PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

Leveraging transcriptomic analysis to identify transcription factors orchestrating cancer progression

Vivek Das

IFOM-IEO campus, Milan

Supervisor: Dr. Giuseppe Testa

IFOM-IEO campus, Milan

Added Co-Supervisor: Dr. Pasquale Laise

IFOM-IEO campus, Milan

Anno accademico 2015 - 2016

Abstract

Table of Figures

Table of Contents

Abstract

Next generation sequencing (NGS) technology is currently employed to explore the molecular profiles associated to different biological contexts. The application of this technology provides at same time a high-resolution and global view of the

genome and epigenome phenomena, enabling us to study the molecular events underlying many human diseases, including cancer. Our lab tries to exploit the utility of high throughput sequencing technologies generating genomic, transcriptomic and epigenomic data from patient's cohort to study the underlying molecular mechanisms that characterize the specific diseases and map the key regulators that can be critical targets for relevant therapeutic measures. I take the advantage of this technology to mainly understand two aggressive cancers: Ovarian Cancer (OC) and Glioblastoma multiforme (GBM) .

OC is a leading cause of cancer-related death for which no significant therapeutic progress has been made in the last decades. Also in this case, despite multimodal treatment its prognosis remains extremely poor. This is due to the fact that the molecular mechanisms underlying OC tumorigenesis and progression are still poorly understood (Vaughan et al., 2011). GBM is the most common and aggressive primary brain malignancy with very poor prognosis (Frattini et al., 2013). The median survival rate is of 12-15 months (Singh et al., 2012) with 5-year survival that is less than 5% despite the multimodal treatment which include surgery, radiotherapy and chemotherapy. To this end, I will be integrating various genomic and transcriptomic analysis to define the key regulatory actors that characterize the disease progression paving. This integrated analysis has been devised in form of a computational workflow that gives way for a discovery pipeline for physiopathologically meaningful epigenetic targets that can lead to therapies.

Chapter 1- Introduction

1.1 A brief account on Cancers

Cancer is one of the most dreaded diseases that a human life can be inflicted with and it has been affecting lives for several years. It's causal can be attributed to uncontrolled cell divisions thus affecting the nearby neighboring cells. This creates an environment that might affect a specific tissue/organ or other tissues. Considering the statistics provided by the World Cancer Research fund International, the age-standardized rate for men and women combined was 182 per 100,000 in 2012 for all cancers (that excludes non-melanoma skin cancer). It was also shown by the same agency that there were among 14.2 million cases recorded in the world in 2012 of which Australia ranked 3rd, USA had sixth highest spot, Italy with a rank of 21st while UK was as high as 23rd among the highest cancer rate for men and women together with age-standardized rate per 1000,000 people. The ratio for male to female inflicted with cancer is around 10:9. The Figure 1 in the heatmap shows the interactive map of cancer incidence in the world till 2012

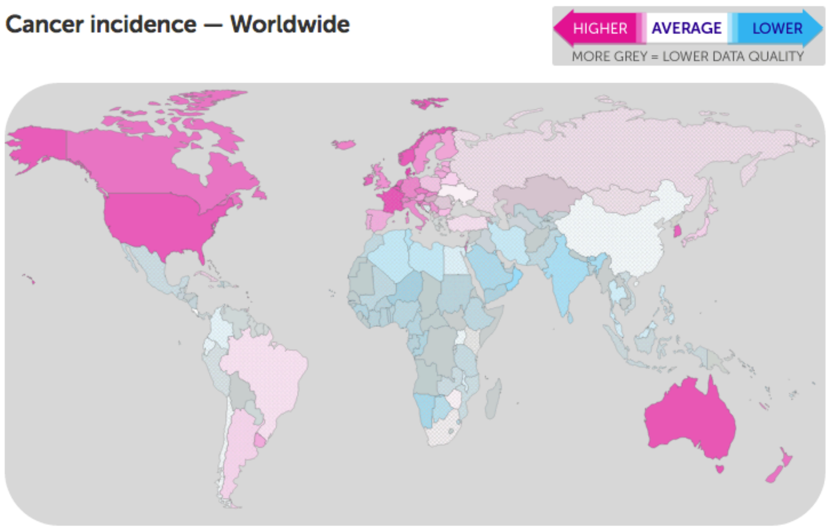


Figure 1: Shows the interactive map of cancer incidence around the world estimated in 2012

It shows the interactive map of cancer incidence around the world estimated in 2012. The color pink denotes a higher incidence of the disease while blue shows a lower followed by white for average or mid-level incidence. The color grey shows that there is not enough data quality to assess for those countries. This figure is adapted from the website of Cancer Research UK and the sources as stated in the websites are from GLOBOCAN 2012 v1.0 and United Nations, Department of Economic and Social Affairs, Population Division (2013). All these data access have been made in 2013.

There have been reportedly more than 200 different cancers and several consortiums and agencies have been working effectively to understand the causal method along with preventive measures that can be made in order to create new therapies. Genomics over the last few years have been able to contribute a lot in making us understand the genomic and epigenomic landscapes of different cancers. In this thesis I have been providing my findings and understanding for two different aggressive cancers, namely high-grade serous **Ovarian carcinoma (OC)** and **Glioblastoma Multiforme (GBM).** Our lab has been able to get patient's samples suffering from both the two above-mentioned disease and generated genomic, transcriptomic and epigenomic profiles to understand the progression and development of the disease in separate projects. There is a dual phase of the OC studies where we have been able to generate the transcriptomes of OC tumors coming from high grade serous ovarian cancer, namely of epithelial origin (EOC) and ascetic fluid (AS) and their possible tissue of origin which are Fimbria (FI) and ovarian surface epithelium

(OSE). I have intended to study the transcriptomic behavior of the tumor and the normal tissues and what molecular mechanisms are associated with the tumor progression. The second study made in the OC project was to study the key genetic players contributing the mutant landscape of the tumor and that these lethal variations in primary tumors were also preserved in reprogrammed tumor derivatives achieved through somatic nuclear reprogramming. In order confirm this paradigm that somatic reprogramming is compatible with keeping the mutant genome intact, we generated exome profiles from both high and low grade OC tumor patients and their tumor-induced pluripotent stem cell (tumor-iPSCs) clones. This study enabled us to study how iPSCs can be used as a tool to reconstruct the developmental history of a disease. In the case of the GBM project we had the resource of collecting samples from patients having Primary and recurrent GBM with collaboration from University of Bonn. We have been able to generate the genome wide expression profiles of the patients with a unique partitioning of cellular compartments of a brain for each patient, which on critical assessments will reveal the core areas that are more prone to attain a relapse. I intended to study the transcriptomic behavior of these topological partitioning of various sections of the GBM tumor in light of capturing the molecular mechanisms underlying the primary GBM evolution to its relapse.

Both these studies presented in my thesis will be able to find out targets that might help in understanding the disease development and also elucidate key genetic and transcriptomic regulators which in turn can help in better prognosis. Farther I provide a brief account of the history of OC and GBM.

1.1.1 A brief account on Ovarian Cancer

Ovarian cancer (OC) is a leading cause of cancer-related death for which no significant therapeutic progress has been made in the last decades. It is considered to be one of the fifth cause of cancer-related death and the most lethal gynaecological malignancies (Bowtell 2010). The diagnosis is usually pretty late as far as the stage is concerned. It is often misconstrued to other gastrointestinal or reproductive system diseases. No major improvements have been made since cis-platin treatment (or carboplatin treatment) was introduced in 1980s and more recently its combination with the taxanes. This is due to the lack of markers and therapeutic targets that reflects: i) our poor understanding of the molecular mechanisms underlying OC biology; ii) the absence of suitable models, since available OC cell lines fail to recapitulate the histopathological origin of the disease (Vaughan et al., 2011). The overall 5-year survival rate is 31%.

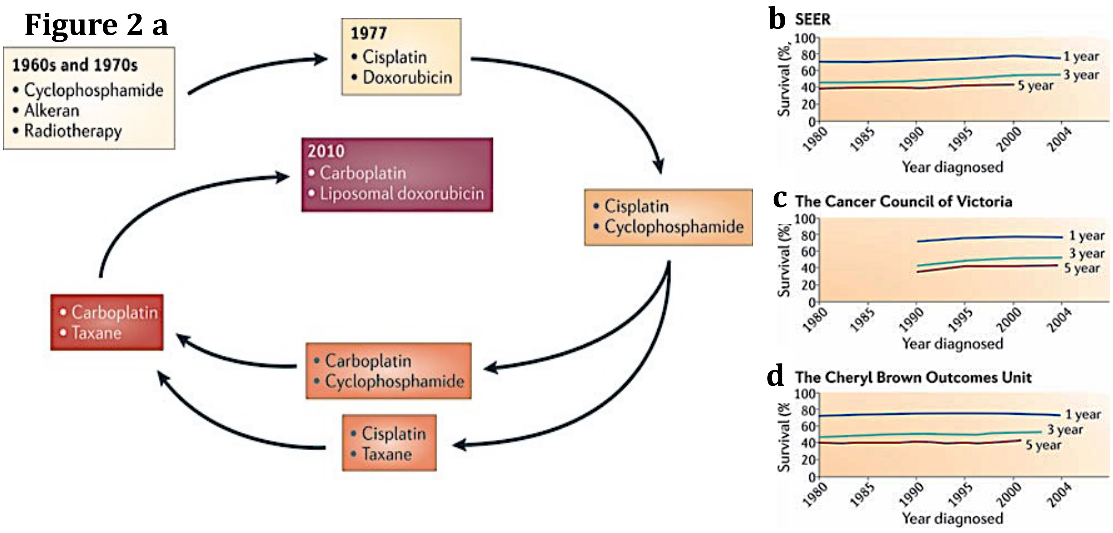


Figure 2: Current treatments of Ovarian Cancer have not lead to great patient's care

a) Shows the different treatments that have been developed for OC patients since 1960s till date while 2 b), c), d) – Shows a panel of disease-free survival curves since 1980 without much of a major improvement over the years and highlighted by 3 major units from USA, Australia and Canada which provides epidemiologic information on the incidence and survival rates of OC here. Adapted and modified from (Vaughan et al., 2011) .

Ovarian cancer refers to a heterogeneous population of tumors rather than a single one, which can arise from 3 different cellular types accounting to development of different types of tumors:

i) Epithelial that can give rise to tumors starting from the cells that cover the outer surface of the ovary

ii) oocytes or germ cells that give rise to Germ cell tumors.

iii) Structural tissue cells or stromal cells which produces female hormones like estrogen and progesterone.

Non-epithelial ovarian tumors roughly accounts to ~40% of all tumors which rarely reaches malignancy. 90% of the tumors are epithelial in origin (will be referred as Epithelial Ovarian Cancer, EOC), which is predominant in nature and is considered to be very heterogeneous. These can be farther classified into serous, endometrioid, mucinous, clear cell, transitional cell, squamous cell, mixed epithelial, and undifferentiated (Iarc, Tavassoéli, & Devilee, 2003). These types of tumors are further classified into benign, malignant and borderline (low malignant potential tumors, LMP), which on tumor subtypes classifications, are termed as low or high-grade.

Table 1

Stage	Description
I	These types of tumors are confined to the ovary or fallopian tubes
II	These types of tumors are extend or metastasize from ovaries and/or fallopian tube to adjacent pelvic structures
III	These types of tumors are metastasis extending out of the pelvis and/or to the regional lymph nodes
IV	These types of tumors include metastases at distant sites that includes patients with metastatic stages of parenchymal liver/splenic and extra-abdominal

Table 1: Different stages of epithelial ovarian cancers based on the International Federation of Gynecological Oncologists (FIGO) system and informs the doctor about the growth and spread of the tumor.

Serous ovarian cancer (SOC) are conferred Type I and Type II (Vang, Shih, & Kurman, 2009) based on their histopathology and patterns of mutation. LMP tumors give rise to Type I SOC which are regarded mostly as low grade or serous borderline and characterized by BRAF and KRAS mutations (Singer et al., 2003) and devoid of TP53 mutations (Wong et al., 2010) while the Type II SOC is frequented with TP53 mutations (A. A. Ahmed et al., 2010) , absence of BRAF and KRAS mutations (Wong et all 2010) and often mutations in BRCA1 and BRCA2 (Hylander et al., 2013) have been associated in this type as well. Thus Type II tumors are chromosomally instable and also referred to as high-grade serous ovarian cancer (HGSOC). As a matter of fact these HGSOC's have been reportedly sensitive to platin- and PARP inhibitors-based treatments (Bowtell, 2010).

The origin cell of ovarian cancer is still very poorly understood and stands largely debatable since the tumors upon diagnosis are found to have already invaded the major areas of patient's abdomen, which includes ovaries as well. Earlier years it was to be believed that the origin was ovarian in nature. However growing evidences over the years cite HGSOCs might originate from distal epithelial cells of the fimbria of fallopian tube. These tubes are usually every close to the ovaries. Gene expression study of HGSOCs demonstrated fallopian tube epithelium as potent origin (Tone et al., 2008). Several other studies also showed many carcinomas in fallopian tube of both invasive and non-invasive nature. These findings made scientists to believe that primary OC might be stimulated from shedding of malignant cells on the ovary from the fallopian tubes (R. J. Kurman, 2013) . All these theories established owing to the evidences clearly hints that both tissues might pose as origin for subset of HGSOCs. Thus identification of signatures might better allow us to understand the developmental niches of the disease namely the cell of origin that precisely will identify critical pathways contributing to OC pathogenesis and better up the future prognosis. There is thus an acute need to identify new therapeutic targets and prognostic biomarkers that can improve OC management. The lack of markers and targets reflects our poor understanding of the disease. Available OC cell lines fail to recapitulate the histopathological origin of the disease this strongly indicate for an important role of studying the developmental aberrations leading to OC pathogenesis.

Moreover, OC is one of the tumor types where we have poor knowledge about the relative contribution of genetic versus epigenetic alterations to the tumor phenotype, preventing the identification of molecular pathways as prognostic signature or therapeutic target.

1.1.2 A brief account on Glioblastoma multiforme

Glioblastoma multiforme (GBM) is one of the most common and aggressive primary brain malignancy with very poor prognostic and less efficient therapeutic measures (Chen et al., 2012) . It has been graded by WHO as Grade IV astrocytoma accounting about 15 percent of all brain tumors where the occurrence age in adults is between 45 to 70 years. The median survival rate is of 12-15 months (Frattini et al., 2013) with 5-year survival rate that is approximately of 4% even though there have been highly advanced and innovative techniques for detection like that of spectroscopy and perfusion. Even the treatments over the years, which include surgery, followed by radiation therapy or a combined radiation therapy and chemotherapy, have not been able to improve the prognosis. This is often attributed to the fact that GBM patients are highly resistant to therapeutic drugs owing to its heterogeneous nature of the disease in different patients. To ensure effective therapies it is important to understand the development and progression of the disease by breaking down the genetic and the transcriptomic background of the disease. GBMs are classified as primary and secondary. Primary GBMs arise de novo as there is no previous clinical history while secondary are believed to be arising from progressive accumulations of genetic alterations in grade III anaplastic astrocytoma or from low-grade diffuse astrocytoma which in turn have over time developed from a low grade tumors (grade II). Both primary and secondary GBMs share identical histopathological features along with wide spread cell proliferation and aggressive invasiveness. These tumoral cells migrate beneath the subdural sheets along with white matter tracts while infiltrating the parenchymal cells. These invasions while progressing into the perivascular space coupled with angiogenesis can lead to hemorrhages. Patients with GBM is also often associated with recurrence due to high resistance of these tumor-infiltrating cells to conventional chemo and radiotherapy thus making it difficult for prognosis.

In recent year several studies have highlighted the molecular characterization of GBM thus outlining different subtypes, which shows over expression of specific subset of genes.

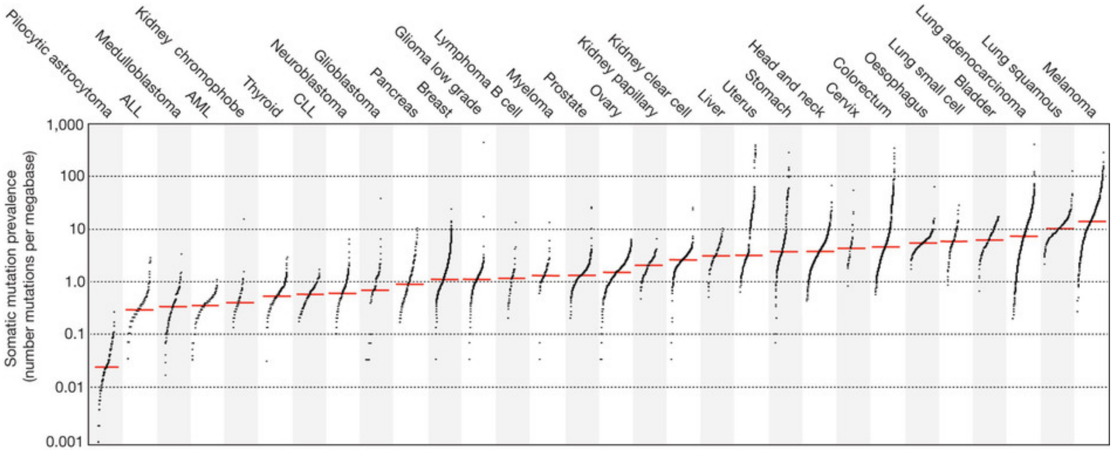
The cell of origin for GBM is still not clearly understood and has varied speculations and various theories have been proposed about it. Ideally it should stand for normal cells, which upon implications of events give rise to a tumor formation. However certain considerations have led to the proposal of theories where initiators are other cells and not astrocytes of oligodendrocytes. This can be stated first due to the difference in de novo versus progressive GBMs (i.e. primary versus secondary GBMs) have given rise to the possibility that specific genetic or epigenetic alterations can act upon different cells which could finally give rise to different diseases of the same order of tumor. The other possibilities are presence of the intra and inter heterogeneity of the GBM patients, which accords for complex cytological subtypes with altering patterns of genetic lesions and transcriptomic profiles. This farther reopened the cell of origin debate and thus it is not clear if a tumor origin can be solely classified based on appearance even they look similar under the microscope. This is due to the fact that same mutations that characterize the tumor may account for different subtypes of the tumor. This has been well characterized while studying whole-genome pathology, clinical aspects of the disease and glioma animal models. This was well documented by (Zong, Verhaak, & Canoll, 2012) in their review which shows neural stem cells (NSCs), progenitors of glial cells which also includes progenitors of oligodendrocytes and lastly astrocytes could be concluded to serve as origin cells for gliomas. In case of the mutations that cells sustain might not transform on its own but their progenitors can actually transform to behave as cells of origin for GBMs. Recently a study published by (Steed et al., 2016) developed a new computational method that could present the origin of different glioblastoma subtypes with the usage of clinical images derived from 217 brain tumor patients. We know that the 4 subtypes of glioblastoma are classical, neural, pro-neural and mesenchymal (Verhaak et al., 2010) . It is found by Chen's team that sub-ventricular zone (SVZ) serves originating region for pro-neural and neural GBMs while other two subtypes are associated to be farther distributed and settled away from SVZ and serve as region containing cells of origin. This could be explained on the basis of the mutations that give cancer in NSCs of the SVZ produced pro-neural and neural GBMs while the similar mutation occurring in a region far away from SVZ in another cell population gave the other subtypes of GBMs namely mesenchymal and classical. SVZ is accorded as the region which seats the neural stem cells and these cells migrate from the center of the brain to its outward side during developmental stages of the brain and thus different cell types are formed that human brain is made up of. This hypothesis that Chen’s team developed was confirmed in animal model that was developed at Cincinnati Children's hospital by pediatric hematologist-oncologist Lionel Chow. Thus more multi-disciplinary approaches in future will be able to define the cellular origin of GBM and help in identifying therapeutic markers that will be able to alter the tumor type and improve the prognosis.

1.2 Somatic changes contributing to cancers development and progression

Cancer is collective group of several different diseases that can develop in any parts of the human body. As we all know that cells are the basic units that forms the basis of human body. They grow and divide to in order to maintain the needs of the body. Cells are often replaced and replenished with new ones when they are either old or damaged however if there are certain genetic changes due to an impairment that disrupts the orderly process then it might give rise to cancer. This happens if the autoimmune system of the body fails to take care of such events. During these phases uncontrollable growth of cells may take place, which might give rise to a tumor mass that can be either malignant or benign. A tumor mass when malignant can grow and spread in other parts of the body while the benign tumor will grow but not reach out to other parts. There are also some forms of cancer like lymphoma and myeloma, which do not form tumor. In this thesis my focus of study is GBM and HGSOC. Mutations in the gene stand out to be one of the major hallmarks of cancer that may be central to its evolution. Multiple mutations can form the basis of cancers and since cancer undergoes cellular inheritance it is likely to be suggestive that tumor progression can be driven by mutagenesis. With the advent of NGS technologies now it is possible to dissect the entire human genome at deeper resolution. Thus we can zero-in at nucleotide sequences and get a more unprecedented power to catalogue more and more mutations. Tumors can be a result of extensive heterogeneity of cancer cells or may be also a result of chemo-resistance. This can be attributed to the fact of specific mutations in genes which might arise due to DNA damage that are un-repairable or incurred errors during DNA synthesis. According to the Hanahan and Weinberg (Hanahan, 2000) proposal genetic alterations in cancer can be catalogued by 6 very distinct and complementary changes that change the physiology of cells thus enabling tumor growth and metastasis. These are most commonly referred to as self-sufficient growth signals, abnormal sensitivity to antigrowth signals, bypass apoptosis, possibilities of illimitable replications, continued angiogenesis, and metastasis and tissue invasiveness. According to the authors suggestions all tumors should attain all these 6 hallmarks while tumor of the same type might be composed of different gene mutations in varied order. Genetic alterations constituting of cancers might not be only dictated or restricted to point mutations or multi-nucleotide variations (MNVs) like INDELs. Even gains and losses of large nucleotide segments in the genome spanning a few to too many kilobases or for that matter the whole chromosomes might give rise to the tumor formation and growth. These alterations or variations lead to aneuploidy and chromosomal aberrations. Copy number alterations (CNAs) that are present in the normal genome in forms of deletions, insertions, or duplications are referred to as germline while those that occur or are acquired during the lifetime of an individual are referred to as somatic CNAs (SCNAs). SCNAs are often dubbed as major contributors to cancer development as well and in particular for solid tumors.

1.2.1 Somatic variations in cancer at point or multinucleotide is critical for its development

Mutations in somatic cells may be a result of i) adulteration in the DNA replication machinery, ii) being subjected to exo- or endogenous mutagens, iii) enzymatic alteration of DNA and iv) imperfect DNA repair. Even long exposure to radiations or UVs or smoking addictions can result in somatic mutations may result in triggering specific cancers. Several projects have been laid in the field of cancer genomics that exploit the power of high throughput sequencing technologies to find somatic mutational signatures across different cancers. This uncovers the diversity and complexity of the process in human carcinogenesis. Better resolution with the mutational signatures referring to a landscape is expected in future with definitive features using higher reductionist approach. This will take shape when more whole-genome sequencing of cancer patients will be added up in the research wagon thus helping in elucidating the mechanistic basis of some signatures which as of now are partially understood. The paper of (Alexandrov et al., 2013) as shown in Figure 3 and Figure 4, which tries to catalogue somatic mutational landscape post analysis of over several cancers thus revealing 20 distinct signature is of high relevance as some of these signatures are redundant in most cancer classes while others are restricted to single cancer classes. It also reveals some signatures association with age of cancer diagnosis or with DNA maintenance impairment/mutagenic exposures and to kataegis.



Adapted from (Alexandrov et al., 2013)

Figure 3: Representation of median mutational scores of patient's across different cancer types

Each and every dot in the figures represents a sample or a patient while red horizontal lines represent the median mutational scores in the respective cancer types. The vertical axis which is log scaled represents the mutation number per megabase while somatic point mutations coming from different cancer represents the horizontal axis with median red line as the median somatic mutational.

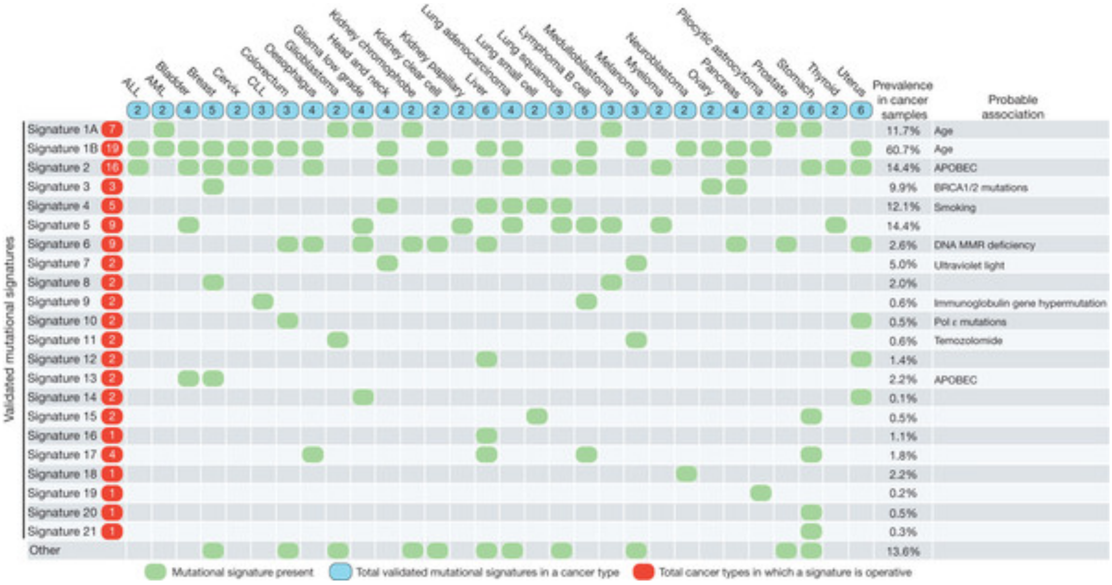
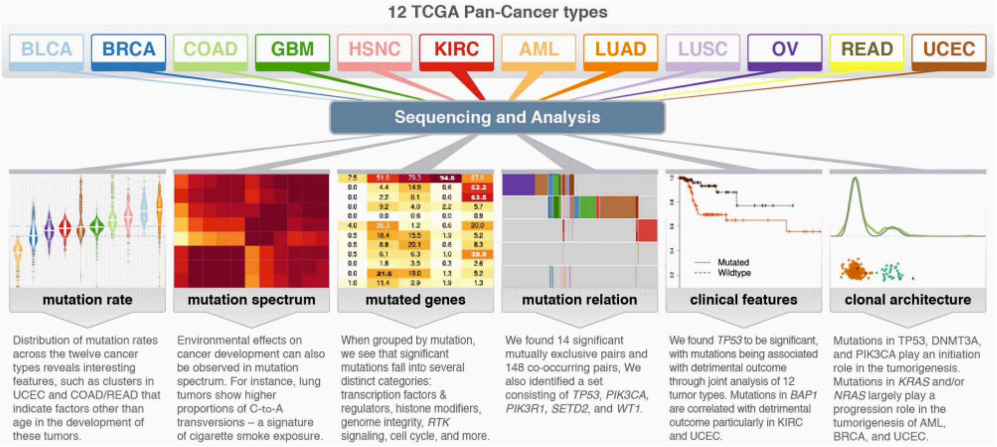


Figure 4: Different cancers are typically ordered in alphabetical manner in horizontal axis with mutational signatures in vertical axis

Another paper (Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye, 2013) that tries to identify significantly mutated genes(SMGs) from the TCGA Pan-Cancer data set from 3281 tumors across 12 cancer types provides us with insightful revelations about 127 SMGs. These SMGs highlights the cellular and enzymatic process are linked to tumor progression. Out of these 127 SMGs 67 was given driver status based on *'ratiometric'* method using COSMIC catalogs. This paper illustrates the mutational frequency distributions shared among different tumor or exclusive to specific tumors, unifies the tissues of origin with SMGs, to DNA impairments, environmental and mutagenic influences. It also presented the mutational drivers required for oncogenic drive and relevance of mutations in tissue specific TFs. Even histone modifiers are seen to be under the mutational burden in these 12 cancers. Finally it scores the importance of knowing the clonal architecture of each patients from SMGs and relevant effect of these genes in survival making it clinically important for future prognosis or panel cancer sequencing studies. The importance of the paper findings can be seen in the below Figure 5



Distribution of mutation rates across the twelve cancer types reveals interesting features, such as clusters in UCEC and COAD/READ that indicate factors other than age in the development of these tumors.

Environmental effects on cancer development can also be observed in mutation spectrum. For instance, lung tumors show higher proportions of C-to-A transversions – a signature of cigarette smoke exposure.

When grouped by mutation, we see that significant mutations fall into several distinct categories: transcription factors & regulators, histone modifiers, genome integrity, *RTK* signaling, cell cycle, and more.

We found 14 significant mutually exclusive pairs and 148 co-occurring pairs, We also identified a set consisting of *TP53*, *PIK3CA*, *PIK3R1*, *SETD2*, and *WT1*.

We found *TP53* to be significant, with mutations being associated with detrimental outcome through joint analysis of 12 tumor types. Mutations in *BAP1* are correlated with detrimental outcome particularly in KIRC and UCEC.

Mutations in TP53, DNMT3A, and PIK3CA play an initiation role in the tumorigenesis. Mutations in *KRAS* and/or *NRAS* largely play a progression role in the tumorigenesis of AML, BRCA, and UCEC.

Figure 5: Schematic representation of the TCGA Pan-Cancer mutation dataset identifying SMGs, cancer-related cellular processes, and genes associated with clinical features and tumour in an orderly fashion

1.2.2 Impact of somatic variations in forms of copy number alterations in cancer initiation and development

A copy number alteration in somatic cells is often contributing to oncogenesis and is of high importance. Often large segments of chromosome might undergo aberrant changes resulting in activations of oncogenes or silencing tumor suppressor genes (TSGs). These acquired changes might be detrimental to the patients and so Pan-Cancer projects tried to characterize these SCNA events in 4934 cancers. The project observed whole-genome abnormalities having increasing frequencies with varied SCNAs, TP53 mutations, CCNE1 amplifications and changes in the PPP2R complexes in 37% of the cancer samples across 11 cancer type (Zack et al., 2013) . Out of 140 regions revealed having significant SCNAs, 102 regions were found devoid of known oncogenes and TSGs targets and 50 SMGs. This really provides even SCNAs affect focal regions other than oncogenes and TSGs that might contribute to tumorigenesis. This study gave mechanistic insights of functional consequences of these SCNAs in tumorigenesis and their phenotypic effects. This study also underlines the critical role of SCNAs in oncogene activation or TSG inactivation that is a great leap in cancer diagnostics and prognosis in addition to the new regions having erroneous copy alterations. However there lies a challenge of segregating the driver and passenger SCNAs in cancers. Thus it is critical not only to assess the positively selected SCNAs but also track the increased generation rates or decreasing negative selections of the same. This will help in understanding underlying mechanisms of their generation and outline functional context. Apart from this, whole-genome sequencing also provides specific information of rearrangements in sequences that forms each SCNA. This information's can assesses the genetic heterogeneity within tumors separating early events from the later. Thus mechanistic insight of SCNAs generation and selective pressures reshaping them can also be known.

1.3 Epigenetics

Epigenetics, coined by embryologist and geneticist, C.H Waddington (Waddington, 1942) is a field of study where variations at cellular and phenotypic level are due to external or environmental factors which results in turning genes "on and off" thus affecting the gene expressions. They do not alter the nucleotide sequences and thus DNA remains unscathed. Although the field is still ever growing and debatable but things like phenotypic changes that includes expression of genes, which can be passed on mitotically, and/or meiotically without affecting an individual's DNA sequence are widely accepted. A molecular event can be epigenetic when it is transmitted to its progeny on its own with a mechanism that is still maintained post DNA replication and rounds of cell divisions and finally result in expression of genes. Histone modifications and **DNA methylations** are *bona fide* marks that alter the gene expression without affecting the parental DNA sequence. The first report of epigenetics being linked to any disease-affecting humans was that of a cancer in the year 1983. It reported that patients with colorectal cancers showed striking differences when their normal tissues were compared to the affected tissues. The level of methylation was higher in normal tissues in the same patients when compared to their diseased tissue counterparts (Feinberg & Vogelstein, 1983) . The consequence that could be derived was the turning off methylated genes results in activating certain other genes that ensues chromatin rearrangement. CpG sites seats the DNA methylation and DNA in proximity to promoter regions have higher concentrations of CpG sites or even defined as CpG islands but they are free of methylation in normal cells. These islands are heavily coded with methylation in cancer cells, which in turn silences some important genes. This is pretty much evidently documented in early cancer development as an epigenetic alteration (Egger, Liang, Aparicio, & Jones, 2004) , (Jones & Baylin, 2002), (Robertson, 2002). In my study however we are trying to use DNA methylation as a developmental tracer where we are trying to find the origin of tissue for unknown primary tumors. (Sproul et al., 2012) and (Moran et al., 2016) previously did this work as described in details in the **Results** section of the thesis. They used information from DNA methylation to trace back the developmental tissue type for different tumors for which either the origin was known or unknown. Taking a cue from those study we developed a similar strategy on those lines. This enabled us to understand the real origin of these OC tumors and then studying the transcriptional landscapes (Detail of the strategy is mentioned in the **Results** section). Epigenetic alteration has also been dubbed to cause mutation even though the DNA sequence is unaltered. Familial or inherited cancers have most of the genes silenced due to methylation, which also in turn shuts down TSGs, and thus the DNA repair machinery is at stake. Some of them include genes like MGMT, MLH1 cyclin-dependent kinase inhibitor 2B (CDKN2B), and RASSF1A. Even hypermethylation have been reported to microsatellites instabilities, which in turn have been involved in many cancers like colorectal, endometrial, ovarian, and gastric cancers (Jones & Baylin, 2002). These are a few examples in cancers where epigenetic changes have been discovered which are important for transforming cells to cancer. Thus understanding these mechanisms will be very important to dissect the key regulator that develop and maintains the cancer progression leading to finding better prognostic markers for therapies. There are even some other mechanisms that change the gene expression like different histone posttranslational modifications, nucleosome repositioning and remodeling and small noncoding RNAs.

Other epigenetic modifications like that of the histones can intervene with the transcriptional landscape regulating its maintenance and transmission like that of Polycomb and Trithorax protein group of proteins mediated histone modifications (Kouzarides 2007, Orkin & Hochedlinger, 2011a), (Laugesen &

Helin, 2014), Steffen and Ringrose 2014*). Most commonly and well characterized is H3K4 and H3K27 methylation, which is in effect of the Trithorax (TrxG) and polycomb (PcG) protein groups. The trimethylation at lysine residues for H3K4 is associated activation of genes (Byrd & Shearn, 2003) while that of H3K27me3 with repression (Kirmizis et al., 2004). There are 4 chromatin states that are critical to tumor considering the genome-wide distribution of these 2 epigenetic marks. These are i) repressed state which is due to only H3K27me at gene promoters, ii) an active state characterized by H3K4me3 genes, iii) bivalent state where both marks are present at the promoters and finally iv) the silent state where both are absent and RNA polymerase is not bound as well. These tight interplays when disrupted associates with tumors. As e.g., in metastatic prostate (Varambally et al., 2002), breast (Kleer et al. 2003, Raaphorst et al. 2003), and bladder cancer (Arisan et al., 2005) EZH2 is often seemed to be over expressed which is a catalytic subunit of polycomb repressive complex2 (PRC2), that can give way to cancer progression through p14 and p16 (Ink4A/ARF locus) (Bracken et al., 2007) silencing. In ovarian cancer EZH2 over expression is found in advanced stage (Rao et al., 2010) and associated with poor survival and cisplatin resistance (L. Hu, McArthur, & Jaffe, 2010). Thus it is indicative that late progression of OC is implicated with histone mark repression. Another interesting revelation in OC epigenetics world was seen in the (Chapman-Rothe et al., 2013) article, which defined, that bivalent marked genes promotes malignancy and leads to chemoresistance. All these findings clearly scores the importance of histone modifications in cancer and epigenetic tweaking might help in understanding the tumor biology and also lead to finding new novel drugs that are otherwise ineffective and posing a strong resistance in prognosis.

1.4 Induced pluripotent stem cell reprogramming as a tool to reconstruct disease history and developmental aberrations

As we know cancer is considered to portray a tight interplay between genetic and epigenetic changes. Thus it is important to dissect this contribution of epigenetic from genetic in cancer pathogenesis. Established cell line, tumor xenografts or even engineered murine models can capture early stages of cancer progression but still not to its entirety. The scope is limited and unable to elucidate heterogeneity among patients from same tumor phenotypes. Even fresh samples cannot grow in culture for indefinite time. We also know that human embryonic stem cells (ESCs) can provide us with plethora of information but its derivation has major ethical bottlenecks. All these can be handled with that of induced pluripotent stem cell reprogramming of somatic tissues. It is a multi-step process that requires epigenetic resetting keeping the genetic background conserved with the establishment of a transcriptional landscape that is very much compatible with pluripotency stages.

