PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Settore disciplinare: Med/04

A Network-based Approach to Breast Cancer Systems Medicine

Eleonora Lusito

IEO, Milan

Matricola n. R09334

Supervisor: Prof. Pier Paolo Di Fiore

IFOM, IEO, Milan

Added Supervisor: Dr. Fabrizio Bianchi

IEO, Milan

Academic year 2013-2014

Abstract

Breast cancer is the most commonly diagnosed cancer and the second leading cause of cancer death in women. Although recent improvements in the prevention, early detection, and treatment of breast cancer have led to a significant decrease in the mortality rate, the identification of an optimal therapeutic strategy for each patient remains a difficult task because of the heterogeneous nature of the disease. Clinical heterogeneity of breast cancer is in part explained by the vast genetic and molecular heterogeneity of this disease, which is now emerging from large-scale screening studies using "-omics" technologies (e.g. microarray gene expression profiling, next-generation sequencing). This genetic and molecular heterogeneity likely contributes significantly to therapy response and clinical outcome. The recent advances in our understanding of the molecular nature of breast cancer due, in particular, to the explosion of high-throughput technologies, is driving a shift away from the "one-dose-fits-all" paradigm in healthcare, to the new "Personalized Cancer Care" paradigm. The aim of "Personalized Cancer Care" is to select the optimal course of clinical intervention for individual patients, maximizing the likelihood of effective treatment and reducing the probability of adverse drug reactions, according to the molecular features of the patient. In light to this medical scenario, the aim of this project is to identify novel molecular mechanisms that are altered in breast cancer through the development of a computational pipeline, in order to propose putative biomarkers and druggable target genes for the personalized management of patients. Through the application of a Systems Biology approach to reverse engineer Gene Regulatory Networks (GRNs) from gene expression data, we built GRNs around "hub" genes transcriptionally correlating with clinical-pathological features associated with breast tumor expression profiles. The relevance of the GRNs as putative cancer-related mechanisms was reinforced by the occurrence of mutational events related to breast cancer in the "hub" genes, as well as in the neighbor genes.

Moreover, for some networks, we observed mutually exclusive mutational patterns in the neighbors genes, thus supporting their predicted role as oncogenic mechanisms. Strikingly, a substantial fraction of GRNs were overexpressed in triple negative breast cancer patients who acquired resistance to therapy, suggesting the involvement of these networks in mechanisms of chemoresistance. In conclusion, our approach allowed us to identify cancer molecular mechanisms frequently altered in breast cancer and in chemorefractory tumors, which may suggest novel cancer biomarkers and potential drug targets for the development of more effective therapeutic strategies in metastatic breast cancer patients.

Acknowledgements

Foremost, I would like to express my gratitude to my supervisor Pier Paolo Di Fiore for giving me the opportunity to work in his group and on a fascinating project. I am particularly grateful to my added co-supervisor Fabrizio Bianchi for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm and knowledge. His guidance helped me in all the time of research and writing of this thesis. I would also like to thank my external co-supervisor Stein Aertz and my internal co-supervisor Michele Caselle for precious suggestions. A special thank goes to Rosalind Gunby for her invaluable and precious help in editing badly written sentences into something readable. My sincere thank also goes to special labmates Rose Mary Carletti, Elisa Dama, Valentina Melocchi, Fabio Dezi, Francesca De Santis and Stefano Marchesi for their precious support, stimulating discussions and for all the fun we have had in the last four years. I would also like to thank Simona Monterisi, Francesca Montani, Michele Caccia, Emanuele Villa, Stefano Freddi, Alessandra Villa, Michela Lupia, Pietro Lo Riso and Francesca Angiolini for suggestions during our lab meetings. Thanks also to all the bioinformaticians past and present, with whom I have shared ideas, troubles, rants, laughs and lunches. In particular a special thank goes to Giovanni D'Ario for his help with R programming language and for having encouraged me during despair moments. Last but not the least, I would like to thank my family for continued and sincere support. Finally, thank you to all those people who believed in me and in this work.

Contents

Acronyms

CMs Cancer Modules

CM-genes Cancer Modules-genes

PhLs Phenotype Labels

GRNs Gene Regulatory Networks

GRNi Gene Regulatory Network inference

MR Master Regulator

MR-GRNs Master Regulator-Gene Regulatory Networks

CM-GRNs Cancer Module Gene Regulatory Network

TNBC Triple Negative Breast Cancer

RD Residual Disease

pCR pathologic Complete Response

TGFB1I1 Transforming growth factor beta 1 induced transcript 1

TCF4 Transcription factor 4

ZFPM2 Zinc finger protein, FOG family member 2

PRRX1 Paired related homeobox 1

ELF4 E74-like factor 4 (ets domain transcription factor)

COL1A1 Collagen, type I, alpha 1

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cancer

Cancer is a multifactorial disease characterized at a macroscopic level by uncontrolled and limitless cell proliferation, self-sufficiency in growth signals, and the ability to invade tissues, spread to distant organs (metastasize) and form new blood vessels for nutrient supply to cancer cells (angiogenesis; Figure 1.1) ([1],[2]). With 1,665,540 new cases and 585,720 deaths estimated in the United States, alone, in 2014 ([3]) cancer remains one of the main causes of death worldwide. Therefore, the identification of innovative cancer biomarkers, as well as more effective strategies for detection and treatment of cancer, is paramount.

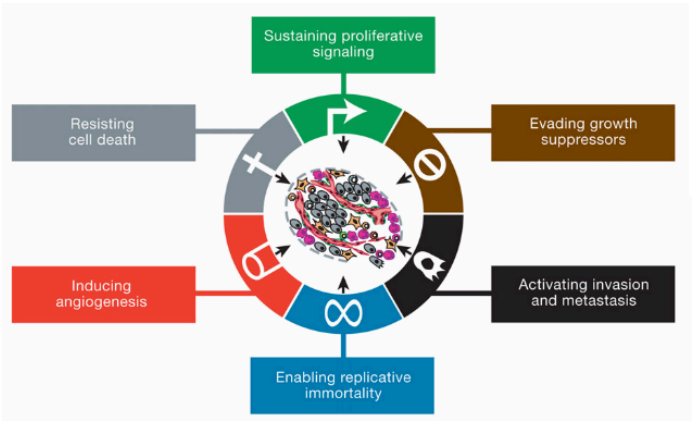

Figure 1.1: **The hallmarks of cancer.**
Schematic representation of the hallmarks of cancer. Recent advances in cancer biology have improved our comprehension of cancer-related mechanisms required to sustain the neoplastic phenotype. Taken from [2].

1.1.1 Cancer classification

There are over 100 different types of cancer classified by the type of cell that is initially affected. Cancer types can be grouped in five main categories according to the histological type and the primary site:

- **Carcinoma** is a type of cancer that originates from epithelial cells lining the inner or outer surfaces of the body. Many histological subtypes of carcinoma have been characterized including basal cell carcinoma, adenosquamous carcinoma, transitional cell carcinoma and squamous cell carcinoma.
- **Sarcoma** is a type of cancer originating in blood vessels, nerves and tendons, muscles, cartilage, bone, fat, connective tissue.
- **Lymphoma and myeloma** originate in the lymph system (lymph nodes and lymphatic vessels) and in general involve cells of the immune system.
- **Leukemia** is a type of cancer that originates in blood-forming tissue (bone marrow) allowing the formation of abnormal blood cells that will be released in the blood.
- **Central nervous system cancers** are cancers originating in brain tissues.

1.1.2 Cancer development

Cancer development may be caused by environmental factors, harmful life habits, and genetic inheritance. The accumulation of mutations in the DNA due to exogenous and endogenous DNA-damaging agents and the resulting genomic instability are at the basis of neoplastic transformation ([4]). More recently, a role for epigenetic alterations was also proposed as an addictive factor that may induce neoplastic transformation ([5],[6],[7],[8]). Genes that hold the potential to promote neoplastic transformation are called oncogenes, while those that oppose transformation are named tumor suppressor genes. Mutations or epigenetic alterations may cause overexpression or reduction/ablation of an oncogene or a tumor suppressor, respectively, thereby contributing to neoplastic transformation [9],[10],[11]). In addition, many cancer-related mutations cause activation or inactivation of specific signaling proteins, resulting in hyper-activation of signaling pathways that promote proliferation, migration or invasion, and ultimately neoplastic transformation ([12],[13],[14]).

1.1.3 Cancer stem cells

An emerging field in cancer biology is related to the identification and characterization of cancer stem cells ([15],[16]). Cancer stem cells are thought to be the engine of the tumor since they are the only cells within the tumor that are able to regenerate tumors in vivo. Cancer stem cells possess self-renewal and differentiation capabilities similar to those of normal adult stem cells ([17],[18]). The elucidation of the mechanisms that control such stem cell properties would shed light on the disrupted pathways responsible for cancer stem cell generation and tumor growth.
The first evidence of the existence of cancer stem cells came from the study of leukemia; it was shown that only a small fraction of leukemia cells proliferated extensively in vivo and in vitro ([15]). The involvement of cancer stem cells in tumorigenesis is further sustained by their distinctive trait to be the only long-lived cell population. This feature makes them preferential targets of initial oncogenic mutations because of their long exposure to genotoxic stresses. In addition, two further observations support the cancer stem cell theory: the first refers to tumor heterogeneity; the second concerns the number of cells required for tumor growth. In the first case, although cancer cells originate from a single transformed cell they display different phenotypic traits that were present in the original normal tissue from which they derive ([19]). In the second case, the cancer stem cell hypothesis is supported by evidence showing that only cells with a high capability of self-renewal, like the cancer stem cells, are able to sustain the intensive proliferation of a tumor.

1.1.4 Cancer metastasis

Cancer cells may invade and colonize other tissues and organs through the lymphatic system and/or blood. This metastatic process is initially triggered by stochastic events that allow the dispersion of cancer cells into the circulation, and is dependent on the ability of a small fraction of cells to survive in distant organs, giving rise to metastases ([20]). The ability of cancer cells to invade distant organ sites is tumor-specific, although in some cases different tumor types are able to colonize the same organ site (Table 1.1).

Table 1.1: **Metastatic relapse sites for solid tumors.**

Tumor typet	**Principal sites of metastasis**
Breast	Bone, lungs, liver, brain
Lung adenocarcinoma	Brain, bones, adrenal gland, liver
Skin melanoma	Lungs, brain, skin, liver
Colorectal	Liver, lungs
Pancreatic	Liver and lungs
Prostate	Bones
Sarcoma	Lungs
Uveal melanoma	Liver

In the case of metastatic invasion through the lymphatic system, cancer cells travel through the lymph system and they may end up in lymph nodes giving rise to a metastatic lymph node tumor. In order to spread to new parts of the body through the lymphatic system, cancer cells have to break away from the original tumor and attach themselves to the outside wall of a lymph vessel. Then, the cells move through the vessel wall to flow with the lymph to a new lymph node. The progression of the tumor towards metastasis through the blood vessels can be summarized by the following steps (Figure A.1):

- the **local invasion**: the cancer cells locally infiltrate through the basement membrane into the surrounding/adjacent tissue.
- the **intravasion**, also called "endothelial transmigration", of tumor cells into vessels: the cancer cells invade the blood or lymphatic vessels through the basal membrane.
- the **hematogenous survival and translocation**: the cancer cells are able to survive in the circulatory system and disseminate through the bloodstream to microvessels of distant tissues. The intravasion together with the hematogenous survival constitute the "hematogenous dissemination" process.
- the **extravasion**: cancer cells exit from the bloodstream.
- the **colonization**: cancer cells colonize distant organs. The cells adapt to the foreign microenvironment of distant site and start proliferating and forming macroscopic secondary tumors in competent organs.

Although the molecular mechanisms at the basis of the metastatic ability of cancer cells are not fully characterized, the functional activity of some genes is associated to the initiation and progression of metastasis. The metastasis initiation genes promote cell motility, epithelial-to-mesenchymal transition (EMT), extracellular matrix degradation and angiogenesis. The key genes that, for example, promote EMT (local tumor invasion) through changes in cell adhesion and migratory properties of tumor cells include the Snail (SNAI1 and SNAI2) ([22],[23]), Zeb (ZEB1 and ZEB2) ([24],[25]) and basic helix-loop-helix (bHLH: E47 and TWIST) ([26]) transcription factor families that contribute to the activation of a plethora of genes involved in the above mentioned EMT pathway.

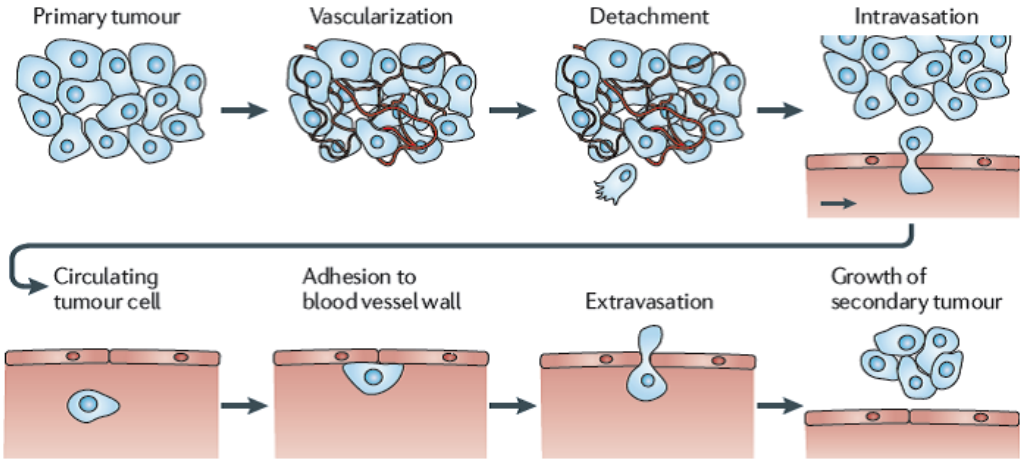

Figure 1.2: **Mechanical processes of a metastatic event.** (Taken from [21]).

1.2 Breast cancer

Breast cancer is the most common cancer diagnosed worldwide in women (second most common cancer overall) with 232,670 new cases and 40,000 deaths estimated in the US in 2014 ([27]). The incidence and overall mortality rates are higher in high-income countries respect to low and middle-income countries mainly because of an increasing adoption of cancer-causing behaviors, like for example overweight/obesity, a sedentary lifestyle and smoking. Although the incidence and the overall mortality rates are lower in low and middle-income countries the fatality rates from breast cancer still remain high mainly because of a scarcity of adequate facilities for detection and diagnosis, as well as poor access to primary treatment ([28],[29],[30]). Breast cancer originates from the epithelial and myoepithelial cells lining the ductal or lobular part of the mammary gland, and it occurs almost entirely in women, although there are rare cases of breast cancer in men. The majority of breast cancers originate in cells lining the ducts: tubes that carry the milk from the lobules to the nipple. Thus, these cancers are named ductal carcinomas. Tumors originating from cells lining the lobules are instead named lobular carcinomas (Figure 1.3). Ductal and lobular carcinomas can be further classified as invasive or in situ carcinoma depending on whether the cancer has spread into the surrounding tissues or to distant sites (i.e., invasive ductal carcinoma, IDC, or invasive lobular carcinoma, ILC), or whether it has remained localized at the site of origin (i.e., ductal carcinoma *in situ*, DCIS, or lobular carcinoma *in situ*, LCIS). IDC accounts for 80% of invasive breast cancers while ILC accounts for 10% of invasive breast tumors ([31]). Generally, in situ carcinomas are classified as early stage (stage 0) tumors and if untreated may become invasive and metastatic breast tumors.

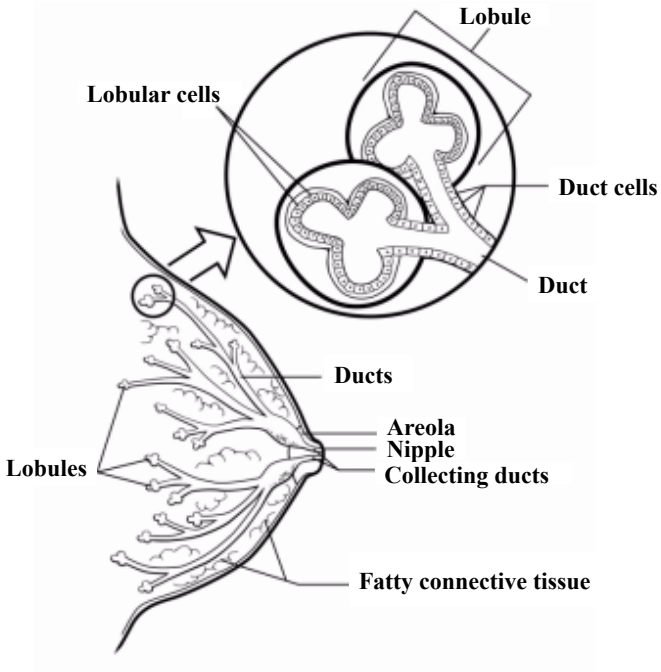

Figure 1.3: **Breast cancer sites of origin.**
Breast cancers may arise from the cells lining the milk lobules (glands) or from the cells lining the milk ducts within the breast lobes. In the first case they are called lobular carcinomas while in the second case they are called ductal carcinomas. Both types of tumours are further classified as invasive or in situ carcinoma according to the site of invasion. Taken from ([31]).

1.2.1 Breast cancer is an heterogeneous disease

Breast carcinoma is a highly heterogeneous disease with multiple tumor subtypes characterized by different histological and molecular features, which likely impact on therapy response and clinical outcome ([32],[33],[34]). Despite recent improvements in prevention, early detection, and treatment of breast cancer have led to a significant decrease in the mortality rate ([27]), the identification of an optimal therapeutic strategy for each patient remains a difficult task because of the heterogeneous nature of the disease. Originally, breast cancer diagnosis and subtype classification was based on specific histological and morphological features (histological heterogeneity) that allow the classification of the disease in 20 major tumor types and 18 minor subtypes ([35]). The recent advances in microarray gene expression profiling, next-generation sequencing (NGS), and high-throughput proteomics, is now allowing a more in-depth molecular characterization of breast cancer at a genomic and proteomic level. This has led to the identification of novel breast cancer subtypes ([36],[37],[38],[39]), and an improvement in the diagnostic and prognostic evaluation of breast cancer patients ([40],[41]). In addition, recent NGS studies of breast tumors revealed a certain level of intra-tumor heterogeneity ([42],[43]). In particular, Navin and colleagues ([43]), through single-nucleus DNA sequencing, identified the presence of intra-tumor distinct clonal subpopulations characterized by distinct genomic alterations. In some cases, an almost identical profile was found in metastatic tumor cells (i.e. synchronous metastatic lesions) respect to specific clonal subpopulations in primary tumours. Indeed, Ding and colleagues ([42]) showed that in basal-like breast tumors, metastatic lesions arise from sub-populations of cancer cells in the primary tumor with a specific repertoire of mutations, which were suggested to be drivers for cancer progression. However, despite these recent advances in the characterization of the genomic profiles of breast cancer, the molecular mechanisms involved in disease onset and progression remain mostly unclear. Further studies are thus necessary to better characterize these breast cancer pathways and to identify reliable cancer biomarkers for improving therapeutic intervention and survival of breast cancer patients ([37],[38],[44],[45],[46],[47],[48]).

1.2.2 Molecular classification of breast cancer

Breast cancer can be classified into three main groups based on the expression of different breast cancer markers, which are detected by immunohistochemistry (IHC) or fluorescence *in situ* hybridization (FISH) assays:

- **The hormone receptor positive group**: these tumors express the ER and/or PgR and account for 68% of breast cancers.
- **The Her2 positive group**: these tumors overexpress the Her2 receptor at the protein level (detected by IHC), or carry an amplification of the *Her2* gene (detected by FISH). They do not express ER and/or PgR. The frequency of occurrence is 11% in the female breast cancer population.
- **The triple negative group**: there tumors lack expression of all three receptors, ER, PgR and Her2, and account for 19% of female breast cancers.

The hormone receptor positive group can be further divided into two additional subtypes, according to the expression of the proliferation marker, Ki-67, and Her2 receptor:

- **The Luminal A subtype**: these tumors express the ER and/or PgR, but not Her2, and poorly express the Ki-67 proliferation marker. These tumors account for 44% of female breast cancers.
- **The Luminal B subtype**: these tumors express the ER and/or PgR, and Her2, and highly express the Ki-67 proliferation marker. The frequency of occurrence is 24% in female breast cancer population.

The **Luminal A** subtype is the only breast cancer subtype that has a good prognosis, high survival rates and low risk of recurrence ([49],[50],[51]). The low level of aggressiveness of these tumors is attributed to their low rate of proliferation and their positivity for ER expression, which allows the use of endocrine therapy (also referred to as hormone therapy): a systemic therapy that blocks tumor cell growth ([52]). In contrast, the **Luminal B** subtype is highly proliferative and poorly-differentiated, and displays features associated with poor prognosis, such as: i) large tumor size; ii) high tumor grade; iii) the presence of tumor cells in the lymph nodes. The Her2-positive and triple negative subtypes are highly metastatic and exhibit the worse clinical outcome, although more effective therapies are available for the former ([49],[53],[54],[55],[56]).

1.2.3 Breast cancer treatment

Breast cancer is generally treated locally through surgery and/or radiation therapy, or systemically using chemotherapeutic agents or hormone therapy. Chemotherapy may be given in a neoadjuvant regimen to reduce tumor burden before surgery, or in an adjuvant regimen after surgery to reduce the risk of recurrence. Recently, it has been documented that chemotherapy treatments including taxane (i.e., paclitaxel, Taxol and docetaxel, Taxotere) and anthracycline (i.e., doxorubicin, Adriamycin and epirubicin, Ellence) in a neoadjuvant setting is an effective strategy to increase overall survival in patients with locally advanced breast cancer ([57],[58]). Other chemotherapeutic agents for breast cancer treatment include: Cyclophosphamide (Cytoxan), Capecitabine (Xeloda) and fluorouracil (5 FU), methotrexate (Rheumatrex, Trexall), lapatinib (Tykerb). Chemotherapy drugs are usually given in 2-4 week cycles, but some may be used on a weekly basis. They can be also given in combinations with two or more drugs. The hormone therapy drugs frequently used in clinical practice to treat early, locally advanced or metastatic ER positive breast cancers are: i) tamoxifen (Nolvadex), a selective ER modulator (SERM) that binds to the receptor and prevents its activation by the ligand, estrogen, thereby inhibiting tumor cell growth; ii) the aromatase inhibitors that act by blocking the biosynthesis of estrogen, thus, reducing the availability of estrogen to cancer cells ([52]). Patients diagnosed with Her2-positive tumors or with triple negative breast tumors are frequently unresponsive to standard chemotherapy. The use of the hormone therapy is not an option since these tumors do not express ER or PgR. For Her2-positive tumors, a molecularly targeted therapy is available based on monoclonal antibodies targeting the extracellular portion of the Her2/neu receptor (i.e. Trastuzumab or Herceptin). Patients with metastatic breast cancer (late-stage), who were treated with Trastuzumab displayed an increase in overall survival of 20 to 25 months ([59]), while in patients with Her2-positive non metastatic cancer (early-stage) Trastuzumab reduce the absolute risk of relapse after the surgery of 9.5% and the absolute risk of death of 3%([60]). Triple negative tumors are typically treated with the combination of surgery radiation therapy and chemotherapy. They cannot be treated with hormone therapy or Trastuzumab (Herceptin) because they are ER-negative and Her2-negative. Target therapies are not available for these tumours because the genes that are linked to this breast cancer subtype are still not well understood. Although new treatments are being studied ([61]), more effective treatments are urgently required for this group of breast cancer patients characterized to have low five-year survival rate.

1.2.4 Personalized Breast Cancer Care

Although metastatic breast cancer still remains an aggressive and incurable disease, early stage breast cancer is curable in most patients. Indeed, the decrease in the mortality rate observed worldwide in the last 10 years ([27]) is, in part, due to the diffusion of preventive mammography screening programs that allow the detection of non-metastatic, early-stage disease that is curable by surgery ([62],[63]). In addition, our increased understanding of breast cancer biology over recent years has led to the development of more effective, molecularly targeted treatments like for example those making use of Lapatinib (Tykerb) and Trastuzumab (Herceptin) for Her2-positive tumours, Tamoxifen ER-positive tumours, that have helped to reduce the mortality of certain breast cancer subtypes ([64]). The deeply understanding of breast cancer due, in particular, to the explosion of "-omics" technologies, is driving a shift away from the "one-dose-fits-all" paradigm to a new paradigm in healthcare, the so-called "Personalized Cancer Care" or "Personalized Medicine". Personalized medicine aims to select the optimal course of clinical intervention for individual patients, maximizing the likelihood of effective treatment and reducing the probability of adverse drug reactions. The major determinant in the success of personalized medicine is the identification of predictive and prognostic molecular biomarkers that reflect the variability of breast cancer patients in terms of therapy response and clinical outcome, respectively. The availability of such cancer biomarkers would allow the stratification of patients in terms of risk of disease recurrence and responsiveness to specific therapies, thereby overcoming the problems of undertreatment and overtreatment of cancer. For instance, biomarkers that identify more aggressive tumors can help avoid undertreatment, since such tumors can be treated with more aggressive therapies. Whereas, biomarkers that are predictive of therapy response, will help to prevent overtreatment of patients who would otherwise receive little benefit from the treatment, whilst being exposed to potentially adverse side-effects ([65],[66],[67],[68],[69]).

1.2.5 Prognostic and predictive biomarkers in breast cancer care

As defined in Clark et al.,([70]) prognostic biomarkers are biological molecules whose modulation, in terms of quantity or function, correlates with prognosis ([71],[72],[73]). These biomarkers can be used in clinical practice to stratify cancer patients and identify the optimal treatment regimens. Predictive biomarkers, instead, correlate with treatment response and are used to predict whether or

not a patient is likely to respond to a specific treatment. Predictive biomarkers may overlap with prognostic biomarkers. For instance, the prognostic biomarkers that are routinely used in the clinic for breast cancer, i.e., Her2 and ER/PgR, are also predictive biomarkers. The levels of ER/PgR are used to predict response to endocrine therapy, with high ER/PgR expressing tumors being more responsive than low ER/PgR expressing tumor ([48],[74],[75],[76]). Likewise, Her2 overexpression, as well as being a risk factor for metastatic disease, is also an indicator of responsiveness to targeted therapy with the anti-Her2 monoclonal antibodies, Trastuzumab and Herceptin ([77],[78]). Cancer genomics is producing a wealth of gene signatures with prognostic and predictive potential. However, only few of them are commercially available and currently employed in clinical practice ([79]): the Oncotype DX signature ([48]) and the MammaPrint signature ([45]). MammaPrint is a 70-gene expression assay that stratifies patients according to high and low risk of distant recurrence, using marker genes associated with proliferation, angiogenesis, stromal invasion and metastases ([45]). It has been shown that this signature is able to predict relapse better than traditional clinicopathological features ([45],[80]). The Oncotype DX classifies patients into two groups: i) those with a low or intermediate-risk of recurrence who benefit significantly from Tamoxifen treatment; ii) those with a high risk of recurrence who may benefit from chemotherapy. The genes in the Oncotype DX signature that have a high predictive value include proliferation genes, such as those encoding cyclin B1 (CCNB1), Ki67, Myb-related protein B (MyBL2), survivin, and serine/threonine protein kinases (STKs), as well as genes encoding the ER and PgR ([81]). Apart from these two examples of prognostic and predictive gene signatures that are currently in clinical use, numerous other signatures have not made it to the clinic. The major reasons why many gene signatures have not been developed into clinical tools are: their poor overlap in terms of common genes, the lack of validation in independent studies and limited improvement in the predictive value with respect to that provided by standard clinicopathologic parameters([79],[82],[83],[84],[85]). Historically, prognostic gene signatures were derived using microarray gene expression profiles. Such profiles allow the identification of the transcriptional variations amongst breast tumors that correlate with clinical outcome and therapy response. Despite the potential of such genomics technologies, the poor overlap between the currently available signatures is mainly due to: i) the large number of the differentially expressed genes that correlate with prognosis; ii) the high tumor genetic heterogeneity added to the intrinsic genetic heterogeneity of individuals of different ethnicities; iii)

the different data analysis techniques; iv) poor experimental design and insufficent sample size ([86]). An alternative strategy towards the discovery of more powerful prognostic and predictive tools is the definition of molecular biomarkers according to disease-related pathways, such as signal transduction pathways directly implicated in disease phenotypes.

1.3 The Systems Biology approach to cancer research: Cancer Systems Biology (CSB)

Systems Biology is a field of biological research ([87]). The major goal of Systems Biology is to discover the general properties that govern biological systems at system-level through the characterization of the functional relationships among biological molecules. Systems Biology integrates multi-scale types of high-throughput biological data (i.e. genomic, transcriptomic, metabolic, proteomic data) and uses mathematical modeling and simulations to understand the biological complexity. A second aim of Systems Biology, when applied to the human health, is to investigate the impact of perturbations on the biological systems and to determine whether these perturbations are linked to a specific disease, and could thus be relevant to the development of novel therapeutic strategies ([88]). The application of System Biology to cancer research is called Cancer Systems Biology (CSB). CSB aims to unveil biological properties of cancer cellular systems through the characterization of molecular mechanisms involved in cancer (ranging from genome-wide regulatory and signalling networks to more detailed kinetic models of key biological reactions) to finally identify molecular therapeutic targets. Traditional approaches to the study of complex diseases like cancer, were based on the gene-centric analysis of constituent parts of the system under study ([89],[90]) and their functional involvement in the pathology. Although this has been a successful approach that has led to the discovery of genes (and mechanisms) involved in tumorigenesis, such as *MYC, TP53, ERBB2*, and *EGFR*, it is unable to fully and comprehensively capture the complex nature of biological systems ([91],[92],[93],[94]) and in particular of highly perturbed systems like cancer cells. CSB aims to gain insights into such complexity using unbiased and genome-wide high-throughput "-omics" data. The deconvolution of the structure and topological properties of the molecular mechanisms actively involved in cancer will increase the understanding of tumor initiation and progression, unveil mechanisms of action of anticancer drugs, and contribute to the elucidation of mechanisms of resistance to pharmacological treatments towards more effective therapeutic strategies. The ineffectiveness of some lifesaving pharmacological treatments making use of anti-cancer drugs, with a failure rate of approximatively 95% ([95]), is, in fact, mainly due to the ability of cancer cells to find alternative mechanisms to escape the effect of anti-cancer molecules ([96],[97]). In light of the flexible behavior of cancer cells it is of crucial importance to shed light on the complex mechanisms governing cancer disease in order to increase the probability of success of pharmacological therapies. A typical approach in

CSB, is the inference of mechanisms of gene expression regulation (i.e. Gene Regulatory Networks (GRNs)) that are altered in cancer and that contribute to the major hallmarks of the disease like the sustained cell proliferation, escape from apoptosis and invasiveness. GRNs are collections of genes and regulators, connected by physical and/or regulatory interactions. Some examples of GRNs are: transcription factor (TF)-target genes network, microRNA-target genes network, and networks deriving from the combinatorial activity of regulators like TFs, microRNA, RNA binding proteins and their target genes. GRNs, and more in general biological mechanisms and systems, are represented by graph diagrams, i.e. networks, in which the functional relationships between the components are represented by edges connecting nodes, i.e. biological molecules (see "Graph Theoretical Models (GTMs)", Subsection 1.4.4).

1.3.1 The Systems Biology pipeline to model cancer systems and computational approaches to Systems Biology and Cancer Systems Biology

The Systems Biology approach to cancer research involves:

1. The massive profiling of the tumor genome, transcriptome, proteome, epigenome and metabolome (DNA/RNA sequencing, microarray gene expression profiling, proteome screening), to qualitatively and quantitatively map the molecular profile of cancer cells.

2. The integration of multi-omics data layers to produce a comprehensive molecular landscape.

3. The modelling of the system through realistic models (e.g. models of gene expression regulation, GRNs), to infer the dynamical properties and key features of the system.

4. The experimental validation of the reliability of the predicted models and their biological relevance.

5. The identification of candidate molecular targets for disease therapy according to the structure and topological properties of biologically relevant models.

Computational approaches to Systems Biology, also applied to CSB, can be divided into two major categories: data mining and simulation-based approaches.

Computational approaches that focuses on data mining aim to extract hidden patterns from high-throughput experimental data (knowledge discovery), while simulation-based approaches test hypothesis from *in silico* experiments, producing predictions to be tested *in vitro* and *in vivo* wet lab experiments ([98]). Data-mining approaches make use of sophisticated machine learning algorithms that are able to deal with high-dimensional data. In contrast, simulation-based analysis methods predict the dynamics of systems and experimentally tests the validity of such predictions in the wet lab. This approach is relies on the interplay between computationally predicted models and experimental observation.

1.3.2 Bioinformatics tools used in Systems Biology and in Cancer Systems Biology

Systems Biology and CSB strongly depend on software tools and resources to achieve the goals of novel biological discovery and design of more effective drugs. Over the last years, we have witnessed the explosion of a plethora of computational tools for Systems Biology and CSB (summarized in Figure 1.4).

	Tools		Standards			Projects
	Software	Resources	Ontologies	File format	Minimum information	
Data and knowledge management	MAGE-TAB, ISA-TAB, KNIME, caGrid, Taverna, Bio-STEER	BioCatalogue	SBO, OBO, NCBO	MGED (MAGE), PSI, MSI	MIAME, MIAPE, MIBBI, ISO MDR, DCMI	
Data-driven network inference	R, MATLAB, BANJO					DREAM Initiative, Sage Bionetworks
Deep curation	CellDesigner, EPE, Jdesigner, PathVISIO	KEGG, Reactome, Panther pathway database, BioModels.net, WikiPathways		SBML, SBGN, CellML, BioPAX, PSI-MI	MIRIAM	
***In silico* simulation**	COPASI, SBW, JSim, Neuron, GENESIS, MATLAB, ANSYS, FreeFEM, ePNK, ina, WoPeD, Petri nets, OpenCell, CellDesigner + COPASI, CellDesigner + SOSlib, PhysioDesigner (formerly *insilico*IDE)			SED-ML, SBRML, PNML, SBML	MIASE	
Model analysis	MATLAB, Auto, XPPAut, BUNKI, ManLab, ByoDyn, SenSB, COBRA, MetNetMaker, DBSolve Optimum, Kintecus, NetBuilder, BooleanNet, SimBoolNet					
Physiological modelling	JSim, PhysioDesigner (formerly *insilico*IDE), CellDesigner (cellular modelling), FLAME, OpenCell, Virtual Physiology (produced by cLabs), GENESIS, Neuron, Heart Simulator, AnyBody			CellML, SBML, NeuroML, MML		IUPS Physiome Project, Virtual Physiological Human, High-Definition Physiology
Molecular interaction modelling	AutoDock Vina, GOLD, eHiTS	RCSB PDB, ZINC, PubChem, PDBbind				

Figure 1.4: **Softwares and computational resources commonly used in Systems Biology and in CSB.** Taken from ([99]).

Computational tools for Systems Biology and CSB can be divided into different categories:

- tools for **data knowledge management** (e.g., MAGE-TAB, ISA-TAB, Taverna) and in particular for the acquisition and storage of data ([100]). These are of crucial importance especially in the current big-data era. A critical challenge that faces the development of such tools concerns the definition of standard formats and identifiers to facilitate data exchange between different sources.

- tools for **data-driven network inference** (e.g., R, MATLAB and Banjo) from high-throughput static and time-course data, which are able to infer causal relationships among biomolecules ([101],[102]).

- tools to build **molecular interaction maps from curated data**(e.g., CellDesigner, PathVISIO). This is an alternative strategy to network inference from data and it is based on the integration of different sources of curated data ([103]). Networks generated using this approach do not necessarily carry information on the causality of the relationships between the molecular entities.

- tools for ***in silico* simulation** (e.g., MATLAB, COBRA, SenSB). These tools are frequently use to model dynamic networks ([104]). This task is not addressed by data-driven network inference methods nor from networks built from curated data because of the static nature of such inferred mechanisms. Dynamic simulations are often made on networks from curated data because of the stoichiometry and the mechanistic information they carry.

- tools for **multi-scale physiological modeling** (e.g., JSim, PhysioDesigner, GENESIS). These tools allow the development of models describing the association between genetic polymorphisms and network dynamics associated with such polymorphisms that are responsible for physiological traits and diseases ([105]).

In light of such diversity in the tools used in systems biology, it is clear that the emergence of analytical platforms and bioinformatics tools is at the core of the development and application of systems biology.

1.4 Data-driven Gene Regulatory Network Inference

1.4.1 Gene expression regulatory mechanisms of eukaryotic cellular systems

The functional activities of a cell originate from information that is encoded in the DNA. However, cell behavior depends ultimately on regulatory mechanisms that influence gene expression at the transcriptional, post-transcriptional, translational, and post-translational level. Regulation of gene expression depends on the cell's functional state and may be also influenced by the environment. Moreover, gene expression levels are mainly determined at the transcriptional level by the activity of thousands of TFs and cofactors, chromatin modification (i.e. DNA methylation), and histone modification (Figure 1.5). Instead, at the post-transcriptional level, RNA editing and non-coding RNAs (i.e. microRNAs, miRNAs; long intergenic non-coding RNAs, lincRNAs) are key regulators of gene expression ([106],[107],[108],[109]). The control of gene expression programs is an essential and vital process for living organisms and its alteration is often associated with diseases, such as cancer ([110],[111],[112],[113]). DNA mutations in CIS-regulatory elements (i.e. enhancers, promoters) or TRANS-regulatory elements (i.e. TFs, co-factors), as well as in chromatin modifiers, can have profound effects on gene expression patterns, with relative pathological consequences. Indeed, the association between mutations in these regulatory elements and cancer has been extensively demonstrated ([114],[115],[116],[117]). For example: i) aberrant overexpression of the TAL1 TF in T cell acute lymphoblastic leukemia leads to an increase in its transcriptional activity and activation of oncogenic pathways ([118],[119]); ii) amplification and overexpression of c-Myc, which controls transcription of genes involved in cell proliferation, cell growth, differentiation and apoptosis, leads to the activation of molecular pathways that are involved in cancer ([120]); iii) loss of RB gene function leads to a pro-tumorigenic activation of the E2F TF activity ([121]).

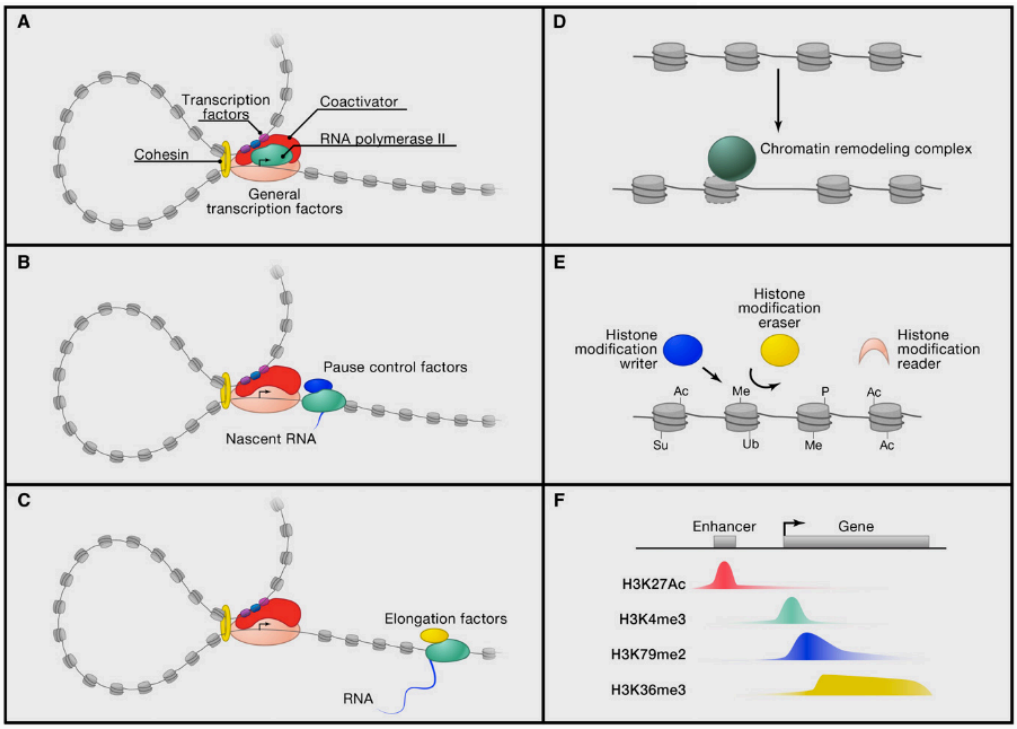

Figure 1.5: **Mechanisms of transcriptional regulation.**
The full set of mechanisms of transcriptional regulation is reported.
Binding of transcription factors to enhancer sequences and to coactivators. RNA polymerase II (RNA Pol II) binds to TF coactivator complexes at the transcriptional start sites. The loop formed between the enhancer sequences and the start site at genomic level is stabilized by cofactors (e.g. mediator complex and cohesin) and is necessary for transcription.
B. Initiation of the transcriptional activity by RNA Pol II. The initiation site is the starting point for RNA polymerase II activity. Pause control factors stop RNA Pol II activity approximately 10 base pairs downstream of the initiation site.
C. Elongation of the mRNA molecule after removal of the pause control factors. Different elongation factors and cofactors allow the RNA Pol II to proceed and elongate the mRNA molecule.
D. Accessibility to the DNA molecule. ATP-dependent remodeling complexes act on the nucleosome allowing the transcriptional complex access to DNA regions to be transcribed.
E. Histone components of nucleosomes are modified by proteins. This modification influences transcriptional activity and can be summarized into five types of modifications: acetylation (Ac), methylation (Me), phosphorylation (P), sumoylation (Su) and ubiquitination (Ub). The modifications are added by proteins called writers and they are removed by proteins called erasers. Readers proteins are able to bind DNA via these modifications.
F. Patterns of transcriptional activity determine the histone modifications. Patterns of histone modifications relative to actively transcribed genes are reported as examples: the histone H3 lysine 27 acetylation (H3K27Ac), histone H3 lysine 4 trimethylation (H3K4me3), histone H3 lysine 79 dimethylation (H3K79me2), and histone H3 lysine 36 trimethylation (H3K36me3).
Taken from ([122]).

