicate synergism, additive effect and antagonism, respectively.

Chapter 8

Results

8.1 Tumor clustering

In order to find sub-classes among 11 tumor types (Table 7.1), we analyzed 2,378 tumor samples considering a list of 760 cancer-related genes. The genes selected for this study were those present in the Ontocancro database that had protein-protein interaction annotations in the BioPlex network. The tumor datasets were clustered based on their gene-gene correlation matrices by applying a hierarchical clustering method. The clustering results indicate the existence of three tumor clusters containing 2, 6 and 3 cancer types: 1) Colon adenocarcinoma (COAD) and Rectum Adenocarcinoma (READ); 2) Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Glioblastoma Multiforme (GBM), Ovarian Serous Cystadenocarcinoma (OV), Breast Invasive Carcinoma (BRCA), and Uterine Corpus Endometrial Carcinoma (UCEC); and 3) Brain Lower Grade Glioma (LGG), Kidney Renal Clear Cell Carcinoma (KIRC), and Kidney Renal Papillary Cell Carcinoma (KIRP) (Figure 8.1).

8.2 Multi-tumor gene signatures

Then, we searched for multi-tumor gene signatures characterizing the tumor clusters by applying a network analysis approach. We superimposed each cluster

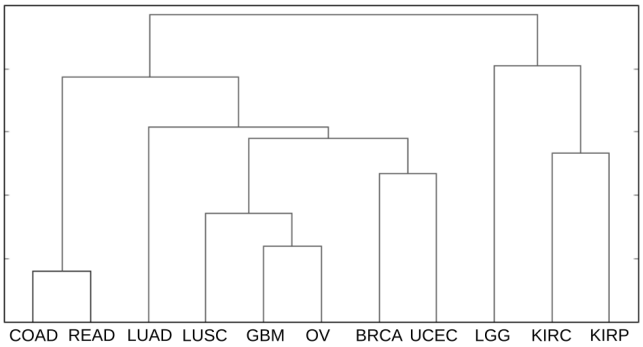
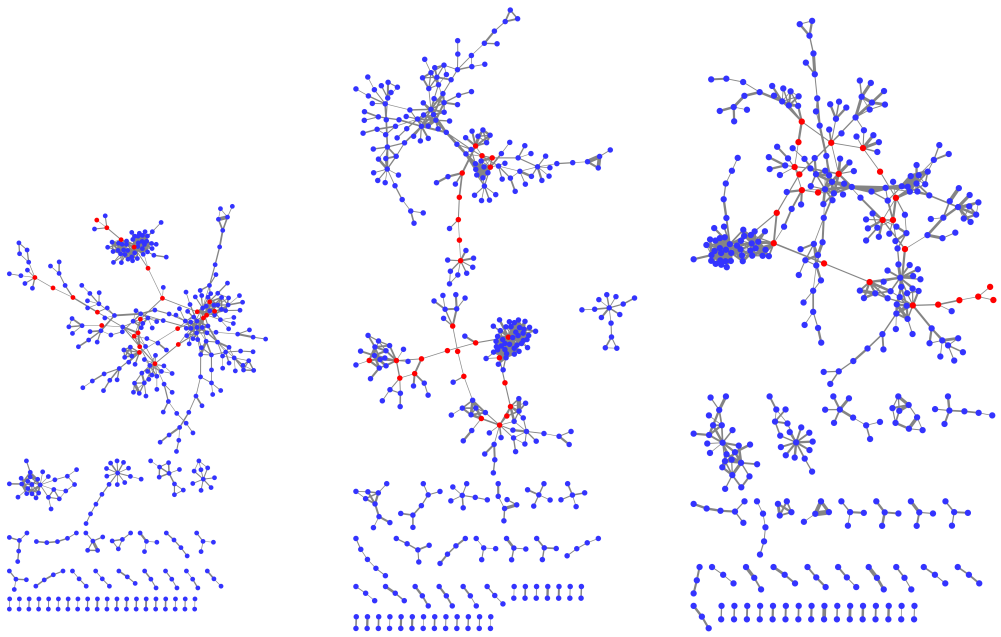


Figure 8.1: Tumor clustering. For each tumor, we produced a matrix from the correlation (Pearson) of the expression profiles among 760 genes. The correlations values were adjusted by the CLR algorithm. Then, we clustered the resulting matrices by euclidean metrics.

Table 8.1: Network Properties.

	BioPlex-Ontocancro	Cluster 1	Cluster 2	Cluster 3
Clustering Coefficient	0.24	0.20	0.19	0.17
Connected Components	24	41	42	41
Network Diameter	16	18	19	18
Avg Path Length	6.52	7.41	6.88	7.31
Avg Degree	3.84	3.2	3.14	2.98
Number of Nodes	511	406	408	410
Number of Edges	981	650	642	612

BioPlex-Ontocancro: BioPlex network considering only genes present in Ontocancro database. Cluster 1, 2, and 3: BioPlex-Ontocacro superimposed with the respective cluster correlation matrix.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Figure 8.2: Tumor cluster networks. The cluster networks were produced by superimposing each cluster gene-gene correlation matrix with the backbone network BioPlex Ontocancro. The red nodes are those defined as the most central in each cluster networks (Spectral Centrality $\geq 90^{\text{th}}$ percentile. Cluster 1: COAD and READ; Cluster 2: LUAD, LUSC, GBM, OV, BRCA, and UCEC; Cluster 3: LGG, KIRP, and KIRP.

gene-gene correlation matrix onto the backbone network BioPlex-Ontocancro. In the resulting networks, genes had links if they either correlated in the gene expression profiles and if their proteins had physical interaction annotations in the BioPlex network. We observed that the cluster networks presented approximately 80% of the nodes and 60% of the edges of the backbone BioPlex-Ontocancro network (Table 8.1, Figure 8.2).

We hypothesized that the most central genes in each network should play a fundamental role in the tumors represented in the cluster. To find the most central genes we measured the Spectral Centrality (SC) topological measure [U+2060], which evaluates changes in the network global diffusivity after node perturbation. We considered as the most central nodes those having SC above the 90^{th} percentile (25, 27 and 24 genes for clusters 1, 2, 3 respectively, Table 8.2 and Figure 8.2). The overlap between central genes in the cluster networks and in the backbone network

Table 8.2: Central genes

BioPlex-Ontocancro	Cluster 1	Cluster 2	Cluster 3
ALOX5 APP C17orf70 CCDC99 CETN2 CSNK2A1 CSNK2A2 EME1 ERCC1 ERCC4 ERCC6L FANCB GAB1 GRB2 H2AFX IL6R MAP4K5 MCM10 MLF1IP MUS81 NFKIA NRP1 PIK3CA PIK3CB PIK3CD PIK3R2 PIK3R3 PLK1 POLA1 POLA2 PRIM1 PRIM2 PSMB3 PSMC3 RAC1 SEC13 TNF TNFRSF1A TRAF6 UBB UBE2T XPA XPC	ALOX5 BTRC BUB1 CDC20 CENPC1 CHUK CUL1 MIS12 MLF1IP NDC80 NFKB1 NFKB2 NFKIA PMF1 PPP2CB PPP2R5D PSMB9 PSMC2 PSMF1 RAD21 REL RELB RPS27 SRC STAG1	BTRC CENPC1 CETN2 DSN1 ERCC1 ERCC4 FANCB FYN H2AFX IL6R MCM10 MIS12 MLF1IP NEDD1 NFKB1 NFKIA NUP43 PARP1 PLK1 PSMB3 PSMC3 RPA2 SRC TNFRSF10B TUBGCP5 TUBGCP6 XPA	AKT2 ALOX5 BAG4 CAPN1 CAPN2 CDC16 CDC27 CDT1 ENDOG FBXW11 FNTA GMNN KIF2B KIF2C LSP1 NEDD1 PRKACG PSMC3 PSMD9 SKP2 TNFRSF1A TUBGCP5 UBB VIM
	3/25	13/27	4/24

The table shows the genes that have high centrality (Spectral Centrality $\geq$ the 90^{th} percentile) in each network. The ratios show the proportion of genes that are also central in the BioPlex-Ontocancro network

BioPlex-Ontocancro is 3/25, 13/27, and 4/24 for clusters 1, 2, and 3 respectively. It shows that the importance of the genes in the BioPlex-Ontocancro network changed according to the gene expression profiles of each tumor cluster.

We defined the most central genes in each cluster network as the cluster gene signatures. The cluster 1 and cluster 2 signatures have 6 genes in common, cluster 1 and cluster 3 have 1 gene in common; and cluster 2 and cluster 3 have three genes in common. As we superimposed the gene expression profiles in the backbone network, some links were differentially removed across the cluster networks, which means that the same genes can have different set of interacting nodes in each network (Figure 8.3).

We observed that all signatures contain genes related to three biological categories: NF-κB signaling pathway, chromosomal instability and ubiquitin-proteasome system (Table 8.3). The chromosomal instability category relates to genes involved in kinetochore formation, microtubule dynamics and chromosome segregation functions. All signatures have at least one substrate recognition component of E3 ubiquitin ligase complexes: BTRC in clusters 1 and 2; and FBXW11 in cluster 3.

Cluster 1 has genes involved in spindle checkpoint (BUB1, CDC20). The cluster 2 signature has many genes related to DNA repair (CETN2, FANCB, H2AFX, ERCC1, ERCC4, PARP1, XPA) and DNA replication (RPA2, MCM10). Moreover, it has three important genes in the signaling path that activates the STAT3 transcription factor: SRC, NFKB1 and IL6R. Indeed, the STAT3 gene expression levels are significantly higher in cluster 2 (ANOVA p-value: 5.58 x 10^{-15}) both in comparison with cluster 1 (T-Test p-value: 1.08 x 10^{-9}) and cluster 3 (T-Test p-value: 1.14 x 10^{-8}) patients (Figure 8.4). The cluster 3 signature contains genes involved in three different apoptotic mechanisms: induced by TNF-α (TNFRSF1A and BAG4), induced by Endoplasmatic Reticulum stress (CAPN1 and CAPN2) and caspase-independent apoptosis (ENDOG).

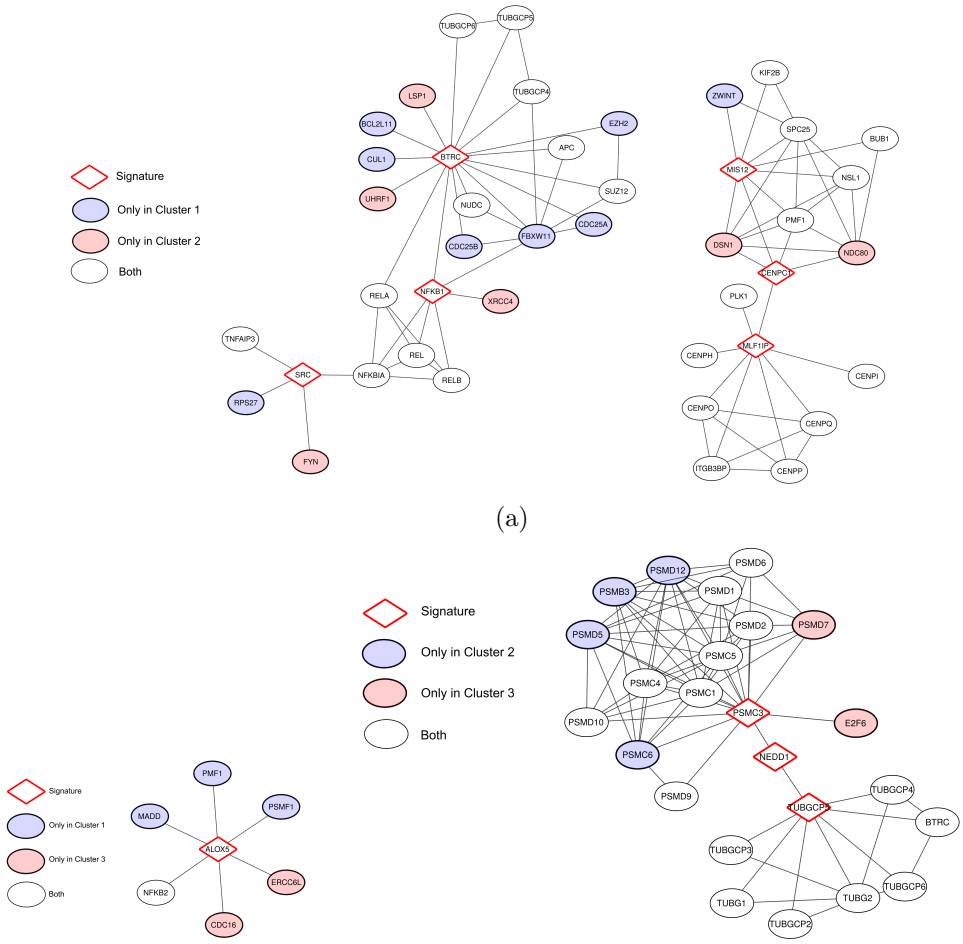


Figure 8.3: Networks showing first neighbors of signature genes that are shared by different tumor clusters. Figure a: signature genes common to cluster 1 and cluster 2; Figure b: signature genes common to cluster 1 and cluster 3; and Figure c: signature genes common to cluster 2 and cluster 3.

Table 8.3: Common biological categories present in the gene signatures

	NFKB signaling	Chromosomal Instability	Proteasome
1	BTRC, CUL1, SRC, NFKBIA, NFKB1, NFKB2, REL, RELB, CHUK	CDC20, BUB1, MLFPIP, CENPC1, MIS12, PMF1, NDC80, RAD21, STAG1	PSMB9, PSMC2, PSMF1
2	BTRC, SRC, NFKBIA, TNFRS10B, IL6R	MIS12, DSN1, MLFPIP, CENPC1, PLK1, NEDD1, TUBGCP5, TUBGCP6	PSMB3, PSMC3
3	FBXW11, AKT2, TNFR1A	CDC16, CDC27, NEDD1, TUBGCP5, KIF2B, KIF2C	PSMC3, PSMD9

All cluster signatures have genes that can be grouped in the following categories: NF-κB signaling, chromosomal instability and ubiquitin-proteasome system.

