Université d'Evry Val d'Essonne

Ecole Doctorale n°423 des Génomes Aux Organismes (GAO)

Thèse

présentée et soutenue à Evry le 9 janvier 2015

par :

Rémy Nicolle

Regulatory networks driving bladder cancer

Composition du jury

Président :	Christophe Ambroise
Rapporteur :	Charles Lecellier
Rapporteur :	Frédéric Devaux
Examinateur :	Céline Lefebvre
Directeur de thèse :	François Radvanyi
Directeur de thèse :	Mohamed Elati

Attention, ce mémoire contient des informations confidentielles susceptibles de faire l'objet d'une demande de brevet. En aucun cas, les informations contenues dans ce mémoire de thèse ne doivent i) être divulguées à qui que ce soit (sauf aux personnes s'étant engagées à garder ces informations confidentielles) et sous aucune forme que ce soit, ii) être utilisées notamment dans le but de commercialiser, directement ou indirectement ou par personne interposeée, quelque produit, méthode ou service qui aurait quelque similarité avec le contenu du mémoire, et ce jusqu'à ce que ces informations soient rendues publiques.

Summary

Carcinogenesis is a consequence of the unceasing activation of cell proliferation. In normal cells, mitogenic stimuli are processed by a complex network of protein interactions and enzymatic reactions, often referred to as pathways, which can eventually trigger the activation of new genes to engage the cell into mitosis. During developmental or wound healing processes, this complex regulation of cellular phenotypes results in a tight control of the number and behavior of cells and therefore contributes to the maintenance of a functional and healthy tissue architecture.

Based on genomic, transcriptomic and proteomic profiles of bladder tumors and transcriptomes of normal urothelial cells at various states of proliferation and differentiation, I devised novel methodologies to characterize the pathways driving bladder cancer.

I first developed a set of tools to identify and visualize sample and subtype-specific transcriptional programs through the inference of a co-regulatory network and the prediction of transcription factor activity. These methods were embedded in a Bioconductor package entitled CoRegNet (bioconductor.org). The measure of transcriptional activity is based on the influence of a transcription factor on the expression of its target genes and was used to characterize the most active regulators of each bladder cancer subtypes. The integration of genomic profiles highlighted two altered transcription factors with driver roles in luminal-like and basal-like bladder cancer, one of which was experimentally validated.

The use of CoRegNet to model the contribution of regulatory programs of normal proliferation and differentiation in bladder cancers underlined a strong preservation of normal networks during tumorigenesis. Furthermore, a regulator of normal proliferation was found to be constitutively activated by genetic alterations and its influence on bladder cancer cell proliferation was experimentally validated. In addition, a master regulator of urothelial differentiation was found to have a loss of activity in nearly all tumors. This was then associated to the discovery of frequent inactivating mutations and further analysis uncovered a major role in differentiated tumors.

In order to characterize signaling pathways from proteomic pull-down assays, I then designed a novel algorithm to grow a densely connected network from a set of proteins and a repository of protein interactions. The proposed algorithm was made available as a Cytoscape application named Pepper for Protein Complex Expansion using Protein-Protein interaction networks (apps.cytoscape.org).

Finally, using both a proteomic pull-down assay of the bladder cancer oncogene *FGFR3* and a transcriptomic profiling of its downstream regulated genes, I applied Pepper to characterize the full *FGFR3* signaling pathway from its protein partners to the downstream transcriptional regulators. In particular, this uncovered a regulatory link between *FGFR3* and the tumor suppressor *TP53*.

Résumé en Français

La carcinogénèse est une conséquence de la continuelle activation de la prolifération cellulaire. Dans les cellules normales, les signaux mitogéniques sont traités par un réseau complexe d'interactions protéiques et de réactions enzymatiques, appelées voies de signalisation. Dans certains cas, le signal peut induire l'activation de nouveaux gènes et ainsi déclencher la mitose. Lors du développement ou de la cicatrisation, cette régulation du phénotype cellulaire contrôle étroitement le nombre et le comportement des cellules contribuant ainsi au maintien d'un tissu fonctionnel sain.

A partir de profils génomiques, transcriptomiques et protéomiques de tumeurs de la vessie ainsi que des transcriptomes de cellules urothéliales normales dans différents états de prolifération et de différenciation, j'ai mis au point de nouvelles méthodologies pour caractériser les voies de signalisation et de régulation responsables des cancers de la vessie.

Dans un premier temps, j'ai développé des outils pour l'identification et la visualisation des programmes transcriptionnels spécifiques à une tumeur ou à un sous-type tumoral et ce, par l'inférence d'un réseau de co-régulation et la prédiction de l'activité des facteurs de transcription. Ces méthodes sont disponibles dans un package Bioconductor, CoRegNet (bioconductor.org). La mesure de l'activité transcriptionnelle est basée sur l'influence d'un facteur de transcription sur l'expression de ses gènes cibles. Cette mesure a été utilisée pour identifier les régulateurs les plus actifs de chaque sous-type de cancer de la vessie. L'intégration de profils génomiques a mis en avant deux facteurs de transcription génétiquement altérés et ayant des rôles oncogènes dans les tumeurs luminales et basales. L'un d'entre eux a été validé expérimentalement dans ce travail.

L'utilisation de CoRegNet a mis en évidence une large utilisation dans les tumeurs, des réseaux normaux de la différenciation et de la prolifération des cellules normales. Un régulateur de la prolifération normale est identifié comme étant activé de façon constitutive par des altérations génétiques dans les tumeurs. Son impact sur la prolifération des cellules tumorales de la vessie a été expérimentalement validé. Par ailleurs, il a été constaté que l'un des régulateurs de la différenciation urothéliale présentant une baisse d'activité dans la quasi-totalité des tumeurs, est fréquemment muté. De plus amples analyses ont mis en avant son rôle majeur dans les tumeurs différenciées.

Dans le but de caractériser les voies de signalisation à partir de données protéomiques d'expériences d'immunoprécipitations, j'ai développé un nouvel algorithme visant à construire un réseau dense à partir d'une liste de protéines d'intérêt et d'un ensemble d'interactions protéiques connues. L'algorithme est proposé sous la forme d'une application Cytoscape et s'intitule Pepper: Protein Complex Expansion using Protein-Protein interaction networks (apps.cytoscape.org)

Enfin, en utilisant à la fois le profil protéomique d'une expérience d'immunoprécipitation de FGFR3 ainsi que le profil transcriptomique des gènes qu'il régule en aval, j'ai

appliqué Pepper pour caractériser la voie de signalisation de *FGFR3* depuis ses partenaires protéiques jusqu'aux facteurs de transcription en aval. Enfin, ce travail a plus particulierement permis d'identifier un lien de régulation entre *FGFR3* et le gène suppresseur de tumeurs *TP53*.

Remerciements

En premier lieu, je souhaite remercier mes deux directeurs de thèse, François Radvanyi et Mohamed Elati, pour m'avoir accueilli au sein de leurs laboratoires et pour m'avoir permis de prendre part à des projets de recherche passionants.

Je tiens à adresser mes sincères remerciements à Charles Lecellier et Frédéric Devaux pour avoir accepté de juger et critiquer mon travail en tant que rapporteurs. Je remercie également Céline Lefebvre et Christophe Ambroise pour leur participation à mon jury de thèse.

Un grand merci aux membres de l'équipe MEGA à l'ISSB et à l'équipe d'Oncologie Moléculaire à Curie. Je remercie tout particulièrement Jennifer pour nos nombreux échanges qui m'ont permis, entre autre, de discerner l'ensemble de ses tics verbaux, Mélanie pour nos discussions incessantes et obsessives à propos de certains gènes et Clémentine pour avoir réussi à me faire parler d'autre chose que de vessie mais surtout pour m'avoir enseigné les rudiments de la culture cellulaire. Un grand merci à Elodie, Verra, Thibault et Céline pour avoir rendu aussi agréable le partage de ce petit espace empli de l'odeur du café brûlé. Merci à Meriem, Florent et tous les autres membres de l'équipe pour leurs nombreuses visites de notre bureau bien que souvent poussés par la gourmandise. Enfin, merci aux Costas pour ne pas avoir hésité à donner leurs avis en toutes circonstances.

Je remercie les personnes avec qui j'ai eu l'occasion de collaborer durant ma thèse et en particulier Etienne, Thomas, Julien, Pierre et Sophie pour avoir partagé avec moi leurs visions des mathématiques.

Enfin un grand merci à Laura pour m'avoir soutenu et surtout supporté durant ces trois dernières années.

Contents

List of Figures

Abbreviations

3C	Chromosome Conformation Capture
ABS	Adult Bovine Serum
aCGH	array Comparative Genomic Hybridization
AP-MS	Affinity Purification followed by Mass-Spectrometry
AUPR	Area Under the Precision Recall Curve
BAC	Bacterial Artificial Chromosome
CIS	Carcinoma in situ
CNA	Copy Number Aberration
CNA	Copy Number Variation
DB	Database
DREAM	Dialogue on Reverse Engineering Assessment and Methods
EGFR	Epidermal Growth Factor Receptor
ENCODE	Encyclopedia of DNA Elements
ER	Estrogen Receptor
FANTOM	Functional Annotation of the Mammalian Genome
FGFR3	Fibroblast Growth Factor Receptor 3
GES	Gene Expression Signature
ICGC	International Cancer Genome Consortium
indel	insertion-deletion
IP	Immuno-Precipitation
kb	kilobase
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
MAPK	Mitogen-Activated Protein Kinase Pathway
MARINa	Master Regulator Inference algorithm
MRES	Multiple Region Epigenetic Silencing
mRNA	messenger RNA
MS	Mass Spectrometry
NCA	Network Component Analysis
NHU	Normal Human Urothelium
NLS	Nuclear Localization Signal
PPARγ	Peroxisome Proliferator-Activator Receptor Gamma
PPI	Protein Protein Interaction
PSSM	Position Specific Scoring Matrix
PTM	Post Translational Modification
PUNLMP	Papillary Urothelial Neoplasm of Low Malignant Potential

PWM	Position Weight Matrix
qPCR	Quantitative Polymerase Chain Reaction
RNA-seq	RNA sequencing
RPPA	Reverse Phase Protein Arrays
RTK	Receptor Tyrosine Kinase
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TFA	Transcription Factor Activity
TFBS	Transcription Factor Binding Site
TSG	Tumor Suppressor Gene
TSS	Transcription Starting Site
UTR	Untranslated Regions
WHO	World Health Organization

Preamble

"Tumors appear to the host in the guise of [...] an unending series of wounds that continually initiate healing but never heal completely" (Dvorak, 1986). Harold Dvorak's catchy description of *tumors as wounds that do no heal* was originally based on an analysis of the tissue surrounding solid tumors and wounds. Decades of research in the biochemistry of signal transduction revealed that this might also be true at the molecular level of the internal regulatory circuitry. I fact, the parallel of tumor growth with normal processes goes beyond wound healing and is relevant to any mechanism involving tissue regeneration and cellular proliferation such as those found in developmental processes.

In appearance, the aberrant proliferative phenotype of cancer cells suggests that they entirely reprogram the control over their own growth and division. The continuous attempt to map the control circuitry of cellular division, migration or death revealed that these circuits, often referred to as pathways, are nearly identical in both cancer and normal/healthy cells during developmental and healing processes. Thereon, in addition to the refinement of pathway maps, the focus is to be made on finding the small alterations of the control machinery that lead to malignant transformation.

Pathways controlling cellular processes are composed of chains of reactions between proteins, metabolites, mRNA and DNA species. These reactions are usually simply defined as interactions, corresponding in fact to a wide variety molecular processes including enzymatic reactions, protein and mRNA degradations, the formation of complexes and various processes altering the steric conformation of one or several of the interacting molecules. The simplification of these reactions into interactions involving pairs of molecules results in the possibility to apprehend a set of complex reactions as a biological network. Networks are used to model the set of interactions involving usually one or two types of molecules (*e.g.* protein-protein, protein-DNA, protein-RNA) at the level of an entire cellular system. The benefit of using network models relies on the availability of a wide variety of algorithmic tools to analyze these abstract representations of the control of cellular functions.

Throughout my PhD, I focused on different strategies to uncover the molecular networks of proliferation and differentiation in normal and cancerous cells of the urinary bladder. The use of large-scale networks, either regulatory networks reconstructed from transcriptomic data or based on repositories of protein interactions, gave me the opportunity to analyze transcriptomic, genomic and proteomic profiles of particular normal and malignant cellular conditions. This integration of both network models and molecular profiles resulted in the discovery of novel mechanisms underlying the neo-plastic transformation of urothelial cells.

In order to present the results of my studies, I will first describe cellular pathways by outlining the way extracellular signals are transduced from cell surface receptors to the nucleus and the effect these have on the regulation of genes. I will also discuss the

various forms of genetic alterations, their impact on the deregulations of cellular circuitry and the means by which these aberrations can lead to neo-plastic transformation. Then, I will describe the set of techniques used to profile molecular species of tumor cells, to identify the interactions between them and finally the current methods to uncover the cancer-driving pathways based on these datasets. In a last introductory section, I will summarize the current knowledge of the carcinogenesis of the urinary bladder.

Introduction

Chapter I

Cell behavior, signaling and transcription

In a complex multicellular organism, cells can engage into a broad variety of phenotypes and behaviors. The commitment of specific cell populations to a particular phenotype during development or the modification of their behavior in response to the environment is tightly controlled in time and space. This control of cell population maintains tissue architecture and activity by replacing missing cells, removing unneeded ones and forcing cells into particular functioning, differentiation, states. Therefore, the control of cell behavior is not defined at the level of individual cells but rather coordinated by the crowd to ensure coherence inside the population of a tissue and to a higher level, of an individual. Indeed, the decision of proliferating cannot be taken by an individual cell but by a consensus of the cell neighborhood in response to a need for cell renewal, following tissue damage for instance. Carcinogenesis is in some way an individual cell making the unconscious decision to divide and proliferate (Nowell, 1976). This population-based regulation of cell behavior relies on messages being broadcasted between cells. Cell-to-cell communication is essential to the cooperation of cells. Letting mechanical forces aside, this communication is mainly mediated by molecules (often proteins) emitted by a specific cell population and received by another (and sometimes by the same). The transmission of a single molecule can trigger major changes in the cells receiving the signal. For instance, in the presence of a growth factor, a population of cells can undergo a rapid transition from a quiescent to a fast proliferative state. Such high impact of extracellular signals on the state of cells implies a complex signal processing machinery with the potential to greatly modify the cells' phenotype. More importantly, the signal needs to be processed so that only the expected cellular processes, if any, are engaged. Decision making being an important component of signal transduction, the process starting from the reception of the signal to the modification of the cells phenotype is often compared to a circuit. In this metaphor, the signal is first received at the surface membrane of the cell, processed in the cytoplasm and finally conveyed to the nucleus where the signal becomes a modification of the set of expressed genes, the transcriptome.

I.1 Cell surface receptors and signal transduction

Cell-to-cell communication is mostly mediated by small proteins secreted outside of the signal-emitting cells. Whether the signal is received by the secreting cell (autocrine signaling) or neighboring cells (paracrine signaling), the delivery of the message requires the presence of a receptor specific to the emitted signal often present at the surface of the receiving cell. Growth factors (Witsch, Sela, and Yarden, 2010) represent a family of secreted proteins that convey signals for cellular proliferation and have important roles in major processes such as wound healing and development (Barrientos et al., 2008). These proteins act as ligands of their cognate receptors that import the signal inside the cell. A particular type of receptor transduces the growth signal through a tyrosine kinase domain, a mechanism that is illustrated in figure I.1. The mode of action of the growth factor/receptor tyrosine kinase (RTK) mainly relies on the dimerization of the receptor, whether it is through the formation of homodimers or heterodimers with another RTK.

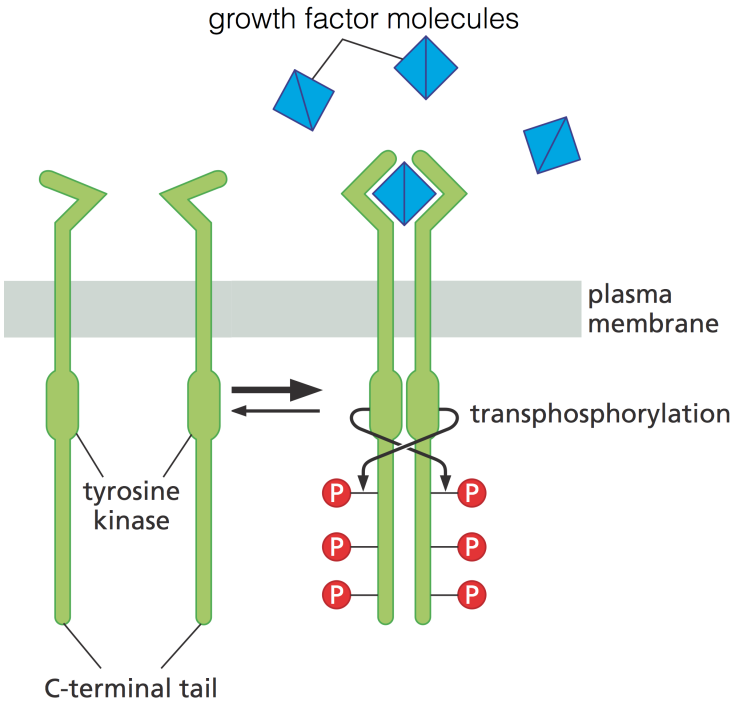


Figure I.1: *Signal reception and Growth factor receptors. In the absence of ligand, the growth factor receptors are present at the membrane in the form of monomers. Binding of corresponding ligand stabilizes a dimer linking the two tyrosine kinase intracellular domains leading to their transphosphorylation. (adapted from Weinberg, 2013)*

Several other families of membrane proteins function as receptor for extracellular signaling. For instance, the transforming growth factor-β family of ligands and receptors

have extremely diverse role depending on the initial cellular state. The notch juxtacrine signaling system functions by detecting ligands immobilized at the surface of the signal-emitting cell. The Wnt/Frizzled, Hedhgehog/Patched, integrins or Cytokine receptors represent some of the numerous other signaling systems each of which have different roles and effects.

While proteins and small molecules are the most studied cell-to-cell communication mediators, the secretion and reception of small non-coding RNA has also been described. In particular, cells can secrete small vesicles, exosomes, which contain non-coding RNA that can be internalized by another cell (or same cell in autocrine secretion). The regulatory function of the exosome-secreted RNA will directly alter the transcriptome of the targeted cell and in some cases alter its phenotype (Zhou et al., 2014). Other cases of secreted microRNA acting as ligand of specific membrane protein receptors have been reported (Fabbri et al., 2012).

Cellular signaling is meant to modify cellular phenotype and control cell behavior. The acquisition of a new cell state implies new cellular functions, which in turn entails the production and activation of new proteins to perform the required processes. New proteins can be synthesized by stabilizing or initiating the translation of pre-existing messenger RNA (mRNA). However, signal transduction usually involves transcriptional regulation and *de novo* synthesis of mRNA as a final step. Three types of signaling pathways can link the reception of the signal at the membrane to the nuclear transcriptional response, each of which are schematically represented in figure I.2. These three mechanisms are more thoroughly described hereafter to show the complexity of these cascades and the numerous cross-talk possibilities.

The first of these depicted strategy is often described as a signaling cascade and illustrates well the complexity of the processing of extra-cellular signals. The most widely studied of these pathways is the Mitogen-Activated Protein Kinase Pathway MAPK which involves three successive kinase reaction (see figure I.3). The MAPK pathway is one of the major signaling cascades activated by the reception of a growth factor signal (Dhillon et al., 2007). Following an RTK ligand binding, the recruitment of the Grb2 protein and the guanine nucleotide exchange factor SOS activates the RAS protein. This central signaling protein constantly switches from an inactive to an active state by binding either a GDP or a GTP. Once RAS is associated to a GTP and active, it can tether RAF, a MAPKKK (mitogen-activated protein kinase kinase kinase), to the cell membrane and activate it thereby triggering the MAPK signaling cascade. The RAF MAPKKK phosphorylates MEK, a MAPKK, which in turns phosphorylates ERK a MAPK which can proceed to the activation, through a final phosphorylation reaction, of transcription regulators enabling their transcriptional activity directly in the nucleus. Three types of MAPK cascade exist in Humans (Dhillon et al., 2007). Moreover, the activation of RAS can trigger several other signaling pathways, the Ral pathway controlling modification of the cytoskeleton and the Phosphatidylinositol 3-kinase (PI3K) including the activation of Akt/mTOR and the inactivation of the GSK-3β signal repressor. The activation of an RTK following ligand

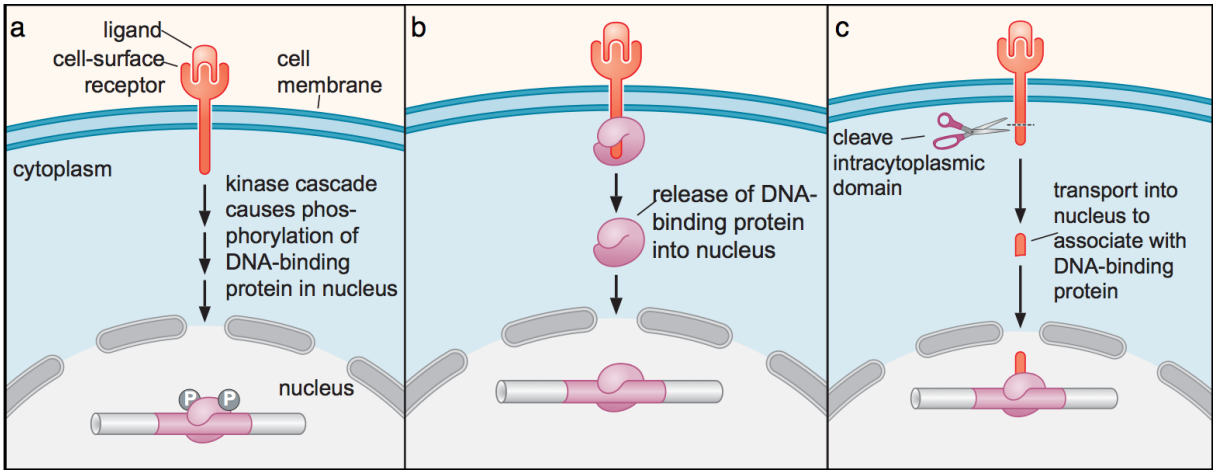


Figure I.2: *Signal transduction mechanisms. a. ligand/receptor binding triggers a signaling cascade leading to the activation of transcriptional regulators. b. Ligand binding releases a transcription factor that is then translocated to the nucleus. c. the intracellular domain of the receptor is cleaved upon ligand binding. The cleaved part of the receptor is imported in the nucleus and activates transcriptional regulators. (from Watson, Baker, and Bell, 2014)*

binding is only one of the events that can activate RAS. For instance, integrins can also activate the MAPK and the PI3K pathways.

The second signal transduction mechanism linking signal reception to transcriptional response requires the activation and translocation of the Transcription Factor (TF) from the cytoplasm to the nucleus. In the case of the Jak-STAT pathway, following the activation of cytokine receptors the jak tyrosine kinase transphosphorylates each other and recruit STAT proteins. The STATs are then phosphorylated and form a dimer, which is translocated to the nucleus where the dimer acts as a TF, which can regulate the expression of key genes. The TGF-β acts in a similar way by activating SMAD proteins forming Smad complexes, which are imported in the nucleus and regulate genes involved in major cellular processes such as proliferation and differentiation. In the case of the Wnt/β-Catenin pathway, the activation of the Frizzled and Dishevelled following the binding of Wnt inactivates GSK-3β otherwise sequestering the β-Catenin protein. Once released, β-Catenin is imported in the nucleus where it activates a set of transcription factors thereby regulating the expression of a broad number of genes.

Nuclear translocation of transcription factor is a crucial step in signal transduction. A Nuclear Localization Signal (NLS), a short sequence of amino acids, often determines the possibility for protein to be imported in the nucleus. The actual import of the protein can either be regulated by post-translational modification inside the NLS (*e.g.* phosphorylation) or by masking the amino acid sequence to the nuclear import mechanism. For instance, the NF-κB transcription factor is associated in the cytoplasm with IκB, which masks the NLS thereby sequestering NF-κB and preventing its nuclear translocation. In response to

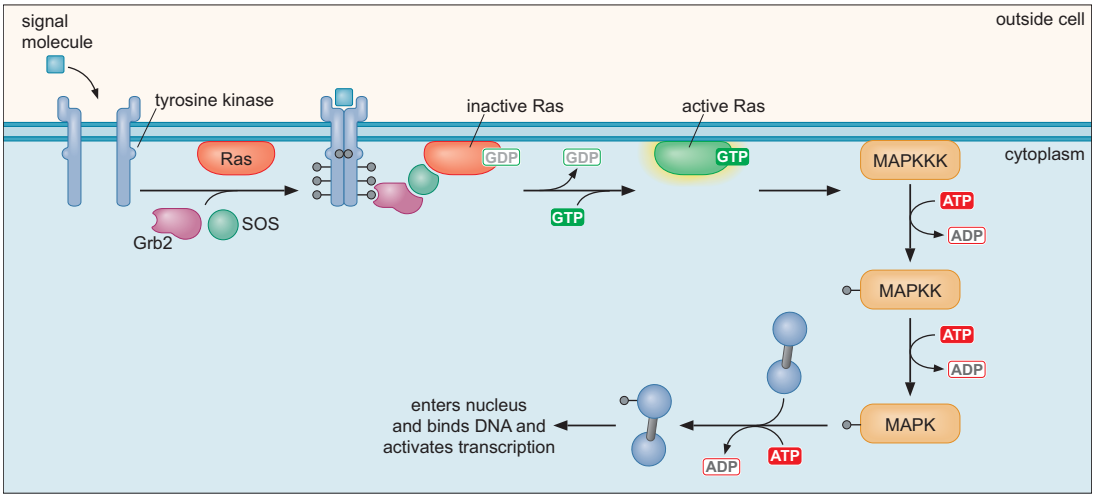


Figure I.3: *Growth factor and MAPK signaling pathways. The dimerization of RTK induces its activation. SOS and GRB2 are then recruited to activate RAS, which in turn fires a MAP kinase cascade. A MAP kinase which is able to phoshorylate another MAP kinase is therefore a MAP kinase kinase. Thus, the three successive phosphorylation steps involve a MAPKKK, MAPKK and finally a MAPK. The most downstream MAP kinase then activates a transcription factor, which can enter the nucleus and regulate gene expression. (from Watson, Baker, and Bell, 2014)*

diverse signals, the activation of IKK initiates the degradation of IκB thus unmasking the NLS of NF-κB, which is then imported to the nucleus where it becomes transcriptionally active (Ganchi et al., 1992).

The last and less common signal transduction strategy relies on the cleavage of the activated membrane receptor and transfer of the cytoplasmic part of the cleaved protein to the nucleus. This mechanism takes place in the Notch pathway in which ligand binding leads to a proteolytic cleavage of the receptor (Schroeter, Kisslinger, and Kopan, 1998). The cytoplasmic fragment is translocated to the nucleus where it associates with and activates a transcription factor complex (reviewed in Kovall, 2008).

An additional mechanism relying on the direct diffusion of signaling molecules to directly regulate the activity of a specific class of transcription factors called nuclear receptor is also discussed in section I.3.

I.2 DNA binding and Transcriptional activation

Independently of the ligand, the receptor or the signal transduction mechanism, nearly all signaling pathways eventually lead to the nucleus and to transcriptional regulation. This is the process by which a given gene is transcribed by the large protein complexes of the basal transcription machinery. Transcriptional activation is in essence the process by which the mediator complex, general transcription factors and eventually the RNA

polymerase II are recruited to the promoter of a gene. In eukaryotes, the accessibility state of the chromatin at the level of the gene promoter is critical to the initiation of its transcription (Li, Carey, and Workman, 2007). The chromatin state is determined by several factors: post-translational modification of histone tails, CpG DNA methylation and the action of epigenetic regulators in the form of proteins and non-coding RNAs. The interaction between transcriptional activators and chromatin modifiers is crucial to transcriptional activation as illustrated in figure I.4.

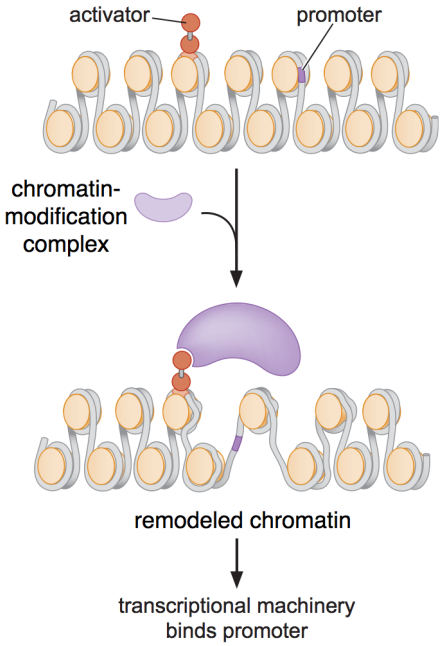


Figure I.4: *Activator directed chromatin remodeling. An activator protein is shown to bind to an activating element upstream of the promoter of its target gene in a region of high chromatin density. The activator then recruits a protein complex to increase the distance between nucleosome making the promoter DNA element accessible to the transcriptional machinery. Post-translational modification (for instance acetylation), interaction with chromatin-remodeling complex or non-coding RNA are examples of events which can alter the structure of the chromatin. (adapted from Watson, Baker, and Bell, 2014)*

Once the chromatin is accessible, activating transcription factor complexes are able to bind the enhancing elements in the promoter of their target genes. These DNA regulating elements are typically bound by several TF acting synergistically to activate the transcription of the gene downstream (Panne, 2008). TF synergy, sometimes termed cooperativity, resumes in the fact that the effect of two TF regulating a target gene in coordination is much greater than the sum of the effect of each of these TF taken separately. This combinatorial control of gene regulation is critical to the integration and processing of cellular signaling. Combinatorial regulation can process several simultaneous signaling

pathways by integrating the activating, and in some cases repressing, action of each of the pathways in the regulation of each gene. Moreover, it makes the response to a stimulus highly dependent on the cellular context by combining the signal-responding TF with the lineage-specific TF. As an example, let two genes A and B be activated by cell-type specific TFs respectively α and β, both acting synergistically with a stimulated TF ε. In this case, the transcriptional response to an ε-activating stimulus will depend on whether the cell receiving the signal expresses α, in which case it will activate gene A, β in which case it will activate B or both lineage specific TF α and β which will result in the activation of both A and B.

The combinatorial regulation is evidently much more complex for real human gene and is often exemplified using the human β-interferon enhanceosome for which the structure is presented in figure I.5. This DNA response element requires the coordinate binding of three different transcription activator complex: the Activating Complex 1 (AP1) composed of CJUN and ATF2, the interferon response factors which respond to interferon stimulus and finally the NF-κB complex composed of two subunits. In this example, the cooperation of these activating complexes is not done by direct protein interaction but rather through their binding to close DNA element and through their interaction with general coactivators CREB-binding protein or P300. Preliminarily to the assembly of the final activating complex, the architectural protein HMGA1 binds to the enhancer element and alters the shape of the double-stranded DNA fiber so that it can be bound simultaneously by so many protein complexes (Panne, 2008).

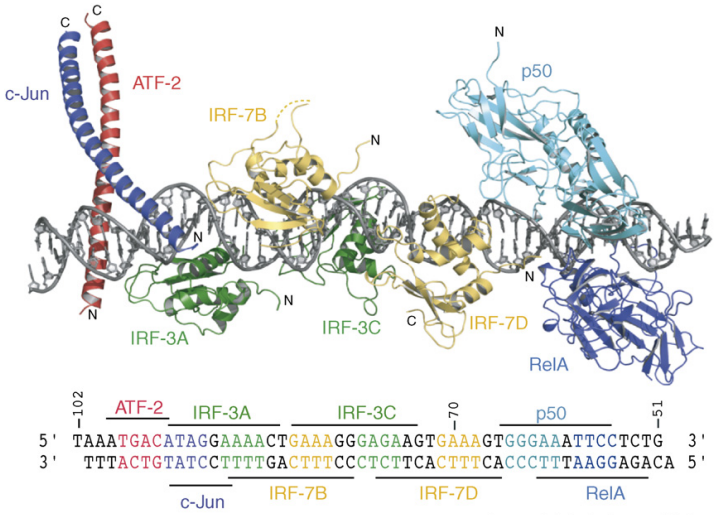


Figure I.5: *Structure of the β-interferon enhanceosome. Each of the DNA binding activators are color coded. Colors correspond to the sequence of their binding site in the lower part of the figure. DNA is shown in grey. (from Panne, 2008)*

The β-interferon enhanceosome is thought not to be an isolated case of higher eukaryote

regulation but rather an example of the regulatory complexity needed to control the diversity of cellular states and behaviors. The required combination of TF for the regulation of genes results in a tight control of the specificity in the transcriptional response following an environmental stimulus.

Transcriptional activation of genes results in the synthesis of a single strand messenger RNA from the DNA sequence template, which can eventually be translated to proteins. Although the translation step is dropped out for functional non-coding RNA, each of these steps describing the *central dogma* of molecular biology are regulated in the cell. The maturation, localization, stability and translation of mRNAs require the association of protein complexes and RNAs, which are regulated by the presence and activity of other proteins. Moreover, non-coding RNAs have been found to play an important role in mRNA maturation and more importantly in stability and translation. The most well described process of mRNA regulation by non-coding RNA is undoubtedly microRNA. These 21 to 23 nucleotide-long RNA associate with the RNA-Induced Silencing Complex RISC to target mRNA in a sequence specific manner. This micro-RNA/RISC/mRNA complex either results in the degradation of the target mRNA or in the inhibition of its translation (Esquela-Kerscher and Slack, 2006).

I.3 The particularity of Nuclear receptors

In order to sense and respond to signaling molecules in the form of protein ligands, a complex signal transduction pathway is necessary to import the signal from the extracellular space to the nucleus. As proteins cannot freely diffuse through cellular membranes, several steps often composed of protein interactions and post-translational modifications are usually necessary to convey the signal to the nucleus but also introduce as many possibilities for regulating the transduction of the signal as they are steps in the pathway. Cellular communications can also be carried by hydrophobic ligands with small molecular weights capable of diffusing through lipid membranes. Molecules such as steroid hormones, fatty acids or vitamin D can convey information without requiring complex signaling cascade between the cell surface receptors and the transcriptional regulators in the nucleus. Nuclear Receptor (NR) is a class of transcription factors able to bind these signaling molecules as a ligand and become transcriptionally active. The ligand-activation of NR can induce by its nuclear translocation, by post translational modifications or its dimerization with another NR as shown in figure I.6. Activated nuclear receptors typically bind to specific DNA response element and regulate their target gene in coordination with specific co-activators.

Nuclear receptors provide a shorten path between a signaling molecule and the targeted cell's transcriptional response. The lack of intermediary signaling cascade reduces the number of steps at which other signals can be integrated in the cell's decision to alter its phenotype. However, the actual transcriptional activation of genes remains dependent

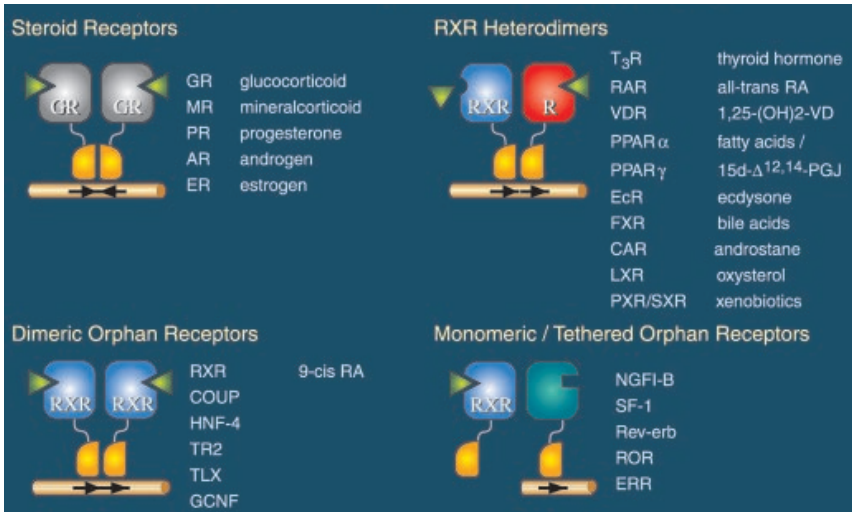


Figure I.6: *Nuclear receptor family. Nuclear receptors contain two main domains, a DNA binding domain composed of two zinc fingers (Yellow part) and a Ligand binding domain (Grey, Blue, Green or Red colored part of the nuclear receptor depending on its subfamily). The four sub-classes of nuclear receptors are based on their type of ligand and their ability to act as monomers, homodimers or heterodimers. (from Olefsky, 2001)*

on many factors. More importantly, cross-talk and competition between NR has been reported in numerous studies (Alimirah et al., 2012). In particular, NR that heterodimerize with the Retinoid X Receptor (RXR) compete in the formation of this functional dimer when RXR is quantitatively limiting (Wood, 2008).

These ligand activated transcription factor regulate many aspects of human physiology and development and more importantly have a major role in many diseases. As nuclear receptors are activated by small liposoluble ligands, a large number of therapeutic molecules have been developed to target and modulate the activity of NR. Chemicals that act as agonists or antagonists are clinically used to either reactivate a silenced NR or conversely inactivate pathologically active NR. For instance, the estrogen receptor (ER) is a master regulator and critical driver of a large sub-type of breast cancer. To inhibit ER activity, 4-hydroxytamoxifen, a chemical similar to its natural ligand estrogen, is used clinically and causes ER to associate with co-repressors instead of co-activators thereby inhibiting its transcriptional activity (Berry, Metzger, and Chambon, 1990).

I.4 Early and late response

In the precedent sections, the various mechanisms by which an extracellular signal is received and conveyed to the nucleus were discussed. The reception of such signals can directly impact cellular processes such as cytoskeleton organization or the metabolism.

However, major cellular changes usually involve the synthesis of new mRNA and proteins. In particular the commitment to a new differentiation state requires *de novo* synthesis of a wide array of proteins to carry out the cellular functions entailed to the newly acquired phenotype.

The production of new proteins species following stimulation is a two step process, the early and late response as depicted in figure I.7. The stimulation by a growth factor, cytokine or any signaling molecule results in the activation of a few transcription factors forming activating complexes. This activation induces the expression of a large number of genes including sometimes the activated TF itself forming a positive feedback loop (*e.g.* Wakabayashi et al., 2009). This first wave of transcriptional activation is called the early response. The first set of genes can also include transcription factor coding genes. These, once translated to functional proteins and activated if needed, can promote the activation of a second set of genes. This second wave represents the late response.

The early and late transcriptional response are particularly important to take into account for regulatory network inference (discussed in section IV.3) which is the task of defining the target gene of transcription factors often solely based on transcriptomic experiments. Measures of the transcriptome only capture modification of the set of transcripts in a given sample. Regulatory network inference therefore can only capture targets of a given TF if the mRNA level of this TF varies. In the case of the early response, the mRNA level of the TF remains unchanged since the activation of the TF is controlled at the post-transcriptional level for instance by its localization or through post-translational modifications, none of which are identifiable through transcriptomic measures. However, the transcriptional regulators at the origin of the late response have observable changes of their transcript level. Therefore, it is possible to infer the regulation of the late responsive genes by their effective regulatory TF.

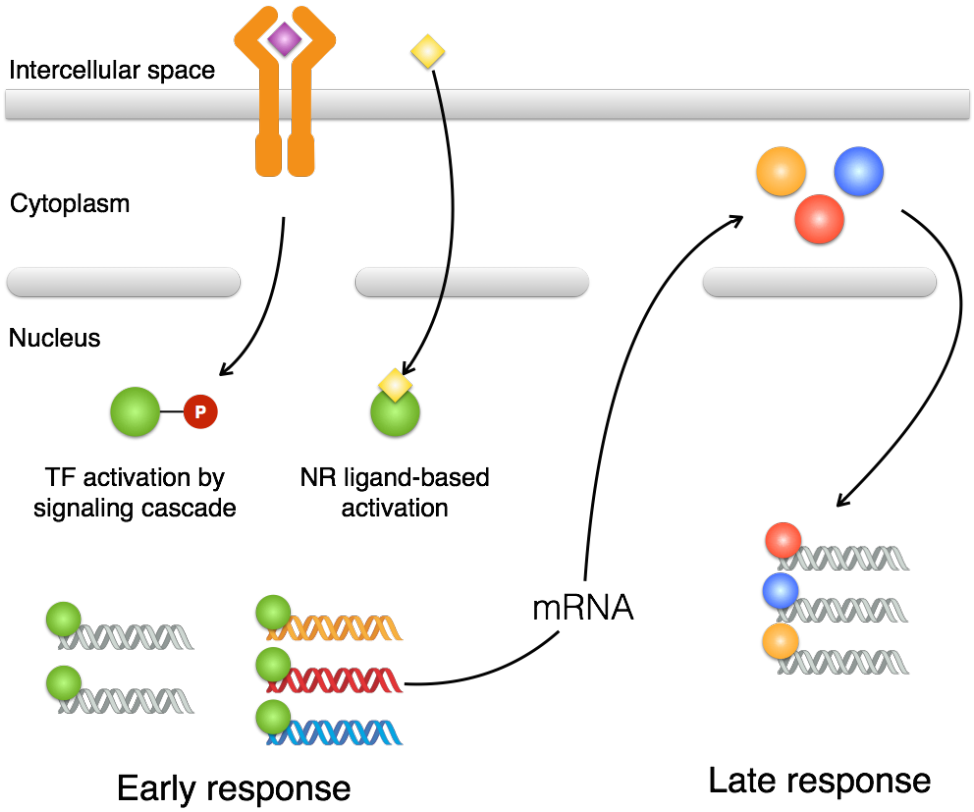


Figure I.7: *Early and late transcriptional response. Extracellular signals activate downstream transcription factor (TF) either through a sequence of reaction in a signaling pathway or by directly binding a nuclear receptor (NR) in the case of small hydrophobic molecules. The subsequent increased expression of genes by the activated TF (green) corresponds to the early response to the stimulus. Among the first set of activated genes are transcriptional regulator coding genes (blue, yellow and red DNA molecules bound by the green activated TF). The transcription and following translation of these genes into functional transcription factors, can eventually activate a second set of gene thereby generating the late response to the stimulus.*

Chapter II

Carcinogenesis and deregulation

II.1 Neo-plastic transformation

In 1909, Peyton Rous discovered a strain of avian virus capable of inducing sarcomas, a neoplasm arising in mesenchymal tissues, in chickens. It was named RSV, standing for Rous Sarcoma Virus. While this finding first brought the cancer research community to debate the classification of cancer as an infectious disease, it eventually was understood as the discovery of a tool to transform normal cells into tumor cells. Chicken embryonic fibroblasts infected with RSV were found to undergo radical phenotypic modifications and acquired characteristics observed in cancer cells (Temin and Rubin, 1958). For instance, these transformed cells continued growing in a Petri dish after reaching confluence whereas normal cells stop proliferating once a monolayer of cells is formed. This process is called contact inhibition in untransformed cells. Cellular transformation also caused an alteration of their morphology, the ability to proliferate in absence of growth factors and without any attachment.

More than a century of research lead to a deeper understanding of the cellular characteristics acquired during neo-plastic transformation, the process by which a normal cell becomes a malignant tumor-forming cell. In particular, an influential review enumerates the capabilities acquired by cancer cells leading to tumor growth and metastasis. These Hallmarks of cancer proposed by Douglas Hanahan and Robert Weinberg in their 2011 reference paper (Hanahan and Weinberg, 2011) are illustrated in figure II.1.

As tumors are masses composed of continuously proliferating cells, the foremost capacity acquired by cancer cells is their ability to proliferate in absence of mitogenic signals. As for normal cells, the mitogenic signals usually originate from the coordination of the cell population, which results in the secretion of growth factors (Thisse and Thisse, 2005). As discussed in chapter I of the introduction, these signals are integrated and processed so that normal cells can decide whether or not to enter the cell division cycle and start to proliferate. Conversely, cancer cells disrupt this high level process by which the organism ensures a

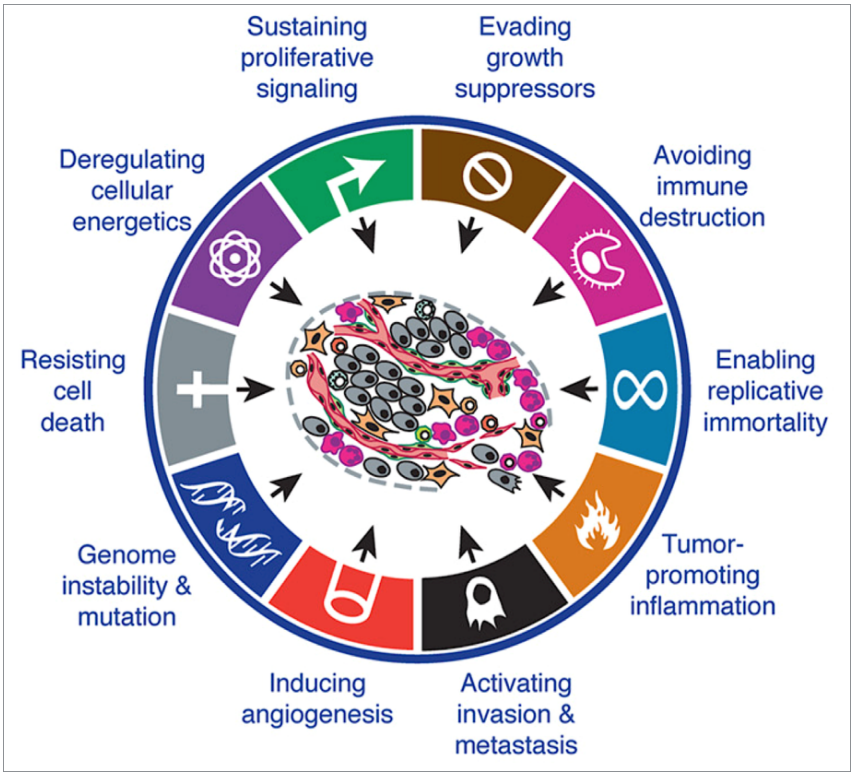


Figure II.1: *Hallmarks of cancer. Illustrated capabilities and characteristics acquired by cancer cells. (From Hanahan and Weinberg, 2011)*

constant and normal number of cells. This major step of malignant transformation is attained either by continuously producing growth signals themselves in an autocrine mode, by forcing other normal cells in their microenvironment to produce these signals or by continuously sustaining the intracellular signaling pathway normally induced by mitogenic signals (Giancotti, 2014).

Other general cellular processes allow cancer cells to acquire malignant characteristics. Tumors most frequently arise from epithelial cells and are termed carcinomas. During development stages as well as following the formation of a wound, epithelial cells can undergo Epithelial-Mesenchymal Transition (EMT) (De Craene and Berx, 2013). Transformed cells that undergo this regulatory program acquire the abilities to invade, form metastasis and to resist to programmed cell death, referred to as apoptosis. EMT is normally a temporary state that is followed by MET, Mesenchymal-Epithelial Transition, and more advanced stages of cell differentiation. However, tumor cells that underwent EMT remain in this mesenchymal phenotype suggesting that a malignant process maintains this cellular state.

Mitogenic signaling and EMT can be observed in normal cells during embryogenic development and following wounding (Thiery et al., 2009). Wound healing is also associated

to tissue inflammation, a process that has been described to promote tumor progression (Grivennikov, Greten, and Karin, 2010). These similarities between tumor progression and wound healing were observed in several different ways. For instance, Harold Dvorak detailed the similarities between the stroma of solid tumors and the tissue surrounding a wound. His observation lead him to describe tumors as *"wounds that do not heal"* emphazing the idea that tumors sustain normal processes to maintain growth (Dvorak, 1986).

Several of the hallmarks proposed by Hanahan and Weinberg rely on normal processes of wound healing, which involves cell proliferation and migration as well as vascularization of the wounded tissue (Hanahan and Weinberg, 2011). The constitutive activation of these normal processes support the concept that cellular mechanisms activated in cancer cells mostly rely on normal regulatory programs. Evidently, it is thought that core processes such as DNA replication and cell division are essentially conserved during neoplastic transformation. Instead of inducing a global reorganization of cellular functions, carcinogenesis implies sustained signals of several characteristics of normal wound healing and development as well as abrogation of negative feedbacks and safeguards, which includes inhibition of apoptosis, avoidance of immune destruction and evading growth suppressors.

Moreover, two of the proposed hallmarks cause the constitutive activation of characteristics relative to growth and inhibition of growth restrictions. These were identified as enabling characteristics and include tumor promoting inflammation and genetic instability of which the impact on carcinogenesis is discussed in the next section.

II.2 Cancer is a genetic disease

The identification by Peyton Rous of tumor initiating virus's lead, years later, to another stunning discovery: the tumor was induced by the transcription of a viral gene that has a close homolog in the host genome. This suggested that regardless of the physiological mechanism underlying neo-plastic transformation, cancer is a disease of genes and it looked like only small modifications of the original copies of a gene could induce tumors.

While the discipline tumor virology was being created, particular chemical and physical agents were discovered to have the ability to induce cancer. Several organic compounds, mostly extracted from coal tar, were found to be highly carcinogenic. The carcinogenicity of X-rays was also described as well as its ability to introduce modifications in the genome of exposed cells. The latter discovery revealed that genetic information could be modified: genomes are mutable. From there, studies of carcinogenic chemicals showed that these were mutagenic compounds as well. Altogether, these observations were indicative that cancer arises from alteration of the genetic information. This new perspective was supported in 1960 by the discovery of an abnormal karyotype (the description of chromosome number and shape in a cell) common to nearly all cases of Chronic Myelogenous Leukemia. This alteration causing an exchange between the two long arms of a copy of chromosome 9

and 22 results in a characteristically shorten 22 chromosome entitled the Philadelphia chromosome (Nowell and Hungerford, 1960).

Since then, we learned that the translocation involved in the formation of the Philadelphia chromosome takes place in the middle of two coding regions (Hermans et al., 1987). A homolog of one of these genes, *abl*, was also found in a murine tumorigenic virus (Daley, Van Etten, and Baltimore, 1990). The study of several other tumor forming viruses revealed that most of these carried a homolog of a human normal gene. The common grounds of tumorigenic viruses and oncogenic genetic alterations became the genes that were encoded in virus or targeted by genetic alterations. The genes carried by these viruses were termed oncogenes. Their normal homologs present in the genome of non-transformed healthy cells are called proto-oncogene, has they have the potential to become oncogenes consequently to specific genetic alterations.

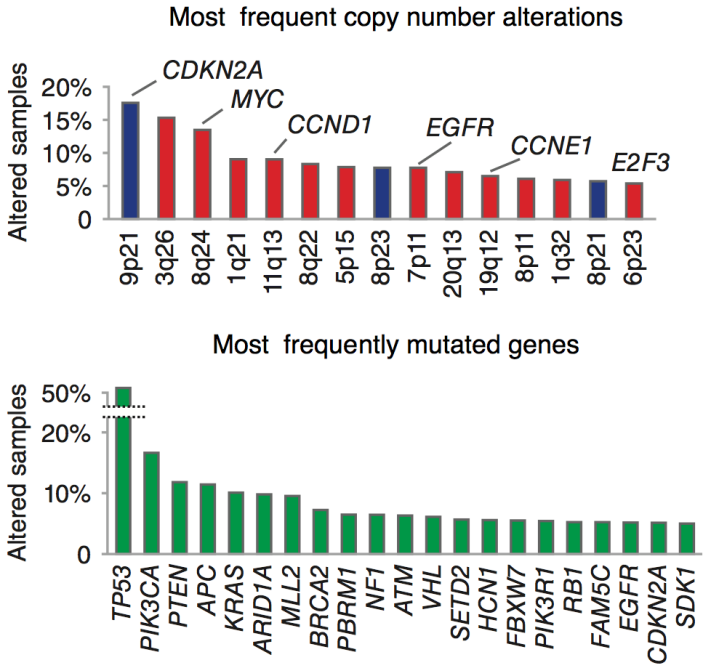


Figure II.2: *Frequent genetic alterations. Based on the profiling of 3,299 tumors from 12 cancer types by the TCGA (The Cancer Genome Atlas) consortium. The upper panel shows segments of chromosome and their relevant associated gene that is most often affected by copy number alterations. Blue bars represent losses. Red bar represent gains in copy number. The lower panel presents the genes most frequently found with point mutation in their coding sequences. (from Ciriello et al., 2013)*

The further study of tumor forming viruses resulted in a broader list of proto-oncogenes encoded in the human genome. In the case of non-viral caused tumors, it is now known that most cancer arises from genetic alterations in normal cells, which promote their malignant transformation. Advances in genome-wide techniques and the analysis of thousands of

cancer genomes also revealed a set of frequently altered genes. Figure II.2 summarizes the most frequent alterations found by the TCGA (The Cancer Genome Atlas, a consortium intending to profile most human cancer types). These alterations can result from, or actually cause, the acquisition of a mutator phenotype of genome instability increasing the probability of chromosome alteration and mutation events (Loeb, 2001). Although these alterations target more or less random regions of the genome, the ones providing the fittest advantage for tumor growth and progression is selected in a Darwinian fashion. This process explains one of the most important enabling characteristic proposed as a Hallmark of cancer as it can, through this selection process, cause the acquisition of most if not all of the other characteristics of neo-plastic transformation.

II.3 The diversity of genetic alterations

Tumor inducing retroviruses can contain the mRNA of an oncogene. In normal cells, the expression of the corresponding proto-oncogene is tightly controlled to respond to given stimuli. This is done to a certain extent by specific DNA elements in its promoter controlling the recruitment of the transcriptional machinery and thus the transcription initiation rate. For instance, housekeeping genes are controlled by promoter elements causing a high and constant level of transcription. In the case of an oncogene acquired by the infection of a retro-virus, the retroviral promoter upstream of its coding sequence results in a much higher and constant level of expression. Oncogenes of retro-viral origin or proto-oncogenes under the control of a promoter of viral origin, are defined by pathological level of expression with a foreseeable effect on their downstream effectors. This is only one example of the mechanisms by which tumor viruses bypass the way a proto-oncogene is normally regulated, usually only being expressed and activated for a short period of time, in most cases determined by the length and strength of the stimulus. Interestingly, this is also one of the effect of tumor-driving genetic alterations.

Genetic alterations are modifications of the information contained in the DNA molecules of a cell. In normal cells, genetic aberration occur randomly and trigger repair mechanisms or cell death when the level of damage is unmanageable. The acquisition and maintenance of genetic alterations in cancer cells therefore greatly participate to malignant transformation. These modifications include point mutations and chromosome alterations.

Point mutations

Point mutations only affect a few, often one, nucleotide in the DNA sequence.

The substitution of a nucleotide inside a gene's coding sequence can have several implications and are classified into four categories:

- Missense mutations resulting in the modification of the amino acid sequence of the translated protein

- Non-sense mutations introducing a stop codon (TAA, TAG, TGA), thereby shortening the protein if it is translated

- Silent mutations, which will have no consequence on the protein sequence yet may impact post-transcriptional processes such as the stability of the mRNA

- Non-stop mutations substituting a stop codon into an amino acid coding codon

As an example of missense mutations and its potential effect, figure II.3 illustrates a mutation frequently found in bladder cancer and affecting the gene coding for the growth factor receptor FGFR3.

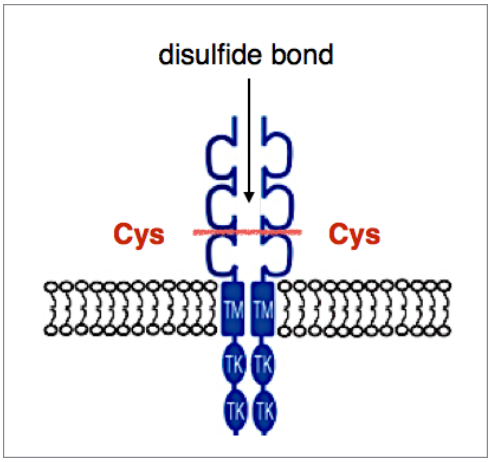


Figure II.3: *Missense FGFR3 mutations. A nucleotide substitution in the 249th (R249C) or 248th (S249C) codon can be found in bladder cancer and melanomas and results in the replacement of the original amino acid by a Cysteine. In normal cells, FGFR3 homodimerizes in the presence of ligand which activates the cytoplasmic Tyrosine Kinase domain (TK, below the trans-membrane domain TM). In cells presenting these mutations, the cysteins are able to form a disulfide bond causing a constitutive dimer and activation of the kinase activity.*

The effect of insertions and deletions, generally referred as indels, of nucleotides in the coding sequence of a gene depends on the number of nucleotides affected. Indels of three nucleotides or multiples of three are called in frame indels and will simply add or remove amino acids from the final synthesized protein which can have variable effects on its function. Indels affecting a number of nucleotides that is not a multiple of three are termed frame shift indels. As expected by their name, these alter the reading frame of the mRNA and change virtually all downstream codons. In most cases frame shift indels result in dysfunctional and/or shorten proteins.

The deletion, insertion or exchange of a nucleotide can also affect many other processes in the regulation of a gene whether they appear inside or outside the translated or even the transcribed sequences. Modifications in the promoter and DNA response element can alter the recruitment of regulatory proteins or RNA and therefore have consequence on the expression of the downstream gene (Labussière et al., 2014). Modifications inside and

around the coding sequence can alter the maturation of the mRNA and in some cases induce aberrant splicing (Liu et al., 2001). Other alterations of the transcribed sequence can affect the affinity for mRNA stabilizing proteins, ribosome-recruiting proteins, regulatory RNAs such as miRNA or simply affect the secondary structures encoded in the mRNA sequence. Beside the direct modification of the coding protein sequence, most of these effects are usually difficult to identify.

In fact, the addition, loss or the substitution of a single nucleotide into another often has no or only little effect. For instance, a non-sense mutation or a frame shift indel is supposed to be of great significance but can occur in a gene that is not expressed neither in normal or pathological conditions of the studied cell type. These inconsequential mutations are frequent especially in cancer with high genomic instability. These were termed passenger mutations and are opposed to driver mutations that have higher and sometimes dominant pathological effects. This concept of passenger mutation is recurrent in the scientific literature of high-throughput sequencing. However, arguments to classify mutations as passenger often rely on weak characteristics such as frequency of occurrence and fail to take into account mutations that do not affect the amino acid sequence of proteins but their abundance.

Chromosome alterations

Larger alterations of the genetic information can involve chromosomal regions of several kilobases - under which the border with indels becomes difficult to define - or an arm of a chromosome and up to an entire chromosome. The maintenance of aberrant forms of chromosomes during cancer cell division often requires some tweak in the mitotic process to allow mitosis to deal with chromosomes altered in form and number. This is particularly true for alterations affecting large portions of chromosomes often resulting in aneuploid cells. Alterations in copy number and chromosome rearrangements define two main categories of alterations. These two classes do not differ in terms of the mechanisms by which they occur in cancer cells, which mainly involves a break in the DNA strand. However, these alteration events have rather different effects on genes and their expression into functional, and sometimes dysfunctional, proteins.

A defect in chromosome segregation during mitosis can result in the gain or loss of a chromosome in the daughter cells. In some cases the faulty segregation affects an entire chromosome resulting in trisomy, in the cell gaining a chromosome, or monosomy in the case of a loss. Copy number aberrations (CNA) can also affect focal regions of a chromosome.

