UNIVERSITÀ DEGLI STUDI DI SALERNO
Dipartimento di Farmacia

Dottorato di ricerca

in Biologia dei Sistemi

Ciclo XIV Anno di discussione 2015

Coordinatore: Chiar.mo Prof. *Antonietta Leone*

PhD thesis
in

“Identification of new genetic alterations and potential biomarkers in papillary thyroid carcinoma”

Subtitle
Computational and experimental analysis of thyroid cancer transcriptome

Settore scientifico disciplinare di afferenza: INF/01

Dottorando	**Tutore**
Dott. *Roberta Esposito*	**Chiar.mo Prof.** *Italia De Feis*

Table of Contents

Abstract (English)

Papillary thyroid carcinoma (PTC) is the most frequent thyroid malignant neoplasia. Oncogene activation occurs in more than 70% of the cases. *BRAF* mutations occur in about 40% of PTCs, whereas RET rearrangements (RET/PTC oncogenes) are present in about 20% of cases. Finally, RAS mutations and *TRK* and *PPARG* rearrangements account for about 5% each of these malignancies.

However, despite the presence of tumor-initiating driver events, cancer results from the progressive accumulation of mutations in genes that confer growth advantage over surrounding cells. A better understanding of molecular alterations of PTC will provide important insights into cancer etiology. It will also lead to advance in their diagnosis, possibly opening the way for developing novel molecular therapies.

Thus, the aim of this PhD project is to deeply explore the transcriptome of PTC in order to identify new driver events in this type of cancer.

In the first part of this study, we used RNA-Sequencing in a discovery cohort of 18 patients with papillary thyroid carcinoma to identify fusion transcripts and expressed mutations in cancer driver genes. Furthermore, we used targeted sequencing on the DNA of these same patients to validate identified mutations. We extended the screening to thyroids of 50 PTC patients and of 30 healthy individuals. Using this approach we identified new somatic mutations in *CBL*, *NOTCH1*, *PIK3R4* and *SMARCA4* genes. We also found mutations in *DICER, MET* and *VHL* genes, previously found mutated in other tumors, but not described yet in PTC. We also identified a new chimeric transcript generated by the fusion of lysine deficient protein kinase 1 (*WNK1*) and beta-1,4-N-acetyl-galactosaminyl transferase 3 (*B4GALNT3*) genes and correlated with an overexpression of *B4GALNT3*.

Moreover, although protein coding genes play a leading role in cancer genetics, in recent years, many studies focused on a novel class of non-coding RNAs, long non-coding RNAs (lncRNAs), which regulate the expression levels of protein coding genes. Since deregulated expression of lncRNAs has been reported in many cancers, it suggests that they may act as potential oncogene or tumor-suppressor.

Thus, to assess if lncRNAs can exert a tumorigenic role in thyroid, in the second part of my PhD project I systematically quantified lncRNAs' expression in PTC *vs* healthy thyroids using our RNA-Seq data. Combining *ab initio* reconstruction to a custom computational pipeline we found that novel and known lncRNAs are significantly altered in PTC, and some of them are possibly associated with cancer driver genes. Then we extensively focused on an un-annotated lncRNA transcribed antisense to *MET* oncogene, named *MET-AS*. Both genes are significantly up-regulated in a sub-class of PTCs - i.e. patients with *BRAF* mutations and *RET* gene rearrangements, compared to other PTCs and "non-tumor" thyroid biopsies. Preliminary data indicate that *MET-AS* knockdown induces down-regulation of *MET*, and produces changes in cell cycle in a PTC cell line, suggesting the novel lncRNA might be a new *MET* regulator. Further studies should be conducted to demonstrate detailed mechanism of our findings.

Finally, our data confirmed the genetic heterogeneity of papillary thyroid carcinoma revealing that gene expression correlates more with the mutation pattern than with tumor staging. Overall, this study provides new information about PTC genetic alterations, suggesting potential pharmacological adjuvant therapies in PTC.

Abstract (Italiano)

Il carcinoma papillare tiroideo (PTC) costituisce circa l'80% di tutti i tumori maligni della tiroide. Ad oggi, sono state identificate mutazioni a carico del gene *BRAF* in circa il 40% di *casi,* mentre riarrangiamenti che coinvolgono il gene *RET* (RET/PTC) sono presenti in circa il 20% dei casi. Infine, mutazioni nei geni *RAS* e riarrangiamenti dei geni *TRK* e *PPARG* occorrono in circa il 5% dei casi ciascuno.

Tuttavia, nonostante la presenza di alterazioni genetiche che possano dare inizio al processo canceroso, il tumore è il risultato del progressivo accumulo di mutazioni in geni che conferiscono un vantaggio di crescita sulle cellule circostanti. Pertanto, una conoscenza più approfondita delle alterazioni molecolari del carcinoma papillare tiroideo è fondamentale per migliorare gli aspetti diagnostici e prognostici, e la risposta individuale ai trattamenti farmacologici.

Alla luce di ciò, il mio progetto di dottorato ha avuto come obiettivo principale l'analisi del trascrittoma del PTC al fine di individuare nuovi eventi molecolari che possano essere coinvolti in questo tipo di cancro.

La prima parte di questo progetto è stata focalizzata sul sequenziamento - mediante RNA-Seq – di 22 RNA isolati da biopsie di tiroide (18 tiroidi di soggetti con carcinoma papillare tiroideo, 4 tiroidi di soggetti in assenza di PTC) per identificare nuovi trascritti di fusione e mutazioni somatiche in geni espressi. I risultati sono stati validati sul DNA dei medesimi pazienti mediante sequenziamento diretto di Sanger. Inoltre, l'analisi mutazionale è stata estesa ad ulteriori 50 pazienti con carcinoma papillare tiroideo e 30 individui in assenza di PTC. Mediante quest'approccio sono state identificate nuove mutazioni puntiformi nei geni *CBL*, *NOTCH1*, *PIK3R4* e *SMARCA4*. Inoltre, l'analisi ha rivelato la presenza di mutazioni somatiche nei geni *DICER1*, *MET* e *VHL,* già note nella patogenesi in altri tipi di cancro, ma ad oggi non note nel PTC. Inoltre, è stato individuato un nuovo evento intra-cromosomico generato dalla fusione tra il primo esone del gene *WNK1* (*lysine deficient protein*

kinase 1) e il secondo esone del gene *B4GALNT3* (*beta-1,4-N-acetyl-galactosaminyl transferase 3).*

I geni codificanti rivestono un ruolo di primo piano nella genetica del cancro, ma negli ultimi anni, molti studi si sono concentrati su una nuova classe di RNA non-codificanti, i *long non-coding RNA* (lncRNAs), che regolano l'espressione dei geni codificanti. I livelli di espressione dei lncRNA sono spesso alterati in diversi tipi di tumori, suggerendo che essi possano agire sia da oncogeni sia da oncosoppressori. Al fine di valutare il loro potenziale ruolo nella tumorigenesi del PTC, la seconda parte di questo progetto è stata focalizzata sull'analisi computazionale dei nuovi lncRNA, e già annotati, nei dataset da me ottenuti mediante RNA-Seq. Attraverso l'utilizzo di approcci per la ricostruzione *ab initio* del trascrittoma e di una *pipeline* computazionale sono stati indentificati i lncRNA significativamente deregolati nei campioni tumorali. Inoltre, per individuare i lncRNA che potessero regolare l'espressione genica *in cis,* alcuni di essi sono stati associati - per vicinanza al TSS (*transcription start site*) - a geni *driver* in diversi tipi di cancro. Infine, mi sono focalizzata su un lncRNA non annotato nei *database* pubblici, associato all'oncogene *MET*, e trascritto in direzione antisenso rispetto al gene *MET*. Questo nuovo lncRNA è stato chiamato *MET-AS*. Entrambi i geni (*MET* e *MET-AS*) sono significativamente sovra-espressi in una sotto-classe di PTC - vale a dire i pazienti con mutazioni del gene *BRAF* e riarrangiamenti dell'oncogene *RET* – chiamati *BRAF-like-*, rispetto ai campioni tumorali PTC, con profilo trascrizionale simile ai campioni mutati nei geni *RAS* – chiamati *RAS-like* - e campioni di tiroide "non-tumorali". Esperimenti preliminari condotti *in vitro* in una linea cellulare di carcinoma papillare tiroideo, TPC-1, indicano che il silenziamento del lncRNA *MET-AS* induce una sotto-regolazione dell'oncogene *MET*, che induce un blocco del ciclo cellulare in fase G1. Ciò potrebbe suggerire che MET-AS sia un nuovo regolatore dell'oncogene *MET*.

In conclusione, i risultati ottenuti in questo lavoro di tesi confermano l'eterogeneità genetica del carcinoma papillare della tiroide rivelando che l'espressione genica correla più con il profilo mutazionale dei pazienti che con la stadiazione del tumore. Inoltre, questo studio fornisce nuove informazioni sulle alterazioni genetiche del PTC, suggerendo potenziali terapie adiuvanti farmacologiche per questo tipo di cancro.

1 Introduction

1.1 Cancer Genetics

Cancer arises as a result of deregulated cell growth. Essentially, the gradual accumulation of mutations in genes that regulate crucial cell processes, like cell cycle or DNA repair, is one of the proposed mechanisms accounting for the increased proliferation rate, which is a typical feature of cancer cells. The acquired genetic alterations are then transmitted to the next generation of cells, which can accumulate also other genetic alterations. Indeed, the vast majority of genetic alterations are acquired somatically.

In 1914, Boveri proposed firstly the hypothesis that cancer can arise from somatic alterations in DNA. He noted abnormal mitotic division and cell masses, very similar to tumors, in eggs of sea urchin fertilized by two sperms (Boveri, 1914). In the last 30 years significant experiments supporting this thesis have been performed in different fields, from molecular biology to epidemiology. Today, we know that the onset and the expansion of a malignant cell population result from multiple (perhaps five, ten or more) genetic alterations that occur in the transition of a cell from a normal to malignant phenotype. Such alterations, that in proof-of-principle can be both somatic or germline, can occur in three major classes of genes, i.e. oncogenes, tumor suppressor genes and DNA damage recognition/repair genes, which play key roles in tumorigenesis.

1.1.1 Proto-oncogenes and oncogenes

Proto-oncogenes are genes that drive normal cells to become cancerous when they are mutated (Adamson, 1987; Weinstein & Joe, 2006). Oncogenes are the mutated version of proto-oncogenes that typically carry dominant mutations, i.e. mutations affecting only one allele of the gene can be sufficient to activate the gene and trigger the neoplastic program. A broad spectrum of genes can be defined as proto-oncogenes even though most of them encode proteins involved in stimulation of cell division, inhibition of cell differentiation, and are responsible of halting cell death and apoptosis. All of these processes are crucial for normal tissues and organs development and maintenance. The oncogenes' activation involves a quantitative or qualitative gain of function. It can result from different genetic mechanisms, which can be schematized as follows: i) point mutations, deletions, or insertions that lead to a hyperactive gene product; ii) point mutations, deletions, or insertions in the promoter region of a proto-oncogene that lead to increased transcription; iii) gene amplifications that lead to extra chromosomal copies of a proto-oncogene; iv) chromosomal translocations that relocate a proto-oncogene to a new chromosomal site, possibly leading to a higher expression of the transposed gene; v) chromosomal translocations that lead to a fusion between a proto-oncogene and another gene, with the result of producing a chimeric protein with an oncogenic activity.

Activated oncogenes typically give rise to increased protein translation and/or activation, with a significant alteration of cellular processes such as increased cell division, decreased cell differentiation, and inhibition of cell death and apoptosis. All of these phenotypes are hallmarks of the cancer cells. Thus, oncogenes are currently a major molecular target for anti-cancer drug design (Chial, 2008).

Among the most extensively studied proto-oncogenes, *RAS* and *RAF*, which encod proteins involved in intracellular signalling, are frequently mutated in human malignancies, like melanoma or thyroid cancer (Bos 1989; Davies 2002). Mutations in these genes determine an increased activation of their protein products with the resulting over-stimulation of the mitogen-activated protein kinase (MAPK) pathway. Other frequent events are the germline mutations in RET proto-oncogenes in familiar medullary thyroid carcinoma (Eng 1999) and in MET gene, often mutated in papillary renal carcinoma.

1.1.2 Tumor suppressor genes

In 1988, Harris hypothesized that − in addition to oncogene activation − the loss of genetic material is a crucial event in tumorigenesis (Harris 1988). Tumor suppressor genes can be defined as genes encoding proteins that inhibit cell proliferation, or that act as the “brakes” for cell cycle. Other tumor suppressor genes encode proteins that promote the apoptosis or that are involved in cell differentiation. A tumor suppressor gene contributes to cancer when both alleles are inactivated by a mutation. Indeed, as long as the cell contains one functional copy of a given tumor suppressor gene, it can inhibit the formation of tumors. Therefore, mutations in tumor suppressor genes are recessive or loss-of-function mutations, and they are often point mutations or small deletions that disrupt the function of the protein encoded by the gene. The requirement of two mutations to promote tumorigenesis was proposed by Dr. Alfred Knudson in 1971 during his studies on retinoblastoma. Knudson proposed that sporadic cases of this tumor require the inactivation of both copies of a particular gene, the retinoblastoma gene (RB1). He formulated the "two-hit theory": a "first hit" inactivates one of the two copies of RB1. Later a "second hit" inactivates the remaining functional copy of RB1 in the same cell or one of its progeny (Knudson 1971). However, in the hereditary forms of cancer the first mutation is inherited from parents, thus only one somatic "hit" is necessary for tumor initiation.

1.1.3 DNA repair

DNA mutations are caused by copying errors during DNA replication, chemical and physical agents and can compromise DNA function. To protect the genome, mammalian cells employ at least eight distinct DNA repair pathways to cope with a multitude of different genotoxic lesions (Dietlein et al., 2014). In this network of genome maintenance pathways, the two major repair system consists of mismatch repair (MMR), and nucleotide-excision repair (NER). MMR recognizes erroneous insertions, deletions, and erroneous bases incorporations of bases. The MutS α (formed by Msh2/Msh6), and MutS β (formed by Msh2/Msh3) complexes detect small mismatches and large mismatches and insertion loops, respectively (Kunkel and Erie 2005). The MutL α complex (formed by MLH1 and PMS2) binds MutS and recruits the exonuclease Exo1; subsequently DNA Pol δ fills the lesion gap. NER is mainly responsible for repairing single or double strand breaks and helix-distorting lesions, which are induced by UV irradiation and platinum-based chemotherapeutics. Cells with defects in MMR encoding genes have a mutation rate 100-1000 fold higher than normal cells. Moreover, these MMR-defective cells display microsatellite instability. Microsatellites are repetitive genetic elements dispersed in the genome, with repeating units of 1-4 bases. Because of the repetitive nature of microsatellites they are prone to DNA polymerase slippage, which is efficiently repaired by the MMR. Defects in MMR result in increasing length and number of microsatellites, which have been observed in different types of cancer (Lengauer 1998). The important function of NER to protect against skin cancer becomes obvious by the rare genetic disease *Xeroderma Pigmentosum*, in which different NER genes are mutated. In animal models, it has been demonstrated that UVB is more effective to induce skin cancer than UVA when this repair system is mutated (Rass and Reichrath, 2008).

1.1.4 Cell cycle

Cellular life span is highly variable between different cell types. At a given time the absolute majority of cells are not dividing, but exists in a resting and metabolic active state called G_0. A cell enters in the cell cycle, duplicating its DNA and dividing itself, in response to external or internal *stimuli*. This process is supervised by checkpoint controls, which act to ensure that identical chromosome copies are transferred to the two daughter cells.

The cell cycle is divided into two basic parts: mitosis and interphase (Figure 1.1 B). Mitosis (nuclear division), or M phase, is the most dramatic stage of the cell cycle, corresponding to the separation of daughter chromosomes and usually ending with cell division (cytokinesis). However, approximately 95% of the cell cycle is spent in interphase, the period between mitoses. The M phase is followed by the G1 phase (gap 1), which corresponds to the interval between mitosis and initiation of DNA replication, which takes place in S phase (synthesis). This phase is followed by the G2 phase (gap 2), during which cell growth continues and proteins are synthesized in preparation for mitosis.

The progression of cells through the division cycle is regulated by extracellular signals from the environment, as well as by internal signals that monitor and coordinate the various processes that take place during different cell cycle phases (Figure 1.1 A). In addition, cell cycles have different checkpoints, which ensure the correct coordination of all phases. Indeed, it is critically important that the cell does not begin mitosis until replication of the genome has been completed. Progression through the cell cycle is also arrested at the G1 and G2 checkpoint in response to DNA damage. This arrest allows time for the damage to be repaired, rather than being passed on to daughter cells. In mammalian cells, arrest at the G1 checkpoint is mediated by the action of p53 protein. Interestingly, the gene encoding p53 is frequently mutated in human cancers. Loss of p53 function as a result of these mutations prevents G1 arrest in response to DNA damage, so the damaged DNA is replicated and passed on to daughter cells instead of being repaired.

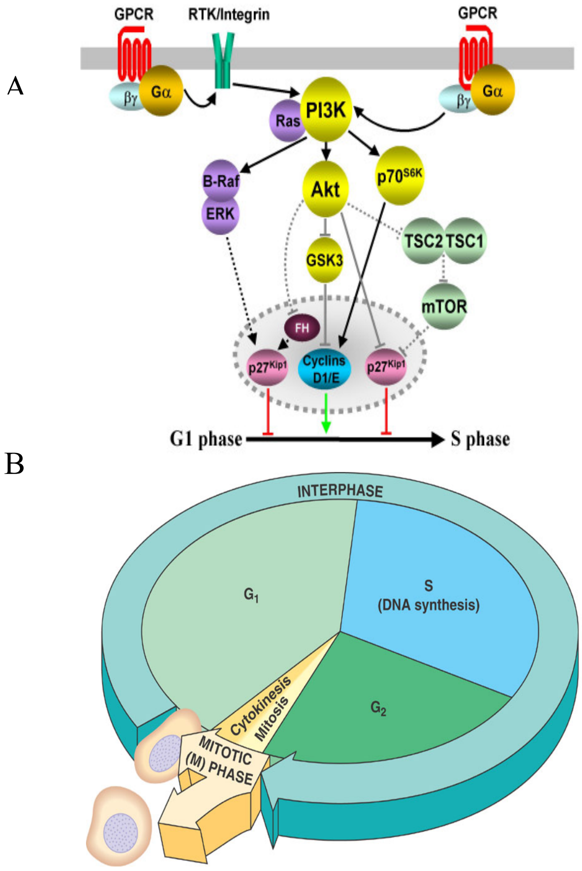


Figure 1.1. A) Gene products and pathways involved in induction of S phase from G1 phase. B) Schematic illustration of cell cycle.

1.1.5 Mutation timing and driver genes

Tumors evolve from benign to malignant lesions by acquiring different mutations over time, a process that has been particularly well studied in colorectal tumors (Nowell, 1976; Fearon and Vogelstein, 1990). The first, or "gatekeeping," mutation provides a selective growth advantage to a normal epithelial cell (Vogelstein et al., 2013). When a second mutation in another gene that can promote or "drive" tumorigenesis occurs, it causes a second round of clonal growth that allows an expansion of cell number (Figure 1.2). The mutations that confer a selective growth advantage to the tumor cell are called "driver" mutations. All the mutations that have no effect on the neoplastic process are called "passenger" mutations (Vogelstein et al., 2013). It has been estimated that a typical tumor contains two to eight of these "driver" gene mutations; the remaining mutations are passengers. Moreover, it is important to point out that there is a fundamental difference between a driver gene and a driver gene mutation. A driver gene is a gene that contains driver mutations (also defined as Mut-Driver gene) or is aberrantly expressed conferring a selective growth advantage to cells (Epi-Driver gene, Vogelstein et al., 2013).

For instance, *BRAF* is a well-known driver gene, but only mutations that result in increased kinase activity of the protein are considered driver mutations. An example is constituted by the V600E mutation that results in a valine (V) to a glutamic acid (E) substitution at position 600 in BRAF. Other missense mutations throughout the gene, as well as protein-truncating mutations in the C-terminal domain, are passenger gene mutations.

Numerous statistical methods to identify driver genes have been published. Some of them are based on the predicted effects of the mutation on the encoded protein (Carter et al., 2009; A. Youn et al., 2011; J. S. Kaminker et al, 2007). Other methods are based on other frequency of mutations in an individual gene in a specific tumor, compared with the mutation frequency of other genes, occurred by chance, in the same or related tumors after correction for sequence context and gene size (Parmigiani et al., 2009; M.

Meyerson et al., 2010). All of these algorithms useful for genes' prioritization are most likely to confer a growth advantage when mutated.

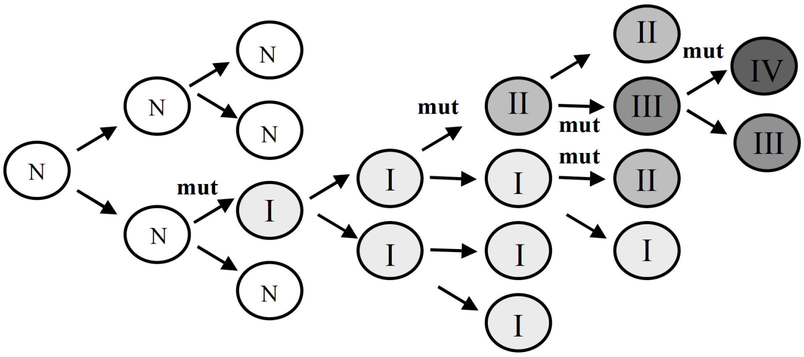


Figure 1.2. Multistep progression toward cancer. A normal cell (N) acquires one mutation (I), which provides the cell a growth advantage. It constitutes a substrate for accumulation of additional mutations resulting in cell clones with increased proliferative capacity (II-IV).

1.1.6 Signaling pathways in tumors

All of the known driver genes can be classified into one or more of 12 pathways (Figure 1.3). These pathways can be categorized into three main cellular processes: cell fate, genome maintenance and cell survival. Cell fate: is regulated by the inverse relationship between cell division and differentiation. Pathways that function through this process include, Hedgehog (HH) pathway, APC and NOTCH signaling, all of which are known to control cell fate. Genes involved in chromatin modifications can also be included in this category. Cell survival pathways are shown as regulators of cell metabolism and cell survival, but examples are also provided where aberrant activity of the pathway may contribute to the induction of apoptosis. *MYC, BCL2 RAS* and *BRAF* are driver genes that directly regulate progression through the cell cycle and apoptosis.

Genome maintenance: cells are frequently exposed to a variety of toxic substances, such as reactive oxygen species, o radiations. These events cause mistakes in DNA replication process or during division. Tumor cells with mutations in DNA damage control pathway, such as mutations that abrogate checkpoints genes – for instance *TP53* and *ATM* -, have a selective growth advantage compared to cells without these mutations.

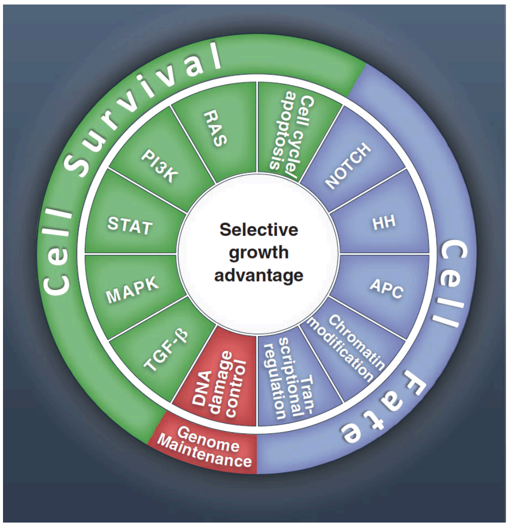


Figure 1.3. Cancer cell signalling pathways and the cellular processes they regulate. All of the driver genes can be classified into one or more of 12 pathways (middle ring) that confer a selective growth advantage (inner circle). These pathways can themselves be further organized into three core cellular processes (outer ring). *Figure from Vogelstein et al., 2013*

1.2 Thyroid Carcinoma

The thyroid is an endocrine gland located in the anterior region of the neck. It consists of two lobes connected with the isthmus. In a healthy adult the thyroid gland weights about 15-35 g, but it can considerably increase in size and weight in pathological conditions. The thyroid gland, which is the largest endocrine organ in humans, regulates systemic metabolism through thyroid hormones, with an important physiological role in skeletal development and brain, as well as in regulating the body's metabolism and the development of skin, subcutaneous tissue and organs.

It is composed of two distinct hormone-producing cell types, follicular and parafollicular C cells. Follicular cells comprise most of the epithelium and are responsible for iodine uptake and thyroid hormone synthesis, triiodothyronine (T3) and thyroxine (T4). The synthesis requires enzymatic activity provided by thyroxine peroxidase (TPO). The hormones are directly secreted in blood vessels bound to the thyroglobulin protein (Tg). T3 is the biologically active hormone and T4 can be converted to T3. C cells are scattered intrafollicular or parafollicular cells that are dedicated to the production of calcitonin, the calcium-regulating hormone.

The activity of thyroid gland is mainly regulated by the secretion of thyroid stimulating hormone (TSH) from the pituitary gland, which is regulated by the thyrotropin-releasing hormone (TRH) from the hypothalamus (Figure 1.4). TSH stimulates growth of follicular epithelium and the synthesis of thyroid hormones. In turn, thyroid hormones exert negative feedback control on the hypothalamus as well as on the anterior pituitary, thus controlling the release of both TRH from hypothalamus and TSH from anterior pituitary gland (Dietrich JW et al. 2012). This control mechanism is compromised in pathological conditions such as the presence of thyroid carcinoma.

Thyroid cancer is the most common malignancy of the endocrine system and accounts for approximately 1% of all newly diagnosed cancer cases (Carlomagno e Santoro, 2011).

In spite of thyroid carcinoma being an uncommon type of tumor, over the past decades, increasing attention has been focused on this malignancy. In 1986, after the Chernobyl nuclear accident, the incidence of thyroid cancer dramatically increased among children who lived in the regions contaminated with radioactive isotopes. This attracted the attention of medical experts, resulting in increased awareness of the disease. Moreover, the thyroid covers a broad spectrum of malignancies, ranging from well-differentiated carcinomas to undifferentiated tumors, thus, thyroid tumors provide an ideal model for studying tumorigenesis in epithelial tissue.

According to the National Cancer Institute, the rate of increase in the incidence of thyroid cancer among women in the United States is more rapid than for all other types of tumors. In the last 30 years it has tripled, and for reasons not yet well understood it is about 3-4 times more common in women than men (Brown et al., 2011). Indeed, this neoplasia is becoming the seventh most common tumor in women (Pillai et al., 2015).

Approximately 95% of thyroid cancers arise from follicular cells - which is the most common endocrine malignancy - and it can be distinguished in papillary carcinoma (PTC), follicular carcinoma (FTC), poorly differentiated thyroid carcinoma, and anaplastic thyroid carcinoma (ATC). Papillary carcinoma is the most common type of thyroid malignancy, comprising about 80–90% of all carcinomas (Ries et al., 2007; Davies et al., 2006).

FTC represents about 15% of thyroid cancers and is very common in geographic regions where the addition of iodine in the diet is inadequate. The tumor is characterized by the presence of follicular structures well defined; the lesion appears surrounded by a capsule but the cells are often able to invade and metastasize through the blood into districts furthest such as bones and lungs (D'Avanzo et al., 2004). Mutations in *RAS* genes are very frequent in this cancer, however, in 30% of cases, rearrangements involving the gene *PPARγ* have been found (Kroll et al., 2000).

ATC constitutes about ~ 2% of cases and is very aggressive. The prognosis is often unfavorable and death occurs after 6 months from diagnosis. This tumor is typically made up of spindle cells mixed in giant cells cancer cells, which have lost partially or fully differentiated phenotype. In more than 50% of the cases were detected mutations in the gene encoding the p53 protein, which plays a crucial role in regulating the cell cycle, DNA repair and apoptosis (Taccaliti and Boscaro, 2009).

The remaining 5% of the cases is represented by medullary thyroid carcinoma (MTC), which is a neoplasm arising from the calcitonin-producing C thyroid cells derived from neural crest (Mears L and Diaz-Cano, 2003; Skinner et al., 2005)*.* In the majority of cases (~ 75%), this cancer occurs as a sporadic tumor, whereas the remaining ones are part of familial disorders (Kloos et al. 2009). The hereditary forms are caused by a mutation in the "rearranged during transfection" *(RET)* gene, of which familial MTC (FMTC), MEN2A and MEN2B variants are discerned. Hereditary forms are transmitted with an autosomal rate in patients with an autosomal dominant pattern with high penetrance (>90%; Weels Jr 2000).

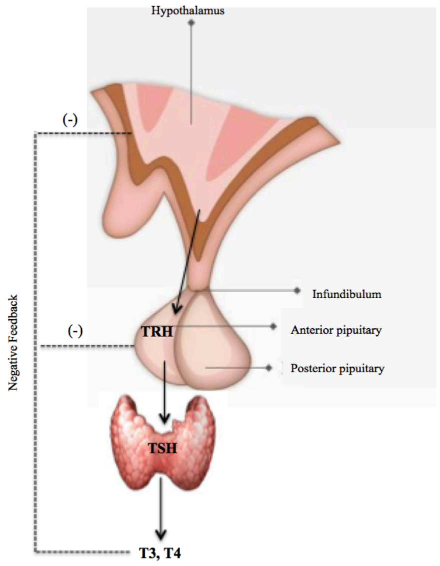


Figure 1.4. Hypothalamic-Pituitary-Thyroid Axis. Solid lines correspond to stimulatory effects, and dotted lines depict inhibitory effects. Conversion of T4 to T3 in the pituitary and the hypothalamus is mediated by 5′ -deiodinase type II. This event also is important throughout the central nervous system, thyroid, and muscle. 5′ -Deiodinase type I (propylthiouracil-sensitive) plays a major role in liver, kidney, and thyroid function. TRH, Thyrotropin-releasing hormone; TSH, Thyroid-stimulating hormone.

1.2.1 Papillary Thyroid Carcinoma

Papillary thyroid carcinoma is the most common malignant tumour of thyroid gland in countries having iodine-sufficient or iodine-excess diets, and comprises about 80–85% of thyroid malignancies. PTCs tend to be biologically indolent and have an excellent prognosis (survival rates of 95% at 25 years). Papillary carcinoma can occur at any age and rarely has been diagnosed as a congenital. Most tumors are diagnosed in patients in the third to fifth decades of life. Women are affected more frequently than men in ratios of 2:1 to 4:1 (Mazzaferri et al., 2002).

The gross appearance of papillary thyroid cancer is quite variable. The lesions may appear anywhere within the gland. By definition, typical papillary carcinomas often average 2–3 cm, although lesions may be quite large. The lesions are solid and usually white in color with an invasive appearance. Lesional calcification is a common feature. In addition, cyst formation may be observed. Indeed, some lesions may be rarely almost completely cystic making diagnosis difficult (Rosai et al., 1992; Carcangui et al., 1985).