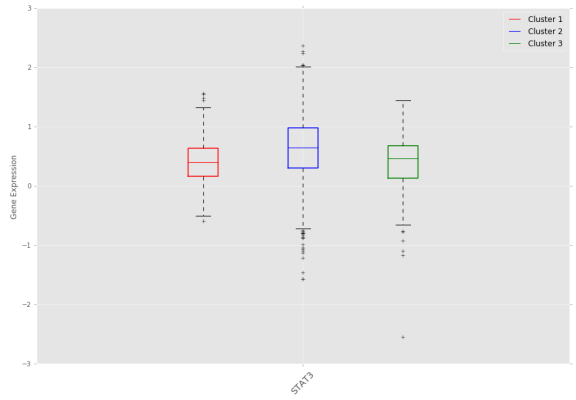


Figure 8.4: Cluster 2 patients presented higher STAT3 gene expression in comparison with cluster 1 (T-Test p-value: 1.08 x 10^{-9}) and cluster 3 (T-Test p-value: 1.14 x 10^{-8}).

Table 8.4: List of genes from the signatures that are also being tested in ongoing clinical trials studies (according ClincalTrials.gov).

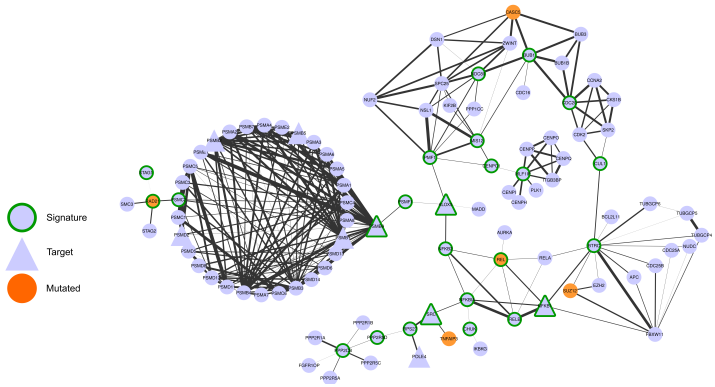
Inhibition target	Number of clinical trials	Cluster signature
ALOX5	18	1, 3
CHUK	9	1
FYN	97	2
IL6R	2	2
NFKB1	40	1, 2
NKFB2	8	1
NKFBIA	8	1, 2
PARP1	106	2
PPP2CB	5	1
PSMB9	25	1
SRC	135	1, 2

8.3 Validation of the multi-tumor gene signatures

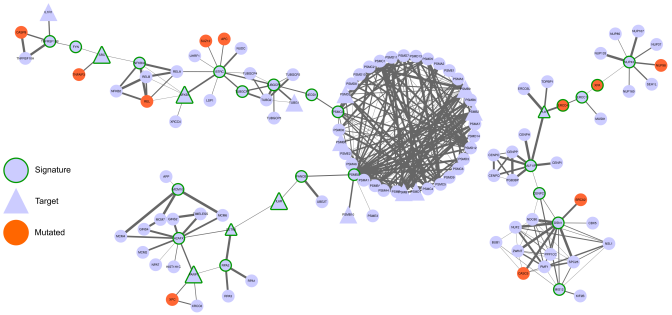
Then, we searched for possible relationships between the gene signatures and genes commonly mutated in the studied tumors. We observed that some signature genes also presented somatic mutations (REL and RAD21 in cluster 1, ERCC4 and XPA in cluster 2, and AKT2 in cluster 3) or that mutated genes were direct neighbors of the signature genes in the network (see Figures 2, 3, 4). A permutation test over the signature labels (see Methods) reveals a significant proximity of signature genes to mutated genes for cluster 1 and cluster 2 (p-value = 8.76 x 10^{-4} and p-value = 6.9 x 10^{-3} respectively) (Figure 8.6). For the particular case of cluster 3, only one mutated gene is present in the network and it is successfully selected as a signature gene.

Since the signature genes are the most central nodes in each cluster, we hypothesized that they might be suitable drug targets. For this purpose we collected, from the Drug Bank database, the drugs that target genes in the signatures and we evaluated in the ClinicalTrials repository if these drugs are under ongoing clinical trials for cancer treatment. We observed that 11 genes from the cluster signatures are being tested: 4 and 3 genes, from cluster 1 and 2, respectively; 3 genes from both cluster 1 and 2; and 1 gene from both cluster 1 and 3 (Table 8.4).

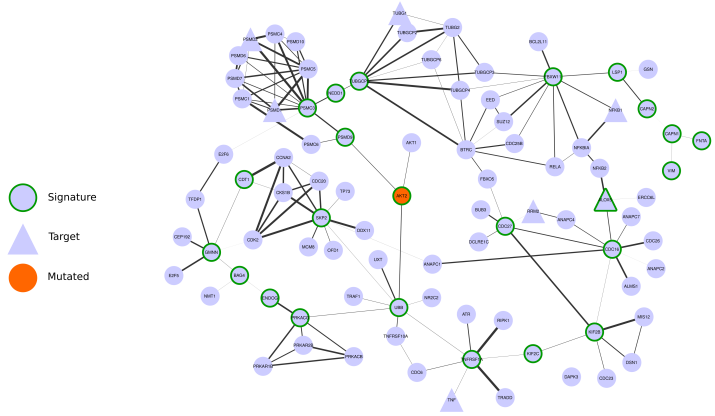
We then asked whether the expression level of the signature genes could predict the patients survival in each cluster, independently of the tumor type. First, we grouped the tumor cluster patients assuming the existence of two groups (one with



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

Figure 8.5: Network composed by the first neighbors of the cluster 1 (a), cluster 2 (b) and cluster 3 (c) signature genes.

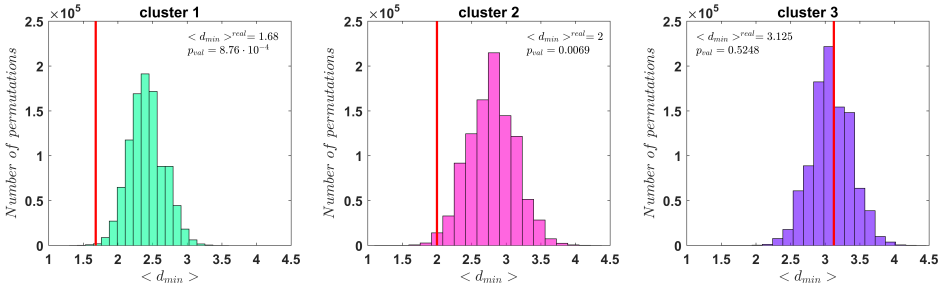


Figure 8.6: Plot of the distribution of the 10^6 permutations for the 3 clusters (from left to right). The insets show the minimum average distances for the signatures (represented in the plots as red vertical lines), and the p-values according to the permutations.

Table 8.5: Combination Indexes for BI6727 and Bortezomib at different concentrations in MCF-7 cell line.

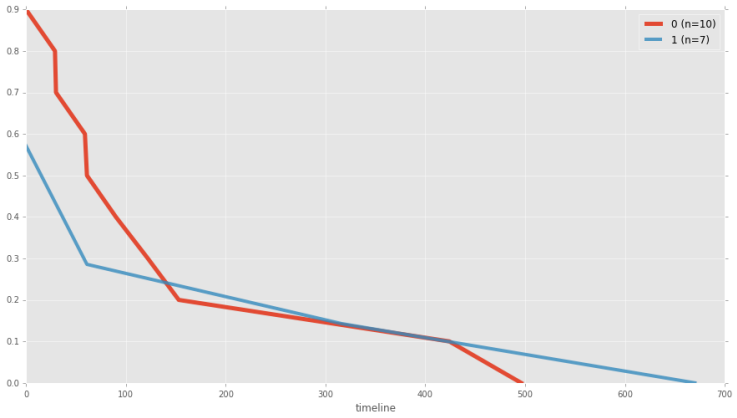
		BI6727 [mM]				
		17.3	**34.5**	**69.2**	**138.4**	**276.8**
Bortezomib [mM]	**0.15**	1.39	1.24	1.16	0.73	1.7
	0.3	1.85	1.19	1.32	0.87	1.57
	0.6	1.22	1.18	1.18	1.38	1.2
	1.2	0.96	1	0.96	0.96	0.75
	2.4	1.09	1.09	1.09	0.91	0.91

good and another with bad survival outcome) and we calculated the Kaplan-Meier survival curves for both groups. The TCGA data portal had clinical information for 17, 448, and 32 patients from the tumor clusters 1, 2, and 3, respectively. The analysis resulted in non-significantly different survival curves for the cluster 1 and 3 (Log-rank test p-values: 0.91 and 0.90, respectively), possibly because we did not had enough clinical information (Figures 8.7a, 8.7c). For the cluster 2, the gene signature significantly separated the patients in two groups according to good or bad survival outcome (Log-rank test p-value = 4.54 x 10^{-3}, Figure 8.7b).

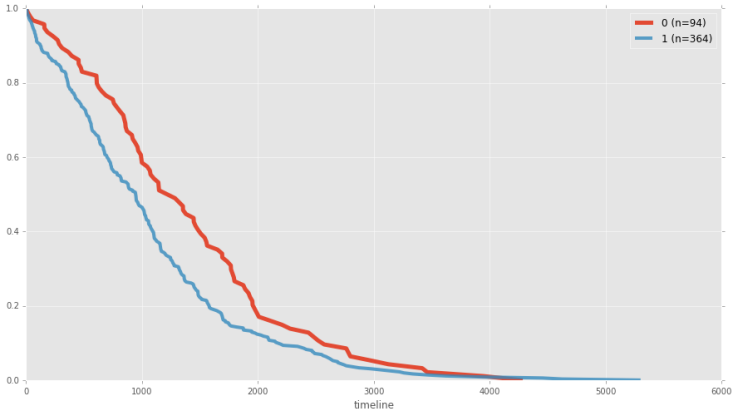
We tested if we could propose a new therapeutic strategy by inhibiting the genes from the cluster signatures. As the tumor cluster 2 contained the largest and most heterogeneous set of tumors, we designed *in vitro* assays to test the inhibition of two genes derived from the cluster 2 signature and the inhibition of one gene that is related to the same biological function as those in the signature. By performing the experiments in this way, we tested if our approach selected the specific genes important for the tumors in the clusters or if it rather selected genes related to the predominant biological categories acting on the tumors. We selected three drugs: i) Bortezomib, to target the gene PSMB3, which is related to the proteasome and NF-κB pathways; ii) BI6727, to target the gene PLK1, related to chromosomal instability; and iii) the PF-00477736 drug, to target the CHK1/2 genes, which is not in the cluster 2 signature but is also related to DNA damage response. We tested these drugs, alone or in combination, in two cancer models: the glioblastoma cell line T98G and the breast adenocarcinoma model MCF-7.

Both cell lines were highly sensitive to Bortezomib, with an IC50 of 200 nM for MCF-7 and 0.6 nM for T98G (Figure 8.8). BI6727 treatment reduced viability in a concentration-dependent manner in both models, with the glioblastoma model showing increased responsiveness (IC50 of 69.2 nM versus 1.8 μM for MCF-7, Figure 8.8). Moreover, both cell lines showed low response to CHK1/2 inhibition, with IC50 of 26.9 μM for MCF-7 and 15.1 μM for T98G (Figure 8.8).

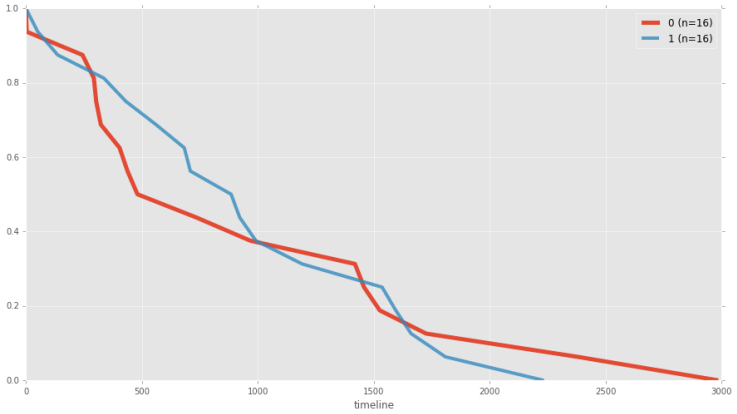
We then asked whether these drugs may synergize in the selected models. Although the combinations of PF-00477736 with either BI6727 or Bortezomib did not show any additive or synergistic effect in MCF-7 and T98G cell lines (data not shown), we observed a cooperation effect between inhibition of PLK1 and proteasome activity (Figure 8.9). Indeed, after the treatment with the drug combination, cell viability was significantly lower compared with single agent treatments in MCF-7 cells (Figure 8.9a, p-value < 0.05), showing a general additive effect (Table 8.5). We observed low Combination Index values (< 1) for both cell lines, indicating synergistic effect for all concentrations tested in the breast cancer model, and for selected concentrations in the glioblastoma model (Figure 8.9, Tables 8.5 and 8.6).



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

Figure 8.7: Kaplan-Meier curves for the two groups of tumor cluster patients defined by the K-means clustering approach (see Methods). Log-rank test p-values: 0.91, 4.54 x 10^{-3}, and 0.90 for the cluster 1 (a), 2 (b), and 3 (c), respectively.

Table 8.6: Combination Indexes for BI6727 and Bortezomib at different concentrations in T98G cell line.