An immediate consequence is the altered number of copies of genes present on the parts of the chromosome affected by this alteration. Normal cells have two copies of each gene present on autosome pairs. The effect of the loss of one of these copies is highly variable. In some cases it will have no effect, either because a single copy is sufficient for the normal functioning of the gene product or simply because none of the two copies are

normally expressed. Other cases range from a 50% to a 100% decrease in expression in the case were only the lost copy was normally expressed. This can be due to differentially expressed alleles of a same gene depending on their parent of origin, a phenomenon called genomic imprinting (Joyce and Schofield, 1998). Conversely the effect of the gain of the part or of the entire chromosome is simpler in that it increases the number of potentially transcribed loci and therefore the potential number of proteins encoded by genes usually present in only two copies. Figure II.4 illustrates the effect of copy number alteration on the expression of 20 genes present in regions of frequent copy number aberrations of 131 urothelial carcinomas.

Finally, loss of heterozygosity (LOH) involves the loss of the arm or of an entire chromosome and the duplication of the remaining copy without any modification of the overall copy number. LOH is a copy-neutral chromosome aberration mainly implying the presence of the two same copies of genes present in the locus subjected to LOH. This mechanism was described in retinoblastoma by the Knudson two-hit hypothesis (Knudson, 1971). The first hit is an inactivating mutation of the *RB1* gene. The second hit is then the loss of part or the whole chromosome 13 carrying the wild type *RB1* gene and a duplication of the concordant locus of the chromosome carrying the mutated *RB1* (Kato et al., 1993). This results in cells homozygous for the mutated form of *RB1* and therefore the double inactivation (of both functional alleles), a necessary step for tumorigenesis.

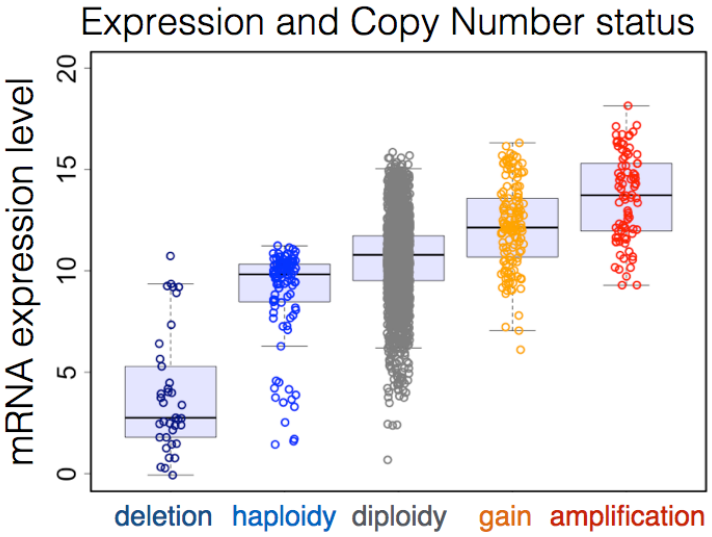


Figure II.4: *Effect of copy number on gene expression. The expression of 20 genes recurrently found in chromosomal regions of copy number aberration in bladder cancer is reported. The values of expression, measured by RNA-sequencing and* $\log_2$ *transformed, of a gene in a sample is plotted as a function of the copy number status of the same gene in the same sample. (from Cancer Genome Atlas Network, 2014)*

Extreme cases of aberrant copy number alterations include the deletion of both copies

of a chromosome locus or the amplification of a locus resulting in often more than 5 copies. The amplicon, the copies of an amplified locus, can remain on the chromosome from which it was copied. Amplicons can also be part of an independent particle able to perpetuate itself throughout cellular divisions. These extra-chromosomal particles are called double-minute chromosomes. Chromosome alterations analysis in nearly 5,000 tumors also revealed that more than one third of tumor cells undergo whole genome duplication (Zack et al., 2013).

Chromosome rearrangements are simply displaced sections of chromosomes. These usually result from breaks in the DNA strand and erroneous ligation of the detached segments. In some cases, the segment is simply reintegrated upside-down at the same position in the chromosome. In cases similar to that of the Philadelphia chromosome, the extremities of the arms of two different chromosomes are exchanged. These events do not affect the number of copies of genes, although cases of amplification of these aberrant chromosome segments have been reported. The particularity of these events is to create breaks and mending back up together sequences of DNA that are not normally near each other.

The main pathological effect of chromosome rearrangement is gene fusion. This happens when the breaks occur inside genes. The reorganized segment of DNA results in a sequence coding for new gene, mRNA and sometimes a new protein that does not exist in normal cells. Gene fusion can create new gene products sharing functions from both of its original genes. The addition of new domains to pre-existing protein or creation of a shorten protein can significantly alter its function. This is well exemplified by the frequent fusion of the *FGFR3* and *TACC3* genes in bladder cancer (Williams, Hurst, and Knowles, 2013) and glioblastoma as illustrated in figure II.5.

Gene fusion can induce neo-plastic transformation without involving the coding sequence itself. Most of the miRNA regulate mRNA by binding the 3' untranslated regions (UTR) of their targets. The fusion between the 5' UTR and coding sequence of a proto-oncogene with the 3' UTR of another gene can result in the resistance of the proto-oncogene to miRNA silencing thereby abnormally increasing the mRNA expression level (Parker et al., 2013). Other cases of gene fusion with an unrelated promoter have been reported. This can substantially modify the expression pattern of a gene. In most cases of Burkitt's Lymphoma, a translocation involving the proto-oncogene *MYC* was reported. This rearrangement causes a copy of *MYC* to be set downstream of the promoter controlling the expression of the immunoglobulin heavy-chain gene (Janz, 2006). In normal cells, *MYC* is activated only in response to specific stimuli and in particular of growth factors. These signals are normally temporary and their absence results in a silencing of *MYC*. However, the immunoglobulin heavy-chain gene is highly expressed in a cell-specific but constant manner. Therefore, the regulation of *MYC* by this constitutive enhancer overflows the cells with this transcription factor of which the major function is to activate cellular proliferation.

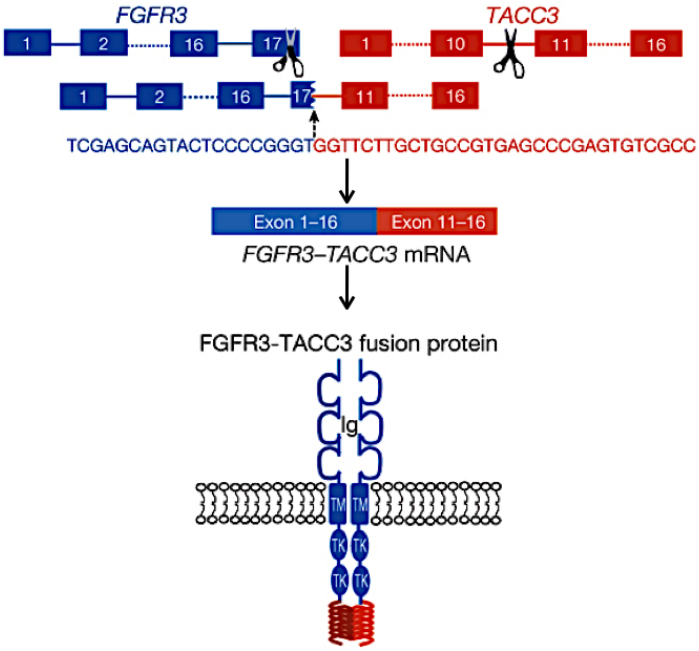


Figure II.5: *FGFR3/TACC3 fusion in bladder cancer. Reported in several tumors of the bladder and in bladder cancer cell lines. The breaks occur in the 17th exon of FGFR3 and the intron between exon 10 and 11 of the TACC3 gene. The new gene product is a protein very similar to the normal FGFR3 protein with an additional helical domain in C-terminal. The addition of this TACC3 originating domain is thought to contribute to a ligand-independent dimerization and subsequent activation of the growth receptor. Ig: Immuno-globulin like domain. TM: trans-membrane domain. TK: tyrosine kinase domain. (from Cancer Genome Atlas Network, 2014)*

Structural and regulatory effect of genetic alteration

Genetic alterations can transform a proto-oncogene into an oncogene either by altering the structure of the encoded protein or its regulation and therefore its quantity. The simultaneous alteration of both the expression and the structure of the encoded protein have been reported. This is the case when the amplified locus carries a gene containing a mutation affecting the structure and function of the encoded protein.

The family of growth factor receptors, also called receptor tyrosine kinase (RTK), is affected by a wide array of genetic alterations. Moreover, their extensive study produced models of the effect of several types of genetic alterations. RTK are activated subsequently to their dimerization or in the case of the insulin receptor family trough stablisation of an active dimer, which is normally caused by the binding of specific ligand. As illustrated in figure II.3, point mutations can induce changes in the amino acid sequence of RTK causing their constitutive dimerization and activation. Again, with the example of *FGFR3*

in bladder cancer, the fusion of an RTK with another gene can result in the addition of a new protein domain causing a constitutive activation of the receptor. The amplification of a growth factor receptor, which was frequently observed for *EGFR* and *ERBB2*, can result in a substantial increase in its expression. A high number of RTK proteins present at the surface of cells increases the chance of random ligand-independent dimerization or can simply reduce the quantity of needed ligands to activate the downstream signaling pathway. In these cases, the amplification also results in constitutive activation of the protein.

Finally, genetic alterations can indirectly activate oncogenes. For instance, the amplification and over-expression of growth factors within a cell can result in autocrine signaling and constitutive activation of the corresponding growth factor receptors. This incidental effect implies that more important than the effect of a mutations on a gene, is the effect on the genes regulated by it, on *downstream* pathways.

II.4 Carcinogenesis or the deregulation of signaling circuits

The conversion of a proto-oncogene into an oncogene requires direct genetic alteration of the encoding gene and causes a constant activation of the protein. In the case of growth-related oncogenes, the constitutive activation state is equivalent to that of found in normal proliferating cells following mitogenic stimuli. This resemblance suggests that most of cancer malignant abilities are acquired by abnormal maintenance of normal processes of proliferation. Based on this, we can assume that the regulatory circuits controlling cell proliferation, which are stimulated during developmental processes or wound healing, are simply sustained during tumorigenesis. Therefore, the abnormal activation by oncogenic alterations of nodes in these signaling pathways should be capable of driving tumor progression.

This hypothesis is greatly supported by the high frequency of genetic alterations targeting central nodes in mitogenic signaling pathways. A simplistic view of mitogenic responding signaling cascades is composed of successive steps of enzymatic reactions activating proteins one after the other. Although these pathways are more thoroughly discussed in section I, the sequence of reactions involved in the MAPK cascade is listed below.

1. a growth factor binds to a corresponding RTK
2. the RTK activates RAS through the GRB2 and SOS proteins
3. RAS activates RAF, a MAPKKK, by transferring a phosphate group
4. RAF activates MEK, a MAPKK, by transferring a phosphate group
5. MEK activates ERK, a MAPK, by transferring a phosphate group
6. ERK phosphorylates MYC, a transcription factor engaging the cell into a proliferation states

As this cascade is here described as a simple sequence of positive regulatory steps, the constitutive activation of any of the proteins involved in any of these steps should be sufficient to maintain a proliferation state. Indeed, these proteins are known proto-oncogenes for which activating alterations were frequently identified (see II.2 for approximate frequency in more than 3,000 tumors). The proportion of alterations of each of the nodes in the pathway is dependent on the tissue-origin of the tumor. However, most cancer types were identified with amplification or mutations of growth factor receptors (*EGFR*, *ERBB2*,...), Ras activating mutations (in *KRAS*, *HRAS* or *NRAS*) and Myc amplifications (of *CMYC*, *NMYC* or *LMYC*) representative of key steps of the MAPK signaling cascade.

The occurrence of genetic alterations inducing constitutive activation of proteins at different levels of these mitogenic signaling pathways support the idea that cancer cells use these circuits to proliferate. More importantly, genetic alteration seems to be the fuel that sustains the unceasing activation of these pathways in cancer.

As stated earlier, these signaling cascades are here described in their simplest form. A more detailed but still incomplete model of pathways involved in carcinogenesis is shown in figure II.6. This portray of cellular circuits barely illustrates the complexity of the set of chemical reactions underlying cell signaling. However, the proposed diagram is an understandable model in which the consequences of the aberrant activation of key nodes, Ras for instance, can lead to the acquisition of several characteristics essential to malignant transformation.

Several of the Hallmarks of cancer proposed by Hanahan and Weinberg were left behind in this section. While the *sustained proliferative signaling* has been here thoroughly discussed and explained by genetic alterations, other characteristics such as resistance to cell death and evading growth suppressors were not addressed.

When DNA suffers a great deal of damage, such as that necessary to induce chromosome alterations, normal cells have a safeguards that halt the progression of the cell cycle until the damage is repaired. This is also the case when other stresses occur in the cell such as insufficient quantities of nucleotides, glucose or other key cellular metabolites. In extreme cases, when the DNA is damaged beyond repair, these gatekeepers can induce apoptosis. A specific protein acts as an input of all the alarm signals, TP53. Given that cancer cells require aberrant cellular conditions to continuously proliferate, a normal version of TP53 should be activated and induce cell death or arrest the proliferation in virtually all transformed cells. By its ability to interfere with tumor progression, *TP53* is a Tumor Suppressor Gene (TSG). Moreover, *TP53* is the most frequently mutated gene in all cancer type (see figure II.2). Unlike mutation described to activate RTK in the previous section, *TP53* bears inactivating mutations such as non-sense mutations.

A full comprehension of neo-plastic transformation implies a thorough understanding of the signaling pathways constitutively activated or repressed by oncogenic alterations. This includes not only the critical nodes that are altered but also the entire downstream effectors that are affected as both of these are potent clinical targets and can explain the

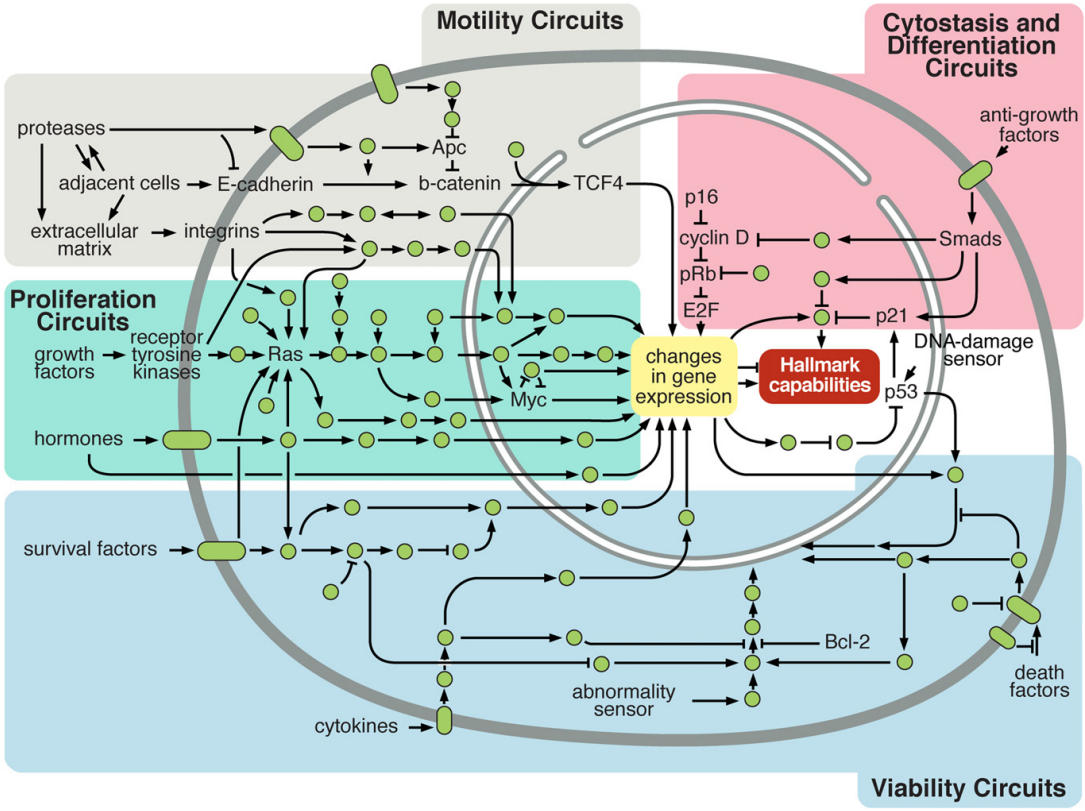


Figure II.6: *Carcinogenic Signaling circuits. A simplified version of the pathways aberrantly regulated to drive neo-plastic transformation. Key nodes only are identified by the name of the protein. The circuit is divided into four categories directly related to Hallmarks of cancer. (from Hanahan and Weinberg, 2011)*

acquired malignant phenotype.

Chapter III

Urothelial carcinoma

The urothelium is the epithelial lining of the urinary bladder. It serves as a barrier to the toxic content of the urine and to the invasion of pathogens. Urothelial cells have a very low turnover rate of approximately 6 weeks in mice Lewis, 2000 and six months to one year in human (Varley et al., 2005). However, in response to injury, the urothelium rapidly enters a highly proliferative state (Kreft et al., 2005; Varley et al., 2005).

The urothelium is composed of three layers of cells, the basal, intermediary and superficial layers of cells as depicted in figure III.1. Depending on the stretched state of the bladder, the thickness of the epithelium ranges from three to eight cells depending on its distension state (Staack et al., 2005). The basal layer of cells is assumed to be germinal cells (Lewis, 2000) although mitosis has been observed in both the basal and intermediary layers (Varley et al., 2005). The superficial layer of cells shows unique differentiation characteristics by expressing four kinds of uroplakins (Ia, Ib, II and III, see III.1) which form rigid plaques called asymmetric unit membrane (AUM) plaques. These highly specialized terminally differentiated urothelial cells are called *umbrella* cells as they make the urothelium one of the most effective and impermeable epithelial barrier (Wu et al., 2009).

III.1 Epidemiology and clinical aspects

In 2012, 118.900 men and 32.900 women were diagnosed with bladder cancer in europe making it the 5th cancer in terms of incidence (Ferlay et al., 2013). Approximately 50% of all bladder cancers are associated with cigarette smoking as the risk of bladder cancer is increased fourfold for smokers (Clavel et al., 1989). Another risk factor is exposure to chemicals such as aromatic amines, benzidine, β-naphthylamine or aromatic hydrocarbons from coal, all found in specialized industries and associated with a thirty-fold increased risk of bladder cancer. Finally, the tropical parasite *Schistosoma haematobium* causing bilharzia (also termed schistosomiasis) can also induce cancers of the urinary bladder and

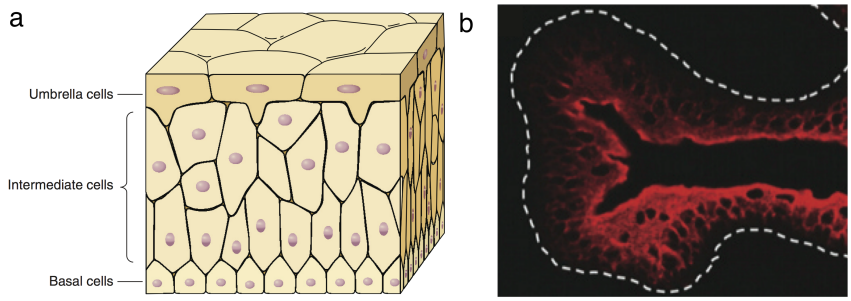


Figure III.1: *Structure of the bladder urothelium. a. Illustration of the multiple epithelial cell layers composing the urothelium. b. Immunohistochemical staining of uroplakins in mouse urothelium, mainly expressed in umbrella highly-differentiated urothelial cells. (from Birder and Andersson, 2013)*

in fact makes bladder cancer the first cancer in Egypt (Sengupta, Siddiqui, and Mumtaz, 2004).

Most tumors of the urinary bladder are urothelial carcinomas (90% to 95%). Based on physical and histological examination, imaging and endoscopy, tumors are systematically classified using the TNM system, standing for Tumor, Nodes and Metastasis. The 2009 consensus staging of bladder cancer is described in figure III.2. Tumors presenting a T2 stage or more are referred to as muscle-invasive cancers as opposed to superficial or non-muscle-invasive Ta and T1 tumors. The TNM status has a significant impact on prognosis with for instance, a five-year survival rate dropping from 85% for T2 to 25% for T4 cancers and a mean survival of six to twelve months for metastatic bladder cancer (Sengupta, Siddiqui, and Mumtaz, 2004).

Non-muscle-invasive tumors represent nearly 80% of cancers at diagnosis and have a variable prognosis depending on the phenotype of tumors cells described by tumor cell grading. The histological description of tumor cells is a grading of its differentiation state. Tumor grade is assigned based on tissue abnormalities such as the polarity of cells, the state of the nucleus and number of observed mitosis. In bladder cancer, two types of grading are used and are referred to as the 1973 and 2004 WHO (World Health Organization) consensus grades. The overlap between these two grading system is depicted in figure III.3.

Histological grading is also of major significance in terms of prognosis and clinical care. While only 35% papillary urothelial neoplasm of low malignant potential (PUNLMP) develop recurrences, 70% of low and high grade Ta tumors do (Holmäng et al., 2001). Moreover, disease progression of PUNLMP was nearly never observed. Low grade Ta tumors rarely progress (approximately 5%) whereas high grade Ta tumors progress in nearly one of four cases (Holmäng et al., 2001). The high recurrence rate of Ta and T1 (80%) tumors requires frequent cystoscopy and lifelong follow-up. Therefore, the cost per patient of bladder cancer from diagnosis to death is the highest of all cancers (Botteman

T - Primary Tumor
The suffix *"is"* is added to indicate presence of *caricnomas in situ.*

TX	Primary tumor cannot be assessed
T0	No evidence of primary tumor
Ta	Non-invasive papillary carcinoma
Tis	Carcinoma *in situ*: 'flat tumor'
T1	Tumor invades subepithelial connective tissue
T2	Tumor invades muscle:
	T2a superficial muscle (inner half)
	T2b deep muscle (outer half)
T3	Tumor invades perivesical tissue
	T3a microscopically
	T3b macroscopically
T4	Tumor invades peri-vesiculare structures:
	T4a prostate stroma, seminal vesicles, uterus or vagina
	T4b pelvic or abdominal walls

N - Regional Lymph Nodes

NX	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastasis
N1	Metastasis in a single lymph node in the true pelvis
N2	Metastasis in multiple lymph nodes in the true pelvis
N3	Metastasis in a common iliac lymph node(s)

M - Distant Metastasis

MX	Distant metastasis cannot be assessed
M0	No distant metastasis
M1	Distant metastasis

Figure III.2: *Urinary Bladder cancer TNM classification. 2009 consensus in bladder cancer TNM classification and histopathological staging.*

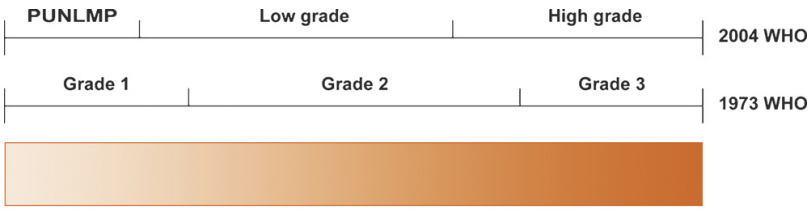


Figure III.3: *Bladder cancer grading system. Tumor grades range from low to high cellular abnormality (left to right). The 1973 grade 1 are mostly reassigned to Papillary Urothelial Neoplasm of Low Malignant Potential (PUNLMP) as well as to low grade tumors. Similarly, 1973 grade 2 are reassigned to low and high grade whereas grade 3 are all assigned to high grade tumors.*
WHO = World Health Organization

et al., 2003). While muscle-invasive tumors are systematically high grade cancers, T1 tumors are high-grade (2004 WHO) or grade 2 and grade 3 (1973 WHO) and generally progress in 60% of cases and show a 35% 10-year survival rate (Eble, 2004). Finally, among non-muscle-invasive bladder cancers, Carcinomas *in situ* (CIS) are high grade flat neoplasms. CIS have a more frequent progression rate than papillary tumors (40% to 50%) and are often associated with in muscle-invasive tumors (Eble, 2004).

Interestingly, PUNLMP present mostly diploid cells whereas non-muscle-invasive high grade papillary tumors and especially CIS cells are often aneuploid. These features of genetic instability are much more important in muscle-invasive bladder cancer which is among the most genetically instable cancer (Lawrence et al., 2014).

Bladder cancer treatment is dependent on the stage of the tumor. Tumor resection followed or not with endovesical chemotherapy or immunotherapy (BCG therapy) are used for non-muscle invasive tumors. In absence of metastasis, cystectomy is performed for muscle-invasive tumors, associated or not with systemic chemotherapy. To this date, no targeted therapies are clinically used for bladder cancer.

III.2 Contrasting bladder cancer progression pathways

Classifying cancer events of the urinary bladder by differentiating non-muscle-invasive superficial tumors and muscle-invasive tumors has major clinical significance in terms of prognosis. Moreover, histo-pathological descriptions of tumors also define two opposite progression pathways of bladder tumors with distinct progression pathways: Ta papillary tumors and carcinoma *in situ* (CIS). The history of progression and recurrence is illustrated in figure III.4.

Papillary tumors are associated with frequent *FGFR3* activating mutations (Bakkar

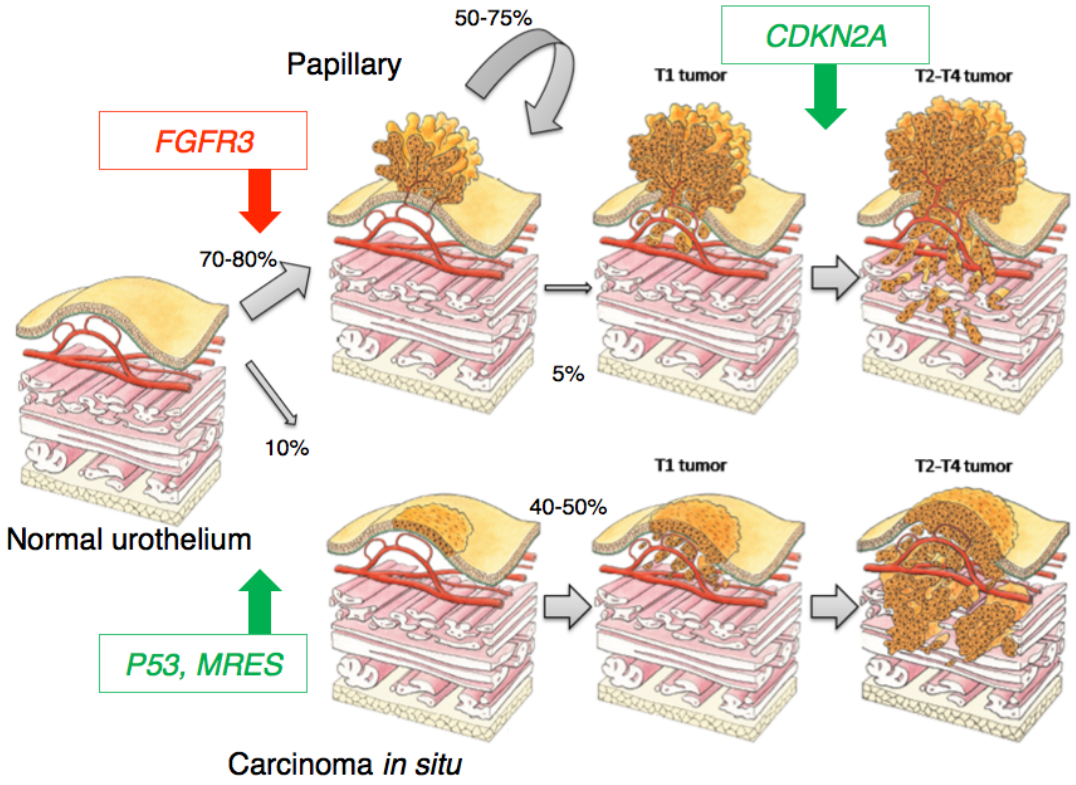


Figure III.4: *Bladder cancer progression pathways. Two major type of tumor arise from the normal urothelium. The most frequent is papillary tumors (70 to 80 %) followed by Carcinoma in situ (CIS: 10%). The remaining is mostly of unknown origin. FGFR3 activating (red) mutation is a frequent event in papillary tumors which themselves recur in 50% to 75% of cases. Progression is rare (5%) and often associated with CDKN2A loss (green). The CIS pathway is associated with TP53 mutations and with an MRES epigenetic phenotype. Progression and invasion in 40% to 50% of cases. (Drawn by Renaud Chavrier based on: Bakkar et al., 2003; Dyrskjøt et al., 2004; Neuzillet et al., 2012; Rebouissou et al., 2012; Vallot et al., 2011)*

et al., 2003; Rhijn et al., 2001). In the papillary pathway, the homozygous deletion of the Tumor Suppressor Gene *CDKN2A* is associated with progression and invasion (Rebouissou et al., 2012). CIS are associated with *TP53* mutations (Rhijn et al., 2004). Moreover, a particular gene expression signature consisting in the characteristic level of 16 mRNA species was found to be specific and predictive of CIS (Dyrskjøt et al., 2004). Finally, a multiple regional epigenetic silencing (MRES) phenotype was associated with invasive tumors, rare FGFR3 mutations and predicted to be CIS tumors (Vallot et al., 2011).

This CIS/papillary dichotomy was further completed by a number of reference studies producing seemingly different molecular classification of urothelial carcinomas (Cancer Genome Atlas Network, 2014; Choi et al., 2014; Rebouissou et al., 2014; Sjodahl et al.,

2012). Despite obvious differences between the proposed classifications, two major subtypes of bladder cancer are emerging from all of these studies. Similarly to breast cancer, it appears that the differentiation state of the tumor is a main contributor to its classification (Damrauer et al., 2014). A first type is composed of cells presenting a highly differentiated phenotype expressing markers similar to the cells composing the superficial layers of the urothelium. Inspired by the biology of breast cancer, this subtype is now often referred to as *luminal* or *luminal-like* tumors, although Urobasal A was originally used (Sjodahl et al., 2012). A second type of bladder cancers was described as *basal* (Cancer Genome Atlas Network, 2014; Choi et al., 2014), *basal-like* (Rebouissou et al., 2014) tumors or Squamous Cell Carcinomas (Sjodahl et al., 2012). Basal-like bladder cancers express markers of the basal layer of the normal urothelium (*e.g.* cytokeratin 5, 14 and 6A) and are particularly aggressive tumors with poor clinical outcome (Rebouissou et al., 2014; Sjodahl et al., 2012). An example of classification is shown in figure III.5 and present, unlike most other studies, both muscle-invasive and non-muscle-invasive tumor classification.

Genome-wide characterization of bladder tumors enabled molecular classification of tumors but also help to define new drivers and therapeutic targets or to associate known targets to subtypes.

FGFR3 is the gene with the highest mutation rate (coding mutations) in bladder cancers with nearly 50% mutated tumors (Hernandez, 2006). *FGFR3*, standing for *Fibroblast Growth Factor Receptor 3*, is a gene encoding a transmembrane protein with an extracellular growth receptor domain and an intracellular kinase domain (illustrated in a previous section in figure I.1). The most frequent mutations appear in the extracellular domain (S249C: 55%; R248C: 8%) and mutations in the transmembrane domain of the proteins (Y375C: 24%). Most of the FGFR3 mutations potentially result in constitutive activation of FGFR3 and therefore on constitutive firing of the downstream signaling pathway (Cappellen et al., 1999). These mutations were found to be oncogenic mutations with transforming potential most probably resulting in ligand independent activation of the kinase activity (Bernard-Pierrot, 2005). Moreover, gains of copy number, over-expression (Neuzillet et al., 2014) and activating gene fusion (Williams, Hurst, and Knowles, 2013) have been observed. The alterations of FGFR3 are more frequent in papillary tumors and both are associated with luminal-like bladder cancer (Cancer Genome Atlas Network, 2014) making it an excellent target for tumors belonging to this type of highly differentiated tumors (Cappellen et al., 2006).

Despite the numerous evidences suggesting to be a major oncogene, *FGFR3*-activating mutations were previously identified as a cause of dwarfism syndromes and a negative regulator of bone growth (Webster and Donoghue, 1997). Moreover, studies also suggested *FGFR3* to have tumor suppressor properties in epithelial cells by reducing proliferation (Lafitte et al., 2013). Overall, the definition of the FGFR3 signaling pathway and its variations in normal and transformed epithelial cells is necessary to further understand its role and rationalize FGFR3-targeted treatments.

In addition to luminal-associated *FGFR3* mutations, a specificity of urothelial

carcinoma is its alteration of the proliferation/differentiation balance (DeGraff et al., 2013). At this point an interesting parallel can be made with non-aberrant phenotypes specific to the normal epithelial counterpart, the urothelium. Normal Human Urothelial cells (NHU) can be cultivated as primary unimmortalized cultures *in vitro* (Southgate et al., 1994; Southgate, Masters, and Trejdosiewicz, 2002). NHU cells initially express markers of basal/intermediate layers and undergo rapid proliferation (Southgate et al., 1994). As in several other epitheliums, the Epidermal Growth Factor Receptor (EGFR), along with its ligand and downstream signaling pathway, is necessary to this highly active proliferation state (Daher et al., 2003). This phenotype is reversed by the activation of a ligand-inducible nuclear receptor (Varley et al., 2010), the Peroxisome Proliferator-Activated Receptor-γ (PPARγ). Similarly to its key role in the differentiation of adipocytes, PPARγ was shown to induce terminal differentiation of NHU cells (Varley et al., 2004). The effect of PPARγ on NHU phenotypes is augmented by the inhibition of EGFR, which in its active form, inhibits PPARγ by an increased phosphorylation and decreased nuclear translocation (Varley and Southgate, 2008; Varley et al., 2004). This complex feedback control reflects the differentiation/proliferation balance characteristic of urothelial and generally of epithelial regeneration.

A recent study identified an activation signature of PPARγ (Choi et al., 2014) in a luminal subtype, while EGFR (Rebouissou et al., 2014) is suggested to be a driver of the basal subtype of bladder cancer. EGFR has been observed to drive several types of human cancer (Normanno et al., 2006), the implication of PPARγ is more difficult to rationalize in particular due to its role in differentiation and its negative effect on NHU cell proliferation (Varley and Southgate, 2008).

Overall, the implication of oncogenic alterations in bladder cancer is only partially understood. In particular, the divergent roles of oncogenic drivers in transformed and normal urothelium suggests major deregulation of their associated pathways.

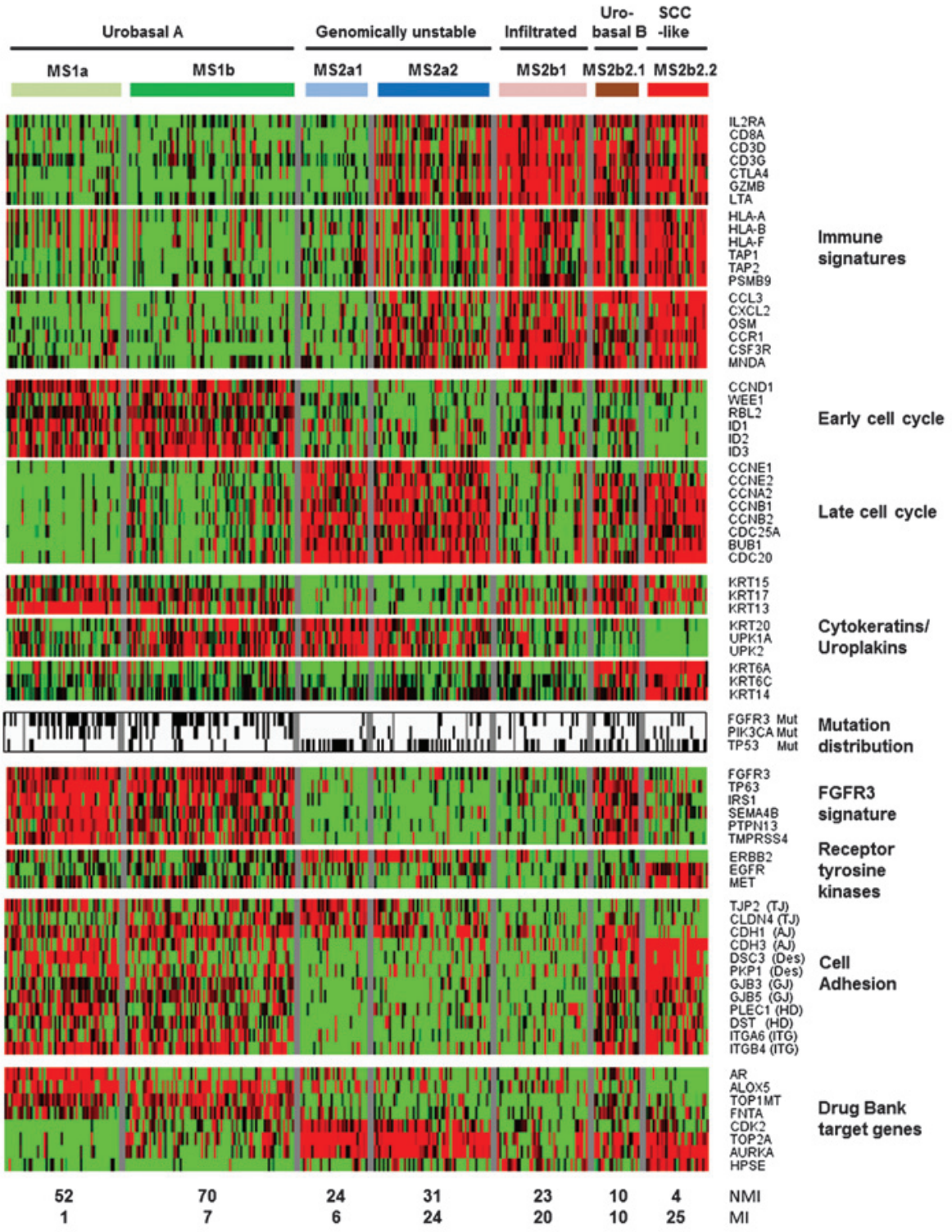


Figure III.5: *Bladder cancer subtypes. 7 sub-type of urothelial carcinomas identified using hierarchical clustering on 308 bladder cancer transcriptomes of all stage and grade. Only genes involved in specific functions only are shown on the right. Numbers of non-muscle-invasive (NMI) and muscle-invasive (MI) are shown for each cluster at the bottom of the heatmap.*
TJ, tight junction; AJ, adherence junction; Des, desmosome; GJ, gap junction; HD, hemidesmosome; ITG, epithelial integrins.
(from Sjodahl et al., 2012)

Chapter IV

Unravelling oncogenic pathways

A challenge of modern cancer research is to develop tools for the diagnosis and prognosis of tumors and more importantly for the identification the most effective therapy: theragnosis. During nearly twenty years now and especially since the sequencing of the Human genome was finalized in 2001, high-throughput technologies have been continuously improving. Since then, a large part of the scientific community tackled these issues by applying genome-wide profiling techniques on what can seem as large cohorts of patients. These studies were quite successful in identifying recurrent alterations. Furthermore, the characterization of whole transcriptomes (mostly messenger RNA) has lead to the definition of subtypes inside tumors arising from the same tissue. The stratification of patients has major clinical implications as different sub-types often have very different prognosis and in some cases are associated with good responses to specific therapies. Several transcriptomic studies of breast cancer observed a strong relationship between the expression of sets of genes and the clinical prognosis of the patient (Veer et al., 2002; Vijver et al., 2002). Other studies aiming at linking these gene expression profiles to the tumor phenotype started to be published in 2000 (Perou et al., 2000). Figure IV.1 shows a partial illustration of the transcriptomic datasets and the extracted sub-types using a heatmap view. For instance, the luminal sub-type of breast cancer is often associated with the activation of the Estrogen Receptor (ER) and potentially benefits from tamoxifen treatment, an inhibitor of ER.

Although genomic data hold great promises, its analysis is accompanied by many difficulties. For instance, most of the data comes from heterogeneous population of cells making it unfit to distinguish tumor sub-clones or cells originating from the stroma. However, the major source of difficulty arises from our inability to reliably search or extract relevant features from datasets containing at least thousands of features within samples that are usually only several hundreds. Furthermore, while recent studies profile more and more tumors, up to several thousands, technologies progress much faster and now can measure simultaneously several hundreds of thousands and even millions of biological signals. This problem is called the *curse of dimensionality* and is referred to for analysis of datasets in which the number of sample n is much smaller than the number of measured

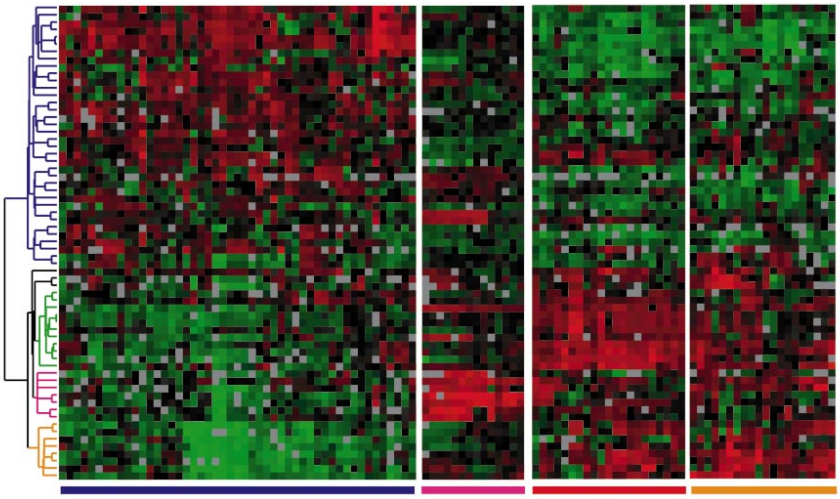


Figure IV.1: *Early transcriptomic-based breast cancer subtypes. The matrix of gene expression is represented using a* heatmap. *Each colored square represent the level of expression of a gene in a sample. Each line represents a sample and each column a gene. A heatmap is color-coded based on the relative level of expression, often compared to a reference level, here the median level of each transcripts in all samples, with red corresponding to over-expression, green under-expression and black no change in expression. The dendrogram on the left of the heatmap represents the hierarchical clustering of samples in which each branch is colored depending on it's assignment to one of the determined sub-type: basal-like, orange; Erb-B2+, pink; normal-breast-like, light green; and luminal epithelial/ER+, dark blue. This color code is also reported for groups of genes of which the over-expression is associated to one of the subtypes. (from Perou et al., 2000)*

features p ($p \gg n$) whereas in fact classical statistical analysis usually requires a large number of samples for relevant results (Clarke et al., 2008).

System-wide characterizations of human cancers studies now often include the measurement of more than one *layer* of molecular biology. That is, one of the three interconnected levels introduced as the *central dogma* by Francis Crick in 1958: $DNA \rightarrow RNA \rightarrow proteins$. In fact, early high-throughput studies often included only messenger RNA or data on chromosome copy numbers. Nowadays, large international consortiums have formed for the nearly complete characterization of most cancer types. Notably, the International Cancer Genome Consortium (ICGC: icgc.org) and The Cancer Genome Atlas (TCGA: cancergenome.nih.gov) projects have already produced and analyzed the mutational, copy number, DNA methylation, messenger RNA and micro RNA profiles of thousands of tumors. Although these efforts clearly do not resolve the problem of dimensions, the simultaneous analysis of genetic alterations and transcript levels is a key step towards a better identification and understanding of oncogenic alterations as well as their role in carcinogenesis.

In parallel to the extensive profiling of cancers, a considerable effort was made to

produce and gather information on the molecular interactions occurring in cells. Given the importance of pathways and signaling cascades in cancer, these studies aim at collecting all the interactions between proteins or between proteins and DNA in the attempt to reconstruct fully functional pathways *in silico*. The large number of interactions identified gave rise to models of celluar pathways in the form of biological networks.

The following sections discuss the current status of high-throughput techniques, network biology and finally of methods integrating both of these types of knowledge to uncover oncogenic driver pathways.

IV.1 Large scale tumor profiling

Most high-throughput technologies aim at quantifying all forms of a particular molecular specie in a biological sample. The complete profile of a given cell is named after the layer it corresponds to in the *central dogma*, extended with the metabolic reaction encoded by enzymes, with the 'ome' suffix:

$$Genome \xrightarrow{transcription} Transcriptome \xrightarrow{translation} Proteome \xrightarrow{enzymatic\ reaction} Metabolome$$

In fact, the most widely used techniques to profile cancer cells are the ones measuring nucleic acids. The main reason is that the progress in microarray technologies and more recently in sequencing resulted in the possibility to identify and quantify most, if not all, of the variations in DNA sequences and of the transcribed species by covering practically the whole genome and transcriptome.

With more than 20 published studies including the multi-genomic and transcriptomic profiling of 12 different cancer types, the TCGA consortium is among the most advanced group in genome-wide cancer analysis. The consortium is laying down standards in cancer genomic research ranging from production to data processing and analysis. Moreover, the TCGA is providing a unique standardized multi-omic pan-cancer dataset used in hundreds of published studies (more than 350 in may 2013, and counting). Figure IV.2 shows the different layers and types of alterations covered by the TCGA which mainly involve analysis of cancer genomes and transcriptomes.

Genome

The first genome based analysis of cancer samples taking over karyotyping was array Comparative Genomic Hybridization (aCGH) and was simultaneously introduced by D. Pinkel (Pinkel et al., 1998) and P. Lichter (Solinas-Toldo et al., 1997). This method is schematically depicted in figure IV.3. aCGH profiles only chromosome alterations which involves a modification of the copy number of large chromosomal regions. Bacterial Artificial Chromosomes (BAC) from genomic DNA library are usually used as probes in the array. Thus, this method is sometimes called CGH BAC array with probes generally

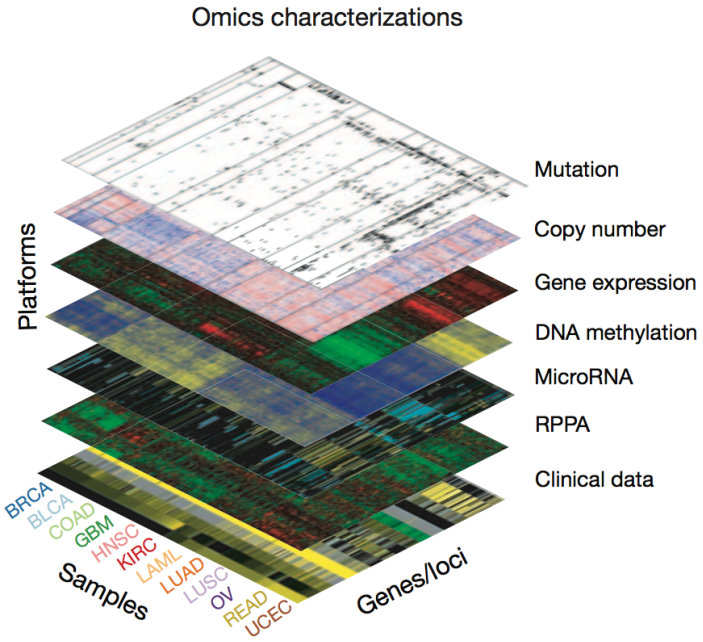


Figure IV.2: *Omics levels covered by the TCGA consortium. Each of the presented level of omic data is available for 12 tumor types at the time of the TCGA-12 pan-cancer analysis. The 12 cancer types include: Breast Invasive Carcinoma (BRCA), Bladder Urothelial carcinoma (BLCA), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Acute Myeloid Leukemia (LAML), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV), Rectum adenocarcinoma (READ) and Uterine Corpus Endometrial Carcinoma (UCEC). RPPA: Reverse Phase Protein Arrays, see section IV.1. (from Weinstein et al., 2013)*

measuring approximately 150kb (kilobase) and arrays often containing less than 5,000 probes.

The next main technological progress in genome analysis was the development of dense microarrays with smaller 25-mer probes and hundreds of thousand to millions of them on the same chip. These chips were used for genotyping by assigning several probes for the same genomic position and corresponding to different nucleotides possibilities, polymorphisms. Thus, the probes were designed around known human Single Nucleotide Polymorphisms (SNP). The higher definition, resolution, of these SNP arrays, with approximately one probe every 2,000 bases on average (McCarroll et al., 2008), enabled the identification of focal copy number aberrations that aCGH could not capture. SNP arrays are also useful to identify regions of LOH (see II.3) for which the normal/tumor log ratio is null but the homozygosity of a large locus is abnormal.

Evidently, SNP arrays are also used to genotype samples. This is conditioned on the fact that the position and the polymorphisms have been previously observed and

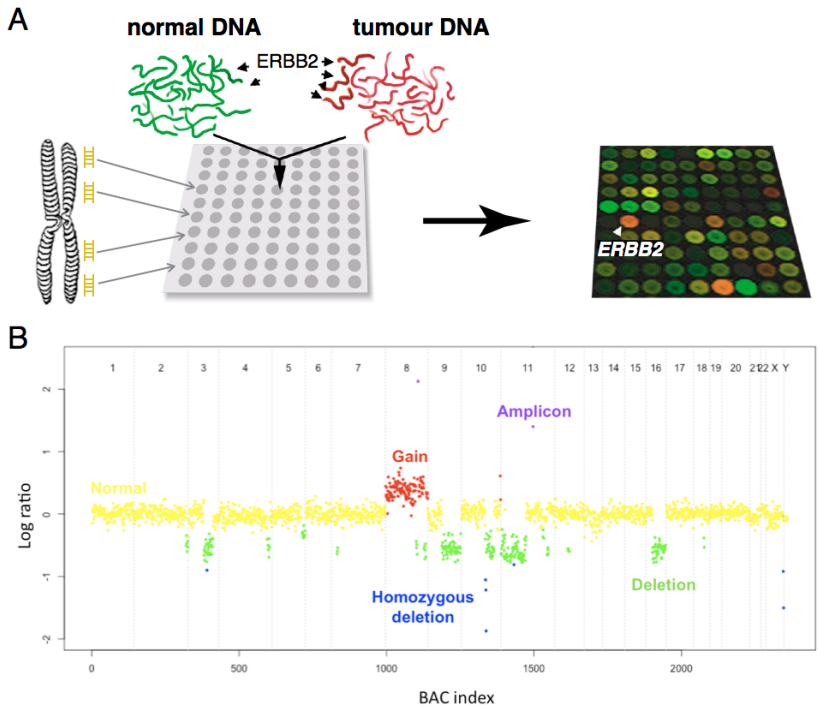


Figure IV.3: *aCGH genome profiling of copy number aberrations. A. DNA extracted from a tumor and a matching normal sample are labeled with a red (cyanine 5) and green (cyanine 3) dyes respectively. Samples are then hybridized to a microarray spotted with long DNA sequences representative of the Human genome. In this example, the tumor presents an amplification of the locus containing the ERBB2 gene. B. Copy number profile of a bladder tumor obtained with a BAC array-CGH. The base 2 logarithm of the normal/tumor fluorescence ratio is plotted for each BAC clone. Chromosomes are indicated at the top. (adapted from Pollack et al., 1999)*

are detectable by the array given its design. This problem of only detecting something that we might expect is gradually vanishing with what is still called, after a decade of innovations, next-generation sequencing. Sequencing technologies have brought to research the possibility to obtain the full sequence of a human genome for now less than a thousand dollars. In cancer research, these technologies have the potential to identify a vast number of alterations leading to neo-plastic transformation. Indeed, the full landscape of point mutations as well as of DNA breakage leading to chromosome alterations and of viral integration is potentially available through high-throughput sequencing.

The amount of data and their complexity requires the development of efficient method to retrieve relevant information from the data. To reduce the complexity, most sequencing approaches use a preliminary enrichment step to only sequence regions of interest in the genome. Since point mutations in coding sequences are the most easy to grasp, a large majority of studies sequences only the exome, the full set of exons in the genome that

represents approximately 1% to 2% of the genome. Exome sequencing consists in three simple steps: capturing and sequencing the DNA sequences corresponding to exons (with surrounding sequences in some cases), mapping the sequences to the human genome and identifying variations with the reference human genome. Putting sequencing and mapping errors aside, this last step is most problematic for several reasons. The number of reads, that is the number of sequences mapped at a given position in the genome, needs to be sufficient to cover both alleles of the normal and cancerous samples. For instance, in the case in which a mutation is found in several reads in the tumor samples but only a few reads are given for the matched normal sample, it becomes difficult to distinguish somatic mutations, difference between tumor DNA and normal DNA from the same patient, from germline mutations that can be found in any of the patient cells. This problem is well illustrated by the lack of concordance between variant calling pipelines, softwares that aim to identify all point mutations from a sequencing experiment. Many methods were proposed to retrieve a list of mutations and several comparison studies revealed a substantial lack of agreement between them with often less than 25% concordance, measured by the proportion of mutations found by all methods (intersection) among all the mutations that were found by at least one method (Kim and Sung, 2012; Roberts et al., 2013). This first problem is only partly resolved by increasing the depth of sequencing, the mean number of reads on the targeted parts of the genome. Sequencing depth ranges from approximately 20 average reads per sequenced position up to 2,000 in what is called ultra deep sequencing which usually targets small predefined regions of the genome. To validate mutations, the most frequent strategies include resequencing the mutations identified by high-throughput and verifying the sequence of mRNA in the cases for which the mutation appears inside a coding sequence and in a gene that is transcriptionally active.

Despite these great difficulties, an increasing number of tumors have been sequenced, progressively revealing the mutational landscape of carcinogenesis. The TCGA in particular has completed it's first pan-cancer analysis step which is composed of more than 3,000 tumors of 12 different human organs. Interestingly, most of the frequently mutated genes were already known (*TP53*, *RAS* family, *PIK3CA*,...). Most of the mutations occurred in signaling pathways, from receptors to transcriptional regulators. Other processes such as maintenance of genome integrity or protein/RNA core processing are also altered (Kandoth et al., 2014). This pan-cancer analysis also revealed the high variation of mutation frequencies between cancer types with nearly two-orders of magnitude difference between Acute Myelogenous Leukemia and Melanoma (see figure IV.4).

The wealth of data provided by cataloging mutations requires method to identify those that are important for carcinogenesis. Most often, only mutations in coding regions are considered, thus the goal becomes the identification of cancer *driver* genes. For instance, a gene that is often found mutated in cancer samples could be considered important and defined as a driver gene. However, Human coding genes have extremely variable sizes of coding sequences ranging within three orders of magnitudes from hundreds of base pairs (*e.g.* Histone H1A: 645 bp) to hundreds of thousands (*e.g.* Titin: 103kb). Therefore, large

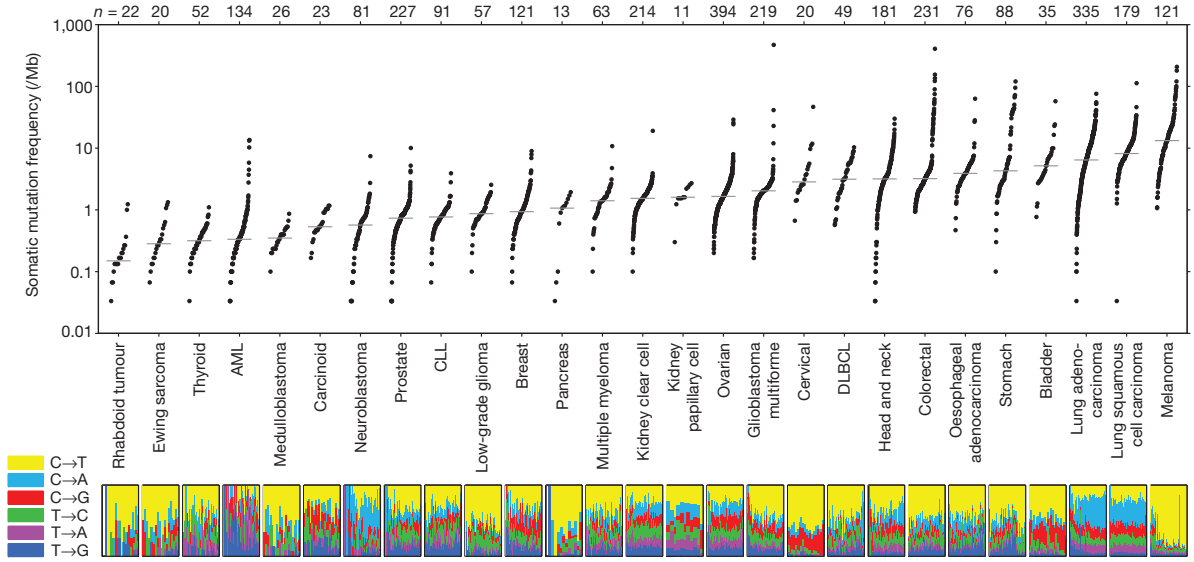


Figure IV.4: *Somatic mutation frequencies in 28 cancer types. Each dot represents one of 3,083 pairs of tumor/normal samples of which the exome was sequenced to identify somatic mutations. Bottom panel plots the proportion of each of the sixe type of possible base pair substitution. AML: Acute myelogenous leukemia; CLL: Chronic Lymphoid Leukemia; DLBLC: Diffuse large B-cell lymphoma. (from Lawrence et al., 2014)*

genes will be found to have more mutations simply by chance and when using a simple frequency criterion, these will more likely be mistakenly considered as driver gene.

To address this problem of identifying *driver* mutated genes, several methods were proposed to identify genes that are altered by mutations and do not seem to appear randomly. All methods are based on a cohort of patients and aim at identifying drivers for a particular cancer type. This can be done by considering the size and composition of the genes to assess whether it's mutation rate is higher to a background mutation rate (Dees et al., 2012). Other methods focus on the effect of a non-silent mutation on the encoded amino-acid sequence and the impact it is predicted to have on the function of the protein (Reimand, Wagih, and Bader, 2012). Another class of method prioritizes altered genes based on the position of the mutations in their coding sequence either through a pattern of highly clustered mutation (Tamborero, Gonzalez-Perez, and Lopez-Bigas, 2013) or based on predefined position corresponding to potential phosphorylation sites (Reimand, Wagih, and Bader, 2012). Finally, a more elaborate procedure is based on the fact that several gene characteristics, such as size, expression level, replication time or the chromosomal opening compartment, influences the random occurrence of mutations. These can be used to derive a more refined background mutation rate to which each gene is compared to depending on its own characteristics and is considered a significantly mutated genes if it deviates from its background (Lawrence et al., 2014).

Exome sequencing delivers more information than simply the list of polymorphisms, somatic mutations and indels. Copy number alterations can also be derived from the coverage of genome locus (Chiang et al., 2008; Cibulskis et al., 2012). Also, sequences that cannot be mapped to the human genome can actually map to viral sequences and therefore help to identify oncogenic viral integration. More recent studies using deep or ultra deep sequencing addressed the problem of intra-tumoral heterogeneity, which is more usually seen as a pitfall in the processing and retrieval of somatic mutations that only appear in a small portion of the sequenced sample (Nordentoft et al., 2014). Sub-clones identification relates to the history of neo-plastic transformation forming a tumor has major clinical implication by affecting treatment effects (Fischer et al., 2014).

Epigenome

Alongside the massive profiling of genome alterations, a divers set of techniques based on similar technologies were developed to model the organization of the genome (most reviewed in Furey, 2012) sometimes referred to as the epigenome.

A first category of methods aims at identifying epigenetic marks, which influence the compaction of the chromatin and eventually the expression of surrounding genes. DNA methylation using microarrays or Histone tail modifications with ChIP-seq (Chromatin Immuno-Precipitation followed by high-throughput sequencing) cover the main epigenetic marks regulating compaction of the chromatin. While most histone modifications have a known effect on chromatin opening and therefore gene expression (or actually other chromosomal structures), DNA methylation is much more ambiguous. It was originally thought that a local high rate of CpG methylation down-regulates surrounding genes. Many cases of promoter hyper-methylation have reported to result in gene under-expression, but the opposite has also been shown (Bahar Halpern, Vana, and Walker, 2014). The upstream or inner position of the methylated CpG can also have an effect not only on gene expression but also on mRNA splicing. Overall, DNA methylation is easily measurable yet has many still misunderstood functions (Suzuki and Bird, 2008).