In recent years there have been great advances in the reprogramming of differentiated somatic cells to pluripotency induced by only four transcription factors (TFs)- OCT3/4, SOX2, KLF4 and c-MYC (Carette et al., 2010) . It is known that oncogenic transformation frequently involves procurement of *de novo* developmental programs that is comparable to cellular reprogramming. This provides an immense self-renewal potential to cells, which is a distinct feature shared with induced pluripotent stem cells (iPSCs) (K. Takahashi & Yamanaka, 2006). This agreement is fortified at a mechanistic level by coordinators and impeding agents shared between the two processes of tumorigenesis and cell fate reprogramming (Kazutoshi Takahashi & Yamanaka, 2006) . Most of the reprogramming TFs are considered to be bona fide oncogenes while some might act as reprogramming barriers which commonly belong to known tumor suppressors e.g. p53 and Ink4A/Arf can take part both in proliferation and apoptosis. In addition to this, chromatin regulators that are also established reprogramming modulators have been observed mediating oncogenesis (Orkin & Hochedlinger, 2011). These proof of concepts indicates the partly recapitulation of epigenetic circuitry essential for cellular reprogramming during transformation of cells. By the same token, the epigenetic rewiring that accompanies cellular reprogramming can thus be used to also erase cancer-specific epigenetic aberration by reprogramming cancer cells from primary tumors into iPSC lines. This in turn would open up access solely to genetic lesions, which remain intact. The utility of this approach has been demonstrated recently for several types of cancer cells, for which the epigenetic rewiring resulted also in reduced tumorigenicity upon differentiation and transplantation in vivo (Stricker et al., 2013). The Table 2 shows the usage of cell derived iPSC lines in different cancers.

Table 2

CANCER TYPES	**CELL LINE OR PRIMARY CELLS**	**REPROGRAMMING METHOD**	**REFERENCE**
Melanoma	Colo	Retroviral mir-302s	(S.-L. Lin et al., 2008)
Prostate cancer	PC-3		
Melanoma	R545	Lentiviral OCT4, KLF4, and c-MYC	(Utikal, Maherali,

			Kulalert, & Hochedlinger, 2009)
Chronic myeloid leukemia (blast crisis stage)	KBM7	Retroviral OSKM	(Carette et al., 2010)
Colorectal cancer	DLD-1, HT-29	Combination of retroviral or lentiviral OSKM, NANOG, LIN28, BCL2, KRAS, and shRNA for tumor suppressors optimized for each cell line	(Miyoshi et al., 2010)
Esophageal cancer	TE-10		
Gastric cancer	MKN45		
Hepatocellular cancer	PLC		
Pancreatic cancer	MIAPaCa-2, PANC-1		
Cholangiocellular cancer	HuCC-T1		
Chronic myeloid leukemia (chronic	Patient-derived bone marrow cells	Episomal OSKM, NANOG, LIN28, SV40 LT	(K. Hu et al., 2011)

phase)			
Lung cancer	A549	Lentiviral OSNL and nondegradable HIFα	(Mathieu et al., 2011)
Chronic myeloid leukemia (chronic phase)	Patient-derived bone marrow cells	Retroviral OSKM	(Kumano et al., 2012)
Breast cancer	MCF-7	Retroviral OSKM	(Corominas-Faja et al., 2013)
Juvenile myelomonocytic leukemia (JMML)	Patient-derived mononuclear cells with E76K missense in PTPN11 gene	Lentiviral OSKM	(Gandre-Babbe et al., 2013)
Pancreatic ductal adenocarcinoma (PDAC)	Patient-derived pancreatic ductal adenocarcinoma	Lentiviral OSKM	(J. Kim et al., 2013)
Glioblastoma multiforme (GBM)	GBM neural stem (GNS) cell lines	PiggyBac driving OCT4 and KLF4	(Stricker et al., 2013)
Osteosarcom	SAOS2, HOS,	Lentiviral OSKM,	(X. Zhang, Cruz,

a	MG63	NANOG, LIN28	Terry, Remotti, & Matushansky, 2013)
Liposarcoma	SW872		
Ewing’s sarcoma	SKNEP		
Myelodysplastic syndromes (MDS)	Two patients with del(7q)-MDS	Lentiviral OSKM	(Kotini et al., 2015)
Li Fraumeni syndrome (LFS)	Three patients with G245D missense in p53 gene	Sendai viral OSKM	(Lee et al., 2015)
Ewing sarcoma (EWS)	CHLA-10	Episomal OKSM	(Moore et al., 2015)

This table is adapted and modified from (J. J. Kim, 2015)

This technology has the potential to be used to model and treat human disease. This can be seen in figure 6 a where the patient has been shown having a neurodegenerative disorder. This figure shows the generation of patient specfici iPSCs generated by ectopic co-expression of TFs in cells isolate dfrom skin biopsy. There are two complementary approaches that have been shown in the figure. In case of diseases which are dominated by a single disease-causing mutations (e.g. familial Parkinson's disease), gene targeting methods could be employed leading to DNA repair (represents the rigt part of the figure 6a). This is followed by differentiation of the corrected gene in specific patients for which the iPSCs are derived into affected neuronal subtype (e.g. midbrain dopaminergic neurons) and transplantation of it in the patients brain. The alternative approach relies on directed differentiation of specific patient iPSCs into affected neuronal subtype (represents the left part of the Figure 6a) and model the disease in-vitro giving way to drug screen which in turn might aid the discovery of new therapeutic drugs. This approach can be in fact be implemented on tumor-iPSCs that can be in be indefinitely expanded and differenetiated in cells post derivation from all three germ layers. This would need usage of relevant protocols of differentation towrads cancer lineages that can be exploited from the of vast number of available differentiation protocols (Cheng et al. 2012, Ye et al. 2013, Sampaziotis et al. 2015). This scores a perfect setting not only capturing the genetic architecture of the parental tissue but also the epigenetic memory. So we can finally also understand the epigenetic mechanisms at the early onset of tumor (Figure 6b). Thus we will also be able to uncover the molecular pathways that are specifically susceptible of altering tumor malignancy and aggressiveness as a result of epigenetic alterations.

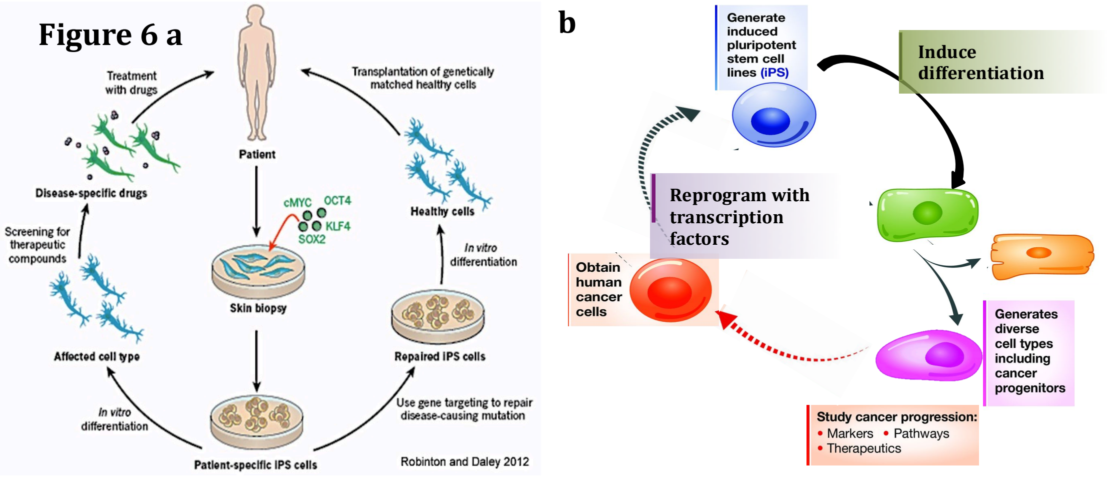


Figure 6 a
Treatment with drugs
Patient
Transplantation of genetically matched healthy cells
Disease-specific drugs
cMYC
OCT4
KLF4
SOX2
Healthy cells
Screening for therapeutic compounds
In vitro differentiation
Skin biopsy
Affected cell type
Repaired IPS cells
Use gene targeting to repair disease-causing mutation
In vitro differentiation
Patient-specific IPS cells
Robinton and Daley 2012
b
Generate induced pluripotent stem cell lines (iPS)
Induce differentiation
Reprogram with transcription factors
Obtain human cancer cells
Generates diverse cell types including cancer progenitors
Study cancer progression:
• Markers • Pathways
• Therapeutics

Modified from Jungsum K et al., The EMBO Journal (2015) embj.201490736

Figure 6: A representation of how iPSC technology can be used to model diseases in humans

However one of the most important bottlenecks in using iPSC directly towards clinical applications for disease modeling is having the risk of genomic instability that is induced by reprogramming factors. This genomic instability at times also includes malignant growth. There have been reports where genetic mutations have been identified in the iPSCs that were supposed to be incorporated in second phase of human clinical iPSC-based therapy. Although direct consequence of these mutations leading to any adverse effects have not been evidenced yet but it still raises a concern about the genomic instability that reprogramming induces. IPSCs have shown and evidenced to be carrying genomic instability in forms of chromosomal aberration, point mutations and CNVs. However they are also pretty much evidenced to bring in the parental variations that existed in a line that were subjected to iPSC reprogramming. The genetic variations of iPSCs are envisaged to be originating from i) pre-existing variations in its parental somatic cells (this is what can be targeted to understand the genetic background while studying tumor mutational landscape that are reprogrammed to tumor iPSCs), ii) mutations arising during the reprogramming process and iii) different passaging brings forth a mutational burden that arises during prolonged culturing conditions (Figure 7 top panel). In my thesis of using tumor-iPSCs as a disease modeling tool, I am more concerned in identifying the parental pre-existing variations in forms of SNV, CNV and chromosomal aberrations that passes from tumor to its tumor-iPSC derivative. However, according to the review of Mahashito Yoshihara et al. we should be take into account that these pre-existing parental variations are randomly captured. These random variations are then passed on to and expanded in subsequent iPSC generations. There is also another second possibility that these pre-existing parental lesions that pass from tumor to tumor-iPSCs can also add to reprogramming and proliferative power of iPSCs via selective advantage mechanisms. WGS studies on iPSCs have revealed numerous point mutations occurring after onset of iPSC reprogramming and also confirmed the heterogeneity of SNVs in a single iPSC clone by studying subclones from a distinct iPSC clone. Ideally the parental lesions in forms of SNVs have ~50% allelic frequency since they are present in one of the allele of all iPSCs generated from one parental cell. However, reprogramming induced mutations are very immediate even before a cell division setting. These mutations can be seen having frequencies of ~50%, ~25% and ~12.5% (Figure 7 lower panel). This is one reason it is difficult to state all the SNVs with 50% frequencies are from parental cell but Suguira et al showed that mutations arising from reprogramming have a pattern of transversion-dominant while that coming from parental and passage-induced show pattern of transition-dominant. Even Gore et al confirmed that mutations happen stochastically within the cell populations and a single iPSC derived from human exhibit additional mutations when these iPSCs are compared at early and later passaging via WES analysis. It was confirmed that prolonged culture conditions are also testament to bring forth new mutations. So one important way to capture the parental tumor clonality and parental point mutational landscape via tumor-iPSC should be creating more iPSC clones from the tumors at similar passaging conditions and then delineate the differences of acquired and pre-existing mutations thus reconstructing the tumor clonality from all the tumor-iPSC clones derived from it.

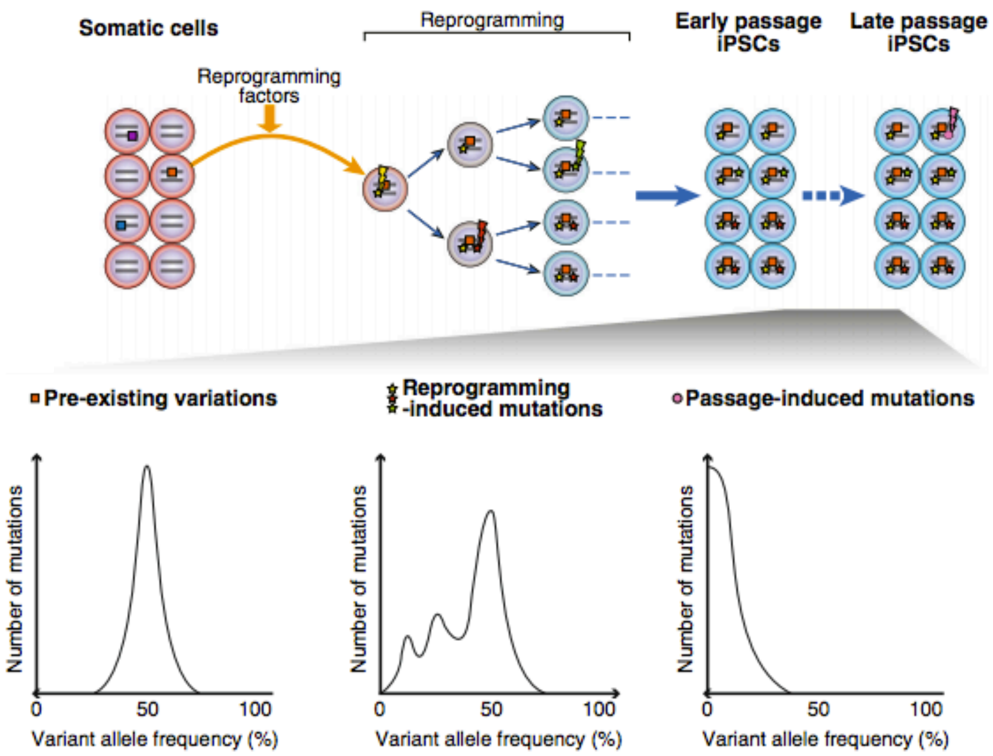


Figure 7: A representation of genetic variations origin in iPSCs adapted from Mahashito Yoshihara et al.

The figure is divided into two panels.

The top panel shows genetic make up of iPSCs can be originating from parental pre-exiting variations, reprograming induced variations during processing of reprogramming and finally variations arising during passaging during prolonged culture. The lower panel shows the distribution of the variant allele frequency in pre-existing variations, variations arising due to reprogramming and variations arising due to passaging in iPSC generations.

1.5 Impact of Transcription Factors in cancer development, progression and maintenance

Transcription factors (TFs) are group of proteins that follow the conversion process or specially transcribe DNA into RNA. They cover a great number of proteins, which is involved in genes transcriptional machinery (initiation and regulation). An important feature of them is the presence of DNA-binding domains that allows them to bind specific regions on the DNA sequences commonly known as enhancers or promoters. These bindings can be referred to as as promoter specific (the binding occurs at the promoters sequences of the DNA) nearby the transcription start sites (TSS) initiating the transcription initiation complex (TIC). Conversely when the binding is at regulatory regions far away from the TSS that controls the expression of other genes, it is known as enhancer specific. These bindings lead to activation or repression of the transcriptional status of the related gene. These regulatory regions can be either upstream or downstream of the gene to be transcribed ranging from a few to over thousands of bases. These TFs have been seen to be dysregulated in cancer. The cancer cells are majorly dependent on them for their development, progression and maintenance. A plethora of human TFs have been identified, validated and recorded as candidates by several databases like JASPAR (Mathelier et al., 2016), TRANSFAC (Wingender, Dietze, Karas, & Knüppel, 1996), AnimalTFdb (H. M. Zhang et al., 2012). These databases records for lists of genes regarded as TFs based on experimental and predicted findings. TFs have also been regarded as either having an oncogenic or tumor suppressing effect. This indicates that TFs that binds other genes promoting the tumorigenicty are oncogenic while others that rescue of the tumor phenotype whose down-regulation again promotes oncogenesis are tumor suppressors. In order to achieve the maximum from this TF mediated implications in cancer, one important discovery is to find these sites on the genes that are being regulated by the TFs. This referred to as motifs (D'haeseleer, 2006), which are recurring DNA or amino-acid sequences that has seats proteins or upstream TFs. This binding allows the gene regulation presumably implicating some biological functions. This is fairly important since gene expressions are often controlled by upstream TFs, which has a consensus-binding site on these genes. This binding of upstream TFs with a target group of genes having motifs leads to regulation of expression of the target genes. This expression can be detrimental in nature to normal cells or initiate tumor cells to develop farther, migrate, proliferate and invade nearby cells giving way to tumorigenesis. So targeting these TFs can be highly efficient in treating specific types of tumors that will eventually target nuclear hormone receptors improving clinical efficiency. Several chemical approaches are being worked upon these days to regulate the effect of these TFs to understand their potential in driving cellular transformation, which accounts for making them favorite targets for drug discovery.

1.6 Aim of the thesis

This thesis is divided into two subparts where I am trying to identify molecular mechanisms underlying to aggressive states of cancer, namely Ovarian Cancer (OC) and Glioblastoma multiforme (GBM) through an approach involving genetic and transcriptomic data that has been acquired and generated in our lab.

The initial part of the thesis involves transcriptomic data and genetic data of patient's samples from OC. This part is sub-divided into two. The initial phase of the OC studies relies on assessing the transcriptomic differences of OC tumors coming from high grade serous ovarian cancer, namely of epithelial origin (EOC) and ascitic fluid (AS) and their possible tissue of origin which are Fimbria (FI) and ovarian surface epithelium (OSE). Since from the availability of the OC tumors at the time of surgery is a multi-order mass which is formed of solid tumor floating in the ascetic fluid and so we chose to profile the transcriptomes of both and compare them to their both possible tissue of origin (FI and OSE). We select both these tissue as normal since the tumors associated to them are in close proximity as their cell of origin is still debatable. Thus study the of transcriptomic behavior between the tumor and the normal tissues will help in identifying molecular mechanisms associated with the tumor progression from either of the candidate normal tissues of origin in OC.

The second phase of the OC project involves whole exome-sequencing (WES) analysis to predictively identify the mutant lesions underlying OC tumors coming from high-grade (HG) and low-grade tumors (LG) and assess the extent these parental lesions are maintained in their tumor- iPSC clones which are generated from their corresponding tumors through induced reprogramming technique. This gives an assurance that somatic reprogramming is able to maintain the mutant landscape in the tumor-IPSCs intact and their epigenetic features are erased. I was primarily interested to identify the genetic lesions associated with the tumors and the derivative tumor-iPSCS that are maintained in both the parental tumor and their iPSC derivatives. This would be able to confirm that reprogramming tumor to its pluripotent state was compatible in keeping the mutant genome intact.

The second part of the thesis is concerning GBM project where we collected samples from patients having primary and recurrent GBM with collaboration from University of Bonn. We generated the genome wide expression profiles of the patients with unique topological compartments of brain sections for each patient (refer to Figure 28), which upon critical assessments will reveal the core areas that are more prone to attain a relapse through the study of underlying molecular mechanisms. This will be providing us with the opportunity to precisely trace human GBM evolution avoiding the confounding effect of inter-individual genetic heterogeneity and capture the molecular events underlying its transition to its recurrent stage. In light of achieving the above I have developed a computational workflow that will perform both the genetic and transcriptomic analysis in the respective tumor types to revealing the key hubs contributing to tumorigenesis.

Chapter 2: Materials and Methods

2.1 Sample selection and collection

2.1.1 Glioblastoma samples

A key asset of this study is the availability of a uniquely large and informative cohort of matched primary and recurrent GBM samples of 37 (in total) collected at tumor resection site from 7 patients (Prof. Scheffler, Bonn). At primary resection GICs were derived from biopsies of the tumor core and from at least two additional biopsy sites of the resection wall. Marking of the peripheral biopsy sites with MRI-detectable radio opaque clips permitted then to establish, upon recurrence, which of the initial biopsied sites sustained relapse, from which GICs were in turn established. Cells were received frozen and kept in liquid N2 until use, already classified in prior with patient number and origin (OC, PGRT, PDGRT, etc.). These cells were thawed and cultured in polyornithine coated plates with F-Complete medium (Neurobasal containing laminin (5ng/mL), glutamine 1:100, B27 1:100 and N2 1:200) including 5% of EF medium (DMEM-F12 with FBS (10%) and FGF and EGF at final concentration of 20 ng/ mL.). Culture medium was changed once a week. Additional feeding with EF medium (5% v/v) was performed every 48 h. Cells were harvested at 80% confluence and between passages 5-15. At the beginning of the project we dissected the transcriptomes of a uniquely rare cohort of 27 human glioma samples from 5 patients derived from the infrequently conducted surgery at disease recurrence. These 27 samples were primarily subjected to mRNA sequencing on which the initial transcriptomic analysis were performed. Currently two more patients have been added to scale up the cohort and the number of patients have raked up to 7 while total number of samples have gone up to 37. For each patient, glioma initiating cells (GICs) samples were derived from the tumor core and peripheral biopsies at resection of both primary and recurrent GBM samples. These cells were cultured in vitro and then the mRNA was extracted from them. This unique cohort thus provides us with the opportunity to precisely trace human GBM evolution avoiding the confounding effect of inter-individual genetic heterogeneity.

2.1.2 Ovarian Cancer samples

We have analyzed a total of 8 samples that were subjected to whole exome sequencing, of which there were, 2 tumors (one from each grade of high and low grade serous epithelial ovarian cancer; hereafter referred to as HG and LG), 2 normal (peripheral blood samples same patients to serve as genetically matched controls) and 4 iPSC(2 clones of tumor-iPSCs derived from each grade of tumor). These HG and LG tumor were subjected to transcription factor induced reprogramming to generate tumor iPSC derivate (HG-iPSC and LG iPSC respectively) in the laboratory of Stem Cell Epigenetic headed by Prof. Giuseppe Testa at the Department of Experimental Oncology of IEO. Two clones of tumor iPSCs were generated per grade of patients used in the study. Additionally, peripheral blood from the same patient was used as genetically matched normal control. The primary tumor samples one from each grade of tumor are made available from a unique panel of human ovarian cancer samples along with the corresponding archive of medical records, characterized in the Unit of Medical Gynecology at IEO, headed by Dr. Nicoletta Colombo, in collaboration with Dr. Ugo Cavallaro from the Molecular Medicine Program of IEO

2.2 Exome Library Preparation

The sequencing facility of the Campus IFOM-IEO processed gDNA that was extracted from tumor, iPSC and blood. The protocol starts with 10ng of DNA. T4 DNA polymerase, E. coli DNA polymerase I large fragment (Klenow polymerase), and T4 polynucleotide kinase were used to convert the DNA overhands into phosphorylated blunt ends. These enzymes underwent 3' to 5' exonuclease activity that removed 3' overhangs and the polymerase activity filled in the 5' overhangs. Polymerase activity of Klenow fragment (3' to 5' exo minus) recruited the addition of single 'A' nucleotide to the 3' end of the blunt phosphorylated DNA fragments. The NDA fragments were thus able to ligate to the adapters that constitutes a single 'T' base overhang at their 3' end. For proper hybridisation of the DNA fragments to the flow cell, adapter ligation to end was performed. DNA was run on a TAE 2% agarose gel to remove excess adaptors and selects a size range of templates; a gel slice containing the material in the 300±50 bp range was cut from the gel and purified with QIAquick Gel Extraction Kit (Qiagen) according to manufacturer instructions. Processed DNA was then subjected to exon enrichment with the TruSeq exome enrichment kit (Illumina) as per instructions of manufacturer's manual. This was followed by gDNA inucubation with capture probes of exonic regions. The captured regions were purified with streptavidin beads followed by a second enrichment round. Post this adapter-modified DNA fragments were selected in the final steps that were enriched by PCR amplification. The DNA library which was enriched with exon underwent dilution to 16 pM and then used for cluster generation and sequencing with the Illumina HiSeq 2000 machine. The coverage obtained for tumors were 70x while that for iPSCs and normal blood samples were 35x paired end with 100 bp fragment length.

2.3 RNA-Seq library preparation

RNA Samples were processed with the TruSeq Stranded Total RNA Library Prep Kit (Illumina). The amount of RNA to start with was 1 μg per sample, as quantified by Agilent RNA 600 Nano kit (RNA integrity number: 0.9-1). Post step of bead-mediated ribosomal RNA depletion (rRNA removal beads, ribo-zero kit), the RNA fragmentation was carried out using divalent cataions at higher temperature and primed for cDNA synthesis with random hexamers. Reverse transcriptase and random primers were used on the primed and cleaved RNA fragments for reversal into first strand cDNA. The double-stranded (ds) cDNA was generated with DNA polymerase I post removal of RNA template. Beads (AMPure XP beads) are used to separate the ds cDNA from the second strand reaction mix. Overhangs resulting from fragmentation were converted into blunt ends using an End Repair Mix: the 3' to 5' exonuclease activity of this mix removes the 3' overhangs and the polymerase activity fills in the 5' overhangs. A single 'A' nucleotide was added to the 3' ends of the blunt fragments for precautionary measure to prevent them ligating to one another during the adapter ligation reaction. A corresponding single 'T' nucleotide on the 3' end of the adapter provided a complementary overhang for ligating the adapter to the fragment. This is a low rate of chimera (concatenated template) formation strategy. Multiple indexing adapters were ligated to the ends of the ds cDNA, thus preparing them for hybridization process onto a flow cell. PCR amplification was performed to enrich the DNA fragments having adapter molecules at both ends and ensure DNA amplification for library generation. There are fragments with one or missing adapters at the ends, which shows the inefficacy of the ligation reaction process. These will not hybridize to the surface-bound primers in the flow cell and thus cannot form clusters, so they are discarded. PCR was performed with a PCR primer cocktail that has annealing effect at the adapter ends. The sequencing was performed with an Illumina HiSeq 2000, with paired end 50 bp reads to achieve a coverage of 35x for the initial runs made in 2015 but post our publication of RNAonthebench paper (Germain et al., 2016), which effectively showed the power of quantification and DE analysis can still be performed over paired end data with as low as 19x coverage so the new sequencing runs were performed with 20x coverage.

2.4 Computational analysis

2.4.1 Exome data of Ovarian Cancer project

Quality check of the raw reads using Fastqc tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Each lane of sequencing data underwent alignment to the hg19 assembly using BWA (Burrows-Wheeler Aligner) algorithm [5] resulting in sorted sequence alignment/mapping file (SAM) format that was converted to binary format (BAM) using SAMtools. The optical duplicate reads have been marked with Picard MarkDuplicates 1.84 (http://picard.sourceforge.net).

2.4.2 Somatic SNV variant detection

High confidence somatic variant calling were done on the GATK 2.3-4 post filtered and processed BAM files using two callers i) VarScan2 with its default setting for the reads coverage while the p-value threshold set to 0.05 and ii) Mutect with default coverage for normal (8 reads) and for tumor and tumor iPSCs (14 reads). Only high confidence somatic variants were considered identified by both the methods where no evidence in the matched germline sample was included.

2.4.3 Somatic Copy Number Variation (CNV) detection

Somatic CNV calling was made with Control-FREEC with window size 500 and step size 250 while the other parameters were considered as mentioned in the manual for exome data using the normal/tumor bam files and normal/tumor-IPSC bam files. The algorithm is a three-step process as described below in the figure how to assign the somatic copy number regions with proper significance.

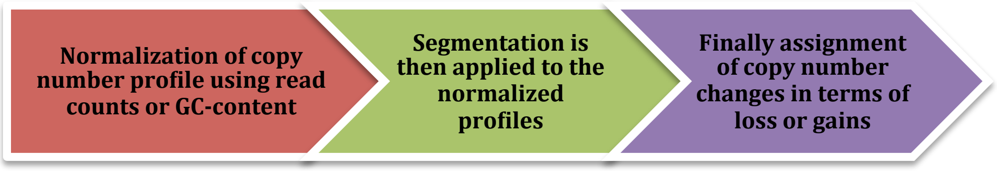

Normalization of copy number profile using read counts or GC-content
Segmentation is then applied to the normalized profiles
Finally assignment of copy number changes in terms of loss or gains

Figure 8: A three-step flow of Control-FREEC tool to infer somatic CNVs from both normal and tumor samples

2.4.4 Driver analysis on the exome data

Driver analysis to distinguish between the passenger and the driver mutation that characterize the tumorigenic drive was done with the help of the tool called IntOGen (Gonzalez-Perez et al., 2013). This tool relies on either BAM file or even somatic variant file (.vcf) to be provided as input and then through a series of rate limiting steps a mutation is assigned with a driver status. The tool is available both as standalone or a web tool. The analysis was performed here with the web-based version as single tumor analysis and post analysis the data was filtered to find mutations that were conferred upon as high-confidence variant by the tool and also by other databases that are embedded with the system having validated mutations as driver thus giving more strength to the analysis. Currently the database hosts around 4,623 exomes from 13 cancer sites and thus help in scoring mutations according to their ranks of driver status for improved clinical decisions. The database comprises of datasets coming from 31 projects that encompasses 13 anatomical sites relying data from repositories like ICGC, TCGA and independent lab data. The mutation pipeline is a resource of several tumor genomes that have results of mutations analyzed with different mutational callers thus having thousands of tumor samples. The pipeline is associated with a workflow management system, which executes all the operations in a proceeding manner. The first state is consequence association of mutation using VEP from ENSEMBLE thus attributing the functional impact scores to non-synonymous mutations using popular methods like SIFT, PolyPhen2 and MutationAssessor(annotation phase). This is followed by transformation of these scores to distinguish the baseline tolerance among genes with the transFIC tool and classifying mutations in-group of impacts, ranging from “None” to “High," in accordance with the consequence type and transFIC Mutation Assessor score. There is also computation of mutational frequency of in sample and across projects and then there are two major algorithms that identify the drivers across the cancer samples by grouping variants of same gene (or pathway). The algorithm OncoDriveFM relies on significant detection of genes that have high functional impact and tend to accumulate together while OncodriveCLUST identifies genes for which mutations cluster in protein sequence region in reference to synonymous mutations, which is the CLUST bias method. Finally the pipeline uses a combinatorial p-value computation from the different p-values that have been calculated for each gene by the different methods. Below is the schematic representation of the pipeline in the figure 9 below.

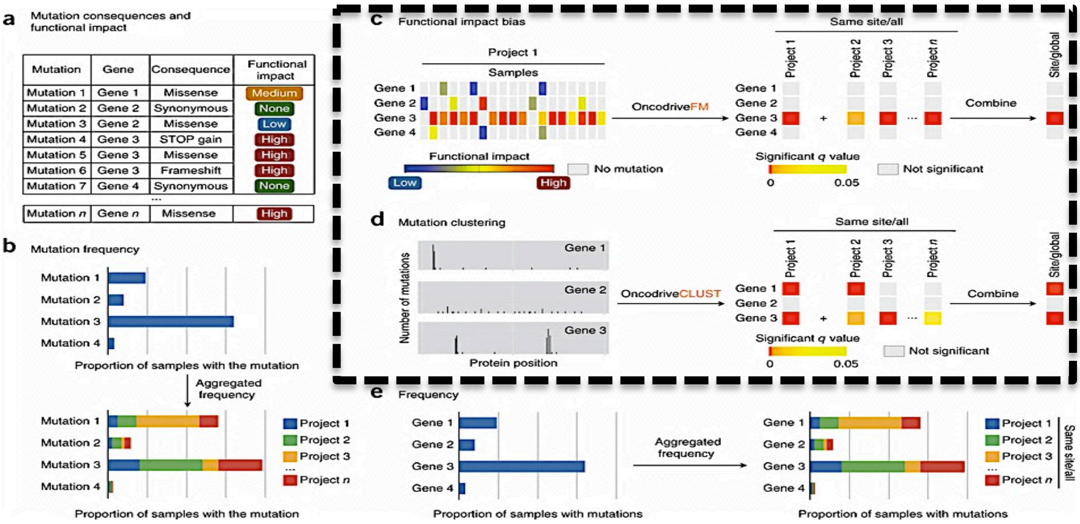


Figure 9: Representative analytical workflow of IntOGen tool conferring driver status from variant datasets

The algorithm works in a series of 5 steps where a) represents the step where the mutational consequences and their functional impact are assessed b) this step is the mutational frequency assessment c,d) this two steps are using the functional impact bias and the mutational clustering algorithms while e) is final step that aggregates the mutations gene-wise in different samples and project and finally identifies the drivers among those genes.

Adapted and modified from (Gonzalez-Perez et al., 2013)

2.4.5 RNAseq analysis

2.4.5.1 Quantification method

Salmon (a tool developed by group of Rob Patro) (Patro, Duggal, & Kingsford, 2015) was used to perform gene-level quantification of RNA-Seq data that has a two step workflow: indexing and quantification. This is a quasi-mapping method, which is highly accurate in terms of errors that might originate in the read or in a variant genome thus providing a pretty robust output. This is pretty fast algorithm that is mapping reads to transcript positions that are computed without performing a base-to-base alignment of the read or fragments to the transcript. The transcriptome index was built on hg19. Salmon was chose for the quantification purpose keeping in mind its high consistencies from our publication RNAontheBench (Germain et al., 2016) which was used to quantify the mRNA abundance at gene level providing TPM estimates across samples in both the projects of the OC and GBM. Few key observations from our benchmarking paper which led to the selection of the Salmon was the yield of relative quantification across heterogeneous samples which is very crucial in projects involving RNA-Seq and their growing importance in deciphering and addressing key events in transcriptomic changes in human diseases. The paper could present the need of relative differences within heterogeneous samples and show that relative quantification of transcriptomes was taking precedence over the absolute quantification across genes in each sample. Another key observation that led to the selection of tool was its speed with which it quantifies each samples in few minutes deviating from the conventional alignment based method. All these key features led to the selection of employing Salmon as the key method for estimating the relative quantification of the transcriptomes across samples in both the project. An important feature while measuring the gene-level differences it was found in our benchmarking paper that count-based methods performed better than other methods thus making it easy to be employable for downstream analysis for tools that are able to accommodate count based gene estimates for assessing the cross samples or cross conditional transcriptomic changes and Salmon is one such tool that performs seemingly well in such gene level counting.

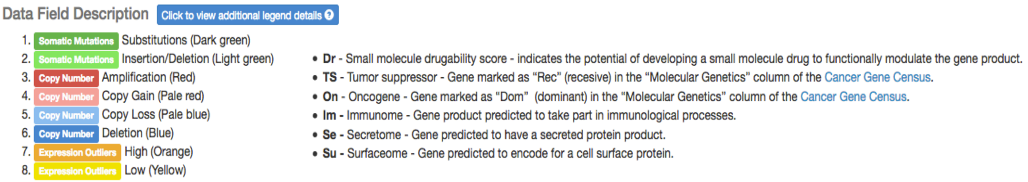
2.4.5.2 Differential analysis and tool selection

The differential analysis was performed was performed with DESeq2 (Love, Huber, & Anders, 2014) with the count-based data that was obtained from Salmon. This tool was selected post assessing the down-sampling methods since the library sizes of the new samples run in 2016 were scaled down to 20x PE reads for both the Ovarian Cancer and the GBM project. All the 3 count-based methods like edgeR/DESeq2/voom was applied on down-sampled data of the ovarian cancer project prior to receiving the new samples of 2016 where limma-voom did not give any differential expressed genes for the standard practices of comparison of normal tissues (FI) and tumor tissues (EOC/ASC). DESeq2 was found to be yielding more differential expressed genes than edgeR so it was selected to be used as standard DE analysis tool for all the for new incoming samples in 2016.

In case of the glioblastoma samples both limma-voom and DESeq2 was employed to characterize the genome-wide expression profile that would not only characterize to uncover the transcriptomic differences between primary and secondary glioblastoma but also in addition would characterize: i) the center of primary tumor; ii) the peripheries of the primary tumor that give rise to the recurrence; iii) the peripheries of the primary tumor that do not give rise to the recurrence; vi) the center of the recurrent tumor; v) the periphery of the recurrent tumor. Limma-voom was found to only yield results of DE genes with that of primary vs secondary GBM samples while DESeq2 was able to find DE genes associated to each and every comparisons that could be used to make for down-stream analysis for DE assessment. Since DESeq2 emerged as front-runner in both the project in terms of univocally representing DE genes computation, thus it was employed in both the project for all the samples for various comparisons that could be exploited for unraveling the various transcriptomic changes associated with different tissues types (Ovarian Cancer) and cellular compartment (GBM project).

2.4.5.3 Extracting Human TFs and Oncogenic TFs in DEGs and association of DEGs with TCGA samples

The AnimalTFDB 1.0 (http://bioinfo.life.hust.edu.cn/AnimalTFDB1.0/) provides with an exclusive list of human TFs, which was obtained and used to extract the genes that are putative Human TF in our list of DEGs. This was sourced in to extract TF DEGs from our various DE analyses, which gave us significantly enriched DEGs between transcriptomes. Association of the genes TF genes to an Oncogene was done with Cancer Gene Census database which was a part of the Oasis-Cancer genomics web portal (Fernandez-Banet et al., 2015) and the various features of point mutation, indels, amplification, deletion, copy-gain, copy-loss, and over/under expression associated to these genes and their frequency in clinical TCGA GBM samples could also be obtained from this oasis-cancer genomics web portal. This web portal currently hosts sample-level annotations and gene-level mutation data, copy number variation (CNV) and expression data from 12,108 primary tumor, 13,007 normal samples and 1,054 cell lines across 55 cancers and 43 tissues from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/), the Cancer Cell Line Encyclopedia (CCLE), the Genotype-Tissue Expression (GTEx) project and 4 published projects on genomics studies of liver, gastric and breast cancers. This is developed by the Pfizer Computational Biology group which is one of the first private pharmaceutical company that aggregated large scale public high throughput genomics data and also their own in house and made it available for public use as a web platform to perform multi-omic multi-seq data mining and inch closer to find drug targets from these analysis which is otherwise very cumbersome owing to the multitude of data produced by several consortiums and trying to source them at one platform. Below is the legend of the PAN-Cancer report that has been used to obtain the Oncogenes in GBM TCGA samples with this platform.