		BI6727 [μM]				
		0.45	**0.9**	**1.8**	**3.6**	**7.2**
Bortezomib [μM]	**0.05**	0.52	0.46	0.61	0.63	0.53
	0.1	0.66	0.51	0.61	0.7	0.48
	0.2	0.55	0.62	0.75	0.67	0.48
	0.4	0.81	0.45	0.67	0.52	0.39
	0.8	0.58	0.69	0.49	0.43	0.35

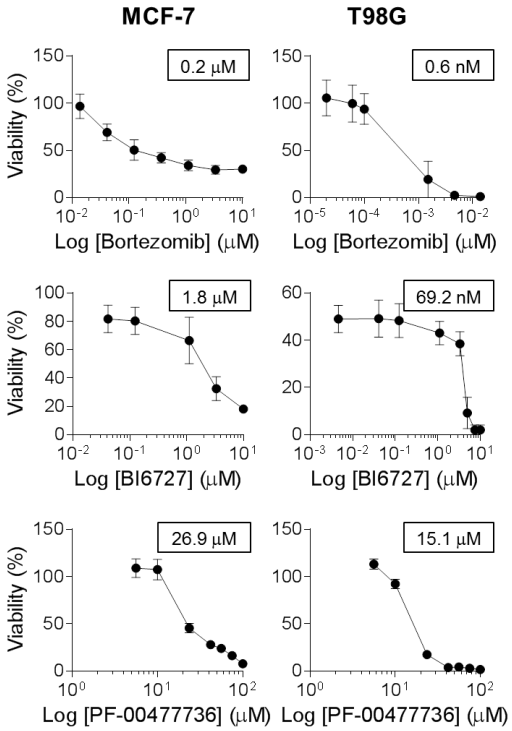
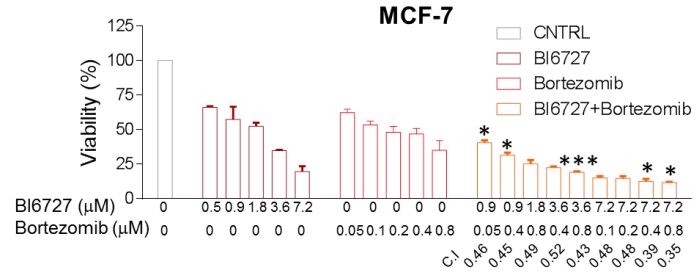
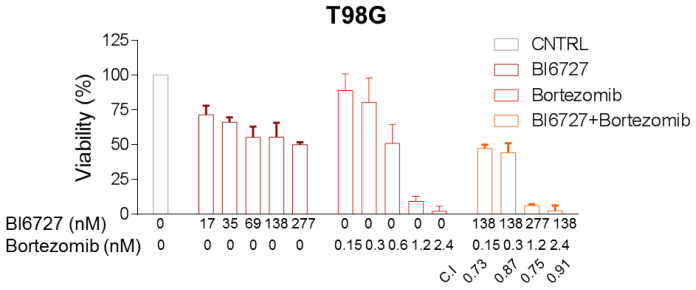


Figure 8.8: *In vitro* response of cancer cell lines from signature 2 to treatment with Bortezomib, BI6727 and PF-00477736 as single agent. MCF-7 and T98G cells were treated with increasing doses of Bortezomib (0.01 to 10 μM for MCF-7, 0.02 to 10 nM for T98G), BI6727 (0.04 to 10 μM for MCF-7, 0.004 to 10 μM for T98G), PF-00477736 (5.6 to 100 μM) and cell viability was measured 72h after drug administration by WST-1 assay (three independent experiments). Cell viability is represented as (mean ± SEM). IC50 values are reported in the inset boxes (GraphPad Prism 6).



(a)



(b)

Figure 8.9: Sensitivity of MCF-7 and T98G cells to combined inhibition of PLK1 and the proteasome. MCF-7 and T98G cells were treated with increasing doses of Bortezomib (0.05 to 0.8 μM for MCF-7, 0.15 to 2.4 nM for T98G) and BI6727 (0.5 to 7.2 μM for MCF-7, 17 to 277 nM for T98G), alone or in combination and cell viability was measured 72h after drug administration by WST-1 assay (three independent experiments). Statistical significance was determined by Student's t test (*, $P < 0.05$; ***, $P < 0.001$). Combination index (C.I.) was calculated by CompuSyn software. (a) MCF-7 cells: combinations with a C.I. lower than 0.5 are shown. (b) T98G cells: combinations showing synergistic effect are shown.

Chapter 9

Discussion

We studied the expression profiles of 11 tumors by considering a selected set of genes from the Ontocancro database and the BioPlex protein-protein interaction network. This knowledge-based selection reduced the dimensionality of the data to a highly curated list of cancer-related genes, involved in pathways that are hallmarks of cancer as cell cycle, inflammation, and apoptosis [11] [U+2060]. This approach also ensured that all studied genes had protein-protein interaction annotations, which are crucial to the understanding of how the signaling transduction propagates in the cell [67].

Tumor Clustering

We clustered tumors by their gene-gene relationships, defined by the Pearson's correlation coefficients, to evaluate the functional relationships between genes and their impact on transcriptome organization [68, 69] [U+2060]. Tumors from the same organ tended to group together, in agreement with previous studies, showing that tissue-of-origin features provide the dominant signals in the identification of cancer subtypes [49, 70] [U+2060] (Figure 8.1). However, the clustering also grouped tumors originated from different tissues, according to similarities in genomic alterations, as in the case of BRCA, OV, LUSC, and UCEC, which share common characteristics as presence of TP53 mutations and multiple recurrent chromosomal gains and losses [51] [U+2060]. In particular, BRCA and UCEC presented the best prognosis when compared to other 10 tumor types in a previous study [49], and we our results they grouped into a well defined sub-cluster. Also interestingly was the fact that Glioblastoma Multiforme (GBM) and Brain Lower Grade Glioma (LGG) clustered in two different groups, showing that tumors from the same tissue of origin may activate different pathways and processes.

Gene Signatures - Common Biological Processes

Network analyses permitted us to integrate different types of biological information, identify functional modules and rank genes according to network properties [71, 72]. Here, we created a network for the tumor clusters by combining a backbone network (BioPlex-Ontocancro) and cluster-specific gene expression correlation profiles. Then, we defined gene signatures based on node ranking by centrality measures, which presented genes mainly involved in three biological processes: NF-κB

signaling, chromosomal instability and the ubiquitin-proteasome system (Table 8.3).

The NF-κB signaling pathway regulates genes that participate in cell proliferation, innate and adaptive immune responses, inflammation, cell migration, and apoptosis regulation processes. The aberrant activity of NF-κB may act as survival factor for transformed cells which would otherwise become senescent or apoptotic [73][U+2060].

The genes classified into the chromosomal instability category involve kinetochore formation, microtubule dynamics and chromosome segregation functions. The dysfunction of these genes may cause cell inability to faithfully segregate chromosomes, generating genomic alterations as DNA mutations, chromosomal translocations, and gene amplification. The mutant genotypes may confer beneficial phenotypic traits to cancer cells, such as sustained proliferative signaling and resistance to cell death [11][U+2060]. Two genes classified into this category have already been related to clinical practice: the prognostic marker KIF2C [74, 75][U+2060]; and the BUB1 gene, which expression correlates with poor clinical diagnosis [76, 77][U+2060].

The ubiquitin-proteasome system is the major degradation machinery that controls the abundance of critical regulatory proteins. Perturbation of the regulatory proteins turnover disturbs the intricate balance of signaling pathways and the cellular homeostasis, contributing to the multi-step process of malignant transformation [78][U+2060]. Proteasome inhibitors have become valuable tools in the treatment of certain types of cancer, mainly because malignant cells show greater sensitivity to the cytotoxic effects of proteasome inhibition than non-cancer cells [79][U+2060].

Gene Signatures - Specific Biological Processes

In addition to common features, cluster 2 signature has several genes related to DNA repair (CETN2, FANCB, H2AFX, ERCC1, ERCC4, PARP1, XPA) and DNA replication (RPA2, MCM10). Interestingly, the tumors in this cluster usually present high rates (50% to 90%) of samples with mutated TP53, which is an important sensor for the cell DNA damage response [44, 46, 49][U+2060]. The cluster 2 signature also presents the genes SRC, NFKB, and IL6R, which participates in the activation of STAT3, a transcription factor that is necessary for cell transformation [80][U+2060]. We observed that STAT3 gene expression is higher in the tumors of cluster 2 when compared with the tumors of clusters 1 and 3 (Anova p-value: 5.58 x 10^{-15}) (Figure 8.4). The cluster 3 signature has genes involved in three apoptotic mechanisms: induced by TNF-α (TNFRSF1A and BAG4), Endoplasmatic Reticulum stress (CAPN1 and CAPN2) and caspase-independent apoptosis (ENDOG). As the regulation of cell death serves as a natural barrier to cancer development, these processes may reflect different strategies that these tumors use in response to various cellular stresses.

Somatic Mutations and Gene Signatures

Since the transcriptional disturbances observed in cancer can sometimes be explained by underlying somatic mutations [81][U+2060] we retrieved TCGA mutational data, and focused on cancer related mutations reported in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Many signature genes re-

sulted also somatically mutated, or first neighbors to mutated genes (Figures 8.5, 8.6), showing their strict relationship and the functional relevance of the biologically processes they are involved in.

Pharmacological and Clinical Evidences

Several genes in the signatures or in their direct network neighborhood are already under clinical investigation in a variety of tumor conditions (as annotated in Clinicaltrials.org database) (Table 8.4). For example, the AKT pathway has been described as a potential drug intervention in clear cell renal carcinoma [82] [U+2060]. Our results confirm its potential for drug treatment, since this gene belongs to the signature of cluster 3 (comprising LGG, KIRC, and KIRP), it is somatically mutated in the tumors of cluster 3 and it has been annotated as drug-target according to the Drug Bank database.

We also asked whether the gene signatures could predict survival outcomes in each cluster, i.e. independently of tumor type. Our results show that in cluster 2 (the only one with enough available samples) the gene signature defined two groups of patients with significantly different Kaplan-Meier survival curves (log-rank test p-value: 4.54 x 10^{-3}) (Figure 8.7b).

***In vitro* experiments**

We tested 3 existing drugs (targeting 2 genes belonging to cluster 2 signature, and 1 involved in a related biological process) on 2 tumor types from the cluster 2, T98G and MCF-7 cell line models (Figures 8.8, 8.9). The CHK1/2 inhibitor (PF-00477736) had poor effect on both cell lines. The PLK and proteasome inhibitors (BI6727 and Bortezomib, respectively) showed a high effect on both cancer cell models, showing a significant synergic action at several dosages and suggesting a novel therapeutic strategy to be further explored for the treatment of cluster 2 tumors.

Conclusions

These observations indicate that our study succeeded in: 1) clustering tumors and highlighting common functional mechanisms related to their transcriptional profile; and 2) selecting genes with a relevant functional role in the studied tumors. The combination of these results may provide the rationale for choosing novel drug targets, drug combinations, and for the design of new drug repurposing strategies. As future perspectives, we believe that the investigation of transcriptional and mutational profiles of single patients, combined with the information provided by our gene signatures, might suggest strategies for personalized therapy approaches.

Part III
Searching for the molecular mechanisms of tumor progression in thyroid cancer by gene expression data analysis

Chapter 10

Introduction

The incidence of thyroid cancers has rapidly increased over the past 30 years [86]. The most prevalent, the Papillary Thyroid Carcinoma (PTC), accounts for up to 80%-85% of the cases and presents 5-year survival rates over 95% [87]. The PTC derives from follicular thyroid cells and are designated as well-differentiated, in contrast with the poorly differentiated (PDTC) and undifferentiated Anaplastic Thyorid Carcinoma (ATC). The ATC comprises the minority of thyroid cancer cases (2%-3%), but, due to its high aggressiveness, is responsible for up to 40% of thyroid cancer related deaths. ATC is not sensitive to radiation and chemotherapy treatments, and, in most cases, patients do not survive more than a year after the cancer diagnosis [88]. The understanding of the molecular pathogenesis of thyroid cancers and the mechanisms leading to the loss of differentiation in the most aggressive carcinomas are crucial to the development of more effective treatment strategies.

PTCs are characterized by high frequency (70%) of activating somatic alterations that deregulate the mitogen-activated protein kinase (MAPK) signaling pathway, as point mutations in BRAF and RAS genes; and gene fusions involving RET and NTRK1 genes [87, 89]. It has been observed that ATCs also present frequent somatic mutations in BRAF and RAS genes, supporting the hypothesis that ATC and PTC share the same tumorigenic origin, being the undifferentiated carcinomas originated from the previous well-differentiated forms [90, 91]. Aggressive recurrent PTC and ATC present the coexistence of multiple genetic alterations that are otherwise mutually exclusive in the well-differentiated forms, indicating that the thyroid cancer progression may occur through the accumulation of multiple somatic alterations that cooperate to amplify oncogenicity [87].

In this study, we aimed to characterize the thyroid transcriptional profiles of normal, PTC, and ATC samples in order to investigate the molecular mechanisms involved in tumor progression. We retrieved gene expression arrays of 50 ATC, 102 PTC, and 127 normal samples available in the GEO database. We characterized the differentially regulated genes considering the PTC vs normal, ATC vs normal, and PTC vs ATC comparisons. Then, we separated genes that resulted as differentially expressed in all comparisons according to two general trends: those having increasing or decreasing expression levels across the different tumor phases. From these lists, we defined signatures representing the most deregulated genes in each transition: normal to PTC and PTC to ATC. Our results support the hypothesis that ATC represent a progression of the PTC forms. The normal to PTC transition involves the activation of genes belonging to pathways related to cellular morphology and

extracellular matrix; while the PTC to ATC transition involves the activation of genes related to cell cycle control. We evaluated the relevance of the signatures by mapping the genes onto protein-protein and transcriptional regulatory networks. Our results highlight new thyroid cancer genes, providing a list of potential markers for cancer prognosis and targeted-therapy strategies.