A second category of techniques aims at directly identifying the state of the chromatin itself by estimating the enrichment in DNA-protein interaction, mostly consisting in DNA-nucleosome interaction. The idea is to create a library of sequences enriched in nucleosome depleted DNA by either inducing cuts using DNaseI (DNAse-seq, Song et al., 2011) which preferentially digest naked DNA, or by cross-linking proteins and DNA using formaldehyde and sequencing the segregated naked DNA sequences (FAIRE-seq, Song et al., 2011). A more recent method named ATAC-seq (assay for transposase-accessible chromatin using sequencing) relies on the fact that the hyperactive Tn5 transposase have a tendency to integrate in open chromatin regions. ATAC-seq has a better signal-to-noise ratio than DNAse-seq and requires only 10^4 cells against 10^7 for DNAse-seq. These techniques also can be used identify very specific regions of open chromatin with non-nucleosome protein binding. In the case of sequence specific binding such as transcription factor, regions of the

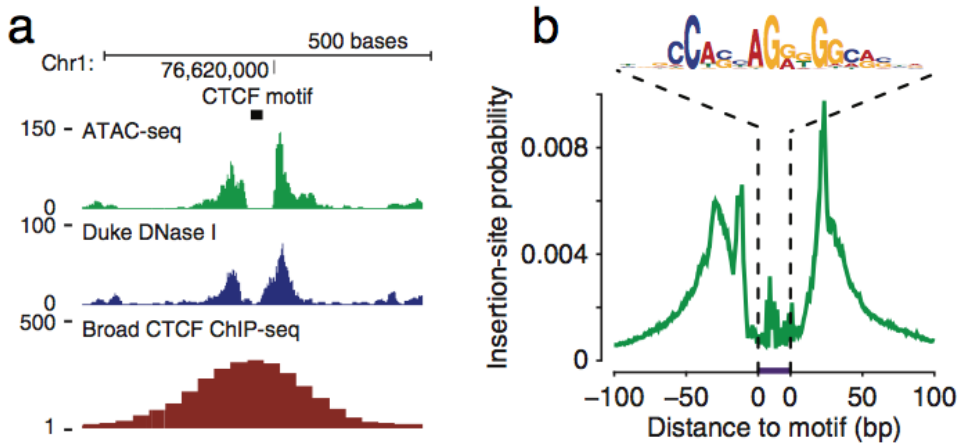


Figure IV.5: *Transcription factor binding inferred from chromatin opening. a. ATAC-seq, DNase-seq and CTCF (a transcription factor) ChIP-seq reads mapped in the same locus located in human chromosome 1. A predefined CTCF binding motif is shown as a black mark. b. Average ATAC-seq reads around CTCF motifs throughout the genome. (from Buenrostro et al., 2013)*

genome with particular ditch between two peaks of mapped reads can sometimes contain a consensus DNA binding sequence corresponding to a particular TF as depicted in figure IV.5. Although this is not as specific as ChIP-seq using a antibody against a given TF, this analysis enables the identification of context specific binding sites for any TF with a known consensus binding sequence.

A last category of high-throughput genomic structure analysis concerns the spatial organization of the genome and the interactions between loci that may be distant in terms of number of base pairs. The main technique is based on Chromosome Conformation Capture (3C) which is represented schematically in figure IV.6 and basically consist in concatenating DNA sequences that are adjacent in the nucleus whether they are in the same genome locus or on different chromosomes. 3C can be complemented by high-throughput sequencing, often referred to as high-C, to identify all captured interactions and thereby map the whole DNA interactome to model the conformation of the chromosomes (Belton et al., 2012). Genomic approaches to uncover the three-dimensional organization of the genome are increasingly drawing attention. More importantly than the data provided by these studies, our understanding of regulatory processes is shifting from simple TF-DNA interaction with sequence specific binding to chromatin interaction, modification, compactness and the dynamics of the three-dimensional shape of the genome regarding the activity of TF and chromatin regulators. To unravel such complex phenomenon, novel methods and in particular combination of long used techniques are being proposed. For instance, to uncover the impact of TF binding on the conformation of chromosomes and on the spatial association of co-regulated genes (for instance into transcription factories, Sutherland and Bickmore, 2009), a method called ChiA-PET (chromatin interaction

analysis by paired-end tag sequencing) combines variants of ChIP 3C and high-throughput sequencing. ChiA-PET can be used to identify the chromatin interactions associated to a particular protein such as the oestrogen-receptor (Fullwood et al., 2009). Another example is the use of DNAse I digestion to improve the result of high-C techniques (Ma et al., 2014). Overall, the analysis of genomes is shifting from the description of its sequence to the analysis of the high-level and dynamic organization.

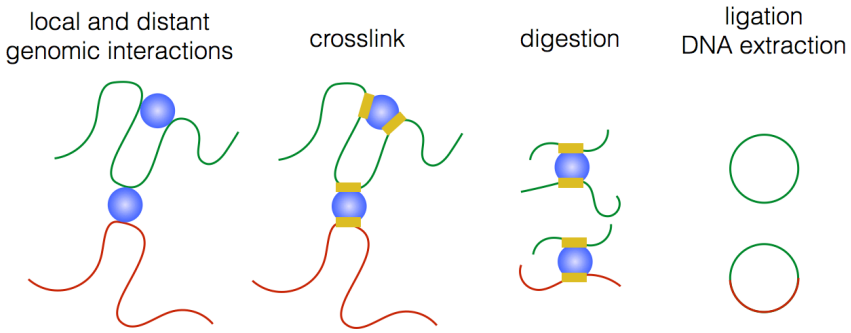


Figure IV.6: *Chromosome Conformation Capture. Interactions inside or between genomic loci (represented by a green or red color, can also be two different chromosomes) through proteins (blue). Formaldehyde is used to crosslink proteins and DNA. Following a digestion step, sequences of the same complex are ligated creating a chimeric DNA composed of spatially adjacent sequences.*

Transcriptome

Arguably, the most widely exploited genome-wide technique is transcriptomics and, until recently, almost exclusively messenger RNA profiling. Since genomic analysis give only little information about cellular phenotypes and proteomics has not met as fast development as nucleic acids based techniques, transcriptome profiling remains the best option to analyze cellular states.

Whether using a nylon or glass substrate or whether the hybridized mRNA is biotinylated or marked with fluorescent dyes, the principles of transcriptomic microarrays remain the same. A set of predefined probes is attached to an array to quantify potential complementary mRNA sequences, which are first reverse transcribed to use more stable cDNA. However, the transcriptome is not only defined by the level of expression of genetic loci. Indeed, the maturation of mRNA and especially the numerous possibility of splicing events can result in different isoforms and thus increase the transcriptomic diversity. Therefore, probes can be designed to hybridize with an exon common to all isoforms, with each exon or with sequences corresponding to the concatenation of consecutive exons characteristic to a particular set of isoforms (exon junction). Given the numerous possible combinations, the detection of these specific transcripts is conditioned on the evolution of microarray technologies to enable the simultaneous probing of millions of sequences

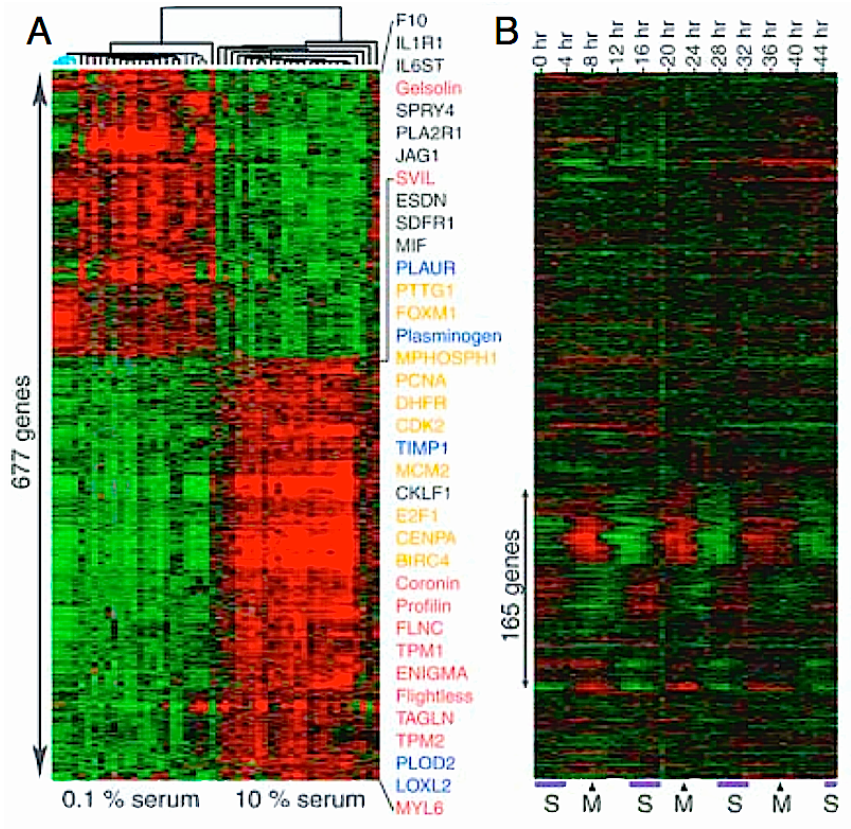


Figure IV.7: *Phenotype-reflected transcriptomes. A. Heatmap visualization of serum responding genes in fibroblasts from ten different anatomic sites. The most differentially expressed genes among the whole set of genes represented on the microarray are shown here in the heatmap. A subset of gene names are added and colored depending on their function. B. Transcriptomes of HeLa cells synchronized by double thymidine block and sampled every 4 hours. Approximately 25% of the genes show a periodical pattern of expression. At the bottom of the heatmap is shown the cell cycle phase transition during the time course.*
Color legend for left panel. Cell cycle progression: orange; Matrix remodeling: blue; Cytoskeleton rearrangement: red; cell-cell signaling: black (from Chang et al., 2004; Whitfield et al., 2002)

corresponding to hundreds of thousand possible isoforms.

Gene expression analysis underwent the same rapid development than DNA sequencing with high-throughput RNA sequencing (RNA-seq). Mapping and quantification of RNA-seq reads can first be used to quantify gene expression similarly to microarrays (Li and Dewey, 2011). However, RNA-seq improves the estimation of RNA expression levels approaching absolute quantification compared to microarrays by estimating a null expression values when no reads were mapped to non-expressed gene locus for instance. Exon inclusion can be deduced from mRNA sequence reads mapped to exon sequences. While reads joining two distant exons uncover splicing events. Moreover, reads that do not map to the

Human genome can reveal the integration and expression of viral genes (Tang et al., 2013). Furthermore, gene fusions due to DNA breakage and abnormal chromosome rearrangement can be identified by a substantial number of reads overlapping two unrelated genes (Supper et al., 2013). Finally, point mutations occurring in transcribed regions of the genome are detectable only in cases in which the mutated locus is stably transcribed in the analyzed sample.

Irrespectively to the technology used for transcriptome profiling, the purpose remains the near-complete analysis of variations in gene expression. As discussed in section I.2, the consequence of transient, and more importantly of malignant constitutive activation of cell signaling pathways is transcriptional control. Indeed, while genetic alterations are used to identify genes with potential implications in carcinogenesis, the transcriptome is representative of the cellular state. Evidently, from a single large-scale tumor profile it is nearly impossible to infer whether a genetic alteration is functional or whether an increased mRNA level will result in an increased protein activity. However, coordinated modification of mRNA levels, noticeable through systemic analysis, can effectively reveal cellular phenotypes as shown in figure IV.7.A by the transcriptomic response to serum treatment or IV.7.B with the cyclic expression of sets of genes during cell cycle.

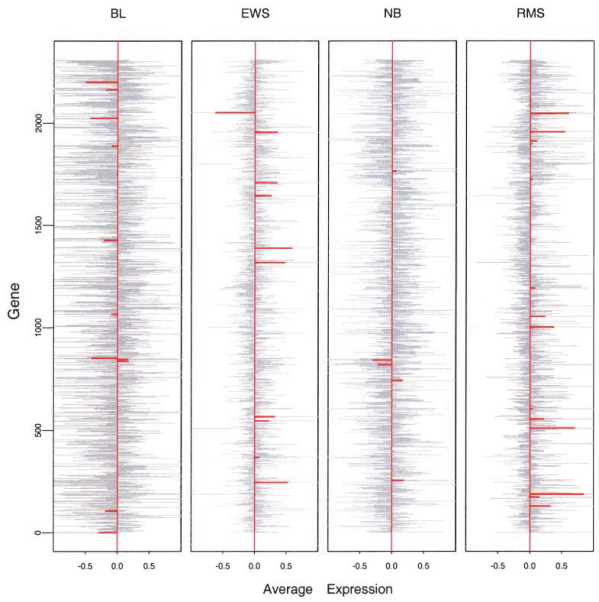


Figure IV.8: *Gene expression signature. Centroids and shrunken centroids of the transcriptomes of 4 classes of small round blue cell tumor. Centroids are shown in grey and represent the average expression of the 2,308 genes for each class. Shrunken centroids are shown in red and represent the corrected average expression of 43 selected genes. Burkitt lymphoma (BL), Ewing sarcoma (EWS), neuroblastoma (NB), or rhabdomyosarcoma (RMS) (from Tibshirani et al., 2002)*

Transcriptomic experiments became the method of choice to identify and understand

links between phenotype and molecular changes. In cancer research, genome-wide mRNA profiles brought the potential to identify subtypes of cancers of the same organ of origin that could have different prognosis or targetable driving pathways. For instance, the previously discussed breast cancer initial subdivision (figure IV.1) described classes which are still used nowadays and in particular by the more recent complete multi-omic breast cancer analysis (mostly genome and transcriptome) of the TCGA consortium (Cancer Genome Atlas Network, 2012).

The main clinical implications of these cancer subtypes lead to a widespread attempt to produce Gene Expression Signature (GES). GES are used to identify a small number of genes that show expression patterns specific to particular tumor subtypes. Indeed, following the descriptive capacity of transcriptome analysis, the next step became the predictive capacities of these technologies. Because microarrays and more recently RNA-seq remain too expensive to be systematically used in clinic, GES usually contain less than 100 genes and therefore their expression can be measured using more reliable low-throughput techniques such as qPCR (Quantitative Polymerase Chain Reaction) or nanoString®. Figure IV.8 illustrates this process through the use of the shrunken centroid method (also called Prediction Analysis of Microarray, Tibshirani et al., 2002) which has been widely used for signatures as it combines efficient and simple feature selection and classification steps.

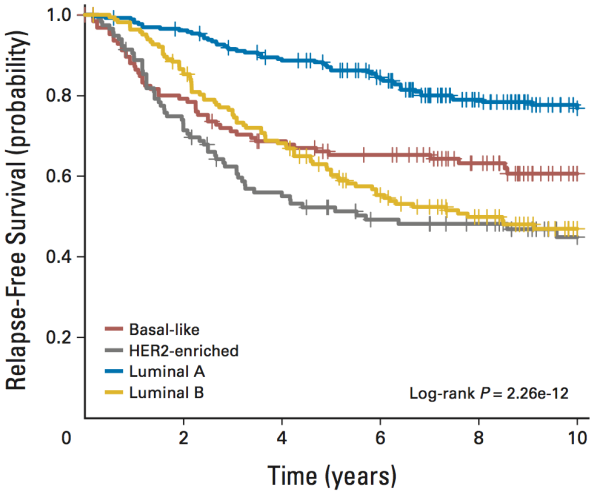


Figure IV.9: *PAM50 subtypes prognosis. Breast cancer subtypes determined by the expression of 50 genes. (from Parker et al., 2009)*

The presumed simple functions and central role of mRNA as templates for functional proteins lead to several important studies with the creation of clinical GES such as the breast cancer subtype and prognosis predictor PAM50 (figure IV.9, Parker et al., 2009). However, non-coding RNA have more recently been brought into the limelight for their key

role in regulating all regulatory processes (Esteller, 2011). During the past decade, several studies of non-protein coding genes revealed their major role in regulating key tumor suppressor genes (Poliseno et al., 2010) or tumor enabling characteristics (De Craene and Berx, 2013). However, because of the seemingly simple regulatory function on mRNA, microRNA are probably the most widely studied non-coding RNA. For instance, the TCGA consortium systematically profiles miRNA similarly to mRNA, thereby producing datasets with measures of both type of RNA species (Jacobsen et al., 2013).

Proteome and beyond

The central dogma of molecular biology illustrates the link between DNA-encoded information and cellular functions. Despite it's simplicity and recurrent critics, this paradigm illustrates well the hope that genomics will identify alterations that will then point towards targetable parts of aberrant cellular states.

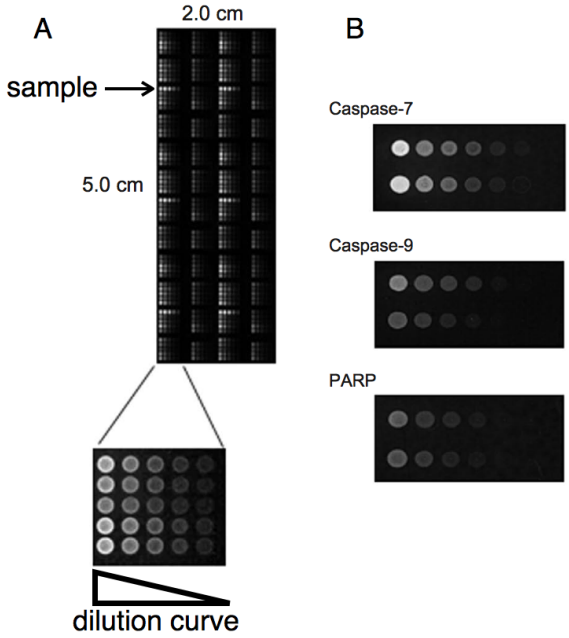


Figure IV.10: *Example of Reverse Phase Protein Array. A. Sampled cell populations are lysed and arrayed onto nitrocellulose slides following a linear dilution curve. The array can then be incubated with a highly specific antibody that is detected by chemiluminescent, fluorescent or colorimetric assays. The horizontal dilution curve serves as an internal control for quantification. B. Example of apoptosis related protein quantification. (adapted from Charboneau et al., 2002)*

In an attempt to follow the fast development of transcriptomics, several microarray technologies were developed to capture protein expression at the proteome level. Protein microarrays can be spotted using any high-affinity substrate for single proteins or protein

families such as antibodies, peptides, small molecules, nucleic acid polymers or phages (Wulfkuhle et al., 2004). However, Reverse Phase Protein Arrays (RPPA), illustrated in figure IV.10, has taken over because of its reliability. RPPA relies on applying high-affinity antibodies to fixed cell lysates. Therefore, only predefined protein species or protein post-translational modifications (PTM) with reliable antibodies can be quantified. The TCGA consortium used this technology to acquire about 200 protein related measurements (expression and PTM levels). The design of RPPA results in a precise analysis of proteins and PTM with key role in signaling pathway. For instance, the TCGA consortium measured the expression of a protein in each level of the Mitogen-Activated Protein Kinase (MAPK) pathway from EGFR to MYC and more interestingly, of the activating phosphorylation of some of these proteins in order to assess the activity of the pathway.

The use of highly specific and sensitive antibodies precludes the systematic quantification of protein and PTM. Firstly, because of the difficulty to obtain such antibodies and more importantly because of the number of possible protein isoforms and PTM sites. Following nearly a century of development, mass-Spectrometry (MS) can now be used to identify and quantify proteins and protein modifications in a high-throughput setting. Given the technological restrictions and the importance of kinase reactions in cell signalling, a preliminary enrichment in phosphorylated peptides (see figure IV.11), using immobilized metal affinity chromatography for instance, is one of the most widely used MS analysis (Mumby and Brekken, 2005). *Phosphoproteomics* is still nowadays an effective analytical tool to identify active signaling pathways and remains one of the most informative peptide enrichment step with immuno- or tag-affinity- precipitation. This latter technique is used to identify protein complexes and is thoroughly discussed in the following section (section IV.2).

Nevertheless, recent technological developments resulted in the possibility to identify and quantify a large portion of the proteome without any pre-enrichment step. Based on these advances and along with the progress of the TCGA, the Clinical Proteomic Tumor Analysis Consortium (CPTAC: proteomics.cancer.gov) aims at systematically profiling the proteomes of most human tumor types. The CPTAC established a proteomic pipeline illustrated in figure IV.11 which was used to quantify more than 7,000 protein species in 95 colorectal cancers also fully analyzed by the TCGA (Zhang et al., 2014a).

Although MS is also the preferred analytical technique for metabolomics, the cell-wide analysis of small molecules is the least studied of the omic levels. A recent investigation used MS to quantify 399 identifiable metabolites in 25 breast cancers to identify differences in metabolites level between subtypes (Tang et al., 2014). Besides MS, Nuclear magnetic resonance or NMR, can also be used as a quantitative method in a somewhat complementary way to MS analysis (Wishart, 2008).

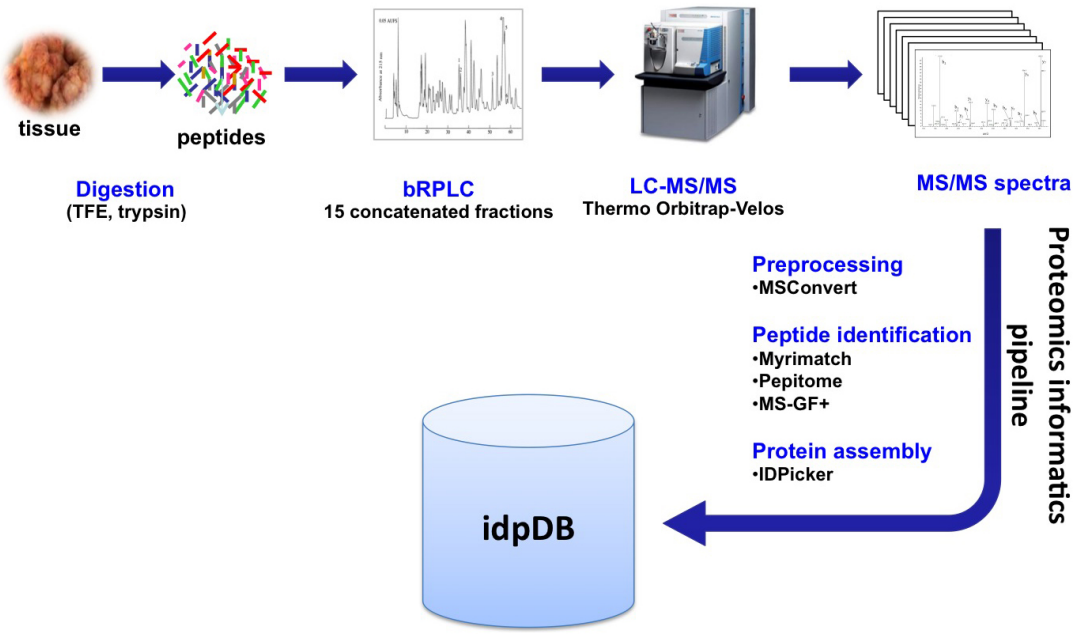


Figure IV.11: *Mass-Spectrometry based proteomics. Proteins are extracted from tissues and digested into smaller peptides. Tryptic peptides are further fractioned using basic reverse-phase liquid chromatography. Collected fractions are used for reverse-phase High-Performance Liquid Chromatography followed Mass-Spectrometry (here Thermo Orbitrap-Velos). From the resulting MS specters, a computing pipeline is used to identify peptides and eventually proteins all of which are stored in an application-specific database. The different processing tools used by the CPTAC are listed. (from Zhang et al., 2014a)*

IV.2 Signaling pathways, from interactions to networks

Large-scale tumor analysis mainly results in an enumeration of molecular aberrations. Whether there is a significant difference in mRNA levels between a normal and cancer sample or an alteration in the tumor DNA sequence, this information is often unpractical. Indeed, the over-expression or the gain of a single gene coding for an RTK is not sufficient to conclude that the downstream pathways are activated and that this sustains aberrant proliferation. This notion of activity is fundamental to our understanding of carcinogenesis and to the development of targeted therapy by means of inactivating drivers of tumor progression using highly specific drugs. The idea is that genetic alteration modifies the regulation of particular genes, which in turn activate oncogenic pathways from which hallmarks of cancer can arise. This is well exemplified by mutations in the RTK EGFR driving non-small cell lung cancers. Mutations in the kinase domain of EGFR activate the Akt and STAT signaling pathways. Small pharmacological inhibitors targeting any of STAT, Akt or EGFR proteins resulted in reduced cell proliferation and more importantly increased apoptosis (Sordella et al., 2004). Knowing that EGFR can activate the STAT or Akt pathways, the concomitant EGFR mutation, STAT and Akt protein activation (often by phosphorylation) and over-expression of the genes activated by these pathways is informative on their activity. These facts can be consequently used clinically to specifically interfere with this constitutive aberrant signaling.

While large-scale profiling is necessary to identify which pathways are active, they do not enable the identification of the composition of the pathways themselves. Having the knowledge of the structure, dynamic and impact of signaling pathways - from signal receptors to effector transcription factors - is critical to the analysis of tumors and to have a clinical impact.

Signaling pathways are in fact mostly composed of proteins passing signals by interacting with and modifying other proteins until reaching a transcriptional regulator. These will in turn interact with chromatin and in some cases DNA to either enhance or inhibit the activity of the transcription machinery towards particular coding loci of the genome.

Signaling pathways are discussed in section I.1 as linear cascades of protein PTM (see example of figure I.3). However, each level, each protein can integrate various signals and can be influenced by the specificity of the cellular context. In some cases, a single protein can be regulated, either in a cooperative or competitive way, by different pathways fired by different signals. This is referred as pathway crosstalk and is illustrated in figure IV.12 around the NF-κB regulator. This complexity clearly dismisses linear representation of pathways yet makes the use of graph models appropriate. Graphs are mathematical models of networks which in their simplest form are composed of a set of vertices or nodes noted V, proteins for instance, and a set of edges E connecting pairs of vertices such as two interacting proteins or a transcription factor binding to a gene promoter.

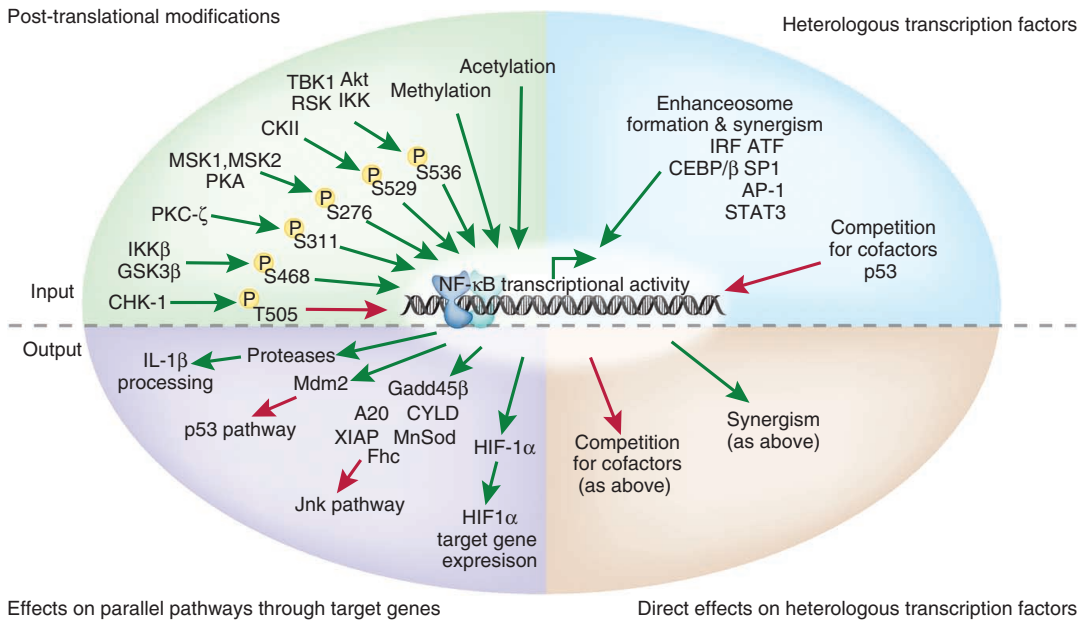


Figure IV.12: *NF-κB crosstalk. The transcriptional activity of NF-κB subunit is regulated by several post-translational modifications, mostly phosphorylation. The various PTM control the possibility of interactions of NF-κB with several co-activators and co-repressors which controls the integration of different contextual* input *signals from various signaling pathways and subsequently determines the specificity of the* ouput *signal: the regulation of target genes (from Oeckinghaus, Hayden, and Ghosh, 2011)*

Biological networks are collections of interactions between biological entities ranging from molecular interactions between proteins and small chemical compounds up to links between individuals in a population. In this section, only molecular interactions assembling into signaling pathway are discussed. The set of interactions included in a network serves as a structure to understand the functionality and consequence of alteration of proteins inside a signaling pathway. For instance, the over-expression of mRNA or proteins interacting in the same pathway is more robust indicator of its activity than the over-expression of any single node in the pathway.

Pathways can be mainly described by two types of interactions: protein-protein interactions (PPI) and regulatory interactions. This distinction is not only due to the nature of the molecules involved, protein-protein in one case and protein-DNA in the other case, but also by it's direct and detectable effect. In the case of PPI regulatory interactions, post-translational modifications are the main observable consequences such as phosphorylation in the MAPK pathway (see section I.1) or the ubiquitination as exemplified by the degradation of TP53 promoted by an extensive ubiquitination by MDM2 (Marine and Lozano, 2009). However, regulatory interactions have an effect on gene expression and

therefore have an impact on mRNA levels.

The next sections discuss these two major types of interactions, protein and regulatory, as well as whole pathway models and lists the available methods and repositories for each of these.

Protein interactions

Composing most of the signaling cascades, interactions between proteins involve a wide variety of molecular reactions. The most widely studied are post-translational modifications such as phosphorylation or ubiquitination. Whether protein interactions are transient or stable, whether an interaction results in the activation or repression of their activity, these involve the physical contact between proteins. In fact, most proteins associate into functional complexes with highly variable sizes and functions.

In order to gain insight into the functional associations between proteins, high-throughput methods have been devised to identify physical interactions. The Yeast-two-hybrid (Fields and Song, 1989) methodology aims at determining whether two proteins are capable of physically interacting and is illustrated in figure IV.13. The two-hybrid system has been extensively improved since the first study in 1989 by using mammalian cells, by testing protein-DNA interaction or by changing the reporting feature. Although these methods can be carried out in a high-throughput manner using automatic and robotic experimental designs, all these systems have several limitations. The synthetic expression level of two proteins which might never be co-expressed in the same cell or present in the same cellular compartment may retrieve false positive interactions which in fact never occur in natural systems. Furthermore, the fusion with technique-specific protein domains (Gal4 DBD and AD in figure IV.13), the lack of particular protein modification especially in yeast systems when testing mammalian proteins as well as the lack of particular necessary co-factors are other possible interfering effects. The numerous difficulties mainly affect the number of false positive interactions, estimated to be as high as 70% by some studies taking into account the expression of the corresponding coding mRNA or by comparing paralogs (Deane et al., 2002). Despite these complications, the yeast-two hybrid technique and its improvements has been widely used to identify edges in organism-wide PPI networks.

In contrast to the unspecific PPI identified by yeast-two-hybrid derived system, western blotting following an immuno-precipitation (IP) can identify two proteins interacting in a particular cellular condition. A generalized affinity capture method can be used in association with a Mass-Spectrometry based identification of co-precipitated proteins to identify interacting partners in a high-throughput procedure. AP-MS (Affinity-Purification followed by Mass-Spectrometry) can be used to identify all proteins interacting with a protein interest and is illustrated in figure IV.14. This method was used in yeast to systematically identify for each protein their corresponding interacting partners (Gavin et al., 2006) and more recently in human to map the protein complexes of the autophagy

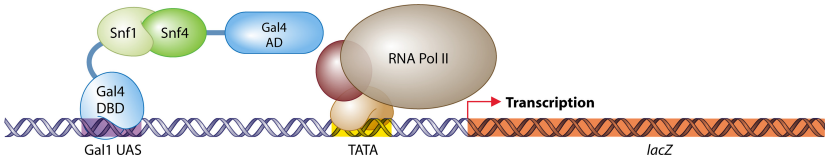


Figure IV.13: *Yeast Two-Hybrid. The first two-hybrid experiment (Fields and Song, 1989). The interaction between two proteins of interest (here Snf1 and Snf4) is tested by fusing one of them, the bait, with the DNA-binding domain (DBD) of Gal4 and the other, the prey, with the activation domain (AD) of Gal4. The Gal4 DBD domain binds to the upstream activating sequences (UAS) of the lacZ reporter gene. The interaction between Snf1 and Snf4 is detected by the tethering of the fused AD to the UAS, which activates the transcription of lacZ, which is detected by chromogenic analysis. Conversely, in the case of non-interacting proteins fused to the AD and DBD of Gal4, the activating domain is not recruited resulting in no specific activation of lacZ and therefore no detectable signal. (from Stynen et al., 2012)*

process (Behrends et al., 2010).

Given the importance of PPI in the understanding of cellular processes, numerous studies aimed at identifying interactions between specific sets of proteins using low-throughput and high-throughput systematic mapping. In order to access easily the pairs of interacting protein from all these studies, several groups endeavor to collect all these published interactions in a unique database (DB). Each of these PPI-DB integrate published results and provides an access to the data in different manners.

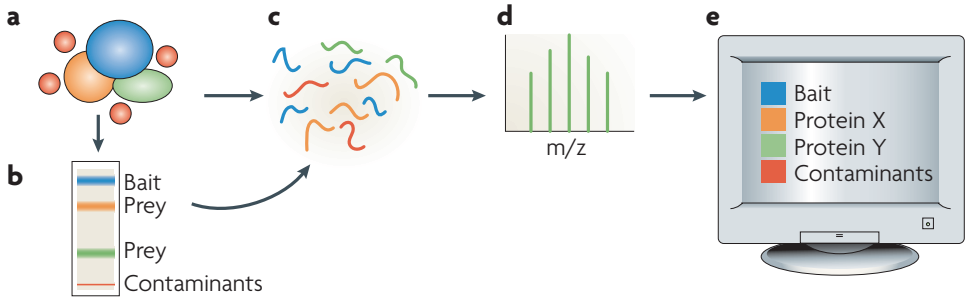


Figure IV.14: *Identifying protein complexes using affinity purification followed by mass spectrometry. a. A protein of interest (blue, can be tagged with a specific epitope) is purified from the lysate of a population of cells. Co-precipitates include binding partners (green and orange) and contaminants (red). b. Proteins in the complex can be separated by SDS-PAGE or by chromatography. c. Protein digestion results in small peptides, which (d.) can then be identified using Mass-Spectrometry. e. The identified proteins are then analyzed to eliminate false positive (contaminants). (from Gingras et al., 2007)*

The Biological General Repository for Interaction Datasets BioGRID (thebiogrid.org) remains of the most up to date repository of PPI (Chatr-aryamontri et al., 2012). The

BioGRID PPI-DB is composed of more than 170,000 human protein interaction extracted from more than 23,000 manually curated publications (BioGRID v3.2.116, September 2014). The interactions recorded in the database were identified by a wide variety of experiments ranging from a simple immuno-precipitation followed by western blot or MS analysis to X-ray crystallography and FRET (Fluorescence Resonance Energy Transfer). An interesting feature offered by the BioGRID repository is the accessibility of the source of the interaction identified between two proteins. For instance, EGFR and GRB2, the epidermal growth factor receptor and the next protein in the MAPK signaling cascade respectively, are identified as interacting protein in BioGRID and are supported by 39 published sources in fall 2014, 23 of which are low-throughput high confidence experiments.

The Human Integrated Protein-Protein Interaction Reference database HIPPIE (cbdm.mdc-berlin.de/tools/hippie) aims at integrating the PPI referenced in other databases and to select only high-confidence interactions (Schaefer et al., 2012). This database was developed to overcome the limitation of database such as BioGRID which directly retrieve PPI from any experiment although these are known to contain large number of false positive as discussed earlier in this section. HIPPIE applies a specific scoring scheme by basically assigning high weights to low throughput high confidence experiments such as X-Ray crystallography or circular dichroism, and assigning low weights to high-throughput or low confidence experiments and data such as co-localization or yeast-two hybrid assay. This results in a lower number of PPI referenced in HIPPIE, approximately 70,000, much lower than the number of PPI in the BioGRID database. Moreover, the web interface to the HIPPIE provides several useful tools to query sets of proteins and obtain the list of interactions between all pairs of the input proteins.

Finally, the STRING database (string-db.org) is most probably the largest collection of putative PPI (Franceschini et al., 2012; Mering, 2004). Additionally to the PPI experimentally identified for human proteins, the project behind the STRING database aims at predicting interaction based on several orthogonal datasets. The first feature used for PPI prediction in STRING is the automatic processing of the full text of nearly 2 million scientific publications. The second contribution to PPI prediction is the transfer of interactions between organisms using orthologous groups of genes. Based on these features, PPI are predicted and scored in the STRING database, which results in almost 5 million interactions. In addition, the web interface to the database provides tools to analyze sets of interacting proteins and by displaying global STRING scores as well as feature-specific scores.

Several other databases of PPI exist, some of which are listed below in chronological order of the latest update:

DIP Database of Interacting Proteins (dip.doe-mbi.ucla.edu) (Xenarios et al., 2002),

MINT Molecular Interaction Database (mint.bio.uniroma2.it/mint) (Chatr-aryamontri et al., 2007)

HPRD Human Protein Reference Database hprd.org (Keshava Prasad et al., 2009)

IntAct an EBI supported interaction database (ebi.ac.uk/intact/) (Kerrien et al., 2011)

GeneMANIA an analytical application merging genetic, protein and predicted interactions (genemania.org) (Warde-Farley et al., 2010; Zuberi et al., 2013)

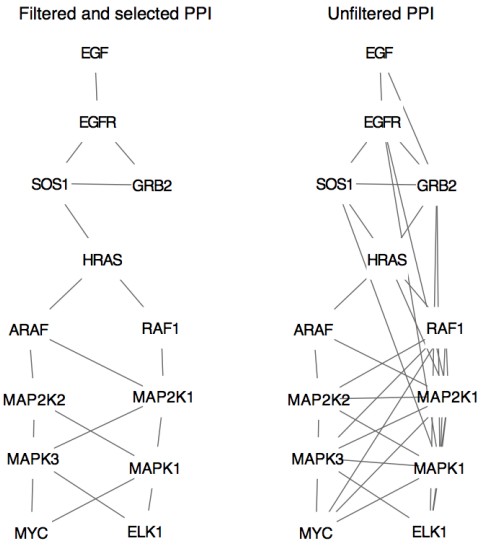


Figure IV.15: *PPI in MAPK signaling pathway. A selected set of proteins involved in the MAPK pathway and the protein interactions between them referenced in the HIPPIE database are shown. The first network contains only high confidence interactions (score above 0.8 in the HIPPIE scoring scheme). (obtained from the HIPPIE database (Schaefer et al., 2012) and using Cytoscape for network manipulation and visualization (Shannon et al., 2003))*

These collections of protein interactions are extensive source of information concerning potential signaling pathways. However, these resources are noisy and enclose inaccurate data either due to the integration of faulty experimental data such as those coming from yeast-two-hybrid experiments, or by errors in the retrieval of interactions from scientific articles. As an example, figure IV.15 shows the interactions found in the HIPPIE database between proteins involved in each level of the MAPK signaling pathway. While the filtered list of protein interactions models the expected signaling cascade (growth factor, receptor, RAS, MAPKKK, MAPKK, MAPK and TF, left panel of the figure), the complete list of PPI shown in the unfiltered network (right panel) shows many unexpected, though possible, interactions between various steps in the cascades. For instance, the surprising interaction between EGF and GRB2 was collected by HPRD and therefore present in HIPPIE. This interaction is annotate to be provided by a study of EGF response in rat hepatocytes. However, the published results do not show any interactions between these

two proteins and consequently the deduced protein interaction is erroneous (Kong et al., 2000). Therefore, PPI network and in fact biological networks in general have modest reliability concerning single interactions. However, a broader view and a systemic analysis can in many case be much more reliable and informative despite local errors.

Transcriptional regulatory interactions

Transcriptional regulation is the second complex component of signaling pathways. The possibility to abrogate the synthesis of specific transcripts and to initiate the transcription of new mRNA species is the mean by which gene regulation affects the proteome and potentially modifies the cells phenotype. Several steps are required to regulate the expression of genes.

First, a transcription factor or a set of factors is activated by a stimulus, an extracellular growth signal for instance. For regulators present in the cytoplasm, this activation can trigger the nuclear translocation of the activated TF.

Transcription factors are then able to bind consensus sequences in the genome for which they have a particularly high affinities, which are called Transcription Factor Binding Sites (TFBS). These sequences are specific to TF or TF-families and are often degenerated. TFBS are usually represented as Position Weight Matrices (PWM) also referred to as Position Specific Scoring Matrix (PSSM). An example of the GATA3 human consensus-binding site is shown in figure IV.16. Association of transcription factor to their cognate binding site upstream of a Transcription Start Site (TSS) is one of the mechanism by which TF regulate the synthesis of transcripts.

In fact, transcription initiation is much more complex. First, TF binding is highly dependent on the accessibility of the target binding site. As discussed and depicted earlier (see section I.2 and figure I.4), the presence and activity of co-factors is necessary for the binding of TF. Moreover, functional TFBS, those that can effectively regulate transcription when bound, can be placed anywhere around the TSS with extremely varying distances. Finally, the presence of specific cooperative factor is often essential for the binding of a TF to effectively recruit the core transcriptional machinery.

Transcriptional regulation, and in particular transcriptional activation, requires a large number of molecules to associate in the right place at the right time. This complexity makes transcriptional response extremely specific to the cellular context and to the stimulus or stimuli.

The first step towards the full comprehension of regulatory processes is the collection of all possible binding sites encoded in the human genome sequence as well as the consensus sites of all DNA binding transcription factors.

Several groups collected published data concerning TF binding and in particular TFBS models. Most of these databases also propose tools to scan DNA sequence, sequences of gene promoters for instance, to identify binding sites. One of the most widely used databases is JASPAR (jaspar.genereg.net) which references TFBS in the form of PWM for

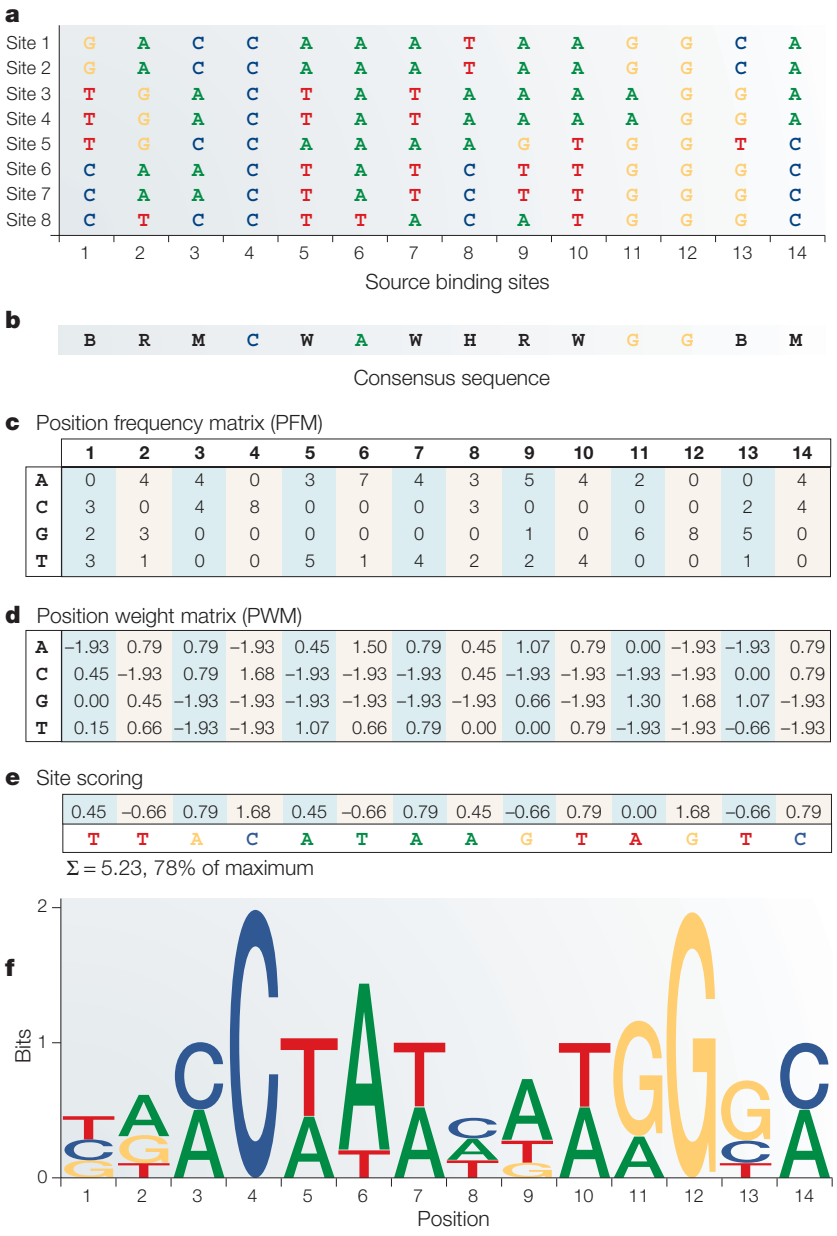


Figure IV.16: *Modeling transcription-factor binding sites. Example of the MEF2 transcription factor. a. Alignment of the DNA sequences of 8 experimentally validated MEF2-binding sites. b. Consensus binding sequence using an extended DNA alphabet. c. Position frequence matrix with 4 lines (one per nucleotide) and as many columns as positions in the binding site. Values in the matrix represent the frequency of the presence of a given nucleotide at a given position. d. A transformation of the frequency matrix into a weight matrix is often used to score a given nucleotide at a given position and therefore score and characterize a sequence (e). The weight of base b at position i is computed as follow:* $W_{b,i} = log_2 \frac{p(b,i)}{p(b)}$, *with* $p(b)$ *the background probability of base b and* $p(b,i)$ *is the (corrected) probability of base b at position i. f. Visual representation of the position frequency matrix in which the height of a nucleotide at a given position reflects its* information content *(Schneider and Stephens, 1990).*
(From Wasserman and Sandelin, 2004)

more than 200 human transcription factors. Another reference database is TRANSFAC which provides a public TFBS dataset as well as a large commercial hand-curated database (Matys et al., 2006). A more recent database, HOCOMOCO (autosome.ru/HOCOMOCO/ Kulakovskiy et al., 2012), collected and curated the most informative PWM for nearly 400 human TF. Finally, the *MotifDB* R/Bioconductor package references most of these TFBS models and can be easily used to scan DNA sequences with a large set of PWM.

The main difficulty in the analysis of gene regulation using consensus sequences is the high frequency of occurrence of potential TFBS throughout the genome. For instance, in a complete random setting, an 8-base long consensus sequence, such as the GATA3 binding site shown in figure IV.16, can approximately appear every 65 kb along the genome which results in about 46,000 putative sites in the Human genome. Therefore, to more accurately map target genes of human TFs, the direct binding of protein to the genome in specific cellular context can be identified using several large-scale experiments such as ChIP-seq or DNase/FAIRE-seq (see IV.1 for short discussion about techniques). In particular, two large consortiums are engaged in this type of full analysis of the Human genome.

The ENCODE project, Encyclopedia of DNA Elements project (encodeproject.org), aims at referencing all functional elements of the human genome. This project extensively uses ChIP-seq (Chromatin ImmunoPrecipitation followed by high-throughput sequencing) using TF or histone modifications, ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing, similar to High-C) as well as DNase- and FAIRE-seq to understand the function of captured sequences.

The Functional Annotation Of Mammalian Genome project, FANTOM (fantom.gsc.riken.jp), focuses on unraveling gene regulatory networks in human cells. For instance, the consortium aims at identifying the promoter sequence of every human gene by accurately referencing all possible TSS. More generally, the goal of the FANTOM project is to map all the possible regulatory interactions governing cellular processes and in particular combinatorial complexity of gene regulation.

Other groups collected published datasets of Protein-DNA interactions to propose databases of potential regulatory interactions. The Human Protein-DNA Interactome hPDI (bioinfo.wilmer.jhu.edu/PDI/) references almost 20,000 regulatory interactions between genome locus and 1,013 human DNA binding proteins including 493 transcription factors. SwissRegulon (swissregulon.unibas.ch) proposes tools and a database of TF bounded genomic regions and in particular to the FANTOM consortium data (Pachkov et al., 2007).

Transcription Factor cooperativity are also referenced in a final type of regulatory interaction repository. Transcription Factor synergy and competition is at the heart of cellular and signal specificity of transcriptional response. Therefore, the FANTOM consortium generated a list of human combinatorial TF interactions (Ravasi et al., 2010). Moreover, the Dragon database of transcription co-factors and TF interacting proteins proposes a curated database of TF-cofactors and TF-TF interactions (cbrc.kaust.edu.sa/tcof/, Schaefer, Schmeier, and Bajic, 2010).

The discussed studies and database of regulatory interactions mostly concern TF-DNA

potential affinity or binding. Tools and data can be used to analyze genes of interest to search for potential binding sites. However, the binding of regulators to promoter elements does not necessary causes modification of the level of transcription of the nearby coding sequence. This is particularly well illustrated by the significantly low intersection between the TF-bound genes and transcriptionally affected target genes. For instance, the FANTOM consortium carried out siRNA depletion of TF followed by transcriptomic analysis to list genes with altered transcriptional levels. These perturbation experiments are used to identify genes transcriptionally affected by the depleted TF. Although perturbation analysis does not only list direct transcriptional targets, the concordance between regulated and bound genes is unexpectedly low as shown in figure IV.17. Overall, gene regulatory networks are highly context-specific and standard repositories can only be used as sets of potential interactions.

	Perturbation	ChIP	Intersection
ETS1	1106	1670	118
STAT1	1022	259	15
MYC	447	2630	66
BCL6	954	474	3

Figure IV.17: *TF target consistency. Number of target genes determined for four transcription factors using two different methods. Perturbation experiments list the genes with a significant different mRNA level after siRNA-mediated depletion of a specific TF. ChIP is the set of genes found near a TF-DNA interaction determined by a ChIP-seq experiment. The last column contains the number of genes found to be targets of each TF in both experiments. For each of these TF the intersection is significantly* lower *than expected randomly (fisher's exact test* $< 1\%$*)*

Pathway models

Signaling pathways are mainly composed of a first set of protein interaction to transduce extracellular signals and a set of regulatory interactions to orchestrate a transcriptional response and induce changes in the targeted cell's behavior. In order to fully understand the operation and impact of cellular signaling, rigorous review and curation processes generated repositories of whole pathway models.

These fundamental investigations started in the late nineties with KEGG the Kyoto Encyclopedia of Genes and Genomes (genome.jp/kegg, Kanehisa and Goto, 1999; Kanehisa et al., 2013). By providing graphical and understandable representations of cellular signaling cascades and of the involved reactions, KEGG is an essential tool to grasp cellular behaviors.

While KEGG provides metabolic pathways as well, recent studies contribute to more complete repositories. For instance, the Reactome pathway database is a collection of more

than 1,500 human pathways precisely describing protein reactions such as the association into complexes, post translational modification or protein cleavage (Croft et al., 2013).

A last highly curated and reliable collection of human pathways is the Pathway Interaction Database (pid.nci.nih.gov) supported by the National Cancer Institute (Schaefer et al., 2009). This database contains *only* 137 human pathways annotated with effect of protein reactions on activity of nodes as well as potential effect of drugs.

IV.3 Unraveling cancer driving pathways

The main objective of genomics in cancer research, which here includes any kind of cell-wide profiling technique, is to identify active and therefore targetable molecules and pathways.

While genome-wide profiles can be informative on the excessive concentration or the altered form of specific biomolecules of each tumor, network representations of signaling pathways model their functioning and potentially the causes and consequences of these alterations. Referenced pathways remain theoretical and contain only possible interactions. Therefore, genomic profiles can potentially enlighten these pathways with some level of context-specificity by pointing towards subparts of cellular networks or pathways with a higher level of relevant alteration. For instance, a routine analysis of transcriptomic profile consists in the selection of pathways or biological processes composed of a significant number of over-expressed genes (*e.g.* Gene Set Enrichment Analysis, Subramanian et al., 2005). These analyses are simplistic and often difficult to analyze because of the number of identified pathways.

In order to identify more than simply enriched processes but active pathways and more importantly vulnerable spots, a great number of studies proposed a wide variety of methods and algorithms. These integrate various levels of prior knowledge to context-specific genomic or transcriptomic data.

Two main categories of analytical systems are discussed in the following sections. In the first section, methodologies attempting to lay genomic data on pre-constructed network, often to identify hotspots, are discussed. The last section focuses on methods relying on inferred and context-specific constructed network often to identify highly influential genes.

Predefined network based method

A network of biological interactions can be mathematically defined as a graph composed of vertices and edges. A wide variety of methods are based on this theoretical framework to extract active hotspots from networks. Generally, methods rely on a PPI network and use it to integrate transcriptomic data and identify highly expressed subparts. Although it can be considered to be generally the case, mRNA levels is not systematically representative of the activity of nodes in a protein network (Chen et al., 2002; Ghazalpour et al., 2011). However, the coordinated up-regulation of genes involved in the same modular subpart of

a functional network can in some cases attest the protein over-expression of the module or pathway and potentially its activity.

One type of analysis of transcriptomic profiles aims at developing predictive models for diagnostic or prognostic purposes. Because of the high number of dimensions of gene expression data and the low number of samples, standard supervised classification models are often unstable and have low reproducibility. To overcome this problem termed the *curse of dimensionality* (see section IV), several groups proposed extension of standard algorithm to take into account prior information in the form of pathways or networks. This *a priori* data is seen as a compilation of the current knowledge of cell biology. Networks have therefore been used to construct a new penalty (Zhu, Shen, and Pan, 2009) or kernel function (Lavi, Dror, and Shamir, 2012; Rapaport et al., 2007) in a Support Vector Machine model for binary classification problem or to filter expression profiles by defining a new metric in an unsupervised framework (Rapaport et al., 2007).

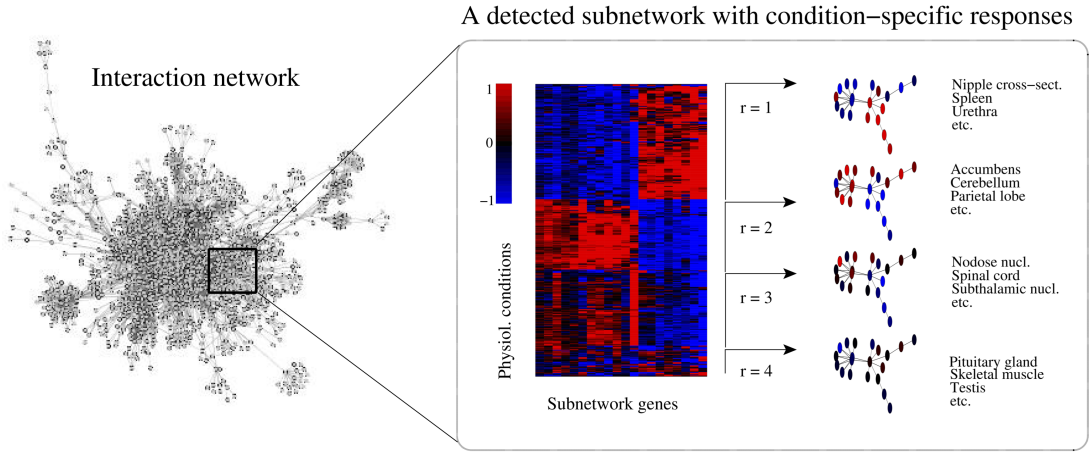


Figure IV.18: *Identification of transcriptionally active sub-networks. Active sub-network extraction consists in the identification of modular subparts of a cell wide interaction network (left) which contain genes with a specific expression level pattern (right). The searched expression patterns are often simply based on the over-expression of genes between two predefined conditions in the transcriptomic experiment. (from Lahti, Knuuttila, and Kaski, 2010)*

Another approach aims at identifying network hotspots or active modular subnetworks in a cell wide interaction network as depicted in figure IV.18. This is sometimes referred to as significant-area-search or active-module detection (Mitra et al., 2013). Most of these methods use the same set of data, a PPI network and transcriptome profiles of two distinct conditions (*e.g.* cancer vs. normal samples) resulting in a list of differentially expressed genes. The extraction of active modules is transformed into an optimization problem in which the most differentially expressed connected module is to be found and can be referred to as the Maximum-Weight Connected Sub-graphs (Dittrich et al., 2008). This problem was first formulated in early 2000's with the work of Trey Ideker on the jActiveModules

software (Ideker et al., 2002). Several solutions based on other search algorithms were introduced and were based on graph clustering (Gu et al., 2010), simulated annealing (Guo et al., 2007) or based on solving a Steiner tree problem to obtain exact solutions (Dittrich et al., 2008). A last method is more relevant to the use of a PPI by searching for active modules through the integration of the results of proteomic studies (Nibbe, Koyutürk, and Chance, 2010).

While these methods differ in their search algorithms and the way they score modules, they are all based on the comparison of two or more phenotypes and often rely on simple univariate statistics on genes (*t-test* for instance). In order to remove potential bias introduced by these often unreliable or incomplete annotations, more global analytical search for network modules of interest were proposed. For instance, the Module Analysis via Topology of Interactions and Similarity Sets *MATISS* (Ulitsky and Shamir, 2007) integrates a PPI network with links of co-expression between genes to find jointly active subnetworks. Another meaningful approach is the non-transcriptomic based *hotnet* algorithm (Vandin, Upfal, and Raphael, 2011) aims at identifying highly mutated sub-network using network heat diffusion. The Mutual Exclusivity Modules in Cancer MEMo also aims at extracting mutated modules but uses an interesting property of mutual exclusivity of alterations of genes in the same pathway (Ciriello et al., 2012). Finally, *Netreponse* (Lahti, Knuuttila, and Kaski, 2010) is a more general approach to identify subnetworks with coordinated expression patterns, termed response, in an unspecified subset of samples as illustrated in figure IV.18. It uses a model based approach which can be used both as a class discovery (clustering) and feature selection approach with a network constraint.

Another class of method aims at identifying sample-specific network activities. Conversely to the precedent methods, which often consist in analyzing datasets with pre-defined sample classification, patient-specific measures have the potential to be used clinically in a personalized treatment framework. In fact, most of these analytical systems are considered as dimension reduction methods as they are based on or related to Principal Component Analysis.