Data Field Description

1. Somatic Mutations Substitutions (Dark green)
2. Somatic Mutations Insertion/Deletion (Light green)
3. Copy Number Amplification (Red)
4. Copy Number Copy Gain (Pale red)
5. Copy Number Copy Loss (Pale blue)
6. Copy Number Deletion (Blue)
7. Expression Outliers High (Orange)
8. Expression Outliers Low (Yellow)

- **Dr** - Small molecule drugability score - indicates the potential of developing a small molecule drug to functionally modulate the gene product.
- **TS** - Tumor suppressor - Gene marked as "Rec" (recesive) in the "Molecular Genetics" column of the Cancer Gene Census.
- **On** - Oncogene - Gene marked as "Dom" (dominant) in the "Molecular Genetics" column of the Cancer Gene Census.
- **Im** - Immunome - Gene product predicted to take part in immunological processes.
- **Se** - Secretome - Gene predicted to have a secreted protein product.
- **Su** - Surfaceome - Gene predicted to encode for a cell surface protein.

Figure 10: Legend of OASIS-cancer genomics listing the annotation of different features that are output of its PAN-Cancer analysis report tool

The legend showing the annotation of the different feature that one can retrieve from the gene list once input in OASIS-cancer genomics and the nomenclature of the terms that are in the output table and the source from which this status of Oncogene and Tumor Suppressor Gene is mirrored.

2.4.5.4 Pathway and GO analysis

Pathway analysis was performed with IPA tool (http://www.ingenuity.com/products/ipa) with the IPA Core Analysis module, which can take gene lists from DE analysis between conditions along with the direction of regulation and try to fetch the canonical pathways that are significantly enriched as a result of changes in gene expression. The association of pathway enrichment is done by finding enrichment of the target genes with the genes that constitute a pathway in their Ingenuity Knowledge Base providing the significance of enrichment to be significantly observed. This p-value calculation is based on right tailed Fischer exact Test which can also be changed to multiple testing correction with B-H multiple testing correction to extract the canonical pathway significance at highest stringency lowering the probability that it is not randomly by chance. The action of activation or inactivation of a pathway in IPA takes into account the effect of directionality of one molecule on another molecule or on a process, and the direction of change of molecules in the dataset. In this way the z-score provided with a sign indicates activation or inactivation of a pathway in a system. IPA takes into account -2 < z-score < 2 to infer significant activation and inactivation and that is what considered here in the results part for the pathways that seems to be activated and inactivated. The ratio metric in the pathway conveys the proportion of target molecules in overall molecules that constitute that pathway. Similarly for a pathway exhibiting up and down regulated genes in forms of red and green color indicates the molecules in the target datasets and the numerical numbers on the right far end for each pathway represents that total molecules in that while the ratio is the proportion of the enrichment.

GO analysis is done by topGO package (Alexa & Rahnenfuhrer, 2010) where the GO enrichment is a wrapper function over the topGO that finds enriched categories in the respective ontology, and related information provided.

The go.enrichment() function has a number of options. Note the following default parameters:

cutoff=0.1: only enrichments with FDR<0.1 are returned. However if there are too many terms the gotreeMapper() will only take the top ones for pictorial representation.

minCatSize=10: all GO terms with less than 10 annotated genes are not tested. This is because these categories are very unlikely to be statistically significant, and discarding them reduces the effect of multiple testing.

maxCatSize=1000: all GO terms with more than 1000 annotated genes are not tested. This is because these categories are typically too broad to be meaningful, and discarding them reduces the effect of multiple testing.

maxResults=200: only the top 200 categories will be returned.

goTreemap() function to plot an enrichment treemap

Pierre-Luc Germain has developed this tool in our lab keeping in mind that we use proper background of genes for the enrichment for specific categories in experimental condition rather than relying on all the genes in a species. The background of genes here relies to the overall genes that are expressed between conditions post quantification and have expression values in at least one sample for the desired comparison of DE analysis thus providing with unique resource of background expressed genes that is representative of that comparison on which the DE analysis is performed.

2.4.5.5 Motif enrichment analysis

Motif enrichment analysis was performed with Pscan tool (Zambelli, Pesole, & Pavesi, 2009) by taking the promoter sequences of the up and down-regulated genes separately and aligning them against hg19 while mirroring the JASPAR 2016 (Mathelier et al., 2016) database with default settings to extract the consensus transcription factor binding sites in these DEGs which results in finding upstream TFs that might target these DEGs and regulate them. Selection of the upstream TFs post assessment was based on the p-value cut-off it provides and assign to each of the upstream TFs and I selected only those that have an error rate of less than 5%. These upstream TFs which are under p-value < 0.05 (obtained by running on promoter sequences of DEGs of a desired comparison) are then overlapped with DEGs in the that comparison to find direct TF targets in the DEGs. Post finding, these targets are tried to be seen in OASIS-cancer genomics portal to find their incidence of dysregulation in OC and GBM TCGA clinical samples and also to retrieve the information of them through the Cancer Gene Census mirror to assign them with status of "Oncogene" or "Tumor Suppressor Gene"

Chapter 3: Results

3.1 Molecular characterization of transcriptional dysregulation in high grade serous OC vis a vis the two candidate tissues of origin

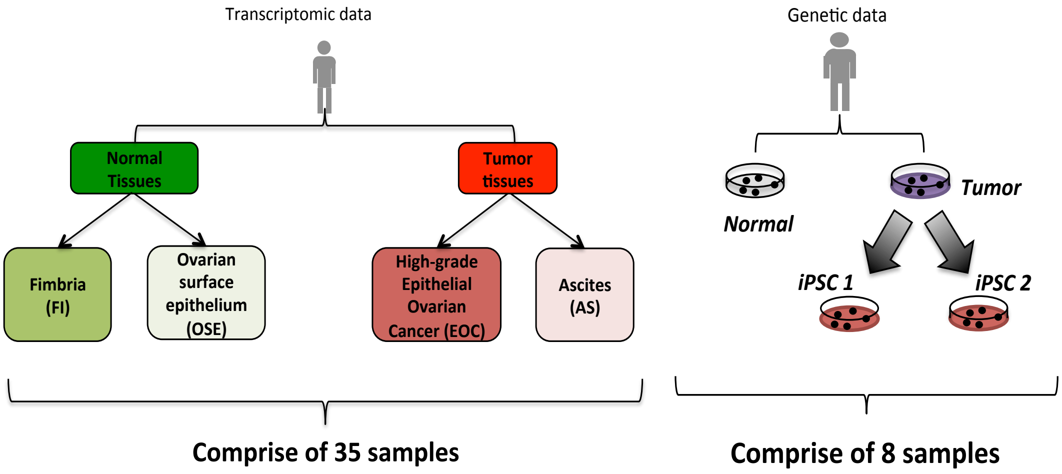


Figure 11: Schematic representation of the datasets used in the ovarian cancer project

The left panel shows the transcriptomic datasets of both normal and tumor tissue samples from patients. The right panel depicts genetic data of the tumor cells that were reprogrammed to IPS. Post reprogramming, I intended to study the genetic background of the tumor and the corresponding tumor-IPSC derivatives. For this reason we performed exome sequenceing on these samples to uncover the mutational background between parental tumor and their reprogrammed derivatives.

Following the elucidation of the transcriptional mechanisms underlying GBM recurrence, I have harnessed the same analytical pipeline to study high-grade serous ovarian cancer (HGSOC). As part of our lab's efforts to understand the developmental origin of HGSOC and reconstruct the transcriptional alterations vis a vis the corresponding tissue of origin, one part of my thesis has been dedicated to study the transcriptional programs associated with tumor development and progression. To this end I focused on the analysis of samples from the two possible OC tissue of origin, namely Fimbria (FI) and Ovarian Surface Epithelium (OSE) along with aggressive high-grade epithelial ovarian cancer (EOC) and fluidic tumor Ascites (AS) representing two most advanced stages of high-grade serous ovarian cancer (HGSOC).

Early years believed that the origin was ovarian in nature. Current theories of OC states that the Fimbria is mostly considered as the origin of HGSOC if not exclusively. This is due to the growing evidences in the recent years based on histopathology and genetic mutations where serous tumors primarily resembled cells that are derived from Müllerian epithelium of the female reproductive tract. Growing evidences from the gene expression studies outlined fallopian tube as the potential origin (Tone et al., 2008) and that primary OC tumors are originating from fallopian tube and the tumors are often localized near the ovaries (Kurman, 2013). These varied theories pose both tissues as potent origin for sub populations of HGSOC. Thus it is crucial to identify the developmental origins to identify key molecular events associated with OC progression from its potential developmental origin (FI or OSE) for better prognosis.

We analyzed the transcriptomes of 4 kinds of tissues (two candidate normal epithelia originating HGSOC; namely OSE and FI) (Ng & Barker, 2015) , (Robert J Kurman & Shih, 2010)) and advanced HGSOC (namely EOC and metastatic AS) from OC patients, amounting to a cohort of 35 samples. This is a relatively large cohort of samples having normal and tumor panels in OC patients. I studied the overall transcriptional landscapes between tumor and the normal samples. Farther, I also studied the classical transcriptomic changes involved between normal FI and the tumors. I also studied the transcriptional landscape between normal OSE and the tumors to elucidate the molecular events involved in OC progression hypothesizing OSE as one of the origin for HGSOC. Finally I also intended to study the transcriptomic differences between the EOC and advanced AS. The origin of ovarian cancer is still debatable (refer to section 1.1.1) and this is one of the reason that I have selected the normal tissues from both candidate tissues of possible origin namely FI and OSE. This helped me in capturing the key molecular events that are involved in OC tumor progression and elicit key transcriptional events that drive tissue-specific oncogenesis. To this end, I performed the following four transcriptomic analyses:

i) Overall transcriptomic changes between both types of candidate tissues of origin and both states of OC (i.e. all samples from FI, OSE vs all samples from EOC, AS), in order to understand the general molecular pathways involved in OC formation and elucidate key TFs that are involved.

ii) The second assessment is aimed at highlighting the transcriptomic between FI (n=8) vs all tumors comprising both EOC (n=10) and AS (n=8). It thus probes the molecular events associated with tumor progression between FI and tumors, aiming at defining the TFs associated with tumor progression, under the currently prevailing assumption that FI is the largely prevalent if not exclusive origin of HGSOC originates form FI.

iii) The third assessment elucidated the transcriptomic behavior that characterizes the major differences between OSE (n=9) tissues against all its tumor counterparts (n=18) and identified key transcriptional programs associated with them in forms of pathways. This analysis was thus meant to test the new hypothesis that OSE may serve instead as epithelium of origin for HGSOC

iv) The fourth and the final assessment involved the key transcriptional differences underlying progression of HGSOC to ascites (EOC, n=10 and AS, n=8) to trace the key transcription upstream of this key transition.

All the above assessments entailed the identification of differentially expressed genes (DEGs), followed by IPA pathway analysis and TF motif enrichment analysis (MEA) . The results of the MEA provides list of candidate master regulators (MRs) that are predicted to be upstream of significant portion of differentially expressed genes and are themselves differentially expressed between the experimental systems under comparison. This helped me in understanding the underlying transcriptional networks associated in OC patients. This analysis allowed systematically assessing the TF-mediated differential expression and defining transcriptional programs involved in OC progression. Finally targeting these regulators might lead to the rescue of the tumor events and reduce the tumorigenicity leading to new therapeutic measures. Figure 12 represents the analytical workflow of the different comparisons that have been pursued in this part of the thesis and strategies employed to capture TF-mediated OC progression.

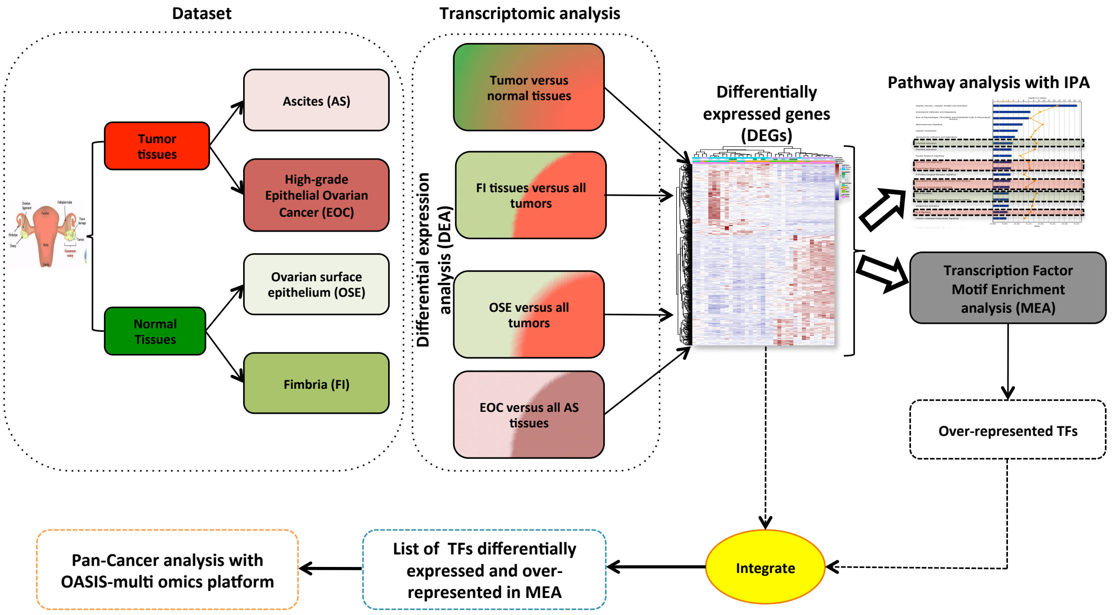

Dataset
Transcriptomic analysis
Ascites (AS)
Tumor tissues
High-grade Epithelial Ovarian Cancer (EOC)
Ovarian surface epithelium (OSE)
Normal Tissues
Fimbria (FI)
Differential expression analysis (DEA)
Tumor versus normal tissues
FI tissues versus all tumors
OSE versus all tumors
EOC versus all AS tissues
Differentially expressed genes (DEGs)
Pathway analysis with IPA
Transcription Factor Motif Enrichment analysis (MEA)
Over-represented TFs
Integrate
List of TFs differentially expressed and over-represented in MEA
Pan-Cancer analysis with OASIS-multi omics platform

Figure 12: Analytical workflow employed in capturing the TF-mediated OC progression

The above figure 12 represents the analytical workflow pursued in the thesis. I interrogated transcriptomic landscapes of OC samples (comprising of high-grade

EOC and advanced AS represented in shades of red) to their candidate tissues of origin (namely FI and OSE represented in shades of green in the above scheme). The first block representing the datasets describes the different tumors and normal tissues that were subjected to RNA-Seq and studied in the thesis. The second block indicates the different transcriptomic comparisons that were performed with differential expression analysis (DEA). The differentially expressed genes (DEGs) obtained from DEA were then subjected to both pathway analyses with IPA tool and motif enrichment analysis (MEA) by pScan (Zambelli et al., 2009). Finally the over-represented TFs were assessed for their differential expression in the comparative systems that were studied. This differentially expressed TFs helped in farther narrowing over-represented TFs obtained from MEA to a list of candidate master regulators (MRs) that captured TF-mediated OC progression

3.1.1 Transcriptomic assessment of normal tissues (all samples from FI, OSE) against the tumor tissues (all samples from EOC, AS)

Differential expression analysis between all the normal tissues versus the tumor tissues revealed 622 differentially expressed obtained with a FDR 0.01 and log2FoldChange of 1.5. ~58% of these genes were down-regulated in tumors.

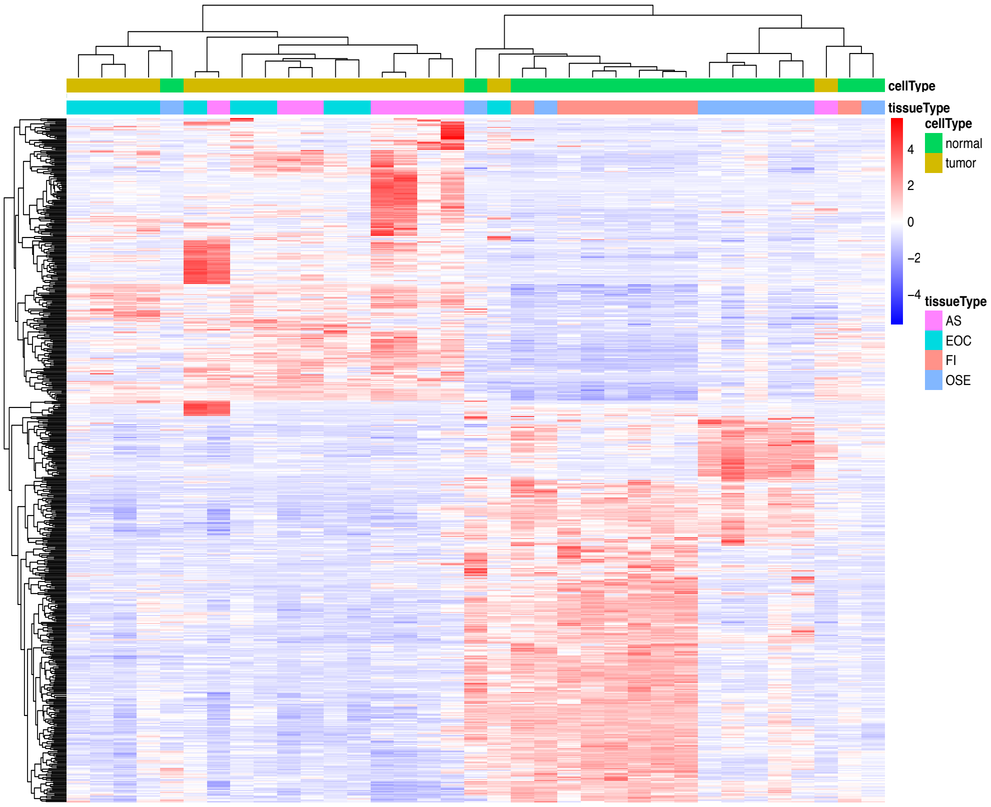


Figure 13: Heatmap of DEGs between tumor and normal tissues in OC patients

The above heatmap represents the differentially expressed genes found between normal and tumor tissues to understand the general molecular pathways involved in OC formation. Up-regulated genes are represented by red color while the blue represents the genes that are down-regulated. The general classification of the samples are normal and tumor (represented as cell type in figure) which are farther divided into FI and OSE representing the normal samples while for tumors they are EOC and AS (represented under tissue Type in the figure).

3.1.1.1 IPA pathway analysis of DEGs between normal and tumor tissues of OC patients

IPA analysis for canonical pathway on the DEGs identified between normal (both candidate tissues of origin) and tumors (solid tumors and ascites) transcriptomes of OC patients revealed a significant enrichment for PPAR signaling inactivation. Figure 14 shows the list of pathways that were enriched among differentially expressed genes.

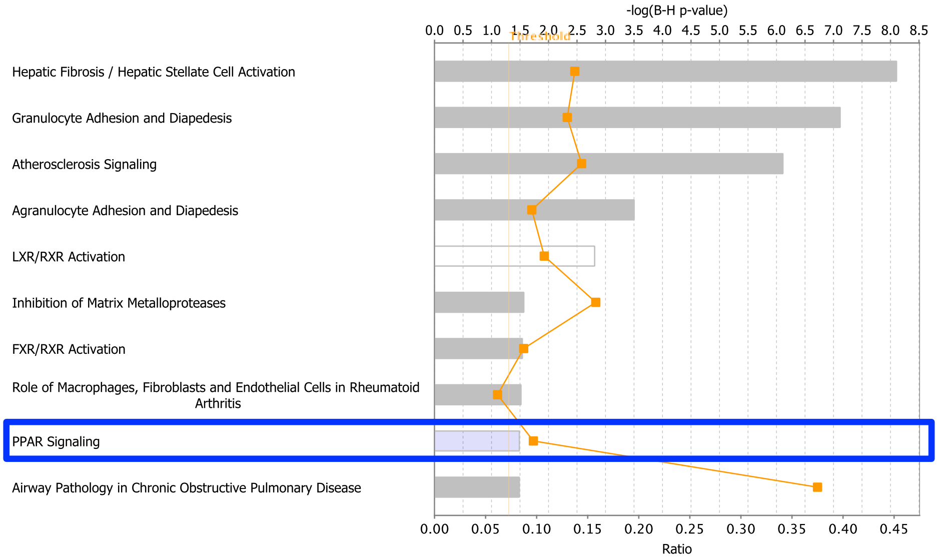


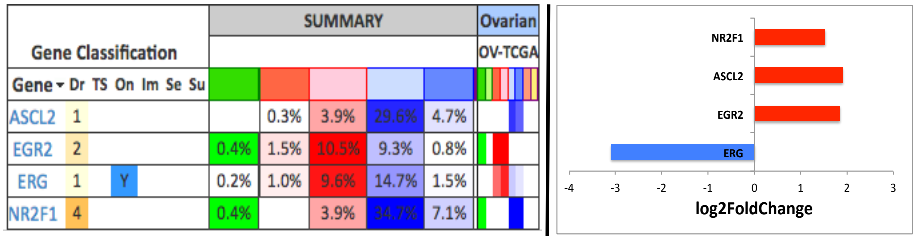
Figure 14: Representations of enriched pathways identified with IPA as a result of DE genes between both candidate tissues of origin and HGSOC tumor samples (both solid tumors and ascites)

The figure 14 represents a set of grey and blue bars for each pathway that are significantly enriched due to DEGs. Blue stands for inactivation of a particular pathway as identified by IPA with its internal scoring system while grey stands for pathways enrich but no assignment of its active/inactive status by the tool. The blue block highlights the pathway that was predicted to be inactive that is PPAR signaling in this figure.

PPAR signaling pathway is inactivated (mild), according to IPA as a result of changes in the gene expression between the normal and tumor patients. PPAR signaling is instrumental in processes like lipid metabolism, cell growth, differentiation, and apoptosis that cater to the physiological development of normal cells. Aberrations of this pathway might thus lead to impairment in the normal cell development. PPARs and their ligands have been earlier linked to cleansing cancer cells via apoptotis whose dysfunction is a clear hallmark of cancer (Elrod & Sun, 2008).Other pathways that have been shown to be enriched due to these DE genes are Hepatic Fibrosis, agranulocyte adhesion, FXR/RXR activation among others however IPA was unable to predict their activation or inactivation status.

3.1.1.2 Motif enrichment analysis predicted oncogenes encoding for TFs controlling the transcriptional programs in OC patients

Motif analysis was performed on the promoters of DEGs to identify the upstream transcription factors that had consensus-binding sites and were thus candidate regulators of these DEGs with pScan. I found 49 upstream TFs on the promoter sequences of these DEGs that co-regulate these differentially expressed genes. These TFs were significant with a p-value < 0.05. 8% (4/49) of these TFs were also DEGs between normal and tumors tissues in OC patients. These 4 DEGs can be thus candidate MRs that can regulate target genes promoting the tumorigenic drive. Among these, 4 genes: ASCL2, EGR2, NR2F1 were up-regulated in tumors; while ERG was down regulated in tumor. The interrogation for these genes in clinical TCGA samples with OASIS-cancer genomics portal confirmed their dysregulation in a large set of TCGA OC samples. In particular ASCL2, ERG, and NR2F1 were associated with copy-loss in TCGA samples with a proportion of ~29%, 14%, 34.7% respectively. These percentages were calculated on a total of 591 TCGA OC samples. EGR2 instead was associated with copy gain in around ~10% of all the TCGA OC samples. Interestingly, the Cancer Gene Census database reports ERG as an oncogene. Figure 15 (below) shows the level of dysregulation of these genes in TCGA (on the left) and in our sample cohort in tumors.



Gene Classification							SUMMARY					Ovarian
												OV-TCGA
Gene	Dr	TS	On	Im	Se	Su						
ASCL2	1							0.3%	3.9%	29.6%	4.7%	
EGR2	2						0.4%	1.5%	10.5%	9.3%	0.8%	
ERG	1		Y				0.2%	1.0%	9.6%	14.7%	1.5%	
NR2F1	4						0.4%		3.9%	34.7%	7.1%	

Figure 15: MEA enrichment reveals over-represented TFs that are differentially expressed between candidate normal and tumor tissues in OC dataset

The heatmap on the left represents the over-represented TFs that are DEGs having high incidence of dysregulation in clinical OC TCGA samples while right bar plot represents the same TF-DEGs fold-change in our in-house OC data sets. Red bar plot represents genes that were up-regulated in tumors while blue represents the gene that was down-regulated in the tumors

The figure legend of the left heatmap is detailed in the methods section.

ASCL2 is a transcriptional regulator, which upon loss of function has been found to be promoting MET pathways in colon cancers via switching on miRNA targets miR-200s. Its higher expression correlates with liver metastasis in colon cancer patients (Tian et al., 2014). Even knockdown studies of this gene have shown to be arresting tumor growth by regulating miR-302 in colon cancer progenitor cells (Zhu et al., 2012). This gene is essentially important in the maintaining the adult intestinal cells. In studies of Gastric cancers (GC) this gene has been shown to be overexpressed and hypomethylated in GC tissues when compared to their normal tissue. The study of the GC revealed that ASCL2 has a crucial role in promoting gastric tumor growth and resistance to chemotherapy. This provides an insight of this gene being an epigenetic target of DNA methylation regulating the gene expression in GC and promoting tumor growth (Kwon et al., 2013). In OC this gene has not been explored much but as our results suggests that this gene is usually associated with copy loss in OC TCGA patients while being up-regulated in our tumor cohorts. This observation of ASCL2 in my study clearly indicates the involvement of a possible epigenetic mechanism orchestrated by DNA methylation.

I also found that the ERG oncogene, frequently mutated in ovarian, skin and lung cancers, was down-regulated in our tumors. This gene belongs to the ETS transcription factor family, which has been seen to be over expressed in tumors like sarcomas, AML and prostate cancers involved in malignant transformation (Sashida, Bazzoli, Menendez, Liu, & Nimer, 2010). It was also identified as a fusion gene with TMPRSS2 with high levels of ERG in ovarian cancers. Recent studies have shown that ERG itself is not a diagnostic target in ovarian cancer as seen in prostrate cancers unless identified as a fusion gene (Huang et al., 2011).

NR2F1 is a transcriptional regulator, for which there is no prior evidence of a role in ovarian cancer. It has only been seen to be possibly regulating metabolism and dedifferentiation process in ovarian cancer cell lines in a report published by Kieback et. al, 1993. It is termed as a master regulator that is associated with cellular dormancy. This gene upon shutting down leads to tumor growth and proliferation of the tumor cells in abnormal way allowing dormant cells to grow throughout the tumors (Sosa et al., 2015). NR2F1 is associated with a frequency of ~34.7% copy loss in OC TCGA patients while in our cohort of OC patients transcriptome it is up-regulated in tumor. This gene identified as a candidate MR post MEA in our OC samples. This gene is having 28 downstream targets with sequence specificity of binding over 90%. Over 60% of its targets are down-regulated in our tumor cohort at the level of gene expression. NR2F1 is usually down-regulated in proliferative tumors since it induces quiescence and promotes tumor growth. In our OC tumors it can be hypothesized that NR2F1 is promoting tumor growth by triggering pathways that are dependent on its downstream target genes that are mostly down-regulated however itself not being repressed. This up-regulation of NR2F1 gene is probably due to epigenetic targeting either in forms of DNA hyopmethylation at its promoter region in tumors leading to is over-expression or recruitment of histone modifiers that activates it. So it would be important to check the status of differential methylation between normal and tumor samples in our OC cohort for this gene or perform ChIP-Seq for specific histone marks to identify the behavior of this gene under a histone modifier. Another important issue to address in this comparison is that this comparison entails a more generic study of transcriptomes between normal and tumor samples in our OC cohort. So the comparative system will have a bias where the tissue heterogeneity (comprising of FI, OSE, EOC and AS) is not fully utilized in the model while estimating differential expression since we are not trying to see the differences based on specific tissues that are sub-categories of normal and the tumors in our OC dataset. This issue is more apparently taken care of in subsequent studies made in the thesis when we compare tumors to candidate tissues of origin separately.

3.1.2 Transcriptomic analysis between Fimbria normal against tumor tissues in the OC patients to capture the events underlying a tumor progression

Differential expression analysis between one of the possible cell origin of OC (FI) and all the tumor samples (EOC, AS) generated in our lab revealed 1288 DEGs with FDR < 0.01 and log2FoldChange 1.5. Among these DEGs more than 61% of them were up-regulated in the tumor tissues. Figure 16 below shows the overall DEGs on the left panel while focusing on a deeper resolution reveals some TFs that are also differentially expressed.

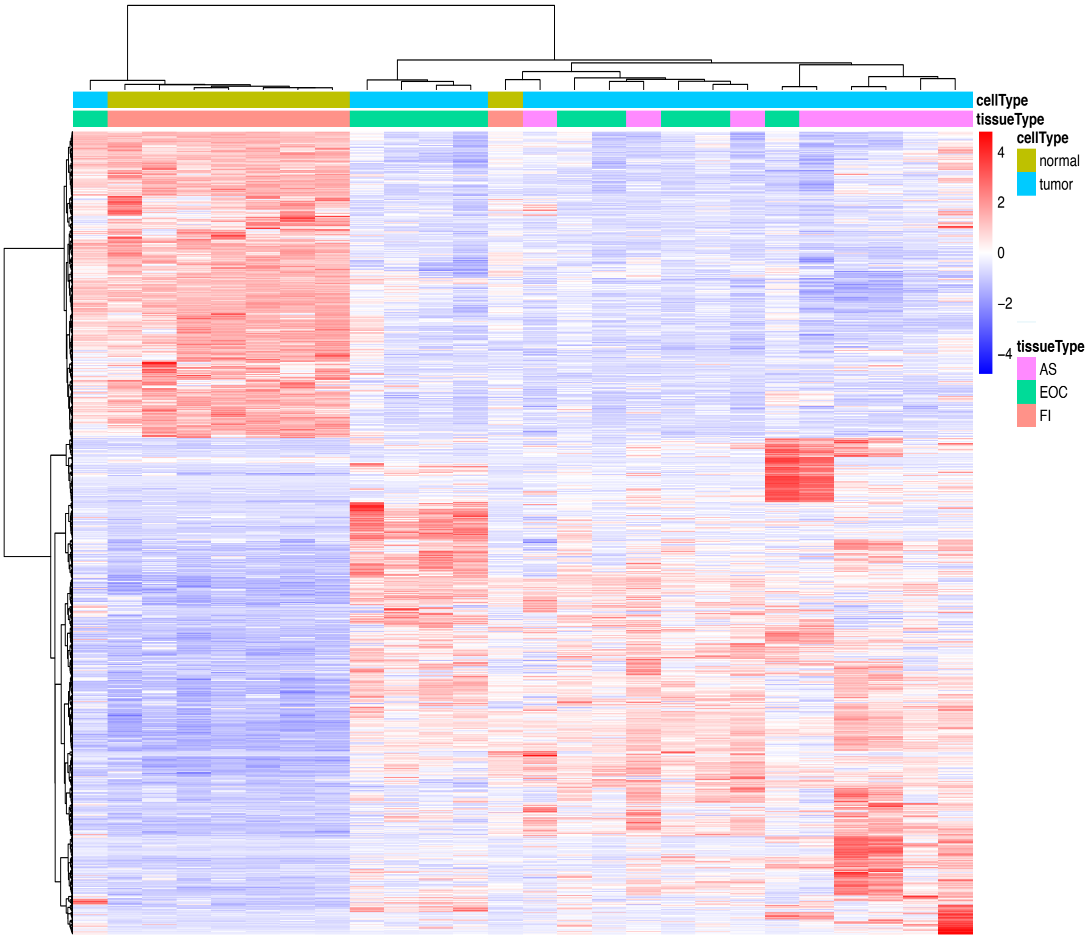


Figure 16: Heatmap representation of DEGs between FI normal and all Tumors (EOC,AS) tissues of OC patients

The color scale of the heatmap represents up-regulated genes in red and down-regulated genes in blue. The cell type classifies the samples as normal and tumor while the tissue type farther sub-classifies the samples into normal FI, high-grade EOC and advanced AS.

3.1.2.1 Canonical pathway analysis of DEGs between FI normal and tumor tissues of OC patients revealed activation of metabolic and inflammatory pathways.

IPA canonical pathway analysis revealed significant pathways that were changing between FI normal versus tumor tissues. Figure 16 shows the list of pathways that were involved as a result of differential gene expression. The canonical pathway analysis predicted the activation of: Acute Phase Response Signaling, VDR/RXR activation, LXR/RXR activation, Complement System, Eicosanoid Signaling among the top pathways. The pathways that got inactivated were: PPAR signaling, Nitric Oxide signaling in cardio-vascular system. Pathways like VDR/RXR and LXR/RXR activation have been earlier linked to tumor malignancies and inflammatory responses (C. Y. Lin & Gustafsson, 2015). This suggested that the tumorigenesis was due to activation of metabolic signals, immune responses and inflammation, which could have provided the tumoral cells with growth advantages and proliferation.

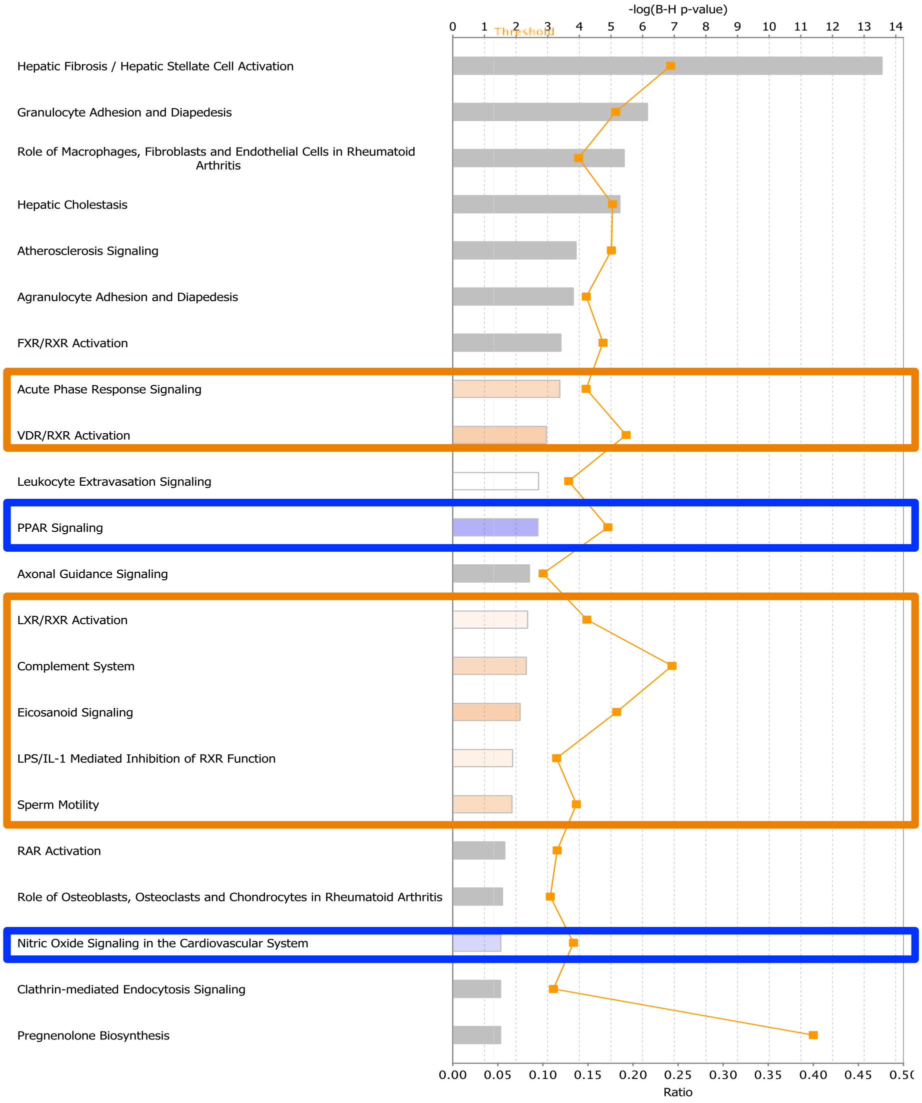
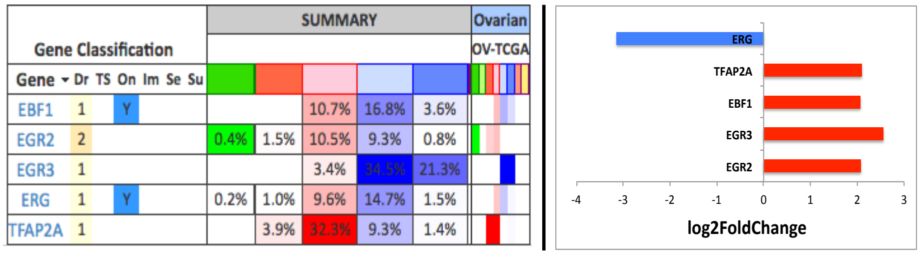


Figure 17: Representations of enriched pathways identified with IPA as a result of DE genes between FI normal and all tumor samples in the OC dataset

Pathways that are predicted to be activated by IPA are represented with orange bars and also summarized under orange blocks. The blue bars represent the inactivated pathways and are summarized under blue blocks.