Chapter 11

Material and Methods

11.1 Data and Processing

In this study, we analyzed gene expression arrays of the Human Genome U133 Plus 2.0 (Affymetrix) platform retrieved from The National Center for Biotechnology Information Gene Expression Omnibus (GEO) database. The dataset contained ATC (n=50), PTC (n=102) and normal thyroid (n=127) tissue samples (Table 11.1).

We performed background correction, quantile normalization and expression calculation using the Robust Multichip Average method implemented in the R/Bioconductor Affy package. After data normalization, samples were clustered in two different approaches: i) first by applying the hierarchical clustering method (correlation distance, single linkage method) based on the 150 probes with highest variance across the entire dataset; and ii) by performing Principal Component Analysis based on the expression levels of all probes.

11.2 Differential Expression Analysis

For each comparison between tissue types (ATC vs PTC, ATC vs normal, and PTC vs normal), the differentially expressed genes were detected using Student's T test followed by false discovery rate (FDR) multi-test correction with the Benjamini and Hochberg's method. The differentially expressed genes (adjusted p-value $<$ 0.05) were characterized by: i) Gene Ontologies (GO) enrichment analysis, using the R/Bioconductor ClusterProfiler package [92]; and ii) Reactome pathways enrichment

Table 11.1: Accession numbers for the thyroid gene expression profiles used in this study

GEO Accession Number	ATC	Well differentiated PTC	Normal
GSE76039	18	0	0
GSE65144	12	0	13
GSE33630	11	0	0
GSE29265	9	20	20
GSE3467	0	9	9
GSE35570	0	32	51
GSE60542	0	27	25
GSE3678	0	7	7
GSE53157	0	7	2
Total	50	102	127

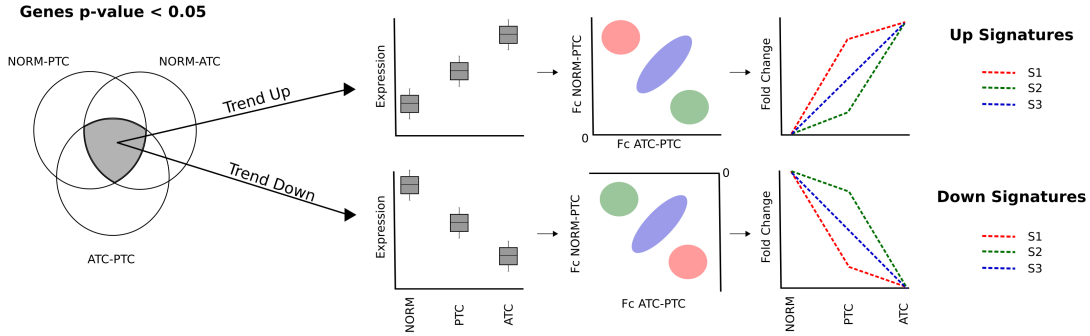


Figure 11.1: Figure showing the steps in the definition of the gene signatures. The genes resulting as significant in all comparisons were divided according to the trends of their expression levels: Trend Up or Trend Down. Subsequently, they were divided into signatures S1, S2 and S3, according to their fold change values in the normal-PTC and ATC-PTC comparisons.

analysis, using the R/Bioconductor ReactomePA package [93].

11.3 Signatures of Tumor Progression

We selected the genes that resulted as significant in all comparisons, and calculated their average expression in each tissue type dataset: normal, PTC, and ATC. Then, a gene g was assigned to the Trend Up list if $\bar{g}_{\text{normal}} < \bar{g}_{\text{PTC}} < \bar{g}_{\text{ATC}}$; and to the Trend Down list if $\bar{g}_{\text{normal}} > \bar{g}_{\text{PTC}} > \bar{g}_{\text{ATC}}$ (Figure 11.1).

We standardized the fold change values (Z) of the Trend Up and Down lists from the PTC-normal and ATC-PTC comparisons. Gene signatures were defined according the absolute values of Z as: signatures S1, the genes having the expression more affected in the PTC-normal transition ($Z_{\text{PTC-NORM}} > 4$ and $Z_{\text{ATC-PTC}} < 2$); signatures S2, genes having the expression more affected in the ATC-PTC transition ($Z_{\text{PTC-NORM}} < 2$ and $Z_{\text{ATC-PTC}} > 4$); and signatures S3, genes having the expression equally affected in both comparisons ($Z_{\text{PTC-NORM}} > 2$ and $Z_{\text{ATC-PTC}} > 4$). The figure 11.1 shows the steps in the definition of the Trend Up and Trend Down genes signatures.

11.4 Analysis of gene signatures in biological networks

We evaluated the genes in the signatures through the analysis of protein-protein interaction (PPI) and transcriptional regulatory networks. The PPI network was built from the interactions of the signature genes and their first neighbors in the BioPlex network [55]. The number of nodes in the networks built from the Up and Down signatures, were, respectively, 386 and 393; while the number of edges were 762 and 987. Then, the signature genes were ranked according to the network topological measures degree and Betweenness Centrality (BC).

We retrieved associations between transcription factors and the genes they activate, from the TRRUST network [94]. Then, for each Up signature, we created a regulatory network built from the signature genes and their incoming edges and nodes. The networks of the Up S1, S2, and S3 signatures presented, respectively, 54, 32, and 57 nodes; and 58, 28, and 54 edges. The network analyses were performed by using Cytoscape, Networkx (Python) and Igraph (R).

Chapter 12

Results

12.1 Clustering Gene Expression Profiles

We analyzed 279 gene expression arrays comprising normal thyroid, papillary thyroid carcinoma (PTC) and anaplastic thyorid cancer (ATC) samples. After data normalization and standardization, we clustered patients according to the expression levels of the 150 probes with highest variance across the entire dataset. We observed that patients did not grouped according to study of origin, supporting the absence of strong batch effects and showing that the assembled dataset was suitable for evaluating the differences and similarities among tissue types. After applying a Principal Component Analysis (PCA) to the entire dataset, the results showed a gradational trend separating normal, PTC and ATC samples along the first component, being the distinction of ATC samples the most well defined (Figure 12.2). The 14 points in the top of the PCA plot represented all samples from the same study (GSE3678) and they were removed from the dataset in the downstream analyses.

12.2 Differential Expression Analysis

We applied the Student's T-Test, followed by False Discovery Rate (BH) multitest correction method, to identify differentially expressed probes among conditions: ATC vs PTC, ATC vs normal, PTC vs normal (Figure 12.3). The number of differentially expressed probes in the PTC vs normal comparison was considerably lower than the ATC vs normal and ATC vs PTC comparisons, confirming a pattern also observed in the PCA results: the PTC expression profiles are more similar to normal rather than to ATC samples.

In order to provide a biological characterization of the differentially expressed genes from the results of each comparison, we performed Gene Ontology and Reactome pathway enrichment analysis. In the ATC vs PTC results, most of the highest enriched GO terms (biological process) and Reactome pathways were related to cell cycle control (Figures 12.4 and 12.5). In the ATC-normal and PTC-normal comparisons, the highly enriched GO terms and Reactome pathways were related to cell morphology and tissue organization (Figures 12.4 and 12.5). Interestingly, in the Reactome enrichment analysis for the differentially expressed genes in the ATC-normal comparison, the results showed enriched pathways related to cell cycle, cell morphology, and tissue organization (Figure 12.5).

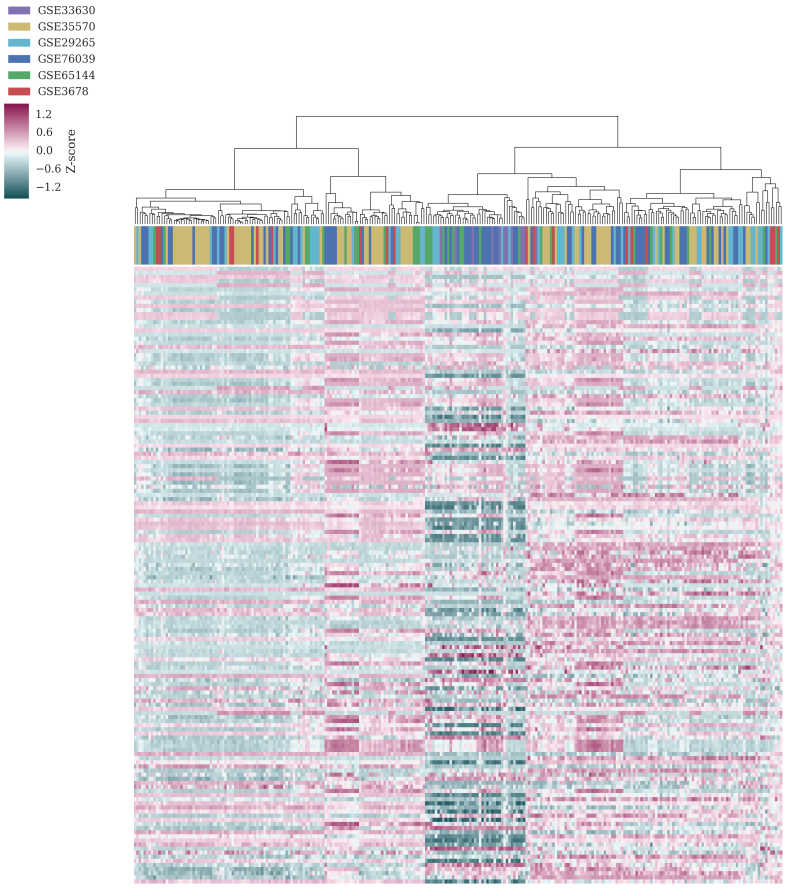


Figure 12.1: Heatmap showing the normalized expression levels of the 150 probes with highest variance across the entire dataset. Columns represent samples and rows represent single probes. The columns are labeled according to the study of origin (GEO accession number) of each sample. The dendrogram shows the results of a Hierarchical Clustering applied to the list of samples using the correlation distance and the single linkage method.

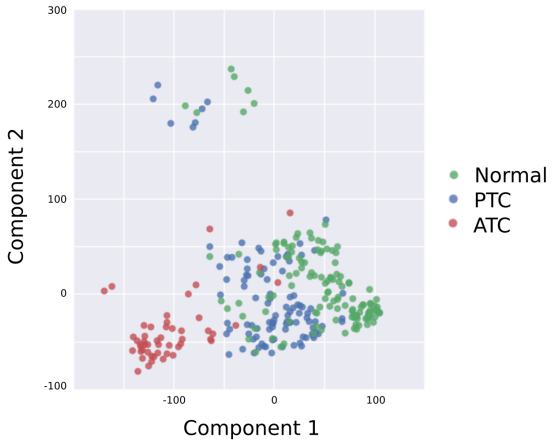


Figure 12.2: Principal Component Analysis applied to the entire gene expression dataset. The first component shows a gradational trend separating ATC, PTC and normal samples.

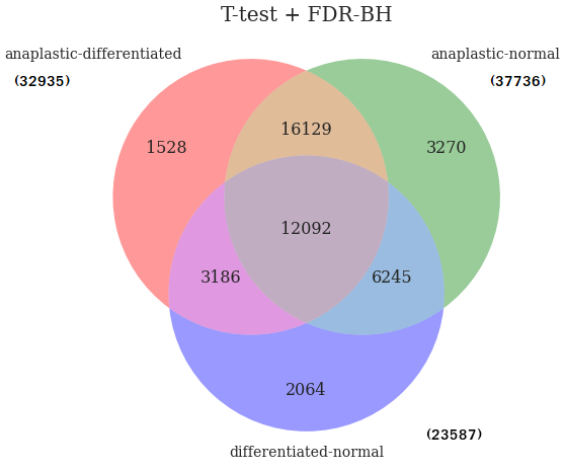


Figure 12.3: Venn diagram showing the number of probes resulting differentially expressed (adjusted p-value < 0.05) in more than one comparison. The numbers inside parenthesis show the total number of significant probes in each comparison.

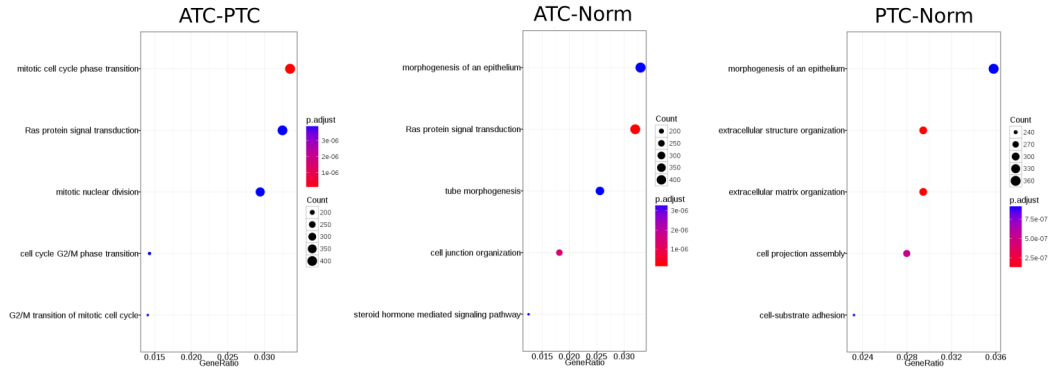


Figure 12.4: Gene Ontology enrichment analysis for genes differentially expressed in the ATC vs PTC (left), ATC vs normal (center), and PTC vs normal (right) comparisons. The figure shows the top 5 enrichments for biological process terms.