1.4.2 Gene Regulatory Networks Inference from microarray gene expression data: limitations and challenges

Recently, the application of Systems Biology approaches to cancer "-omics" data has resulted in the identification of Gene Regulatory Networks (GRNs) representing molecular mechanisms involved in cancer. The inference of GRNs is of crucial importance to explain the homeostasis of a cell and, most importantly, to understand the effect of genomic alterations on the disruption of these regulatory networks which results in the onset and progression of diseases. Although many regulatory molecular mechanisms have been already well characterized at the biochemical and biophysical level, the availability of "-omics" data is now allowing a more comprehensive data-driven characterization of them as well as of, more generally, complex cellular systems ([123], [124],[125],[126],[127],[128]). Microarray gene expression profiles represented the commonly and widely used "-omic" data source for Gene Regulatory Network Inference (GRNi) at mRNA level. The inference of regulatory mechanisms that control the mRNA levels of a cell is based on the assumption that the functional relationship between expressed molecules generates statistical relations in the observed data. This simplification allows the application of mathematical and statistical techniques to network inference in an unbiased way, i.e. without priori knowledge on the functional relationships between the expressed genes. Specifically, if groups of genes are expressed in a cell at the same time, there is a possible functional relationship between these genes, that might be explained by statistical correlations. Different statistical frameworks have been successfully applied to infer networks of interacting genes from microarray gene expression data ([129],[130],[131],[132],[133]). Despite the great contribution of powerful statistical methods to the inference of regulatory mechanisms from microarray gene expression data, some technical issues limits the reliability of the inferred networks. In particular the available data sets lack the quantitative and statistical power to infer GRNs, i.e. the number of possible inferred interactions greatly exceeds the number of independent measurements. This is the "underdetermination" problem, also called "the curse of dimensionality" problem. To gain the statistical power necessary to generate data-driven accurate maps of regulatory mechanisms, hundreds of biological samples are needed. Consequently, it is difficult to derive reliable regulatory network models from the available data, even for small size networks according to data requirements for statistical significance. An innovative strategy to decipher GRNs was introduced by Segal et al. ([134],[135],[136]) and is based on the definition of modules of

co-expressed genes that constitute the building blocks of the GRN. The principle behind this approach is that genes that are grouped together into modules share a common regulatory program. Grouping together functionally regulated genes strongly reduces the complexity of the system that has to be modeled, and further increases the statistical power needed for regulatory network inference. Another crucial factor that affects the inference of gene regulatory networks from microarray data and in general network inference biology is the lack of benchmarking studies for biological data. Benchmarking studies are powerful tools that allow the identification of the best mathematical and statistical framework for finding true and realistic relationships between genes. Consequently, the evaluation of the accuracy of the methods for regulatory network inference is measured through simulated data that even if they represent the only possible way for the validation of the methods they do not capture the true variability of biological systems. To address all these problems, collaborative efforts have been made worldwide through public initiatives, such as the Dialogue of Reverse Engineering Assessments and Methods (DREAM) ([137],[138],[139],[140]) and Sage Bionetworks (`http://sagebase.org/`). The aim of such projects is to catalyze worldwide efforts towards the standardization and rigorous assessment of methods for cellular network inference and quantitative model building in systems biology. In particular, the DREAM project (`http://www.the-dream-project.org/`) is a promising initiative that through a yearly competition allows algorithm developers to present their own methods for network inference and it provides an unbiased assessment of these methods. From the recent DREAM competitions, it has emerged that different algorithms for reverse engineering cellular networks, highly complement each other ([141]), and that a community-based, consensus-driven, reverse-engineering approach can lead to high quality network inference. The reason why integration of reverse engineering algorithms is superior to the selection of the best performing algorithm from a pool of proposed methods, is mainly due to the compensatory effect of using multiple algorithms to balance the strengths and weaknesses of each single algorithm. In conclusion, in spite of the theoretical and technical limitations of network inference methods and strategies, network biology offers an unprecedented opportunity to interpret and reinterpret experimental findings in a global view, to unveil novel interactions and molecular regulatory processes.

1.4.3 Co-expression networks and transcription-regulatory networks

Two types of GRNs can be inferred from high-throughput gene expression data: co-expression networks and transcription-regulatory networks (Figure 1.6) [142]. In co-expression networks, nodes represent genes and edges represent connections between genes. Genes are connected to one another if they share similar expression patterns under various biological conditions. The degree of similarity between two genes can be formalized using statistical weights. Co-expression networks allow the identification of highly connected subgraphs, also called "cliques", corresponding to modules of genes having the same transcriptional profile. In transcription-regulatory networks, networks are represented as bipartite graphs, in which it is possible to identify a set of nodes representing transcription factors and a set of nodes representing target genes (i.e., modules of genes under the control of transcription factors). While in co-expression networks the relationships between genes are undirected for large scale networks, in the case of transcription-regulatory networks the edges are often directed reflecting a causal relationship between genes determined by the transcription factor regulatory program. Causal relationships in transcription-regulatory networks indicate that the observed transcriptional correlation in a module of co-expressed genes, is caused by the expression and regulation of a transcription factor on nodes representing target genes. When a set of genes is under the control of multiple transcription factors, a transcriptional program is defined.

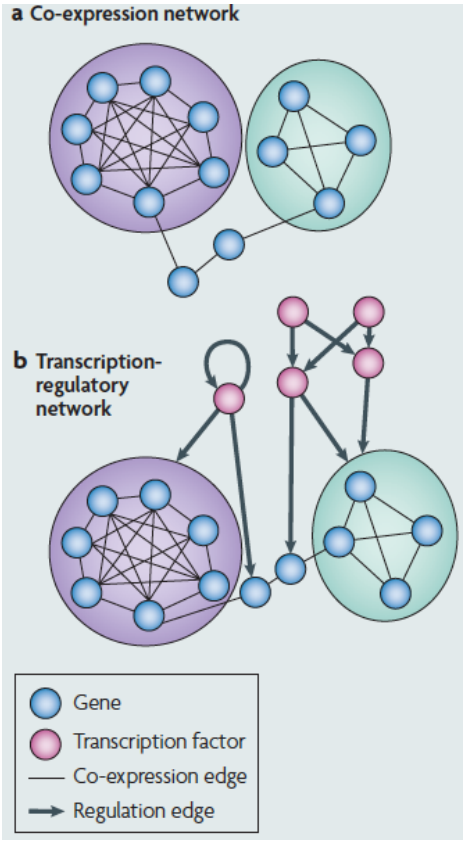

Figure 1.6: **Co-expression networks and transcriptional-regulatory networks.**
The two types of GRNs are reported: a) co-expression networks in which genes showing the same transcriptional pattern are grouped together forming modules of genes; b) transcription-regulatory networks in which regulators (i.e., transcription factors) and their target genes are distinguishable. Adapted from [142].

1.4.4 Model Gene Regulatory Networks

Microarray technology have produced a plethora of gene expression data at mRNA level ([143]), providing an unprecedented opportunity to decipher the functional regulatory mechanisms (GRNs) that control gene expression in a cell. Specifically, microarray technology allows a quantitative and simultaneous measure of the transcriptome and relative fluctuations upon a genetic perturbation ([144]), drug-induced perturbations ([145],[146]) or according to a disease state. Whereas direct experimental investigation of the functional relationship between genes is labor-intensive and time-consuming, computational analysis of gene expression profiles, through the use of statistical inference algorithms, offers a reliable alternative to explore the structure of GRNs that control molecular mechanisms in the cell. In recent years there has been an explosion in the number of computational and mathematical methods to model complex GRNs from different sources of data. Here, the generally used methods to model GRNs are reported. The mRNA cellular levels measured through the microarray technology represent the data source to model regulatory networks.

Graph Theoretical Models (GTMs)

Graph Theoretical Models (GTMs) belong to the group of qualitative network models together with the Boolean Network models, that will be discussed later, because they do not yield any quantitative prediction of gene expression in the system. GTMs are the most frequently used models to explore the structure of regulatory networks from gene expression data ([147],[148]). GTMs are used to decipher the topological structure of biological regulatory networks and are well suited for networks graphical representation or for the representation of the dynamical evolution of the networks (i.e. the topological evolution of biological networks according to the time, cellular context and conditions). In a graph structure, the network $G(N, E)$ is made up of genes as nodes $N = \{1, 2, ..., n\}$ connected by thousands of edges $E = \{(i, j)|i, j \in N\}$ which represent the relationships between the genes (Figure 1.7). The types of the relationships between the genes range from physical-interactions, i.e. the protein-protein interaction networks (PPI networks) and the DNA-protein interaction networks, to gene expression correlations for co-expression networks. Graphs can be directed (oriented) or undirected (unoriented). In the first case the gene node from which the edge starts is the precursor of the node towards the edge is directed and are represented usually by arrows. Directed graph from microarray gene expression data, can be built, for example using time-series data, because using temporal

information associated with gene expression profiles allows the inference of the causality of the connections ([149]). In the case of undirected graphs it is not possible to assign a direction, i.e. causal relationship to the edges connecting the network nodes. A graph of such type is also called "unoriented graph" in contrast with the first case in which the graph is called "oriented graph". Since undirected graphs do not imply a direct causality between the nodes the network can be built by using static (i.e. stationary) gene expression measurements. Finally, in both directed and undirected graphs, the edges can be weighted, where the weights indicate the strength of the connections.

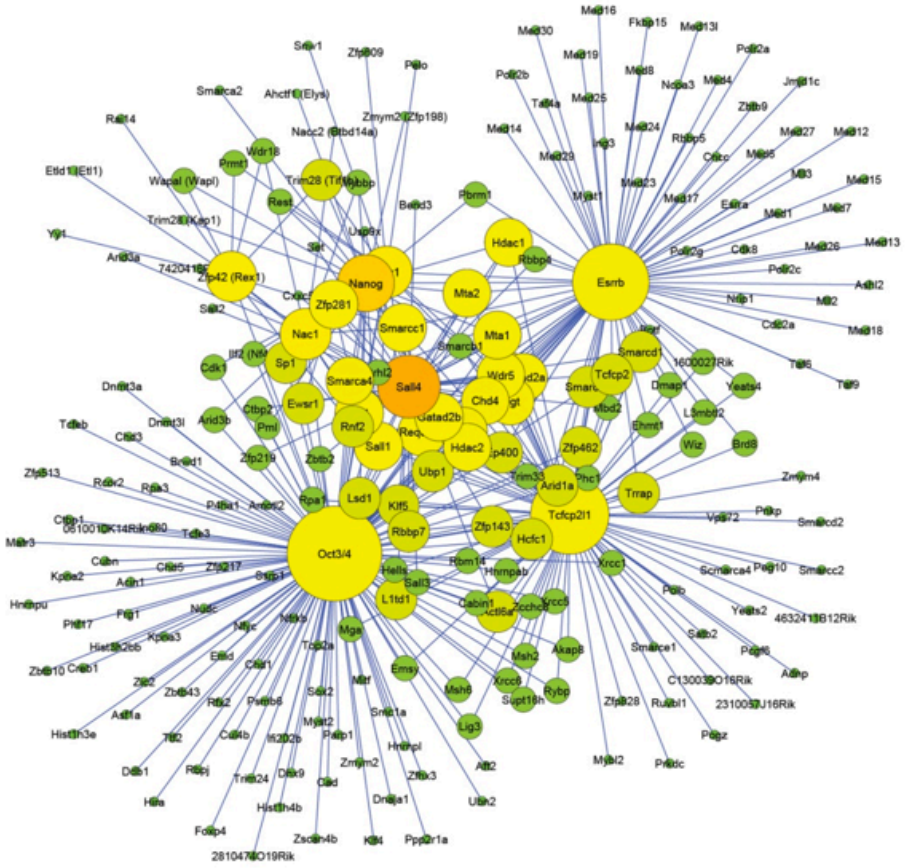

Figure 1.7: **A typical representation of a biological network with nodes and edges.** A representation of a typical biological network is reported with circles representing the genes (nodes) and blue lines (edges) representing the relationships between genes. The size of the circles varies according to the number of connections each gene establishes with the other genes. The higher is the number of connections, the biggest is the circle size.

Bayesian networks

Besides GTMs, another model that is used to explore the structure of regulatory networks from gene expression data is based on Bayesian networks, which combine probability and graph theory, i.e. they are a class of graphical probabilistic models. A Bayesian network consists of an annotated Directed Acyclic Graph (DAG) where the nodes $x_i \epsilon X$ are random variables representing genes' expression values and the edges indicate the probabilistic dependencies between the nodes. The relationships between the connected nodes are specified by a conditional probability distribution ([150],[129]) for each node given its parents: $P(x_i|Parents(x_i))$. A Bayesian network implicitly encodes the *Markov Assumption* where each gene (child node) is conditionally independent from each non-descendants given the parents nodes (genes) of the network. Besides the set of dependencies (children nodes depend on parent nodes) a Bayesian network implies a set of independencies too. The Bayesian networks allow to infer causalities between genes but at the same time they need an additional level of information with respect to the GTMs, that is the prior knowledge on the conditional relationships between the genes. In addition, Bayesian networks need the discretization of gene expression measures into a few values, generally -1 for the underexpression, 0 for no expression and +1 for overexpression. Although Bayesian networks are a valuable tool to infer functional relationships between genes and expression values, the available data suffer from the dimensionality curse (i.e. number of genes, n $\gg$ number of experiments, m) so that gene expression data are insufficient for the accurate gene network inference.

Boolean networks

The explosion of computational and mathematical methods to model complex GRNs is the result of recent progress in molecular biology and the explosion of genome-wide technologies. However a pioneering work in modeling GRNs dates back to 1960s by Stuart Kauffman and colleagues. In their work they considered an idealized random gene network because of the absence of experimental data ([151]). In this modeling attempt Kauffman defined genes as equivalent entities able to receive inputs from a variable number of K neighbors. According to the inputs that each gene receives it can be in only one of two states: the ON (1) or OFF (0). A Boolean function that governs the ON/OFF state of each gene is a statement that uses logical operators AND, OR and NOT. The output is 1 if the statement is true and 0 if it is false. A Boolean network is a directed graph $G(X, E)$ where the nodes $x_i \in X$ are boolean variables (e.g. mRNA measurements). At any given time, the state of each node represents the state of

the network. Although the dynamic properties of Boolean networks, and the reproducibility and the possibility of a finite number of states that make them attractive to model GRNs of living systems, the strong discretization of gene expression values into only two states (up and down) limits the representation of the realistic regulatory mechanisms where expression values are continuous. Another drawback concerns the high computing times (NP-complete problem, [152]) required to build biological networks. As a consequence it performs well for networks with a limited number of nodes and a small in degree value (i.e. with low levels of node connectivity). Finally, the network states are synchronous while realistic biological states of a network are generally asynchronous.

Differential Equations (DEs)

Differential Equations (DE) are used to quantitatively model complex systems. DEs describe gene expression changes as a function of the expression of the other genes and environmental factors. They are well suited to model the non-linear dynamic behaviour of GRNs in a quantitative manner through a continuous and deterministic modelling formalism. DEs are highly flexible models that allow to describe even complex relations among components. The general form of equations for the modelling of gene expression dynamics that apply Ordinary Differential Equations (ODEs) is:

$$\frac{dx}{dt} = f(x, p, u, t)$$

where $x(t) = (x_1(t), ..., x_n(t))$ is the gene expression vector of the genes 1,...., n at time t, f is the function that describes the rate of change of the state variable x_i according to the parameter set p and the external perturbations u. In network inference the function f and parameters p are derived from measured x, u and t. The identification of the model structure (f) and model parameters (p) requires specifications of the function f and constraints like prior knowledge, approximations and simplifications because without constraints there are multiple solutions to ODEs system. One of the possible constraints is for example the prior assumption of the linearity or non-linearity of the f function. Generally, regulatory processes are characterized by complex and non-linear dynamics. However, many GRNs are modeled by using linear models because of the complexity to model non-linear f function ([153]). Other variants of DEs include stochastic differential equations that consider the stochasticity of gene expression occurring especially for low cellular levels of TFs molecules ([154]).

1.4.5 Reverse engineering Gene Regulatory Networks from "genome-wide" expression data

Different approaches have been proposed for Gene Regulatory Network inference (GRNi) using "genome-wide" expression data, and a consensus on the best strategy to adopt to optimize and standardize analyses is still lacking. The heterogeneity of methodologies proposed for GRNi is strictly dependent on the variability of biological systems, which determines the adoption of different theoretical assumptions for statistical inference. However, all GRNi methods are based on the common biological assumption that mRNA measurements, can predict protein activities, and, thus, the function of regulatory mechanisms. The key concept at the basis of the gene regulatory network inference (GRNi) is that tightly co-expressed genes, i.e., genes having the same transcriptional pattern, may be functionally related. Different types of functional relationships may exist for co-expressed genes. For example, co-expressed genes may be part of the same protein complex, or may be indirectly involved in the same pathway, or they may share similar regulatory DNA sequence motifs (i.e. transcriptional regulatory sites) that allow genes to respond similarly to developmental or environmental changes. Traditionally, groups of co-expressed genes have been identified by using clustering algorithms ([155]). Gene expression clustering is still the widely used tool to analyze and visualize gene expression data. Genes sharing a similar gene expression profile are clustered together to highlight a possible functional relationship that can be direct or indirect ([156]). The degree of similarity of the expression profiles between different genes is often measured through distance metrics. One of the most frequently used metric is the Pearson correlation coefficient. Other measures of association are: the Euclidean distance, the Spearman rank correlation coefficient, the partial correlation coefficient and the Mutual Rank. The most frequently used clustering methods are: the Hierarchical Clustering (HCL) ([156]), the k-means ([157]), the Self Organizing Map (SOM) ([158]), the Principal Component Analysis (PCA) ([159]) and the Expectation Maximization algorithm ([134],[160]). The major limitation of clustering methods concerns the best number of clusters needed to partition the gene expression data. Usually, the number of clusters is not known in advance. One frequently used strategy regarding, for example, the k-means method, to overcome this limitation is to iteratively try different numbers of clusters (k) and then choose the k number that best fits the data. In addiction, another limitation of genome-wide clustering of expression profiles concerns the biological interpretation of clusters of genes. Indeed the clustering method groups together genes

that exhibit similar transcriptional responses, under different cellular conditions; however it cannot distinguish direct pair-wise transcriptional interactions from non-direct transcriptional interactions, because most of the similarity measures used by this method are linear. Despite the limitations of the cluster analysis in identifying the functional relationships between genes, it allows the user to hypothesize functions of unknown genes based on genes with known functions in the same cluster ([161],[162]). Moreover, the clustering of gene expression patterns, although not suitable for network inference of functionally related genes, allows the identification of signatures of co-differentially expressed genes according to a cell's phenotype ([45],[163],[164]). To overcome the intrinsic limitations relating to the identification of functional relationships based on cluster analysis, model networks methods are frequently applied to infer causal relationships among genes from their activity profile (See Subsection 1.4.4). Although graphical models for network inference are powerful probabilistic tools to infer conditional relationships between genes, they are affected by some limitations:

- they are based on the assumption of parametric probability distribution.
- they are powerful only with small lists of genes. For largest lists of genes ($\sim$10,000 genes) many observations are required to reliably estimate the conditional dependencies between genes, which are not available from gene expression profiling studies.
- The set of models that can be inferred from multidimensional data (like microarray gene expression data) grows superexponentially, thus, only a subset of them can be reasonably tested considering the computational power of a well-equipped research lab.
- they are based on the assumption of conditional independence that may generate unrealistic models.

To overcome such limitations, methods based on Mutual Information (MI) for GRNi were recently proposed. MI captures non-linear dependence relationships between quantitative variables, in addition to positive and negative correlations ([165],[166]). Specifically, it computes the differential entropy between differentially expressed genes. For two random variables it computes:

$$I_{ij} = S(X_i) + S(X_j) - S(X_i, X_j)$$

where $S(t)$ is the entropy of an arbitrary variable (t). For a discrete variable, the entropy is computed as follows:

$$S(t) = -\langle logp(t_i)\rangle = -\sum_i p(t_i)logp(t_i)$$

where: $p(t_i)$ is the probability of each discrete state (value) of the variable. Like many other correlation metrics, the MI measures the statistical dependency between two variables. The advantage of using MI is that it remains invariant after re-parametrization, unlike, for example, the Pearson correlation measure. In addition, using linear correlation metrics (such as Pearson correlation) the correlation coefficient might be 0 even for clearly dependent variables (non-linear relationships), while the MI is always different from 0 for statistically dependent variables ([167]). The first attempt to build networks by using MI was by Butte and Kohane (RELNET, RELevance NETworks; [168]). In this approach, genes are connected with edges only if they correlate with an MI score above a threshold established by using a permutation test. In such a way, the genes that interact indirectly will still have high values of MI scores. Recently, in the ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) algorithm ([133]) to overcome the problem of indirect interactions, the Data Processing Inequality (DPI) measure was introduced. Briefly it states that if g_1 and g_3 interact only through a third gene g_2, then:

$$I(g_1, g_3) \leq min[I(g_1, g_2); I(g_2, g_3)]$$

The MI score calculated between g_1 and g_3 will result from an indirect interaction. The advantage of DPI is that it allows the identification of indirect interactions between genes, even if they have high MI scores (i.e. significant MI scores). The ARACNE algorithm starts assigning pairwise connections to genes according to the computed MI score. Before assigning connections, a threshold of significance of the computed MI score is established. Then, the algorithm compares all the possible triplets and removes the edge with the smallest value representing the indirect interaction. The major advantage of the ARACNE algorithm (based on MI and DPI measures) with respect to graph models, is that it does not assume any restraints on the network model, even if the interaction network is greatly simplified to the unrealistic pair-wise interactions while biological interactions are multivariate and of a higher order. Other approaches have been proposed to discriminate between direct and indirect interactions. The Context Likelihood

of Relatedness (CLR) ([169])) algorithm, for example, modifies the MI score according to the empirical distribution of all MI values while the Minimum Redundancy Network (MRNET) ([170]) algorithm uses a feature selection method based on a maximum relevance/minimum redundancy criterion. All of them are based on robust and well defined mathematical and statistical formulations and are commonly used to infer reliable GRNs from high-throughput gene expression data.

1.5 Rationale of the project

Breast cancer is a complex disease. Although most breast cancer cases are curable, with a 5-year survival rate of 96% for localized (i.e. early stage) disease and 77% for regional disease, metastatic breast cancer is still incurable with a 5-year survival rate of ∼23% (SEER Cancer Statistics Review (CSR), `http://seer.cancer.gov/publications/csr.html`). The curability of early-detected and locally-recurrent non-metastatic disease is mainly due to effective treatment regimens involving surgical resection and therapeutic interventions (i.e., radiation therapy and pharmacological treatment), while for metastatic disease, pharmacological treatment is often the only option to control the growth of the tumor. Despite a steady decrease in the breast cancer mortality rate observed over the last decades, attributable to early detection and more effective treatment strategies employing predictive and prognostic biomarkers and targeted therapies, breast cancer remains the leading cause of cancer death in women worldwide ([27]). The high level of intra- and inter-tumor molecular heterogeneity of breast cancers is one of the main determinants of the failure of current therapeutic strategies that follow the "one-dose-fits-all" paradigm ([171]). In particular, the lack of knowledge, on the molecular mechanisms that underlie disease progression, on a patient-by-patient basis, represents a major hurdle to development of more effective personalized therapies for breast cancer. In this study, we developed a bioinformatic approach to identify altered transcriptional regulatory networks using a "pathway-centric" approach in order to get more insights in the biology of breast cancer. Our computational pipeline exploited the huge amount of publicly available "genome-wide" transcriptional (steady-state) data on breast cancer, and identified, through the application of sophisticated computational algorithms used in Systems Biology for the reverse engineering the Gene Regulatory Networks (GRNs), a set of GRNs that correlate, at the gene expression level, with clinical-pathological parameters of breast cancer patients (i.e. tumor grade, stage, estrogen status, prognosis, response to therapy). Specifically, we:

- identified modules of genes (i.e. Cancer Modules (CMs)) that transcriptionally correlated with several clinical-pathological parameters starting from oncogenic gene-signatures.
- Built cancer-related GRNs by assuming each gene in CMs, i.e. each CM-gene, as the "hub" gene (i.e. the highly interconnected gene) of the network. Probabilistic dependencies were inferred from the mRNA levels of the hub

gene and of the expressed genes in cancer cells, generating networks that indirectly represent much more sophisticated molecular and biochemical mechanisms of gene expression regulation.

- Performed the mutational annotation of the of the inferred GRNs in order to gain insights into the oncogenic role of the networks in breast cancer biology.
- Predicted candidate transcriptional Master Regulator (MR) genes of GRNs by performing an in-deep network deconvolution analysis.
- Described the transcriptional state of each network in breast cancer patients defining an active/inactive state according to the expression regulation of the gene neighbours with respect to the hub gene.
- Further investigated the transcriptional correlation of the active networks with chemotherapy response in Triple Negative Breast Cancer (TNBC) patients.

Using this approach we hypothesize that an in-depth characterization of the transcriptional regulatory networks that are associated with breast cancer, will allow the identification of novel predictive and prognostic biomarkers for high-resolution patient stratification, as well as the identification of new molecular targets for the development of more effective pharmacological treatments.

Chapter 2

Materials & Methods

2.1 Cancer Modules (CMs) identification

2.1.1 Oncogenic gene signatures selection

We retrieved a total of 23 transcriptional gene sets representing different oncogenic events form the Molecular Signatures Database (MSigDB) and other previously published studies (Table 2.1). Most of the gene signatures were downloaded from C2 curated gene sets (MSIGdb v.3.0; [172]), with the following exceptions (Table 2.2): VEGF signature ([173]); EGFR signature ([174]); Chromosomal Instability (CIN) ([175]); E1A signature ([176]); JAP/TAZ ([177]); JAG1/NOTCH signature ([178]).

Table 2.1: **MSigDB Gene sets details.**

Gene Sets Name (Up/Down)	Cancer gene(s)	Genes			Sample type	Organ(s) and Tissue(s)	Organism(s)
		Up	Down	Total			
M14590	ZEB1	-	-	29	Cell Lines	Breast	*H.sapiens*
M7062/M6189	HIF1A/HIF2A	41	104	146	Cell Lines	Breast	*H.sapiens*
M7160/M12455	TGFB1	106	35	141	Cell Lines	Pancreas	*H.sapiens*
M16229/M11403	TP53	48	16	64	Cell Lines	Lung	*H.sapiens*
M3456	MYC	-	-	176	Cell Lines	Blood	*H.sapiens*
M17742/M1217	TERT	128	71	199	Tissue Sample	Breast	*H.sapiens*
M16737/M3464	BRCA1	33	38	71	Tissue Sample	Breast	*H.sapiens*
M2706/M2704	E2F3	238	34	272	Tissue Sample	Breast, Ovary, Lung	*H.sapiens*
M2714/M2713	SRC	8	53	61	Tissue Sample	Breast, Ovary, Lung	*H.sapiens*
M2703/M2702	BCAT	11	73	84	Tissue Sample	Breast, Ovary, Lung	*H.sapiens*
M12029	HRAS	-	-	321	Tissue Sample	Breast, Ovary, Lung	*H.sapiens*
M2776/M6315	KRAS/PTEN	228	429	657	Tissue Sample	Lung	*M.Musculus*
M9118/M9362	KRAS	196	141	337	Tissue Sample	Lung	*M.Musculus*
M366-M3102/M8901-M1219	ERBB2	306	163	469	Tissue Sample	Breast	*M.Musculus*
M18438/M15346	E2F1	63	65	128	Tissue Sample	Liver	*H.sapiens*/*M.Musculus*
M4420/M5636	MYC/E2F1	57	66	123	Tissue Sample	Liver	*H.sapiens*/*M.Musculus*
M3432/M17372	MYC/TGFA	61	66	127	Tissue Sample	Liver	*H.sapiens*/*M.Musculus*

The description of the 17 gene sets retrieved from MSigDB is reported. Specifically are reported: the MSigDB systematic gene set name (Gene Sets Name). For each gene set two systematic gene set names are reported if the genes are divided into Up-regulated and Down-regulated genes, otherwise only one systematic name is reported; the relative official gene symbol (Cancer gene(s)) of the gene(s) on which the experimental perturbation was applied to generate the transcriptional signature; the number of genes in each gene sets (Genes) divided in up-regulated genes (Up) and down-regulated genes (Down) together with the total amount of genes in the signature (Total); the experimental source of sample (Sample type); the sample type organ and tissue of origin (Organ(s) and Tissue(s)) and the sample type organism of origin (Organism(s)).

Table 2.2: **Literature derived gene sets details.**

Cancer gene(s)	Genes			Sample type	Organ(s) and Tissue(s)	Organism(s)
	Up	Down	Total			
VEGF	-	-	58	Tissue Sample	Blood Vessels	*H.sapiens*
EGFR	-	-	487	Cell Lines	Breast	*H.sapiens*
CIN	-	-	70	Tissue Sample	Breast, Ovary, Lung	*H.sapiens*
E1A	473	16	348	Cell Lines	Breast	*M.Musculus*
JAP/TAZ	-	-	93	Cell Lines	Breast	*H.sapiens*/*M.Musculus*
JAG1/NOTCH	250	206	456	Cell Lines	Breast	*H.sapiens*/*M.Musculus*

The description of the 6 gene sets retrieved from the literature is reported. Specifically are reported: the official gene symbol (Cancer gene(s)) of the gene(s) on which the experimental perturbation was applied to generate the transcriptional signature; the number of genes in each gene sets (Genes) divided in up-regulated genes (Up) and down-regulated genes (Down) together with the total amount of genes in the signature (Total); the experimental source of sample (Sample type); the sample type organ and tissue of origin (Organ(s) and Tissue(s)) and the sample type organism of origin (Organism(s)).

2.1.2 Microarray gene expression datasets selection, quality control and normalization

Breast cancer microarray data sets (Discovery and Validation sets) and the associated clinical information were downloaded from Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`) database. We retrieved microarray data from 9 independent cohorts of patients with breast cancer. The cohorts were subdivided in a Discovery set (5 cohorts of patients) and a Validation set (4 cohorts of patients) (see Table 2.3). We then applied the following quality control procedure on .CEL files to identify flawed arrays, using Relative Log Expression (RLE) values and Normalized Unscaled Standard Error (NUSE) values ([179],[180]):

- We computed the median value and the IQR of both the NUSE and the RLE statistics for each array. This gave four values for each chip: M_{NUSE}, M_{RLE}, IQR_{NUSE} and IQR_{RLE}.
- We compared each of these values with a corresponding cutoff value. If any value exceeded the cutoff, the chip was tagged as "dubious".
- We calculated IQRs of M_{NUSE}, M_{RLE}, IQR_{NUSE} and IQR_{RLE} across the arrays. This gave four values: $IQR_{M_{NUSE}}$, $IQR_{M_{RLE}}$, $IQR_{IQR_{NUSE}}$ and $IQR_{IQR_{RLE}}$.
- If any of these IQR values was greater than $q_3 + 1.5IQR$ or less than $q_1 - 1.5IQR$ where q_1 and q_3 are the first and the third quartile of the distribution, we considered such values as outliers and flagged the corresponding array as "rejected".
- We made diagnostic plots for both the dubious and the rejected arrays for a successive visual analysis.

Cutoff values were chosen heuristically, with a preference for overestimating the number of poor quality chips rather than failing to identify a compromised chip. The four cutoff values are shown in Table 2.4. Arrays identified as defective were removed from the dataset before normalization. Further details on the number of microarrays flagged as: dubious, rejected and accepted for each dataset are reported in the next Subsection 2.1.3. After the filtering, a total amount of 1019 transcriptional profiles for the full collection of the Discovery datasets and 916 for the Validation datasets were retained. The quality control procedure was

implemented using the R (version 2.15.1) language for statistical computing and the libraries `affy`, `affyPLM` and `geneplotter` of the Bioconductor suite.

Table 2.3: **The datasets used in the breast cancer microarray screening.**

datasets	Year	Samples		Platform used	GEO acc.	Reference
		raw	filtered			
Discovery datasets						
Ivshina	2006	289	239	Aymetrix HG-U133A	GSE4922	([183])
Pawitan	2005	159	149	Aymetrix HG-U133A	GSE1456	([46])
TRANSBIG	2007	198	189	Aymetrix HG-U133A	GSE7390	([185])
Wang	2005	286	286	Aymetrix HG-U133A	GSE2034	([186])
EORTC 10994BIG 00-01 clinical trial	2007	161	156	Hu-X3P	GSE6861	([187])
Validation datasets						
Minn	2005	121	115	Aymetrix HG-U133A	GSE2603	([188])
Sotiriou(KIU)	2006	64	63	Aymetrix HG-U133A	GSE2990	([189])
Hatzis	2011	508	437	Aymetrix HG-U133A	GSE25066	([190])
Kao	2011	301	301	Affymetrix HG-U133 Plus 2.0 Array	GSE20685	([191])

The year of publication, number of samples before and after the filtering (in parenthesis), the microarray gene expression platform selected, the relative GEO accession number and pubblication.

After the quality control procedure, for each dataset the accepted microarrays were normalized using the RMA algorithm ([181]) with default parameters except for the GSE2034 dataset. In this case Affymetrix MAS5.0 algorithm was used (see "The Wang dataset" in Subsection 2.1.3.4). Microarray gene expression normalization was performed using the R (version 2.15.1) library affy of the Bioconductor suite. For each normalized gene expression dataset a gene-level annotation of the Affymetrix probe sets was performed using the probeset-gene mapping tables provided at NetAffx, the official gene-level annotation of Affymetrix probesets ([182]).

Table 2.4: **Cutoff chosen in the quality control procedure.**

$\mathbf{M}_{NUSE}$	$\mathbf{M}_{RLE}$	$\mathbf{IQR}_{NUSE}$	$\mathbf{IQR}_{RLE}$
1.10	0.2	0.10	1.0

The chosen cutoff values for the four quantities defined in step 1 of the quality control procedure.

2.1.3 Gene Expression datasets

2.1.3.1 The Ivshina dataset

The Ivshina et al. dataset included gene expression data from two cohorts of patients with primary invasive breast cancer, referred to as the Uppsala and Singapore cohorts ([183]). Affymetrix HG-U133A and HG-U133B microarrays were used for the expression profile, for a total of 578 arrays. We restricted the analysis on HG-U133A microarrays (i.e. 289 arrays) because the HG-U133B arrays were not available for all the other breast cancer data sets we selected originally. Samples from the Singapore cohort (40 samples) were also excluded from analysis due to the lack of clinical information. After the quality control analysis, 14 arrays were flagged as "dubious" of which 7 were removed after visual inspection and 3 were rejected. The clinical parameters we considered for subsequent analysis were:

- Elston (NGS) histologic Tumor Grade.
- Estrogen Receptor (ER) status.
- Disease-Free Survival (DFS).
- Disease-Free Survival in Lymph Node negative (N-) ER+ patients (N- ER+ DFS).

The number of patients per clinical information is reported in Table 2.5

Table 2.5: **Ivshina dataset clinical information detail.**

Tumor Grade	**ER status**	**N- ER+ DFS**	**DFS**
G1 (66)	ER+ (204)	Relapse (33)	Recurrence or Death (85)
G2 (121)	ER- (34)	Non-Relapse or Censored (33)	Censored (157)
G3 (55)	-	-	-

The number of patients (in parenthesis) relative to the following clinical information: the Tumor Grade (Grade 1 [G1], Grade 2 [G2], Grade 3 [G3]), the ER status (ER+, ER-), the Disease-Free Survival in Lymph Node negative (N-) ER+ patients (Relapse, Non-relapse or Censored), the Disease-Free Survival (Recurrence or Death, Censored).

2.1.3.2 The Pawitan dataset

The Pawitan dataset consists of 159 patients with primary invasive breast cancer operated at the Karolinska Hospital from January 1994 to December 1996 ([46]). We applied the quality control procedure to the dataset: 10 arrays were fagged as "dubious" of which 9 were excluded after visual inspection and one as "rejected". The clinical information (see Table 2.6) we considered for subsequent analysis refers to:

- Elston (NGS) Tumor Grade.
- Breast Cancer Relapse.
- Death due to Breast Cancer.

Table 2.6: **Pawitan dataset clinical information detail.**

Tumor Grade	**Breast Cancer Relapse**	**Death due to Breast Cancer**
G1 (28)	Relapse (38)	Dead from Breast Cancer (27)
G2 (58)	Non-relapse or Censored (112)	Alive or Censored (123)
G3 (61)	-	-

The number of patients (in parenthesis) relative to the following clinical information: the Tumor Grade (Grade 1 [G1], Grade 2 [G2], Grade 3 [G3]), the Breast Cancer Relapse (Relapse, Non-relapse or Censored) and Death due to Breast Cancer (Dead from Breast Cancer, Alive or Censored).

2.1.3.3 The TRANSBIG dataset

TRANSBIG is an international network that was launched in 2004 to promote the scientific collaboration in translational research ([184]). It comprises 39 world class institutions in 21 countries. The dataset stored in the GEO data base with accession ID GSE7390 contains 198 expression profiles of primary lymph node negative untreated breast cancer patients. A complete description of the dataset can be found in (Buyse 2006) and ([185]). We applied our quality control procedure, which led to the exclusion of 9 arrays from the dataset, leaving 189 for the successive analysis. We also checked for the presence of batch effects by plotting an array-wise boxplot using the raw data. No obvious batch effects were observed (see Figure 2.1).

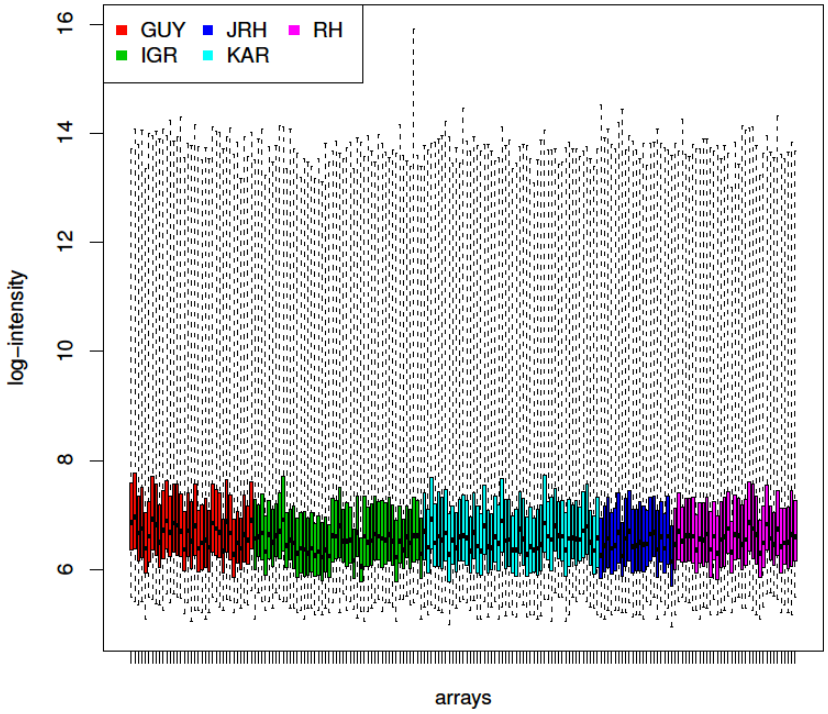

Figure 2.1: **Batch effect inspection of TRANSBIG dataset.**
Different colors indicate the different hospitals that provided patient samples.

The available clinical data for the TRANSBIG dataset includes information on:

- Elston (NGS) Tumor Grade.
- Estrogen Receptor (ER) status.
- Relapse due to Distant Metastasis (event of time to distant metastasis [E.TDM]).

- Disease-Free Survival in Lymph Node negative (N-) ER+ patients (Relapse in N- ER+).
- Overall Survival (OS).

Overall Survival was defined as time from diagnosis to death from any cause. See Table 2.7 for the number of patients with associated relative clinical information we considered for subsequent analysis.

Table 2.7: **TRANSBIG dataset clinical information detail.**

Tumor Grade	**ER status**	**E.TDM**	**Relapse in N- ER+**	**OS**
G1 (27)	ER+ (128)	Relapse (51)	Relapse (28)	Dead (56)
G2 (80)	ER- (61)	Non-Relapse (138)	Non-Relapse (100)	Alive (133)
G3 (80)	-	-	-	-

The number of patients (in parenthesis) relative to the following clinical information: the Tumor Grade (Grade 1 [G1], Grade 2 [G2], Grade 3 [G3]), the ER status (ER+, ER-), the Relapse due to Distant Metastasis (Relapse, Non-Relapse), the Relapse in Lymph Node negative (N-) ER+ patients (Relapse, Non-Relapse), the Overall Survival (Dead, Alive).

2.1.3.4 The Wang dataset

The Wang dataset ([186]) consists of 286 samples from LN-patients. This dataset presents some technical issues. First, the raw data were not available on GEO database. Thus no quality control procedure was possible. Second, the clinical information was limited to:

- The ER status.
- The occurrence of the Relapse in Lymph Node negative (N-) ER+ patients (Relapse in N- ER+).
- The occurrence of the relapse.