The first of these proposed methods is called Network Component Analysis (Liao et al., 2003). Figure IV.19 shows the use of NCA in a simple context of the analysis of hemoglobin species from absorbance spectra. This example is simply transposable to regulatory network in which the observed values are changes in gene expression, the *a priori* connectivity diagram is a large scale regulatory network derived from ChIP experiments or TFBS promoter scanning analysis and the sought values are the level of activity of transcription factor. This is fundamentally different than the previous method as the goal is not to use a predefined structure to identify gene that might have a common function but to identify the true level of activity of a molecule based on observation of downstream entities. Indeed, the mRNA level of a transcription factor is not representative of its true activity on its target gene as this is dependent not only on the level of translation but also on the cellular localization of the regulator protein, on the level of post-translational modifications and on the vicinity and the activation states of co-factors. As all these are

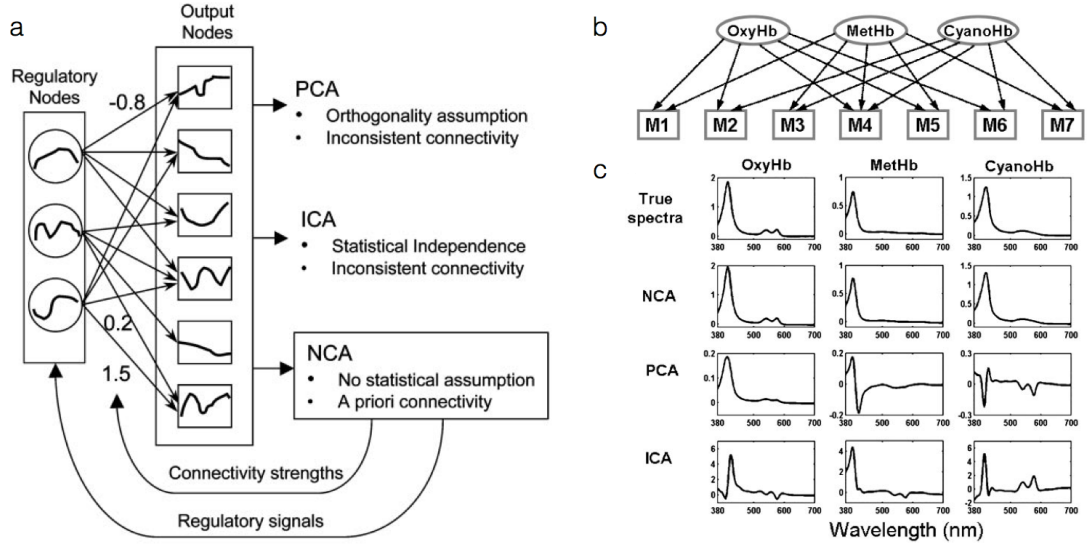


Figure IV.19: *Network component analysis. a. Schematic representation of a simple regulatory system with output nodes dependent on combination of regulatory node values. Unlike Principal or Independent Component Analysis (PCA and ICA respectively), Network Component Analysis (NCA) takes into account the known connectivity between the sought regulatory nodes and the observable output nodes. b. and c. Example of the use of NCA on absorbance spectra of hemoglobin solutions. b.Connectivity diagram between the pure components contained in the solution (regulatory nodes) and the seven measured mixtures (output nodes). c. Comparison of NCA, PCA and ICA estimated regulatory signals with the true signals. (from Liao et al., 2003)*

difficult to measure and to analyze, the NCA proposes to infer the activation state of regulator by the level of transcription of its putative target genes. NCA takes as an input a connectivity diagram containing only potential interactions, which in fact did not needed to be actual interaction happening in the analyzed samples, and a set of transcriptomes in which the transcription factor activities (TFA) would be inferred. The original NCA method required the number of samples to be greater than the number of regulatory nodes. Several improvements overcame this limitation (Galbraith, Tran, and Liao, 2006; Noor et al., 2013) and other approaches based on Partial Least Square were also proposed (Boulesteix and Strimmer, 2005). A particularly interesting feature of NCA is to take only putative interactions between regulatory and output nodes while the method re-weights these interactions based on the observed data.

Finally, a last type of technique relies on full pathway models to infer active and driver genes/proteins. A first method termed DriverNet (Bashashati et al., 2012) was developed to identify drivers among genes presenting genetic alterations, mostly copy number and point mutation, which have a measurable effect on their target or downstream genes. To this end, DriverNet uses both protein interactions and regulatory interactions to link genetic alterations to coordinated gene expression modification among co-regulated genes.

The main intention of DriverNet is to consider infrequently altered genes as significantly altered genes on the basis of the effect of their aberrations on their downstream pathways.

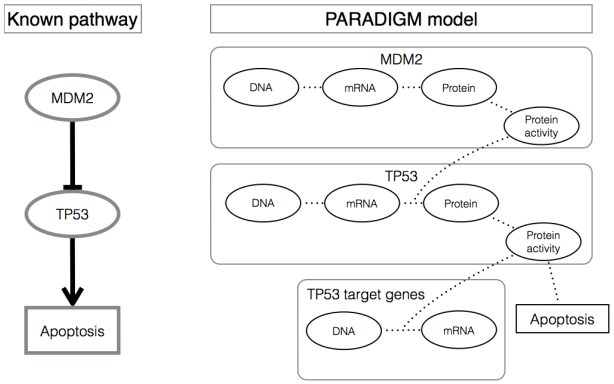


Figure IV.20: *PARADIGM pathway modeling. Based on a set of directed interaction with known effects between parent and child node, which includes various biomolecules and higher level cellular processes (left), paradigm builds a corresponding factor graph (right). The pathway model includes observable variables (DNA, mRNA and for some cases Protein level) and infers unobservable variables (protein activity, apoptosis) based on the coherence between the measurable data of child and parent nodes. (adapted from Vaske et al., 2010)*

PAthway Recognition Algorithm using Data Integration on Genomic Models also known as PARADIGM is an inference system using thoroughly annotated pathway models from the Pathway Interaction Database (Schaefer et al., 2009) to model their function and activity in a patient/sample specific manner. The model used by PARADIGM consist in the transformation of known pathways into factor graphs as illustrated in figure IV.20. The goal is to infer the values of hidden variables such as protein or cellular process activity by searching for the set of possible values with the highest consistency regarding the observed value. For instance in the model depicted in figure IV.20, given an amplification of the *MDM2* gene, an over-expression of it's mRNA and an under-expression of all the genes described to be activated by TP53, the TP53 protein is expected to be inactive and that the MDM2 protein is active. This highly successful approach was used in several of the TCGA marker paper of full cancer genome profiles (Cancer Genome Atlas Network, 2012, 2014; Kandoth et al., 2013). It was also used to identify functional mutations by underlining mutation-driven inconsistency in pathways (Ng et al., 2012) and for patient subtype identification (Sedgewick et al., 2013). Overall, PARADIGM is an influential algorithm for it tackles several of the most important challenges in cancer genomics by integrating all levels of data available and by returning sample specific activities of nodes in signaling pathways and therefore designating potentially active therapeutic targets with high specificity. However, it requires an exceptionally accurate description of cellular pathways which themselves are condition unspecific and error prone.

Network identification method

As thoroughly discussed in the previous section, integrating repositories of biomolecular interactions with genome-wide datasets has the main drawback of taking into account protein or regulatory interactions that might not occur in the studied cells. To overcome this problem, a whole area of research is dedicated to the development of methods aiming at the construction of context-specific networks.

Based on genomic alterations, a first type of method uses an interesting property of genetic aberrations in cancer to identify mutated pathways and driver mutations. It was observed and sometimes used in pathway discovery methods (*e.g.* Ciriello et al., 2012) that genes in the same pathway are rarely mutated in the same sample. Therefore, based on the simple fact that only one mutation in a pathway is sufficient to alter its function, the Dendrix algorithm (Vandin, Upfal, and Raphael, 2012) identifies altered signaling pathways without any prior knowledge by searching for sets of genes with mutual exclusive and frequent mutations as depicted in figure IV.21.

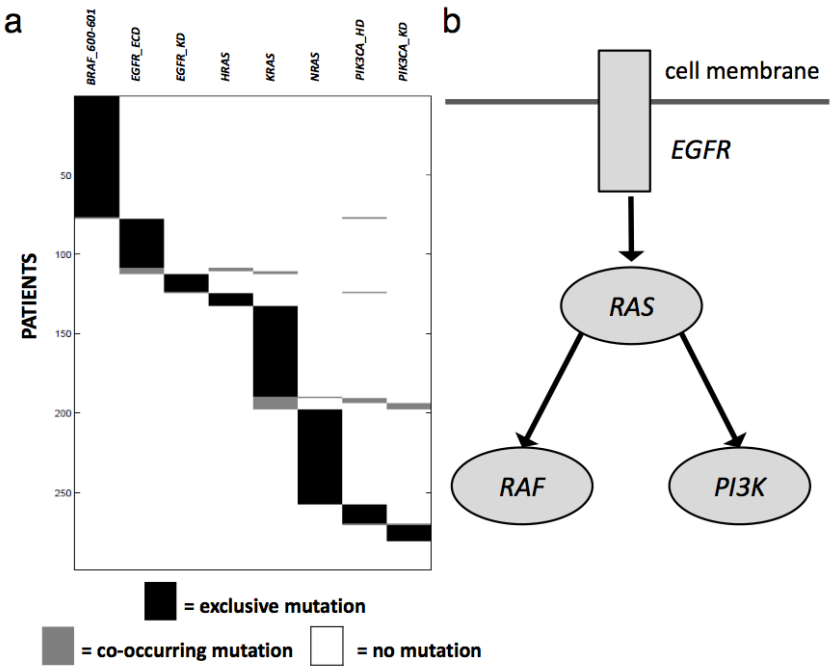


Figure IV.21: *Mutually exclusive mutations. a. Selection of six genes for which their combined alterations are mutually exclusive and cover a large portion of samples. Black bars indicate alterations in genes defined at the top of the plot in a specific sample. Grey bars are co-occurring mutations. b. Known pathway composed of these mutually mutated genes.*
ECD: Extra-Cellular Domain; KD: Kinase Domain; HD: Helical Domain.
(from Vandin, Upfal, and Raphael, 2012)

However, the dominant area of research of context-specific network discovery is undoubtedly transcriptomic-based. Often referred to as network inference or reverse engineering, the multitude of proposed learning systems usually proceed in a similar manner. These all use a normalized transcriptomic dataset composed of thousands of mRNA levels measured in a set of samples resulting in a matrix of gene expression data. Then, a measure is used to identify pairs or sets of genes that have significantly correlated or co-dependent mRNA transcription rate levels. This last step is particularly variable between methods in the way that either pairs or association of sets of genes are scored. However, these all result in the construction of a large scale regulatory network containing the highest ranked (or most statistically significant) associations.

In an ideal experimental design, gene regulation would be inferred by the dependency between transcription rates and several measures of regulatory proteins such as post-translational modification and cellular localization. However, protein levels are still difficult to obtain for large number of samples. Therefore, regulatory network discovery can meanwhile only rely on transcriptomic experiments. This strong limitation is handled in two different ways by inference systems. First, by searching for abstract networks. These methods do not aim at identifying real interactions but higher-level dependencies between mRNA levels. This is generally the case when no prior information is used and is sometime called co-expression network (Carter et al., 2004). Second, by constructing real regulatory networks between transcription factors and target genes (*e.g.* Fletcher et al., 2013; Lefebvre et al., 2010). These methods require to preselect regulatory genes and assume that the mRNA levels of regulators are representative of their activity. Given the wide number of possibility for an over-expressed regulatory gene to not be active (translational regulation, post-translational modifications, cellular localization,...) this is a highly precarious assumption. However, as discussed in section I.4, the activation of a signaling pathway does result in the activation of a first set of regulator at their protein level but also usually activates a late response by triggering the transcriptional activation of new TF.

Despite these great difficulties, network inference aims at deriving functional and context-specific regulatory models from gene expression data only. A network representation of the regulation actually taking place in cancer cells holds the promise of identifying key nodes in the network, their role in tumorigenesis and predicting the effect of their silencing by small inhibitors. In short, to accurately identify oncogenes, tumor suppressor genes and more importantly effective therapies.

Nearly all network inference method identifies regulatory interactions by scoring gene-gene or TF-gene pairs. Therefore, Pearson's or Spearman's correlations are often used to construct what are then called co-expression network (Carter et al., 2004; Langfelder and Horvath, 2008). More successful approaches are based on Mutual Information, an information-theoretic measure of co-dependency (Butte and Kohane, 2000). In particular, two broadly appreciated algorithms build up on Mutual Information to infer robust regulatory networks. The *context likelihood of relatedness* CLR algorithm (Faith et al.,

2007) selects the best TF-gene interactions compared to a background Mutual Information score and remove insignificant pairs. The *Algorithm for the Reconstruction of Accurate Cellular Networks* ARACNE (Margolin et al., 2006a,b) is probably one of the most successful algorithm in terms of biological application and findings (discussed later in this section). It is also based on Mutual Information to score TF-gene pairs but it applies a filter to remove putative indirect interactions using the Data Processing Inequality concept (Margolin et al., 2006a).

Another category of reverse engineering methods assumes a linear dependency between regulators and targets. The simplest methods consider the expression of a gene as a weighted sum of its regulator and infers these weights using a simple linear regression model (D'haeseleer et al., 1998). More recent algorithms basically use the same linear model but introduce a penalization term to induce sparsity in the inferred network using Least Angle Regression (LARS) (Haury et al., 2012) or Least Absolute Shrinkage and Selection Operator (LASSO) (Someren et al., 2006). Penalization terms have also been used to infer networks with specific structures such as modular networks (Chiquet et al., 2009).

Finally, other types of methods were developed and based on various areas of mathematics and machine learning to solve the problem of constructing regulatory networks. Based on Bayesian networks (Husmeier, 2003) or ordinary differential equations (Quach, Brunel, and Buc, 2007), some methods aim at identifying the true underlying mathematical model of gene regulation. Similarly to the penalized regression methods, another type of statistical algorithm tackles the problem of identifying the regulators of a gene as a feature selection problem. The idea is to identify variables which best explains a given random variable, *i.e.* the expression of a target genes. Therefore, a popular supervised machine learning algorithm with an intrinsic feature selection capability, Random Forest (Breiman, 2001), was used as to select for each gene the most influential variables thereby predicting its regulators (Huynh-Thu et al., 2010). Unlike penalized regression methods, this tree-based algorithm called *GENIE3* (GEne Network Inference with Ensemble of trees) does not assume linearity between regulator and target gene mRNA levels.

Most of the proposed learning systems identify pairs of interaction, although it is known, especially in higher eukaryotes, that genes are regulated by a complex set of competing and synergistic regulators (see section I.2). To overcome this problem, an algorithm based on frequent itemset mining called LICORN (Elati et al., 2007) aims at identifying for each gene the best combination of TF.

As a great number of methods were proposed to reconstruct genome-wide regulatory network from gene expression data, it is necessary to compare them it terms of prediction accuracy. Usually, a golden standard is used as ground truth to compare predictive models. However, in the case of gene regulation, very little interactions are highly reliable and more importantly they are often context-specific. For instance, ChIP experiments identify TF bound genomic locus. However, binding does not necessarily result in regulation of gene expression (see figure IV.17). Similarly, perturbation experiment (siRNA target

depletion for instance) results in a list of differentially expressed genes which might be in fact regulated indirectly, that is involving an intermediary TF or regulatory molecule, by the depleted TF.

The Dialogue on Reverse Engineering Assessment and Methods (DREAM) project is a community challenge to reconstruct large-scale regulatory network. It provides to challengers a transcriptomic dataset on which network inference algorithms can be applied and compares their results with benchmark data. In early challenges (Marbach et al., 2010), synthetic gene expression data were generated *in silico* to ensure the quality of the gold standard. Later, real gene expression datasets for *Escherichia coli*, *Staphylococcus aureus*, *Saccharomyces cerevisiae* were used and the predictions were compared to unfortunately unreliable ChIP data or TFBS predictions (Marbach et al., 2012b). The main outcome of these challenges is the extremely poor precision of all these methods with less than 10% AUPR (Area Under the Precision Recall curve) for *E. coli* and less than 5% in yeast.

Although these extremely low quality metrics should be put into the perspective of the lack of exact gold standard, it can also be explained by the incomplete information included in transcriptomes and strong assumptions as discussed earlier. Overall, the conclusion is that a single interaction taken out of an inferred network has an extremely low reliability with a high probability of being false.

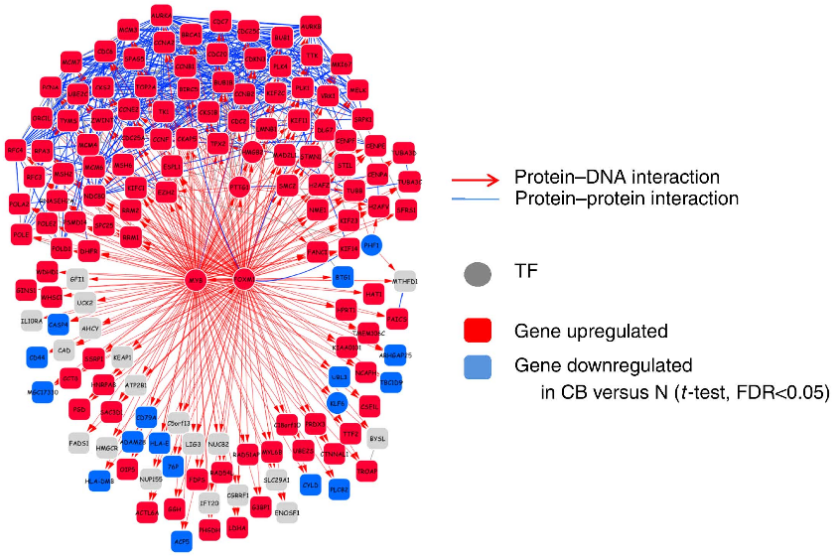


Figure IV.22: *Master Regulator Inference. Sub-network of the FOXM1 and MYB inferred as master regulators of from a Human B-Cell proliferation. (Lefebvre et al., 2010)*

However, this does not undermine the potential of network inference algorithm to provide an informative model (Della Gatta et al., 2012). Evidently, this requires the development of new methods to analyze network and retrieve functional information from the error-prone network. Significant foundation in regulatory network analysis was laid

by an extension of the Gene Set Enrichment Analysis (Subramanian et al., 2005) in the master regulator inference algorithm (MARINA, Lefebvre et al., 2010). Using as an input a network and a set of gene of interest, for instance the most differentially expressed genes between two conditions of interest, MARINA identifies the most specific regulators of these genes. This method was first used to infer the master regulators of human B-cell proliferation (Lefebvre et al., 2010) as depicted in IV.22. An extension of the algorithm was applied to prostate cancer of mice and men in a *cross-species* regulatory network analysis to identify synergistic regulators of prostate malignancy (Aytes et al., 2014). Another application of this algorithm is particularly compelling in the context of pathway analysis and discovery. The MARINA algorithm was improved to identify the downstream regulators of the *FGFR2* gene, which was only known to be a risk factor of breast cancer (Fletcher et al., 2013). This was done by exogenously activating *FGFR2* in breast cancer cell lines and searching for the master regulators of the gene deregulated by the induced perturbation. Although no intermediary connections between *FGFR2* and the discovered downstream regulators are identified, this is a major step towards the complete definition of cancer driving pathways.

Conclusion Overall, the goal of network-based model is to identify targetable oncogenes as active nodes in large-scale networks. Relying on previous knowledge allows to identify network hotspots (Netresponse has a nice and easy general purpose R package at bioconductor.org, Lahti, Knuuttila, and Kaski, 2010) or to use functional pathway models to discover weak points (*e.g.* PARADIGM available at sbenz.github.io/Paradigm, Vaske et al., 2010). Another possibility is to rely only on data to identify master regulators from context-specific reconstructed network (*e.g.* LICORN for co-regulation models in the CoRegNet R package available in bioconductor.org, ARACNE, MARINA and extensions are well implemented in the RTN R package at bioconductor.org, Fletcher et al., 2013) or by integrating both genomic and transcriptomic data to discover genes with genetic alteration that have an impact on the cell phenotype (Akavia et al., 2010, provides software at c2b2.columbia.edu/danapeerlab). However, better solutions might arise from methods using knowledge as an *a priori* instead of a strict framework to work on. The network component analysis provides a good example of using prior knowledge of gene regulation without forcing all provided regulations as true interactions. Finally, as network reconstruction methods showed the highest analytical value so far, methods integrating prior knowledge in the inference step also hold great promises (*e.g.* a interesting study by Setty et al., 2012 and the CoRegNet package implementing network refinement solutions based on regulatory datasets).

Results

In order to identify and understand the regulatory networks and signaling pathways driving bladder cancer, a project entitled INSight (Identification of Networks Specifically altered during tumorigenesis) was initiated by the *laboratoire d'oncologie moléculaire* and funded by the french *Institut National du Cancer*. It also involved two computational teams from the ISSB (Institute of Systems and Synthetic Biology,issb.genopole.fr) and the LIPN (Laboratoire d'Informatique de Paris Nord,lipn.univ-paris13.fr). The project aimed at developing algorithms to infer the regulatory network of normal urothelial growth and differentiation in order to identify its disruptions leading to cancers of the urinary bladder. Based on the resources of the INSight project and on a proteomic analysis of the proteins participating in the signaling pathway of FGFR3, the objective of my study was twofold:

1. Develop methods to reconstruct and analyze networks by integrating context-specific tumor profiles (transcriptomic and/or proteomic)

2. Propose and validate dysregulated and driver networks of bladder cancer

I will first present an analytical approach based on an algorithm to infer large-scale cooperative regulatory networks termed LICORN (Elati et al., 2007). First, improvements of the search algorithm were proposed in a joint work with the team of Céline Rouveirol at the LIPN to increase the accuracy of LICORN on Human data. This work was published in 2014 (Chebil et al., 2014, article available in appendix A). Using the inferred regulatory network, I then devised a procedure to estimate the influence of transcription factors on their target genes. I first proposed this method as a robust dimension reduction approach for transcriptome data analysis (Nicolle, Elati, and Radvanyi, 2012, article available in appendix B). I then designed and implemented several additional methods to both improve the reliability of the predicted network using external data and to facilitate its analysis through a visualization tool of the inferred transcriptional programs. Finally, I integrated the entire pipeline, from the inference of the network to its visualization, in a Bioconductor package termed CoRegNet. This work was submitted to the journal *Bioinformatics* (article available in appendix C).

I will then describe the use of the CoRegNet package through the characterization of the transcriptional programs of bladder cancer. This analysis showed the specific association of the activity of distinct transcriptional programs to each bladder cancer subtypes. The integration of transcriptional activity with genomic alteration profiles highlighted driver transcriptional programs. This emphasized the role of $PPAR\gamma$ as a driver of luminal-like bladder cancers and identified a potential driver of the basal-like bladder cancers, *FOXM1*, for which I experimentally assessed the impact on cellular proliferation.

In a third chapter, I will present my study of the networks controlling the proliferation and differentiation of normal urothelial cells and the contribution of these normal transcriptional programs to urothelial carcinogenesis. This work illustrated the extent to which normal regulatory programs are active and sustained during neo-plastic transformation. Moreover, it identified two transcription factors with major role in bladder cancer. First, a constitutively activated master regulator of proliferation for

which I experimentally validated its impact on bladder cancer cell proliferation. Second, a master regulator of the urothelial terminal differentiation program for which we discovered frequent mutations and a dual role in carcinogenesis of the bladder.

I will then present a novel algorithm to reconstruct protein complexes and signaling pathways using both proteomic data from an affinity purification followed by mass spectrometry and a repository of protein-protein interactions. The algorithm entitled Pepper is proposed as a Cytoscape application (apps.cytoscape.org/apps/pepper) and was published in the journal *Bioinformatics* in 2014 (Winterhalter et al., 2014, article available in appendix D).

The last chapter describes my study of the signaling pathway of FGFR3, a growth factor receptor frequently mutated in bladder cancer. This work aimed at constructing the entire pathway downstream of FGFR3, including the downstream transcription factors. Based on a proteomic analysis of the protein partners of FGFR3 in a bladder cancer cell line, Pepper was used to extend the signaling pathway and identify transcription factors linking the signal transduction proteins to the transcriptional impact of the activation of FGFR3.

I also included in appendix E three articles in which I was involved. In the first work by Elati et al., 2013, I developed a multi-view classifier fusion algorithm. In the second study by Mahmood et al., 2013, I analyzed the relationship between copy number and gene expression in several cancers. Finally, in the work of Ho et al., 2012, I analyzed expression patterns of a normal human urothelium transcriptomic dataset.

Chapter 1

CoRegNet: reconstruction and integrated analysis of co-regulatory networks

1.1 Introduction

Cancer cell behavior is often sustained by aberrant molecular signaling such as those induced by growth factors and signal transducers. By bridging cellular signaling to the control of gene expression and ultimately cellular phenotypes, Transcription Factors (TF) play crucial role in the maintenance of a malignant cellular state. The capabilities of regulators of transcription to driver a cellular phenotype is well exemplified by the recent breakthroughs in cellular reprogramming in which only several TF are needed to trans-differentiate a particular cell type (usually fibroblasts) into another (Lee and Young, 2013).

The transcriptional regulation of genes involves the cooperation of a large number of proteins to regulate the nuclear translocation, association with DNA in particular positions of the genome and finally transcriptional activation (Hill and Treisman, 1995; Panne, 2008). Thus, the activity of a particular TF is dependent on the availability of specific co-factors and other regulators to define its target genes. For instance, the transduction of Gata4 in mouse fibroblasts can lead to either cardiomyocyte or hepatocyte reprogramming depending on its co-induction with either Tbx5-Mef2c or Hnf1a-Foxa3 respectively (Huang et al., 2012; Ieda et al., 2010). These examples of engineered cellular phenotypes through TF induction underline the importance of identifying the set of active co-regulators leading to a phenotype-specific and more interestingly a disease-driving transcriptional program. Furthermore, the potentially small number of overactive TF and their direct role in maintaining a malignant cellular phenotype make transcriptional regulators appealing for targeted therapies (Darnell, 2002).

Transcriptional programs and more generally transcriptional regulation is usually modeled using gene regulatory networks composed of edges linking regulators to their

target genes. Large-scale network models are used to gain insight into complex biological processes (Bandyopadhyay et al., 2010; Behrends et al., 2010), identify driver pathways of clinically relevant cancer subtypes(Dutta et al., 2012) and overall holds great promise in the understanding of diseases (Goodarzi, Elemento, and Tavazoie, 2009) and cellular behavior (Karr et al., 2012).

Numerous methods discussed in introductory chapter IV.3 were proposed to reconstruct, use or characterize phenotype-related networks or network-modules. Overall, the most successful approaches in terms of biological discovery were those that focused on context-specific network or that aimed at characterizing sample-specific network activation measures.

Given the importance of identifying the set of active transcription factors driving the behavior of a particular cell, I propose a global analytical system composed of a set of methods to *a*) reconstruct a context-specific large-scale regulatory network; *b*) estimate the sample-specific activity of each of the TF using a novel measure of regulatory *influence*; *c*) infer sets of cooperative regulators as part of an active transcriptional program; *d*) integrate prior evidences on regulatory interactions such as TFBS (TF binding sites) or ChIP-seq/ChIP-on-chip data; and *e*) identify sample- or subtype-specific active transcriptional programs through a visualization using both a network of cooperative regulators and TF/gene specific data such as the expression, *influence* and genomic alterations when available. Each of these methods are further detailed in the following sections.

Altogether, the goal of this network-based strategy is to ease the analysis of large-scale gene expression data by modeling the effect of master regulators on the transcriptome. The idea is to identify the transcription factors responsible for a particular transcriptomic state and to some extent, to a particular cellular phenotype. The identification of highly influent and active regulators in normal and cancer samples as well as their validation on corresponding cell lines with the most representative transcriptional programs is shown in the next chapters.

The entire analytical pipeline is embodied in an R/Bioconductor package entitled CoRegNet (bioconductor.org). The package is divided in three parts described in the next sections. First, a method to infer large-scale regulatory network from transcriptomic data is proposed to identify cooperative TF forming specific transcriptional programs. Second, the activity of these TF is estimated sample by sample to identify sample- or subtype-specific active sets of regulators. Third, a visualization tool is used to analyze a given network with a given transcriptomic dataset to easily apprehend the set of active regulators in each sample (or subtype of samples) with the possibility to easily integrate additional data (DNA copy number alterations, TFBS...).

A first version of the network inference method was presented at the IEEE International Conference on Bioinformatics and Biomedicine in 2013 (BIBM 2013) and published in the IEEE Transactions on NanoBioscience:
Chebil I, Nicolle R, Santini G, Rouveirol C and Elati M (2014) Hybrid Method Inference

for the Construction of Cooperative Regulatory Network in Human. IEEE Transactions on NanoBioscience, 13: 97-103.

The measure of transcriptional regulatory influence was presented at the 11th International Conference on Machine Learning and Applications (ICMLA 2012):
Nicolle R, Elati M and Radvanyi F (2012) Network Transformation of Gene Expression for Feature Extraction. IEEE 11th International Conference on Machine Learning and Applications (ICMLA), 1: 108-113.

This work including the set of methods of the CoRegNet package and the visualization tool has been submitted to the Bioinformatics journal of the Oxford Publishing Group.

1.2 Reconstruction of large-scale cooperative regulatory networks using LICORN

Network inference methods are used to reconstruct the regulatory network underlying a particular set of transcriptomes. This task is sometimes called reverse-engineering as it attempts to describe the regulation events which resulted in the observed transcriptome. In the simplest cases, the regulatory events that are searched are described as a pair composed of a Transcription Factor responsible for the variation of the expression of a target gene. In fact, most of the available methods only try to extract these pairs (Marbach et al., 2012b). However, our goal is not only to enumerate all the regulatory interactions at work in a given set of samples but also to identify sets of cooperative regulators. The inference method proposed in the CoRegNet package is based on an algorithm that proposes to search for sets of TF regulating target genes. To do so, the package implements an improved version of the LICORN (Learning cooperative regulation network) algorithm (Elati et al., 2007). LICORN uses a discretized version of the transcriptomic data in which each gene is associated to a vector of expression values g_d defined in $\{-1, 0, 1\}$, respectively under-expression, normal expression and over-expression. The discrete values are obtained from the vectors of continuous mRNA level usually defined by microarray or RNA-seq techniques and which is noted g_c (for *c*ontinuous as opposed to g_d for *d*iscrete). LICORN uses a frequent itemset mining approach (Agrawal, Imieliński, and Swami, 1993) to first enumerate all putative sets of co-regulators (sets of cooperative TF) and then extract for each gene the best sets of co-activators A and the set of co-inhibitors I. LICORN enumerates for each genes local gene regulatory network models $GRN = (A, I, g)$ in which A and I are non-intersecting sets of regulators where both cannot be empty.

In the CoRegNet package, the discretization of the data is user-defined. However, a standard threshold-based procedure is implemented for default use. As in previous applications of LICORN (Chebil et al., 2014; Elati et al., 2007), the value of a given gene in a given sample is set to 1 if it is above a predefined threshold, to -1 if it is below another threshold and 0 otherwise. The simplest use of such discretization technique requires to center the expression of each gene either on the mean of a set of reference sample when

relevant or on the mean of each gene in all samples. In the case of a ratio with a reference sample, a fold change can be used as a discretization threshold. Otherwise, in order to adapt to each dataset, the default threshold is set to the standard deviation of the entire values of gene expression resulting in a balanced discrete dataset with frequent non-zero values for highly variable genes and a constant 15% of 1 and 0 (approximately) for all g_d.

In order to identify the most plausible combinations of cooperative regulators, LICORN introduces a scoring scheme to compare the observed expression of a gene g_d to the expected expression $\hat{g}_d$ given the proposed GRN model. The computation of $\hat{g}_d$ is represented in figure 1.1. It was designed to favor the discovery of co-regulators for which the combination is necessary for the regulation of the target gene. First, the values of the q activators in A ($q = |A|$) and the p inhibitors ($p = |I|$) are aggregated using an extended AND logical function which sets the value of A (or I) to 1 or -1 only if all activators (or inhibitors) have a 1 or -1 expression value respectively, as in the following equation for a given sample:

$$E\_AND(A) = \begin{cases} 1, & \text{if } \forall a_i \in A : a_i = 1 \\ -1, & \text{if } \forall a_i \in A : a_i = -1 \\ 0, & otherwise \end{cases}$$

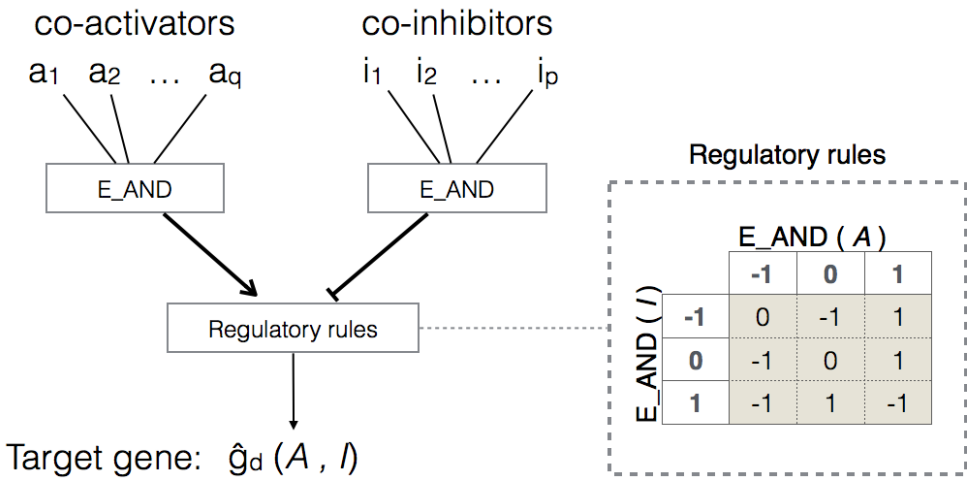


Figure 1.1: *LICORN regulatory rules. The expected expression $\hat{g}_d$ is determined for each sample by an aggregation of the expression of its co-activators in A and co-inhibitors in I. The E_AND function aggregates the expression of co-regulators by setting to 1 or -1 only if all co-regulators are set to 1 or -1 and 0 otherwise. The expression of co-activators and co-inhibitors is then aggregated using a set of regulatory rules in which if on of the co-regulators is set to 0 the expression of the other is determinant. (from Elati et al., 2007)*

Given $E\_AND(A)$ and $E\_AND(I)$, $\hat{g}_d$ is determined by a set of rules illustrated in figure 1.1. Given the vectors of observed and estimated expression, each GRN models

are then scored by the differences between $\hat{g}_d$ and g_d using the Mean Absolute Error: $\sum |g_d - \hat{g}_d|$.

Using the previously described computation on discretized expression data, LICORN is able to identify putative *GRN* in a time efficient manner. The search algorithm is specifically designed to extract sets of cooperative regulators, which *together* regulate the target genes whereas each of the regulators have no effect individually. The major drawback is that since the data is discretized in three values, a large number of possible *GRN* will end up with the same best score. For instance, on a bladder cancer transcriptomic experiment (Stransky et al., 2006), LICORN identifies a mean of 45 *GRN* per gene that have the best score (minimum Mean Absolute Error). Figure 1.2 shows the distribution of the number of *best GRN* gene showing that many putative *GRN* cannot be differentiated solely based on the discretized expression data.

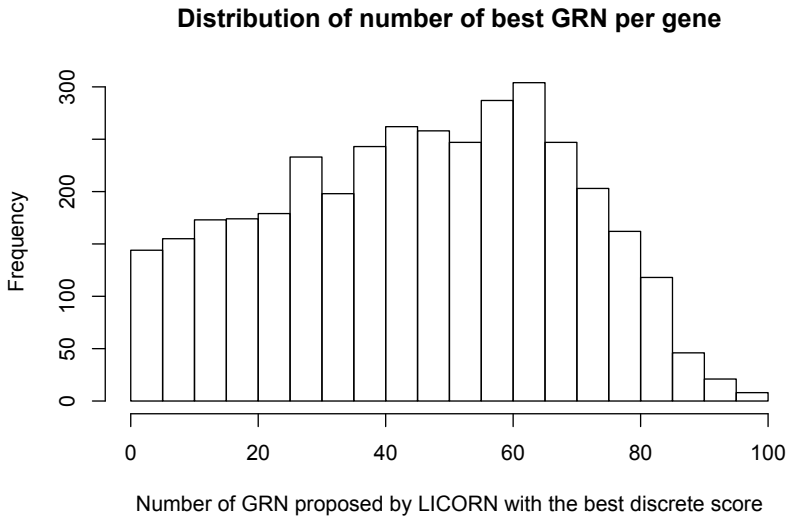


Figure 1.2: *LICORN identifies many putative GRN with identical scores. For each gene in a transcriptomic dataset, the 100 best GRN were extracted using LICORN and the number of GRN with the best score was kept. This histogram represents the distribution of the number of best GRN per gene.*

Transcription Factors

At this point, it is important to note what are here referred to as transcription factors. As discussed in the following sections, the large-scale regulatory network is inferred using a set of genes in the transcriptomic data that identified as Transcription Factors. The set of transcription factor used by the methods of the CoRegNet package and throughout the following studies is defined mostly as proteins that have been described to have a direct

role in the regulation of transcription although not necessarily mediated by DNA binding. The list of 1,988 transcriptional regulators from the FANTOM consortium (Ravasi et al., 2010) was used as a reference from which Histone proteins were removed and to which missing TF found in the TRANSFAC database (Matys et al., 2006) were added. These added TF were mostly from the Krüppel-like family of transcription factors, zinc-fingers and zinc-fingers SCAN domain containing families of genes. This resulted in a list of 2,020 genes encoding for proteins which have a role in transcriptional regulation. This list in fact includes many protein that do not bind to DNA to activate downstream gene loci. Many transcriptional co-factors, such as transcription-involved DNA topoisomerase (*e.g. TOP2B*) or enhancers of transcription factors (*e.g. TP53BP1*), are included in the list.

1.3 Hybrid-LICORN

A three value discretization of the expression dataset allows to easily encode and search for cooperative regulation models. However, it does not allow to differentiate between several proposed models and overall induces a loss of information during the discretization process. To overcome this, each local *GRN* models are tested in the space of continuous gene expression values by fitting a linear model using all co-regulators in A and I as predictor variables and the target gene g as the response.

We first proposed this approach in collaboration with Ines Chebil as a novel hybrid algorithm of network inference using both the discretized and continuous space of gene expression. The published article is available in appendix A. The objective of the original method entitled h-Licorn (hybrid LICORN, Chebil et al., 2014) was to select pairs of TF-gene in order to compare the performances of the algorithm to other methods in predicting correct regulatory interactions. For each *GRN* proposed by LICORN, a linear model is used to estimate the expression of the target as follow:

$$\hat{g}_c = \beta + \sum_{i=1}^{q+p} \alpha_i * r_i$$

with q the number of co-activators $q = |A|$, p the number of co-inhibitors $p = |I|$ and r_i belonging to the set of regulators $r_i \in A \cup I$. Here $\hat{g}_c$ is used as opposed to $\hat{g}_d$ to denote that it is an estimate of the continuous expression of g.

In this classical setting of linear regression the target gene expression is the response variable and the explanatory variables are the p co-activators and the q co-inhibitors. The parameters α and β are estimated using the Ordinary Least Squares method. In a bootstrapped setting, the α are then used to weight the confidence of the method in the prediction of a regulatory link between a regulator and a target gene.

The performance of h-Licorn was tested on *in silico* dataset of the 5th DREAM challenge (the-dream-project.org, Marbach et al., 2012b). In this project the set of regulatory interactions are considered as known and the provided gold standard can be

Performance measure
Most measures of prediction performance are based on four terms representing counts of correct and wrong predictions similar to the type 1 and type 2 errors in statistics. These require the predictions of the tested algorithm to be assigned classes (instead of weights, scores or probabilities). These are based on the comparison of the set of instances predicted as positive or negative and the true class of each of these instances and are defined as follow :

TP True Positive is the number of instances predicted as positive and that are positive
TN True Negative is the number of instances predicted as negative that are negative
FP False Positive is the number of instances predicted as positive that are negative
FN False Negative is the number of instances predicted as negative that are positive

Based on these four measures, a number of metrics are often used to estimate the overall performances and are defined as follow :

Recall or **TPR** (True Positive Rate) $\frac{TP}{TP+FN}$
Precision $\frac{TP}{TP+FP}$
FPR (False Positive Rate) $\frac{FP}{FP+TN}$

A last type of performance measure requires the predictions to be scored (probabilities for instance). By sliding the decision threshold on these scores from the lowest to highest values, some metrics are computed at each step and their distributions are analyzed. Two plots are often used to visualize the performance of a prediction algorithm.

ROC The Receiver Operating Characteristic curve plots all TPR and FPR values computed at each variations of the decision threshold
PR Similar computation for the Precision Recall curve

Both of these curves can be used to graphically represent the performances but the area under these curves is also a broadly used measure of performance. The AUC, the area under the ROC curve, relates to the wilcoxon signed and ranked test. The AUPR, area under the PR curve, is better suited in cases in which the class distribution is skewed (*i.e.* far more negative instances than positive).

Figure 1.3: *Box : classification performance measures.*

used to calculate the prediction error and estimate the performance of proposed algorithms. The measures of performance of binary classification are used here and briefly explained in the Box 1.3 *Performance measure*.

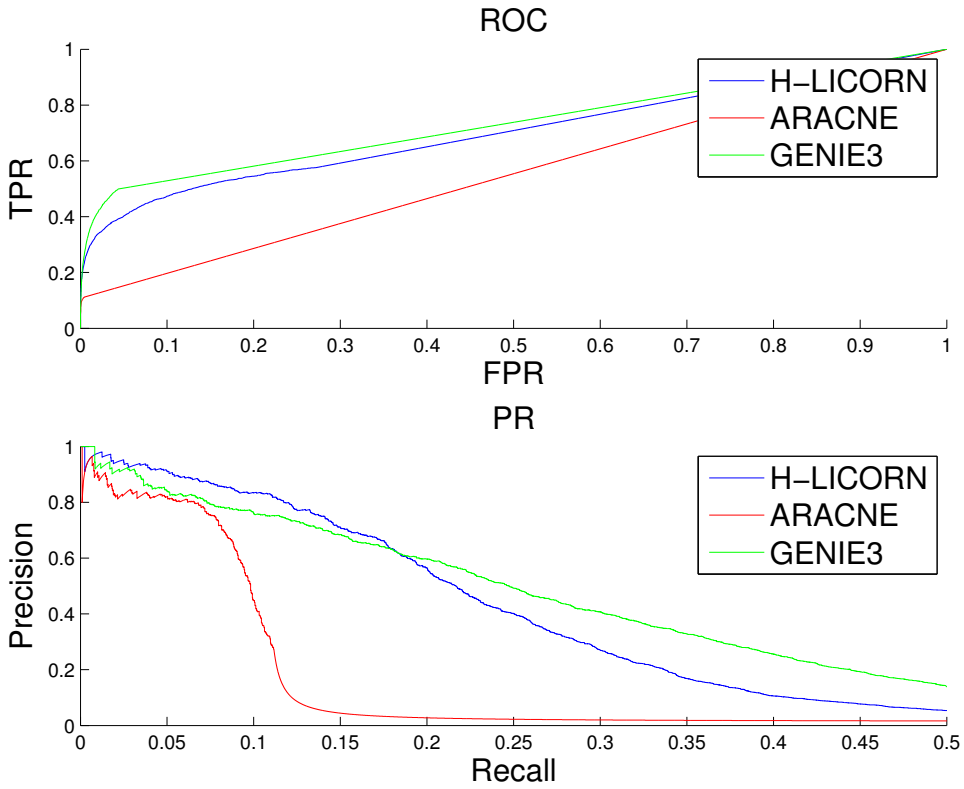


Figure 1.4: *Hybrid-LICORN performance. The ROC and PR curve are plotted using the prediction of the ARACNE (Margolin et al., 2006a), GENIE3 (Huynh-Thu et al., 2010) and the proposed* h-Licorn *(Chebil et al., 2014) algorithms on the DREAM5 in silico dataset (Marbach et al., 2012b). (from Chebil et al., 2014)*

The ROC and PR curves on this dataset are represented in figure 1.4. Along with h-Licorn, two algorithms were tested, the winner of the DREAM5 challenge GENIE3 (Huynh-Thu et al., 2010), which is based on the Random Forest machine learning method (Breiman, 2001) and ARACNE (Margolin et al., 2006a), which is based on Mutual Information and was applied successfully to resolve original biological and clinical problems (Aytes et al., 2014; Lefebvre et al., 2010).

In an *in silico* setting, the performances of h-Licorn are comparable to those of GENIE3, a state of the art network inference method (Marbach et al., 2012b). We also showed that h-Licorn is more robust to sub-sampling and therefore more efficient in experimental settings in which only a small number of samples are available (Chebil et al., 2014).

TFBS

n interactions	Random	GENIE3	ARACNE	h-LICORN
50	6.3	2	3	10 .
100	12.7	8	4	14
500	63.4	57	27	49
1000	126.8	109	81	107
5000	634.1	556	587	565
10000	1268.2	1138	1286	1200
50000	6341	6086	6534 **	6769 ***

ChIP

n interactions	Random	GENIE3	ARACNE	h-LICORN
50	3.1	6 .	1	9 **
100	6.2	11 *	1	17 ***
500	31	44 *	20	64 ***
1000	61.9	80 *	49	108 ***
5000	309.7	395 ***	343 *	417 ***
10000	619.3	761 ***	762 ***	819 ***
50000	3096.6	3380 ***	3482 ***	3787 ***

Figure 1.5: *ChIP and TFBS supported inferred regulatory interaction. Significance code (Fisher's test) :* $p < 0.1^{\bullet}$, $p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$ *and no symbol for* $p \geq 0.1$. *(from Chebil et al., 2014)*

Beyond pure *in silico* performances, the same algorithms were also tested on human transcriptomics and against real regulatory evidences. This comparison was carried out on a bladder cancer data set (Stransky et al., 2006) and the inferred regulatory networks were compared to the sets of genes bounded by TFs based on a TFBS promoter scanning and ChIP-on-chip datasets from the TRANSFAC database (Matys et al., 2006). As discussed in section IV.2, these datasets are mainly unreliable on their own (see their low overlap in figure IV.17 of the introduction section). This is mostly explained by the non-context-specificity of TFBS and more generally by the fact that TF binding does not systematically imply gene regulation. Therefore, the comparison of the algorithms is done by computing the enrichment of the inferred regulatory links in corresponding TFBS or ChIP data. The number of TFBS or ChIP defined interactions, hereafter referred to as regulatory evidences, found among the n best regulatory interactions of GENIE3 (Huynh-Thu et al., 2010), ARACNE (Margolin et al., 2006a) and of h-Licorn (Chebil et al., 2014) are reported in figure 1.5. h-Licorn is more efficient in retrieving regulatory interactions supported by predicted (TFBS) or experimentally observed TF binding. Interestingly, as the network gets larger, ARACNE performs well suggesting that the links with the highest Mutual

Information score are not necessarily more relevant yet the algorithm is suitable for higher eukaryotes. GENIE3 predicted regulatory interactions correspond to ChIP evidences yet are as enriched in TFBS than randomly picked interactions. However, our proposed algorithm h-Licorn is both enriched in TFBS and ChIP supported evidences. Overall, this work suggests that a hybrid method using both a discretized and continuous gene expression dataset is relevant in the context of regulatory network inference.

The CoRegNet package implements a bootstrapped version of h-Licorn. However, the goal is to select sets of co-activators and co-inhibitors to identify transcriptional programs, the linear model is used to select relevant GRN instead of TF-gene pairs. To do so, interaction are added to the linear model and the mean adjusted coefficient of determination over 100 bootstrap iteration is used to score a GRN. The linear model is the following:

$$\hat{g}_c = \beta + \sum_{j=1}^{q+p} \alpha_j * r_j + \alpha_a \prod_{k=1}^{q} a_k + \alpha_i \prod_{l=1}^{p} i_l$$

in which regulators $r_j \in A \cup I$, activators $a_k \in A$, inhibitors $i_l \in I$ and the product of expression of the set of co-activators and co-inhibitors is used to model the interaction between these co-regulators.

The adjusted coefficient of determination used to score each GRN and noted $\bar{R}^2$ is computed as follow:

$$\bar{R}^2 = 1 - \frac{VAR_{err}}{VAR_{tot}}$$
$$with\ VAR_{err} = \frac{\sum (g_c - \hat{g}_c)^2}{n - q - p - 1}\ and\ VAR_{tot} = \frac{\sum (g_c - \bar{g})^2}{n - 1}$$

Given all these described calculations, the network inference process proposed in the CoRegNet package runs as follow. The LICORN search algorithm is used on a discretized version of the transcriptomic data. In order to ensure the diversity of extracted GRN and based on the observation that for at least 99% of genes, the first 100 GRN contain all the best possible GRN (see figure 1.2), the 100 first GRN are selected using the Mean Absolute Error for each gene. The linear model is then fitted on a set of samples drawn randomly with replacements and the adjusted coefficient of determination ($\bar{R}^2$, see previous equation) is computed. This bootstrap procedure is repeated 100 times and the values of the $\bar{R}^2$ are averaged. This results in a robust evaluation in the original continuous gene expression data of all the GRN extracted by the original LICORN. This score is then be used to select the most relevant GRN per gene.

1.4 Regulatory influence

In order to identify the transcriptional programs that are active in a sample or a set of samples of interest, a new measure of Transcription Factor activity (TFA) is devised. This work was originally presented at the 11th International Conference on Machine Learning and Applications and the associated published article is in appendix B.

The proposed method aims at estimating the *influence* of a transcription factor on it's target genes. It simply measures the difference between the activated and repressed target genes of a TF. Figure 1.6 shows an example of the expression of the targets of a given TF in a given sample in which it is expected to be highly active.

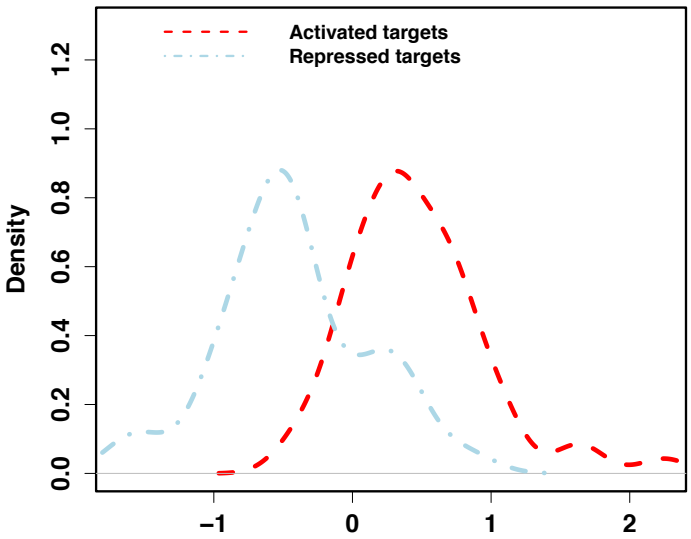


Figure 1.6: *Expression of FOXA1 target genes. Example of distribution of the centered (non scaled) expression of FOXA1 target genes in a bladder cancer sample. Targets were inferred from a bladder cancer data set (Stransky et al., 2006).*

Transcription Factor Activity is calculated for a single TF in a single sample and is based on Welch's t statistics, computed as follow:

$$\frac{\overline{X}_a - \overline{X}_i}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_i^2}{n_i}}}$$

with $\bar{X}$ the mean and s^2 the variance of expression of the target genes, n their number, and the subscripts $_a$ or $_i$ for the activated and inhibited targets respectively. This measure

is only computed for TF with a minimum number of both activated and repressed targets in the provided network (set to 10 by default in CoRegNet).

The computation of this measure for all TF of a provided large-scale regulatory network and in all samples of a transcriptomic dataset results in a new dataset with the same number of sample but a highly reduced number of features dropping from several thousands of gene expression measures to usually several hundreds of TF influence measures.

The first use of this new view over transcriptomic data was feature extraction. This relates to feature selection, that is, the selection of a set of genes with relevant expression patterns to discriminate two types of samples. Feature extraction only differs in the fact that it aims at selecting transformed versions of features, principal components from PCA or transcription factor activities for instance. The set of selected genes is usually termed Gene Expression Signature, in the case of feature selection. In terms of predictability, GES hold acceptable performance (Haury, Gestraud, and Vert, 2011). However, they usually perform as well as a random list of genes and GES designed to predict the same prognostic feature can be very different in terms of gene content (Fan et al., 2006). This instability questions the biological relevance of the selected features and challenges the field for more reliable models.

In the published paper, I showed that using the *influence* to classify muscle-invasive bladder cancers has similar prediction performances than using the original gene expression dataset. More importantly, the reproducibility of feature selection was much higher. Gene Expression signatures were constructed using the shrunken centroid method (Tibshirani et al., 2002). The overlap between the signatures built in two different datasets were compared using the Kuncheva stability measure (Kuncheva, 2007) which corrects for the number of selected features as follow:

$$stability = \frac{2\sum_{N_s}^{i=1}\sum_{N_s}^{j=i+1} F(f_i, f_j)}{N_s(N_s-1)}$$

The Kuncheva stability of the features selected in two bladder cancer datasets (Dyrskjøt et al., 2004; Stransky et al., 2006) are reported in figure 1.7.

Overall, the transformation of gene expression into a higher-level transcription factor influence is more representative of the cellular state. Therefore, using this network-transformation of transcriptomes, the analysis of two distinct datasets of the same type of samples, here bladder cancers, is more reliable.

Other methods were previously proposed to estimate transcription factor activity using a large scale regulatory network and a set of transcriptome profiles. These methods are discussed in the introductory section IV and are based on linear models. The CoRegNet package aims at inferring active TFs based on an inferred network, which often contains a large number of false regulatory interactions (Marbach et al., 2012b). Therefore, these algorithms must be able to estimate TFA using noisy networks. TFA was computed using: the *influence* method embedded in the CoRegNet package, the ROBNCA (Noor et al., 2013) algorithm (a newer and more robust version of the original Network Component

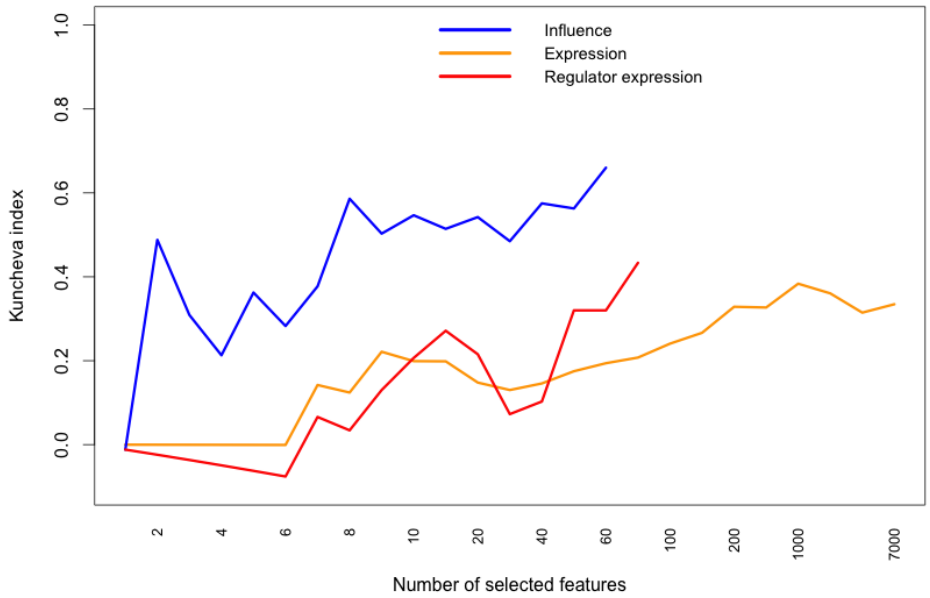
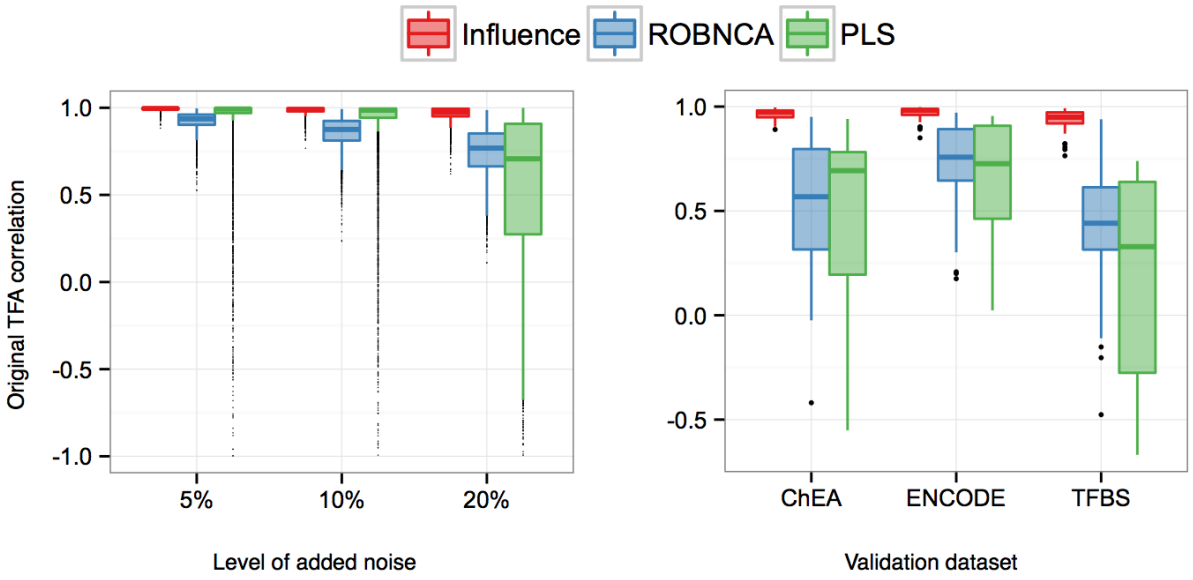


Figure 1.7: *Stability of expression and influence signatures. Plots of the Kuncheva stability measures for signatures built on the gene expression, the TF influence and the TF expression. The measure is computed for an increasing number of selected features. (adapted from Nicolle, Elati, and Radvanyi, 2012)*

Analysis) and a PLS (Partial Least Square) based method (Boulesteix and Strimmer, 2005). The TFA was computed on the CIT bladder cancer dataset (CIT : "Carte d'Identité des Tumeurs", a french initiative for tumor profiling) using a network inferred by CoRegNet on the same data and following the addition of noise in the network. This noise is added by permuting the targets of each TF in the network in order to preserve the topology. To compare the *influence* to the two other TFA methods in terms of robustness to noise, the correlation between the original TFA with the noisy TFA was computed and is reported in the top left panel of figure 1.8 in which the overall noise-resistance process (network permutation, TFA computation and correlation to original TFA) was repeated 10 times. To further compare these methods, the original TFA of all TF in the network for which TFBS or public ChIP-seq data is available was correlated to the TFA computed using only the targets genes with corresponding regulatory evidences. The distribution of the correlation between the original TFA and the validated TFA is reported in the top right panel of figure 1.8.

Finally, the three TFA methods were tested on a dataset in which the activation status of the *PPAR*γ TF is known. In essence, urothelial cells were cultivated with a PPARγ agonist (Roziglitazone) in combination with the PD153035 EGFR inhibitor to



Noise	Influence	ROBNCA	PLS
5%	0.99	0.92	0.93
10%	0.98	0.86	0.91
20%	0.97	0.75	0.54

Evidences	Influence	ROBNCA	PLS
ChEA	0.96	0.51	0.49
ENCODE	0.97	0.72	0.67
TFBS	0.93	0.42	0.19

Figure 1.8: *Comparison of Transcription factor activity measures. Top left panel: Pearson correlation of all TF activities in all samples with the activity computed on the same data using a network in which noise was added. Top right panel: Pearson correlation of all TF activities in all samples with the activity computed on the same data using a network containing only interactions found in the CHEA2 (Kou et al., 2013) and ENCODE (Gerstein et al., 2012) ChIP-seq data as well as on the interaction identified by scanning gene promoters with TFBS mostly from the HOCOMOCO (Kulakovskiy et al., 2012) and JASPAR (Portales-Casamar et al., 2009) databases. Bottom tables : average correlations between TFA computed using the original and modified network.*

prevent an EGFR-dependent phosphorylation and inhibition of *PPARγ* (Böck et al., 2014; Varley et al., 2008). The cells were sampled at various time after the activation of *PPARγ* resulting in a small time series (6 hours, 24 hours, 3 days and 6 days). In this experimental setting, *PPARγ* exhibits null-to-weak activation at confluence (starting

at day 3) in non-treated cells, a modest activation as soon as 6 hours and to reach full transcriptional activation at 24 hours and maintain this state in treated cells. Based on these transcriptomes, the activity of *PPARγ* was computed using the three tested methods, including the *influence* of the CoRegNet package. The result is shown in figure 1.9.