These results indicate that the transition from the normal tissue to the well differentiated PTC involves the deregulation of processes related to cell morphology and tissue organization, while the transition from PTC to ATC involves deregulating processes controlling cell cycle and cell replication. The fact that the significant genes from the ATC vs normal comparison presented pathways also enriched in the PTC-normal and ATC-PTC comparisons supports the idea of a common tumorigenic origin for PTC and ATCs.

12.3 Trends

Then, we aimed to investigate the genes most deregulated during the progression from normal to PTC, and from PTC to ATC tissue types. To approach this problem, we first defined two lists, Trend Up and Trend Down, composed by the genes that resulted as differentially expressed in all comparisons and that presented a increasing (or decreasing) trend in their expression levels when assuming the normal-PTC-ATC as the progression steps of tumorigenesis (Figure 11.1). The Trend Up

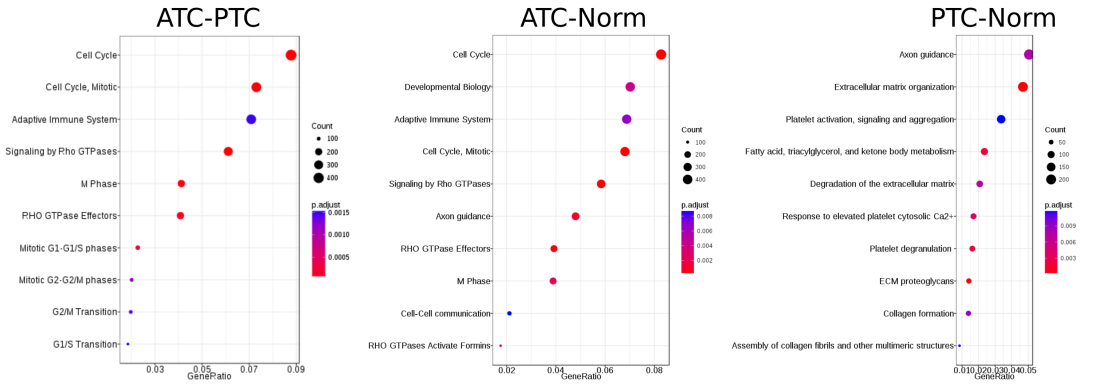


Figure 12.5: Reactome pathway enrichment analysis. The figure shows the top 10 enriched pathways (Reactome) for genes differentially expressed in the ATC vs PTC (left), ATC vs normal (center), and PTC vs normal (right) comparisons.

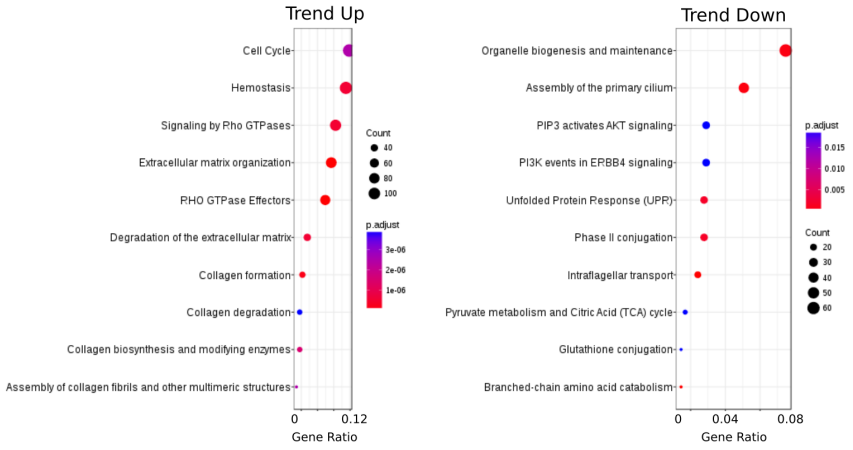


Figure 12.6: Reactome pathway enrichment analysis. The figure shows the top 10 enriched pathways (Reactome) for genes in the Trend Up and Trend Down lists.

and Trend Down lists presented 2,571 and 3,108 genes, respectively.

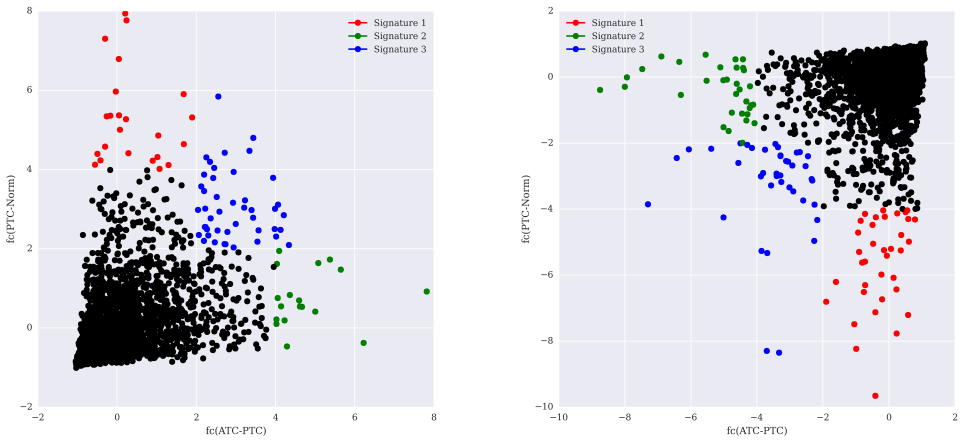
Reactome pathway enrichment analysis (Figure 12.6) showed that genes having their expression levels increased in the tumor progression (Trend Up) are mostly related to cell cycle and extracellular matrix, while those having their expression levels decreased in the tumor progression (Trend Down) are related mainly to signaling and metabolism pathways. We observed that the enriched pathways in the Trend Up genes were very concordant to those enriched in the whole list of differentially expressed genes from each comparison (Table 12.1). These results indicate that the processes leading to gene activation might be more important in the tumor progression than those related to gene repression.

We performed a further separation of genes in the Trend Up and Trend Down lists based on their absolute fold change values in the PTC vs Normal and ATC vs PTC comparisons. Genes with high values in the PTC-Normal and low values in ATC-PTC were defined as S1, those with the inverse pattern were defined as S2, and those equally high in both comparisons were defined as S3 (Figure 12.7, Table 12.2). The S1, S2, and S3 signature reflect the relative importance of the genes in each transition: S1 are those strongly deregulated in the normal to PTC transition;

Table 12.1: Common enriched pathways in the Trend Up and Trend Down lists compared to the differentially expressed genes in each comparison

	ATC-PTC (92)	ATC-Norm (53)	PTC-Norm (31)
Trend Up (70)	23	20	14
Trend Down (29)	5	3	5

The numbers in parenthesis represent the total number of significantly enriched pathways considering the list of differentially expressed genes from each comparison (ATC vs PTC, ATC vs normal, and PTC vs normal), and considering the genes in the Trend Up and Trend Down lists.



(a) Trend Up

(b) Trend Down

Figure 12.7: We defined three signatures for each list of Trend Up and Trend Down gene lists. The axes show standardized fold change in the comparisons PTC-Norm (Y axis) and ATC-PTC (X axis).

S2 are those strongly deregulated in the PTC to ATC transition; and S3 are those strongly deregulated in both steps (Figure 11.1).

We evaluated the genes in the signatures based on a protein-protein interaction network built from their interactions in the BioPlex network (See methods). The Up signature genes ICAM1 (S1), ASPM (S2) and PLAUR (S3); and the Down signature genes FAM167A (S1), SNRPN (S2) and FGFR2 (S3); presented high degree and centrality (Tables 12.3, 12.4). This suggests that they may have an important role in the cell signaling transduction and, consequently, might be good candidates for targeted-therapy strategies.

We also built transcriptional regulatory networks (See methods) in order to evaluate which are the regulators that activate the genes in the Up signatures. We observed that four transcription factors that activate the signature gene ICAM1 (S1) were up-regulated when comparing PTC and normal tissues (Figure 12.8). The network also showed that RELA was up-regulated in the PTC-normal results and this gene is responsible for activating the following S1 genes: FN1, NOX4, PLAU, and TNC (Figure 12.8).

Four and five transcriptional factors that activate the S2 signature genes MMP1 and SERPINE1, respectively, were up-regulated in the comparison between ATC and PTC samples (Figure 12.9). Again, the transcription factor RELA, which is responsible for the activation of three S2 signature genes, was also up-regulated in

Table 12.2: Trend Up and Trend Down gene signatures

Up signatures			Down Signatures		
S1 (n=25)	S2 (n=17)	S3 (n=29)	S1 (n=29)	S2 (n=23)	S3 (n=31)
ADM	ANLN	ADAM12	ABAT	BEX1	AIF1L
ALDH1A3	ASPM	BCAT1	ATP8A1	CLCNKB	ALDH1A1
BID	CDKN3	BUB1B	BEX2	COL23A1	AQP4
CDKN2B	CEP55	CCL20	CD24	COL9A3	C16orf89
COL10A1	CXCL5	CENPK	CDH1	CRABP1	CLDN8
CTSC	DLGAP5	COL11A1	CLIC3	CSGALNACT1	DIO1
EDIL3	DUXAP10	COL1A1	EPB41L4B	DPP6	DIO2
EVA1A	E2F7	COL1A2	FAM189A2	FAM167A	EDN3
FN1	MELK	COL5A1	FOXE1	GPM6A	FCGBP
FXYD5	MMP1	CTHRC1	GPX3	HGD	FGFR2
ICAM1	MMP12	CXCL8	HHEX	IGFBPL1	FOLR1
LOC100506403	PBK	INHBA	KLHL14	KCNIP4	FREM2
LOC101928269	SERPINE1	KIAA0101	NEBL	LOC646736	HLF
MRC2	SRPX2	LOXL2	NTRK2	LRP1B	HSD17B6
NOX4	TMEM158	MS4A4A	PAX8	MPPED2	IYD
NTM	TRIP13	NUSAP1	PDE8B	MUM1L1	KCNAB1
PLAU	UBE2C	PLAUR	PLEKHH1	PKHD1L1	KIAA1456
RAB27B		PMAIP1	PWAR6	PPARGC1A	LRRC2
RUNX1		POSTN	SLC6A13	RYR2	MATN2
RUNX2		RRM2	SNRPN	SLC4A4	PCP4
SPOCK1		SCD	TG	SPX	PRDM16
TNC		SPP1	TSHR	STXBP5L	PRTG
TREM1		TDO2	ZBED2	TCEAL2	PTCSC1
TYMS		TGFBI		TDRD9	SCUBE3
ULBP2		THBS1		TFCP2L1	SLC26A4
		TNFAIP6		TFF3	SLC26A4-AS1
		TOP2A		TMEM139	SMAD9
		UHRF1		TNFRSF11B	SORBS2
		VCAN		WSCD2	SORD
					TPO
					TXNL1

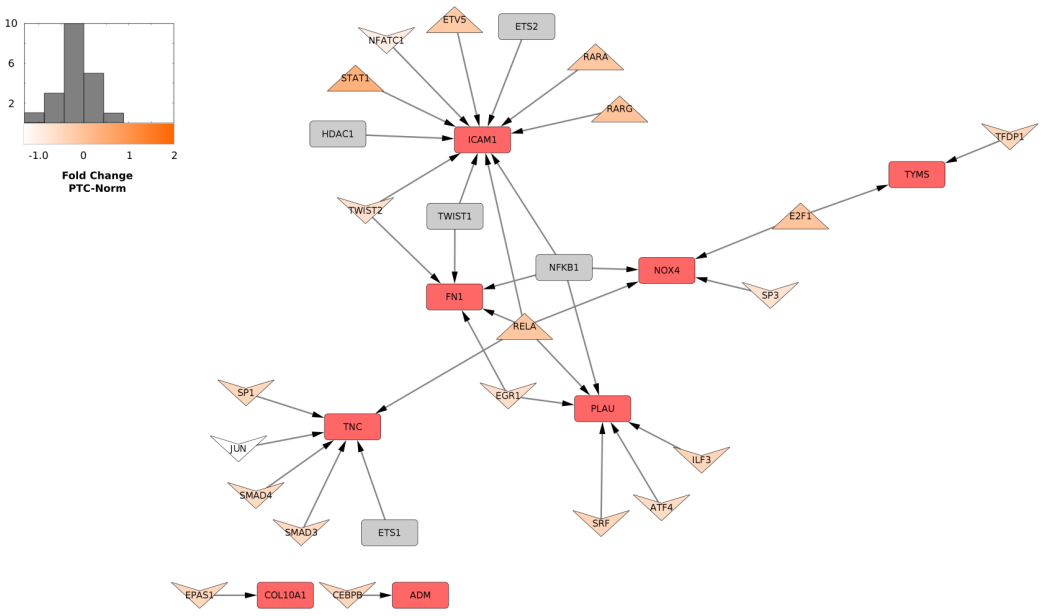


Figure 12.8: Genes that activate Up S1 genes (Red nodes). Orange colored nodes represent genes resulting as significant in the PTC vs normal comparison. Triangles represent significantly up-regulated genes and inverted triangles represent significantly down-regulated genes. The histogram shows the distribution (and color grade) of the fold change values for orange colored nodes.