See Table 2.8 for the number of patients with associated relative clinical information we considered for subsequent analysis.

Table 2.8: **Wang dataset clinical information detail.**

ER status	**Relapse in N- ER+**	**Relapse**
ER+ (209)	Relapse (80)	Relapse (107)
ER- (77)	Non-Relapse (129)	Non-Relapse(179)

The number of patients (in parenthesis) relative to the following clinical information: the ER status (ER+, ER-), the occurrence of the relapse in Lymph Node negative (N-) ER+ patients (Relapse, Non-Relapse), the occurrence of Relapse (Relapse, Non-Relapse).

Another technical issue concerned the gene expression arrays normalization. As a matter of fact the available gene expression matrix was normalized by using MAS5 algorithm, while we used RMA to normalize all the other datasets. To evaluate the effect of a different normalization method on gene expression data, we renormalized our previously collected gene expression datasets with the MAS5 method, repeated the analyses, and compared the results with those obtained with the RMA normalized data. We found that the lists of significantly regulated probe sets were almost identical with similar p-values and fold changes. The addiction of this dataset in our analysis allowed us to have three datasets with information on ER status (not available in the Pawitan and EORTIC datasets).

2.1.3.5 The EORTC 10994BIG 00-01 dataset

The EORTIC dataset is composed by 161 arrays and is part of EORTC 10994 phase III breast cancer clinical trial comparing non-taxane regimen (5-fluorouracil, cyclophosphamide, epirubicin) with a taxane regimen (epirubicin, docetaxel) in women with estrogen receptor negative breast cancer ([187]). The gene expression was measured using Affymetrix Hu-X3P chip containing the whole exome. After the quality control procedure 5 arrays were flagged as dubious and retained after visual inspection while the remaining 156 arrays were classified as good quality arrays. Finally we retained 156 arrays for subsequent analysis. The clinical information we considered were:

- Elston (NGS) Tumor Grade:
  - G1 (0).
  - G2 (36).
  - G3 (68).

2.1.3.6 The Minn dataset

The Minn dataset ([188]) is composed by an initial set of 121 expression profiles. 22 expression profiles were of breast cancer cell lines and 99 were of primary breast tumors. We performed the quality control procedure only on the 99 samples of primary breast tumors. The expression profiles of the cell lines were removed because no clinical information is possible for cell lines. After quality control 4 arrays were flagged as dubious and removed after visual inspection and 2 arrays were flagged as rejected. 93 arrays were retained as good quality arrays for GSEA analysis. Gene expression analysis was performed using HG-U133A GeneChip (Affymetrix). The clinical information we considered for GSEA analysis were (see Table 2.9):

- Estrogen Receptor (ER) status
- Relapse due to Metastatic Event

Table 2.9: **Minn dataset clinical information detail.**

ER status	Relapse due to Metastatic Event
ER+ (52)	Relapse (26)
ER- (41)	Non-Relapse (52)

The number of patients (in parenthesis) relative to the following clinical information: the ER status (ER+, ER-), the occurrence of relapse due to the Metastatic Event (Relapse, Non-Relapse).

2.1.3.7 The Sotiriou dataset

The Sotiriou dataset contains information on 189 patients with primary operable invasive breast cancer. The frozen tumor specimens were obtained from two institutes: the John Radcliffe Hospital (Oxford, UK) and the Uppsala University Hospital (Uppsala, Sweden). RNA samples from Oxford (101 total samples) were processed at the Jules Bordet Institute in Brussels, Belgium ([189]). For the Uppsala samples (88 in total), RNA was extracted at the Karolinska Institute and processed at the Genome Institute of Singapore. Some of the patients were treated with tamoxifen while others were not. Table 2.10 shows the partition of the samples with respect to the institute of origin and treatment.

Table 2.10: **Sotiriou dataset institutes of origin.**

Origin (label)	Samples	Treatment
Uppsala (KIT)	24	YES
Oxford (OXFT)	40	YES
Uppsala (KIU)	64	NO
Oxford (OXFU)	61	NO

Partition of the samples in the Sotiriou dataset with respect to the institute of the origin and the treatment.

Gene expression analysis was performed with Affymetrix Human Genome U133A microarray platform. We performed the quality control procedure on the Sotiriou dataset, which led to the exclusion of 5 arrays, leaving 184 for the successive analysis. The fact that the samples were from different institutes and had undergone

different manipulations was a reason of concern. We checked for the presence of batch effects by analyzing the distribution of the measured signal before and after the normalization procedure (raw and normalized signal). Unfortunately even after the normalization, the difference between the Oxford and the Uppsala data sets was still present: the Uppsala and the Oxford samples form two perfectly separated groups. Specifically the OXFT and the OXFU form two separate groups while there is no difference between the KIT and the KIU groups. We, therefore, concluded that it was inappropriate to treat the data as a single data set. Thus, we analyzed only the arrays from the Uppsala group that had passed the initial quality control and specifically the 63 (on 64 total) samples from the KIU group. We excluded the samples of the KIT group because of the clinical information were not available. The clinical information we considered for GSEA analysis were (see Table 2.11):

- Elston (NGS) Tumor Grade.
- Estrogen Receptor (ER) status.
- Relapse Free Survival [RFS].

Table 2.11: **Sotiriou dataset clinical information detail.**

Tumor Grade	**ER status**	**RFS**
G1(26)	ER+ (53)	Relapse (11)
G2(27)	ER- (10)	Non-Relapse (52)
G3(9)	-	-

The number of patients (in parenthesis) relative to the following clinical information: the Tumor Grade (Grade 1 [G1], Grade 2 [G2], Grade 3 [G3]), the ER status (ER+, ER-), the occurrence of Relapse (Relapse, Non-Relapse).

2.1.3.8 The Hatzis dataset

The Hatzis dataset ([190]) is composed by 508 total patients with newly diagnosed ERBB2 (Her2 or Her2/neu) negative invasive breast cancer treated with taxane and anthracycline (and endocrine therapy for estrogen receptor positive patients) in neoadjuvant and adjuvant regime in a multicentric study. They are divided into the Discovery and Validation set. The Discovery set comprises 310 patients while the Validation set comprises 198 patients. All gene expression microarrays were profiled in the Department of Pathology at the M. D. Anderson Cancer Center (MDACC), Houston, Texas with Affymetrix Human Genome U133A microarray platform. We performed the quality control on the full set of patients: the discovery and the validation set. Three chips were classified as bad quality chips and they were rejected while 68 were classified as doubt chips. We decided to reject all the dubious chips because this dataset was used to validate the cancer modules identified through the GSEA analysis. This allowed us to avoid the effects of technical artifacts on the validation step of our pipeline. Finally we retained 437 arrays for subsequent analysis. The clinical information we considered were (see Table 2.12):

- Elston (NGS) Tumor Grade.
- Estrogen Receptor (ER) status.
- Distant Relapse Free Survival [DRFS].

The Distant Relapse Free Survival was defined as "the interval from initial diagnostic biopsy until diagnosis of distant metastasis or death from breast cancer, non breast cancer or unknown causes".

Table 2.12: **Hatzis dataset clinical information detail.**

Tumor Grade	**ER status**	**DRFS**
G1(30)	ER+ (245)	Relapse (100)
G2(146)	ER- (186)	Non-Relapse (128)
G3(224)	-	-

The number of patients (in parenthesis) relative to the following clinical information: the Tumor Grade (Grade 1 [G1], Grade 2 [G2], Grade 3 [G3]), the ER status (ER+, ER-), the Distant Relapse Free Survival [DRFS] (Relapse, Non-Relapse).

2.1.3.9 The Kao dataset

The Kao dataset ([191]) is composed by 327 total patients. 312 breast cancer patients were diagnosed and treated between 1991 and 2004 at the Koo Foundation Sun-Yat-Sen Cancer Center (KFSYSCC) while 15 lobular breast carcinoma samples were collected between 1999 and 2004 at the KFSYSCC. All patients received radiotherapy, adjuvant chemotherapy, and/or hormonal therapy if indicated, after the surgical resection of the tumor. Patients with locally advanced disease received the neoadjuvant chemotherapy too. Gene expression analysis was performed with Affymetrix Human Genome U133 Plus 2.0 platform. After the quality control analysis, 26 arrays were classified as doubt chips and were removed. Finally we retained 301 good quality chips. The clinical information we considered were (see Table 2.13):

- Relapse (Metastatic Event).
- Survival.

Table 2.13: **Kao dataset clinical information detail.**

Relapse (Metastatic Event)	**Survival**
Relapse (74)	Dead (74)
Non-Relapse (227)	Alive (227)

The number of patients (in parenthesis) relative to the following clinical information: the Relapse due to a metastatic Event (Relapse, Non-Relapse), and the Survival (Dead, Alive).

2.1.4 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) was performed using the R-GSEA program (GSEA-P-R.1.0). The Signal2Noise (S2N) metric was used for gene ranking according to the gene expression regulation across the phenotype labels. The statistical significance of the Enrichment Score (ES) was assessed through an empirical phenotype-based procedure consisting of 1,000 permutations of the Phenotype Labels (PhLs). In addition, we randomly selected a set of 20 lists of genes by using, as universe, the 4,697 unique genes derived from all the collected 23 signatures. These random gene lists were used as an additional control for the significance of our results (see Results, Subsection 3.1.1). The initially collected 23 oncogenic gene signatures and the 20 randomly generated gene sets represent our *a priori* defined set of genes S while each gene expression cohort previously retrieved, preprocessed, normalized and annotated for gene symbols represents the Expression datasets. All the remaining parameters were as by default.

2.2 GRN inference analysis on Cancer Modules genes (CM-genes)

Network inference on cancer module genes (CM-genes) was performed using ARACNE algorithm ([133],[131]). CUDA-MI ([192]) and WGCNA ([193]) algorithms were used to benchmark the pairwise interactions inferred using ARACNE. The Loi et al. breast cancer dataset ([194]) was used for Gene Regulatory Network inference (GRNi). This dataset was an independent transcrptional dataset with respect to those previously used for the GSEA analysis in order to avoid data overfitting, and it was large enough ($\geq$ 100 samples) to account for ARACNE network inference analysis requirements ([133]). Microarray gene expression profiles (Affymetrix HG-U133A array 2.0 array) were retrieved from GEO database, (`http://www.ncbi.nlm.nih.gov/geo/`) at the following accession number: `GSE6532`. We applied the quality control procedure previously described (see Subsection 2.1.2) retaining 327 microarrays classified as good quality microarrays for subsequent analysis. No "doubt chips" or "rejected chips" were flagged. Microarray chips were normalized by using MAS5 algorithm (Affymetrix 2002) because, as demonstrated by Lim et al. ([195]) it provides the most faithful cellular network reconstruction (i.e. with a reduced fraction of False Positive (FP) inferred interactions) respect to the normalization performed with RMA, GCRMA and Li-Wong ([196]) representing the frequently used normalization methods for microarray data. The normalized data were log2 transformed and Affymetrix probesets were annotated to Unique Gene Symbols (see Subsection 2.1.2). Gene expression relative to 13,211 unique gene symbols and 1,258 not annotated probe sets was finally considered for network inference analysis. For each network inferred around each single CM-gene by using ARACNE, CUDA-MI and WGCNA algorithms, the first 100 best correlating genes (neighbors) having the highest measured correlation score with respect to the hub gene (CM-gene) were considered. The concordance analysis was performed to measure the extent of the overlap of the inferred networks by using the three independent algorithms. The Cohen test (R, version 2.15.1) was used to test significance of the concordance analysis results.

2.2.1 GRN inference analysis using ARACNE

The adaptive partitioning algorithm was used for gene expression values discretization. The statistical p-value for mutual information (MI) threshold was set to 1e-7, while the data processing inequality (DPI) tolerance was set to 0.1 (default settings). One thousand bootstrap iterations were performed to assess the statistical significance of the inferred transcriptional pairwise gene interactions. The p-value threshold to construct the consensus network after bootstrap analysis was set to 1e-6. A total of 1,516 genes (i.e. representative of the 7 major Cancer Modules) out of 1,652 total genes in Cancer Modules were used as marker list (hubs) for network inference. We further reduced the gene list to 1516 genes because the FANCD2 gene was not represented on the Affymetrix HG-U133A 2.0 chip but it is represented on the Affymetrix HG-U133B 2.0 array. The analysis was performed on a Linux cluster of 12 nodes with a total of 288 hyperthreaded cores and 1.2 TB RAM memory.

2.2.2 GRN inference analysis using CUDA-MI

The number of bins for gene expression data discretization and the order of spline functions were set to 10 and 3, respectively. The algorithm was run on a NVIDIA Tesla C3050 GPU machine.

2.2.3 GRN inference analysis using WGCNA

Soft thresholding power estimate for scale-free topology approximation was set to 8. The resulting adjacency matrix was used. WCGNA R package was used to infer GRNs.

2.3 Cancer Modules (CMs) somatic mutation annotation

CM-genes were annotated for the presence of somatic mutations using the Catalogue Of Somatic Mutations in Cancer (COSMIC `http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/`) and The Cancer Genome Atlas (TCGA `http://cancergenome.nih.gov/`) repositories. All the statistical tests were performed using the R (version 2.15.1) library `stats`.

2.3.1 Cancer Modules (CMs) somatic mutation annotation: COSMIC

Only the breast cancer mutations relative to primary tumors annotated in the Cancer Gene Census of COSMIC v61 database were considered for the annotation for a total of 5,416 mutations (238 genes). The following criteria were applied in order to further identify a subset of candidate mutations for the annotation:

- Available PubMed ID.
- Available genomic position.
- Validation of the mutation as somatic mutation.
- Mutation identified in primary tumors (i.e. tumor origin).

From the initial set of 5,416 breast cancer mutations in Cancer Gene Census, 1,781 final somatic mutations were retained for a total amount of 238 unique genes. The final set of mutated genes was used for the annotation of 1,652 genes composing the Cancer Modules. To check for the statistical significance of the enrichment of mutated genes in Cancer Modules, a set of 1,000 random lists of 1,652 genes length was generated through sampling with replacement using 19,639 genes as starting universe (i.e. all the genes annotated in the original microarray platforms used for CMs identification: Affymetrix HG-U133A and HuX3P chip). A Shapiro-Wilk test to check for the normality of the distribution of annotated mutated genes in random lists was performed with a p-value < 0.05 (2.2e-16). The enrichment of mutated genes in CMs was evaluated by performing a proportion test between the total amount of mutated genes of each random list and the

total amount of mutated genes in the Cancer Modules. A Benjamini-Hochberg multiple testing correction was applied on the full set of comparisons. Only the comparisons showing an FDR$<$ 0.01 were considered statistically significant.

2.3.2 Cancer Modules (CMs) somatic mutation annotation: TCGA

The mutational annotation was performed by using the breast invasive carcinoma [BRCA] dataset (CGA Network, 2012, `http://www.cbioportal.org/public-portal/`). A total amount of 7,136 unique mutated genes on 463 total complete tumors (available data relative to: mRNA, miRNA, methylation, CNA, whole exome-sequencing) were used for the mutational annotation. The statistical analysis pipeline to check for a the significance of the enrichment of mutated genes in Cancer Modules was as previously described for COSMIC mutational annotation (see Subsection 2.3.1).

2.3.3 Mutational annotation of GRNs and mutual mxclusivity analysis

Mutual exclusivity analysis was performed on the gene neighbors of the GRNs found to be mutated according to the mutational annotation performed using TCGA [BRCA] data (CGA Network, 2012, `http://www.cbioportal.org/public-portal/`). Mutational data relative to only complete tumors (available data relative to: mRNA, miRNA, methylation, CNA, whole-exome sequencing) were used, for a total amount of 463 tumor samples matching the selection criteria and 7,136 unique mutated genes. For the generation of the 1,000 random gene lists we used the genes represented on the Affymetrix HG-U133A platform as universe prior to random lists generation because the networks from CM-genes were inferred from Affymetrix HG-U133A transcriptional profiles (see Subsection 2.2.1). This allowed us to be consistent with the network inference analysis we performed, avoiding the effect of the overrepresentation of genes in the random set due to a different universe of genes. We firstly annotated the 22.215 Affymetrix HG-U133A probe sets to 14,469 unique gene symbols that were subsequently mapped on the TCGA dataset gene symbol annotation. This was done in order to remove the discrepancies of gene names annotation between two different sources of data. Finally, the universe was composed by a curated list of

13,397 genes. Both, random lists and the inferred networks were annotated for the presence of mutated genes. Multiple mutations per gene were considered only once to avoid the bias of different frequency of mutation per gene. We performed a two-tailed proportion test (with 95% CI) followed by a BH correction for multiple testing in order to investigate for a statistically significant enrichment (i.e. in terms of number of genes) of mutually exclusive mutated genes in the inferred GRNs respect to the random gene lists. Specifically we compared the number of genes having a mutually exclusive pattern in each GRN with the number of genes with a mutually exclusive pattern in each one of the 1,000 gene lists. The statistical tests were performed using R version 2.15.1. Only the comparisons showing an FDR< 0.01 were considered statistically significant.

2.4 The Concordance analysis

The occurrence of gene expression patterns defining the transcriptional activity (activation/inhibition) of the inferred GRNs and MR-gene networks was determined by a scoring strategy. By using this approach, it was possible to determine the concordance of the GRNs in terms of the gene expression regulation of the gene neighbors. The concordance (i.e. the transcriptional activity of the networks) was evaluated on the transcriptional profiles of 997 breast tumors (Metabric study Discovery set [36]). The pre-processed gene expression matrix was used. The Illumina Human WG-v3 probes were annotated to human gene symbols by using HUGO gene name annotation. The absolute intensities expressed on a logaritmic scale were transformed to obtain a log ratio of the gene expression measures by subtracting raw-wise (gene-wise) the median expression from the measurement in each sample. For further details on the scoring strategy see Results, Subsection 3.5.1.

2.5 Gene set analysis of transcriptionally active networks in Triple Negative Breast Cancer (TNBC) patients

The gene expression dataset for the enrichment analysis was retrieved form GEO at the following accession ID GSE25066 (`http://www.ncbi.nlm.nih.gov/geo/`). It contains a discovery and a validation cohort of patients. In both cohorts, patients were treated with neoadjuvant taxane-anthracycline chemotherapy (NACT). The gene expression profiles refer to human breast tissues before any systemic therapy. For our analysis both cohorts of patients were considered (508 total patients) to increase statistical power. Raw data were downloaded and preprocessed (see Subsection 2.1.2 for the quality control procedure, normalization and gene symbol annotation). After the quality control procedure we retained 437 total expression profiles. The gene expression matrix was raw-wise (gene-wise) centered on the median gene expression value across all samples by subtracting for each gene and for each patient the median gene expression value from the log2 normalized expression measures. The gene set analysis was performed by using two independent computational methods: the GSEA ([197]) and the GSA ([198]). GSEA analysis was run by using the R-GSEA program (GSEA-P-R.1.0) and the GSA analysis by using the R software package `GSA`. The analysis was performed on data normalized according to three different normalization procedures: the RMA, the MAS5 normalization and the normalization reported in [190] on a set of 152 Triple Negative Breast Cancer (TNBC) transcriptional profiles passing the quality control procedure. TNBC patients were selected, from the associated clinical data, according to the histopathological negativity to ER, PgR and Her2 receptors. The pathologic response following NACT was considered as clinical variable for gene ranking. Specifically, the enrichment of the transcriptionally active networks was evaluated respect to the pathologic complete response (pCR; 52 patients) and Residual Disease (RD; 100 patients) based on RECIST criteria. For the GSEA analysis parameters setting see Subsection 2.1.4. For the GSA analysis we used the *maxmean* method for summarizing a gene set as by default and we performed 1,000 permutations to estimate false discovery rates.

Chapter 3

Results

A synthetic representation of the computational pipeline used to infer breast cancer-related GRNs from microarray gene expression data.

To infer breast cancer-related GRNs from microarray gene expression data we used a pathway-centric approach (Figure 3.1) in which we firstly retrieved publicly available oncogenic gene sets representing collections of genes whose expression levels are modulated according to an experimentally induced genetic perturbation on known oncogenes and tumor suppressor genes. We then investigated, through the Gene Set Enrichment Analysis (GSEA), the enrichment of the oncogenic gene sets in breast cancer microarray expression profiles from tumor patients (Discovery set) according to a set of clinical-pathological parameters (i.e., tumor grade, ER status, nodal status and prognosis) used as Phenotype Labels (PhLs). We then identified, by clustering, groups of "core genes" (i.e. Cancer gene Modules, CMs) from the enriched oncogenic gene sets whose expression significantly correlated with the pathological variables we used. We confirmed the enrichment of the CMs according to the clinical-pathological variables we used before, by performing the GSEA analysis of CMs on an independent set of breast cancer microarray expression profiles (Validation set). GRNs were then inferred by assuming each gene in CMs (i.e. CM-gene) as marker or 'hub" gene of the network and by searching for all possible gene neighbours among all the expressed genes in cancer cells. Networks were inferred by using primarily ARACNE algorithm ([133]). We also performed a network inference analysis by using two independent algorithms (CUDA-MI ([192]) and WGCNA ([193]) in order to confirm the transcriptional statistical interactions predicted by the ARACNE algorithm. A statistical concordance analysis was then performed

in order to compare the gene neighbours of the CM hub gene predicted by the three independent algorithms. A mutational annotation was then performed on CM-genes and on the GRN genes, in order to gain insights into the possible role of the CM-genes as oncogenic gene drivers and of the inferred GRNs as possible oncogenic mechanisms. The mutational annotation was performed by using the mutational data from COSMIC (Census) and TCGA databases. We then performed a mutual exclusivity analysis of the mutated genes composing the GRNs in order to investigate the role of the breast cancer mutated genes (i.e. the mutational landscape of breast cancer) in a network context of predicted interacting genes. We then selected two subsets of putative biologically relevant networks in breast cancer biology: the first set of networks contains GRNs inferred from mutated CM-genes while the second set contains GRNs enriched in mutually exclusive mutated genes. From the two sets of networks, we performed an in deep network deconvolution analysis in order to identify the transcriptional "hub" of the network (i.e. the transcriptional candidate regulator), to overcome the initial bias of assuming each one of the CM-genes as the hub genes of the network. We then inferred GRNs (MR-GRNs) from the predicted transcriptional hub gene we called candidate Master Regulator (MR) gene of each CM-GRN. A transcriptional analysis of the CM and MR-GRNs inferred from the two sets of CM-networks was performed by using the Metabric breast cancer gene expression dataset in order to investigate the transcriptional profile of the networks. A concordance scoring strategy was identified to define the transcriptional activity of each one of the networks according to the number of gene neighbours having the same or the opposite gene expression modulation (up or down-regulation) with respect to the expression of the hub gene. From the concordance analysis, a subset of CM and MR-networks was selected as transcriptionally significantly active. Through the GSEA analysis, the enrichment of them was evaluated in gene expression profiles from Triple Negative Breast Cancer (TNBC) patients with Residual Disease (RD) pathological condition after neoadjuvant chemotherapy versus patients with Pathological Complete Response (pCR) in order to propose, for further investigation, putative mechanisms of gene expression regulation associated with the clinical condition and novel candidate biomarkers and drug targets for therapy.

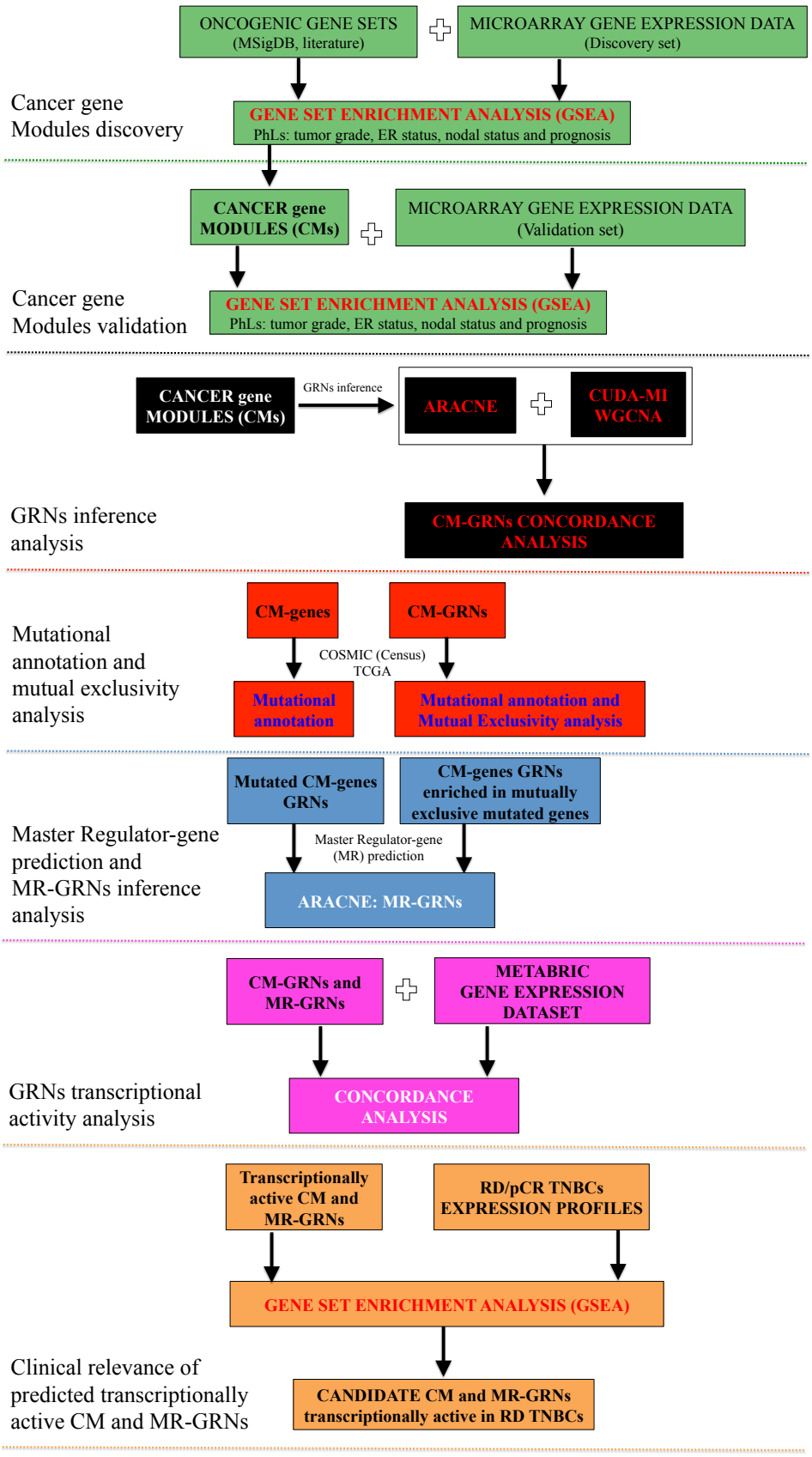

Figure 3.1: **The computational pipeline used to infer breast cancer-related GRNs from microarray gene expression data.**

3.1 Breast Cancer gene Modules (CMs)

3.1.1 Oncogenic gene sets enrichment analysis

We characterized the expression profiles of 23 oncogenic gene sets in independent cohorts of breast cancer patients, with the aim of identifying groups of genes (gene modules) whose expression significantly correlated with clinical-pathological parameters (i.e., tumor grade, ER status, nodal status, and prognosis, Figure 3.2). We preferred to use the Gene Set Enrichment Analysis (GSEA) algorithm for this characterization, since it analyzes the expression profiles of entire gene sets, representative of specific biological functions, rather than the expression profiles of individual genes. We retrieved microarray gene expression data from 5 independent cohorts of breast cancer patients (the "Discovery Set"), making a total of 1,019 patients (Table 2.3), with complete clinical-pathological information (Table 3.1). In the Discovery Set we performed GSEA to assess the enrichment (ES, enrichment score) of the 23 oncogenic gene sets in breast cancer patients stratified by tumor grade (e.g., G3 or G1), ER status (ER+ or ER-), survival (dead or alive), and relapse (regional or distant). Importantly, the GSEA enrichment analysis of the 23 gene sets was performed independently in the 5 cohorts of breast cancer patients, in order to preserve cohort-dependent biological variability. Clinical-pathological characteristics of patients were defined as "Phenotype Labels (PhLs)" and the ESs were calculated using the Weighted Kolmogorov-Smirnov test that reflects the degree of differential expression of a gene set according to the selected PhLs.

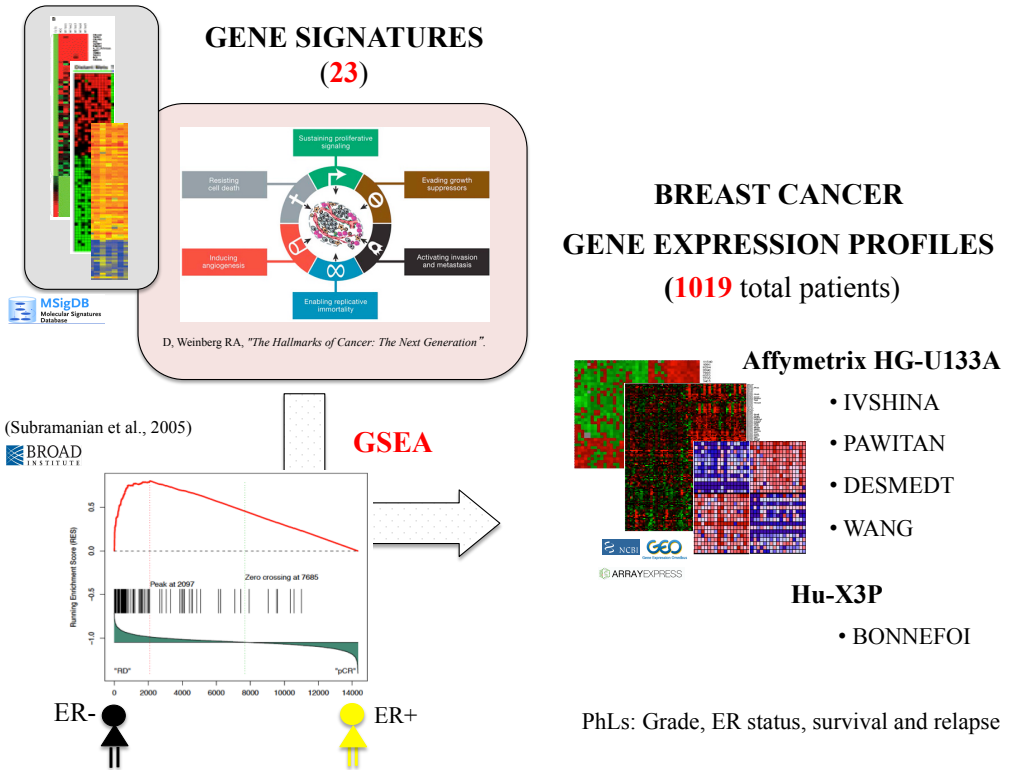

Figure 3.2: **Oncogenic gene sets enrichment strategy.**
The *in silico* pipeline relative to the enrichment analysis of the 23 oncogenic gene sets on expression profiles is reported. The transcriptional correlation of the oncogenic gene sets (i.e. groups of genes showing a coordinate change in gene expression upon experimentally induced stimuli on oncogenes and tumor suppressor genes representing the major hallmarks of cancer) was evaluated through the GSEA algorithm on 5 breast cancer gene expression datasets. 4 out of 5 expression profiles were performed on Affymetrix HG-U133A platform while 1 was performed on Hu-X3P platform. The GSEA analysis was performed by considering the following PhLs: tumor grade, ER status, survival and relapse. A typical enrichment plot is reported showing the running ES of a gene sets when comparing ER+ tumours *versus* ER- tumours.

The statistical significance of the computed ESs was estimated by a 1,000 time permutation test (as by default) in which the PhLs associated with each microarray expression profile were randomly shuffled. This allowed us to create a robust empirical null distribution of the ES measures for the nominal p-value calculation. Furthermore, to prevent errors in the inference of significantly enriched gene sets, we used the False Discovery Rate (FDR) q-value, instead of the nominal p-value, to account for the multiple testing problem. Gene sets were deemed significantly enriched if the FDR was less than 25%. In addition, we applied a normalization procedure on the full set of GSEA enriched gene sets (FDR $< 25\%$), in order to control the effect of non-homogeneous distribution of clinical-pathological parameters in the datasets (see Table 3.1). Briefly, for each gene set and for each

clinical-pathological condition (PhL), we counted the number of patient cohorts in which the gene set was enriched. This number was then divided by the total number of cohorts with the relative clinical-pathological condition (PhL). The normalization rule is summarized in the following equation:

$$S = \frac{S_{xj}}{N_j}$$

where S represents the Normalization Score; S_{xj} represents, for each significantly q-value $< 25\%$) enriched gene set (x= 1,...23), the number of cohorts in which it was enriched (j = 1,...5); and N_j represents the total number of cohorts in which the clinical-pathological variable was available (N_j = 1...5 for the Discovery Set).

We then established a threshold for the selection of enriched gene sets after the normalization procedure. To do this, we performed GSEA on a set of 20 randomly generated gene sets in the Discovery Set. After the enrichment analysis and the normalization procedure, three random gene sets appeared to be enriched (q-value $< 25\%$) according to tumor grade (in 1 out of the 11 cohorts, considering each grade independently; 9%), and to ER status (in 1 out of 4 cohorts; 25%), (Table 3.2). Therefore, according to the "Normalization Score, (S)" obtained using random gene sets, we established cut-off for selection of the final core of significantly enriched oncogenic gene sets, which should have an "S" above the 25%, i.e. the S relative to enrichment of random signatures in the ER pathological condition. We found 18 gene sets with S$>$ 25% that we considered significantly enriched (Table 3.2). This "core set" represents 78% of the initial set of oncogenic gene sets (18 out of 23). Specifically, the gene sets composing the "core set" are: E1A, ZEB1, TP53, TERT, MYC/TGFA, E2F1, MYC, KRAS, HRAS, YAP/TAZ, HIF1A/HIF2A, ERBB2, BCAT, BRCA1, CIN, E2F3, EGFR, MYC/E2F1.

Table 3.1: **Breast cancer Clinical-pathological conditions and cohort-specific samples distribution relative to the Discovery set.**

Clinical-pathological parameter	Ivshina	Pawitan	TRANSBIG	Wang	EORTC	Total
G1	66	28	27	-	-	121
G2	121	58	80	-	36	295
G3	55	61	80	-	68	264
ER+	204	-	128	209	-	541
ER-	34	-	61	77	-	172
Relapse	85*	38	51	107	-	281
Non-Relapse	157*	112	138	179	-	586
N0 ER+ Relapse	33	-	28	80	-	141
N0 ER+ Non-Relapse	33	-	100	129	-	262
Dead (Survival)	85*	27	56	-	-	168
Alive (Survival)	157*	123	133	-	-	413

Table reports the number of patients associated with specific clinical-pathological characteristics, used as phenotype labels (PhLs) in the GSEA analysis, for each of the 5 cohorts (Ivshina, Pawitan, TRANSBIG, Wang, EORTC) of the Discovery Set. Clinical-pathological parameters reported include: tumor grade (G1, G2, G3); estrogen receptor (ER) status, ER-positive (ER+) vs. ER-negative (ER-); relapse status (relapse vs. non-relapse); local/distant relapse (node-negative (N0) ER+ relapse vs. N0 ER+ non-relapse primary tumors; survival (dead or alive). *patients were assigned to both relapse and survival clinical pathological variables due to ambiguous definition by the authors.

Table 3.2: **Significantly enriched gene sets after the normalization procedure.**

Gene set	**Grade(%)**	**ER(%)**	**Relapse(%)**	**N0 ER+ Relapse(%)**	**Survival(%)**
E1A	77	100	40	75	100
ZEB1	-	25	-	-	33
TP53	55	62	20	-	33
TERT	55	100	-	-	-
MYC/TGFA	-	50	-	-	-
E2F1	9	50	-	-	-
MYC	86	58	40	25	67
KRAS	-	25	-	-	33
HRAS	-	75	-	-	-
YAP/TAZ	55	75	-	-	-
HIF1A/HIF2A	63	63	40	25	67
ERBB2	36	25	20	-	33
BCAT	18	-	20	-	33
BRCA1	55	88	-	-	33
CIN	82	100	60	100	68
E2F3	9	50	-	-	-
EGFR	-	50	-	-	-
MYC/E2F1	64	75	20	25	33
RANDOM 12	9	-	-	-	-
RANDOM 20	9	25	-	-	-
RANDOM 6	9	25	-	-	-
Number of cohorts per clinico-pathological parameter	11	4	5	4	3

The GSEA significantly enriched gene sets, along with associated normalization score (S, shown as a percentage), is reported for each clinical-pathological condition. For each condition, the number of cohorts in the "Discovery Set", for which data on the selected condition was available, is also reported. For the clinical variable "Grade" , the G1, G2 and G3 levels were considered independently and the availability of the clinical information across the cohorts of the "Discovery Set" was considered for each level. In red are highlighted the rejected gene sets according to the cut-off derived from the enrichment observed in the random gene lists RANDOM 12, RANDOM 20 and RANDOM 6 (see main text). The absence of enrichment after GSEA analysis (q-value $> 25\%$) is reported as "-" symbol.

3.1.2 Definition of Cancer Modules

We subsequently selected those genes in the 18 "core" gene sets that contributed most to the enrichment of these gene sets in the PhLs analyzed. This group of genes, identified as "core genes" by GSEA analysis is made up of 1,652 unique genes. We summarized the enrichment results for each PhL in an N x M binary matrix, in which the rows (N) represent the full set of 1,652 genes and the columns (M) represent the full list of PhLs considered. We numerically indexed the matrix with two integers: 1, indicating the gene-wise significant enrichment in a particular PhL; 0, indicating the gene-wise non-significant enrichment in a particular PhL. We then applied the Hierarchical Cluster Analysis (HCA) to the binary matrix in order to cluster together the core genes according to the relative PhLs enrichment. The cluster analysis allowed us to identify 15 groups of core genes (clusters) enriched according to each single PhL or according to combinations of PhLs (Figure 3.3). Of the full set of 15 gene clusters, we selected 7 "major clusters" that contained a number of genes greater than 5% of the total (1,652 genes). By using this criterion we wanted to prioritize more "biologically" informative clusters of genes in the specific PhLs (Figure 3.3). This subset of clusters represents our set of "Cancer Modules (CMs)", i.e., groups of genes whose expression correlates with specific clinical-pathological conditions associated with breast cancer. The number of the genes relative to each one of the 7 selected CMs is reported in Table 3.1.2 (i.e., 1,516 out of 1,652, 92%).

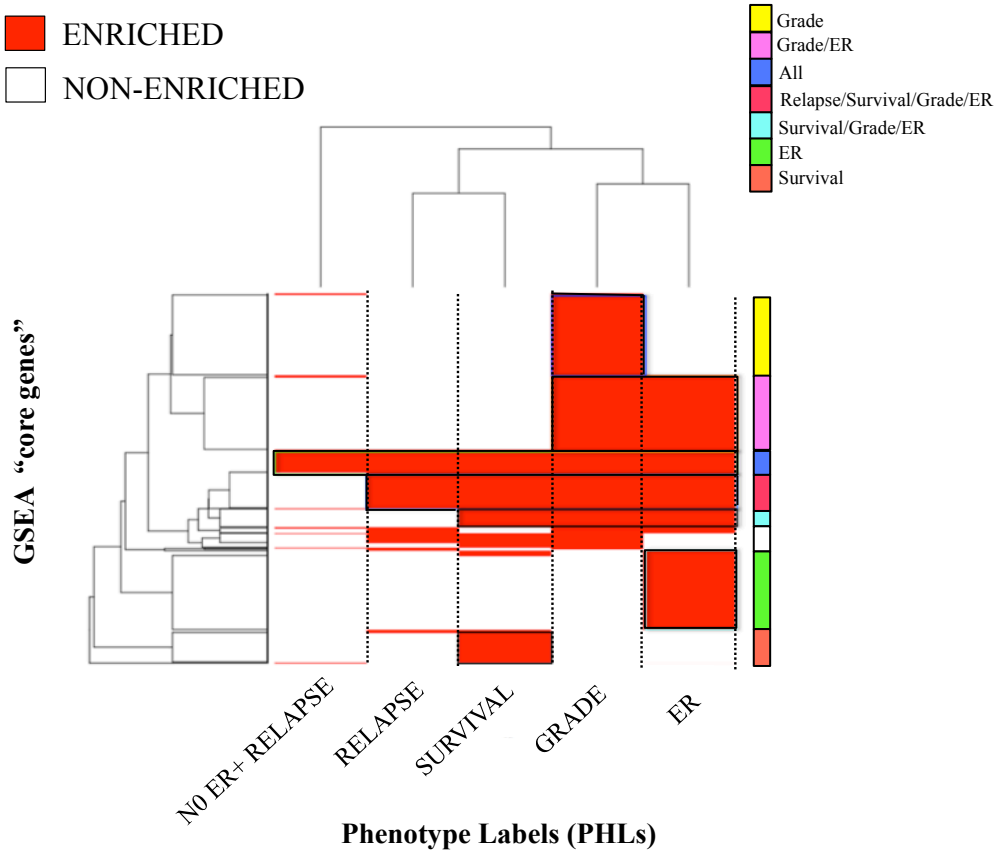

Figure 3.3: **Definition of Cancer Modules.**
The heatmap of the enrichment of the GSEA "core genes" (y-axis) belonging to the 18 enriched "core" gene sets, with respect to the PhL considered (reported on the x-axis). The "core genes" were clustered (HCL) according to the enrichment observed in each PhL. Red color indicates the enrichment of a core gene in the PhL considered (x-axis). The color bar on the right-side of the y-axis highlights the 7 CMs. From the top to the bottom the CMs are defined as: Grade, Grade/ER, All, Relapse/Survival/Grade/ER, Survival/Grade/ER, ER, Survival.

