UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE

INDIRIZZO di GENETICA E BIOLOGIA MOLECOLARE DELLO SVILUPPO

CICLO XXV

THE INTEGRATION OF GENE AND miRNA EXPRESSION USING PATHWAY TOPOLOGY: A CASE STUDY ON EPITHELIAL OVARIAN CANCER

Direttore della Scuola : Ch.mo Prof. Giuseppe Zanotti

Coordinatore d'indirizzo : Ch.mo Prof. Paolo Bonaldo

Supervisore : Dott.ssa Chiara Romualdi

Dottorando : Enrica Calura

Declaration of Authorship

I, CALURA ENRICA, declare that this thesis titled, "The integration of gene and miRNA expression using pathway topology: a case study on epithelial ovarian cancer" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Enrica Calura

Date: 31th January 2013

"Misura ciò che è misurabile,

e rendi misurabile ciò che non lo è."

Gallileo Gallilei [1564-1642]

UNIVERSITA' DEGLI STUDI DI PADOVA

Abstract

Department of Biology

Ph.D. School in Biosciences and Biotechnology

Curriculum of Genetics and Molecular Biology of Development

The integration of gene and miRNA expression using pathway topology: a case study on epithelial ovarian cancer.

by Enrica Calura

Pathways are formal descriptions of the biological processes involving finely regulated structures by which a cell converts molecules or processes signals. The study of gene expression in terms of pathways is defined as pathway analysis and aims at identifying groups of functionally related genes that show coordinated expression changes. Recently, pathway analysis moved from algorithms using merely gene list to ones exploiting the topology that define gene connections. A crucial, and unfortunately limiting step for these novel methods are the availability of the pathways as gene networks in which nodes are genes and edges are the relations between two elements.

To this aim, we develop a pathway data interpreter, called *graphite*, able to uniformly store, process and convert pathway information into gene networks. *graphite* has been made publicly available as R package within the Bioconductor platform. In the field of the topological pathway analysis, *graphite* fills the existing gap lying between technical and methodological aspects. *graphite* i) allows performing more informative analysis on *omics* data and ii) allows developing new methods based on the increased accessibility of biological knowledge. However, the pathways of the four main public resources integrated into *graphite* (KEGG, Reactome, Biocarta and PID), still lack of crucial interactors: the microRNAs.

The microRNAs are small non-coding RNAs that post-transcriptionally regulate gene expression, their function on the messenger target is repressive but their effect on the transcription is dependent of the topology of the pathway in which the miRNA is involved. In the last decade, many targets have been discovered and experimentally validated, dedicated databases are available providing these information. Thus, I worked on an extension of *graphite* package able to integrate microRNAs in pathway topology, i) linking the non-coding RNAs to their validated target genes, ii) providing integrated networks suitable for the topological pathway analyses. The feasibility of this approach

has been validated on a specific biological context, the early stage of Epithelial Ovarian Cancer (EOC).

EOC has long been considered as a single disease. The emerging opinion, however, sees ovarian cancer as a general term that encloses a group of histo-pathological subtypes sharing a common anatomic location. In collaboration with the Mario Negri institute, 257 stage I EOC tumour biopsies were collected and stratified into training and validation sets. miRNA microarray data was used to generate the most highly reproducible signatures for each histotype through a dedicated resampling inferential strategy. qRT-PCR was used to validate the results in both the training and validation set. The results indicate that the clear cell histotype is characterized by high expression levels of miR-30a and miR-30a*, while mucinous patients by high levels of miR-192 and miR-194, interestingly as well as mucinous non-ovarian tissues. Then, the integrative approach that combines mRNA and miRNA profiles using *graphite* has been applied to identify the mucinous specific regulatory circuits. Taken together our findings demonstrate that EOC histotypes have discriminant regulatory circuits that drive the differentiation of the tumour environment. Our approach successfully guides us towards important biological results with interesting therapeutic implications in EOC.

UNIVERSITA' DEGLI STUDI DI PADOVA

Abstract

Dipartimento di Biologia

Scuola di dottorato in Bioscienze e Biotecnologie

Curriculum di Genetica e Biologia Molecolare dello Sviluppo

Integrazione dell'espressione di geni e microRNA usando la topologia dei pathway: un caso studio sul tumore epiteliale ovarico.

di Enrica Calura

I pathway sono descrizioni formali ed accurate dei processi biologici, che sono serie di eventi, finemente regolati, attraverso i quali la cellula opera trasformazioni su molecole o altri processi. Lo studio dell'espressione genica attraverso i pathway è definita come "analisi di pathway" e cerca di identificare gruppi di geni correlati da un punto di vista funzionale con cambiamenti coordinati nell'espressione. Recentemente, la sua più promettente è stato il passaggio da algoritmi che considerano i pathway semplici elenchi di geni, ad algoritmi che sfruttano la topologia del pathway stesso, ovvero le reti delle connessioni geniche. Uno dei limiti di questi nuovi metodi di analisi è la mancanza dei pathway sotto forma di reti, in cui i nodi sono geni e gli archi sono le relazioni che intercorrono tra di essi.

Durante il mio dottorato abbiamo sviluppato un interprete per i pathway, denominato *graphite*, in grado di memorizzare in modo uniforme, elaborare e convertire le informazioni contenute in essi. *graphite* è stato reso pubblicamente disponibile come pacchetto R all'interno della piattaforma Bioconductor. Nel campo d'indagine delle analisi topologiche, *graphite* colma il divario esistente tra gli aspetti tecnici e metodologici. *graphite* i) permette di eseguire analisi più avanzate dei dati omici e ii) consente lo sviluppo di nuovi metodi offrendo una maggiore e più facile accessibilità ai dati. Tuttavia, i pathway provenienti dalle quattro principali risorse pubbliche che sono state integrate in *graphite* (KEGG, Reactome, BioCarta e PID), mancano ancora di interattori cruciali: i microRNA.

I microRNA sono piccoli RNA non codificanti che regolano l'espressione genica post-trascrizionale, la loro funzione sul target messaggero è repressiva ma il loro effetto generale sull'espressione va interpretato a seconda delle interazioni del contesto biologico in cui sono descritti, cioè della topologia del pathway in cui coinvolto il miRNA. Negli ultimi dieci anni, molti geni target di miRNA sono stati scoperti e validati sperimentalmente, informazioni a riguardo sono fornite in diversi database pubblici. Durante il mio dottorato ho lavorato su un'estensione del pacchetto *graphite* per integrare i microRNA all'interno dei pathway, i) collegando i miRNA ai loro geni target validati, ii) fornendo reti bipartite adatte per le analisi di pathway topologiche. Questo approccio è stato applicato allo studio di dati di espressione ottenuti da pazienti allo stadio iniziale di tumore ovarico epiteliale.

Il tumore ovarico epiteliale è stato a lungo considerato come una singola malattia. Recentemente, tuttavia, si ritiene che ”tumore ovarico epiteliale” sia solo un termine generale che racchiude un gruppo di sottotipi isto-patologici che condividono solo una comune localizzazione anatomica. In collaborazione con l'Istituto ”Mario Negri” di Milano, sono state raccolte le biopsie di 257 tumori ovarici epiteliali di stadio I e sono state stratificate in un *training set* e in un *validation set*. I dati di espressione di microRNA del *training set*sono stati analizzati con una specifica strategia di ricampionamento inferenziale con l'intento di trovare le firme molecolari specifiche di ogni istotipo. La qRT-PCR è stata utilizzata per validare i risultati sia nel *training set* che nel *validation set*. I nostri risultati mostrano come l'istotipo a cellule chiare sia caratterizzato da elevati livelli di espressione di miR-30a e miR-30a*, mentre l'istotipo mucinoso da elevati livelli di miR-192 e miR-194. Questi ultimi due microRNA sono marcatori riconosciuti e validati di tumore al colon, a sostegno dell'ipotesi che l'istotipo mucinoso sia una patologia molto diversa da gli altri istotipi ovarici piuttosto che di altri tumori mucinosi. E' stato applicato, poi, un approccio integrativo che utilizzando i profili di espressione sia dei miRNA sia dei geni, sugli stessi campioni, ha permesso di identificare un circuito regolativo che si ritiene essere specifico dei pazienti mucinosi. Nel loro insieme, i nostri risultati dimostrano che i diversi istotipi di tumore ovarico epiteliale hanno circuiti regolatori distinti, inoltre ci permettono di affermare che attraverso l'uso di *graphite* abbiamo saputo affrontare in modo nuovo l'analisi integrata dei dati biologici, rivelando anche interessanti implicazioni terapeutiche per il tumore ovarico epiteliale.

Contents

List of Figures

List of Tables

Alla mia famiglia,
ineguagliabile dispensatrice di supporto,
protezione, comprensione,
coraggio e pazienza…

…tanta pazienza.

Chapter 1

Introduction

The research project of my PhD has two main cores: the development of new bioinformatic tools for the analysis and the integration of gene and miRNAs expression data, and the application of these tools to tackle a given biological question. Specifically, I contributed to the development of a new computational framework for the storage, interpretation and analysis of pathways. Than, I applied this new framework to characterize transcriptional and post-transcriptional alterations in early stage Epithelial Ovarian Cancer. This introduction is divided in three main sections, Epithelial Ovarian Cancer, microRNAs and pathways, that represent the main dealt areas.

1.1 Epithelial Ovarian Cancer: The Silent Killer

Despite the increasing molecular knowledge of tumor biology which underpins the development of new therapeutic and clinical management strategies, Epithelial Ovarian Cancer (EOC) is the most common cause of death in gynecological diseases, with a five-years survival rate virtually unchanged in the past 30 years (Kurman and Shih, 2010; Vaughan et al., 2011). The EOC etiopathogenesis is totally unknown. More than the 90% of ovarian malignancies are called surface epithelial cancer, even if the epithelial origin of these tumor is a controversial issue (Gilks, 2010).

1.1.1 The staging

Nowadays, the only way to determine the diagnosis and the staging of the tumor is surgery. EOC are staged accordingly to International Federation of Gynecology and Obsterics (FIGO) staging system outlined in Table 1.1 (excerpt from Jelovac and Armstrong (2011)).

Figo stage	Characteristics	Stage Distribution	10 year survival rate
I	Disease confined to the ovaries.	20%	73%
IA	One ovary, capsule intact, no ascites.		
IB	Both ovaries capsule intact, no ascites.		
IC	Stage IA and IB plus ascites or washing capsule ruptures, tumor on ovarian surface.		
II	Disease spread confined to the pelvis.	5%	45%
III	Disease confined to the abdominal cavity, including surface of the liver, pelvic, including surface of the liver, pelvic, inguinal, momentum, bowel, para-aortic lymph nodes.	58%	21%
IIIA	Negative lymph nodes, plus microscopic seeding of peritoneal surface.		
IIIB	Negative lymph nodes, peritoneal implants <2 cm.		
IIIC	Positive lymph nodes and/or abdominal implants >2 cm.		
IV	Spread to liver parenchyma, lung, pleura, or other extra-abdominal sites.	17%	>5%

Table 1.1: EOC staging criteria excerpt from Jelovac review.

It should be noted that EOC stage I is the only stage confined to the ovaries; only the stage I is considered ”early”, while II,III and IV stages are considered ”advanced”.

Studies on stage I EOC are important especially to improve diagnosis. In fact, EOC is rarely diagnosed at early stages given the absence of specific symptoms. On the contrary, with the spread of the tumor into the pelvis and upper abdomen (stage III and IV) patients feel pain, pressure, early satiety and abdominal swelling. The screening of asymptomatic women is performed through three screening tests: bimanual pelvic examination, measurements of CA125 cancer antigene, and transvaginal ultrasound. None of them have sufficient efficiency to detect the tumor at early stage, that is a crucial step to enhance the survival chances.

The low survival of patients with EOC is proportional to the dissemination of the tumor that is dependent of the time of diagnosis. In patients with disease limited to the ovaries survival is close to 80%. However, in cases in which the disease involves the upper abdomen or beyond, only about 20% of patients survive at 5 years and this is due to the inability to surgically remove the total amount of tumor mass.

1.1.2 The main classes of stage I EOC

All the attempts to understand EOC were certainly complicated by the heterogeneity of this disease. The ovarian cancer can be divided in at least 15 type of tumors, each of them characterized by histo-patological features, molecular alterations, risk factors, different chemotherapy responses and resistances.

The main classification is in histologic subtypes. Low-grade serous, mucinous, endometrioid and clear cell histotypes represent the great majority of stage I ovarian cancer histotypes. They are characterized by a slow progression rate, are generally confined to the ovary, they lack in TP53 mutations and show a lineage with the corresponding benign neoplasm through an intermediate step called borderline tumor stage (Kurman and Shih, 2010). Although this notion is not universally accepted, there is a general consensus on the different molecular and clinical characteristics of the EOC histotypes (Prat, 2012a). In a retrospective study, Köbel et al. (2008) tested 21 candidate markers in a cohort of 500 advanced stages ovarian carcinomas, demonstrating that the association between biomarker expressions and survival rates varies among subtypes. The results support the hypothesis that different histological types of ovarian cancer are distinct diseases suggesting that the juxtaposition of different histotypes in a single cohort may not only confound survival analysis, but also lead to erroneous conclusions (Köbel et al., 2008).

1.1.2.1 Endometrioid ovarian cancer

Endometrioid ovarian cancer (End) is usually of stage I and II. It shows a relative low morbidity and mortality and seems to have a better survival compared to other histotypes. Few endometrioid studies are available in literature, some studies of immunoprofiling exist and demonstrated a similarity between endometrioid EOC and endometrial cancer (Jelovac and Armstrong, 2011).

1.1.2.2 Mucinous ovarian cancer

Mucinous ovarian carcinoma (Muc) develops almost always within the ovary. It is indistinguishable by mucinous non-ovarian carcinoma or metastatic carcinoma with mucinous differentiation (cervix, colon/rectum, appendix cancers) (Kelemen and Köbel, 2011). Although confined in the ovary, Muc tends to be the largest epithelial ovarian neoplasm with a diameter of 18-20 cm (Jelovac and Armstrong, 2011). Mucinous ovarian carcinoma is characterized by KRAS mutations and HER2 amplifications similar to breast cancer. Due to the low response to treatments, the development of subtype-specific treatment trial is a priority in mucinous ovarian carcinoma (Gilks, 2010).

1.1.2.3 Clear cell ovarian cancer

Early stage clear cell ovarian carcinoma (Cc) is usually considered of high grade (grade 3). Treatments are particularly inefficient independently of the low mitotic rate of this histotype. Gene expression studies detect similarities with renal clear cell carcinoma suggesting the possibility to reuse the therapy of renal cancer for the Cc. Furthermore, co-occurrences of clear-cell and endometriosis has been observed. Due to the low response to standard treatments, the clear cells subtype is a priority for the development of subtype-specific treatment trial, as well as mucinous ovarian carcinoma (Gilks, 2010).

1.1.2.4 Serous ovarian cancer

Serous ovarian cancer (Ser) accounts for the 75% of EOCs and it is divided in low-grade and high-grade (Malpica et al., 2004).

High-grade serous ovarian cancer is the most common representation of EOC. For this reason, Ser is considered the ”true ovarian cancer” and it is supposed to derive by tubal intraepithelial cells. High-grade of Ser is mass poorly differentiated cells, with the only variant of the presence of concentric rings of calcification called psammoma bodies (Jelovac and Armstrong, 2011).

Low-grade of serous ovarian carcinoma seems to represent the natural progression of non-invasive borderline tumor of type serous. Both this type are characterized by a young age at diagnosis, prolonged clinical history and similar histology. Moreover, in a large number of patients the two types co-exist. Ser is characterized by KRAS and BRAF mutations, wild type TP53 and chromosome stability (Vaughan et al., 2011). It is conceivable, instead, that low-grade and high-grade of serous carcinomas are the products of two different tumorigenic pathways, with few and rare intersections, as reflected by different chemotherapeutic sensitivities and prognoses (Vaughan et al., 2011).

1.1.2.5 Relapsing

In advanced EOCs, approximately 10%-15% of treated patients achieve and maintain complete responses to therapy, most of the patients have persistent disease or they eventually relapsed (Armstrong, 2002). On the contrary, fewer than 20% of patients with stage I EOC have aggressive disease and relapsed within 5 years of primary surgery. How to discriminate between curable patients and those who will relapse after adjuvant chemotherapy is still an unresolved clinical issue (Marchini et al., 2008).

Prediction of relapse on the basis of current clinical knowledge and pathological features is difficult (Cannistra, 2004). However, predicting relapse might be possible with a better knowledge of the molecular and genetic mechanisms that are associated with each tumor stage. The knowledge of the molecular pathways that are altered during neoplastic transformation might help to expedite the discovery of biomarkers for early disease detection, prediction of clinical response and guidance of treatment.

Given that only 10% of all patients with stage I EOC shows relapse, this makes extremely difficult the recruitment of sizable cohorts of patients.

1.1.2.6 Grading

The grade is a measure of the gravity of the cell transformation in a debulked tumor. Three levels are possible: grade 1, grade 2 and grade 3. Grade 1 is used when the cells are similar to normal cells, the cancer cells grow and multiply quite slowly and are not so aggressive; Grade 3 when the cells are abnormal and unrecognizable and the cancer is aggressive. Grade 2 is in the middle between the two situation presented above.

1.2 microRNAs

MicroRNAs (miRNAs) are fundamental regulatory elements of gene expression in animals and plants. They are 17-24 nucleotide long and regulate eukaryotic gene expression post-transcriptionally.

Primary miRNA (pri-miRNA) transcripts with stem-loop regions are usually produced by RNA polymerase II, but occasionally by RNA polymerase III. The stem-loop precursor, pre-miRNA, is released by a cleavage event, which is catalysed by the nuclear Microprocessor complex that contains the RNase III Drosha and Pasha. Pre-miRNAs are actively exported from the nucleus into the cytoplasm by the Exportin 5. A distinct RNase III, Dicer, subsequently produces a ~22 base-pair duplex RNA, that is the mature miRNA. In miRNA duplexes, usually the strand with the weakest 5'-end base pairing is selected and loaded onto RISC (RNA-induced silencing complex) which contains the Argonaute (Ago) protein. miRNAs use base-pairing to guide RISC to specific messenger RNAs (mRNAs) with fully or partially complementary sequences located especially in 3' untranslated regions (UTRs). WatsonCrick base-pairing of 2-7 miRNA nucleotides called "seed" is crucial for the targeting. The couple miRNA-mRNA enables translational inhibition or exonucleolytic mRNA decay. Unfortunately the factors that govern the prevalence of one specific mechanism remain unknown.

Each miRNA regulates numerous target genes and a lot of computational algorithms were developed to predict the resulting expression down-regulation. The development of algorithms goes hand in hand with the understanding of miRNA mode of action and function (Bartel, 2004, 2009).

The miRBase is the reference database of published miRNA sequences and annotations (http://www.mirbase.org/). At the time of writing, the current release of miRBase (version 19) contains more than 15000 microRNA gene loci in more than 140 species and at least 17000 distinct mature microRNAs, 1600 of which discovered in human (Griffiths-Jones et al., 2008; Griffiths-Jones, 2010). These numbers are increasing day by day far exceeding the forecasts (Bentwich et al., 2005; Berezikov et al., 2005) and also the the number of targeted genes, considered to be the 30% of human genes, has probably been underestimated (Bartel, 2004).

1.2.1 Identification of miRNA target genes

1.2.2 Experimental study of miRNA mechanism and identification of target genes

We can divide experimental identification of miRNA targets in direct and indirect methods (Ørom and Lund, 2010): (i) direct methods allow the validation of miRNA-mRNA couples and are often based on the quantification of a reporter construct; (ii) indirect methods do not test the binding between miRNA and mRNA but screen all the interactions suggesting a set of candidates. These last methods use high-throughput techniques.

High-throughput approaches can be subdivided, as proposed by Ørom and Lund (2010), in transcriptome analyses, proteome analyses and biochemical approaches.