Table 12.3: Signatures Up in BioPlex Network - Ranked by Degree and Betweenness Centrality (BC)

Gene	Signature	Degree	BC
ICAM1	s1	33	10307
FN1	s1	18	4201
ULBP2	s1	14	3885
MRC2	s1	12	3753
RAB27B	s1	11	3374
EDIL3	s1	7	1242
COL10A1	s1	7	1106
ALDH1A3	s1	6	1582
NTM	s1	6	876
TYMS	s1	5	1303
NOX4	s1	2	100
PLAU	s1	2	17
ASPM	s2	31	10188
PBK	s2	12	2431
MELK	s2	12	2135
CEP55	s2	11	2521
TRIP13	s2	8	1990
DLGAP5	s2	7	3428
E2F7	s2	1	0
CDKN3	s2	1	0
PLAUR	s3	59	18180
LOXL2	s3	21	3265
BCAT1	s3	19	6322
COL1A1	s3	12	2747
UHRF1	s3	11	2726
NUSAP1	s3	10	2193
RRM2	s3	8	673
TOP2A	s3	7	980
COL5A1	s3	7	275
BUB1B	s3	5	201
THBS1	s3	4	488
COL1A2	s3	4	179
VCAN	s3	3	721
KIAA0101	s3	2	0

Table 12.4: Signatures Down in BioPlex Network - Ranked by Degree and Betweenness Centrality (BC)

Gene	Signature	Degree	BC
FAM167A	s1	48	19278
GPM6A	s1	19	3995
LRP1B	s1	17	5444
KCNIP4	s1	10	3346
CSGALNACT1	s1	9	1053
TMEM139	s1	2	369
TCEAL2	s1	2	369
DPP6	s1	2	36
TFCP2L1	s1	1	0
STXBP5L	s1	1	0
SNRPN	s2	89	12983
FAM189A2	s2	48	10372
TSHR	s2	8	1707
KLHL14	s2	7	1030
NEBL	s2	4	1104
PLEKHH1	s2	2	369
CDH1	s2	2	100
CLIC3	s2	1	0
ATP8A1	s2	1	0
FGFR2	s3	38	12564
FREM2	s3	15	3371
HSD17B6	s3	11	2603
SMAD9	s3	5	1104
SORD	s3	2	1101
KIAA1456	s3	2	369
MATN2	s3	2	0
PRTG	s3	2	0
FCGBP	s3	1	0
HLF	s3	1	0

the ATC-PTC comparison (Figure 12.9).

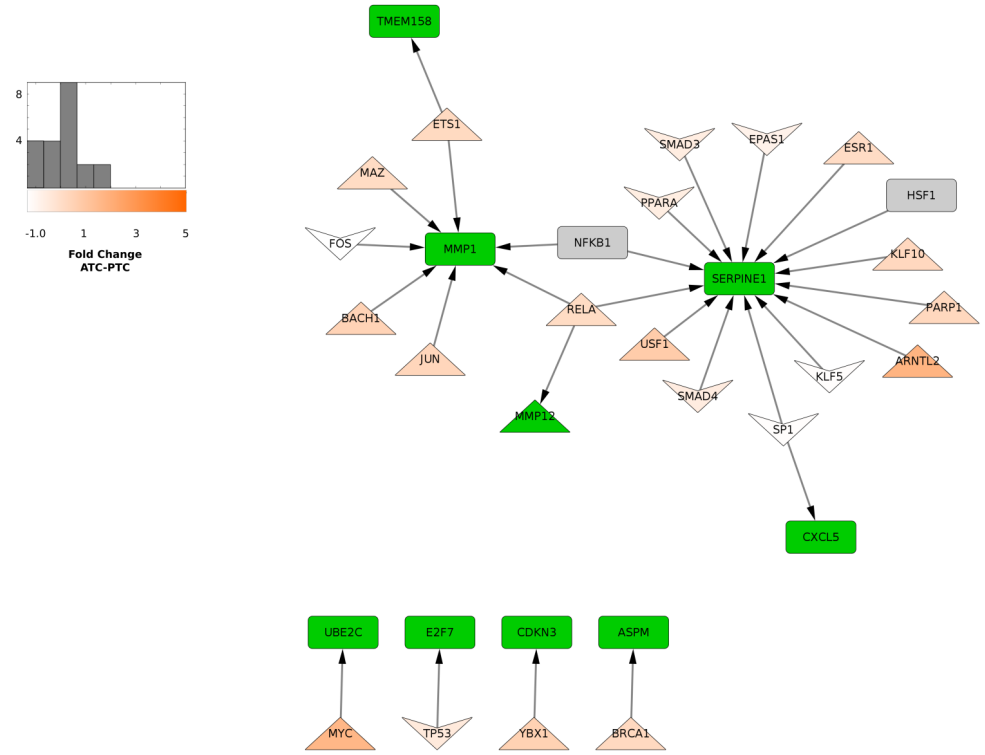


Figure 12.9: Genes that activate Up S genes (Red nodes). Orange colored nodes represent genes resulting as significant in the PTC vs normal comparison. Triangles represent significantly up-regulated genes and inverted triangles represent significantly down-regulated genes. The histogram shows the distribution (and color grade) of the fold change values for orange colored nodes.

Chapter 13

Discussion

In this study, in order to investigate the molecular mechanisms involved in the thyroid cancer progression, we analyzed gene expression profiles of normal thyroid, papillary thyroid carcinoma (PTC), and anaplastic thyroid carcinoma (ATC) samples.

A principal component analysis applied to entire dataset of expression profiles showed a gradational trend separating normal, PTC, and ATCs samples. We observed that many deregulated pathways in the comparison between ATC and normal samples were also deregulated in the ATC-PTC and PTC-normal comparisons. These results indicate common oncogenic alterations, supporting the hypothesis that ATCs originate from preexisting PTCs. Recently, a study demonstrated that the allelic frequency of TERT mutations increase from PTC to ATC [95], indicating that the tumor progression may result from the expansion of subclones in the advanced disease.

Then, we separated differentially expressed genes based on two general trends: Trend Up and Trend Down, representing genes with increasing and decreasing expression levels along tumor progression, respectively. For each trend, we defined three signatures representing the highest deregulated genes in the normal to PTC (S1) or PTC to ATC (S2) transitions, and those highly deregulated in both steps (S3).

Weinberger *et. al.* [96] proposed a list of ATC-specific genes, based on the meta-analysis of 85 samples that were also evaluated by our study. We observed that 13/17 genes defined by us as highly up-regulated in the PTC-ATC transition (Trend Up signature S2) were concordant with their list. Our results also confirmed their observation about the importance of cell cycle controlling processes in the ATC tumor biology. However, they reported genes that we observed as signatures for both normal-PTC and PTC-ATC transitions: 2/36 were defined as Trend Up signature S3, and 8/40 and 4/40 genes were defined as Trend Down signatures S1 and S3, respectively. As they did not performed all possible comparisons between tissue types, some genes that they reported as ATC signature were also deregulated in the normal to PTC transition. Again, these observations suggest that PTC and ATCs have common transcriptional alterations originated from a unique tumorigenic origin.

Trend Up Signatures

The genes in the Trend Up signatures (Table 12.2) agree with the role of MAPK signaling pathway in thyroid cancer tumorigenesis. For example, in PTCs, this pathway contributes to cancer malignancy by activating the urokinase plasminogen activator receptor (PLAUR) [97, 98], which was reported in the signature S3. Additionally, the signatures S1 and S2 also presented other genes associated with this receptor: MRC2 and PLAU in the S1; SRPX2 and SERPINE1 in the S2.

Aberrant MAPK signaling permits the release, in the extracellular matrix, of proteins that interact with integrins and non-integrin membrane receptors, resulting in autocrine and paracrine loops that promote tumor progression and metastasis [99, 100]. Our results indicate that most of the genes up-regulated in the normal-PTC transition were associated with the extracellular matrix organization. Specifically, the integrin ligand ICAM1, presented in the Trend Up signature S1, resulted as highly central in the protein-protein interaction network (Table 12.3), suggesting it as candidate for targeted therapy strategies. In fact, recent studies demonstrated ICAM1 prognosis potential in different tumor types [101–103], and its downregulation attenuated the metastatic ability of human breast cancer cell lines [104].

The analysis of the transcriptional regulatory networks (Figures 12.8 and 12.9) showed that the transcriptional factor RELA activates 4 and 3 genes from the signatures S1 and S2, respectively. The RELA protein integrates the NFKB transcription factor, which controls proliferative and anti-apoptotic signaling pathways in thyroid cancer cells [105, 106]. It has been demonstrated that the RELA expression differentiated PTC according to clinicopathological parameters, indicating that it may contribute to tumor growth and aggressiveness [107].

Trend Down Signatures

The trend down signatures (Table 12.2) presented the following genes related to iodide-handling machinery: TSHR and TG in the S1; and TPO and SCL26A4 in the S3. It has been demonstrated that BRAF mutated thyroid tumors have decreased expression of these genes, which may explain the inefficiency of the radioiodine treatment in ATC tumors. In fact, TSHR signaling seems to be protective against malignant transformation of thyroid cells, and low serum levels of TSH are associated with common genetic variants that predispose to increased risk of thyroid cancer (Reviewed in [87]).

The trend down signature genes FAM167A, SNRPN, and FGRF2 resulted as highly connected in the protein-protein interaction network (Table 12.4), indicating their importance in the cell signaling transduction. In fact, FGFR2 has been reported as downregulated in thyroid cancer [108–110] and further research about the role of these genes in thyroid tumors may indicate the most important silencing processes that takes place in tumor development and progression.

Conclusion

We believe that our study provides evidence for the understanding of mechanisms leading to thyroid cancer progression. We provided gene signatures related to each progression step: from normal tissue to PTC, and from PTC to ATC. The signatures present genes already reported in the literature as related to thyroid can-

cer, but they also propose, for the first time, several genes that have never been investigated before for their role in thyorid tumors. As a future perspective, we will validate the changes in the expression of the signature genes in cases which patients presented both the PTC and ATC cancer forms.

List of Publications

- Network integration of multi-tumour omics data suggests novel targeting strategies. (Submitted)
  - **do Valle ÍF**, Menichetti G, Simonetti G, Bruno S, Zironi I, Durso DF, Mombach JCM, Martinelli G, Castellani G, Remondini D.
- Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC Bioinformatics. **2016**.
  - **do Valle ÍF**, Giampieri E, Simonetti G, Padella A, Manfrini M, Ferrari A, Papayannidis C, Zironi I, Garonzi M, Bernardi S, Delledonne M, Martinelli G, Remondini D, Castellani G.
- Acceleration of leukocytes' epigenetic age as an early tumor- and sex-specific marker of breast and colorectal cancer. Oncotarget. **In press**.
  - Durso DF, Bacalini MG, Sala C, Pirazzini C, Marasco E, Bonafè M, **do Valle ÍF**, Gentilini D, Castellani G, Faria AMC, Franceschi C, Garagnani P, Nardini C.
- Aberrant methylation patterns in colorectal cancer: a meta-analysis. Oncotarget. **2017**.
  - Durso DF, Bacalini MG, **do Valle ÍF**, Pirazzini C, Bonafé M, Castellani G, Faria AMC, Franceschi C, Garagnani P, Nardini C.
- Stochastic neutral modelling of the Gut Microbiota's relative species abundance from next generation sequencing data. BMC Bioinformatics. **2016**.
  - Sala C, Vitali S, Giampieri E, **do Valle ÍF**, Remondini D, Garagnani P, Bersanelli M, Mosca E, Milanesi L, Castellani G.
- Systems medicine of inflammaging. Briefings in Bioinformatics. **2015**.
  - Castellani G, Menichetti G, Garagnani P, Bacalini MG, Pirazzini C, Franceschi C, Collino S, Sala C, Remondini D, Giampieri E, Mosca E, Bersanelli M, Vitali S, **do Valle ÍF**, Liò P, Milanesi L.

List of Figures

1.1 Examples of genomic alterations in cancer . . . 10
1.2 Hallmarks of cancer . . . 12

2.1 Acquisition of somatic mutations by cancer cells . . . 15
2.2 Somatic mutational frequency and mutations profiles in cancer . . . 16

3.1 Pipeline of sSNV detection in sequencing data of cancer samples . . . 20

4.1 Comparison of GATK and MuTect results prior and after applying the GATK-LOD_N filter . . . 27
4.2 Number of False Negatives and True Positives from MuTect and GATK-LOD_N results at different coverage levels . . . 30

6.1 A schematic illustration of the molecular information layers of a cell . 37

7.1 The Context Likelihood of Relatedness (CLR) algorithm . . . 39
7.2 BioPlex Ontocancro Network . . . 41

8.1 Tumor clustering . . . 44
8.2 Tumor cluster networks . . . 45
8.3 Networks showing first neighbors of signature genes that are shared by different tumor clusters . . . 47
8.4 STAT3 gene expression . . . 48
8.5 Networks composed by the first neighbors of the cluster signatures . . 49
8.6 Permutation test . . . 49
8.7 Kaplan-Meier Curves . . . 51
8.8 *In vitro* response of cancer cell lines from signature 2 to treatment with Bortezomib, BI6727 and PF-00477736 as single agent . . . 52
8.9 Sensitivity of MCF-7 and T98G cells to combined inhibition of PLK1 and the protesome . . . 53

11.1 Figure showing the steps in the definition of the gene signatures . . . 61

12.1 Heatmap showing the normalized expression levels of the 150 probes with highest variance across the entire dataset . . . 64
12.2 Principal Component Analysis applied to the entire gene expression dataset . . . 64
12.3 Differentially expressed probes in the thyroid cancer dataset . . . 65
12.4 Gene Ontology enrichment analysis . . . 65
12.5 Reactome pathway enrichment analysis . . . 66

12.6 Reactome pathway enrichment analysis for Trend Up and Trend Down lists 66
12.7 Trend Up and Trend Down gene signatures 67
12.8 Transcriptional regulatory network for genes in the Trend Up S1 signature 68
12.9 Transcriptional regulatory network for genes in the Trend Up S2 signature 71

List of Tables

1.1 Categories of Genomic Alteration and Exemplary Cancer Genes Exemplary . . . 11

3.1 List of parameters and thresholds used for SNV hard filtering . . . 22
3.2 MuTect filters . . . 23
3.3 Artifical Tumor Samples . . . 25
3.4 Number of SNVs submitted to experimental validation . . . 25

4.1 MuTect False Negatives . . . 28
4.2 MuTect results with lowered decision threshold values . . . 28
4.3 MuTect and GATK-LOD_N applied to artificial tumor samples . . . 29
4.4 The GATK-LOD_N method increases the GATK performance for both mutation detection and classification . . . 31

7.1 List of tumors and the respective number of gene expression arrays analyzed . . . 38

8.1 Network Properties. . . . 45
8.2 Central genes . . . 46
8.3 Common biological categories present in the gene signatures . . . 47
8.4 List of genes from the signatures that are also being tested in ongoing clinical trials studies (according ClincalTrials.gov). . . . 48
8.5 Combination Indexes for BI6727 and Bortezomib at different concentrations in MCF-7 cell line. . . . 50
8.6 Combination Indexes for BI6727 and Bortezomib at different concentrations in T98G cell line. . . . 51

11.1 Accession numbers for the thyroid gene expression profiles used in this study . . . 60

12.1 Common enriched pathways in the Trend Up and Trend Down lists compared to the differentially expressed genes in each comparison . . . 67
12.2 Trend Up and Trend Down gene signatures . . . 68
12.3 Signatures Up in BioPlex Network - Ranked by Degree and Betweenness Centrality (BC) . . . 69
12.4 Signatures Down in BioPlex Network - Ranked by Degree and Betweenness Centrality (BC) . . . 70

A1 AML samples coverage . . . 90
A2 ALL samples coverage . . . 91

A3 GIST samples coverage 91
A4 Lung Adenocarcinoma samples coverage 92

Bibliography

[1] American Cancer Society, "Global Cancer Facts & Figures 3rd Edition.," *American Cancer Society*, no. 800, pp. 1–64, 2015.