Table 3.3: **Gene content of Cancer Modules.**

CMs	**Gene(s) content**
Grade	366
Grade/ER	328
All	105
Relapse/Survival/Grade/ER	162
Survival/Grade/ER	83
ER	338
Survival	134
Tot	1516

The number of genes in each Cancer Module (CMs) is reported.

The composition of the CMs in terms of the enriched gene sets is reported in Figure 3.4. The most representative gene sets per CM are:

- Grade: MYC, E1A, ERBB2.
- Grade/ER: MYC, E1A, BRCA1, TERT, TP53.
- All: YAP/TAZ, E1A, CIN, MYC.
- Relapse/Survival/Grade/ER: MYC, E1A, HIF1A/HIF2A.
- Survival/Grade/ER: BRCA1, ERBB2, E1A, MYC/E2F1, E2F1, MYC/TGFA.
- ER: HRAS, EGFR, HIF1A/HIF2A.
- Survival: ERBB2, BCAT.

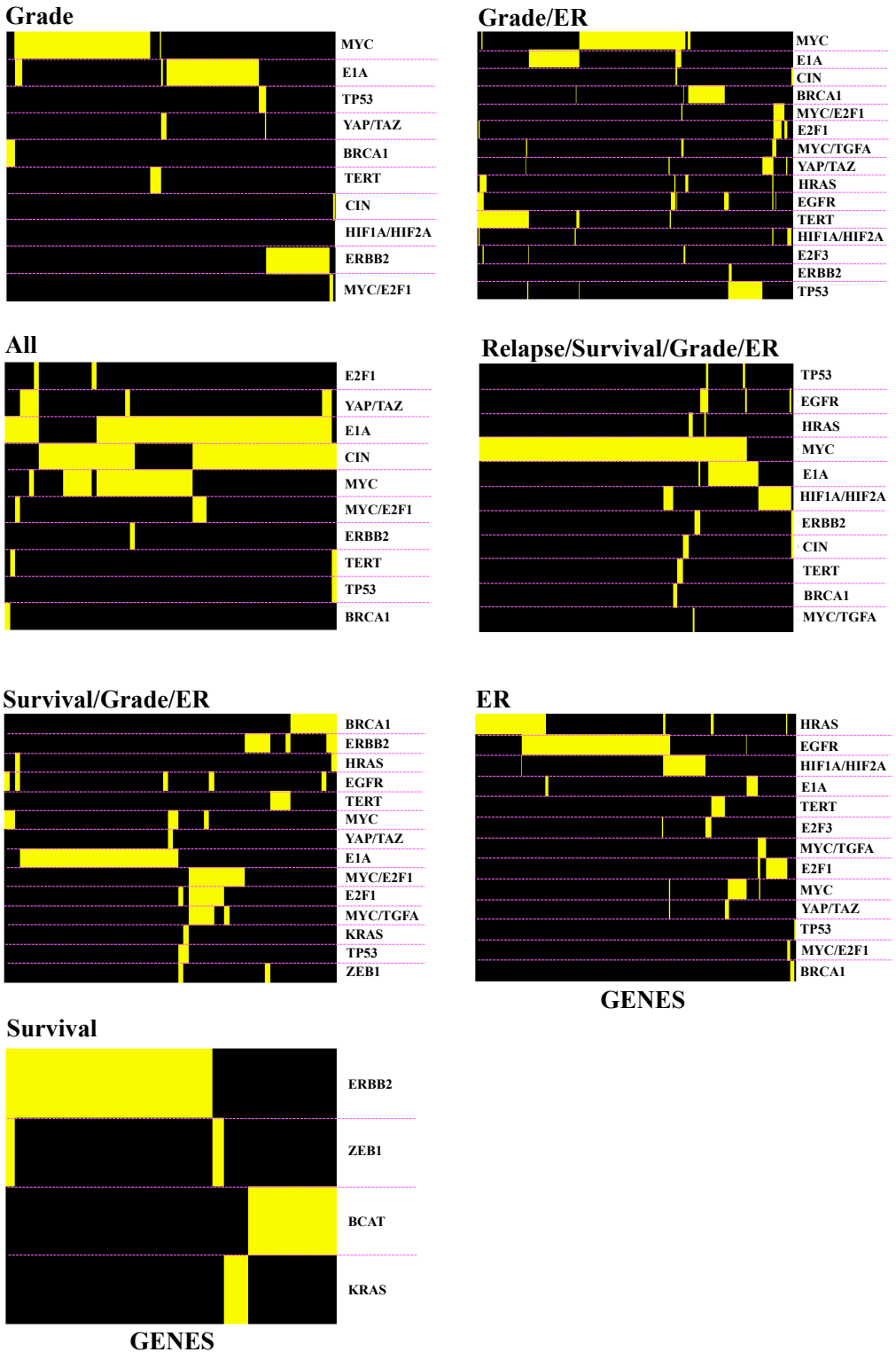

Figure 3.4: **Oncogenic gene set distribution across the Cancer Modules (CMs).** The most representative gene sets for each Cancer Module (CM) are reported. On the x-axis the genes of the relative Cancer Module are reported. On the y-axis the enriched gene sets are shown. Yellow bars indicate the fraction of the genes in each Cancer Module relative to enriched gene set reported on the y-axis. Overlapping bars indicate that the enriched gene sets share a fraction of genes that are in common.

3.1.3 Independent validation of Cancer Modules

We performed an *in silico* validation of the CMs using GSEA on an independent set of 4 breast cancer patient cohorts (the "Validation set") for a total of 916 individuals (Figure 3.5; Tables 2.3 and 3.4). This *in silico* validation was performed in order to assess the robustness of the observed CM enrichment with respect to the PhLs in an independent set of patients ("Validation set"). As for the Discovery Set, after a 1,000 times permutation test of PhLs, we defined only the CMs with an FDR q-value of less than 25% as significantly enriched. We were able to confirm 6 out of 7 CMs as significantly enriched in at least one dataset of the "Validation Set", (Table 3.5 and Figure 3.5). The enrichment we observed confirmed the association, at the level of gene expression, of the genes composing the CMs with the considered PhLs. The schematic computational pipeline used to identify and validate CMs is reported in Appendix A in green.

Table 3.4: **Clinical-pathological characteristics of breast cancer patients belonging to the 4 cohorts constituting the Validation Set.**

Clinical-pathological condition	Minn	Sotiriou	Hatzis	Kao	Total
G1	-	26	30	-	56
G2	-	27	146	-	173
G3	-	9	224	-	233
ER+	52	53	245	-	298
ER-	41	10	186	-	196
Relapse	26	11	100	74*	211
Non-Relapse	52	52	128	227*	459
N0 ER+ Relapse	-	-	-	-	-
N0 ER+ Non-Relapse	-	-	-	-	-
Dead (Survival)	-	-	-	74*	74
Alive (Survival)	-	-	-	227*	227

Table reports the number of patients associated with specific clinical-pathological characteristics, used as phenotype labels (PhLs) in the GSEA analysis, for each of the 4 cohorts (Minn, Sotiriou, Hatzis, Kao) of the Validation Set. Clinical-pathological parameters reported include: tumor grade (G1, G2, G3); estrogen receptor (ER) status, ER-positive (ER+) vs. ER-negative (ER-); relapse status (relapse vs. non-relapse); local/distant relapse (node-negative (N0) ER+ relapse vs. N0 ER+ non-relapse primary tumors; survival (dead or alive).
*patients were assigned to both relapse and survival clinical pathological variables due to ambiguous definition by the authors.

Table 3.5: **Computational validation of Cancer Modules (CMs) enrichment.**

Cancer Module	ER	Grade	Relapse	Survival
Grade	-	2	-	-
Grade/ER	3	2	-	-
All	3	2	1	0
Relapse/Survival/Grade/ER	3	2	1	0
Survival/Grade/ER	3	2	-	1
ER	3	-	-	-
Survival	-	-	-	0
Tot	3	2	4	1

Table reports the number of cohorts in which each Cancer Module (CM) shows a statistically significant enrichment with respect to the total (Tot) number of cohorts, in the Validation set, possessing the relative pathological information. The absence of a statistically significant enrichment is reported as 0, while "-" indicates the absence of datasets with the specific PhL (reported as columns).

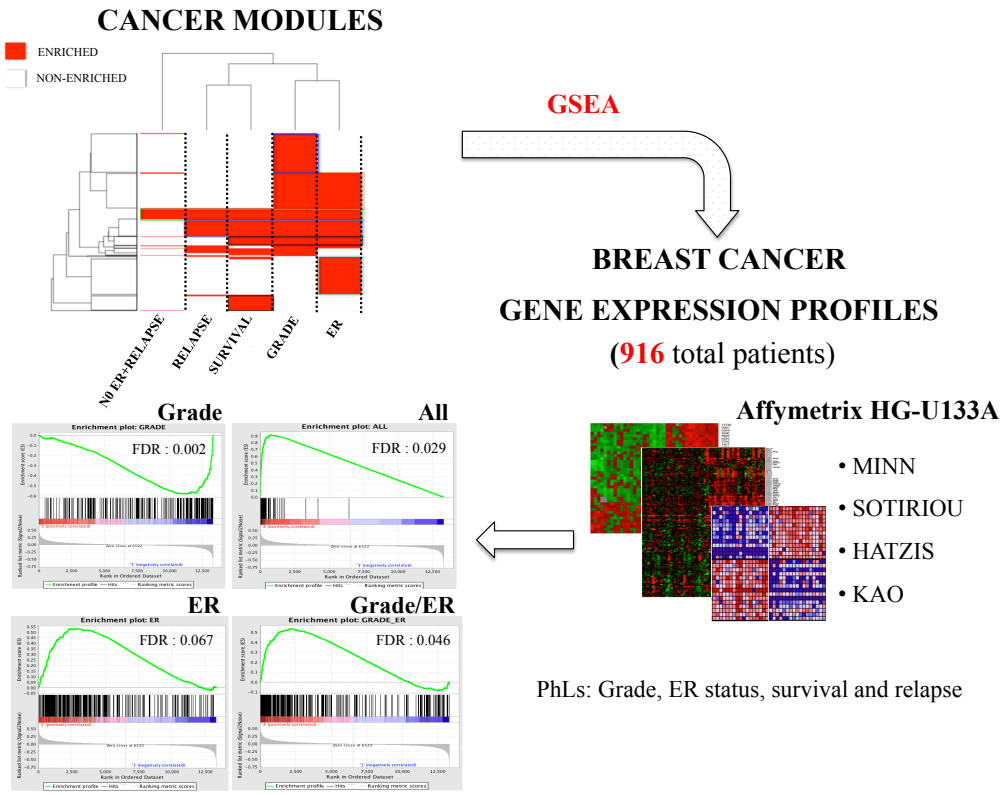

Figure 3.5: **Validation strategy for the Cancer Modules and Enrichment Results.** The *in silico* validation pipeline for the 7 CMs is reported. GSEA analysis was performed using the CMs as gene sets on 4 independent cohorts of patients, the Validation Set. The phenotype labels (PhLs) used in the GSEA analysis are: tumor grade, ER status, survival, relapse. Four GSEA enrichment plots are reported as representative examples of CM validation: Grade, All, ER and Grade/ER with associated FDR q-values. (Hatzis et al. dataset([190]), 437 patients).

3.2 Reverse Engineering Gene Regulatory Networks

3.2.1 Reverse engineering of Gene Regulatory Networks using ARACNE algorithm

We inferred networks of transcriptionally correlating genes starting from the full list of genes of the 7 CMs (see the computational pipeline in Appendix A, black section). We performed the network inference analysis in order to identify biological processes relevant to breast cancer progression that can be used for the selection of cancer biomarkers and possible novel drug targets. Specifically, we performed a network inference analysis by centering network growth on each of the CM-genes (1,516 genes in total) that represent the hub/marker gene around which the transcriptional correlations (i.e. the edges connecting two genes of a network) were inferred. Using this strategy, we built a total of 1,516 gene networks to be further investigated for their breast cancer relevance. The GRN inference (GRNi) analysis (also called network deconvolution analysis) was performed using the data-driven Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) ([131], [133])(see Subsection 2.2.1), which is based on an information-theoretic method for system-wide reconstruction of complex transcriptional networks from gene expression data. The transcriptional correlation among genes was calculated through the Mutual Information (MI) measure that estimates the amount of information one variable (gene X) contains about another (gene Y), i.e., their mutual dependency. The total number of genes in the full set of 1,516 GRNs was 14,293 genes and the MI measures spanned from a minimum of 0.10 to a maximum of 1, where the higher the MI the stronger the transcriptional correlation between two genes (i.e., the hub gene and its neighbor). In order to select, for each network, the core genes that displayed the highest degree of transcriptional correlation with the hub gene (i.e. the gene neighbors of the hub gene with the highest MI score), we investigated the shape of the distribution of the MI values for each inferred network. Due to the low degree of overlap between the distributions of the MI scores across the networks (Figure 3.6), the use of a unique MI based cut-off for the selection of the best neighbor genes with respect to the hub gene was not feasible. We then used a ranking strategy for the selection of the best correlating genes. For each network, we ranked the neighbors by their MI values, from the highest to lowest, and selected as best neighbors, the first 100 genes for each network with the highest ranked MI measure. The total number of genes in the full set of 1,516 networks was thus reduced to 11,721 unique genes.

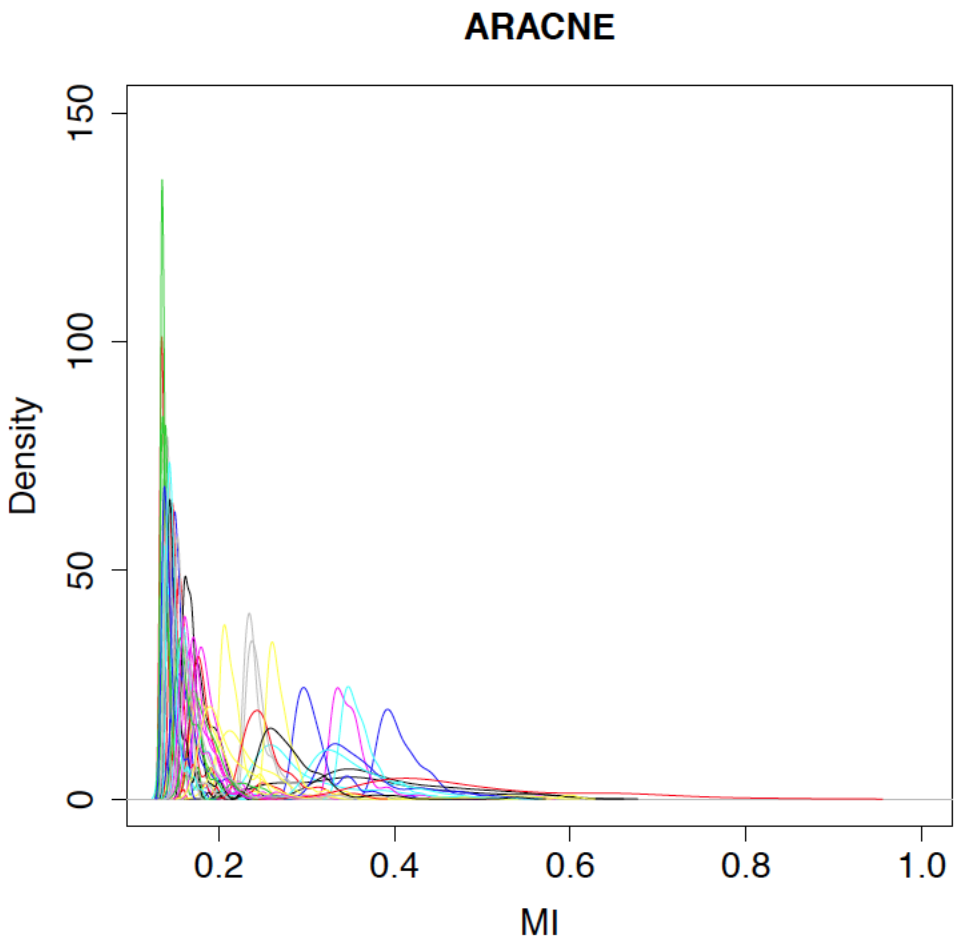

Figure 3.6: **Distributions of Mutual Information measures.**
The distributions of the Mutual Information (MI) measures of a representative subset of the full set of 1,516 GRNs is reported. Different colours highlight the distribution of MI measures relative to each single GRN. The distributions are clearly non-overlapping for MI values greater than 0.2.

3.2.2 *In silico* validation of the transcriptional correlations predicted by ARACNE

To assess the robustness of the identified GRNs, in terms of their gene content and predicted pair-wise transcriptional relationships, we performed GRNi analysis using two other independent methods: the Compute Unified Device Architecture-Mutual Information computation (CUDA-MI) and the Weighted Correlation Network Analysis (WGCNA) algorithms (see the computational pipeline in Appendix A, black section). As for the network inference analysis we performed by using ARACNE algorithm, we assumed that each CM-gene was a hub gene around which we could build the network of transcriptionally correlating genes. Specifically, the CUDA-MI software implements the MI estimation using B-spline functions proposed by Daub et al., to infer transcriptional correlations from high-throughput gene expression data ([199]). The B-spline approach is an alternative to the kernel-based MI estimation proposed by Margolin et al. ([133]); it propose the use of polynomial B-spline functions to deal with the problem of assign data points (expression values) to one bin or to the nearest one, in the discretization (binning) phase of continuous gene expression measurements, when they are extremes of the numerical range of discrete intervals (bins). For data points near to the border of a bin, in fact, small fluctuations due to biological or measurement noise might shift these points to neighbouring bins affecting the resulting mutual information, especially for datasets of small or moderate size ([199]) and generating unstable gene networks. To overcome such limitations, the gene expression data point in the B-spline approach can be assigned simultaneously to weighted multiple bins by a set of B-spline functions. The CUDA-MI software is based on the Compute Unified Device Architecture (CUDA) programming model on a graphics processing unit (GPU), in order to accelerate the B-spline function for MI estimation in the case of large datasets. The WGCNA algorithm builds weighted gene co-expression networks by "soft-thresholding" the correlation coefficient for gene-to-gene interaction predictions. Specifically, it thresholds the pair-wise connection between two genes by a number in [0,1]. Thus, the transcriptional correlation becomes a connection strength. The advantage of using such a methodology is that it preserves the continuous nature of correlation information, avoiding information loss due to dichotomization (1 = connected genes, 0 = unconnected genes), as well as sensitivity issues relating to choosing a statistical threshold. To build networks with WGCNA, we set the soft-thresholding power β to 8 after visual inspection of network indices, in order to approximate

the scale-free topology (Figure 3.7). The distributions of the correlation measurements calculated using the three independent methods, ARACNE, CUDA-MI and WGCNA, is reported for a subset of networks, as representative example (Figure 3.8). As already observed in Subsection 3.2.1 the distributions of the correlation measurements relative to each network are non-overlapping also according to CUDA-MI and WGCNA algorithms, highlighting that the inferred networks are heterogeneous between them, in terms of the degree of transcriptional correlations with the gene neighbours, despite the computational method used to infer them. Moreover, although ARACNE and CUDA-MI algorithms are both based on MI computation, we observed that the distributions of the MI measurements substantially differ between them, i.e., they are not overlapping. This behaviour might be explained not only by the different way they perform data discretization but also by the Data Processing Inequality (DPI) procedure that is implemented in ARACNE and not in CUDA-MI algorithm. By removing putative indirect interactions through the DPI (i.e. according to the DPI threshold, see Methods, Subsection 2.2.1), the set of gene neighbours of the hub genes predicted by the two algorithms may differ as well as, consequently, the MI measurements distributions. For each GRN inferred using ARACNE, CUDA-MI and WGCNA, we subsquently selected the top 100 neighbors after ranking interaction scores with the hub gene, as reported in Subsection 3.2.1.

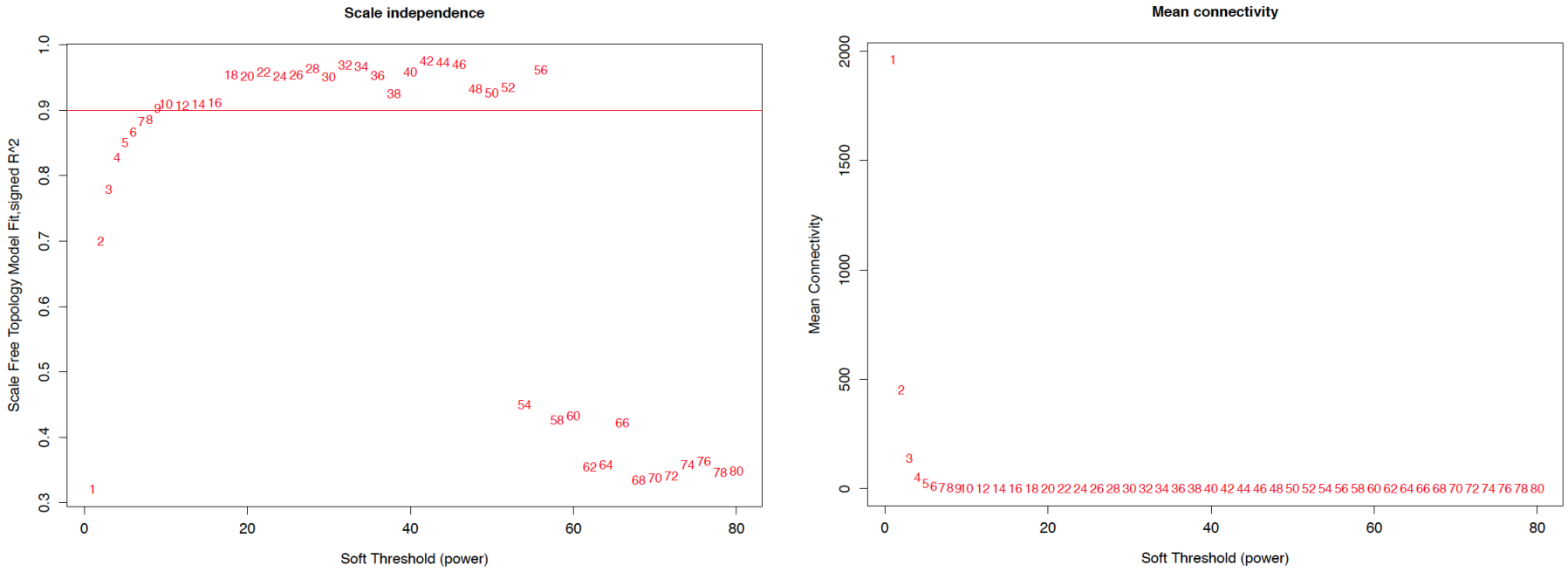

Figure 3.7: **Network topology for various soft-thresholding power indices.**
Left panel, the graph shows the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). Right panel, the mean connectivity degree (y-axis) is reported as a function of the soft-thresholding power (x-axis).

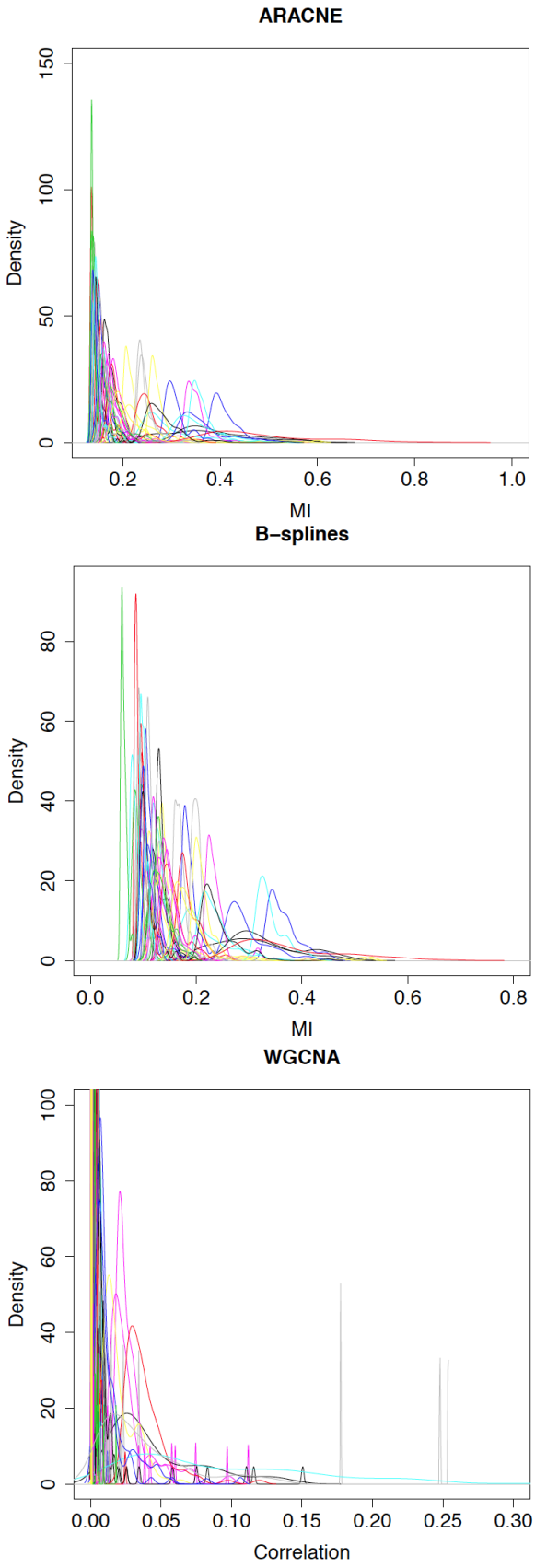

Figure 3.8: **Distribution of the correlation measurements computed by ARACNE, CUDA-MI and WGCNA algorithms for a representative subset of GRNs.**
The distributions of the correlation measurements computed respectively by ARACNE, CUDA-MI and WGCNA methods relative to a representative subset of GRNs are reported.

To assess the robustness of GRNs derived, we evaluated the concordance (i.e. gene content and predicted pair-wise transcriptional relationships) of the "first 100 network neighbors" inferred using the three independent methods. We applied the Cohen test on each triplet of networks inferred from each CM-gene. In total, we performed comparisons for 1,516 triplets. Based on the Cohen's estimated coefficients for agreement, 1,498 networks were found to be concordant (p-value < 0.05), while 18 were not (p-value > 0.05). Among the set of 1,498 statistically significant coefficients, 1,124 (75%) triplets showed a "good agreement" (i.e. with "almost perfect", "substantial" or "moderate" agreement; Figure 3.9), while 374 (25%) showed a "bad agreement" (i.e. with "slight" or "fair" agreement; Figure 3.9). Based on this concordance analysis, it emerges that a large fraction of GRNs share the same set of transcriptionally correlating genes with respect to the relative hub gene, independently of the method used for GRNi (i.e., ARACNE, CUDA-MI, or WGCNA) thus highlighting the strength of the transcriptional correlations inferred.

Figure 3.9: **Concordance analysis agreement distribution.**
The concordance analysis agreement distribution is reported. On the x-axis the number of GRNs is reported (Nets); on the y-axis the extent of agreement is reported (Agreement). Percentages indicate the number of GRNs showing the observed agreement with respect to the total set of networks (see main text). Green circles group together networks having a bad agreement, while red triangles group together networks with a good agreement.

3.3 Mutational annotation for the identification of cancer-mutated genes and mutated GRNs

3.3.1 Mutational annotation of Cancer Module genes and enrichment tests

The mutational annotation of CM-genes was performed in order to identify those genes mutated in breast cancer, which may ultimately affect biological functions predicted by our GRNi analysis. We focused on somatic mutations annotated in the COSMIC-Census relative to Breast cancer (238 genes; see Methods, Subsection 2.3.1 for details) and TCGA datasets (7136 total mutated genes; whole-genome sequencing, TCGA[BRCA], CGA Network, 2012. See Methods, Subsection 2.3.2). On a total of 1652 CM-genes, 49 were mutated according to the COSMIC-Census set of mutated genes, while 812 genes were mutated according to TCGA dataset (see the computational pipeline in Appendix A, red section). These mutated genes represent ∼0.2% and ∼4%, respectively, of the entire genome (∼20.000 genes:`http://www.ncbi.nlm.nih.gov/`), or ∼3% and ∼49%, respectively, of the full list of CM-genes (1,652 total genes). This means an enrichment in mutated genes of 10- to 15-fold in CMs with respect to the entire genome. The mutational analysis was intentionally repeated twice using the COSMIC and TCGA databases, to balance the benefits and limitations of the two approaches. Indeed, the COSMIC dataset contains mainly experimentally validated mutations, but since it is a literature-curated database it may be biased by the number of papers focused on specific sets of cancer genes. In contrast, the TCGA dataset is not biased by the literature, since it is the result of unsupervised screening by next-generation sequencing analysis; however, this dataset is likely to contain false positives because many of the mutations reported have not been experimentally validated. Importantly, 73% (35 out of 49; $p=4x10^{-38}$) of mutated CM-genes according to COSMIC, were present also in the TCGA mutational annotation. We next tested the statistical significance of the enrichment of mutated genes in the set of CM-genes. We generated a set of 1,000 random gene lists that we annotated for mutations using both COSMIC and TCGA. We then performed 1,000 runs of proportion tests (with a 99% confidence interval) to compare the number of mutated genes found in the CMs with respect to the number of mutated genes in each one of the 1,000 randomly generated gene lists. The computed p-values were adjusted for multiple comparisons by Benjamini-Hochberg correction and the relative FDR calculated (see Tables 3.6 and 3.7 for

the statistical test results relative to the COSMIC-Census and TCGA datasets, respectively).

Table 3.6: **Proportion test results for COSMIC-Census mutational annotation.**

Random gene list	Number of mutated genes	Q-value
R_1	29	0.0227
R_2	27	0.0116
R_3	26	0.0083
R_4	25	0.0058
R_5	24	0.0039
R_6	23	0.0026
R_7	22	0.0017
R_8	21	0.0011
R_9	20	0.0007
R_{10}	19	0.0004
R_{11}	18	0.0002
R_{12}	17	0.0001
R_{13}	16	8.51e-05
R_{14}	15	4.72e-05
R_{15}	14	2.55e-05
R_{16}	13	1.34e-05
R_{17}	12	6.90e-06
R_{18}	11	3.47e-06
R_{19}	10	1.73e-06
R_{20}	9	8.51e-07
R_{21}	8	4.29e-07
R_{22}	7	2.30e-07
R_{23}	6	1.57e-07

The results of the proportion tests relative to the COSMIC-Census mutational annotation are reported. For each random gene list, the number of mutated genes and the adjusted q-value are reported. Each proportion test was performed by comparing the mutated gene content in each random list versus the number of mutated genes found in the CMs (i.e. 49). In this table, only the results relative to 23 random gene lists are reported on the full set of 1000, because, for some of them, the number of mutated genes is equal; hence, here, "replicated" comparisons (i.e. referring to gene lists with the same content of mutated genes) are represented once.

Table 3.7: **Proportion test results for TCGA mutational annotation.**

Random gene list	Number of mutated genes	Q-value
$R1$	503	2.86E-18
$R2$	515	1.51E-18
$R3$	520	1.09E-18
$R4$	547	4.11E-19
$R5$	497	2.96E-19
$R6$	510	1.53E-19
$R7$	505	7.87E-20
$R8$	544	4.02E-20
$R9$	536	2.88E-20
$R10$	515	2.06E-20
$R11$	522	1.46E-20
$R12$	543	1.04E-20
$R13$	484	7.39E-21
$R14$	491	5.25E-21
$R15$	506	3.72E-21
$R16$	505	2.64E-21
$R17$	497	1.86E-21
$R18$	514	1.32E-21
$R19$	499	9.33E-22
$R20$	511	6.58E-22
$R21$	530	4.64E-22
$R22$	483	3.26E-22
$R23$	516	2.30E-22

The results of the proportion tests (23 out of 98 comparisons are shown as an example) relative to the TCGA mutational annotation are reported. Each proportion test was performed by comparing the mutated gene content in each random list versus the number of mutated genes found in the CMs (i.e. 812). As for the Table 3.6 "replicated" comparisons (i.e. proportion tests performed on random gene lists having the same content of mutated genes) are reported once.

The distribution of the number of mutated genes in the random gene lists was plotted against the number of mutated genes in CMs for both the COSMIC-Census and TCGA mutational annotations (see Figure 3.10 **A** and **B**, respectively). A Shapiro-Wilk test was performed to check the normality of the mutated gene content distribution of the random gene lists. For both mutational annotations, COSMIC-Census and TCGA, the Shapiro-Wilk test demonstrated that the number of mutated genes in the random gene lists was normally distributed with a p-value < 0.05. In contrast, the number of mutated genes in CM-genes is significantly higher with respect to the number of mutations in the random lists (square boxes in Figure 3.10). The statistically significant enrichment of cancer-related mutated genes in the full set of CM-genes further reinforced their relevance to breast cancer.

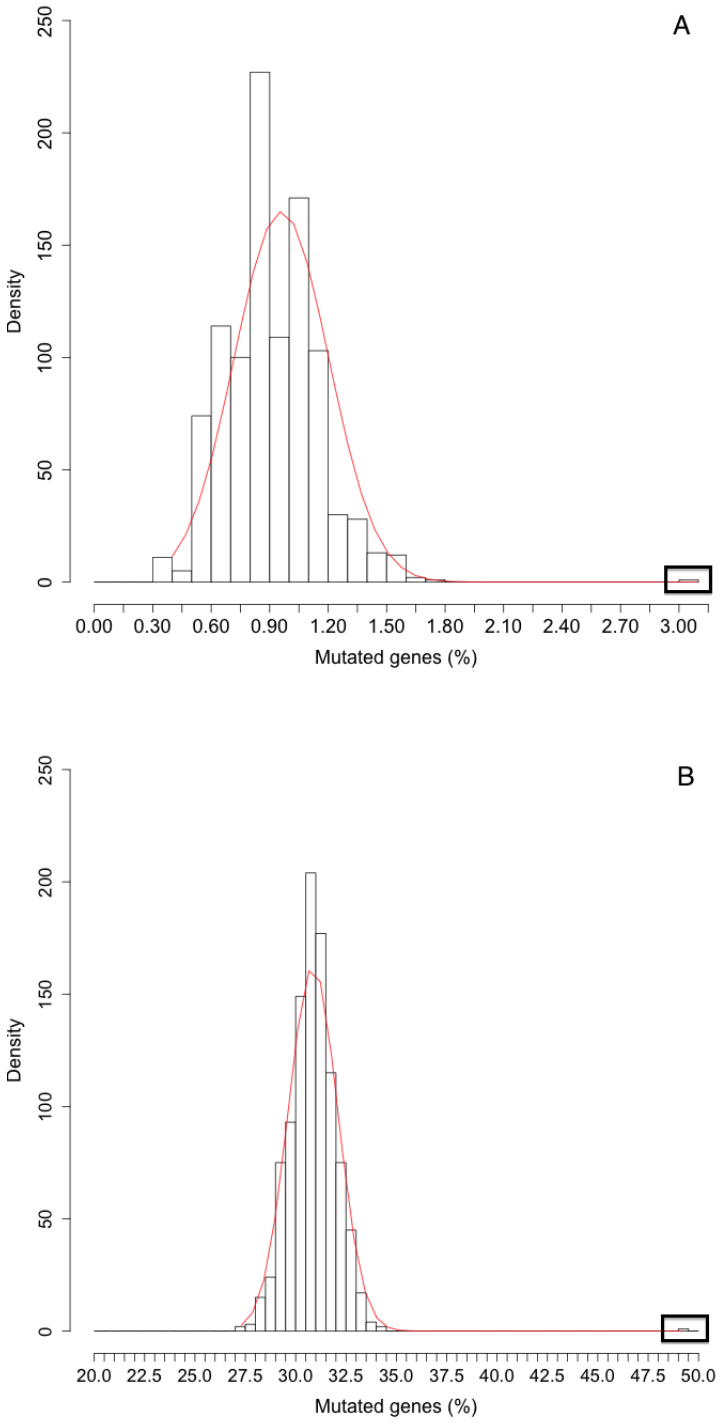

Figure 3.10: **Distribution of mutated genes after the mutational annotation in random gene lists and in CMs.**
The distribution of the number of mutated genes in percentage (x-axis) in the 1000 random gene lists and in the CMs is reported according to:
A: COSMIC-Census mutational annotation;
B: TCGA [BRCA] (CGA Network, 2012) mutational annotation.
The mutated genes content relative to the 1,000 random gene lists is normally distributed (red line). The mutated genes content relative to CMs is highlighted by square boxes outside the distribution.

3.3.2 Mutational annotation and mutual exclusivity analysis of GRNs

We extended the mutational annotation to the "top-100" gene neighbors of the 1,516 GRNs inferred by ARACNE (see Subsection 3.2.1 and the computational pipeline in Appendix A, red section). The mutational annotation and the mutual exclusivity analysis of the inferred GRNs were performed to: i) place breast cancer mutated genes in a network context in order to unveil functional relationships and regulation among cancer genes; ii) investigate the functional role of mutually exclusive and low frequently mutated genes in the breast cancer population of patients in a mechanistic context; iii) to prioritize driver genes in high-throughput cancer mutation data. Because of the huge amount of genes from the inferred networks to be annotated for the presence of mutated genes (11,721 total unique genes), we performed the mutational annotation by using the comprehensive list of 7,136 mutated genes (with respect to the COSMIC-Census mutational data on 238 mutated genes; see Subsection 3.3.1) from the TCGA [BRCA] dataset (CGA Network, 2012 `http://www.cbioportal.org/public-portal/`; see Methods, Subsection 2.3.3). For each network we firstly mapped the gene names on the TCGA database gene symbols removing networks genes not recognized by the TCGA annotation. A total of 10,936 out of 11,721 networks genes were recognized as valid gene symbols.The remaining 785 not annotated genes ($\sim$6%) according to TCGA dataset gene symbols annotation, were part of the not annotated Affymetrix HG-U133A probe sets we included in ARACNE GRNi analysis (see Methods, Subsection 2.2). The distribution of the mutated genes in the 1,516 GRNs (5,849 out of 10,936 genes;$\sim$53%) is reported in Figure 3.11. We next investigated the presence of patterns of occurrence of mutated genes in the GRNs across breast tumors, and, in particular, the presence of mutually exclusive mutated genes. As shown in Figure 3.12 (the EFNA3 network is reported as an example), the GRNs inferred from CM-genes contain genes mutated at low and high frequency in breast cancer population ($< 1\%$ and $\sim$37%, respectively). Importantly, some GRNs displayed a clear mutually exclusive mutational pattern of mutated gene neighbours as from the visual inspection of the mutational profile of breast cancer patients . This, might suggest an oncogenic role of the GRNs according to the recent observations that genes commonly involved in the same cancer pathway tend not to be mutated together in the same patient ([200],[201],[202]).

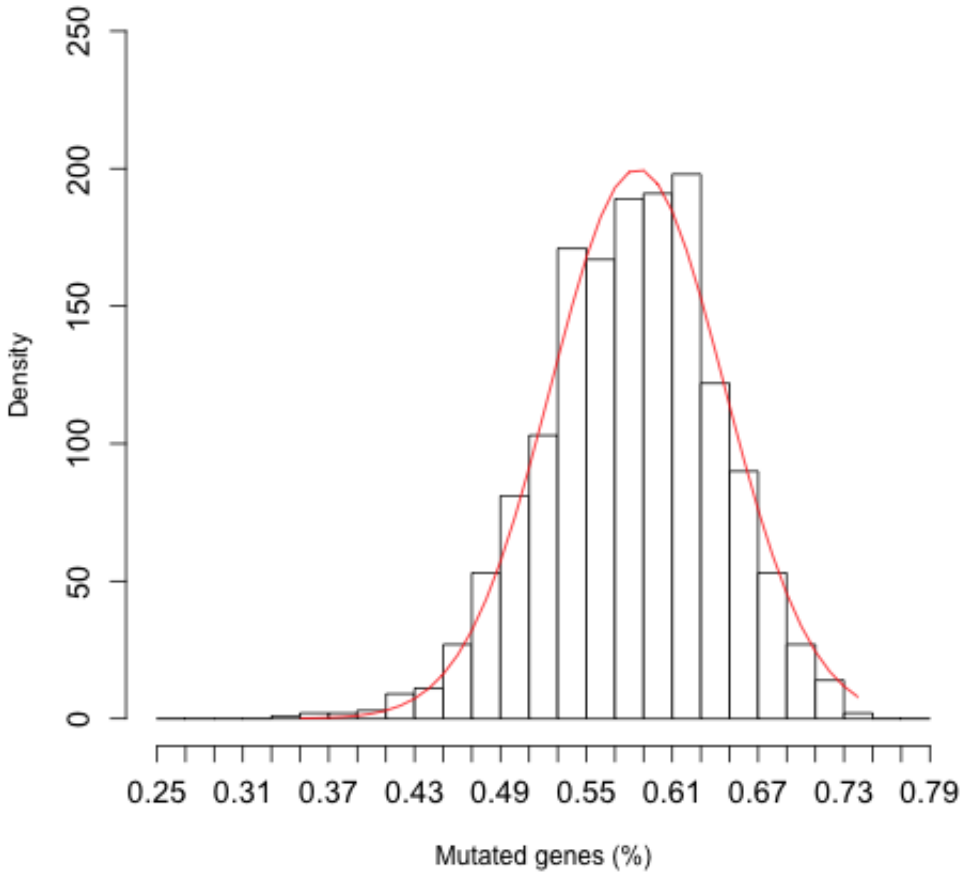

Figure 3.11: **Distribution of mutated genes in the GRNs.**
The distribution of the number of mutated genes, reported as a percentage (x-axis), in the GRNs (1,516 total networks) is reported according to TCGA [BRCA] mutational data (CGA Network, 2012).