Among the biochemical approaches we can find the immuno-precipitation of Ago complex and the HITS-CLIP approach. The first is based on the isolation of functional Ago-miRNA-mRNA complexes using antibody anti-Ago followed by either microarray or sequencing analysis of the two set of RNAs. The second one uses the UV light to cross-link Ago protein with the associated miRNA and mRNA that will be subsequently identified by sequencing. Both techniques give similar results, they identify a pool of miRNAs and a pool of mRNAs reducing the possible miRNA-mRNA couples but they are not able to identify the direct interaction.

Among proteome analyses we can find SILAC techniques that use stable light and heavy isotope labeled amino-acids in cell culture to distinguish protein level variations after miRNA over-expression or inhibition. Proteins are identified using mass spectrometry and the protein quantity is considered proportional to the protein peak intensity. This technique is the first example of high-throughput proteomic data, however it is not able to screen all the proteome and to discriminate between direct and indirect interaction effects.

The transcriptome analyses can be divided in three categories depending on the experimental design: forced up-regulation or down-regulation of a single miRNA with subsequent measure of genome-wide expression, or measurement of both miRNA and gene expression on the same biological samples.

The miRNA up-regulation is performed transfecting high quantity of a specific miRNA in a cell line and looking for down-regulated genes. The miRNA down-regulation is performed inhibiting the miRNA of interest with complementary oligonucleotides and looking for up-regulated genes. These two approaches are widely used and several datasets based on this strategies are deposited in databases as Gene Expression Omnibus (GEO, Barrett et al. (2011)) and Array Express (AE, Parkinson et al. (2011)). However, miRNA up-regulation and down-regulation strategies have several drawbacks: (i) they fail to distinguish direct and indirect relations between miRNA and target genes; (ii) the expression of the miRNA and its target is not necessary anti-correlated, but it depends on the topology of the regulatory circuits in which they are involved; (iii) cell cultures do not have necessarily the same behavior of the cells in tissue (Ruepp et al., 2010); (iv) they are unable to measure miRNA activity performed exclusively inhibiting the translation; (v) overexpression of miRNA levels could cause multiple artifacts, such as saturation of RISC complexes, preventing the access of endogenous miRNAs (Khan et al., 2009); (vi) some inhibitors of a miRNA can have low level of efficacy and in some cases it is not possible to design an inhibitor blocking all members of a miRNA families (Peter, 2009).

The third category of transcriptome analysis is based on the use of miRNA-mRNA matched datasets. This technique depicts a more natural situation without experimental artifacts and it is suitable for studies in which an high number of patients samples is required. Despite the pros, this type of experiments are less frequently used than the previous one and more difficult to find in public repositories.

1.2.2.1 Bioinformatic identification of miRNA target genes

Despite the increase of experimentally validated miRNA targets the majority of them are unknown. The *in silico* predictions remain the only solution to pre-investigate the data in a rapid way.

In the last years several algorithms have been developed using strategies like the sequence alignment between 3' region of the genes and the seed sequence of the miRNA, the sequence conservation through species, the target site accessibility and the binding stability.

Some of the most famous target prediction algorithms are DIANA-microT (Maragkakis et al., 2009), ElMMO (Gaidatzis et al., 2007), miRSVR (Betel et al., 2010), Pictar (Lall et al., 2006), PITA(Kertesz et al., 2007), RNA22 (Miranda et al., 2006) and TargetScan (Friedman et al., 2009). All these algorithms differs for the features considered and for the strategy adopted to perform predictions. Evaluating and comparing these tools present several difficulties given the absence of a true solution.

Alexiou et al. (2009), in one of the latest reviews, compare the results of each algorithm with data retrieved from miRNA overexpression experiments, and with a collection of experimentally validated targets. The authors concluded that, despite some features are more useful and some programs like DIANA-MicroT and TargetScan are more accurate than others, in general all the programs fail to identify most of the targeted genes. The biggest issue is that we dont know the proportion of miRNAs that follow the rules used by the predictors. Then, when we make comparisons with experimental data we can eliminate false positive predictions but the total amount of the false negatives remain unknown (Ørom and Lund, 2010).

1.2.2.2 miRNA dedicated databases

An increasing amount of databases, which collect information about miRNAs and their targets, exists. These databases contain information coming from literature, concerning both *in silico* and experimental approaches in physiological and disease conditions. miRbase (Griffiths-Jones, 2010) is the most important web resource, especially for nomenclature and sequences. HMDD (Human MicroRNA-associated Disease Database) (Lu et al., 2008), miR2Disease (Jiang et al., 2009) and PhenomiR (Ruepp et al., 2010) are dedicated to miRNAs in diseases. miRGator (Nam et al., 2008), miRGen (Megraw et al., 2007), Argonaute (Shahi et al., 2006) contain *in silico* target gene predictions. miRecords (Xiao et al., 2009) and Tarbase (Vergoulis et al., 2012) contains validated miRNA targets and information about the experimental validation methodologies.

1.2.3 miRNA and disease

miRNAs are involved in a large number of processes and the deregulation of their expression could lead to dysfunctions and diseases. Cancer is responsible for about the 25% of all deaths in the U.S. and is one of the major public health problems in the world. For these reasons, cancer is one of the first and more studied pathology of modern biology, and miRNAs behavior were extensively studied in this disease. PhenomiR collection, which contains information about miRNAs associated with diseases, reveals that cancers are the most investigated (81%) followed by muscular (4,3%) and cardiovascular (4,1%) disorders. Among cancers the most investigated are leukaemia (16.7%), colorectal cancer (10.6%) and breast cancer (9.5%) (Ruepp et al., 2010).

Despite the high variety of tumor types the hallmarks of cancer are (i) self-sufficiency of tumor cells, (ii) insensitivity to anti-growth signals, (iii) absence of apoptosis, (iv) unlimited replicative potential, (v) induction and maintenance of angiogenesis, (vi) invasion and metastasis (Hanahan and Weinberg, 2011). Multiple aspects of cancer are regulated by miRNAs and they are aberrantly expressed in an high number of different cancers. It has been demonstrated that miRNAs are considered ideal candidates as diagnostic and prognostic markers to distinguish between type of cancer, stage and other clinical variables, even better than mRNAs. Moreover, they are attractive therapeutic targets for the relative facility of their overexpression or inhibition (Lee and Dutta, 2009).

Despite the general low expression of miRNAs in cancer samples compared to normal tissues miRNAs can act both as oncomir, miRNAs that cause tumor, or oncosuppressor, miRNas that protect against tumor.

miRNA aberrant expression can occur through several mechanisms, such as (i) chromosomal abnormalities like deletion, duplication and translocation, (ii) Single Nucleotide Polimorphisms (SNPs) either in miRNA locus or in binding site for miRNA or (iii) alternative splicing. In addition, also virus are involved in tumorigenesis (iv) encoding viral miRNAs or (v) through viral insertion near miRNA loci (Lee and Dutta, 2009).

The miRNA-associated disease network, obtained connecting diseases that share at least one miRNA associated to the pathology, shows that all cancer are highly connected together and clearly separated from any other diseases. Therefore, different cancers may have the same miRNAs acting as oncomir or oncosoppressor although not necessarily involved in the same onco-mechanism (Lu et al., 2008).

1.2.4 miRNAs in Epithelial Ovarian Cancer

The role of microRNAs in high-grade ovarian cancer has been extensively studied in the last decade, highlighting a strong involvement of both the non-coding RNAs and the enzymes of the miRNA processing machinery (Iorio et al., 2007; Ventura and Jacks, 2009). Studies of stage III and IV EOC suggest that miRNAs are down-regulated in tumour samples compared with normal tissue, and that their altered expression affects response to chemotherapy. Genes involved in the biogenesis of miRNAs were also altered in EOC. Furthermore, Dicer and Drosha mRNAs and protein low concentrations have been associated with poor outcome in a cohort of 108 stage III and stage IV ovarian tumors (Pampalakis et al., 2010).

Numerous miRNAs associated with several aspects of EOC have been detected for their ability to target multiple oncogenes (P53, PTEN, RAS, BRCA1 and BRCA2, VEGF, TUBB3) at the same time.

Specifically, our group, in collaboration with the "Mario Negri" Institute, revealed that miR-200c is an independent prognostic factor of EOC. Examining a large cohort of patients of stage I EOC, we found that the loss of miR-200c and the increasing of the expression of its targets TUBB3 and VEGFA, produces a decreasing of overall and progression free survival (Marchini et al., 2011).

Recently, this finding has been used to develop a therapy for EOC by Cittelly et al. (2012). They show how the restoration of miR-200c expression into ovarian cancer cells elicits an increased anoikis sensitivity (a particular type of apoptosis induced by anchorage-dependent cells detaching from the surrounding extracellular matrix, the name derived by Greek that means "...the state of being without a home") and a reduction of *in vitro* adherence to biologic substrates. Since anoikis-resistance is a critical steps in EOC, restoration of miR-200c expression in intraperitoneal xenograft model, an in vivo preclinical model of ovarian cancer, decreased tumor formation and tumor burden. The study demonstrates also that a restoration of miR-200c expression in combination with paclitaxel, a mitotic-inhibitor chemotherapeutic agent, enhances the decrease of the tumor (Cittelly et al., 2012).

1.2.5 Multiple miRNAs cooperations

An indication of synergism in miRNA mode of action is highlighted by the analysis of chromosomal distribution of miRNAs in relation to the studied pathology. Clustered miRNAs (multiple miRNAs that share the same promoter because originated by a long primary transcript) are not only co-expressed but also act in concerted way. The fraction of disease-associated miRNAs within a cluster are on average 1.4 times higher than the background frequency and polycistronic loci are on average 3.5 times more disease-associated than expected by chance (Ruepp et al., 2010). Malumbres (2012) assertes that the deregulation of miRNA clusters occurs systematically in human diseases, and that not only one miRNA acts as the causative, but multiple miRNAs act synergistically on cellular processes.

The cells express multiple miRNAs at the same time and it is a matter of fact that a single gene can be targeted by multiple miRNAs. This new layer of complexity is supposed to be considered to understand the miRNA function (Krek et al., 2005; Lewis et al., 2005; Peter, 2010).

In general targets down-regulation, yielded by a single miRNA, is quantitatively moderated at the protein level, (Baek et al., 2008) suggesting that multiple miRNA contributions determine whether a gene is expressed. The single contribution can be weak but the sum of several weak contribution can have a significant impact on the system, leading to combinatorial diversity and synergy in biological effects (Peter, 2010).

Moreover, while miRNAs control the gene expression, the expression of the genes is also influenced by other genes, in an highly connected regulatory network composed of miRNA-gene and gene-gene edges. The complexity of the topology of this network highlights the needs of studying the cell signals with holistic approaches, that are able to solve and decipher the biological complexity. It is increasingly clear that we can not continue studying miRNAs and genes separately outside the context of their network interactions.

Finally, Poliseno et al. (2010), Tay et al. (2011), Salmena et al. (2011) demonstrated that miRNA co-operate in a combinatorial manner both on coding and non-coding mRNA transcripts, and that,*viceversa*, mRNAs competitively act as a decoy for microRNAs. In a seminal paper Poliseno et al. (2010), studied PTEN and its pseudogene, that share the same 3'UTR and the seeds for a set of miRNAs. They demonstrated that an up-regulation of the pseudogene induces a recruitment of miRNAs that, as a consequence, increases the presence of the PTEN mRNAs. This is because the pseudogene acts as a sponge for their inhibitors. Generalizing this theory, we can assert that all the coding and non-coding, long and short transcripts forms a large-scale regulatory network across the transcriptome.

In this perspective, one of the new and biggest challenge in biology is the combination of different kind of data, such as gene and miRNA expression, and the development of new methods for the integration and interpretation of biologic complex systems.

1.2.6 miRNAs in signaling pathways

Especially for their rapid action and multi-genes regulatory capacity, miRNAs are the best candidates to play pivotal role in the modulation of the signal transductions in time and space.

A signal transduction is a mechanism that convert a signal (stimulus) in a change of behavior of the cell, i.e. alteration of metabolism, proliferation or apoptosis, regulation of transcription of genes, cell commitment etc.

The signaling pathways are characterized by two signal transduction mechanisms: (i) context-dependent transcriptional activation and (ii) inhibition of a default repression. Both ensure that the change is operated only in the presence of the signal, maintaining passively or actively turned off the signal in absence of the stimulus.

Although miRNAs have the role to down-regulate gene expression, their function is not only repressive but is dependent on pathway topology. In fact, there are examples of miRNAs involvement in both kind of transduction (Inui et al., 2010).

miRNAs can influence the interpretation of the signal, in fact not always a stimulus elicits an unambiguous on or off situation, often the cell has to distinguish between real signal and too week or too transient inputs. miRNAs in these situations play their role amplifying or repressing the response, so that the signal can or can not pass the threshold of sensitivity of the system.

Another fundamental consideration is that the miRNA presence is strictly dependent on the cell-type in which the reaction occurs, determining in this way modulation of the signal, which depends on the context. The unique miRNA milieu of each cell type gives great plasticity to the system, suggesting why different cell types, with so different functions, can share the same cascades of signal. Moreover, it can explain how the cell is able to perceive quantitatively the signal generating a response tailored to the intensity and the duration of the stimulus (Inui et al., 2010).

Analyzing combined gene and miRNA expressions we see, as expected, that miRNAs and their target genes are anti-correlated, this means that when miRNA is highly expressed the level of expression of the target gene is low and *vice versa*. However, unexpectedly, we also find that many miRNA-mRNA couple have positive correlated expression, either high or low expressed at the same time. Both these situations can be explained using the theory of networks, and in literature we can find several examples of both situations involving miRNAs (Inui et al., 2010).

A feed-forward loop in biology is a well-known network motif. The module is composed of at least three elements (A, B and C), in which A interacts with B and C, and B interacts with C. This system can be coherent or incoherent, in the first case all the interactions are in agreement each others, in the second case two signals with opposite effects on the same element are present. As shown by Inui et al. (2010), an example of coherent loop using miRNAs in signaling pathways is obtained when a signal activates a gene and represses a miRNA that is also the repressor of the gene itself, in this case expression of miRNA and its target gene will be anti-correlated.

On the other hand, an example of incoherent loop is obtained when a signal activates both the miRNA and its target gene. In this situation, when the signal is turned on we will see incoherent expressions of the two elements. In biology this apparently illogic situation has fundamental advantage: it prevents undesired pathway activation by random signal fluctuations, placing the threshold of the system sensitivity at a higher level. In fact, to turn on the signal, the stimulus has to overcome the repression operated by the miRNA. At the same time, the system can easily maintain the steady-state (Herranz and Cohen, 2010). Moreover, the temporal difference to produce miRNAs and proteins (miRNA is faster processed) allows miRNA to affect gene expression more rapidly than what is done by transcription factors.

In this way miRNAs are fundamental elements of signaling pathway conferring temporal, as well as quantitative precision.

1.3 Pathways

In a famous commentary regarding systems biology, Lazebnik (2002) using the analogy between biological pathways and electronic circuits, proposed the use of standard procedures through which even a biologist - without any specific knowledge - could fix a radio. One of the most challenging goals of modern biology is to decipher and describe the complexity of cell systems, and what Lazebnik pointed out is that without the integration of knowledge coming from different fields of science, the efforts of reverse engineering the cell are destined to fail (Beltrame et al., 2011).

Recently, research on system biology has been characterized by an increasing number of efforts to define common languages for sharing information in multidisciplinary areas (Abbott, 1999) with the aim to develop tools, to describe accurate models, run effective simulations, visualize, analyze and integrate high-throughput data. Networks describing the interactions occurring within cell macromolecules are key elements for this research.

According to the nature of nodes and interactions, existing biological networks can be classified into three categories: metabolic pathways, gene regulatory networks and signaling pathways (Li and Davidson, 2009; Wang et al., 2007). Metabolic pathways are chain of chemical reactions catalyzed by enzymes, while regulatory networks are composed by relations of expression regulations, that occurs between the transcription factors and the regulated genes. Finally, signaling pathways are formal descriptions of the signaling processes by which a cell converts certain signals into others, involving interconnected, finely regulated structures that may present a high level of redundancy.

There are many public resources which store and share representations of these networks, however currently there is no gold standard on how biological pathways should be represented. This shortcoming affects particularly signaling pathways: without solid, consistent and unambiguous representations, hypotheses and analyses are not effective.

Furthermore, a proper representation of a pathway is important to enable efficient knowledge management and integration of data coming from multiple sources. Recent efforts on the pathways representation have followed two main trends: a proper graphical representation and a machine-readable format. According to the presence of graphical and machine-readable formats, pathway representations can be classified into three categories (Pan et al., 2003): static, providing a non-modifiable graphical representation; semi-dynamic, representing information not only as a graphical map, but also has a machine-readable format, which is not, however, strongly interconnected with the graph; dynamic, where the graphical representation format depends directly on the underlying data model, and thus any modification in the latter can be immediately translated to the former. At the time of writing, all pathway representations stored in public databases are either static or semi-dynamic (Beltrame et al., 2011).

The most recent example of a pure graphical representation is the System Biology Graphical Notation (SBGN; Le Novère et al. (2009)). SBGN splits the representation of a biological network into three different levels (the process definition, the entity relationship and the activity flow language). The three representations are constructed in order to capture different aspects of the biological systems, defining a set of glyphs and constraints to reduce ambiguity and improve interpretation. The resulting representations are highly informative, and SBGN quickly achieved a broad consensus in the scientific community. However, despite ongoing efforts (Czauderna et al., 2010), an SBGN-dedicated pathway repository does not exist yet, and the conversion from the existing pathway representations to SBGN format is still difficult (Beltrame et al., 2011).

Machine-readable formats, on the other hand, aim at creating a representation of the pathway that can be read and interpreted by computer programs and used to perform analyses or simulations. Many formats have been proposed: the Systems Biology

Markup Language (SBML) (Hucka et al., 2003) especially dedicated to quantitative simulations; the Biological Pathways eXchange (BioPAX; Luciano (2005)) and KGML (Kanehisa and Goto, 2000) extensively treated in the next paragraphs; BCML (Beltrame et al., 2011) a pathway format focused on the tissue specificities, created to be simpler than the existing formats but at the same time richer in details. I personally contributed to the construction of BCML as a subproject of my Ph.D..

1.3.0.1 Pathway Repositories

A variety of databases containing information on cell signaling pathways have been developed in conjunction with methodologies to access and analyze the data (Bauer-Mehren et al., 2009). Pathway databases serve as repositories of current knowledge on cell signaling. They present pathways both in a graphical format, comparable to the representation present in text books, both in machine-readable formats. This last form allows the exchange between different software platforms and further processing by network analyses, visualization and modeling tools. At the present day, there exist a vast variety of databases containing biochemical reactions, such as signaling pathways or protein-protein interactions. The Pathguide resource serves as a good overview of current pathway databases (Bader et al., 2006). It lists more than 200 pathway repositories; over 60 of those are specialized on reactions of the human species. However, only half of them provide pathways and reactions in computer-readable formats needed for automatic retrieval and processing. Although these initiatives are successful results of the joint efforts of a wider community, they are still incomplete, different databases are characterized by different annotations and only a part of the interactions are confirmed by all the repositories. On the other hand, Cerami et al. (2011) have recently developed a web repository aiming at collecting and integrating all public pathway data available in standard formats. It currently contains data from nine databases with over 1400 pathways and 687,000 interactions.

1.3.0.2 Machine-redable pathway formats

Aware of the complicate scenario described above, we decide to focus our attention to BioPAX and KGML because, togethers, they recover the vast majority of public available pathways. Pathway data of Reactome, NCI and BioCarta are available in BioPAX, while KEGG pathways in KGML. BioPax and KGML languages use ontologies to describe pathways.

An ontology is a formal description of a concept for the sharing and reuse of knowledge among software entities. It is composed of objects, with properties and relations with other objects, and restrictions, and it often uses a controlled vocabulary. Furthermore an ontology is often arranged into a hierarchy, with parent classes representing more general concepts and child the more specific ones. The scope of the ontology is to specify an abstract data modeling representation that can be queried, exported, translated, and unified across independently developed systems and services (Demir et al., 2010; Gruber, 1995).