3.1.2.2 Motif enrichment analysis predicted a core set of oncogenes encoding for TFs controlling the transcriptional programs in FI normals and tumor samples in OC patients

I interrogated the DE genes between normal fimbria (FI) and tumors by performing the TF motif enrichment analysis (MEA) as done with the earlier comparisons. This analysis primarily aimed at providing a mechanistic interpretation of the transcriptomic changes in tumors by identifying putative TFs master regulators (MRs) through interrogation of binding sties at the promoter of the DEGs. Motif analysis was performed on promoter sequences of these overall 1288 DEGs obtained as a result of DE analysis between FI normals and tumors with Pscan with default settings. This revealed 57 (p-value threshold < 0.05) upstream TFs that have consensus binding sites in our DEGs. Among these upstream over-represented TFs, 5 were also differentially expressed between FI normals and tumor samples of OC patients whose transcriptomes we assessed. Upon investigation of these 5 genes in clinical OC TCGA samples I identified their level of dysregulation. I found genes like TFAP2A, EBF1, EGR2 and EGR3 as up-regulated in tumor while ERG was down-regulated. Figure 18 below shows the incidence of 5 of these targets in OC TCGA clinical samples and confers the status of Oncogene/TSG to them.



Gene Classification							SUMMARY					Ovarian
												OV-TCGA
Gene	Dr	TS	On	Im	Se	Su	🟩	🟥				
EBF1	1		Y						10.7%	16.8%	3.6%	
EGR2	2						0.4%	1.5%	10.5%	9.3%	0.8%	
EGR3	1								3.4%	34.5%	21.3%	
ERG	1		Y				0.2%	1.0%	9.6%	14.7%	1.5%	
TFAP2A	1							3.9%	32.3%	9.3%	1.4%	

Figure 18: MEA enrichment reveals over-represented TFs differentially expressed between FI normal and tumor tissues in OC dataset

The left heatmap represents the level of dysregulation of the genes in OC TCGA patients. The legend can be found in methods section. The right bar plot represents the status of up and down-regulation of the same genes in our tumor transcriptomes. Red stands for up-regulated in tumor while blue represents down-regulated gene.

Upon interrogation of these 5 DEGs in OASIS-cancer genomics portal I was able to identify two oncogenes in this list of MRs that were DEGs. These were EBF1 and ERG. EBF1 was up regulated in tumors while ERG was down-regulated. The other 3 genes were EGR2 (associated with ~10% copy gain in Ovarian TCGA samples), EGR3 (associated with 34.5% copy loss in TCGA OC samples) and TFAP2A (primarily associated with 32.3% copy gain in TCGA OC samples).

Mutations or INDELS along with in-frame deletions in EBF1 have been observed in cancers such as intestinal cancer, skin cancer, and stomach cancer. These observations were pretty insightful in revealing the contribution of TFs in promoting oncogenesis between tumor and normal FI tissues seating oncogenes among them.

ERG is often associated with fusions or mutations such as silent or missense. This is observed in lung cancer, ovarian cancer, and skin cancer and its role as oncogene in prostate cancer as a fusion gene is widely acknowledged while in ovarian cancer it is not a prognostic marker (Huang et al., 2011).

3.1.3 Transcriptomic changes between OSE and tumors identify cAMP pathway signaling inactivation

Differential gene expression analysis of OSE (normal tissues) against the tumor revealed very few 248 DEGs that were changing expression. Among them over ~70% genes were down-regulated in tumor which shows the oncogenesis was mediated as a result of greater proportion of down-regulated genes in the tumor tissues.

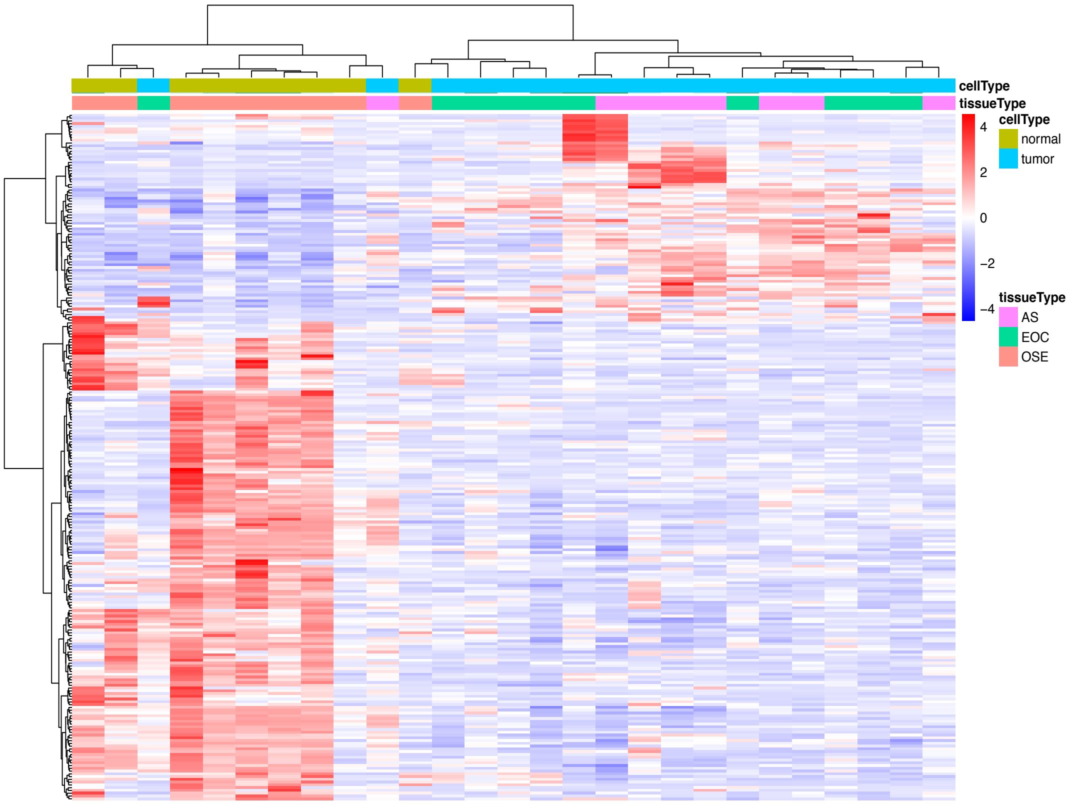


Figure 19: Heatmap representation of DEGs between OSE normal and all Tumors (EOC, AS) tissues of OC patients

The up-regulated genes are represented by color red in the heatmap while down-regulated genes are represented by color blue. The cell type refers to the normal (OSE) and tumors (EOC and AS) while tissue type farther details the normal and tumors interrogated. Here we have only OSE so there is only one tissue type represented while for tumors we have both high-grade EOC and advanced AS.

3.1.3.1 Canonical pathway analysis with IPA revealed cAMP signaling inactivated between OSE normal and tumor tissues on OC patients

IPA pathway analysis revealed a significant enrichment for cAMP signaling, based on the directionality (fold change) of the DEGS and was predicted to be inactive. This pathway is crucial in maintaining processes like immune function, growth, differentiation, gene expression and metabolism in normal physiological environment. The role of cAMP processes and their downstream targets activation is indeed a complex problem in cell biology and its role in cancer is still a matter of debate. In particular, its unclear if it is having stimulating or inhibiting effects in cancer cells. However, it has been reported to be having both positive and negative effects on cell growth or survival at specific cellular or tissue context. Thus, their aberrations may play important role in oncogenesis (Fajardo, Piazza, & Tinsley, 2014).

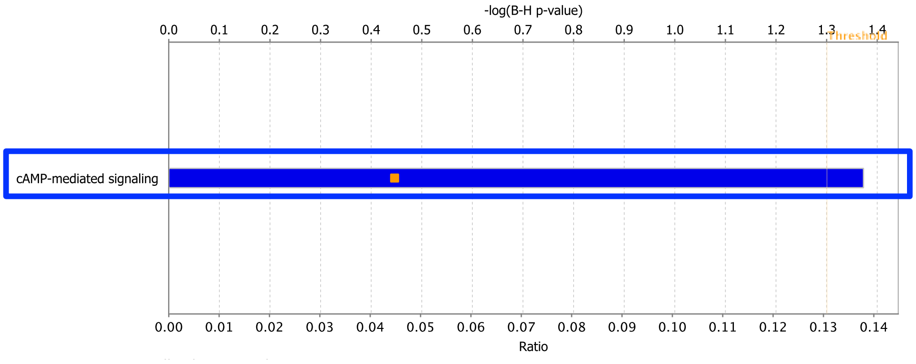


Figure 20: Representation of enriched pathways identified with IPA as a result of DE genes between OSE normal and tumor patients

IPA canonical pathway analysis only revealed cAMP-mediated signaling pathway to be significantly enriched. Blue bar refers to the status of inactivation of the pathway and the blue block highlights the same pathway that is inactivated.

The results from MEA revealed a set of 8 TFs which are predicted to regulate the DEGs, however none of them were differentially expressed in the comparison between OSE and tumor tissues.

3.1.4 Genome-wide DNA methylation analysis predicts HGSOC cell of origin

As a complementary approach to the analysis I have outlined above, we have also harnessed the power of genome-wide DNA methylation profiling of normal and tumor cells to classify a well characterized cohort of HGSOC cases in our OC cohort on the basis of their tissues of origin. This enabled us to build on convergent evidence how tumors preserve DNA methylation signatures (DMS) from their cells of origin (Sproul et al., 2012), (Moran et al., 2016). This strategy allowed us to associate tumors to their normal counterpart, thus increasing precision and stringency in the identification of relevant altered pathways in HGSOC for patients which had been profiled for both DNA methylation and RNA-Seq. DNA methylation analysis identified a set of 92 CpG sites with a delta-beta variation of at least 40%, whose power in distinguishing FI-like from OSE-like tumor samples was validated on the two largest Fimbria and Ovarian cancer biopsies datasets (Klinkebiel, Zhang, Akers, Odunsi, & Karpf, 2016),(Patch et al., 2015). This analysis confirmed the predictive power of this core signature we had defined on cultured cells in distinguishing FI and OSE samples directly sourced from biopsies. Therefore, this analysis also confirmed that our in vitro culture conditions for FI and OSE-derived epithelial cells recapitulate their salient *in vivo* features. Next, we used these 92 CpG's to classify 147 HGSOC samples (both cultured cells and biopsies), coming from our IEO cohort (n = 24) as well as from two independent cohorts (Karpf n = 10, Bowtell n = 113) (Klinkebiel et al., 2016)(Patch et al., 2015), into FI-like and OSE-like HGSOC. Next, having used DNA methylation as a tool to identify the cell of orgin for the tumors, we harnessed this classification to perform RNAseq-based differential expression analysis, comparing FI (n=8) vs. FI-like tumors (n=5) and OSE (n=9) vs. OSE-like tumors (n=10) to identify differentially expressed genes specifically associated to the oncogenic transformation of the original normal tissue followed by MEA analysis to find out over-represented TFs that are also differentially expressed in the same transcriptomic comparisons of FI vs FI-like and OSE vs OSE-like. Figure 21 shows the analytical workflow that we followed, starting from the use of DNA methylation signatures(DMS) for assigning tumors to their tissue of origin, followed by DEA and MEA analysis of tumors to their respective normal source of origin to identify candidate master regulators.

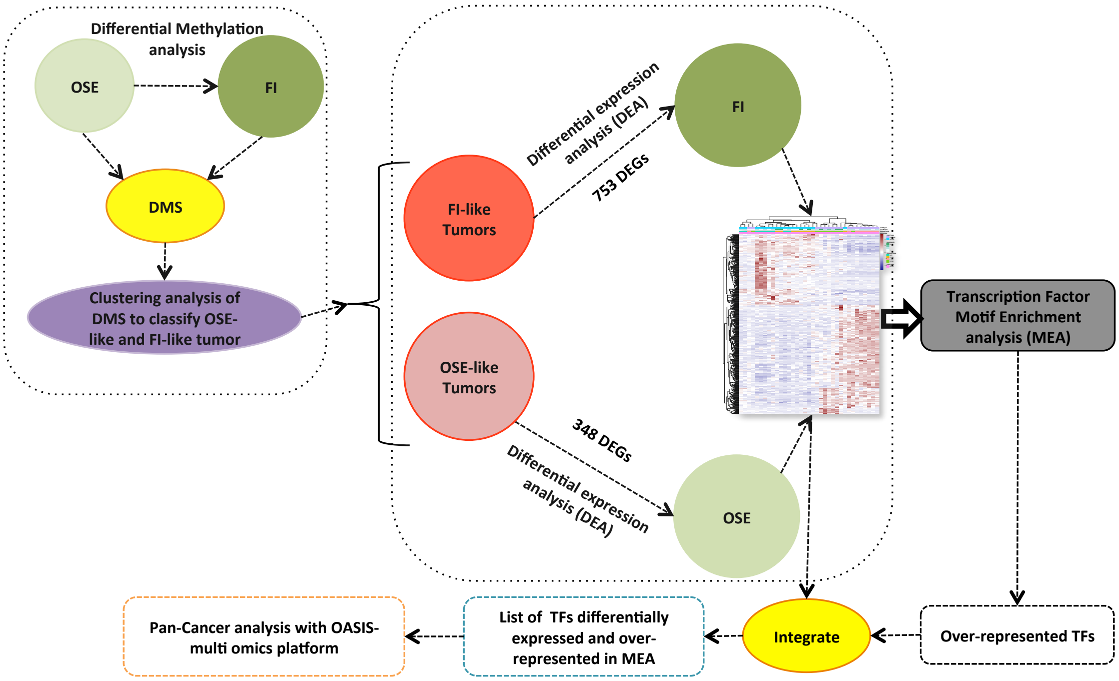

Differential Methylation analysis
OSE
FI
DMS
Clustering analysis of DMS to classify OSE-like and FI-like tumor
FI-like Tumors
OSE-like Tumors
Differential expression analysis (DEA)
753 DEGs
FI
348 DEGs
Differential expression analysis (DEA)
OSE
Transcription Factor Motif Enrichment analysis (MEA)
Over-represented TFs
Integrate
List of TFs differentially expressed and over-represented in MEA
Pan-Cancer analysis with OASIS-multi omics platform

Figure 21: Analytical workflow describing the classification of tumor to its origin by DMS followed by transcriptional and MEA analysis of tumors to its corresponding tissue of origin

The above figure schematizes the stratification of OC samples (high grade EOC and advanced AS) to their respective tissue of origin (FI and OSE). The differentially methylated signature that brings out the differences between the two normal tissues (FI and OSE) was used to stratify our tumors (EOC and AS) to their corresponding cell of origin using unsupervised clustering approach. Once we obtained the stratification of the tumor they were termed as FI-like and OSE-like. This was followed by an analysis of the transcriptomic changes between FI normal versus FI-like tumors and OSE normal versus OSE-like tumors. Post DEA I followed the pathway analysis and MEA to obtain candidate master regulators to capture TF-mediated OC progression for tumors coming from its specific cell of origin.

DEA analysis of FI vs FI-like detected 753 DEGs with FDR 0.01 and log2FC 1.5 while for OSE vs OSE-like tumors it revealed 348 DEGs using the same threshold for significant DEGs selection.

Glutamate receptors earlier have been shown as potential growth factor that fuels the migratory and propagating behavior of tumor cells (Stepulak, Rola, Polberg, & Ikonomidou, 2014). This is now seen also in our datasets that DEGs in FI vs FI-like tumors enrich this pathway and activates it. Even these migratory movements are farther supported by calcium transport and ILK signaling pathways. There are evidences of such activation in advanced ovarian cancers (Bruney, Liu, Grisoli, Ravosa, & Stack, 2016). ILK-activation dependent tumor formation and propagation promotes oncogenesis via invasive and migratory properties in cells. This mechanism have already been established in transgenic mice (Bruney et al., 2016). These key findings provide new insight on the pathogenic pathways underlying FI-like tumors. MEA identifies 3 over-represented TFs that are predicted to control the DEGs in FI vs FI-like tumors, and that are also differentially expressed in the same comparisons. These genes are: EHF, TFAP2A and ZIC1, which are all up-regulated in our FI-like tumors. Among these 3 genes TFAP2A and ZIC1 are associated with copy gain of ~32.3% ~45.9% respectively in OC clinical TCGA samples as reported by OASIS genomics web portal.

In case of OSE vs OSE-like tumors, the DEGs enriched for pathways mainly concerning diapedesis and adhesion. Interestingly earlier when I analyzed the

OSE normal against all tumors without stratifying the tumors to their origin I found cAMP signaling to be inactivated. This however is no more significantly enriched in the DEGs between OSE and OSE-like tumors. I also could not find any overlap between differentially expressed genes and the over represented TFs from MEA analysis while comparing transcriptomes of OSE normal against all tumors. However, comparing transcriptomes of OSE normal versus OSE-like tumors, I found EGR1 as an over-represented TF post MEA that is also a DEG in OSE vs OSE-like tumors. This vindicates the value of using an epigenetic tracer, in this case DNA methylation, to precisely assign the cell of origin for the tumors in our study. Thus studying the transcriptional commitment of these newly classified tumors (FI-like and OSE-like tumors) to their normal tissues gave us better understanding of the underlying transcriptional programs associated in OC. EGR1 gene is usually found to be associated with ~9.3 % copy gain in clinical OC TCGA samples and was up-regulated in our OSE-like tumors. EGR1 have been linked to several cancers earlier. In gastric and colorectal cancers its activity has been linked as a tumor suppressor gene which when mutated promotes tumor development (Choi, Yoo, Kim, An, & Lee, 2016). In non-small cell lung cancer also it has been associated with tumor suppressor properties (H. Zhang et al., 2014). In others like prostrate caner its been shown to be over-expressed in tumors (Parra Villegas, Ferreira, & Ortega, 2011), (Gregg & Fraizer, 2011). It is often associated with multi-functional transcriptional activity that can promote or decrease tumor activity. This gene has not yet been explored much in the OC field, while our analysis of OSE-like tumors against OSE normal tissues pointing to it as candidate MR provides now a sound basis for its more systematic investigation in this tumor.

3.1.5 Transcriptomic analysis involving two biological tumors spread (solid EOC and fluidic AS) in the OC patients identified inflammatory and immune response signaling switch between them

It is familiar that patients harboring solid HGSOC are often diagnosed with AS. This is seen in one-third of the OC patients (Ayantunde & Parsons, 2007; Kipps, Tan, & Kaye, 2013). As a matter of fact these ascites confer to phenomena of chemo-resistance and relapse. These AS often provide rich source of tumor micro-environment stimulating the tumor cell growth along with chemo-resistance (N. Ahmed & Stenvers, 2013) . This was particularly interesting to capture the inherent differences between the solid EOC and fluidic AS at the level of transcriptome in our patients as well as a surrogate to identify features that could lead to predict relapse. There have been lines of evidences that states HGSOC being heterogeneous also have specific mechanisms of how these ascites spread and settle (Auer et al., 2015). So my goal was to identify the DE genes between these two tumoral sources and find molecular events specific to either. This could help in assessing the tumor spread and also give important information regarding the pathways associated with inflammation, tumor-micro environment that are characteristics of AS.

DEA identified 135 genes that were significantly changing between EOC and highly aggressive AS that are OC infiltrating cells circulating near the peritoneum. Both tumor sources typically share certain features that differ their expression profiles. The classification power of these 135 DEGs is shown in Figure 22 by the unsupervised clustering analysis, where the tumor cluster according to their transition states. More than ~70% of the genes were up-regulated in the AS.

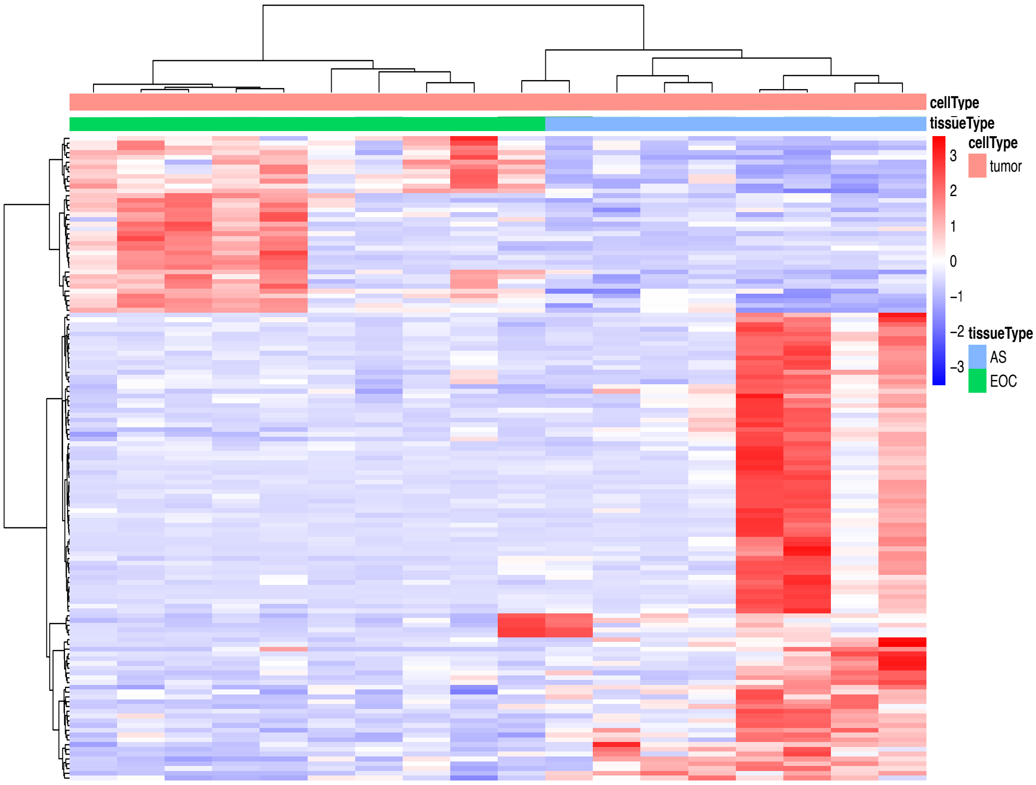


Figure 22: Heatmap representation of DEGs between 2 tumoral sources, EOC and AS present in our OC datasets

The above heatmap shows the DEGs between EOC and AS in our OC tumor cohort. The up-regulated genes are represented in red while the down-regulated are represented in blue. The cell type represents the class of the samples which is tumor while tissue type refers to both the transition stages that sub categorize the tumors in EOC and AS

3.1.5.1 Canonical pathway analysis of DEGs between EOC and AS reveal inflammatory and immune response signals

IPA canonical pathway analysis of these 135 DEGs revealed a number of signaling pathways significantly enriched. The pathways found enriched by IPA analysis revealed that most of them were inactivated. Among the top significant pathways predicted to be inactivated by IPA were: CD28 Signaling in T Helper Cells, Role of NFAT in Regulation of the Immune Response, Calcium-induced T

Lymphocyte Apoptosis, iCOS-iCOSL Signaling in T Helper Cells, LXR/RXR Activation, Dendritic Cell Maturation, IL-8 Signaling from among others. Most of them were related to inflammation and immune responses but their inactivation status was due to the fact that they were down-regulated in the EOC or conversely these pathways were enriched in AS where they were up-regulated. So we can say the pathways are silenced from AS to EOC.

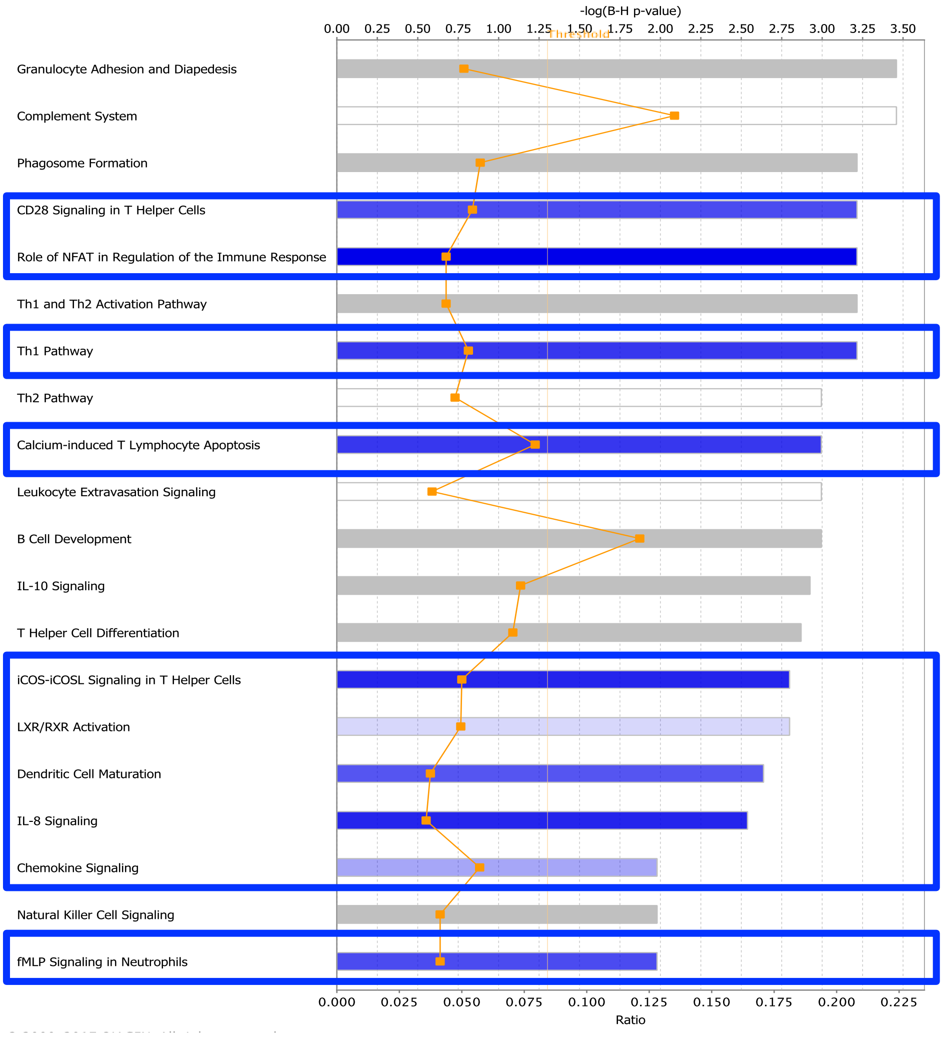


Figure 23: Pathway identified with IPA as a result of DE genes between EOC and AS transcriptomes in OC datasets

IPA pathway analyses reveal mostly inflammatory and immune response signaling pathways to be inactivated between EOC and AS. The blue bars represent the pathways that are inactivated and the blue blocks highlight those inactive pathways.

MEA revealed 89 over-represented TFs (p-value <0.05), however none of them were differentially expressed. So candidate MR that were also DEGs could not be could not be obtained. This is possibly due to the fact that the number of DEGs obtained were also very less with the thresholds that have been considered for analysis, or for the different turnover of the TF proteins that could be uncorrelated with their gene expression. Since our analysis is based on gene expression I focused only on those TFs whose differential gene expression and differential protein activity were positively correlated.

This analysis could trace the pathways that in particular list out the signaling pathways that are triggered due to differences in EOC and AS which are mostly inflammatory or immune signaling pathways. These gives us an idea of the typical pathways associated with AS.

3.2 Genetic assessment of induced pluripotent stem cell clones in context of Ovarian Cancer

Cancer is now widely held to result from a tight interplay of genetic and epigenetic aberrations, pointing to the need to dissect the genetic vs the epigenetic contribution in cancer pathogenesis. In addition, there is limited availability of suitable models that can recapitulate its phenotype, a dearth that is particularly relevant in OC. Even though cell lines, tumor xenograft models may provide a better fit but their scope is limited when it concerns the capturing of intra-patient heterogeneity in the primary tumor samples. Most of these issues are proposed to be overcome with the multi-step reprogramming process.

This multi-step reprograming process resets the epigenetic landscape and allows the setting of compatible transcriptional landscape. This indefinite expansion and differentiation of tumor-iPSCs would thus be able to capture the genetic variations and associate them to early developmental tumor. This makes it a perfect fit for not only capturing the genetic landscape that builds up the mutant genome of the parental tumor but also assess the epigenetic informations that were encoded in it which gets cleansed in its reprogrammed derivatives. In light of this my primary task was to capture the genetic lesions in the parental tumor and to map their match in their tumor iPSCs. In simple terms it tracks the extent of parental genetic lesions of primary tumor that are preserved in reprogrammed tumor-iPSCs are compatible to reprogramming. WES can be one of the ways to address the extent of maintenance of parental lesions between tumor and their reprogrammed derivatives and confirm if iPSC was indeed tumor derived. To this end, I performed 3-tier approach where the first approach was to extract the somatic mutations that could be potentially representing the key genetic lesions in tumor (1 high grade serous ovarian carcinoma: HG and 1 low grade OC tumor: LG) and preserved in their derivative tumor-iPSCs (2 clones from each of the HG and LG tumor hereby named as HIPS1 and HISP2 for high grade and LIPS1 and LIPS2 for LG tumor derivatives respectively). The second approach was to identify from the above-defined somatic mutations, key drivers that are mostly found to be recurrent in our tumors and other cancer types. Also to track the extent of tumor drivers that were preserved in tumors iPSC clones indicating that these reprogramed derivatives are driven by parental driver mutagenic variations. The final approach was to find the somatic CNAs in our data and to understand to what extent these somatic SNAs were found both in tumor and retained in the tumor-iPSCs.

3.2.1 Somatic SNV analysis revealing iPSCs were tumor derived

I used two complementary strategies to capture the somatic tumor and tumor-iPSCs SNVs with both VarScan2 (Koboldt et al., 2012) and Mutect2 (Cibulskis et al. 2013) to derive the high confidence mutations identified by both platforms. There were high inconsistencies between the number of mutations found by the two algorithms which primarily score the importance of both algorithms and the mutations that were commonly obtained by both were designated as the most confident ones that could verify the extent of genetic background lesions preserved between both tumor and their derivative iPSCs. This suggested clearly that at least a fraction of mutations found in tumor were retained in the iPSC suggesting the reprogramming of a tumoral subclone rather than of a normal tumor-associated cell. VarScan2 is an open source software for variant detection having compatibility with several short read aligners with immense ability to identify SNPs and indels of high-sensitivity and specificity, in both individual and pooled samples. Mutect is a method developed for accurate identification of somatic point mutations in next generation sequencing data of cancer genomes. It uses Bayesian classifiers to find mutations with very low allelic frequencies and few supporting reads, using fine tuned filters to remove artifacts ensuring high specificity. Using this pipeline I analyzed the detection of somatic mutations for both control (here matched peripheral blood sample of the patients) vs. tumor and control vs. tumor-iPSC comparisons and obtained rare somatic events (not present in dbSNP database which is a free public archive that catalogues genetic variation within and across different species developed and hosted by the NCBI in collaboration with the NGHRI). Upon comparing the somatic SNVs between tumor and tumor-IPSCs, I found of 25% somatic mutations (with VarScan2) shared between the iPSC and the parental tumor while fraction was 9% shared somatic mutations (with Mutect) between iPSCs and their parental tumor.

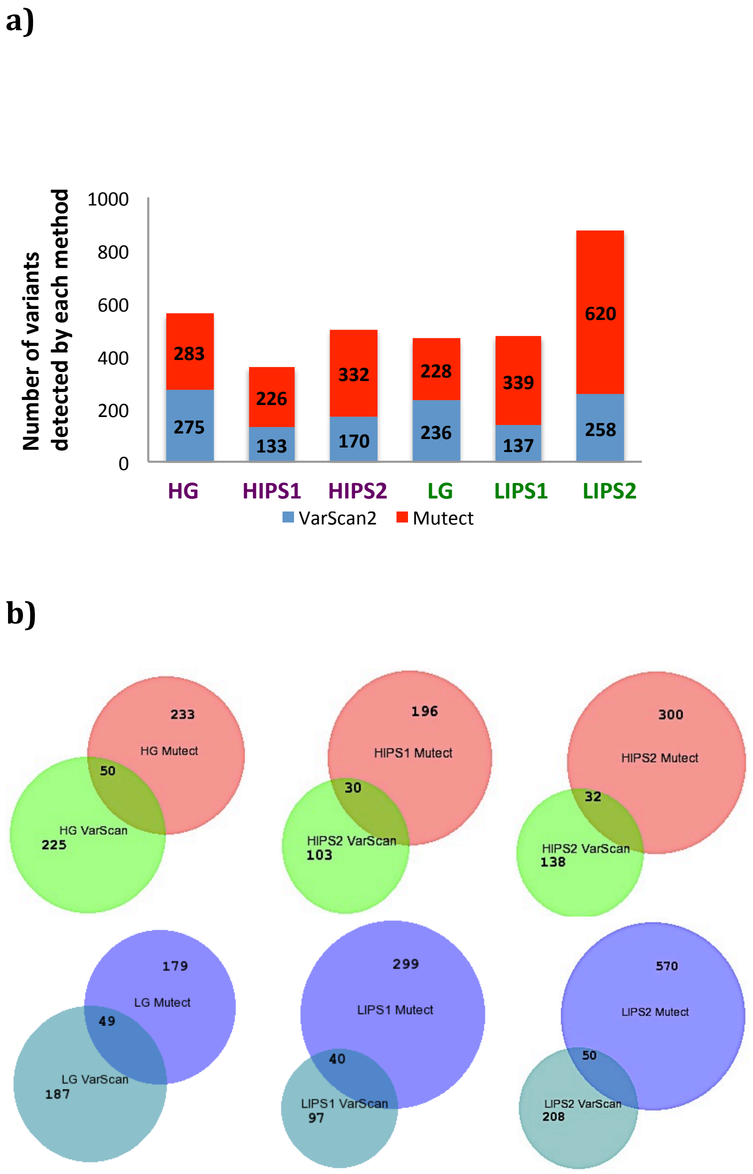


Figure 24: SNV data of exome analysis

a) Represents the number of somatic SNVs called by both the tool on each of the tumor samples and their corresponding tumor-iPSCS. Red represents

Mutect while blue represents VarScan2 b) Represents the overlap of the somatic SNVs in same samples identified both by Mutect and VarScan2.

3.2.2 Somatic CNV and driver analysis show that iPSCs were tumor derived

Subsequently, I used Control-FREEC, a bioinformatics tool that can efficiently detect somatic chromosomal rearrangements at copy level from matched tumor/control samples applying the normalization and segmentation on the tumor/control copy profiles generated from exome data. From the somatic copy number variations (sCNV) analysis of each samples [Figure 25 a] has emerged, on the one hand, a correlation of 50% [Figure 25b] between the high-grade iPSC and the high-grade tumor. One the other hand a correlation of 60% [Figure 25b] between sCNVs of low-grade tumor and its derivative iPSCs, confirms that the iPSC were indeed tumor-derived, and were not the result of inadvertent reprogramming of potentially contaminating stromal cells. Figure 26 shows the representative chromosomes in tumors and its iPSCs showing similar copy alteration profiles.

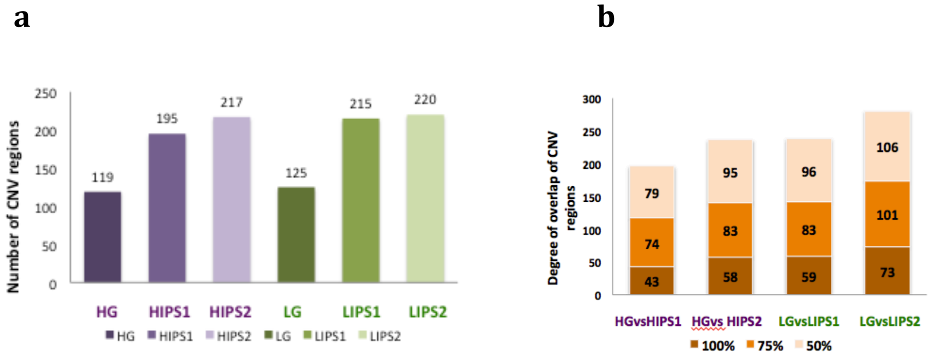
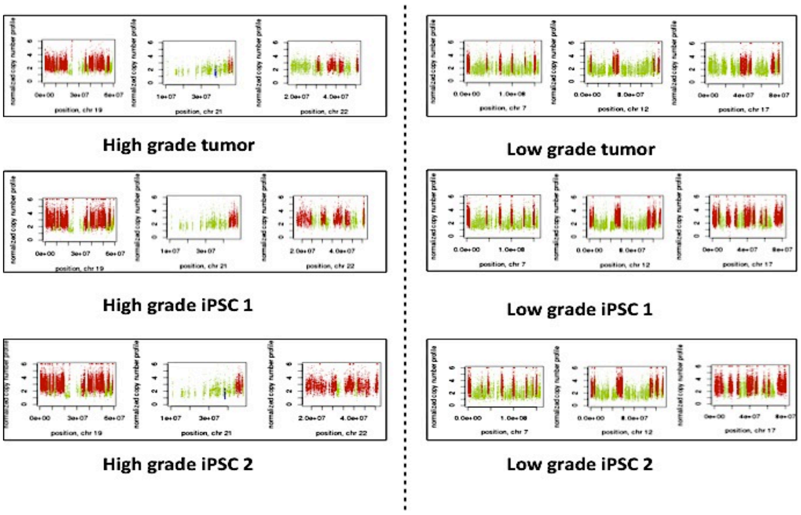


Figure 25: CNV data of exome analysis

a: Represents the number of CNV regions detected across tumor and its associated iPSCs. **b:** Degree of overlap of CNV regions between tumor and its corresponding iPSC derivatives. At 100% identity of CNVs regions shared upon reprogramming the overlap is 50-60% between tumor and iPSCs, upon relaxing the identity stringency to 75% or 50% the degree of overlap of CNV regions between tumor and its iPSC counterparts increases over 70%. HIPS1 and HIPS2 stands for iPSC clone 1 and clone 2 of high grade and LIPS1 and LIPS2 for iPSC clone 1 and clone 2 for low-grade tumors.