[2] B. W. Stewart and C. P. Wild, "World cancer report 2014," *World Health Organization*, pp. 1–2, 2014.

[3] D. von Hansemann, "Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung," *Virchows Arch. Path. Anat*, vol. 119, no. 299, 1890.

[4] T. Boveri, *Zur frage der entstehung maligner tumoren.* 1914.

[5] C. J. Tabin, S. M. Bradley, C. I. Bargmann, R. A. Weinberg, A. G. Papageorge, and et al, "Mechanism of activation of a human oncogene.," *Nature*, vol. 300, pp. 143–9, nov 1982.

[6] E. P. Reddy, R. K. Reynolds, E. Santos, and M. Barbacid, "A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene.," *Nature*, vol. 300, pp. 149–52, nov 1982.

[7] L. E. MacConaill and L. A. Garraway, "Clinical implications of the cancer genome," *Journal of Clinical Oncology*, vol. 28, no. 35, pp. 5219–5228, 2010.

[8] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer.," *Nature reviews. Genetics*, vol. 14, no. 10, pp. 703–18, 2013.

[9] H. Easwaran, H. C. Tsai, and S. B. Baylin, "Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance," *Molecular Cell*, vol. 54, no. 5, pp. 716–727, 2014.

[10] H. Davies, G. R. Bignell, C. Cox, P. Stephens, S. Edkins, and et al, "Mutations of the BRAF gene in human cancer.," *Nature*, vol. 417, pp. 949–54, jun 2002.

[11] D. Hanahan and R. a. Weinberg, "Hallmarks of cancer: the next generation.," *Cell*, vol. 144, pp. 646–74, mar 2011.

[12] L. A. Garraway and E. S. Lander, "Lessons from the cancer genome," *Cell*, vol. 153, no. 1, pp. 17–37, 2013.

[13] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, and et al, "Accurate whole human genome sequencing using reversible terminator chemistry.," *Nature*, vol. 456, pp. 53–9, nov 2008.

[14] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, and et al, "Genome sequencing in microfabricated high-density picolitre reactors.," *Nature*, vol. 437, pp. 376–80, sep 2005.

[15] M. L. Metzker, "Sequencing technologies - the next generation.," *Nature reviews. Genetics*, vol. 11, pp. 31–46, jan 2010.

[16] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome.," *Nature*, vol. 458, pp. 719–24, apr 2009.

[17] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, and et al, "Mutational heterogeneity in cancer and the search for new cancer-associated genes.," *Nature*, vol. 499, pp. 214–8, jul 2013.

[18] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, and et al, "A survey of tools for variant analysis of next-generation genome sequencing data," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 256–278, 2013.

[19] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, "Genotype and SNP calling from next-generation sequencing data.," *Nature reviews. Genetics*, vol. 12, pp. 443–51, jun 2011.

[20] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, and et al, "SOAP2: An improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[21] N. D. Roberts, R. D. Kortschak, W. T. Parker, A. W. Schreiber, S. Branford, and et al, "A comparative analysis of algorithms for somatic SNV detection in cancer," *Bioinformatics*, vol. 29, no. 18, pp. 2223–2230, 2013.

[22] D. H. Spencer, M. Tyagi, F. Vallania, A. J. Bredemeyer, J. D. Pfeifer, and et al, "Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data," *Journal of Molecular Diagnostics*, vol. 16, no. 1, pp. 75–88, 2014.

[23] H. Xu, J. DiCarlo, R. V. Satya, Q. Peng, and Y. Wang, "Comparison of somatic mutation calling methods in amplicon and whole exome sequence data.," *BMC genomics*, vol. 15, p. 244, jan 2014.

[24] Q. Wang, P. Jia, F. Li, H. Chen, H. Ji, and et al, "Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers.," *Genome medicine*, vol. 5, no. 10, p. 91, 2013.

[25] S. Hwang, E. Kim, I. Lee, and E. M. Marcotte, "Systematic comparison of variant calling pipelines using gold standard personal exome variants," *Scientific Reports*, vol. 5, no. December, p. 17875, 2015.

[26] A. R. Carson, E. N. Smith, H. Matsui, S. K. Brækkan, K. Jepsen, and et al, "Effective filtering strategies to improve data quality from population-based whole exome sequencing studies.," *BMC bioinformatics*, vol. 15, p. 125, 2014.

[27] H. Li, "Toward better understanding of artifacts in variant calling from high-coverage samples," *Bioinformatics*, vol. 30, pp. 2843–2851, oct 2014.

[28] J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, and et al, "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.," *Genome medicine*, vol. 5, no. 3, p. 28, 2013.

[29] M. Bodini, C. Ronchini, L. Giac, A. Russo, G. E. M. Melloni, and et al, "Perspectives The hidden genomic landscape of acute myeloid leukemia : subclonal structure revealed by undetected mutations," *Blood*, vol. 125, no. 4, pp. 600–606, 2015.

[30] G. Kang, H. Yun, C.-h. Sun, I. Park, J. Kwon, and et al, "Integrated genomic analyses identify frequent gene fusion events and VHL inactivation in gastrointestinal stromal tumors," *Oncotarget*, pp. 1–14, 2015.

[31] J. S. Seo, Y. S. Ju, W. C. Lee, J. Y. Shin, J. K. Lee, and et al, "The transcriptional landscape and mutational profile of lung adenocarcinoma," *Genome Research*, vol. 22, no. 11, pp. 2109–2119, 2012.

[32] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1754–60, jul 2009.

[33] M. a. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, and et al, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.," *BMC genomics*, vol. 13, p. 341, jan 2012.

[34] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, and et al, "Characterizing and measuring bias in sequence data.," *Genome biology*, vol. 14, no. 5, p. R51, 2013.

[35] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, and et al, "Sequence-specific error profile of Illumina sequencers.," *Nucleic acids research*, vol. 39, p. e90, jul 2011.

[36] M. a. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, and et al, "A framework for variation discovery and genotyping using next-generation DNA sequencing data.," *Nature genetics*, vol. 43, pp. 491–8, may 2011.

[37] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, and et al, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.," *Nature biotechnology*, vol. 31, pp. 213–9, mar 2013.

[38] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.," *Nucleic acids research*, vol. 38, p. e164, sep 2010.

[39] M. R. Breese and Y. Liu, "NGSUtils: A software suite for analyzing and manipulating next-generation sequencing datasets," *Bioinformatics*, vol. 29, no. 4, pp. 494–496, 2013.

[40] N. F. Hansen, J. J. Gartner, L. Mei, Y. Samuels, and J. C. Mullikin, "Shimmer: Detection of genetic alterations in tumors using next-generation sequence data," *Bioinformatics*, vol. 29, no. 12, pp. 1498–1503, 2013.

[41] A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, and et al, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, sep 2015.

[42] S. Y. Kim, L. Jacob, and T. P. Speed, "Combining calls from multiple somatic mutation-callers," *BMC Bioinformatics*, vol. 15, no. 1, p. 154, 2014.

[43] G. J. Lyon and K. Wang, "Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress," *Genome Medicine*, vol. 4, no. 7, p. 58, 2012.

[44] E. a. Collisson, J. D. Campbell, A. N. Brooks, A. H. Berger, W. Lee, and et al, "Comprehensive molecular profiling of lung adenocarcinoma.," *Nature*, vol. 511, no. 7511, pp. 543–50, 2014.

[45] J. Reumers, P. De Rijk, H. Zhao, A. Liekens, D. Smeets, and et al, "Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing," *Nature Biotechnology*, vol. 30, no. 1, pp. 61–68, 2011.

[46] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, and et al, "Mutational landscape and significance across 12 major cancer types.," *Nature*, vol. 502, no. 7471, pp. 333–9, 2013.

[47] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, and et al, "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.," *Genome research*, vol. 22, pp. 568–76, mar 2012.

[48] D. M. Muzny, M. N. Bainbridge, K. Chang, H. H. Dinh, J. a. Drummond, and et al, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–337, 2012.

[49] K. a. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, and et al, "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin," *Cell*, vol. 158, no. 4, pp. 929–944, 2013.

[50] C. G. A. R. Network, G. S. C. B. Institute, G. Getz, S. B. Gabriel, K. Cibulskis, and et al, "Integrated genomic characterization of endometrial carcinoma.," *Nature*, vol. 497, no. 7447, pp. 67–73, 2013.

[51] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and et al, "Emerging landscape of oncogenic signatures across human cancers.," *Nature genetics*, vol. 45, no. 10, pp. 1127–1133, 2013.

[52] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, and et al, "Systematic identification of genomic markers of drug sensitivity in cancer cells.," *Nature*, vol. 483, no. 7391, pp. 570–5, 2012.

[53] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. a. Margolin, and et al, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–7, 2012.

[54] V. Gligorijevic and N. Przulj, "Methods for biological data integration: perspectives and challenges.," *Journal of the Royal Society,*, vol. 12, no. 112, pp. 20150571–, 2015.

[55] E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, and et al, "The BioPlex Network: A Systematic Exploration of the Human Interactome," *Cell*, vol. 162, no. 2, pp. 425–440, 2015.

[56] E. Saccenti, M. Suarez-diez, C. Luchinat, C. Santucci, and L. Tenori, "Probabilistic Networks of Blood Metabolites in Healthy Subjects As Indicators of Latent Cardiovascular Risk," *Journal of Proteome Research*, vol. 14, no. 1101-1111, 2015.

[57] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, and et al, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, pp. 0054–0066, 2007.

[58] P. E. Meyer, F. Lafitte, and G. Bontempi, "minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information.," *BMC bioinformatics*, vol. 9, p. 461, 2008.

[59] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, and et al, "Cytoscape: a software environment for integrated models of biomolecular interaction networks.," *Genome research*, vol. 13, pp. 2498–504, nov 2003.

[60] S. D. Pauls and D. Remondini, "Measures of centrality based on the spectrum of the Laplacian," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 85, no. 6, p. 066127, 2012.

[61] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, and et al, "COSMIC: Exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Research*, vol. 43, no. D1, pp. D805–D811, 2015.

[62] A. H. Wagner, A. C. Coffman, B. J. Ainscough, N. C. Spies, Z. L. Skidmore, and et al, "DGIdb 2.0: mining clinically relevant drug-gene interactions.," *Nucleic acids research*, vol. 44, pp. D1036–44, jan 2016.

[63] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, and et al, "DrugBank: a comprehensive resource for in silico drug discovery and exploration.," *Nucleic acids research*, vol. 34, pp. D668–72, jan 2006.

[64] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. New York: Springer, 2nd ed., 2009.

[65] J. M. Bland and D. G. Altman, "Survival probabilities (the Kaplan-Meier method).," *BMJ (Clinical research ed.)*, vol. 317, p. 1572, dec 1998.

[66] J. M. Bland and D. G. Altman, "The logrank test.," *BMJ (Clinical research ed.)*, vol. 328, p. 1073, may 2004.

[67] a. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, and et al, "A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction," *Science Signaling*, vol. 4, no. September 2011, pp. rs8–rs8, 2011.

[68] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets.," *Genome research*, vol. 14, pp. 1085–94, jun 2004.

[69] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 14863–8, dec 1998.

[70] E. Martínez, K. Yoshihara, H. Kim, G. M. Mills, V. Treviño, and et al, "Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects," *Oncogene*, vol. 34, no. 21, pp. 2732–2740, 2015.

[71] T. Ideker and R. Sharan, "Protein networks in diseases," *Genome Research*, vol. 18, pp. 644–652, 2008.

[72] G. C. Castellani, G. Menichetti, P. Garagnani, M. Giulia Bacalini, C. Pirazzini, and et al, "Systems medicine of inflammaging.," *Briefings in bioinformatics*, no. April, pp. 1–14, 2015.

[73] B. Hoesel and J. a. Schmid, "The complexity of NF-κB signaling in inflammation and cancer.," *Molecular cancer*, vol. 12, no. 1, p. 86, 2013.

[74] L. Bie, G. Zhao, Y.-p. Wang, and B. Zhang, "Kinesin family member 2C (KIF2C/MCAK) is a novel marker for prognosis in human gliomas.," *Clinical neurology and neurosurgery*, vol. 114, pp. 356–60, may 2012.