BioPAX is a community process started in 2002, it is defined as the pathway standard ontology. The last release of BioPax (level 3) includes all the mayor concepts familiar to biologists. With BioPax, we can represent both metabolic both signaling pathways. The BioPAX language is based on OWL (Ontology Web Language), that is an RDF/XML-based language but with a larger vocabulary and stronger syntax. More details and documentation are available at http://www.biopax.org/ and Demir et al. (2010).

KGML, the KEGG-dedicated markup language, is an XML language with a schema dedicated to KEGG data model. KGML is simpler than BioPAX, however less detailed and more ambiguous. In KGML, the pathway element specifies one graph object with "entries" that are elements (nodes such as genes, proteins, complexes, compounds) and with "relation" and "reaction" elements (edges). More details at http://www.genome.jp.

1.3.0.3 Topological pathways analysis

A great deal of effort has been directed towards the study of gene sets (hereafter GSA) in the context of microarray data analysis. The aim of these analyses is to identify groups of functionally related genes with possibly moderated, but coordinated, expression changes. Several GSA tests, both univariate and multivariate, have been recently developed. See Ackermann and Strimmer (2009) for a comprehensive review, Goeman and Mansmann (2008), Nam and Kim (2008) and Dinu et al. (2009) for a detailed description and a critical investigation of the tested hypotheses.

These approaches, although effective, miss the information of the topological properties of the pathways. To this end, the seminal paper by Draghici et al. (2007) proposed a radically different approach (called impact analysis, *SPIA*) attempting to capture several aspects of the data: the fold change of differentially expressed genes (DEGs), the pathway enrichment and the topology of signaling pathways. In particular, *SPIA* enhances the impact of a pathway if the DEGs tend to lie near its entry points. Massa et al. (2010) introduced TopologyGSA an alternative approach that is based on a correlation structure test. Specifically, the graphical model theory is used to decompose the overall pathway into smaller cliques, with the aim of exploring in detail small portions of the entire model. Recently, Isci et al. (2011) proposed a Bayesian Pathway Analysis that models each biological pathway as a Bayesian network (BN) and considers the degree to which observed experimental data fits the model. L. Jacob and Dudoit (2010) developed a graph-structured two-sample test of means for problems in which the distribution shift is assumed to be smooth on a given graph. Finally, an evolution of TopologyGSA has been published in 2012, called *CliPPER*, introducing for the first time a great innovation in the field of pathway analysis. *CliPPER* is able to recognize not only the significantly deregulated pathways, but also the portion of the pathways with the greatest association with a specific phenotype (Martini et al., 2013).

During my phD project, I used, tested and evaluated several of these approaches. However in this thesis for reason of brevity, I would like to focus the reader attention on two of them: *SPIA* and *CliPPER*. These two algorithm has been extensively treated in Chapter 6, along with the EOC expression data analyses.

1.3.1 Pathway conversion

The methods cited above need a graph structure. In this perspective, the retrieval of pathway information and the subsequent conversion into a gene/protein network is crucial.

The extraction of the topological information from a biological pathway and their interpretation to obtain a network is not a trivial task and are still extremely dependent on the level of detailed information provided by the data format (Alves et al., 2006; Draghici et al., 2007; Massa et al., 2010; Beltrame et al., 2011).

Pathway annotations comprise a myriad of interactions, reactions, and regulations which are often too rich for the conversion into a network. In particular, challenges are posed by the presence of chemical compounds mediating interactions and by different types of gene groups (e.g. protein complexes or gene families) that are usually represented as single nodes. R packages for pathway conversion are available, such as *KEGGgraph* (Zhang and Wiemann, 2009) and *NCIgraph*, but share some drawbacks: i) they are focused on a single pathway database; ii) they do not consider gene connections through chemical compounds; iii) they do not handle the various kinds of biological gene groups.

During my Ph.D. thesis I worked on the development of *graphite* (GRAPH Interaction from pathway Topological Environment) a bioinformatic tool that fill all these gaps.

1.4 Aims and organization of the thesis

The aim of my thesis is the development of methodologies and tools to integrate gene and microRNA expressions to dissect the circuits in Epithelial Ovarian Cancer (EOC) exploiting and improving topological pathway analysis methods.

In particular, I contributed to the development of a computational tool, later called *graphite*, that i) facilitates the access and the integration of pathway data and microRNA information and ii) allows the application of topological analysis on these pathways, then, I applied this tool to analyze and identify cell circuits differentiating Epithelial Ovarian Cancer histotypes.

This thesis is divided in two main parts, the first part (chapters 2- 4) is dedicated to the development of computational and bioinformatic tools and the second part (chapters 5 and 6) is dedicated to the application of this methodologies to early stage of EOCs.

Specifically, chapter 2 describes the technical details and rules defined for the conversion of pathway topologies to gene network; chapter 3 describes the functionalities of *graphite* Bioconductor package that provides pathway data as networks; chapter 4 is dedicated to the test and application of *graphite* to simulated and benchmark datasets.

Chapter 5 is dedicated to the analysis of miRNA expression profile on a large set of early stage EOC patients. A resampling strategy is presented aiming at identifying a robust miRNA signature differentiating EOC subtypes. Validations with qRT-PCR of the proposed signature will be presented, as well.

Chapter 6 represents the fusion of the two cores of the thesis: the application of *graphite* to the identification of EOC subtype-specific biological circuits composed of genes and miRNAs. An expanded version of *graphite*, including miRNAs in pathway topology, is presented and applied. In this chapter we show how to move from single and isolated markers to the characterization of more informative and therapeutically attractive histotype-specific circuits.

The work presented in this thesis is the results of an ongoing collaboration, started in 2010, with the "Mario Negri" Institute, in particular with the groups of Dr. Maurizio D'Incalci, head of Oncology Department and Cancer Pharmacology unit and Dr. Sergio Marchini, head of the Translational Genomic Unit.

Chapter 2

From pathway to network: the conversion procedure.

The procedure we developed into convert pathway topology in to networks is called *Graphite* (GRAPH Interaction from pathway Topological Environment).

Graphite is divided in two main steps:

- The conversion procedure that creates the networks from pathway topology
- The Bioconductor R package that allows the usage and the analysis of these networks.

This chapter is dedicated to the explanation of the the first part of *Graphite*.

We would like to warn the reader that this chapter is dedicated to the technical description of the rules used for pathway conversion and that it is not necessary for the comprehension of the thesis as a whole. For the description of the resulting Bioconductor package see Chapter 3

2.1 The conversion procedure

A network is a simplified structure in which nodes represents genes. That are connected by edges representing their biological relations.

The pathway data formats that *Graphite* is able to manage and interpret are:

- **KGML**

- **BioPax**

The conversion procedure is based on two steps:

- **Acquisition** – The reading, parsing and storing of the pathway data into a unifyied data model (hereafter called Full Model), able to standardize the heterogeneous information derived from different pathway formats. Each pathway format (KGML and BioPax) has a dedicated acquisition procedure.
- **Simplification** – The simplification of the complex information stored in the Full Model into a network model.

At the end of the conversion *Graphite* provides for each pathway:

- **Network data** – A network is represented as list of nodes and edges. It provides for each edge the source, the destination, the direction and the type.
- **Warnings** – A series of warnings occurred during the pathway conversion phase. Warnings are produced whenever the program is not able to convert an element.

Pathway conversion is not always possible. The pathway data must satisfy three main requirements:

1. the pathway elements must have IDs;
2. a pathway must not contain duplicated IDs;
3. the pathway elements IDs do not contain a number sign (#).

If the data does not comply with these rules a global error is generated and the pathway is not converted.

2.2 The ”Full model”

2.2.1 Pathway definition

The KEGG database provides separate KGML files, one for each pathway. A pathway is thus defined by all the reactions defined within each file. For all the other databases based on BioPax format we identify a pathway upon encountering a ”pathway” element.

2.2.2 The full model components

For each pathway the following components are defined:

- **E**NTITIES
- **I**NSTANCES
- **R**ELATIONS

2.2.2.1 The Entity

The Entity is the abstract and not-redundant representation of a pathway element. The entity defines a set of general attributes qualifying all the physical representations of that element in the pathway. The entity corresponds to the entity Gene or PhysiscalEntity defined in BioPax. A full model entity can be a gene, a compound, a group or a generic type called "other".

Each entity is defined by:

- **I**D
- **X**REFS – A list of external references and articles linked to the entity.
- **S**OURCES – A list of sources that are all the native element ids of the pathway data used to generate the described entity. In other words, the series of original IDs that are transformed to generate the full model entity.

For the Entity of type Group, the following additional attributes are defined:

- **G**ROUP TYPE – Two type are allowed: AND and OR. The OR groups contain a set of possible alternative members. These groups are generally gene families, sets of genes with similar sequences and biochemical functions. The AND groups correspond to protein complexes (proteins linked by protein-protein interactions).
- **G**ROUP COMPONENTS – The list of entity IDs contained in the group.

2.2.2.2 The Instance

The Instance is the physical representation of an entity in the pathway. Each instance points to only one entity and multiple instances can refer to the same entity. The

Instance corresponds to the Physical Entity defined in BioPax. Nested instances are not allowed in the full model and if encountered are simplified, flattening all the members, maintaining AND and OR relations starting from the innermost element.

An Instance is described by:

- **I**D
- **E**NTITY POINTER – The entity referred by the instance.
- **L**OCATION – Information about the biological location of that particular instance in the pathway.
- **E**NTITY SOURCES – A list of sources that are all the native element ids of the pathway data used to generate the described entity. In other words, the series of original IDs that are transformed to generate the full model entity.

2.2.2.3 The Relation

The Relation describes the connection between two instances or an instance and another relation of the pathway. The Relation can be:

- a process that connects two instances;
- a catalysis that connects the catalysts and the process catalyzed;
- a modulation that connects the modulator and the modulated catalysis.

A Relation is described by:

- **I**D
- **X**REFS – A list of external references and articles linked to the relation.
- **S**OURCES – A list of sources that are all the native element IDs of the pathway data used to generate the described relation.
- **E**VIDENCES – A list of experimental evidences validating the relation. The evidences are provided by the pathway data.
- **T**YPE – The relation type, is free text describing the type of relation (e.g. activation, inhibition phosphorylation etc.).

Processes have the following additional attributes:

- **D**IRECTION – Can be direct or undirected (without direction).
- **L**EFT – The instance IDs or the list of instance IDs from which the relation starts.
- **R**IGHT – The instance IDs or the list of instance IDs into which the relation ends.
- **I**S EQUIVALENCE – A flag to mark self-loops.

Catalyses and modulations have the following additional attributes:

- **C**ONTROLLERS – One or more instance IDs or a list of instance IDs from which the relation starts.
- **C**ONTROLLED – It is the controlled relation ID that can be anything but a modulation. In case of catalysis is a process ID, in case of modulation is a catalysis.

2.3 The conversion of pathway data to the full model

2.3.1 BioPax conversion

The BioPax pathway data is read and parsed, the pathway elements (pathways, pathway components, entities, unificationXRefs, bioSources, cellular locations, evidences) are stored in different categories and errors are reported. The pathway conversion proceeds only if no global errors have been generated.

Pathways have a hierarchical structure. A pathway contains one or more pathway components and a pathway component contains one or more interactions. The conversion procedure traverses the pathway tree recursively looking for the interactions. Empty pathways are ignored, nested pathways are solved, self-contained pathways are reported as warnings. For a graphical example see Figure 2.1 where the P1 Pathway has a tree-like structure and contains P2, P3 and P4. P2 contains two interactions I1 and I2, P4 is an empty pathway and P3 contains erroneously P1. The results of this process of conversion are: P1=P2{I1,I2} and two warnings on P3 and P4.

2.3.1.1 Interaction rearrangement procedure

After the pathway interaction collection procedure, the interactions are processed according to the three main BioPax type of interactions: Template Reaction, Control and Conversion.

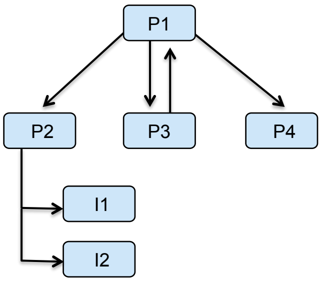


Figure 2.1: Example of pathway structure.

- **T**EMPLATE REACTION – It generates a FULL MODEL PROCESS. At least one right element is required, otherwise the conversion fails and a warning is generated.
- **C**ONTROL – It generates a FULL MODEL CATALYSIS or MODULATION. Following the BioPax definitions, catalyses are those reactions whose right element is a process, while those those have a catalysis as right element are considered modulations. If this pattern is not observed a warning is generated.
- **C**ONVERSION – It generates a FULL MODEL PROCESS. This particular kind of interaction is flagged as EQUIVALENCE.

2.3.1.2 Physical entity simplification and instance generation procedures

After the interaction rearrangement, the conversion of physical entities is executed. The BioPax Physical Entities converted into the FULL MODEL INSTANCES as follows. A Physical entity has to be composed by at least one element and has to contain at least one unificationXRef. Otherwise, a warning is generated.

If the physical entity has only one component, it is translated into a full model instance that refers a full model entity of type gene, compound or other in case of respectively protein, small molecules and RNA/DNA references.

If the physical entity contains more than one element, it is translated into a full model istance that refers to a full model group entity. In the BioPax format only one group type is described: protein ”complexes” (group AND) with the element ”complex”. However, it often happens that a protein tag contains multiple references pointing to alternative elements of the process. These entities are stored in the full model as an OR group. The nested physical entities are solved recursively. The sources of the generated instance keep track of the elements used to create the instance itself. Cellular location consistencies are checked across nested physical entities.

If the physical entity is a BioPax generic reference a warning is generated because there is not sufficient information to drive the conversion.

2.3.1.3 The procedure of correction of multiple IDs referring to the same biological entity

The unificationXref is an attribute of the BioPax entity. Each entity has multiple unificationXRef and an unificationXRef can be referred by multiple entities. Entities with equal unificationXRefs are considered the same biological entity and, for this reason, collapsed into an unique entity (the first in the lexicographical order) called the master entity. All the other entities, that are synonyms of the master entity, are replaced by the master entity.

2.3.2 KGML conversion

The KGML conversion into the full model works differently. First of all, a KGML pathways do not follow a hierarchical structure. During the KGML conversion, the pathway data is read and parsed, and the pathway elements of interest are stored in different categories (map, relation, reaction, compound, gene, group) and errors are reported.

2.3.2.1 Interaction rearrangement procedure

The KGML interactions are:

- **R**ELATION – It generates two FULL MODEL PROCESSES. The relation occurs from a source to a destination element trough a mediating element and is translated in two processes, (i) from the source to the mediator element and (ii) from the mediator to the destination element. At least one source element and one destination element are required, otherwise the conversion fails and a warning is generated. In the case there are no mediating elements only one process is generated from the source to the destination element. The type of relation is reported according to the full model relation.
- **R**EACTION – It generates a FULL MODEL PROCESS from the substrate to the product element. At least one source element and one destination element are required, otherwise the conversion fails and a warning is generated. The type of reaction is reported according to the full model relation.

2.3.2.2 Instance generation procedure

KGML entities (gene, compound, group) are converted to instances. KGML entities have to be composed by at least one element and have to contain at least one name that allows the biological identification of the element.

If the KGML entity has only one component, it is translated in a full model instance that refers a full model entity of type gene or compound, depending on the KGML annotation.

If the KGML entity of type gene contains more than one element it is translated in a full model instance that refers to a full model group entity of type OR.

If the KGML entity of type group is translated in a full model instance that refers to a full model group entity of type AND.

The sources of the generated instances keep track of the elements used to create the instance itself. The nested KGML entities is not allowed and the cellular locations are not reported.

2.3.2.3 The procedure of correction of multiple IDs referring to the same biological entity

Multiple KGML entities pointing to the same biological entity can be created. Entities with equal names are considered the same biological entity and for this reason collapsed into a unique entity (the first in the lexicographical order) called the master entity. All the other entities, synonyms of the master entity, are replaced by the master entity.

2.4 The full model simplification procedure

Once the Full model is generated for all pathways, will start the second step of simplification.

2.4.1 Conversion of gene instance into nodes

Each instance of type gene is converted into a node that has as ID the biological identifier of the gene provided by the pathway data format, called hereafter native identifier. For a graphical example see Figure 2.2.

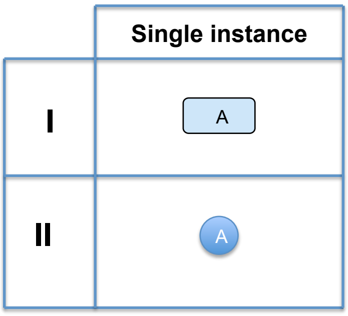


FIGURE 2.2: Example of a node creation starting from a gene instance

2.4.2 Process group instances

Nodes within a group are solved at the beginning of the conversion process through a three-step procedure: (i) the simplification of the mixed group, (ii) the removal of the redundant elements and (iii) the replacement of group nodes with the contained single nodes and their relations.

In the full model we have three kind of groups: groups with only genes, groups with only compounds and mixed groups composed of both genes and compounds.

The first step is the simplification of mixed groups that consists in the elimination of compounds. This step is a preparatory phase for the signal propagation procedure through compounds, see Section 2.4.5. It has been seen that the presence of compounds in mixed groups is only a generator of meaningless redundant edges. For a graphical example see Figure 2.3.

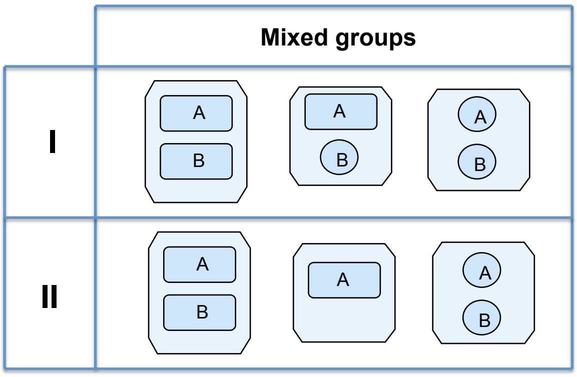


FIGURE 2.3: Example of simplification of mixed groups that consists in the removing of the compounds

The second step is dedicated to the removal of redundant elements (both genes or compounds). After this step only a single element is maintained inside the group.

The third step, is the replacement of the group with the single nodes it contains. After this step the group disappears. This phase takes into account the different biological nature of groups: they can be AND groups, representing protein complexes in which all elements are physically linked together; or they can be OR groups, pools of genes that are alternative member of a process (gene families or group of different protein isoforms). The AND groups are solved connecting all group elements together. This structure is called, in graph theory, a clique. On the contrary, OR groups are transformed in as many nodes as the elements belonging to the group without introducing edges among each other. For a graphical example see Figure 2.4.

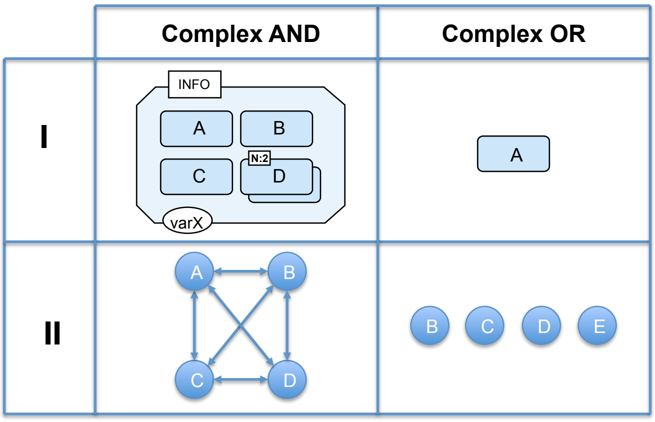


Figure 2.4: Complexes simplification procedure: the AND groups are solved connecting all the group elements togethers, the OR group are transformed in many nodes many elements we have in the group without edges.