High grade tumor

Low grade tumor

High grade iPSC 1

Low grade iPSC 1

High grade iPSC 2

Low grade iPSC 2

Figure 26: Representative chromosomes that have similar copy profiles between its tumor and corresponding iPSC clones

Chr 19,20 and 21 capture similar copy variations across HG OC and HG-OC-iPSCs while similar trend is also obtained in LG OC and LG-OC-iPSCs across Chr 7,12 and 17.

The driver mutation analysis on the tumors and their iPSC derivatives was performed using IntOGen (Gonzalez-Perez et al., 2013), a bioinformatic tool that relies on the application of statistical methodologies aimed at filtering out alterations that are expected by chance, and selecting only those that are statistically significant (driver mutations). From these analysis mutations in the genes F8, ROBO2 and TCF4 genes were assigned driver status in LG tumor and its reprogrammed counterpart. F8 has already been found to harbor driver mutation in other OC TCGA samples by this tool. ROBO2 and TCF4 have not only been assigned with driver status in other cancers like Cutaneous Melanoma, endometriod carcinoma, Medulloblastoma by IntOGen but also as key drivers in Pan-Cancer project across 20 different tumor types (Weinstein et al., 2013a). In addition I found two oncogenes NSD1, GFI1B from the copy variations analysis that are shared between HG OC and HG OC-iPSCs and ETV3 and ETS2 oncogenes from copy regions shared between LG OC and LG OC-iPSC. The combined set of exon mutations and CNA that have been generated from the usage of the above mentioned tools provides information about the order of magnitude of the lesions associated with ovarian malignancy and provide a first identification of the candidate genes implicated in OC development as well.

These results from exome sequencing analysis, show that the mutations and copy number variations harbored in iPSCs are coming from the primary tumor, and suggest that genetic background between iPSCs and the primary tumor is fractionally preserved during reprogramming.

3.2.3 Meta analysis of the candidate genes revealed key genetic players associated with aberrations in publicly available TCGA datasets

In order to compare the results obtained from our mutation and CNA analysis with the published datasets I used Cbioportal database [http://www.cbioportal.org/]. This database provides a great resource for researchers to search, analyze and visualize data sets of different cancer consortiums across the world where wealth of information's of tumor data are preserved in forms of clinical information, genomic characterization data, and high level sequence analysis. Currently the database hosts 21441 tumor samples from 91 cancer studies. I interrogated the genes that were mutated and copy number altered between HG OC and HG OC-IPSCs (an indication of somatic variations preserved upon reprogramming) and found some of them to be dysregulated in a cohort of 316 HG OC patients with a frequency of above 15% [Figure 27]. I found that the genes PTP4A3, C8ORF33 and MBD1 associated to copy variations in our dataset are also associated with copy number aberrations, up regulation and mutation in 316 patients of HG OC TCGA (Weinstein et al., 2013b) with a frequency of alteration ranging between 33-34%[Figure 27]. Interestingly these results suggest key genetic players associated with aberrations both in HG OC and OC-iPSCs of high grade and that are shared between the two corroborating the fact that genetic lesions are preserved upon reprogramming.

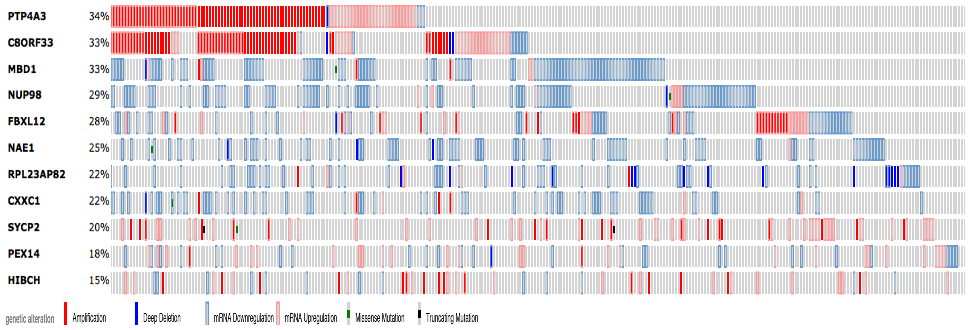
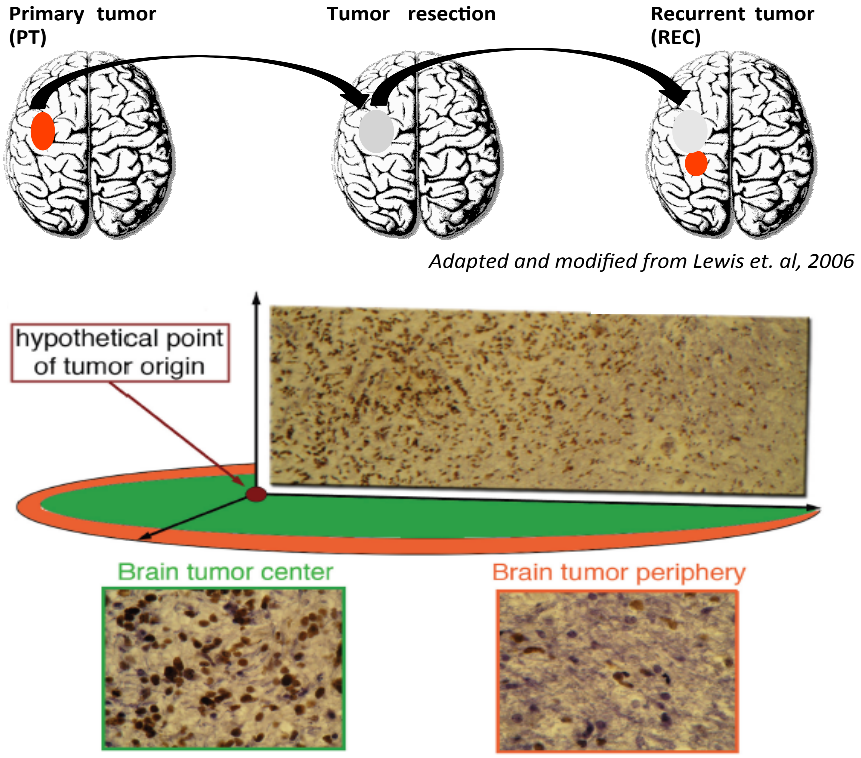


Figure 27: Incidence of SNVs and CNVs shared between HG tumor and its iPSCs across TCGA

The above heatmap represents the percentage of genes dysregulated in OC TCGA cohort in the cbioportal for the genes that were common SNVs and CNVs between HG tumors and its tumor-iPSC derivative. Amplification (copy-gain) is highlighted in bright red while deletion (copy-loss) is highlighted in bright blue. The pale red represents mRNA up-regulation while mRNA down-regulation is represented by pale. The green bar represents missense mutation while black highlights truncating mutations.

3.3 Molecular characterization of primary and recurrent glioblastoma through transcriptomic analysis

We dissected the transcriptomes of a rare cohort of 37 human samples from 7 patients derived from the infrequent surgery at disease recurrence. For each patient, glioma-initiating cells (GIC's) were extracted along with the biopsies around the GICs that represented the peripheries. This was done for both primary and recurrent samples. The peripheral biopsies of the primary tumor were marked with MRI–detectable clips. This enabled our collaborators to establish, upon recurrence, the biopsied sites that sustained relapse. This provided us with a unique opportunity to trace the human GBM evolution avoiding confounding effect of the inter-individual genetic heterogeneity. The details of samples extraction are provided in the Materials and Methods section 2.1.1. We performed RNA-Seq on these topological compartments to characterize the genome-wide expression profile representing: i) the center of primary tumor (PC); ii) the peripheries of the primary tumor that give rise to the recurrence (PGRT); iii) the peripheries of the primary tumor that do not give rise to the recurrence (PDGRT); iv) the center of the recurrent tumor (RC); v) the periphery of the recurrent tumor giving rise to a tertiary tumor (RPGRT); vi) the periphery of the recurrent tumor that do not giving rise to a tertiary tumor (RPDGRT). Figure 28 and Figure 29 show the topological compartments and the datasets considered for our experimental design and downstream analyses.



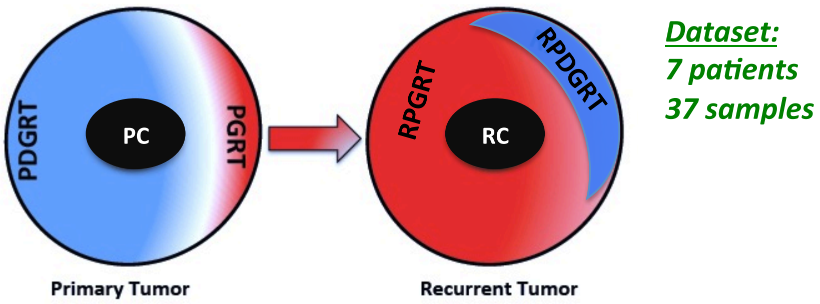
Adapted and modified from Lewis et. al, 2006

Adapted from Glas et al., Ann Neurol, 2010

Figure 28: Representation of the patient's brain with tumor depicting the primary tumor and the recurrent tumor

a) Representation of the patient's brain with tumor depicting the primary tumor and the recurrent tumor.

b) Representation of topological compartments of the brain tumor at a sub-cellular level.



PC= Centers of Primary Tumor
PGRT = Peripheries that give rise to the recurrent tumor
PDGRT= Peripheries that do not give rise to the recurrent tumor
RC = Centers of the recurrence
RPGRT = Recurrent peripheries giving rise to a tertiary tumor
RPDGRT = Recurrent peripheries that do not gives rise to a tertiary tumors

Figure 29: Representation of various topological compartments of the primary and recurrent tumor in a GBM patient along with nomenclature

Upon profiling the transcriptomes of primary and recurrent GBM samples we performed the following 3 key comparisons by following the analytical workflow shown in figure 30:

i) The transcriptomic difference between the primary vs recurrent tumor

ii) The difference between tumorigenic vs non-tumorigenic peripheries in the primary tumor through evaluation of the contribution with that of the primary center and finally the

iii) The transcriptomic differences between primary peripheries giving rise to recurrent tumor vs the centers of the recurrent tumors.

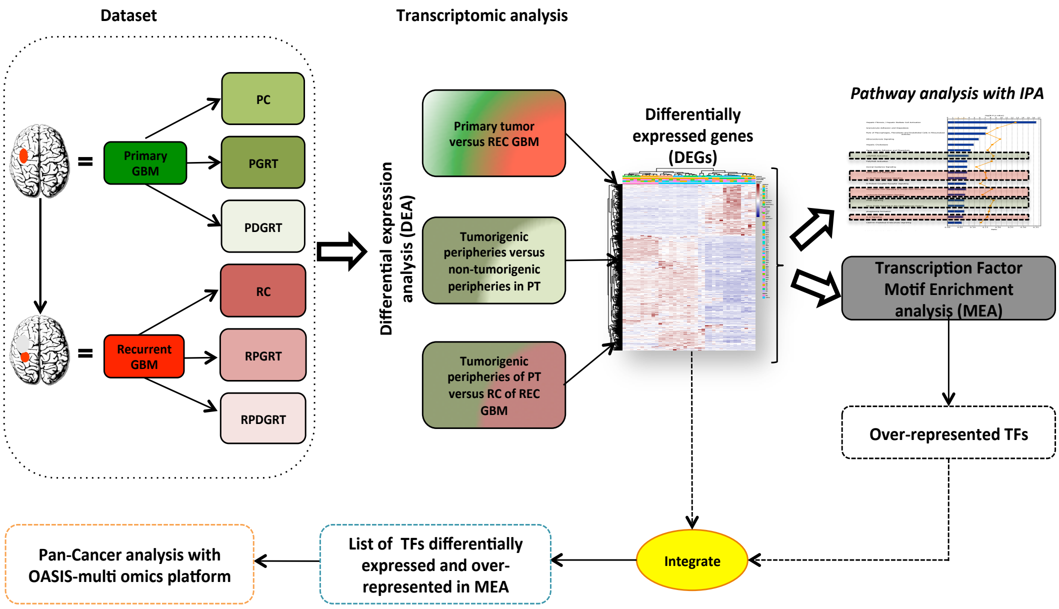

Dataset
Transcriptomic analysis
Primary GBM
Recurrent GBM
PC
PGRT
PDGRT
RC
RPGRT
RPDGRT
Differential expression analysis (DEA)
Primary tumor versus REC GBM
Tumorigenic peripheries versus non-tumorigenic peripheries in PT
Tumorigenic peripheries of PT versus RC of REC GBM
Differentially expressed genes (DEGs)
Pathway analysis with IPA
Transcription Factor Motif Enrichment analysis (MEA)
Over-represented TFs
Integrate
List of TFs differentially expressed and over-represented in MEA
Pan-Cancer analysis with OASIS-multi omics platform

Figure 30: Analytical workflow employed in capturing the TF mediated GBM progression

3.3.1 Transcriptomic analysis of patients with primary and recurrent GBM

Differential expression analysis patients with primary tumor (PT) and recurrent tumor (REC GBM) revealed 1702 differentially expressed genes (1163 up in PT while 539 genes down regulated) with log2FC 1.5 and FDR < 0.01. An unsupervised clustering analysis based on the 1702 DEGS showed that PT and REC GBMs segregate in 2 different clusters, highlighting a major change in the transcriptional programs underlying REC GBM.

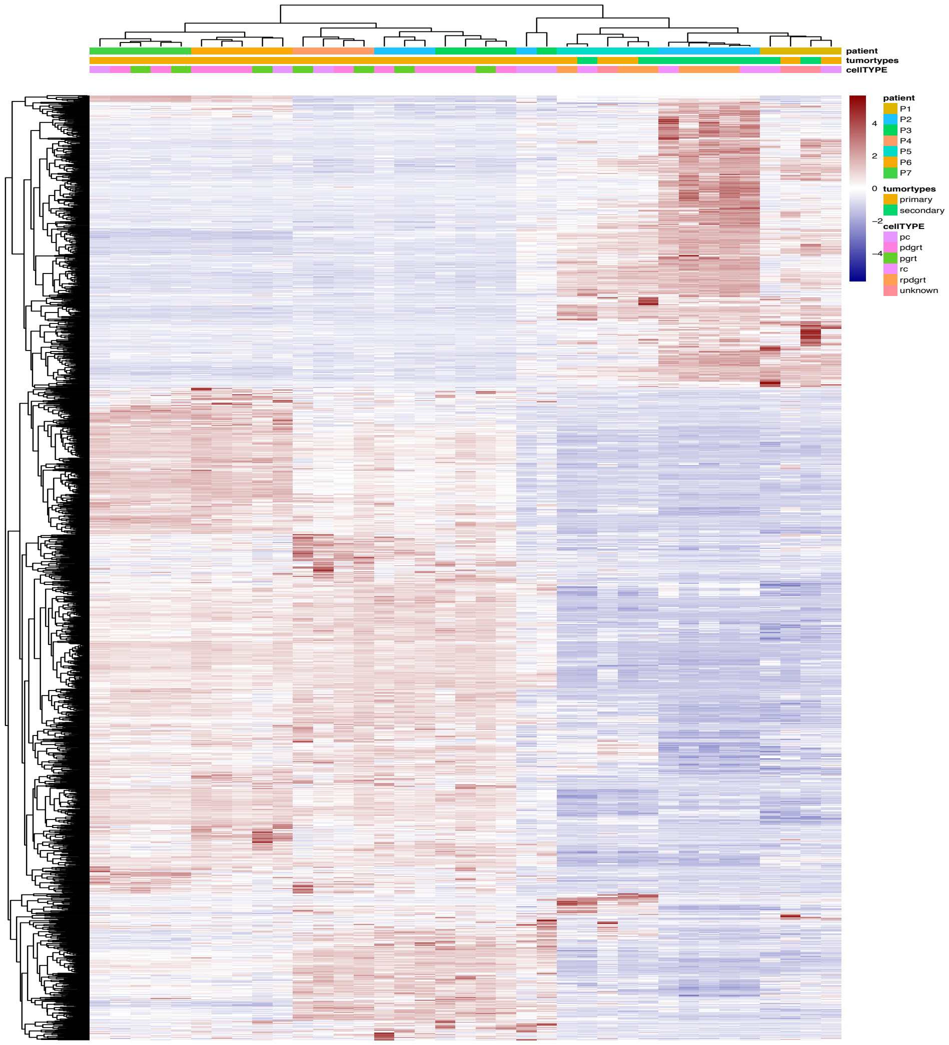


Figure 31: Heatmap of DEGs between primary and recurrent GBM samples

3.3.1.1 Canonical pathways enriched as a result of differential expression of genes between primary and recurrent GBM patients

A canonical pathway analysis of these DEGs revealed pathways like cAMP-mediated signaling, endothelin-1 signaling, intrinsic Prothrombin activation pathway, Gai Signaling, etc were seen to be inactivated while CDK5 signaling and complement system pathway of innate immunity are activated. The figure below shows the most significant pathways that were altered.

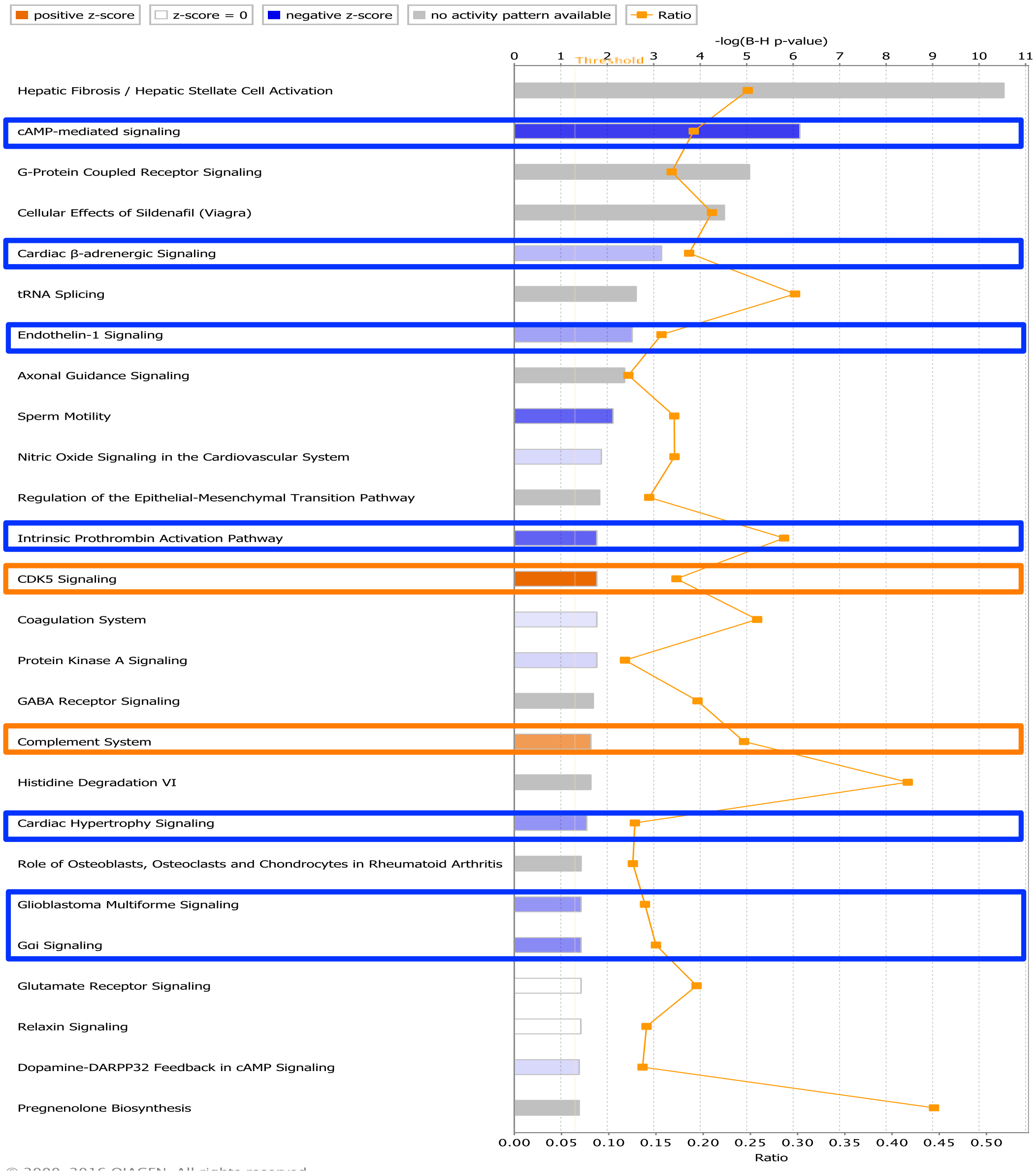


Figure 32: Representation of enriched pathways found by IPA as a result of changes in gene expression between PT and REC GBM

The color orange represents pathways predictively activated as found by IPA and blue represents pathways predictively inactivated as found by IPA

Suppression of the cAMP pathway has been commonly seen in many different types of cancers, including in GBM in which this pathway has been seen to be suppressed and thereby counteract apoptosis (Daniel, Filiz, & Mantamadiotis, 2016). This pathway was seen to be inactivated in recurrent GBM samples in our datasets thus indicating apoptosis evasion as a possible mechanism at work also in recurrent stage of the disease. Several lines of evidence have been cited that indicate the importance of endothelin-1 receptor signaling or ET-1R in cancer. Endothlein-1 (ET-1) signaling is often considered crucial for cancer cell proliferation either as stand-alone factor or in a cooperative manner with other tumor growth factors. Cell proliferation, metastasis, angiogenesis and drug resistances have been often evidenced to be regulated by ET-1R (Rosanò & Bagnato, 2016). More specifically, its dysregulation has been linked to development and progression in many cancers.

In addition to the canonical pathways analysis, we interrogated the set of 1702 DEGs between PT and REC GBMs with the gene set signatures that distinguish the 4 GBM subtypes: classical, mesenchymal, proneural and neural (Verhaak et al., 2010). These gene set signatures were defined by the TCGA consortium in a large cohort of GBM patients and showed a high prognostic value.

The comparison of these gene set signatures with our DEGs showed an overlap of 16 genes with the gene set signature enriched in the classical subtype, 49 with mesenchymal, 14 pro-neural and 9 with Neural. The overlap analysis using hypergeometric test revealed a significant overlap with the genes characterizing the mesenchymal subtype (p< 2.400e-12) of which 92% (45/49) were up-regulated in PT (figure 33).

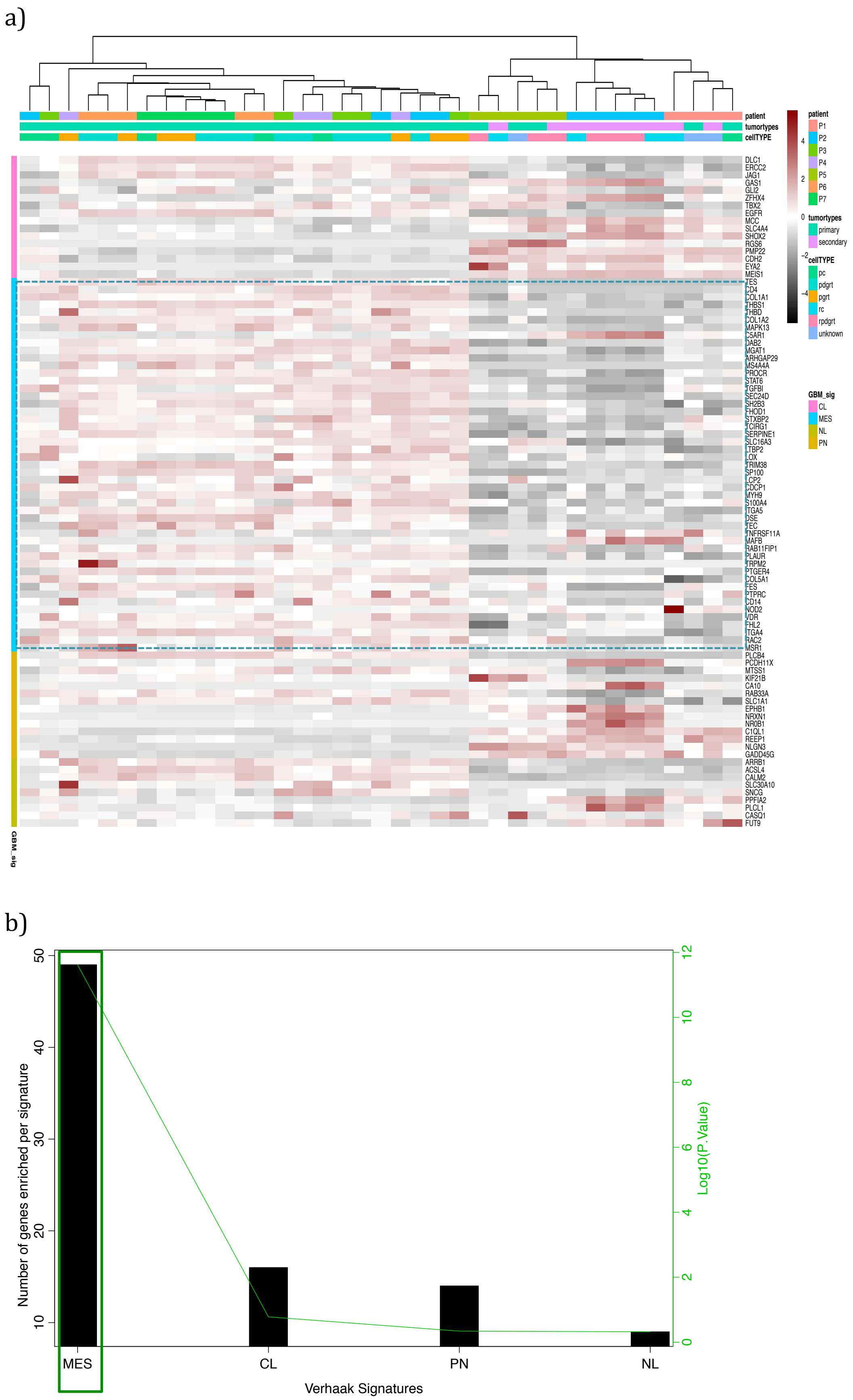


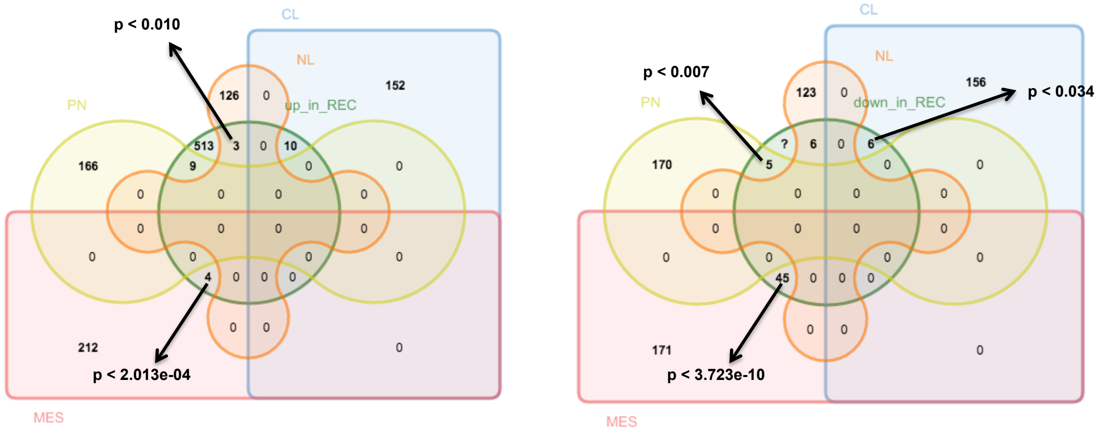
Figure 33: DEGs between PT and REC GBM are enriched for mesenchymal signature genes (Verhaak et al., Cancer Cell, 2010)

a) Heatmap of DEGs overlapping for specific Verhaak GBM signature between PT and REC GBM samples

b) Two way plot of the DEGs enriched for each signature. x-axes represents the 4 molecular subtypes while left y axes represents the number of genes enriched for each subtype and right y axis shows the –log10(p-value) enrichment score for those signatures. Only the genes under mesenchymal subtype are significantly enriched also marked in green box.

Figure 34 a) represents the overlap between our DEGs (defined in PT Vs REC) and the 4 gene set signatures enriched in the 4 GBM molecular subtypes defined by Verhakk et al. when assessed separately based upon their direction of regulation.

a)



b)

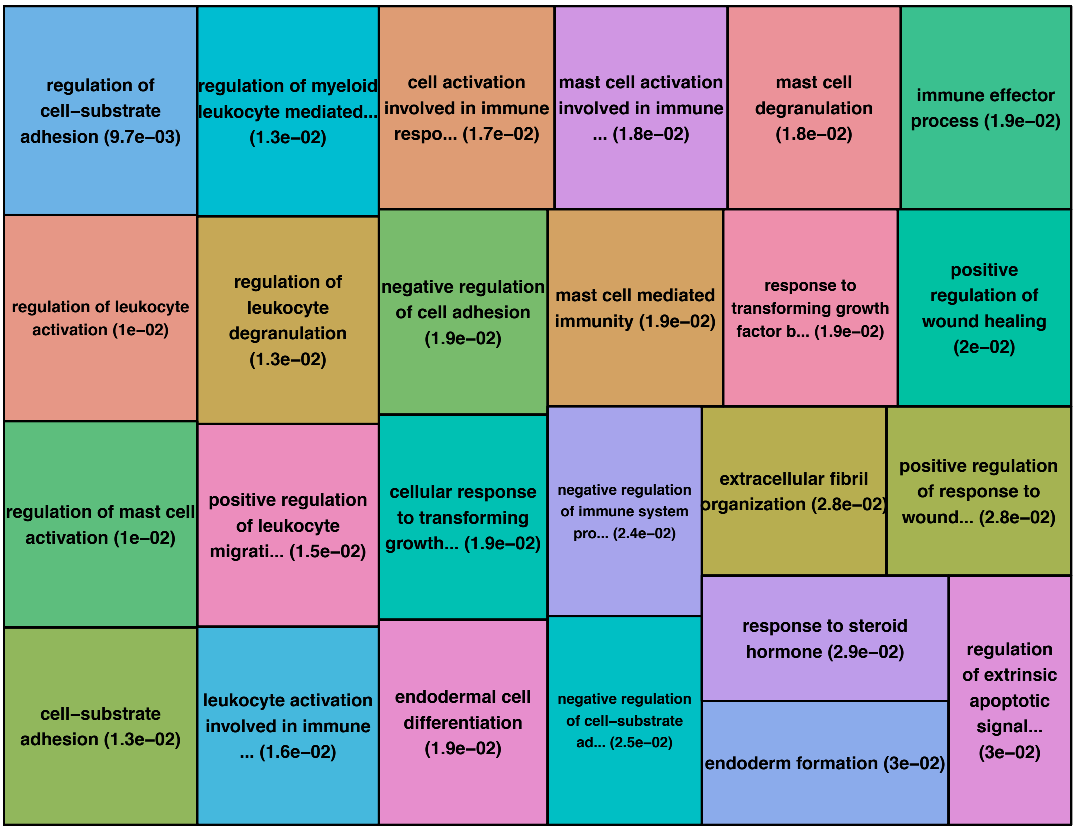


Figure 34: Enrichment of up and down-regulated genes separately to that of the Verhaak's GBM signatures

a) Down-regulated genes in REC GBM (i.e. up-regulated in PT) are the most significantly enriched for Verhaaks' mesenchymal signature genes.
b) Only genes down-regulated in REC GBM showing significant enrichment for mesenchymal's signature shows significant enrichment for GO categories specifically for Biological Processes.

The Venn diagram in figure 34a (left) represents the overlap between the gene set signatures of the 4 molecular subtypes and the DEGs up-regulated in REC GBM while the right represents the DEGs down-regulated in REC GBM. Among the DEGs that were down-regulated in REC GBM even showed enrichment for gene-ontology categories for biological process that were associated with mesenchymal signatures. Above figure 34b shows the various Biological processes that are involved with mesenchymal signatures. Processes like response to transforming growth factor, cellular response to transforming growth factor, negative regulation to immune system and various migratory and adhesion process were seen to be enriched for these 45 DEGs that were up-regulated in PT and significantly enriched for mesenchymal signatures.

3.3.1.2 DEGs involved TFs between PT versus REC GBM

Among these DEGs I found 103 human TFs (the list of human TFs were extracted from AnimalTFdb database http://bioinfo.life.hust.edu.cn/AnimalTFDB1.0/) that were differentially expressed as shown in the barplot in figure 35.

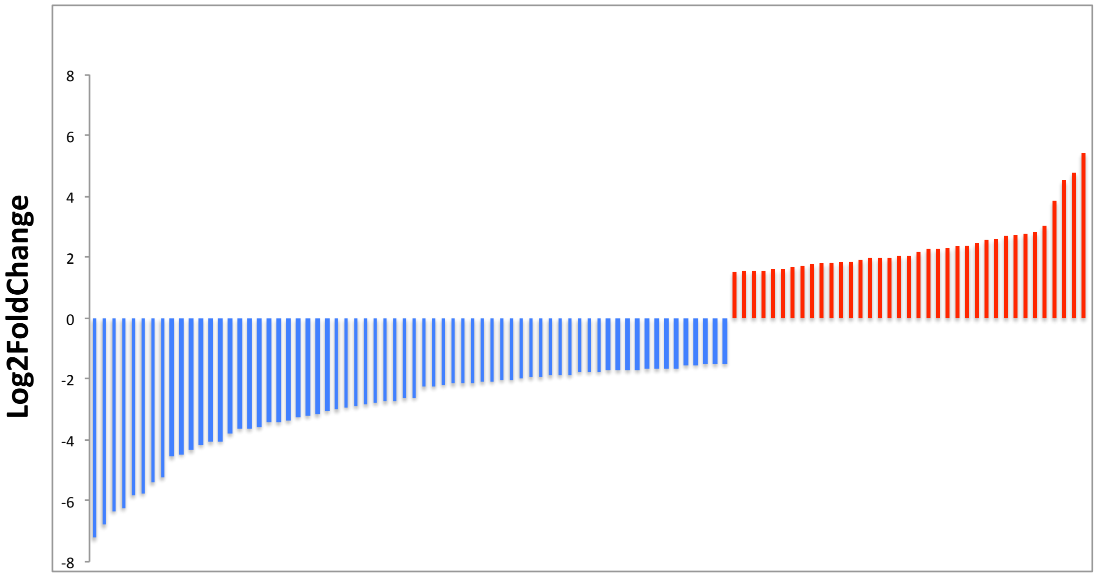


Figure 35: Barplot representing TFs involved as DEGs between PT and REC GBM

IPA pathway analysis of these 103 core TFs that could characterize the contribution of TF mediated trancriptomic differences underlying primary vs recurrent tumor revealed Adipogenesis pathway, Sonic-hedgehog (SHH) signaling pathway and TGF-B signaling pathway. These pathways were however not predicted to be activated or inactivated in the REC GBM by IPA but were found altered due to the differential expression of these genes. Interestingly, the alteration of these pathways was due to the enrichment for genes specifically down regulated in REC GBM samples. Figure 36 (below) shows two panels for the pathways enriched for TF DEGs between PT and REC GBM. The former panel shows only the pathways while the second panel indicates the TF DEGs that enriched these pathways are mostly down-regulated in REC GBM.

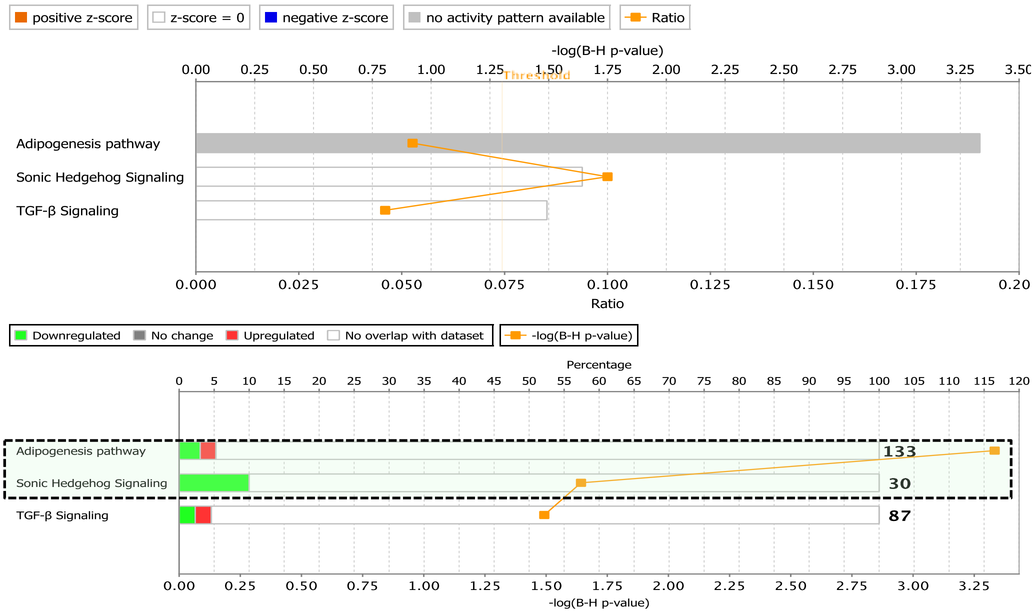
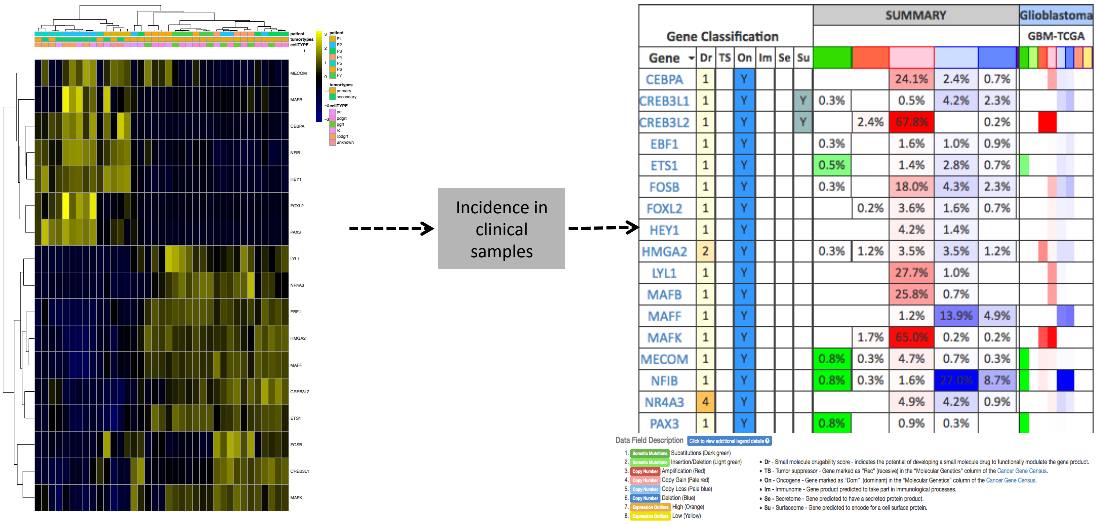


Figure 36: Pathways involved between TF DEGs between PT and REC GBM shows they are dominated by down regulated genes in REC GBM

The first panel shows pathways that were enriched by IPA (however none was predictively assigned to be activated/inactivated by IPA). The second panel shows the same pathways but with an alternative representation revealing the contribution of down and up-regulated genes in them. Green color represents genes down regulated in REC GBM while red signifies those that were up regulated.