Microscopically, papillary carcinomas share common features. The neoplastic papillae contain a central core of fibro-vascular (occasionally just fibrous) tissue lined by one or occasionally several layers of cells with crowded oval nuclei (LiVolsi et al., 2011; Hawk W and Hazard J; 1976).

Papillary thyroid tumors will be composed mostly of papillary areas (Figure 1.5), but a large number will also contain follicular areas. The tumor cells are usually cuboidal or columnar. About the 80% of such lesions contain clear nuclei, in the 80–85% are seen intranuclear inclusions, whereas, nuclear grooves are seen in almost all the cases (Baloch et al., 2008; Scopa et al., 1993; Deligeorgi-Politi H, 1987).

Moreover, psammoma bodies, formed by calcium deposits, are found in about 40-50% of cases, which within the cores of papillae, in the tumor stroma, or in lymphatic vessels, but not within the neoplastic follicles (LiVolsi, 2011). The evidence of psammoma bodies in a cervical lymph node is indicative of a papillary carcinoma in the thyroid. Psammoma bodies are rare in benign thyroid (only 1% of psammoma bodies are in benign glands).

The primary tumor can invade lymphatic vessels leading to multifocal lesions and to regional node metastases. Whether the lymphatic invasion itself causes metastases in distinct *foci* within the thyroid or whether these *foci* represent independent clonal proliferations is still debated. On the other hand, venous invasion is rare; indeed, metastases outside the neck are unusual, occurring only in 5–7% of cases, predominately in lung and bones (LiVolsi et al., 2011). Despite the presence of multiple metastases, in ordinary papillary carcinoma, death is uncommon.

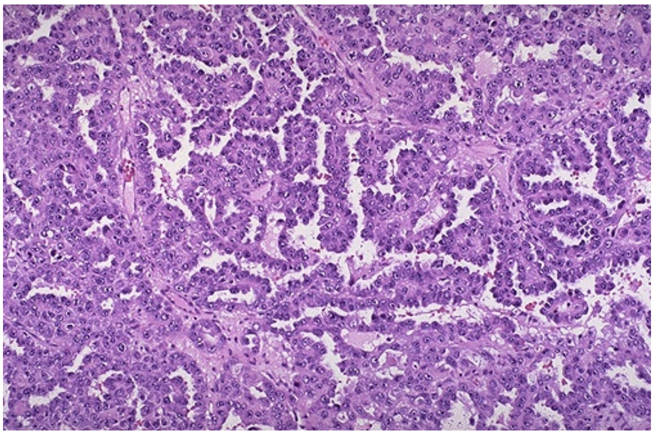


Figure 1.5. Microscopic appearance of a papillary carcinoma of the thyroid. The fronds of tissue have thin fibrovascular cores. The fronds have an overall papillary pattern.

1.2.2 Molecular genetics alteration in Papillary Thyroid Carcinoma

During the past decade there has been an increasing number of publications about genetic alteration in thyroid tumors (Xing 2013). More than 70% of PTCs carry mutations in two genes coding for Mitogen-Activated Protein Kinase (MAPK) signaling pathway effectors - a serine-threonine kinase, *BRAF* and a GTP-binding protein, *RAS* – and rearrangements in two tyrosine kinases receptors – *RET* and *NTRK1* (also known as *TRKA*), which play a role in the regulation of growth, differentiation and programmed cell death of neurons in the peripheral and the central nervous system (Teng and Hempstead, 2004). These alterations are mutually exclusive in PTCs patients, suggesting that the alteration, leading to the constitutive activation, in one of these genes is sufficient for cell transformation and hyper-activation of MAPK pathway and is essential for papillary tumor initiation (Santoro and Carlomagno, 2013; Kimura et al., 2003; Soares et al., 2003; Frattini et al., 2004).

BRAF gene

BRAF gene, located on the long arm of chromosome 7, encodes a serine/threonine protein kinase involved in the epidermal growth factor receptor (EGFR)-mediated MAPK pathway, where it is activated by RAS small GTPase (Lavoie H and Therrien M, 2015). Moreover, BRAF can affect other key cellular processes, such as cell migration (through RHO small GTPases), apoptosis (through the regulation of BCL-2), and survival (through the HIPPO pathway; Matallanas D et al., 2011). Thus, it is not a surprise that *BRAF* is constitutively activated by mutation in 15% of all human known cancer types. Several mutations, affecting different regions of the protein, have been identified. However, despite more than 40 mutations have been so far identified in *BRAF* gene, the vast majority (up to 80%) of mutated BRAF-related tumors carry the 1799T>A (Davies et al., 2002). This mutation frequently occurs in thyroid cancer (Xing 2005) and causes V600E (valine with glutamic acid) amino acid change in the BRAF protein, resulting in the constitutive BRAF kinase activation, with a high oncogenic ability (Davies et al., 2002; Dhomen et al., 2007; Fukushima et al., 2003; Wan et al., 2004). BRAF V600E mutation can both initiate tumorigenesis in normal thyroid follicular cells and maintain and promote thyroid cancer progression (Nucera et al., 2009).

RAS gene

RAS genes encode proteins involved in key intracellular signal transducers that can activate several downstream pathways, RAF-MEK-ERK and PIK3K pathway (Peyssonnaux C and Eychene, 2001). Mutations in *RAS* genes – *NRAS*, *HRAS*, and *KRAS* – usually occur in codons 12, 13 or 61 of any of the three genes. These alterations - common in FTA and FTC and less frequent, in PTC (Bos, 1989) - produce constitutively active RAS proteins. The mutations associated to PTC predominantly involve codons 61 of NRAS and, to a less extent, of HRAS (Vasko et al., 2003; Zhu et al., 2003; Di Cristofaro et al., 2006)

RET/PTC rearrangements

The rearrangements of *RET* involve its fusion to heterologous genes (Nikiforov, 2002) and result in constitutive activation of tyrosine kinase domain leading to the formation of tumorigenic chimeric proteins.

The *RET* proto-oncogene is localized on chromosome 10q11.2 and is composed of 21 exons spanning a region of 55,000 bp. It encodes a single-pass trans-membrane tyrosine kinase receptor (Takahashi, 1988). It is constituted by three functional domains: an extracellular ligand binding domain, a hydrophobic transmembrane domain, composed of 22 amino acids, among which S649 and S653 mediate self-association and dimerization of RET, and an intracellular tyrosine kinase (TK) domain. RET has several autophosphorylation sites (Kawamoto et al. 2004). RET tyrosine 1062 (Y1062) is a multidocking site for signalling molecules, which, in turn, contribute to the activation of RAS-MAPK and PI3K (phosphatidyl inositol 3 kinase)-AKTpathways. These signalling cascades is involved in cell survival, proliferation, and motility (Alberti et al. 1998; Murakami et al. 1999; Segouffin-Cariou and Billaud 2000; Melillo et al. 2001).

The ligands of RET receptor are neurotrophic growth factors belonging to the glial cell line-derived neurotrophic factor family (GNDF) (Sugg et al., 1998). GNDF binding causes receptor dimerization, autophosphorylation of Y1062, and activation of the signalling cascade.

In the thyroid gland the *RET* gene is expressed at high levels only in the parafollicular cells, but not in the follicular cells in which it can be activated by the fusion of the 3’ portion of the RET gene (from exon 12 to the 30-end) to the 5’ portion and to the promoter sequence of to various heterologous genes (Grieco et al. 1990; Nikiforov and Nikiforova 2011). Such chromosomal aberrations result in chimeric oncogenes, known as RET/PTC. To date, at least 11 different RET/PTC fusions have been reported. Common RET/PTC rearrangements (90% of the cases) are RET/PTC1 and RET/PTC3, involving *RET* and *CCDC6* or *NCOA4* genes (both localized in chromosome 10), respectively (Santoro and Carlomagno, 2013).

RET/PTC1, RET/PTC3 are generated through a paracentric inversion of the long arm of chromosome 10 (Grieco et al. 1990; Santoro et al. 1994). Other RET/PTC variants are either rare (RET/PTC2) or identified only in single cases of radiation-induced PTC, and are generated by translocations between different chromosomes (Figure 1.6).

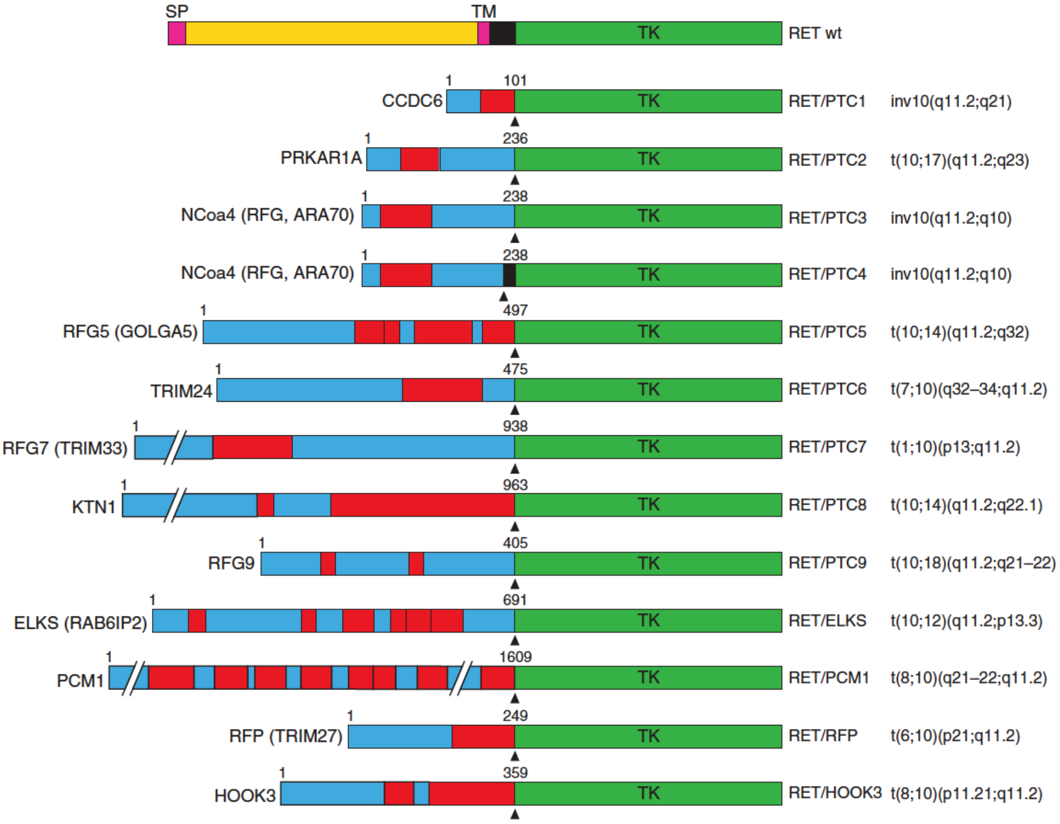


Figure 1.6. Schematic representation of RET/PTC oncoproteins. On the top, wild-type RET protein is illustrated. For each RET/PTC rearrangement, the name of the fusion partner is indicated on the left and the corresponding chromosomal alteration is indicated on the right. The fusion points are indicated by arrowheads. The length in amino acids of the partner protein portion is also indicated. Boxes in red indicate dimerization (coiled-coil) domains. SP, Signal peptide; TM, transmembrane domain; TK, tyrosine kinase domain.

RET/PTC fusions are tumorigenic in follicular cells; indeed, they transform thyroid cells in culture (Santoro et al., 1993) and give rise to thyroid carcinomas in transgenic mice (Santoro et al., 1996).

RET/PTC is found in 20–40% of adult sporadic papillary carcinomas, but the percentages are even higher among children affected by papillary carcinomas after the Chernobyl nuclear accident (about 80% of tumors).

PAX8/PPARG and NTRK1 rearrangements

Oncogenic rearrangements of *PPARG* and *NTRK1* genes are also found in PTC. *PPARG*, localized on chromosome 3, encodes a member of the peroxisome proliferator-activated receptor (PPAR) subfamily of nuclear receptors. Translocations involving the DNA-binding domains of the transcription factor *PAX8* (2q13) and the A-to-F domains of the peroxisome proliferator-activated receptor γ (PPARG1) were found by Kroll and co-workers in FTC (Kroll et al., 2000). This event causes the loss of proper *PAX8* and *PPARG* transcriptional function in the rearranged *PAX8/PPARG* fusion, that can act as dominant-negative (Kroll et al., 2000).

PAX8/PPARG fusions were detected in FTC but not in FTA, PTC, or multinodular hyperplasias, but recently they have been also described in relatively high percentage of cases of the follicular variant of PTC (37.5%, Castro et al., 2006).

The *NTRK1* gene, localized in chromosome 1, codes for nerve growth factor (NGF) receptor, and its activation has been linked to the activation of the RAF-MEK-ERK pathway (Miller and Kaplan, 2001). *NTRK1* rearrangements are rare, usually found in less than 10% of cases of sporadic PTC (Musholt et al., 2000; Kuo et al., 2000). The most common fusion type was between exon 4 of *ETV6* gene and exon 14 of *NTRK3*, significantly more common in tumors associated with exposure to (131) I from the Chernobyl accident (Leeman-Neill RJ et al, 2013). The rearrangement results in a fusion protein constituted by of SAM domain of ETV6 and the tyrosine kinase domain of NTRK3, which lead to a constitutively active tyrosine kinase (Lannon and Sorensen, 2005).

1.3 RNA-Sequencing

Ten years ago, the idea that all of the genes altered in cancer could be identified at base-pair resolution would have seemed like science fiction. Today, such genome-wide analysis, through exome, whole genome, or transcriptome sequencing is ordinary.

The introduction of Next-Generation Sequencing (NGS) technologies has significantly impacted cancer research (Costa et al., 2013; Hoadley et al., 2014).

High-throughput sequencing technologies are widely used in biomedical research. Indeed, NGS technologies overcame many of the limitations dictated by previous technologies, such as cross-hybridization background, signal saturation-induced and range limitation in array technology (Costa et al., 2013; Hoadley et al., 2014). Moreover, these high-throughput technologies produce complex datasets at single nucleotide resolution and at reduced cost, offering the opportunity to investigate on a wide scale both genomics, epigenetics and transcriptional aspects of cells and tissues, in a deep and comprehensive manner.

In May 2008, five articles introducing a new technique that has been upsetting microarray were published online on Science, Cell, Nature and Nature Methods. The method named RNA-Sequencing (RNA-Seq) provides higher resolution snapshot of the transcriptome than what was the standard before.

RNA-Sequencing is perhaps one of the most complex next-generation applications. It consists in a set of experimental procedure that starting from entire RNA molecules generates cDNA sequences, followed by library construction and massively parallel deep sequencing. It allows in a single experiment to analyse expression levels, differential splicing, allele-specific expression, RNA editing and fusion transcripts for both coding and non coding RNAs in disease-related studies (Costa et al., 2010).

Gene expression is known to be tissue-, cell-type-, time- and stimulus-dependent, and many transcripts are only expressed under very specific conditions. RNA-Seq allows the quantification of abundance level of each transcript during defined developmental stages, under specific treatment or in physiological and/or pathological conditions (Costa et al., 2013). In contrast to microarray, it is not limited to the interrogation of selected probes on an array and can be also applied in species, for which the whole reference genome is not assembled yet.

Moreover, RNA-Seq can also be exploratory. Recently, it was appreciated that 85% of the human genome is transcribed and in contrast, only 2-3% of the genome encodes protein-coding genes (Hangauer et al., 2013), and a lot of other non-coding RNAs classes have been discovered. For instance, RNA-Seq allowed the discovery of a novel class of long non coding RNAs, named “enhancer RNA”, a class of transcript directly transcribed from the enhancer region, involved in epigenetic gene regulation. In addition, RNA-Seq allows the analysis of transcriptional start sites (TSSs) revealing alternative promoter usage, and premature transcription termination at the 3’ of untranslated regions (UTRs), which is critical from mRNA stability (Griffith M et al. 2010; Picardi et al., 2010; Wang 2008).

More recently, RNA-Seq has been used to identify allele-specific expression, disease-associated single nucleotide polymorphisms (SNPs) and mutation, as well as gene fusions and alterations involved in cancer pathogenesis (Maher et al., 2009; Supper et al., 2013).

This technology relies heavily on deep sequencing which means that every RNA molecule in the samples is sequenced hundreds or thousands of times (Meyerson, et al., 2010). In general, RNA population (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments with adaptors attached to one (single-end sequencing) or both ends (paired-end sequencing). Each molecule is then sequenced in a high-throughput manner to obtain short sequences typically ranging from 30 to 400 bp, depending on the DNA-Sequencing technology used. Various sequencing platforms are supported including Illumina, Life Sciences, Roche 454, Applied Biosystems and Helicos Biosciences. After sequencing, the resulting reads can be both aligned to a reference genome or reference transcriptome or assembled de novo if the genomic sequence in unknown (Wang et al., 2009).

1.3.1 Illumina sequencing technology

In the project described in this PhD thesis we have used Illumina technology for the sequencing of RNA samples. The single molecule amplification step for the Illumina starts with an Illumina-specific adapter library, takes place on the oligo-derivatized surface of a flow cell, and is performed by an automated device called a Cluster Station (Figure 1.7). Illumina sequencing is based on standard dideoxy method. cDNA fragments are immobilized on a surface of a flow cell to produce multiple DNA copies, or clusters, that each represent the single molecule that initiated the cluster amplification (Metzker et al., 2009).

This system utilizes a sequencing-by-synthesis approach in which the flow cell channels receives a DNA polymerase cocktail with different fluorescently labelled nucleotides A, T, G and C. Different fluorescent molecules are attached to the four nucleobases that thus emit four different wavelengths. Specifically, the nucleotides carry a base-unique fluorescent label and the 3′

-OH group is chemically blocked such that each incorporation is a unique event. The cycles of sequencing are regulated by this block so fluorescent signal can be read correctly. After each imaging step, the 3′ blocking group is chemically removed to prepare each strand for the next incorporation. This series of steps continues for a specific number of cycles, as determined by user-defined instrument settings, which permits discrete read lengths of 25–35 bases. A base-calling algorithm assigns sequences and associated quality values to each read and a quality-checking pipeline evaluates the Illumina data from each run (Metzker et al., 2008; Figure 1.7).

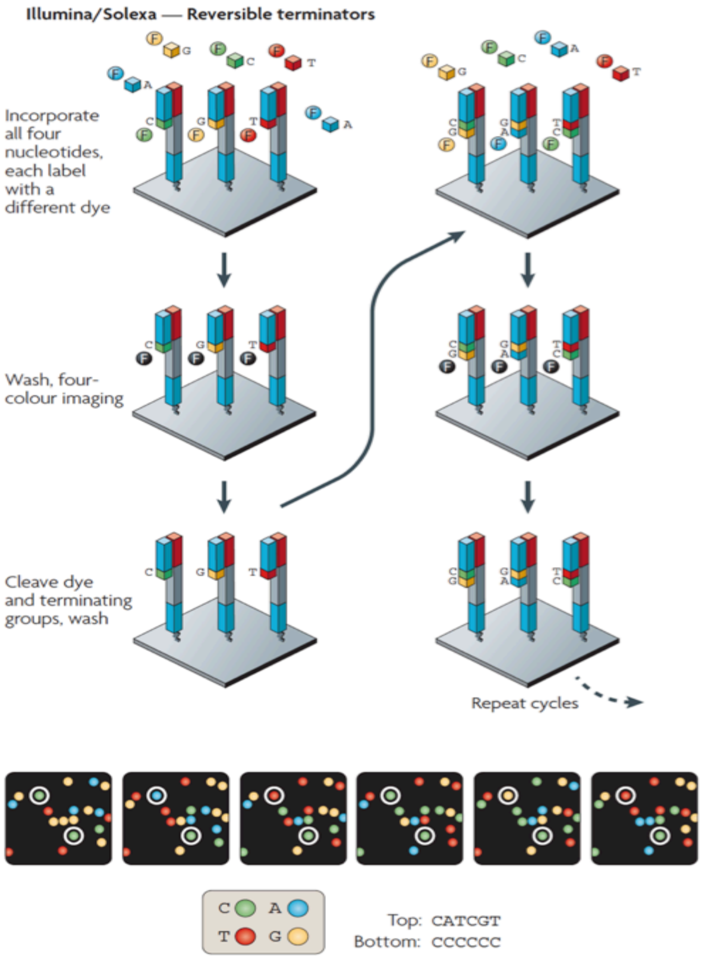

Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
C A
T G
Top: CATCGT
Bottom: CCCCCC

Figure 1.7. Four-colour and one-colour cyclic reversible termination methods. A) The four-colour cyclic reversible termination (CRT) method uses Illumina/Solexa’s 3′-O-azidomethyl reversible terminator chemistry using solid-phase-amplified template clusters. Following imaging, a cleavage step removes the fluorescent dyes and regenerates the 3′-OH group using the reducing agent tris(2-carboxyethyl)phosphine (TCEP). B) The four-colour images highlight the sequencing data from two clonally amplified templates. *From Metzker, 2010.*

1.3.2 RNA-Sequencing data analysis

The unprecedented level of data produced by NGS platforms requires a considerable effort in the development of new bioinformatics tools to deal with these massive data files since data analysis is very expensive in term of memory and computational time. The RNA-Seq data generation is an ever-evolving process, which requires a parallel development in sequencing technologies, experiment designs, and computational algorithms. In light of this, bioinformatics tools with improved performances are emerging constantly.

After image and signal processing the output of a RNA-Seq experiment consists of ten to thousand of millions of short reads (or raw reads). The raw reads are the starting material of the computational analysis that include quality assessment, reads mapping, gene quantification and differential gene expression – in a standard RNA-Seq analysis – and/or alternative splicing identification/quantification, variants' calling and gene fusion detection, depending on the experimental purpose.

Quality assessment

Since raw reads derive from a multiple-step process involving sample preparation, fragmentation, amplification, and sequencing, the quality assessment represents the first step of the bioinformatics workflow of RNA-Seq. Often, it is necessary to filter data, trimming low-quality bases, adaptors, or overrepresented sequences to remove undesirable biases in the analysis.

Reads mapping

Once high-quality data are obtained from pre-processing, the next step consists of mapping the sequence reads to a reference genome (and/or to known annotated transcribed sequences) if available, or *de novo* assembling to produce a genome-scale transcriptional map. This procedure refers to the classic bioinformatics problem of obtaining the more accurate mapping possible in a speed- and memory-efficient manner. The introduction of algorithms that are based on transcriptome mapping before a genome mapping step avoid erroneous mapping of the reads to pseudogenes, generally improving the overall alignment accuracy.

Reads counting

The number of RNA-Seq reads that map to a gene is a direct measure of the gene’s expression level. Such approach can both help in quantifying known elements (i.e., genes or already annotated exons) and/or in detecting new transcribed regions, defined as transcribed segments of DNA that are not yet annotated as exons in public databases (Costa et al., 2010). This step provides the expression of a given gene as the total number of reads mapping to the coordinates of each annotated element. After getting the read counts, data normalization is one of the most crucial steps in data processing, as it is essential to ensure accurate inference of gene expression. Reads counts can be normalized for the length of the transcribed element and the number of mapped reads for each sample. Marioni and colleagues proposed a quantitative normalized measure of gene expression, i.e. the Reads Per Kilobase per Million of mapped reads (RPKM), to compare both different genes within the same sample and the same gene across distinct biological conditions (Marioni et al., 2008).

Differential expression analysis

An important application of RNA-Seq is transcriptomes' comparison between pathological and physiological conditions, across different developmental stages, or between specific experimental *stimuli*. This type of analysis requires identification of genes and/or transcripts with different expression through the comparison of two or multiple samples (Costa et al., 2010). It is essential for interpreting the functional elements of the genome and uncovering the transcriptome complexity, providing important insights in the biological mechanisms of development and diseases.

Detection of fusion genes

Gene fusions have gained attention because of their relationship with cancer. Different tools have been developed to analyze fusion events in RNA-Seq data.

After the reads' mapping step, there will be a pool of unmapped short reads (i.e reads not mapping within an exon or to exon-exon junctions). These unmapped reads can be processed by specific algorithms to determine whether they match an exon-exon junction where the exons come from different genes. This would be evidence of a possible fusion event.

Detection of nucleotide variants

RNA-Seq data are generally used to study gene expression, or to perform novel gene/isoforms' identification and quantification. However, very recently RNA-Seq data have been also used for the identification of expressed mutations, especially in tumor samples. In this approach there are many limitations, such as the unbalanced coverage between different genes. Among many variants calling and annotation methods the best practical workflow to identify mutations from RNA-seq data has been provided by GATK although it is still far from perfect and under heavy development (http://gatkforums.broadinstitute.org/discussion/3891/calling-variants-in-rnaseq).

1.3.3 RNA-Seq in Papillary thyroid carcinoma

In recent years, the introduction and the rapid development of new technologies for the sequencing of nucleic acids, has revolutionized the study of cancer genetics (Costa et al., 2010; Costa et al., 2013). In particular, RNA-Seq has provided a tool to simultaneously investigate all genomic as well as transcriptome alterations occurring in the same cancer cells.

These technologies have recently been used to study mutations in different types of cancers, in order to improve diagnostic and prognostic abilities, as well as the individual response to treatment.

Recent NGS-based studies have explored the mutational landscape and gene expression profiles of PTCs. Smallridge and colleagues performed RNA-Seq to explore the transcriptome of BRAF- and not BRAF-mutated PTCs.

Such analysis revealed different gene expression between the two groups of patients; they found that about 50 of differentially expressed genes were related to immune functions. Moreover, through NGS they also identified 4

fusion genes in PTC samples (i.e. *CKLF3-CMTM4*, *ETV6-NTRK3*, *MKRN1-BRAF* and *PPIP5K1-CATSPER2*). Notably, *CKLF3-CMTM4*, *ETV6-NTRK3*, *MKRN1-BRAF* gene fusions have been found in three different not BRAF-mutated PTC samples, indicating that these may potentially represent new driver events, although with a very rare occurrence (Smallridge et al., 2014).
Similarly, Leeman-Neill and colleagues performed RNA-Seq to identify new chromosomal rearrangements in patients exposed to ionizing radiations. They found that ETV6-NTRK3, RET/PTC and PAX8-PPARγ rearrangements are significantly more common than point mutations in PTCs associated with exposure to ^{131}I.
Recently, the seminal work of The Cancer Genome Atlas (TCGA) Research Network has explored more than 400 PTCs (Cancer Genome Atlas Research Network 2014). In this study, the authors have described a comprehensive multiplatform analysis of the genetic landscape of PTC, performed by SNP arrays, exomes, RNA-Seq, miRNA-Seq, DNA methylation and targeted sequencing. One of the most significant advances was the identification of somatic mutations (single nucleotide variants, INDELSs and gene fusions) as potential new tumor-initiating events - i.e. the "dark matter" - in patients without any known driver lesion. In particular, the authors identified *EIF1AX, PPM1D* and *CHEK2* as new driver genes in PTC, and also discovered *TERT* promoter mutations in a subset of aggressive and less-differentiated PTCs, strongly correlated to a high risk of recurrence. Gene and miRNA expression analysis also allowed defining clinically relevant subclasses potentially correlated to loss of differentiation and tumor progression (e.g. over-expressed miR-21 in association with aggressive tall cell variant of PTC).

1.4. Long Non coding RNAs

Over the past decade, evidence from numerous high-throughput genomic platforms suggests that complexity of the organism is mainly due to the expansion of the non coding portions of the genome (Mattick, 2004). Indeed, the portion of the genome responsible for protein coding constitutes approximately 2%, while many noncoding regulatory elements are transcribed into non coding RNA (ncRNA). Non-coding RNAs are divided into two major classes based on their size: 1) small ncRNAs (sncRNAs, 20-30 nt) which are critical post-transcriptional regulators of target RNAs via RNA interference (RNAi), and/or able to modify other RNAs, including the widely-studied class of microRNAs (miRNAs), piwi-interacting RNAs (piRNAs) and small nucleolar RNAs (snoRNAs), and 2) the heterogeneous group of long ncRNAs (lncRNAs).

With the term lncRNA we define a class of transcripts longer than 200 bp without the protein coding capacity (Derrien and Jhonson et al., 2012).

These transcripts are characterized by relatively low levels of evolutionary conservation, fewer exons than protein coding genes in average, and high tissue-specificity (Guttman and Rinn, 2012; Kapranov et al., 2007; Clark and Mattik, 2011). On the other hand, they exhibit some similarities with protein coding transcripts; for instance, they are transcribed by RNA polymerase II and can be capped, polyadenylated and spliced. They can localize both in *nucleus*, acting mainly as epigenetic modulators, and in cytoplasm, where they can act as post-transcription regulators (Fatica and Bozzoni, 2014; Vance and Ponting, 2014).

According to the GENCODE Consortium (Derrien and Jhonson et al., 2011), lncRNAs can be classified with respect to protein-coding genes in "antisense" (if they intersect protein-coding loci on the opposite strand), "lincRNA" (long intergenic non-coding RNA), "sense overlapping" (that overlap intron and exon of a coding gene on the same strand), "sense intronic" (within the intron of a coding gene on the same strand), "processed transcript" (without ORF and not classified in the other categories because of their complexity; Figure 1.8)

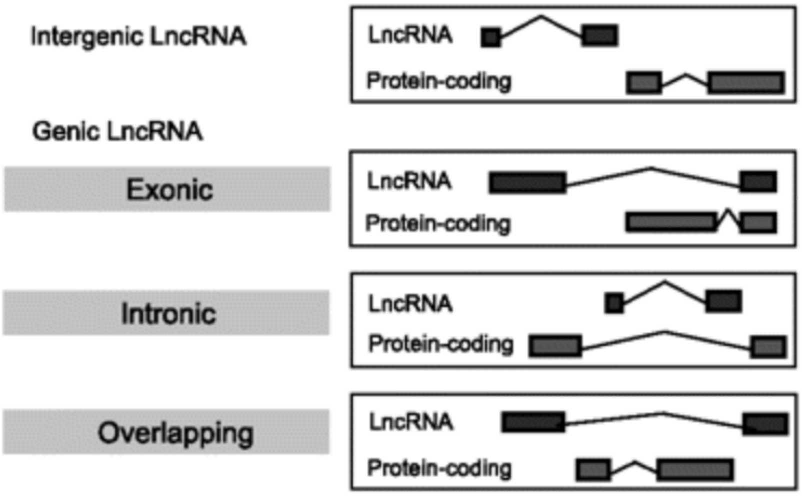


Figure 1.8. LncRNAs classification according to GENCODE catalogue (Derrien et al., 2011)

While the nomenclature is still evolving, lncRNAs typically refers to polyadenylated lncRNAs that are transcribed by RNA polymerase II and associated with epigenetic signatures common to protein-coding genes, such as trimethylation of histone 3 lysine 4 (H3K4me3) at the transcriptional start sites (TSSs) and trimethylation of histone 3 lysine 36 (H3K36me3) throughout the gene body (Guttman et al., 2009).

1.4.1 Functions and mechanisms of lncRNAs

Like protein-coding genes, long ncRNAs cover a broad spectrum of functions. Compared with coding transcripts, lncRNAs are expressed at 10-fold lower levels on average, and their expression in different tissues and cell types has generally been found to be more cell type specific (Clark and Blackshaw, 2014).