The *influence* measure is concordant with the expected state of *PPARγ* activity whereas the ROBNCA method does not detect *PPARγ* activation at 24 hours nor a small increase at confluence in non-treated cells. The PLS-based measure of TFA shows less difference between the two types of cultures (treated and non-treated), especially at day 6.

1.5 Transcriptional programs

Transcription factors have a major role in determining cellular phenotype and behavior. The joint activation of a set of specific master regulators can greatly alter the cell and reprogram it's function (Lee and Young, 2013). Interestingly, most of cell reprogramming experiments require a combination of TF to be simultaneously induced. For instance, the Estrogen Receptor α (*ESR1*) is a major driver of Estrogen dependent breast cancer cells. The driver function of *ESR1* was shown to necessitate the presence of two major co-regulators, *GATA3* and *FOXA1* (Kong et al., 2011). As shown in this study of the *ESR1* program in breast cancer and by the combined action of TF in cell reprogramming in general (reviewed in Lee and Young, 2013), regulator cooperation is key to the understanding and modeling of transcriptional programs. While transcriptional co-regulation is clearly modeled at the level of a few extensively studied promoters (Panne, 2008), cell-wide and context specific co-regulatory networks are still difficult to reconstruct. In an attempt to resolve combinatorial TF activity, the FANTOM consortium enumerated the direct physical interactions between the products of the cDNA encoding human TF using a Mammalian-2-Hybrid (similar to Yeast-2-Hybrid with mammalian cells) systematic screening (Ravasi et al., 2010). This study brought simple yet interesting insight into the link between TF interactions and cell fate. More importantly it provides a basis for general human TF co-regulation as cooperation is partly driven by TF protein interaction although these are described in an unspecific cellular context.

In order to identify context-specific transcriptional programs, the CoRegNet package uses the Hybrid LICORN algorithm to extract for each gene the sets of co-regulators previously termed local GRN. The idea is to directly extract the sets of regulators that are together necessary for the regulation of their target genes and thereby identifying global cooperative regulators in a set of transcriptomic profiles. All pairs of regulators that were found by h-Licorn to be co-activators or co-inhibitors of at least one gene are considered as potential co-regulators in the studied context. Then, only those pairs that have a significant overlap of target genes using Fisher's exact test are predicted as co-regulators (with a 1% FDR control).

In order to verify the proposed algorithm, the pairs of co-regulators predicted by the

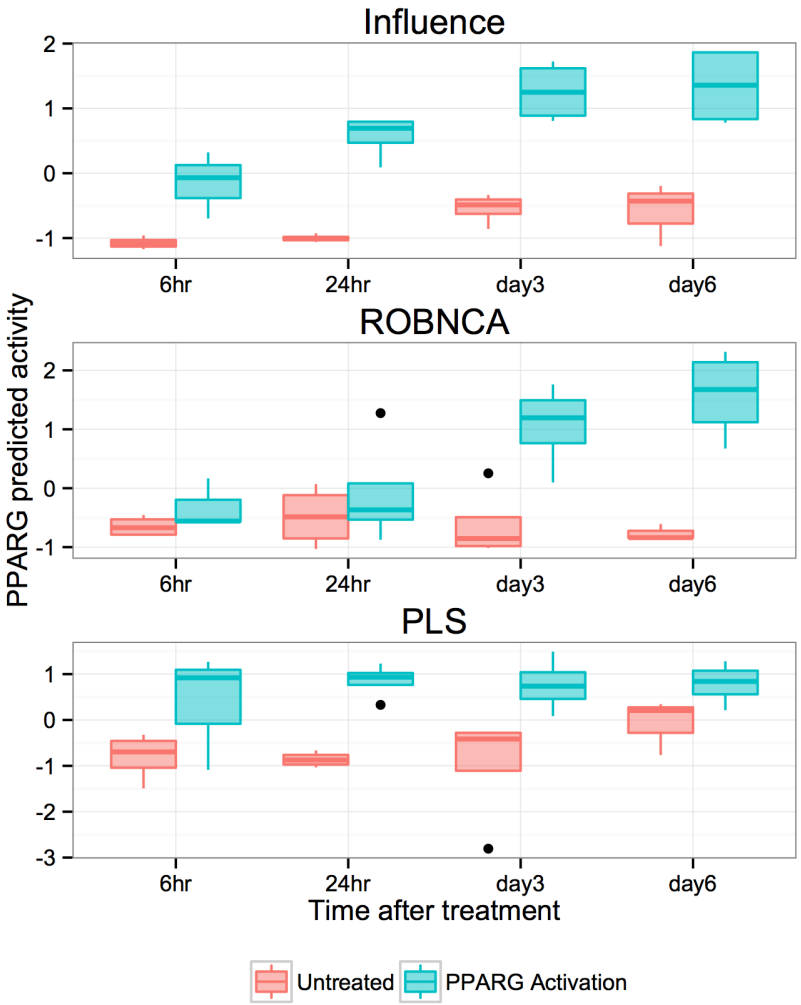


Figure 1.9: *PPARγ predicted activity. Transcription factor activity predicted at 4 time points in a set of sample with PPARγ treatment and in untreated samples. Each time points contain 3 to 4 replicates. In this experimental setting, PPARγ is thought to exhibit no activation without treatment in the first days with a minimal activation at confluence (from day 3). A modest activation is observed in as soon as 6 hours in treated cells and is thought to reach full transcriptional activation at 24 hours and maintain this state.*

CoRegNet package are compared to the pairs of TF with significant *regulon* overlap (intersection between sets of target gene) using the *RTN* package. This package implements the ARACNE algorithm (Margolin et al., 2006a) and was previously used to identify what is called Transcriptional Modules which aims at describing the sets of TF involved in the same transcriptional programs (Fletcher et al., 2013). Only significant pairs of TF identified by the *RTN* package were considered as co-regulators using the same statistical selection

(Fisher's test and multiple hypothesis testing correction). The enrichment of both of these inferred co-regulator networks in real protein-protein interaction is reported in table 1.10. These results show that the CoRegNet-inferred co-regulators better correspond to real protein interactions between TF.

Unfiltered regulatory networks

Algorithm	Dataset	FANTOM	HIPPIE	HPRD	STRING
CoRegNet	CIT	2.43	2.6	2.75	3.75
RTN		1.44	1.49	1.5	1.8
CoRegNet	TCGA	1.26 †	1.37	1.34	1.82
RTN		1.32	1.28	1.4	1.55

Refined regulatory networks

Algorithm	Dataset	FANTOM	HIPPIE	HPRD	STRING
CoRegNet	CIT	3.32	3.36	3.45	5.32
RTN		2.03	1.98	2.31	3.23
CoRegNet	TCGA	3.2 †	3.16	3.67 †	5.13
RTN		1.9	1.56	1.83	3.42

Figure 1.10: *Co-regulation enriched in protein interaction. Table of enrichment, computed as Odds Ratio, of Protein-Protein interactions found among inferred TF-TF cooperation. The co-regulators were identified by the* CoRegNet *package using either the network inference algorithm* h-Licorn *internal to the package or the ARACNE algorithm (Margolin et al., 2006a) implemented in the bioconductor package RTN (Fletcher et al., 2013). The top table contains the Odds Ratio for the entire regulatory network inferred with either algorithm. The bottom table contains filtered network using the* refine *function of the* CoRegNet *package (selecting the best GRN per gene based on the adjusted coefficient of determination) or using the Data Processing Inequality for the* RTN *package (Margolin et al., 2006a). The inferred pairs of co-regulators were compared to the pairs of TF with protein interactions referenced in four studies: the FANTOM screen for combinatorial TF (Ravasi et al., 2010), the HIPPIE (Schaefer et al., 2012), HPRD (Keshava Prasad et al., 2009) and STRING (Franceschini et al., 2012) protein interaction databases. All enrichment are significant (Fisher's exact test* $\alpha = 10\%$*) except the Odds Ratio marked with* †*.*

In order to further investigate the performance of both type of co-regulator predictions, inferred pairs were ordered by the number of shared targets for the *RTN* package and by the number of shared *GRN* for the CoRegNet package. This score was used to draw the Precision Recall curve (see box 1.3 listing classification performance analysis

measures) which is shown in figure 1.11. The precision and recall were computed using the protein interactions found in the STRING database (Franceschini et al., 2012) as ground truth. It is to be noted that STRING references experimentally identified protein interactions as well as predicted interactions based on several type of protein analysis such as phylogeny-based or literature mining. Therefore, STRING is considered as referencing highly relevant functional relationships, which have been previously used for biological predictions of operon for instance (Taboada, Verde, and Merino, 2010).

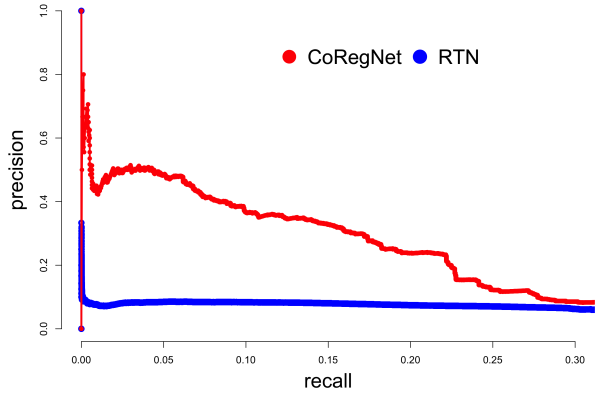


Figure 1.11: *Precision Recall curve of TF interaction predictions. The area under the PR curve is of approximately 0.14 and 0.5 for the co-regulators predicted using the regulatory inference method of the* CoRegNet *and RTN package respectively.*

The h-Licorn algorithm implemented in the CoRegNet package extracts sets of cooperative regulators instead of simple TF-gene pairs. By directly identifying functional regulatory sets of TF, the methods implemented in the CoRegNet package have overall higher performances in identifying relevant co-regulators to build a context-specific co-regulation network of functionally related TF.

1.6 Integration of regulatory evidence

In order to obtain a trustful regulatory network, additional data can be integrated to the fully inferred network. The data is in the form of regulatory (TF-gene) and co-regulatory interactions (TF-TF) and is used to support the predictions of h-Licorn. More specifically, these supporting regulatory evidences are used two ways, first to refine the predicted network and second to validate the measure of transcription factor influence by computing the activity of TF using only targets which are validated by external data. The validation of TF influence is quite simple. Basically, the influence is recomputed by using for each TF only the target genes that were also found in one of the regulatory evidence database (for instance, all genes with a binding site of the TF in their promoter).

The following section detail the origin of the regulation data as well as the way these can be used to refine an inferred network.

Origin of the regulatory evidence data

The additional regulatory evidences originated from several sources.

The ENCODE project aims at identifying and describing the regulatory DNA-elements of the Human genome. To do so, the consortium performs a large number of ChIP-seq experiments using TF-directed antibodies (Gerstein et al., 2012). The ENCODE ChIP-seq data was recovered from the UCSC genome browser (Human hg19 February 2009 genome assembly) by selecting all narrow ChIP-seq peak (ENCODE chip V3) within -5000 bp to 2000 bp around a Transcription Start Site of a gene with a non-null Human genome organization Gene Nomenclature Committee (HGNC, genenames.org) symbol.

In addition to this, the ChIP-seq and ChIP-on-chip from previously published studies are available through the ChEA2 database which aggregates and processes published ChIP data (Kou et al., 2013).

Transcription Factor Binding Sites (TFBS) models in the form of Position Weight Matrices (PWM) were recovered through the MotifDB R/Bioconductor package which references models from several studies (Jolma et al., 2013; Portales-Casamar et al., 2009; Xie et al., 2010). This was complemented by the HOCOMOCO database of human TFBS (Kulakovskiy et al., 2012). When several models were available for the same Transcription Factor (TF), the PWM with the highest Information Content (in bits) was kept. The promoter sequences (using the same coordinate that were used for the ENCODE ChIP-seq) were scanned for these sequences using the PWMEnrich R/Bioconductor package.

Protein-Protein Interactions (PPI) were downloaded from several databases such as: HIPPIE (Schaefer et al., 2012), STRING (Franceschini et al., 2012), HPRD (Keshava Prasad et al., 2009) as well as from the FANTOM study of TF physical interaction through Mammalian-2-Hybrid assays (Ravasi et al., 2010).

Regulatory network refinement

In order to refine large-scale regulatory networks using external regulatory interactions, the CoRegNet package implements functions introduced by the modENCODE consortium (Marbach et al., 2012a) and applies them to the selection of local GRN models.

In essence, the goal is to score each GRN (each interactions in the original method) using both the transcriptomic data and the integrated evidences to select the set of best GRN models. Each GRN is scored by each of the integrated dataset. The transcriptomic data used to infer the network scores GRN by the adjusted coefficient of determination provided by h-Licorn and noted $\bar{R}^2$. The regulatory evidences score each GRN by the proportion of predicted interactions found in the integrated data. The number of intersecting interactions in a given GRN is divided by the total number of predicted

interactions ($|A| + |I|$). For cooperative evidences (TF-TF) such as protein interactions, all possible pairs of activators ($\frac{|A|\times(|A|-1)}{2}$) are compared to the pairs of TF found in the data and similarly for inhibitors.

Following this, to each *GRN* is associated as many scores as their are integrated regulatory datasets all which range from O to 1, plus the network inference $\bar{R}^2$ score, which is defined on $[-\infty, 1]$ although most values are non-negative. The original study (Marbach et al., 2012a) proposes two approaches to merge the scores, an unsupervised and a supervised approach. While both are implemented in the CoRegNet package, the unsupervised approach is preferred as it was shown to have better performances in the original study. It is simply an unweighted average of each of the scores.

Finally, for each gene, the *GRN* with the maximum merged score is selected. Figure 1.12 shows the enrichment in five different regulatory evidence datasets (ChIP data were merged) of networks refined using various evidence datasets.

	Regulatory networks					
Enriched datasets	Raw	Refined	ChIP	TFBS	Regulation	All
TFBS	0.95 †	1.16	1.38	1.15	5.94	2.28
ChIP	1.12	1.03 †	9.68	11.69	8.07	2.56
Hippie	2.51	2.46	3.47	3.7	3.57	17.76
Fantom	2.43	3.53	2.82 †	4.07	2.85 †	20.22
STRING	3.75	5.36	5.23	6.10	5.15	57.9

Figure 1.12: *Enrichment of the refined network. Table of enrichment, computed as Odds Ratio, of known interactions in a predicted network. Each column corresponds to (in same order) the original network, refined using only the score of* h-Licorn *($\bar{R}^2$), using the ChIP and the TFBS data, the merge of both regulatory interactions (ChIP and TFBS) or the entire set of evidences listed in the first column (noted All). Each line corresponds to the dataset used to test the enrichment. All enrichments are significant (Fisher's exact test $\alpha = 5\%$) except the Odds Ratio marked with* †.

As expected, except for TFBS, the integration of a given dataset increases the chances of finding an interaction from the corresponding dataset. Interestingly, the integration of TFBS does not increase the chance of finding a TFBS among the predicted interaction. However, it increase the number of corresponding ChIP interactions. This is in part due to the fact that these datasets are only available for widely studied TF. Therefore, *GRN* containing these TF are more likely to have higher scores whereas less studied TF will less often be selected. Overall, while these results tend to indicate that the resulting network is more reliable, it introduces a bias towards well-studied TF.

1.7 Visualization of transcriptional programs

Transcriptomic measurements usually contain thousands of features, gene expression levels, for each sample. The high number of dimensions makes analysis and direct visualization of transciptomes difficult to perceive. One of the objective of the measure of transcriptional influence is to summarize changes in the expression of several genes by the activity of their common regulators. This abstraction of the expression of thousands of genes into the activity of only several hundreds of transcription factor can be used to visualize the entire dataset in a reduced and as expressive view of the samples. Since the regulation of sets of genes is usually the result of a coordinated action of several transcription factors, the co-regulation network built from the pairs of cooperative TF can be viewed as a blueprint of transcriptional programs composed of sets of co-regulators, which can be activated together to maintain a particular phenotype.

Based on this, an interactive visualization tool is embedded in the CoRegNet package to analyze several layers of information, including transcription factor influence, through the co-regulation network. Figure 1.13 shows snapshots of the visualization application to illustrate its use and capabilities.

The visualization tool of the CoRegNet package is based on a shiny application (shiny.rstudio.com) with a Cytoscape javascript widget. The widgets in the shiny webpage can trigger R functions thereby resulting in an interactive application.

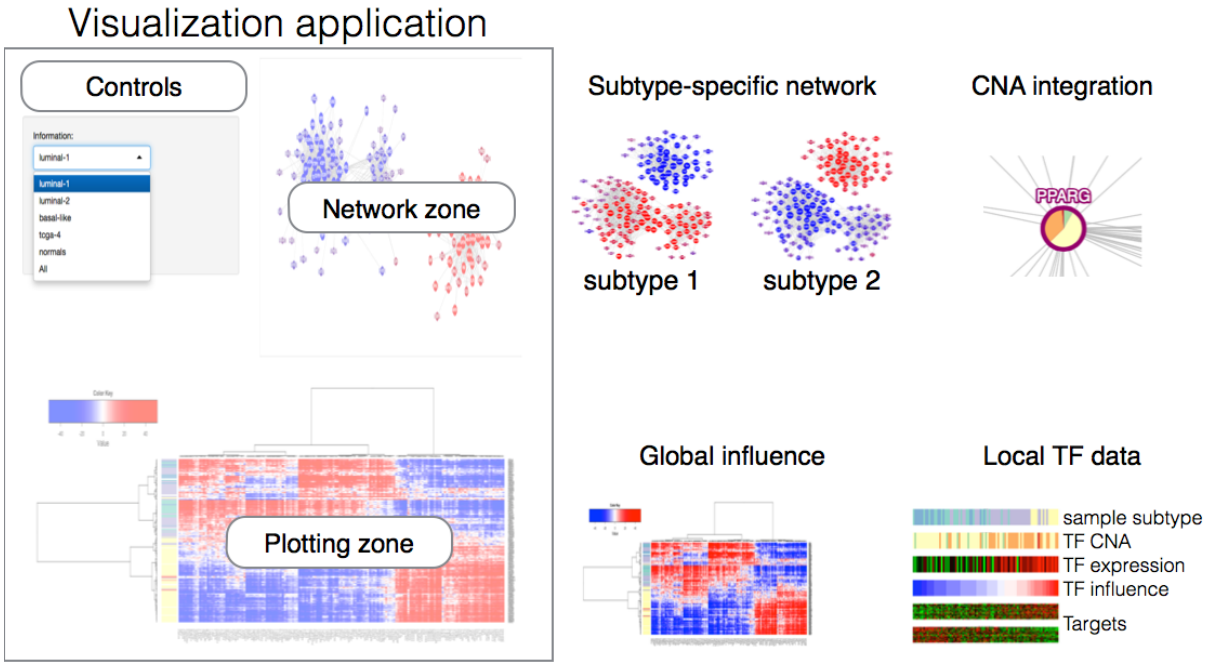


Figure 1.13: *Visualization application of the* CoRegNet *package. The main page of the application is divided in three parts. The control part can be used to select sample subtypes, modify the cooperative threshold or search for a TF in the network. The network part is an interactive widget in which node color reflects the activity of the TF in the selected sample(s). When Copy Number Aberration data is available, nodes are pie charts representing the proportion of the status of the TF (gain, loss, ...) in all samples or in the selected sample(s). The plotting part displays either a heatmap of TF influence with all TF or only the TF selected in the network, or it can display data related to the single TF selected in the network. The representation of the data related to a single TF is represented as a multi-layer heatmap with sample related features which can include sample classification, TF copy number status, TF expression, TF influence and the expression of the activated and repressed genes.*

1.8 Discussion

Being among the most effective pharmacological treatments of cancer, targeted therapy aims at blocking the activity of growth driving molecules and pathways. The success of such clinical approaches mainly relies on our capability to assess the influence of the targeted pathways on the growth of the tumor. Given that regulation of gene expression is the final effect of signaling pathway activation and that transcriptome measures are yet more effective than proteomics, mRNA signature is a compelling tool to predict response to therapy (Bild et al., 2006).

As a first step towards the prediction of signaling pathway activity, this chapter presents a novel approach termed CoRegNet to model the transcriptional programs driving a particular cellular behavior. The idea behind this package is to take advantage of the advances in network inference methods to reconstruct large-scale context-specific gene regulatory models and use these to highlight the combination of active transcription factors. CoRegNet proposes a new inference algorithm, hybrid LICORN, which is particularly suited to identify sets of cooperative regulators thereby forming building blocks to the identification of transcriptional programs. In order to identify the most active regulators in a particular sample, a novel method for Transcription Factor Activity prediction is embedded in the package. This measure of TF *influence* is particularly suited to cope with a recurrent problem found in inferred networks, the lack of reliability of single regulatory interactions.

Several large consortium, such as ENCODE and FANTOM, were formed worldwide with the intent to experimentally identify the entire regulatory functions of the human genome and the function of all regulatory molecules, including non-coding RNA, chromatin modifiers and transcription factors. While these projects are far from achieving their goal of modeling the entire human cell regulatory systems, they provide large amounts of data. The main difficulty of using such knowledge originates from the lack of consistency between the information obtained in different though similar cellular systems, thereby highlighting the importance of context-dependency. Indeed, the regulatory information uncovered by these large-scale experiments are specific to the cellular systems used by the consortium and therefore do not necessarily apply in other circumstances.

The CoRegNet package implements methods to integrate the large amount of regulatory knowledge produced by such projects. The idea is that the integration of highly context-specific data, the transcriptome, and objective-specific data, regulatory-related knowledge will produce more realistic and effective models of the regulation at work in the studied samples.

Finally, the complexity of the models proposed by the CoRegNet package or by any other network-related method necessitates tools to analyze them. In this work, the reduction of transcriptomic variations to the activity of upstream regulators as well as

the construction of a cooperative TF network allowed to propose a visualization tool embedding all the produced information.

Overall, the CoRegNet package is meant to be the basis of a set of tools to model and analyze active transcriptional programs and signaling pathways in a sample-specific manner. Further improvements are now in development among which: a. the inference algorithm, using more efficient models of the cooperativity of transcriptional factors based on novel constraints of the LASSO algorithm b. the network analysis methods, using factor graphs to consider simultaneously the influence of all transcription factors on all target genes and c. the visualization tool, using more advanced network visualization tools and methods to quickly identify and visualize the data corresponding to a set of co-regulators. Moreover, while the methods are here focused on the transcriptional regulation component of pathways, an algorithm is proposed in chapter 4 to identify proteins taking part in the signal transduction part of pathways and chapter 5 presents a complete analysis linking both signaling pathway and transcriptional programs.

Chapter 2

Transcriptional Programs driving bladder cancer

2.1 Introduction

Several studies recently proposed a molecular classification of bladder cancer with the consensus discovery of two main subtypes (Cancer Genome Atlas Network, 2014; Choi et al., 2014; Damrauer et al., 2014; Rebouissou et al., 2014; Sjodahl et al., 2012). The luminal-like bladder cancer subtype shows high level of terminal differentiation markers such as uroplakins and includes most *FGFR3*-mutated tumors. The other reproducibly identified subtype is designated as basal-like. It expresses basal markers, includes tumors with squamous histology and is likely to be sensitive to *EGFR*-targeted therapies (Rebouissou et al., 2014).

In this work, I propose a network analysis and a model of the transcriptional programs driving bladder cancer subtypes. Using the CoRegNet package, a bladder cancer regulation network is constructed from the CIT dataset (Carte d'Identité des Tumeurs, from the French National Cancer League) containing 179 primary bladder tumor samples and 4 normal samples (Rebouissou et al., 2014). The active transcriptional programs of each of these samples and of the available urothelial cell lines are extracted using a method to estimate TF activities and a network of cooperative regulators. The muscle-invasive tumors of the same cohort were used to distinguish transcriptional programs driven by genomic alterations and identified $PPAR\gamma$ as a driver of the luminal-like subtype. Functional validation and knockdown transcriptomic profiles determined an aberrant $PPAR\gamma$-driven activation of the highly energetic beta-oxidation pathway.

2.2 Bladder cancer subtype specific transcription factor influence

The TCGA consortium recently released a classification of muscle-invasive bladder cancer describing four sub-types (Cancer Genome Atlas Network, 2014). Two tumor subtypes were particularly discussed and well described: a highly differentiated class of tumors with frequent papillary histology termed luminal-like and a basal-like subtype of poorly differentiated tumors enriched in characteristic squamous tumors. This classification was applied to the CIT dataset and the most active transcription factors of each of these subtypes were retrieved. The top 5 most active regulator of each of the subtypes are listed in figure 2.1 and a snapshot of the bladder cancer co-regulation network specific to each of the corresponding sub type is showed.

An interesting use of the *influence* measure from the CoRegNet package is to be able to compute the activity of the TF from a given network in related samples. In the case of the analysis of Bladder cancer, this can be used to identify cell lines driven by similar transcriptional program. For instance, figure 2.1 shows two cell lines SD48 and 1207 resembling the luminal-like and basal-like phenotypes. Interestingly, *PPARγ* is the most active TF in the SD48 cell line suggesting it to be an interesting model to study the role of *PPARγ* in cancer cell survival.

In order to assess the reliability of these predicted master regulators of bladder cancer subtype, the same analysis was performed on an independent bladder cancer dataset. The RNA-seq gene expression data from the TCGA consortium was used to infer a bladder cancer large-scale regulatory network. Using this new network, the influence of all TF was computed in the TCGA samples and for each of the four types of tumors the activity of each TF was averaged. Using only TF found in both TCGA and CIT networks, each of the mean subtype *influence* of the CIT network significantly ($\alpha = 10^3$) correlated (using Pearson's correlation) the activity computed in the TCGA dataset (Luminal-like: 0.76, TCGA II: 0.38, TCGA IV: 0.81, Basal-like: 0.82).

The basal-like bladder cancer subtype was recently shown to be associated with frequent *EGFR* gains and amplifications and to be sensitive to anti *EGFR* therapies. *EGR2* was found to be the most active TF in the basal-like subtype (figure 2.1) and is also known to be a downstream effector of EGFR (Chandra et al., 2013). Similarly, the *SOX9* transcription factor is particularly active in basal-like tumors (Student's test, $\alpha = 10^{-6}$) and was determined to mediate *EGFR*-dependent cellular migration in urothelial cancer (Ling et al., 2011). Furthermore, twist and snail, two major regulators of the epithelial-mesenchymal (Thiery et al., 2009) were also predicted to be highly active in basal-like bladder tumors.

The luminal-like bladder cancer subtype was suggested by the TCGA consortium to highly express urothelial terminal differentiation and luminal breast cancer markers. *PPARγ* is predicted to be one of the most highly active TF of the luminal-like subtype and was described as an initiator of adipocyte and more recently of urothelial differentiation

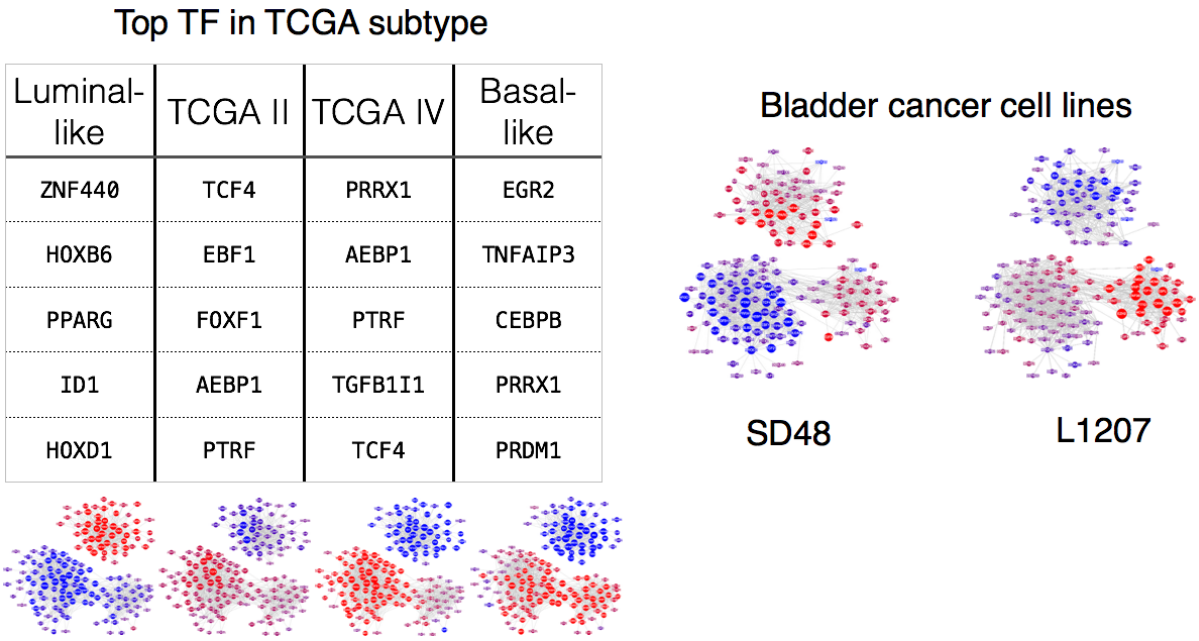


Figure 2.1: *Bladder cancer subtype and cell lines co-regulation network. The table presents the 5 regulators with the highest mean influence in samples of a specific bladder cancer subtype defined by the TCGA. Below each column of each subtype is shown a snapshot of the entire network of co-regulators in which an edge is set between two regulators if these were predicted to be significant cooperative regulators. Each node, regulator, is colored to represent its influence in the given subtype (blue: low activity, red: high activity). The same was done for two bladder cancer cell lines (right panel) to illustrate the ability of the influence method to characterize each and every sample of a dataset, as opposed to other network methods, which usually compare two subtypes of samples. The SD48 cell line shows to use similar transcriptional programs than the luminal-like subtype. The L1207 cell line uses similar transcriptional programs than the basal-like subtype.*

through *FOXA1* activation (Varley et al., 2008). Furthermore, the two key factors of luminal breast cancer, *GATA3* and *FOXA1*, are among the most significant co-regulators of *PPARγ* in the co-regulation network inferred from the CIT dataset (see 2.2).

In the TCGA network, *EGR2*, *SOX9*, *TWIST1* and Snail were also found to be master regulators of the basal-like subtype by being significantly more active in the basal-like subtype than in other samples ($p < 10^{-5}$) as well as *PPARγ* , *GATA3* and *FOXA1* in the luminal-like samples, which were also predicted to be significant cooperative regulators in the TCGA-inferred co-regulation network.

Best co-regulators

PPARγ		*FOXA1*		*GATA3*	
co-regulator	# GRN	co-regulator	# GRN	co-regulator	# GRN
GATA3	143	*PPARG*	53	*PPARG*	143
MSX2	107	*FOXQ1*	47	*MSX2*	88
FOXQ1	106	*GATA3*	39	*FOXQ1*	85
TBX3	70	*TBX3*	31	*TBX3*	50
ZNF626	66	*ID4*	26	*ZNF440*	49
ZNF440	65	*ZNF626*	24	*ZNF626*	45
FOXA1	53	*MSX2*	24	*HOXB6*	42
HOXB6	49	*ZNF440*	21	*FOXA1*	39
ID1	48	*ID1*	19	*TBX2*	35
ID3	43	*SPOCD1*	18	*ID3*	31

Figure 2.2: *PPARγ, FOXA1 and GATA3 best co-regulators. Listing the 10 best co-regulators of PPARγ, FOXA1 and GATA3 as predicted in the CIT network. # GRN is the number local gene regulatory network in which the pair of co-regulator was found. All the shown co-regulators have significant targets overlap (Fisher's exact test).*

2.3 Bladder cancer driver transcriptional programs

Transcription factors are key components of cellular states and were shown to be able to alter cell phenotypes by simply inducing their expression (Lee and Young, 2013). In the case of carcinogenesis, the maintenance of an aberrant phenotype necessarily requires the sustained activation of a set of regulators. In order to maintain a proliferative state it is common that a signaling pathway, the growth factor receptor tyrosine kinase and MAPK pathway for instance, is constitutively activated by genetic alterations. Given that this will result in the activation of a set of responding transcription factors, it is also possible to imagine that these alterations may directly target transcription factors themselves and in a similar manner lead to tumorigenesis.

In order to identify driver transcriptional programs of bladder cancer I investigated whether active bladder cancer regulatory programs were driven by genomic alterations. Copy number data was obtained from CGH BAC arrays profiling regional losses and gains in the genome of 86 tumors for which the transcirptomic profiles and subsequently the TF *influence* are available. Figure 2.3 plots both the subtype specific activity and the concordance between a high *influence* and gain or amplification of the locus containing the gene coding for the tested TF. Only representative TF are noted on the plot: *PPARγ* with both high CNA-*influence* concordance and high influence in the luminal-like subtype, *FOXM1* with high influence in the basal-like and moderate CNA-*influence* concordance and finally *TOP2B* and *POGK* with high CNA-*influence* concordance yet moderate influence in luminal-like tumors.

Several regulators showed significantly higher transcriptional activity in samples with

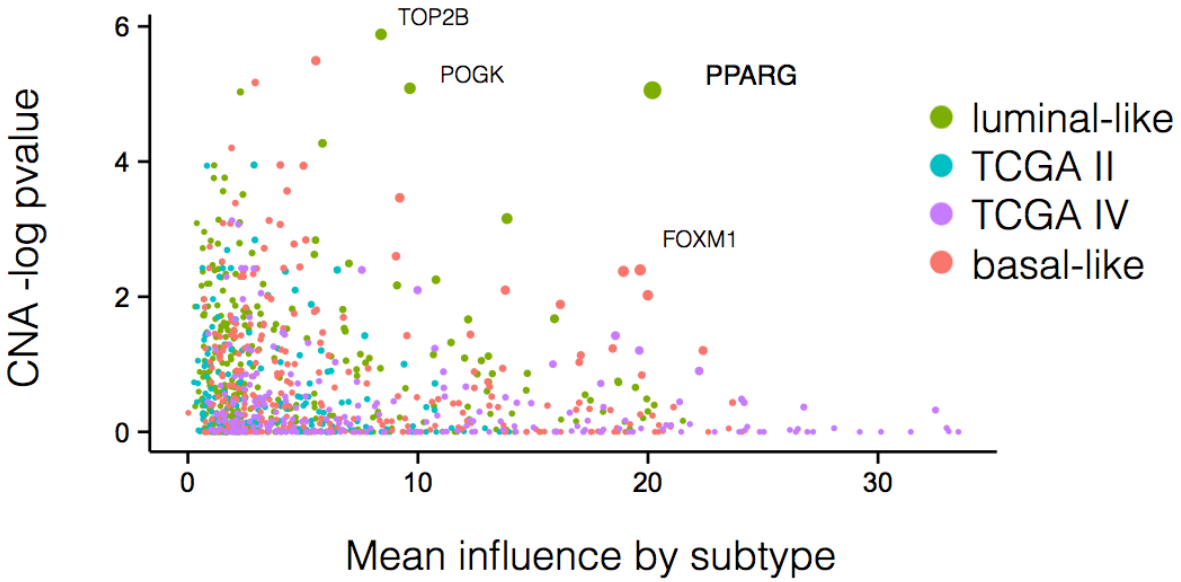


Figure 2.3: *Bladder cancer subtype specific driver TF. For each TF of the network, the effect of Copy Number Aberration (CNA) on the influence was tested by a one tail Student's test of an increase of the influence in samples presenting a gain or amplification (3 or more copies of the TF gene locus). The y-axis reports the negative log transformed p-value for which high values represents a high correlation between CNA and influence. The x-axis plots the mean influence of each TF in a given subtype. Only representative TF names are given.*

gains of copy number of their coding gene as shown in figure 2.4. However, *PPARγ* was both among the TF with highest activity and the highest concordance between high transcriptional activity and abnormal high copy number suggesting *PPARγ* as a driver of the bladder cancer luminal-like phenotype. *PPARγ* copy number status was significantly correlated with both its influence and it's expression as shown in figure 2.4. This result was confirmed in the TCGA dataset in which *PPARγ* is the regulator with the most significantly correlated influence and copy number (kruskal-walis test $p - value = 10^{-6}$).

A recent study identified *PPARγ* as a master regulator of the luminal-like subtype of bladder cancer(Choi et al., 2014). The high correlation between the genetic alterations of *PPARγ* and it's transcriptional influence further suggest *PPARγ* to be a driver of this subtype. Moreover, a recently published study from our group (Biton et al., 2014, *in press*) demonstrated by specific siRNA knockout assay the necessity of *PPARγ* in the survival and proliferation of certain bladder cancer cell lines. In particular, the two cell lines predicted to have the highest *PPARγ* influence, SD48 and UMUC9, are the most sensitive to *PPARγ* siRNA knockdown with a 50% to 60% decrease in cell viability. While these results confirm the driver role of *PPARγ*, the relation between the predicted transcriptional activity and the effect of the KD on cell line survival and proliferation was tested and

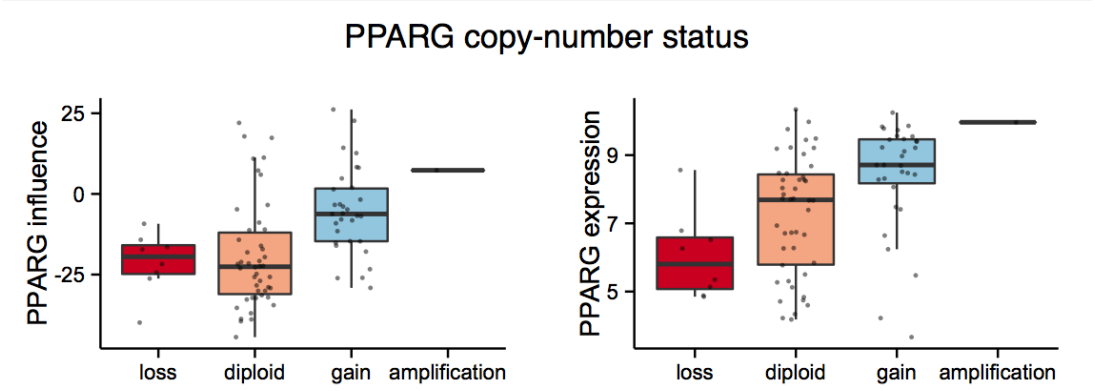


Figure 2.4: *PPARγ Copy Number, influence and expression in bladder cancer. Distributions of the influence (left) and expression (right) of PPARγ relatively to its copy number aberration (CNA) status. The levels of influence and expression were statistically different as measured by a kruskal-walis test ($\alpha = 10^{-3}$)*

shown in figure 2.5. The influence of *PPARγ* significantly correlates ($\alpha = 0.01$) with the effect of the KD suggesting that the proposed *influence* measure allow both to identify driver regulator as well as samples in which these are active and therefore targetable.

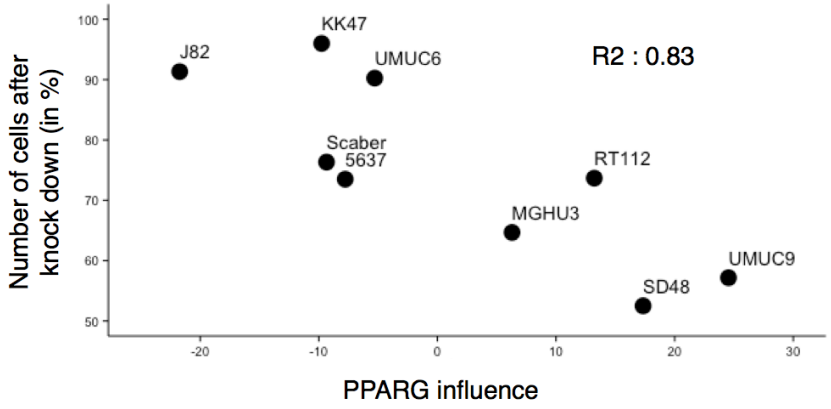


Figure 2.5: *Relation between PPARγ influence and its effect on cell survival. The mean effect of PPARγ siRNA knockdown on the survival of 9 bladder cancer cell lines (y axis) was plotted against its predicted transcriptional activity. Pearson's correlation is shown in the top right corner ($R^2 = 0.83$).*

No other TF showed such high activity and copy number abnormality in any other bladder cancer subtype. However, FOXM1, a known oncogene described in several studies as a master regulator(Lefebvre et al., 2010; Raychaudhuri and Park, 2011), is highly active in the basal-like subtype and presents a slight gain of activity concomitant to abnormal

gains of copy of the coding gene. *FOXM1* is predicted to be the *second* most active TF in the Scaber bladder cancer cell line. To assess the activity and potential oncogenic role of FOXM1 in bladder cancer, I performed an siRNA KD of *FOXM1* in the Scaber cell line. Using two different siRNA, this resulted in a 40% to 50% decrease in cell viability as shown in figure 2.6.

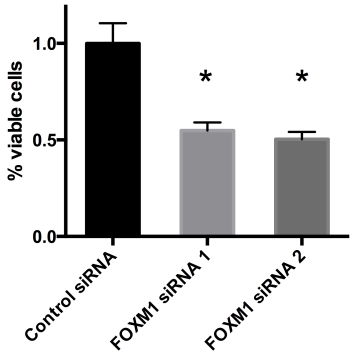


Figure 2.6: *FOXM1 knockdown in the Scaber bladder cancer cell line. Quantification of viable cells following the transfection of a control or FOXM1 targeting siRNA was determined by colorimetric MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] assays. The data shown are numbers of viable cells relative to the control siRNA. Experiments were conducted twice, each in triplicates. * : Student's test* $\alpha = 1\%$.

2.4 Characterization of *PPARγ*-driven carcinogenesis

Previous studies described *PPARγ* as a key regulator of urothelial cell differentiation (Varley et al., 2008; Varley and Southgate, 2008). This is somehow consistent with the strong expression of terminal differentiation markers in luminal-like bladder cancers. Despite these descriptions, no hypothesis was expressed about the oncogenic role of *PPARγ* , which remains unexpected from a master regulator of terminal cellular differentiation.

In order to determine the role of *PPARγ* in luminal-like bladder cancer progression, the siRNA knockdown performed in the SD48 cell line was followed by transcriptomic profiling. This identified 1026 differentially expressed genes due to *PPARγ* silencing (limma, FDR10%). The set of targets of *PPARγ* predicted by the CoRegNet package was 1.8 fold enriched in this set of *PPARγ* responsive genes (fisher's exact test : 2.10^{-5}). As expected the PPAR signaling pathway was found to be one of the most significantly enriched in genes activated by *PPARγ* (KEGG pathway: $\alpha = 1\%$). Interestingly, the lipid metabolism, regulated by *PPARγ* in several other tissues (mainly liver and adipocytes) was also highly disrupted after the KD (KEGG pathway: $\alpha = 1\%$). Several enzymes of

the lipid degradation/beta-oxidation pathway shown figure 2.7 were found to be targets of $PPAR\gamma$. Several of these enzymes were predicted to be activated by $PPAR\gamma$ in both the CIT and TCGA datasets and where confirmed by external ChIP-seq from the encode project. Finally, the UMUC9 cell lines treated with a $PPAR\gamma$ agonist (Roziglitazone) (Choi et al., 2014) also showed significantly higher expression of the lipid metabolism pathway (KEGG pathway:$\alpha = 1\%$) and confirmed the activation of enzymes involved in Beta-oxidation (see figure 2.7).

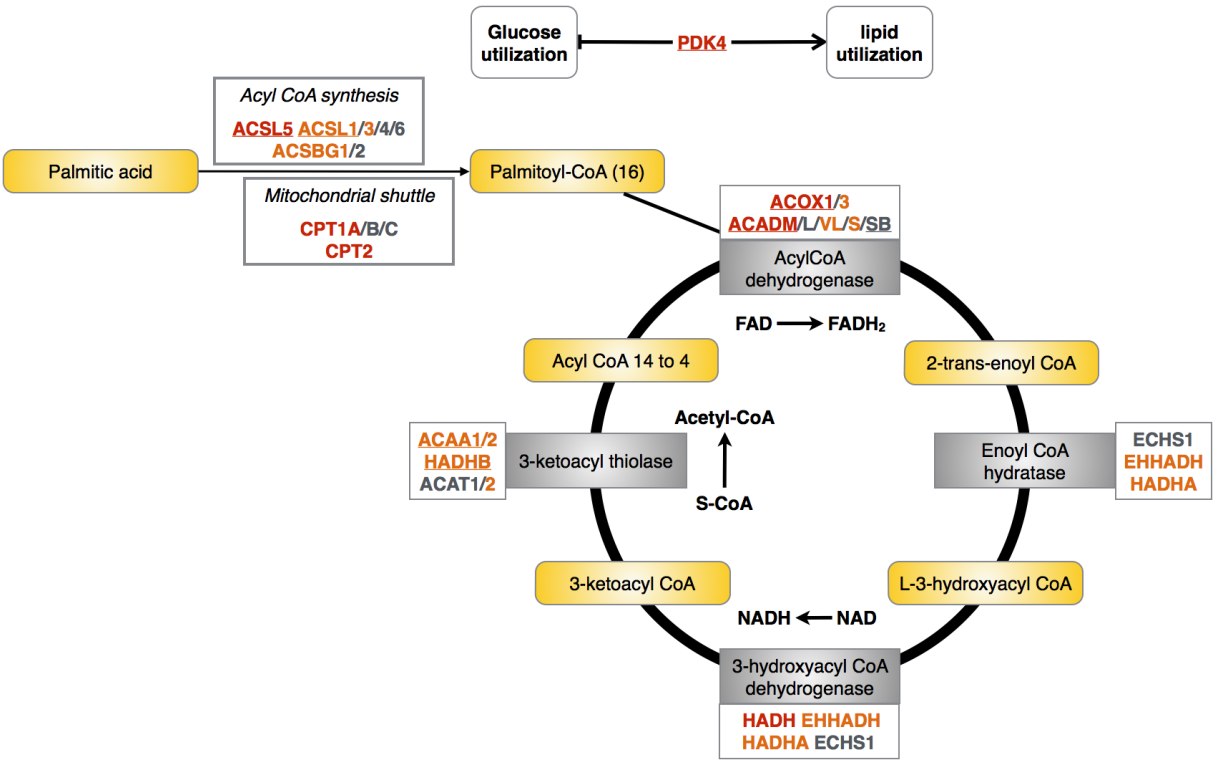


Figure 2.7: *$PPAR\gamma$ regulated lipid metabolism. Metabolic network of the lipid degradation metabolism from the KEGG database. Genes encoding for the enzymes are shown for each step of the degradation pathway. Some steps are encoded by several enzymes, which in some cases are specific to the number of carbons. Genes in red were found to be under-expressed following the depletion of $PPAR\gamma$ in the SD48 cell line and to be over-expressed following the activation of $PPAR\gamma$ by the roziglitazone in the UMUC9 cell line. Genes in orange were only found over expressed in the roziglitazone-treated UMUC9. Underlined genes were found to be bound by Genes for which $PPAR\gamma$ ChIP-seq or Chip-on-chip experiments identified a binding site in their promoter are underlined. The first metabolic step of lipid degradation is Palmitic acid (16 carbons), which is activated with an Acyl-CoA and thereby is transported inside the mitochondrion. Palmitoyl-CoA then enters the β-Oxidation cycle, in which at each end of cycle the fatty acyl-CoA is shorten by two carbons and produces NADH, FADH2 and Acetyl-CoA each of which can be latter use to produce energy either directly or through the citric acid cycle.*

These transcriptomic experiments suggest that *PPARγ* fuels the elevated energy requirement of fast growing cancer cells by activating the enzymes of the high energy yield beta-oxidation pathway. Although several enzymes of the lipid degradation pathway also have a function in the reverse lipid biosynthesis pathway, most of the *PPARγ* activated enzymes specifically direct the metabolic flux towards the production of acetyl-CoA and the citrate cycle such as ACOX1 and the limiting carnitine palmitoyltransferases (CPT1 and CPT2). Furthermore, PDK4 favors lipid over glucose as an energy source (Zhang et al., 2014b) and is shown by both the transcriptomic profiles and the inferred network to be activated by *PPARγ* (under-expressed after KD and over-expressed after agonist treatment).

2.5 Discussion

This work constitutes the first network analysis of bladder cancer. The application of the CoRegNet package to a bladder cancer transcriptomic dataset shows its ability to identify sample- and subtype-specific transcriptional programs. More importantly, the proposed measure of transcription factor *influence* can be used to identify the most relevant cell lines to perform validation experiments.

The aim of the proposed approaches is to eventually identify active signaling pathways and therefore effective targeted therapy for each of the analyzed samples. Evidently, the available data, mostly of the genome and transcriptomes of tumors, only allows, for now, to identify active transcriptional programs. Therefore, to assess the validity of the concept of inferring large-scale and potentially error-prone networks in order to identify active pathways, I first focused on genetic events that directly impacted the transcriptional regulation level. This identified several transcription factors for which the activity was predicted to be high and to be potentially caused by gains in copy number. Among these, I was particularly interested in two TF, *FOXM1*, which was previously described as a master regulator of proliferation, and *PPARγ*, which had clear gains of copy number associated with a remarkably high transcriptional activity. The inhibition of *FOXM1* resulted in a significant decrease of cell proliferation in a cell line in which it was predicted to be highly active. While more experimental validations are to be performed by the team, this illustrates the exploratory capacity of CoRegNet and the possibility to easily design experimental validations.

There is emerging evidence that *PPARγ* is an oncogene in bladder cancer. However, its role remains unclear, especially as it is mostly described as a major activator of the differentiation of the normal urothelium. While the discovery of *PPARγ* as a driver of luminal-like bladder cancer is very recent (Biton et al., 2014, *in press*), the relevance of its predicted activity correlating with the impact of *PPARγ* knockdown on cell line proliferation supports the algorithm itself. Moreover, the prediction of the target genes of *PPARγ* in bladder cancer samples along with transcriptomic profiles following knockdown or agonist treatment provided a hypothesis explaining the role of *PPARγ* in the carcinogenesis of the bladder.

Altogether, the network analysis of bladder cancer demonstrates the benefit of using genome scale models to rationally identify active and potentially targetable effectors of the oncogenic pathways. Ongoing work by the team includes testing of *PPARγ* antagonists and the effect of using drugs targeting known co-regulators.

Chapter 3

Deregulation of normal Transcriptional Programs in bladder cancer

The project detailed in this chapter implicated several collaborators. All NHU-related transcriptomic data and experimental results on NHU cell cultures were obtained from Jennifer Southgate and her team at the university of York (york.ac.uk). Sequencing and knockdown of *ELF3* in bladder cancer cell lines was performed by Clémentine Krucker (Oncologie Moléculaire team, Institut Curie).

3.1 Introduction

The epithelium acts as a protective layer of the underlying tissues. In response to physical damage, a complex process of wound healing is initiated to both replace missing cells and restore the function of the epithelium. Therefore, wound healing involves a complex balance between regenerative repair involving cell proliferation and migration, and restitution of tissue function by specializing cells through a specific differentiation process. This proliferation/differentiation balance is tightly controlled in time and space in order to maintain tissue function and integrity.

Harold Dvorak described the high similarities between the tumor stroma and the surrounding of an injured tissue during wound healing. Dvorak resumed this parallel by appropriately describing tumors as "wounds that do not heal" (Dvorak, 1986). This statement along with several more recent studies (Riss et al., 2006; Schäfer and Werner, 2008; Velnar, Bailey, and Smrkolj, 2009) underlines these similarities and suggests that tumor cells are somehow persisting into a regenerative phenotype. At the molecular level, the aberrant maintenance of an infinite proliferation state implies not only constitutive activation of mitogenic-associated signaling but also loss of negative feedback controls. The identification of the disrupted molecular processes of normal wound healing is critical to the understanding of the driver events and pathways of tumorigenesis.

The normal counterpart of bladder cancer is the urothelium, a highly specialized epithelium lining the urinary bladder. It functions as a self-repairing barrier to urine. It responds to injury by rapidly switching from a mitotically quiescent state (the urothelium being one of the most quiescent tissues in the body) to a highly proliferative state. This switch is driven partly by activation of the *EGFR* pathway (Daher et al., 2003; Varley et al., 2005). Our partners at the university of York (york.ac.uk) have developed robust methods to grow Normal Human Urothelial (NHU) cells as finite non-immortalized cell lines *in vitro* (Southgate, Masters, and Trejdosiewicz, 2002).

In order to identify the transcriptional programs of normal urothelial cell proliferation and differentiation as well as their disruptions in bladder cancer, this study greatly relies on transcriptomic profiles of NHU cell cultures (Fleming et al., 2012; Varley et al., 2008). NHU cells were extracted from the ureter or the bladder of patients admitted for non-cancerous disease. Cells were grown either in normal medium or in the presence of differentiation agents. Urothelial differentiation can be partly induced by the activation of $PPAR\gamma$, which itself is repressed by *EGFR* (Varley and Southgate, 2008; Varley et al., 2004, 2006). Therefore, differentiation was induced either by a combination of the $PPAR\gamma$ activator troglitazone and an *EGFR* inhibitor (PD153035, treatment noted TZ/PD) or by Adult Bovine Serum (ABS). These cultures were sampled at four different time points after seeding and profiled using affymetrix microarrays. Time points were chosen to reflect specific proliferation and differentiation phases as illustrated by figure 3.1. Two samples were taken the first day, 6 hours and 24 hours following seeding, times at which the first downstream regulators of differentiation (Varley et al., 2008) are expressed and proliferation is highest. Two additional samples were taken at day 3 and 6 at which cells reach confluence, become quiescent and reach terminal differentiation in the presence of TZ/PD or ABS.

3.2 Reconstruction of the normal urothelial cell proliferation and differentiation regulatory network

To identify transcriptional programs of normal growth and differentiation, the network inference algorithm of the CoRegNet package was applied to the NHU gene expression data set. The regulatory network inferred by CoRegNet, composed of the predicted co-regulators of 7,179 genes and was then refined by integrating additional regulatory related data. Regulatory interactions ($TF \rightarrow gene$) were supported by ChIP-seq and ChIP-on-chip data from the ENCODE project (Gerstein et al., 2012) and the ChEA2 database (Kou et al., 2013). Systematic promoter sequence analysis using the PWMEnrich R/Bioconductor package was carried out using known transcription factor binding sites (TFBS) models from the MotifDB R/Bioconductor package referencing data from several studies (Jolma et al., 2013; Portales-Casamar et al., 2009; Xie et al., 2010) and complemented by the HOCOMOCO database of human TFBS (Kulakovskiy et al., 2012). When several

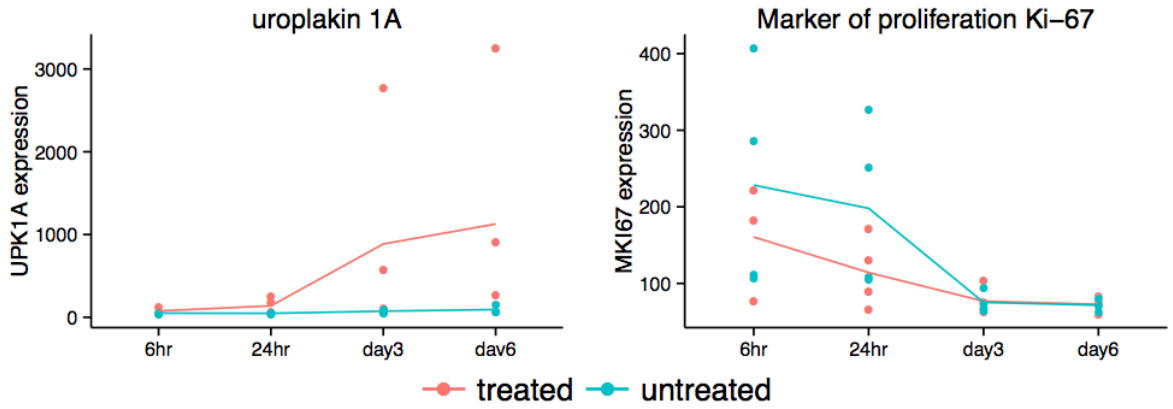


Figure 3.1: *Expression of proliferation and differentiation markers in NHU. Microaray measured expression of the uroplakin 1A (UPK1A, a marker of urothelial differentiation) and of the marker of proliferation Ki-67.*

models were available for the same Transcription Factor (TF), the PWM with the highest Information Content (in bits) was kept. Co-regulatory interactions ($TF \leftrightarrow TF$) were supported by the high-confidence protein-protein interactions of the HIPPIE database (Schaefer et al., 2012).

The original CoRegNet-inferred network was refined using the default functions of the package (see section 1.6). This resulted in a large-scale regulatory network of normal urothelial proliferation and differentiation, containing 36,994 regulatory interactions and supported by 2,895 ChIP (odds ratio: 1.8, fisher's exact test: $p < 10^{-20}$) and 1,293 TFBS-based interactions (odds ratio: 1.6, fisher's exact test: $p < 10^{-20}$).

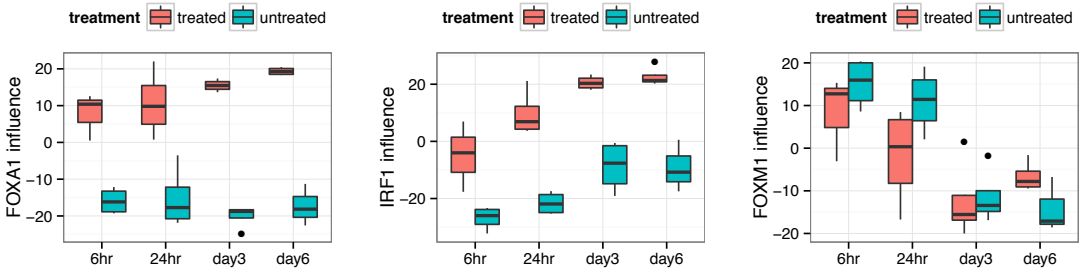


Figure 3.2: *Activity of differentiation and proliferation regulators. The transcriptional activity of two differentiation-associated transcription factor FOXA1, IRF1 and of the proliferation-associated transcription factor FOXM1 was predicted in the NHU samples using the* CoRegNet *bioconductor package. The activity (y axis) is expressed in arbitrary units representing the measure influence.*

As a first insight into the predicted transcriptional programs of normal differentiation and proliferation, the activity of three transcriptional factors was predicted using the measure of *influence* (see section 1.4). *FOXA1* and *IRF1* are the first transcription factors activated by *PPARγ* to drive urothelial differentiation (Varley et al., 2008). Figure 3.2 is a plot of the predicted activity of these two TF which show an increasing activity from the first to the last time points following differentiation treatment. The influence of *FOXM1*, a master regulator of tumor proliferation and more generally of cell cycle, is also plotted in figure 3.2 showing only high activity during the first day, especially in untreated undifferentiated cells.