[75] M. Sanhaji, C. T. Friel, L. Wordeman, F. Louwen, and J. Yuan, "Mitotic centromere-associated kinesin (MCAK): a potential cancer drug target," *Oncotarget*, vol. 2, pp. 935–947, dec 2011.

[76] L. Bie, G. Zhao, P. Cheng, G. Rondeau, S. Porwollik, and et al, "The accuracy of survival time prediction for patients with glioma is improved by measuring mitotic spindle checkpoint gene expression.," *PloS one*, vol. 6, no. 10, p. e25631, 2011.

[77] P. Finetti, N. Cervera, E. Charafe-Jauffret, C. Chabannon, C. Charpin, and et al, "Sixteen-kinase gene expression identifies luminal breast cancers with poor prognosis.," *Cancer research*, vol. 68, pp. 767–76, feb 2008.

[78] A. A. Sahasrabuddhe and K. S. J. Elinitoba-Johnson, "Role of the ubiquitin proteasome system in hematologic malignancies," *Immunological Reviews*, vol. 263, pp. 224–239, 2015.

[79] I. Saez and D. Vilchez, "The Mechanistic Links Between Proteasome Activity, Aging and Age-related Diseases.," *Current genomics*, vol. 15, pp. 38–51, feb 2014.

[80] D. Iliopoulos, H. A. Hirsch, and K. Struhl, "An Epigenetic Switch Involving NF-KB, Lin28, Let-7 MicroRNA, and IL6 Links Inflammation to Cell Transformation," *Cell*, vol. 139, no. 4, pp. 693–706, 2009.

[81] A. Shlien, K. Raine, F. Fuligni, R. Arnold, S. Nik-Zainal, and et al, "Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer," *Cell Reports*, vol. 16, no. 7, pp. 2032–2046, 2016.

[82] C. J. Creighton, M. Morgan, P. H. Gunaratne, D. a. Wheeler, R. a. Gibbs, and et al, "Comprehensive molecular characterization of clear cell renal cell carcinoma.," *Nature*, vol. 499, no. 7456, pp. 43–9, 2013.

[83] C. Lee, A. Fotovati, J. Triscott, J. Chen, C. Venugopal, and et al, "Polo-like kinase 1 inhibition kills glioblastoma multiforme brain tumor cells in part through loss of SOX2 and delays tumor progression in mice.," *Stem cells (Dayton, Ohio)*, vol. 30, pp. 1064–75, jun 2012.

[84] S. Thaler, G. Thiede, J. G. Hengstler, A. Schad, M. Schmidt, and et al, "The proteasome inhibitor Bortezomib (Velcade) as potential inhibitor of estrogen receptor-positive breast cancer.," *International journal of cancer*, vol. 137, pp. 686–97, aug 2015.

[85] D. Yin, H. Zhou, T. Kumagai, G. Liu, J. M. Ong, and et al, "Proteasome inhibitor PS-341 causes cell growth arrest and apoptosis in human glioblastoma multiforme (GBM).," *Oncogene*, vol. 24, pp. 344–54, jan 2005.

[86] A. Y. Chen, A. Jemal, and E. M. Ward, "Increasing incidence of differentiated thyroid cancer in the United States, 1988-2005.," *Cancer*, vol. 115, pp. 3801–7, aug 2009.

[87] M. Xing, "Molecular pathogenesis and mechanisms of thyroid cancer," *Nat Rev Cancer*, vol. 13, no. 3, pp. 184–199, 2013.

[88] X. M. Keutgen, S. M. Sadowski, and E. Kebebew, "Management of anaplastic thyroid cancer.," *Gland surgery*, vol. 4, pp. 44–51, feb 2015.

[89] N. Agrawal, R. Akbani, B. A. Aksoy, A. Ally, H. Arachchi, and et al, "Integrated Genomic Characterization of Papillary Thyroid Carcinoma," *Cell*, vol. 159, no. 3, pp. 676–690, 2014.

[90] Y. E. Nikiforov, "Genetic alterations involved in the transition from well-differentiated to poorly differentiated and anaplastic thyroid carcinomas.," *Endocrine pathology*, vol. 15, no. 4, pp. 319–27, 2004.

[91] L. Santarpia, A. K. El-Naggar, G. J. Cote, J. N. Myers, and S. I. Sherman, "Phosphatidylinositol 3-kinase/akt and ras/raf-mitogen-activated protein kinase pathway mutations in anaplastic thyroid cancer.," *The Journal of clinical endocrinology and metabolism*, vol. 93, pp. 278–84, jan 2008.

[92] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters.," *Omics : a journal of integrative biology*, vol. 16, pp. 284–7, may 2012.

[93] G. Yu and Q.-Y. He, "ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization.," *Molecular bioSystems*, vol. 12, pp. 477–9, feb 2016.

[94] H. Han, H. Shim, D. Shin, J. E. Shim, Y. Ko, and et al, "TRRUST: a reference database of human transcriptional regulatory interactions.," *Scientific reports*, vol. 5, p. 11432, 2015.

[95] I. Landa, T. Ibrahimpasic, L. Boucai, R. Sinha, J. A. Knauf, and et al, "Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers," *Journal of Clinical Investigation*, vol. 126, no. 3, pp. 1052–1066, 2016.

[96] P. Weinberger, S. R. Ponny, H. Xu, S. Bai, R. C. Smallridge, and et al, "Cell Cycle M-phase Genes are Highly Upregulated in Anaplastic Thyroid Carcinoma," *Thyroid*, vol. 27, no. 2, p. thy.2016.0285, 2016.

[97] T. S. Nowicki, H. Zhao, Z. Darzynkiewicz, A. Moscatello, E. Shin, and et al, "Downregulation of uPAR inhibits migration, invasion, proliferation, FAK/PI3K/Akt signaling and induces senescence in papillary thyroid carcinoma cells," *Cell Cycle*, vol. 10, no. 1, pp. 100–107, 2011.

[98] T. S. Nowicki, N. T. Kummer, C. Iacob, N. Suslina, S. Schaefer, and et al, "Inhibition of uPAR and uPA reduces invasion in papillary thyroid carcinoma cells.," *The Laryngoscope*, vol. 120, pp. 1383–90, jul 2010.

[99] C. Nucera, A. Porrello, Z. A. Antonello, M. Mekel, M. A. Nehs, and et al, "B-Raf(V600E) and thrombospondin-1 promote thyroid cancer progression.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 10649–54, jun 2010.

[100] C. Nucera, J. Lawler, and S. Parangi, "BRAFV600Eand microenvironment in thyroid cancer: A functional link to drive cancer progression," *Cancer Research*, vol. 71, no. 7, pp. 2417–2422, 2011.

[101] G. Cai, X. Ma, W. Zou, Y. Huang, J. Zhang, and et al, "Prediction value of intercellular adhesion molecule-1 gene polymorphisms for epithelial ovarian cancer risk, clinical features, and prognosis.," *Gene*, vol. 546, pp. 117–23, aug 2014.

[102] X. Shi, J. Jiang, X. Ye, Y. Liu, Q. Wu, and et al, "Prognostic prediction and diagnostic role of intercellular adhesion molecule-1 (ICAM1) expression in clear cell renal cell carcinoma.," *Journal of molecular histology*, vol. 45, pp. 427–34, aug 2014.

[103] K. Erturk, D. Tastekin, E. Bilgin, M. Serilmez, H. U. Bozbey, and et al, "Serum activated leukocyte cell adhesion molecule and intercellular adhesion molecule-1 in patients with gastric cancer: Can they be used as biomarkers?," *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, vol. 77, pp. 86–91, feb 2016.

[104] D. Di, L. Chen, L. Wang, P. Sun, Y. Liu, and et al, "Downregulation of human intercellular adhesion molecule-1 attenuates the metastatic ability in human breast cancer cell lines.," *Oncology reports*, vol. 35, pp. 1541–8, mar 2016.

[105] F. Pacifico and A. Leonardi, "Role of NF-kappaB in thyroid cancer.," *Molecular and cellular endocrinology*, vol. 321, pp. 29–35, may 2010.

[106] X. Li, A. B. Abdel-Mageed, D. Mondal, and E. Kandil, "The nuclear factor kappa-B signaling pathway as a therapeutic target against thyroid cancers.," *Thyroid : official journal of the American Thyroid Association*, vol. 23, pp. 209–18, feb 2013.

[107] Y. Zhang, Z. Meng, M. Zhang, J. Tan, W. Tian, and et al, "Immunohistochemical evaluation of midkine and nuclear factor-kappa B as diagnostic biomarkers for papillary thyroid cancer and synchronous metastasis.," *Life sciences*, vol. 118, pp. 39–45, nov 2014.

[108] R. St Bernard, L. Zheng, W. Liu, D. Winer, S. L. Asa, and et al, "Fibroblast growth factor receptors as molecular targets in thyroid carcinoma.," *Endocrinology*, vol. 146, pp. 1145–53, mar 2005.

[109] K. Kasaian, S. M. Wiseman, B. A. Walker, J. E. Schein, Y. Zhao, and et al, "The genomic and transcriptomic landscape of anaplastic thyroid cancer: implications for therapy," *BMC Cancer*, vol. 15, no. 1, p. 984, 2015.

[110] A. Redler, G. Di Rocco, D. Giannotti, F. Frezzotti, M. G. Bernieri, and et al, "Fibroblast growth factor receptor-2 expression in thyroid tumor progression: Potential diagnostic Application," *PLoS ONE*, vol. 8, no. 8, pp. 1–9, 2013.

Appendices

A Part I

Table A1: AML samples coverage

Sample	Coverage	Sample	Coverage
a1010d	79.18	b2030d	135.70
a1010s	94.93	b2030s	67.26
a1015d	105.30	b2031d	87.43
a1015s	52.12	b2031s	73.57
a1024d	146.72	b2033d	110.66
a1024s	63.22	b2033s	64.00
a1025s	79.17	b2035d	46.90
b1001d	99.34	b2035s	30.29
b1001s	75.51	b2036d	76.52
b1006d	68.84	b2036s	41.76
b1006s	65.74	b2038s	32.96
b1014d	118.81	b2039d	63.53
b1014s	55.87	b2039s	30.68
b1026d	106.22	b2040d	48.33
b1026s	67.58	b2040s	27.33
b1028d	124.08	b2042d	50.20
b1028s	68.94	b2042s	30.43
b1034d	95.15	b2043d	82.97
b1034s	71.51	b2043s	47.77
b1041d	80.63	b2045d	96.68
b1041s	40.84	b2045s	33.30
b2002d	72.04	c0017d	60.52
b2002s	26.51	c0017s	54.01
b2004d	68.14	c0018d	118.56
b2004s	38.80	c0018s	82.50
b2005d	69.98	c0022d	40.15
b2005s	78.20	c0022s	73.52
b2007d	78.49	c0046d	43.85
b2007s	60.39	c0046s	39.47
b2008d	109.54	d0027d	162.91
b2008s	81.69	d0027s	62.02
b2009d	115.84	n0187s	64.23
b2009s	58.90	n0195d	62.33
b2023d	106.21	n6364d	71.65
b2023s	80.28	n6364s	33.63

Names ending with *d* refer to tumor samples and those ending with *s* refer to saliva samples

Table A2: ALL samples coverage

Sample	Coverage	Sample	Coverage
0010jnd	137.37	01082s	126.88
0010jns	98.15	01833d	135.79
0011ltd	121.64	01833t	123.13
0011lts	144.29	02903d	116.14
003fkd	124.72	02903t	131.85
003fks	102.23	07298d	132.96
00461l	120.99	07298t	142.72
00461s	110.88	10601d	127.22
004pjd	148.67	10601t	132.14
004pjs	137.90	10861d	100.13
005djd	118.63	10861t	127.39
005djl	103.19	11232d	49.10
006mjd	140.46	11232t	125.49
006mjs	122.98	11876d	134.49
00798d	127.93	11876t	147.47
00798t	116.11	11950d	134.55
007tkd	171.06	11950t	178.08
007tks	60.83	12572d	132.03
00808d	119.51	12572t	131.69
00808t	138.27	13921d	126.43
00889d	98.31	13921t	136.60
00889s	83.58	16661d	96.59
008pbd	195.29	16661t	107.33
008pbs	62.19	24631d	108.55
00928d	80.01	24631t	128.74
00928s	91.38	25171d	115.19
00963l	144.49	25171t	74.65
00963s	124.14	30836d	123.61
00994d	122.67	30836t	117.87
00994s	77.25	33121d	138.74
01059d	61.55	33121t	136.23
01059s	60.61	34070d	141.15
01061d	143.52	34070t	132.70
01061s	128.78	41761d	154.76
01067d	76.80	41761t	129.73
01067s	57.70	48471d	142.82
01076d	134.96	48471t	136.68
01076s	139.03	85112d	107.70
01079d	59.66	85112t	134.78
01079s	65.16	98978d	125.41
01082l	139.26	98978t	128.71

Names ending with *d*, *r* or *l* refer to tumor samples; those ending with *s* refer to saliva samples

Table A3: GIST samples coverage

Sample	Coverage
SRR1299146	70.31
SRR1299147	78.41
SRR1299145	82.47
SRR1299144	70.28
SRR1299141	84.99
SRR1299140	84.48
SRR1299139	68.67
SRR1299138	61.43
SRR1299137	74.17
SRR1299136	75.23
SRR1299135	69.11
SRR1299134	78.23
SRR1299133	77.91
SRR1299132	72.11
SRR1299131	87.47
SRR1299130	83.92

The sample names are their respective SRA accession number

Table A4: Lung Adenocarcinoma samples coverage

Sample (SRA acession number)	Coverage
ERR166339	159.23
ERR160136	71.30
ERR166338	211.30
ERR160124	91.04

The sample names are their respective SRA accession number