Long ncRNAs are involved in transcriptional regulation of mRNA processing, which is reminiscent of miRNAs and may indicate a similar sequence-based mechanism to miRNA binding to seed sequences on target mRNAs. However, unlike miRNAs, long ncRNAs show a wide spectrum of biological contexts that demonstrate greater complexity to their functions.

They can act as positive and negative modulators of gene expression (Numata and Kiyosawa, 2012; Su et al., 2012; Johnsson et al., 2013), involved in different functions, such as X inactivation (Brown et al., 1992; Lee 2009), imprinting, epigenetic regulation (Gupta et al., 2010; Tollervey et al., 2012) and can affect any step within the biogenesis or the mobilization of target mRNA, including transcription, splicing, nuclear and cytoplasmic trafficking and translation (Chen and Carmichael, 2010).

LncRNAs can impact genes localized on the same chromosome (*cis*-acting lncRNAs) or on other chromosomes (*trans*-acting lncRNAs); but their function is strictly related to their localization, in nucleus, where they can modulate gene expression at pre-transcriptional, co-transcriptional and post-transcriptional level, or in cytosol, where they act at post-transcriptional level (Figure 1.9).

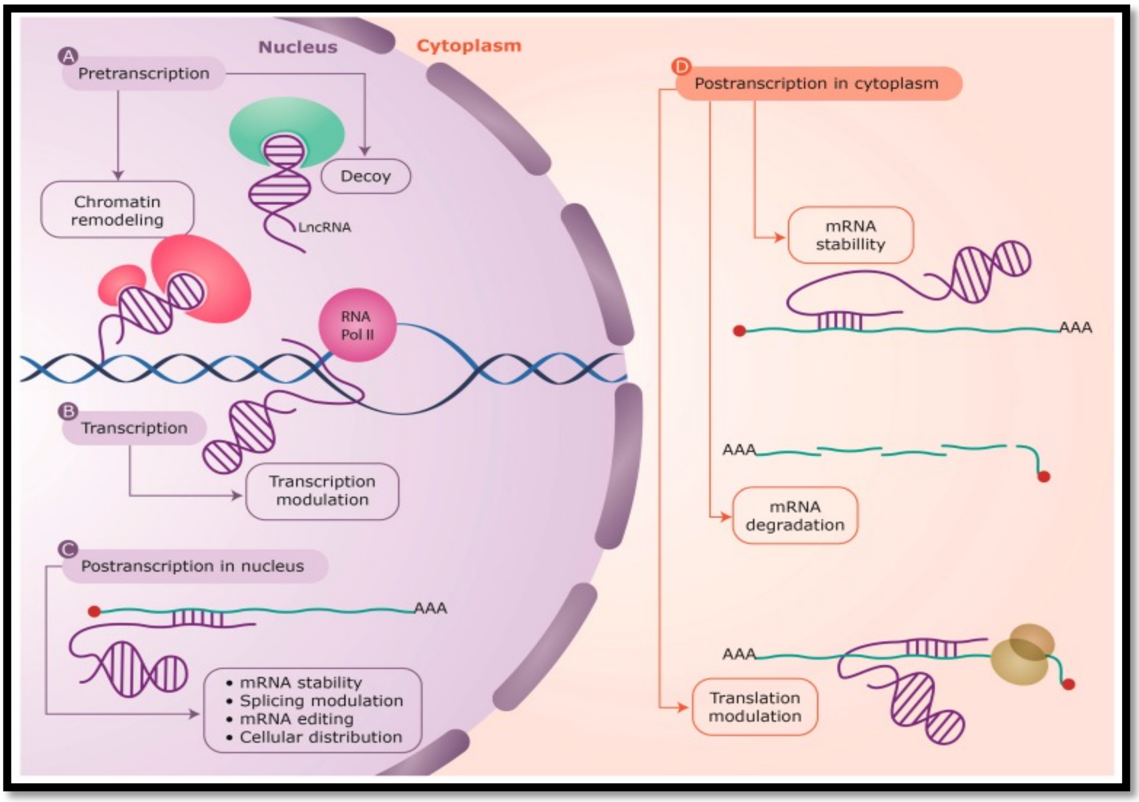


Figure 1.9. LncRNAs have been found to act at every level of gene regulation: **A**) Pretranscriptional, as protein guides or acting as decoys holding proteins away from chromatin; **B**) Transcriptional, as modulators of transcription; **C**,**D**) Posttranscriptional, altering sense mRNA structure or cellular compartmental distribution either in the nucleus or the cytoplasm. LncRNAs are depicted in purple, with the interacting protein factors in green and light red. The mRNAs are shown as green lines and the base pair interactions highlighted by short purple lines. Also shown is the transcribing RNA polymerase II (RNA Pol II) on genomic

DNA (blue helix) and the translating ribosome (yellow) on the mRNA. Figure from Villegas and Zaphiropoulos, 2015.

Epigenetic transcriptional regulation

Studies about nuclear lncRNAs have mainly focused on their potential epigenetic regulation of target genes. Such a regulation typically results in gene transcriptional repression or activation. The first class of lncRNAs to be characterized has been the one containing lncRNA with repressive functions, including *ANRIL*, *HOTAIR*, *H19*, *KCNQ1OT1*, and *XIST* (Gibb et al., 2011; Yap et al., 2010; Rinn et al., 2007). These lncRNAs achieve their repressive function by coupling with histone modifying or chromatin remodelling protein complexes.

The most common protein partners of lncRNAs are the polycomb repressive complexes 1 and 2 (PRC1 and PRC2). These complexes can transfer the repressive post-translational modifications to specific amino acid positions on histone proteins, thereby inducing chromatin folding and heterochromatin formation in order to repress gene transcription. PRC1 may be comprised of numerous proteins, including BMI1, RING1, RING2 and Chromobox (CBX) proteins, which act as a multi-protein complex to ubiquitinate histone H2A at lysine 119 (Margueron and Reinberg, 2011). PRC2 is classically composed of EED, SUZ12, and EZH2, the latter of which is a histone methyltransferase enzymatic subunit that trimethylates histone 3 lysine 27 (Margueron and Reinberg, 2011). Both *EZH2* and *BMI1* genes are up-regulated in numerous common solid tumors, leading to tumor progression and aggressiveness (Margueron and Reinberg, 2011).

Even if, PRC1 and PRC2 are perhaps the most known partners of lncRNAs, numerous other epigenetic complexes are implicated in lncRNA-mediated gene regulation. For example, the 3’ domain of *HOTAIR* contains a binding site for the LSD1/CoREST, a histone deacetylase complex that facilitates gene repression by chromatin remodelling (Tsai et al., 2010). Similarly, *AIR* interacts with G9a, an H3K9 histone methyltransferase (Nagano et al., 2008).

LncRNAs have been also observed in activating epigenetic complexes. For instance, *HOTTIP* interacts with WDR5 to mediate the recruitment of the MLL histone methyltransferase to the distal HoxA *locus*. MLL transfers methyl groups to H3K4me3, generating open chromatin structures that promote gene transcription (Wang et al., 2011).

Moreover, lncRNAs can act as scaffold, serving as central platforms upon which relevant molecular components are assembled. By this mechanism, a lncRNA would bind its multiple effector partners, forming ribonucleoprotein complexes, which may have transcriptional activating or repressive activities, at the same time in the same space (Figure 1.10).

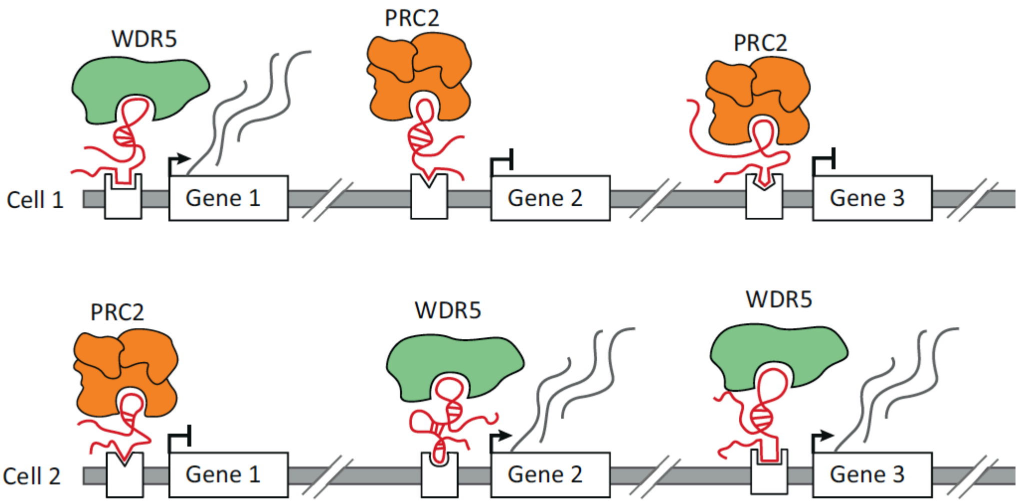


Figure 1.10. Long noncoding RNA (lncRNA)-mediated regulation of gene expression through the recruitment of chromatin regulatory proteins. (A) Different cell types express distinct lncRNAs that can differentially recruit these same chromatin regulatory proteins, including the repressive Polycomb Repressive Complex 2 (PRC2) and the activating WDR5 chromatin-modifying protein, to specific genes. Inset: lncRNAs can recruit these complexes by binding to target sites through three mechanisms: tethering to its nascent transcription locus (top panel); directly hybridizing to genomic targets (middle panel); or interacting with a DNA-binding protein (bottom panel). From Quinodoz and Guttman, Trends in Cell Biology, 2014

lncRNAs in post-transcriptional regulation: mRNA processing, stability, and translation

LncRNAs also play a rule in post-transcriptional processing of mRNAs, which is also critical to gene expression. LncRNAs have been implicated in alternative splicing; for instance, Bernard and colleagues have shown that *MALAT1* (metastasis-associated lung adenocarcinoma transcript 1) localizes serine/arginine (SR) splicing factors to a compartment called nuclear speckles, which are postulated to serve as storage sites for mRNA prior to its export to the cytoplasm for translation (Bernard et al. 2010). *MALAT1* is associated with proper relocation of these splicing factors to sites where splicing occurs, and thus may have a role in controlling alternative splicing of target mRNA precursors (Tripathi et al., 2010).

Another lncRNA, Gomafu/MIAT, which localizes to a novel nuclear domain and has a neuron-restricted expression, may hinder spliceosome formation and affect the splicing of a subset of mRNAs by sequestering splicing factor 1 (SF1) (Sone et al. 2007; Tsuiji et al., 2011).

Moreover, lncRNAs may even affect translational regulation. PU.1, an important TF involved in hematogenesis, has an overlapping natural antisense that was found to negatively influence PU.1 protein level but not mRNA level. The antisense RNA seems to compete with the sense transcript for binding to the translation initiation factor eIF4A, decreasing the translation (Ebralidze et al. 2008).

In a more intricate example, the protein Staufen1, STAU1, a RNA degradation protein, is involved in the regulated decay of ~1% of coding transcripts, a process named “STAU1-mediated decay”. The mechanism of action of STAU1 involves lncRNAs containing ancestral Alu repeats. It was discovered that a subset of target mRNA contains Alu element in its 3’ UTRs that can base pair with a group of cytoplasmic and polyadenylated lncRNAs, named half-STAU1-binding site RNAs (1/2-sbsRNAs), to form the double stranded RNA structure that then recruit STAU1 to implement RNA degradation (Gong and Maquat 2011).

Finally, Pandolfi and colleagues recently suggested another model for mRNA regulation. According to their theory, transcribed pseudogenes, including *PTENP1* and *KRASP1*, serve as decoy for miRNAs that target the protein-coding mRNA transcripts of their cognate genes (Poliseno et al., 2010). Thus, pseudogenes can regulate the gene expression level of the protein coding mRNA indirectly, having miRNA-binding sites in their 3' UTRs and may therefore serve as "sponges" to sequester miRNAs away from their mRNA targets.

This model, named "competing endogenous RNAs" (ceRNAs) model, more broadly suggests that all long ncRNAs, as well as other protein-coding mRNAs, may function as molecular "sponges" that bind and sequester miRNAs in order to control gene expression indirectly (Figure 1.11).

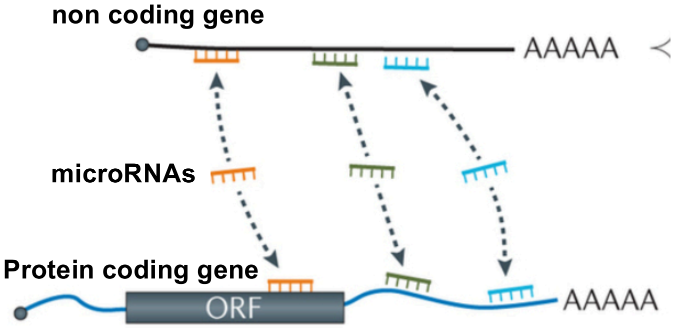


Figure 1.11: Base pairing is also the mode of action of competing endogenous RNAs. In this case, the complementarity is between microRNAs (miRNAs) and different targets (including circular RNAs (circRNAs), lncRNAs, pseudogene transcripts and mRNAs). The effect of these interactions is that protein-coding RNAs and non-coding RNAs can crosstalk to each other by competing for miRNA binding through their miRNA recognition motifs. ORF: Open reading Frame. *Modified from Fatica and Bozzoni, 2014.*

1.4.2 Methods to discover lncRNAs

ncRNAs have historically been more difficult to detect than protein-coding genes. This is largely because ncRNAs are more tissue-specific in their expression (Cabili et al., 2011), exhibit lower levels of expression (Trapnell et al., 2010; Guttman et al., 2010), and cannot be predicted by computational algorithms scanning for open reading frames (ORFs) in the human genome.

However, the advance of high-throughput RNA profiling methods has enabled more precise and accurate cataloguing of ncRNAs. While the Human Genome Project emphasized only protein-coding genes in their computational analysis of DNA, groups investigating RNA had long observed a great number of unannotated transcripts (Matsubara et al., 1993; Liang et al., 2000). Many of these transcripts were discovered cloning and sequencing – by Sanger method – Expressed Sequence Tags (ESTs), that identify fragments of genomic regions that were being actively transcribed. In addition to sequencing advances, new technologies (in particular DNA microarray) were emerging to identify new genes and to understand the regulation of gene expression. In parallel, the first complete human chromosome 22 sequence was released in 1999 (Dunham et al., 1999). The combination of microarrays and draft genome sequences provided the first insight into pervasive transcription of noncoding RNAs. Two independent studies estimated that there might be as many lncRNA genes as protein-coding genes (Kapranov et al, 2002; Rinn et al., 2003).

A major advance came when Guttman et al. combined the microarray technologies with the logic of epigenetics (Guttman et al., 2009). Here, the authors reasoned that ncRNAs could have the same structure and epigenetic characteristics of protein-coding genes. Thus, lncRNAs could be polyadenylated and spliced and they could have a gene promoter marked by H3K4me3 and a gene body marked by H3K36me3. By using ChIP-Seq data of these epigenetic marks as well as RNA polymerase II data, the authors observed by DNA tiling arrays several thousand regions of unannotated transcription overlapping these epigenetic marks (Guttman et al., 2009). These new transcriptional entities were defined lncRNAs, based on their length and the lack of a coding potential.

The advent of RNA-Seq led to the ability to sequence all RNA species in a cell at an unprecedented scale and throughput (Costa et al., 2010). In addition to full-length reconstruction algorithms, several applications have emerged from RNA-Seq. For instance, a method termed 3-Seq targets and sequences the polyadenylated tail of cDNA to quantitatively measure the abundance of transcripts (Beck et al., 2010). Moreover, a variant of this method can be employed to precisely map 3’ ends of transcripts (Jan et al., 2011). These and other emerging technologies are providing deeper insights into the dynamic transcriptome.

RNA-Seq is now the gold standard method to discover lncRNAs. Recent studies employed RNA-Seq to identify distinct classes of large RNA genes. For instance, Cabili and colleagues identified 8,000 lincRNAs in the human genome by integrating different annotation sources in combination with RNA-Seq data (Cabili et al., 2011). Furthermore, in 2011 Derrien and colleagues released the most complete human lncRNA annotation, produced by the GENCODE consortium within the ENCODE project and comprising 9277 manually annotated genes producing 14,880 transcripts (Derrien and Jhosnon et al., 2011).

1.4.3 lncRNA and cancer

Recent researches point out the need to expand the tumor-suppressor and oncogenes classes to non coding RNAs (ncRNAs), defined as ‘tumor-suppressor ncRNAs’ and ‘oncogenic ncRNAs’.

Indeed, numerous profiling and characterization studies of a well investigated ncRNA class, i.e. the microRNAs (miRNAs), have identified key roles for ncRNAs in cancer (Ruan et al., 2009; Calin et al., 2007).

Alterations in miRNAs expression levels have been linked to the initiation and progression of different human cancers. Furthermore, miRNA-expression profiling in human tumors has identified signatures associated with diagnosis, prognosis, staging, progression, and specific treatment (Cho et al., 2007) miRNAs can act as tumor-suppressor or oncogenes depending on their target genes (Zhang et al., 2007).

In addition to the well-characterized miRNAs, the growing knowledge of the mammalian non-coding transcriptome is revealing the presence of thousand of lncRNAs, which could have a major role in the development and progression of cancer, although their mechanisms of function remain less well understood (Ponting et al., 2009; Mercer et al., 2009; Guttman et al., 2009; Wang et al., 2010).

The lncRNA have a key regulatory role in gene expression and, therefore, it has been speculated that they may be involved in the etiology of various diseases. To date, the most studied pathogenic mechanisms in which are involved lncRNA are related cancer (Tsai et al., 2011). Indeed, an altered expression of many long non-coding RNA has been reported in different types of tumors (Rinn and Chang, 2012).

Besides genetic alterations of protein coding tumor suppressor or oncogenes, recent evidences indicate that epigenetic alterations can also contribute to tumor transformation and cancer (Jones P.A., Baylin, 2007).

Chromatin-regulatory complexes are linked with the aberrant proliferation of cancer cells. For instance, SUZ12, a subunit of polycomb repressive complex 2 is overexpressed in colon and breast cancers and EZH2 is up-regulated in many tumors, including Hodgkin lymphoma, prostate and breast cancer (Simon J.A., Lange, 2007). Collectively, these findings point to an important interplay between ncRNAs, chromatin regulation and cancer, representing a new dimension in our understanding of cellular transformation.

1.4.4 Oncogenic long non coding RNAs

Similar to protein-coding oncogenes, long ncRNAs can also promote cellular pathways that lead to tumorigenesis. An example of oncogenic long intergenic RNA (lincRNA) is *HOTAIR*. It is expressed from the HOXC locus and negatively regulates HOXD genes. This repressive regulation is conferred by the interaction between *HOTAIR* and PRC2 complex (Rinn et al., 2007; Figure 1.12 panel A). *HOTAIR* was found significantly overexpressed in breast tumors (Gupta et al., 2010). Furthermore, its expression level in primary breast tumors is considered a predictor of patient's outcomes such as metastasis formation and death (Gupta et al., 2010). Thus, *HOTAIR* underscores the importance of understanding the relationship between epigenetic regulation by ncRNAs and cancer, and demonstrates how oncogenic lncRNAs can drive the epigenetic machinery to reshape the epigenetic landscape leading to cancer.

In addition, global transcriptome analyses have shown that up to 70% of protein-coding transcripts have antisense genes. Interestingly, the perturbation of the antisense RNAs can alter the expression of their sense genes (Katayama et al., 2005). Some of these genes encode tumor-suppressor or oncogenic proteins that can become epigenetically silenced or hyper-activated by the antisense ncRNA. Thus, misregulation of these antisense transcripts can lead to cellular transformation.

An example is provided by the antisense ncRNA ANRIL, which controls expression of the INK4A/ARF locus comprising the tumor-suppressor genes INK4n/ARF/INK4a, p16/CDKN2A and p15/CDKN2B. ANRIL mediates gene silencing of the locus by interaction and recruitment of CBX7, a component of polycomb repressive complex 1 (PRC1), histone 3 lysine 27-methyltransferase complex (Yu et al., 2008; Yap et al., 2010; Figure 1.12 panel B).

For instance, in prostate cancer ANRIL overexpression results in the down-modulation of INK4n/ARF/INK4a and p15/CDKN2B.

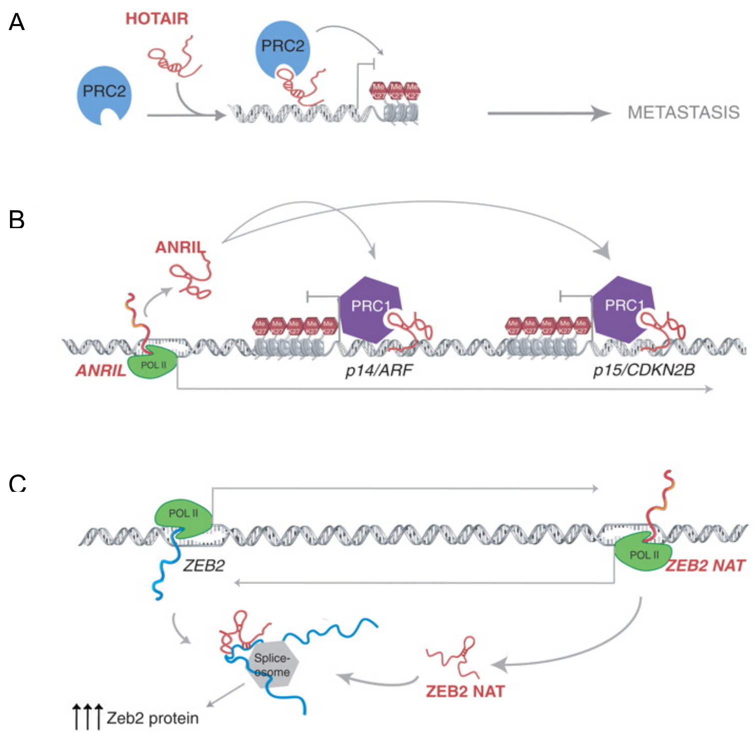


Figure 1.12. Mechanisms of gene regulation by oncogenic long ncRNAs. **A)** lincRNA HOTAIR recruits PRC2 to specific gene promoters for methylation of lysine 27 of histone 3 (H3K27me), inducing gene repression that leads to breast tumor metastasis. **B)** Large ncRNA ANRIL is transcribed antisense of the p14/ARF and p15/CDKN2B genes. ANRIL mediates gene silencing of the locus by interaction with, and recruitment of, CBX7, a component of PRC1 histone 3 lysine 27-methyltransferase complex. **C)** The ncRNA expressed antisense of the Zeb2 gene (Zeb2 NAT) overlaps with the 5' splice site of one Zeb2 intron. Zeb2 NAT inhibits the splicing of the intron, which contains an IRES sequence. In this way, Zeb2 protein translation is upregulated. *Modified from Huarte and Rinn, 2010*.

Moreover, some antisense transcripts can also fine tune gene expression at the post-transcriptional level. E-cadherin encoding gene (*CDH1*) is correlated with cancers of different organs, such as stomach, breast, colon, thyroid and ovary. Its down-modulation has been linked to cancer progression by increasing proliferation, invasion and/or metastasis (Berx and van Roy, 2009). A strong association has been demonstrated between the expression level of a particular natural antisense transcript (NAT) and human tumors with low E-cadherin expression (Beltran et al., 2008). NAT overlaps with an intronic 5' splice site of the *Zeb2* gene and prevents its splicing. The retained intron contains an internal ribosome entry site (IRES) necessary for the increased translation of Zeb2 protein, which can subsequently function as a transcriptional repressor of E-cadherin (Beltran et al., 2008) (Fig. 1.12 panel C). Collectively, these studies provide strong *impetus* for further investigation of antisense ncRNAs in cancer as they are likely to modulate the expression and/or the function of other genes involved in cancer etiology or progression.

MALAT1 is another lncRNA whose deregulation has been associated with cancerous process. This lncRNA has a predominantly nuclear localization and regulates alternative splicing of several genes, interacting and modulating the activity of factors involved in this mechanism, such as the nuclear family SR phosphoprotein (rich in serine and arginine). It has been observed that in metastatic lung cancer (NSCLC, non-small-cell lung cancer) biopsies, the expression of *MALAT-1* is about three-fold higher than in tumors that do not metastasize (Ji et al., 2003). So, although it is not yet clear how *MALAT-1* can modulate and alter SR proteins' phosphorylation, or what a mechanism governs its contribution to cancer development, it has been speculated that altered *MALAT-1* expression is a prognostic marker for the development of metastases and for patients' survival (Tseng et al., 2009).

1.4.5 Tumor-suppressor ncRNAs

“Tumor-suppressor lncRNAs” are lncRNAs that control protein-coding genes involved in tumor-suppressor pathways, and when their function is compromised, cells are prone to develop cancer. An example is provided by lincRNAs that are induced by the p53 tumor-suppressor pathway (Guttman et al., 2009; Huarte et al., 2010). Indeed, under stress conditions, p53 coordinates a tumor-suppressor program, by activating or silencing the expression of different genes. Among them there are many lncRNAs, in particular lincRNA-p21, that interact with the protein hnRNP-K to orchestrate transcriptional programs that maintain cellular homeostasis.

Another tumor-suppressor ncRNA is the lncRNA CCND1-associated. It is involved in the regulation of cyclin D1 (*CCND1*) gene expression. Cyclin D1 is a cell cycle regulator and it has been found frequently mutated or overexpressed in different tumors (Diehl 2002). In response to DNA damages, this ncRNAs is transcribed from the 5' regulatory regions of *CCND1* gene, mediating its transcriptional repression. This mechanism involves the interaction between the lncRNA with TLS protein - which is a sensor of DNA damage -, inducing its allosteric modification. In turn, this conformational change allows the association of TLS to the *CCND1* gene promoter, which inhibits transcriptional induction by histone acetyltransferases such as CBP and p300 (Wang et al., 2008).

Also the lncRNA *GAS5* (Growth Arrest-Specific 5) is a tumor-suppressor gene that plays a role in normal cell growth arrest (Kino et al., 2010). Reduced expression of *GAS5* in cancer breast cancer cells compared to healthy epithelial cells has been documented (Mourtada-Maarabouni et al., 2009). This down-regulation has been linked to the alteration of normal apoptotic process. Indeed, *GAS5* shows some regions that are able to bind to the DNA binding domain (DBD) of the glucocorticoid receptors (GR), preventing their interaction with the recognition sequences at the DNA level. Specifically, the GAS5 RNA conformation mimics gluticorticoid responsive element (GRE) DNA, acting as a sponge for GR, blocking their ability to bind gene promoters to induce their transcription, thus to induce the expression of cIAP2 (cellular inhibitor of apoptosis 2) and caspase 3, 7 and 9, involved in inhibition of the apoptotic process (Kino et al., 2010; Figure 1.13). Therefore, the increase of *GAS5* expression may induce an increase of apoptosis, while the down-regulation this lncRNA, found in breast cancer, can induce an inhibition of the apoptotic process.

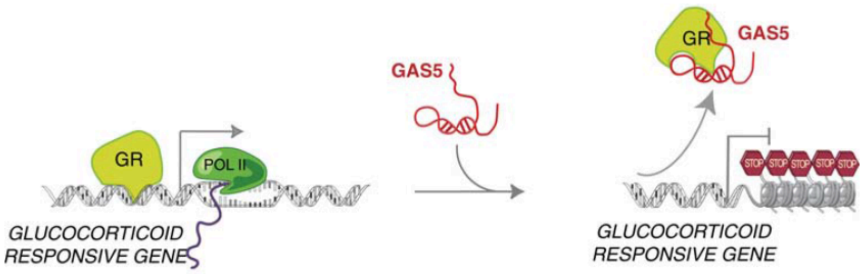


Figure 1.13. GAS5 mimics the conformation of DNA GREs, binding to GR. In this manner, GR loses the ability to activate transcription of target genes.

1.4.6 lncRNAs and papillary thyroid cancer

As already described, it has been shown that lncRNAs play a crucial role in several types of tumors, also in papillary thyroid cancer. Several studies have focused on the role of lncRNAs in papillary thyroid carcinoma pathogenesis (Jendrzejewski et al., 2012; Wang et al., 2014; Yoon et al., 2007).

For instance, Jendrzejewski and colleagues identified a lncRNA, named Papillary Thyroid Cancer Susceptibility Candidate 3 (*PTCSC3*), highly specific of thyroid tissue and down-modulated in cancerous tissues.

Interestingly, this lncRNA is located 3.2 kb downstream the single nucleotide polymorphism (SNP) rs944289 on chromosome 14q.13.3 (Jendrzejewski et al., 2012). In a previous study, the SNP rs944289 has been significantly associated - by genome-wide association studies (GWAS) - to papillary thyroid cancer (Li and Wang, 2012)

Moreover, Fan and colleagues have shown that the over expression of this non-coding RNA in thyroid cell lines inhibits cell proliferation and induces growth arrest and apoptosis (Fan et al., 2013).

It is not clear how *PTCSC3* fulfills its functions, but it has been hypothesized that it may act, according to the ceRNA model, as RNAs competing with oncogenic protein-coding genes for the binding to endogenous miRNAs. This hypothesis is partially supported by the inverse correlation between the expression levels of *PTCSC3* and the one of a miRNA with a proven oncogenic role in other cancers, i.e. miR-574-5p. Therefore, *PTCSC3* would affect the distribution of this miRNA and it may in turn regulate the growth of cancer cells (Fan et al., 2013).

Moreover, a gene expression study by Yoon and colleagues identified genes down-modulated in tumors (PTC) *vs* paired non-tumor tissue, including a novel lncRNA, named *NAMA* (noncoding RNA associated with the MAPK pathway and growth arrest) in patients carrying the mutation BRAF V600E. *NAMA* expression is highly associated with induction of cell cycle arrest (Yoon et al., 2007).

Similarly, the long non-coding *BANCR* (BRAF-activated lncRNA) is strongly associated with BRAF V600E mutation (Wang et al., 2014). In their study

Wang and colleagues demonstrated that *BANCR* expression levels are elevated in PTC tissues compared to healthy counterparts. Moreover, *in vitro* analysis confirmed the role of *BANCR* in inhibiting the apoptosis and increasing proliferation in cell cultures (Wang et al., 2014).

Therefore, the study of lncRNA in thyroid cancer is a fascinating field to outline new markers for PTC diagnosis and prognosis, and to discover potentially new therapeutic targets.

2 Principal aim of the project

Well-differentiated papillary thyroid carcinoma (PTC) constitutes about 85% of all thyroid malignancies. To date, different genetic alterations have been reported in PTC, among which high frequent point mutations in two genes of the Mitogen-activated Protein Kinase (MAPK) pathway, *BRAF* and *RAS* (*HRAS*, *KRAS* and *NRAS*). Moreover, different kind of gene fusion have been identified, i.e. genomic rearrangements involving *RET* gene (RET/PTC) in about 70% of PTC cases, TRK oncogenes rearrangements and *PPARG* gene fusions in ~5% of cases. However, despite the presence of tumor-initiating driver events, cancer results from the progressive accumulation of mutations in genes that confer growth advantage over surrounding cells. Thus, a deeper genetic characterization of PTC will improve clinicians' ability to establish diagnosis and to predict prognosis and individual response to treatments.