Figure 3.3 shows a global representation of the activity of the normal transcriptional programs in the 30 NHU transcriptomes. This illustrates well the ability of the *influence* measure to summarize the entire transcriptomic variations by the activity of a potentially involved transcription factors. For instance, the known regulators of differentiation *PPARγ* , *IRF1* and *FOXA1* (Varley et al., 2008) as well as *GRHL3* (Yu et al., 2009) are all predicted to be active in differentiated samples. Interestingly, the *HOXB* gene family is associated with the ureter-originating urothelium whereas the HOXA cluster is associated with the bladder-originating samples. A large number of TF are predicted to be active during proliferation, among which *FOXM1* is described as a master regulator of bladder cancer in the previous chapter. Moreover, *SOX9* and *SNAI2* (snail) are here predicted to be active in undifferentiated NHU and were also predicted to be master regulators of the basal-like bladder cancer subtype (see section 2.2). Interestingly, *SMAD3*, a downstream effector of the TGFβ pathway, is predicted to be part of the differentiation transcriptional program while *SMAD7*, a negative regulator of the TGFβ pathway, is specific to undifferentiated NHU cells. This observation is supported by the recent discovery of the role of the TGFβ pathway as necessary for the proliferation of differentiated NHU cells and a weak inhibitor of proliferation in undifferentiated NHU cells (Fleming et al., 2012).

3.3 Global contribution of normal urothelial regulatory networks to bladder cancer

Differential network analysis, described as the study of the dissimilarities between the interactions mapped in two conditions of interest, is considered as an innovative concept and seen as the next prevalent type of network analysis (Ideker and Krogan, 2012). However, based on biological observations, such as those of H. Dvorak (Dvorak, 1986), of the high similarity between normal regenerative process and tumor growth, I argue that the aberrant misuse of normal networks, that is the abnormal constitutive activation (or repression) of a normal network with a mainly conserved structure, is also of potential interest. Although it is likely that some transcription factors have differences in their sets of target genes between a normal and cancerous condition, it is also more likely that the transcriptional network of normal cellular proliferation is simply activated instead

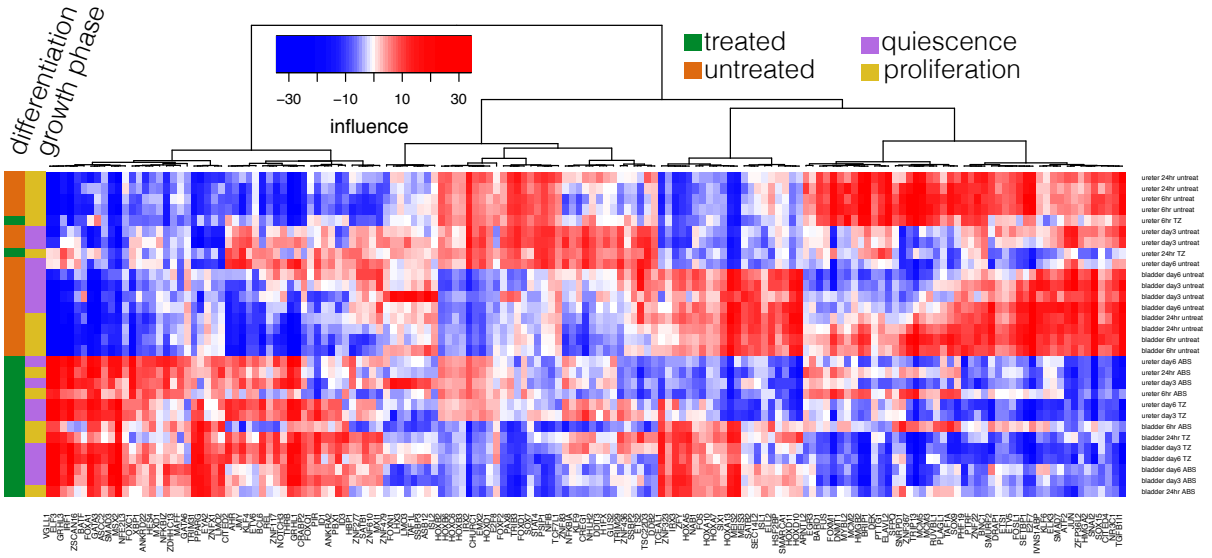


Figure 3.3: *Activity of normal transcriptional programs in NHU. Heatmap representation of the influence of all transcription factors of the inferred regulatory network with at least 10 activated and 10 repressed target genes. 6 hours and 24 hours samples are considered as proliferating samples (yellow) whereas day 3 and 6 are considered as quiescent samples (purple). Differentiation-inducing treatments are also color-coded (green trated, orange untreated). A dendrogram of the hierarchical clustering (correlation distance and ward's method) of transcription factors is shown above the heatmap.*

of being entirely reorganized during neo-plastic transformation. Therefore, the task of comparing the functioning, at the transcriptional regulation level, of normal and malignant cell proliferation, can be carried out by identifying which and how normal processes are aberrantly controlled instead of searching for differences in interactions.

In order to define to what extent the proposed large-scale regulatory network of NHU cell proliferation and differentiation is conserved in and contributing to bladder cancer, a first analysis of gene regulation was carried out. Each of the local gene regulatory networks, that is the set of co-activators and co-inhibitors of each gene, extracted from the NHU transcriptomes by CoRegNet were tested in a bladder cancer dataset of 179 bladder cancer transcriptomes, hereafter referred to as the CIT dataset (Carte d'Identité des Tumeurs, Rebouissou et al., 2014). Each local network was fitted using the same linear regression model used to refine the network inferred using CoRegNet. In essence, the expression of co-regulators was used as predictor variables and the expression of the target gene as a response variable. The linear model was fitted on 90% of the samples to predict the expression of the target gene on the remaining 10% in a cross-validation setting. A threshold of 0.5 was used on the cross-validated coefficient of determination to select the best *normally regulated* genes in cancer cells. This arbitrary threshold, which selects genes for which at least 50% of the variation in expression is explained by

CIT

	Function	# genes (%)	Enrichment	p-value	FDR
GO	cell cycle phase	81 (30%)	8.4	10^{-20}	$< 10^{-20}$
	nuclear division	60 (22%)	10.4	$< 10^{-20}$	$< 10^{-20}$
	DNA replication	34 (13%)	7.6	$< 10^{-20}$	10^{-17}
KEGG	Cell cycle	19 (7%)	6.9	10^{-11}	10^{-8}
	DNA replication	12 (5%)	12.3	10^{-10}	10^{-7}

TCGA

	Function	# genes (%)	Enrichment	p-value	FDR
GO	M phase	64 (26%)	13.4	$< 10^{-20}$	$< 10^{-20}$
	cell cycle process	75 (31%)	9.1	$< 10^{-20}$	$< 10^{-20}$
KEGG	DNA replication	12 (5%)	23.2	10^{-13}	10^{-9}
	Cell cycle	17 (7%)	9.5	10^{-12}	10^{-9}

Figure 3.4: *Cellular functions with conserved normal regulation. Functional analysis of the genes for which the normal co-regulators identified in the NHU transcriptomic dataset are conserved in two bladder cancer data set is shown. 254 and 255 genes for the CIT and TCGA datasets respectively were selected for having a coefficient of determination over 50% ($R^2 > 0.5$). Enrichment is computed as the odds-ratio of finding genes annotated with a specific cellular function among the set of selected genes.*
GO: Gene Ontology, biological process only.

normal regulators, is used simply to investigate whether genes for which the regulators are conserved during tumorigenesis are involved in specific cellular function. Table 3.4 lists the top cellular functions found to be significantly over-represented among the annotation of these normally regulated genes. Core processes involved in cell proliferation are well represented suggesting that indeed, mitosis is regulated the same way in normal and malignant cells. This was verified by reproducing the same analysis in an independent series of bladder cancer transcriptomes from the TCGA analysis (Cancer Genome Atlas Network, 2014).

To further analyze the conservation of the normal regulatory network in cancer samples, the influence of the entire NHU network was compared to the influence of 100 randomly generated networks with similar topology (permuting target genes of every TF in the network). The global influence of the network is measured by the sum of the absolute influence of all TF in the network. Figure 3.5 shows the distribution of the global influences of the NHU network and of the random networks for each sample of the CIT dataset. The NHU network had a higher global influence in any tumor sample than any of the random networks in any of the tumor samples. This significantly higher NHU network influence (Student's t test: $p < 10^{-20}$) suggests that a large part of the normal network is conserved

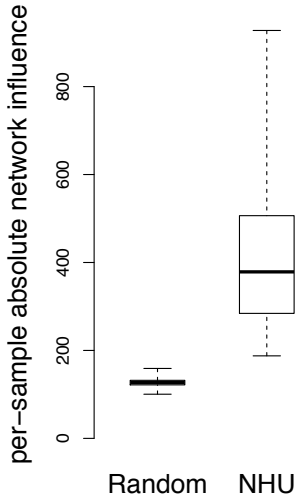


Figure 3.5: *Influence of NHU network is conserved in bladder cancer. The absolute value of the influence of all TF in the networks (NHU or random) is computed in each sample of the CIT dataset. For each sample, the absolute influence of all TF is summed. The ranges of the boxplots are minimal and maximum values.*

in tumors.

3.4 Contribution of normal Master Regulators to bladder cancer

In order to gain further insight into which and how normal transcriptional programs are conserved during tumorigenesis, the influence of the transcription factors of the NHU network was computed in the cancer samples. It is to be noted that the influence, calculated using the transcriptional targets predicted in the NHU transcriptomes, is computed using the cancer gene expression data centered on the mean expression of the normal samples (biopsies of non-cancerous patients, not NHU cell cultures) of the dataset. Therefore, the measured influence represents the extent to which a TF is accountable for the variation of its target genes expression from the normal to cancerous samples.

In order to model the process of neo-plastic transformation, the transcriptional programs for which the regulatory structure of the normal network was most highly conserved were extracted from bladder cancer transcriptomic datasets. The idea is to model which part of regulatory network of the normal urothelial proliferation and differentiation are particularly conserved, whether the associated transcription factor are silenced or activated. To do so, the absolute value of the influence of the normal regulators was used as a measure of network conservation. Figure 3.7 shows the high overlap between the 30 first most

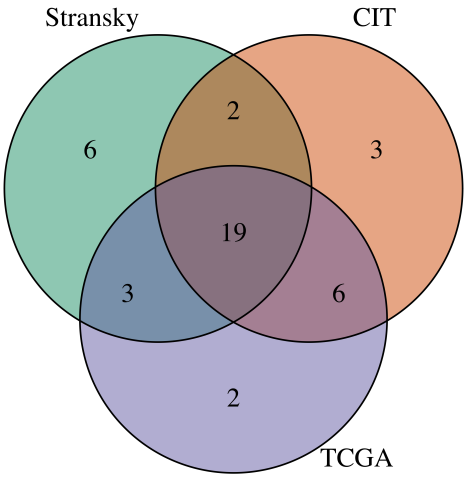


Figure 3.6: *Reproducible identification of conserved master regulator. Overlap of the 30 first Master regulators of the NHU network in three different bladder cancer dataset: CIT (179 bladder tumor samples, affymetrix U133 plus2 chips , Rebouissou et al., 2014), TCGA (211 bladder tumor samples, RNAseq, Cancer Genome Atlas Network, 2014) and Stransky (79 bladder tumor samples, affymetrix U95 chips, Stransky et al., 2006).*

conserved TF in three different bladder cancer datasets. The preservation of the normal regulatory network structure is highly reproducible. For instance, the sum of the absolute influence in all samples of each dataset, used here as a measure of network conservation, highly correlates between these three datasets (Pearson's correlation, CIT and TCGA: 0.91, CIT and Stransky: 0.95, TCGA and Stransky: 0.89). These results show that the importance of each of the normal regulators, using the absolute influence as a ranking, is highly reproducible.

The transcription factors with the highest absolute value of influence in all cancer samples are illustrated in figure 3.7. The TF in the figure are sorted by their sum of influence in their absolute value but the distributions represented as boxplots are the real valued influence and hence showing both highly active TF and inactive TF. Interestingly, TF previously described as master regulators urothelial differentiation (*e.g. IRF1, PPARG, FOXA1, GRHL3*) have a loss of activity in cancer samples while master regulators of cellular proliferation (*e.g. FOXM1* in the study of the bladder cancer network in section 2.1 and in Lefebvre et al., 2010; Raychaudhuri and Park, 2011) are predicted to have an increased activity.

This characteristic transcription factors suggest that transcriptional programs involved in urothelial differentiation are silenced during carcinogenesis while proliferation programs are activated, or that their high level of activity is preserved. Figure 3.8 shows the relation between the influence of regulators measured in proliferating or differentiated NHU and the influence in the CIT dataset. The plot shows that TF with the highest predicted activity

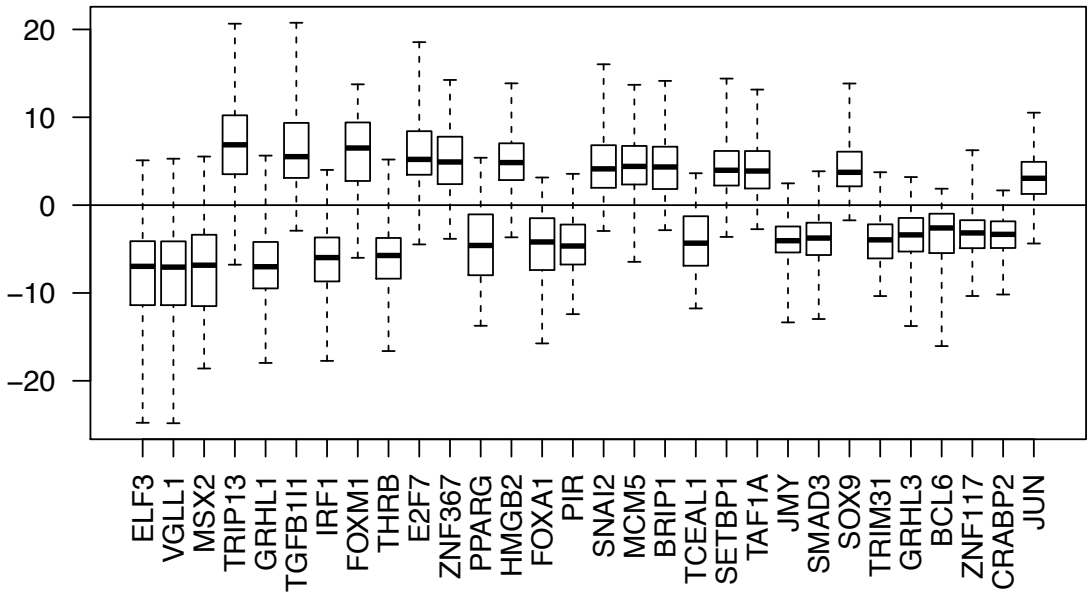


Figure 3.7: *Top influent normal regulators in bladder cancer. 30 regulators of the NHU network with the highest sum of influence in all cancer samples (in absolute value). TF are sorted by the sum, in all samples, of their influence absolute value. Ranges of boxplots correspond to minimal and maximal values.*

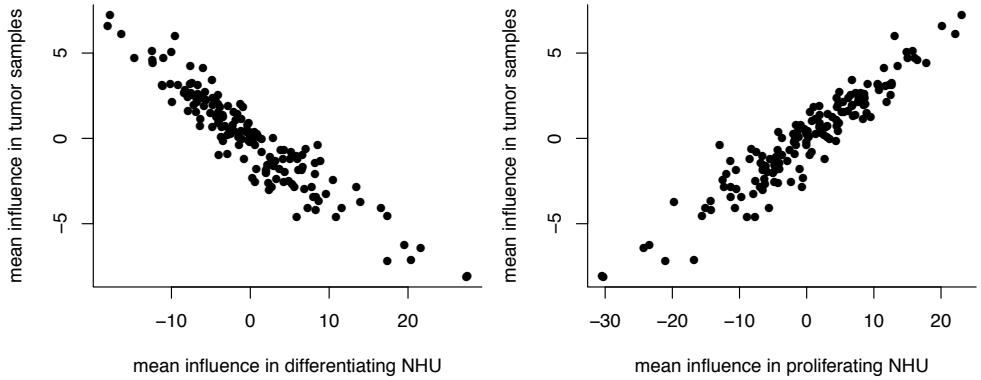


Figure 3.8: *Relation between TF influence in cancer and differentiated or proliferating NHU. The data used is the mean influence of each TF either in all CIT cancer samples, (left) in the two last time points of the differentiating NHU (ABS or TZ/PD treatment, day 3 and 6) or (right) in the growth phase of undifferentiated NHU (no treatment, 6 or 24 hours).*

in cancer samples are also those with the lowest predicted activity in differentiating NHU (Pearson's correlation: -0.93). However, TF with the highest influence in proliferating NHU are also the most active TF in cancer (Pearson's correlation: 0.94). Although expected, this result illustrates well the obvious main characteristic of cancer, cellular proliferation. Moreover, cellular differentiation is also a process generally known to be lost in cancer. For instance, all invasive bladder cancers are high-grade tumor for which histo-pathological analysis show a loss of cellular differentiation.

Such broad conservations of normal urothelial specific programs in cancerous cells brings to think that the cellular organization of cancerous cells is mostly unchanged and that normal programs of proliferation may not be reorganized but instead may be irrevocably set into a particular tumorigenic state through molecular alterations such as Somatic Mutations or Copy Number Aberrations.

To identify regulator of proliferation and differentiation that are genetically altered and which might lead to an upstream constitutive activation, or repression, of a broader transcriptional program, CGH arrays were used to identify copy number aberrations (CNA) in 87 muscle-invasive bladder tumors of the CIT dataset. As CNA of transcription factor encoding genes can only affect transcriptional activity if it is reflected by an increased expression levels, only normal transcriptional regulators for which an increase in copy number was reflected by an increase in its expression were selected (right tail Student's t test of the expression between samples with gain or amplification and other samples, FDR: 1%).

Among the transcription factors for which the copy number status corresponded to increase in gene expression, *MYBL2* was the regulator that was the most frequently found in genomic regions of gain of copy (40% of muscle-invasive samples) and is also found in regions of high amplification (3% of muscle-invasive samples). The same analysis in the TCGA data also identified *MYBL2* as the transcription factor with corresponding increased copy number and expression with the most frequent gains (60% and 3% amplification). *MYBL2* was previously observed to be highly expressed in proliferating cells and was found to regulate and be regulated by the cyclin A1 in acute myeloid leukemia (Muller-Tidow et al., 2001).

In order to assess the role of *MYBL2* in bladder cancer cell proliferation, a knockout experiment was carried in a representative bladder cancer cell line. The effect on cellular proliferation and viability was determined by MTT colorimetric assay following *MYBL2*-targeting siRNA transfection. Figure 3.9 shows the effect of silencing *MYBL2* on the survival of the Scaber bladder cancer cell line, which originally shows one of the highest *MYBL2* influence (4th among 35 cell lines). *MYBL2* knockout results in a 40% to 50% decrease in cell number, which, in addition to its frequent gain of copy and increase expression, supports its role as a constitutively activated regulator involved in a transcriptional program driving urothelial cell proliferation.

These results support the idea of a constitutive activation of transcriptional programs of urothelial cell proliferation. However, several results, such as the loss of activity of

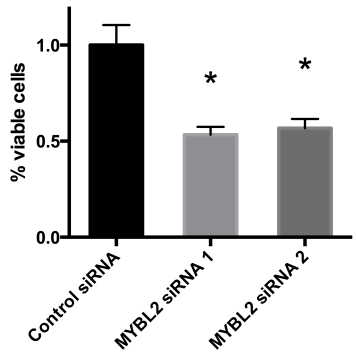


Figure 3.9: *MYBL2 knockout in the Scaber bladder cancer cell line. Quantification of viable cells following the transfection of a control or MYBL2 targeting siRNA was determined by colorimetric MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] assays. The data shown are numbers of viable cells relative to the control siRNA. Experiments were conducted twice, each in triplicates. *: Student's test* $\alpha = 1\%$.

the differentiation-related TF (see 3.8), show that tumorigenesis is accompanied by a loss of cellular differentiation at the transcriptional level. One evident explanation is the constitutive activation of *EGFR* in basal-like bladder cancers (Rebouissou et al., 2014) leading to the inactivation of the driver of urothelial differentiation *PPARγ* (Varley et al., 2004). However, basal-like bladder cancers only represent approximately 20% of tumors (Rebouissou et al., 2014).

3.5 Defining the role of *ELF3*, a master regulator of differentiation in bladder cancers

The partial release of the TCGA exome sequencing of bladder cancers identified several mutations of the *ELF3* gene. *ELF3* is a small gene encoding for an epithelial-specific transcription factor and that is predicted to be the TF with the most important decrease in transcriptional activity in bladder cancer samples (see figure 3.7). Figure 3.10 shows the influence and expression of *ELF3* in each stage of bladder cancer showing a marked decrease during tumor progression.

In order to assess the genetic status of *FGFR3* in bladder cancer samples, we sequenced the exons of *ELF3* in 106 samples. We found *ELF3* mutations in 18 samples (approximately 17%), one with two different mutations. 10 mutations were deleterious, 1 nonsense and 9 frameshifts due to the insertion or deletions. The somatic status of the identified mutation was verified for 9 samples, those for which normal DNA was available. All mutations were heterozygous mutations. The mutations are mapped on the *ELF3* encoded protein in figure 3.11. *ELF3* mutations were not associated to any bladder cancer subtype or any

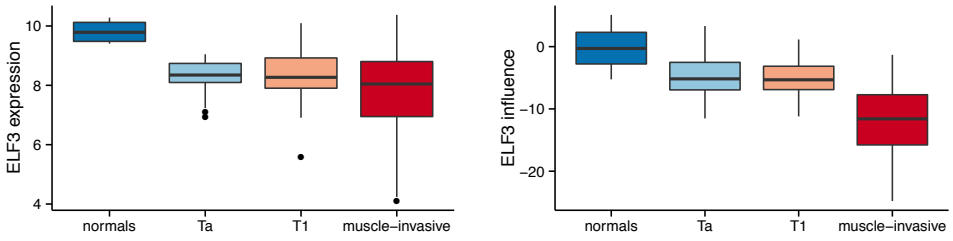


Figure 3.10: *ELF3 expression and influence by stages of bladder cancer. Expression levels are log2 transformed RMA normalized expression of affymetrix u133 plus2 chips. Influence levels are as computed by the* CoRegNet *package using an NHU-inferred regulatory network and computing the TF activity in the CIT bladder cancer transcriptome dataset.*

tumor stage.

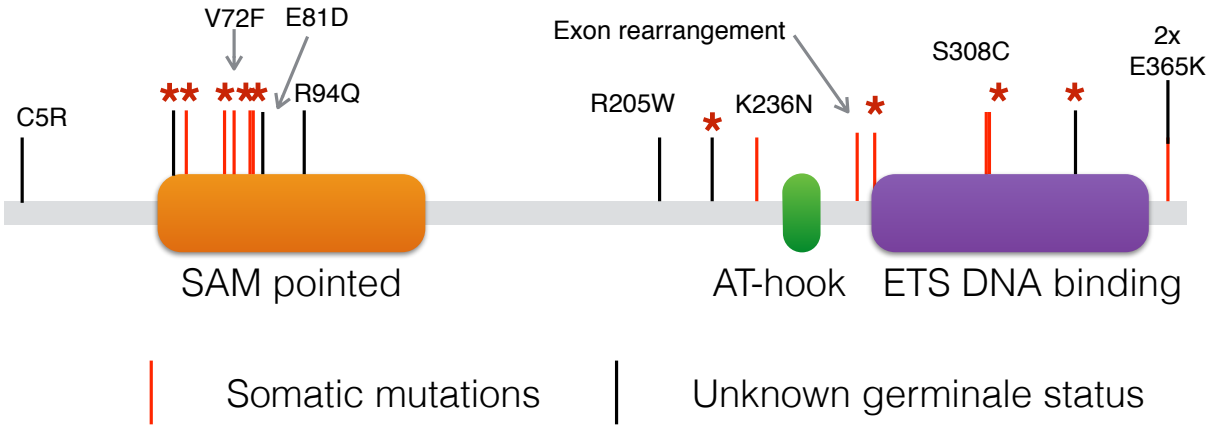


Figure 3.11: *ELF3 mutations. Mapping of the 19 mutations found in 18 bladder cancer samples. Red stars identify deleterious (nonsense and framshift) mutations. Other missense mutations are identified by the amino-acid replacement. Mutations E81D and R94D were found in the same sample.*

The final release of the TCGA data indeed identified *ELF3* as a significantly mutated gene (Cancer Genome Atlas Network, 2014) mostly because of the low probability of finding so many deleterious mutations (mutation found in 8% of samples: 7 framshifts, 1 nonsense, 1 splice site and 2 mi sense) in such a small gene (coding for 371 amino acids). Moreover, a recent study of NHU differentiation identified *ELF3* as necessary to maintain trans-epithelial resistance, characteristic of the barrier function of the differentiated urothelium, in NHU cell cultures (Böck et al., 2014). Overall, these results suggest that the deleterious mutations found in the *ELF3* coding sequence silences the transcriptional program of normal urothelial differentiation in bladder cancer cells.

In order to further understand the function of *ELF3* as a master regulator of urothelial

differentiation, our partners at the university of York performed a knockdown assay of *ELF3* in NHU cell cultures with a differentiation inducing media (TZ/PD). Resulting in an approximately 60% decrease in the expression of *ELF3*, the knockdown resulted in the under-expression of several genes involved in urothelial differentiation and in particular transcription factors that are effectors of differentiation (*IRF1*, *FOXA1* Varley et al., 2008, *GRHL3* Yu et al., 2009 and *KLF5* Bell et al., 2011). Figure 3.12 shows the impact of *ELF3*-knockdown on the expression of 8 genes, including *ELF3* itself and its known target the claudin 7 (Kohno et al., 2006). These results suggest that *ELF3* is necessary to the activation of a large part of the differentiation transcriptional programs although not directly activating differentiation markers such as uroplakins.

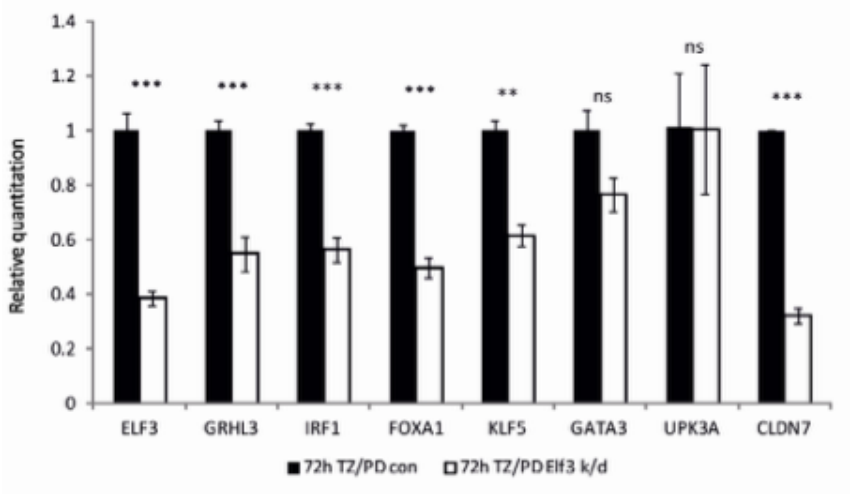


Figure 3.12: *qPCR expression measurements following ELF3 knockdown in NHU. cDNA qPCR measurement of 8 genes following shRNA-mediated knock down of ELF3 in NHU cultured until 70-80% confluence after induction of differentiation with 1μM troglitazone (TZ) and 1μM PD153035 (PD) for 72h. Values (y axis) are relative quantities compared to control shRNA after GAPDH normalization. Statistical analysis were carried out using an analysis of variance (ANOVA,* $***: p < 0.001; **: p < 0.01$*).*

NHU cell cultures transfected with *ELF3*-directed or scramble shRNA were sampled at four different time points after TZ/PD induction of differentiation to reflect the various stages of early and late differentiation (12hr, 24hr, 2 and 3 days, all in triplicate). No genes were found to be differentially expressed after 12 hours (moderate t test, FDR 1%). The effect of *ELF3* knockdown on differentiation-related genes was observable as soon as 24 hours after differentiation induction. For instance, *FOXA1* and *PPAR*γ were under-expressed as soon as 24 hours after TZ/PD treatment compared to the NHU cultures transfected with a scramble shRNA.

Out of the 33,692 profiled genes, 3,171 and 2,980 were identified to be significantly over and under-expressed, respectively (moderate t test, FDR 1%) due to *ELF3* knockdown.

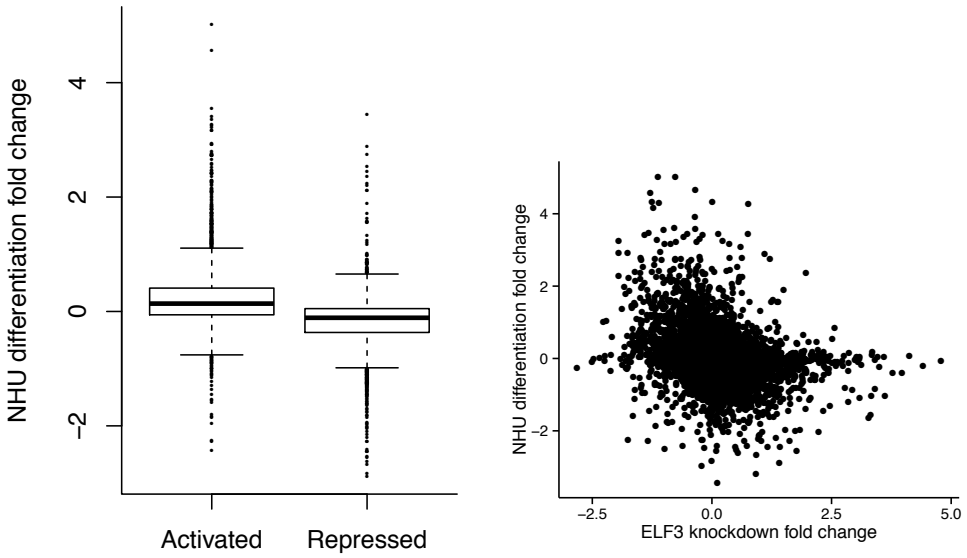


Figure 3.13: *Expression of ELF3 regulated genes in normal NHU differentiation. left. Distribution of the expression of ELF3 regulated genes identified in ELF3 knockdown (under-expressed: activated, over-expressed: repressed) during NHU differentiation. Values shown (y axis) are mean log fold change between the two last time points of NHU cell cultures (day 3 and 6) with versus without differentiation induction (ABS or TZ/PD). right. Comparison of expression following ELF3 knockdown and expression in differentiated NHU. x axis, mean log fold change between ELF3 and scramble shRNA. y axis, mean log fold change between the two last time points of NHU cell cultures (day 3 and 6) with versus without differentiation induction (ABS or TZ/PD).*

Figure 3.14 shows a comparison between the expression of genes following *ELF3* knockdown and of normal NHU differentiation. These result show that the genes activated by *ELF3* have an overall significantly higher expression in differentiated NHU samples than genes repressed by *ELF3* ($p < 10^{-20}$). Moreover, a comparison of the two sets of transcriptomes (in form of mean log fold change, right panel of figure 3.13) shows a significant anti-correlation (Pearson's $R^2 : -0.27$, $p < 10^{-20}$) between the NHU differentiation and the *ELF3* knockdown log fold changes. These results highlight the importance of *ELF3* as a Master Regulator of the normal transcriptional program of human urothelial differentiation.

Epithelial-Mesenchymal Transition (EMT) is a process inducing stem cells and migratory properties, prevents apoptosis and senescence, and normally occurs during development and wound healing (Thiery et al., 2009). The acquisition of EMT feature by cancer cell is generally considered as a pro-tumorigenic event as it confers to the cell several of the hallmarks of a malignant phenotype. Interestingly, several markers of EMT such as fibronectin, vimentin or the n-cadherin show a significantly higher level of expression, up to a seven-fold increase, in *ELF3* knockdown NHU (log fold changes: *FN1* 2.07, *MMP2*

1.14, *CDH2* 2.86, *VIM* 1.42, all significant under 1% FDR).

Overall, these results show the importance of *ELF3* in the regulation of the urothelial differentiation transcriptional program. This major role of *ELF3* suggest that its frequent heterozygous and deleterious mutations in bladder cancer cause a partial dedifferentiation of urothelial similar to EMT. *ELF3* was actually shown to be down-regulated in EMT and to actively drive to opposite process of Mesenchymal to Epithelial Transition (reviewed in De Craene and Berx, 2013).

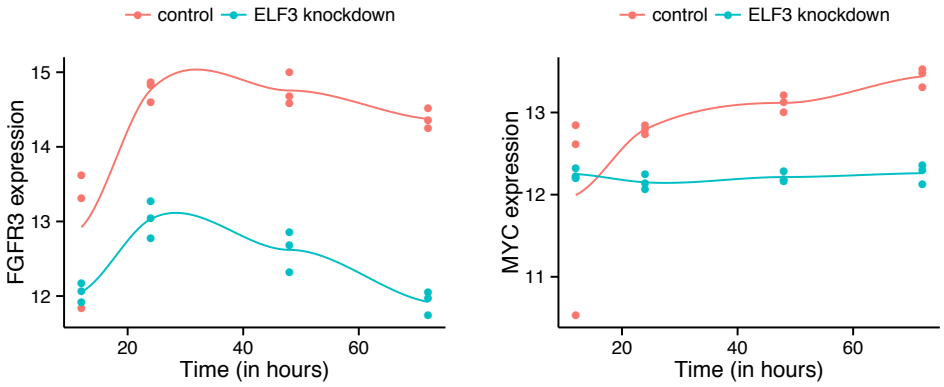


Figure 3.14: *Expression of FGFR3 and MYC in the ELF3 knockdown experiment. Microarray measurements of the expression FGFR3 in MYC following shRNA-mediated knock down of ELF3 in NHU cultures after induction of differentiation with 1µM troglitazone (TZ) and 1µM PD153035 (PD) for 12, 24, 48 and 72 hours.*

Interestingly, one of the genes with the highest decrease in expression following the depletion of *ELF3* in differentiating NHU was *FGFR3* with a four-fold decrease in expression as shown in figure 3.14. The figure also illustrates the marked decrease in the expression of *MYC*, recently identified as a downstream TF of the *FGFR3* pathway (unpublished result). *FGFR3* is one of the most frequently mutated genes in bladder cancer, most frequently presenting activating mutation (Cappellen et al., 1999). *FGFR3* is a driver of 30% to 60% of bladder cancer, when including non-muscle-invasive tumors. Recently, *FGFR3* activating mutations were associated with the luminal-like subtype of bladder cancer, which has the specificity to show particularly high levels of expression of urothelial differentiation markers although lower than normal urothelium (Cancer Genome Atlas Network, 2014). Although *ELF3* is necessary to the expression of *FGFR3* in NHU samples, luminal-like bladder cancers are mutated for *ELF3* (24% mutated samples in our cohort, 14% in the TCGA).

These unexpected evidences along with the fact that all identified *ELF3* mutations are heterozygous suggest that *ELF3* is only partially altered in luminal-like *FGFR3*-dependent bladder cancer cells and that a minimal level of activity is necessary for *FGFR3* mediated

tumor progression in bladder cancer. Previous studies demonstrated the role of *ELF3* as a necessary effector of the *ERBB2* oncogenic pathway in breast cancer (Coppe et al., 2010) and of the oncogenic NF-κB regulator in prostate cancer (Longoni et al., 2013).

Given the major role of *ELF3* in urothelial differentiation and in oncogenic pathways of cancers in other tissue, we propose a model of the subtype-dependent role of *ELF3* in bladder cancer. Given that the a full activation of *ELF3* triggers the urothelial differentiation program, *ELF3* is required to be inactivated in all bladder cancer samples, either by deleterious mutations or under expression as shown in figure 3.10. However, given that a complete inactivation of *ELF3* may cause a down-regulation of the *FGFR3* oncogene, a partial inactivation is necessary in luminal-like *FGFR3*-dependent bladder cancer.

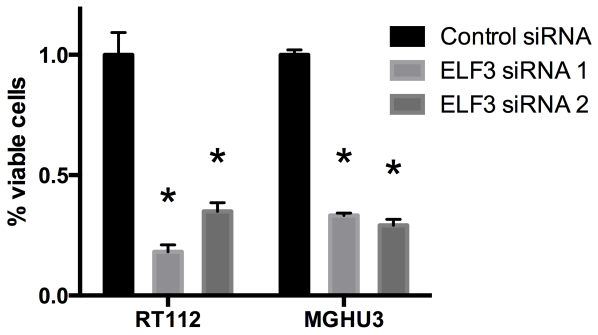


Figure 3.15: *Effect of ELF3 knockout in FGFR3-dependent bladder cancer cell lines. Quantification of viable cells following the transfection of a control or ELF3 targeting siRNA was determined by colorimetric MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] assays. The data shown are numbers of viable cells relative to the control siRNA. Experiments were conducted twice, each in triplicates.*

In order to investigate this hypothesis of the ambiguous and context-specific role of *ELF3*, the effect of siRNA-mediated knockout of *ELF3* on cell viability was assessed on two *FGFR3*-dependent cell lines: MGHU3 which is mutated for both *FGFR3* and *ELF3*, and RT112. The results of this functional validation are presented in figure 3.15 and show a strong decrease on cell viability. Despite the mutation of *ELF3* in the MGHU3 cell line (RT112 is ELF3 wild type), its complete inactivation leads in a nearly 70% decrease in cell number.

Altogether, these results highlight the complexity of the role of *ELF3* in urothelial carcinogenesis and more generally of a potential differentiation-related progression pathway. Further experiments are to be conducted, in particular, the re-expression of *ELF3* in both *FGFR3*-dependent and basal-like related cell line will estimate to what extent *ELF3* is to be silenced to promote urothelial malignancy.

3.6 Discussion

Through the inference of the regulatory networks controlling normal urothelial proliferation and differentiation and by the quantification of the use of such network in bladder cancer sample, I could provide some computational evidences that cancer cells use normal regulatory circuits. This analysis of the "*unending wound*" at the level of cellular control circuitry provides a model to identify the alterations of the normal proliferation and differentiation processes, whether it is through constitutive activation or inhibition of normal regulatory programs or through their disruption and the creation of cancer-specific networks.

This study was conducted with the understanding that the inferred regulatory networks lack accuracy in terms of single interaction predictions. Therefore, no direct comparison of an error-prone NHU network and another error-prone bladder cancer network was performed. To ensure that errors in the network would have minimal impact on the results, the *influence* measure proposed in the CoRegNet package was extensively used.

The main contribution of this work is the identification of two main transcriptional programs, one regulating differentiation and the other driving normal proliferation, both of which are highly active in normal urothelial cells and genetically altered in bladder cancer. While the proliferation program is clearly constitutively activated by aberrant number of copies of the *MYBL2* transcription factor, the analysis of the differentiation program revealed unexpected results. Although cellular differentiation is a process that is thought to be lost during tumor progression, at least at the microscopic level of histo-pathology, the identification of *PPARγ* as a major oncogene revealed that some pathways related to differentiation are drivers of distinct bladder cancer subtypes.

The analysis of the normal programs showed that the seemingly necessary loss of differentiation is partly imputable to undoubtedly damaging mutations of *ELF3*, a recently characterized master regulator of normal urothelial differentiation. However, our investigations showed that total abrogation of *ELF3* leads to the substantial decrease of proliferation of *FGFR3*-dependent bladder cancer cell lines. These results indicate that *ELF3* is potentially a major effector of transcriptional programs related to differentiation, whether these are driving normal cellular differentiation or the proliferation of partially differentiated bladder cancer cells.

The hypothesis of differentiation-associated carcinogenesis and more specifically of *ELF3* as both having a tumor suppressor role (activating differentiation) and an oncogene function (necessary for *FGFR3*-driven proliferation) can have a major impact on our understanding of bladder tumor progression. Clémentine Krucker at the Institut Curie will perform further experimental investigations. At first, the effect of the re-expressing *ELF3* in bladder cancer cell lines on their proliferation will detail its role as a tumor suppressor and potentially explain its frequent and heterozygous inactivating mutations. Then, the depletion of *ELF3* in particular in *PPARγ*-dependent cell lines will help understand its role as an oncogene, or more specifically as a regulator with minimal-tumor-required-activity

in bladder cancers partly retaining their normal differentiation phenotypes and regulatory programs.

Chapter 4

PEPPER: Protein Complex Expansion using Protein-Protein interaction networks

The project that I will describe in this chapter implicated several collaborators. In particular, the integration of the algorithms into a Cytoscape application was done by Charles Winterhalter.

4.1 Introduction

Most cellular processes require a large number of proteins to assemble into functional complexes to perform their activity. Therefore, describing functional protein complexes taking part in given processes is critical to the underlying molecular mechanism understanding. Experimental protocols such as Affinity Purification followed by Mass-Spectrometry (AP-MS) have been devised to precipitate or *pull down* a protein of interest (*bait*) together with all the interacting proteins within the same protein complex (*preys*). However, these sets of *preys* may contain both false positives, proteins detected despite not actually interacting with the *bait*, and omit false negatives (Gingras et al., 2007), proteins interacting in the cellular context studied but not detected. Effective control experiments and usage of contaminants repositories can remove some false positives. However, false negative interacting partners identification, thereby the definition of the entire protein complex, remains challenging. Protein-Protein Interaction (PPI) data represents abundant information that can be employed for this purpose.

Protein complexes extraction from PPI networks is a very active area of research and many methodologies have been developed to tackle this problem. These computational methods generally model protein complexes as dense subnetworks within the complete set of PPIs and thus try to solve a graph-clustering problem or to identify dense regions. Clustering approaches were shown to be efficient either on large PPI networks or with large-scale experimental settings in which big numbers of *baits* result in context-specific

PPI networks (Bader and Hogue, 2003; Nepusz, Yu, and Paccanaro, 2012). However, these algorithms were not developed for use in small-scale AP-MS experiments (*e.g.* using only a single *bait* protein) and are unable to integrate experimental data with repositories of PPI.

We reasoned that although not all the protein partners may be detected in a given AP-MS experiment, these proteins might have been previously identified as interacting with either the *bait* or some of the *preys* of the experiment. Based on this hypothesis, we developed PEPPER, which addresses the problem of finding protein complexes by combining the experimental results of a single AP-MS assay with the available information from protein interactions in a global PPI network. PEPPER solves this non-trivial problem by using a multi-objective evolutionary algorithm, which was tested to demonstrate the relevance of our integrative approach. To do so, we used publicly available AP-MS datasets for yeast and human species and compared PEPPER's results with those of state-of-the-art protein complex discovery methods. Our findings highlight the relevance of integrating PPI repositories to the analysis of AP-MS experiments. We propose PEPPER as a Cytoscape application to further refine protein complex predictions through functional and topological analyses.

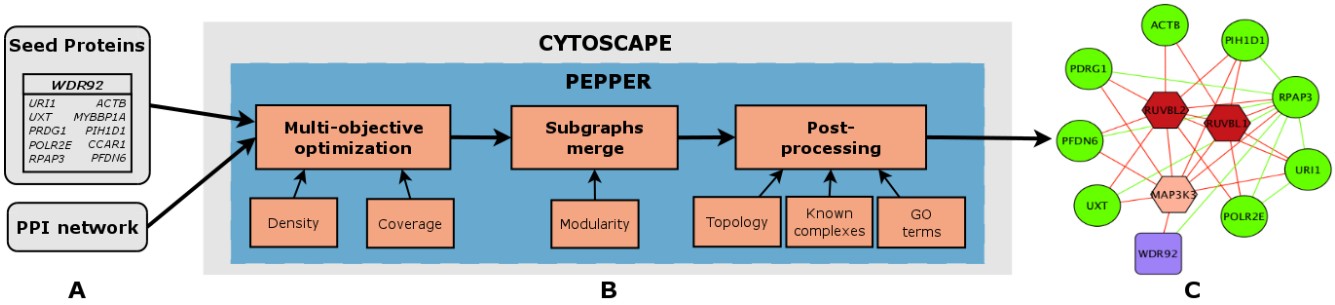


Figure 4.1: *Schematic representation of the plugin. (A) Example of input data, a large-scale PPI and the results of an AP-MS experiment with the bait and a list of prey proteins. (B) Context-specific protein complex extraction pipeline. (C) Output subnetwork representing a putative protein complex using only interactions from the input PPI network: example of WDR92. Purple squares and green circles correspond to bait and prey proteins, respectively. Hexagons indicate the expansions proposed by PEPPER and are shown in various shades of red, according to their post-processing score. Dark red indicates a high predicted relevance to the solution. The edges shown in the graph are exclusively those found in the input PPI network. Green edges are set between seed proteins. All edges involving an expansion protein are red.*

In the context of a single AP-MS experiment, PEPPER aims to identify a dense subnetwork within the PPI network connecting as many of the proteins identified in this experiment as possible, referred to hereafter as the list of *seed* proteins. PEPPER solves this problem by maximizing two objective functions: i) coverage; a solution must contain as many proteins from the *seed* protein list as possible, ii) density; a solution must contain

as many interactions as possible. These objectives are often conflicting thus, no single solution can be considered to dominate over the others. Instead, the optimal solution is a Pareto optimal set with multiple solutions. SPEA2 (Zitzler, Laumanns, and Thiele, 2001), a popular Multi-Objective Evolutionary Algorithm, is used for the simultaneous optimization of the two objective functions and to identify solutions approximating the set of Pareto optimal solutions. These solutions are merged into a final predicted protein complex by maximizing the modularity with a greedy search.

Pepper was developed as a Cytoscape application which uses a *seed* list of proteins and a large-scale PPI network as inputs (figure 4.1A). In addition to the aforementioned subnetwork extraction procedure, Pepper includes a topological and function-based post-processing pipeline for ranking the added proteins (*expansions*) according to their relevance (figure 4.1B). The predicted complex and each of the proteins are annotated based on their cellular localization or function annotation specificity. Enrichment analysis is complemented by matching the solutions to a collection of reference protein complexes, and *expansions* are scored according to their co-occurrence with the *seeds* in these complexes. Topological scoring is based on the impact of the *expansions* on the overall connectivity of the subnetwork. Pepper uses these scores to rank *expansions* and to facilitate results visualization and interpretation (figure 4.1C).

We assessed the performance of Pepper and two network clustering algorithms for protein complex discovery - MCODE (Bader and Hogue, 2003) and ClusterONE (Nepusz, Yu, and Paccanaro, 2012) - on a benchmark dataset of 135 yeast and 9 human single-bait AP-MS experiments and using a set of hand-curated protein complexes as gold standards. For network clustering methods, performance was assessed for each AP-MS experiment by selecting the predicted complex which best matched the *seed*. For each experiment, the reference complex from the gold standard best matching the *seed* was used as the ground truth in a binary classification task. Compared to both of the clustering methods tested, the complexes predicted by Pepper scored higher in all of the performance measures for both organisms with notably an average increase of 16% of the geometric accuracy in Human and 12% in Yeast.

As an example, we describe here the results obtained for the human WDR92 protein. In the initial list of *preys*, WDR92 was identified as interacting with only one protein. Pepper expanded the *seed* with three new proteins (figure 4.1C) and greatly increased the overall density of the original solution (22% to 47%). The new *expansion* proteins were ordered on the basis of post-processing score. The first two proteins, RUVBL1 and RUVBL2 have both a high topological and Gene Ontology score. The lower scored protein, MAP3K3, still remains relevant according to its high topological score (connected to more than 90% of the predicted complex proteins). AP-MS experiments using RUVBL1 or RUVBL2 as *baits* both identified WDR92 as a *prey* protein (Choi et al., 2010). Moreover, in the raw WDR92 experimental data, the set of *preys* with lower processing scores (based on peptide counts) than the threshold contains RUVBL1. Thus, the application of Pepper to this experiment led to the recovery of proteins that would not have been

identified otherwise (potential false negatives).

Overall, these results demonstrate the feasibility of expanding the protein complexes identified in an AP-MS experiment through the use of PPI networks and the value of PEPPER for this purpose.

This study was published in the journal *Bioinformatics* in which the two first authors contributed equally: C Winterhalter et al. (Aug. 2014). "Pepper: cytoscape app for protein complex expansion using protein-protein interaction networks." In: *Bioinformatics*

4.2 Methods

Problem Formulation

Given an unweighted and undirected graph $G = (V, E)$ (PPI network) with V the set of graph vertices (*i.e* proteins), $E \subseteq V \times V$ the set of edges (*i.e* protein interactions) and a list of interesting vertices P (*seed* list of proteins), the problem to be solved is to identify subgraphs $G' = (V', E')$ that are densely connected and include proteins in P.

Graph based objective The first objective is formulated as a maximization of the subgraph density $f_{density}(G')$. Computed as the ratio between the number of interactions found between proteins of the subgraph over the total of all possible interactions, it is defined as follows:

$$f_{density}(G') = \frac{2|E'|}{|V'|(|V'| - 1)}$$

where V' is the set of proteins in a given solution G' and E' is a subset of E containing only interactions between proteins in V'.

***Seed* list based objective** The second objective seeks to include as many *seed* vertices (proteins of interest) of P as possible in G' and is referred to as the coverage:

$$f_{coverage}(G') = \frac{|P \bigcap V'|}{|P|}$$

which is maximized whenever all proteins of the *seed* $|P|$ are chosen.

Both $f_{density}(G')$ and $f_{coverage}(G')$ functions are ranged in $[0, 1]$. Adding a large number of irrelevant proteins, up to an extreme solution in which all proteins are chosen ($V' = V$), will not degrade the coverage function. However, in practice increasing the number of irrelevant proteins in V' rapidly degrades the density ($f_{density}$) of the solution G'. In the same manner, small solutions may have high density (*e.g.* local cliques) but will rarely include many *seed* proteins from P.

Multi-objective optimization for relevant subgraph extraction

Finding dense subgraphs is an NP-hard task (Feige, Peleg, and Kortsarz, 2001). Optimization of multiple objectives makes the problem become even more intractable. To address the problem of extracting subgraphs satisfying the density and coverage criteria, PEPPER uses a Multi-Objective Genetic Algorithm (MOGA) approach to extract a set of solutions maximizing both objective functions. MOGA belong to a family of meta-heuristic optimization algorithms that mimic biological evolution and natural selection to evolve candidate solutions then determine the *fittest* individual - representing a solution - relatively to defined fitness functions. The Strength Pareto Evolutionary Algorithm 2 (*SPEA2*, Zitzler, Laumanns, and Thiele, 2001) was used to optimize simultaneously the Graph- and *seed* list-based objective functions. The MOGA components and operators are described in the following subsections.

Solutions representation and fitness function Given a PPI network $G = (V, E)$, a candidate solution is encoded into a binary chromosome of size $|V|$ representing the indexed set of V. A 1 value at position i corresponds to the presence of the i^{th} protein. PEPPER uses the *SPEA2* implemented in the open-source *JMetal* platform (Durillo and Nebro, 2011). Based on the notion of non-dominance for fitness evaluation, the algorithm searches a set of *Pareto* optimal solutions. A solution is Pareto optimal when no other solution is better in all fitness functions and therefore any of the objectives cannot be improved without degrading another. For instance in our problem, a set of proteins is a Pareto optimal solution when no other set has both a higher density ($f_{density}$) and higher coverage ($f_{coverage}$) and any other Pareto optimal solution with a higher value in one of these function will necessarily have a lower value in the other. The output of the MOGA is a set S of m solutions $S = G^1, G^2, ..., G^m$, which represents an estimation of the whole set of Pareto optimal solutions, also called Pareto front. All the solutions in S are available for custom visualization and post-processing in the Cytoscape app.

Initialization MOGA requires the initialization of a predefined number of chromosomes (population). The initial population is constructed with chromosomes composed of random proteins in P and as many proteins randomly picked in the neighborhood of P in G within a radius of 2.

Genetic operators and parameter settings PEPPER's MOGA optimises $f_{density}(G')$ and $f_{coverage}(G')$ objectives by performing changes driven by mutations (adding or removing a protein) and crossing-over (interchanging sets of proteins in two independent solutions) operators among chromosomes. At each iteration of the algorithm, random chromosome pairs are exposed to these operations and generate offspring sets. Given the two objectives, *fittest* chromosomes are then selected by binary tournament to evolve the population.

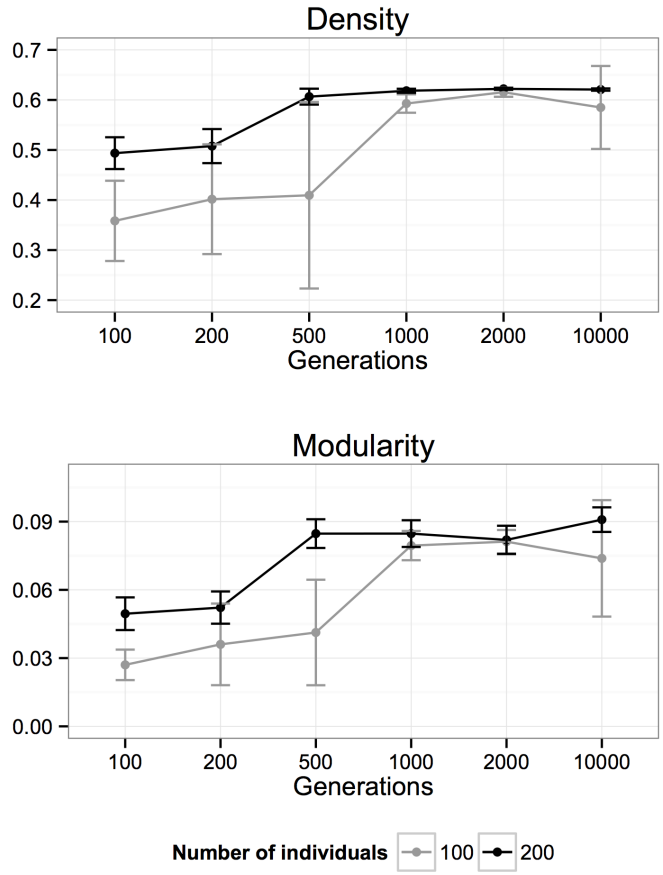


Figure 4.2: *Effect of parameters on density and modularity. Density and modularity of the merged optimal Pareto solutions as a function of the number of generations and size of the population. Results obtained on the ATG10 protein of the autophagy proteomic study. The error bar correspond to the standard deviation of 5 replicates.*

The MOGA requires several parameters that mainly impact the rate of convergence. These parameters include size of the population (number of individuals in a generation), number of generations (iterations), mutation rate, crossover rate and size of the Pareto front to return. In order to define a set of default parameters, we tested the MOGA on a Human proteomics dataset of the autophagy system (Behrends et al., 2010) using the Hippie protein network (Schaefer et al., 2012) after removing protein interactions originating from the proteomics study. Figure 4.2 shows density and modularity values of the merged pareto solutions (see next section) as a function of their size and the number of generations. Using 200 individuals per generation allowed to converge rapidly and with lower variance, as early as 500 generations for the example given in figure 4.2. Based

on these results and on general observations of other proteins from the same dataset, in PEPPER , default number of generations was set to 1,000 and number of individuals to 200. Maximum coverage is virtually always obtained independently of the parameters and is therefore not shown. The other parameters of the MOGA were set to the standard SPEA2 in the JMetal platform and are reported in table 4.1.

Parameter	**Value**
Population size	200
Crossover	$P_C = 0.90$
Mutation	$P_M = 0.10$
Number of generations	1000
Pareto front	10 solutions

Table 4.1: *Parameters of the multi-objective evolutionary algorithm. Crossover and Mutation, the genetic algorithm operators, are expressed as probabilities.*

Merging Pareto optimal solutions. From the union of Pareto optimal solutions $\cup_S$, we devised a simple algorithm to build a consensus *modular* subnetwork noted S_f. The use of modularity is based on a common observation that functional processes are often found in modular subparts of biological networks. This inspired clustering algorithms to use this measure that has led to successful application in particular in protein complex discovery problems (Nepusz, Yu, and Paccanaro, 2012). The modularity is defined here for a subset of proteins as the ratio between the number of interactions that occur only between these proteins against the number of other interactions involving these proteins in the whole PPI (G). It can be computed for a subgraph $G' = (V', E')$ as follow :

$$f_{modularity}(V', E') = \frac{\sum_{i,j \in V', i \neq j} E'(i,j)}{2 \times \sum_{i \in V', j \notin V'} E(i,j)}$$

with $G = (V, E)$ and $G' \subseteq G$.

The merge algorithm starts from all the proteins in P that were found in at least one solution ($\cup_S \cap P$) and iteratively tries to add one of the remaining *expansion* proteins ($\cup_S \setminus P$). At each step the *expansion* protein, which maximizes the overall modularity is kept until it cannot be increased anymore. This greedy algorithm has two characteristics. First, it keeps all the proteins from the initial set P which were identified in at least one of the Pareto optimal solutions of S. Second, it helps removing non-specific proteins which could have been added because of their high connectivity (which will increase density) but low specificity to the subnetwork of interest (which will decrease modularity), typically hub proteins of the network.

The final consensus network S_f is the protein complex predicted by PEPPER. Union of all optimal solutions ($\cup_S$) and predicted complex are the networks that are first generated

by PEPPER in Cytoscape, showing a mixture of expanded and initial proteins as well as a *bait* protein if provided. All solutions in S are also available for visualization and analysis in Cytoscape Results panel.

Assessment of *predicted* protein complexes

A set of methods is used to analyze the specificity of the predicted *expansions* to the solution and the initial list of proteins of interest (P). This allows indicating the relevance of the overall predicted protein complex and of each of the *expansion* proteins. Four scoring measures are browsable in Cytoscape Results panel for a given predicted complex and viewable as a color code for the *expansion* proteins based on:

- topological connectivity to assess the importance of a protein in connecting the predicted complex
- co-occurrence in a repository of hand-curated protein complexes
- similar functional annotation, particularly in terms of cellular localization and function

The aforementioned analysis is proposed as an integrated pipeline automatically performed following the evolutionary-base network extraction and merge steps.

Topological considerations. Four topological properties are used: degree and clustering coefficient, known to be good assessment factors in cellular biology and proteomics studies (Aittokallio, 2006; Glaab et al., 2010; Ozgur et al., 2008); modularity, which is used in the merge algorithm and more generally used for protein complex discovery (Nepusz, Yu, and Paccanaro, 2012); and closeness centrality, a measure used as an indicator of the overall similarity of a network nodes (Ozgur et al., 2008). These measures were calculated for:

1. The whole predicted subnetwork, *i.e.* the solution given by PEPPER noted S_f
2. Only proteins of interest used as an input to PEPPER present in the final solution ($P \cap S_f$)

These measures are reported in Cytoscape Results panel, in which differences between the original list and the final solution serves as a first indicator of PEPPER's predictions importance. Then, each of these topological measures are computed for each of the *expansion* proteins and are summarized in a global topological score ranged in $[0, 1]$ using the following formula:

$$Score_{topology}(X) = \frac{\sum_{\pi \in \Pi} \frac{X_\pi}{max(\pi)}}{|\Pi|}$$

with X a given protein, X_π the measure π associated to the protein X, $max(\pi)$ the maximum observed for measure π in the subnetwork and Π the set of all the topological measures used: degree, clustering coefficient, modularity or closeness centrality.

Overlap with known protein complexes. PEPPER was developed to solve problems encountered in proteomics studies, in particular for protein complex discovery. Therefore, the second measure of similarity takes into account the co-occurrence of predicted proteins in large collections of hand-curated protein complexes. These are available inside the plugin and were retrieved from the CYC2008 database for *S. cerevisiae* (Pu et al., 2009) and the CORUM database (Ruepp et al., 2009) for mammals.

Predicted complexes are evaluated using the matching score (also known as overlap score in Bader and Hogue, 2003):

$$MS(S,R) = \frac{|S \bigcap R|^2}{|S| * |R|}$$

where S and R respectively correspond to the sets of proteins in the predicted and reference complexes. The latter matching score is computed for any reference protein complex that presents at least one protein in the predicted subnetwork. For each match between a predicted and reference complex, PEPPER also generates and displays its associated performances in terms of sensitivity, precision and geometric accuracy (*cf.* section 4.3).