2.4.3 Creating edges

In the full model we can find three kinds of relations: processes (divided in equivalences and non-equivalences), catalyses and modulations. The simplification of the relations into edges has a dedicated procedure for each relation type. Each process is substituted by an edge from the source to the destination. Although direction and type are maintained as close as possible to the original relation described in the pathway data, some new types have been introduced. A similar procedure is applied when the source element and/or the destination element are groups sharing components. For a graphical example see Figure 2.5.

Catalyses, that are relations pointing to other processes, are replaced with two relations of type control, one that enters in the catalyst and the other that exits from the catalyst. After the conversion, the type of the new catalysis maintains all the information about their origins (if it is the IN or OUT arrow and the type of the catalysis, inhibition, activation, etc.). The directions of the two control relations follow the direction of the catalyzed process. For an example, see figure 2.6 two edges are generated: control(In(INHIBITION)) from A to C, control(Out(INHIBITION)) from C to B.

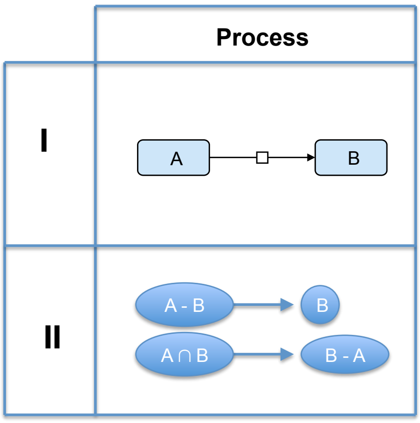


FIGURE 2.5: Example of process simplification

Sequential catalyses are treated as single and independent catalyses, see Figure 2.7.

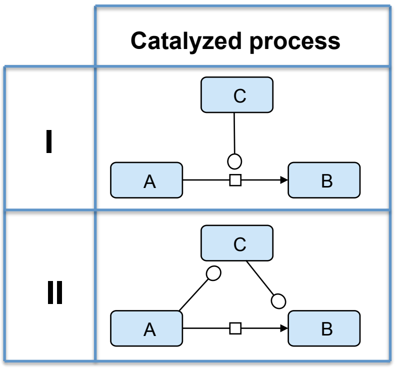


FIGURE 2.6: Catalysis of a process is simplified replacing the catalysis arrow with other two control processes, one that go in and the other out from the catalyst, two edges are generated: control(In(INHIBITION)) from A to C, control(Out(INHIBITION)) from C to B.

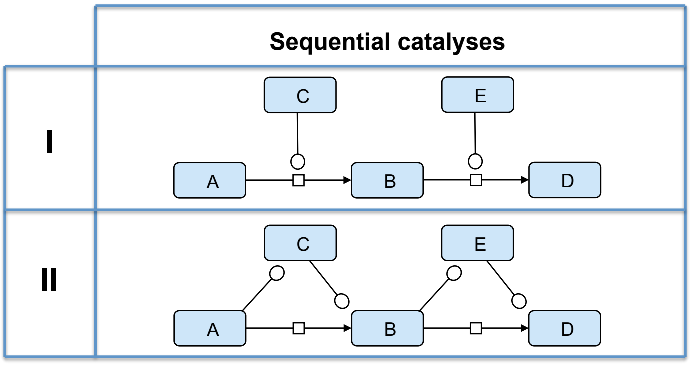


FIGURE 2.7: Simplification on sequential catalyses

The relation of type modulation, that starts from the modulator and points to the catalysis process, is replaced with a relation of type control, from the modulator to the catalyst. After the conversion, the type of the new control relation maintains the type of the catalysis (inhibition, activation, etc.). For a graphical example, see Figure 2.8.

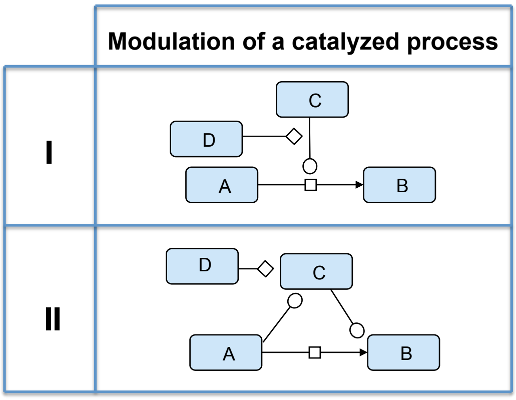


Figure 2.8: Modulation of a catalysis is simplified replacing the modulation relation with a control relation from the modulator to the catalyst.

2.4.4 Collapsing relation of equivalence

All the relations that have source elements equal to the destination elements are called equivalences. This situation usually represents transport, complex association or dissociation, see Figure 2.9. If translated, these relations would generate meaningless self-edges; for this reason they are ignored and no edges are produced.

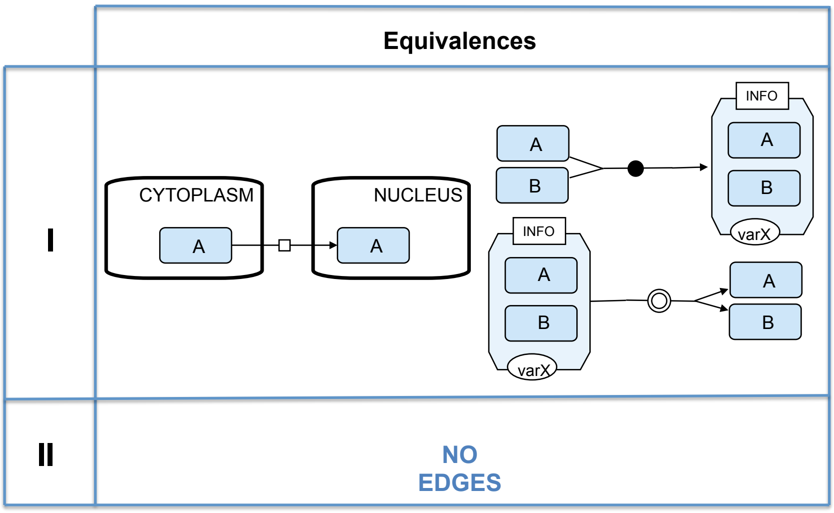


Figure 2.9: Examples of processes considered equivalences, that have source elements equal to the destination elements.

2.4.5 Elimination of compounds propagating the signal

The procedure of elimination of compounds maintaining the signaling chain is one of the most innovative part of the conversion system.

Compound-mediated interactions are interactions for which a compound acts as a bridge between two elements. As chemical compounds are not usually measured with high-throughput technologies, they should be removed from the network. However, the trivial elimination of the compounds will strongly bias the topology interrupting the signals passing through them. If element A is linked to compound c and compound c is linked to element B, then A should be linked to B. Moreover, to best fit the biological model we take into account cell compartment membership: the connection among genes A and B is established only if the shared compound c has the same localization in both the reactions.

Compounds are replaced by an undirect process of type indirect from the gene upstream to the gene downstream the removed compound or chain of compounds, see Figure 2.10. If multiple paths connect the same two genes, only shortest paths are retuned.

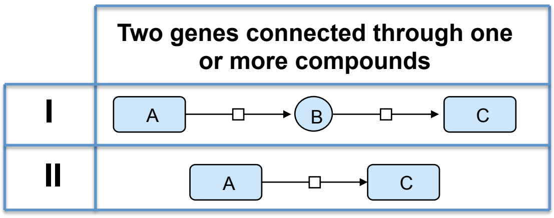


FIGURE 2.10: Two genes connected trough one or more compound

Different processes that share the same compound are merged only if they share the same cellular location, otherwise are kept separated, see Figure 2.11.

The propagation through compounds between genes located in two different cell compartments is performed only if explicitly present in the original pathway, see Figures 2.12 and 2.13.

During the propagation sequential catalyses have dedicated rules. First, sequential catalyses are simplified as a single catalysis generating two type of relations (controls and processes). Then, the propagation path are constructed taking into account the type of relations. While for catalyses the propagation can follow any type of edges, from processes the propagation can follow only other relations of process type, see Figure 2.14.

Not all compounds are involved in signal propagation. Some compounds are too frequently used and totally aspecific (examples of these type of compounds are ATP, GTP, NADH, etc.). They are simply eliminated.

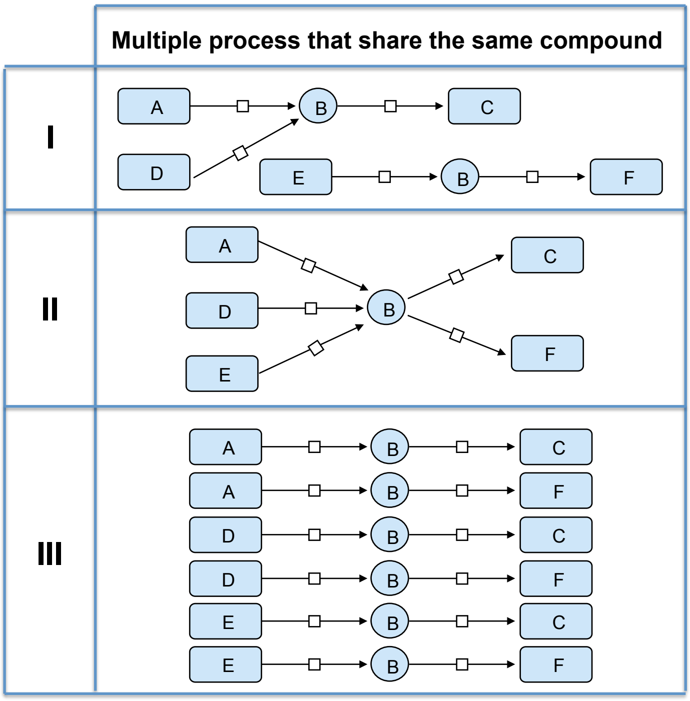


FIGURE 2.11: Multiple process that share the same compound

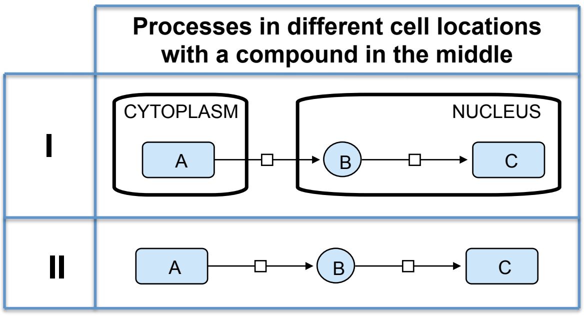

Processes in different cell locations with a compound in the middle
I
CYTOPLASM
NUCLEUS
A
B
C
II
A
B
C

FIGURE 2.12: Processes in different cell locations with a compound in the middle

2.4.6 Remove duplicated and self-edges

At the end of the conversion procedure all the duplicated edges and self-loops are removed.

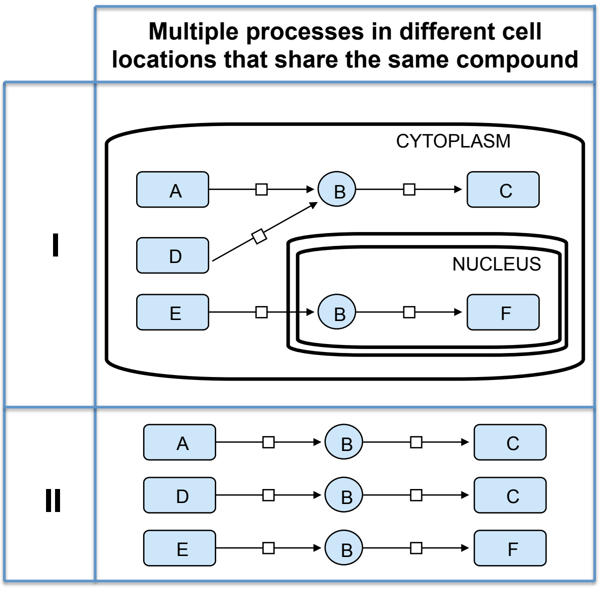

Multiple processes in different cell locations that share the same compound
I
CYTOPLASM
A
B
C
D
NUCLEUS
E
B
F
II
A
B
C
D
B
C
E
B
F

Figure 2.13: Multiple processes in different cell location that share the same compound

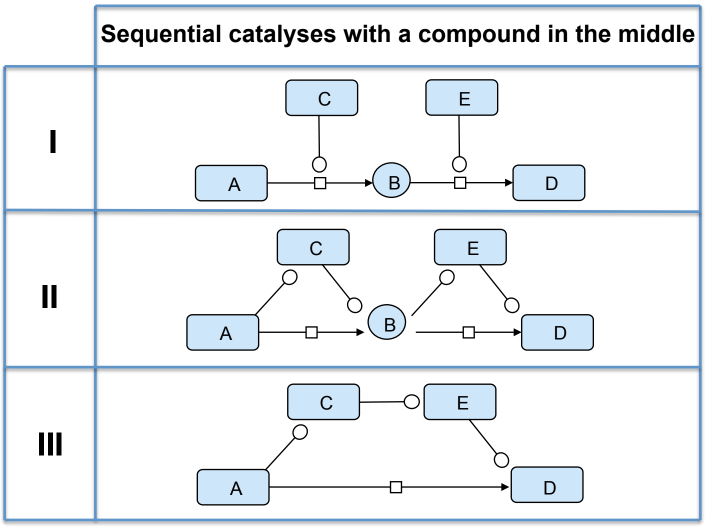


Figure 2.14: Sequential catalyses with a compound in the middle: during propagation, processes follow only other edges of type processes, instead control edges can be propagated following any kind of edges.

Chapter 3

graphite, the Bioconductor R package for topological pathway analyses

3.1 *graphite*

In order to gather curated information about human pathways, we have collected data from the four public databases that have emerged as reference points for the systems biology community: Reactome, KEGG, NCI, BioCarta. Reactome (Vastrik et al., 2007) uses the BioPax format, is backed by the EBI and is one of the most complete repository. Reactome is frequently updated and provides a semantically rich description of each pathway. KEGG (Kanehisa and Goto, 2000) uses KGML format, provides maps for both signaling and metabolic pathways. Finally BioCarta (www.biocarta.com) and NCI (NCI/Nature Pathway Interaction Database) (Schaefer et al., 2009), both published their data using the BioPax format. We transformed pathway data format into networks following the principles and rules mentioned in the previous chapter. Network data are available in a software package called *graphite*.

graphite was implemented using the statistical programming language R and the package is included in the open-source Bioconductor project (Gentleman, 2005). *graphite* has been constructed to act as pathway provider in R environment and functions as a bridge between pathway data and existing methods of topological pathway analysis, such as *SPIA* (Draghici et al., 2007; Tarca et al., 2009), *DEGraph* (L. Jacob and Dudoit, 2010), and *topologyGSA* (Massa et al., 2010; Martini et al., 2013).

This chapter of the thesis is dedicated to the description of the package and its functionalities.

3.2 Pathway functions recovery

In *graphite* a pathway database is a list of pathways. We can access to the database simply through its name. For instance, the names of the first three pathways can be simply retrieved using the following commands:

```
> names(biocarta)[1:3]
[1] "acetylation and deacetylation of rela in nucleus"
[2] "actions of nitric oxide in the heart"
[3] "activation of camp-dependent protein kinase pka"
```

In the same way we can access the Reactome , KEGG and NCI databases (through the *reactome*, *kegg* and *nci* lists, respectively).

Using *graphite* a pathway network can be retrieved using the name of the pathway:

```
>  p <- biocarta[["acetylation and deacetylation of rela in nucleus"]]
>  p
"acetylation and deacetylation of rela in nucleus" pathway from BioCarta
Number of nodes =6
Number of edges =9
Type of identifiers = native
Retrieved on        = 2011-05-12
```

or its position in the list of pathways:

```
>  p <- biocarta[[1]]
```

The object pathway is represented using the class *Pathway* appositely created. The pathway class allows the user-friendly visualization like the one showed above, in which are summarized the number of nodes, the number of edges, the data of its conversion from the pathway data format and the type of identifiers that have the network. "Native" identifiers are those of the original pathway definition. The class pathway has 3 slots:

- *title* – the name of the pathway;

- *nodes* – the list of nodes of the network;
- *edges* – the table of edges of the network.

We can access to each slot as follow:

```
>  p@title
[1] "acetylation and deacetylation of rela in nucleus"

>  nodes(p)
[1] "EntrezGene:4792"   "EntrezGene:5970"
[3] "EntrezGene:8841"   "EnzymeConsortium:2.3.1.48"
[5] "p50_0-0"             "ubiquitin"
```

Nodes can have heterogeneous IDs in their ”native” form:

The list of network edges is a table with four columns:

- *src* – SOURCE NODE
- *dest* – DESTINATION NODE
- *direction* – DIRECTION (directed or undirected)
- *type* – TYPE OF THE EDGE (phosphorylation, activation inhibition, control, etc..)

```
>  edges(p)
             src                   dest          direction     type
1 EntrezGene:4792                   p50_0-0   undirected  binding
2 EntrezGene:5970           EntrezGene:4792   undirected  activation
3 EntrezGene:5970  EnzymeConsortium:2.3.1.48   undirected  binding
4 EntrezGene:5970 ...
```

3.3 Graph

The pathway object can be also translated into a graphNEL, that is the most used R object for a graph. Many topological pathway analyses use graphNEL objects and for this reason we provided a function able to perform an easy conversion.

The function *pathwayGraph* builds a *graphNEL* object from a pathway object *p*:

```
> g <- pathwayGraph(p)
>  g
A graphNEL graph with directed edges
Number of Nodes = 6
Number of Edges = 14

>  edgeData(g)[1]
$ EntrezGene:4792|p50_0-0
$ EntrezGene:4792|p50_0-0 $weight
[1] 1
$ EntrezGene:4792|p50_0-0 $edgeType
[1] "binding"
```

3.4 Identifiers

Gene annotations databases are widely used as public repositories of biological information. Our current knowledge on biological elements is spread out over a number of databases (such as: Entrez Gene, RefSeq, backed by the NCBI http://www.ncbi.nlm.nih.gov/, UniProt, ENSEMBL backed by the EBI http://www.ebi.ac.uk/ to name just a few), specialized on a subset of specific biological entities (for instance, UniProt focuses on proteins while Entrez Gene focuses on genes). Key identifiers (IDs) uniquely represent biological entities, thus biological entities can be identified by heterogeneous IDs according to the selected database they refer to. Due to their different origins and specificity, switching from an ID to another is possible but not trivial: there could be either no correspondence between them or many-to-many relations. For our purposes, we have chosen EntrezGene IDs and Gene Symbols because of their widespread use and simplicity.

The function *converterIdentifiers* allows the user to map such variety of IDs to a single type. This mapping process, however, may lead to the loss of some nodes (not all identifiers may be recognized) and has an impact on the topology of the network (one ID may correspond to multiple IDs in another annotation or *vice versa*).

```
>  pEntrez <- convertIdentifiers(p, "entrez")
>  pEntrez
"acetylation and deacetylation of rela in nucleus" pathway from BioCarta
Number of nodes     = 8
Number of edges     = 27
Type of identifiers = Entrez Gene
```

```
Retrieved on        = 2011-05-12

>  nodes(pEntrez)
[1] "4792" "5970" "8841" "1387" "2033" "2648" "8850" "9575"

>  pSymbol <- convertIdentifiers(p, "symbol")
>  nodes(pSymbol)
[1] "NFKBIA" "RELA"   "HDAC3"  "CREBBP" "EP300"  "KAT2A"  "KAT2B"  "CLOCK"
```

3.5 Cytoscape Plot

Several pathways have a huge number of nodes and edges, thus there is the need of an efficient system of visualization. To this end *graphite* uses Rcytoscape package to export the network to Cytoscape, see Figures 3.1.