Moreover, although the vast majority of studies have focused on the role of protein coding genes in cancer, Next Generation Sequencing technology has revealed that about the 80% of transcription in mammalian is associated to to non-coding RNAs (ncRNAs), implicated in a wide spectrum of biological functions. In particular, in recent years, deregulated long non coding RNAs (lncRNAs) expression has been reported in many cancers, highlighting that they may act as potential oncogene or tumor-suppressor. A better understanding of the mechanisms that control synthesis and activity of these ncRNAs opens new frontiers in molecular oncology.

In this *scenario*, the principal aim of my PhD project is the identification of new genetic alterations and potential biomarkers in papillary thyroid carcinoma. Since RNA-Sequencing (RNA-Seq) technology has revolutionized cancer research, improving our ability to investigate tumor mutations' landscape, we used this technology to explore PTC transcriptome. The project has been organized in two main parts, which have included both bioinformatics analysis and *in vitro* studies.

The first part focuses on a comprehensive analysis of papillary thyroid carcinoma transcriptome data deriving from the sequencing of PTC and healthy thyroid biopsies. This section is structured into two objectives:

1) The identification of fusion transcripts in PTC through the analysis of RNA-Seq data;
2) The identification of missense mutations that could be implicated in cancer.

The second part aims to investigate the involvement of long non coding RNAs in papillary thyroid cancer, through the following objective:

1) Definition of lncRNAs expressed in PTC and healthy thyroid biopsies. The analysis aims to study not only annotated lncRNAs, but also new putative ones, and to address if and how their expression is significantly altered in patients with PTC, and therefore if they are potential candidate in the papillary thyroid carcinoma etiology.
2) Identification of new still unexplored connections between lncRNA and protein-coding genes, with a particular focus on their impact on already known oncogenes and/or oncosuppressors.

3. Identification of new somatic mutations and WNK1-B4GALNT3 gene fusion in papillary thyroid carcinoma

3.1 Methods

3.1.1 Patients and RNA samples preparation

Thyroid biopsies of PTC and healthy control thyroid were obtained from the Service d'Anatomo-Pathologie, Centre Hospitalier Lyon Sud, France and kindly provided by Prof. Alfredo Fusco of IEOS, CNR. Informed written consent was obtained from patients of both cohorts. The entire Project, funded by AIRC to Prof. Alfredo Ciccodicola, was approved by the ethic committee of University of Naples "Federico II" and Lyon Sud Hospital Center. Total RNA was extracted from biopsies using Trizol standard procedure and RNA integrity was assessed using digital gel electrophoresis (Experion®) and spectrophotometry (NanoDrop®).

3.1.2 Library preparation and RNA-Sequencing data analysis

PolyA+ paired-end libraries were prepared using TruSeq RNA Sample Preparation Kit (Illumina) according to manufacturer's instruction and sequenced on Illumina HiSeq2000.

Details about the bioinformatics analysis are provided below.

Read's quality assessment

Reads' quality was evaluated using FastQC software, freely available for the download at (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Briefly, FastQC is a stand-alone interactive application for the immediate analysis of a small number of FastQ files. It provides a quality check report, with different outputs, indicating the reads' quality. In particular, the "per base sequence quality" output shows an overview of the range of quality values across all bases at each position in the FastQ file. The "per sequence quality score" output report allows to check if a subset (or the entire set) of sequenced reads has globally low quality values. In the samples processed in this PhD Thesis, both the "per base" and the "per sequence" quality graphs revealed an overall high quality of the sequenced reads, trimming procedure was not necessary in our case.

Mapping reads with TopHat2

Reads were aligned using TopHat2 v2.0.10 (Kim et al., 2013), a gapped aligner capable of discovering splice junctions *ab initio,* based on using Bowtie2 (Langmead et al., 2012). It provides major accuracy improvements over previous versions and over other RNA-seq mapping tools. Briefly, in a first step, TopHat2 was used to map RNA-Seq reads against a known transcriptome annotation (Gencode v19). The transcriptome-mapping step improves the overall sensitivity and accuracy of reads mapping, avoiding the unwanted alignment of reads to the pseudogenes, very abundant in the human genome. Indeed, the presence of processed pseudogenes, from which some or all introns have been removed, may cause many exon-spanning reads to map incorrectly. This is particularly acute for the human genome, which contains over 14,000 pseudogenes (Pei et al., 2012). Then, TopHat2 aligns unmapped or potentially misaligned reads against the human genome (hg19), potentially representing unannotated transcripts or genes deriving from fusion events. TopHat2 outputs the reads that successfully map to either the genome or the splice junction reference in SAM format for further analysis. In this Thesis, the SAM files generated by TopHat have been converted to BAM format (binary version of SAM) using SAMtools (Li et al., 2009).

The parameters used for the mapping with TopHat2 were: -p 12 -N 2 -g 10 -r 200 -a 15 -m 1 -i 100 --library-type fr-unstranded --fusion-search --segment-mismatches 3 --read-edit-dist 2 --transcriptome-index.

Gencode v19 track downloaded from UCSC Table Browser (http://genome.ucsc.edu) was used as reference for transcripts mapping and quantification. Only uniquely mapped reads (about 95% of sequenced reads, Table 3.1) and with a maximum of 2 mismatches, were used for further analyses. Coverage files (bedgraph format) were produced using BEDTools v2.17.0. Visual inspection of reads and coverage files on UCSC Genome Browser was used to assess the overall quality of the RNA-Seq experiment, and to inspect gene-specific features of interest.

Table 3.1 Summary of data generated from paired-end RNA-Sequencing and results of mapping procedure.

Sample id	N° of reads	N° of fragments	Uniquely mapped reads	% of uniquely mapped reads
S110	107.168.970	53.584.485	97.499.398	91.0
S111	129.387.234	64.693.617	116.552.563	90.1
S112	123.056.226	61.528.113	112.010.077	91.0
S113	111.741.882	55.870.941	101.768.707	91.0
S114	112.356.482	56.178.241	100.998.996	89.9
S115	132.342.772	66.171.386	120.456.084	91.0
S116	109.612.414	54.806.207	99.716.914	91.0
S117	118.011.794	59.005.897	104.854.674	88.8
S118	151.418.292	75.709.146	137.861.418	82.0
S119	168.167.376	84.083.688	149.849.046	89.1
S120	134.011.794	68709146	121.027.091	90.3
S121	51.495.606	25.747.803	46.953.037	91,2
S122	64.766.780	32.383.390	59.983.274	92,6
S123	44.472.128	22.236.064	41.329.664	92,9
S124	64.547.635	32.273.818	59.009.673	91,4
S125	53.749.964	26.874.982	49.701.258	92,5
S126	58.995.744	29.497.872	54.237.927	91,9
S127	64.557.056	32.278.528	59.909.442	92,8
S128	70.098.721	35.049.361	65.036.396	92,8
S129	54.383.090	27.191.545	49.965.488	91,9
S130	97.514.456	48.757.228	89.589.008	91,9
S131	135.610.671	67.805.335	125.073.775	92,2

3.1.3 Analysis of single nucleotide variants

Variant calling for the discovery of single nucleotide variations has been performed using GATK best practices recommendations for calling variants on RNA-Seq data (http://gatkforums.broadinstitute.org/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail). These recommendations are based on classic DNA-focused Best Practices, with key differences in the early data processing steps (focus on handling splice junctions correctly), as well as in the calling step (Van der Auwera et al., 2013).

Add read groups, sort, mark duplicates, and create index.

After the mapping step, it is necessary to add read group information, sort reads, mark duplicates and create index. All these steps have been performed by using Picard tools v1.93 (http://broadinstitute.github.io/picard/).

Split'N'Trim and reassign mapping qualities.

Next, we used a new GATK tool called “SplitNCigarReads” developed specially for RNA-Seq reads, which splits reads into exon segments and hard-clip any sequences overhanging into the intronic regions.

Base Recalibration.

We performed base quality recalibration (BQSR), by using GATK tool. This tool recalibrates base quality scores of sequencing-by-synthesis reads in aligned BAM files. After recalibration, the quality scores in the quality field in each read in the output BAM are more accurate.

Variant calling.

For variant calling step we used HaplotypeCaller tool (GATK), which take into account the information about intron-exon split regions that is embedded in the alignment BAM file by SplitNCigarReads tool.

Variant filtering.

In order to filter the resulting callset, variants with clusters of at least 3 SNPs that were within a window of 35 bases and variants with a Quality By Depth values (QD < 2.0) were filtered out. Moreover, the variants with a recalibrated score < 30 and the predictions supported exclusively by variants located in the beginning or the end of the reads were filtered out.

Variant annotation.

The final filtered list of high quality variants was processed using ANNOVAR, an efficient command line Perl program to functionally annotate genetic variants from high-throughput sequencing data (Wang et al., 2010). In order to remove germline variants we initially filtered out common population variants from in dbSNP v138 (http://www.ncbi.nlm.nih.gov/SNP/), 1000 Genome Project data (The 1000 Genomes Project Consortium, 2012) and SNVs identified through the above-described procedure in normal healthy thyroids. Moreover, we removed nucleotide variants located in super-duplicated regions. However, we retained those variants annotated as somatic mutations in COSMIC database (Forbes et al., 2014). Then, we selected protein-altering point mutations (missense and nonsense mutations) and frameshift alterations that originate from INDELs. Avsift and MA-score algorithms, implemented in ANNOVAR, were used to assess the damaging potential of the variants identified. A list of selected candidate nucleotide variants was analyzed using IntOGen, an integrative platform that summarize somatic mutations, genes and pathway involved in tumorigenesis (Gonzalez-Perez et al., 2013).

3.1.4 Analysis of fusion transcripts

Fusion transcripts discovery was performed using two different algorithms: TopHat Fusion (Kim and Salzberg, 2011) and Chimerascan (Iyer et al., 2011). TopHat Fusion is an enhanced version of TopHat with the ability to align reads across fusion points, which results from the breakage and re-joining of two different chromosomes, or from rearrangements within a chromosome. TopHat-Fusion engine is incorporated into TopHat2 with the name of --fusion-search option. TopHat-Fusion outputs consists in a list of potential fusions and a modified SAM alignment that contains a parameter that allows "fusion" alignment. This file was processed by tophat-fusion-post, a tool implemented in TopHat fusion in Perl language, with the following parameters: --num-fusion-reads 5 --num-fusion-pairs 4. Chimerascan was launched with default parameters. Then, we selected only fusions identified by both algorithms; moreover, fusions with less than 7 spanning reads (reads that map across the fusion breakpoint) were filtered out. Additionally, we removed fusion events observed in adjacent and/or overlapping genes as well as fusions involving *HLA*, *IGH* genes and other involving genes from repeated families.

3.1.5 Reverse Transcription

For mRNA analysis, reverse transcription was performed on total isolated RNA with SUPERSCRIPT II Reverse Transcriptase (Invitrogen). Reaction mix of 1 µg RNA, 1 µL Oligo(dT)12-18 (500 µg/mL) and 1 µL dNTP mix (10 mM each) in a final volume of 12 μL was heated at 65°C for 5 min and then chilled on ice. Subsequently, 4 μL 5X First-Strand Buffer, 2 μL DTT (0.1 M) and 1 μL RNase OUT (40 units/μL) were added to the reaction mix and incubated at 42°C for 2 min. After the addition of 1 μL of SuperScript II RT enzyme, the reaction mix was incubated at 42°C for 50 min and then the enzyme was inactivated heating the mixture at 70°C for 15 min.

3.1.6 RT-PCR assay, cloning and Sanger sequencing

Reverse-Transcription Polymerase-Chain-Reaction (RT-PCR) and Sanger sequencing were used to analyze the novel candidate fusion transcripts and the novel mutations in PTC and healthy samples. cDNA synthesis and PCR amplification were performed using standard protocols that come with Superscript II Reverse Transcriptase (Invitrogen) in a 20 μl reaction according to provided protocol. PCR primers were designed to amplify 200-400 bp fragments containing the putative nucleotide variant or the gene fusion boundary, as indicated by RNA-Seq. Where multiple PCR products were detected, we cloned these amplicons into Topo Vector II plasmid (Invitrogen) according to manufacturer's instructions. PCR products - and plasmids containing PCR amplicons - were analyzed by direct Sanger Sequencing. Analysis of Sanger chromatograms was performed using ApE software (http://biologylabs.utah.edu/jorgensen/wayned/ape/). Refinement of chimeric transcripts' structure was performed using UCSC Blat tool. The primers used for PCR validations are reported in Table 3.2.

Table 3.2. Oligolucleotides used for validations of fusion transcripts and mutations in known and new driver genes. *Published in Costa et al., 2015.*

Gene	Forward primer	Reverse primer
BRAF	CATAATGCTTGCTCTGATAGG	TCTAGTAACTCAGCAGCATCT
CBL	GTGGGTTTTTACTGATTTGCTT	AGGGCAATGAAAATGGAAGTG
DICER1	CTGAGGAGGATGAAGAGAAAG	CTAAAGGGAGCCAACAATACC
HRAS	CCGGAAGCAGGTGGTCATTG	GCCAGCCTCACGGGGTTCA
MET	TCCCCACAATCATACTGCTG	CCATCTTTCGTTTCCTTTAGC
NOTCH1	GCAGCCTGGGTTGGAGTAGG	TCAACACCTGCGGGGGATGG
SMARCA4	CGGTGTTGGGTGTTCCTTCA	TGGGATTACAGGCACGAACC
VHL	CTGGATCGCGGAGGGAATG	AGGCGGCAGCGTTGGGTAG
B4GALNT3	-	CTCTGGGGGATGGTAGAACTGG
WNK1	CGGTCTACAAAGGTCTGGAC	GCGGTGAATGATAGGTGGAG
WNK1-B4GALNT3	CGGTCTACAAAGGTCTGGAC	GGCGGTCCACTCCTTTCCA
PIK3R4	CTATCTGTATGGGGAAAAATTG	AGATTGCATGGAAGTATTTGAG

3.2 Results

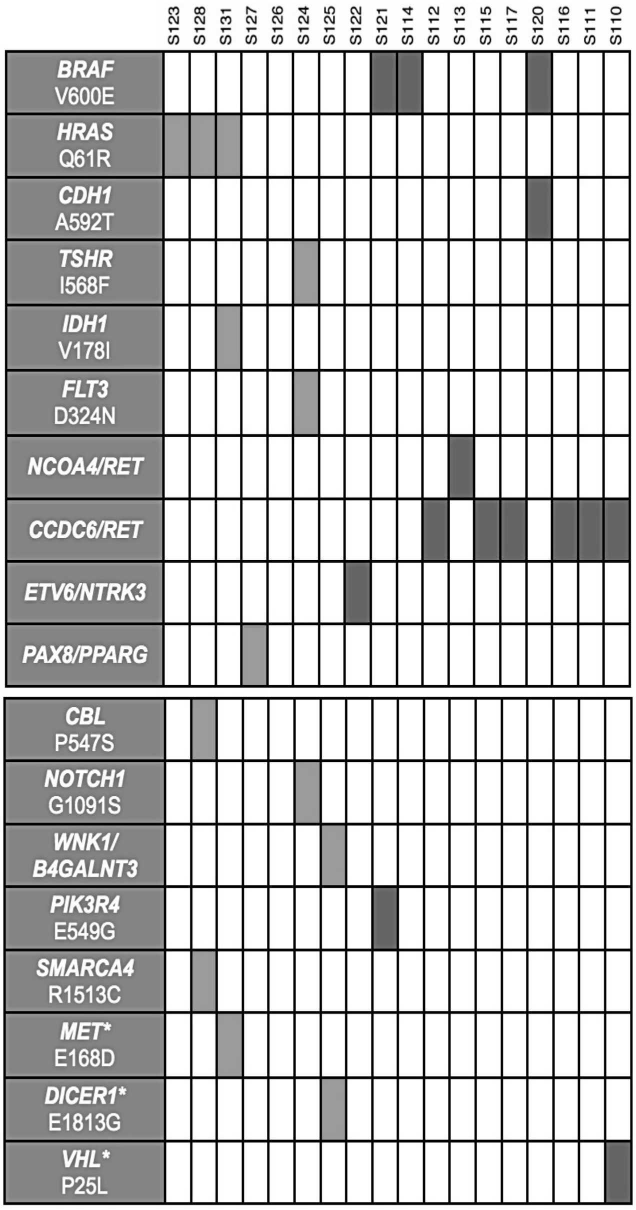
3.2.1.Analysis of known driver PTC alterations

In this pilot study, we employed RNA-Sequencing to profile the transcriptome of PTC samples and to compare it with the expression profile of healthy thyroids, at different levels. The study has been carried out on a cohort of 22 patients, (18 PTC biopsies, and 4 deriving from non-tumor thyroid) randomly chosen from well-characterized cohort of 80 PTC patients. Using Illumina HiSeq 2000 platform we generated a total of 1,6 billion of paired end reads (75+75 bp and 100+100 bp) We performed short read gapped alignment by using TopHat2 (Kim et al., 2013) and recovered about 1,5 billion of mapped reads.

The first aim of my analysis was the identification of genetic alterations in driver genes using the workflow of the software GATK optimized for the variant calling (http://gatkforums.broadinstitute.org/discussion/3891/calling-variants-in-rnaseq) and the algorithm TopHat Fusion (Kim and Salzberg, 2011) and ChimeraScan (Iyer et al., 2011) for fusion genes identification.

In line with literature data, we found that ~65-70% of papillary thyroid tumors had at least one known driver mutation or gene rearrangement (Santoro and Carlomagno, 2013). This analysis allowed the identification of known somatic driver alterations in PTC, which have been validated by targeted Sanger sequencing (Figure 3.1). Most of them (~38%) were *RET* gene fusions. In detail, 6 patients had CCDC6-RET (RET/PTC1, ~33%) and 1 NCOA4-RET (RET/PTC3, ~5%) fusions with a PTC1/PTC3 ratio quite similar to that described in literature for patients not exposed to ionizing radiations. RNA-Seq data confirmed *RET* overexpression in patients carrying RET/PTC fusion (FDR <0.01; Figure 3.2).

Moreover, we found other two already known driver alterations: *PAX8-PPARG* and *ETVN6-NTRK3* gene fusions (~5% frequency each). Similarly, RNA-Seq data confirmed the overexpression of *PPARG* and *NTRK3* (FDR <0.05; Figure 3.2). Notably, *PAX8-PPARG* chimeric gene is usually associated to follicular carcinomas, but has been reported with low frequency in follicular variant of papillary thyroid carcinoma (Cancer Genome Atlas Research Network, 2014). RT-PCR on cDNAs and targeted sequencing validated all described gene fusions detected by RNA-Sequencing, confirming the reliability of the computational analysis. We also focused on the identification of point mutations in our biopsies. We could identify $\mathrm{BRAF}_{\mathrm{V600E}}$ and $\mathrm{HRAS}_{\mathrm{Q61R}}$ in 3 patients each (~16% frequency for each mutation; Figure 3.1 panel A). Notably, as described in literature, we found that mutations in *BRAF* and *HRAS* genes, as well as RET/PTC and other rearrangements, were mutually exclusive events in the etiopathogenesis of PTC (Soares et al., 2003). Sanger sequencing confirmed the presence of mutations even on patients' DNA (where available). Such analysis was extended also to negative patients, confirming again the *bona fide* of the SNP calling procedure on RNA-Seq data.



	S123	S128	S131	S127	S126	S124	S125	S122	S121	S114	S112	S113	S115	S117	S120	S116	S111	S110
BRAF V600E									■	■					■			
HRAS Q61R	▨	▨	▨															
CDH1 A592T															■			
TSHR I568F						▨												
IDH1 V178I			▨															
FLT3 D324N						▨												
NCOA4/RET												■						
CCDC6/RET											■		■	■		■	■	■
ETV6/NTRK3								■										
PAX8/PPARG				▨														

	S123	S128	S131	S127	S126	S124	S125	S122	S121	S114	S112	S113	S115	S117	S120	S116	S111	S110
CBL P547S		▨																
NOTCH1 G1091S						▨												
WNK1/ *B4GALNT3*							▨											
PIK3R4 E549G									■									
SMARCA4 R1513C		▨																
*MET** E168D			▨															
*DICER1** E1813G							▨											
*VHL** P25L																		■

Figure 3.1 Schematic representation of protein-altering mutations and gene fusions identified in PTC samples. Each vertical column represents a PTC patient. In the upper panel, known missense mutations and fusion transcripts associated with papillary thyroid carcinoma are shown. In the lower panel are depicted newly identified somatic mutations and other somatic alterations in cancer driver genes reported in other tumors but described for the first time in PTC (indicated by asterisks). Red boxes indicate HRAS-mutated patients or those with a RAS-like transcriptional profile. Green boxes indicate BRAF-mutated or RET/PTC patients with a BRAF-like transcriptional profile *Published in Costa et al., 2015.*

Finally, we identified other known somatic mutations in our samples in crucial genes, such as E-cadherin (CDH1$_{A592T}$), in thyroid stimulating hormone receptor (TSHR$_{I568F}$), in isocitrate dehydrogenase 1 (IDH1$_{V178I}$) and in fms-related tyrosine kinase 3 (FLT3$_{D324N}$) genes (Figure 1A). Interestingly, these mutation are not exclusive events in PTC patients; indeed, CDH1$_{A592T}$ co-occurs with *BRAF*$_{V600E}$ mutation, and IDH1$_{V178I}$ co-occur with and *HRAS*$_{Q61R}$ mutations, respectively (Figure 3.1). Mutation frequencies are in line with those reported in the COSMIC database (~2-5%).

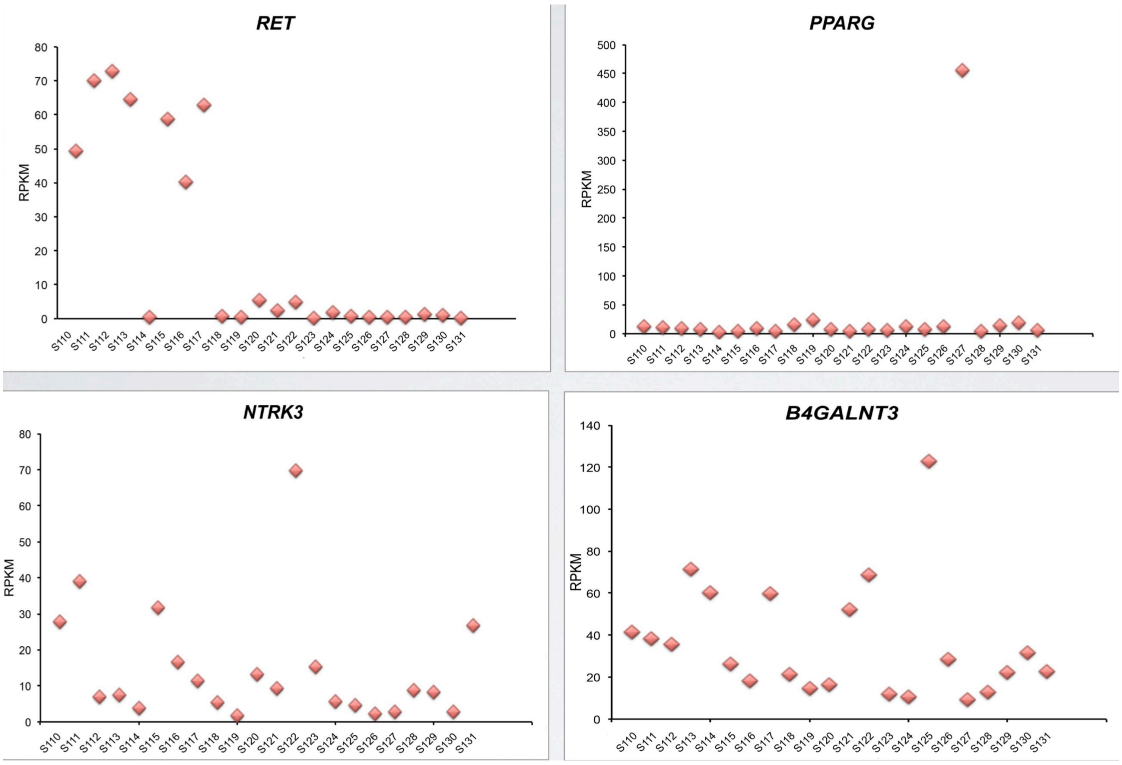


Figure 3.2. Scatter chart with RPKM values (y axis) of *RET*, *PPARG*, *NTRK3*, and *B4GALNT3* genes in PTC samples (y axis). All these genes, partners of the gene fusion, are significantly activated in PTC samples carrying the fusion.

3.2.2 Single nucleotide variants in cancer driver genes

After the SNP calling performed by GATK, we applied a stringent filtering procedure: SNPs annotated in dbSNPv138 (http://www.ncbi.nlm.nih.gov/SNP/), in 1000Genome project, in healthy thyroids and falling in super-duplicated regions were filtered out. After, we retrieved ~7430 missense, stop gain/loss point mutations and insertions/deletions (INDELs). Then, we combined these data with IntoGen (Gonzalez-Perez et al., 2013) and COSMIC databases (Forbes et al., 2011). In particular, we focused on 125 "Mut-driver" genes, defined by Vogelstein and colleagues (Vogelstein et al., 2013) as those containing driver mutations. We found 44 variants in 32 genes.

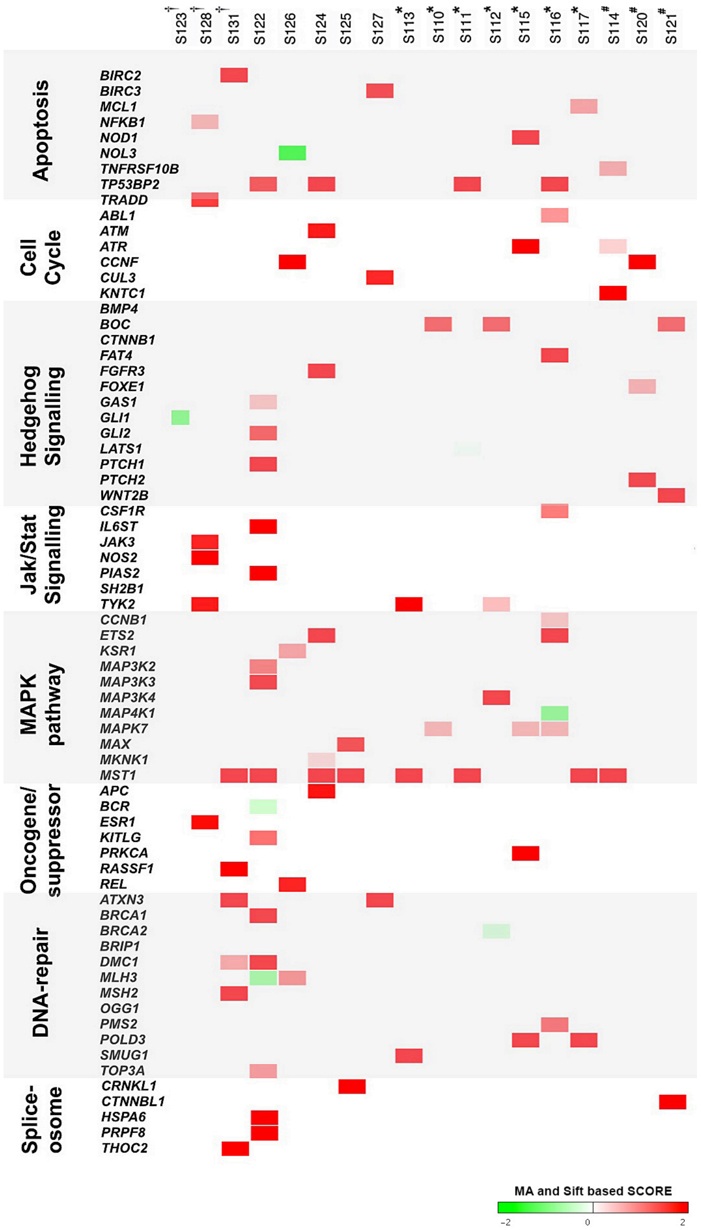


Figure 3.3 Co-occurrence of protein-altering nonsense and missense mutations identified in PTC patients (n=1 8) by RNA-Sequencing. Most relevant shared mutations in biological pathways associated to tumorigenesis are shown. Each vertical column represents a PTC patient. *HRAS*$_{\text{Q61R}}$, *BRAF*$_{\text{V600E}}$ and RET/PTC patients are indicated by †, * and $^{\#}$, respectively. The severity of the amino acid change is proportional to the intensity of red and green boxes (according to MA, "Mutation Assessor", and Sift scores). *Published in Costa et al., 2015.*

Moreover, we investigated mutations in known cancer driver genes and their interacting partners in the 12 pathways reported in Figure 3.3, commonly associated to tumorigenesis. Thus, we searched for damaging mutations affecting JAK-STAT signaling, MAPK, apoptosis, cell cycle, Hedgehog, onco-suppressors and oncogenes, DNA-repair and spliceosome pathways. We found 61 mutated genes in these pathways (Figure 3.3). Interestingly, we identified damaging mutations in *ATR*, *BRCA1/2*, *MAP4K1*, *CUL3* and *MAX*, already reported as somatic mutations in other cancer types in COSMIC database (Forbes et al., 2008), but not described yet as mutated in thyroid cancer.

Although most of them are not classified as cancer drivers, some mutations discovered here in PTC for the first time - MST1$_{\mathrm{R703C}}$ and BOCQ$_{\mathrm{915H}}$ – are annotated as "somatic" mutations in COSMIC in other cancer types.

3.2.3 Identification of new mutations in PTC

From the analysis above described, we identified - for the first time in PTC - mutations in 4 cancer driver genes: *CBL*, *NOTCH1*, *SMARCA4*, *PIK3R4.* In detail, we found the following missense mutations (and amino acid changes): c.C1639T (p.P547S) in *CBL* (proto-oncogene E3 ubiquitin protein ligase), c.G3271A (p.G1091S) in *NOTCH1*, c.A1646G (p.E549G) in *PIK3R4* (phosphoinositide-3-kinase regulatory subunit 4 gene) and c.C4537T (p.R1513C) in *SMARCA4* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4).

As shown in Figure 3.1, low frequency mutations in *CBL* and in *SMARCA4* co-occur with HRAS$_{\mathrm{Q61R}}$, in *PIK3R4* with BRAF$_{\mathrm{V600E}}$, whereas in *NOTCH1* with TSHR$_{\mathrm{I568F}}$ and FLT3$_{\mathrm{D324N}}$ in one patient. Targeted sequencing on DNAs of positive and negative biopsies validated the *bona fide* of these findings. Notably, new mutations in *CBL*, *NOTCH1*, *PIK3R4* and *SMARCA4* genes are not annotated as single nucleotide polymorphisms (SNPs) in dbSNP v138 and in the 1000GenomeProject (1000 Genomes Project Consortium, 2012), nor in COSMIC database. To strengthen these findings, we screened, by targeted sequencing, 80 alleles from healthy donors and we did not detect any of these mutations. Mutation frequencies are shown in Table 3.3.