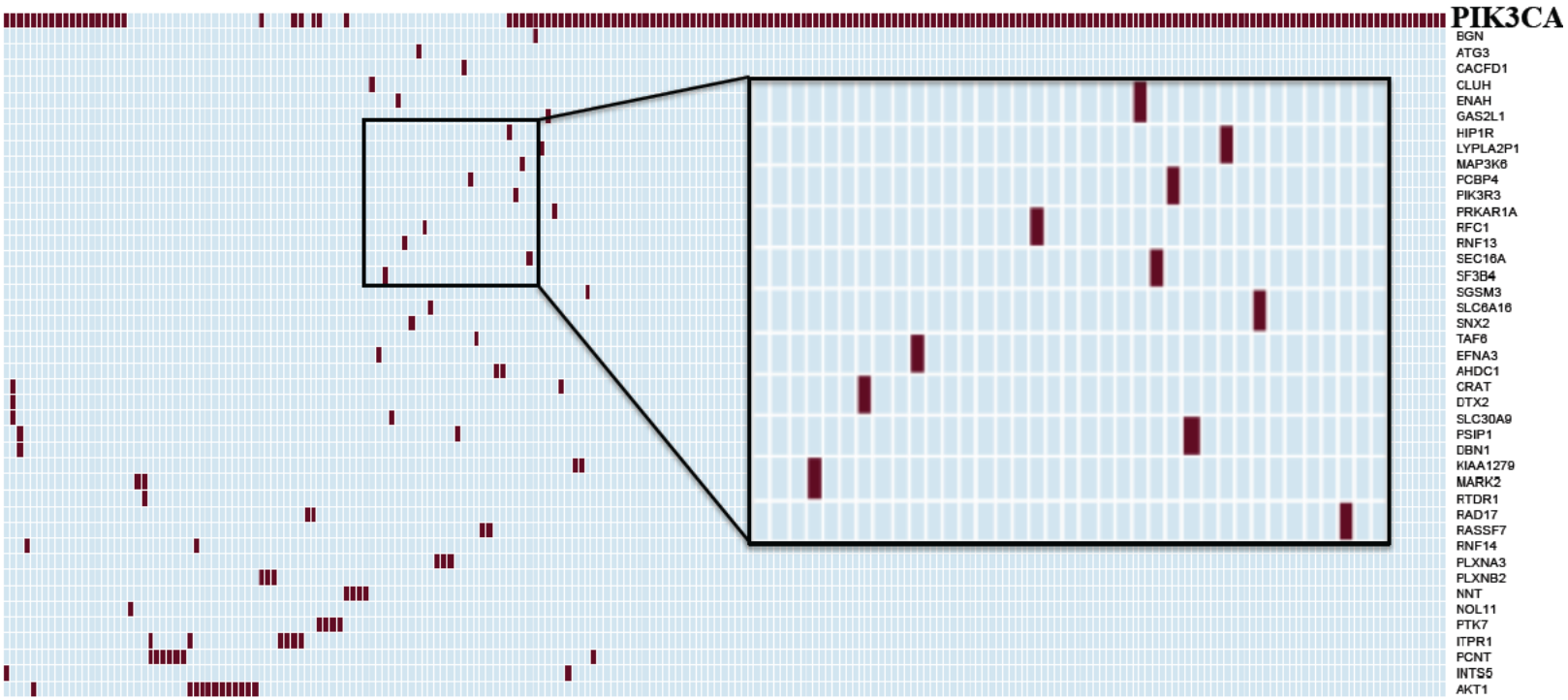

Figure 3.12: **Mutual Exclusivity pattern of networks mutated genes.**
The Mutual Exclusivity pattern of the EFNA3 mutated network neighbors is reported. The set of patients in which the gene neighbors of the network were found to be mutated is reported on the x-axis. The list of the mutated genes of the network according to TCGA mutational annotation is reported on the y-axis. The box highlight the mutually exclusive pattern of the low frequency mutated genes (dark red dots) between them and respect to the high frequency mutated gene (PIK3CA).

We then investigated whether the mutually exclusive pattern observed in some GRNs was statistically significant. To do this, we generated 1,000 random gene sets of 100 genes each (the same size as our "top-100" gene neighbors) and performed a mutational annotation using the TCGA[BRCA] dataset, as previously described for the GRNs. The distribution of mutated genes in these random lists is shown in Figure 3.13. Finally, we checked for the presence of mutually exclusive mutated genes in these random gene lists, and analyzed whether there was a significant difference in the number of mutually exclusive mutated genes between the GRNs and the random gene sets. For this analysis, we applied the statistical proportion test (see Methods, Subsection 2.3.3), running a total of 1,516,000 tests. A schematic representation of the comparisons performed is reported in Figure 3.14.

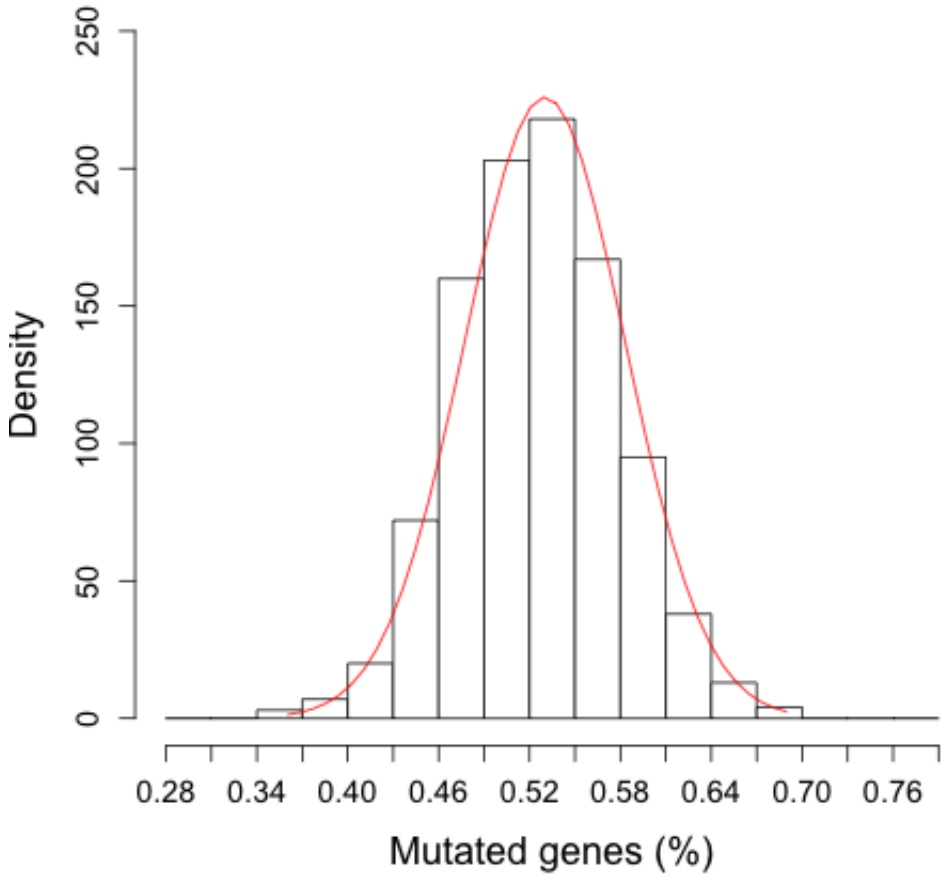

Figure 3.13: **Distribution of mutated genes in the 1,000 random gene lists.** The distribution of the number of mutated genes, reported as a percentage (x-axis), in the 1,000 random gene lists is reported according to TCGA [BRCA] mutational data (CGA Network,2012).

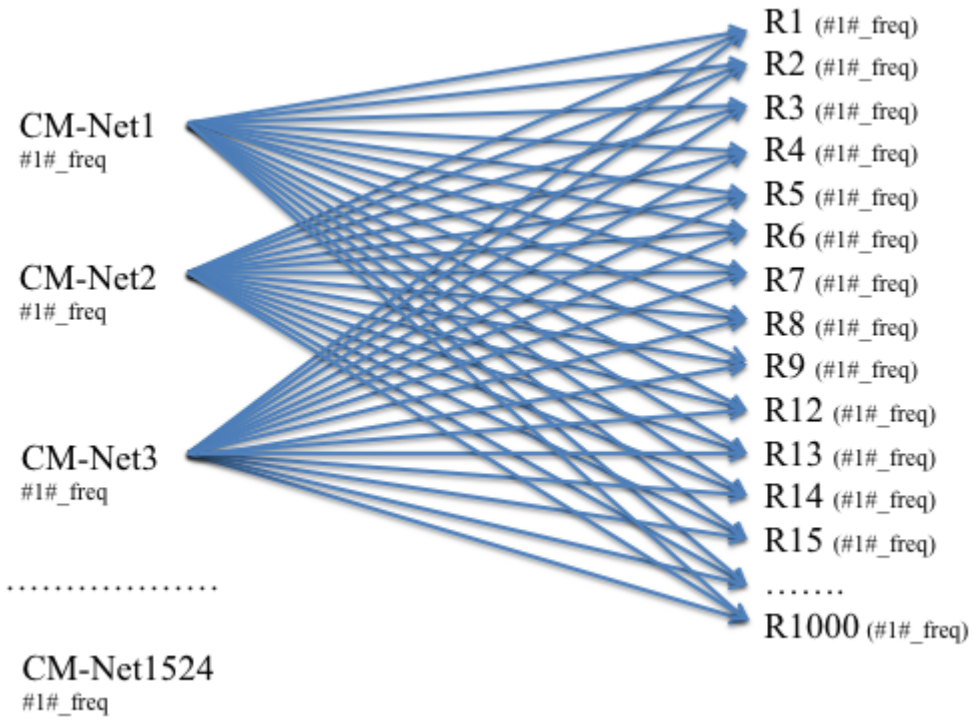

Figure 3.14: **Schematic representation of the comparisons performed by the proportion tests.**
A schematic representation of the comparisons performed by the proportion tests is reported. CM-Net indicates the GRN inferred using the ARACNE algorithm. "R" indicates the random gene list and "#1#_freq" indicates the presence of one mutated gene of the network per breast cancer patient.

The results obtained can be summarized as follows:

- for each network (1,516 total GRNs), we found a fraction of statistically significant proportion tests (i.e., the GRN was significantly enriched in mutually exclusive mutated genes compared with the random lists) and a fraction of not statistically significant proportion tests (i.e. the GRN was not significantly enriched in mutually exclusive mutated genes). The number of statistically significant proportion tests range from 992 (i.e., the GRN is enriched in mutually exclusive mutated genes with respect to 992 random gene lists while it is not with respect to 8 random lists) to 10 (i.e., the GRN is enriched in mutually exclusive mutated genes with respect to only 10 random gene lists while it is not with respect to 990 random lists).

- we observed a "negative correlation" between the number of not statistically significant proportion tests for some GRNS, and the total number of patients in which at least one gene of the GRN was found to be mutated in a mutually exclusive way (Figure 3.15). Specifically, genes that

are highly frequently mutated in the population and that are also mutated in a mutually exclusive way with respect to the other genes in the network can influence the proportion tests results. When these genes are present in GRNs they will drive the significance towards them; on the contrary they will favour the significance of random gene lists when they are represented in random gene lists instead of in GRNs. Since we observed an increase in the number of not statistically significant proportion tests as the mutually exclusive mutated genes were infrequently mutated in the population for some GRNs, we argued that this might be one possible reason of the lack of significance of our GRNs. Different normalization procedures exists in order to control the effect of the frequency of mutation. In our case we did not normalized according to the frequency of mutation since we were interested to investigate the role of low frequency mutually exclusive mutated genes with respect to, also, high frequency mutated genes (as for the case of EFNA3 network, see Figure 3.12) as putative cancer driver genes playing a role in a network context.

- the statistically significant difference in the number of mutually exclusive mutated genes between the GRNs and the 1,000 random networks (i.e. the FDR q-value was < 0.01 after multiple correction on the set of 1,000 random gene lists), might be the result of the left-tail effect of the two-tailed proportion test. Indeed, the statistical significance of the comparisons (q-value < 0.01), might result from a higher content of mutually exclusive mutated genes in the random gene lists rather than in the GRN. According to this left-tail effect of the proportion test, there is still a significant difference in the number of mutually exclusive mutated genes between the GRN and the random lists, but the enrichment is relative to the random gene lists instead of the GRNs.

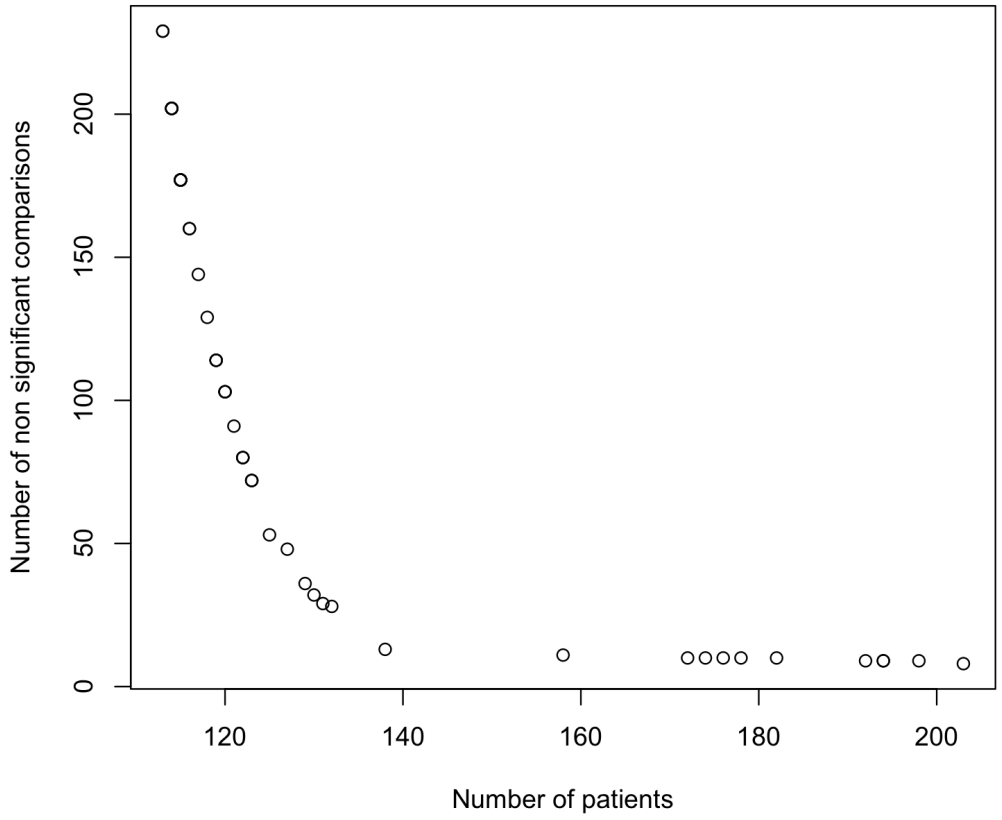

Figure 3.15: **The number of non significant comparisons as a function of the sample size.**
The number of non significant comparisons as a function of the sample size is reported. As the sample size (i.e. the number of breast cancer patients with at least one gene of the network mutated) increase, the number of non significant comparisons decrease.

To select the GRNs significantly enriched in mutually exclusive mutated genes from the proportion tests considering both, the number of statistically significant comparisons and the left-tail effect of the two-tailed proportion test on the significance, we used the following Enrichment Score ES:

$$ES = \frac{NSC + NF}{1000}$$

Where:
-"NSC" represents the fraction of "Non-statistically Significant Comparisons", i.e. the number of proportion tests run on GRNs *versus* random networks, which did not pass the significance threshold; FDR = 0.01;
-"NF" represents the "Negative Frequencies", i.e., the number of statistically

significant comparisons resulting from a higher content of mutually exclusive mutated genes in random gene lists with respect to the GRNs.

The score ES, was then normalized on the total number of random gene lists (i.e., 1,000). According to this scoring system, the lower the score (ES), the higher is the significance of the enrichment of mutually exclusive mutated genes in the GRNs. The networks were then sorted according to ES and only the networks with a score < 0.25 were finally considered as significant. Based on these statistical tests and selection criteria, we identified a core of 50 GRNs that were significantly enriched in mutually exclusive mutated gene neighbors (see Table 3.8). Interestingly, this set of networks was unique with respect to the 48 networks prioritized according to the COSMIC-Census mutational annotation (see Subsection 3.3.1), with the exception of one network MSH2 that was in common.

Table 3.8: **List of 50 GRNs significantly enriched in mutually exclusive mutated genes.**

GRN	MGs	SCs	NSCs	NFs	PFs	Score
RASA2	50	992	8	0	992	0.008
EFNA3	55	991	9	0	991	0.009
FADS2	49	991	9	0	991	0.009
TXNIP	45	991	9	0	991	0.009
NIPAL3	45	991	9	0	991	0.009
MYBBP1A	46	990	10	0	990	0.01
G3BP1	57	990	10	0	990	0.01
SMARCC1	62	990	10	0	990	0.01
AATF	53	990	10	0	990	0.01
RRP1	46	990	10	0	990	0.01
DDB2	57	989	11	0	989	0.011
OSBPL8	59	987	13	8	979	0.021
FOXC1	61	972	28	8	964	0.036
CDH3	62	971	29	8	963	0.037
SR140	54	968	32	9	959	0.041
ARL4C	58	964	36	9	955	0.045
ITPR2	57	952	48	9	943	0.057
KDM4B	52	947	53	9	938	0.062
IVD	55	928	72	9	919	0.081
DHCR7	52	928	72	9	919	0.081
CAMP	55	928	72	9	919	0.081
ADM	52	920	80	9	911	0.089
PMAIP1	56	920	80	9	911	0.089
SKP2	57	920	80	9	911	0.089
CEBPB	58	909	91	9	900	0.1
GSTP1	58	897	103	9	888	0.112
NPY1R	54	897	103	9	888	0.112
ENO1	58	897	103	9	888	0.112
NDN	61	886	114	10	876	0.124
NQO1	55	886	114	10	876	0.124
SLC7A5	57	886	114	10	876	0.124
HMGCS2	51	871	129	10	861	0.139
PNP	58	856	144	10	846	0.154
PHGDH	56	840	160	10	830	0.17
IL8	59	840	160	10	830	0.17
ANGPTL2	57	823	177	10	813	0.187
PLAU	62	823	177	10	813	0.187
CXCL12	58	823	177	10	813	0.187
LUM	55	823	177	10	813	0.187
ACTN1	65	823	177	10	813	0.187
XPO4	62	823	177	10	813	0.187
KRT15	56	823	177	10	813	0.187
SERPINH1	53	798	202	10	788	0.212
PMP22	58	798	202	10	788	0.212
ASAH1	65	798	202	10	788	0.212
PENK	65	798	202	10	788	0.212
MSH2	51	798	202	10	788	0.212
POLB	58	798	202	10	788	0.212
KCNN4	61	771	229	10	761	0.239
THOP1	51	771	229	10	761	0.239

The list of 50 GRNs significantly enriched in mutually exclusive mutated genes is reported followed by: the number of Mutated Genes (MGs) according to TCGA mutational annotation; the number of Significant Comparisons (SCs), i.e. the number of statistically significant proportion tests; the number of Non-Significant Comparisons (NSCs), i.e. the statistically non-significant proportion tests; the Negative Frequencies (NFs), representing the number of random gene lists enriched in mutually exclusive mutated genes; the Positive Frequencies (PFs), representing the number of random lists not enriched in mutually exclusive mutated genes compared to the GRNs; the Score calculated as described in the main text.

3.4 Identification of higher order regulatory mechanisms

We performed an in-depth GRN deconvolution analysis to identify putative transcriptional Master Regulator (MR) hub genes for the 48 and 50 networks prioritized by the COSMIC-Census mutational annotation and the mutual exclusivity analysis, respectively (see Subsections 3.3.1 and 3.3.2 and the computational pipeline in Appendix A, blue section). A hub gene is defined as a highly interconnected gene within a complex network, that has a key biological role in the cell ([203],[204]). In the case of gene expression, the connections with other genes represent the degree of correlation between expression measurements. Typically, for GRNs, a hub gene is a transcription factor (TF) or co-factor that is the MR of expression of neighboring genes, making it highly interconnected. Specifically, the TF or co-factor is at the top of a of a regulatory hierarchy. Besides TFs, MRs can also be genes whose expression is critical to the transcriptional activation of multiple downstream genes (including TFs) involved in cell signaling cascades ([205]). Since the expression of CM-genes correlates with pathological conditions associated with breast cancer, we attempted to identify cancer-related mechanisms at the transcriptional level. For the full list of CM-genes (1516 total genes), we performed a network inference analysis in which we assumed that each CM-gene was a hub gene, i.e. a transcriptionally highly interconnected gene (Figure 3.16, panel 'a'). From this analysis, we inferred 1,516 GRNs (CM-gene GRNs), however, we were unable to determine whether a gene was likely to be a real hub or not. In addition, we observed that for some of these CM-gene GRNs, the hub genes shared common neighbors supporting the possibility of a higher-order regulatory program under the control of a common putative transcriptional MR (Figure 3.17). To overcome the bias of assuming that all CM-genes were hub genes in the network inference analysis, and to identify potential MRs of transcriptional programs, we performed a transcriptional network deconvolution analysis using the ARACNE algorithm (Figure 3.16, panel 'b'). In this analysis, we considered that for each one of the 48 and 50 GRNs (CM-gene GRNs) each gene neighbor within the network is a hub gene around which to build a new "transcriptional" network. The MR was then defined as the gene neighbor that occurs most frequently among the new networks (Figure 3.16, panel 'b'). Once the MR gene was identified, we finally built a GRN around the MR hub gene. The networks inferred from the MRs were called MR-gene GRNs (or MR-networks). Using this strategy, we were able to identify in an unbiased way, the MR of transcriptional mechanisms identified through the network inference analysis performed

on CM-genes. For some networks inferred from CM-genes as hub genes, the transcriptional MR appeared to be the CM-gene itself, confirming the relevance of CM-genes in these putative oncogenic networks. In contrast, for other networks the MR was a neighboring gene of the original network hub. Interestingly, some GRNs inferred from CM-genes shared a common MR (Figure 3.18), allowing these networks to be grouped together, representing branches of a higher-order regulatory mechanism. The full list of MRs for the set of 48 and 50 GRNs is reported in Table 3.9 and 3.10. We identified 23 and 31 unique MRs for the set of 48 and 50 GRNs respectively, with the exception of the RUNX1T1 gene that was in common. 14 out of 54 total MR genes are described as transcription factors in TRANSFAC database (`http://www.gene-regulation.com/pub/databases.html/`). They are: ATF6B, CREBL2, E2F4, ESR1, FOXM1, FOXO1, HOXA5, MAX, MLXIPL, MYBL2, NFAT5, RUNX1T1, TCF4 and TCF7L1.

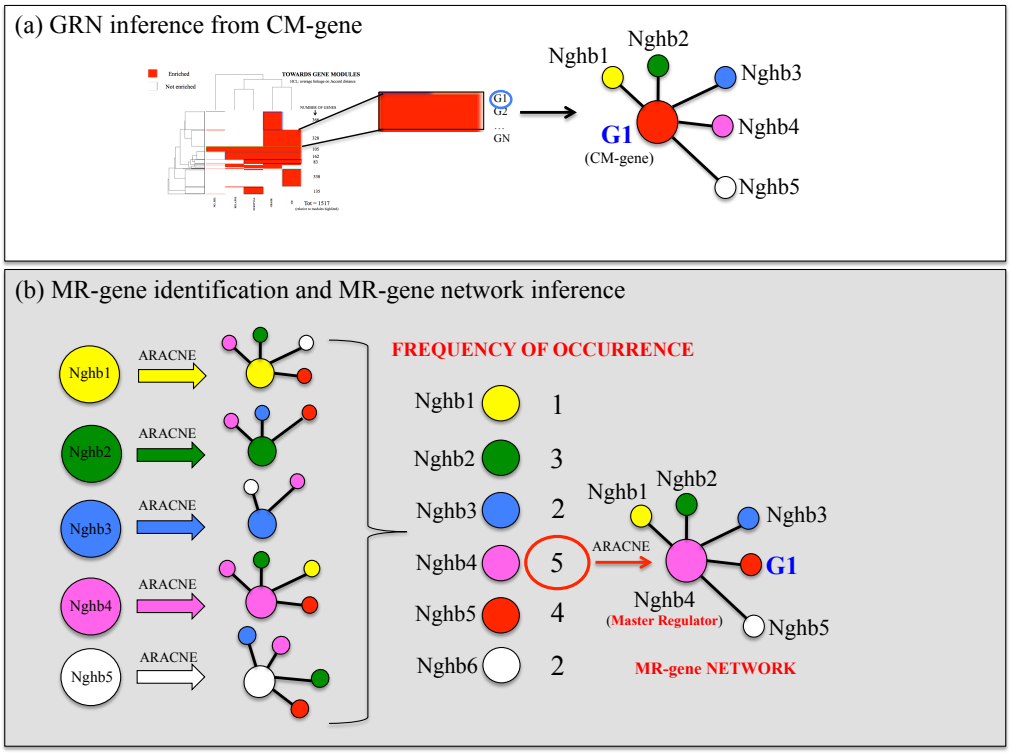

Figure 3.16: **Identification of higher order regulatory mechanisms.** Schematic representations of the network inference analysis performed on CM-genes **(a)** and the network analysis performed for the identification of higher-order regulatory mechanisms **(b)**. In panel **(a)**, the network inference analysis performed to build GRNs from CM-genes (CM-gene GRNs) is shown. Each CM-gene was assumed to be a hub gene to identify transcriptionally correlating genes, i.e. gene neighbors (Nghb). In panel **(b)**, the network inference analysis to investigate for the presence of higher-order regulatory mechanisms is shown. Each neighboring gene for each GRN inferred from CM-genes (shown in 'a') was assumed to be a hub gene; the network was then built around the hub gene based on gene expression data using the ARACNE algorithm. The master regulator (MR) was then identified as the gene with the highest frequency of occurrence, as a neighbor gene, across all networks inferred from the original GRN (shown in 'a'). Once identified, a new network composed of genes from the original GRN, was built around the MR gene.

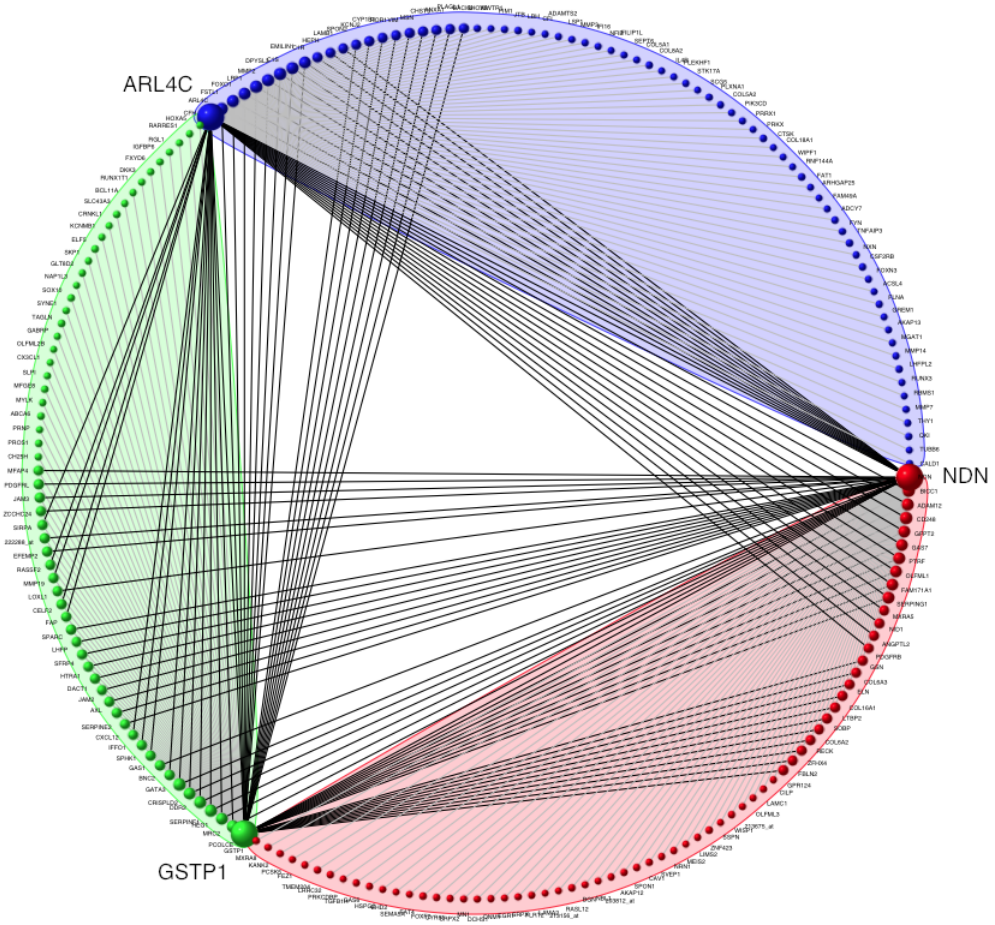

Figure 3.17: **Overlapping neighbors of the ARL4C, NDN and GSTP1 CM-gene GRNs.**
The GRNs inferred from ARL4C, NDN and GSTP1 CM-genes are reported. The gene neighbors of the networks and the relative hub genes are highlighted in blue, red and green for ARL4C, NDN and GSTP1, respectively. For each network, the edges connecting the hub gene with its neighbors are in gray, while the edges connecting the hub genes with neighbors of other networks (i.e. shared neighbors) are highlighted in black.

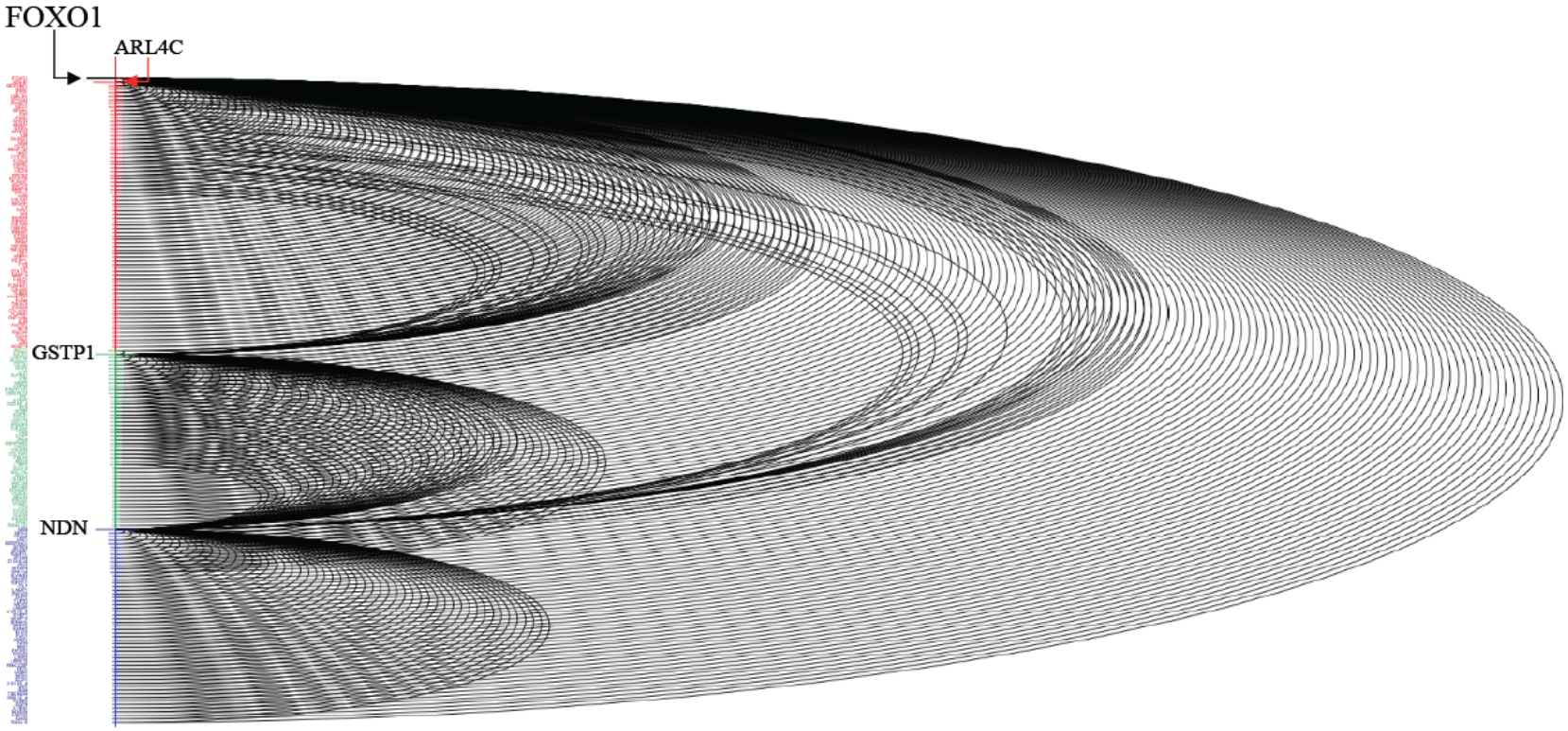

Figure 3.18: **FOXO1 gene as Master Regulator (MR) of ARL4C, NDN and GSTP1 CM-gene GRNs.**
FOXO1 is a master regulator of the ARL4C, NDN and GSTP1 CM-gene GRNs. FOXO1 expression correlates at a transcriptional level with the expression of all the genes composing the networks inferred from the CM-genes ARL4C, NDN and GSTP1. FOXO1 therefore represents the transcriptional master regulator of a higher-order regulatory mechanism. On the left of the figure, the genes of each network are listed. In red are reported the gene neighbours relative to the ARL4C network, in green those relative to the GSTP1 network, and in blue the gene neighbours relative to the NDN network. Black curved lines represent pairwise transcriptional correlations inferred by ARACNE algorithm between the CM-gene as hub gene and the relative neighbours listed in the gene list. At the top of the list the gene FOXO1 is connected with all the genes in the list as Master Regulator (MR) of ARL4C, NDN and GSTP1 CM-gene GRNs.

Table 3.9: **Master regulators relative to the set of 48 CM-gene GRNs identified through the COSMIC-Census mutational annotation of CM-genes.**

Master Regulator (MR)	Full gene name	CM-gene hub(s)
ACRV1	acrosomal vesicle protein 1	CASC5, HOXD13
AGPAT1	1-acylglycerol-3-phosphate O-acyltransferase 1	EWSR1
CAV1	caveolin 1, caveolae protein, 22kDa	GPC3
CCDC9	coiled-coil domain containing 9	FGFR2
CD2	CD2 molecule	SOCS1
CD48	CD48 molecule	JAK2
CDK1	cyclin-dependent kinase 1	BRCA1
DAZAP1	DAZ associated protein 1	TOP1
DMWD	dystrophia myotonica, WD repeat containing	BAP1, CDK12, CDKN2A, MYD88, ETVC
DPYSL3	dihydropyrimidinase-like 3	EXT2
EGR1	early growth response 1	JUN
FAM171A1	family with sequence similarity 171, member A1	RET
HSPA8	heat shock 70kDa protein 8	MSH6
KCNMB1	potassium large conductance calcium-activated channel	MET
MMP2	matrix metallopeptidase 2	COL1A1
MYO15B	myosin XVB pseudogene	CDK4
PABPN1	poly(A) binding protein, nuclear 1	PMS1
PSG6	pregnancy specific beta-1-glycoprotein 6	EGFR
RUNX1T1	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)	CCND2
TPX2	TPX2, microtubule-associated, homolog (Xenopus laevis)	BRCA2, BRIP1, BUB1B,
TRBC1	T cell receptor beta constant 1	POU2AF1
TROAP	trophinin associated protein (tastin)	BLM
UBE2C	ubiquitin-conjugating enzyme E2C	RECQL4,WHSC1
ZNF160	zinc finger protein 160	APC

Table lists the master regulators relative to the set of 48 CM-gene GRNs identified through the COSMIC-Census mutational annotation of CM-genes, followed by their full gene name and relative CM-gene hub(s).

Table 3.10: **Master regulators relative to the set of 50 CM-gene GRNs identified through the mutual exclusivity analysis.**

Master Regulator (MR)	Full gene name	CM-gene hub(s)
ACTL6A	Actin-Like 6A	TXNIP
ASH1L	ash1 (absent, small, or homeotic)-like (Drosophila)	XPO4
ATF6B	activating transcription factor 6 beta	NIPAL3
BCL11A	B-Cell CLL/Lymphoma 11A (Zinc Finger Protein)	PHGDH
CHD3	cadherin 3, type 1, P-cadherin (placental)	MYBBP1A
CREBL2	cAMP responsive element binding protein-like 2	ASAH1
E2F4	E2F transcription factor 4, p107/p130-binding	THOP1
ELK3	ELK3, ETS-domain protein (SRF accessory protein 2)	ITPR2
ESR1	estrogen receptor 1	IVD
FOXM1	forkhead box M1	DHCR7, HMGCS2, NPY1R
FOXO1	forkhead box O1	ARL4C, GSTP1, NDN
GATA3	GATA binding protein 3	IL8, KDM4B, PMAIP1,POLB
HOXA5	homeobox A5	KRT15
IFI16	interferon, gamma-inducible protein 16	ADM
LRPPRC	leucine-rich PPR-motif containing	MSH2
MAX	MYC associated factor X	G3BP1
MLXIP	MLX interacting protein	KCNN4
MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	PNP, SLC7A5
NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive	RASA2
PRRX1	paired related homeobox 1	PLAU, SERPINH1
PTTG1	pituitary tumor-transforming 1	CAMP
RBMS1	RNA binding motif, single stranded interacting protein 1	OSBPL8
RUNX1T1	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)	PENK
SOX10	SRY (sex determining region Y)-box 10	CDH3, FOXC1
TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor	EFNA3
TCF4	transcription factor 4	CXCL12
TCF7L1	transcription factor 7-like 1 (T-cell specific, HMG-box)	NQO1
TGFB1I1	transforming growth factor beta 1 induced transcript 1	ACTN1, ANGPTL2
TP53BP1	tumor protein p53 binding protein 1	RRP1
ZFPM2	zinc finger protein, multitype 2	LUM, PMP22
ZNF45	Zinc Finger Protein 45 (A Kruppel-Associated Box (KRAB) Domain	FADS2
ZNF302	zinc finger protein 302	SR140

Table lists the master regulators relative to the set of 50 CM-gene GRNs identified through the mutual exclusivity analysis, followed by their full gene name and relative CM-gene hub(s).

3.5 Clinical relevance of GRNs

3.5.1 Transcriptional activity of GRNs: the concordance analysis

We then assessed the clinical and pathological relevance of the 48 and 50 sets of GRNs (CM-gene GRNs) prioritized by the COSMIC-Census mutational annotation and the mutual exclusivity analysis respectively and of the relative MR-gene GRNs (inferred as described in Section 3.4) to breast cancer (see the computational pipeline in Appendix A, pink section). We first established a scoring system to predict the activation/inhibition of the inferred networks in breast cancer (see Methods, Section 2.4). The transcriptional activity of the networks was evaluated using an independent microarray gene expression dataset of 997 breast tumors (the Metabric study, Discovery set ([36])). The log2 median centered gene expression matrix was sorted across breast cancer patients according to the transcriptional profiles of the CM-gene 'hub' or the MR gene, around which the networks were built. The sorted dataset was then transformed into a binary matrix assigning +1 to positive gene expression measurements (genes up-regulated) of the gene neighbors of the network and of the hub gene, and -1 to the negative gene expression measurements (genes down-regulated). From the binarized data, a Concordance Score (CS) was computed as follows:

$$CS_{up} = \forall_{PHup}(Ng_{up} - Ng_{dn})$$

Where:
-CS_{up} is the network concordance score in the case of up-regulated hubs;
- $(Ng_{up} - Ng_{dn})$ is the difference between the number of up-regulated genes and the number of down-regulated genes of the network, computed for each breast cancer sample ($\forall_{PHup}$) when the hub gene is up-regulated.

or

$$CS_{dn} = \forall_{PHdn}(Ng_{dn} - Ng_{up})$$

Where:
-CS_{dn} is the network concordance score in the case of down-regulated hubs;
-$(Ng_{dn} - Ng_{up})$ is the difference between the number of down- regulated genes and the number of up-regulated genes of the network, computed for each breast cancer sample ($\forall_{PHdn}$) when the hub gene is down-regulated.