The enrichment of the SHH pathway is mostly due to genes that are down-regulated genes in REC GBM, while the enrichment for the TGF-Beta pathway comprises both up and down-regulated genes. This was not so informative about the status of pathways that characterize the recurrent tumor but gave an indication of the direction of dysregulation in these pathways. I also found oncogenes among these 103 TFs that were differentially expressed between PT vs REC GBM. Upon interrogation of these 103 TF DEGs with oasis-cancer genomics portal developed by the Pfizer Computational Biology Group (Fernandez-Banet et al., 2015) I found 17 oncogenes. This analysis revealed that some of these TF-Oncogenes, which were DEGs between PT vs REC GBM had high level of dysregulation in 577 TCGA GBM samples. Genes like CREB3L2 have frequency of over ~67% times having a low-level gain of copy in 577 TCGA GBM samples while MAFK represented over ~60% of the similar copy gain dysregulation in total number of 577 GBM TCGA samples. The figure 37 below represents the TF-Oncogenes that were DEGs between PT vs REC GBM along with their level of dysregulation in clinical TCGA samples. This figure scores the importance of these DE TF-oncogenes and their relevance not only in our datasets but also in clinical TCGA GBM patients.



Gene Classification							SUMMARY					Glioblastoma GBM-TCGA
Gene	Dr	TS	On	Im	Se	Su	🟩	🟥	🟥	🟦	🟦	
CEBPA	1		Y						24.1%	2.4%	0.7%	
CREB3L1	1		Y			Y	0.3%		0.5%	4.2%	2.3%	
CREB3L2	1		Y			Y		2.4%	67.8%		0.2%	
EBF1	1		Y				0.3%		1.6%	1.0%	0.9%	
ETS1	1		Y				0.5%		1.4%	2.8%	0.7%	
FOSB	1		Y				0.3%		18.0%	4.3%	2.3%	
FOXL2	1		Y					0.2%	3.6%	1.6%	0.7%	
HEY1	1		Y						4.2%	1.4%		
HMGA2	2		Y				0.3%	1.2%	3.5%	3.5%	1.2%	
LYL1	1		Y						27.7%	1.0%		
MAFB	1		Y						25.8%	0.7%		
MAFF	1		Y						1.2%	13.9%	4.9%	
MAFK	1		Y					1.7%	65.0%	0.2%	0.2%	
MECOM	1		Y				0.8%	0.3%	4.7%	0.7%	0.3%	
NFIB	1		Y				0.8%	0.3%	1.6%	27.0%	8.7%	
NR4A3	4		Y						4.9%	4.2%	0.9%	
PAX3	1		Y				0.8%		0.9%	0.3%		

Data Field Description Click to view additional legend details

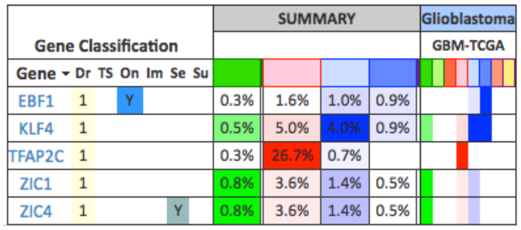
1. Somatic Mutations Substitutions (Dark green)
2. Somatic Mutations Insertion/Deletion (Light green)
3. Copy Number Amplification (Red)
4. Copy Number Copy Gain (Pale red)
5. Copy Number Copy Loss (Pale blue)
6. Copy Number Deletion (Blue)
7. Expression Outliers High (Orange)
8. Expression Outliers Low (Yellow)

- **Dr** - Small molecule drugability score - indicates the potential of developing a small molecule drug to functionally modulate the gene product.
- **TS** - Tumor suppressor - Gene marked as "Rec" (recesive) in the "Molecular Genetics" column of the Cancer Gene Census.
- **On** - Oncogene - Gene marked as "Dom" (dominant) in the "Molecular Genetics" column of the Cancer Gene Census.
- **Im** - Immunome - Gene product predicted to take part in immunological processes.
- **Se** - Secretome - Gene predicted to have a secreted protein product.
- **Su** - Surfaceome - Gene predicted to encode for a cell surface protein.

Figure 37: TF DEGs involves a set of oncogenes that shows high level of dysregulation in clinical GBM TCGA samples

3.3.1.3 Motif enrichment analysis predicted a core set of oncogenes and tumor suppressor genes encoding for TFs controlling the transcriptional programs in REC GBM

After the definition of the transcriptomic changes underlying GBM recurrence, I interrogated the differentially expressed between PT and REC GBMs by performing the TF motif enrichment analysis (MEA). This analysis aims at providing a mechanistic interpretation of the transcriptomic changes in recurrent GBM by identifying putative TFs master regulators through the analysis of the binding sites at the promoter regions of the DEGs. Master regulators TFs are key molecules of a cellular network that are supposed to control the cell-type specific transcriptional programs. Indeed, as was recently argued, the aberrant activity of master regulators is *"both necessary and sufficient for tumor cell state implementation and maintenance"* (Andrea Califano & Mariano J. Alvarez, Nature Reviews Cancer 2016). Motif enrichment analysis (MEA) was done on the promoter sequences of entire 1702 DEGs obtained as a result of DE analysis between PT vs REC GBM with Pscan tool (Zambelli et al., 2009) with default settings. MEA revealed 61 over-represented TFs with a p-value less than 0.05 where 5 of them happened to be differentially expressed in our comparison: TFAP2C, ZIC4, ZIC1, EBF1 and KLF4. These 5 TFs were found to be down-regulated in the REC GBM samples. Figure 38 shows the incidence of 5 of these targets in GBM TCGA and confers the status of Oncogene to them.



Gene Classification							SUMMARY				Glioblastoma
											GBM-TCGA
Gene	Dr	TS	On	Im	Se	Su					
EBF1	1		Y				0.3%	1.6%	1.0%	0.9%	
KLF4	1						0.5%	5.0%	4.0%	0.9%	
TFAP2C	1						0.3%	26.7%	0.7%		
ZIC1	1						0.8%	3.6%	1.4%	0.5%	
ZIC4	1				Y		0.8%	3.6%	1.4%	0.5%	

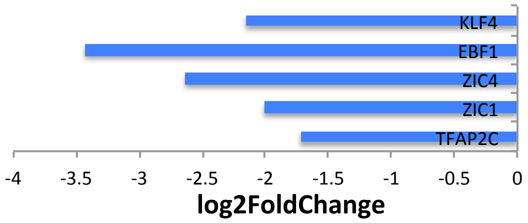


Figure 38: Representation of dysregulation of over-represented TFs that are DEGs in clinical GBM TCGA samples

The left panel shows the list of over-represented TF targets found from MEA analysis that are differentially expressed in our GBM datasets. The table shows their dysregulation in clinical GBM TCGA samples, while the barplot on the right shows their down-regulation in REC GBM.

Among these gene sets I found the oncogene EBF1 and a tumor-suppressor gene (TSG) KLF4, both down-regulated in REC GBM when compared to PT. KLF4 is regarded as a key transcriptional target in breast-cancers and colon-cancers. Its over-expression has been shown to be associated with reduced tumorigenicity (Dang et al., 2003), (Wang et al., 2015). This can be an important transcription factor when assessed in vitro to see if its over-expression in REC GBM could be associated with reduced tumorigenicity in our GBM model. The tumor suppressor role of EBF family genes as transcription factors has already been associated in GBMs (Liao, 2009), (Guilhamon et al., 2013).Particularly, EBF1 has been reported to associate with TET2, a member of the TET enzyme family that cause oxydation of 5-methyl-cytosine (5mC) that will eventually lead to DNA-demethylation at specific loci. Inhibition of EBF1 might be an alternative mechanism to achieve aberrant DNA hypermethylation during gliomagenesis (Guilhamon et al., 2013). Since the orchestrated regulation of DNA methylation is crucial for neuronal differentiation (Mohn et al., 2008),inactivation of dna demethylation could alter physiological differentiation and thereby contribute to the oncogenic phenotype.

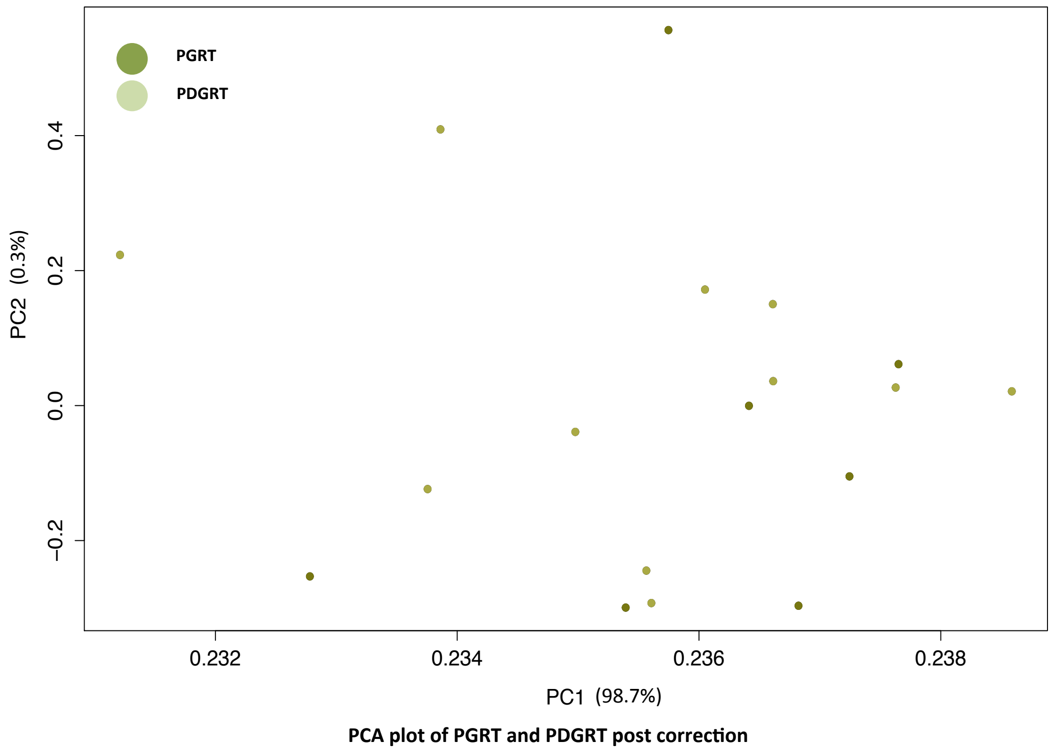
Note: KLF4 is not described, as a tumor suppressor here since the version of Cancer Gene Census used by OASIS-cancer genomics is v70 while the current upgraded version of Cancer Gene Census in COSMIC database is v78, which confers it as a tumor suppressor.

3.3.2 Capturing the transcriptomic differences between tumorigenic versus non-tumorigenic peripheries in the primary tumor of GBM patients through evaluation of the contribution with that of the primary center

I studied differences between the tumorigenic (peripheries giving rise to recurrent tumor: PGRT) and the non-tumorigenic peripheries (peripheries accorded as not giving rise to recurrent tumor based on their topology: PDGRT) in the GBM samples of the PT. This was done by assessing their differential regulation upon a base reference of the primary center (PC) thus extracting only the DEGs that were exclusively altered in the PGRT and the PDGRT of the PT from the PC. The genes that were altering in PGRT might seat for molecular events that prepared the tumor microenvironment that could lead to a recurrence. On the other hand the genes altered in the PDGRT could be regarded as turnover genes whose over-expression could lead to reduced tumorigenicity.

The PCA projection of the samples belonging to PGRT and PDGRT revealed that both the tumorigenic and non-tumorigenic cellular compartments had strikingly no difference at the level of gene expression. Upon projecting the first two principal components (PC1 and PC2) that accounted for the majority of the variability as seen in the below figure 39 a, it was evident that the samples were not clustering based on their cell type. Upon performing the differential expression analysis as well there was no DEGs as seen in (Figure 39 b).

a)



b)

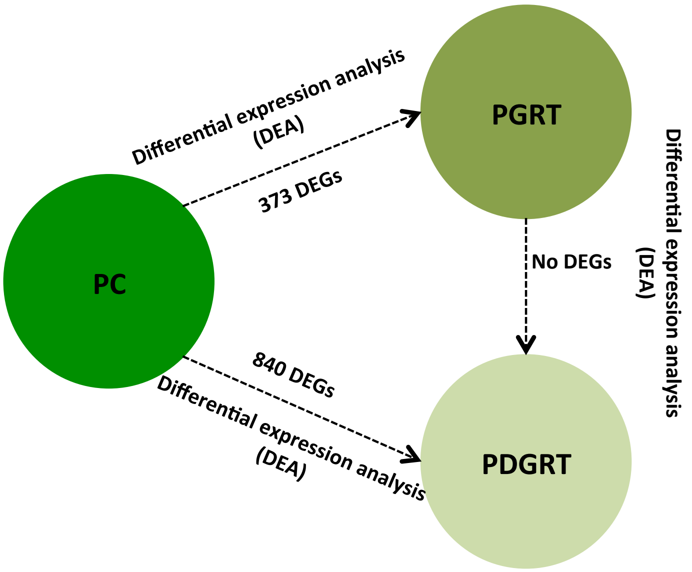

Differential expression analysis (DEA)
373 DEGs
PGRT
PC
No DEGs
Differential expression analysis (DEA)
840 DEGs
Differential expression analysis (DEA)
PDGRT

Figure 39: The PCA plot on the shows that upon performing batch correction this difference between tumorigenic and non-tumorigenic peripheries at transcriptional level are majorly compromised

a) Represents the samples clustering for PGRT and PDGRT in orthogonal space.

b) Represents the number of DEGs obtained while comparing the various cellular compartments.

So I devised a strategy to account for the differences between tumorigenic and non-tumorigenic peripheries by assessing their differential regulation upon a base reference of the primary center (PC). This was done as shown in the figure 13 where the DEA was performed between PC and the peripheries (PGRT and PDGRT here) separately that extracted only exclusive DEGs between tumorigenic and non-tumorigenic peripheries. This gave me 48 DEGs that exclusively altered between PC and PGRT while 515 DEGs altering exclusively between PC and PDGRT. Upon farther assessing their directionality through their contribution to PC, genes specifically altering in tumorigenic and non-tumorigenic peripheries were obtained.

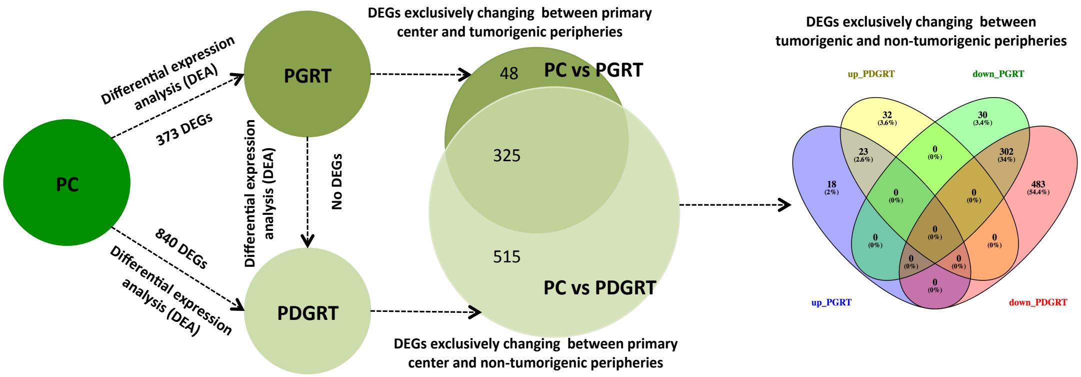

DEGs exclusively changing between primary center and tumorigenic peripheries
DEGs exclusively changing between tumorigenic and non-tumorigenic peripheries
Differential expression analysis (DEA)
PGRT
373 DEGs
PC
Differential expression analysis (DEA)
No DEGs
840 DEGs
Differential expression analysis (DEA)
PDGRT
48
PC vs PGRT
325
515
PC vs PDGRT
DEGs exclusively changing between primary center and non-tumorigenic peripheries
up_PDGRT
down_PGRT
32
(3.6%)
30
(3.4%)
23
(2.6%)
0
(0%)
302
(34%)
18
(2%)
0
(0%)
0
(0%)
483
(54.4%)
0
(0%)
0
(0%)
0
(0%)
0
(0%)
0
(0%)
0
(0%)
up_PGRT
down_PDGRT

Figure 40: Representative figure of extracting exclusive DEGs changing between tumorigenic and non-tumorigenic peripheries through their contribution with that of primary centers of the PT

The figure 41 (left heatmap) shows the 515 genes exclusively altered between PC and PDGRT; and (right heatmap) the 48 exclusive DEGs altering between PC and PGRT.

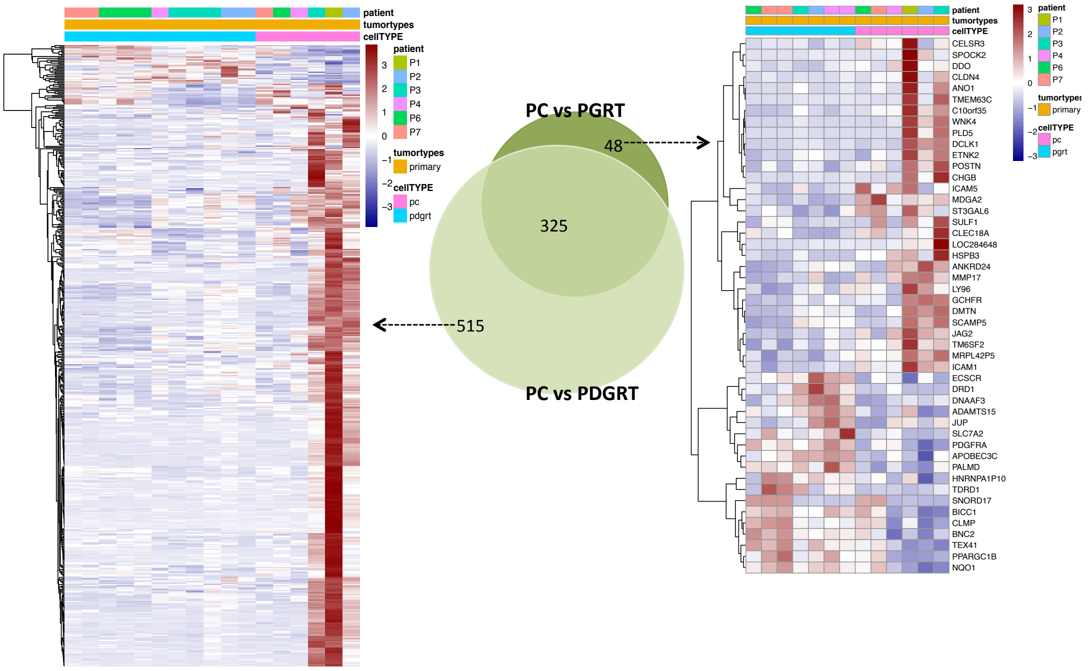


Figure 41: Left heatmap represents the exclusive DEGs between PC and non-tumorigenic peripheries. While the right heatmap represents the exclusive DEGs between PC and tumorigenic peripheries from their baseline primary center of PT

Among the 515 exclusive DEGs between PC and PDGRT there were few oncogenes that were differentially expressed. Oncogenes like ACKR3, FGFR3, FOXO1 whose expressions level were progressively going down in PDGRT not only from the PC but also from the PGRT. However the expression of these genes went up in the center of the relapse tumors. These genes could be fairly important as well contributing to the relapse.

The genes that were DEGs (n=48) exclusive to the comparison between PGRT and PC of the PT could also relevant, especially for understanding the key molecular events specific to transition from primary tumor to relapse. These exclusive 48 DEGs that differ between the cellular compartments of PC and PGRT of PT samples in GBM serve as important subunits underlying the development of a more aggressive phenotype leading to a relapse of GBM. This can be attributed due to the specific molecular events these 48 DEGs trigger and also since they show high dysregulation in clinical TCGA GBM samples. Out of these 48 exclusive DEGs, some of them were found to have high frequency of dysregulation in clinical GBM TCGA samples. In particular, BICC1 and TDRD1 have been reported with copy loss in GBM samples over ~25% of samples. Both of these genes were down-regulated in our PC while their expression went up in the PGRT which topologically neighbors the REC GBM. The expression of TDRD1 was farther down in RC of REC GBM centers while that of BICC1 was still partially maintained in our RC of REC GBM samples. BNC2 (a human TF) was found as up-regulated in PGRT whose expression was also maintained in the RC of REC GBM samples. BNC2 is often been reported to be associated with a chromosomal loss in tumors like HCC (Wu, Zhang, Liu, Lu, & Chen, 2016). Its relapse has not been associated with GBM as of now however on retrospective inspection of this gene in clinical TCGA samples of GBM it was found to be associated with copy loss in ~27% of them. This analysis also revealed a gene SPOCK2 whose expression went down from PC to PGRT while it is expression was up in the RC of REC GBM samples. This gene has been associated with a copy loss in over ~45% of TCGA GBM samples. Some other genes have also been identified from this analysis like CLDN4 (associated with ~67% copy gain in TCGA GBM), SNORD17, ICAM1 and ICAM5 (associated with ~27 % copy gain in TCGA GBM). CLBN4, ICAM1 and ICAM5 expression was going down from PC to PGRT while they went up in the RC of REC GBM. Finally I also found an oncogene PDGFRA that is up-regulated in PGRT with comparison to PC and was farther maintained in RC of REC GBM. The analysis of the gene association with clinical TCGA GBM samples is done with OASIS-Cancer genomics platform. This analysis thus revealed some key important genes that attributed to contributing properties of invasiveness and migration that could be potentially responsible for the relapse.

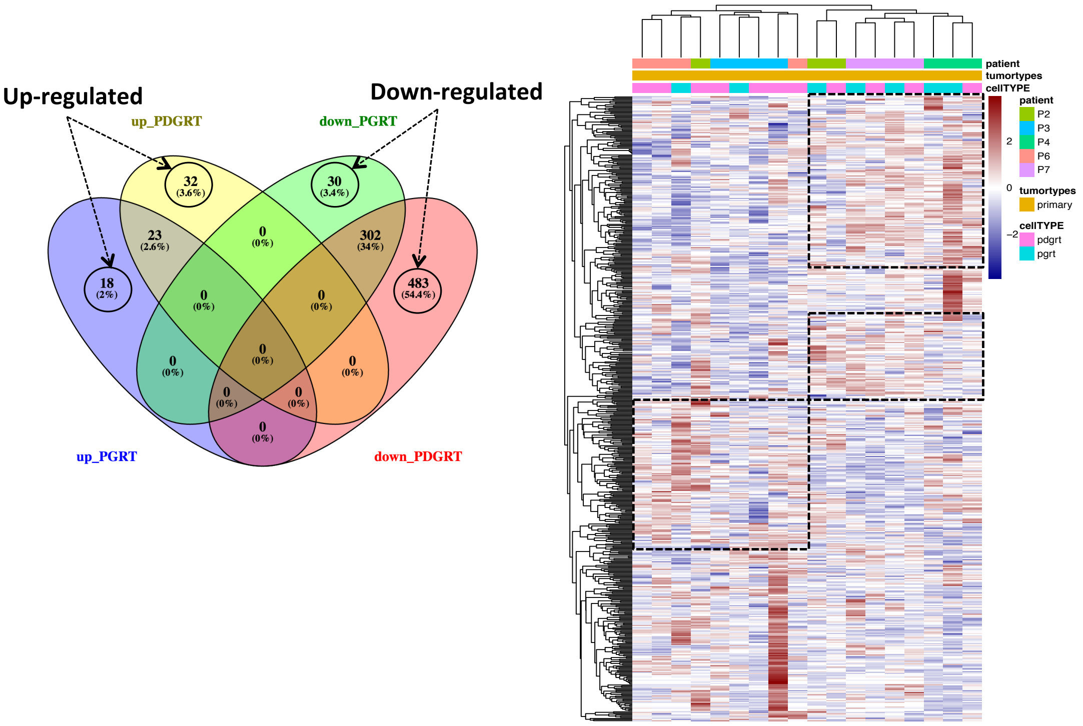


Figure 42: DEGs between PGRT and PDGRT of GBM primary tumors

a) The Venn diagram represents the DEGs specifically up and down regulated in tumorigenic and non-tumorigenic peripheries. b) Heatmap representation of these DEGs specifically between tumorigenic and non-tumorigenic peripheries.

The above Venn diagram on the left represents the genes specifically up and down-regulated in PGRT and PDGRT while assessing their differences from the PC. These genes (48 and 515) independently classified the tumorigenic and non-tumorigenic peripheries on two clusters when projected as seen in the above heatmap on the right. Thus I was able to obtain DEGs that were specifically changing between tumorigenic and non-tumorigenic peripheries which otherwise was not visible when DEA was directly applied on these two cellular compartments. Thus the difference between tumorigenic and non-tumorigenic peripheries at the level of gene expression was obtained from this analysis.

Pathway analysis was done on these genes that were specifically up and down in PGRT and PDGRT. IPA canonical pathway analysis only revealed axonal guidance signaling pathway to be significantly enriched as a result of the changes in gene expression between tumorigenic and non-tumorigenic cellular compartments.

MEA analysis on these DEGs revealed around 93 over-represented TFs that were having binding sites with the DEGs. Upon integration of these upstream TFs to the DEGs of this comparison, I found EGR2 as an upstream TF that was also differentially expressed between PGRT and PDGRT. This gene was reportedly found to be associated with ~45.06% of copy loss while the deletion level associated is around 20% in clinical TCGA GBM samples by OASIS-Cancer genomics platform. Role of EGR2 in cancers have been published already. In metastatic Leiomyosarcoma this gene was overexpressed (Davidson et al., 2014). It was also been found to be associated with functioning of the immune system. It was rather appearing to be acting as a molecular switch for the immunomodulators that gets affected in harsh hypoxic environments (Barbeau et al., 2014). This analysis was able to find an important TF that is associated with immunomodulatory functions.

3.3.3 Transcriptomic analysis between primary tumorigenic peripheries and recurrent tumor centers of the GBM patients to capture the transitory events underlying a relapse.

Finally, I performed a DE analysis in order to identify the transcriptional changes that characterize the transition between tumorigenic peripheries of PT and the recurrent centers. This analysis allows understanding the molecular events associated with relapse at a deeper partitioning of cellular compartments. The analysis revealed near about 3446 DEGs that were up and down regulated between PGRT vs RC with FDR < 0.01 and log2FoldChange 1.5. Among these DEGs ~56% of them were up-regulated in RC of the REC GBM. The heatmap in Figure 43 shows the DEGs between PGRT of PT and RC of REC GBM.

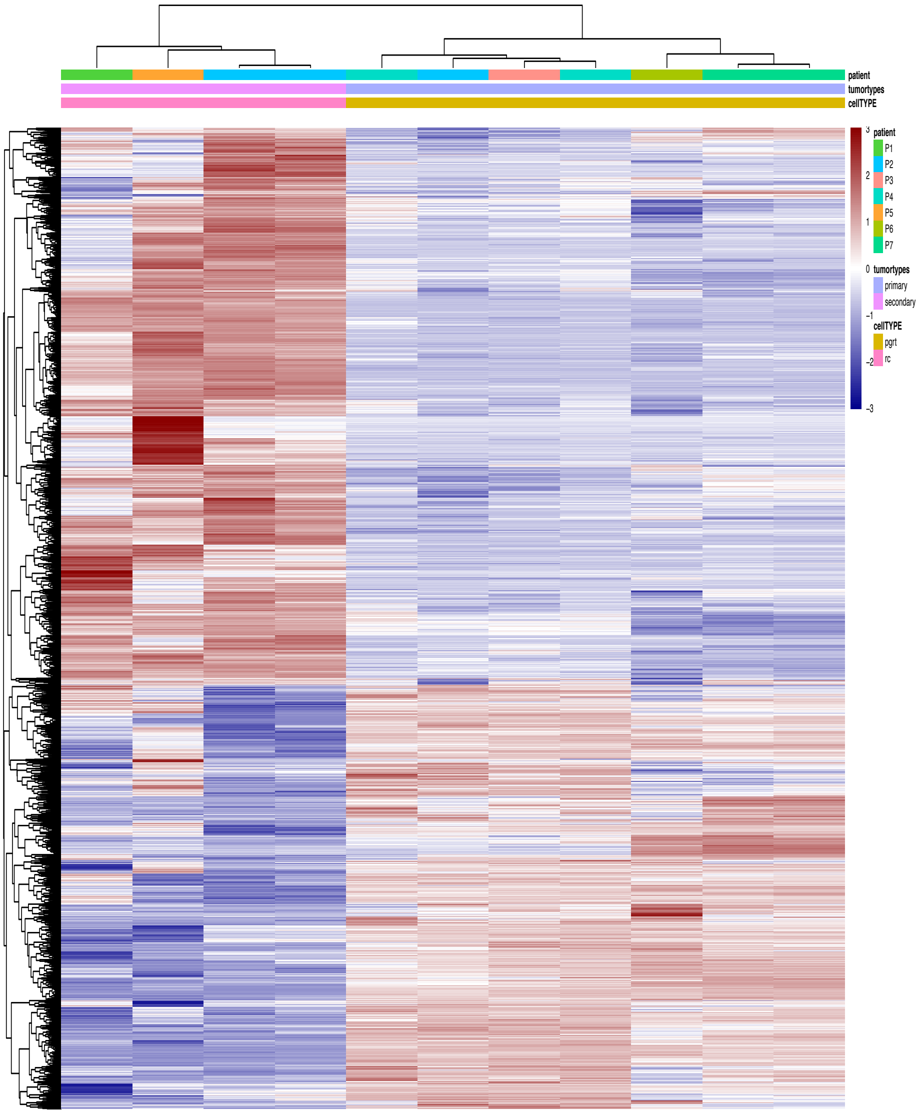


Figure 43: Heatmap representation of DEGs between PGRT of PT and RC of REC GBM

3.3.3.1 Canonical pathway analysis of DEGs between tumorigenic peripheries of PT and recurrent centers of RECG GBM reveals metabolic and growth pathways

IPA canonical pathway analysis of the overall DEGs is able to provide us the pathways that were activated or inactivated as a result of the changes in gene expression. Some of the most significant pathways that were altered are highlighted in figure 44.

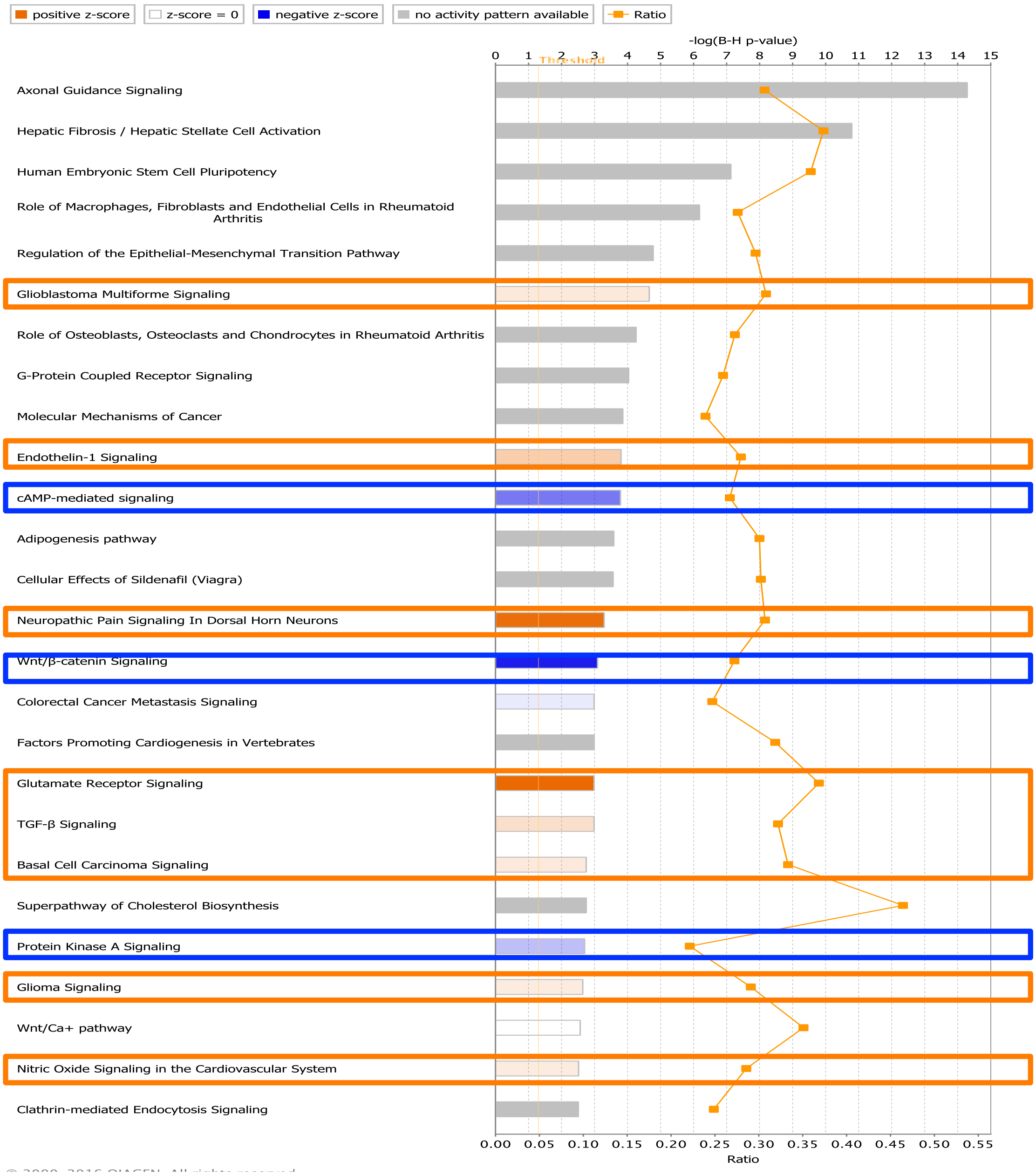


Figure 44: Representation of enriched pathways found by IPA as a result of changes in gene expression between PGRT of PT and RC of REC GBM

Orange blocks represent activation while blue block represents inactivation. The orange line represent the log (B-H) p-value with the orange dots reports the fraction of target molecules enriched for a particular pathway against all the molecules that builds it up.

The canonical pathway analysis predicted the activation of the following: GBM signaling, Neuropathic Pain Signaling in Dorsal Horn Neurons, Glutamate Receptor Signaling, TGF-Beta Signaling, Basal Cell Carcinoma. The pathways that got inactivated were: cAMP-mediated signaling, Wnt/Beta-catenin signaling and Protein-Kinase A signaling. This suggested that the relapse was due to activation of metabolic and growth-signaling pathways, which could have provided immunosuppression, proliferation and stemness to tumor cells.

In addition, I also performed an enrichment analysis for the gene expression signatures associated to the 4 molecular subtypes of GBM (Verhaak et al., 2010) in our DEGs between PGRT of PT and RC of REC GBM. I found 60 DEGs associated to classical, 71 associated with mesenchymal, 72 proneural and 29 Neural. All these enrichments with the signatures were statistically significant. The heatmap in the figure 45 shows the DEGs enriched for each of the 4 molecular subtypes.