Moreover, according to the same *criteria*, we selected and validated on patients' DNA also a stop gain mutation in *BRCA1* and a missense mutation in *ATM* genes. The loss of function mutation potentially identified in *BRCA1* was discarded from further analyses because it was a false positive, whereas, *ATM* mutation was discarded because it was found also in the healthy tissue counterpart, indicating a potential germline mutation. However, we cannot exclude that such a nucleotide variation may represent a common SNP, since for this gene we did not extend the sequencing analysis to the control cohort.

Finally, we selected and validated on genomic DNAs from patients' thyroids the missense mutations MET_{E168D}, $DICER1_{E1813G}$ and VHL_{P25L}. Interestingly, these specific mutations are annotated as "driver" in other cancer types in COSMIC database, but have not yet been reported in papillary thyroid cancer. Noteworthy, also these mutations identified in RNA-Seq data have been validated on DNA. Therefore they are not generated by RNA editing.

Interestingly, our analysis revealed that, most of the genes that we found mutated for the first time in PTC (i.e. *CBL*, *NOTCH1*, *SMARCA4*, *MET* and *VHL*) are "Mut-driver" genes. There is a still on-going mutational screening on a larger cohort of patients that will definitely help us to establish the frequency of these new mutations in PTC.

Gene	Genomic position	Nucleotide change	AA change	Frequency	*Status*	Other cancer
CBL*	11:119155974	C1639T	P547S	0.02/50	heterozygous	-
NOTCH1*	9:139402738	G3271A	G1091S	0.021/47	homozygous	-
PIK3R4*	3:130447468	A1646G	E549D	0.055/18	heterozygous	-
SMARCA4*	19:11169467	C4447T	R1483C	0.055/18	heterozygous	-
MET#	7:116339642	G561T	E187D	0.02/49	heterozygous	Hematopoietic, lymphoid, endometrium, lung
DICER1#	14:95557629	A5438G	E1813G	0.021/47	heterozygous	Brain, uterus
VHL#	3:10183605	C74T	P25L	0.055/18	heterozygous	Kidney

Table 3.3. Confirmed mutations in PTC samples. * indicate completely new mutations; # indicate known mutations never described in PTC.

3.2.4 *In silico* analysis of newly identified mutations

Despite the computational challenge and the higher false positive rate for SNP calling in RNA-Seq data compared to exome sequencing, one of the most interesting features is that it allows uncovering whether or not a mutation is expressed. Indeed, not all DNA mutations occur in actively transcribed genes, and without any direct proof one cannot be sure that the mutated allele is expressed in that cell/tissue. Moreover, even in presence of active transcription from that gene *locus*, we cannot exclude that allelic imbalance occurs, leading to very low (or conversely abundant) fraction of mRNA carrying that specific mutation. Clearly, it directly influences the protein product and, therefore, the altered (or not) function of the given protein.

As RNA-Seq is a sequencing-based approach and we have computationally identified mutations in these nucleotide sequences, we expected that all mutated genes were expressed in our datasets. Thus, the next step was to predict the effects of these new mutations on translated protein, using *in silico* methods.

In detail, our analysis revealed that the nucleotide change C1639T in E3 ubiquitin protein ligase CBL leads to P547S amino acidic change, which occurs in a proline stretch (PPPPPPDR) of the Proline-rich domain of Cbl protein (Figure 3.4 panel A). Since the mutated residue is smaller and less hydrophobic than the wild type it is predicted to affect local folding. Moreover, the proline stretch in this domain is highly conserved in homologous sequences along the evolutionary scale, and neither the mutant nor other residues have been observed at this position. Thus, conservation analysis and structural scores indicate the mutation as damaging.

The glycine-to-serine (G1091S) mutation in Notch1 involves a conserved glycine in a highly-conserved functional region, the EGF-like domain 28 (Figure 3.4 panel B). Sift and Polyphen scores indicate this mutation as very damaging for protein functionality. Indeed, wild-type and mutant residues differ in size, charge and hydrophobic properties.

The glutamic-acid-to-glycine (E549G) mutation in the phosphoinositide-3-kinase regulatory subunit 4 (PI3KR4) falls in a highly conserved "Armadillo-like helical" domain (Figure 3.4 panel C). This multi-helical fold, with extensive solvent-accessible surface, is suited to bind large substrates such as proteins and nucleic acids. The wild-type and mutant amino acids have different electric charge and hydrophobic properties; moreover the presence of glycine - instead of glutamic acid - is predicted to significantly reduce chain rigidity. *In silico* data indicate that the mutation is potentially damaging to PIK3R4 activity.

The arginine-to-cystein mutation in SMARCA4 occurs in a critical functional region (Figure 3.4 panel D), the bromodomain (BRD). Structural 3D analysis revealed that the wild-type residue (arginine) forms salt bridges with Asp1506, Glu1512 and Asp1528 and that the mutated residue loses these interactions (Figure 3.4 panel D).

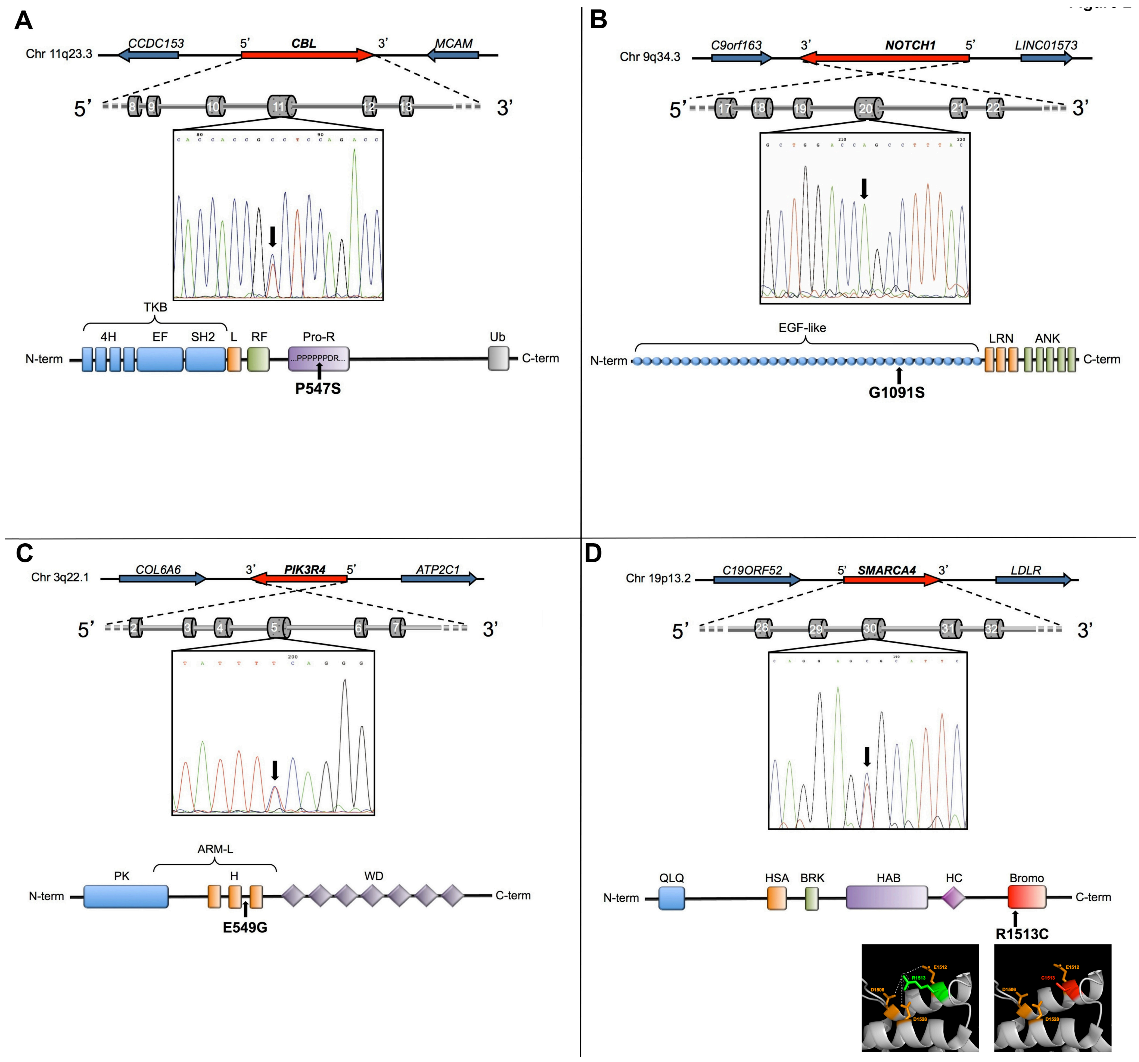


Figure 3.4. New mutations in PTC. In figure are schematized the genomic localization and the exon/intron structure of each mutated gene. In each panel, the ectropherogram shows the nucleotide variation identified by RNA-Seq, and the protein graphic representation shows the functional domains affected. In panel **D,** a detail of the three-dimensional structure of SMARCA4 bromodomain highlights the salt interactions among wild-type residue (colored in green) and the surrounding amino acids (colored in orange). These interactions are lost in the mutated protein (the mutated residue colored in red).

3.2.5 WNK1-B4GALNT3: identification of a novel gene fusion

Another aim of this study was the identification of new oncogenic driver gene fusions in papillary thyroid carcinoma. After applying two different algorithms to detect fusion genes from RNA-Seq data, I obtained 6 different fusion candidates identified by both software, specifically expressed only in tumor thyroids and with a positive prediction score (Table 3.4).

5’ Gene	3’ Gene	5’Chr	3’ Chr	Number of samples
CCDC6	*RET*	10	10	5
NCOA4	*RET*	10	10	1
PAX8	*PPARG*	2	3	1
ETV6	*NTRK3*	12	15	1
WNK1	*B4GALNT3*	12	12	1
KANSL1	*ARL17A*	17	17	4

Table 3.4. Filtered gene fusions identified in our RNA-Seq datasets.

As stated before, computational analysis of chimeric transcripts revealed the presence of known gene fusions (*CCDC6-RET*, *NCOA4-RET*, *PAX8-PPARG* and *ETVN6-NTRK3).*

Moreover, we identified other two fusions: *WNK1-B4GALNT3* chimeric transcript in one patient negative for known PTC-causing genetic alterations and *KANSL1-ARL17A* in 4 patients. After visual inspection and literature search, the *KANSL1-ARL17A* fusion was discarded because it is part of a known polimorphic region that often undergoes benign genomic rearrangements.

Thus, we focused on *WNK1-B4GALNT3* chimeric transcript. RNA-Seq data indicated that the new chimeric transcript originates by fusion of the exon 1 of *WNK1* and the exon 2 of *B4GALNT3* (Figure 3.5). RT-PCR, cloning and Sanger sequencing confirmed the fusion breakpoint in the transcript. Interestingly, in the PTC biopsy of the positive patient we observed 2 different splicing isoforms of this fusion gene: the longest one formed by the fusion of exon 1 of *WNK1* gene and exon 2 of *B4GALNT3* gene, and a shorter isoform, skipping the exons 2 and 3 of *B4GALNT3,* and thus constituted by a fusion between the exon 1 of *WNK1* and the exon 4 of *B4GALNT3* (Figure 3.5). Notably, sequence and ORF analysis revealed that the longest fusion transcript is out-of-frame, whereas the alternative isoform keeps the ORF intact (Figure 3.6 panel B). We demonstrated that this patient does not carry reciprocal gene fusions. In addition, since RNA-Seq indicated that the canonical mRNAs of *WNK1* and *B4GALNT3* genes are actively transcribed in this patient, we confirmed it by RT-PCR (Figure 3.6 panel A). Using the same approach, we validated the absence of gene fusions in negative patients (Figure 3.6 panel A).

Analysis of expression levels - measured as fragments per kilobase of exon per million fragments mapped (FPKM) - of each individual gene (*WNK1* and *B4GALNT3*) was performed on patient samples from RNA-Seq reads. Expression levels of WNK1 were distributed in a range starting from 110.3 FPKM up to 199.6 FPKM. In the patient carrying the fusion, its expression was 191.1 FPKM. On the other hand, expression levels of *B4GALNT3* gene were between 9.2 and 56.4 FPKM, in the positive sample harboring the gene fusion it was 122.7 FPKM. Thus, in line with the over-expression of *RET*, *PPARG* and *NTRK3* in patients with gene rearrangements, RNA-Seq data showed *B4GALNT3* over-expression in this patient ($p < 0.05$; Figure 3.2). The expression of *WNK1* was not affected.

Fusion partners map on chromosome 12 (chr12p13.33), are transcribed from the same strand (5'-3' orientation) and are separated by ~220 Kb (Figure 3.5). It indicates that the fusion derives from an intrachromosomal paracentric rearrangement. We could not identify in this patient, or in other patients of the discovery cohort, additional fusions involving genes mapping in the same genomic region.

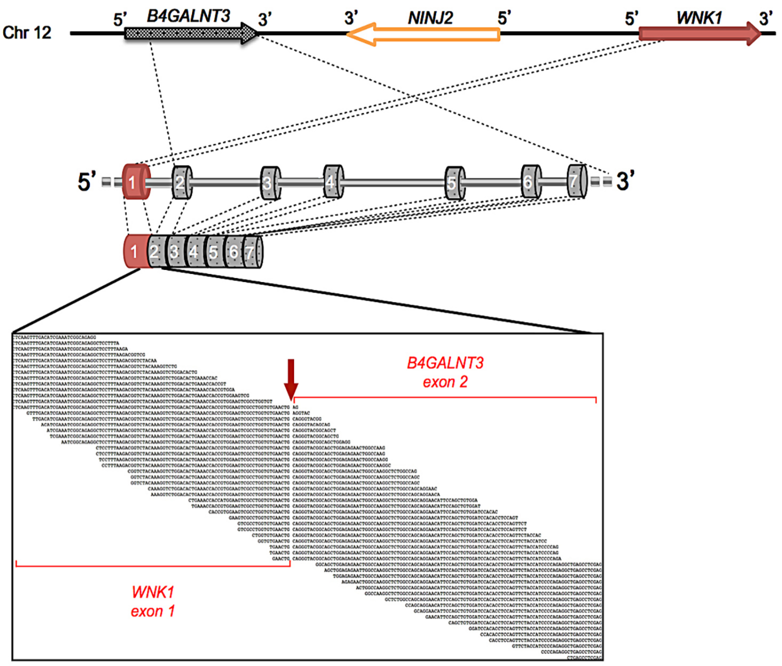


Figure 3.5. Schematic representation of the localization of the fusion partners, *WNK1* and *B4GALNT3,* on chromosome 12. The exons of *WNK1* and *B4GALNT3* genes that are involved in the fusion are indicated in red and grey, respectively. The RNA-Sequencing reads that map across the fusion breakpoint are shown in the black box. The red arrow indicates the exact fusion breakpoint.

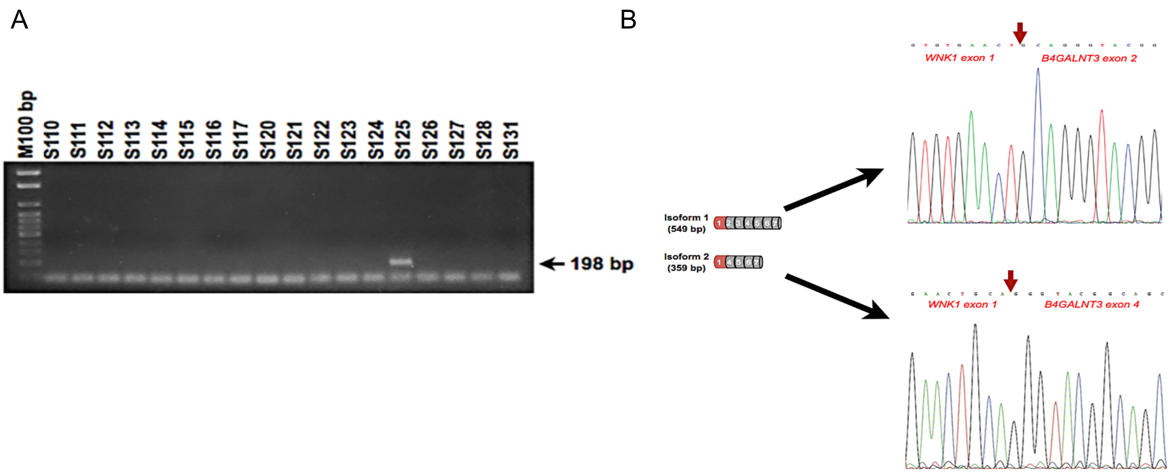


Figure 3.6. A) RT-PCR validation of the *WNK1-B4GALNT3* fusion performed on the RNA of 18 PTC samples of the discovery cohort. **B)** Schematic mRNA structure of the two isoforms of *WNK1-B4GALNT3* fusion gene. The electropherograms show the nucleotide sequences of the breakpoint (indicated by red arrows).

3.2.6 WNK1-B4GALNT3 in colon cancer sample

Recently, Huang's group published two papers about the up-regulation of *B4GALNT3* gene in colon cancer. Briefly, they observed the up-regulation of this gene in ~ 70% of tissues deriving from colon cancer biopsies. Moreover they demonstrated that the over-expression of this gene enhances the malignant phenotype of colon cancer cells and modulate cancer stemness through EGFR signaling pathway (Huang et al., 20XX; Che et al., 2015). These results encouraged us to search for the same gene fusion in colon cancer patients, in order to explain *B4GALNT3* over-expression. I performed RT-PCR on cDNAs deriving from 22 paired tissues (colon cancer tumor and healthy tissues counterparts). The analysis revealed the presence of *WNK1-B4GALNT3* fusion in one tumor sample, but not in its healthy counterpart, indicating the somatic nature of this event.

3.3 Discussion

The first part of my PhD thesis describes the analysis of genetic alterations (mutations and rearrangements) in 18 PTC samples. During this analysis we identified a novel gene fusion, new somatic mutations in well-established cancer driver genes and known mutations (reported in other cancer types) not yet described in PTC (Costa et al., 2015, published on *Oncotarget*).

We confirmed that driving somatic mutations ($\mathrm{BRAF}_{\mathrm{V600E}}$ and $\mathrm{RAS}_{\mathrm{Q61R}}$) and rearrangements (RET/PTC) are mutually exclusive in PTC.

Interestingly, the new *WNK1-B4GALNT3* fusion has been identified in a patient negative for known driver events in PTC. Noteworthy, a significant over-expression of *B4GALNT3* gene was found in this patient, whereas the expression of the fusion partner was not affected (Figure 3.2). *B4GALNT3* has been described both as tumor suppressor in neuroblastoma (Hsu et al., 2011) and as oncogene in the colon cancer. Huang and colleagues in their work observed that *B4GALNT3* over-expression increases the malignant phenotype of colon cancer cells through enhanced integrin and MAPK signaling (Huang et al., 2007). Similarly, Che and colleagues (2014) described that its expression positively correlates with metastasis and poor survival in patients with colorectal cancer. The identification of this new fusion gene involving *B4GALNT3* suggests a new role of *B4GALNT3* as oncogene also in PTC. Clearly, we need to expand our patients' cohort - particularly focusing on the about 30% of PTC cases without any known genetic etiology - in order to assess the frequency of this new rearrangement in the PTC. Moreover, further *in vitro* studies in thyroid cancer cell are necessary to definitely clarify the role of *B4GALNT3* over-expression in the etiology of PTC. Indeed, *B4GALNT3* over-expression may play a crucial role in promoting malignant behaviors of thyroid cancer, like cell adhesion, migration, and invasion, similarly to colon cancer.

Notably, the PTC patient positive for *WNK1-B4GALNT3* fusion also carries a somatic mutation in *DICER1, a* well-established cancer driver gene. The mutation $\mathrm{DICER1}_{\mathrm{E1813G}}$ - never described till now in PTC - affects the metal binding site of RNase IIIb domain and has been recently identified as somatic variation in non-epithelial ovarian cancers (Heravi-Moussavi et al., 2012). This is a very relevant finding as Heravi-Moussavi and colleagues observed an impaired RNase IIIb activity and retention of RNase IIIa activity in tumors with the mutation DICER1$_{E1813G}$. The altered DICER1 activity in the RNase IIIb domain could arise an oncogenic miRNA profile in patients carrying this mutation. Since specific miRNAs are crucial in cell differentiation and cell-fate determination, aberrant miRNA processing resulting from DICER1 mutations could be considered a key oncogenic event.

We also found another mutation in the oncogene *MET* (E168D). The mutation falls in the SEMA domain, crucial for the interaction with plexins, and it has been previously described in small cell lung cancer (Ma et al., 2003). Interestingly, this mutation impairs the affinity for HGF and alters MET functionality (Ma et al., 2003). However, this is the first time that such mutation is described in PTC. The c-MET Sema domain is conserved among all semaphorins and is also found present in the semaphorins receptors that are plexins. Semaphorins are a large family of secreted and transmembrane signaling proteins regulating neuronal axonal guidance and mediating scattering signaling in epithelial cells. Interestingly, semaphorin signaling may have a role in tumor progression, it would be useful to further determine the functional implication of the E168D mutation in PTC.

We also identified a nucleotide variation in *ATM* gene. Although we could not verify if it is a somatic or germline variation due to the lack of the healthy tissue counterpart, we could exclude it to be a SNP. This finding is relevant since PTC is the most frequent radiation-sensitive tumor, and ATM is a fundamental kinase that triggers the DNA damage checkpoint, determining cell cycle arrest, DNA repair or apoptosis.

Interestingly, starting from RNA-Seq data we could also discover a completely new somatic *NOTCH1*$_{G1091S}$ mutation in a patient negative for *BRAF*/*RAS* mutations and RTKs rearrangements. Many *NOTCH1* driver mutations have been reported in hematopoietic tumors, head and neck squamous cell carcinoma and other malignancies (Sharma et al., 2013). The presence of inactivating mutations indicates this gene as tumor suppressor, rather than oncogene, in solid tumors (Sharma et al., 2013; Yamashita et al., 2031). Interestingly, most of the mutations in solid tumors are clustered within EGF-like repeats. Accordingly, the newly discovered *NOTCH1*$_{\mathrm{G1091S}}$ mutation falls in the EGF-like domain 28 and affects a highly conserved residue. Although 3D model revealed this domain does not directly bind Notch ligands, Sharma and colleagues reported it to interact with EGF-like 11-15 domains, crucial for receptor activity (Sharma et al., 2013). Thus, this mutation may affect Notch1 protein functionality and Notch signaling that is directly linked to PTC cell proliferation (Yamashita et al., 2013). These data suggest that this pathway should be taken into account as adjuvant therapy for treating PTC, when *NOTCH1* is mutated. The same mutation co-occurs with two low-frequency mutations (*TSHR*$_{\mathrm{I568F}}$ and *FLT3*$_{\mathrm{D324N}}$*)* previously reported in PTC.

Additional mutations have been discovered in *BRAF*- and *RAS*-mutated patients. Among these, *CBL*$_{\mathrm{P547S}}$ was found in an *HRAS*-mutated patient. This mutation affects the proline-rich region, responsible of the binding with SH3 domain of Grb2 protein that indirectly recruits it to RTKs via GRB2 adaptor protein (Tan et al., 2010). In lung cancer, mutations in *CBL* and in other driver genes usually co-occur (Tan et al., 2010). Interestingly, the same patient carried a new missense mutation in *SMARCA4* gene, a tumor suppressor gene frequently mutated in lung cancer and small cell ovarian carcinoma (Jelinic et al., 2014; Medina et al., 2008). The mutation SMARCA4$_{\mathrm{R1513C}}$ disrupts salt interactions with charged residues in the BRD domain, a functional domain that allows the recognition of acetyl lysine marks on H3 and H4 tails (Muller et al., 2011). SMARCA4 protein, associating with Rb proteins, induces cell cycle arrest through HDAC-dependent transcriptional repression. Mutations, rearrangements or over-expression of BRD-containing proteins have been reported in tumors, and BRD inhibitors have been developed to induce cycle arrest and apoptosis of carcinoma cells (Muller et al., 2011). Therefore, a similar pharmacological approach could be adopted in the treatment of PTC cases with *SMARCA4* mutations.

4 Genome-wide analysis of lncRNAs involved in papillary thyroid carcinoma

4.1 Methods

4.1.1. *Ab initio* assembly

Reads aligned with TopHat were assembled into sample-specific transcriptomes with Cufflinks, version 2.0.2 (Trapnell et al., 2012). Cufflinks assembles exonic and splice-junction reads into transcripts using their alignment coordinates. The option –g (–GTF-guide) was used with Gencode v19 human reference genes in GTF format. This option, when provided, tells Cufflinks to use the supplied reference annotation to guide the assembly. Reference transcripts are tiled with faux-reads to provide additional information in assembly. Output include all reference transcripts as well as any novel genes and isoforms that are assembled. Moreover, to limit false positive assemblies we used a maximum intronic length of 300kb, corresponding to the 99.93° percentile of known introns. The other parameters were default.

The resulting GTFs were merged using Cuffmerge, version 2.0.2 (Trapnell et al., 2012), using option –g Gencode v19 human reference genes in GTF format as reference. Cuffmerge produces a GTF file, named *merged.gtf*, which merges together the input assemblies. Finally in order to compare, for each sample, assembled transcripts to the reference annotation, we used Cuffcomapare tool version 2.0.2 to distinguish known and novel transcripts. Cuffcompare produces different output files. In particular the *.traking* file matches transcripts up between samples. Each row contains a transcript structure that is present in one or more input GTF files. Cuffcompare examines the structure of the transcripts, matching transcripts that agree on the coordinates and order of all of their introns, as well as strand. Matching transcripts are allowed to differ on the length of the first and last exons, since these lengths will naturally vary from sample to sample due to the random nature of sequencing. This file contains also, for each assembled transcript, a "class code", which indicates the type of match between the transcripts considered and the reference transcript. Because of the lack of strand specificity in the sequencing protocol, we focused exclusively on intergenic or antisense lncRNAs without exon shearing with protein coding genes, by filtering out all transcripts showing any overlap with protein coding genes. Thus, we selected novel transcripts with “class clode” = “u” or “class clode” = “x”, which indicate unknown intergenic or antisense transcripts.

4.1.2 Identification of novel lncRNAs

Using the output of Cuffcompare, only putative novel transcripts with at least two exons were retained; this step is necessary for a stringent analysis and to filter a lot of false positive novel transcripts. The GTF annotation file, containing the annotation of novel lncRNAs, was converted in FASTA format by using the online platform Galaxy (https://usegalaxy.org). The resulting file, was used as input for the software CPAT version 1.2.2 (Wang et al., 2013), in order to check the protein-coding potential of the novel transcripts. For each analyzed transcript, CPAT returns the length in bp and a score, defined "potential coding". Since lncRNAs are described as transcripts longer then 200 bp, novel reconstructed transcripts were filtered for minimal length of 200 bp. Moreover, we selected transcripts with a potential coding less then 0.364 (threshold recommended by CPAT software developers to discriminate between coding and non-coding transcripts). The annotation file, containing the novel lncRNAs was uploaded on UCSC genome browser, for immediate data visualization.

4.1.3 Expression analysis

Read counts were then calculated per gene from the alignment bam files using HTSeq (v0.5.4p2) with options -m union --stranded no. As transcript annotation reference was used the GTF file containing Gencodev19 annotation supplemented with the GTF file of novel lncRNAs expressed in thyroid tissue. Genes were then filtered for minimal expression (mean counts >= 5 across all conditions).

Using the output of Cuffcompare, the transcripts were classified into 3 categories: known mRNAs, known lncRNAs (Gencode version19 annotation as reference) and novel lncRNAs.

Differential expression analysis. Data counts were fitted to a statistical model based on the negative binomial distribution using the R package EdgeR (Robinson et al., 2010), which is useful for detecting significant RNA-Seq variation with biological replicates (Anders and Huber, 2010b). To perform the normalization and differential expression analysis, raw read counts per gene were normalized to the relative size of each library. The difference between the means of tumor *vs* non-tumor samples and the means of BRAF-like *vs* RAS-like samples was then calculated using a negative binomial test.

For each gene, 'adjusted *p*-value' (also known as *q*-value) has been calculated to calculate the expected false discovery rate (FDR) (i.e. the proportion of positives returned which are false positives) to control differential expression. Thus, *p*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Genes with an adjusted *p*-value of <=0.05 were considered to be differentially expressed.

Finally, pathway and gene ontology analysis was performed using DAVID Functional Annotation Tool (https://david.ncifcrf.gov/).

4.1.4 Selection of novel lncRNAs

Since it is known that lncRNAs can act *in cis* regulating the expression of neighbor protein coding genes, in order to select cis-acting lncRNAs, we associated each lncRNA to the nearest protein-coding gene.

CPAT output file was converted in BED format to extract the coordinates of transcription start sites (TSSs) of novel lncRNAs. The BED file - with TSS genomic coordinates of all the genes (both coding and non-coding) annotated in GENCODEv19 - was downloaded from the tool "Table Browser" of UCSC genome browser (www.genome.ucsc.edu).

Thus, we associated TSS of both known and novel lncRNAs to the TSS of the closest protein-coding gene by using the function "closestBed" of BedTools. This analysis allowed us to identify gene-lncRNAs pairs.

Furthermore, in order to assess whether new putative lncRNAs could play a role in the pathogenesis or progression of PTC, only pairs with differentially expressed genes and differentially expressed lncRNAs have been selected. This list was intersected with a list of genes with a proven role in different type of cancer and defined as “cancer driver genes”. For this purpose, we used a list of 114 driver genes published by Vogelstein and colleagues (Vogelstein et al., 2013).

4.1.5 Subcellular fractionation

To further detect the cellular location of lncRNA *MET-AS,* cytosolic and nuclear fractions of thyroid cancer cell lines TPC-1 were isolated and collected Cytoplasmic and Nuclear RNA Purification Kit (Norgen, Biotek, Corp) as the manufacturer’s instructions. After that, total RNA was extracted from the collections of both cytoplasm and nucleus and cDNA was synthesized for the evaluation of *MET-AS*. PPIA and U2 non-coding RNA gene were used as control of cytosolic and nuclear fractions, respectively.

Table 4.1. Primer used for cytosolic, nuclear and chromatinic fractions.

	Forward	**Reverse**
PPIA	TACGGGTCCTGGCATCTTGT	GGTGATCTTCTTGCTGGTCT
U2	CATCGCTTCTCGGCCTTTTG	TGGAGGTACTGCAATACCAGG

4.1.6 Chromatin states

To analyze the presence of chromatin marker peaks at promoters, we used available public data published as part of the ENCODE project (The ENCODE Project Consortium, 2012), and downloadable from table browser of UCSC Genome Browser (www.genome.ucsc.edu). A marker was considered present if a non-empty intersection could be detected between the TSS of lncRNA region and a marker peak, in any of the replicates. The intersections were detected using the window command of the BEDTools program (Quinlan and Hall, 2010), version 2.17.0, with option -w 1000.