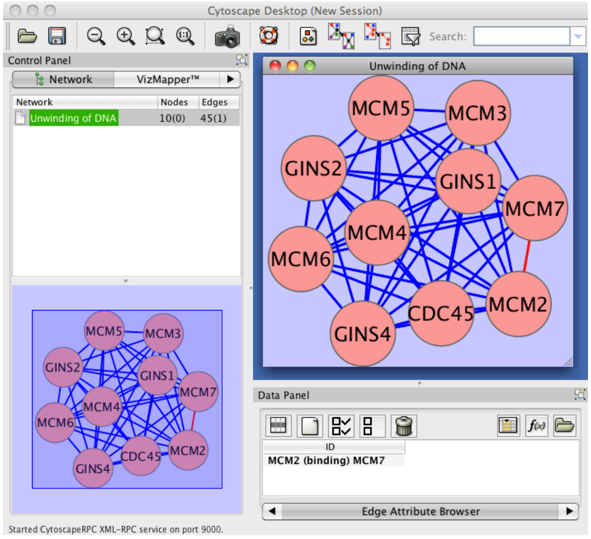


FIGURE 3.1: Screenshot of a *graphite* network imported in Cytoscape using RCytoscape

Cytoscape is a Java based software specifically built to manage biological network complexity and for this reason it is widely used by the biological community (Smoot et al., 2011). The command used to import a *graphite* network in Cytoscape is:

```
> cytoscapePlot(convertIdentifiers(reactome$ Unwinding of DNA , "symbol"))
```

3.6 Topological pathway analysis

graphite gives access to three types of topological pathway analyses recently proposed. More details on the results obtained by these methods are presented in the corresponding R package manuals.

3.6.1 *SPIA*

The analysis with *SPIA* requires the conversion of the networks in a series of adjacency matrices. This conversion is performed by the function *prepareSPIA* that must be executed before the analysis command *runSPIA*. The *SPIA* data will be saved in the current working directory; every time you change it you should re-run *prepareSPIA*. Edges type not included in *SPIA* have been coerced into the admitted *SPIA* types. Compound mediated interactions annotated in *graphite* with "indirect" type are mapped into the *SPIA* edge type "indirect effect" by default set to zero. To use the signal propagated through compounds the user has to type 1 in "indirect effect".

For a detailed description of *SPIA* see Chapter 6.

```
>  prepareSPIA(biocarta[1:2], "biocartaEx")
>  runSPIA(de=DE_Colorectal, all=ALL_Colorectal, "biocartaEx")
Done pathway 1 : acetylation and deacetylation ...
Done pathway 2 : actions of nitric oxide in the ...
...
Name pSize
1 actions of nitric oxide in the heart 43

N tA pPERT pG pGFdr pGFWER Status
12 0.1456427 -0.5368375 0.680 0.3280366 0.4110914 0.6560732 Inhibited

Name pSize
2 acetylation and deacetylation of rela in nucleus  7

N tA pPERT pG pGFdr pGFWER Status
3 0.1527022 -0.3137486 0.903 0.4110914 0.4110914 0.8221828 Inhibited
```

For more details see the *SPIA* package (Draghici et al., 2007; Tarca et al., 2009).

3.6.2 *DEGraph*

DEGraph implements recent hypothesis testing methods which directly assess whether a particular gene network is differentially expressed between two conditions.

```
> library(DEGraph)
Scalable Robust Estimators with High Breakdown Point (version 1.3-01)
> data("Loi2008_DEGraphVignette")
> p <- convertIdentifiers(
                 biocarta[["actions of nitric oxide in the heart"]],
                    "entrez")
> res <- runDEGraph(p, exprLoi2008, classLoi2008)
> res$ 1

$p.value
       T2    T2 (1 Fourier components)
0.4801202  0.4510231

$graph
A graphNEL graph with directed edges
Number of Nodes = 2
Number of Edges = 3

$k [1] 1
```

For more details see the *DEGraph* package (L. Jacob and Dudoit, 2010).

3.6.3 *topologyGSA*

topologyGSA uses graphical models to test the pathway components and to highlight those involved in its deregulation.

```
> library(topologyGSA)
> data(examples)
> p <- convertIdentifiers(kegg[["Fc epsilon RI signaling pathway"]], "symbol")
> runTopologyGSA(p, "var", exp1, exp2, 0.05)

$alpha.obs
```

```
[1] 0.007421451

$cli.moral
$cli.moral[[1]]
[1] "GRB2"

$cli.moral[[2]]
[1] "SYK"   "BTK"   "PLCG2"

$cli.moral[[3]]
[1] "SYK" "LYN" "BTK"

$check
[1] TRUE

$graph
A graphNEL graph with undirected edges
Number of Nodes = 5
Number of Edges = 5

$lambda.obs
[1] 26.02199

$lambda.theo
[1] 18.30704
```

For more details see the *topologyGSA* package (Massa et al., 2010).

The *graphite* package is in continuous development, a new version will be released soon with the possibility to run analysis also with *CliPPER* method (Martini et al., 2013).

Chapter 4

graphite in practice

In this chapter we will provide two practical examples of pathway conversions and some statistics about *graphite*.

We will highlight the innovations provided by our package *graphite*, critically comparing *graphite* with other available existing R/Bioconductor packages.

Finally, we will show a simulation study to demonstrate the efficacy of our signal propagation strategy in terms of topological analyses and an example of topological gene set analysis using benchmark real data.

4.1 *graphite* in numbers

At time of writing, *Graphite* contains more than 1300 human pathways belonging to the four most famous pathway repositories (KEGG, Reactome, NCI and BioCarta). Table 4.1 and Figure 4.1 report respectively pathway summary statistics and nodes/edges distributions for the four pathway databases after the conversion.

Database	N. of pathways	Mean (Median) nodes	Mean (Median) edges
KEGG	232	71.86 (54)	211.12 (75.5)
Reactome	1070	33.22 (14)	780.64 (33)
BioCarta	254	15.18 (14)	36.88 (28)
NCI	177	76.79 (48)	165.18 (81)

Table 4.1: Number of pathways converted to networks with average number of edges and nodes according to the selected database.

Compound-mediated interactions are interactions for which a compound acts as a bridge between two elements. As chemical compounds are not usually measured with high-throughput technologies, they should be removed from the network during the analyses.

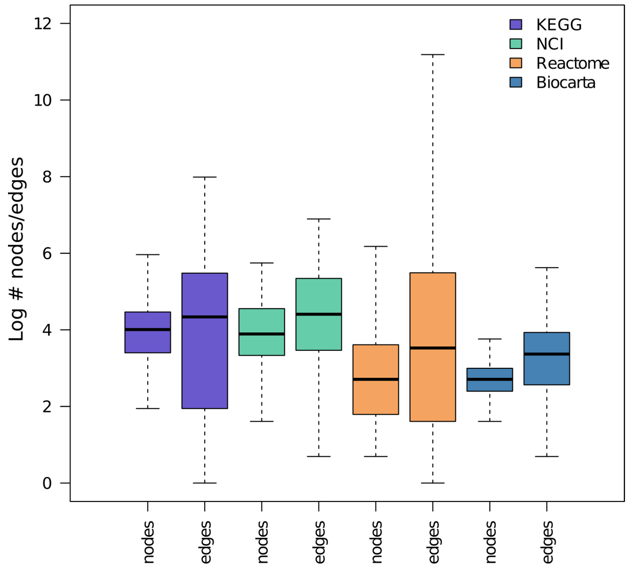


Figure 4.1: Edges and nodes distribution of networks after pathway conversion according to the selected database.

However, the trivial elimination of the compounds will strongly bias the topology interrupting the signals passing through them. Signal propagation, provided by *graphite* package, solve this issue. After parsing all the BioPax and KGML data we obtain compound chains whose length distribution are reported in Table 4.2.

Chain length	KEGG	Reactome	Biocarta	NCI
2	19790	55155	502	2790
3	0	874	9	134
4	0	736	8	11
5	0	140	0	0
6	0	39	0	0
7	0	6	0	0
8	0	17	0	0
9	0	1	0	0

Table 4.2: Frequency of compound chains that we propagate according to different databases.

4.2 *graphite* practical conversion procedures and its competitor

KEGGgraph a competitor of *graphite* providing pathway topology, shows some drawbacks: i) it is specific of KEGG; ii) it does not consider gene connections through chemical compounds; iii) it does not handle differently the diverse groups. In the next sections we will present two examples of pathway conversions, the first one using KGML code, the second one with a BioBax pathway. The first one will be compared with *KEGGgraph* results.

4.2.1 KGML pathway conversion example 1 – Insulin signaling pathway and KEGGgraph comparison

Figure 4.2 represents an example of pathway simplification in which the elimination of compounds leads to an incorrect network topology. An incorrect network topology is exactly what we obtain using KEGGgraph.

Insulin is an hormone controlling the balance between mobilization and storage of energy molecules. Insulin binds the Insulin Receptor (IR) and through phosphorilation of the IRS adaptors is able to recruit and activate PI3K. PI3K is a kinase that converts PIP2 in PIP3 which is a secondary messenger involved in the regulation of various processes. The conversion between PIP3 into PI(3,4)P2 or PI(4,5)P2 operated by phosphatases like SHIP1/2 or PTEN induce a depletion of PIP3 levels and of consequence a reduced activity on its downstream targets Ooms et al. (2009). PIP3 associates with the inner lipid bilayer of the plasma membrane to promote the recruitment of proteins with pleckstrin homology (PH) domains, like PDPK and AKT, which is a crucial mediator of various cell process, such as apoptosis, cell cycle, protein synthesis, regulation of metabolism Ruggero and Sonenberg (2005). Among other functions, AKT activates also the cyclic nucleotide phosphodiesterases (PDEs), that is a group of enzymes able to regulate the localization, duration, and amplitude of the cyclic nucleotides. Signaling PDEs are therefore important regulators of signal transduction mediated by these second messenger molecules Kitamura et al. (1999). In this pathway, PDE, depleting cAMP, indirectly inhibits the PKC mediated phosphorilation, and the activation of LIPE that is a lipase able to mobilize lipid energy stores. PDE acts, in this way, as a anti-lipolytic agents Hołlysz et al. (2011). This hormonal mediated signaling cascade, from the insulin receptor to the inhibition of HSL, involves two compounds (PIP3 and cAMP) crucial for the transduction of the signal.

In panel A of Figure 4.2 we report a part of the insulin signaling pathway of KEGG (hsa4910) that contains three groups OR (PDE3, AKT and PKA), and two compound mediated interactions (through PIP3 and cAMP). This is a clear examples of a signal cascade in which the propagation of the signal through compounds is crucial to keep the whole signaling path.

In panel B we report *graphite* reconstructed signal cascade while in panel C the *KEGGgraph* partially reconstructed signal.

From the *XML* code entry 2 (SKIP) and entry 3 (SHIP) are linked to compound 15 (PIP3) while there is no direct interaction between compound 15 (PIP3) and entry 62 (PDK1/2). This is why *KEGGgraph* misses the signal, while *graphite* captures it by splitting the relation between entry 52 (protein complex P13K) and 62 (PDK1/2) through compound 15 (PIP3) into both 52 to 15 and 15 to 62. This dissection allows the reconstruction of the signal, otherwise impossible.

4.2.2 BioPax pathway conversion example 2 – catalysis and cleavage of Notch 1 by Gamma-Secretase Complex

We selected the reaction 1784.3 from the Reactome pathway called "A third proteolytic cleavage releases NICD". Gamma secretase is a multi-subunit protease complex, itself an integral membrane protein, that cleaves single-pass transmembrane proteins at residues within the transmembrane domain. Here represented the processing of the Notch 1 protein. The Gamma-secretase complex is composed of Presenilin homodimer (PSEN1 variant 1 or 2 or 3 or 4 or 5 and PSEN2 variant 1 or 2), Nicastrin (NCSTN variant 1 or variant 2), APH1 (APH1A or APH1B) and PEN2. Maturation of the Notch receptor involves a cleavage of the protein, the intracellular domain is liberated from the plasma membrane that can enter into the nucleus to engage other DNA-binding proteins regulating gene expression. The cleavage is catalyzed and performed by Gamma-secretase complex.

Figure 4.3 shows Reactome representation of the reactions (Panel A), the BioPax information as it is stored in owl model and in Cytoscape plug-in for BioPax (respectively panel B and C) and the *graphite* final network (panel D). In the *graphite* network the nodes are annotated using the BioPax unificationXRefs informations while edges preserve the type of the reaction annotated the OWL model. Distinction between OR complexes (formed by all the possible variants of each protein) nested inside the AND complex of the Gamma-secretase are topologically preserved in the resulting graph.

4.3 *graphite* for topological analyses

4.3.1 Simulation study: compound propagated signals improve topological analysis

In order to verify our signal propagation strategy we perform a simulation study. Using the insulin signaling pathway of the KEGG database we select as differentially expressed 22 genes lying on the signal paths highlighted in Figure 4.4 A. These genes are connected

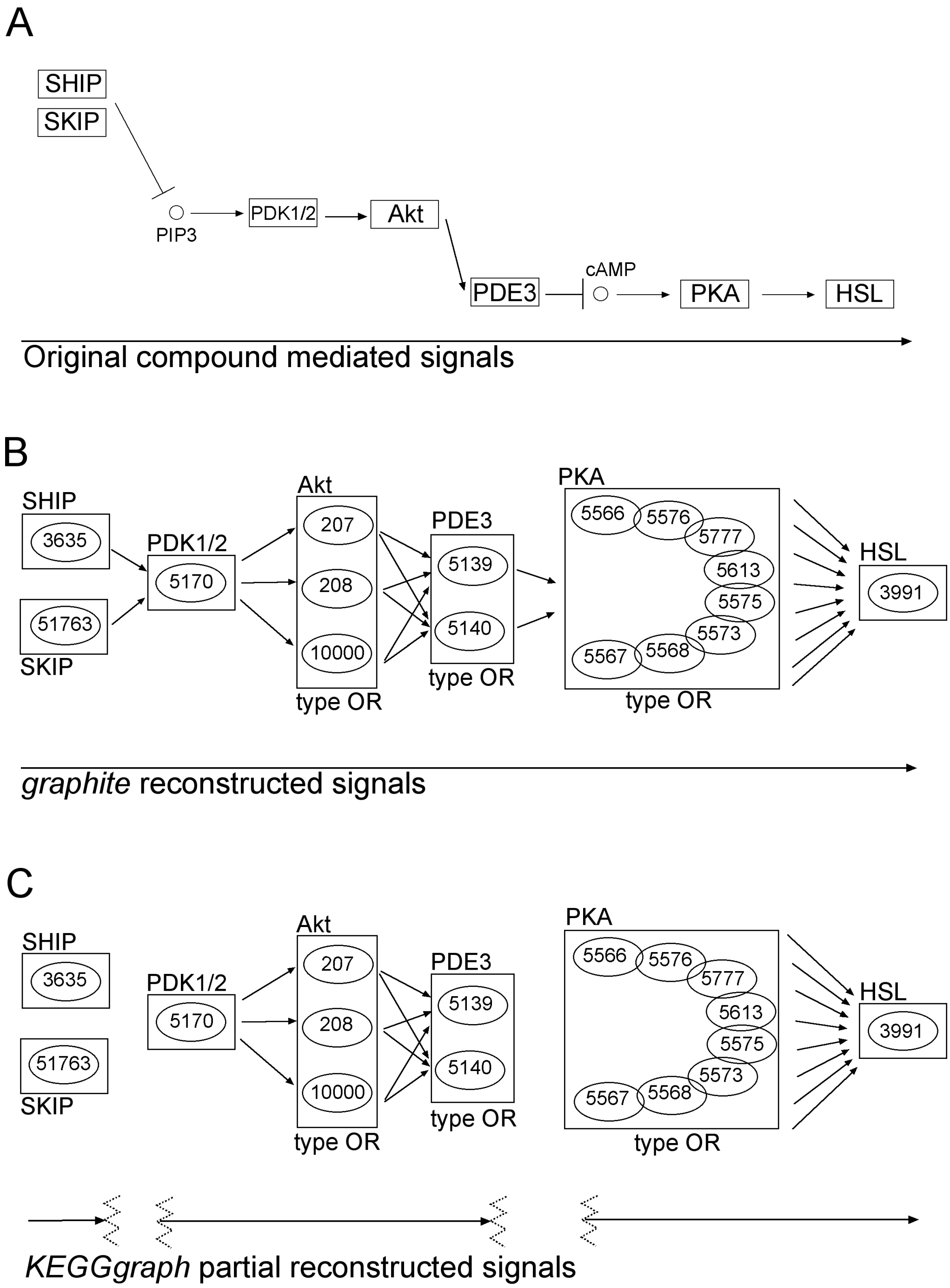


FIGURE 4.2: Differences in signal reconstruction of a selected portion of the insulin signaling pathway of KEGG (hsa04910). Panel A. The original signal cascade. Panel B. *graphite* signal reconstruction through chemical compound propagation. Numbers represent EntrezGene IDs. Panel C. *KEGGgraph* signal reconstruction.