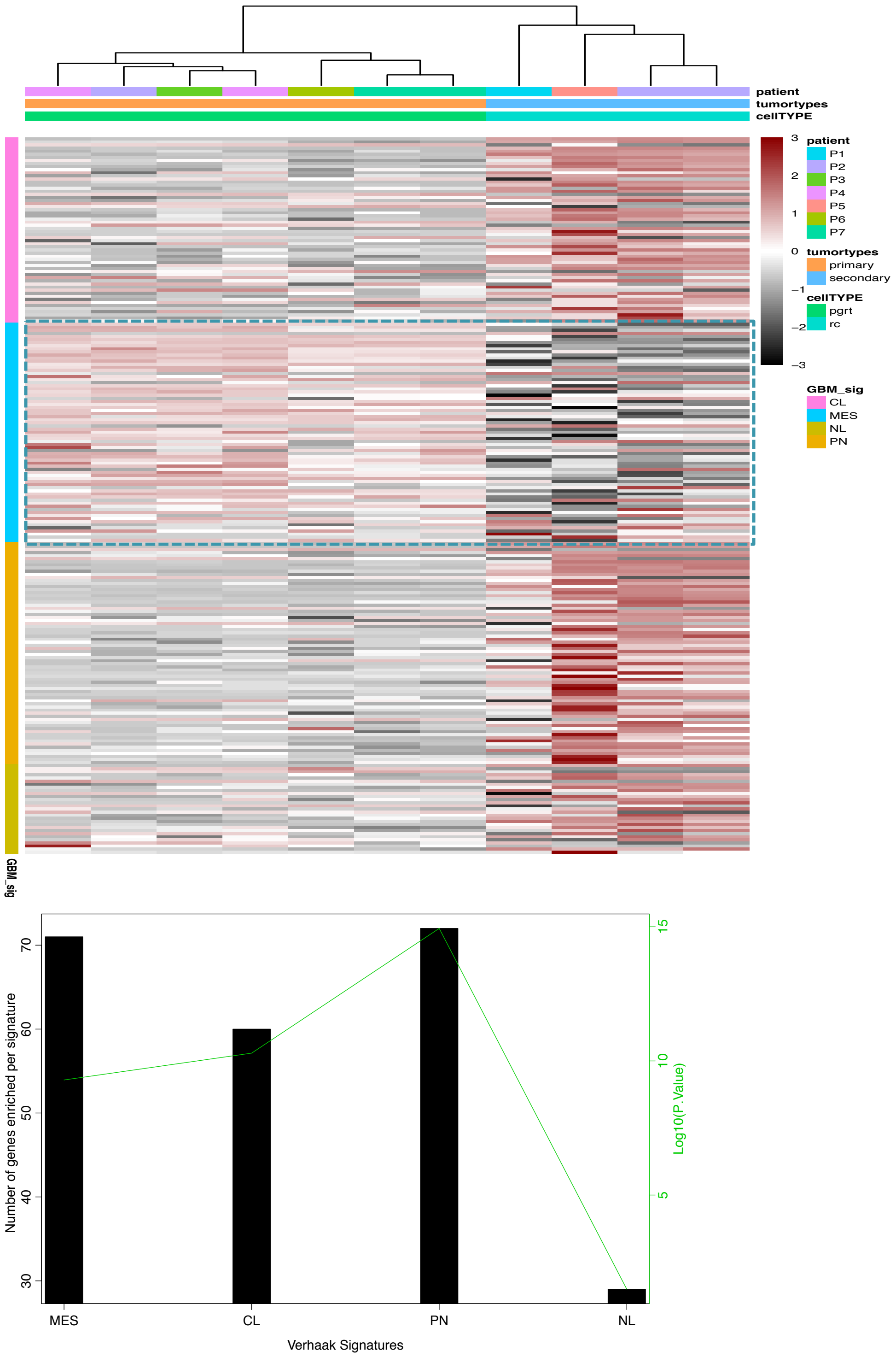
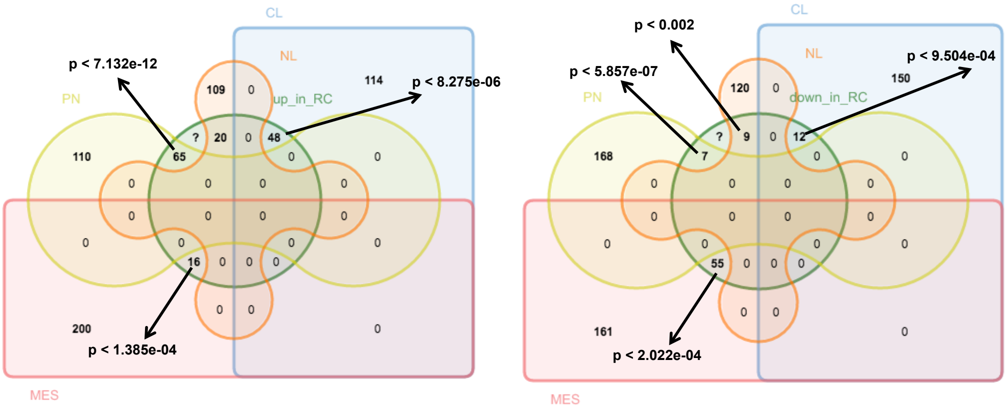


Figure 45: DEGs between PGRT of primary tumor (PT) and RC of REC GBM are enriched for Verhaaks' GBM signature genes

a) Heatmap of DEGs overlapping for specific Verhaak GBM signature between PGRT of PT and RC of REC GBM samples. b) Two way plot of the DEGs enriched for each signature. x-axes represents the 4 molecular subtypes while left y axes represents the number of genes enriched for each subtype and right y axis shows the –log10(p-value) enrichment score for those signatures. Enrichment was significant for all the 4 signatures.

Among these, only 55 genes up-regulated only in PGRT showed enrichment for mesenchymal signature with a p < 2.02e-04 (Figure 46 a right Venn diagram). This showed even at a deeper cellular compartment we find an enrichment of mesenchymal genes only up-regulated in PGRT of PT (conversely down-regulated in RC of REC GBM). There were some other interesting observation as well which revealed among these DEGs, specially those that were only up-regulated in the RC of the REC GBM had enrichment for mostly Proneural (65 genes with p < p < 7.132e-12) followed by Classical (48 genes p < 8.275e-06) as seen in Figure 46 a left Venn diagram. These genes there were then assessed separately upon their direction to find the significant enrichment of up and down genes in each of the signatures.

a)



b)

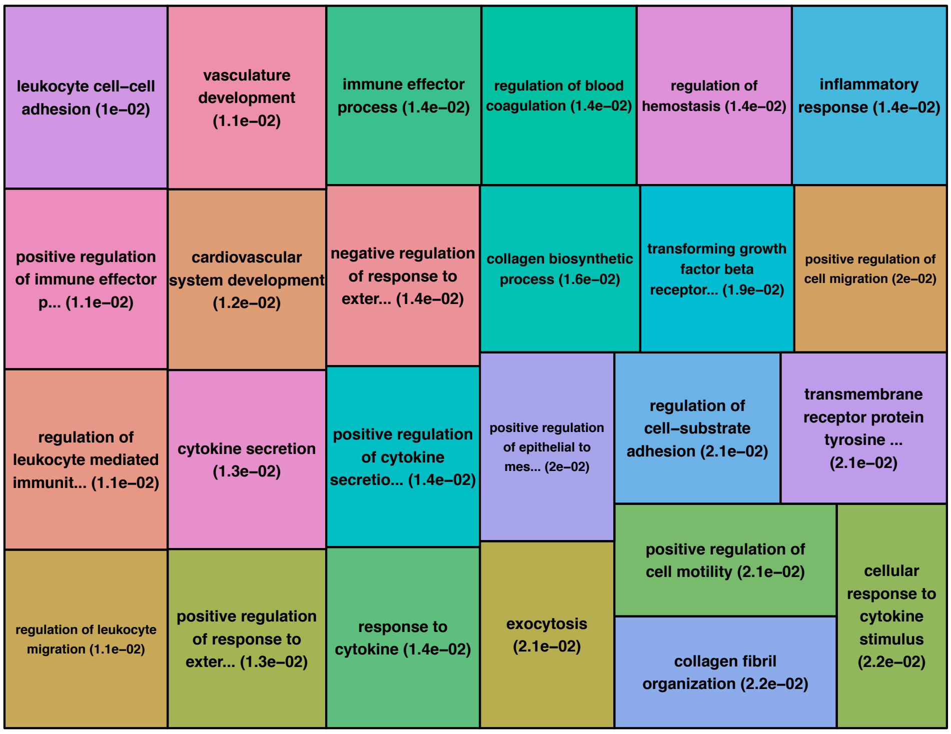


Figure 46: Enrichment of up and down-regulated genes in RC of REC GBM are enriched for Verhaaks' GBM molecular subtype signature genes

a) Venn-diagram of DEGs overlapping for specific Verhaak GBM signature between PGRT of PT and RC of REC GBM samples based on directionality. b) Only genes down-regulated in RC of REC GBM (i.e. up-regulated in PGRT) showing significant enrichment for mesenchymal's signature also showed significant enrichment for GO categories specifically for Biological Processes.

Venn diagram 46 a) represents the significance of the DEGs enriched for each of the 4 molecular subtypes defined by Verhakk et al. when assessed separately based upon their direction of regulation. In Figure 46a) the left Venn diagram represents the overlap of Verhaak signatures with DEGs up-regulated in REC GBM while the right represents the DEGs down-regulated in REC GBM. Among the DEGs that were down-regulated in REC GBM (conversely up-regulated in PGRT) even showed enrichment for gene-ontology categories for biological process that were associated with mesenchymal signatures. The GO figure 46 b) represents the various biological processes that were involved with mesenchymal signatures (up-regulated in PGRT of PT). Processes like response to transforming growth factor, cellular response to transforming growth factor, response to cytokine, negative regulation to immune system and various migratory and adhesion process were seen to be enriched for these 45 DEGs that were up-regulated in PGRT of PT and significantly enriched for mesenchymal signatures.

3.3.3.2 DEGs between PGRT versus RC of REC GBM involved a set of human TFs that classified their transcriptomes

I was interested to understand what are the TFs involved in these DEGs and to see their contribution in the GBM relapse. Upon inspection I found among 3446 DEGs, there were 261 human TFs that were able to classify the transcriptomic differences between the PGRT and the RC. Thus I was able to highlight the differences between primary and relapsed tumor at a more microscopic level of cellular partitioning that provided a deeper resolution to TFs mediated tumorigenic progression. This was obtained by interrogating the DEGs with the human TFs annotated in the AnimalTFDB database. Barplot figure 47 shows the overall TF-DEGs between PGRT vs RC.

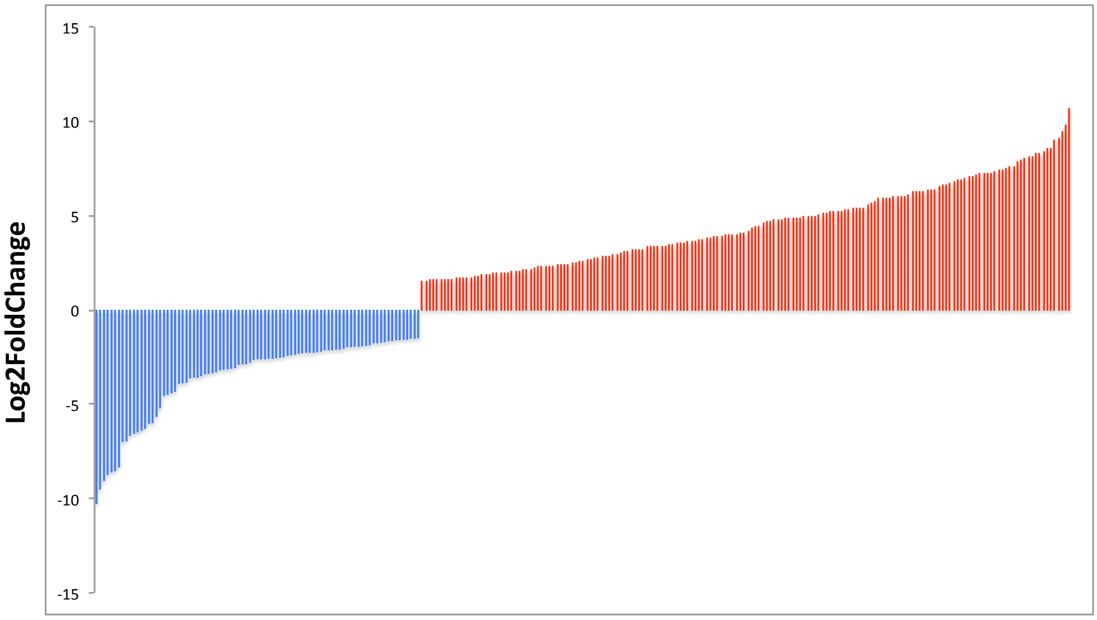


Figure 47: Barplot representing TFs as a subset of DEGs between PGRT of PT and RC of REC GBM

From this analysis a lot of TFs as DEGs in this comparison I was interested in understanding the pathways that were involved as a result of these 261 DEG TFs with IPA between PGRT of primary tumor and RC of REC GBM. Some of these pathways, that are associated primarily to cancer progression and maintenance are seen up-regulated in RC GBM samples which identified molecular events triggered during the relapse that could characterize the core of a recurrence tumor GBM. The below figure shows the pathways that were triggered as a result of the TF-DEGs, some which were activated and inactivated as predicted by IPA.

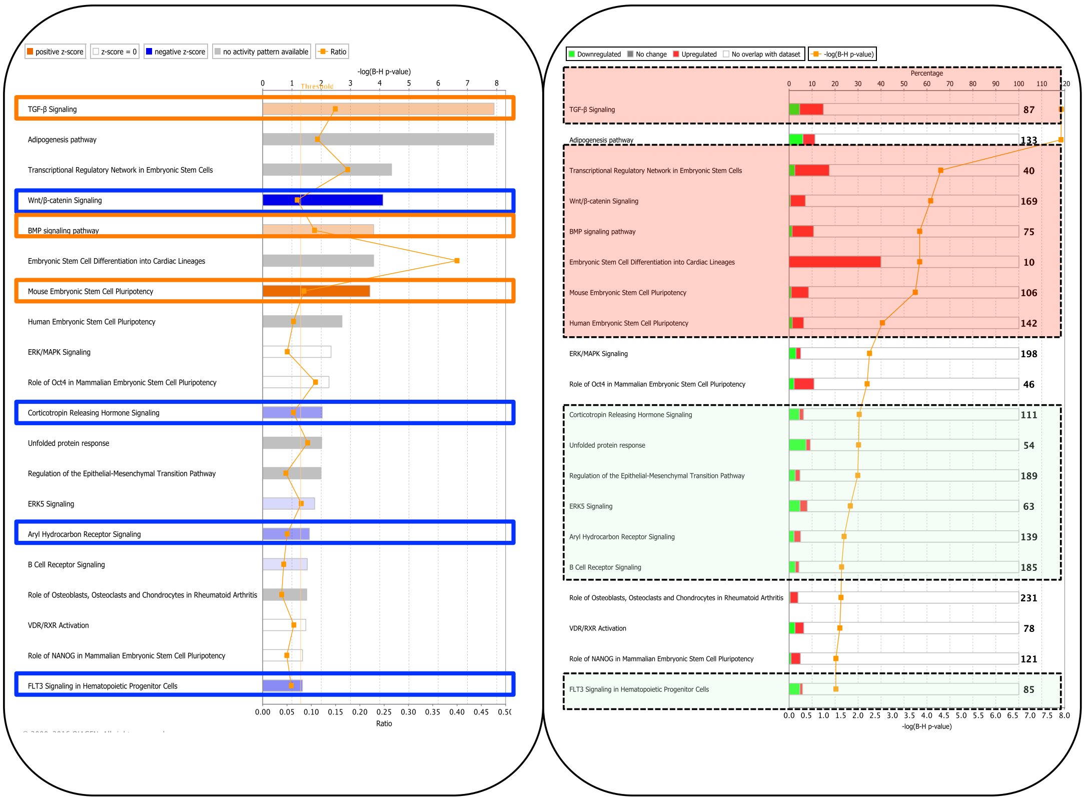


Figure 48: A two-way representation of the pathways found by IPA canonical pathway analysis with TF DEGs between PGRT of PT and RC of REC GBM

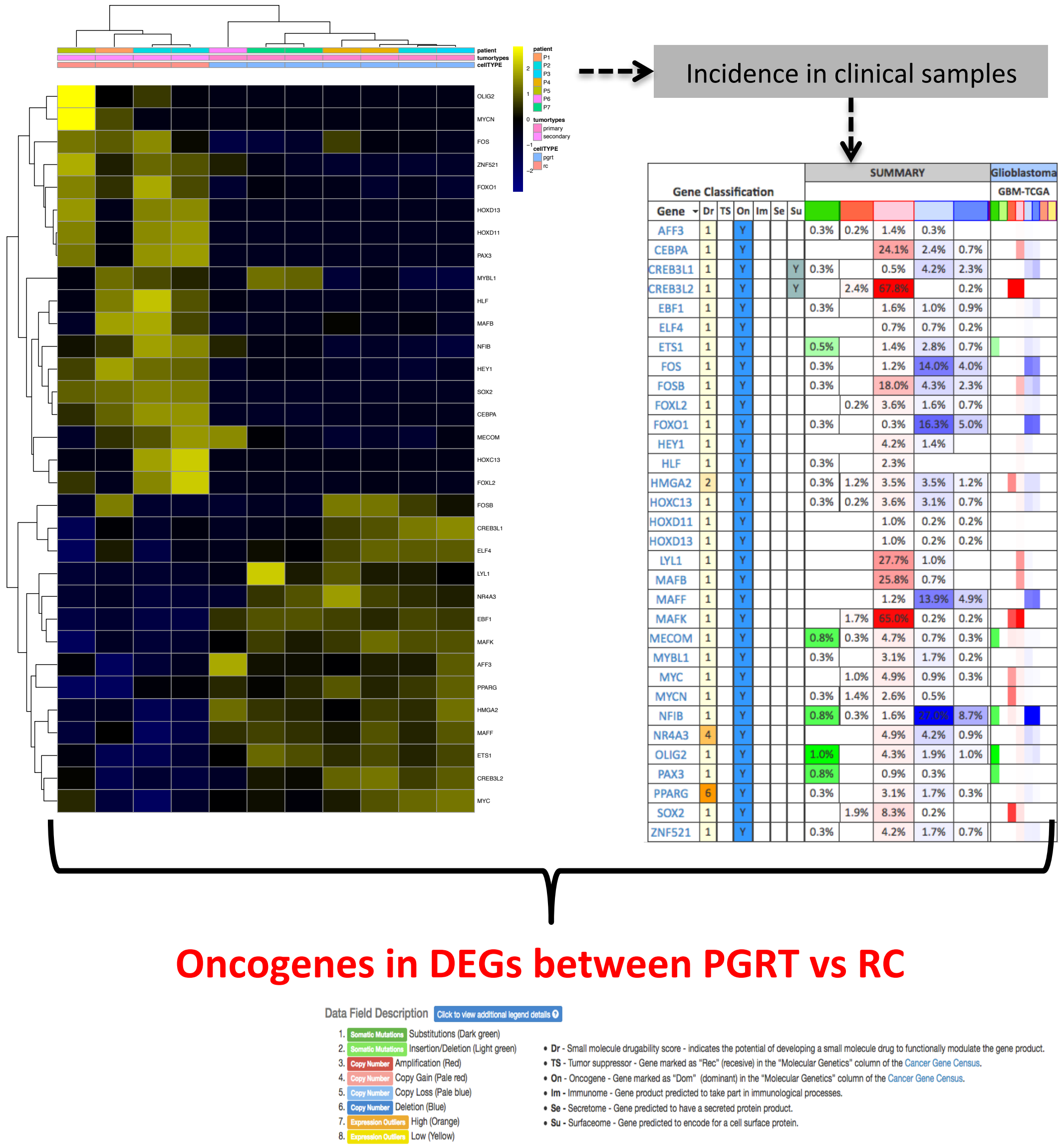
The left panel shows the pathways that were predictively activated and inactivated by IPA while the right panel highlights the contribution of direction of genes in those pathways.

Interestingly TGF-Beta signaling was seen to be activated in the RC of the REC GBM where it was largely dominated by genes up-regulated in RC of REC samples along with BMP signaling pathway and Mouse Embryonic Stem cell pluripotency pathway. This characterizes that REC GBM tumor centers were having predictively activated pathways associated with over expression of growth-factors, genes involved in pluripotency and signal transduction. This pluripotency pathway that was activated due to up-regulated genes in the RC of the relapsed GBM samples thus highlighted the importance of aberrant pluripotent factors that were relevant in tumor drive constituting a transitory phase from primary to relapse. There were some major pathways that were inactivated in the RC of the REC GBM like Wnt/Beta-Catenin, Corticotrophin releasing hormone signaling, ERK5 signaling and FLT3 signaling in hematopoietic progenitor cells. Majority of these pathways inactivated in RC of REC GBM were predominantly having genes down-regulated in the RC of the REC GBM apart from Wnt/Beta-Catenin that was inactivated due to up-regulated genes in the RC of the REC GBM

EMT pathway was also seen to be triggered due to change in gene expression between PGRT of PT and RC of REC GBM, however IPA did not predict its effect. This was possibly due to lack of annotated information of this pathway in IPA knowledge base that associates pathways from genes and then find the association of these pathways to activation and inactivation based on z-score calculation of genes direction from input gene list versus the direction annotated in their database. However it seemed to be enriched and dominated by genes down regulated in the RC of the REC GBM.

3.3.3.3 Interrogation of TF DEGs in OASIS-Cancer genomics reveals oncogenes association and their dysregulation in clinical TCGA samples

These TFs upon farther interrogation with the clinical TCGA samples (n=577) with oasis-cancer genomics portal revealed certain key oncogenes among the cohort. These genes scored their relevance in tumorigenic recruitment, progression and maintenance. The heatmap in Figure 49 shows the oncogenes between PGRT of PT and RC of REC GBM. The right panel lists oncogenes from these TF DEGs that were found to be dysregulated in TCGA clinical GBM samples. This analysis with TCGA samples of GBM was performed with the help of oasis-cancer genomics web platform to identify the oncogenes from the TF DEGs. Farther these genes were assessed for level of dysregulation at different feature of point mutation, indels, amplification, deletion, copy-gain, copy-loss, and over/under expression.



Gene Classification							SUMMARY					Glioblastoma
												GBM-TCGA
Gene	Dr	TS	On	Im	Se	Su	🟩	🟥	🟥	🟦	🟦	
AFF3	1		Y				0.3%	0.2%	1.4%	0.3%		
CEBPA	1		Y						24.1%	2.4%	0.7%	
CREB3L1	1		Y			Y	0.3%		0.5%	4.2%	2.3%	
CREB3L2	1		Y			Y		2.4%	67.8%		0.2%	
EBF1	1		Y				0.3%		1.6%	1.0%	0.9%	
ELF4	1		Y						0.7%	0.7%	0.2%	
ETS1	1		Y				0.5%		1.4%	2.8%	0.7%	
FOS	1		Y				0.3%		1.2%	14.0%	4.0%	
FOSB	1		Y				0.3%		18.0%	4.3%	2.3%	
FOXL2	1		Y					0.2%	3.6%	1.6%	0.7%	
FOXO1	1		Y				0.3%		0.3%	16.3%	5.0%	
HEY1	1		Y						4.2%	1.4%		
HLF	1		Y				0.3%		2.3%			
HMGA2	2		Y				0.3%	1.2%	3.5%	3.5%	1.2%	
HOXC13	1		Y				0.3%	0.2%	3.6%	3.1%	0.7%	
HOXD11	1		Y						1.0%	0.2%	0.2%	
HOXD13	1		Y						1.0%	0.2%	0.2%	
LYL1	1		Y						27.7%	1.0%		
MAFB	1		Y						25.8%	0.7%		
MAFF	1		Y						1.2%	13.9%	4.9%	
MAFK	1		Y					1.7%	65.0%	0.2%	0.2%	
MECOM	1		Y				0.8%	0.3%	4.7%	0.7%	0.3%	
MYBL1	1		Y				0.3%		3.1%	1.7%	0.2%	
MYC	1		Y					1.0%	4.9%	0.9%	0.3%	
MYCN	1		Y				0.3%	1.4%	2.6%	0.5%		
NFIB	1		Y				0.8%	0.3%	1.6%	27.0%	8.7%	
NR4A3	4		Y						4.9%	4.2%	0.9%	
OLIG2	1		Y				1.0%		4.3%	1.9%	1.0%	
PAX3	1		Y				0.8%		0.9%	0.3%		
PPARG	6		Y				0.3%		3.1%	1.7%	0.3%	
SOX2	1		Y					1.9%	8.3%	0.2%		
ZNF521	1		Y				0.3%		4.2%	1.7%	0.7%	

Oncogenes in DEGs between PGRT vs RC

Data Field Description Click to view additional legend details

1. Somatic Mutations Substitutions (Dark green)
2. Somatic Mutations Insertion/Deletion (Light green)
3. Copy Number Amplification (Red)
4. Copy Number Copy Gain (Pale red)
5. Copy Number Copy Loss (Pale blue)
6. Copy Number Deletion (Blue)
7. Expression Outliers High (Orange)
8. Expression Outliers Low (Yellow)

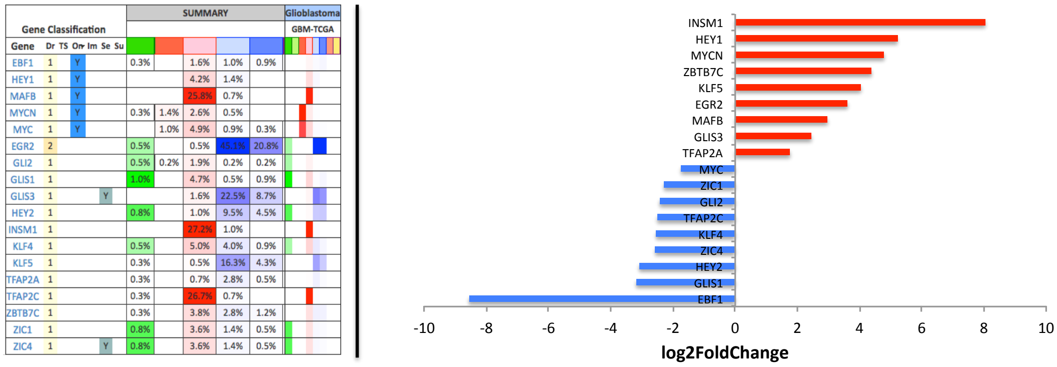
- **Dr** - Small molecule drugability score - indicates the potential of developing a small molecule drug to functionally modulate the gene product.
- **TS** - Tumor suppressor - Gene marked as "Rec" (recesive) in the "Molecular Genetics" column of the Cancer Gene Census.
- **On** - Oncogene - Gene marked as "Dom" (dominant) in the "Molecular Genetics" column of the Cancer Gene Census.
- **Im** - Immunome - Gene product predicted to take part in immunological processes.
- **Se** - Secretome - Gene predicted to have a secreted protein product.
- **Su** - Surfaceome - Gene predicted to encode for a cell surface protein.

Figure 49: DEGs include a set of oncogenes encoding for TFs dysregulated in clinical GBM TCGA samples

The right heatmap in Figure 49 revealed CREB3L2 and MAFK dysregulated in over ~60% of TCGA samples at low-level of copy gain highlighting their relevance in GBM and it was also previously found as key oncogenic TF that was differentially expressed between PT vs REC GBM samples as well. This highlighted the importance of these 2 genes in GBM tumorigenesis and more specifically in relapse.

3.3.3.4 Motif enrichment analysis predicted a core set of oncogenes and tumor suppressor genes encoding for TFs controlling the transcriptional programs in recurrent centers of REC GBM

Post defining the transcriptomic changes underlying the GBM recurrent centers, I interrogated the DE genes between tumorigenic peripheries of PT (PGRT) and recurrent centers (RC) of REC GBM by performing the TF motif enrichment analysis (MEA) as done with the earlier comparisons. This analysis primarily aimed at providing a mechanistic interpretation of the transcriptomic changes in recurrent centers (RC) of REC GBM by identifying putative TFs master regulators (MRs) through interrogation of binding sties at the promoter of the DEGs. Motif analysis was performed on promoter sequences of these overall 3446 DEGs obtained as a result of DE analysis between PGRT of PT and RC of REC GBM with Pscan with default settings. This revealed 79 (p-value threshold < 0.05) upstream TFs that have consensus binding sites in our DEGs among which 18 over-represented TFs were also differentially expressed between PGRT of PT and RC of REC GBM. I checked the basis of these 18 genes in clinical GBM TCGA samples to understand their level of dysregulation and to retrieve information of them regarding their status of being an oncogene or a tumor suppressor. Figure below represents the incidence of these direct TF targets (which are not only upstream of DEGs but themselves differentially expressed) in clinical TCGA GBM samples along with their direction of change in our sample cohort that I analyzed. Figure 50 shows the incidence of 18 of these targets in GBM TCGA clinical samples and confers the status of Oncogene/TSG to them.



Gene Classification							SUMMARY					Glioblastoma
												GBM-TCGA
Gene	Dr	TS	On	Im	Se	Su						
EBF1	1		Y				0.3%		1.6%	1.0%	0.9%	
HEY1	1		Y						4.2%	1.4%		
MAFB	1		Y						25.8%	0.7%		
MYCN	1		Y				0.3%	1.4%	2.6%	0.5%		
MYC	1		Y					1.0%	4.9%	0.9%	0.3%	
EGR2	2						0.5%		0.5%	45.1%	20.8%	
GLI2	1						0.5%	0.2%	1.9%	0.2%	0.2%	
GLIS1	1						1.0%		4.7%	0.5%	0.9%	
GLIS3	1				Y				1.6%	22.5%	8.7%	
HEY2	1						0.8%		1.0%	9.5%	4.5%	
INSM1	1								27.2%	1.0%		
KLF4	1						0.5%		5.0%	4.0%	0.9%	
KLF5	1						0.3%		0.5%	16.3%	4.3%	
TFAP2A	1						0.3%		0.7%	2.8%	0.5%	
TFAP2C	1						0.3%		26.7%	0.7%		
ZBTB7C	1						0.3%		3.8%	2.8%	1.2%	
ZIC1	1						0.8%		3.6%	1.4%	0.5%	
ZIC4	1				Y		0.8%		3.6%	1.4%	0.5%	

Figure 50: The heatmap represents the dysregulation of over-represented TFs that are DEGs in clinical GBM TCGA

The left panel shows the list of over-represented TF targets found from MEA analysis that are differentially expressed in lab GBM datasets between PGRT and RC. The heatmap represents their dysregulation in clinical GBM TCGA while right barplot depicts direction of these over-represented TFs in RC of REC GBM.

a) Represent 18 direct targets of upstream TFs in our DEGs and their level of dysregulation in clinical GBM TCGA samples.

b) Represents the log2foldchange of these 18 direct TFs in RC of REC GBM.

Among these 18 direct-TF targets there were 5 oncogenes retrieved from this analysis. KLF4 was not assigned as tumor-suppressor as the tool used v70 of Cancer gene census to confer this status but according to the latest v78 it is annotated as an oncogene/TSG. The right panel shows the direction of regulation of these 18 direct targets in our samples where red represents its up-regulated in RC of REC GBM while blue represents it is down-regulated.

This analysis could identify direct target-TFs (over-represented by MEA analysis and differentially expressed). Some these were oncogenes that were up-regulated (MYCN, MAFB, HEY1) in RC of the relapsed GBM samples while 2 of them are down (EBF1 and MYC). Role of EGR2 had been associated both as oncogenic and tumor suppressing. This TF was seen to be over-represented and differentially expressed in this comparison, rather up-regulated in RC of REC GBM. Earlier as mentioned it was also seen in the comparison of PGRT and PDGRT as a TF that was differentially expressed. Since EGR2 is associated with immune-modulators it could be highlighting the fact that the mechanisms relating to negative immunoresponse might trigger a niche population of cells in PGRT to be immunosuppressed. These farther might acquire cancer properties and cooperatively with other growth factors possibly give way to relapse. These predictive targets could be putative master regulator candidates. These candidates could be self sufficient in controlling the regulation of other genes and contribute to tumor progression and invasiveness giving rise to relapse. This analysis also revealed down regulation of KLF4 gene (a direct target of upstream TF) that is usually considered as a TSG in renal cell carcinoma (Li et al., 2013). Hypermethylation of KLF4 promoter is seen in renal cell carcinoma (Li et al., 2013) which indicates that this can be epigenetically guided factor. This scores the importance of epigenetic drive that might be responsible for the down regulation of this target in the recurrent tumor. KLF4 was also seen to have context-dependent oncogene and tumor-suppressor functions in carcinomas (Rowland, Bernards, & Peeper, 2005) .

Thus it would be really important to center on these direct target-TFs and try to find the proportion of DEGs that are being regulated between PGRT of PT and RC of the REC GBM by them. Finally it will be good to cross-validate the TFs with other tools or even with TF that are a result of ChIP-ChIP. Even one can score the precedence of the most important TF by

a) Simply taking into account the proportion of DEGs each TF is controlling,

b) Direction of DEGs controlled by each of these TF,

c) Distribution of the fold changes and

d) Finally extracting the most relevant pathways or the biological processes that gets enriched as a result of these DEGs regulated by each of the upstream TFs.

Thus we will be able to assign a key regulators which might be relevant for validation with *in vitro* experiments to father assess their nature to drive or reduce the tumorigenicity in GBM relapse or even the primary GBM. These can also act as candidate prognostic markers that characterize the aggressive behavior of the relapsed tumor to that of the primary whose targeted over/under-expression in relevant tumor types might lead to reduce tumorigenicity.

The analysis between primary and relapsed tumor samples were able to find the key processes that were involved as a result of transcriptional change between a primary GBM and its relapse. This analysis was more specifically centered on core TF mediated network that was mediating this transition.

The analysis between the tumorigenic (PGRT) and non-tumorigenic (PDGRT) peripheries highlighted the TF BNC2 up-regulated in tumorigenic periphery or PT along with a growth factor gene PDGFRA that indicate the how a tumor micro-environment was harbored that involved tumor-migration and invasion of these cells to neighboring cells leading to a relapse. It also revealed EGR2, a TF that was over-represented by MEA and differentially expressed between PGRT and PDGRT.

Finally a topological assessment of a tumorigenic periphery of PT versus the RC of REC GBM elucidated the key transcriptional programs that were mediating the tumor progression. This analysis also revealed key oncogenes and TSGs. It will be interesting now to assess the strength of these TFs networks by studying the underlying biological processes they are involved in. Thus we could score the most relevant candidate master regulator from them that could be crucial for the transition of primary tumor cells to GBM relapse.

Thus the studies in the GBM patients were able to highlight the core transcriptional programs involved in transition of the primary tumor to a relapsed stage. It farther revealed a set of oncogenes/TSG that were involved among these direct-TF targets. These could in turn indicate the tumorigenic drive that allowed specific cellular compartments to acquire a more aggressive or invasive tumor properties thus paving a way for relapse.

Discussion

The work presented in this thesis improved our understanding on the TF-mediated cancer progression through the dissection of the transcriptional networks underlying two of the most aggressive disease with high mortality and poor prognosis: and i) High Grade Serous Ovarian Caner (HGSOC) and ii) Glioblastoma Multiforme (GBM)). Despite the multimodal treatment, GBM and HGSOC patients have a very poor prognosis (Bowtell, 2010) , (Frattini et al.,

2013) . This is due to the fact the molecular mechanisms underlying the initiation and the progression of these tumors are still poorly understood.

The first part of the thesis involves study of HGSOC. In HGSOC, a major problem resides on the uncertainty on the cellular origin of this tumor. The molecular events thus associated with the tumor initiation and progressions specific to its origin have not yet been studied thoroughly. Since the transcriptional programs (gene regulatory networks involving MRs or upstream TFs) associated with tumor and its specific origin counterparts are not much explored. This limits the identification of new molecular pathways, key TF hubs that drive the tumor evolution and serves as prognostic signatures or rationale therapeutic targets for tumors associated specific origin.

In studying HGSOC, I have used the integration of transcriptomic and genomic analysis. The transcriptomic part of the HGSOC study dealt with transcriptomic assessment of the high-grade OC tumors vis-à-vis two-candidate tissues of origin. This was followed by MEA analysis on the promoters of the DEGs for specific comparisons to find upstream TF's that were also differentially expressed (DE). Post MEA analysis, I found which DE TFs were also Oncogene/TSG and their dysregulation in clinical TCGA samples. This is indeed an important finding that helped us to outline the key oncogenic/tumor suppressing TF mediated networks implicated in malignant drive. Future assessment of these TF networks will enable us to outline molecular events relevant for tumor initiation and vis-à-vis two candidate tissues of origin. Such findings can be pathologically relevant as upon perturbations of these key regulators *in vitro* will enable us to record their role in tumor development. Role of TFs in oncogenic drive has been already studied in few cancer types. Over/under-expression of these TFs a.k.a candidate MRs is catalytic in induction of downstream pathways that either promotes or reduces tumor development. In light of this in our OC samples I was able to find such candidate MRs while assessing the overall transcriptomic differences between tumors vis-à-vis two candidate tissues of origin.

The transcriptomic assessment between all the normal samples and tumor samples irrespective of the tissue origin consideration revealed key upstream TF's a.k.a candidate MRs that were differentially expressed. We found ASCL2 and NR2F1 as candidate MRs, which were up-regulated in tumors. Among these MRs, ASCL2 is a key transcriptional regulator reported in colon cancer studies (Tian et al., 2014). We also obtained the gene NR2F1, an MR associated with cellular dormancy. Shutting down of this gene leads to tumor development via aberrant growth of dormant cells throughout the tumors (Sosa et al., 2015).