In order to evaluate and rank *expansion* proteins, each *expansion* is scored based on its occurrence in reference complexes associated to the solution given by PEPPER. This score is weighted by the matching score to give higher ranks to proteins that occur in reference complexes, which are more relevant to the solution. It is computed as follows:

$$Score_{complex}(S,X) = \frac{\sum_{r \in R} |X \cap r| \times MS(P,r)}{|R|}$$

where S is a PEPPER predicted complex, R is the set of reference complexes with at least one protein shared in S and X is one of the *expansions* in S. Known protein complexes matching thus results in a detailed list of overlapping complexes with PEPPER predictions but also provides a score translating *expansions* importance in those complexes.

Common functions and co-localization. Proteins associating in a complex are necessarily co-localized in the cell and are likely to share a given biological function. Based on this, gene annotations of cellular function and localization were used to estimate the relevance of the predicted complex and each of the *expansion* proteins. This was computed based on the Gene Ontology (GO) annotations. A hypergeometric test is used to identify Biological Process and Cellular Component annotations that are significantly associated with the predicted protein complex (with $\alpha < 5\%$).

To evaluate *expansion* proteins individually, each of them are scored by the number of annotations they share with those found to be specific to the overall predicted complex as follows:

$$Score_{GO}(X) = \frac{|X_{GO} \cap S_{GO}|}{|S_{GO}|}$$

where S_{GO} is the set of GO terms (Biological Process: *gobp*; or Cellular Component: *gocc*) associated with a solution of PEPPER S (fisher's exact test $\alpha = 5\%$), X is the protein contained in S that is scored and X_{GO} is the set of annotations of X. Each *expansion* protein is scored by the number of functional and localization terms it shares with the overall predicted complex.

Protein *expansions* global score. The integrated post-processing pipeline provides four distinct scores related to: (i) topology, (ii) reference complexes, (iii) Biological Process GO terms and (iv) Cellular Component GO terms. To summarize the information at a higher level, *expansion* proteins must be characterized by an integrated post-processing score. Several aggregation methods can be used to merge and normalize scores in a $[0, 1]$ range: mean, max or min. In order to identify the best aggregation method, we compared the ranking of the *expansion* proteins using these methods with an approximation of the optimal ranking. The reference ranking is defined by the ranked list that minimizes the distance with the rank given by each of the individual scores. This *optimal ranked list* was obtained using the R package RankAggreg Pihur, Datta, and Datta, 2009, which uses a Cross-Entropy method to identify the ranking minimizing the sum of absolute differences with the ranks of each individual score. Because of the computational time required to obtain this optimal ranking, RankAggreg was used only for comparison and was not directly integrated in the pipeline.

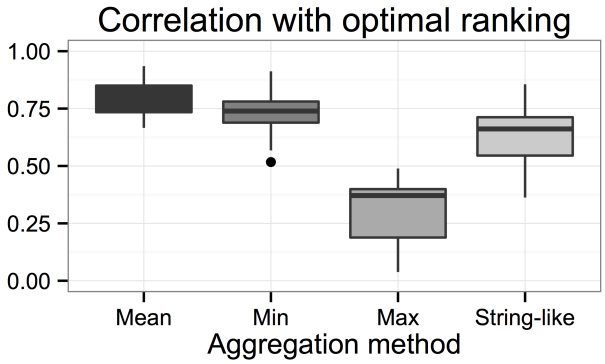


Figure 4.3: *Ranking aggregation methods. Distribution of Spearman correlations between optimal rankings of expansion proteins identified RankAggreg and ranking obtain by several score aggregation methods.*

Figure 4.3 shows the distribution of Spearman correlations of several score aggregation methods with the optimal ranking from RankAggreg for 10 sets of *expansions* from 10 pull-down assays led in the Human autophagy system (Behrends et al., 2010). Besides mean, max and min, we also used the score integration function introduced in the String database (Mering, 2004) $(1 - \prod_i (1 - S_i)$, with S_i corresponding to each individual scores).

These results show that the mean of the scores is the closest to the optimal ranking of the expansion proteins. The min function also appears to be an efficient aggregation method; however, a large set of *expansions* present at least one null score (approximately 30%).

In order to bring more flexibility to the global score calculation, PEPPER gives the possibility to add weights to each score and compute a weighted arithmetic mean such as:

$$Score_{postProcess}(X) = \frac{\sum_{i \in \zeta} \omega_i x_i}{\sum_{i \in \zeta} \omega_i}$$

where X represents a specific protein expansion, x its associated score to a post-processing feature, ω the weight given to the latter feature within ζ the set of post-processing assessment criteria:

$$\zeta = \{Score_{topology}, Score_{complex}, Score_{gobp}, Score_{gocc}\}$$

Each score weight has a default value equal to 1, which summarizes equally the information from each post-processing feature into a common mean $\frac{1}{n}\sum_{i=1}^{n} x_i$, where n is the overall number of assessment criteria. Choice to modify weights individually is left to users in PEPPER post-processing panel "Overview" tab. Tuning those parameters is a way to make abstraction of certain properties, which may be helpful for results visualization and interpretation since the plugin automatically updates the overall post-processing score as users modify weights. PEPPER dynamically translates each *expansion* overall score into a red color gradient (the darker the higher) in Cytoscape graphs.

4.3 Performance comparison

Comparison principles

To evaluate the ability of PEPPER to find relevant protein complexes, we applied it to real Affinity-Purification followed by Mass-Spectrometry (AP-MS) data and compared the results to gold standard sets of hand-curated reference protein complexes. Each of the AP-MS experiments performed on a single *bait* protein resulted in a list of *preys* and the union of both was used as a *seed* list of proteins. For each *seed*, the best matching reference complex from the gold standard was considered as the complex to be predicted. Therefore, only AP-MS with high matching reference complexes and for which no ambiguity was possible (only one highly matching reference complex) were selected. PEPPER was directly applied to each of these selected *seeds*. State-of-the-art protein complex discovery algorithms based on graph clustering, namely MCODE (Bader and Hogue, 2003) and ClusterONE (Nepusz, Yu, and Paccanaro, 2012), were used for performances comparison. Network clustering methods do not aim at finding protein complexes from a *seed* protein list of interest but rather enumerate all protein complexes in a PPI network. Therefore, these two methods were applied to the PPI network also used in PEPPER and the extracted

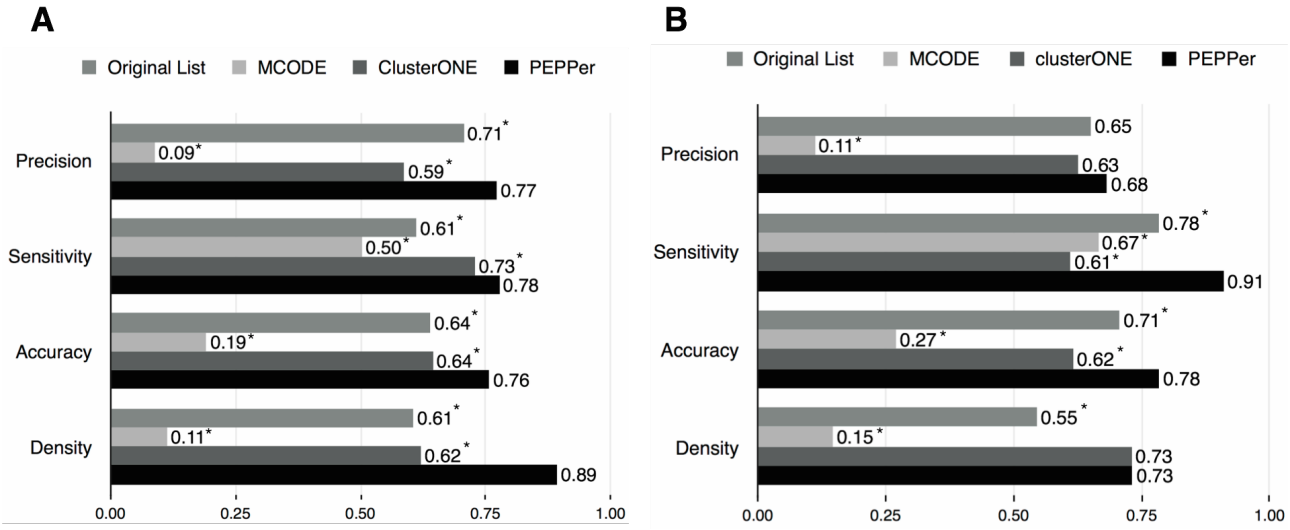


Figure 4.4: PEPPER*, MCODE and ClusterONE performances. (A) Protein complexes predicted from 135 single Bait AP-MS experiments in Yeast. (B) Protein complexes predicted from 9 single Bait AP-MS experiments in Human. The statistical significance is shown for comparison between MCODE and* PEPPER *as well as ClusterONE and* PEPPER *(* $*$ *:* $\alpha = 5\%$*)*

complex with the highest overlap with the *seed* was considered as its associated prediction. For fair comparison, we tested several overlapping measures (intersection, Jaccard and Matching-Score) and reported only the results of the measure with highest performance, which was obtained with the absolute size of the intersection.

Assessing performances

For each of the protein complexes predicted by PEPPER, ClusterONE or MCODE, the overlap of the set of predicted proteins with the known complex was computed as well as four common prediction performance measures:

- True Positive, TP: the number of proteins of the predicted complex that are found in the reference complex.
- True Negative, TN: the number of proteins that are not in the predicted complex and that are not found in the reference complex.
- False Positives, FP: the number of proteins of the predicted complex that are not found in the reference complex.
- False Negative, FN: the number of proteins that are not in the predicted complex that are in the reference complex.

Because the total number of proteins in the PPI network is several orders of magnitude higher than the number of proteins in the predicted or reference complexes, the number of TN provides little information. Therefore, we chose the following measures commonly

used in information retrieval:

• sensitivity, also called True Positive Rate (TPR), which evaluates how well positives are predicted,

$$Sn = \frac{TP}{TP + FN} \tag{4.1}$$

• precision, also called Positive Predictive Value (PPV),

$$Prec = \frac{TP}{TP + FP} \tag{4.2}$$

• geometric accuracy,

$$Acc = \sqrt{Sn * Prec} \tag{4.3}$$

• density which measures subgraphs connectivity degree

$$Density(V) = \frac{2|E|}{|V|(|V| - 1)} \tag{4.4}$$

where V and E respectively stand for the set of vertices (proteins) and edges (interactions) in a graph.

Results

A gold standard of manually curated protein complexes was used as a reference for *Saccharomyces cerevisiae* (Pu et al., 2009) and *Homo sapiens* (Ruepp et al., 2009). Single *bait* AP-MS experiments were obtained from a large-scale study in Yeast (Gavin et al., 2006) and Human (Choi et al., 2010). For each experiment, the *bait* and its associated set of *preys* were used as the *seed* list of proteins. Data for Yeast was already a set of curated proteins. In Human, only high-confidence proteins (SAINT score greater or equal to 99%) were kept as a list of *preys*. In order to assess the quality of predictions, only experiments for which a reference gold standard is available were selected. To this end, *seed* lists were selected based on the overlap with one of the complexes in the gold standard according to two criteria:

• the *seed* should contain more than 5 proteins in the same gold reference protein complex

• more than 50% of the *seed* should be contained in the same reference complex

From this filtering, 135 and 9 lists of *seeds* were selected for Yeast and Human respectively. The PPI networks used for the analysis were the default Yeast Biogrid network (Stark et al., 2010) and the HIPPIE database (Schaefer et al., 2012) for Human.

Performances of our method are reported in figure 4.4 alongside with those of MCODE, ClusterONE and of the original list of proteins used as seeds. Significant differences between Pepper and MCODE or ClusterONE were computed using Student's two-sided t-test with $\alpha = 5\%$. The higher performance of Pepper is statistically significant except

for density and precision in Human species. Moreover, unlike ClusterONE, MCODE identifies protein complexes for only a small portion of the total number of proteins. In Yeast, only approximately 40% of the proteins of the PPI network (2,449 out of 5,968 proteins) where assigned to a predicted complex whereas ClusterONE predicted a complex for nearly 80% of the proteins (4,742). The Human PPI networks being less connected, these proportions drop to 17% for MCODE and 36% for ClusterONE. Therefore, many seed lists of proteins that can be assigned to a known protein complex cannot however be mapped to an MCODE predicted complex. The results obtained with Pepper in Yeast or Human always showed an increase in all of the classification performance measures as compared to the original list of proteins or to the two tested methods. Interestingly, this increase in performance is associated with an increase of the density in Yeast. In Human, however, ClusterONE and Pepper find protein complexes with very similar densities. Yet, Pepper significantly outperforms ClusterONE with an average increase of 16% in accuracy and of 30% in sensitivity. These results suggest that extracting solutions solely based on optimising topological measures can be improved by integrating the context specificity of real experimental data.

4.4 Case study

An example of usage of Pepper is shown in figure 4.5 for a particular application on the Human protein *WDR92*.

Input data

The experimental results of an AP-MS assay performed using *WDR92* as a *bait* protein were obtained from a previously published study (Choi et al., 2010). From the raw list of proteins identified in the assay, 10 high confidence *prey* proteins (SAINT score greater or equal to 99%) were selected (list available as Supplementary File). In all, 8 interactions were found between all the proteins identified by the assay and three of the preys (*ACTB*, *CCAR1* and *MYBBP1A*) were not connected to any of the other preys.

Main prediction results

From the *seed* list, Pepper generated a consensus subgraph covering 75% of the initial *seed* proteins with density and modularity values of respectively 0.47 and 0.032 (figure 4.5B). The complex predicted by Pepper and the original network built from the *seed* list are shown in figure 4.5F-G. Pepper predicted three *expansions*: *RUVBL1*, *RUVBL2* and *MAP3K3*.

The *expansion* proteins predicted by Pepper greatly increased the connectivity of the initial solutions, which was measurable for several topological features described in

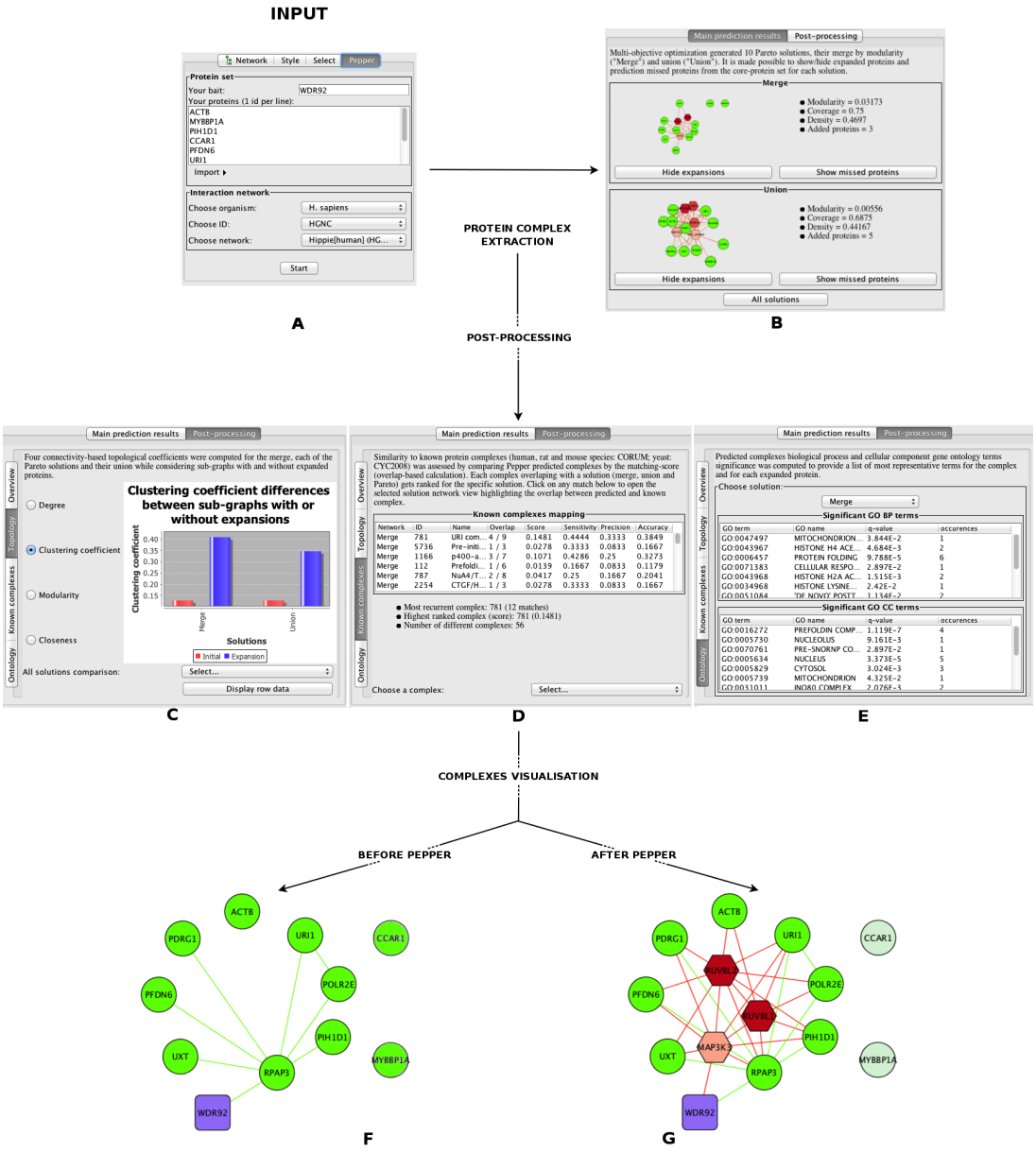


Figure 4.5: PEPPER *user interface. A view of the user interface of* PEPPER *in Cytoscape and the application to the WDR92 case study. (A) Necessary inputs are the organism (with default PPI networks for Human, Yeast and Mouse) and the list of seed proteins. (B) The predicted protein complexes, including first the final merged solution then all the Pareto optimal solutions and its union, are directly visible in the first tab of the result panel. The second tab shows the results of the post-processing scores including (C) topology feature differences when considering extracted subgraphs without or with* PEPPER *expansions, (D) occurrence of proteins of the solutions in reference protein complexes and (E) enriched GO terms. The set of proteins of a given solution that co-occur in a particular reference protein complex or are annotated with a specific GO term, are highlighted clicking on the annotation of interest in the result panel. The network formed by adding known interactions between proteins of the seed list (F) and between the proteins predicted by* PEPPER *to form a complex (G) are visible at the end of the run. Green nodes are prey proteins, the squared purple node is the bait and hexagonal nodes are expansions predicted by* PEPPER *with a color code (light to dark red) increasing with significance. Green edges represent interactions between seed proteins whereas red edges connect expansion proteins added by* PEPPER*, both originating from the input PPI network.*

Section 4.2 (example of the clustering coefficient in figure 4.5C). For instance, degree and modularity values showed more than a two-fold increase. The *expansions* also slightly increased the overall subnetwork closeness centrality. Finally, the clustering coefficient was more than three times higher with than without *expansions*. Proteins added by PEPPER significantly affected the apparition of cliques (fully connected components) in the subnetwork, thus increasing the complex connectivity.

No protein complex was found to be associated with the original list of *seed* proteins. Matching PEPPER's predictions to the reference protein complexes resulted in 56 mapped known complexes when considering all solutions (figure 4.5D). Among these, 43 were found to overlap with at least one protein of the final solution. The best matching complex was the *URI complex (Unconventional prefolding RPB5 Interactor)* with a matching score of 0.1481, associated to two of the *seed* proteins and two of the *expansions*. While the interpretation necessitates further experiments to validate whether this complex is formed in the studied system, PEPPER provides directions for validation by prioritizing the candidates.

Among the significantly enriched cellular processes, several functions related to histones acetylation and methylation were found and respectively linked to *RUVBL1/2* (*expansions*) and *WDR92* (*bait*) proteins. Annotation of proteins and their cellular localization provided information about the possible localization of the complex in the nucleus, which is consistent with the putative association with the *URI complex* and the possible function in histone modifications (figure 4.5E).

Finally, *expansion* proteins were found with post-processing scores of 0.21 for *MAP3K3* and 0.42 and 0.44 respectively for *RUVBL2* and *RUVBL1* proteins. *RUVBL1* actually appears as a *prey* protein of *WDR92* with a very low number of unique peptide (only one) and a low SAINT score (0%) and therefore did not pass the detection threshold. Furthermore, AP-MS using both *RUVBL1* and *RUVBL2* also identified *WDR92* as a *prey*. Altogether, these results strongly suggest that *WDR92* forms a complex with both of these proteins predicted as *expansions* by PEPPER.

4.5 Discussion

The proposed method is a contribution to several aspects of science. First to graph theory, as an algorithm to extract dense subgraphs with a maximum number of nodes of interest. Second to bioinformatics, as a novel method for gene set analysis for which a highly connected functional network can be associated. And third to biology, as a pipeline to analyze proteomic data.

The use of a multi-objective genetic algorithm enabled us to solve the difficult problem of searching in a large space of possible solutions to extract several dominant solutions maximizing both the density and the coverage of the initial list of interest. Evolutionary algorithms have the main disadvantage of being random in nature and therefore providing

results with small variations when ran multiple times. PEPPER partially solves this problem by merging several solutions using an efficient and deterministic pruning algorithm.

The extensive evaluation of PEPPER on real protein complex datasets of Yeast and Human supports its ability to retrieve relevant biological structures in protein interaction networks. This is further exemplified by the analysis of the FGFR3 signaling pathway driving bladder cancer growth in the next chapter.

Further improvements of the algorithm includes the investigation of other objective functions to optimize the relevance of the solution to the network structure, for instance, the clustering coefficient may be more suitable than the density to identify complexes or pathways with high local densities. PEPPER and the multi-objective genetic algorithm it uses provide an excellent basis to solve other complex biological problems. Indeed, the existing objectives can be easily complemented by a third function to be maximized, for instance the number of alterations found in the proteins of the solution in order to identify highly altered and context-specific signaling pathways. The addition or replacement of the objective functions can potentially result in numerous other applications such as modeling entire signaling pathways including the transcriptional regulation step or integrating phospho-proteomic, genomic or transcriptomic data. All these potential applications only require the design of new appropriate objective functions.

Chapter 5

Joint proteomic and transcriptomic characterization of the FGFR3 signaling pathway driving bladder cancer

The project that I will describe in this chapter implicated numerous collaborators. The results were obtained in close collaboration with the platform for proteomics of the Institut Curie, directed by Damarys Loew. Cellular constructions and validations were done by Johannes Aubertin and Mélanie Mahé.

5.1 Introduction

Fibroblast growth factors receptors (FGFRs) are implicated in fundamental cellular processes such as proliferation, migration, differentiation, angiogenesis and wound healing (Korc and Friesel, 2009). Therefore, and as expected, the alteration of these receptor tyrosine kinase (RTK) leads to diseases. Activating germline mutations of FGFR3 results in dwarfism and severe skeletal disorders caused by an inhibition of bone cell growth (Naski et al., 1996). The same mutations and constitutive activation of FGFR3 in their somatic form drives specific human tumors and in particular is a major driver of bladder cancer (Cappellen et al., 1999) in which it is among the most frequently mutated genes (30% to 60%). The ambiguous role of FGFR3 in the regulation of cellular proliferation and its obvious dependency on cellular context is underlined by a recent description of *FGFR3* as a tumor suppressor gene in transformed epithelial cells (Lafitte et al., 2013). The ambiguity in the consequence of *FGFR3* mutations suggests a diversity of co-factors and effectors leading to opposing cell fates. Therefore, a thorough definition of its downstream signaling pathway is key to the understanding of its oncogenic function in cancers.

In order to determine the oncogenic pathway driven by *FGFR3* mutations in bladder cancer, we devised a novel approach based on a joint proteomic and transcriptomic

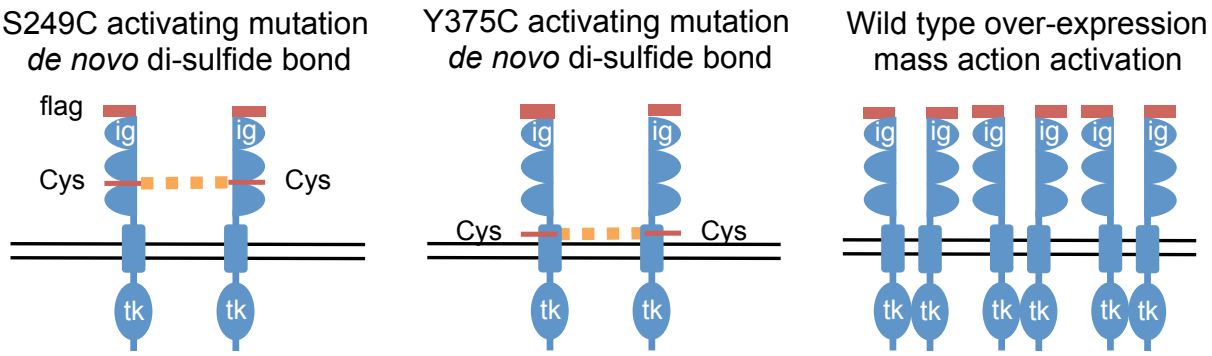
analysis. The idea is to resolve the entire pathway by identifying a protein-signaling pathway connected to a set of transcription factors that convey the signal to the nucleus. We first propose a model of the FGFR3 signaling pathway by extracting a set of FGFR3-associated proteins from a Immuno-Precipitation (IP) followed by Mass-Spectrometry (MS) experiment. The identified proteins are connected into a functional network using a reference protein interaction network and an algorithm for pathway extension, PEPPER (Winterhalter et al., 2014). Then, using an siRNA-mediated knockout of FGFR3, the genes responding the activation of FGFR3 are used to identify upstream regulators of the transcriptomic response signature. Finally, the transcription factors linking the FGFR3 proteomic and transcriptomic response are discussed and their role are further analyzed to understand the oncogenic function of FGFR3 in bladder cancer.

5.2 Deriving FGFR3-associated signaling proteins

In order to identify the downstream effectors of FGFR3 oncogenic signaling in bladder cancer, we established three model systems of FGFR3 constitutive activation. A tagging FLAG sequence was introduced in the extracellular N-terminal domain of the FGFR3 coding sequence to avoid interference with the intracellular binding partners. These FGFR3-tagged clones were stably expressed in the FGFR3-dependent bladder cancer cell lines RT112 (Qing et al., 2009). Three types of coding sequences were transfected and stably expressed as illustrated in figure 5.1. First, an isoform for which the serine at position 249 is replaced by a cysteine, which is thought to constitute a di-sulfide bond, thereby constitutively forming an active $FGFR3^{S249C}$ homodimer. Second, an isoform for which the tyrosine at position 375 is replaced by a cysteine, which is though to have the same effect on the activity of the receptor. Finally, the wild type isoform (often noted FGFR3-II or FGFR3-b) was transfected and expressed in high exogenous levels, which when resulting in high protein level is thought to form homodimers by effect of mass-action.

Immunoprecipitates of the FLAG-FGFR3 protein using anti-FLAG antibodies were analyzed by Mass-Spectrometry (MS). Six replicates of each construct (2 removed for the wild-type construction after quality check), as well as six control clones with non-FGFR3 coding transfections, were analyzed by MS. In all, 2,315 proteins for which at least one unique peptide was correctly mapped were identified. Any proteins for which one peptide was found in any of the control experiments were removed resulting in a set of 1,093 putative partners. The top proteins of each experiment are listed in table 5.1.

We first tested whether the three constructs identified different proteins, that is, whether various activation of FGFR3 resulted in different signaling protein partners. Firstly, no protein had a significantly different distribution of number of peptides between each set of experiment (kruskal-wallis, FDR 10%). Therefore, no proteins could be identified as a predictor of the type of FGFR3 transfected. Second, the jaccard coefficient of similarity ($jaccard(a, b) = \frac{a \cap b}{a \cup b}$) was computed between each pairs of experiment to



Co-precipitated proteins identified by Mass-Spectrometry

S249C		Y375C		WT	
Protein	peptides	Protein	peptides	Protein	peptides
FGFR3	51	FGFR3	46	TUBB3	40
TUBB3	23	TUBB3	46	FGFR3	38
CDC37	13	TUB4A	29	HNRNPH2	27
LRP1	9	HNRNPH2	23	TUB1A1	21

Figure 5.1: *The three FGFR3-tag constructions analyzed using Mass-Spectrometry. Top. Three tagged (red) FGFR3 constructions were stably expressed in a bladder cancer cell line. Rough positions of the introduced mutations in two first constructions (S249C and Y375C) are shown by a red line. Putative di-sulfide bonds are shown by a thick yellow dashed line. The plasmid membrane is shown by two black parallel lines. ig: immuno-globulin domain; tk: tyrosine kinase domain.*
Bottom. Top co-precipitated proteins identified by Mass-Spectrometry in each construction. Showing as an example the 4 proteins to which the most unique peptides were matched.

measure the similarity in terms of protein content (all proteins with at least 3 unique peptides). Figure 5.2 shows the distribution of the similarity of the set of identified proteins among each experimental setting and between all pairs of single replicate experiments. The sets of identified proteins are as similar (or dissimilar) among each type of clones (intra-experiments) than between each experiment independently of the type of FGFR3 expressed (inter-experiments). As no difference were observed between each constructions, these were pooled and the final list of co-precipitated proteins was defined as the set of proteins for which at least three peptides were found in at least two experiments. This resulted in a set of 60 high confidence proteins, hereafter referred to as the core proteins.

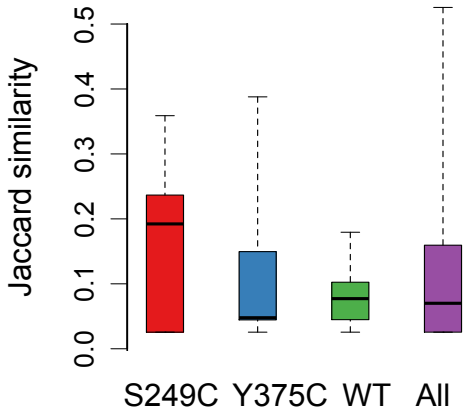


Figure 5.2: *Distribution of jaccard coefficients between each MS replicate. Jaccard coefficients were computed between all pairs of experimentally identified sets of FGFR3-co-precipitated proteins (6 S249C, 6 Y375C and 4 WT). The distribution of the jaccard coefficients are shown either for each pairs of experiments of the same construction (noted: S249C, Y375C or WT) or for all pairs of experiments.*

5.3 Protein interaction-based FGFR3 signaling pathway expansion

To analyze the set of identified proteins in a comprehensive way, the previously reported physical interactions between these proteins were retrieved from a high confidence database, HIPPIE (Schaefer et al., 2012). Two proteins only were already known to interact with FGFR3, the solute carrier SLC25A6 identified in a large-scale proteomic screen and the co-chaperone CDC37, which has been described as necessary to the stability and function of FGFR3 (Laederich et al., 2011). In all, 20 interactions between the set of 60 high confidence proteins were previously reported (see figure 5.3). Notably, interactions between and with tubulin subunits (TUBA1A, TUBB6 and TUBB3), between proteins involved in endocytosis (TFRC and RAB5B) with a potential role in transport of the receptor and interactions between subunits of the SWI/SNF family of chromatin remodelers (SMARCA2 and SMARCA4) suggesting FGFR3 to eventually have an impact on the expression of downstream genes. Overall, the small number of interaction found between the identified proteins suggests that the list of partners is incomplete as it is often the case in proteomic MS experiments (Gingras et al., 2007).

In order to obtain a comprehensive FGFR3 signaling pathway, we applied PEPPER (Winterhalter et al., 2014) on the list of 60 high-confidence core proteins. PEPPER aims at constructing a signaling pathway by densely connecting a set of seed proteins in a large-scale protein interaction network. This resulted in a small network of 136 proteins containing 57 of the core proteins and 1,738 referenced interactions. Interestingly, the 79

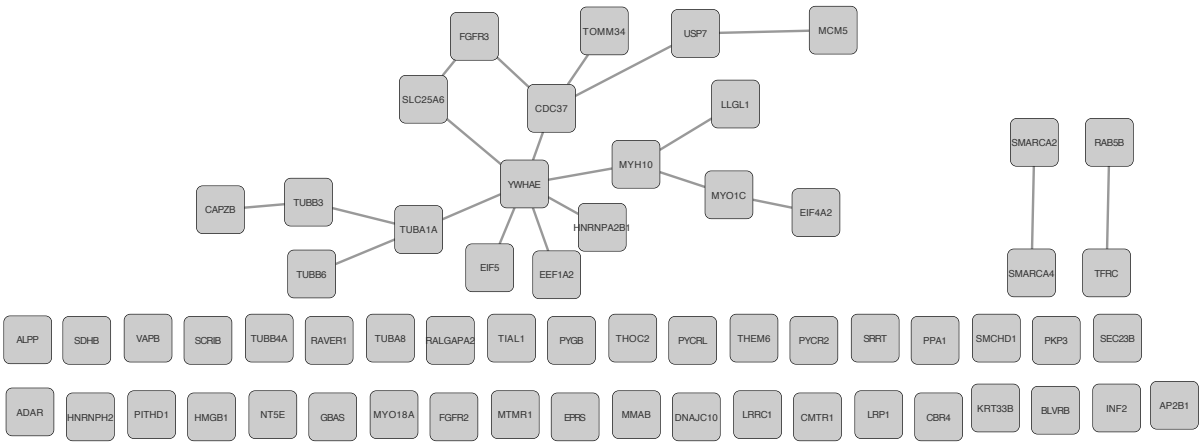


Figure 5.3: *Interaction previously reported between the 60 high confidence FGFR3 co-precipitated proteins. The interactions originate from the HIPPIE database of curated protein-protein interactions.*

added proteins significantly overlapped with the proteins identified in the MS experiments (38 overlapping with all proteins with at least one peptide, fisher's exact test $< 1\%$; 12 overlapping with all proteins with at least one peptide and none found in any control experiments, fisher's exact test $< 1\%$). The pathway extracted using PEPPER is shown in figure 5.4 with proteins grouped by general cellular functions each of which contains at least a high-confidence protein (square). The two most significantly represented functions are related to general Receptor Tyrosine Kinase (RTK) signaling pathways such as the Mitogen Activated Protein Kinase pathway (KEGG, fdr $< 10^{-7}$), to the cell cycle (KEGG, fdr $< 10^{-4}$) and to the regulation of apotosis (GO, fdr $< 10^{-10}$). The identification of proteins involved in translation, post-translational modification and transport are most probably a reflection of the route of FGFR3 from its synthesis to its activity as a growth factor receptor. Interestingly, the presence of general and specific transcription factors, co-factors and chromatin remodelers, suggest that FGFR3 controls a signaling pathway that eventually leads to a gene-specific transcriptional control.

Constitutive activating mutation of *FGFR3* is a major driver of cell survival and proliferation in bladder cancer cell lines. Therefore, we can assume that a large portion of the proteins in the downstream pathway should have an impact on the cell viability and proliferation. To determine the functional relevance of the FGFR3 signaling pathway, we used the Achilles (Cheung et al., 2011) in vitro short hairpin RNA (shRNA) screen for cancer gene vulnerability, which measures the impact of down-regulating 11,194 genes on the survival of cancer cell lines and in particular of the RT112 bladder cancer cell line. The impact of the entire FGFR3 signaling pathway on the survival of RT112 was significantly higher (Student's test, $< 10^{-5}$) than the controls. The protein interaction network in figure 5.4 shows the mean vulnerability score of all the proteins for which the

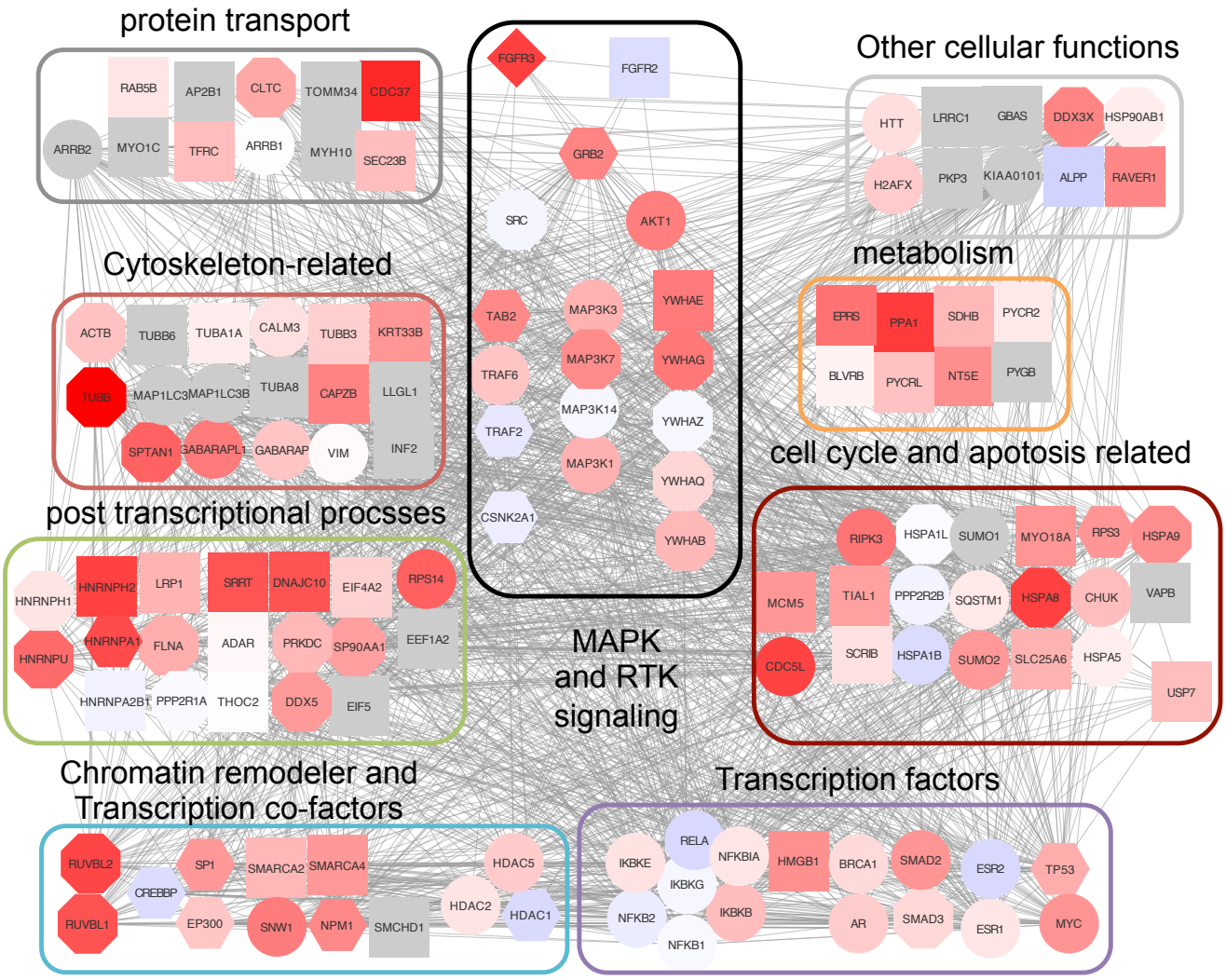


Figure 5.4: *FGFR3 signaling pathway. Interaction network of the proteins predicted to be part of the FGFR3 signaling pathway. Proteins are grouped by cellular functions based on pathway and ontology annotations. Nodes are shaped depending on their identification in the MS experiment and colored by their impact on cell survival and viability (colored using a red to blue gradient denoting red for high impact and blue low, gray no data). Square nodes (and a diamond for FGFR3) are high confidence proteins for which at least 3 unique peptides were found in at least two experiments and none in the control experiments. Hexagons and octagons have a few peptides in the FGFR3 experiment while only octagons also have peptides in the control experiments. Finally, round nodes are proteins for which no peptides were identified in the FGFR3 experiment.*

coding gene was tested in the screen.

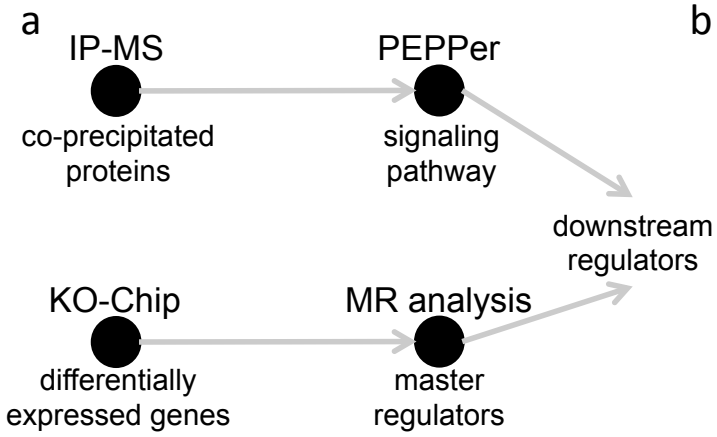
5.4 Master regulators of the FGFR3 signaling pathway

The activation of growth factor receptors and its downstream signaling pathway eventually impacts cellular behavior by regulating the transcriptional rate of specific genes. In order to identify the final effectors of signal transduction, we then aimed at identify the transcription factors responsible for the response to the activation of FGFR3 in bladder cancer cell lines. The strategy we employed to link the FGFR3 signaling pathway with the transcriptional response to the pathway activation is described in figure 5.5a.

In essence, we derived a response signature of genes that have a different transcriptional rate following an abrogation of the FGFR3 signaling using siRNA transient gene knockout. Then, we searched for transcription factors (TF) potentially regulating these genes. Regulons, defined as the set genes specifically regulated by a transcription factor, were derived from public ChIP-seq and ChIP-on-chip datasets from the ENCODE project (Gerstein et al., 2012) and the CHEA2 repository (Kou et al., 2013). The predicted upstream TF were more than five-fold enriched in the set of TF identified in the FGFR3 pathway (fisher's exact test $p = 0.001719$) with 10 overlaping TF. Figure 5.5b lists the transcription factors that were predicted to be part both of the signaling pathway based on the proteomic experiment and to be downstream regulators of FGFR3 signaling based on transcriptomic experiments. The NF-κB regulator complex was previously described as a downstream regulator of FGFR3 (Salazar et al., 2014). Moreover, MYC and FGFR3 were identified as cooperating oncogenes (Zingone et al., 2010)

To further validate the list of putative downstream regulators, a public dataset of FGFR3 shRNA silencing following transcriptomic profiling was used to derive a replicate response gene signature. The estrogen nuclear receptors ESR1 and ESR2 were identified as significant regulators of both gene signatures and were predicted to be part of the signaling pathway from the MS experiment. While few biological experiment has directly proven the estrogen receptor to be driving bladder cancer, major studies suggested its implication in a subtype of luminal-like bladder cancer strongly resembling the luminal subtype of estrogen-driven breast cancer (Cancer Genome Atlas Network, 2014; Choi et al., 2014). Moreover, the FGFR3 homologous receptor FGFR2 has been recently implicated as a co-factor of the estrogen pathway in luminal breast cancer and is among the high-confidence partners of FGFR3 in our results (Fletcher et al., 2013). This is further substantiated by the small, although significant, increase of the phosphorylation of ESR1 at serine 118 (relative to its transcriptional activation, Hasbi et al., 2004) in bladder cancer samples with an *FGFR3* mutation as presented in figure 5.6.

SMAD3, a downstream effector of the TGFβ signaling pathway, is predicted to be a regulator of the transcriptional response by both knockout experiments. Moreover, the prediction of SMAD3 as part of the signaling pathway is supported by the identification of 24 corresponding unique peptides in the 16 FGFR3 IP experiments and only 4 in the



b

	siRNA	shRNA	MS-1	MS-2
TP53	✔	✔	✔	✔
SMARCA4	✔	✔	✔	✔
EP300	✔		✔	✔
SMAD3	✔	✔		✔
SMAD2	✔	✔		
ESR1	✔	✔		
ESR2	✔	✔		
MYC	✔			
NFKB1	✔			
HDAC2	✔			

MS-1: Pulled IP-MS proteins
MS-2: Pulled IP-MS proteins also identified in control proteins

Figure 5.5: *Joining FGFR3 proteomic and transcriptomic analysis at the level of transcriptional regulators.*
a. FGFR3 Immuno-Precipitation (IP) followed by Mass-Spectrometry (MS) analysis identified a set of proteins of the FGFR3 signaling pathway extended using the PEPPer algorithm. FGFR3 knockout followed by transcriptomic chip profiling identified sets of genes responding to the silencing, hence the activation, of FGFR3. Master Regulator (MR) analysis determined a set of putative upstream regulators of these genes.
b. The set of regulators identified by IPMS is significantly enriched in the set of master regulators identified by the transcriptomic analysis (fisher's exact test $\alpha = 1\%$) . siRNA and shRNA, significant regulators (fdr $< 1\%$) upstream of the gene response signature obtained by siRNA and shRNA silencing. A tick is added in the table if a corresponding transcriptional regulator is identified in the column-specified experiment. For siRNA and shRNA experiments, a tick corresponds to a significant overlap between the targets (regulon) of a given TF and the set of genes in the corresponding FGFR3 response signature, independently of whether the TF is activated or repressed by FGFR3-depletion.

control experiments. The relation between TGFβ and FGFR3 has been superficially studied in chondrocytes in which cross-talk between the two signaling pathways was suggested. Interestingly, TGFβ was shown to be required for the proliferation of normal urothelial cells presenting a terminal differentiation phenotype (Fleming et al., 2012), the untransformed counterpart of the frequently FGFR3 mutated luminal-like bladder cancers. Moreover, bladder cancer samples with an *FGFR3* mutation have significantly higher levels of SMAD3 proteins (see figure 5.7).

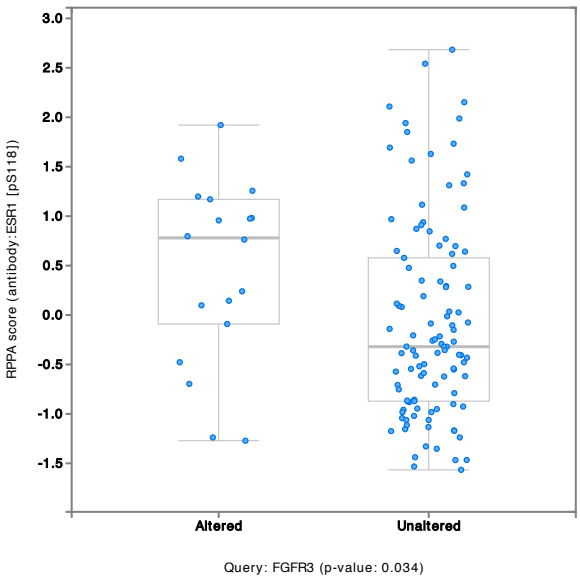


Figure 5.6: *ESR1 phosphorylation in FGFR3 wild type and altered samples. TCGA Reverse Phase Protein Array quantification of the phosphorylation of ESR1 at Serine 118 depending on the genetic status of FGF3 in 244 bladder cancer samples. Alteration of FGFR3 are here point mutations or gene fusion.*

SMARCA4, a chromatin remodeler, and TP53 were experimentally identified as proteins involved in the signaling pathway of FGFR3 as well as significant regulators of its downstream transcriptional response based on independent knockout and transcriptomic experiments (see figure 5.5b). While it is difficult to comprehend the role of epigenetic regulators, the implication of the key tumor suppressor TP53 in FGFR3's signaling is of major functional relevance.

5.5 Regulation of TP53 by the FGFR3 signaling pathway

As a key regulator of DNA-damage induced apoptosis, *TP53* is altered and silenced in most cancers and however infrequent in FGFR3 mutated tumors. The co-precipitation of TP53 with FGFR3 indicates their implication in the same pathway and in particular the potential regulation of TP53 by FGFR3.

In order to quantify the impact of FGFR3 on the activity of TP53, we tested its transcriptional activity following the knockout. We selected refined sets of TP53 activated and repressed targets from a previous study (Mirza et al., 2003) and deduced an increase transcriptional activity of TP53 following FGFR3 silencing by a general decreased

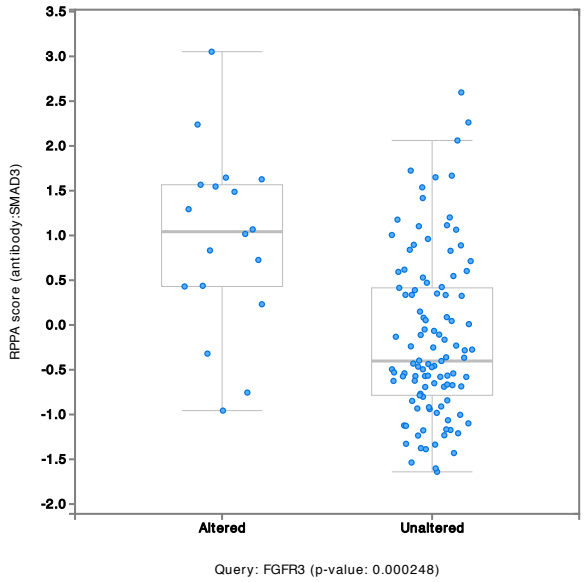


Figure 5.7: *SMAD3 protein expression in FGFR3 wild type and altered samples. TCGA Reverse Phase Protein Array quantification of the SMAD protein depending on the genetic status of FGF3 in 244 bladder cancer samples. Alteration of FGFR3 are here point mutations or gene fusion.*

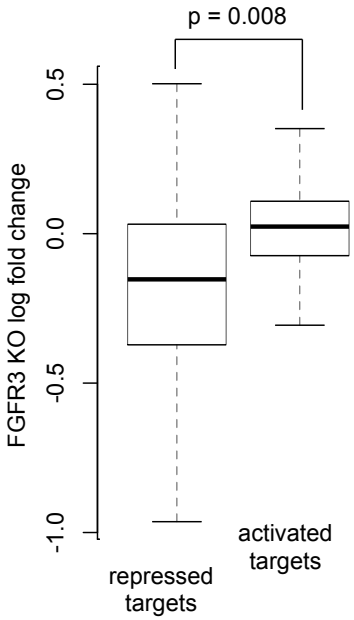


Figure 5.8: *Expression of TP53 targets following FGFR3 knockout. Mean fold-change of genes repressed and activated by TP53 following the siRNA knockout of FGFR3.*

expression of its repressed target genes and a mild over-expression of the genes activated by TP53 (figure 5.8). Therefore, the inhibition of FGFR3 results in an increased TP53 transcriptional activity revealing a regulatory role of FGFR3 on TP53.

The identification of several TP53 peptides following FGFR3 precipitation indicates the possibility of a physical interaction between these two proteins. To verify this physical interaction, we immunoprecipitated the endogenous FGFR3 in two FGFR3-dependent bladder cancer cell lines and tested the interaction western-blot. As indicated by the small number of peptides identified in the IP-MS experiment, this low-throughput validation revealed no direct interaction between FGFR3 and TP53 (figure 5.9). However, the ubiquitin-specific-processing protease 7, USP7, a thoroughly described regulator of TP53 (Epping et al., 2010; Sarkari, Sheng, and Frappier, 2009), was identified as a high-confidence FGFR3-associated protein in the mass spectrometry screen. The direct and physical FGFR3-USP7 interaction was confirmed by the IP-western-blot experiment (figure 5.9) implying that the regulation of TP53 by FGFR3 is potentially mediated by UPS7.

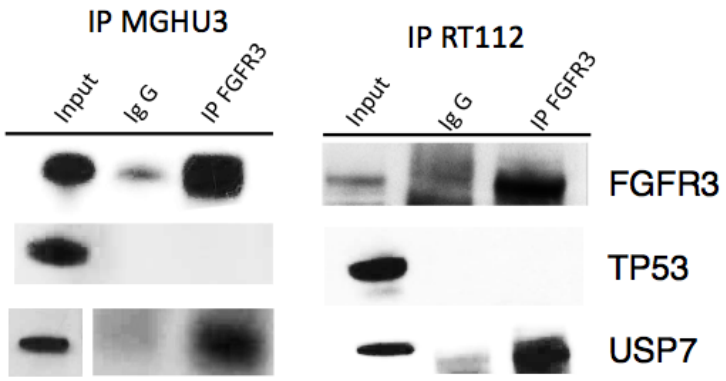


Figure 5.9: *Western-blot analysis of the FGFR3-USP7 interaction. Endogenous FGFR3 was immunoprecipitated from MGHU3 and RT112 cells with an antibody against FGFR3 and the western-blot was stained with anti-FGFR3, anti-TP53 and anti-USP7 antibodies.*

5.6 Discussion

The results show the benefit and relevance of integrating several large-scale datasets from proteomic, transcriptomic and functional (shRNA screen) experiments. The integration of massive and seemingly unrelated datasets revealed significantly overlapping results in a common model of a driving signaling pathway. This work provides an example of a procedure to effectively integrate large-scale profiles using the underlying biological structure from which all the datasets are indirectly imputable to. Moreover, the construction of a signaling pathway provides a rationalized approach to the

discovery of new therapies. For instance, the extensive identification of Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein (YWHAx) in the protein-based signaling pathway and the strong impact of their silencing on the survival of a related cell line suggest that these proteins, and in particular YWHAE, are potentially effective targets. Therefore, the use of Difopein, a peptide inhibitor of YWHA proteins, is a potential drug to be used in FGFR3-dependent bladder cancers.

Through the integration of the proteomic identification of protein partners of FGFR3 and the transcriptomic impact of the activation of FGFR3, this work results in a first model of the entire FGFR3 signaling pathway in bladder cancer. The proposed pathway provides new leads to understand the carcinogenic role of the ambiguous growth factor receptor FGFR3. In particular, the results implicate ESR1 as well as the TGFβ pathway in the oncogenic role of FGFR3. The presented experimental validation shows the implication of a long acknowledged regulator of TP53, USP7, directly interacting with FGFR3. Supplementary experiments are scheduled at the Institut Curie to further understand the role of FGFR3 in the regulation of TP53. First investigations will aim at measuring the effect of the inhibition of FGFR3 on: the sub-cellular localization of TP53, the regulation of TP53 target genes and the activation of apoptosis.

5.7 Material and methods

FGFR3 co-precipitated proteins

The FGFR3b isoform was epitope-tagged with a DYKDDDDK (FLAG^{TM})-sequence inserted after the 27th amino acids. The FLAG-FGFR3 was inserted into a pIRESpuro3 vector under a CMV-promoter and site directed mutagenesis was used to obtain the S249C and Y375C mutated forms. Following immunoprecipitation with anti-FLAG antibodies coupled to agarose beads, elution was performed using high concentrations of FLAG-peptides, migrated on 1D-acrylamide gel which was digested and from which pepitdes were extracted and analyzed using LC-MS/MS. Spectrums and peptides were identified using the MASCOT (Matrix Science) software aligning on the uniprot database (peptide FDR 0.1%). Each construct was analyzed in 6 replicates. Two replicates of the wild-type construction lacked the identification of any FGFR3 peptide and therefore were removed.

In all, 2,014 proteins were identified from which 921 putative contaminant proteins with at least one peptide identified in a control experiment were removed resulting in 1,093 proteins. 60 Proteins with at least three peptides in at least two replicates (of any construction) were considered as high-confidence protein partners of FGFR3. FGFR3 and two other high confidence proteins were confirmed by Western Blotting (see figure 5.10).

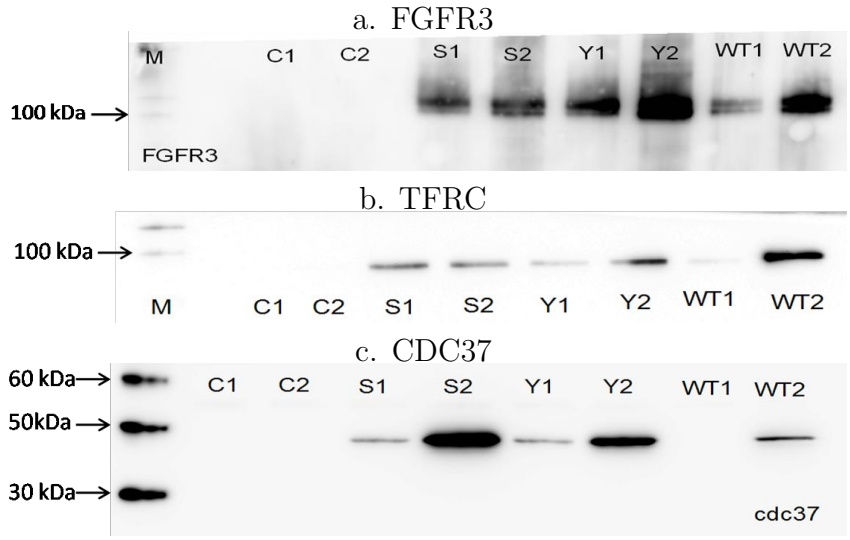


Figure 5.10: *Western blot validation of Mass Spectrometry results. a. Anti-FGFR3 western blot of anti-FLAG immunoprecipitates. In both controls (C1 and C2), no FGFR3 was detected. In contrast, in all immunoprecipitates from FLAG-FGFR3b expressing RT-112 cells, FGFR3 was detected. b. Anti-TFRC western blot of anti-FLAG immunoprecipitates. In both controls (C1 and C2), no TFRC (95 kDa) was detected. In contrast, in all immunoprecipitates from FLAG-FGFR3b expressing RT-112 cells, TFRC was detected. c. Anti-CDC37 western blot of anti-FLAG immunoprecipitates. In both controls (C1 and C2), no CDC37 (44 kDa) was detected. In contrast, in all immunoprecipitates from FLAG-FGFR3b expressing RT-112 cells, except WT1, CDC37 was detected. C1 and C2 = controls (empty pIRESpuro3); S1 and S2 = FLAG-FGFR3b S249C; Y1 and Y2 = FLAG- FGFR3b Y375C; WT 1 and 2: FLAG-FGFR3b wild type.*

Pathway extension using PEPPer

The 60 high-confidence proteins were used with the latest highly curated HIPPIE protein interaction network in the PEPPer cytoscape application to extend the FGFR3 signaling pathway. Default parameters were used and to prevent any loss of information the non-refined sub-network was used as the predicted pathway (called union in the application). The presented sub-network was reformatted to have nodes representing proteins to be shaped depending on their identification in the MS experiment (no peptides precipitated with FGFR3: round; FGFR3 diamond, high-confidence: squares; no peptides in control: hexagon; peptides in control: octagon).

Gene title for the RT112 shRNA Achilles screen was downloaded from the broad data portal. For each gene, the values of all targeting shRNA was averaged and used to color nodes in the sub-network. The relevance of the entire proposed pathway was tested by comparing the distribution of the gene title values of all the genes in the sub-network to the values of the shRNA targeting control genes (RFP, Lac2, GFP and Luciferase coding genes).

Master regulator analysis

Three replicate siRNA knockout experiments were profiled using affymetrix Human exon 1 arrays along with 5 control lipofectamine-only experiments. The signature of FGFR3 responsive genes was obtained by selecting differentially expressed genes using the limma bioconductor package (FDR : 5%) (Smyth, 2005).

TF regulons were derived from two repositories of human and mouse ChIP-seq and ChIP-on-chip data. The ENCODE ChIP-seq data was recovered from the UCSC genome browser (Human hg19 February 2009 genome assembly) by selecting all narrow ChIP-seq peak (ENCODE chip V3) within -5000 bp to 2000 bp around a Transcription Start Site of a gene with a non-null Human genome organization Gene Nomenclature Committee (HGNC, genenames.org) symbol. Additional ChIP-seq or ChIP-on-chip data was directly downloaded from the ChEA2 database (amp.pharm.mssm.edu/ChEA2).

Upstream TF were identified by testing the enrichment of all TF regulons in genes differentially expressed following FGFR3 knockdown using Fisher's exact test and corrected for multiple hypothesis testing (keeping FDR at 1%).