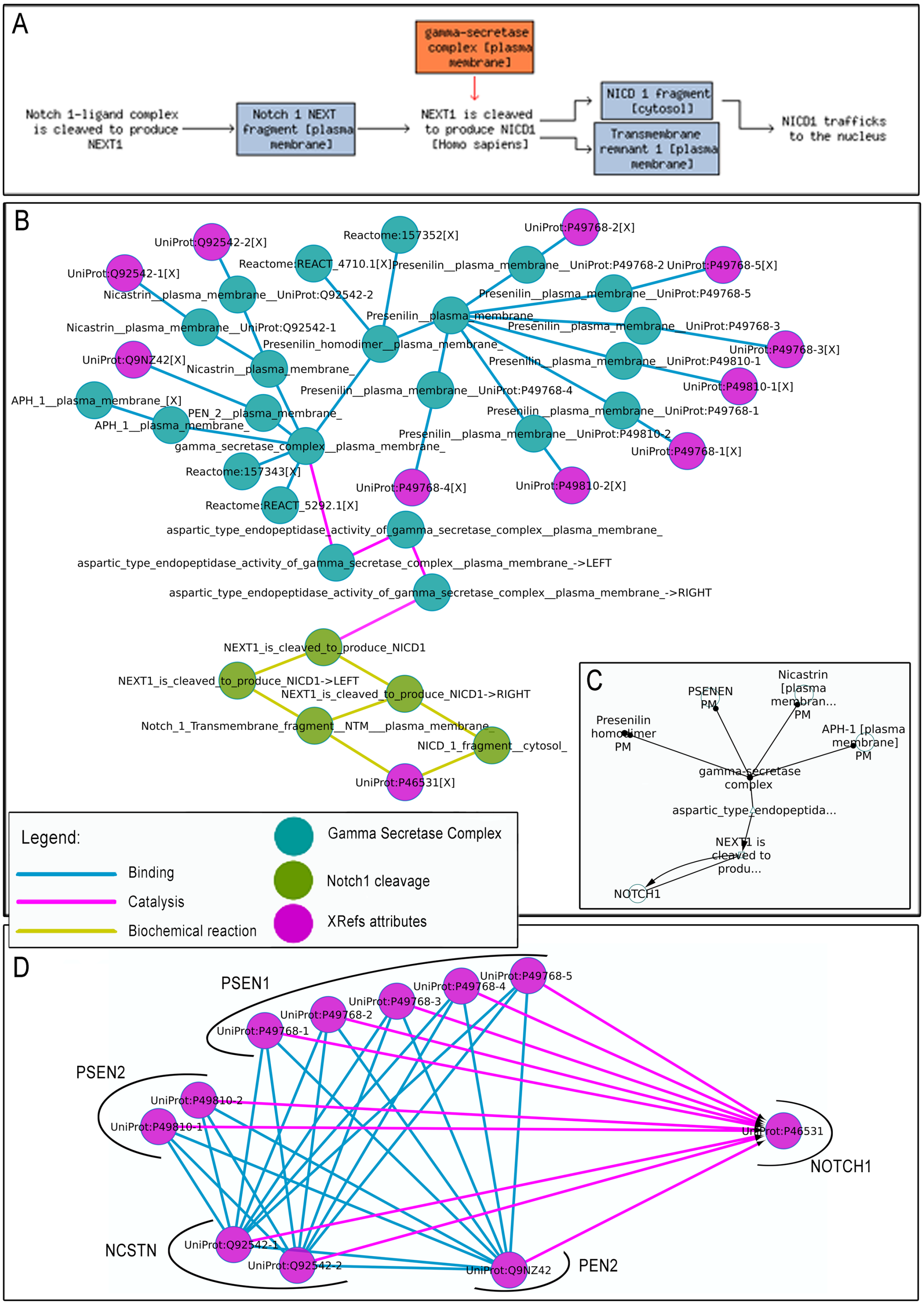

A
gamma-secretase complex [plasma membrane]
Notch 1-ligand complex is cleaved to produce NEXT1
Notch 1 NEXT fragment [plasma membrane]
NEXT1 is cleaved to produce NICD1 [Homo sapiens]
NICD 1 fragment [cytosol]
Transmembrane remnant 1 [plasma membrane]
NICD1 trafficks to the nucleus
B
UniProt:Q92542-2[X]
UniProt:Q92542-1[X]
Reactome:157352[X]
Reactome:REACT_4710.1[X]
UniProt:P49768-2[X]
Presenilin__plasma_membrane__UniProt:P49768-2
UniProt:P49768-5[X]
Nicastrin__plasma_membrane__UniProt:Q92542-2
Presenilin__plasma_membrane__UniProt:P49768-5
Nicastrin__plasma_membrane__UniProt:Q92542-1
Presenilin__plasma_membrane_
Presenilin__plasma_membrane__UniProt:P49768-3
Presenilin_homodimer__plasma_membrane_
UniProt:P49768-3[X]
UniProt:Q9NZ42[X]
Nicastrin__plasma_membrane_
Presenilin__plasma_membrane__UniProt:P49810-1
UniProt:P49810-1[X]
APH_1__plasma_membrane_[X]
Presenilin__plasma_membrane__UniProt:P49768-4
PEN_2__plasma_membrane_
Presenilin__plasma_membrane__UniProt:P49768-1
APH_1__plasma_membrane_
Presenilin__plasma_membrane__UniProt:P49810-2
gamma_secretase_complex__plasma_membrane_
UniProt:P49768-1[X]
Reactome:157343[X]
UniProt:P49768-4[X]
UniProt:P49810-2[X]
Reactome:REACT_5292.1[X]
aspartic_type_endopeptidase_activity_of_gamma_secretase_complex__plasma_membrane_
aspartic_type_endopeptidase_activity_of_gamma_secretase_complex__plasma_membrane_->LEFT
aspartic_type_endopeptidase_activity_of_gamma_secretase_complex__plasma_membrane_->RIGHT
NEXT1_is_cleaved_to_produce_NICD1
NEXT1_is_cleaved_to_produce_NICD1->LEFT
NEXT1_is_cleaved_to_produce_NICD1->RIGHT
Notch_1_Transmembrane_fragment__NTM__plasma_membrane
NICD_1_fragment__cytosol_
UniProt:P46531[X]
Legend:
Binding
Catalysis
Biochemical reaction
Gamma Secretase Complex
Notch1 cleavage
XRefs attributes
C
PSENEN PM
Nicastrin [plasma membran... PM
Presenilin homodimer PM
APH-1 [plasma membrane] PM
gamma-secretase complex
aspartic_type_endopeptida...
NEXT1 is cleaved to produ...
NOTCH1
D
PSEN1
UniProt:P49768-5
UniProt:P49768-4
UniProt:P49768-3
UniProt:P49768-2
UniProt:P49768-1
PSEN2
UniProt:P49810-2
UniProt:P49810-1
UniProt:P46531
NOTCH1
NCSTN
UniProt:Q92542-1
UniProt:Q92542-2
UniProt:Q9NZ42
PEN2

FIGURE 4.3: Catalysis and cleavage of Notch 1 by Gamma Secretase Complex. Reactome representation of the reactions (Panel A), BioPax information as it is stored in owl model and in Cytoscape plug-in BioPax dedicated (respectively panel B and C) and the *graphite* final network (panel D).

if propagation is employed, otherwise they are disconnected (see Figure 4.4 C and D for propagation and non-propagation respectively). We expect that propagation will lead to better results in terms of topological analyses.

Our simulation is based on the following steps: 1) we randomly generate $\mu_{FC} \sim U(2, 10)$; 2) we randomly generate log fold change values (δ_i for $i = 1, \ldots, 22$) of the differentially expressed genes as $\delta_i \sim N(\mu_{FC}, 2)$ (interactions of the signal paths selected are characterized all by activation, thus, fold changes have the same sign); 3) we run the *SPIA* algorithm on the Insulin signaling pathway with and without signal propagations and we take the p-value of the topological analysis (P_{PERT}); 4) we repeat from step 1 10,000 times.

As shown in Figure 4.4 B the distribution of the topological significance p-values in case of signal propagation is shifted towards lower values with respect to the case of non-propagation. Propagation p-value distribution is not only centered on 0.1 (while the one with non-propagation is centered on 0.3) but is also less variable. As expected the same results are obtained simulating negative fold changes (data not shown). This finding demonstrate that compound mediating signal propagation improves topological analyses giving more reliable results.

4.3.2 Example of topological analysis with real data: B-lineage Adult Acute Lymphocytic Leukemia

4.3.2.1 Data

The dataset published by Chiaretti et al. (2005), characterizes gene expression signatures in acute lymphocytic leukemia (ALL) cells associated with known genotypic abnormalities in adult patients. Several distinct genetic mechanisms lead to acute lymphocytic leukemia (ALL) malignant transformations deriving from distinct lymphoid precursor cells that have been committed to either T-lineage or B-lineage differentiation. Chromosome translocations and molecular rearrangements are common events in B-lineage ALL and reflect distinct mechanisms of transformation. The relative frequencies of specific molecular rearrangements differ in children and adults with B-lineage ALL. The BCR breakpoint cluster region and the c-abl oncogene 1 (BCR/ABL) gene rearrangement occurs in about 25% of cases in adult ALL, and much less frequently in pediatric ALL.

Data is available at the Bioconductor site (www.bioconductor.org). Expression values, appropriately normalized according to *rma* and *quantile* normalization, derived from Affymetrix single channel technology, consist of 37 observations from one experimental

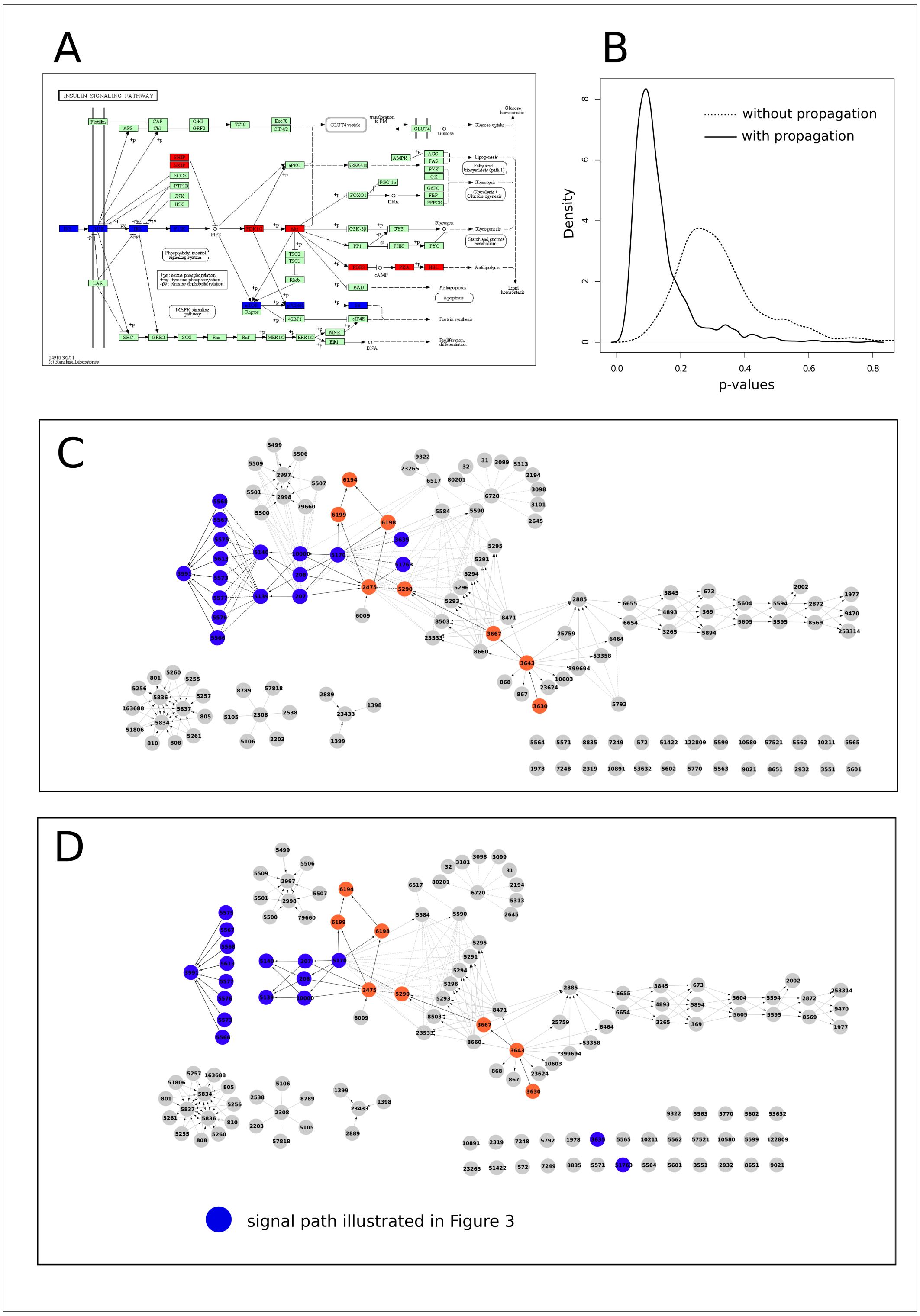


FIGURE 4.4: Results of the simulation study on the Insulin signaling pathway compound mediated signal propagation. Panel A. Signal paths selected to be differentially expressed. Panel B. p-value distribution of the topological analysis *SPIA* (P_{PERT}) with and without propagation. Panel C. *graphite* network obtained from insulin pathway with propagation. Panel D. network obtained from insulin pathway without propagation.

condition ($n_1 = 37$, BCR; presence of BCR/ABL gene rearrangement) and 41 observations from another experimental condition ($n_2 = 41$, NEG; absence of rearrangement). Probes platform have been annotate using EntrezGene custom CDF version 14 (Dai et al., 2005). Given the involvement of BCR and ABL genes in the chimera rearrangement, we expect these genes playing a central role in the gene set analysis; thus, most of the pathways containing BCR and/or ABL genes should be found as significant.

4.3.2.2 Results

We report the results obtained by *SPIA* (Draghici et al., 2007) and *topologyGSA* (Massa et al., 2010) on the *graphite* networks. These statistical tests are based on completely different null hypotheses; while *SPIA* needs the list of differentially expressed genes, *topologyGSA* performs two statistical tests (to compare the mean and the variance of the pathway between two groups) on the entire list of genes belonging to a pathway. Here, differentially expressed genes required for *SPIA* package have been identified using RankProd test (Hong et al., 2006) ($FDR < 0.01$), while the test on the mean has been chosen for *topologyGSA* package.

Table 4.3 and Table 4.4 reports the list of significant pathways identified by the above approaches; pathways marked with $\surd$ are those containing BCR and/or ABL genes. It is interesting to observe that several pathways containing either BCR and ABL genes were identified as deregulated especially with *topologyGSA*. Then, as expected, several additional pathways associated to cancer progression, apoptosis, cell cycle, cell proliferation and inflammation have been selected as significant.

Leaving the comparison between topological analyses aside (because it is out of the scope of the present work), the results testify the feasibility of performing analyses using *graphite* and the ability to obtain reliable results independently of the chosen analysis method. In addition, for the first time, thanks to *graphite* all the topological methods gain the access to pathway repositories previously not considered.

Our results highlight that the hierarchical pathway structure and the reduced dimension of the pathways characterizing respectively the Reactome and Biocarta databases jointly with the specialized cancer pathways of the NCI databases allow the user to have deeper insight into the data.

To highlight the usefulness of topological analysis in the context of transcriptomic data interpretation, we report two *graphite* networks identified as significantly altered in the previous analysis.

Chronic myeloid leukemia pathway includes both genes, BCR and ABL1, and was identified as differentially expressed between BCR/ABL positive and negative patients by *topologyGSA*. Figure 4.6 shows the chronic myeloid leukemia *graphite* network from KEGG database with differentially expressed genes mapped with different colors according to fold change sign. It is interesting to note the presence of several OR groups (e.g. PI3K, AKT, IKK, CBL gene families), single members of which resulted to be differentially expressed. Two clear deregulated paths starting from BCR and ABL1 genes towards apoptosis and NFKB pathways highlight the power of topological analysis to deeper investigate signal cascades within large pathways.

	Name	FDR	Signal	Database	BCR	ABL
1	Leishmaniasis	0.03	Activated	KEGG		
2	Phase 1 - Functionalization of compounds	0.02	Activated	Reactome		
3	Syndecan-4-mediated signaling events	0.00	Activated	NCI		
4	Regulation of RAC1 activity	0.00	Activated	NCI		
5	RAC1 signaling pathway	0.00	Activated	NCI		
6	RhoA signaling pathway	0.00	Activated	NCI		
7	Regulation of RhoA activity	0.00	Activated	NCI		
8	Noncanonical Wnt signaling pathway	0.00	Activated	NCI		
9	Wnt signaling network	0.00	Activated	NCI		
10	BCR signaling pathway	0.00	Inhibited	NCI		
11	IL6-mediated signaling events	0.00	Inhibited	NCI		
12	Hypoxic and oxygen homeostasis regulation of HIF-1-alpha	0.00	Inhibited	NCI		
13	Stabilization and expansion of the E-cadherin adherens junction	0.00	Activated	NCI		
14	E-cadherin signaling in the nascent adherens junction	0.00	Activated	NCI		
15	E-cadherin signaling events	0.00	Activated	NCI		
16	HIF-1-alpha transcription factor network	0.00	Inhibited	NCI		
17	ALK1 signaling events	0.01	Activated	NCI		
18	Canonical Wnt signaling pathway	0.02	Activated	NCI		
19	ALK1 pathway	0.02	Activated	NCI		
20	S1P2 pathway	0.02	Inhibited	NCI		
21	Regulation of nuclear SMAD2/3 signaling	0.02	Activated	NCI		
22	Regulation of cytoplasmic and nuclear SMAD2/3 signaling	0.02	Activated	NCI		
23	TGF-beta receptor signaling	0.02	Activated	NCI		
24	C-MYB transcription factor network	0.02	Activated	NCI		
25	Osteopontin-mediated events	0.02	Inhibited	NCI		
26	Direct p53 effectors	0.02	Inhibited	NCI		
27	Validated transcriptional targets of AP1 family members Fra1 and Fra2	0.03	Activated	NCI		
28	Regulation of nuclear beta catenin signaling and target gene transcription	0.03	Activated	NCI		
29	S1P4 pathway	0.03	Inhibited	NCI		
30	amb2 Integrin signaling	0.03	Activated	NCI		
31	p38 MAPK signaling pathway	0.04	Activated	NCI		
32	Posttranslational regulation of adherens junction stability and disassembly	0.04	Activated	NCI		
33	N-cadherin signaling events	0.04	Activated	NCI		
34	Lissencephaly gene (LIS1) in neuronal migration and development	0.05	Activated	NCI		√
35	C-MYC pathway	0.06	Inhibited	NCI		
36	p53 pathway	0.06	Activated	NCI		

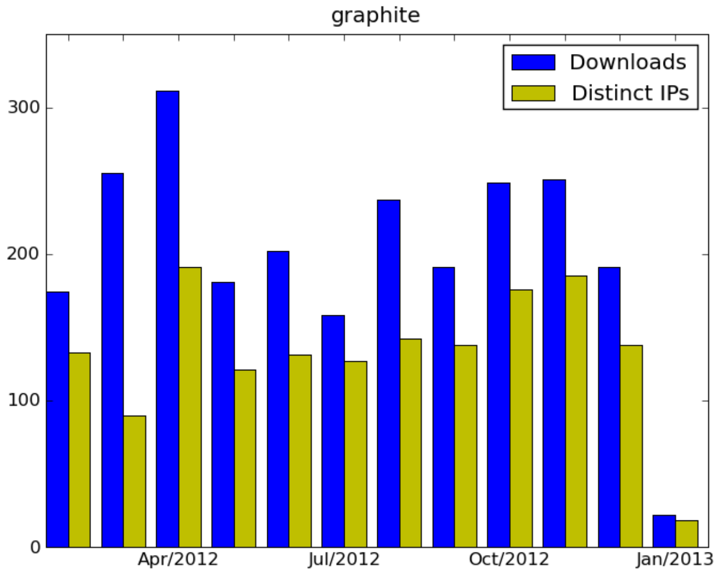
Table 4.3: Pathway analysis performed using *SPIA* statistical test on *graphite* networks.

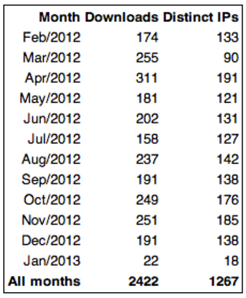
4.4 The first year of *graphite*

graphite has been published the 31 January 2012 (Sales et al., 2012). At the time of writing, 1 year later, more than 1200 different IPs have performed the approximately 2500 downloads, see Figure 4.5.

	Name	FDR	Database	BCR	ABL
1	CDO in myogenesis	0.00	Reactome		✓
2	Regulation of cytoskeletal remodeling and cell spreading by IPP complex components	0.00	Reactome		
3	Role of Abl in Robo-Slit signaling	0.00	Reactome		✓
4	NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10	0.01	Reactome		
5	TNF signaling	0.01	Reactome		
6	G1 Phase	0.02	Reactome		
7	mTOR signalling	0.02	Reactome		
8	PI3K Cascade	0.02	Reactome		
9	Cyclin D associated events in G1	0.02	Reactome		
10	PI-3K cascade	0.03	Reactome		
11	E2F mediated regulation of DNA replication	0.04	Reactome		
12	Cyclin A/B1 associated events during G2/M transition	0.04	Reactome		
13	Intrinsic Pathway for Apoptosis	0.04	Reactome		
14	Extrinsic Pathway for Apoptosis	0.05	Reactome		
15	Lissencephaly gene (LIS1) in neuronal migration and development	0.00	NCI		✓
16	ErbB4 signaling events	0.01	NCI		
17	Regulation of retinoblastoma protein	0.00	NCI		✓
18	Canonical NF-kappaB pathway	0.01	NCI		
19	p73 transcription factor network	0.01	NCI		✓
20	Atypical NF-kappaB pathway	0.02	NCI		
21	Neurotrophic factor-mediated Trk receptor signaling	0.00	NCI		✓
22	Pathogenic Escherichia coli infection	0.00	KEGG		✓
23	Chronic myeloid leukeamia	0.00	KEGG	✓	✓
24	Cell cycle	0.0	KEGG		✓
25	Axon guidance	0.00	KEGG		✓
26	Neurotrophin signaling pathway	0.00	KEGG		✓
27	mtor signaling pathway	0.01	Biocarta		
28	nf-kb signaling pathway	0.01	Biocarta		
29	tnf/stress related signaling	0.02	Biocarta		
30	p53 signaling pathway	0.03	Biocarta		
31	tnfr1 signaling pathway	0.02	Biocarta		
32	integrin signaling pathway	0.02	Biocarta		
33	erk and pi-3 kinase are necessary for collagen binding in corneal epithelia	0.02	Biocarta		
34	rb tumor suppressor/checkpoint signaling in response to dna damage	0.03	Biocarta		
35	egf signaling pathway	0.04	Biocarta		
36	tgf beta signaling pathway	0.04	Biocarta		
37	role of mitochondria in apoptotic signaling	0.04	Biocarta		
38	inhibition of cellular proliferation by gleevec	0.04	Biocarta		
39	atm signaling pathway	0.05	Biocarta		✓
40	influence of ras and rho proteins on g1 to s transition	0.05	Biocarta		

TABLE 4.4: Pathway analysis performed using *topologyGSA* statistical test on *graphite* networks.