CS_{dn} is the network concordance score in the case of down-regulated hub, $(Ng_{dn} - Ng_{up})$ is the difference between the number of down-regulated genes and the number of up-regulated genes of the network computed for each breast cancer sample ($\forall_{PHdn}$) when the hub gene is down-regulated. According to the computed difference Δ, the extent of the concordance was defined as follows:

$$CS_{up} = \begin{cases} \text{Positive Concordance if } \Delta_{up} > 0 \\ \text{Negative Concordance if } \Delta_{up} < 0 \\ \text{Non Concordance if } \Delta_{up} = 0 \end{cases}$$

$$CS_{dn} = \begin{cases} \text{Positive Concordance if } \Delta_{dn} > 0 \\ \text{Negative Concordance if } \Delta_{dn} < 0 \\ \text{Non Concordance if } \Delta_{dn} = 0 \end{cases}$$

We then defined three distinct patterns of transcriptional activation/inhibition of the networks in breast cancer according to the scoring system:

- Absence of transcriptional activation/inhibition of the network (Non Concordance): 50% of the neighbors of the network have the same transcriptional profile with respect to the hub gene, while the remaining 50% of neighbors have the opposite transcriptional regulation (Figures 3.19 and 3.20 box a).
- Positive transcriptional activation/inhibition (Positive Concordance): $>$ 50% of gene neighbors in the network have the same transcriptional regulation as the hub gene (up-regulated/activated or downregulated/inhibited) (Figures 3.19 and 3.20 box b).
- Negative transcriptional activation/inhibition (Negative Concordance): $>$ 50% of gene neighbors in the network have the opposite transcriptional regulation with respect to the hub gene (Figures 3.19 and 3.20 box c).

The gene expression values of the network neighbors plotted with respect to the computed CSs for the three transcriptional patterns is shown in Figure 3.21. For the absence of a transcriptional activation/inhibition pattern, the gene expression profiles of the CM-gene AATF network neighbors (reported as a representative example) is plotted against the computed CSs. As shown, when half of the neighbors are up-regulated and half are down-regulated, the computed CS is close to

0. In contrast, in the case of positive transcriptional activation/inhibition (as displayed by the CM-gene PLAU network), when the majority of the neighbors have the same direction of gene expression regulation (either up-regulated or down-regulated) as the hub gene, the CS is greater than 0. Finally, when the majority of the neighbors of the network have an opposite gene expression regulation compared to the hub gene, (as disaplyed by the CM-gene CEBP network), the CS is less than 0. The full list of the networks, grouped by transcriptional activation/inhibition patterns is reported in Table 3.11 for the set of 48 networks and in Table 3.12 for the set of 50 networks relative to networks inferred from CM-genes and MR-genes. For subsequent analyses, we focused on the set of networks with positive and negative transcriptional activation patterns, because they could indicate activation/inhibition of molecular mechanisms represented by the networks in breast cancer.

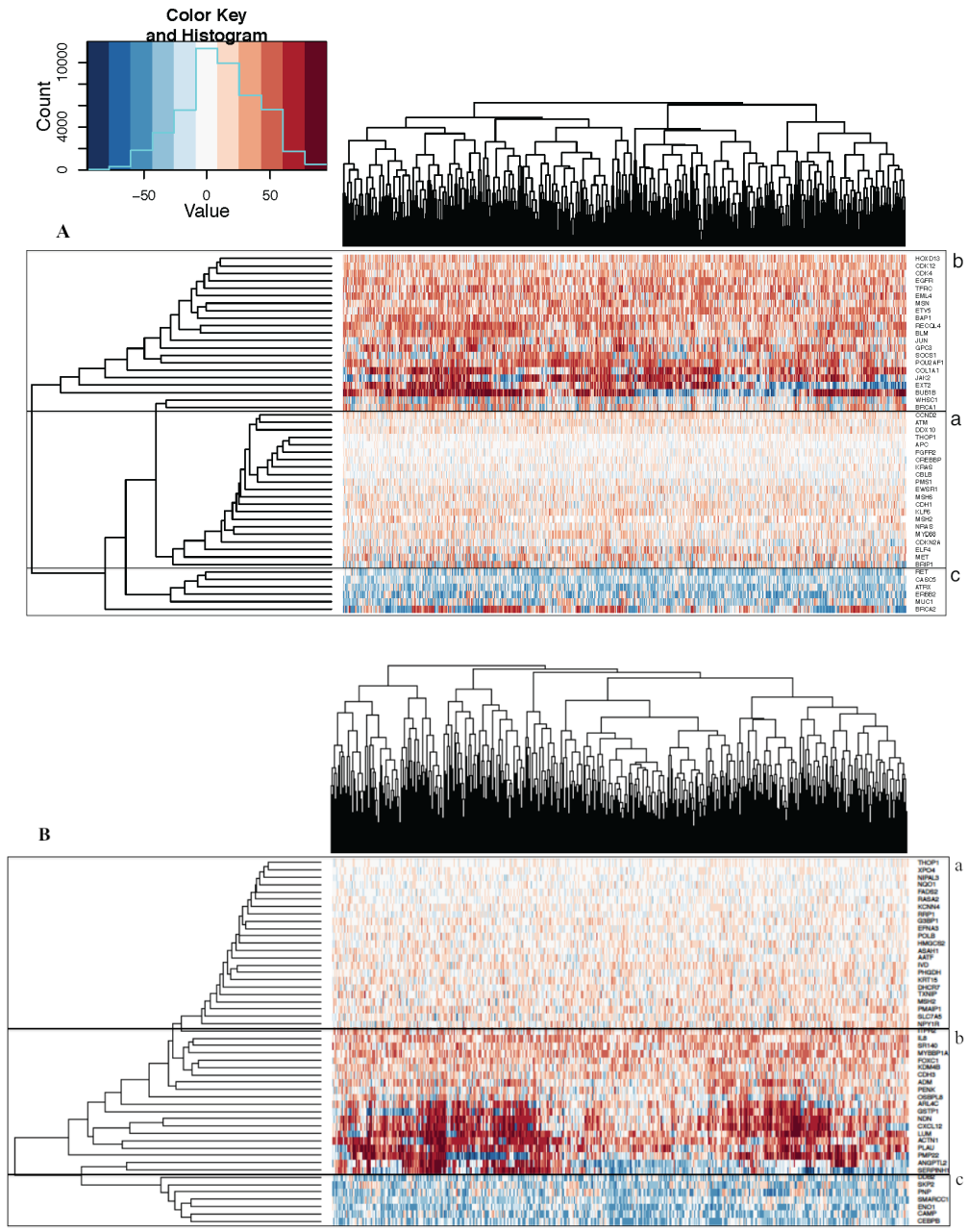

Figure 3.19: **The transcriptional activation of CM-gene GRNs on Metabric cohort of breast cancer patients.**
Hierarchical cluster analysis of the Concordance Scores (CSs) computed on the CM-gene GRNs inferred from the set of 48 networks resulting from the COSMIC-Census mutational annotation and from the set of 50 networks resulting from mutual exclusivity analysis. Columns represent breast cancer patients from the Metabric cohort, while rows represent the networks scores. The heatmaps refer to the set of 48 CM-gene GRNs **(A)** and to the set of 50 CM-gene GRNs **(B)**. The boxes highlight the three transcriptional activation patterns of the networks resulting from the concordance analysis: box a, the absence of a transcriptional activation/inhibition; box b, the presence of a positive transcriptional activation/inhibition pattern; box c, the presence of a negative transcriptional activation/inhibition pattern.

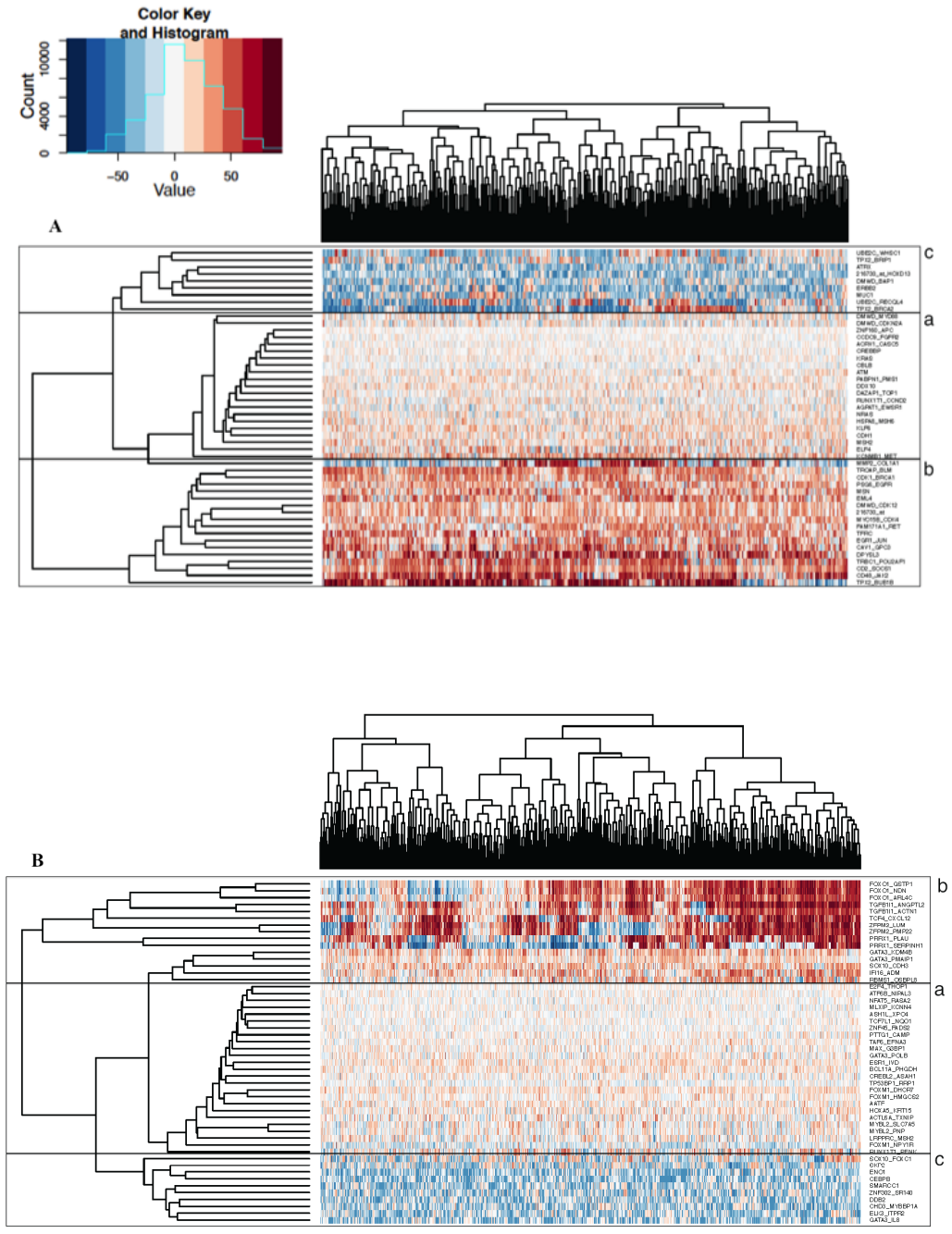

Figure 3.20: **The transcriptional activation of MR-gene GRNs on Metabric cohort of breast cancer patients.**
Hierarchical cluster analysis of the Concordance Scores (CSs) computed on the MR-gene GRNs inferred from the set of 48 networks resulting from the COSMIC-Census mutational annotation and from the set of 50 networks resulting from mutual exclusivity analysis. Columns represent breast cancer patients from the Metabric cohort, while rows represent the networks scores. The heatmaps refer to the set of 48 MR-gene GRNs **(A)** and to the set of 50 MR-gene GRNs **(B)**. The boxes highlight the three transcriptional activation patterns of the networks resulting from the concordance analysis: box a, the absence of a transcriptional activation/inhibition; box b, the presence of a positive transcriptional activation/inhibition pattern; box c, the presence of a negative transcriptional activation/inhibition pattern.

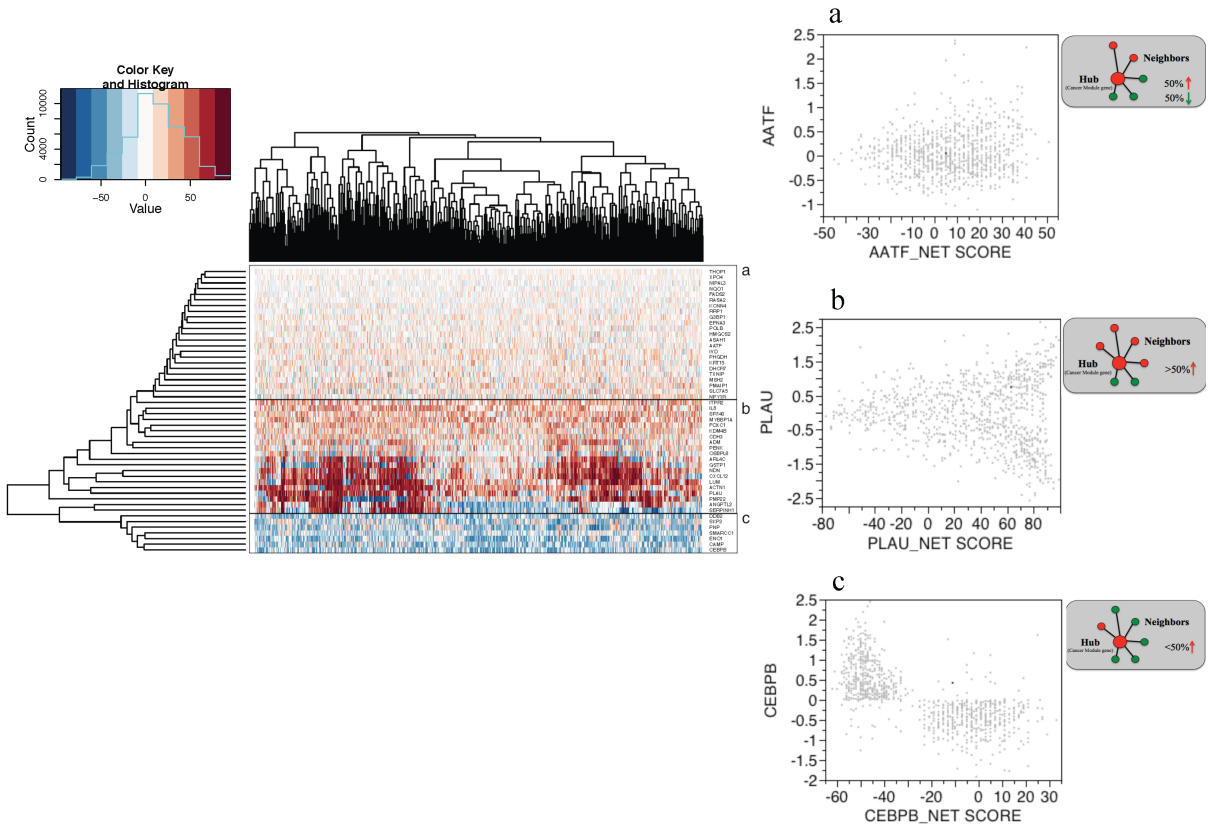

Figure 3.21: **Transcriptional patterns of CM-gene GRNs: gene expression vs. concordance.**
The CM-gene GRN expression profiles plotted against the GRN Concordance Scores (NET SCORE) for three networks representative of the three transcriptional activation patterns: **a)** the absence of a transcriptional activation pattern (CM-gene AATF network); **b)** the positive transcriptional activation pattern (CM-gene PLAU network); **c)** the negative transcriptional activation pattern (CM-gene CEBP network).

Table 3.11: **The concordance analysis results relative to the 48 CM-gene and MR-gene networks.**

Non Concordance	Positive Concordance	Negative Concordance
CM-gene GRN		
CCND2, ATM, DDX10, THOP1, APC, FGFR2, CREBBP, KRAS, CBLB, PMS1, EWSR1, MSH6, CDH1, KLF6, MSH2, NRAS, MYD88, CDKN2A, ELF4, MET, BRIP1	HOXD13, CDK12, CDK4, EGFR, TFRC, EML4, MSN, ETV5, BAP1, RECQL4, BLM, JUN, GPC3, SOCS1, POU2AF1, COL1A1, JAK2, EXT2, BUB1B, WHSC1, BRCA1	RET, CASC5, ATRX, ERBB2, MUC1, BRCA2
MR-gene GRN		
ZNF160_APC, CCDC9_FGFR2, ACRV1_CASC5, PABPN1_PMS1, DAZAP1_TOP1, RUNX1T1_CCND2, AGPAT1_EWSR1, HSPA8_MSH6, KCNMB1_MET	TPX2_BUB1B, TROAP_BLM, CDK1_BRCA1, PSG6_EGFR, DMWD_CDK12, MYO15B_CDK4, MYO15B_CDK4, EGR1_JUN, CAV1_GPC3, TRBC1_POU2AF1, CD2_SOCS1, CD48_JAK2, MMP2_COL1A1, FAM171A1_RET, DPYSL3_EXT2	UBE2C_WHSC1, TPX2_BRIP1, ACRV1_HOXD13, DMWD_BAP1, UBE2C_RECQL4, TPX2_BRCA2, DMWD_MYD88, DMWD_CDKN2A

The concordance analysis results relative to the set of 48 CM-gene and MR-gene networks are reported. The networks are grouped according to the transcriptional activation pattern : Non Concordance, Positive Concordance, or Negative Concordance patterns. The CM-gene networks are represented by their hub gene (i.e. the CM-gene imposed as the hub gene of the network). The MR-gene networks are represented by two gene symbols separated by underscore. The first gene name refers to the MR-gene. The second gene symbol refers to the original CM-gene.

Table 3.12: **The concordance analysis results relative to the 50 CM-gene and MR-gene networks.**

Non Concordance	Positive Concordance	Negative Concordance
CM-gene GRN		
THOP1, XPO4, NIPAL3, NQO1, FADS2, RASA2, KCNN4, RRP1, G3BP1, EFNA3, POLB, HMGCS2, ASAH1, AATF, IVD, PHGDH, KRT15, DHCR7, TXNIP, MSH2, PMAIP1, SLC7A5, NPY1R	ITPR2, IL8, SR140, MYBBP1A, FOXC1, KDM4B, CDH3, ADM, PENK, OSBPL8, ARL4C, GSTP1, NDN, CXCL12, LUM, ACTN1, PLAU, PMP22, ANGPTL2, SERPINH1	DDB2, SKP2, PNP, SMARCC1, ENO1, CAMP CEBPB
MR-gene GRN		
E2F4_THOP1, ATF6B_NIPAL3, NFAT5_RASA2, MLXIP_KCNN4, ASH1L_XPO4, TCF7L1_NQO1, ZNF45_FADS2, PTTG1_CAMP, TAF6_EFNA3, MAX_G3BP1, GATA3_POLB, ESR1_IVD, BCL11A_PHGDH, CREBL2_ASAH1, TP53BP1_RRP1, FOXM1_DHCR7, FOXM1_HMGCS2, HOXA5_KRT15, ACTL6A_TXNIP, MYBL2_SLC7A5, MYBL2_PNP, LRPPRC_MSH2, FOXM1_NPY1R	FOXO1_GSTP1, FOXO1_NDN, FOXO1_ARL4C, TGFB1I1_ANGPTL2, TGFB1I1_ACTN1, TCF4_CXCL12, ZFPM2_LUM, ZFPM2_PMP22, PRRX1_PLAU, PRRX1_SERPINH1, GATA3_KDM4B, GATA3_PMAIP1, SOX10_CDH3, IFI16_ADM, RBMS1_OSBPL8	GATA3_IL8, ELK3_ITPR2, CHD3_MYBBP1A, ZNF302_SR140, SOX10_FOXC1, RUNX1T1_PENK

The concordance analysis results relative to the set of 50 CM-gene and MR-gene networks are reported. The networks are grouped according to the transcriptional activation pattern : Non Concordance, Positive Concordance, or Negative Concordance patterns. The CM-gene networks are represented by their hub gene (i.e. the CM-gene imposed as the hub gene of the network). The MR-gene networks are represented by two gene symbols separated by underscore. The first gene name refers to the MR-gene. The second gene symbol refers to the original CM-gene.

3.5.2 Gene set enrichment analysis of transcriptionally active networks in triple-negative breast cancer (TNBC) patients

Triple-Negative Breast Cancer (TNBC) is an aggressive breast cancer subtype that lacks the expression of the hormone receptors, ER and PgR, and does not overexpress Her2/neu. Consequently, these tumors are unresponsive to the currently available targeted therapies for breast cancer, such as endocrine therapies (e.g. Tamoxifen) and anti-Her2 agents (e.g. Trastuzumab). Indeed, treatment options are limited for TNBC and chemotherapy remains the mainstay of treatment ([206]). Although chemotherapy can delay tumor progression in TNBC patients, it is not curative and the development of chemoresistance, resulting in disease progression, is common. Therefore, a better understanding of the mechanisms of chemoresistance in TNBC could help in the identification of novel molecular targets for the development of more effective breast cancer therapies. To determine whether the transcriptionally active networks that we identified in breast cancer (71 positively and 27 negatively concordant networks derived from the CM-gene and MR-gene GRNs) could be relevant to chemoresistance, we performed an enrichment analysis of these networks in the Hatzis et al.,([190]) cohort of 152 TNBC patients ([190]). In this cohort, the 152 TNBC patients had received neoadjuvant taxane-anthracycline chemotherapy (NACT; see Methods, Subsection 2.1.3.8 and Section 2.5). The response to this treatment varied from a pathologic complete response (pCR) with significant improvements in both disease-free survival and overall survival, to residual invasive disease (RD) with no benefit in terms of survival rate. Through this enrichment analysis, we aimed to identify networks "enriched" in chemorefractory tumors, in order to predict their involvement in chemoresistance mechanisms. Our ultimate goal was to identify putative molecular targets for the development of novel treatment strategies for TNBC. To identify networks associated with chemoresistance, we performed the Gene Set Enrichment Analysis (GSEA) and the Gene Set Analysis (GSA) in pCR and RD patients from the 152 TNBC cohort, using the transcriptionally active GRNs as gene sets. The GSEA and GSA algorithms are two of many computational tools, known as Functional Class Scoring (FCS) methods, available to investigate the transcriptional enrichment of gene sets with respect to different phenotype conditions. Briefly, the GSEA analysis tests whether the distribution of the ranks of genes in a gene set differs from an empirical null distribution, using a weighted Kolmogorov-Smirnov statistical test. The genes in the gene list are ranked by the strength of association with the phenotype, defined by, for example, the t-test, signal-to-noise ratio, correlation coefficients, or fold-change etc. Instead, the

GSA analysis uses the *maxmean* statistic to determine whether the strongest evidence for a particular gene set is the up-regulation or the down-regulation. Both methods are optimized to investigate harmonized changes of expression levels of genes, within a particular gene set, according to a desired condition (phenotype). For the enrichment analysis, we used the two methods GSEA and GSA, on gene expression data normalized according to three different normalization methods (i.e. RMA, MAS5 and MAS5-based normalization reported in Hatzis et al., 2011. See Methods, Section 2.5). We chose three normalization techniques in order to control the effect of the normalization methods on the enrichment results, and two methods for the enrichment analysis to verify the robustness of them. The rationale behind this choice was that if the active networks represent biologically relevant transcriptional mechanisms in RD TNBC, we expect to observe the enrichment, independently of the different theoretical formulations of the method (i.e. GSEA or GSA) used to perform the enrichment analysis. The statistical significance of the enrichment of the transcriptionally active networks was evaluated according to the nominal p-value resulting from the GSEA and GSA analysis output. We did not consider the FDR q-value to identify the statistically significantly enriched networks, since our aim was not to select the most enriched networks (i.e. it was not to perform network selection or to compare networks between them) from the set of networks used as inputs, but, instead, to evaluate the enrichment of each network as single mechanisms. The FDR computation, in fact, considers the distribution of the enrichment scores computed across all the gene sets ([197]). Moreover, we considered as significantly enriched, those networks found to be enriched (p-value$\leq$ 0.1) according to at least one normalization method and according to both enrichment analysis algorithms. Using this approach, we observed the enrichment of 6 transcriptionally active networks in RD TNCB: TCF4, TGFB1I1, ZFPM2, PRRX1, ELF4, COL1A1 (Table 3.13; Figure 3.22 and 3.23; Appendix A, orange section). A typical enrichment plot from the GSEA analysis is shown in Figure 3.22 for the TGFB1I1 transcriptional network. The gene neighbors of the network were ranked according to the Signal2Noise (S2N) metric, which is defined as the difference of means of expression levels of the two phenotype classes (PhLs), RD and pCR, scaled by their standard deviation. The genes of the TGFB1I1 network were sorted from high S2N (high association with RD pathological condition) to low S2N (low association with RD pathological condition) values. A representative set of the core genes of the enrichment analysis, i.e. the genes that contributed most to the enrichment of the entire TGFB1I1 network are also reported in Figure 3.22, including the PDGFRB receptor, a putative druggable target gene that functions as a cell

surface kinase receptor. A more comprehensive list of the GSEA core genes for each one of the 6 enriched networks is reported in Appendix B.

Table 3.13: **Enrichment results of the transcriptionally active networks in RD TNBC.**

Transcriptionally active network in BC	**(RMA)** **NES; p-val**	**(MAS5)** **NES; p-val**	**(MAS5, Hatzis 2011)** **NES; p-val**
GSEA			
TCF4	-	-	1.54; 0.074
TGFB1I1	1.43; 0.051	1.54; 0.052	1.63; 0.022
ZFPM2	1.39; 0.072	1.52; 0.061	1.56; 0.023
PRRX1	1.44; 0.021	1.64; 0.031	1.63; 0.011
ELF4	1.53; 0.044	1.72; 0.045	1.80; 0.002
COL1A1	1.46; 0.044	1.54; 0.054	1.61; 0.025
GSA			
TCF4	-	-	0.60; 0.100
TGFB1I1	0.78; 0.025	0.81; 0.020	0.82; 0.021
ZFPM2	0.96; 0.015	1.05; 0.005	1.03; 0.004
PRRX1	1.27; 0.002	1.24; 0.004	1.24; 0.002
ELF4	0.72; 0.023	0.74; 0.012	0.77; 0.006
COL1A1	1.69; 0.004	1.69; 0.004	1.64; 0.004

The enrichment results of transcriptionally active networks in TNBC patients resistant to neoadjuvant taxane-anthracycline chemotherapy (RD). The normalized enrichment score (NES) followed by the nominal p-value (p-val) is reported for the GSEA and GSA analyses, relative to the three normalization methods: RMA, MAS5 and the MAS5-based normalization proposed in Hatzis et al 2011 (MAS5 Hatzis 2011)([190]). Networks with an enrichment p-value ≥ 0.1 are indicated with the "-" symbol. The hub genes TCF4, TGFB1I1, ZFPM2 and PRRX1 are MR-hub genes relative to the original set of 50 GRNs enriched in mutually exclusive mutated genes. ELF4 and COL1A1 are CM-hub genes from the original set of 48 COSMIC-Census mutated genes.

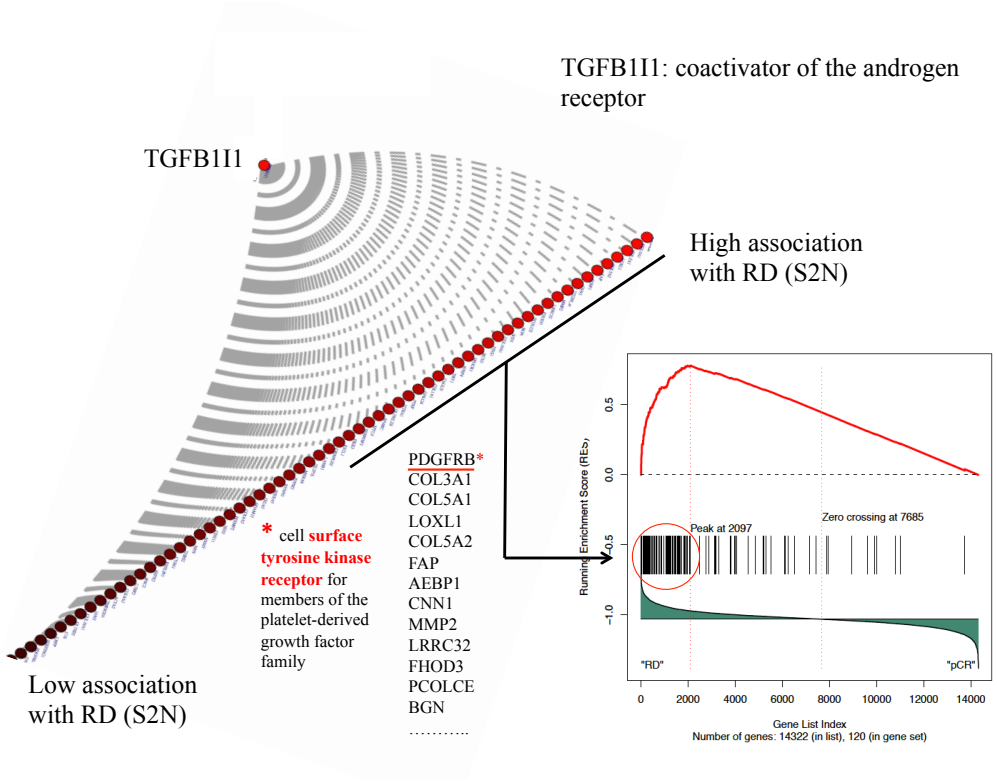

Figure 3.22: **Enrichment analysis plot for the TGFB1I1 network in RD TNBC tumors.**
The GSEA enrichment plot relative to the TGFB1I1 network enrichment analysis is reported. The TGFB1I1 gene neighbors in the network representation were sorted according to the S2N, from high (high association with RD pathological condition) to low (low association with RD pathological condition) values. Red circles and corresponding vertical black lines in the graph, highlight a subset of core genes that contributed most to the enrichment of the entire network in RD TNBC. For some of these genes, the Hugo gene name is reported next to the enrichment plot. The PDGFRB receptor is highlighted with a red line.

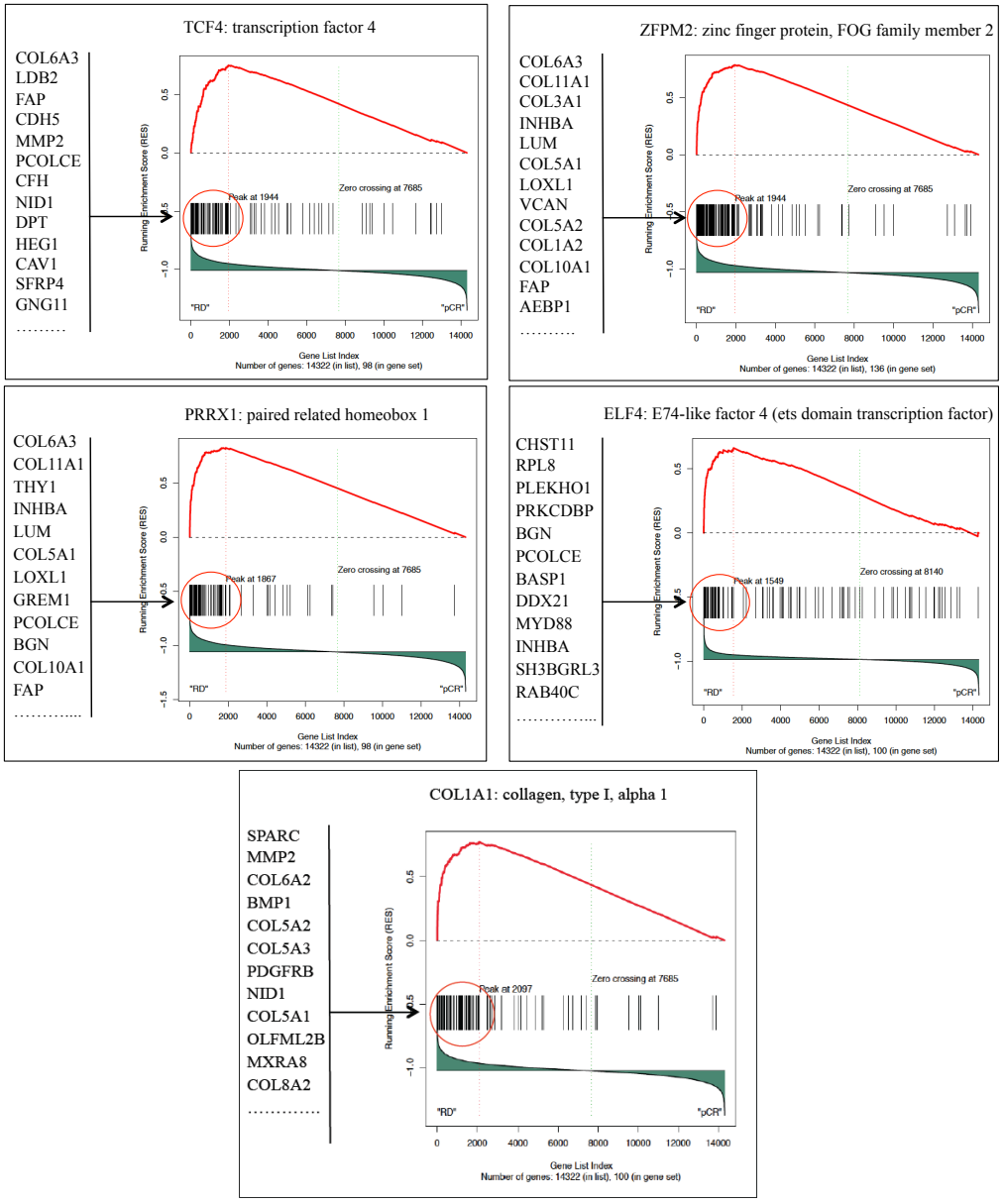

Figure 3.23: **GSEA plots relative to the enrichment analysis in RD TNBC of the transcriptionally active networks TCF4, ZFPM2, PRRX1, ELF4 and COL1A1.** The GSEA enrichment plots relative to the TCF4, ZFPM2, PRRX1, ELF4 and COL1A1 networks enrichment analysis are reported. Red circles highlight the core genes of the networks that mostly contributed to the Running Enrichment Score (RES). A representative set of 12 core genes ranked according to the S2N metric from high to low values are reported.

Chapter 4

Discussion

4.1 Summary

In this project, we aimed to identify GRNs associated with breast cancer from microarray gene expression data, in order to predict cancer biomarkers and novel druggable targets. Initially, we identified oncogenic gene sets whose collective expression profiles significantly correlated with different clinico-pathological parameters of breast cancers (e.g., ER status, tumor grade and prognosis). We then derived cancer gene modules (i.e., CMs) from these gene sets, by identifying those genes that contributed most to the correlation between the gene set expression profiles and the clinico-pathological parameters. Finally, regulatory networks were inferred from these CMs by assuming that each CM-gene was a hub gene around which to build the network. For the network inference analysis, we used a statistically representative gene expression dataset (the Loi et al., dataset [194]), in terms of number of tumors screened, which was derived using the Affymetrix HG-U133A chip. From the full set of 1,652 CM-genes, networks were inferred on a subset of 1,516 genes that were represented on the Affymetrix HG-U133A chip. For the remaining 136 genes, network inference was not possible (these genes are represented on the Affymetrix HG-U133B chip). The inference analysis was performed by using the ARACNE algorithm along with two other independent methods for network inference analysis: CUDA-MI and WGCNA. For each CM-gene from which a network was inferred, we observed a good agreement of the pair-wise transcriptional correlations according to the three methods supporting the robustness of the inferred networks. To gain insights into the functional relevance of CM-genes in breast cancer, we performed a mutational annotation of the full set of CM-genes (1,652 total genes). We identified 49

and 812 CM-genes mutated according to mutational annotation based on the COSMIC-Census and TCGA datasets, respectively. The significance of the of mutated genes in the set of CM genes was statistically verified. We prioritized the set of 49 mutated genes with respect to the 812 mutated genes for subsequent analyses, because of its small size and the fact that mutations in the COSMIC-Census dataset are documented in the literature. We considered networks derived from 48 CM-mutated genes, instead of the complete set of 49 genes, because for one gene, network inference analysis was not possible since it is represented on the Affymetrix HG-U133B chip. The mutational annotation was also performed on the gene neighbors of the full set of 1,516 networks in order to identify mutated genes that might impair the function of entire networks. This analysis allowed us to identify an additional set of 50 networks enriched in mutually exclusive mutated gene neighbors. The set of 48 networks and the set of 50 networks were finally considered as potentially functionally relevant networks for subsequent investigation in breast cancer. Using these two sets of networks, we performed an additional phase of network inference analysis, aimed at identifying the putative transcriptional MR-gene of each network. This analysis was performed to overcome the bias of *a priori* assuming CM-genes as transcriptional hub genes of networks. The identification of the MR-genes was performed as follows: for each CM-gene network, each neighboring gene was assumed to be a hub gene around which a new network was built using ARACNE. In these new networks, the set of transcriptionally interacting genes were the genes represented in the original CM-gene network. The most frequently occurring gene neighbor in the full set of new networks was considered as the putative transcriptional MR-gene. Once the MR-gene was identified, a final network was built, in which the MR-gene was assumed to be the hub gene and the full set of CM-genes in the original network were considered as putative transcriptionally interacting genes. Using this strategy, we identified MRs that were unique to specific CM-gene networks and MRs that were in common to different CM-gene networks. This analysis allowed us not only to identify putative transcriptional MR-genes, but also to group together CM-networks with common transcriptional regulatory programs that were under the control of the same MR-gene. By considering the set of 48 and 50 networks (i.e. the networks inferred from CM-genes [98 total networks] and the new networks inferred from MR-genes [76 total networks. CM-gene networks with the same MR gene were not collapsed to a unique network]), we subsequently evaluated the transcriptional activity of the entire networks in breast cancer tumors. We defined networks as transcriptionally active if more than half of the neighboring genes had the same or the opposite transcriptional modulation as

the corresponding hub gene (positive and negative concordance, respectively). In contrast, transcriptionally inactive networks were those in which half of the neighboring genes have the same transcriptional modulation as the hub gene, while the remaining half have the opposite transcriptional regulation. Using this strategy, we were able to further prioritize our set of networks to those most relevant to breast cancer in terms of transcriptional activity (98 total active networks: 71 positively and 27 negatively concordant networks). The clinical relevance of our computational predictions, in terms of putative transcriptional mechanisms deregulated in breast cancer, was assessed by evaluating the correlation between the expression profiles of transcriptionally active networks and the occurrence of RD in TNBC patients after neoadjuvant taxane-anthracycline chemotherapy. Interestingly, we identified six active networks (i.e. TCF4, TGFB1I1, ZFPM2, PRRX1, ELF4, COL1A1) whose transcriptional profiles correlated with RD in TNBC patients. These networks represent putative mechanisms responsible for the chemoresistance in these patients. Moreover, the genes of these networks could represent candidate biomarkers of therapy response, as well as putative druggable targets for the development of more effective therapeutic strategies.

4.2 Cancer Modules (CMs) definition from oncogenic gene sets: a biased approach

Carcinogenesis is mainly caused by genetic alterations, i.e. somatic mutations in oncogenes and tumor-suppressor genes ([207],[208],[209],[210]). Oncogenes control cell proliferation and apoptosis, while tumor-suppressor genes (also called anti-oncogenes) normally inhibit or "suppress" abnormal cell proliferation and induce apoptosis of abnormal cells. Genomic alterations (i.e., mutations) in proto-oncogenes and tumor-suppressor genes, cause such normal genes to become cancer-causing genes. Given the role of these genes in cancer, we identified modules of genes transcriptionally related to breast cancer (CMs) from publicly available oncogenic gene sets. An oncogenic gene set is a collection of genes showing a coordinate gene expression modulation upon the perturbation of known oncogenes and tumor suppressor genes. Such perturbations are generally experimentally-induced down-regulation, in the case of tumor-suppressor genes, or overexpression, in the case of oncogenes, which causes alterations in the expression of downstream genes in the oncogenic pathway. The induced perturbations in gene expression and subsequent alteration in the activity of such genes, recapitulates, in a biased way ([211]), the transcriptional regulatory events that are downstream of oncogenes and tumor suppressor genes in cancer cells. We collected experimentally-derived oncogenic gene sets representative of the major hallmarks of cancer, such as: i) cell proliferation, apoptosis and differentiation (i.e., MYC, MYC/TGFA, MYC/E2F1, MYB, TGFB1, SRC, JAG1/NOTCH, EGFR, KRAS, KRAS/PTEN, BCAT, HRAS, ERBB2, BRCA1, TERT, E2F1, E2F3, TP53, E1A); ii) EMT (i.e., ZEB1, JAP/-TAZ); iii) angiogenesis (i.e., VEGF, HIF1A/HIF2A); chromosomal instability (i.e., CIN). GSEA analysis allowed us to investigate if the genes belonging to these oncogenic gene sets were also transcriptionally modulated (up-/down-regulated) in breast tumors as the result of inactivation of the normal function of proto-oncogenes and tumor suppressor genes. We found that the genes belonging to 18 of the original 23 oncogenic gene sets were up-/down-modulated in tumors and also according to different pathological conditions that characterize breast cancer: i.e., ER status, tumor grade and prognosis. Further analysis revealed that genes coming from the full list of significantly enriched oncogenic gene sets identified by GSEA, clustered into 7 "major" gene modules, the CMs: Grade, Grade/ER, All, Relapse/Survival/Grade/ER, Survival/Grade/ER, ER, and Survival. These findings support the notion that the experimentally-derived oncogenic gene sets recapitulate the activation of oncogenic pathways in breast tumors and also that

the related genes are transcriptionally modulated according to the pathological condition of the disease (Figure 3.2). For example, we observed the enrichment of the genes belonging to the E1A, as well as to the CIN or MYC gene sets, in different and non-overlapping CMs (Figure 3.3). This result suggests a differential functional involvement of the genes originating from a common oncogenic lesion, according to the state of the tumor, thus, helping in the identification of the molecular determinants of the clinical state of the disease. Importantly, our predictions (i.e., the CMs) were *in silico* verified in an independent set of breast cancer datasets, confirming that the observed enrichment was not cohort-dependent. Although the selected gene sets represent only a small fraction of all the cancer-related oncogenic events occurring in a tumor cell, this approach represents a flexible strategy towards the deconvolution of the altered molecular mechanisms responsible for the disease and may be easily extended as new oncogenic gene sets become available.