Conclusion

This manuscript describes the analysis I carried out during my PhD to uncover the pathways and networks driving the initiation and progression of bladder tumors. This work fundamentally relies on a general model of signaling pathways in which signals are transduced from receptors to the nucleus through a complex interplay of protein interactions and results in context-dependent control of the expression of genes. The overall strategy was to first focus on the latter well-studied step of gene regulation and then to move up to the computationally less explored signal transduction level.

By taking advantage of both the high availability of transcriptomic experiments and the numerous developments in regulatory network inference, I was able to develop and improve current algorithms to characterize driver transcriptional programs. However, despite the ample interest to this field, my major concern was the low reliability of each of the predicted regulatory interactions when taken separately. Therefore, I focused on developing approaches that can cope with this local uncertainty by taking advantage of higher-level information.

The lack of reference algorithmic systems to explore the space of signal transduction pathways required the design of entirely novel methods to analyze proteomic experiments. The use of evolutionary computations allowed me to easily implement new search algorithms with simply designed objectives. Although the proposed method is clearly shown to be effective by its direct application to bladder cancer, this approach is at its infancy and leaves plenty of room for improvements, hopefully for efficient deterministic algorithms.

The set of proposed algorithms was used to delineate pathways controlling malignant proliferation and differentiation in bladder cancers as well as those of normal urothelial cells. Firstly, this illustrate the usability and benefit of the methods developed during this work. Second, it allowed the discovery of new cancer genes and to improve our understanding and knowledge of the signaling pathway downstream of formerly known or recently discovered oncogenes. In particular, the biological results recurrently point towards a puzzling conclusion that differentiation-dependent pathways and transcriptional programs drive a specific type of bladder cancer.

This work resulted in three main contributions to computational biology and cancer biology:

• **Providing tools to identify and analyze cancer-driving networks.** I developed several algorithms for network reconstruction and analysis, which were embedded in two software packages. CoRegNet is a Bioconductor package to infer gene regulatory network, extract transcriptional programs from them, integrate external regulatory data, estimate transcription factor activity and to visualize the initial and produced knowledge in a single data/network visualization tool. Pepper is a Cytoscape application to extract the densely connected interaction network that is the most relevant to a particular set of proteins of interests, finding its application in both identifying protein complexes and signaling pathways.

• **Highlight new pathways and producing information on known pathways driving specific subtypes of bladder cancer.** The use of the aforementioned tools

on bladder cancer datasets - genomic, transcriptomic and proteomic profiles - resulted in the discovery of new bladder cancer driver genes responsible for the activation of subtype-specific proliferation-related transcriptional programs. Moreover, these tools and strategies further specified the signaling pathways and generally the downstream effectors and carcinogenic functions of bladder cancer oncogenes.

• **Presenting evidences that tumors use very similar circuits than the one found in normal cells.** Based on the analysis of the regulatory circuits contributing to normal and malignant proliferation and differentiation as well as the means by which these are blocked in such states, my results support the parallel between tumors and normal regenerative processes at the level of cellular regulatory circuits.

Among the most promising outcomes of cancer genomics and its related molecular profiling, is the definition of diagnostic and prognostic tools in the form of predictive genetic, transcriptional or protein (expression or post-translational level) markers. Even more compelling by its direct clinical benefit, systematic profiling hopes to identify new pairs of therapeutic targets and companion diagnostics, which aims at predicting the response to a particular therapy and is sometimes called theragnosis.

With approximately 20,000 coding genes, 250,000 to one million proteins and now more than 100 million potential Single Nucleotide Polymorphisms reported in dbSNP (ncbi.nlm.nih.gov/SNP/), the ultra-high dimensionality of human molecular profiles impugns the capacity of classical statistical approaches until the number of profiled tumors exceeds the number of measured signals. Until then, pragmatic approaches are required both to design large-scale experiments and to analyze these datasets in an integrative and rational way regarding the underlying biological process. For instance, while searching for a biomarker of the activation of a particular pathway, its seems more likely to reproducibly identify an increase in mRNA level of a gene that is actually regulated by the studied pathway than simply the gene with the best statistics in a given dataset. The complete definition of sample-specific driving pathways has huge clinical applicative potential. The identification of a driving genetic event can point towards all the downstream effectors as potential biomarkers, targets for drug as well as therapeutic pitfalls as exemplified by the ineffective anti-EGFR treatment in RAS mutated tumors.

This work illustrates the necessity of developing computational approaches that bear in mind the underlying mechanistic of the studied biological processes. As a closing remark, I also hope it will form an algorithmic basis for, or at least persuade of the benefit of, joining proteomic, transcriptomic and genomic analysis.

Bibliography

Agrawal, R, T Imieliński, and A Swami (1993). "Mining association rules between sets of items in large databases". In: *ACM SIGMOD Record* 22.2, pp. 207–216. (Cited on page 81).

Aittokallio, T (May 2006). "Graph-based methods for analysing networks in cell biology". In: *Briefings in Bioinformatics* 7.3, pp. 243–255. (Cited on page 138).

Akavia, U. D. et al. (Dec. 2010). "An Integrated Approach to Uncover Drivers of Cancer". In: *Cell* 143.6, pp. 1005–1017. (Cited on page 74).

Alimirah, F. et al. (Nov. 2012). "Crosstalk between the peroxisome proliferator-activated receptor". In: *Experimental Cell Research* 318.19, pp. 2490–2497. (Cited on page 13).

Aytes, A. et al. (May 2014). "Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy". In: *Cancer Cell* 25.5, pp. 638–651. (Cited on pages 74, 86).

Bader, G. and C. Hogue (2003). "An automated method for finding molecular complexes in large protein interaction networks". In: *BMC Bioinformatics* 4.1, p. 2. (Cited on pages 132, 133, 139, 141).

Bahar Halpern, K, T Vana, and M. D. Walker (July 2014). "Paradoxical role of DNA methylation in activation of FoxA2 gene expression during endoderm development". In: *Journal of Biological Chemistry*. (Cited on page 46).

Bakkar, A. A. et al. (Dec. 2003). "FGFR3 and TP53 gene mutations define two distinct pathways in urothelial cell carcinoma of the bladder." In: *Cancer Research* 63.23, pp. 8108–8112. (Cited on pages 34, 35).

Bandyopadhyay, S. et al. (Dec. 2010). "Rewiring of genetic networks in response to DNA damage." In: *Science (New York, N.Y.)* 330.6009, pp. 1385–1389. (Cited on page 80).

Barrientos, S. et al. (Aug. 2008). "Growth factors and cytokines in wound healing." In: *Wound Repair And Regeneration* 16.5, pp. 585–601. (Cited on page 6).

Bashashati, A. et al. (Dec. 2012). "DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer". In: *Genome Biology* 13.12, R124. (Cited on page 68).

Behrends, C. et al. (Jan. 2010). "Network organization of the human autophagy system". In: *Nature* 466.7302, pp. 68–76. (Cited on pages 58, 80, 136, 140).

Bell, S. M. et al. (Oct. 2011). "Kruppel-like factor 5 is required for formation and differentiation of the bladder urothelium". In: *Developmental Biology* 358.1, pp. 79–90. (Cited on page 125).

Belton, J.-M. et al. (Nov. 2012). "Hi-C: A comprehensive technique to capture the conformation of genomes". In: *Methods* 58.3, pp. 268–276. (Cited on page 47).

Bernard-Pierrot, I (Oct. 2005). "Oncogenic properties of the mutated forms of fibroblast growth factor receptor 3b". In: *Carcinogenesis* 27.4, pp. 740–747. (Cited on page 36).

Berry, M, D Metzger, and P Chambon (Sept. 1990). "Role of the two activating domains of the oestrogen receptor in the cell-type and promoter-context dependent agonistic activity of the anti-oestrogen 4-hydroxytamoxifen." In: *The EMBO Journal* 9.9, pp. 2811–2818. (Cited on page 13).

Bild, A. H. et al. (Jan. 2006). "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." In: *Nature* 439.7074, pp. 353–357. (Cited on page 101).

Birder, L. and K.-E. Andersson (2013). "Urothelial signaling". In: *Physiological reviews* 93.2, pp. 653–680. (Cited on page 32).

Biton, A. et al. (Nov. 2014). "Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes". In: *Cell Reports* in press. (Cited on pages 107, 112).

Böck, M. et al. (Feb. 2014). "Identification of ELF3 as an early transcriptional regulator of human urothelium". In: *Developmental Biology* 386.2, pp. 321–330. (Cited on pages 92, 124).

Botteman, M. F. et al. (2003). "The health economics of bladder cancer: a comprehensive review of the published literature." In: *PharmacoEconomics* 21.18, pp. 1315–1330. (Cited on page 32).

Boulesteix, A.-L. and K. Strimmer (2005). "Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach." In: *Theoretical biology & medical modelling* 2, p. 23. (Cited on pages 68, 91).

Breiman, L. (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32. (Cited on pages 72, 86).

Buenrostro, J. D. et al. (Oct. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10.12, pp. 1213–1218. (Cited on page 47).

Butte, A. J. and I. S. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". In: *Pacific Symposium on Biocomputing* 5, pp. 418–429. (Cited on page 71).

Cancer Genome Atlas Network (Oct. 2012). "Comprehensive molecular portraits of human breast tumours." In: *Nature* 490.7418, pp. 61–70. (Cited on pages 51, 69).

— (Jan. 2014). "Comprehensive molecular characterization of urothelial bladder carcinoma". In: *Nature*, pp. 1–8. (Cited on pages 24, 26, 35, 36, 69, 103, 104, 118, 120, 124, 127, 155).

Cappellen, D et al. (Sept. 1999). "Frequent activating mutations of FGFR3 in human bladder and cervix carcinomas." In: *Nature Genetics* 23.1, pp. 18–20. (Cited on pages 36, 127, 149).

Cappellen, D et al. (2006). *Fibroblast growth factor receptor 3; measuring for the presence of a carcinoma-associated FGFR3 point mutations in cellular samples of the bladder or cervix*. Tech. rep. US7135311 B1. US Patent Office. (Cited on page 36).

Carter, S. L. et al. (Sept. 2004). "Gene co-expression network topology provides a framework for molecular characterization of cellular state". In: *Bioinformatics* 20.14, pp. 2242–2250. (Cited on page 71).

Chandra, A et al. (July 2013). "Epidermal Growth Factor Receptor (EGFR) Signaling Promotes Proliferation and Survival in Osteoprogenitors by Increasing Early Growth Response 2 (EGR2) Expression". In: *Journal of Biological Chemistry* 288.28, pp. 20488–20498. (Cited on page 104).

Chang, H. Y. et al. (2004). "Gene Expression Signature of Fibroblast Serum Response Predicts Human Cancer Progression: Similarities between Tumors and Wounds". In: *PLoS Biology* 2.2, e7. (Cited on page 49).

Charboneau, L. et al. (2002). "Utility of reverse phase protein arrays: applications to signalling pathways and human body arrays". In: *Briefings in functional genomics & proteomics* 1.3, pp. 305–315. (Cited on page 52).

Chatr-aryamontri, A et al. (Jan. 2007). "MINT: the Molecular INTeraction database". In: *Nucleic Acids Research* 35.Database, pp. D572–D574. (Cited on page 59).

Chatr-aryamontri, A et al. (Dec. 2012). "The BioGRID interaction database: 2013 update". In: *Nucleic Acids Research* 41.D1, pp. D816–D823. (Cited on page 58).

Chebil, I et al. (2014). "Hybrid Method Inference for the Construction of Cooperative Regulatory Network in Human." In: *IEEE transactions on nanobioscience*. (Cited on pages 77, 81, 84, 86, 87).

Chen, G. A. et al. (Apr. 2002). "Discordant protein and mRNA expression in lung adenocarcinomas". In: *Molecular & cellular proteomics : MCP* 1.4, pp. 304–313. (Cited on page 65).

Cheung, H. W. et al. (2011). "Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer". In: *Proceedings of the national academy of sciences of the United States of America* 108.30, pp. 12372–12377. (Cited on page 153).

Chiang, D. Y. et al. (Nov. 2008). "High-resolution mapping of copy-number alterations with massively parallel sequencing". In: *Nature Methods* 6.1, pp. 99–103. (Cited on page 46).

Chiquet, J et al. (Jan. 2009). "SIMoNe: Statistical Inference for MOdular NEtworks". In: *Bioinformatics* 25.3, pp. 417–418. (Cited on page 72).

Choi, H. et al. (Dec. 2010). "SAINT: probabilistic scoring of affinity purification–mass spectrometry data". In: *Nature Methods* 8.1, pp. 70–73. (Cited on pages 133, 143, 144).

Choi, W. et al. (Feb. 2014). "Identification of Distinct Basal and Luminal Subtypes of Muscle-Invasive Bladder Cancer with Different Sensitivities to Frontline Chemotherapy". In: *Cancer Cell* 25.2, pp. 152–165. (Cited on pages 35–37, 103, 107, 110, 155).

Cibulskis, K. et al. (Apr. 2012). "Absolute quantification of somatic DnA alterations in human cancer". In: *Nature Biotechnology* 30.5, pp. 413–421. (Cited on page 46).

Ciriello, G et al. (Feb. 2012). "Mutual exclusivity analysis identifies oncogenic network modules". In: *Genome Research* 22.2, pp. 398–406. (Cited on pages 67, 70).

Ciriello, G. et al. (Sept. 2013). "Emerging landscape of oncogenic signatures across human cancers". In: *Nature Genetics*, pp. 1–9. (Cited on page 20).

Clarke, R. et al. (Jan. 2008). "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data". In: *Nature reviews. Cancer* 8.1, pp. 37–49. (Cited on page 40).

Clavel, J et al. (1989). "Tobacco and bladder cancer in males: increased risk for inhalers and smokers of black tobacco". In: *International journal of cancer* 44.4, pp. 605–610. (Cited on page 31).

Coppe, J.-P. et al. (2010). "ERBB receptor regulation of ESX/ELF3 promotes invasion in breast epithelial cells". In: *Open Cancer Journal* 3, pp. 89–100. (Cited on page 128).

Croft, D et al. (Dec. 2013). "The Reactome pathway knowledgebase". In: *Nucleic Acids Research* 42.D1, pp. D472–D477. (Cited on page 65).

Daher, A. et al. (Sept. 2003). "Epidermal Growth Factor Receptor Regulates Normal Urothelial Regeneration". In: *Laboratory Investigation* 83.9, pp. 1333–1341. (Cited on pages 37, 114).

Daley, G. Q., R. A. Van Etten, and D Baltimore (Feb. 1990). "Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome." In: *Science (New York, N.Y.)* 247.4944, pp. 824–830. (Cited on page 20).

Damrauer, J. S. et al. (Feb. 2014). "Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology". In: *Proceedings of the National Academy of Sciences* 111.8, pp. 3110–3115. (Cited on pages 36, 103).

Darnell, J. E. (Oct. 2002). "Transcription factors as targets for cancer therapy". In: *Nature Publishing Group* 2.10, pp. 740–749. (Cited on page 79).

De Craene, B. and G. Berx (Feb. 2013). "Regulatory networks defining EMT during cancer initiation and progression". In: *Nature Publishing Group* 13.2, pp. 97–110. (Cited on pages 18, 52, 127).

Deane, C. M. et al. (May 2002). "Protein interactions: two methods for assessment of the reliability of high throughput observations." In: *Molecular & cellular proteomics : MCP* 1.5, pp. 349–356. (Cited on page 57).

Dees, N. D. et al. (Aug. 2012). "MuSiC: Identifying mutational significance in cancer genomes". In: *Genome Research* 22.8, pp. 1589–1598. (Cited on page 45).

DeGraff, D. J. et al. (Aug. 2013). "When urothelial differentiation pathways go wrong: Implications forbladder cancer development and progression". In: *URO* 31.6, pp. 802–811. (Cited on page 37).

Della Gatta, G. et al. (Feb. 2012). "Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL". In: *Nature Medicine* 18.3, pp. 436–440. (Cited on page 73).

D'haeseleer, P et al. (Dec. 1998). "Linear modeling of mRNA expression levels during CNS development and injury." In: *Pacific Symposium on Biocomputing*, pp. 41–52. (Cited on page 72).

Dhillon, A. S. et al. (2007). "MAP kinase signalling pathways in cancer". In: *Oncogene* 26.22, pp. 3279–3290. (Cited on page 7).

Dittrich, M. T. et al. (June 2008). "Identifying functional modules in protein-protein interaction networks: an integrated exact approach". In: *Bioinformatics* 24.13, pp. i223–i231. (Cited on pages 66, 67).

Durillo, J. J. and A. J. Nebro (Oct. 2011). "Advances in Engineering Software". In: *Advances in Engineering Software* 42.10, pp. 760–771. (Cited on page 135).

Dutta, B et al. (Mar. 2012). "A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes." In: *British journal of cancer* 106.6, pp. 1107–1116. (Cited on page 80).

Dvorak, H. F. (1986). "Tumors: Wounds That Do Not Heal". In: *The New England journal of medicine* 315.26, pp. 1650–1659. (Cited on pages 1, 19, 113, 116).

Dyrskjøt, L. et al. (June 2004). "Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification." In: *Cancer Research* 64.11, pp. 4040–4048. (Cited on pages 35, 90).

Eble, J. N. (Jan. 2004). *Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs*. IARC. (Cited on page 34).

Elati, M et al. (Sept. 2007). "LICORN: learning cooperative regulation networks from gene expression data". In: *Bioinformatics* 23.18, pp. 2407–2414. (Cited on pages 72, 77, 81, 82).

Elati, M. et al. (Feb. 2013). "PreCisIon: PREdiction of CIS-regulatory elements improved by gene's positION". In: *Nucleic Acids Research* 41.3, pp. 1406–1415. (Cited on page 78).

Epping, M. T. et al. (Dec. 2010). "TSPYL5 suppresses p53 levels and function by physical interaction with USP7." In: *Nature Cell Biology* 13.1, pp. 102–108. (Cited on page 159).

Esquela-Kerscher, A. and F. J. Slack (Apr. 2006). "Oncomirs — microRNAs with a role in cancer". In: *Nature reviews. Cancer* 6.4, pp. 259–269. (Cited on page 12).

Esteller, M. (Dec. 2011). "Non-coding RNAs in human disease". In: *Nature Reviews Genetics* 12.12, pp. 861–874. (Cited on page 52).

Fabbri, M. et al. (2012). "MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response". In: *Proceedings of the national academy of sciences of the United States of America* 109.31, E2110–E2116. (Cited on page 7).

Faith, J. J. et al. (2007). "Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles". In: *PLoS Biology* 5.1, e8. (Cited on page 71).

Fan, C. et al. (July 2006). "Concordance among gene-expression-based predictors for breast cancer." In: *The New England journal of medicine* 355.6, pp. 560–569. (Cited on page 90).

Feige, U, D Peleg, and G Kortsarz (Mar. 2001). "The Dense k -Subgraph Problem". In: *Algorithmica* 29.3, pp. 410–421. (Cited on page 135).

Ferlay, J et al. (Apr. 2013). "Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012". In: *European Journal of Cancer* 49.6, pp. 1374–1403. (Cited on page 31).

Fields, S and O Song (July 1989). "A novel genetic system to detect protein-protein interactions." In: *Nature* 340.6230, pp. 245–246. (Cited on pages 57, 58).

Fischer, A. et al. (June 2014). "High-Definition Reconstruction of Clonal Composition in Cancer". In: *CellReports* 7.5, pp. 1740–1752. (Cited on page 46).

Fleming, J. M. et al. (Dec. 2012). "Differentiation-Associated Reprogramming of the Transforming Growth Factor β Receptor Pathway Establishes the Circuitry for Epithelial Autocrine/Paracrine Repair". In: *PLoS ONE* 7.12, e51404. (Cited on pages 114, 116, 156).

Fletcher, M. N. C. et al. (2013). "Master regulators of FGFR2 signalling and breast cancer risk." In: *Nature Communications* 4, p. 2464. (Cited on pages 71, 74, 94, 95, 155).

Franceschini, A et al. (Dec. 2012). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration". In: *Nucleic Acids Research* 41.D1, pp. D808–D815. (Cited on pages 59, 95–97).

Fullwood, M. J. et al. (May 2009). "An oestrogen-receptor-". In: *Nature* 461.7269, pp. 58–64. (Cited on page 48).

Furey, T. S. (Nov. 2012). "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions." In: *Nature Reviews Genetics* 13.12, pp. 840–852. (Cited on page 46).

Galbraith, S. J., L. M. Tran, and J. C. Liao (Aug. 2006). "Transcriptome network component analysis with limited microarray data". In: *Bioinformatics* 22.15, pp. 1886–1894. (Cited on page 68).

Ganchi, P. A. et al. (Dec. 1992). "I kappa B/MAD-3 masks the nuclear localization signal of NF-kappa B p65 and requires the transactivation domain to inhibit NF-kappa B p65 DNA binding." In: *Molecular biology of the cell* 3.12, pp. 1339–1352. (Cited on page 9).

Gavin, A.-C. et al. (Jan. 2006). "Proteome survey reveals modularity of the yeast cell machinery". In: *Nature* 440.7084, pp. 631–636. (Cited on pages 57, 143).

Gerstein, M. B. et al. (Sept. 2012). "Architecture of the human regulatory network derived from ENCODE data". In: *Nature* 488.7414, pp. 91–100. (Cited on pages 92, 97, 114, 155).

Ghazalpour, A. et al. (June 2011). "Comparative Analysis of Proteome and Transcriptome Variation in Mouse". In: *PLoS Genetics* 7.6, e1001393. (Cited on page 65).

Giancotti, F. G. (Aug. 2014). "Deregulation of cell signaling in cancer". In: *FEBS Letters* 588.16, pp. 2558–2570. (Cited on page 18).

Gingras, A.-C. et al. (Aug. 2007). "Analysis of protein complexes using mass spectrometry." In: *Nature Reviews Molecular Cell Biology* 8.8, pp. 645–654. (Cited on pages 58, 131, 152).

Glaab, E et al. (Apr. 2010). "TopoGSA: network topological gene set analysis". In: *Bioinformatics* 26.9, pp. 1271–1272. (Cited on page 138).

Goodarzi, H., O. Elemento, and S. Tavazoie (Dec. 2009). "Revealing Global Regulatory Perturbations across Human Cancers". In: *Molecular Cell* 36.5, pp. 900–911. (Cited on page 80).

Grivennikov, S. I., F. R. Greten, and M. Karin (Mar. 2010). "Immunity, Inflammation, and Cancer". In: *Cell* 140.6, pp. 883–899. (Cited on page 19).

Gu, J. et al. (2010). "Identification of responsive gene modules by network-based gene clustering and extending: application to inflammation and angiogenesis". In: *BMC Systems Biology* 4.1, p. 47. (Cited on page 67).

Guo, Z et al. (Sept. 2007). "Edge-based scoring and searching method for identifying condition-responsive protein protein interaction sub-network". In: *Bioinformatics* 23.16, pp. 2121–2128. (Cited on page 67).

Hanahan, D. and R. A. Weinberg (Mar. 2011). "Hallmarks of Cancer: The Next Generation". In: *Cell* 144.5, pp. 646–674. (Cited on pages 17–19, 29).

Hasbi, A. et al. (May 2004). "Real-Time Detection of Interactions between the Human Oxytocin Receptor and G Protein-Coupled Receptor Kinase-2". In: *Molecular Endocrinology* 18.5, pp. 1277–1286. (Cited on page 155).

Haury, A.-C., P. Gestraud, and J.-P. Vert (Dec. 2011). "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures". In: *PLoS ONE* 6.12, e28210. (Cited on page 90).

Haury, A.-C. et al. (2012). "TIGRESS: Trustful Inference of Gene REgulation using Stability Selection." In: *BMC Systems Biology* 6, p. 145. (Cited on page 72).

Hermans, A et al. (Oct. 1987). "Unique fusion of bcr and c-abl genes in Philadelphia chromosome positive acute lymphoblastic leukemia." In: *Cell* 51.1, pp. 33–40. (Cited on page 20).

Hernandez, S (May 2006). "Prospective Study of FGFR3 Mutations As a Prognostic Factor in Nonmuscle Invasive Urothelial Bladder Carcinomas". In: *Journal of Clinical Oncology* 24.22, pp. 3664–3671. (Cited on page 36).

Hill, C. S. and R Treisman (Jan. 1995). "Transcriptional regulation by extracellular signals: mechanisms and specificity." In: *Cell* 80.2, pp. 199–211. (Cited on page 79).

Ho, J. R. et al. (2012). "Deregulation of Rab and Rab Effector Genes in Bladder Cancer". In: *PLoS ONE* 7.6, e39469. (Cited on page 78).

Holmäng, S et al. (Apr. 2001). "Stage progression in Ta papillary urothelial tumors: relationship to grade, immunohistochemical expression of tumor markers, mitotic frequency and DNA ploidy." In: *JURO* 165.4, 1124–8–discussion 1128–30. (Cited on page 32).

Huang, P. et al. (Apr. 2012). "Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors". In: *Nature* 475.7356, pp. 386–389. (Cited on page 79).

Husmeier, D (Nov. 2003). "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks". In: *Bioinformatics* 19.17, pp. 2271–2282. (Cited on page 72).

Huynh-Thu, V. A. et al. (Sept. 2010). "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods". In: *PLoS ONE* 5.9, e12776. (Cited on pages 72, 86, 87).

Ideker, T. and N. J. Krogan (Jan. 2012). "Differential network biology". In: *Molecular Systems Biology* 8, pp. 1–9. (Cited on page 116).

Ideker, T. et al. (2002). "Discovering regulatory and signalling circuits in molecular interaction networks." In: *Bioinformatics* 18 Suppl 1, S233–40. (Cited on page 67).

Ieda, M. et al. (Aug. 2010). "Direct Reprogramming of Fibroblasts into Functional Cardiomyocytesby Defined Factors". In: *Cell* 142.3, pp. 375–386. (Cited on page 79).

Jacobsen, A. et al. (Nov. 2013). "Analysis of microRNA-target interactions across diverse cancer types." In: *Nature structural & molecular biology* 20.11, pp. 1325–1332. (Cited on page 52).

Janz, S. (Sept. 2006). "Myc translocations in B cell and plasma cell neoplasms". In: *DNA Repair* 5.9-10, pp. 1213–1224. (Cited on page 25).

Jolma, A. et al. (Jan. 2013). "DNA-Binding Specificitiesof Human Transcription Factors". In: *Cell* 152.1-2, pp. 327–339. (Cited on pages 97, 114).

Joyce, J. A. and P. N. Schofield (Aug. 1998). "Genomic imprinting and cancer". In: *Journal of Clinical Pathology-Molecular Pathology* 51.4, pp. 185–190. (Cited on page 24).

Kandoth, C. et al. (May 2013). "Integrated genomic characterization of endometrial carcinoma." In: *Nature* 497.7447, pp. 67–73. (Cited on page 69).

Kandoth, C. et al. (Apr. 2014). "Mutational landscape and significance across 12 major cancer types". In: *Nature* 502.7471, pp. 333–339. (Cited on page 44).

Kanehisa, M. and S Goto (Dec. 1999). "KEGG: kyoto encyclopedia of genes and genomes." In: *Nucleic Acids Research* 28.1, pp. 27–30. (Cited on page 64).

Kanehisa, M. et al. (Dec. 2013). "Data, information, knowledge and principle: back to metabolism in KEGG". In: *Nucleic Acids Research* 42.D1, pp. D199–D205. (Cited on page 64).

Karr, J. R. et al. (July 2012). "A Whole-Cell Computational Model Predicts Phenotype from Genotype". In: *Cell* 150.2, pp. 389–401. (Cited on page 80).

Kato, M. V. et al. (July 1993). "Loss of heterozygosity on chromosome 13 and its association with delayed growth of retinoblastoma." In: *International Journal of Cancer* 54.6, pp. 922–926. (Cited on page 24).

Kerrien, S et al. (Dec. 2011). "The IntAct molecular interaction database in 2012". In: *Nucleic Acids Research* 40.D1, pp. D841–D846. (Cited on page 60).

Keshava Prasad, T. S. et al. (Jan. 2009). "Human Protein Reference Database–2009 update". In: *Nucleic Acids Research* 37.Database, pp. D767–D772. (Cited on pages 60, 95, 97).

Kim, D.-H. and S. Sung (Feb. 2012). "Environmentally coordinated epigenetic silencing of FLC by protein and long noncoding RNA components". In: *Current Opinion in Plant Biology* 15.1, pp. 51–56. (Cited on page 44).

Knudson, A. G. (Mar. 1971). "Mutation and cancer: statistical study of retinoblastoma." In: *Proceedings of the national academy of sciences of the United States of America* 68.4, pp. 820–823. (Cited on page 24).

Kohno, Y. Y. et al. (2006). "Expression of claudin7 is tightly associated with epithelial structures in synovial sarcomas and regulated by an Ets family transcription factor, ELF3". In: *Multiple values selected* 281.50, pp. 38941–38950. (Cited on page 125).

Kong, M et al. (Nov. 2000). "Epidermal Growth Factor-induced Phosphatidylinositol 3-Kinase Activation and DNA Synthesis: identification of Grb2-associated binder 2 as the major mediator in rat hepatocytes". In: *Journal of Biological Chemistry* 275.46, pp. 36035–36042. (Cited on page 61).

Kong, S. L. et al. (Aug. 2011). "Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state". In: *Molecular Systems Biology* 7, pp. 1–14. (Cited on page 93).

Korc, M and R. E. Friesel (Aug. 2009). "The role of fibroblast growth factors in tumor growth." In: *Current cancer drug targets* 9.5, pp. 639–651. (Cited on page 149).

Kou, Y. et al. (2013). "ChEA2: Gene-Set Libraries from ChIP-X Experiments to Decode the Transcription Regulome". English. In: *Availability, Reliability, and Security in Information Systems and HCI*. Ed. by A. Cuzzocrea et al. Vol. 8127. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 416–430. (Cited on pages 92, 97, 114, 155).

Kovall, R. A. (Sept. 2008). "More complicated than it looks: assembly of Notch pathway transcription complexes". In: *Oncogene* 27.38, pp. 5099–5109. (Cited on page 9).

Kreft, M. E. et al. (May 2005). "Urothelial injuries and the early wound healing response: tight junctions and urothelial cytodifferentiation". In: *Histochemistry and Cell Biology* 123.4-5, pp. 529–539. (Cited on page 31).

Kulakovskiy, I. V. et al. (Dec. 2012). "HOCOMOCO: a comprehensive collection of human transcription factor binding sites models". In: *Nucleic Acids Research* 41.D1, pp. D195–D202. (Cited on pages 63, 92, 97, 114).

Kuncheva, L. (2007). "A stability index for feature selection". In: *Proceedings of the 25th IASTED International Multi-Conference Artificial Intelligence and Applications*. (Cited on page 90).

Labussière, M et al. (Nov. 2014). "TERT promoter mutations in gliomas, genetic associations and clinico-pathological correlations." In: *British journal of cancer* 111.10, pp. 2024–2032. (Cited on page 22).

Laederich, M. B. et al. (May 2011). "Fibroblast Growth Factor Receptor 3 (FGFR3) Is a Strong Heat Shock Protein 90 (Hsp90) Client: implication for therapeutic manipulation". In: *Journal of Biological Chemistry* 286.22, pp. 19597–19604. (Cited on page 152).

Lafitte, M. et al. (July 2013). "FGFR3 has tumor suppressor properties in cells with epithelial phenotype". In: *Molecular Cancer* 12.1, pp. 1–1. (Cited on pages 36, 149).

Lahti, L., J. E. A. Knuuttila, and S. Kaski (Oct. 2010). "Global modeling of transcriptional responses in interaction networks." In: *Bioinformatics* 26.21, pp. 2713–2720. (Cited on pages 66, 67, 74).

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1, p. 559. (Cited on page 71).

Lavi, O., G. Dror, and R. Shamir (June 2012). "Network-Induced Classification Kernels for Gene Expression Profile Analysis". In: *Journal of Computational Biology* 19.6, pp. 694–709. (Cited on page 66).

Lawrence, M. S. et al. (Apr. 2014). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457, pp. 214–218. (Cited on pages 34, 45).

Lee, T. I. and R. A. Young (Mar. 2013). "Transcriptional Regulation and Its Misregulation in Disease". In: *Cell* 152.6, pp. 1237–1251. (Cited on pages 79, 93, 106).

Lefebvre, C. et al. (June 2010). "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers". In: *Molecular Systems Biology* 6, pp. 1–10. (Cited on pages 71, 73, 74, 86, 108, 120).

Lewis, S. A. (May 2000). "Everything you wanted to know about the bladder epithelium but were afraid to ask." In: *American Journal of Physiology - Renal Physiology* 278.6, F867–F874. (Cited on page 31).

Li, B., M. Carey, and J. L. Workman (Feb. 2007). "The Role of Chromatin during Transcription". In: *Cell* 128.4, pp. 707–719. (Cited on page 10).

Li, B. and C. N. Dewey (Aug. 2011). "RSEM: accurate transcript quantification fromRNA-Seq data with or without a reference". In: *BMC Bioinformatics* 12.1, p. 323. (Cited on page 49).

Liao, J. C. et al. (2003). "Network component analysis: reconstruction of regulatory signals in biological systems". In: *Proceedings of the National Academy of Sciences* 100.26, p. 15522. (Cited on pages 67, 68).

Ling, S et al. (May 2011). "An EGFR-ERK-SOX9 Signaling Cascade Links Urothelial Development and Regeneration to Cancer". In: *Cancer Research* 71.11, pp. 3812–3821. (Cited on page 104).

Liu, H.-X. et al. (2001). "A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes". In: *Nature Genetics* 27.1, pp. 55–58. (Cited on page 23).

Loeb, L. A. (Apr. 2001). "A mutator phenotype in cancer." In: *Cancer Research* 61.8, pp. 3230–3239. (Cited on page 21).

Longoni, N et al. (July 2013). "ETS Transcription Factor ESE1/ELF3 Orchestrates a Positive Feedback Loop That Constitutively Activates NF- B and Drives Prostate Cancer Progression". In: *Cancer Research* 73.14, pp. 4533–4547. (Cited on page 128).

Ma, W. et al. (Dec. 2014). "Fine-scale chromatin interaction maps reveal the". In: *Nature Methods*, pp. 1–14. (Cited on page 48).

Mahmood, S. F. et al. (Oct. 2013). "PPAPDC1B and WHSC1L1 are common drivers of the 8p11-12 amplicon, not only in breast tumors but also in pancreatic adenocarcinomas and lung tumors." In: *The American journal of pathology* 183.5, pp. 1634–1644. (Cited on page 78).

Marbach, D et al. (Apr. 2010). "Revealing strengths and weaknesses of methods for gene network inference". In: *Proceedings of the National Academy of Sciences* 107.14, pp. 6286–6291. (Cited on page 73).

Marbach, D et al. (July 2012a). "Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks". In: *Genome Research* 22.7, pp. 1334–1349. (Cited on pages 97, 98).

Marbach, D. et al. (July 2012b). "Wisdom of crowds for robust gene network inference". In: *Nature Methods* 9.8, pp. 796–804. (Cited on pages 73, 81, 84, 86, 90).

Margolin, A. A. et al. (2006a). "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context". In: *BMC Bioinformatics* 7.Suppl 1, S7. (Cited on pages 72, 86, 87, 94, 95).

Margolin, A. A. et al. (July 2006b). "Reverse engineering cellular networks". In: *Nature Protocols* 1.2, pp. 662–671. (Cited on page 72).

Marine, J.-C. and G Lozano (June 2009). "Mdm2-mediated ubiquitylation: p53 and beyond". In: *Cell Death and Differentiation* 17.1, pp. 93–102. (Cited on page 56).

Matys, V et al. (Jan. 2006). "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes." In: *Nucleic Acids Research* 34.Database issue, pp. D108–10. (Cited on pages 63, 84, 87).

McCarroll, S. A. et al. (Sept. 2008). "Integrated detection and population-genetic analysis of SNPs and copy number variation". In: *Nature Genetics* 40.10, pp. 1166–1174. (Cited on page 42).

Mering, C von (Dec. 2004). "STRING: known and predicted protein-protein associations, integrated and transferred across organisms". In: *Nucleic Acids Research* 33.Database issue, pp. D433–D437. (Cited on pages 59, 140).

Mirza, A. et al. (June 2003). "Global transcriptional program of p53 target genes during the process of apoptosis and cell cycle progression". In: *Oncogene* 22.23, pp. 3645–3654. (Cited on page 157).

Mitra, K. et al. (Sept. 2013). "Integrative approaches for finding modular structure in biological networks." In: *Nature Reviews Genetics* 14.10, pp. 719–732. (Cited on page 66).

Muller-Tidow, C et al. (Mar. 2001). "Cyclin A1 directly interacts with B-myb and cyclin A1/cdk2 phosphorylate B-myb at functionally important serine and threonine residues: tissue-specific regulation of B-myb function." In: *Blood* 97.7, pp. 2091–2097. (Cited on page 122).

Mumby, M. and D. Brekken (2005). "Phosphoproteomics: new insights into cellular signaling." In: *Genome Biology* 6.9, p. 230. (Cited on page 53).

Naski, M. C. et al. (May 1996). "Graded activation of fibroblast growth factor receptor 3 by mutations causing achondroplasia and thanatophoric dysplasia." In: *Nature Genetics* 13.2, pp. 233–237. (Cited on page 149).

Nepusz, T., H. Yu, and A. Paccanaro (Mar. 2012). "Detecting overlapping protein complexes in protein-protein interaction networks". In: *Nature Methods* 9.5, pp. 471–472. (Cited on pages 132, 133, 137, 138, 141).

Neuzillet, Y. et al. (Dec. 2012). "A Meta-Analysis of the Relationship between FGFR3 and TP53 Mutations in Bladder Cancer". In: *PLoS ONE* 7.12, e48993. (Cited on page 35).

Neuzillet, Y. et al. (June 2014). "FGFR3 mutations, but not FGFR3 expression and FGFR3 copy-number variations, are associated with favourable non-muscle invasive bladder cancer". In: *Virchows Archiv* 465.2, pp. 207–213. (Cited on page 36).

Ng, S et al. (Sept. 2012). "PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis". In: *Bioinformatics* 28.18, pp. i640–i646. (Cited on page 69).

Nibbe, R. K., M. Koyutürk, and M. R. Chance (Jan. 2010). "An Integrative -omics Approach to Identify Functional Sub-Networks in Human Colorectal Cancer". In: *PLoS Computational Biology* 6.1, e1000639. (Cited on page 67).

Nicolle, R., M. Elati, and F. Radvanyi (2012). "Network Transformation of Gene Expression for Feature Extraction". In: *11th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 108–113. (Cited on pages 77, 91).

Noor, A. et al. (Sept. 2013). "ROBNCA: robust network component analysis for recovering transcription factor activities." In: *Bioinformatics* 29.19, pp. 2410–2418. (Cited on pages 68, 90).

Nordentoft, I. et al. (June 2014). "Mutational Context and Diverse Clonal Development in Early and Late Bladder Cancer". In: *CellReports* 7.5, pp. 1649–1663. (Cited on page 46).

Normanno, N. et al. (Jan. 2006). "Epidermal growth factor receptor (EGFR) signaling in cancer". In: *Gene* 366.1, pp. 2–16. (Cited on page 37).

Nowell, P. C. (Sept. 1976). "The clonal evolution of tumor cell populations." In: *Science (New York, N.Y.)* 194.4260, pp. 23–28. (Cited on page 5).

Nowell, P. C. and D. A. Hungerford (Apr. 1960). "A minute chromosome in Human Chronic Granulocytic Leukemia". In: *Science (New York, N.Y.)* 131.3409, pp. 1316–1322. (Cited on page 20).

Oeckinghaus, A., M. S. Hayden, and S. Ghosh (July 2011). "Crosstalk in NF-κB signaling pathways". In: *Nature Immunology* 12.8, pp. 695–708. (Cited on page 56).

Olefsky, J. M. (Oct. 2001). "Nuclear Receptor Minireview Series". In: *Journal of Biological Chemistry* 276.40, pp. 36863–36864. (Cited on page 13).

Ozgur, A et al. (June 2008). "Identifying gene-disease associations using centrality on a literature mined gene-interaction network". In: *Bioinformatics* 24.13, pp. i277–i285. (Cited on page 138).

Pachkov, M et al. (Jan. 2007). "SwissRegulon: a database of genome-wide annotations of regulatory sites". In: *Nucleic Acids Research* 35.Database, pp. D127–D131. (Cited on page 63).

Panne, D. (Apr. 2008). "The enhanceosome". In: *Current Opinion in Structural Biology* 18.2, pp. 236–242. (Cited on pages 10, 11, 79, 93).

Parker, B. C. et al. (Jan. 2013). "The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma". In: *Journal of Clinical Investigation*. (Cited on page 25).

Parker, J. S. et al. (Mar. 2009). "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes". In: *Journal of Clinical Oncology* 27.8, pp. 1160–1167. (Cited on page 51).

Perou, C. M. et al. (Aug. 2000). "Molecular portraits of human breast tumours." In: *Nature* 406.6797, pp. 747–752. (Cited on pages 39, 40).

Pihur, V., S. Datta, and S. Datta (2009). "RankAggreg, an R package for weighted rank aggregation". In: *BMC Bioinformatics* 10.1, p. 62. (Cited on page 140).

Pinkel, D et al. (Oct. 1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." In: *Nature Genetics* 20.2, pp. 207–211. (Cited on page 41).

Poliseno, L. et al. (June 2010). "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology". In: *Nature* 465.7301, pp. 1033–1038. (Cited on page 52).

Pollack, J. R. et al. (Sept. 1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." In: *Nature Genetics* 23.1, pp. 41–46. (Cited on page 43).

Portales-Casamar, E et al. (Dec. 2009). "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles". In: *Nucleic Acids Research* 38.Database, pp. D105–D110. (Cited on pages 92, 97, 114).

Pu, S et al. (Feb. 2009). "Up-to-date catalogues of yeast protein complexes". In: *Nucleic Acids Research* 37.3, pp. 825–831. (Cited on pages 139, 143).

Qing, J. et al. (May 2009). "Antibody-based targeting of FGFR3 in bladder carcinoma and t(4;14)-positive multiple myeloma in mice". In: *Journal of Clinical Investigation* 119.5, pp. 1216–1229. (Cited on page 150).

Quach, M, N Brunel, and F d'Alche Buc (Nov. 2007). "Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference". In: *Bioinformatics* 23.23, pp. 3209–3216. (Cited on page 72).

Rapaport, F. et al. (2007). "Classification of microarray data using gene networks". In: *BMC Bioinformatics* 8.1, p. 35. (Cited on page 66).

Ravasi, T. et al. (Mar. 2010). "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man". In: *Cell* 140.5, pp. 744–752. (Cited on pages 63, 84, 93, 95, 97).

Raychaudhuri, P and H. J. Park (June 2011). "FoxM1: A Master Regulator of Tumor Metastasis". In: *Cancer Research* 71.13, pp. 4329–4333. (Cited on pages 108, 120).

Rebouissou, S. et al. (July 2014). "EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype." In: *Science Translational Medicine* 6.244, 244ra91–244ra91. (Cited on pages 35–37, 103, 117, 120, 123).

Rebouissou, S. S. et al. (June 2012). "CDKN2A homozygous deletion is associated with muscle invasion in FGFR3-mutated urothelial bladder carcinoma." In: *The Journal of Pathology* 227.3, pp. 315–324. (Cited on page 35).

Reimand, J., O. Wagih, and G. D. Bader (Dec. 2012). "The mutational landscape of phosphorylation signaling in cancer." In: *Scientific Reports* 3, pp. 2651–2651. (Cited on page 45).

Rhijn, B. W. G. van et al. (Mar. 2004). "FGFR3 and P53 characterize alternative genetic pathways in the pathogenesis of urothelial cell carcinoma." In: *Cancer Research* 64.6, pp. 1911–1914. (Cited on page 35).

Rhijn, B. van et al. (2001). "The fibroblast growth factor receptor 3 (FGFR3) mutation is a strong indicator of superficial bladder cancer with low recurrence rate". In: *Cancer Research* 61.4, p. 1265. (Cited on page 35).

Riss, J et al. (2006). "Cancers as wounds that do not heal: differences and similarities between renal regeneration/repair and renal cell carcinoma". In: *Cancer Research* 66.14, pp. 7216–7224. (Cited on page 113).

Roberts, N. D. et al. (Sept. 2013). "A comparative analysis of algorithms for somatic SNV detection in cancer." In: *Journal of Gerontology* 29.18, pp. 2223–2230. (Cited on page 44).

Ruepp, A et al. (Dec. 2009). "CORUM: the comprehensive resource of mammalian protein complexes–2009". In: *Nucleic Acids Research* 38.Database, pp. D497–D501. (Cited on pages 139, 143).

Salazar, L. et al. (Jan. 2014). "Fibroblast Growth Factor Receptor 3 Interacts with and Activates TGFβ-Activated Kinase 1 Tyrosine Phosphorylation and NFκB Signaling in Multiple Myeloma and Bladder Cancer". In: *PLoS ONE* 9.1, e86470. (Cited on page 155).

Sarkari, F., Y. Sheng, and L. Frappier (Dec. 2009). "USP7/HAUSP promotes the sequence-specific DNA binding activity of p53." In: *PLoS ONE* 5.9, e13040–e13040. (Cited on page 159).

Schaefer, C. F. et al. (Jan. 2009). "PID: the Pathway Interaction Database". In: *Nucleic Acids Research* 37.Database, pp. D674–D679. (Cited on pages 65, 69).

Schaefer, M. H. et al. (Feb. 2012). "HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores". In: *PLoS ONE* 7.2, e31826. (Cited on pages 59, 60, 95, 97, 115, 136, 143, 152).

Schaefer, U, S Schmeier, and V. B. Bajic (Dec. 2010). "TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins". In: *Nucleic Acids Research* 39.Database, pp. D106–D110. (Cited on page 63).

Schäfer, M. and S. Werner (2008). "Cancer as an overhealing wound: an old hypothesis revisited". In: *Nature Reviews Molecular Cell Biology* 9.8, pp. 628–638. (Cited on page 113).

Schneider, T. D. and R. M. Stephens (Oct. 1990). "Sequence logos: a new way to display consensus sequences." In: *Nucleic Acids Research* 18.20, pp. 6097–6100. (Cited on page 62).

Schroeter, E. H., J. A. Kisslinger, and R. Kopan (1998). "Notch-1 signalling requires ligand-induced proteolytic release of intracellular domain". In: *Nature* 393.6683, pp. 382–386. (Cited on page 9).

Sedgewick, A. J. et al. (July 2013). "Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM." In: *Bioinformatics* 29.13, pp. i62–i70. (Cited on page 69).

Sengupta, N, E. Siddiqui, and F. H. Mumtaz (Aug. 2004). "Cancers of the bladder." In: *Journal of the Royal Society for the Promotion of Health* 124.5, pp. 228–229. (Cited on page 32).

Setty, M. et al. (Aug. 2012). "Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma". In: *Molecular Systems Biology* 8, pp. 1–16. (Cited on page 74).

Shannon, P. et al. (Nov. 2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." In: *Genome Research* 13.11, pp. 2498–2504. (Cited on page 60).

Sjodahl, G et al. (June 2012). "A Molecular Taxonomy for Urothelial Carcinoma". In: *Clinical Cancer Research* 18.12, pp. 3377–3386. (Cited on pages 35, 36, 38, 103).

Smyth, G. (2005). "limma: Linear Models for Microarray Data". English. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Ed. by R. Gentleman et al. Statistics for Biology and Health. Springer New York, pp. 397–420. (Cited on page 162).

Solinas-Toldo, S. S. et al. (Nov. 1997). "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances." In: *Genes, chromosomes & cancer* 20.4, pp. 399–407. (Cited on page 41).

Someren, E. P. van et al. (Feb. 2006). "Least absolute regression network analysis of the murine osteoblast differentiation network". In: *Bioinformatics* 22.4, pp. 477–484. (Cited on page 72).

Song, L et al. (Oct. 2011). "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity". In: *Genome Research* 21.10, pp. 1757–1767. (Cited on page 46).

Sordella, R. et al. (2004). "Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways". In: *Science (New York, N.Y.)* 305.5687, pp. 1163–1167. (Cited on page 55).

Southgate, J. J. et al. (Sept. 1994). "Normal human urothelial cells in vitro: proliferation and induction of stratification." In: *Laboratory Investigation* 71.4, pp. 583–594. (Cited on page 37).

Southgate, J., J. R. W. Masters, and L. K. Trejdosiewicz (Apr. 2002). "Culture of Human Urothelium". In: *Culture of Epithelial Cells, Second Edition*. New York, USA: John Wiley & Sons, Inc., pp. 381–399. (Cited on pages 37, 114).

Staack, A. et al. (Apr. 2005). "Molecular, cellular and developmental biology of urothelium as a basis of bladder regeneration". In: *Differentiation* 73.4, pp. 121–133. (Cited on page 31).

Stark, C et al. (Dec. 2010). "The BioGRID Interaction Database: 2011 update". In: *Nucleic Acids Research* 39.Database, pp. D698–D704. (Cited on page 143).

Stransky, N. et al. (Nov. 2006). "Regional copy number–independent deregulation of transcription in cancer". In: *Nature Genetics* 38.12, pp. 1386–1396. (Cited on pages 83, 87, 89, 90, 120).

Stynen, B et al. (June 2012). "Diversity in Genetic In Vivo Methods for Protein-Protein Interaction Studies: from the Yeast Two-Hybrid System to the Mammalian Split-Luciferase System". In: *Microbiology and Molecular Biology Reviews* 76.2, pp. 331–382. (Cited on page 58).

Subramanian, A. et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, p. 15545. (Cited on pages 65, 74).

Supper, J. et al. (Jan. 2013). "Detecting and visualizing gene fusions". In: *Methods* 59.1, S24–S28. (Cited on page 50).

Sutherland, H. and W. A. Bickmore (July 2009). "Transcription factories:gene expression in unions?" In: *Nature Reviews Genetics* 10.7, pp. 457–466. (Cited on page 47).

Suzuki, M. M. and A. Bird (June 2008). "DNA methylation landscapes: provocative insights from epigenomics". In: *Nature Reviews Genetics* 9.6, pp. 465–476. (Cited on page 46).

Taboada, B, C Verde, and E Merino (July 2010). "High accuracy operon prediction method based on STRING database scores". In: *Nucleic Acids Research* 38.12, e130–e130. (Cited on page 96).

Tamborero, D., A. Gonzalez-Perez, and N. Lopez-Bigas (Sept. 2013). "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes." In: *Bioinformatics* 29.18, pp. 2238–2244. (Cited on page 45).

Tang, K.-W. et al. (2013). "The landscape of viral expression and host gene fusion and adaptation in human cancer." In: *Nature Communications* 4, p. 2513. (Cited on page 50).

Tang, X. et al. (Aug. 2014). "A joint analysis of metabolomics and genetics of breast cancer." In: *Breast cancer research : BCR* 16.4, p. 415. (Cited on page 53).

Temin, H. M. and H. Rubin (1958). "Characteristics of an assay for Rous sarcoma virus and Rous sarcoma cells in tissue culture". In: *Virology* 6.3, pp. 669–688. (Cited on page 17).

Thiery, J. P. et al. (Nov. 2009). "Epithelial-Mesenchymal Transitions in Development and Disease". In: *Cell* 139.5, pp. 871–890. (Cited on pages 18, 104, 126).

Thisse, B. and C. Thisse (Nov. 2005). "Functions and regulations of fibroblast growth factor signaling during embryonic development". In: *Developmental Biology* 287.2, pp. 390–402. (Cited on page 17).

Tibshirani, R. et al. (May 2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." In: *Proceedings of the National Academy of Sciences* 99.10, pp. 6567–6572. (Cited on pages 50, 51, 90).

Ulitsky, I. and R. Shamir (2007). "Identification of functional modules using network topology and high-throughput data." In: *BMC Systems Biology* 1, p. 8. (Cited on page 67).

Vallot, C et al. (Jan. 2011). "A Novel Epigenetic Phenotype Associated With the Most Aggressive Pathway of Bladder Tumor Progression". In: *JNCI Journal of the National Cancer Institute* 103.1, pp. 47–60. (Cited on page 35).

Vandin, F, E Upfal, and B. J. Raphael (Feb. 2012). "De novo discovery of mutated driver pathways in cancer". In: *Genome Research* 22.2, pp. 375–385. (Cited on page 70).

Vandin, F., E. Upfal, and B. J. Raphael (Mar. 2011). "Algorithms for Detecting Significantly Mutated Pathways in Cancer". In: *Journal of Computational Biology* 18.3, pp. 507–522. (Cited on page 67).

Varley, C. L. et al. (Aug. 2008). "FOXA1 and IRF-1 intermediary transcriptional regulators of PPARγ-induced urothelial cytodifferentiation". In: *Cell Death and Differentiation* 16.1, pp. 103–114. (Cited on pages 92, 105, 109, 114, 116, 125).

Varley, C. et al. (May 2005). "Autocrine regulation of human urothelial cell proliferation and migration during regenerative responses in vitro". In: *Experimental Cell Research* 306.1, pp. 216–229. (Cited on pages 31, 114).

Varley, C. L. and J. Southgate (Sept. 2008). "Effects of PPAR agonists on proliferation and differentiation in human urothelium". In: *Experimental and Toxicologic Pathology* 60.6, pp. 435–441. (Cited on pages 37, 109, 114).

Varley, C. L. et al. (Apr. 2004). "Role of PPARgamma and EGFR signalling in the urothelial terminal differentiation programme." In: *Journal of Cell Science* 117.Pt 10, pp. 2029–2036. (Cited on pages 37, 114, 123).

Varley, C. L. et al. (Aug. 2006). "PPARgamma-regulated tight junction development during human urothelial cytodifferentiation." In: *Journal of cellular physiology* 208.2, pp. 407–417. (Cited on page 114).

Varley, C. L. et al. (Dec. 2010). "Activation of peroxisome proliferator-activated receptor-gamma reverses squamous metaplasia and induces transitional differentiation in normal human urothelial cells." In: *The American journal of pathology* 164.5, pp. 1789–1798. (Cited on page 37).

Vaske, C. J. et al. (June 2010). "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM". In: *Bioinformatics* 26.12, pp. i237–i245. (Cited on pages 69, 74).

Veer, L. v. et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer". In: *Nature*. (Cited on page 39).

Velnar, T, T Bailey, and V Smrkolj (Oct. 2009). "The Wound Healing Process: An Overview of the Cellular and Molecular Mechanisms". In: *Journal of International Medical Research* 37.5, pp. 1528–1542. (Cited on page 113).

Vijver, M. J. van de et al. (Dec. 2002). "A gene-expression signature as a predictor of survival in breast cancer." In: *The New England journal of medicine* 347.25, pp. 1999–2009. (Cited on page 39).

Wakabayashi, K. i. et al. (June 2009). "The Peroxisome Proliferator-Activated Receptor /Retinoid X Receptor Heterodimer Targets the Histone Modification Enzyme PR-Set7/Setd8 Gene and Regulates Adipogenesis through a Positive Feedback Loop". In: *Molecular and Cellular Biology* 29.13, pp. 3544–3555. (Cited on page 14).

Warde-Farley, D et al. (June 2010). "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function". In: *Nucleic Acids Research* 38.Web Server, W214–W220. (Cited on page 60).

Wasserman, W. W. and A. Sandelin (Apr. 2004). "Applied bioinformatics for the identification of regulatory elements". In: *Nature Reviews Genetics* 5.4, pp. 276–287. (Cited on page 62).

Watson, J. D., T. A. Baker, and S. P. Bell (2014). *Molecular Biology of the Gene*. Pearson. seventh edition. Benjamin Cummings. (Cited on pages 8–10).

Webster, M. K. and D. J. Donoghue (1997). "FGFR activation in skeletal disorders: too much of a good thing". In: *Trends in Genetics* 13.5, pp. 178–182. (Cited on page 36).

Weinberg, R. (May 2013). *The Biology of Cancer, Second Edition*. Garland Science. (Cited on page 6).

Weinstein, J. N. et al. (Sept. 2013). "The Cancer Genome Atlas Pan-Cancer analysis project." In: *Nature Genetics* 45.10, pp. 1113–1120. (Cited on page 42).

Whitfield, M. L. et al. (June 2002). "Identification of genes periodically expressed in the human cell cycle and their expression in tumors." In: *Molecular biology of the cell* 13.6, pp. 1977–2000. (Cited on page 49).

Williams, S. V., C. D. Hurst, and M. A. Knowles (Feb. 2013). "Oncogenic FGFR3 gene fusions in bladder cancer." In: *Human Molecular Genetics* 22.4, pp. 795–803. (Cited on pages 25, 36).

Winterhalter, C et al. (Aug. 2014). "Pepper: cytoscape app for protein complex expansion using protein-protein interaction networks." In: *Bioinformatics*. (Cited on pages 78, 134, 150, 152).

Wishart, D. S. (Mar. 2008). "Quantitative metabolomics using NMR". In: *TrAC Trends in Analytical Chemistry* 27.3, pp. 228–237. (Cited on page 53).

Witsch, E, M Sela, and Y Yarden (Apr. 2010). "Roles for Growth Factors in Cancer Progression". In: *Physiology* 25.2, pp. 85–101. (Cited on page 6).

Wood, R. J. (Feb. 2008). "Vitamin D and adipogenesis: new molecular insights". In: *Nutrition Reviews* 66.1, pp. 40–46. (Cited on page 13).

Wu, X.-R. et al. (June 2009). "Uroplakins in urothelial biology, function, and disease." In: *Kidney international* 75.11, pp. 1153–1165. (Cited on page 31).

Wulfkuhle, J. et al. (Nov. 2004). "Genomic and proteomic technologies for individualisation and improvement of cancer treatment". In: *European Journal of Cancer* 40.17, pp. 2623–2632. (Cited on page 53).

Xenarios, I. et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions". In: *Nucleic Acids Research* 30.1, pp. 303–305. (Cited on page 59).

Xie, Z et al. (Jan. 2010). "hPDI: a database of experimental human protein-DNA interactions". In: *Bioinformatics* 26.2, pp. 287–289. (Cited on pages 97, 114).

Yu, Z. et al. (June 2009). "The epidermal differentiation-associated Grainyhead gene Get1/Grhl3 also regulates urothelial differentiation". In: *The EMBO Journal* 28.13, pp. 1890–1903. (Cited on pages 116, 125).

Zack, T. I. et al. (Sept. 2013). "Pan-cancer patterns of somatic copy number alteration". In: *Nature Genetics*, pp. 1–10. (Cited on page 25).

Zhang, B. et al. (July 2014a). "Proteogenomic characterization of human colon and rectal cancer". In: *Nature*, pp. 1–21. (Cited on pages 53, 54).

Zhang, S. et al. (Feb. 2014b). "The pivotal role of pyruvate dehydrogenase kinases in metabolic flexibility". In: *Nutrition & Metabolism* 11.1, pp. 1–9. (Cited on page 111).

Zhou, W. et al. (Apr. 2014). "Cancer-Secreted miR-105 Destroys Vascular Endothelial Barriers to Promote Metastasis". In: *Cancer Cell* 25.4, pp. 501–515. (Cited on page 7).

Zhu, Y., X. Shen, and W. Pan (2009). "Network-based support vector machine for classification of microarray samples". In: *BMC Bioinformatics* 10.Suppl 1, S21. (Cited on page 66).

Zingone, A et al. (Apr. 2010). "leu201050a". In: *Leukemia* 24.6, pp. 1171–1178. (Cited on page 155).

Zitzler, E., M. Laumanns, and L. Thiele (2001). "SPEA2: Improving the strength Pareto evolutionary algorithm". In: *Technical Report*, pp. 1–21. (Cited on pages 133, 135).

Zuberi, K et al. (June 2013). "GeneMANIA Prediction Server 2013 Update". In: *Nucleic Acids Research* 41.W1, W115–W122. (Cited on page 60).