4.1.7 RT-PCR and quantitative Real-Time assays

Previously isolated RNA samples (1 μg) were reverse transcribed using SuperScript II Reverse Transcriptase with oligo dT primers (Invitrogen, Carlsbad, CA, USA) according to manufacturer's instructions. To check the amplifiable template RNA/cDNA, RT-PCR amplification of a housekeeping gene (Peptidylprolyl Isomerase A, *PPIA*) was performed in all samples. Each amplification reaction was set-up using AmpliTaq Gold (Life Technologies, Carlsbad, CA, USA). PCR products were analyzed by electrophoresis on agarose gel at 1.5%. Then cDNAs obtained were used as tamplate for Quantitative Real Time PCRs (qRT-PCR).

All quantitative qRT-PCRs were performed on the CFX Connect Real-Time PCR Detection System (Bio-Rad). For mRNA quantification, 10 μL final volume reaction mix was prepared using 1 μL cDNA from 1/5th of RT reaction with 5 μL iTaq Universal SYBR Green Supermix (Bio-Rad) and 0.4 μL of each primer (10 μM). After an initial polymerase activation step at 95°C for 2 min, 40 2-steps cycles of amplification were run at 95°C for 5 sec and 60°C for 30 sec. Melt-curve analysis was performed from 65°C to 95°C with a 0.5°C increment and 5 sec/step to verify PCR product specificity. The housekeeping *PPIA* gene was used as reference gene for data normalization and relative gene expression was measured with the 2-ΔΔCt method, comparing the Ct values of the samples of interest with the control.

Gene specific primers used for quantitative Real-Time amplification of mRNAs and lncRNA are listed in Table 2.

Table 4.2. Primers used for qRT-PCR

	Forward	Reverse
MET	TCTGCCCCACCCTTTGTTCA	ATCCAAAGTCCCAGCCACATA
MET-AS	TAACATAAATCCGCAAATCACA	GTGGGACCTCAAAGCCTAT
PPIA	TACGGGTCCTGGCATCTTGT	GGTGATCTTCTTGCTGGTCT

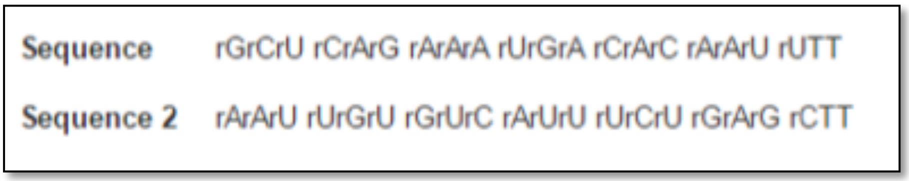
4.1.8 Cell culture

TPC-1 human papillary thyroid carcinoma cell line was kindly provided by Prof. Alfredo Fusco. Cells were grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10 % fetal bovine serum (FBS). 2mM glutamine, 100 units/mL penicillin and 100 units/mL streptomycin. Cultures were maintained at 37°C and 5% CO_2.

4.1.9 RNA interference

RNA interference transfection in TPC-1 cell line was carried out using Oligofectamine transfection reagent (Life Technologies) according to the manufacturer's protocol. Approximately $6x10^4$ cells were plated in six-well plates overnight. The next day, the cells were transiently co-transfected with 2 different custom designed short interference RNAs (50 nM) targeting 3' *MET-AS* and with a control siRNA purchased from Origene (Figure 4.1). The transfected cells were harvested at 24, 48, and 72 h for further analyses. The efficiency of the siRNA transfection showed a significant reduction in *MET-AS* RNA expression level (P<0.001). Each assay was carried out in triplicate in at least three independent experiments.

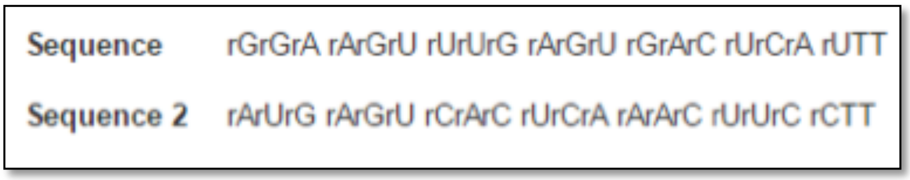
siRNA_1



Sequence rGrCrU rCrArG rArArA rUrGrA rCrArC rArArU rUTT

Sequence 2 rArArU rUrGrU rGrUrC rArUrU rUrCrU rGrArG rCTT

siRNA_2



Sequence rGrGrA rArGrU rUrUrG rArGrU rGrArC rUrCrA rUTT

Sequence 2 rArUrG rArGrU rCrArC rUrCrA rArArC rUrUrC rCTT

Figure 4.1. Duplex siRNAs sequence

4.1.10 Cell Cycle analysis

Seventy-two hours after transfection, cells were harvested and fixed in 70% pre-cooled ethanol. Then they were treated with RNAase A 50 µg/ml for 30 minutes and then stained with 50 μg/ml of propidium iodide (PI). Their DNA contents were analyzed by flow cytometry on a FACS system (Becton Dickinson FACSCanto A). The percentage of cells in each cell cycle phase was used as indications of cell cycle progression.

4.1.10 Cell viability assay

2x10^3 TPC-1cells were plated in 96-well white opaque plates. Viable cell growth was measured using the Cell Titer-Glo luminescent cell viability assay kit (Promega) according to manufactures' instruction after 24, 48 and 72 hours.

4.1.11 Statistical analysis

All experiments were repeated at least in triplicate. All data were presented as the mean±standard error of the mean. A statistical significance was determined by a Student's t test, and the differences with p-values of < 0.05 were accepted as statistically significant.

4.2 Results

4.2.1. Global gene expression of papillary thyroid carcinoma

We first characterized global gene expression in papillary thyroid samples through the analysis of both the coding and non-coding transcriptome. Moreover, we set up a computational workflow, depicted in Figure 4.2, to identify novel lncRNAs that are potentially able to modulate (and/or interfere with) the expression of cancer driver genes.

Low abundance and condition- (i.e. tumor-) specific transcripts may be lost or underestimated in standardized analyses, and particularly lncRNAs that exhibit high tissue-specificity and transcriptional levels that are generally lower than protein-coding genes. Thus, we first applied a *de novo* transcript assembly procedure by using Cufflinks (version 2.0.2) on mapped reads of the 22 RNA-Seq datasets from thyroid biopsies to define a thyroid model transcriptome. Using Cuffmerge tool (version 2.0.2) the transcriptome of each sample was merged and then compared to Gencodev19 annotation by using Cuffcompare (version 2.0.2).

As we aim to systematically and extensively study the lncRNA fraction, both annotated and novel, we computationally selected new transcripts resulting from Cuffcompare output that were multi-exonic, with a coding potential <0.364 (according to CPAT parameters) and longer than 200 nt. Moreover, we selected only new transcripts that do not overlap already annotated *loci.* The resulting GTF annotation file, containing 454 novel putative lncRNAs was added to Gencode v19 annotation GTF, in order to obtain a complete annotation file of the thyroid transcriptome. This novel transcriptome model was used to quantify gene expression in all 22 PTC samples (Figure 4.2)

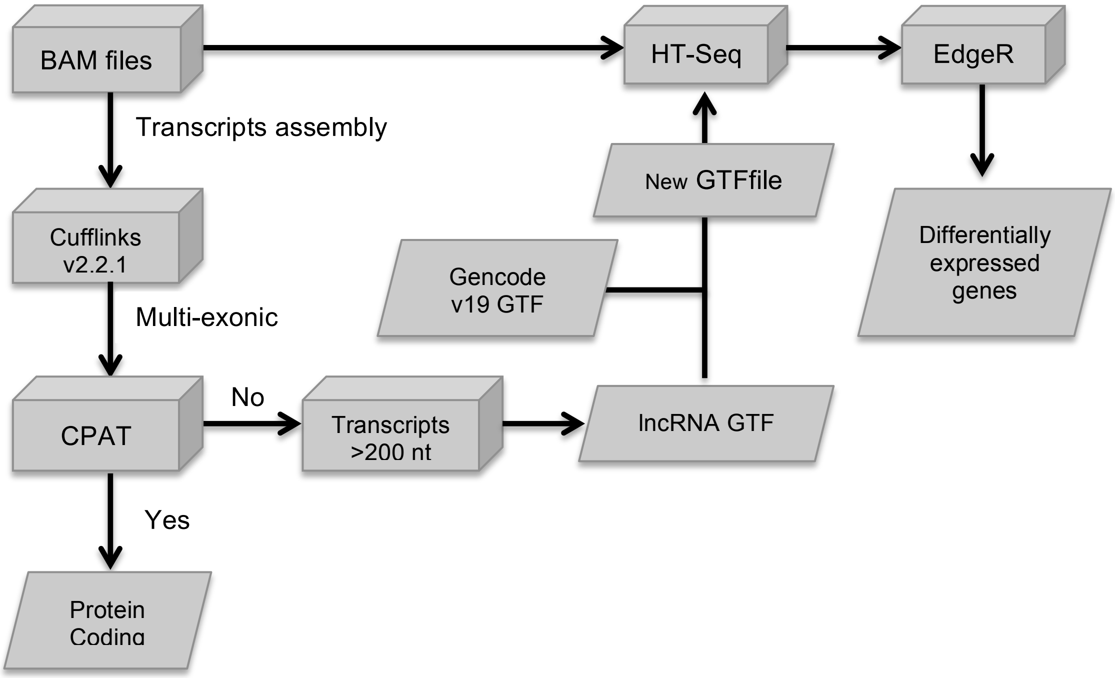


Figure 4.2. Pipeline of gene expression data analysis. The first step of the analysis, consisted in *de novo* assembly of TopHat output alignment file. To identify novel un-annotated lncRNAs, only multi-exonic transcripts longer than 200 bp and without potential coding were selected. Reads count was performed with HTSeq; analysis of differentially expressed genes was performed with EdgeR.

This analysis reconstructed 18137 multi-exonic transcripts, of which 15404 correspond to Gencode v19-annotated protein-coding genes (Figure 4.3). Moreover, our lncRNA annotation pipeline identified 2733 multi-exonic lncRNAs (>200 bp). There were 2279 Gencodev19-annotated lncRNAs and 454 novel unannotated lncRNAs, encompassing all known lncRNA locus-types (Figure 4.3).

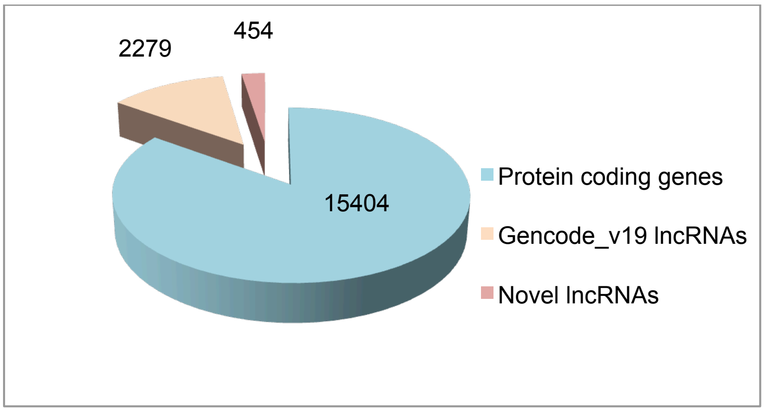


Figure 4.3. Pie chart showing composition of PolyA$^+$ transcriptome, Gencode mRNAs (blue) Gencode long non-coding RNAs (orange) and novel long non-coding RNAs (red). Transcript numbers in each group are indicated.

To verify the non-coding nature of our novel lncRNA candidates, we used the CPAT-coding potential score and found that these novel transcripts have minimal protein-coding potential, comparable with Gencode-annotated lncRNAs (Figure 4.4). Furthermore, novel lncRNAs and Gencode lncRNAs were expressed at significantly lower levels than coding genes (Figure 4.4).

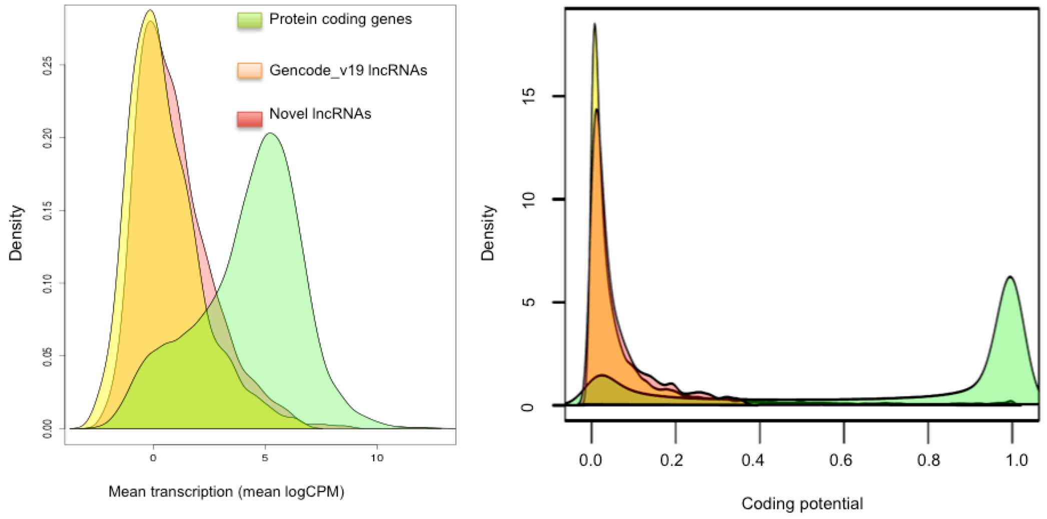


Figure 4.4. On the left, the kernel density plots of transcripts' abundance indicate that novel lncRNAs behave a distribution similar to already annotated ones, confirming the bona fide of de novo assembly and of the criteria chosen for the selection of such transcripts. On the right, CPAT-coding potential score of the three categories were plotted.

4.2.2 Differential expression analysis

Further analysis of differential expression at a false discovery rate (FDR) of <0.05 revealed that 3686 genes are significantly deregulated in PTC compared to healthy thyroids, 2901 (1332 up-regulated and 1569 down-regulated) of which are Gencodev19-annotated protein coding genes, 402 (142 up-regulated and 260 down-regulated) are Gencodev19-annotated lncRNAs and 109 (49 up-regulated and 60 down-regulated) are newly identified lncRNAs (Figure 4.5). Unsupervised hierarchical clustering of the normalized expression values for protein-coding, annotated and novel lncRNAs segregates PTCs from healthy thyroids, indicating that also lncRNAs can be considered potential PTC biomarkers (Figure 4.5).

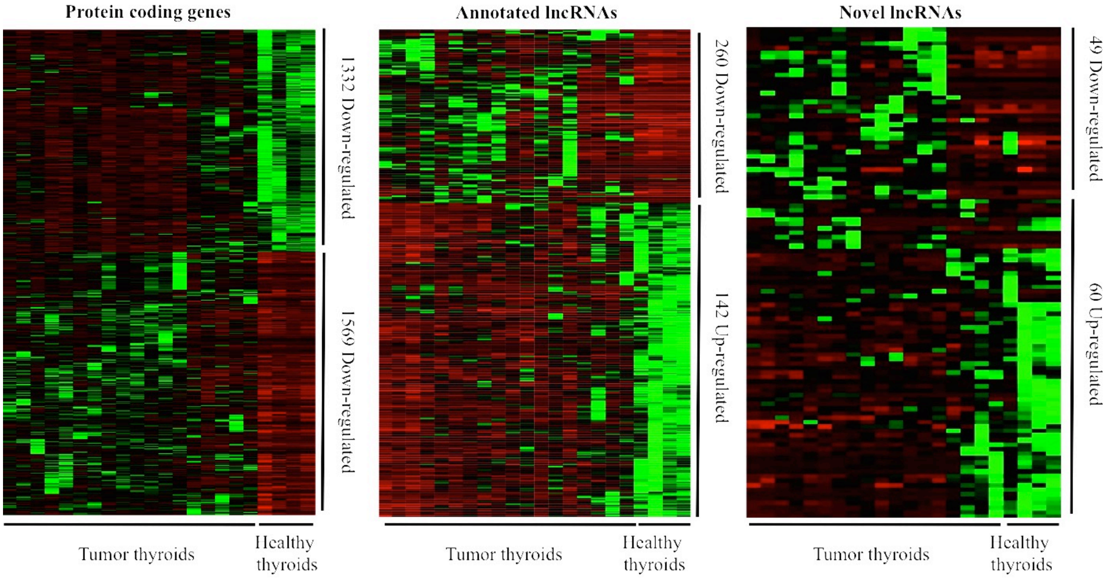


Figure 4.5. Heatmaps showing hierarchical clustering of differentially expressed transcripts within the three RNA classes comparing PTC and normal thyroids.

Taking into account gene expression of all detected protein-coding genes, pathway analysis revealed that the most affected pathways were cell cycle, axon guidance, p53 signaling and cytokine-cytokine receptor interaction, in line with the cancer phenotypes (Figure 4.6).

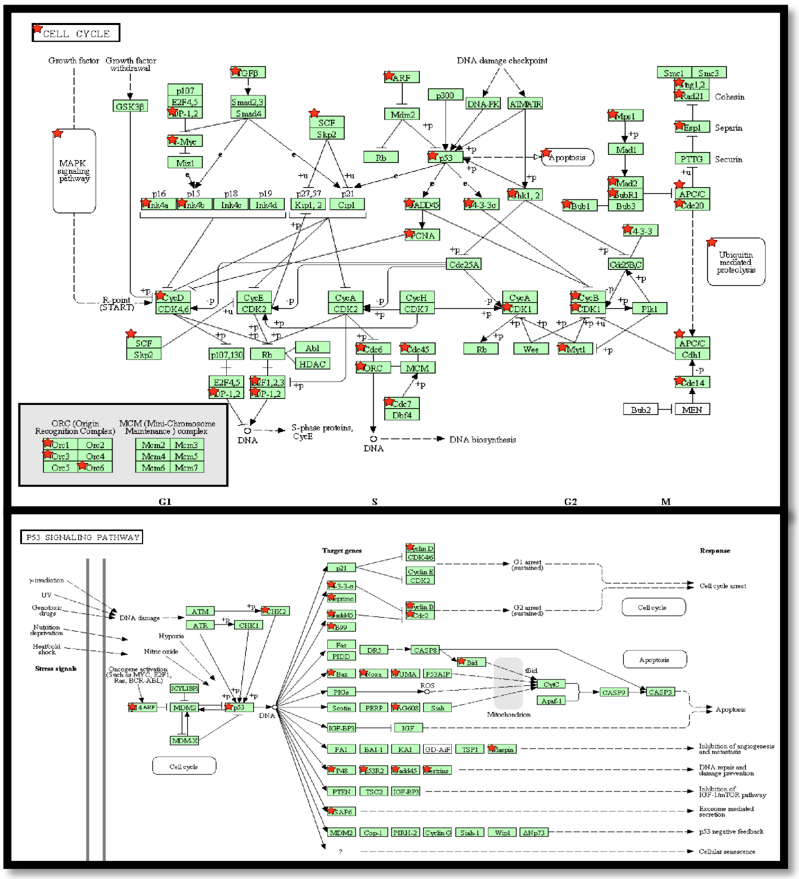


Figure 4.6. Cell cycle (upper panel) and p53 signaling (lower panel) affected pathways in papillary thyroid samples. Red stars indicate differentially expressed genes.

Recently, the seminal work of The Cancer Genome Atlas (TCGA) Research Network, published only few months before our *Oncotarget* publication (Costa et al., 2015), has defined a peculiar gene expression signature in samples mutated in *BRAF* and *RAS* genes. Mutually exclusive driver mutations in *BRAF* and *RAS* genes determine the differential signaling consequences on the activation of MAPK and PI3K signaling. Patients with *BRAF* mutations have a major activation of MAPK pathway, compared to RAS-mutated patients, which present the hyper-activation of PI3K pathway (Cancer Genome Atlas Research Network, 2014). In light of these results, taking advantage of our RNA-Seq data, we correlated global gene expression profiles to known mutations and rearrangements. We confirmed that BRAF-mutated and RET/PTC samples have very similar gene expression patterns and that they differ from RAS-mutated patients (~2230 differentially expressed genes; FDR <0.05).

Extending the analysis to PTC patients without any known mutation we found RAS- and BRAF-like gene signatures (Figure 4.7). These findings are in agreement with the notion the *BRAF*$_{V600E}$ and RET over-expression activate MAPK pathway more than *HRAS*$_{Q61R}$ and with the recent results of TCGA Consortium (Cancer Genome Atlas Research Network, 2014). Indeed, we found a significant over-expression - in BRAF- *vs* RAS-like PTCs - of DUSP genes (*DUSP2*, *DUSP5* and *DUSP6*) that are induced through the stimulation of ERK signaling *via* MAPK. Conversely, RAS-mutated patients over-expressed anti-apoptotic genes, including *BCL2*.

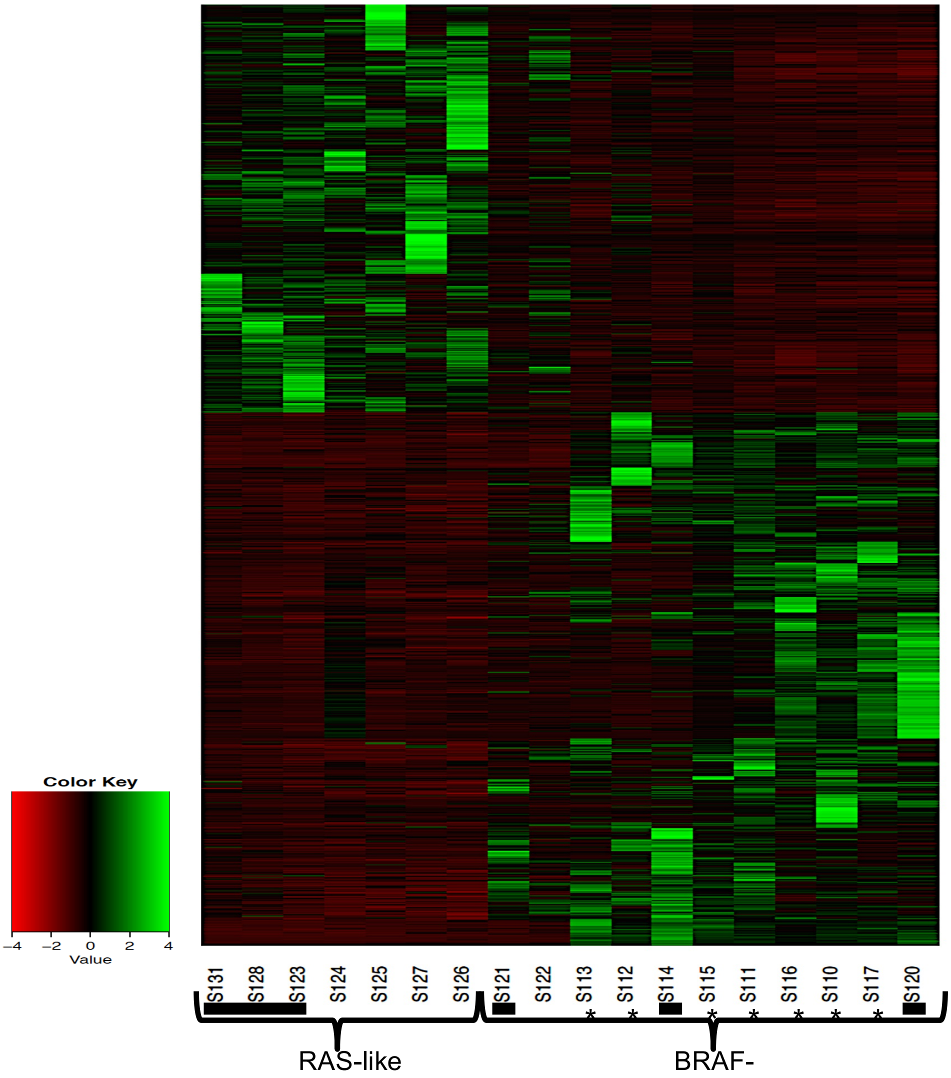


Figure 4.7. Heatmap of the hierarchical clustering of differentially expressed genes between BRAF-like and RAS-like PTC samples. Black bars indicate samples with point mutations in HRAS and BRAF genes. * indicates samples with *RET* gene fusions.

Similarly, we divided differentially expressed genes between *BRAF*-like and *RAS*-like samples in protein-coding genes, annotated and novel lncRNAs (Figure 4.8).

We found a signature of 202 lncRNAs (77 over- and 125 under-expressed) whose expression significantly differs when comparing *BRAF*- and *RAS*-like carcinomas. In detail, 59 over-expressed lncRNAs are annotated by Gencode and 18 are completely novel (Figure 4.8). On the opposite, 83 and 42 (annotated by Gencode and novel, respectively) are down-modulated in *BRAF*-like *vs RAS*-like PTCs.

This analysis confirms that, even using *de novo* assembled transcriptome, gene expression correlates to a specific mutation pattern rather than tumor staging, indicating that this class of ncRNAs could be implicated in the differential activation of MAPK or PI3K pathways according to the mutation pattern.

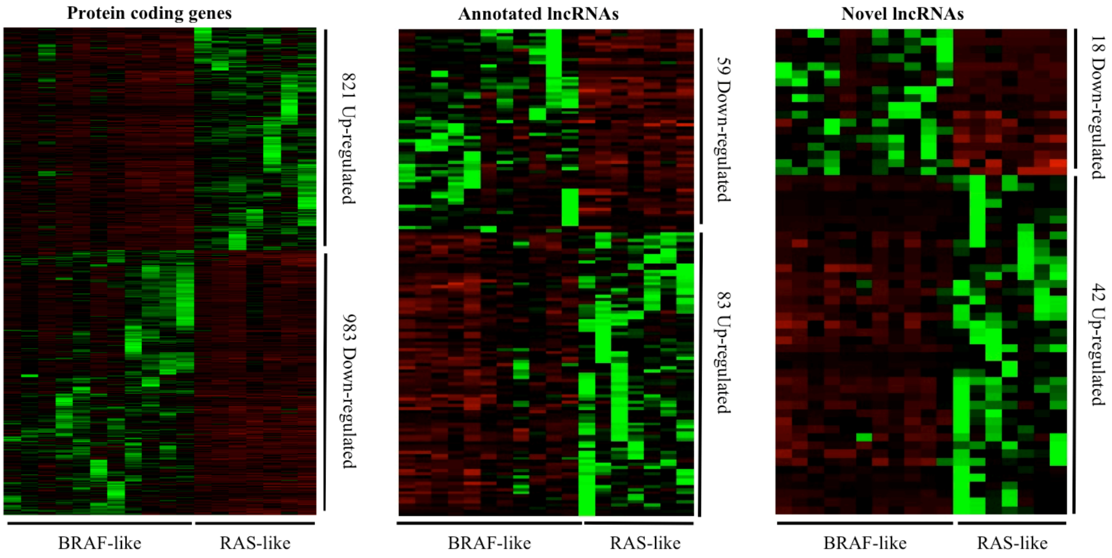


Figure 4.8. Heatmaps showing hierarchical clustering of differentially expressed transcripts within the three RNA classes comparing BRAF-like and RAS-like tumors.

4.2.3 Identification of new long non-coding RNAs altered in PTC

Finally, to identify lncRNAs with aberrant expression in PTC that may act *in cis,* modulating the expression (i.e. activate and/or interfere with) of neighbor protein-coding genes, we defined pairs of genes and lncRNAs that are localized in close proximity using the "nearest transcription start site" (TSS) method, as schematized in Figure 4.9.

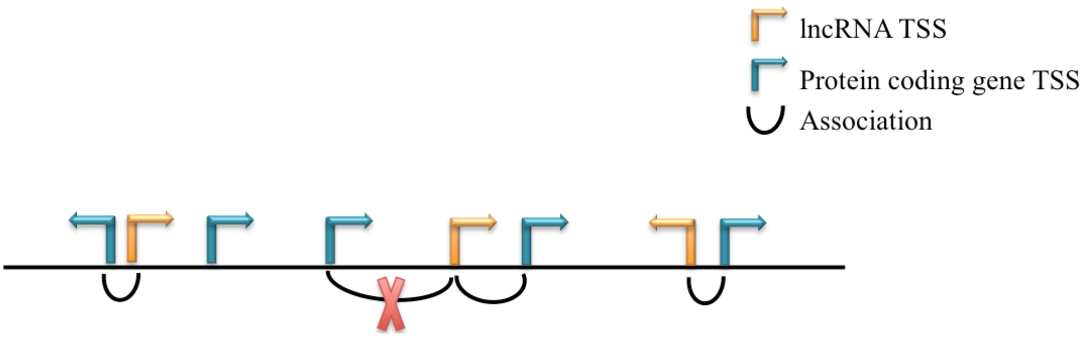


FIGURE 4.9. TSSs of lncRNAs were "associated" to TSSs of nearest protein coding genes.

Subsequently, comparing the transcriptome of tumor and healthy thyroid samples we selected only differentially expressed genes associated to newly identified differentially expressed lncRNAs. Using this approach we could identify 365 lncRNAs-protein coding gene pairs. Moreover, in order to identify lncRNAs potentially involved in cancer, we selected only the genes defined as "cancer driver genes" in a recent review of Vogelstein and colleagues (Vogelstein et al., 2013). Using this approach we found 5 pairs in which both the cancer driver gene and the lncRNA are differentially expressed in PTC samples: *CHD4*, *EZH2*, *TSC1*, *CDKN2A* and *MET* genes (Figure 4.10).