4.3 Gene Regulatory Networks inference analysis: identification of cancer-related mechanisms

Network inference from CM-hub-genes

In this project, we applied a semi-unsupervised data-driven approach to reverse engineer GRNs, starting from publicly available gene expression microarray datasets. Data-driven network inference analysis methods can be classified into two groups: "unsupervised" and "supervised". Unsupervised methods infer functional relationships between genes directly from the data. The unsupervised reconstruction of regulatory networks from "genome-wide" expression data is computationally intensive because of the high-dimensional space of transcriptional data (i.e. it considers the full set of genes expressed in a cell under a particular condition). Moreover, the biological interpretation of the predicted functional relationships between genes is often unfeasible because of the huge amount of inferred interactions. In order to reduce the complexity of the analysis and to simplify the biological interpretation of data, alternative "supervised" methods have been proposed for GRNi. These methods consist of inferring GRNs from a set of genes whose regulatory interactions are already known; this set of genes is used as the training set. Although this approach reduces the complexity of the inference analysis, it is biased towards well characterized genes, which are weighted in the analysis more than less studied genes. In addition, for the vast majority of the human genes the regulatory interactions are still unknown. To overcome the

limitations of the unsupervised and supervised methods for GRNi, we adopted a semi-unsupervised network inference approach. Specifically, networks were built starting from the CMs that we identified using published gene sets, representing known oncogenic stimuli; i.e., the GRNi analysis started from a set of cancer-relevant genes (CM-genes) that we used as hub genes around which functional relationships were predicted among genes found correlating with the hub gene expression profile (i.e. the gene neighbors). These CM-genes worked as "routers" to guide the inference of regulatory interactions in the multidimensional transcriptional space, in an unsupervised way using high-throughput expression data. Regulatory network inference at a global expression level in mammalian cellular contexts (i.e. considering the full set of expressed genes in a human or mouse primary cell) is a complex task because of the complexity of the regulatory programs characterizing such living organisms, at high end of the evolutionary ladder. Indeed, the majority of algorithms were developed to infer GRNs in simpler model organisms ([168],[170],[212]). Nevertheless, several algorithms have been developed to infer regulatory mechanisms from the full set of gene expression data in mammalians ([131],[133],[193],[213],[214]). These algorithms, however, suffer from a number of limitations. The first limitation concerns the inference analysis from static expression data for the vast majority of them. Cells, in fact, are adaptive systems with dynamic properties and "static measurements" of expression levels do not incorporate temporal, spatial and conditional information except indirectly. The modeling of regulatory networks that take into consideration the dynamic properties of cellular systems (i.e. the evolution of cellular systems over the time), requires time-series gene expression data. The time-related changes of expression levels allow the inference of causal relationships between biological molecules and of more accurate regulatory mechanisms. In particular, time-series expression data might be useful to gain insights into the transcriptional programs that govern the cellular behavior of highly perturbed and highly evolving living systems like cancer cells. Unfortunately, with the exception of model organisms and cell lines systems, the generation of expression profiles at different time points directly from human cancer tissues is unfeasible because of ethical reasons and because individuals with cancer quickly undergo surgical resection or pharmacological treatments. Thus, it is not possible to monitor changes in gene expression levels over several time points directly from tumor biopsies. Static expression profiles from cancer tissues before treatment are, therefore, the only source of data from which to infer regulatory probabilistic mechanisms. In this project, GRNi analysis was performed using primarily the ARACNE algorithm

([133]). This algorithm was preferred because of its literature documented reliability as a "genome-wide" network inference algorithm in the mammalian context ([131]), consisting in the experimentally validated subset of gene-to-gene interactions. In addition, from a theoretical perspective, the ARACNE algorithm infers transcriptional interactions also from non-linearly dependent variables (genes) through the MI measure, extending the possible pair-wise associations with respect to the correlation-based methods, such as those used in the clustering, which instead infer interactions only from linearly-dependent variables. Finally, the ARACNE algorithm implements the DPI procedure that allows the removal of putative indirect interactions. Although the ARACNE algorithm infers reliable (i.e. experimentally-proved) pair-wise interactions and despite the fact that we performed 1,000 bootstrap replications to prove the statistical significance of the transcriptional correlations, the ability of reverse engineering methods, including ARACNE, to infer realistic gene-to-gene associations from microarray gene expression data faces two limitations. The first limitation concerns the impact of measurement noise, especially for genes expressed at very low levels, which affects the reliability of the predictions. The second limitation concerns the dimensionality curse phenomena (i.e. the number of genes is higher than samples) that prevents to accurately recover the pair-wise gene interactions from their expression level. Two possible strategies to verify the reproducibility and hence the robustness of the inferred interactions are: i) to perform the network inference analysis using one or more independent dataset of expression profiles, i.e. the validation datasets; ii) to perform network inference analysis on the same dataset, but using different network inference methods. In our case, the first strategy was unfeasible because gene expression datasets with the same or comparable number of expression profiles were not available (327 breast tumor transcriptional profiles). Thus, we applied the second strategy under the assumption that if an ARACNE-predicted pair-wise interaction is statistically robust (i.e. it does not represent a false positive finding, but instead a true finding), it will also be inferred using alternative algorithms, based on alternative measures to score the gene-to-gene transcriptional correlation. We therefore performed GRNi analyses using the CUDA-MI ([192]) and the WGCNA ([193]) algorithms. CUDA-MI algorithm implements the MI computation through the use of the B-spline functions, while WGCNA builds weighted gene co-expression networks by "soft thresholding" the Pearson correlation coefficient for gene-to-gene interaction predictions. We chose these two alternative algorithms because in the first case, although the correlation is estimated through the MI measure, data discretization is performed using B-spline functions instead of the ARACNE adaptive partitioning,

while in the second case, the correlation is measured using a totally different measure. Notably, using the Cohen test to compare the networks inferred from each CM-gene using the three methods, we observed that for the vast majority of ARACNE-inferred networks (i.e. 75% of the 1,498 statistically significant Cohen tests, from a total of 1,516 networks), the pair-wise transcriptional interactions were the same as those inferred using the two alternative methods. The agreement we observed between the three algorithms, reinforced the relevance of the inferred transcriptional networks to breast cancer.

4.4 Mutational annotation of CM-genes and the mutual exclusivity analysis

4.4.1 Mutational annotation of CM-genes

Cancers arise mainly as the result of the acquisition of a number of somatic genomic alterations, such as point mutations, copy number alterations, epigenetic changes and karyotypic rearrangements, which confer a selective advantage characterized by uncontrolled cell proliferation with respect to normal cells and escape from apoptotic control ([215],[216]). Recent advances in massively parallel, high-throughput sequencing of DNA (exome and whole genome sequencing) has allowed a comprehensive characterization of DNA somatic mutations through the sequencing of a large number of tumor samples, and provided an unprecedented opportunity to gain biological insights to the origin and evolution of cancer. Much of the available mutational data (and genomic data in general) comes from a handful of large international collaborations: The Cancer Genome Atlas (TCGA, NCI and NHGRI), The Catalogue of Somatic Mutations in Cancer (COSMIC, UK Cancer Genome Project) and the more general Cancer Genome Consortium (ICGC) that allows access to both TCGA and COSMIC data. The aim of such collaborative efforts is to comprehensively understand the molecular basis of cancer, not only at mutational level, but more generally at a genomic level; gene expression data, methylation data and other genomic data are also available. To better characterize the functional role of CM-genes previously predicted to be associated at the gene expression level (GSEA analysis) to breast cancer, we performed a mutational annotation of the CM-genes representing the hub genes around which the networks were built. This analysis allowed us, not only to gain insights into the involvement of CM-genes in cancer, but also to

reduce the set of 1,516 networks to those most biologically relevant. The mutational annotation was primarily performed (on the full set of 1,652 CM-genes) using the Cosmic Cancer Gene Census ([217]) dataset relative to breast cancer, because this collection of mutated genes has been causally implicated in cancer. According to this analysis, 49 CM-genes appeared to be mutated. We also performed the mutational annotation by using TCGA mutational data relative to breast cancer. This analysis was performed to overcome the bias of the mutational annotation performed according to a subset of well-characterized mutated genes. Many mutated genes, in fact, still lack biological validation, but might nevertheless be equally involved in breast cancer. TCGA data allows an unbiased whole-genome mutational annotation. According to TCGA mutational annotation, 812 CM-genes appeared to be mutated in breast cancer. Importantly, for both mutational annotations, we observed a statistically significant enrichment of mutated genes in CM-genes with respect to the empirical null distribution we generated from a collection of 1,000 random gene lists (p-value$<$ 0.001). This significant enrichment suggests that the CM-genes might be involved in breast cancer, not only at the gene expression level, but also at the mutational level. Moreover, this finding indicates that the strategy we used for evaluating the enrichment of oncogenic gene sets in breast cancer PhLs, might be relevant to select putative key cancer-related genes. Undoubtedly, further investigations are needed to clarify the relationship between the presence of the mutational event in CM-genes and their transcriptional association with the disease. Although we subsequently focused our attention on the networks derived from the smaller set of 48-CM-mutated genes (for one of the set of 49 genes, network inference was not possible), the set of 812 TCGA mutated genes might contain novel, biologically uncharacterized mutated genes, which could have an important role in cancer.

4.4.2 Mutual Exclusivity analysis

Cancer genomes contain "driver" mutations and "passenger" mutations. The former are causative of the tumor, while the latter are neutral mutations that occur randomly, during cell division, without functional consequences. Large-scale cancer genomic projects, like the TCGA (`http://cancergenome.nih.gov/`) and ICGC (`https://dcc.icgc.org/`), with their high-resolution view of molecular defects, at the DNA level, in different types of tumors, and whole-genome sequencing that allows the analysis of more than 20,000 protein-coding genes, offer an unprecedented opportunity to determine which mutations are drivers and which mutations are passengers. According to a large fraction of sequencing

projects, the mutational profile, for most cancer types, consists of a small number of genes altered in a high percentage of tumors, the so-called "mountains" and a large faction of genes altered infrequently in the population ($< 1\%$), the so-called "hills" ([218]). High frequency mutations in the population (i.e. the mountains), such as mutations in the TP53, MYC, KRAS, ATM, APC, EGFR, PIK3CA, BRAF, JAK2, and FGFR2 genes, confer a selective growth advantage to cancer cells. Although the role of these high frequency mutations as drivers in cancer has been extensively demonstrated through genome sequencing of a large number of individuals and experimental validation, the role of low frequency mutations (i.e. the "hills"), as drivers, is still not fully characterized. The main reason for this, is that their low frequency of occurrence in the population, resembles the frequency expected for neutral passenger mutations. One possibility for elucidating the role of low frequency mutations in individuals and to address inter-tumor heterogeneity, is to investigate the function of these infrequently mutated genes in a pathway-context. Indeed, it is well-known that different gene mutations can target the same pathway ([219],[220]). Moreover, the presence of a single mutated gene is sufficient to perturb the entire pathway ([220], [201]), such that the mutation of key genes belonging to the same pathway exhibit a mutually exclusive behavior. In this project, we inferred networks of transcriptionally correlating genes from CM-genes, predicted to be associated with breast cancer at the transcriptional level. Hence, we have network tools (i.e. pathway tools) to investigate for the presence and to predict the role of low frequency mutations in individual breast cancer patients. From the mutational annotation of the gene neighbors of each network (TCGA[BRCA] mutational data), we observed that some genes were mutated with a frequency of $< 1\%$ in the population, i.e. low frequency mutations or "hills", while others were mutated with a frequency of $\sim$37%, i.e. high frequency mutations genes or "mountains", as expected. Moreover, we observed a mutually exclusive behavior of a fraction of mutated gene neighbors of the inferred networks (Figure 3.11). For a subset of 50 networks, out of the total set of 1,516 networks, the enrichment of mutually exclusive mutated genes was statistically significant. Interestingly, among the set of mutually exclusive mutated genes, we observed that, not only high frequency mutated genes (e.g. PIK3CA) exhibited a mutually exclusive mutational behavior, but also low frequency mutated genes. This result suggests that the presence of a mutational event in these latter genes might confer the same selective advantage as known high frequency driver mutations in the population. To further assess the robustness of the mutually exclusive relationships of genes, we plan to benchmark our *in silico* predictions by applying computational methods

based on different mathematical and statistical frameworks like those reported in: [200],[221],[222],[223].

4.5 Identification of putative Transcriptional Master Regulators

The network inference analysis to identify cancer-related mechanisms was primarily performed by assuming that each CM-gene was the hub gene of the network. This assumption derived from the observed transcriptional regulation of CM-genes in breast cancer gene expression profiles, according to the pathological state of the disease, thus sustaining their centrality as cancer-related genes. Despite the biological rationale behind this assumption, the transcriptional centrality of such CM-genes as regulators of transcriptional programs required additional investigation. For this reason, we performed an additional network inference analysis in order to assess the role of these CM-genes as transcriptional hubs, and to identify the putative transcriptional MRs at the top of the transcriptional regulation hierarchy. The inference analysis was performed by assuming iteratively each gene neighbor of each CM-gene network (i.e. from the set of 48 GRNs identified through the COSMIC-Census mutational annotation and from the set of 50 GRNs significantly enriched of mutually exclusive mutated genes), as the hub gene for a new network. Then, the most frequently occurring gene neighbor, i.e. the most highly interconnected gene in the new networks was considered to be the putative transcriptional MR-gene. For a set of CM-gene networks (i.e., 78% on 98 total CM-gene networks), the MR-gene was different with respect to the CM-gene, i.e. it was a neighbor from the original CM-gene hub network. Among this group, for a subset of CM-gene networks, the putative inferred MR-gene was the same, meaning that we were able to group together CM-gene networks representing sub-mechanisms of a global regulatory transcriptional network. Notably, for a set of CM-gene networks (i.e., 22% on 98 total networks), the putative MR-gene was confirmed as the CM-gene. This observation confirmed the centrality and the putative transcriptional relevance of the CM-genes, as well as the validity of the approach used to identify cancer-relevant genes. Although the identification of the MR-genes was limited to the set of the gene neighbors composing the initial CM-gene network, our approach has the advantage of simplifying the search space of candidate MRs, which would otherwise be infeasible considering the full transcriptome of a cell. Using this approach, we have intermediate levels of regulation, among which to investigate for the presence of MRs.

4.6 Clinical relevance of breast cancer-related networks

TNBC is a subtype of breast cancer characterized by the lack of the ER, the PgR and Her2 ([206]). Drugs like Tamoxifen or Trastuzumab, which are routinely used in clinic as pharmacological treatments for ER-, PgR- or Her2-positive breast cancers cannot be used to treat TNBC. In addition, TNBCs are clinically characterized by high malignancy, high risk of the local recurrence, poor prognosis (i.e. poor disease-free survival), and poor cancer-specific survival ([224],[225]). Molecularly, they are characterized by high proliferation and mitotic rates. The risk of recurrence in TNBC is higher in the 3-5 years after diagnosis with respect to ER-positive breast cancers ([226],[227]). Few therapeutic strategies are available for TNBC and chemotherapy is the only effective treatment for TNBC patients after surgery ([228]). Several studies have shown that TNBCs are much more sensitive to adjuvant or neoadjuvant chemotherapy than other subtypes of breast cancer ([229]). pCR to neoadjuvant chemotherapy (mainly anthracyclines alone or in combination with taxanes) correlates with a better prognosis in responsive TNBC patients, with an overall survival similar to that of the non-TNBCs. Despite the therapeutic benefits achieved in TNBC patients showing a pCR to neoadjuvant chemotherapy, accounting for approximately 30% of TNBCs, there is a fraction of TNBC patients who arae resistant to neoadjuvant chemotherapy. These patients present RD after neoadjuvant chemotherapy and are characterized by high rates of metastatic recurrence (because of the presence of viable cancer cells in breast or in lymph nodes), and overall poor clinical outcome ([227],[230]). Thus, new treatments are urgently needed to treat TNBCs resistant to chemotherapy. To achieve this, it is crucial to elucidate the molecular mechanisms responsible for chemoresistance, in order to molecularly sensitize cancer cells to pharmacological treatments in a targeted way. In this project, we inferred gene networks representative of putative breast cancer-related mechanisms. The case of unresponsive RD TNBC patients represented a good opportunity to evaluate the clinical relevance of our predictions, i.e. to "translate" our findings to the clinic to meet an unmet clinical need. The clinical relevance of the inferred cancer mechanisms was assessed by investigating the transcriptional profiles of the breast cancer active networks (GSEA analysis) in RD TNBCs. Active networks were defined as mechanisms in which the vast majority of neighboring genes had the same or the opposite transcriptional regulation with respect to the hub gene around which the network was built (i.e., the CM-gene or MR-gene), as assessed by the concordance analysis. The importance of defining the transcriptional activity of a

network lies in the possibility to pharmacologically revert the cancer-associated transcriptional profile of an entire mechanism to a non-cancer profile. More specifically, in the case of RD TNBC, to identify alternative drug targets for overcoming mechanisms of chemoresistance. From the initial full set of active networks, we identified a small subset of 6 networks (TCF4, TGFB1I1, ZFPM2, PRRX1, ELF4, COL1A1) transcriptionally regulated (up-/down-modulated) in RD TNBCs with respect to pCR TNBCs. These networks represent putative mechanisms associated with chemoresistance, and although further investigation is needed to define their precise role in RD TNBC, evidence in literature supports the functional involvement of at least some of these networks in the RD TNBC biology. For example, PDGFRB is one of the core genes of the TGFB1I1 network that contributed most to the enrichment of this network in RD TNBCs (i.e. it highly correlated at the transcriptional level with the RD pathological condition with respect to the pCR condition; see Figure 3.21). Notably, it has been recently demonstrated that PDGFR signaling contributes to TGFB-induced EMT in oncogenic mammary epithelial cells, and, consequently, to the metastatic potential of these cells ([231]). Moreover, the vast majority of the enriched core genes in the TGFB1I1 network are involved in remodeling of the extracellular matrix (COL6A3, COL3A1, COL5A1, COL5A2, LOXL1, MMP2, PCOLCE), supporting the involvement of the TGFB1I1 network as a whole in the EMT process in highly metastatic TNBCs. These genes represent, also, a fraction of the enriched core genes of the remaining 5 networks (Figure 3.22), i.e. they are in common between all the 6 networks, suggesting a cooperative involvement of all the 6 networks in the EMT process as branches of a more global molecular mechanism and offering multiple putative alternatives to pharmacologically target the entire EMT process. Using our strategy, we identified putative molecular mechanisms that, at the transcriptional level, might provide predictive biomarkers of resistance to currently available therapies for TNBC. In contrast to typical biomarker studies that screen for aberrant expression of single genes in tumors to derive standalone tumor markers, our approach has the advantage that we first defined the association of well-known cancer mechanisms (i.e. proliferation, apoptosis, angiogenesis, tissue invasion and metastasis mechanisms) with specific pathological conditions and then derived biomarkers involved in these mechanisms. As reported in [79], there are approximately 150,000 papers documenting the identification of thousands of biomarkers, however, only $\sim$100 of them have been validated and approved for clinical practice. There are many reasons that account for this failure in translation to the clinic. One main reason concerns

the lack of an in-depth knowledge on the molecular mechanisms involved in disease. In this project, we developed a computational pipeline to infer molecular mechanisms functionally related to breast cancer that might be used, not only as sources of novel cancer biomarkers, but also of new druggable targets.

4.6.1 Ongoing work and future plans

Experimental validation of the transcriptionally regulated active networks in RD TNBCs

To determine the biological relevance of the *in silico* predicted networks that transcriptionally correlate with RD TNBC, we are now performing RNA interference (RNAi) experiments on the hub genes of the networks. Specifically we firstly selected a panel of basal-like breast cancer cell lines expressing the hub genes: TCF4, TGFB1I1, ZFPM2, PRRX1, ELF4, and COL1A1. The expression of them was assessed through the quantitative real-time RT-PCR (RT-qPCR). By using the short hairpin RNAs (shRNAs) we are now perturbing the expression of the hub genes. We expect to observe a gene expression modulation of the gene neighbours of the entire network upon the loss-of-function of the hub gene, if the statistically inferred dependencies, biologically reflect molecular and biochemical mechanisms of gene expression regulation acting in breast cancer cells. We are using the high-throughput OpenArray technology to investigate the changes in the mRNA levels of the gene neighbours of each candidate network upon interference with the expression of the hub gene.

Computational analysis

We will build CM and MR-GRNs from breast cancer TCGA RNA-Seq expression data in order to confirm the gene expression interactions predicted by using Affymetrix expression data. We will perform an in-depth *in silico* molecular characterization of the networks that we predicted to correlate with RD TNBC expression profiles through the integration of multi-level large "-omics" data in order to gain insights into the molecular mechanisms that might govern the transcriptional state of them. In particular, we will annotate the genes of the networks for the presence of targets of known TFs through the integration of publicly available ChIP-Seq data on breast cancer. We will, also, annotate the genes, for the presence of predicted targets of miRNA that might control the expression levels of them as well as for the presence of breast cancer related patterns of DNA methylation that, together with miRNA, might represent mechanisms of epigenetic gene expression regulation of the networks associated with RD clinical condition in TNBCs. TCGA methylation data and miRNA expression data as well as Metabric miRNA expression profiles will be used to investigate for the presence of epigenetic mechanism of gene expression regulation. We will also annotate the genes of the networks for the presence of protein-protein interactions, taking advantage of recently generated proteomic profiling data ([232]),

to investigate for the presence of physical interactions that might give rise to molecular events responsible for the resistance to the neoadjuvant pharmacological treatment. Finally, we will interrogate databases of drugs and bioactive molecules (Mantra 2.0 `https://http://mantra.tigem.it/`, Connectivity Map 2, `https://www.broadinstitute.org/cmap/`) in order to *in silico* predict putative molecules (agents) able to target the genes of the networks associated with RD pathological condition thus representing candidate alternative pharmacological treatments in TNBC patients that do not respond to the neoadjuvant taxane-anthracycline chemotherapy.

Appendix A

Appendix A: Computational pipeline

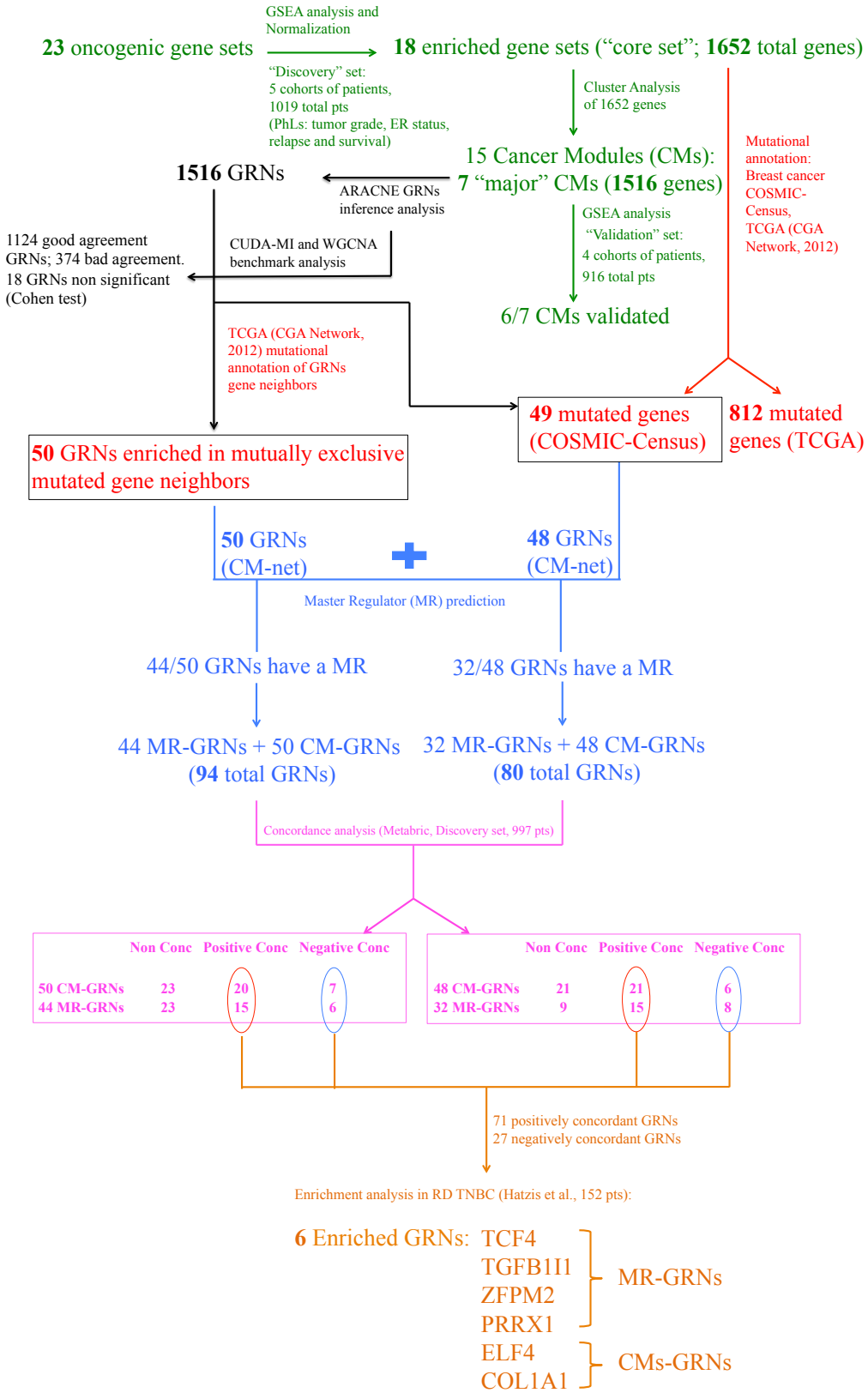

Figure A.1: **Computational pipeline used to identify transcriptional breast cancer networks.**

The computational pipeline used to infer breast cancer relevant transcriptional networks from microarray gene expression data.

Green section: to identify groups of genes (gene modules) whose expression correlate with clinical-pathological features of breast cancer disease, we performed the Gene Set Enrichment Analysis (GSEA) of 23 oncogenic gene sets on microarray gene expression profiles relative to 5 cohorts of breast cancer patients ("Discovery set") for a total of 1,019 patients. From the GSEA, 18 out of 23 oncogenic gene sets (the "core set") resulted significantly enriched in breast tumours according to the following clinical-pathological variables (Phenotype Labels, PhLs): the tumor grade, the ER status, the survival and the relapse. We then performed the cluster analysis of the 1,652 genes, composing the 18 enriched oncogenic gene sets, in order to cluster together the core genes according to the relative PhLs enrichment. We identified 15 total groups of genes (i.e. clusters) we called "Cancer gene Modules" (CMs). Of the full set of CMs 7 were "major" CMs (1,516 total genes), i.e. they represented clusters with a higher content of genes with respect to the other clusters. We further validated the enrichment of the 7 CMs by performing GSEA on an independent set of microarray gene expression profiles relative to 4 cohorts of breast cancer patients ("Validation set", 916 total patients). The clinical-pathological variables we considered were the same we used for the GSEA analysis performed on the "Discovery set", i.e. he tumor grade, the ER status, the survival and the relapse. We finally validated 6 out of 7 CMs.

Black section: to identify breast cancer relevant mechanisms at gene expression level, we performed the GRNs inference analysis using ARACNE algorithm on the full set of 1,516 genes composing the 7 CMs. We assumed each gene of the full set of 1,516 genes to be the "hub" gene of the network. We inferred 1,516 total networks (CM-GRNs). We also performed the network inference analysis by using two independent algorithms with respect to ARACNE, i.e. CUDA-MI and WGCNA, to assess the *in silico* reliability of the predicted networks. We performed the Cohen test in order to evaluate the statistical significance of the overlap of the neighbours of each network inferred by using the three algorithms. 1,498 networks on 1,516 total GRNs were found to be statistically significantly concordant of which, 1,124 showed good agreement and 374 bad agreement. 18 networks were found to be not statistically significantly concordant.

Red section: we performed the mutational annotation (COSMIC-Census and TCGA, CGA Network 2012, mutational data) of the full set of 1,652 genes composing the 15 total CMs to further assess their association with breast cancer disease not only at transcriptional level. We found 49 mutated genes according to the COSMIC-Census mutational annotation and 812 according to the TCGA annotation. We also performed the mutational annotation of the gene neighbours of the 1,516 networks inferred by ARACNE in order to investigate their relevance in breast cancer disease as oncogenic mechanisms. We identified 50 GRNs significantly enriched in mutually exclusive mutated gene neighbours. We then considered the set of 50 GRNs and 48 GRNs (98 total CM-GRNs) we prioritized through the mutual exclusivity analysis and the mutational annotation for further investigation. We reduced the set of 49 GRNs to 48 GRNs because for one hub gene the network inference analysis by ARACNE was not possible.

Blue section: we then performed an in-depth GRN deconvolution analysis to identify putative transcriptional Master Regulator (MR) hub genes for the set of 48 and 50 CM-GRNs. For 44 out of 50 GRNs and for 32 out of 48 GRNs we identified a putative MR hub gene different from the CM gene we previously assumed to be the hub gene of the transcriptional network. New networks were then inferred considering each MR-gene as the hub gene of the GRN.

Pink section: we performed the concordance analysis (i.e. we characterized the transcriptional regulation of the gene neighbours of each GRN with respect to the hub gene) in order to characterize the transcriptional activity of the set of 94 GRNs (44 MR-GRNs and 50 CM-GRNs) and of the set of 80 GRNs (32 MR-GRNs and 48 CM-GRNs) in breast tumours. The concordance analysis was performed on Metabric cohort of patients (Discovery set, 997 patients). For the set of 50 CM-GRNs we found 23 Non Concordant GRNs, 20 Positively Concordant GRNs and 7 Negatively Concordant GRNs; For the set of 44 MR-GRNs we found 23 Non Concordant GRNs, 15 Positively Concordant GRNs and 6 Negatively Concordant GRNs; For the set of 48 CM-GRNs we found 21 Non Concordant GRNs, 21 Positively Concordant GRNs and 6 Negatively Concordant GRNs and for the set of 32 MR-GRNs we found 9 Non Concordant GRNs, 15 Positively Concordant GRNs and 8 Negatively Concordant GRNs.

Orange section: we predicted the clinical relevance of our findings by investigating the transcriptional correlation of the full set of 71 Positively Concordant networks (relative to the 50 CM-GRNs, 44 MR-GRNs and to the 48 CM-GRNs and 32 MR-GRNs) and 27 Negatively concordant networks (relative to the 50 CM-GRNs, 44 MR-GRNs and to the 48 CM-GRNs and 32 MR-GRNs) with

the expression profiles of 152 TNBC with Residual Disease (RD) pathological condition after neoadjuvant taxane-anthracycline chemotherapy. We identified 6 enriched GRNs: TCF4, TGFB1I1, ZFPM2, PRRX1, ELF4 and COL1A1 representing putative mechanisms involved, at transcriptional level, in the metastatic process of RD TNBC tumours. The first four networks were inferred from the MR gene (MR-GRNs) while the last two were inferred from CM-genes (CM-GRNs). The 6 GRNs represent our final candidate set of breast cancer related mechanisms for the experimental validation of their biological relevance.

Appendix B

Appendix B: Transcriptionally active networks enriched in RD TNBC

Table B.1: **A representative list of 42 core genes (GSEA analysis) of the TCF4 network enriched in RD TNBC.**

Gene Symbol	Description
TCF4	transcription factor 4
CRISPLD2	cysteine-rich secretory protein LCCL domain containing 2
FSTL1	follistatin-like 1
SPARC	secreted protein, acidic, cysteine-rich (osteonectin)
FAP	fibroblast activation protein, alpha
CFH	complement factor H
PCOLCE	procollagen C-endopeptidase enhancer
MXRA5	matrix-remodelling associated 5
DPT	dermatopontin
FILIP1L	filamin A interacting protein 1-like
HEG1	HEG homolog 1 (zebrafish)
CAV1	caveolin 1, caveolae protein, 22kDa
ADAMTS5	ADAM metallopeptidase with thrombospondin type 1 motif, 5
OLFML2B	olfactomedin-like 2B
A2M	alpha-2-macroglobulin
MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
CXCL12	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
CILP	cartilage intermediate layer protein, nucleotide pyrophosphohydrolase
CFI	complement factor I
COL15A1	collagen, type XV, alpha 1
ITGBL1	integrin, beta-like 1 (with EGF-like repeat domains)
SPARCL1	SPARC-like 1 (hevin)
EFEMP2	EGF-containing fibulin-like extracellular matrix protein 2
NDN	necdin homolog (mouse)
ZCCHC24	zinc finger, CCHC domain containing 24
IGF1	insulin-like growth factor 1 (somatomedin C)
GPR124	G protein-coupled receptor 124
SFRP4	secreted frizzled-related protein 4
COL6A3	collagen, type VI, alpha 3
CDH5	cadherin 5, type 2 (vascular endothelium)
LDB2	LIM domain binding 2
SERPINF1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1
NID1	nidogen 1
HTRA1	HtrA serine peptidase 1
GNG11	guanine nucleotide binding protein (G protein), gamma 11
SLIT2	slit homolog 2 (Drosophila)
LAMB1	laminin, beta 1
ZFPM2	zinc finger protein, multitype 2
MEIS2	Meis homeobox 2
PTRF	polymerase I and transcript release factor
DCN	decorin
IGFBP6	insulin-like growth factor binding protein 6

A representative list of 42 core genes (GSEA analysis) of the TCF4 network enriched in RD TNBC is reported. The HUGO Gene Symbol is reported, followed by the full gene name (i.e. Description). Only 42 genes are reported for simplicity.

Table B.2: **A representative list of 42 core genes (GSEA analysis) of the TGFB1I1 network enriched in RD TNBC.**

Gene Symbol	Description
TGFB1I1	transforming growth factor beta 1 induced transcript 1
PDGFRB	platelet-derived growth factor receptor, beta polypeptide
COL3A1	collagen, type III, alpha 1
COL5A1	collagen, type V, alpha 1
LOXL1	lysyl oxidase-like 1
COL5A2	collagen, type V, alpha 2
FAP	fibroblast activation protein, alpha
AEBP1	AE binding protein 1
CNN1	calponin 1, basic, smooth muscle
MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
LRRC32	leucine rich repeat containing 32
FHOD3	formin homology 2 domain containing 3
PCOLCE	procollagen C-endopeptidase enhancer
BGN	biglycan
COL6A3	collagen, type VI, alpha 3
NID1	nidogen 1
COL16A1	collagen, type XVI, alpha 1
HEG1	HEG homolog 1 (zebrafish)
RCN3	reticulocalbin 3, EF-hand calcium binding domain
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1
SFRP4	secreted frizzled-related protein 4
THY1	Thy-1 cell surface antigen
TAGLN	transgelin
COL1A1	collagen, type I, alpha 1
GPR124	G protein-coupled receptor 124
PTRF	polymerase I and transcript release factor
HTRA1	HtrA serine peptidase 1
OLFML2B	olfactomedin-like 2B
SPARC	secreted protein, acidic, cysteine-rich (osteonectin)
MYL9	myosin, light chain 9, regulatory
SERPINF1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1
DKK3	dickkopf homolog 3 (Xenopus laevis)
FSTL1	follistatin-like 1
CRISPLD2	cysteine-rich secretory protein LCCL domain containing 2
LAMB1	laminin, beta 1
FILIP1L	filamin A interacting protein 1-like
MXRA8	matrix-remodelling associated 8
ZFPM2	zinc finger protein, multitype 2
SYNPO	synaptopodin
MXRA5	matrix-remodelling associated 5
ITGA5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)

A representative list of 42 core genes (GSEA analysis) of the TGFB1I1 network enriched in RD TNBC is reported. The HUGO Gene Symbol is reported, followed by the full gene name (i.e. Description). Only 42 genes are reported for simplicity.

Table B.3: **A representative list of 42 core genes (GSEA analysis) of the ZFPM2 network enriched in RD TNBC.**

Gene Symbol	Description
ZFPM2	zinc finger protein, multitype 2
COX7A1	cytochrome c oxidase subunit VIIa polypeptide 1 (muscle)
POSTN	periostin, osteoblast specific factor
COL3A1	collagen, type III, alpha 1
TAGLN	transgelin
LOXL1	lysyl oxidase-like 1
COL10A1	collagen, type X, alpha 1
LRRC15	leucine rich repeat containing 15
FAP	fibroblast activation protein, alpha
CFH	complement factor H
PCOLCE	procollagen C-endopeptidase enhancer
DPT	dermatopontin
COL1A1 c	ollagen, type I, alpha 1
COMP	cartilage oligomeric matrix protein
AEBP1	AE binding protein 1
WISP1	WNT1 inducible signaling pathway protein 1
RCN3	reticulocalbin 3, EF-hand calcium binding domain
FBN1	fibrillin 1
MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
THY1	Thy-1 cell surface antigen
COL16A1	collagen, type XVI, alpha 1
VCAN	versican
SFRP4	secreted frizzled-related protein 4
COL6A3	collagen, type VI, alpha 3
NOX4	NADPH oxidase 4
COL5A2	collagen, type V, alpha 2
BGN	biglycan
C20orf103	chromosome 20 open reading frame 103
SULF1	sulfatase 1
ASPN	asporin
LUM	lumican
NID1	nidogen 1
SNAI2	snail homolog 2 (Drosophila)
GNG11	guanine nucleotide binding protein (G protein), gamma 11
INHBA	inhibin, beta A
ODZ3	odz, odd Oz/ten-m homolog 3 (Drosophila)
COL11A1	collagen, type XI, alpha 1
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1
CNN1	calponin 1, basic, smooth muscle
COL5A1	collagen, type V, alpha 1
DCN	decorin
COL1A2	collagen, type I, alpha 2

A representative list of 42 core genes (GSEA analysis) of the ZFPM2 network enriched in RD TNBC is reported. The HUGO Gene Symbol is reported, followed by the full gene name (i.e. Description). Only 42 genes are reported for simplicity.

Table B.4: **A representative list of 42 core genes (GSEA analysis) of the PRRX1 network enriched in RD TNBC.**

Gene Symbol	Description
PRRX1	paired related homeobox 1
POSTN	periostin, osteoblast specific factor
COL3A1	collagen, type III, alpha 1
MMP11	matrix metallopeptidase 11 (stromelysin 3)
SPARC	secreted protein, acidic, cysteine-rich (osteonectin)
TAGLN	transgelin
LOXL1	lysyl oxidase-like 1
COL10A1	collagen, type X, alpha 1
LRRC15	leucine rich repeat containing 15
FAP	fibroblast activation protein, alpha
PCOLCE	procollagen C-endopeptidase enhancer
MYL9	myosin, light chain 9, regulatory
COL1A1	collagen, type I, alpha 1
GREM1	gremlin 1, cysteine knot superfamily, homolog (Xenopus laevis)
AEBP1	AE binding protein 1
WISP1	WNT1 inducible signaling pathway protein 1
UNC5B	unc-5 homolog B (C. elegans)
OLFML2B	olfactomedin-like 2B
FBN1	fibrillin 1
MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
FN1	fibronectin 1
LRRC32	leucine rich repeat containing 32
THY1	Thy-1 cell surface antigen
VCAN	versican
COL6A3	collagen, type VI, alpha 3
NOX4	NADPH oxidase 4
COL5A2	collagen, type V, alpha 2
BGN	biglycan
SULF1	sulfatase 1
ASPN	asporin
LUM	lumican
NID1	nidogen 1
HTRA1	HtrA serine peptidase 1
TIMP3	TIMP metallopeptidase inhibitor 3
INHBA	inhibin, beta A
MMP13	matrix metallopeptidase 13 (collagenase 3)
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1
WNT2	wingless-type MMTV integration site family member 2
COL11A1	collagen, type XI, alpha 1
PTRF	polymerase I and transcript release factor
COL5A1	collagen, type V, alpha 1
COL1A2	collagen, type I, alpha 2

A representative list of 42 core genes (GSEA analysis) of the PRRX1 network enriched in RD TNBC is reported. The HUGO Gene Symbol is reported, followed by the full gene name (i.e. Description). Only 42 genes are reported for simplicity.

Table B.5: **A representative list of 42 core genes (GSEA analysis) of the ELF4 network enriched in RD TNBC.**

Gene Symbol	Description
ELF4	E74-like factor 4
ARHGAP25	Rho GTPase activating protein 25
PLEKHO2	pleckstrin homology domain containing, family O member 2
CRISPLD2	cysteine-rich secretory protein LCCL domain containing 2
LOXL1	lysyl oxidase-like 1
SERPING1	serpin peptidase inhibitor, clade G (C1 inhibitor), member 1
IL15RA	interleukin 15 receptor, alpha
C1S	complement component 1, s subcomponent
CCR1	chemokine (C-C motif) receptor 1
PCOLCE	procollagen C-endopeptidase enhancer
ADCY7	adenylate cyclase 7
SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
GFPT2	glutamine-fructose-6-phosphate transaminase 2
GREM1	gremlin 1, cysteine knot superfamily, homolog (Xenopus laevis)
TCIRG1	T-cell, immune regulator 1, ATPase, H+ transporting, lysosomal V0 subunit A3
ITGA5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
ITGAX	integrin, alpha X (complement component 3 receptor 4 subunit)
COL6A2	collagen, type VI, alpha 2
OLFML2B	olfactomedin-like 2B
THY1	Thy-1 cell surface antigen
HSPG2	heparan sulfate proteoglycan 2
SLC7A4	solute carrier family 7 (cationic amino acid transporter, y+ system), member 4
CHST11	carbohydrate (chondroitin 4) sulfotransferase 11
COL6A3	collagen, type VI, alpha 3
CD97	CD97 molecule
CD74	CD74 molecule, major histocompatibility complex, class II invariant chain
COL5A2	collagen, type V, alpha 2
BGN	biglycan
CTSZ	cathepsin Z
NNMT	nicotinamide N-methyltransferase
ARPC1B	actin related protein 2/3 complex, subunit 1B, 41kDa; similar to Actin-related protein 2/3 complex subunit 1B (ARP2/3 complex 41 kDa subunit) (p41-ARC)
SERPINF1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1
IL4R	interleukin 4 receptor
LEPRE1	leucine proline-enriched proteoglycan (leprecan) 1
C3AR1	complement component 3a receptor 1
ADAP2	ArfGAP with dual PH domains 2
INHBA	inhibin, beta A
XBP1	X-box binding protein 1
ANGPTL2	angiopoietin-like 2
BIN2	bridging integrator 2
COL5A1	collagen, type V, alpha 1
TGFBI	transforming growth factor, beta-induced, 68kDa
EGFL6	EGF-like-domain, multiple 6

A representative list of 42 core genes (GSEA analysis) of the ELF4 network enriched in RD TNBC is reported. The HUGO Gene Symbol is reported, followed by the full gene name (i.e. Description). Only 42 genes are reported for simplicity.