Transcriptomic assessment of FI normal samples against all tumors revealed 5 candidate MRs that were also DEGs. Two oncogenes were present in this list of 5 candidates MR's where EBF1 was up regulated in tumor while ERG was down-regulated. Other 3 genes were EGR2 (associated with ~10% copy gain in Ovarian TCGA samples), EGR3 (associated with 34.5% copy loss in TCGA OC samples) and TFAP2A (primarily associated with 32.3% copy gain in TCGA OC samples).

Variations in forms of mutation or INDELS in EBF1 have been observed in cancers such as intestinal cancer, skin cancer, and stomach cancer. Comparing OSE normal against all tumors did not give any candidate MRs that was differentially expressed. We also stratified tumors based on DNA methylation to their cellular origin and then compared the tumor transcriptome to their corresponding origin (normal). DNA-methylation served as developmental tracer for tumors and provided us with origin of tissue for the tumors and thus we classified FI-like tumors (coming from normal FI) and OSE-like tumors (coming from OSE normal). This was followed by transcriptomic study of both newly classified FI vs FI-like tumors and OSE vs OSE-like tumors. This study outlined specific transcriptional programs associated to tumor development from its specific tissues of origin defined via DNA methylation signatures (DMS). DEA and MEA in FI vs FI-like tumors revealed these 3 up-regulated MRs: EHF, TFAP2A and ZIC1. TFAP2A and ZIC1 are associated with copy gain of ~32.3% ~45.9% respectively in OC clinical TCGA samples.

In OSE vs OSE-like tumors we were able to find a candidate MR post comparing specific tumors to its corresponding normal. We could not do it in the earlier comparison when we had no stratification and compared all tumors to OSE normal tissues. We found EGR1 as a candidate MR that is up-regulated in OSE-like tumors. This gene is usually found to be associated with ~9.3 % copy gain in clinical OC TCGA samples. It has been linked as a TSG in gastric and colorectal cancers which when mutated promotes tumor development (Choi et al., 2016). Also in NSCLC it has been reported with tumor suppressor properties (H. Zhang et al., 2014). However, in prostrate caner its been shown to be over-expressed in tumors (Parra Villegas et al., 2011), (Gregg & Fraizer, 2011). It is often associated with multiple functions that can both promote or decrease tumor activity.

Finally studying the transcriptional programs between two tumoral tissues (EOC and AS) in our dataset helped in understanding the role of inflammatory and immune signals that are activated from EOC to AS.

Thus key candidates MRs along with key molecular events orchestrating tumorigenic events were obtained as a result of this study. This MRs on careful assessment for network analysis and biological activity post perturbation can lead to new therapeutic targets.

The final study was to understand the preservation of genetic background between tumor and its reprogrammed derivate. The results obtained in this work shows that employing WES analysis on OC data and their derived OC-iPSC with a 3-tier computational approach enabled to establish that the reprogramming counterparts were indeed coming from tumors and not their normal. This was confirmed, as tumor aberrations that build up the genetic background of the parental tumor can be tracked as well in tumor-iPSCs. OC specific lesions were fractionally tractable in OC-iPSCs. This computational workflow showed both at the SNV and CNV level, OC-iPSC shared a fraction of the parental tumor genetic aberrations more specifically by layers of copy number alterations. Moreover driver mutations were shared between OC tumor and OC-iPSCs of the low grade establishing the fact that reprogrammed clones were indeed tumor derived. The somatic SNV shared only a fraction but this could be attributed to that fact that the sequencing coverage was not deep enough to find larger fractions of genetic aberrations. Another hypothesis can be derived that reprogramming itself might have induced some mutational burden in the iPSCs that over time created some selective pressure. This selective pressure made us only to retrieve partially the somatic SNVs that were shared between the parental tumor and the iPSCs derived from them.

IPSCs are single clones owing to their generation from a single cell in culture. They should in theory have higher resolution of the representative genetic composition of tumor subclones. A tumor is on the other hand is a heterogeneous bulk comprising of polyclonal mutations that confers it a heterogeneous mass. In order to simplify the tumor genetic heterogeneity it would be ideal to perform in-depth analysis of mutations and CNVs (gain and loss of copy number is genes) on a larger cohort of tumor iPSCs per tumor to reconstruct the genetic landscape of the subclones in parental tumor. This would in addition pave way to track the tumor evolution in either of the two ways. First by outlining the fact that mutations are less represented in parent tumor if their derivative iPSCs being clonal have consistent increase in the genetic alteration frequencies from tumor to iPSCs. Second, if these frequencies remain unaltered it would be indicative of the fact that the mutations represent the vast majority of the tumoral cells. This in turn would capture the early events in tumor pathogenesis. Finally iPSC clonality in theory would also help in assignment of defined haplotypes to tumor subclones, thus reconstructing the genetic evolution of the tumor in study.

In the case of the GBM I inferred a TF-mediated model based on a core set of TFs controlling the transcriptional programs underlying GBM progression from primary to relapsed tumor. In particular I found interesting hits of TFs that were also differentially expressed in comparisons of Primary tumor versus REC GBM and in the tumorigenic peripheries of PT versus the recurrent centers of the REC GBM. One interesting observation was the enrichment of DEGs (PT versus REC GBM and PGRT of PT versus RC of REC GBM) with mesenchymal signature that was only up-regulated in PT. Among these over-represented TFs found by MEA analysis, which also happened to be DEGs between PT and REC GBM, I found an oncogene EBF1 and tumor suppressor gene KLF4 to be down regulated in REC GBM. Role of KLF4 has already been observed as both oncogene and tumor suppressor genes in few cancer types scoring its importance (Rowland et al.,

2005) , (Wang et al., 2015) . Particularly the suppression of KLF4 in renal cell carcinoma has been attributed to the fact that it is epigenetically silenced by the hypermethylation of the CpG promoters that leads to its suppression. *In vitro* assay reports have stated that overexpressing KLF4 led to inhibition of renal cancer cells migration and also suppressed EMT pathways. Cells associated with EMT are often regarded as having aggressive and invasive properties. While *in vivo* assay reports have shown to inhibit the oncogenic progression and put on hold metastasis in renal carcinoma by ectopic expression of KLF4 (Li et al., 2013). This underlines the role of DNA hypermethylation in epigenetically inhibiting the expression of KLF4 giving us an epigenetic target. So we can hypothesize this gene KLF4 to be one of the driving factors in GBM relapse. It will be particularly relevant to perform DNA methylation analysis on these samples to understand if this suppression in our GBM patients were carried out due to methylated CpG sites at promoters or any other epigenetic factors. Also perturbing the expression of the KLF4 in cell cultures followed by RNA-Seq can lead us to understand the tumor turnover if there is any. Role of EBF family genes as transcription factors have already been associated in GBMs as a tumor suppressor (Liao, 2009) . Particularly EBF1 has been to be undergoing loss of genomic regions or associated with somatic mutation (Liao, 2009). Since EBF1 is crucial for neuronal differentiation process therefore its inactivation could put on hold the developmental processes of the normal cells while that could lead to promote oncogenic environment. So the down-regulation of this gene in our GBM recurrent tumor might be due to an upstream mutation or deletion. This could lead to an impact of the secondary tumorigenic drive leading to a relapse. I was also able to characterize the transcriptomic differences associated with tumorigenic and non-tumorigenic peripheries in the PT and able to highlight exclusive genes differentially expressed in primary tumorigenic peripheries that were able to classify the primary tumorigenic peripheries from the centers of the primary tumor. In doing so I was able to find a transcription factor BNC2 and an oncogene PDGFRA up-regulated in only tumorigenic periphery. This gives an idea of key genes whose up-regulation could be crucial in maintaining the tumor environment in these topographical sections of the brain. We can hypothesize that the tumorigenic periphery compartments prepares the seating of oncogenes and other growth-factor genes that initiates a transitory processes leading to relapse. These processes might also induce a site for the initiation of recurrence by conferring the neighboring cells with tumor infiltrating cells leading to invasiveness and tumor propagation. The final transcriptomic comparison was carried out between tumorigenic peripheries of PT (PGRT) and recurrent centers (RC) of REC GBM. This comparison also revealed a fraction of DEGs belonged to mesenchymal signatures that were associated with only up-regulation in PGRT. This observation was pretty consistent with the DEGs of PT versus REC GBM where mesenchymal genes were enriched in DEGs that were only up-regulated in the PT samples. Pathway analysis of revealed key activation of pathways like TGF-beta signaling and basal cell carcinoma signaling in the recurrent centers and inactivation of Wnt/beta catenin signaling, protein kinase signaling and cAMP mediated signaling. This is of relevant importance which highlights relapse was mediated via metabolic, growth signaling pathways and immunosuppression. It can also be hypothesized that these afore-mentioned pathways provided stemness and led to tumor cell proliferation along with immunosuppression. These observations that I was able to record were pretty insightful in giving a first hand print of the molecular mechanisms involved in relapse of GBM. Finally with motif analysis I found few targets of upstream TFs in our DEGs that were differentially expressed and termed as candidate MRs. These upstream TFs were also oncogenes that were up-regulated in RC of the REC GBM while some of them were down regulated. EBF1 and KLF4 was again observed to be down regulated TF in the RC of REC GBM scoring their importance of their recruitment in tumorigenic transition from primary to recurrence GBM. Thus I constructed a TF mediated network that was able to partition specific events associated with GBM evolution to its relapse. Finally upon *in vitro* targeting these candidate MRs will be important to see the effect of regulation of their gene targets and the molecular mechanisms that gets triggered as a result of this perturbation. This system of perturbation might provide meaningful insights in finding new novel therapeutic insights. It will also be relevant to perform ChIP-Seq of specific histone marks associated with promoter repression or activation. This ChIP-Seq of specific histone marks would find the chromatin states associated with recurrence. Upon integration of ChIP-Seq results with RNA-Seq we can also find epigenetically driven candidate MRs. Thus, we can associate the contribution of epigenetic drive orchestrated by histone modifiers targeting the candidate MRs promoting relapse. Thus opening up a new angle of epigenetic recruitment of repressor or activators by histone modifiers mediating regulatory networks in GBM leading to its recurrence

In conclusion, here we advance significantly our understanding of TFs-mediated contribution to cancer progression. This TF-model approach was already successfully applied in my host lab to study the transcriptional networks controlled by Polycomb during gliomagenesis in a mouse model (Signaroldi et al., 2016). This was now extended here to interrogate the molecular networks in human set of samples derived from two of the most aggressive cancers. Specifically, I combined the informative power of human cancer cells of differential tumorigenic potential derived from HGSOC and GBM with a computational pipeline for the analysis and integration of multi-omic NGS data sets to generate mechanistic testable hypotheses on cancer progression.

The future scope of this work seems to be quite abundant and important as the TF networks associated with both the OC and GBM tumor revealed in this study can be farther extended to find association of epigenetic repressors or activators on these candidate MRs. This will pave a way for defining epigenetic circuitry involved in mediating oncogenesis. Farther *in vitro* assessment of these TFs (either epigenetically driven or as a result of genetic factors) might lead to altering or aggravating the tumor phenotype thus leading the discovery of clinically relevant therapeutic targets.

Bibliography

Ahmed, A. A., Etemadmoghadam, D., Temple, J., Lynch, A. G., Riad, M., Sharma, R., … Brenton, J. D. (2010). Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *Journal of Pathology*, *221*(1), 49–56. http://doi.org/10.1002/path.2696

Ahmed, N., & Stenvers, K. L. (2013). Getting to know ovarian cancer ascites: opportunities for targeted therapy-based translational research. *Frontiers in Oncology*, *3*, 256. http://doi.org/10.3389/fonc.2013.00256

Alexa, A., & Rahnenfuhrer, J. (2010). topGO: topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0. *October*.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. a J. R., Behjati, S., Biankin, A. V, … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–21. http://doi.org/10.1038/nature12477

Arisan, S., Buyuktuncer, E. D., Palavan-Unsal, N., Caskurlu, T., Cakir, O. O., & Ergenekon, E. (2005). Increased expression of EZH2, a polycomb group protein, in bladder carcinoma. *Urol Int*, *75*(3), 252–257. http://doi.org/10.1159/000087804

Auer, K., Bachmayr-Heyda, A., Aust, S., Sukhbaatar, N., Reiner, A. T., Grimm, C., … Pils, D. (2015). Peritoneal tumor spread in serous ovarian cancer-epithelial mesenchymal status and outcome. *Oncotarget*, *6*(19), 17261–75. http://doi.org/10.18632/oncotarget.3746

Ayantunde, A. A., & Parsons, S. L. (2007). Pattern and prognostic factors in patients with malignant ascites: a retrospective study. *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO*, *18*(5), 945–9. http://doi.org/10.1093/annonc/mdl499

Barbeau, D. J., La, K. T., Kim, D. S., Kerpedjieva, S. S., Shurin, G. V, & Tamama, K. (2014). Early growth response-2 signaling mediates immunomodulatory effects of human multipotential stromal cells. *Stem Cells and Development*, *23*(2), 155–66. http://doi.org/10.1089/scd.2013.0194

Bowtell, D. D. (2010). The genesis and evolution of high-grade serous ovarian cancer. *Nat Rev Cancer*, *10*(11), 803–808. http://doi.org/10.1038/nrc2946

Bracken, A. P., Kleine-Kohlbrecher, D., Dietrich, N., Pasini, D., Gargiulo, G., Beekman, C., … Helin, K. (2007). The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells.

Genes and Development, *21*(5), 525–530. http://doi.org/10.1101/gad.415507

Bruney, L., Liu, Y., Grisoli, A., Ravosa, M. J., & Stack, M. S. (2016). Integrin-linked kinase activity modulates the pro-metastatic behavior of ovarian cancer cells. *Oncotarget*, *7*(16). http://doi.org/10.18632/oncotarget.7880

Byrd, K. N., & Shearn, A. (2003). ASH1, a Drosophila trithorax group protein, is required for methylation of lysine 4 residues on histone H3. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11535–40. http://doi.org/10.1073/pnas.1933593100

Carette, J. E., Pruszak, J., Varadarajan, M., Blomen, V. A., Gokhale, S., Camargo, F. D., … Brummelkamp, T. R. (2010). Generation of iPSCs from cultured human malignant cells. *Blood*, *115*(20), 4039–4042. http://doi.org/10.1182/blood-2009-07-231845

Chapman-Rothe, N., Curry, E., Zeller, C., Liber, D., Stronach, E., Gabra, H., … Brown, R. (2013). Chromatin H3K27me3/H3K4me3 histone marks define gene sets in high-grade serous ovarian cancer that distinguish malignant, tumour-sustaining and chemo-resistant ovarian tumour cells. *Oncogene*, *32*(38), 4586–92. http://doi.org/10.1038/onc.2012.477

Chen, J., Li, Y., Yu, T.-S., McKay, R. M., Burns, D. K., Kernie, S. G., & Parada, L. F. (2012). A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature*, *488*(7412), 522–526. http://doi.org/10.1038/nature11287

Choi, E. J., Yoo, N. J., Kim, M. S., An, C. H., & Lee, S. H. (2016). Putative Tumor Suppressor Genes &lt;b&gt;&lt;i&gt;EGR1 &lt;/i&gt;&lt;/b&gt;and&lt;b&gt;&lt;i&gt; BRSK1&lt;/i&gt;&lt;/b&gt; Are

Mutated in Gastric and Colorectal Cancers. *Oncology*, *91*(5), 289–294. http://doi.org/10.1159/000450616

Corominas-Faja, B., Cufí, S., Oliveras-Ferraros, C., Cuyàs, E., López-Bonet, E., Lupu, R., … Menendez, J. A. (2013). Nuclear reprogramming of luminal-like breast cancer cells generates Sox2-overexpressing cancer stem-like cellular states harboring transcriptional activation of the mTOR pathway. *Cell Cycle*, *12*(18), 3109–3124. http://doi.org/10.4161/cc.26173

Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye, B. N. and C. L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *503*(7471), 333–339. http://doi.org/10.1007/s13398-014-0173-7.2

D'haeseleer, P. (2006). How does DNA sequence motif discovery work? *Nature Biotechnology*, *24*(8), 959–961. http://doi.org/10.1038/nbt0806-959

Dang, D. T., Chen, X., Feng, J., Torbenson, M., Dang, L. H., & Yang, V. W. (2003). Overexpression of Krüppel-like factor 4 in the human colon cancer cell line RKO leads to reduced tumorigenecity. *Oncogene*, *22*(22), 3424–30. http://doi.org/10.1038/sj.onc.1206413

Daniel, P. M., Filiz, G., & Mantamadiotis, T. (2016). Sensitivity of GBM cells to cAMP agonist-mediated apoptosis correlates with CD44 expression and agonist resistance with MAPK signaling. *Cell Death and Disease*, *7*(12), e2494. http://doi.org/10.1038/cddis.2016.393

Davidson, B., Abeler, V. M., Førsund, M., Holth, A., Yang, Y., Kobayashi, Y., … Wang, T. L. (2014). Gene expression signatures of primary and metastatic uterine leiomyosarcoma. *Human Pathology*, *45*(4), 691–700. http://doi.org/10.1016/j.humpath.2013.11.003

Egger, G., Liang, G., Aparicio, A., & Jones, P. a. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, *429*(6990), 457–463. http://doi.org/10.1038/nature02625

Elrod, H. A., & Sun, S. Y. (2008). PPAR?? and apoptosis in cancer. *PPAR Research*. http://doi.org/10.1155/2008/704165

Fajardo, A. M., Piazza, G. A., & Tinsley, H. N. (2014). The role of cyclic nucleotide signaling pathways in cancer: Targets for prevention and treatment. *Cancers*. http://doi.org/10.3390/cancers6010436

Feinberg, A. P., & Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, *301*(5895), 89–92. http://doi.org/10.1038/301089a0

Fernandez-Banet, J., Esposito, A., Coffin, S., Horvath, I. B., Estrella, H., Schefzick, S., … Kan, Z. (2015). OASIS: web-based platform for exploring cancer multi-omics data. *Nature Methods*, *13*(1), 9–10. http://doi.org/10.1038/nmeth.3692

Frattini, V., Trifonov, V., Chan, J. M., Castano, A., Lia, M., Abate, F., … Iavarone, A. (2013). The integrated landscape of driver genomic alterations in glioblastoma. *Nature Genetics*, *45*(10), 1141–9. http://doi.org/10.1038/ng.2734

Gandre-Babbe, S., Paluru, P., Aribeana, C., Chou, S. T., Bresolin, S., Lu, L., … Weiss, M. J. (2013). Patient-derived induced pluripotent stem cells recapitulate hematopoietic abnormalities of juvenile myelomonocytic leukemia. *Blood*, *121*(24), 4925–4929. http://doi.org/10.1182/blood-2013-01-478412

Germain, P. L., Vitriolo, A., Adamo, A., Laise, P., Das, V., & Testa, G. (2016). RNAontheBENCH: Computational and empirical resources for benchmarking RNAseq quantification and differential expression methods.

Nucleic Acids Research, *44*(11), 5054–5067. http://doi.org/10.1093/nar/gkw448

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., … Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, *10*(11), 1081–1082. http://doi.org/10.1038/nmeth.2642

Gregg, J., & Fraizer, G. (2011). Transcriptional Regulation of EGR1 by EGF and the ERK Signaling Pathway in Prostate Cancer Cells. *Genes & Cancer*, *2*(9), 900–9. http://doi.org/10.1177/1947601911431885

Guilhamon, P., Eskandarpour, M., Halai, D., Wilson, G. A., Feber, A., Teschendorff, A. E., … Beck, S. (2013). Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2. *Nature Communications*, *4*, 2166. http://doi.org/10.1038/ncomms3166

Hanahan, D. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70. http://doi.org/10.1016/S0092-8674(00)81683-9

Hu, K., Yu, J., Suknuntha, K., Tian, S., Montgomery, K., Choi, K.-D., … Slukvin, I. I. (2011). Efficient generation of transgene-free induced pluripotent stem cells from normal and neoplastic bone marrow and cord blood mononuclear cells. *Blood*, *117*(14), e109–e119. http://doi.org/10.1182/blood-2010-07-298331

Hu, L., McArthur, C., & Jaffe, R. B. (2010). Ovarian cancer stem-like side-population cells are tumourigenic and chemoresistant. *British Journal of Cancer*, *102*(8), 1276–83. http://doi.org/10.1038/sj.bjc.6605626

Huang, L., Schauer, I. G., Zhang, J., Mercado-Uribe, I., Deavers, M. T., Huang, J., & Liu, J. (2011). The oncogenic gene fusion TMPRSS2: ERG is not a diagnostic or prognostic marker for ovarian cancer. *International Journal of Clinical and Experimental Pathology*, *4*(7), 644–650.

Hylander, B. L., Punt, N., Tang, H., Hillman, J., Vaughan, M., Bshara, W., … Repasky, E. a. (2013). Origin of the vasculature supporting growth of primary patient tumor xenografts. *Journal of Translational Medicine*, *11*, 110. http://doi.org/10.1186/1479-5876-11-110

Iarc, Tavassoéli, F. ., & Devilee, P. (eds. (2003). Pathology and Genetics of Tumours of the Breast and Female Genital Organs. *Pathology and Genetics of Tumours of the Breast and Female Genital Organs*.

Jones, P. a, & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews. Genetics*, *3*(6), 415–28. http://doi.org/10.1038/nrg816

Kim, J., Hoffman, J. P., Alpaugh, R. K., Rhimm, A. D., Reichert, M., Stanger, B. Z., … Zaret, K. S. (2013). An iPSC Line from Human Pancreatic Ductal Adenocarcinoma Undergoes Early to Invasive Stages of Pancreatic Cancer Progression. *Cell Reports*, *3*(6), 2088–2099. http://doi.org/10.1016/j.celrep.2013.05.036

Kim, J. J. (2015). Applications of iPSCs in cancer research. *Biomarker Insights*, *2015*, 125–131. http://doi.org/10.4137/BMI.S20065

Kipps, E., Tan, D. S., & Kaye, S. B. (2013). Meeting the challenge of ascites in ovarian cancer: new avenues for therapy and research. *Nat Rev Cancer*, *13*(4), 273–282. http://doi.org/10.1038/nrc3432

Kirmizis, A., Bartley, S. M., Kuzmichev, A., Margueron, R., Reinberg, D., Green, R., & Farnham, P. J. (2004). Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes and Development*, *18*(13), 1592–1605. http://doi.org/10.1101/gad.1200204

Klinkebiel, D., Zhang, W., Akers, S. N., Odunsi, K., & Karpf, A. R. (2016). DNA Methylome Analyses Implicate Fallopian Tube Epithelia as the Origin for High-Grade Serous Ovarian Cancer. *Molecular Cancer Research*, *14*(9), 787–794. http://doi.org/10.1158/1541-7786.MCR-16-0097

Kotini, A. G., Chang, C.-J., Boussaad, I., Delrow, J. J., Dolezal, E. K., Nagulapally, A. B., … Papapetrou, E. P. (2015). Functional analysis of a chromosomal deletion associated with myelodysplastic syndromes using isogenic human induced pluripotent stem cells. *Nature Biotechnology*, *33*(6), 646–55. http://doi.org/10.1038/nbt.3178

Kumano, K., Arai, S., Hosoi, M., Taoka, K., Takayama, N., Otsu, M., … Kurokawa, M. (2012). Generation of induced pluripotent stem cells from primary chronic myelogenous leukemia patient samples. *Blood*, *119*(26), 6234–42. http://doi.org/10.1182/blood-2011-07-367441

Kurman, R. J. (2013). Origin and molecular pathogenesis of ovarian high-grade serous carcinoma. *Annals of Oncology*, *24*(SUPPL.10). http://doi.org/10.1093/annonc/mdt463

Kurman, R. J., & Shih, I.-M. (2010). The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory. *The American Journal of Surgical Pathology*, *34*(3), 433–43. http://doi.org/10.1097/PAS.0b013e3181cf3d79

Kwon, O., Park, J., Baek, S., Noh, S., Song, K., Kim, S., & Kim, Y. S. (2013). demethylation promotes the growth and resistance to 5-fluorouracil of gastric cancer cells, *104*(3), 391–397. http://doi.org/10.1111/cas.12076

Laugesen, A., & Helin, K. (2014). Chromatin repressive complexes in stem cells, development, and cancer. *Cell Stem Cell*.

http://doi.org/10.1016/j.stem.2014.05.006

Lee, D. F., Su, J., Kim, H. S., Chang, B., Papatsenko, D., Zhao, R., … Lemischka, I. R. (2015). Modeling familial cancer with induced pluripotent stem cells. *Cell*, *161*(2), 240–254. http://doi.org/10.1016/j.cell.2015.02.045

Li, H., Wang, J., Xiao, W., Xia, D., Lang, B., Yu, G., … Xu, H. (2013). Epigenetic alterations of krüppel-like factor 4 and its tumor suppressor function in renal cell carcinoma. *Carcinogenesis*, *34*(10), 2262–2270. http://doi.org/10.1093/carcin/bgt189

Liao, D. (2009). Emerging roles of the EBF family of transcription factors in tumor suppression. *Molecular Cancer Research : MCR*, *7*(12), 1893–1901. http://doi.org/10.1158/1541-7786.MCR-09-0229

Lin, C. Y., & Gustafsson, J. A. (2015). Targeting liver X receptors in cancer therapeutics. *Nat Rev Cancer*, *15*(4), 216–224. http://doi.org/10.1038/nrc3912

Lin, S.-L., Chang, D. C., Chang-Lin, S., Lin, C.-H., Wu, D. T. S., Chen, D. T., & Ying, S.-Y. (2008). Mir-302 reprograms human skin cancer cells into a pluripotent ES-cell-like state. *RNA (New York, N.Y.)*, *14*(10), 2115–24. http://doi.org/10.1261/rna.1162708

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. http://doi.org/10.1186/s13059-014-0550-8

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., … Wasserman, W. W. (2016). JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *44*(D1), D110–D115. http://doi.org/10.1093/nar/gkv1176

Mathieu, J., Zhang, Z., Zhou, W., Wang, A. J., Heddleston, J. M., Pinna, C. M., … Ruohola-Baker, H. (2011). HIF induces human embryonic stem cell markers in cancer cells. *Cancer Res*, *71*(13), 4640–4652. http://doi.org/10.1158/0008-5472.CAN-10-3320

Miyoshi, N., Ishii, H., Nagai, K., Hoshino, H., Mimori, K., Tanaka, F., … Mori, M. (2010). Defined factors induce reprogramming of gastrointestinal cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 40–5. http://doi.org/10.1073/pnas.0912407107

Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., Stadler, M. B., … Schübeler, D. (2008). Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. *Molecular Cell*, *30*(6), 755–766. http://doi.org/10.1016/j.molcel.2008.05.007

Moore, J. B., Loeb, D. M., Hong, K. U., Sorensen, P. H., Triche, T. J., Lee, D. W., … Arceci, R. J. (2015). Epigenetic reprogramming and re-differentiation of a Ewing sarcoma cell line. *Frontiers in Cell and Developmental Biology*, *3*(March), 15. http://doi.org/10.3389/fcell.2015.00015

Moran, S., Martínez-Cardús, A., Sayols, S., Musulén, E., Balañá, C., Estival-Gonzalez, A., … Esteller, M. (2016). Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *The Lancet Oncology*. http://doi.org/10.1016/S1470-2045(16)30297-2

Ng, A., & Barker, N. (2015). Ovary and fimbrial stem cells: biology, niche and cancer origins. *Nature Reviews. Molecular Cell Biology*, *16*(10), 625–38. http://doi.org/10.1038/nrm4056

Orkin, S. H., & Hochedlinger, K. (2011). Chromatin connections to pluripotency and cellular reprogramming. *Cell*. http://doi.org/10.1016/j.cell.2011.05.019

Parra Villegas, E., Ferreira, J., & Ortega, A. (2011). Overexpression of EGR-1 modulates the activity of NF-??B and AP-1 in prostate carcinoma PC-3 and LNCaP cell lines. *International Journal of Oncology*, *39*(2), 345–352. http://doi.org/10.3892/ijo.2011.1047

Patch, A.-M., Christie, E. L., Etemadmoghadam, D., Garsed, D. W., George, J., Fereday, S., … Bowtell, D. D. L. (2015). Whole–genome characterization of chemoresistant ovarian cancer. *Nature*, *521*(7553), 489–494. http://doi.org/10.1038/nature14410

Patro, R., Duggal, G., & Kingsford, C. (2015). Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*, 21592. http://doi.org/10.1101/021592

Rao, Z.-Y., Cai, M.-Y., Yang, G.-F., He, L.-R., Mai, S.-J., Hua, W.-F., … Xie, D. (2010). EZH2 supports ovarian carcinoma cell invasion and/or metastasis via regulation of TGF-beta1 and is a predictor of outcome in ovarian carcinoma patients. *Carcinogenesis*, *31*(9), 1576–83. http://doi.org/10.1093/carcin/bgq150

Robertson, K. D. (2002). DNA methylation and chromatin - unraveling the tangled web. *Oncogene*, *21*(35), 5361–5379. http://doi.org/10.1038/sj.onc.1205609

Rosanò, L., & Bagnato, A. (2016). β-arrestin1 at the cross-road of endothelin-1 signaling in cancer. *Journal of Experimental & Clinical Cancer Research*, *35*(121). http://doi.org/10.1186/s13046-016-0401-4

Rowland, B. D., Bernards, R., & Peeper, D. S. (2005). The KLF4 tumour suppressor is a transcriptional repressor of p53 that acts as a context-dependent oncogene. *Nature Cell Biology*, *7*(11), 1074–82. http://doi.org/10.1038/ncb1314

Sashida, G., Bazzoli, E., Menendez, S., Liu, Y., & Nimer, S. D. (2010). The oncogenic role of the ETS transcription factors MEF and ERG. *Cell Cycle*. http://doi.org/10.4161/cc.9.17.13000

Signaroldi, E., Laise, P., Cristofanon, S., Brancaccio, A., Reisoli, E., Atashpaz, S., … Testa, G. (2016). Polycomb dysregulation in gliomagenesis targets a Zfp423-dependent differentiation network. *Nature Communications*, *7*, 10753. http://doi.org/10.1038/ncomms10753

Singer, G., Oldt, R., Cohen, Y., Wang, B. G., Sidransky, D., Kurman, R. J., & Shih, I.-M. (2003). Mutations in BRAF and KRAS Characterize the Development of Low-Grade Ovarian Serous Carcinoma. *JNCI Journal of the National Cancer Institute*, *95*(6), 484–486. http://doi.org/10.1093/jnci/95.6.484

Singh, D., Chan, J. M., Zoppoli, P., Niola, F., Sullivan, R., Castano, A., … Iavarone, A. (2012). Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science (New York, N.Y.)*, *337*(6099), 1231–5. http://doi.org/10.1126/science.1220834

Sosa, M. S., Parikh, F., Maia, A. G., Estrada, Y., Bosch, A., Bragado, P., … Aguirre-Ghiso, J. a. (2015). NR2F1 controls tumour cell dormancy via SOX9- and RARβ-driven quiescence programmes. *Nature Communications*, *6*, 6170. http://doi.org/10.1038/ncomms7170

Sproul, D., Kitchen, R. R., Nestor, C. E., Dixon, J. M., Sims, A. H., Harrison, D. J., … Meehan, R. R. (2012). Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biology*, *13*(10), R84. http://doi.org/10.1186/gb-2012-13-10-r84

Steed, T. C., Treiber, J. M., Patel, K., Ramakrishnan, V., Merk, A., Smith, A. R., … Chen, C. C. (2016). Differential localization of glioblastoma subtype: implications on glioblastoma pathogenesis. *Oncotarget*, 1–9. http://doi.org/10.18632/oncotarget.8551

Stepulak, A., Rola, R., Polberg, K., & Ikonomidou, C. (2014). Glutamate and its receptors in cancer. *Journal of Neural Transmission*, *121*(8), 933–944. http://doi.org/10.1007/s00702-014-1182-6

Stricker, S. H., Feber, A., Engström, P. G., Carén, H., Kurian, K. M., Takashima, Y., … Pollard, S. M. (2013). Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner. *Genes and Development*, *27*(6), 654–669. http://doi.org/10.1101/gad.212662.112

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–676. http://doi.org/10.1016/j.cell.2006.07.024

Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, *126*(4), 663–676. http://doi.org/10.1016/j.cell.2006.07.024

Tian, Y., Pan, Q., Shang, Y., Zhu, R., Ye, J., Liu, Y., … Wang, R. (2014). MicroRNA-200 (miR-200) cluster regulation by achaete scute-like 2 (Ascl2) impact on the epithelial-mesenchymal transition in colon cancer cells. *Journal of Biological Chemistry*, *289*(52), 36101–36115. http://doi.org/10.1074/jbc.M114.598383

Tone, A. A., Begley, H., Sharma, M., Murphy, J., Rosen, B., Brown, T. J., & Shaw, P. A. (2008). Gene expression profiles of luteal phase fallopian tube epithelium from BRCA mutation carriers resemble high-grade serous carcinoma. *Clinical Cancer Research*, *14*(13), 4067–4078. http://doi.org/10.1158/1078-0432.CCR-07-4959

Utikal, J., Maherali, N., Kulalert, W., & Hochedlinger, K. (2009). Sox2 is dispensable for the reprogramming of melanocytes and melanoma cells into induced pluripotent stem cells. *J Cell Sci*, *122*(Pt 19), 3502–3510. http://doi.org/10.1242/jcs.054783

Vang, R., Shih, I.-M., & Kurman, R. J. (2009). Ovarian low-grade and high-grade serous carcinoma: pathogenesis, clinicopathologic and molecular biologic features, and diagnostic problems. *Advances in Anatomic Pathology*, *16*(5), 267–82. http://doi.org/10.1097/PAP.0b013e3181b4fffa

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., … Chinnaiyan, A. M. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, *419*(6907), 624–629. http://doi.org/10.1038/nature01075\rnature01075 [pii]

Vaughan, S., Coward, J. I., Bast, R. C., Berchuck, A., Berek, J. S., Brenton, J. D., … Balkwill, F. R. (2011). Rethinking ovarian cancer: recommendations for improving outcomes. *Nature Reviews. Cancer*, *11*(10), 719–25. http://doi.org/10.1038/nrc3144

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., … Hayes, D. N. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, *17*(1), 98–110. http://doi.org/10.1016/j.ccr.2009.12.020

Wang, B., Zhao, M.-Z., Cui, N.-P., Lin, D.-D., Zhang, A.-Y., Qin, Y., … Chen, B.-P.

(2015). Krüppel-like factor 4 induces apoptosis and inhibits tumorigenic progression in SK-BR-3 breast cancer cells. *FEBS Open Bio*, *5*, 147–54. http://doi.org/10.1016/j.fob.2015.02.003

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., … Stuart, J. M. (2013a). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–20. http://doi.org/10.1038/ng.2764

Weinstein, J. N., Collisson, E. a, Mills, G. B., Shaw, K. R. M., Ozenberger, B. a, Ellrott, K., … Stuart, J. M. (2013b). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–20. http://doi.org/10.1038/ng.2764

Wingender, E., Dietze, P., Karas, H., & Knüppel, R. (1996). TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Research*. http://doi.org/10.1093/nar/24.1.238

Wong, K.-K., Tsang, Y. T. M., Deavers, M. T., Mok, S. C., Zu, Z., Sun, C., … Gershenson, D. M. (2010). BRAF Mutation Is Rare in Advanced-Stage Low-Grade Ovarian Serous Carcinomas. *The American Journal of Pathology*, *177*(4), 1611–1617. http://doi.org/10.2353/ajpath.2010.100212

Wu, Y., Zhang, X., Liu, Y., Lu, F., & Chen, X. (2016). Decreased expression of BNC1 and BNC2 is associated with genetic or epigenetic regulation in hepatocellular carcinoma. *International Journal of Molecular Sciences*, *17*(2). http://doi.org/10.3390/ijms17020153

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., … Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*(10), 1134–1140. http://doi.org/10.1038/ng.2760

Zambelli, F., Pesole, G., & Pavesi, G. (2009). Pscan: Finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Research*, *37*(SUPPL. 2). http://doi.org/10.1093/nar/gkp464

Zhang, H., Chen, X., Wang, J., Guang, W., Han, W., Zhang, H., … Gu, Y. (2014). EGR1 decreases the malignancy of human non-small cell lung carcinoma by regulating KRT18 expression. *Scientific Reports*, *4*, 5416. http://doi.org/10.1038/srep05416

Zhang, H. M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., & Guo, A. Y. (2012). AnimalTFDB: A comprehensive animal transcription factor database. *Nucleic Acids Research*, *40*(D1). http://doi.org/10.1093/nar/gkr965

Zhang, X., Cruz, F. D., Terry, M., Remotti, F., & Matushansky, I. (2013). Terminal differentiation and loss of tumorigenicity of human cancers via pluripotency-based reprogramming. *Oncogene*, *32*(18), 2249–2260. http://doi.org/10.1038/onc.2012.237

Zhu, R., Yang, Y., Tian, Y., Bai, J., Zhang, X., Li, X., … Wang, R. (2012). Ascl2 knockdown results in tumor growth arrest by mirna-302b-related inhibition of colon cancer progenitor cells. *PLoS ONE*, *7*(2). http://doi.org/10.1371/journal.pone.0032170

Zong, H., Verhaak, R. G. W., & Canoll, P. (2012). The cellular origin for malignant glioma and prospects for clinical advancements. *Expert Review of Molecular Diagnostics*, *12*(4), 383–94. http://doi.org/10.1586/erm.12.30