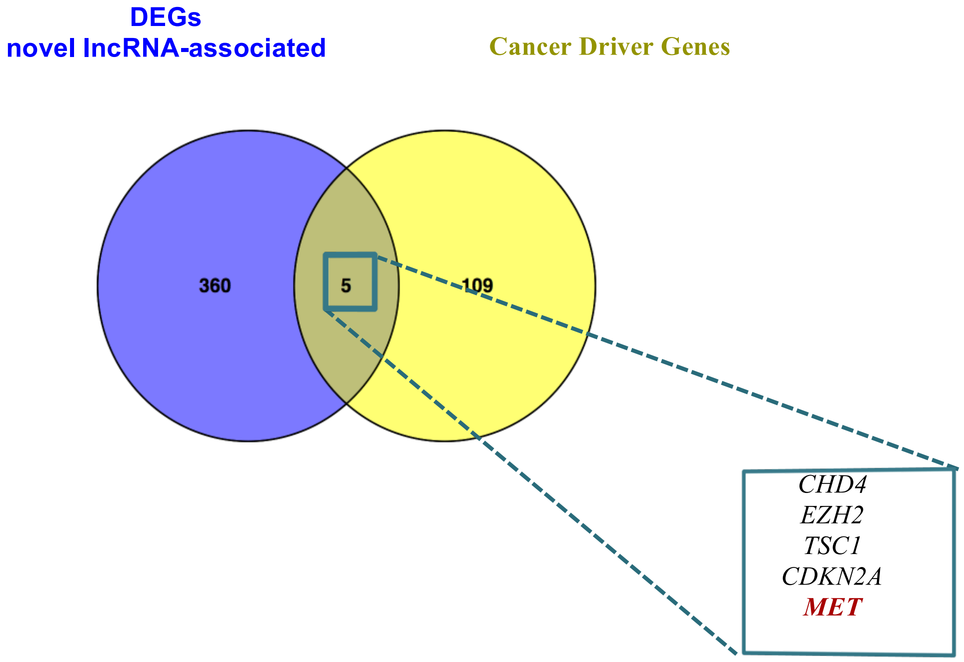


Figure 4.10. Venn diagram intersecting the differentially expressed genes associated to differentially expressed lncRNAs (in blue), and cancer driver genes (in yellow)

Taking advantage of available RNA-Seq datasets (Aversa et al., 2015; Costa et. al 2010) and public RNA-Seq data of cell lines from the ENCODE Project we found that two of these gene-lncRNA pairs are likely to represent false positives, mainly corresponding to misannotated 3'UTRs and families of repeated sequences. Subsequently, we focused our attention on *MET* oncogene since it encodes for a tyrosine kinase receptor (c-Met) that interacts with the cytokine HGF/SF, acting on MAPK pathway, which is significantly affected in PTC. Other than mediating cell proliferation, c-Met has been demonstrated to increase tumor cell motility and invasion. In the thyroid, c-Met overexpression is postulated to play a role in tumorigenesis by conferring a more aggressive and invasive behavior to PTCs (Nardone et al., 2003).

4.2.4 Characterization of a new lncRNA antisense to MET oncogene

De novo transcriptome reconstruction from RNA-Seq data revealed that the novel lncRNA associated to *MET* oncogene maps on chromosome 7q31.2 and, as shown in detail in Figure 2A, this transcript partially overlaps the

GENCODE entries AC006159.3 and AC06159.4 (corresponding to the RefSeq *LINC01510*).

As depicted in Figure 2A, *de novo* assembly revealed the presence of 4 transcripts, 1 of which corresponds exactly to *LINC01510*. Another short transcript, here named *LINC001510*_var (Figure 4.11), derives from the skipping of *LINC01510* exon 2. The remaining two lncRNA transcripts are constituted by 4 and 5 exons, respectively. The first exon of the longer isoforms is transcribed by the first intron of *MET*, but from the opposite strand, indicating that it is an antisense long non-coding RNA of *MET* oncogene. Thus, it was named *MET-AS*. We submitted and annotated these two transcripts in GenBank as *MET-AS1* (GenBank *LN812953*) and *MET-AS2* (GenBank *LN812954*).

Using a combination of RT-PCR, cloning and direct Sanger sequencing we could experimentally confirm their presence and their exon/intron structure. However, we also identified a longer transcript, with two additional exons (Figure 4.11), that we named MET-AS_L.

We examined if the novel lncRNA can be specified by the presence of distinct chromatin marks and/or DNA methylation in the genomic region encompassing its TSS. In addition, as a relatively novel class of lncRNAs associated with active enhancer states able to modulate gene expression both in *cis* and *trans* has been discovered (i.e. enhancer-associated lncRNAs; Marques AC et al., 2013), we also scanned the TSS of this novel lncRNA for H3K27Ac, H3K4me1 and p300 epimarks. Taking advantage of freely available ENCODE ChIP-Seq and chromatin state segmentation data we found that *MET-AS* is characterized by marks of open chromatin, active enhancer states and transcription. The presence of such marks, together with the sequence-based evidence of its transcription (RNA-Seq data and qRT-PCR), its length (>> 200 nt) and the lack of ORF suggest it is likely to be a novel enhancer-associated lncRNA, whose expression is significantly deregulated in PTC.

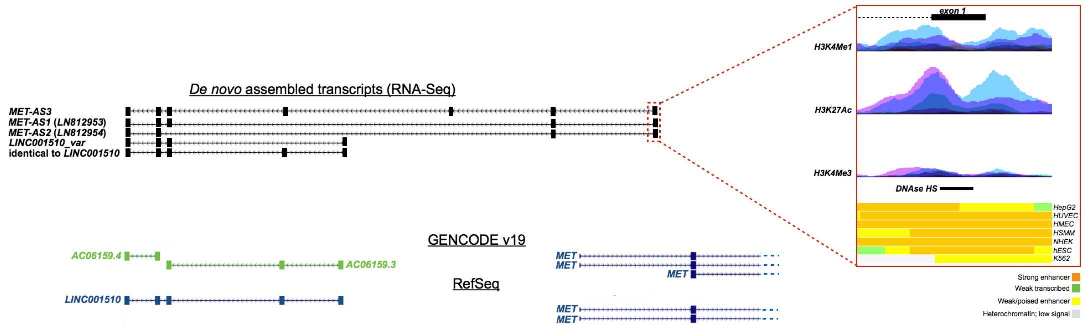


Figure 4.11. *MET* and *MET-AS locus.* In black are shown 5 different isoforms of the novel lncRNA associated to *MET* oncogene*.* In blue are shown *MET* gene (the first two exons) on the right, and the annotated lncRNAs *LINC015210* according to RefSeq annotation on the left. Boxes indicate in exons, dashed lines indicate introns. The three long isoforms of *MET-AS* have the first exon transcribed from the first intron of *MET* oncogene, but from the opposite strand. Thus, it is a novel antisense lncRNA of *MET.* In red boxes on the right are shown H3K27Ac, H3K4me1 and p300 epimarks overlapping the TSS of *MET-AS*.

Using BLAT algorithm we found that all lncRNA transcripts of *MET-AS locus* are conserved in primates, but not in other vertebrates. Subsequently, to define the coding probability (CP) of these novel lncRNAs we analyzed the CP score and compared it to already known lncRNAs and protein-coding genes. The longest transcript of *MET-AS* is constituted by 1943 bp, has a maximum ORF length of 162 bp and reached a coding probability of 0.024, a score quite similar to *XIST* (CP score = 0.027) and *ANRIL* (CP score = 0.039), and very different from those of the protein-coding genes *MET* (CP score= 1) and *BRAF* (CP score = 0.999).

LncRNAs are preferentially localized in the nucleus (Derrien et al., 2012), where they can exert their functions (both in *cis* and in *trans*), even though they can also localize in the cytosolic fraction. Thus, to assess the intracellular localization of *MET-AS* we used the RNA fractionation coupled to RT-PCR in TPC-1 thyroid cell line. The analysis revealed that *MET-AS* has a preferential enrichment in the cytosolic RNA fraction (Figure 4.12).

Moreover, in order to test the tissue-specificity of MET-AS we attempted to amplify this gene in different cancer cells and cell lines. We used different cDNAs available in our laboratory retro-transcribed from both non-tumorigenic epithelial cell lines (i.e. HEK293, MCF10) and tumorigenic cell lines (i.e. MCF7 and MDA-MB-231 from breast cancer, Caco-2 from colorectal adenocarcinoma and J82 from bladder cancer). Interestingly, we were not able to detect the longer isoforms of *MET-AS* in these cell lines.

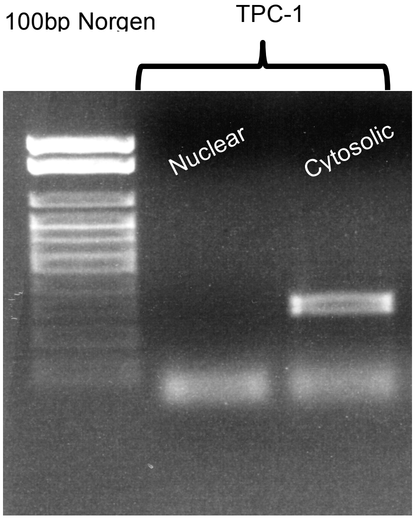


Figure 4.12 Agarose gel showing the cytosolic localization of MET-AS

4.2.5 Both MET and MET-AS expression highly correlates with somatic alterations in PTC

As previously stated, both *MET* and *MET-AS* are differentially expressed in PTCs compared to healthy tissues.

Taking advantage from RNA-Seq data, we analyzed the expression of *MET-AS* and its sense gene in thyroid samples. The expression of *MET* and *MET-AS* shows the same trend in all the analyzed samples although – in line with other independent studies on antisense lncRNAs - the expression of *MET-AS* is clearly lower than *MET* (Figure 4.13). Furthermore, this analysis indicated that the samples defined as BRAF-like show a significantly higher expression of both genes compared to RAS-like and control samples (Figure 4.13).

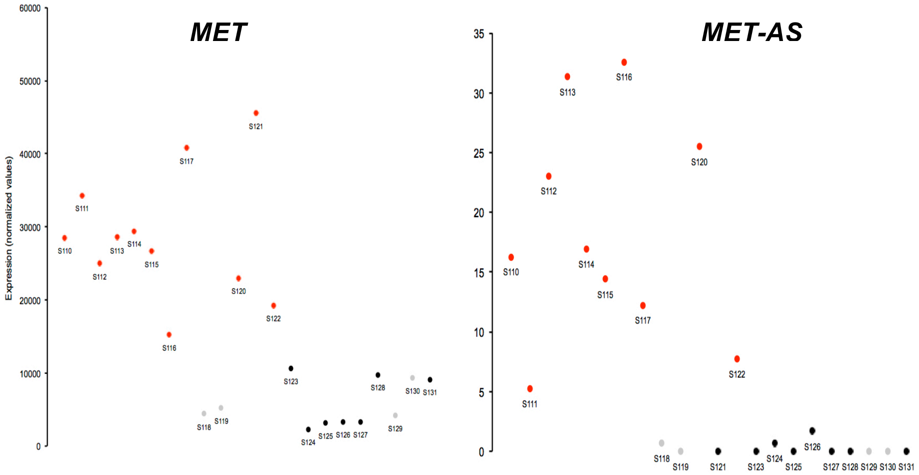


Figure 4.13. Scatter chart with CPM values (y axis) from RNA-Seq data in 22 samples showing MET and MET-AS expression in PTC samples. These genes are over-expressed in BRAF-like (in red) compared to RAS-like (in black) PTCs and control thyroids (in gray).

Subsequently, these results have been validated by qRT-PCR on an independent cohort of 46 PTC tissues and 11 normal tissues types. More in detail, as we found a significant association between the mutational status and the expression of this gene-lncRNA pair, we first characterized the mutational *status* of each tumor samples screening them for the mutations BRAF$_{\text{V600E}}$, and in codons 12, 13 and 61 of *HRAS* and *KRAS* genes, and for the presence of *RET* rearrangements. From this analysis we could classify the patients in BRAF-like (n=28) or RAS-like (n=18). Then, *MET* and *MET-AS* expression was evaluated in the three groups of samples: BRAF-like, RAS-like and Control thyroids (Figure 4.14).

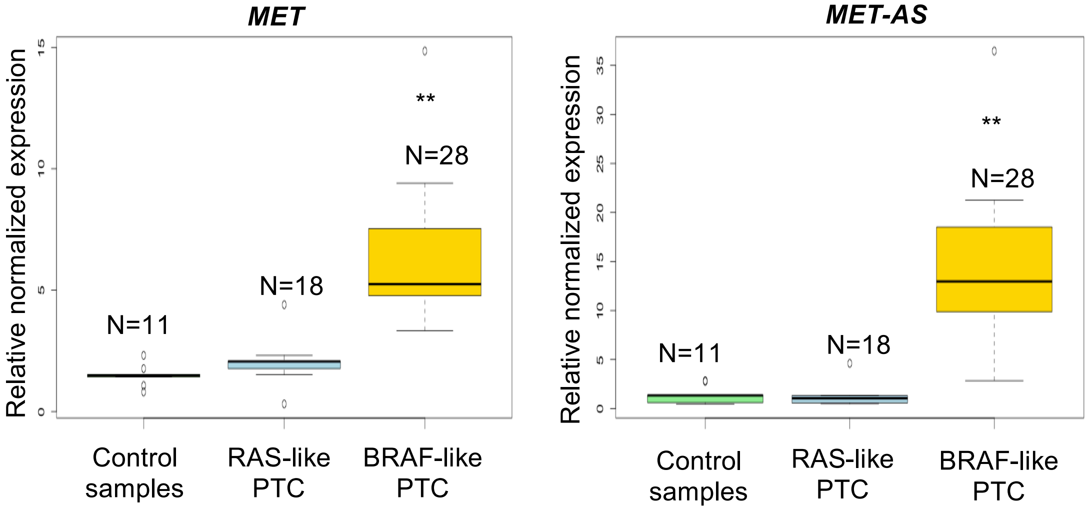


Figure 4.14. qRT PCR for *MET* gene and the novel lncRNA *MET-AS.* BRAF-like group is indicated by yellow boxes, RAS-like and control groups are indicated in light blue and green, respectively. All experiments are normalized to *PPIA* gene.

Real-Time PCR confirmed RNA-Seq data. Both *MET* gene and the novel lncRNA *MET-AS* are up-regulated (*pval*<=0.01) in the BRAF-like, compared to RAS-like and control groups.

4.2.6 Regulating role of *MET-AS* on *MET* expression level

Antisense lncRNAs are functionally very diverse. They can be regulators of gene expression acting as positive and negative modulators of protein-coding genes. Since antisense lncRNAs usually regulate the expression of their sense genes we decided to use RNA interference to knockdown the expression of *MET-AS* and to measure *MET* expression. We used TPC-1 (papillary thyroid carcinoma cell line carrying RET/PTC1). We co-transfected two different siRNAs targeting the 3' region of *MET-AS* transcript and used a scrambled siRNA provided by the manufacturer (Origene) as negative control. The levels of *MET-AS* and *MET* have been then measured in these two cell lines 24, 48 and 72 hours after transfection.

qRT-PCR assay in *MET-AS*$_{siRNA}$ cells confirmed the knock-down of the lncRNA (Figure 4.15). Interestingly, *MET-AS*$_{siRNA}$ cells also displayed a significant down-regulation of MET oncogene, at least at the mRNA level

(Figure 4.15). These results suggest that *MET-AS* may exert a positive effect on *MET* expression.

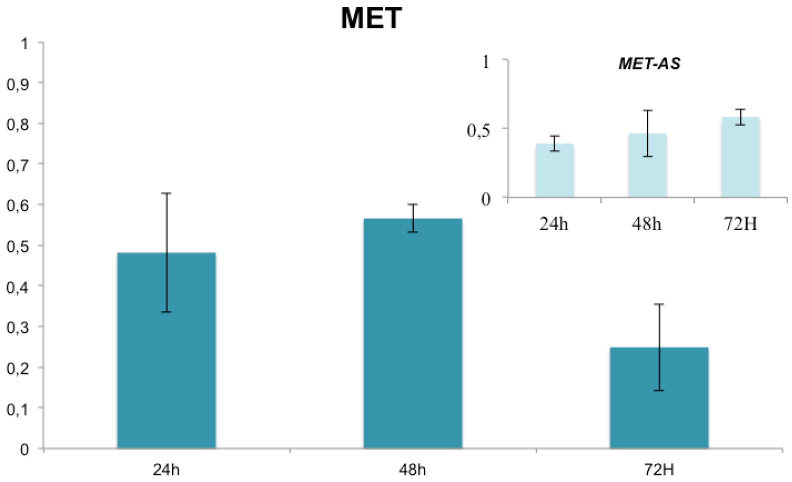


Figure 4.15. We measured both *MET-AS* and *MET* by qRealTime-PCR. Expression values have been normalized using *PPIA* as housekeeping gene. The results indicate a down-modulation of *MET* gene (pval<0.05) using siRNA for *MET-AS* compared to a siRNA control set to 1 in the graph, suggesting a putative role of the lncRNA on *MET* expression. N=5.

4.2.7 *MET-AS* regulates cell cycle progression and cell proliferation

The significant down-regulation of *MET-AS* in BRAF-like biopsies and cells prompted us to explore the potential biological functions of *MET-AS* in carcinogenesis. Therefore, we analyzed cell cycle by FACS analysis 72 hours after *MET-AS* knockdown, where *MET* mRNA levels were significantly down-modulated (75-80% of MET reduction; pval<0.05). Interestingly, we found a significant decrease in the percentage of *MET-AS*$_{siRNA}$ cells in S phase compared to control TPC-1 cells, transfected with scrambled siRNA cocktail (Figure 4.16). These results indicate that the down-regulation of the lncRNA *MET-AS* is able to induce a cell cycle arrest at the G1 phase through the negative modulation of MET oncogene.

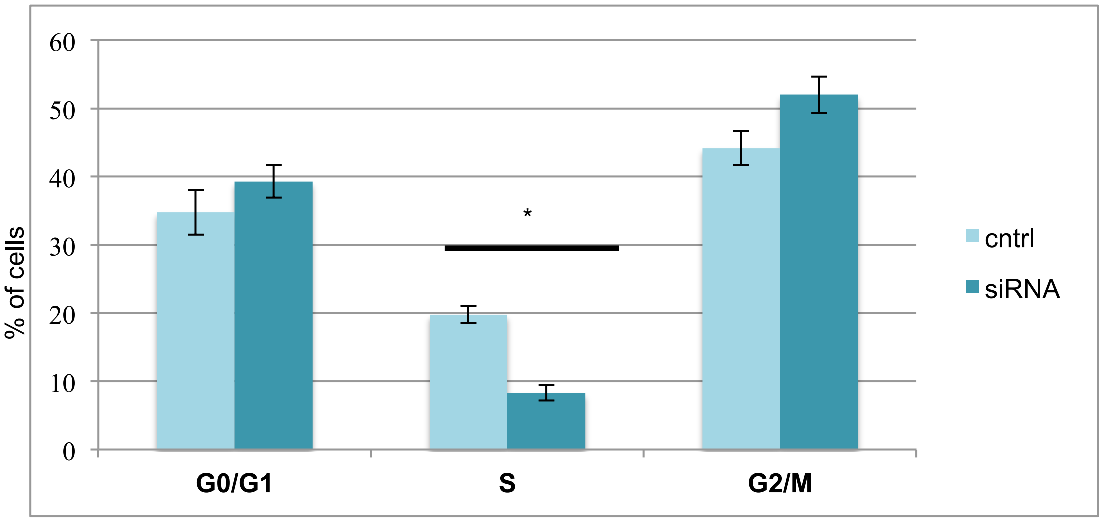


Figure 4.16. Distribution of cells in each cycle phase. The percentage of cells in S phase was decreased in in cells co-transfected with 2 siRNAs targeting *MET-AS*. Control cells have been transfected with a scrambled siRNAs. *, p < 0.05. N=3.

Moreover, we investigated the effect of MET-AS knockdown on cells viability. As shown in Figure 4.17, in accordance to cell cycle results, co-transfection of siRNAs directed against *MET-AS* resulted in a time-dependent decrease in cellular proliferation compared to cells transfected with siRNA scrambled (indicated as “control” in Figure 4.17).

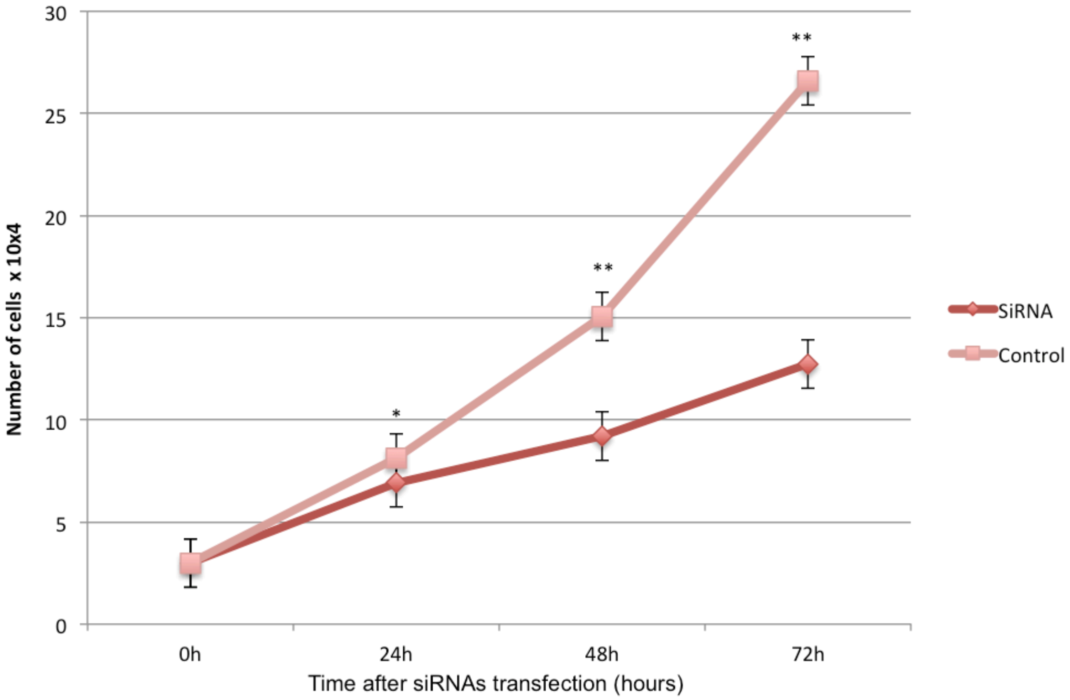


Figure 4.17. Reduction of MET-AS expression in TPC-1 cells using two independent siRNAs results in a significant decrease in cellular proliferation. The number of viable cells after the treatment was measured using the luminescent Cell Titer-Glo assay and expressed as percentage viable cells. Data represents the mean±standard error of the mean of three independent experiments (n=3), each of which was replicated four times.** indicats p-value < 0.01; * indicates pvalue < 0.05. N=3.

4.3 Discussion

Recently, the TCGA Consortium Network has demonstrated that distinct driver mutations in *RAS* and *BRAF* genes lead to striking differences in the activation of signaling pathways in papillary thyroid carcinomas. In particular, in patients with $\mathrm{BRAF}_{\mathrm{V600E}}$ mutation a preferential activation of the mitogen-activated protein kinases pathway has been documented, whereas RAS mutations activate also the phosphoinositide 3-kinase pathways (Cancer Genome Atlas Research Network, 2014).

Since the relative simplicity of the PTC genome, with few dominant mutually exclusive driving events, taking advantage of RNA-Sequencing we could confirm that BRAF- and RAS-mutated tumors have distinct gene expression profiles and that the expression patterns of PTC tumors carrying RET/PTC mutation behave very similar to those with $\mathrm{BRAF}_{\mathrm{V600E}}$ mutation.

The analysis described in this section of the PhD thesis revealed that PTCs can be roughly classified in two main subgroups: BRAF-like and RAS-like. Such a classification is purely based on gene expression similarity of a given patient to samples with BRAF or RET/PTC alterations and RAS mutations, respectively. Interestingly, we found that gene expression is mostly correlated to specific genetic alterations rather than tumor stage, suggesting the importance of the genetic characterization of PTC patients.

Moreover, since the discovery of the transcription of thousands of long RNAs in eukaryotic genomes with no clear coding potential, one of the main challenges has been the understanding of the biological functions associated to these novel transcripts. Long non-coding RNAs are emerging as key regulatory components of gene regulatory networks. However, little is known about the roles of these molecules in disease-relevant organs. Leveraging the power of genome-wide sequencing techniques, joint to the effort of large consortia, like the ENCODE project, is quickly generating a comprehensive catalogue of lncRNAs involved in human diseases, and particularly in cancer.

To date, several examples of lncRNAs able to influence the cellular transcriptional program have been described. They can act either at the pre-transcriptional level by influencing the chromatin remodeling (Tsai et al., 2010, Gupta RA, et al.; 2010) or at the post-transcriptional level by controlling mRNA stability (Liu et al., 2012), cellular localization (Yang et al. 2011), or translation (Carrieri et al. 2012). Growing evidences are showing lncRNAs as important players in cancer, since they are able to regulate both tumor-suppressor and oncogenetic pathways (Huarte and Rinn 2010).

In light of these considerations, to address the rule of lncRNAs in papillary thyroid cancer we have systematically identified and characterized the thyroid long non-coding transcriptome in both pathologic (i.e. papillary thyroid tumors) and physiological conditions (i.ie. healthy thyroids). My PhD project provides a genome-wide screening of lncRNA expression profile in PTC, revealing the presence of hundreds of novel still unannotated lncRNAs in thyroid. Furthermore, as largely described in the Results section, our bioinformatics analysis has revealed that thousands lncRNAs have a different expression patterns in PTC compared to noncancerous thyroids and that *BRAF*-like and *RAS*-like PTCs display significant differences in their expression. These results reinforce the idea that lncRNA are linked to the onset and/or progression of papillary thyroid cancer. In addition, this study also revealed the presence of new lncRNAs that are aberrantly expressed in papillary thyroid cancer and, therefore, may represent potential novel candidates to explore the cancerous process.

Since lncRNA may act *in cis* and exert transcriptional activation or repression of genes transcribed from the same *locus*, the analysis focused on new differentially expressed lncRNAs that are transcribed in close proximity of differentially expressed protein-coding genes considered “drivers genes” in different types of cancers (Vogelstein et al., 2013), During this study, we identified 5 gene/lncRNA pairs potentially involved in the pathogenesis of papillary thyroid carcinoma.

Of note, among them *MET* oncogene revealed to be the best candidate. It encodes a tyrosine kinase receptor for Hepatocyte Growth Hactor (HGF) also known as Scatter Factor (SF; Giordano et al., 1989; Naldini et al., 1991). HGF regulates proliferation and differentiation of epithelial and endothelial tissues of many organs, through the activation of different signaling pathways, including that of MAPK and PI3K. Hepatocyte growth factor and MET control a complex biological program defined as “invasive growth” (Trusolino and Comoglio, 2002). This program coordinates cell proliferation with cell invasion, and provides protection from apoptosis usually occurring in cells removed from their physiological context. MET-driven invasive growth is a physiological program, taking place during embryonic development and post-natal tissue growth and regeneration (Birchmeier and Gherardi, 1998; Boccaccio and Comoglio, 2006; Trusolino et al., 2010). Alterations of the expression level of *MET* oncogene have been reported in a wide variety of human tumors, where it is involved in pathological invasive growth, leading to cancer aggressiveness and metastatic dissemination (Comoglio et al., 2008). Indeed, it was observed that tumor cells with aberrant *MET* expression show an increased ability to cross the endothelial barrier, are characterized by a strong uncontrolled proliferation, and have increased metastatic capacity (Trusolino et al., 2002). Furthermore, in thyroid, c-Met protein overexpression is postulated to play a role in tumorigenesis by conferring a more aggressive, invasive behavior to PTCs. Using a combination of computational and molecular biology, we could identify a new lncRNA antisense to *MET* oncogene and called herein *MET-AS*. It produces at least five different alternative transcripts. The expression analysis showed that *MET-AS* and *MET* expressions are closely associated with a specific mutation profile. Indeed, PTC biopsies with $\text{BRAF}_{\text{V600E}}$ mutation or with rearrangements in the *RET* gene (BRAF-like) displayed a significant over-expression of both genes (*MET* and *MET-AS*) compared to samples with mutations in RAS, or to those with gene expression profiles similar to RAS-mutated samples (Ras-like) as well as to non-tumor samples.

These results suggest to further investigate if the altered expression of this new lncRNA could be related to the deregulation of *MET*, and thus if it might compromise its activity in the thyroid, potentially worsening tumor phenotype.

5 Conclusions

In conclusion, the results of this PhD project confirmed that RNA-Sequencing is a powerful approach to analyze gene expression profiles, gene fusions and even mutations in cancer. Clearly, one of the main limitations - beyond the technical aspect regarding the computational analysis - of using this application to profile the mutational landscape is that only variations in expressed genes can be reliably detected. Conversely, it is intuitive that DNA mutations occurring in "gene deserts" or within genes that are not expressed are often difficult to causally link to the phenotype under examination. However, it must be taken into account that most of the mutations responsible of tumor phenotype fall in actively transcribed genes and lead to the translation of mutated proteins. Data herein described also confirm the genetic heterogeneity of PTC and the possibility to stratify patients according to the driver mutations rather than tumor staging. Notably, using a combination of RNA-Seq and more standard targeted resequecing approach using Sanger methods, this work reliably identified a new gene fusion event involving G4GALNT3 gene, known to act as oncogene in colon cancer, as well as new mutations in candidate driver genes. In addition, where healthy thyroid counterparts were available, we could also validate these variations as somatic (*DICER1* gene).

Such findings pave the way to the development of new potential pharmacological adjuvant therapies in PTC, based on the presence of new affected pathways, such as Notch signaling and chromatin remodeling.

Furthermore, this thesis has investigated new aspects of papillary thyroid cancer biology and has discovered genes not previously known. As such, the implications of this work are broad and suggest that numerous aspects of cancer biology remain still unrevealed. While it is now well known the functional importance of a variety of proteins in cancer biology, now the wide-ranging reports of lncRNAs in multiple cell types raise new questions and challenges for the field of lncRNA research and cancer biology.

In this regard, it is now established that lncRNAs expression signature can differentiate between tumors and the corresponding normal tissues (Yan et al., 2015). In line with this notion, unsupervised hierarchical cluster analysis of our samples distinctly differentiated PTCs from normal tissues, also indicating the presence of a "mutation- specific" lncRNA signature within tuomrs. Although we could not confirm the thyroid-specific expression of these lncRNA, a recent milestone study from Yan and colleagues (2015) has demonstrated that the fraction of lncRNAs displaying tissue-specific expression is about two-fold the one of protein-coding genes. This finding, together with other several reports from independent groups and Consortia, reveals that lncRNAs can be reliably considered new cancer biomarkers, with a predictive value higher than protein-coding genes.

Overall, the second part of this PhD thesis describes the identification - through a combination of computational and experimental approaches - of a novel lncRNA, that we named *MET-AS* as it is transcribed antisense to *MET* oncogene. We found that this new lncRNA is up-regulated in a subgroup of papillary thyroid carcinoma patients, the BRAF-like patients, i.e. patients carrying $BRAF_{V600E}$ mutations or RET/PTC rearrangements as well as those with similar gene expression profiles. In this subgroup of patients we observed that the up-regulation of *MET-AS* co-occurred with high expression levels of MET, in line with the notion that constitutive activation of *BRAF* and *RET* genes significantly enhances and sustains MAPK pathway activation.

MET oncogene up-regulation in PTC increases the malignant phenotype of thyroid cancer cells. Thus, our finding that MET-AS knockdown is able to reduce MET expression and to significantly reduce cell proliferation in a cell model of thyroid cancer is a promising result.

Further studies should be carried to systematically investigate the detailed molecular mechanisms that causally link MET-AS to MET oncogene. However, the results described in this PhD thesis are likely to be promising, not only as we found new players of PTC etiology, but also because we have shown a combined computational and experimental approach that can be reliably extended to other cancer types as well as to study other human diseases.

6 References

Alberti L, Borrello MG, Ghizzoni S, Torriti F, Rizzetti MG, Pierotti MA. Grb2 binding to the different isoforms of Ret tyrosine kinase. Oncogene. 1998 Sep 3;17(9):1079-87.

Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics 2014;pii:btu638.

Aversa R, Sorrentino A, Esposito R, Ambrosio MR, Amato A, Zambelli A, Ciccodicola A, D'Apice L, Costa V. Alternative Splicing in Adhesion- and Motility-Related Genes in Breast Cancer. Int J Mol Sci. 2016 Jan 16;17(1). pii: E121.

Baloch ZW, LiVolsi VA, Asa SL, et al. Diagnostic terminology and morphologic criteria for cytologic diagnosis of thyroid lesions: a synopsis of the National Cancer Institute Thyroid Fine-Needle Aspiration State of the Science Conference. Diagn Cytopathol 2008;36:425–437.

Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, et al. 2010. 3'-end sequencing for expression

Beltran M., Puig I., Pena C., Garcia J.M., Alvarez A.B., Pena R., Bonilla F., de Herreros A.G. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes Dev. 2008;22:756-769.

Benjamini Yoav, Hochberg Yosef. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1. (1995), pp. 289-300.

Berx G., van Roy F. Involvement of members of the cadherin superfamily in cancer. Cold Spring Harb. Perspect. Biol. 2009;1:a003129.

Birchmeier C, Gherardi E. Developmental roles of HGF/SF and its receptor, the c-Met tyrosine kinase. Trends Cell Biol. 1998 Oct;8(10):404-10.

Boccaccio C, Comoglio PM. Invasive growth: a MET-driven genetic programme for cancer and stem cells. Nat Rev Cancer. 2006 Aug;6(8):637-45.

Bos JL. Ras oncogenes in human cancer: a review. Cancer Res 1989;49:4682-9.

Boveri. Zur frage der entstehung maligner tumoren. G. Fischer, 1914

Brown, C.J.; Hendrich, B.D.; Rupert, J.L.; Lafrenière, R.G.; Xing, Y.; Lawrence, J.; Willard, H.F. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 1992, 71, 527–542.

Cabili MN, Trapnell C, Goff L, KoziolM, Tazon-Vega B, et al. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25:1915–27

Calin G.A., Pekarsky Y., Croce C.M. The role of microRNA and other non-coding RNA in the pathogenesis of chronic lymphocytic leukemia. Best Pract. Res. Clin. Haematol. 2007;20:425-437.

Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. Cell 2014; 159: 676-90.

Carcangui ML, Zampi G, Pupi A, et al. Papillary carcinoma of the thyroid: a clinico-pathologic study of 241 cases treated at the University of Florence, Italy. Cancer 1985;55:805–828.

Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, Forrest AR, Carninci P, Biffo S, Stupka E, Gustincich S. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature. 2012 Nov 15;491(7424):454-7.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009 Aug 15;69(16):6660-7.

Castro P, Rebocho AP, Soares RJ, Magalhães J, Roque L, Trovisco V, et al. PAX8-PPARg rearrangement is frequently detected in the follicular variant of papillary thyroid carcinoma. J Clin Endocrinol Metab 2006;91:213-20.

Che MI, Huang J, Hung JS, et al. β1, 4-N-acetylgalactosaminyltransferase III modulates cancer stemness through EGFR signaling pathway in colon cancer cells. Oncotarget 2014;5:3673-84.

Chen, L.-L.; Carmichael, G.G. Decoding the function of nuclear long non-coding RNAs. Curr. Opin. Cell Biol. 2010, 22, 357–364

Chial, H. (2008) Proto-oncogenes to oncogenes to cancer. Nature Education

1(1):33

Cho W.C. OncomiRs: the discovery and progress of microRNAs in cancers. Mol. Cancer 2007;6:60.

Clark MB, Mattick JS. Long noncoding RNAs in cell biology. *Seminars in cell & developmental biology* 2011; **22**(4): 366-76.

Clark, B.S.; Blackshaw, S. Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease. Front. Genet. 2014, 5, 164.

Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol. 2010;2010:853916.

Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur J Hum Genet 2013; 21: 134-42.

Costa V, Esposito R, Ziviello C, Sepe R, Bim LV, Cacciola NA, Decaussin-Petrucci M, Pallante P, Fusco A, Ciccodicola A. New somatic mutations and WNK1-B4GALNT3 gene fusion in papillary thyroid carcinoma. Oncotarget. 2015 May 10;6(13):11242-51.

D'Avanzo A, Treseler P, Ituarte PH, Wong M, Streja L, Greenspan FS, Siperstein AE, Duh QY, Clark OH. Follicular thyroid carcinoma: histology and prognosis. Cancer. 2004 Mar 15;100(6):1123-9. PubMed PMID: 15022277.

Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417(6892):949–954.

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix- Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JW, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA 2002 Mutations of the BRAF gene in human cancer. Nature 417:949–954.

Davies L, Welch HG. Increasing incidence of thyroid cancer in the United States, 1973–2002. JAMA 295:2164–7, 2006.

Deligeorgi-Politi H. Nuclear crease as a cytodiagnostic feature of papillary thyroid carcinoma in fine-needle aspiration biopsies. Diagn Cytopathol 1987;3:307–310.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012 Sep;22(9):1775-89.

Dhomen N, Marais R 2007 New insight into BRAF mutations in cancer. Curr Opin Genet Dev 17:31–39.

Di Cristofaro J, Marcy M, Vasko V, Sebag F, Fakhry N, Wynford-Thomas D, et al. Molecular genetic study comparing follicular variant versus classic papillary thyroid carcinomas: association of N-ras mutation in codon 61 with follicular variant. Hum Pathol 2006;37:824-30.

Diehl, J.A. (2002) Cycling to cancer with cyclin D1. Cancer Biol. Ther., 1, 226–231.

Dietlein F, Thelen L, Reinhardt HC. Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. Trends Genet. 2014 Aug;30(8):326-39. doi: 10.1016/j.tig.2014.06.003.

Dietrich JW, Landgrafe, G, Fotiadou, EH (2012). "TSH and Thyrotropic Agonists: Key Actors in Thyroid Homeostasis". *Journal of Thyroid Research* **2012**.

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, et al. 1999. The DNA sequence of human chromosome 22. Nature 402:489–95.

Ebralidze AK, Guibal FC, Steidl U, Zhang P, Lee S, Bartholdy B, Jorda MA, Petkova V, Rosenbauer F, Huang G, Dayaram T, Klupp J, O'Brien KB, Will B, Hoogenkamp M, Borden KL, Bonifer C, Tenen DG. PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. Genes Dev. 2008 Aug 1;22(15):2085-92.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.

Fan M, Li X, Jiang W, Huang Y, Li J, Wang Z. A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. Exp Ther Med. 2013; 5(4):1143-1146.

Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet. 2014 Jan;15(1):7-21.

Fearon E. R., B. Vogelstein, Cell 61, 759 (1990).

Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014;pii:gku1075.

Frattini M, Ferrario C, Bressan P, Balestra D, De Cecco L, Mondellini P, Bongarzone I, Collini P, Gariboldi M, Pilotti S, Pierotti MA, Greco A 2004 Alternative mutations of BRAF, RET and NTRK1 are associated with similar but distinct gene expression patterns in papillary thyroid cancer. Oncogene 23:7436–7440

Fukushima T, Suzuki S, Mashiko M, Ohtake T, Endo Y, Takebayashi Y, Sekikawa K, Hagiwara K, Takenoshita S. BRAF mutations in papillary carcinomas of the thyroid. Oncogene. 2003 Sep 25;22(41):6455-7.

Gibb EA, Brown CJ, Lam WL. The functional role of long non-coding RNA in human carcinomas. Mol Cancer. 2011 Apr 13;10:38.

Giordano S, Ponzetto C, Di Renzo MF, Cooper CS, Comoglio PM. Tyrosine kinase receptor indistinguishable from the c-met protein. Nature. 1989 May 11;339(6220):155-6.

Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature. 2011 Feb 10;470(7333):284-8.

Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods 2013;10:1081-2.

Grieco M, Santoro M, Berlingieri MT, Melillo RM, Donghi R, Bongarzone I, Pierotti MA, Della Porta G, Fusco A, Vecchio G. PTC is a novel rearranged form of the ret proto-oncogene and is frequently detected in vivo in human thyroid papillary carcinomas. Cell. 1990 Feb 23;60(4):557-63.

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT, Marra MA. Alternative expression analysis by RNA sequencing. Nat Methods. 2010 Oct;7(10):843-7.

Gupta R.A., Shah N., Wang K.C., Kim J., Horlings H.M., Wong D.J., Tsai M.C., Hung T., Argani P., Rinn J.L., et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010;464:1071-1076.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010 May;28(5):503-10.

Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012 Feb 15;482(7385):339-46.

Guttman M., Amit I., Garber M., French C., Lin M.F., Feldser D., Huarte M., Zuk O., Carey Cassady B.W., J.P., et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223-227.

Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013 Jun;9(6):e1003569.

Hawk W, Hazard J. The many appearances of papillary carcinoma of the thyroid. Cleveland Clin Q 1976;43:207–216.

Heravi-Moussavi A, Anglesio MS, Cheng SW, et al. Recurrent somatic DICER1 mutations in nonepithelial ovarian cancers. N Engl J Med 2012;366:234-42.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014 158: 929-44.

Hsu WM, Che MI, Liao YF, et al. B4GALNT3 expression predicts a favorable prognosis and suppresses cell migration and invasion via β_1 integrin signaling in neuroblastoma. Am J Pathol 2011;179:1394-404.

Huang J, Liang JT, Huang HC, et al. Beta1,4-N-acetylgalactosaminyltransferase III enhances malignant phenotypes of colon cancer cells. Mol Cancer Res 2007;5:543-52.

Huarte M, Rinn JL. Large non-coding RNAs: missing links in cancer? Hum Mol Genet. 2010 Oct 15;19(R2):R152-61.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M. et al. (2010) A large intergenic non-coding RNA induced by p53 mediates global gene repression in the p53 response. Cell, 142, 409–419.

Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27:2903-4.

Jan CH, Friedman RC,Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature 469:97–101

Jelinic P, Mueller JJ, Olvera N, et al. Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. Nat Genet 2014;46:424-6.

Jendrzejewski J, He H, Radomska HS, et al. Thepolymorphism rs944289 predisposes to papillary thyroidcarcinoma through a large intergenic noncoding RNAgene of tumor suppressor type. ProcNatlAcad Sci U S A.2012;109:8646-51.

Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel WE, Serve H, Müller-Tidow C. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene. 2003 Sep 11;22(39):8031-41.

Johnsson, P.; Ackley, A.; Vidarsdottir, L.; Lui, W.-O.; Corcoran, M.; Grandér, D.; Morris, K.V. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nat. Struct. Mol. Biol. 2013, 20, 440–446.

Jones P.A., Baylin S.B. The epigenomics of cancer. Cell 2007;128:683-692.

Kaminker JS, Zhang Y, Watanabe C, Zhang Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res. 2007.

Kapranov P, Cawley SE,Drenkow J, Bekiranov S, StrausbergRL, et al. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–19

Kapranov P, Willingham AT, Gingeras TR. Genome-wide transcription and the implications for genomic organization. Nature reviews Genetics 2007; 8(6): 413-23.

Katayama S., Tomaru Y., Kasukawa T., Waki K., Nakanishi M., Nakamura M., Nishida H., Yap C.C., Suzuki M., Kawai J., et al. Antisense transcription in the mammalian transcriptome. Science 2005;309:1564-1566.

Kawamoto Y, Takeda K, Okuno Y, Yamakawa Y, Ito Y, Taguchi R, Kato M, Suzuki H, Takahashi M, Nakashima I. Identification of RET autophosphorylation sites by mass spectrometry. J Biol Chem. 2004 Apr 2;279(14):14213-24.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;14:R36.

Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011;12:R72.

Kimura ET, Nikiforova MN, Zhu Z, Knauf JA, Nikiforov YE, Fagin JA 2003. High prevalence of BRAF mutations in thyroid cancer: genetic evidence for constitutive activation of the RET/PTC-RAS-BRAF signaling pathway in papillary thyroid carcinoma. Cancer Res 63:1454–1457

Kino, T., Hurt, D.E., Ichijo, T., Nader, N. and Chrousos, G.P. (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci. Signal, 3, ra8.

Kroll TG, Sarraf P, Pecciarini L, Chen CJ, Mueller E, Spiegelman BM, Fletcher JA. PAX8-PPARgamma1 fusion oncogene in human thyroid carcinoma [corrected]. Science. 2000 Aug 25;289(5483):1357-60. Erratum in: Science 2000 Sep 1;289(5484):1474.

Kunkel, T.A. and Erie, D.A. (2005) DNA mismatch repair. Annu. Rev. Biochem. 74, 681–710

Kuo CS, Lin CY, Hsu CW, Lee CH, Lin HD. Low frequency of rearrangement of TRK protooncogene in Chinese thyroid tumors. Endocrine 2000;13:341-4.

Langmead Ben, Salzberg Steven L. Fast gapped-read alignment with Bowtie 2 Nat Methods. 2012 March 4; 9(4): 357–359.

Lannon CL, Sorensen PH. ETV6-NTRK3: a chimeric protein tyrosine kinase with transformation activity in multiple cell lineages. Semin Cancer Biol. 2005

Jun;15(3):215-23.

Lavoie H, Therrien M. Regulation of RAF protein kinases in ERK signalling. Nat Rev Mol Cell Biol. 2015 May;16(5):281-98.

Lee JT. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev. 2009 Aug 15;23(16):1831-42.

Leeman-Neill RJ, Kelly LM, Liu P, Brenner AV, Little MP, Bogdanova TI, Evdokimova VN, Hatch M, Zurnadzy LY, Nikiforova MN, et al. ETV6-NTRK3 is a common chromosomal rearrangement in radiation-associated thyroid cancer. Cancer 2014; 120: 799-807.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Li X, Wang Z. The role of noncoding RNA in thyroid cancer. Gland Surg. 2012; 1(3):146-50.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. Nat Genet. 2000 Jun;25(2):239-40. Erratum in: Nat Genet 2000 Dec;26(4):501.

Liu X, Li D, Zhang W, Guo M, Zhan Q. Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. EMBO 2012 J 31(23):4415–4427.

LiVolsi VA. Papillary thyroid carcinoma: an update. Mod Pathol. 2011 Apr;24 Suppl 2:S1-9. doi: 10.1038/modpathol.2010.129.

Ma PC, Kijima T, Maulik G, et al. c-MET mutational analysis in small cell lung cancer: novel juxtamembrane domain mutations regulating cytoskeletal functions. Cancer Res 2003;63:6272-81.

Maher CA, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009;458(7234):97–101.

Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature. 2011 Jan 20;469(7330):343-9.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008 Sep;18(9):1509-17.

Matallanas D, Birtwistle M, Romano D, Zebisch A, Rauch J, von Kriegsheim A, Kolch W. Raf family kinases: old dogs have learned new tricks. Genes Cancer. 2011 Mar;2(3):232-60.

Matsubara K, Okubo K. Identification of new genes by systematic analysis of cDNAs and database construction. Curr Opin Biotechnol. 1993 Dec;4(6):672-7.

Mattick JS. RNA regulation: a new genetics? Nat Rev Genet. 2004 Apr;5(4):316-23.

Mazzaferri EL, Massoll N. Management of papillary and follicular (differentiated) thyroid cancer: new paradigms using recombinant human thyrotropin. Endocr Relat Cancer 2002;9:227–247.

Mears L, Diaz-Cano SJ. Difference between familial and sporadic medullary thyroid carcinomas. Am J Surg Pathol 27(2):266–7,2003.

Medina PP, Romero OA, Kohno T, et al. Frequent BRG1/SMARCA4-inactivating mutations in human lung cancer cell lines. Hum Mutat 2008;29:617-22.

Melillo RM, Santoro M, Ong SH, Billaud M, Fusco A, Hadari YR, Schlessinger J, Lax I. Docking protein FRS2 links the protein tyrosine kinase RET and its oncogenic forms with the mitogen-activated protein kinase signaling cascade. Mol Cell Biol. 2001 Jul;21(13):4177-87.

Mercer T.R., Dinger M.E., Mattick J.S. Long non-coding RNAs: insights into functions. Nat. Rev. Genet. 2009;10:155-159.

Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46.

Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010 Oct;11(10):685-96.

Miller FD, Kaplan DR. Neurotrophin signalling pathways regulating neuronal apoptosis. Cell Mol Life Sci 2001;58:1045-53.

Mourtada-Maarabouni M, Pickard MR, Hedge VL, Farzaneh F, Williams GT. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. Oncogene. 2009 Jan 15;28(2):195-208.

Muller S, Filippakopoulos P, Knapp S. Bromodomains as therapeutic targets. Expert Rev Mol Med 2011;13:e29.

Murakami H, Iwashita T, Asai N, Shimono Y, Iwata Y, Kawai K, Takahashi M. Enhanced phosphatidylinositol 3-kinase activity and high phosphorylation state of its downstream signalling molecules mediated by ret with the MEN 2B mutation. Biochem Biophys Res Commun. 1999 Aug 19;262(1):68-75.

Musholt TJ, Musholt PB, Khaladj N, Schulz D, Scheumann GF, Klempnauer J. Prognostic significance of RET and NTRK1 rearrangements in sporadic papillary thyroid carcinoma. Surgery 2000;128:984-93.

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science. 2008 Dec 12;322(5908):1717-20.

Naldini L, Vigna E, Narsimhan RP, Gaudino G, Zarnegar R, Michalopoulos GK, Comoglio PM. Hepatocyte growth factor (HGF) stimulates the tyrosine kinase activity of the receptor encoded by the proto-oncogene c-MET. Oncogene. 1991 Apr;6(4):501-4.

Nardone HC, Ziober AF, LiVolsi VA, Mandel SJ, Baloch ZW, Weber RS, Mick R, Ziober BL. c-Met expression in tall cell variant papillary carcinoma of the thyroid. Cancer. 2003 Oct 1;98(7):1386-93. PubMed PMID: 14508824.

Nikiforov YE, Nikiforova MN. Molecular genetics and diagnosis of thyroid cancer. Nat Rev Endocrinol. 2011 Aug 30;7(10):569-80.

Nikiforov YE. RET/PTC rearrangement in thyroid tumors. Endocr Pathol 2002;13:3-16.

Nowell PC. The clonal evolution of tumor cell populations. Science. 1976 Oct 1;194(4260):23-8.

Nucera C, Goldfarb M, Hodin R, Parangi S. Role of B-Raf(V600E) in differentiated thyroid cancer and preclinical validation of compounds against B-Raf(V600E). Biochim Biophys Acta. 2009 Apr;1795(2):152-61.

Numata, K.; Kiyosawa, H. Genome-wide impact of endogenous antisense transcripts in eukaryotes. Front. Biosci. 2012, 17, 300–315.

Parmigiani G, Boca S, Lin J, Kinzler KW, Velculescu V, Vogelstein B. Design and analysis issues in genome-wide somatic mutation studies of cancer.

Genomics.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, Reymond A, Hubbard TJ, Harrow J, Gerstein MB. The GENCODE pseudogene resource. Genome Biol. 2012 Sep 26;13(9):R51.

Peyssonnaux C, Eychene A. The Raf/MEK/ERK pathway: new concepts of activation. Biol Cell 2001;93:53-62.

Picardi E, Horner DS, Chiara M, Schiavon R, Valle G, Pesole G. Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. Nucleic Acids Res. 2010 Aug;38(14):4755-67.

Pillai S, Gopalan V, Smith RA, Lam AK. Diffuse sclerosing variant of papillary thyroid carcinoma--an update of its clinicopathological features and molecular biology. Crit Rev Oncol Hematol. 2015 Apr;94(1):64-73.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010 Jun 24;465(7301):1033-8.

Ponting C.P., Oliver P.L., Reik W. Evolution and functions of long noncoding RNAs. Cell 2009;136:629-641.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2.

Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biol. 2014 Nov;24(11):651-63.

Rass K, Reichrath J. UV damage and DNA repair in malignant melanoma and nonmelanoma skin cancer. Adv Exp Med Biol. 2008;624:162-78. doi:10.1007/978-0-387-77574-6_13. Review. PubMed PMID: 18348455.

Ries LAG, Melbert D, Krapcho M, et al. SEER Cancer Statistics Review, 1975–2004. Bethesda, MD: National Cancer Institute 2007.

Rinn J.L., Kertesz M., Wang J.K., Squazzo S.L., Xu X., Brugmann S.A., Goodnough L.H., Helms J.A., Farnham P.J., Segal E., et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 2007;129:1311-1323.

Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145-66. doi: 10.1146/annurev-biochem-051410-092902.

Rinn JL, Euskirchen G, Bertone P, Martone R, LuscombeNM, et al. 2003. The transcriptional activity of human Chromosome 22. Genes Dev. 17:529–40

Robinson MD, McCarthy DJ and Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26, pp. -1.

Rosai J, Carcangui ML, DeLellis RA. Tumors of the Thyroid Gland. Atlas of Tumor Pathology, Fascicle 5. Armed Forces Institute of Pathology: Washington, DC, 1992.

Ruan K., Fang X., Ouyang G. MicroRNAs: novel regulators in the hallmarks of human cancer. Cancer Lett. 2009;285:116-126.

Santoro M, Carlomagno F. Central role of RET in thyroid cancer. Cold Spring Harb Perspect Biol. 2013;5:a009233

Santoro M, Chiappetta G, Cerrato A, Salvatore D, Zhang L, Manzo G, Picone A, Portella G, Santelli G, Vecchio G, Fusco A 1996 Development of thyroid papillary carcinomas secondary to tissue-specific expression of the RET/PTC1 oncogene in transgenic mice. Oncogene 12:1821–1826.

Santoro M, Dathan NA, Berlingieri MT, Bongarzone I, Paulin C, Grieco M, Pierotti MA, Vecchio G, Fusco A. Molecular characterization of RET/PTC3; a novel rearranged version of the RETproto-oncogene in a human thyroid papillary carcinoma. Oncogene. 1994 Feb;9(2):509-16.

Santoro M, Melillo RM, Grieco M, Berlingieri MT, Vecchio G, Fusco A 1993 The TRK and RET tyrosine kinase oncogenes cooperate with ras in the neoplastic transformation of a rat thyroid epithelial cell line. Cell Growth Differ 4:77–84.

Scopa CD, Melachrinou M, Saradopoulou C, et al. The significance of the grooved nucleus in thyroid lesions. Modern Pathol 1993;6:691–694. 34

Segouffin-Cariou C, Billaud M. Transforming ability of MEN2A-RET requires activation of the phosphatidylinositol 3-kinase/AKT signaling pathway. J Biol Chem. 2000 Feb 4;275(5):3568-76.

Sharma A, Rangarajan A, Dighe RR. Antibodies against the extracellular domain of human Notch1 receptor reveal the critical role of epidermal-growth-factor-like repeats 25-26 in ligand binding and receptor activation. *Biochem J* 2013;449:519-30.

Simon J.A., Lange C.A. Roles of the EZH2 histone methyltransferase in cancer epigenetics. Mutat. Res. 2008;647:21-29.

Skinner MA, Moley JA, Dilley WG, et al. Prophylactic thyroidectomy in multiple endocrine neoplasia type 2A. N Engl J Med 353(11):1105–13, 2005.

Smallridge RC, Chindris AM, Asmann YW, Casler JD, Serie DJ, Reddi HV, Cradic KW, Rivera M, Grebe SK, Necela BM, et al. RNA sequencing identifies multiple fusion transcripts, differentially expressed genes, and reduced expression of immune function genes in BRAF (V600E) mutant vs BRAF wild-type papillary thyroid carcinoma. J Clin Endocrinol Metab 2014; 99: E338-47.

Soares P, Trovisco V, Rocha AS, Lima J, Castro P, Preto A, Maximo V, Botelho T, Seruca R, Sobrinho-Simoes M 2003 BRAF mutations and RET/PTC rearrangements are alternative events in the etiopathogenesis of PTC. Oncogene 22:4578–4580

Soares P, Trovisco V, Rocha AS, Lima J, Castro P, Preto A, Máximo V, Botelho T, Seruca R, Sobrinho-Simões M. BRAF mutations and RET/PTC rearrangements are alternative events in the etiopathogenesis of PTC. Oncogene. 2003 Jul 17;22(29):4578-80.

Sone M, Hayashi T, Tarui H, Agata K, Takeichi M, Nakagawa S. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. J Cell Sci. 2007 Aug 1;120(Pt 15):2498-506.

Su, W.-Y.; Li, J.-T.; Cui, Y.; Hong, J.; Du, W.; Wang, Y.-C.; Lin, Y.-W.; Xiong, H.; Wang, J.-L.; Kong, X.; et al. Bidirectional regulation between WDR83 and its natural antisense transcript DHPS in gastric cancer. Cell Res. 2012, 22, 1374–1389.

Sugg SL, Ezzat S, Rosen IB, Freeman JL, Asa SL. Distinct multiple RET/PTC gene rearrangements in multifocal papillary thyroid neoplasia. J Clin Endocrinol Metab 1998;83:4116-22.

Supper J, Gugenmus C, Wollnik J, et al. Detecting and visualizing gene fusions. Methods. 2013;59(1):S24–8.

Taccaliti A, Boscaro M. Genetic mutations in thyroid carcinoma. Minerva

Endocrinol. 2009 Mar;34(1):11-28.

Takahashi M, Buma Y, Iwamoto T, Inaguma Y, Ikeda H, Hiai H. Cloning and expression of the ret proto-oncogene encoding a tyrosine kinase with two potential transmembrane domains. Oncogene. 1988 Nov;3(5):571-8.

Tan YH, Krishnaswamy S, Nandi S, et al. CBL is frequently altered in lung cancers: its relationship to mutations in MET and EGFR tyrosine kinases. *PLoS One* 2010;5:e8972.

Teng KK, Hempstead BL. Neurotrophins and their receptors: signaling trios in complex biological systems. Cell Mol Life Sci 2004;61:35-48.

Tollervey, J.R.; Lunyak, V.V. Epigenetics: Judge, jury and executioner of stem cell fate. Epigenetics 2012, 7, 823–840.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar 1;7(3):562-78.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, Blencowe BJ, Prasanth SG, Prasanth KV. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell. 2010 Sep 24;39(6):925-38.

Trusolino L, Bertotti A, Comoglio PM. MET signalling: principles and functions in development, organ regeneration and cancer. Nat Rev Mol Cell Biol. 2010 Dec;11(12):834-48.

Trusolino L, Comoglio PM. Scatter-factor and semaphorin receptors: cell signalling for invasive growth. Nat Rev Cancer. 2002 Apr;2(4):289-300.

Trusolino L, Comoglio PM. Scatter-factor and semaphorin receptors: cell signalling for invasive growth. Nat Rev Cancer. 2002 Apr;2(4):289-300.

Tsai M-C, et al. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science 329(5992):689–693.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010 Aug 6;329(5992):689-93. doi: 10.1126/science.1192002.

Tseng JJ, Hsieh YT, Hsu SL, Chou MM. Metastasis associated lung adenocarcinoma transcript 1 is up-regulated in placenta previa increta/percreta and strongly associated with trophoblast-like cell invasion in vitro. Mol Hum Reprod. 2009 Nov;15(11):725-31.

Tsuiji H, Yoshimoto R, Hasegawa Y, Furuno M, Yoshida M, Nakagawa S. Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. Genes Cells. 2011 May;16(5):479-90.

Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;1110:11.10.1-11.10.33.

Vance KW, Ponting CP. Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends Genet. 2014 Aug;30(8):348-55.

Vasko V, Ferrand M, Di Cristofaro J, Carayon P, Henry JF, de Micco C. Specific pattern of RAS oncogene mutations in follicular thyroid tumors. J Clin Endocrinol Metab 2003;88:2745-52.

Villegas VE, Zaphiropoulos PG. Neighboring gene regulation by antisense long non-coding RNAs. Int J Mol Sci. 2015 Feb 3;16(2):3251-66.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science* 2013;339:1546-58.

Wan PT, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D, Good VM, Jones CM, Marshall CJ, Springer CJ, Barford D, Marais R; Cancer Genome Project. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell. 2004 Mar 19;116(6):855-67.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470-6.

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

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011 Apr 7;472(7341):120-4.

Wang X., Song X., Glass C.K., Rosenfeld M.G. The long arm of long noncoding RNAs: roles as sensors regulating gene transcriptional programs. Cold Spring Harb. Perspect. Biol 2010. Epub ahead of print 23 June 2010.

Wang Y, Guo Q, Zhao Y, Chen J, Wang S, Hu J, Sun Y. BRAF-activated long non-coding RNA contributes to cell proliferation and activates autophagy in papillary thyroid carcinoma. Oncol Lett. 2014;8(5):1947-1952.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63.

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic acids research. doi:10.1093/nar/gkt006

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K. and Kurokawa, R. (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature, 454, 126–130.

Weels Jr SA, Franz C. Medullary carcinoma of the thyroid gland. World J Surg 2000; 24: 952-956

Xing M. BRAF mutation in thyroid cancer. Endocr Relat Cancer. 2005 Jun;12(2):245-62.

Xing Mingzhao. Molecular pathogenesis and mechanisms of thyroid cancer. Nature review cancer. 13: 184-199 2013

Yamashita AS, Geraldo MV, Fuziwara CS, et al. Notch pathway is activated by MAPK signaling and influences papillary thyroid cancer proliferation. *Transl Oncol* 2013;6:197-205.

Yan X, Hu Z, Feng Y, Hu X, Yuan J, Zhao SD, Zhang Y, Yang L, Shan W, He Q, Fan L, Kandalaft LE, Tanyi JL, Li C, Yuan CX, Zhang D, Yuan H, Hua K, Lu Y, Katsaros D, Huang Q, Montone K, Fan Y, Coukos G, Boyd J, Sood AK, Rebbeck T, Mills GB, Dang CV, Zhang L. Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. Cancer Cell. 2015 Oct 12;28(4):529-40.

Yang L, Lin C, Liu W, Zhang J, Ohgi KA, Grinstein JD, Dorrestein PC, Rosenfeld MG. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell. 2011 Nov 11;147(4):773-88.

Yap K.L., Li S., Munoz-Cabello A.M., Raguz S., Zeng L., Mujtaba S., Gil J., Walsh M.J., Zhou M.M. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol. Cell 2010;38:662-674.

Yoon H, He H, Nagy R, et al. Identification of a novel noncoding RNA gene, NAMA, that is downregulated in papillary thyroid carcinoma with BRAF mutation and associated with growth arrest. Int J Cancer 2007;121:767-75.

Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics. 2011 Jan 15;27(2):175-81.

Yu W., Gius D., Onyango P., Muldoon-Jacobs K., Karp J., Feinberg A.P., Cui H. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature 2008;451:202-206.

Zhang B., Pan X., Cobb G.P., Anderson T.A. microRNAs as oncogenes and tumor suppressors. Dev. Biol. 2007;302:1-12.

Zhu Z, Gandhi M, Nikiforova MN, Fischer AH, Nikiforov YE. Molecular profile and clinical pathologic features of the follicular variant of papillary thyroid carcinoma. An unusually high prevalence of ras mutations. Am J Clin Pathol 2003;120:71-7.

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.