Table B.6: **A representative list of 42 core genes (GSEA analysis) of the COL1A1 network enriched in RD TNBC.**

Gene Symbol	Description
COL1A1	collagen, type I, alpha 1
COL3A1	collagen, type III, alpha 1
CRISPLD2	cysteine-rich secretory protein LCCL domain containing 2
MMP11	matrix metallopeptidase 11 (stromelysin 3)
FSTL1	follistatin-like 1
SPARC	secreted protein, acidic, cysteine-rich (osteonectin)
LOXL1	lysyl oxidase-like 1
DKK3	dickkopf homolog 3 (Xenopus laevis)
COL10A1	collagen, type X, alpha 1
LRRC15	leucine rich repeat containing 15
FAP	fibroblast activation protein, alpha
MXRA8	matrix-remodelling associated 8
PCOLCE	procollagen C-endopeptidase enhancer
PDGFRB	platelet-derived growth factor receptor, beta polypeptide
ADAM12	ADAM metallopeptidase domain 12
MXRA5	matrix-remodelling associated 5
MYL9	myosin, light chain 9, regulatory
HEG1	HEG homolog 1 (zebrafish)
GREM1	gremlin 1, cysteine knot superfamily, homolog (Xenopus laevis)
AEBP1	AE binding protein 1
COL6A2	collagen, type VI, alpha 2
UNC5B	unc-5 homolog B (C. elegans)
WISP1	WNT1 inducible signaling pathway protein 1
OLFML2B	olfactomedin-like 2B
MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
LRRC32	leucine rich repeat containing 32
THY1	Thy-1 cell surface antigen
COL16A1	collagen, type XVI, alpha 1
VCAN	versican
COL6A3	collagen, type VI, alpha 3
COL5A2	collagen, type V, alpha 2
BGN	biglycan
ASPN	asporin
NID1	nidogen 1
HTRA1	HtrA serine peptidase 1
INHBA	inhibin, beta A
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1
CTSK	cathepsin K
NBL1	neuroblastoma, suppression of tumorigenicity 1
PTRF	polymerase I and transcript release factor
COL5A1	collagen, type V, alpha 1

A representative list of 42 core genes (GSEA analysis) of the COL1A1 network enriched in RD TNBC is reported. The HUGO Gene Symbol is reported, followed by the full gene name (i.e. Description). Only 42 genes are reported for simplicity.

Bibliography

[1] Hanahan D, Weinberg RA: *The hallmarks of cancer.* Cell 100: 57-70, (2000).

[2] Hanahan D, Weinberg RA: *Hallmarks of cancer: the next generation.* Cell 144(5): 646-74, (2011).

[3] Edwards BK, Noone AM, Mariotto AB: *Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer.* Cancer 120(9): 1290-314, (2014).

[4] Kasparek TR, Humphrey TC: *DNA double-strand break repair pathways, chromosomal rearrangements and cancer.* Semin Cell Dev Biol 22: 886-97, (2011).

[5] Yu DH, Waterland RA, Zhang P: *Targeted p16Ink4a epimutation causes tumorigenesis and reduces survival in mice.* Semin J Clin Invest 124(9): 3708-12, (2014).

[6] Gazzoli I, Loda M, Garber J, et al: *A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the MLH1 gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor.* Cancer Res 62(14): 3925-8, (2002).

[7] Robert MF, Morin S, Beaulieu N: *DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells.* Nat Genet 33(1): 61-5, (2003).

[8] Dalgliesh GL, Furge K, Greenman C, et al: *Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes.* Nature 463(7279): 360-3, (2010).

[9] Thomas RK, Baker AC, Debiasi RM: *High-throughput oncogene mutation profiling in human cancer.* Nat Genet 39(3): 347-51, (2007).

[10] Montagna C, Andrechek ER, Padilla-Nash H, et al: *Centrosome abnormalities, recurring deletions of chromosome 4, and genomic amplification of*

HER2/neu define mouse mammary gland adenocarcinomas induced by mutant HER2/neu. Oncogene 21(6): 890-8, (2002).

[11] Holstege H, Joosse SA, van Oostrom CT, et al: *High incidence of protein-truncating TP53 mutations in BRCA1-related breast cancer.* Cancer Res 69(8): 3625-33, (2009).

[12] Ji Z, Mei FC, Xie J, et al: *Oncogenic KRAS activates hedgehog signaling pathway in pancreatic cancer cells.* J Biol Chem 282(19): 14048-55, (2007).

[13] Muraoka RS, Koh Y, Roebuck LR, et al: *Increased malignancy of Neu-induced mammary tumors overexpressing active transforming growth factor beta1.* Mol Cell Biol 23(23): 8691-703, (2003).

[14] Holland EC, Hively WP, DePinho RA, et al: *A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice.* Genes Dev 12(23): 3675-85, (1998).

[15] Bruce WR, Van Der Gaag H: *A quantitative assay for the number of murine lymphoma cells capable of proliferation in vivo.* Nature 199: 79-80, (1963).

[16] Dick JE: *Breast cancer stem cells revealed.* Proc Natl Acad Sci U S A 100(7): 3547-9, (2003).

[17] Salmon SE, Hamburger AW, Soehnlen B, et al: *Quantitation of differential sensitivity of human-tumor stem cells to anticancer drugs.* N Engl J Med 298(24): 1321-7, (1978).

[18] Bhatia M, Wang JC, Kapp U, et al: *Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice.* Proc Natl Acad Sci U S A 94(10): 5320-5, (1997).

[19] Clarkson B, Ohkita T, Ota K, et al: *Studies of cellular proliferation in human leukemia. I. Estimation of growth rates of leukemic and normal hematopoietic cells in two adults with acute leukemia given single injections of tritiated thymidine.* J Clin Invest 46(4): 506-29, (1967).

[20] Vanharanta S and Massagué J: *Origins of metastatic traits.* Cancer Cell 24(4): 410-21, (2013).

[21] Wirtz D, Konstantopoulos K, Searson PC: *The physics of cancer: the role of physical interactions and mechanical forces in metastasis.* Nat Rev Cancer 11(7): 512-22, (2011).

[22] Batlle E, Sancho E, Franc C, et al: *The transcription factor Snail is a repressor of E-cadherin gene expression in epithelial tumour cells.* Nat Cell Biol 2(2): 84-9, (2000).

[23] Cano A, Pérez-Moreno MA, Rodrigo I, et al: *The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression.* Nat Cell Biol 2(2): 76-83, (2000).

[24] Comijn J, Berx G, Vermassen P, et al: *The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion.* Semin Mol Cell 7(6): 1267-78, (2001).

[25] Eger A, Aigner K, Sonderegger S, et al: *DeltaEF1 is a transcriptional repressor of E-cadherin and regulates epithelial plasticity in breast cancer cells.* Semin Oncogene 24(14): 2375-85, (2005).

[26] Yang J, Mani SA, Donaher JL, et al: *Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis.* Semin Cell 117(7): 927-39, (2004).

[27] Siegel R, Ma J, Zou Z, et al: *Cancer statistics, 2014.* CA Cancer J Clin 64(1): 9-29, (2014).

[28] Porter P: *"Westernizing" women's risks? Breast cancer in lower-income countries.* N Engl J Med 358(3): 213-6, (2008).

[29] Sloan FA, Gelband H: *Cancer Control Opportunities in Low- and Middle-Income Countries.* The National Academies Collection: Reports funded by National Institutes of Health (2007).

[30] Shulman LN, Willett W, Sievers A, et al: *Breast cancer in developing countries: opportunities for improved survival.* J Oncol 2010: 595167, (2010).

[31] Sharma GN, Dave R, Sanadya J, et al: *Various types and management of breast cancer: an overview.* J Adv Pharm Technol Res 1(2): 109-26, (2010).

[32] Weigelt B and Reis-Filho JS: *Histological and molecular types of breast cancer: is there a unifying taxonomy?* Nat Rev Clin Oncol 6(12): 718-30, (2009).

[33] Stephens PJ, Tarpey PS, Davies H, et al: *The landscape of cancer genes and mutational processes in breast cancer.* Nature 486(7403): 400-4, (2012).

[34] Lawrence MS, Stojanov P, Polak P, et al: *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature 499(7457): 214-8, (2013).

[35] Ellis P, Schnitt SJ, Sastre-Garau X, et al: *Invasive breast carcinoma.* F.A. Tavassoli, P. Devilee (Eds.), WHO Classification of Tumours Pathology and Genetics of Tumours of the Breast and Female Genital Organs, Lyon Press, Lyon 9-110, (2003).

[36] Curtis C, Shah SP, Chin SF, et al: *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature 486(7403): 346-52, (2012).

[37] Perou CM, Sørlie T, Eisen MB, et al: *Molecular portraits of human breast tumours.* Nature 406(6797): 747-52, (2000).

[38] Sørlie T, Perou CM, Tibshirani R, et al: *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A 98(19): 10869-74, (2001).

[39] Desmedt C, Haibe-Kains B, Wirapati P, et al: *Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes.* Clin Cancer Res 14(16): 5158-65, (2008).

[40] Ginestier C, Cervera N, Finetti P, et al: *Prognosis and gene expression profiling of 20q13-amplified breast cancers.* Clin Cancer Res 12(15): 4533-44, (2006).

[41] Geiger T, Madden SF, Gallagher WM, et al: *Proteomic portrait of human breast cancer progression identifies novel prognostic markers.* Cancer Res 72(9): 2428-39, (2012).

[42] Ding L, Ellis MJ, Li S, et al: *Genome remodelling in a basal-like breast cancer metastasis and xenograft.* Nature 464(7291): 999-1005, (2010).

[43] Navin N, Kendall J, Troge J, et al: *Tumour evolution inferred by single-cell sequencing.* Nature 472(7341): 90-4, (2011).

[44] Sorlie T, Tibshirani R, Parker J, et al: *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A 100(14): 8418-23, (2003).

[45] van 't Veer LJ, Dai H, van de Vijver MJ, et al: *Gene expression profiling predicts clinical outcome of breast cancer.* Nature 415(6871): 530-6, (2002).

[46] Pawitan Y, Bjöhle J, Amler L, et al: *Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.* Breast Cancer Res 7(6): R953-64, (2005).

[47] Finak G, Bertos N, Pepin F, et al: *Stromal gene expression predicts clinical outcome in breast cancer.* Nat Med 14(5): 518-27, (2008).

[48] Paik S, Shak S, Tang G, et al: *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* N Engl J Med 351(27): 2817-26, (2004).

[49] Voduc KD, Cheang MC, Tyldesley S, et al: *Breast cancer subtypes and the risk of local and regional relapse.* J Clin Oncol 28(10): 1684-91, (2010).

[50] Metzger-Filho O, Sun Z, Viale G, et al: *Patterns of Recurrence and outcome according to breast cancer subtypes in lymph node-negative disease: results from international breast cancer study group trials VIII and IX.* J Clin Oncol 31(25): 3083-90, (2013).

[51] Arvold ND, Taghian AG, Niemierko A, et al: *Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy.* J Clin Oncol 29(29): 3885-91, (2011).

[52] Lumachi F, Brunello A, Maruzzo M, et al: *Treatment of estrogen receptor-positive breast cancer.* Curr Med Chem 20(5): 596-604, (2013).

[53] Yang XR, Sherman ME, Rimm DL, et al: *Differences in risk factors for breast cancer molecular subtypes in a population-based study.* Cancer Epidemiol Biomarkers Prev 16(3): 439-43, (2007).

[54] Millikan RC, Newman B, Tse CK, et al: *Epidemiology of basal-like breast cancer.* Breast Cancer Res Treat 109(1): 123-39, (2008).

[55] Potemski P, Kusinska R, Watala C, et al: *Prognostic relevance of basal cytokeratin expression in operable breast cancer.* Oncology 69(6): 478-85, (2005).

[56] Millar EK, Graham PH, O' Toole SA, et al: *Prediction of local recurrence, distant metastases, and death after breast-conserving therapy in early-stage invasive breast cancer using a five-biomarker panel.* J Clin Oncol 27(28): 4701-8, (2009).

[57] Goldhirsch A, Winer EP, Coates AS, et al: *Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International*

Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 24(9): 2206-23, (2013).

[58] Earl HM, Vallier AL, Hiller L, et al: *Effects of the addition of gemcitabine, and paclitaxel-first sequencing, in neoadjuvant sequential epirubicin, cyclophosphamide, and paclitaxel for women with high-risk early breast cancer (Neo-tAnGo): an open-label,* 2×2 *factorial randomised phase 3 trial.* Lancet Oncol 15(2): 201-12, (2014).

[59] Hudis CA: *Trastuzumab–mechanism of action and use in clinical practice.* N Engl J Med 357(1): 39-51, (2007).

[60] Moja L, Tagliabue L, Balduzzi S, et al: *Trastuzumab containing regimens for early breast cancer.* Cochrane Database Syst Rev 4: CD006243, (2012).

[61] Yehiely F, Moyano JV, Evans JR, et al: *Deconstructing the molecular portrait of basal-like breast cancer.* Trends Mol Med 12(11): 537-44, (2006).

[62] Nelson HD, Tyne K, Naik A, et al: *Screening for breast cancer: an update for the U.S. Preventive Services Task Force.* Ann Intern Med 151(10): 727-37, (2009).

[63] Moss SM, Cuckle H, Evans A, et al: *Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial.* Lancet 368(9552): 2053-60, (2006).

[64] Nielsen DL, Kmler I, Palshof JA, et al: *Efficacy of HER2-targeted therapy in metastatic breast cancer. Monoclonal antibodies and tyrosine kinase inhibitors.* Breast 22(1): 1-12, (2013).

[65] Azim HA Jr, de Azambuja E, Colozza M, et al: *Long-term toxic effects of adjuvant chemotherapy in breast cancer.* Ann Oncol 22(9): 1939-47, (2011).

[66] Hassett MJ, O' Malley AJ, Pakes JR, et al: *Frequency and cost of chemotherapy-related serious adverse effects in a population sample of women with breast cancer.* J Natl Cancer Inst 98(16): 1108-17, (2006).

[67] Shapiro CL and Recht A: *Side effects of adjuvant treatment of breast cancer.* N Engl J Med 344(26): 1997-2008, (2001).

[68] Fan HG, Houédé-Tchen N, Yi QL, et al: *Fatigue, menopausal symptoms, and cognitive function in women after adjuvant chemotherapy for breast cancer: 1- and 2-year follow-up of a prospective controlled study.* J Clin Oncol 23(31): 8025-32, (2005).

[69] Martin M, Pienkowski T, Mackey J, et al: *Adjuvant docetaxel for node-positive breast cancer.* N Engl J Med 352(22): 2302-13, (2005).

[70] Clark GM, Zborowski DM, Culbertson JL, et al: *Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib.* J Thorac Oncol 1(8): 837-46, (2006).

[71] Cheang MC, Voduc D, Bajdik C, et al: *Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype.* Clin Cancer Res 14(5): 1368-76, (2008).

[72] Shao ZM, Nguyen M, Barsky SH: *Human breast carcinoma desmoplasia is PDGF initiated.* Oncogene 19(38): 4337-45, (2000).

[73] Rubin BP, Schuetze SM, Eary JF et al: *Molecular targeting of platelet-derived growth factor B by imatinib mesylate in a patient with metastatic dermatofibrosarcoma protuberans.* J Clin Oncol 20(17): 3586-91, (2002).

[74] Byar DP, Sears ME, McGuire WL: *Relationship between estrogen receptor values and clinical data in predicting the response to endocrine therapy for patients with advanced breast cancer.* Eur J Cancer 15(3): 299-310, (1979).

[75] Viale G, Regan MM, Maiorano E, et al: *Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98.* J Clin Oncol 25(25): 3846-52, (2007).

[76] Dowsett M, Allred C, Knox J, et al: *Relationship between quantitative estrogen and progesterone receptor expression and human epidermal growth factor receptor 2 (HER-2) status with recurrence in the Arimidex, Tamoxifen, Alone or in Combination trial.* J Clin Oncol 26(7): 1059-65, (2008).

[77] Slamon DJ, Clark GM, Wong SG, et al: *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.* Science 235(4785): 177-82, (1987).

[78] Mass RD, Press MF, Anderson S, et al: *Evaluation of clinical outcomes according to HER2 detection by fluorescence in situ hybridization in women with metastatic breast cancer treated with trastuzumab.* Clin Breast Cancer 6(3): 240-6, (2005).

[79] Poste G: *Bring on the biomarkers.* Nature 469(7329): 156-7, (2011).

[80] Knauer M, Mook S, Rutgers EJ, et al: *The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer.* Breast Cancer Res Treat 120(3): 655-61, (2010).

[81] Paik S, Tang G, Shak S, et al: *Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer.* J Clin Oncol 24(23): 3726-34, (2006).

[82] Reis-Filho JS, Pusztai L: *Gene expression profiling in breast cancer: classification, prognostication, and prediction.* Lancet 378(9805): 1812-23, (2011).

[83] Weigelt B, Pusztai L, Ashworth A, et al: *Challenges translating breast cancer gene signatures into the clinic.* Nat Rev Clin Oncol 9(1): 58-64, (2011).

[84] Lee JK, Coutant C, Kim YC, et al: *Prospective comparison of clinical and genomic multivariate predictors of response to neoadjuvant chemotherapy in breast cancer.* Clin Cancer Res 16(2): 711-8, (2010).

[85] Borst P and Wessels L: *Do predictive signatures really predict response to cancer chemotherapy.* Cell Cycle 9(24): 4836-40, (2010).

[86] Sotiriou C and Piccart MJ: *Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?* Nat Rev Cancer 7(7): 545-53, (2007).

[87] Kitano H: *Systems biology: a brief overview.* Science 295(5560): 1662664, (2002).

[88] Pujol A, Mosca R, Farrés J, et al: *Unveiling the role of network and systems biology in drug discovery.* Trends Pharmacol Sci 31(3): 115-23, (2010).

[89] Hollstein M, Sidransky D, Vogelstein B, et al: *p53 mutations in human cancers.* Science 253(5015): 49-53, (1991).

[90] Dalla-Favera R, Bregni M, Erikson J, et al: *Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells.* Proc Natl Acad Sci U S A 79(24): 7824-7, (1982).

[91] Lander ES, Linton LM, Birren B, et al: *Initial sequencing and analysis of the human genome.* Nature 409(6822): 860-921, (2001).

[92] Rual JF, Venkatesan K, Hao T, et al: *Towards a proteome-scale map of the human protein-protein interaction network.* Nature 437(7062): 1173-8, (2005).

[93] Barrett T, Troup DB, Wilhite SE, et al: *NCBI GEO: mining tens of millions of expression profiles–database and tools update.* Nucleic Acids Res D760-5, (2007).

[94] Ewing RM, Chu P, Elisma F, et al: *Large-scale mapping of human protein-protein interactions by mass spectrometry.* Mol Syst Biol 3: 89, (2007).

[95] Morgan G, Ward R, Barton M: *The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies.* Clin Oncol (R Coll Radiol) 16(8): 549-60, (2004).

[96] Milojkovic D and Apperley J, et al: *Mechanisms of Resistance to Imatinib and Second-Generation Tyrosine Inhibitors in Chronic Myeloid Leukemia.* Clin Cancer Res 15(24): 7519-7527, (2009).

[97] Druker BJ, Guilhot F, O' Brien SG, et al: *Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia.* N Engl J Med 355(23): 2408-17, (2006).

[98] Toni T and Stumpf MP: *Simulation-based model selection for dynamical systems in systems and population biology.* Bioinformatics 26(1): 104-10, (2010).

[99] Ghosh S, Matsuoka Y, Asai Y, et al: *Software for systems biology: from tools to integrated platforms.* Nat Rev Genet 12(12): 821-32, (2011).

[100] Kröger P and Bry F: *A Computational Biology Database Digest: Data, Data Analysis, and Data Management.* Distributed and Parallel Databases 13: 7-42, (2003).

[101] Morley M, Molony CM, Weber TM, et al: *Genetic analysis of genome-wide variation in human gene expression.* Nature 430(7001): 743-7, (2004)

[102] Zhu J, Zhang B, Smith EN, et al: *Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.* Nat Genet 40(7): 854-61, (2008).

[103] Oda K, Matsuoka Y, Funahashi A, et al: A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1:2005.0010. (2005).

[104] López-Avilés S, Kapuy O, Novák B, et al: *Irreversibility of mitotic exit is the consequence of systems-level feedback.* Nature 459(7246): 854-61, (2009).

[105] Chen Q, Kirsch GE, Zhang D, et al: *Genetic basis and molecular mechanism for idiopathic ventricular fibrillation.* Nature 392(6673): 293-6, (1998).

[106] Valente L and Nishikura K: *ADAR gene family and A-to-I RNA editing: diverse roles in posttranscriptional gene regulation.* Prog Nucleic Acid Res Mol Biol 79: 299-338, (2005).

[107] Nishikura K: *Editor meets silencer: crosstalk between RNA editing and RNA interference.* Nat Rev Mol Cell Biol 7(12): 919-31, (2006).

[108] Wu L and Belasco JG: *Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs.* Mol Cell 29(1): 1-7, (2008).

[109] Matzke MA and Matzke AJ: *Planting the seeds of a new paradigm.* PLoS Biol 2(5): E133, (2004).

[110] De Launoit Y, Baert JL, Chotteau-Lelievre A, et al: *The Ets transcription factors of the PEA3 group: transcriptional regulators in metastasis.* Biochim Biophys Acta 1766(1): 79-87, (2006).

[111] Chakrabarti R, Hwang J, Andres Blanco M, et al: *Elf5 inhibits the epithelial-mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2.* Nat Cell Biol 14(11): 1212-22, (2012).

[112] Chen X, Johns DC, Geiman DE, et al: *Krppel-like factor 4 (gut-enriched Krppel-like factor) inhibits cell proliferation by blocking G1/S progression of the cell cycle.* J Biol Chem 276(32): 30423-8, (2001).

[113] Yori JL, Johnson E, Zhou G, et al: *Kruppel-like factor 4 inhibits epithelial-to-mesenchymal transition through regulation of E-cadherin gene expression.* J Biol Chem 285(22): 16854-63, (2010).

[114] Jiang Y, Shen H, Liu X, et al: *Genetic variants at 1p11.2 and breast cancer risk: a two-stage study in Chinese women.* PLoS One 6(6): e21563, (2011).

[115] Huang FW, Hodis E, Xu MJ, et al: *Highly recurrent TERT promoter mutations in human melanoma.* Science 339(6122): 957-9, (2013).

[116] Demichelis F, Setlur SR, Banerjee S, et al: *Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk.* Proc Natl Acad Sci U S A 109(17): 6686-91, (2012).

[117] Pientong C, Wongwarissara P, Ekalaksananan T, et al: *Association of human papillomavirus type 16 long control region mutation and cervical cancer.* Virol J 10: 30, (2013).

[118] Wang D, Zhu G, Wang N, et al: *SIL-TAL1 rearrangement is related with poor outcome: a study from a Chinese institution.* PLoS One 8(9): e73865, (2013).

[119] Sanda T, Lawton LN, Barrasa MI, et al: *Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia.* Cancer Cell 22(2): 209-21, (2012).

[120] Lin CY, Loven J, Rahl PB, et al: *Transcriptional amplification in tumor cells with elevated c-Myc.* Cell 151(1): 56-67, (2012).

[121] Sage J, Mulligan GJ, Attardi LD, et al: *Targeted disruption of the three Rb-related genes leads to loss of G(1) control and immortalization.* Genes Dev 14(23): 3037-50, (2000).

[122] Lee TI and Young RA: *Transcriptional Regulation and its Misregulation in Disease.* Cell 152(6): 1237-51, (2013).

[123] Schena M, Shalon D, Davis RW, et al: *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science 270(5235): 467-70, (1995).

[124] Pollack JR, Perou CM, Alizadeh AA, et al: *Genome-wide analysis of DNA copy-number changes using cDNA microarrays.* Nat Genet 23(1): 41-6, (1999).

[125] Albertson DG, Ylstra B, Segraves R, et al: *Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene.* Nat Genet 25(2): 144-6, (2000).

[126] Lindblad-Toh K, Tanenbaum DM, Daly MJ, et al: *Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays.* Nat Biotechnol 18(9): 1001-5, (2000).

[127] Berns K, Hijmans EM, Mullenders J, et al: *A large-scale RNAi screen in human cells identifies new components of the p53 pathway.* Nature 428(6981): 431-7, (2004).

[128] Xiong Y, Chen X, Chen Z, et al: *RNA sequencing shows no dosage compensation of the active X-chromosome.* Nat Genet 42(12): 1043-7, (2010).

[129] Friedman N, Linial M, Nachman I, et al: *Using Bayesian networks to analyze expression data.* J Comput Biol 7(3-4): 601-20, (2000).

[130] Pe'er D, Regev A, Elidan G, et al: *Inferring subnetworks from perturbed expression profiles.* Bioinformatics 17 Suppl 1: S215-24, (2001).

[131] Basso K, Margolin AA, Stolovitzky G, et al: *Reverse engineering of regulatory networks in human B cells.* Nat Genet 37(4): 382-90, (2005).

[132] Ideker T, Ozier O, Schwikowski B, et al: *Discovering regulatory and signalling circuits in molecular interaction networks.* Bioinformatics 18 Suppl 1: S233-40, (2002).

[133] Margolin AA, Nemenman I, Basso K, et al: *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics 7 Suppl: 1:S7, (2006).

[134] Segal E, Shapira M, Regev A, et al: *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet 34(2): 166-76, (2003).

[135] Ihmels J, Friedlander G, Bergmann S, et al: *Revealing modular organization in the yeast transcriptional network.* Nat Genet 31(4): 370-7, (2002).

[136] Roy S, Wapinski I, Pfiffner J, et al: *Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules.* Genome Res 23(6): 1039-50, (2013).

[137] Stolovitzky G, Monroe D, Califano A: *Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway.* Ann N Y Acad Sci 1115: 1-22, (2007).

[138] Stolovitzky G, Prill RJ, Califano A: *Lessons from the DREAM2 Challenges.* Ann N Y Acad Sci 1158: 159-95, (2009).

[139] Marbach D, Schaffter T, Mattiussi C: *Generating realistic in silico gene networks for performance assessment of reverse engineering methods.* J Comput Biol. 16(2): 229-39, (2009).

[140] Marbach D, Prill RJ, Schaffter T: *Revealing strengths and weaknesses of methods for gene network inference.* Proc Natl Acad Sci U S A 107(14): 6286-91, (2009).

[141] Michoel T, De Smet R, Joshi A: *Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks.* BMC Syst Biol 3: 49, (2009).

[142] De Smet R and Marchal K: *Advantages and limitations of current network inference methods.* Nat Rev Microbiol 8(10): 717-29, (2010).

[143] Hughes TR, Marton MJ, Jones AR, et al: *Functional discovery via a compendium of expression profiles.* Cell 102(1): 109-26, (2000).

[144] Kemmeren P, Sameith K, van de Pasch LA, et al: *Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors.* Cell 157(3): 740-52, (2014).

[145] Zhao X, He L, Li T, et al: *Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs.* CPT Pharmacometrics Syst Pharmacol 3: e102, (2014).

[146] Babcock JJ, Du F, Xu K, et al: *Integrated analysis of drug-induced gene expression profiles predicts novel hERG inhibitors.* PLoS One 8(7): e69513, (2013).

[147] Pavlopoulos GA, Secrier M, Moschopoulos CN, et al: *Using graph theory to analyze biological networks.* BioData Min 4: 10, (2011).

[148] Huber W, Carey VJ, Long L, et al: *Graphs in molecular biology.* BMC Bioinformatics 8: Suppl 6:S8, (2007).

[149] Bansal M, Della Gatta G, di Bernardo D: *Inference of gene regulatory networks and compound mode of action from time course gene expression profiles.* Bioinformatics 22(7): 815-22, (2006).

[150] Pearl J: *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, Calif.

[151] Kauffman SA: *The Origins of Order: Self Organization and Selection in Evolution.* Oxford University Press (1993).

[152] Akutsu T, Miyano S, Kuhara S: *Algorithms for inferring qualitative models of biological networks.* Pac Symp Biocomput 293-304, (2000).

[153] de Jong H: *Modeling and simulation of genetic regulatory systems: a literature review.* J Comput Biol 9(1): 67-103, (2002).

[154] Kaern M, Elston TC, BlakeWJ, et al: *Stochasticity in gene expression: from theories to phenotypes.* Nat Rev Genet 6(6): 451-464, (2005).

[155] D'haeseleer P, Liang S, Somogyi R: *Genetic network inference: from co-expression clustering to reverse engineering.* Bioinformatics 16(8): 707-26, (2000).

[156] Eisen MB, Spellman PT, Brown PO, et al: *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A 95(25): 14863-8, (1998).

[157] Tavazoie S, Hughes JD, Campbell MJ: *Systematic determination of genetic network architecture.* Nat Genet 22(3): 281-5, (1999).

[158] Tamayo P, Slonim D, Mesirov J, et al: *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc Natl Acad Sci U S A 96(6): 2907-12, (1999).

[159] Raychaudhuri S, Stuart JM, Altman RB: *Principal components analysis to summarize microarray experiments: application to sporulation time series.* Pac Symp Biocomput 455-66, (2000).

[160] Dempster AP, Laird NM, Rubin DB: *Maximum likelihood from incomplete data via the EM algorithm.* J R Stat Soc B 39: 1-38, (1977).

[161] Hume DA, Summers KM, Raza S: *Functional clustering and lineage markers: insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations.* Genomics 95(6): 328-38, (2010).

[162] Su AI, Cooke MP, Ching KA: *Large-scale analysis of the human and mouse transcriptomes.* Proc Natl Acad Sci U S A 99(7): 4465-70, (2002).

[163] Tuller T, Atar S, Ruppin E: *Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases.* Genes Immun 14(2): 67-82, (2013).

[164] Navab R, Strumpf D, Bandarchi B: *Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer.* Proc Natl Acad Sci U S A 108(17): 7160-5, (2011).

[165] Herzela H and Großeb I: *Measuring correlations in symbol sequences.* Physica A 216(4): 518-42, (1995).

[166] Steuer R, Kurths J, Daub CO, et al: *The mutual information: detecting and evaluating dependencies between variables.* Bioinformatics 18 Suppl 2: S231-40, (2002).

[167] Shannon CE: *A mathematical theory of communication.* Bell Syst Tech J 27: 379-423, (1948).

[168] Butte AJ and Kohane IS: *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.* Pac. Symp Biocomput 418-429, (2000).

[169] Faith JJ, Hayete B, Thaden JT, et al: *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol 5(1): e8, (2007).

[170] Meyer PE, Kontos K, Lafitte F, et al: *Information-theoretic inference of large transcriptional regulatory networks.* EURASIP J Bioinform Syst Biol 79879, (2007).

[171] Polyak K: *Heterogeneity in breast cancer.* J Clin Invest 121(10): 3786-8, (2011).

[172] Liberzon A, Subramanian A, Pinchback R, et al: *Molecular signatures database (MSigDB) 3.0.* Bioinformatics 27(12): 1739-40, (2011).

[173] Schweighofer B, Testori J, Sturtzel C, et al: *The VEGF-induced transcriptional response comprises gene clusters at the crossroad of angiogenesis and inflammation.* Thromb Haemost 102(3): 544-54, (2009).

[174] Amit I, Citri A, Shay T, et al: *A module of negative feedback regulators defines growth factor signaling.* Nat Genet 39(4): 503-12, (2007).

[175] Carter SL, Eklund AC, Kohane IS, et al: *A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers.* Nat Genet 38(9): 1043-8, (2006).

[176] Nicassio F, Bianchi F, Capra M, et al: *A cancer-specific transcriptional signature in human neoplasia.* J Clin Invest 115(11): 3015-25, (2005).

[177] Dupont S, Morsut L, Aragona M, et al: *Role of YAP/TAZ in mechanotransduction.* Nature 474(7350): 179-83, (2011).

[178] Sethi N, Dai X, Winter CG, et al: *Tumor-derived JAGGED1 promotes osteolytic bone metastasis of breast cancer by engaging notch signaling in bone cells.* Cancer Cell 19(2): 192-205, (2011).

[179] Bolstad BM: *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.* PhD thesis, University of California, Berkeley Spring (2004).

[180] Gentleman RC, Carey VJ, Bates DM, et al: *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol 5(10): R80, (2004).

[181] Irizarry RA, Hobbs B, Collin F, et al: *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics 4(2): 249-64, (2003).

[182] Liu G, Loraine AE, Shigeta R, et al: *NetAffx: Affymetrix probesets and annotations.* Nucleic Acids Res 31(1): 82-6, (2003).

[183] Ivshina AV, George J, Senko O, et al: *Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer.* Cancer Res 66(21): 10292-301, (2006).

[184] Buyse M, Loi S, van't Veer L, et al: *Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer.* J Natl Cancer Inst 98(17): 1183-92, (2006).

[185] Desmedt C, Piette F, Loi S, et al: *Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series.* Clin Cancer Res 13(11): 3207-14, (2007).

[186] Wang Y, Klijn JG, Zhang Y, et al: *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet 365(9460): 671-9, (2005).

[187] Bonnefoi H, Potti A, Delorenzi M, et al: *Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial.* Lancet Oncol 8(12): 1071-8, (2007).

[188] Minn AJ, Gupta GP, Siegel PM, et al: *Genes that mediate breast cancer metastasis to lung.* Nature 436(7050): 518-24, (2005).

[189] Sotiriou C, Wirapati P, Loi S, et al: *Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.* J Natl Cancer Inst 98(4): 262-72, (2006).

[190] Hatzis C, Pusztai L, Valero V, et al: *A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer.* JAMA 305(18): 1873-81, (2011).

[191] Kao KJ, Chang KM, Hsu HC, et al: *Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization.* BMC Cancer 11: 143, (2011).

[192] Shi H, Schmidt B, Liu W, et al: *Parallel mutual information estimation for inferring gene regulatory networks on GPUs.* BMC Res Notes 4: 189, (2011).

[193] Langfelder P and Horvath S: *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics 9: 559, (2008).

[194] Loi S, Haibe-Kains B, Desmedt C, et al: *Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade.* J Clin Oncol 25(10): 1239-46, (2007).

[195] Lim WK, Wang K, Lefebvre C, et al: *Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks.* Bioinformatics 23(13): i282-8, (2007).

[196] Li C and Wong WH: *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A 98(1): 31-6, (2001).

[197] Subramanian A, Tamayo P, Mootha VK, et al: *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A 102(43): 15545-50, (2005).

[198] Efron B and Tibshirani R: *On testing the significance of sets of genes.* Annals of Applied Statistics 1(1): 107-29, (2007).

[199] Daub CO, Steuer R, Selbig J, et al: *Estimating mutual information using B-spline functions-an improved similarity measure for analysing gene expression data.* BMC Bioinformatics 5: 118, (2004).

[200] Miller CA, Settle SH, Sulman EP, et al: *Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors.* BMC Med Genomics 14: 4:34, (2011).

[201] Yeang CH, McCormick F, Levine A: *Combinatorial patterns of somatic gene mutations in cancer.* FASEB J 22(8): 2605-22, (2008).

[202] Masica DL and Karchin R: *Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival.* Cancer Res 71(13): 4550-61, (2011).

[203] Jeong H, Mason S P, Barabási AL, et al: *Lethality and centrality in protein networks.* Nature 411: 41-42, (2001).

[204] Yook SH, Oltvai ZN, Barabâsi AL: *Functional and topological characterization of protein interaction networks.* Proteomics 4(4): 928-42, (2004).

[205] Salter MW and Kalia LV: *Src kinases: a hub for NMDA receptor regulation.* Nat Rev Neurosci 5(4): 317-28, (2004).

[206] Bosch A, Eroles P, Zaragoza R, et al: *Triple-negative breast cancer: molecular features, pathogenesis, treatment and current lines of research.* Cancer Treat Rev 36(3): 206-15, (2010).

[207] Tsujimoto Y, Gorham J, Cossman J, et al: *The t(14;18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining.* Science 229(4720): 1390-3, (1985).

[208] Dalla-Favera R, Bregni M, Erikson J, et al: *Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells.* Proc Natl Acad Sci U S A 79(24): 7824-7, (1982).

[209] Capon DJ, Seeburg PH, McGrath JP, et al: *Activation of Ki-ras2 gene in human colon and lung carcinomas by two different point mutations.* Nature 304(5926): 507?513, (1983).

[210] McCoy MS, Toole JJ, Cunningham JM, et al: *Characterization of a human colon/lung carcinoma oncogene.* Nature 302(5903): 79-81, (1983).

[211] Bianchi F, Nicassio F, Di Fiore PP: *Unbiased vs. biased approaches to the identification of cancer signatures: the case of lung cancer.* Cell Cycle 7(6): 729-34, (2008).

[212] Huynh-Thu VA, Irrthum A, Wehenkel L, et al: *Inferring regulatory networks from expression data using tree-based methods.* PLoS One 5: e12776, (2010).

[213] Belcastro V, Gregoretti F, Siciliano V, et al: *Reverse engineering and analysis of genome-wide gene regulatory networks from gene expression profiles using high-performance computing.* IEEE/ACM Trans Comput Biol Bioinform 9(3): 668-78, (2012).

[214] Reverter A and Chan EK: *Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks.* Bioinformatics 24(21): 2491-7, (2008).

[215] Yang H, Zhong Y, Peng C, et al: *Important role of indels in somatic mutations of human cancer genes.* BMC Med Genet 11: 128, (2010).

[216] Watson IR, Takahashi K, Futreal PA, et al: *Emerging patterns of somatic mutations in cancer.* Nat Rev Genet 14(10): 703-18, (2013).

[217] Futreal PA, Coin L, Marshall M, et al: *A census of human cancer genes.* Nature Reviews. Cancer 4: 177-183, (2004).

[218] Wood LD, Parsons DW, Jones S, et al: *The genomic landscapes of human breast and colorectal cancers.* Science 318(5853): 1108-13, (2007).

[219] Hahn WC and Weinberg RA: *Modelling the molecular circuitry of cancer.* Nat Rev Cancer 2(5): 331-41, (2002).

[220] Vogelstein B and Kinzler KW: *Cancer genes and the pathways they control.* Nat Rev Cancer 10(8): 789-99, (2004).

[221] Ciriello G, Cerami E, Sander C, et al: *Mutual exclusivity analysis identifies oncogenic network modules.* Genome Res 22(2): 398-406, (2012).

[222] Vandin F, Upfal E, Raphael BJ: *De novo discovery of mutated driver pathways in cancer.* Genome Res 22(2): 375-85, (2012).

[223] Zhao J, Zhang S, Wu LY, et al: *Efficient methods for identifying mutated driver pathways in cancer.* Bioinformatics 28(22): 2940-7, (2012).

[224] Lara-Medina F, Pérez-Sánchez V, Saavedra-Pérez D, et al: *Triple-negative breast cancer in Hispanic patients: high prevalence, poor prognosis, and association with menopausal status, body mass index, and parity.* Cancer 117(16): 3658-69, (2011).

[225] Russo AL, Arvold ND, Niemierko A, et al: *Margin status and the risk of local recurrence in patients with early stage breast cancer treated with breast-conserving therapy.* Breast Cancer Res Treat 140: 353-61, (2013).

[226] Schneider BP, Winer EP, Foulkes WD, et al: *Triple negative breast cancer: risk factors to potential targets.* Clin Cancer Res 14: 8010-8, (2008).

[227] Liedtke C, Mazouni C, Hess KR, et al: *Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer.* J Clin Oncol 26: 1275-81, (2008).

[228] Bae YH, Ryu JH, Park HJ, et al: *Mutant p53-Notch1 Signaling Axis Is Involved in Curcumin-Induced Apoptosis of Breast Cancer Cells.* Korean J Physiol Pharmacol 17: 291-7, (2013).

[229] Xu Y, Diao L, Chen Y, et al: *Promoter methylation of BRCA1 in triple-negative breast cancer predicts sensitivity to adjuvant chemotherapy.* Ann Oncol 24(6): 1498-505, (2013).

[230] Kuerer HM, Newman LA, Smith TL, et al: *Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy.* J Clin Oncol 17: 460-9, (1999).

[231] Jechlinger M, Sommer A, Moriggl R, et al: *Autocrine PDGFR signaling promotes mammary cancer metastasis.* J Clin Invest 116(6): 1561-70, (2006).

[232] Kim MS, Pinto SM, Getnet D, et al: *A draft map of the human proteome.* Nature 509(7502): 575-81, 2014.