Month	Downloads	Distinct IPs
Feb/2012	174	133
Mar/2012	255	90
Apr/2012	311	191
May/2012	181	121
Jun/2012	202	131
Jul/2012	158	127
Aug/2012	237	142
Sep/2012	191	138
Oct/2012	249	176
Nov/2012	251	185
Dec/2012	191	138
Jan/2013	22	18
All months	**2422**	**1267**

FIGURE 4.5: Statistics of Downloads provided by Bioconductor website.

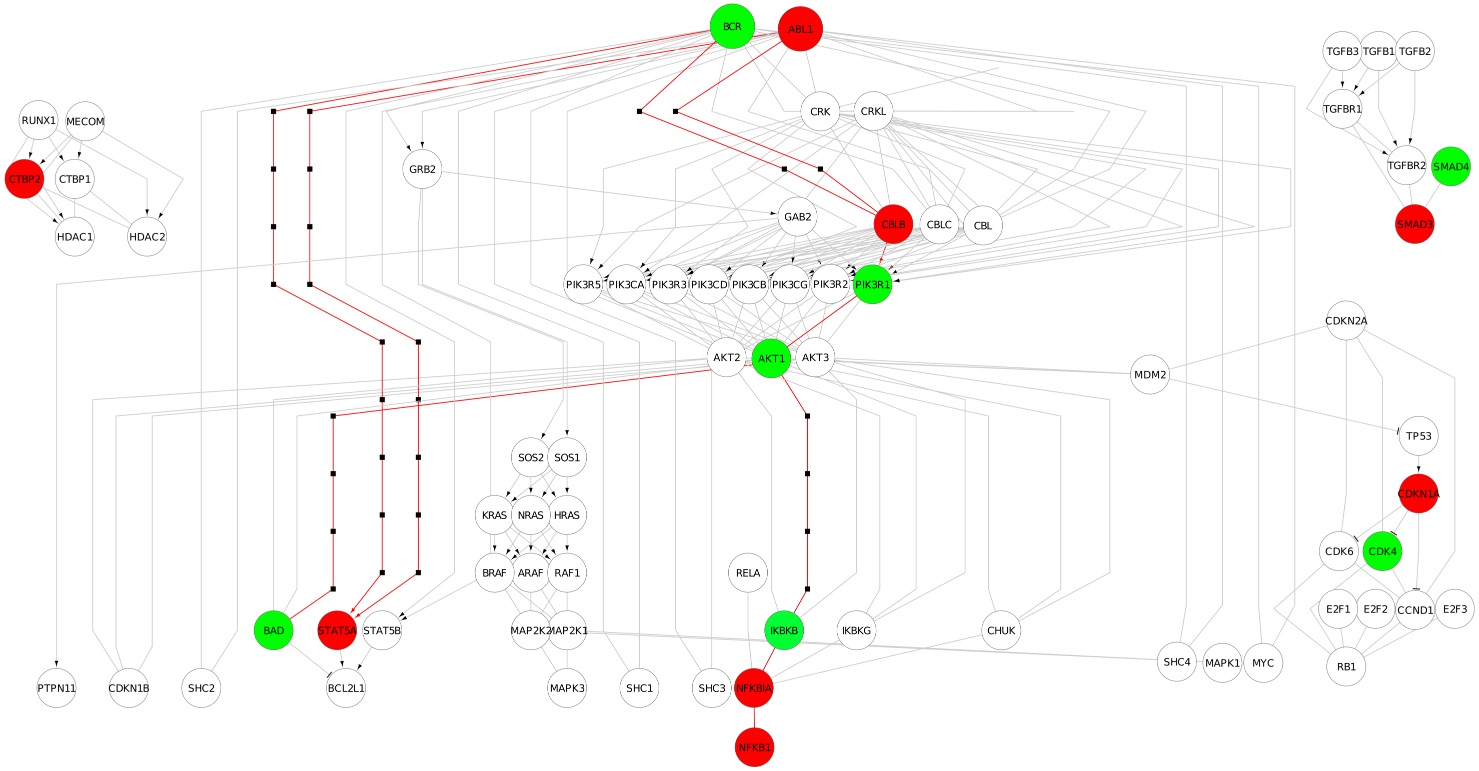


Figure 4.6: Visualization of the chronic myeloid leukemia network of *graphite*, that contain BCR and ABL1 genes. Colors represent up or down regulated genes between positive and negative BCR/ABL rearrangement.

Chapter 5

miRNAs Specificities in Epithelial Ovarian Cancer histotypes

Since the introduction of platinum compounds in the adjuvant treatment of EOC in the 1980s, there has been little improvement in treatment outcome. Only Paclitaxel and, more recently, Bevacizumab, have slightly, but by no means dramatically, improved the long-term survival of patients. One of the possible explanations of the failure of virtually all clinical trials with new drugs, is that most of them did not consider histotypes as an inclusion criterion, nor were patients stratified based on histotype. It has been long known that different histotypes are characterized by differences in risk factors (Risch et al., 1996), grades of nuclear atypia and dissemination, frequency of mutations of genes related to cell proliferation (Madore et al., 2010), responses to standard platinum-based chemotherapy (Itamochi et al., 2002; Polverino et al., 2005; Alexandre et al., 2010) and gene expression profiles (Marquez et al., 2005). Despite this heterogeneity, all stage EOC histotypes are being treated equally using surgical debulking and carboplatin-paclitaxel chemotherapy.

miRNAs are highly tissue-specific and have recently been identified as attractive targets for therapeutic intervention. During my Ph.D. I investigated the expression and the role of miRNAs in stage I EOC histotype, with the aims to better understand the pathology and to suggest new putative targets for histotype-specific treatments.

5.1 Tissue sample collection

We analysed a collection of 257 snap frozen tumour biopsies obtained from primary surgery on stage I EOC patients naïve to chemotherapy, gathered together from three independent tumour tissue collections (A,B and C).

Collection A was composed of 40 frozen biopsies belonging to a tissue bank containing 300 frozen samples collected between March 2003 and January 2011 at the "A. Nocivelli" Institute for Molecular Medicine, Division of Gynaecologic Oncology, University of Brescia, Italy. Samples were recovered from patients who underwent surgery for EOC at the Obstetrics and Gynaecology Department, Spedali Civili, Brescia, Italy.

Collection B was composed of 167 tissue samples from a frozen tissue bank containing more than 1600 samples collected between September 1992 and March 2010, located at the Department of Oncology, "Mario Negri" Institute, Milano, Italy. Biopsies were collected from patients who underwent surgery for EOC at the Obstetrics and Gynaecology Department, San Gerardo Hospital, Monza, Italy.

Collection C was composed of 50 biopsies belonging to a tissues collection containing 600 frozen samples that were collected between January 1992 and December 2005, and available at the Department of Gynaecology-Oncology, University of Torino, Torino, Italy.

Patients underwent radical surgical tumor debulking and a complete staging procedure according to the International Federation of Gynaecological and Obstetrics criteria (FIGO) (Trimbos et al., 2003). A written informed consent was obtained from all the patients enrolled in the study and the local scientific ethical committees approved the collection and the use of tumor samples. Tumor grade and histological type were determined following World Health Organisation (WHO) standards. The tumor content of the specimens was assessed by haematoxylin and eosin staining to check epithelial purity by the respective pathology units. Only specimens containing more than 70% of epithelial tumor cells were used. Tumor tissue samples, collected at the time of surgery, were identified, sharp dissected and snap frozen in liquid nitrogen within 15 min from resection and then stored at -80C. Clinical and anatomo-pathological patient information was registered, and follow-up data were obtained from periodic gynecological and oncological check-ups.

Table 5.1 shows the clinical and histo-patological distribution of patients involved in this study. Median ages in the three collections were similar, as well as the distribution between histotypes and grades of nuclear differentiation. As expected, in the light of the good prognosis of stage I EOC, the recurrence rate was low around 20%.

Both univariate and multivariate analyses did not reveal any difference in survival rate between different histotypes, confirming results published in literature. As expected, the most significant prognostic feature is the grade of tumors: with increasing grade at time of diagnosis decreases the survival of patients (see Appendix B).

	Training set			Validation set		
Annotations	Collection B1 143 patients	Collection A 40 patients	Total	Collection B2 24 patients	Collection C 50 patients	Total
Clear Cell	23	9	32 (17.5 %)	-	6	6 (8.2 %)
Endometrioid	36	14	50 (27.3 %)	7	20	27 (36.4 %)
Mucinous	41	10	51 (27.9 %)	7	9	16 (21.6 %)
Serous	43	7	50 (27.3 %)	10	15	25 (33.8 %)
Borderline	22	16	38 (20.8 %)	12	3	15 (20.3 %)
1	28	6	34 (18.6 %)	1	21	22 (29.7 %)
2	41	7	48 (26.2 %)	5	10	15 (20.3 %)
3	52	11	63 (34.4 %)	6	16	22 (29.7 %)
A	47	26	73 (39.9 %)	8	21	29 (39.2 %)
B	8	4	12 (6.5 %)		5	5 (6.7 %)
C	88	10	98 (53.6 %)	16	21	37 (50 %)
Unknown	-	-	0	-	3	3 (4.1 %)
Age Median years(range)			51 (16-87)			57 (16-81)

Table 5.1: Main characteristics of patients and tissue samples.

As a general comment, these data confirm that our cohort of patients is consistent with data reported in literature for stage I EOC and this is conducive to downstream analysis.

The entire cohort of patients has been subdivided into a training set (n=183) and a validation set (n=74). In order to have similar histotype and grade proportions between training and validation sets, a subset of collection B (B2) has been dedicated to the validation set with collection C (Table 5.1). The training set was used to: (i) generate miRNA expression measurements and marker identification procedure (A+B1); (ii) integrate miRNA profiles with gene expression data in a subset of patients, and (iii) validate gene and miRNA signature by qRT-PCR. The validation set was used only to re-confirm by qRT-PCR technology miRNA markers and their putative targets previously identified by the analyses of the training set.

5.2 miRNA microarray experiments and analyses

To generate the entire miRNA landscape, our cohort of 183 patients with stage I EOC (training set) have been profiled using Agilent array technology. Experiments have been performed at ”Mario Negri” Institute by the Oncology Group of Maurizio D’Incalci and all the methods to extract DNA, perform hybridization and fluorescence quantification are deeply described in Appendix A

Raw data, submitted to ArrayExpress (series number E-MTAB-1067), were pre-processed to filter out those probes with more than 40% of measurements below the signal-to-noise threshold. Pre-processed data were normalized using quantile method (Bolstad et al., 2003). After normalization, ComBat algorithm (Walker et al., 2008) was used to adjust for time-batch effects.

Figure 5.1 shows the heat-map of the expression value of the 250 miRNAs obtained after the filtering process. Hierarchical cluster analysis was performed using Euclidean distances and complete linkage. On a global scale, a large part of the entire set of miRNAs is similar across samples and it does not help to separate samples by histotype and grade. Then, to identify histotypes-specific miRNAs, the expression levels have been analyzed. Empirical Bayes test, implemented in *Limma* Bioconductor package (Smyth, 2005; Gentleman, 2005), has been used to identify differentially expressed miRNAs among histotypes. False discovery rate (FDR; the expected number of false positives in the list of differentially expressed; Reiner et al. (2003)) was used to assess for the multiple testing using $FDR \leq 0.01$. The complete list of differentially expressed miRNAs of each histotype comparison can be found at Appendix B.

It is noteworthy that among all the different comparisons, the mucinous histotype showed the highest number of differentially expressed features when compared with the other histotypes (Figure 5.1 B). If we consider the number of differentially expressed miRNAs a measure of the difference between histotypes, our results support the hypothesis that the mucinous histotype is markedly different from all the others.

Although all the miRNAs reported above have significant adjusted p-values ($FDR \leq 0.01$), due to patient variability, only some of them can be exploited as histotype-specific markers. Histotype-specific markers are defined as miRNAs with the capability of almost perfectly predicting the histotypes of independent samples.Then, with the intention of identifying miRNA markers among those differentially expressed, we performed a resampling strategy.

5.3 Resampling strategy to identify marker miRNAs

The power of a statistical test (the probability of rejecting the null hypothesis when it is effectively false) is dependent from the sample size; higher the sample size, stronger the statistical power. Then, in case of large sample size, although significant several differentially expressed miRNAs can be characterized by patient variability within the same histotype. On the other hand a miRNA is defined as a histotype-specific marker if its expression profile is sufficient to predict the histotype class in an unknown sample.

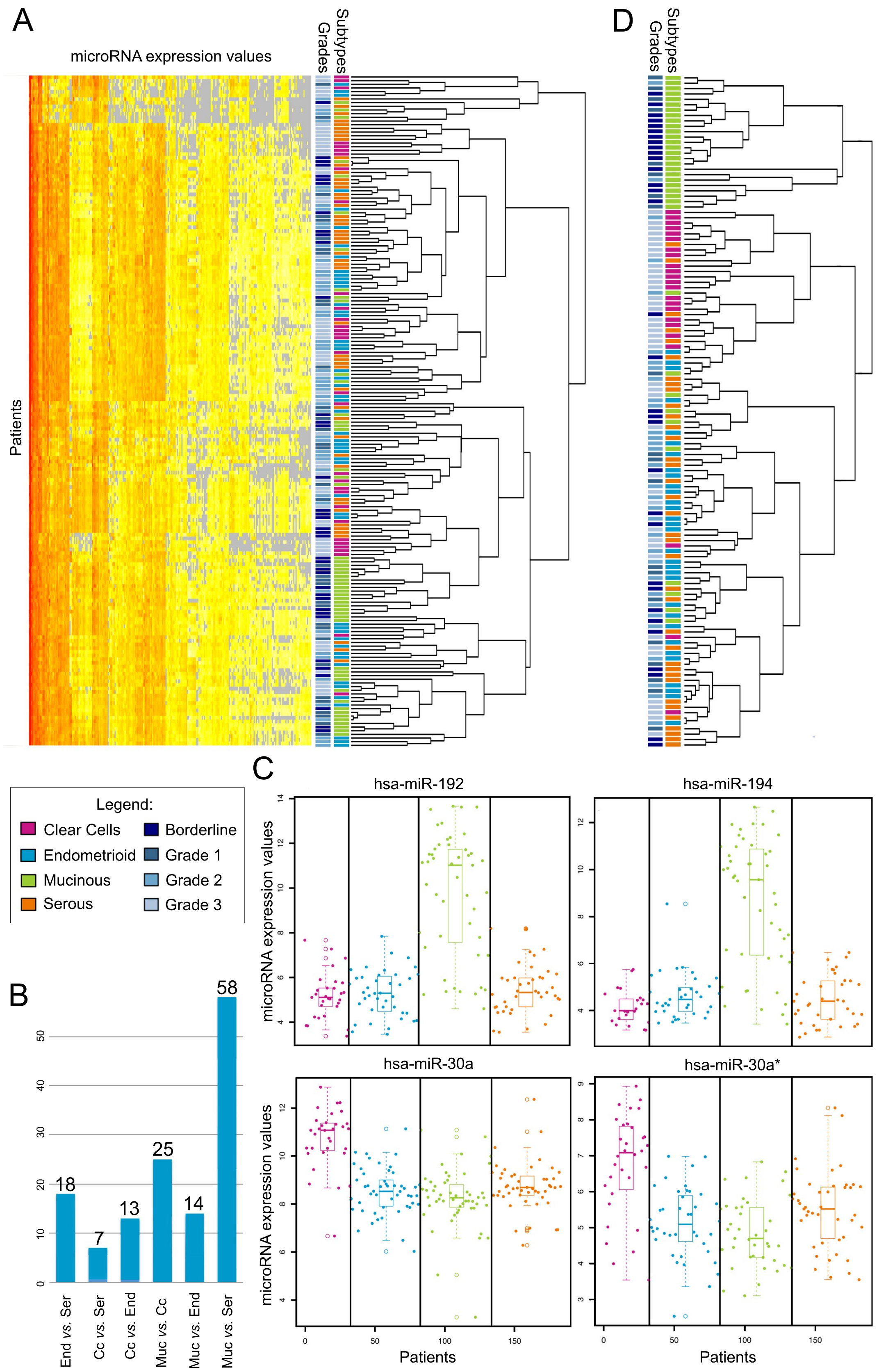


FIGURE 5.1: Microarray Data Analysis. Panel A: Heat-map with two-way cluster analysis on miRNA and samples using all the expression profiles. Panel B: Bar-plot of the number of differentially expressed genes across histotype comparisons. Panel C: Scatterplots and box-plots of patients miRNA marker expression values divided by histological histotypes. Panel D: Sample cluster analysis obtained using only the four histotype-specific miRNA markers (miR-194, miR-192, miR-30, miR-30a). Grade and histotypes are reported in different colors.

Then with the intention to identify miRNA markers among those differentially expressed we perform a resampling strategy based on the following steps:

- Random selection (without replacement) of a subset of patients within each histotype. The dimension of the subset has been set to 60% of the histotype sample size.
- Identification of differentially expressed miRNAs using empirical Bayes statistical test, with a FDR threshold of 0.05.
- Define S_ij a vector of dichotomous values with i=1,...,N, and j=1,...,B where N is the total number of miRNAs, and B the total number of resampling; 1 for differentially express miRNA and 0 for equally expressed.
- Repeat from step 1, B times.
- Define the resampling score RS as:

$$RS_i \sum_{j=1}^{B} S_j$$

$$with 0 \leq RS_i \geq B$$

here B was set to 500

Then, five hundred subsets of the original 183 samples were randomly selected and used to identify histotype-specific differentially expressed miRNAs. The resampling score ($0 \leq RS \geq B$) is the number of times that a miRNA is identified as differentially expressed in the 500 analysis runs. miRNAs with the highest RS (RS=500) has been considered as marker miRNA. The procedure is implemented using the R programming language (R version 2.14) (Team, 2010), and the BioConductor software suite (version 2.9) (Gentleman et al., 2004). The complete list of differentially expressed miRNAs on each histotype comparison ordered by resampling score can be found at Appendix B. In order to evaluate the significance of RS score we used a permutational-based approach. Randomly permuting sample labels each run we performed our resampling strategy. Then for each random permutation we have an simulation resampling score (SRS). We set to 1000 the number of permutation. Then we take the max SRS for each gene in the 1000 SRS. As reported in Table 5.2 the SRSs are always much lower than the observed RS, highlighting the robustness of our signature.

Only ten miRNAs were found commonly deregulated across all possible comparisons of histotypes, and only three miRNAs reached the maximum score (RS=500): miR-192 and miR-194 were highly expressed in the mucinous histotype, and miR-30a was highly

microRNAs	Class comparison	$\log_2$(FC)	adj.P.Val	max SRS	RS
Mucinous					
hsa-miR-192	Muc vs. Cc	4.58	2.36E-19	28	500
	Muc vs. End	4.46	1.80E-21	29	500
	Muc vs. Ser	4.29	1.93E-20	26	500
hsa-miR-194	Muc vs. Cc	4.49	1.27E-16	15	500
	Muc vs. End	4.01	6.09E-17	55	500
	Muc vs. Ser	4.15	2.75E-18	122	500
hsa-miR-338-3p	Muc vs. Cc	1.75	8.35E-06	37	490
	Muc vs. End	1.80	1.16E-06	13	491
	Muc vs. Ser	1.27	0.000384	12	422
hsa-miR-1274a	Muc vs. Cc	-1.62	2.65E-06	76	477
	Muc vs. End	-0.92	0.007193	101	168
	Muc vs. Ser	-0.82	0.006782	95	224
Clear cell					
hsa-miR-30a	Cc vs. End	2.18	6.94E-13	112	500
	Cc vs. Muc	2.47	1.27E-16	35	500
	Cc vs. Ser	1.99	8.08E-11	11	500
hsa-miR-30a*	Cc vs. End	1.62	4.67E-07	99	494
	Cc vs. Muc	1.91	8.94E-10	178	500
	Cc vs. Ser	1.28	8.72E-05	19	408
hsa-miR-193b	Cc vs. End	1.15	7.86E-05	83	429
	Cc vs. Muc	0.93	0.001861	41	288
	Cc vs. Ser	0.85	0.009579	58	175
Serous					
hsa-miR-34b*	Ser vs. Cc	2.28	2.49E-05	18	484
	Ser vs. End	1.34	0.006583	29	189
	Ser vs. Muc	1.91	2.99E-05	52	482
hsa-miR-575	Ser vs. Cc	-1.49	0.009579	53	159
	Ser vs. End	-1.44	0.003786	71	235
	Ser vs. Muc	-1.55	0.000521	79	400
hsa-miR-29b	Ser vs. Cc	1.25	0.009579	63	184
	Ser vs. End	1.52	0.000286	112	364
	Ser vs. Muc	1.31	0.000521	137	404

Table 5.2: Selection of most differentially expressed miRNAs across histotypes. Mucinous (Muc), Clear cell (Cc), Serous (Ser), Endometrioid (End), Fold Change is in log2 scale (log2(Fold Change)), P-value adjusted for multiple testing (adj.P.Val), simulation resampling score (SRS), resampling score (RS).

expressed in clear cell EOCs (Table 5.2). Thus, we reasoned that miR-192 and miR-194 could be considered markers for the mucinous EOCs and miR-30a a marker for the clear cell histotype. Although characterized by a high resampling score, miR-30a* (RS=494 for Cc vs. End; RS=500 for Cc vs. Muc; RS=408 for Cc vs. Ser) does not reach the maximum RS in all comparisons. However, due to its physical association with miR-30a, it has been included in the list of miRNAs selected for downstream validations. In Figure 5.1 C we report the distribution of miRNA expression levels of miR-192, miR-194, miR-30a and miR-30a* in the four histotypes. Box-plots clearly show that the median expression values for miR-192 and miR-194 are significantly higher in the mucinous compared to the other histotypes, while the median expression values of miR-30a and miR-30a* are higher in the clear cell compared to the others. Otherwise, no miRNA can be identified as potential marker for endometrioid or serous EOC histotypes.

We next used the expression profiles of the four markers to re-cluster the entire set of patients, and Figure 5.1 D shows the clustering obtained. The classification of histotypes dramatically improves, separating mucinous and clear cell histotypes from the others, while endometrioid and serous histotypes generate a single and heterogeneous cluster. Interestingly, the mucinous cluster in Figure 5.1 D is preferentially characterized by the presence of low-grade patients (Grade 1 and Borderline). Given this evidence, we investigated the presence of potential variability in expression level across grades within histotypes. We found that, although not significant, miR-192 and miR-194 expression levels slowly decrease with the increasing grade and that these decrease follows the decreasing of survival of mucinous patients through the increasing grade (Figure 5.2).

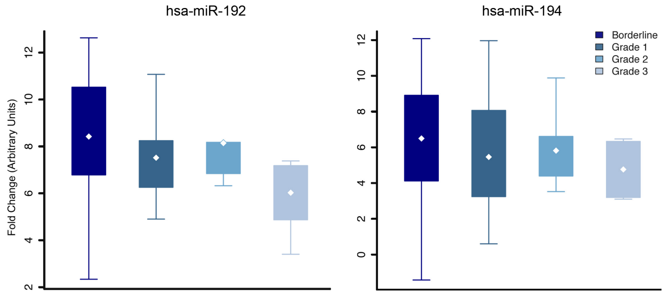


Figure 5.2: Box plots depicting the median plus/minus the IQ range of expression values for miR-192 and miR-194 mucinous histotypes stratified according to their grade in the entire cohort of patients (n=257).

5.4 Validation of markers using qRT-PCR

To assess the robustness and the reproducibility of the array signature identified so far, we measured by qRT-PCR the expression levels of miRNA markers in the training set and then in an independent validation set. To avoid potential errors due to batch effects, new batches of snap frozen material for the entire cohort of samples were used.

Box-plots in Figure 5.3 depict the expression values (measured as fluorescent intensity signals normalized) of miR-192, miR-194, miR-30a and miR-30a* in the training set, stratified for their histological histotypes. For each class comparison we reported in Table 5.3 the fold changes, as $\log_2$(class 1/class 2) and the p-values.

Within the training set, data shows that miR-192 and 194 are roughly 5 to 8 folds over-expressed in mucinous compared to the other histotypes and miR-30a and miR-30a* mirror the same trend, being 5 to 7 folds over-expressed in the clear cell compared to

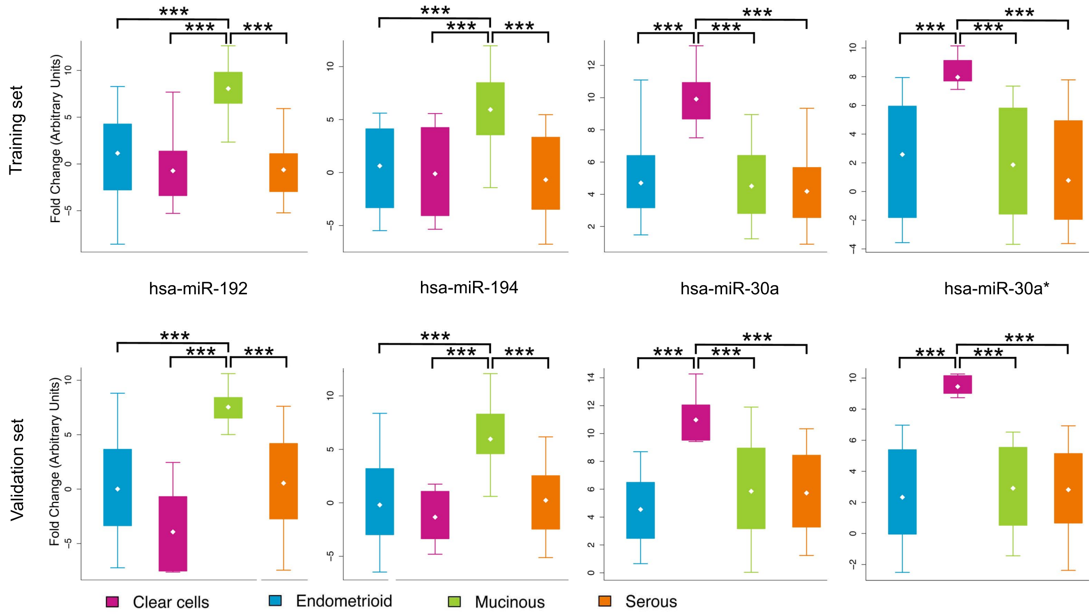


FIGURE 5.3: Box plots depicting the qRT-PCR expression values (measured as fluorescent intensity signals normalized) of miR192, miR-194, miR-30a and miR-30a* in training and validation sets. Mean are indicated with a white dot in the middle of box plot.

		Training set		Validation set	
microRNAs	Class comparison	log$_2$(FC)	p-value	log$_2$(FC)	p-value
Mucinous					
hsa-miR-192	Muc vs. Cc	8.79	1.02E-22	11.46	5.36E-05
	Muc vs. End	6.90	6.22E-17	7.52	2.54E-07
	Muc vs. Ser	8.69	1.02E-28	6.98	8.59E-08
hsa-miR-194	Muc vs. Cc	6.05	3.63E-11	7.30	0.0003
	Muc vs. End	5.31	1.32E-11	6.16	2.42E-06
	Muc vs. Ser	6.62	7.18E-17	5.73	7.33E-06
Clear cell					
hsa-miR-30a	Cc vs. End	5.21	7.47E-20	6.43	1.80E-06
	Cc vs. Muc	5.40	1.88E-20	5.12	0.0045
	Cc vs. Ser	5.73	3.81E-22	5.25	5.16E-05
hsa-miR-30a*	Cc vs. End	5.38	3.44E-09	7.12	1.80E-06
	Cc vs. Muc	6.10	1.37E-11	6.54	2.68E-05
	Cc vs. Ser	7.19	1.96E-15	6.63	2.72E-06

TABLE 5.3: qRT-PCR analysis of miRNAs for both training and validation set. The logarithmic fold change of average expression value (measured as fluorescent intensity signals normalized) of class 1 versus average expression value of class 2 (log2(FC)) of selected miRNAs. p-values were considered as significant when lower than 0.05. Mucinous (Muc), Clear cell (Cc), Serous (Ser), Endometrioid (End).

the other histotypes. The statistical analysis confirmed these differences with adequate significance (p $\leq$ 0.001). Two-sided student's t-test (for training set) and Wilcoxon test (for validation set) were used to verify among groups mean differences. Differences with p-value $\leq$ 0.05 were considered statistically significant. In conclusion, differences measured by qRT-PCR coherently mirrored those previously reported by array technology.

Expression values of selected miRNAs were evaluated also in an independent validation set. Data confirmed miR-192 and miR-194 as up regulated in the mucinous compared to the other histotypes, as well as miR-30a and miR-30a* up regulated in the clear cell subgroup, with p-value lower than 0.005.

5.5 The miRNA markers of EOC histotypes: the importance of results.

The results of the analysis of histotype in stage I EOC presented above indicate an unambiguous miRNA signature for clear cell (high levels of miR-30a and miR-30a*) and mucinous histotypes (high levels of miR-192 and miR-194). This is particularly interesting as these two histotypes, although less frequently found at advanced stages compared to serous EOCs, are rarely curable, showing a low response rate to standard chemotherapy, unlike serous and endometrioid histotypes.

5.5.1 Clear cell histotype markers: miR-30a and miR-30a*

miR-30a and miR-30a* have been identified as markers of the clear cell histotype. Both are 5-fold more highly expressed in clear cell histotype than in the others. miR-30a* has hitherto not been well characterized. In contrast, miR-30a is known to negatively regulate Beclin-1, a positive regulator of the autophagy pathway (Zhu et al., 2009). Autophagy is a tightly regulated catabolic process considered a key pathways in cancer, with a pivotal role for the balance between protein degradation and synthesis, between non-apoptotic programmed cell death and the cell growth (Kreuzaler and Watson, 2012). Since the role of autophagy in carcinogenesis and tumour progression has not been fully elucidated, the potential consequences of these data are not clear. Nevertheless, these findings hint tantalisingly at a possible role of autophagy in determining the sensitivity and resistance to clear cell EOC therapy.

5.5.2 Mucinous histotype markers: miR-192 and miR-194

miR-192 and miR-194 were found to be 5-fold more highly expressed in mucinous than in the other histotypes. miR-192 and miR-194 are co-localized in a miRNA cluster located in chromosome 11. Low levels of miR-192/194 are known to be associated with a more tumorigenic status in a panel of solid tumors (Braun et al., 2008; Hino et al., 2008; Ma et al., 2011; Meng et al., 2010; Song et al., 2008). Interestingly, these miRNAs are considered markers in colon and gastric tissues (Hino et al., 2008; Meng et al., 2010) and have a primary role in tumors of these tissues. The molecular similarity between mucinous EOC and the colon and gastric environment tentatively supports the idea that mucinous ovarian cancer should not be treated with the same regimes used for the other ovarian cancer histotypes. There are some reported common pathological features of mucinous ovarian cancer and colorectal cancer such as a high frequency of microsatellite instability and of K-RAS mutations (Cheng et al., 2009), which reinforce the notion that these tumors are not only morphologically similar, but also biologically.

Chapter 6

The Integration of miRNA and Gene Expression

In the last years, genome-wide expression studies of genes and miRNAs have given a strong impulse in the comprehension of the regulatory mechanisms involved in cancer diseases. Moreover, it has been increasingly clear that the integration of different *omic* data, although challenging, is a successful approach to have a wider perspective of the complexity of the system. In the previous chapter, I analyzed a set of miRNA expression profiles identifying a subtype-specific signature. Although highly interesting from a clinical and biological point of view, a better comprehension of the regulatory circuits in which these miRNA are involved, can be obtained integrating gene and miRNAs expression data. In this perspective, in collaboration with the "Mario Negri" Institute, the gene expression of a subset of patients of the training set have been profiled using microarray. In the following Sections the integration of miRNA and gene expression data is introduced and the results obtained on EOC samples have been discussed.

6.1 Gene and miRNA expressions integrated analyses: an introduction

6.1.1 Classical approach

In silico target identification is based on (i) sequence similarity search, possibly considering target site evolutionary conservation and (ii) thermodynamic stability. However, it is known that the results of target prediction algorithms are characterized by very low specificity (Alexiou et al., 2009). This is caused both by the limited comprehension of the molecular basis of miRNA-target pairing and by the context-dependency of post-transcriptional regulation due to the cooperative interactions of different miRNAs. The integration of target predictions with miRNA and gene expression profiles using correlation measures has been recently proposed to improve the detection of functional miRNA-target relationships (Sales et al., 2010; Bisognin et al., 2012).

6.1.2 A new appraoch: gaining power including miRNAs into pathway annotations

The use of correlation measures, although effective, does not contextualize the putative miRNA-target interactions. From this point of view the visualization of miRNAs within biological pathways would strongly enhance the data analyses and the interpretation of results. However, all pathways annotations in the four databases mentioned in the previous chapters totally lack of miRNA elements.

Taking advantage of our tool *graphite*, we decide to expand its pathway annotations, including experimentally validated miRNAs. This new *graphite* will allow us to perform topological pathway analyses on pathways including both genes and miRNAs. As far as I known, this is the first attempt in this field.

6.1.2.1 The new *graphite* with miRNAs

Many miRNA target genes has been validated with direct methods, they can be found in literature and are collected in public databases.

I decided to select only experimentally validated miRNA-target interactions from Tarbase (Vergoulis et al., 2012) and miRecords (Xiao et al., 2009). Then, a miRNA has been introduced within a pathway only if its experimentally validated target are still present in the pathway.

In this new version of *graphite* network nodes are both genes and miRNAs, and relations of type inhibition have been added between miRNAs and target genes.

After the conversion of *graphite* genes into Entrez Gene ID, as expected, the addition of miRNAs increases the size of the pathways, especially for cancer related pathways, for details see Table 6.1

In Table 6.2 the top twenty KEGG pathways ordered by the ratio between number of miRNAs and the original size of the pathway (with only genes) are reported. Cancer related pathways triple their original size, this is due to the increasingly higher number of miRNAs studies on cancer diseases.

		Pathways with only genes		Pathways with genes and miRNAs	
Database	N of pathways	Nodes Mean (Median)	Edges Mean (Median)	Nodes Mean (Median)	Edges Mean (Median)
KEGG	232	71.86 (54)	211.10 (75.5)	92.65 (67.5)	240.87 (101)
Reactome	1070	28.10 (11)	338.48 (23)	34.46 (15)	346.11 (28)
BioCarta	254	24.98 (16.5)	74.50 (24)	39.32 (30.5)	90.94 (41.5)
NCI	177	67.85 (44)	124.36 (53)	111.84 (79)	188.49 (114)

Table 6.1: Number of pathways converted into networks with the average number of edges and nodes before and after the miRNA addition.

Pathway Name	Number of genes	Number of added miRNAs	miRNAs/genes
Bladder cancer	42	145	3.5
Glioma	65	167	2.6
Melanoma	71	178	2.5
Thyroid cancer	29	72	2.5
Chronic myeloid leukemia	73	181	2.5
Pancreatic cancer	70	172	2.5
p53 signaling pathway	69	168	2.4
Prostate cancer	89	209	2.3
Endometrial cancer	52	121	2.3
Non-small cell lung cancer	54	125	2.3
Colorectal cancer	62	142	2.3
Dorso-ventral axis formation	25	53	2.1
Small cell lung cancer	85	177	2.1
Renal cell carcinoma	70	145	2.1
ErbB signaling pathway	87	174	2.0
Acute myeloid leukemia	58	115	2.0
mTOR signaling pathway	52	98	1.9
Adherens junction	73	136	1.9
VEGF signaling pathway	76	139	1.8

Table 6.2: Pathways ordered by the ratio of the contents of miRNAs and genes

Accordingly, the ranking of target genes by the number of their experimentally validated miRNAs, identifies the most famous cancer genes, as shown in Table 6.3.

Gene symbol	Number of miRNAs that regulate the gene
CDKN1A	37
VEGFA	32
BCL2	22
CDK6	17
MCL1	16
PTEN	15
CCND1	15

Table 6.3: Targeted genes ordered by the number of miRNAs.

6.2 miRNA and gene expression integration in Epithelial Ovarian Cancer

Seventysix patients of the training set cohort has been profiled for gene expression, and their histotype stratification is reported in Table 6.4. Materials and methods of array and qRT-PCR validations described in this chapter are available in the Appendix A.

Annotations	Number of Patients n=76
Clear Cell	16 (21%)
Endometrioid	19 (25%)
Mucinous	17 (22%)
Serous	24 (32%)

Table 6.4: Histotype stratification of patients with expression measurements of both genes and miRNAs.

Gene expression raw data, were pre-processed to filter out those probes with more than 40% of measurements below the signal-to-noise threshold. Filtered data were normalized using *quantile* algorithm (Bolstad et al., 2003).

Pathway analysis using gene and miRNA expression profiles has been performed using *SPIA* and *CliPPER* methods. Both methodologies require the definition of two classes of patients. Our datasets contain four classes (histotypes). Then, we have tested each histotype versus the others. However, in this thesis, for reasons of brevity, we decide to show only the results of mucinous histotype circuits, for which experimental validation have been performed.

6.2.1 Mucinous EOC histotype analyses using *SPIA*

6.2.1.1 Method

Published in 2007, *SPIA* has been the first topological method for pathway analysis and, currently is one of the most used.

SPIA needs as input the list of differentially expressed genes with their log fold changes and the complete list of gene names in the platform (Appendix B).

Then, *SPIA* calculates 1) the classical hypergeometric enrichment p-values, P_{NDE}, and 2) a perturbation factor as a linear function of the perturbation factors of all genes in a given pathway, whose significance is calculated through a bootstrap approach, P_{PERT}. P_{NDE} represents the probability of obtaining a number of DEGs on a given pathway at least as large as the observed one by chance. P_{PERT} represents the impact of the deregulated genes on the whole pathway. Specifically, the perturbation analysis consists on the propagation of measured expression changes across the pathway topology, taking into account the position of the genes in the pathway. Then, P_{PERT}, is the probability to observe a total accumulated perturbation value of the pathway by chance. The perturbation of each single gene is calculated using the expression change of the gene between the two conditions ($log_2(FoldChange)$), corrected for a perturbation factor. The perturbation factor of each gene is derived by the sum of perturbation factors of the upstream genes divided by the number of downstream genes, moreover each interaction is weighted by the edge type considered (e.g. +1 for activation, -1 for repression and inhibition). According to the sign of the perturbation score the pathway is defined as activated (positive perturbation score = positively perturbed) otherwise the pathway is inhibited (or negatively perturbed). Finally, these two independent p-values, P_{NDE} and P_{PERT} are combined to generate a global probability, called P_G. Adjusted P_G were calculated using popular FDR algorithm (Reiner et al., 2003). It has been demonstrated that P_G, P_{NDE} and P_{PERT} are independent of the size of the pathway.

Then, the results of a *SPIA* analysis is a table in which for each pathways the following information are reported:

- Name of the pathway;
- Size – pathway size;
- NDE – number of differentially expressed genes;
- P_{NDE} – number of differentially expressed genes contained in the pathway;
- t_A – the total accumulated perturbation of the pathway;
- P_{PERT} – see above;
- P_G – see above;
- P_{GFDR} – P_G corrected for multiple testing;
- The pathway status that can be activated or inhibited;

6.2.1.2 Results

To identify subtypes-specific miRNAs and genes between mucinous histotype and the other samples, we used the empirical Bayes test, implemented in *Limma* Bioconductor package (Smyth, 2005; Gentleman, 2005). False discovery rate (Reiner et al., 2003) was used to assess for the multiple testing using False Discovery Rate (FDR) with 0.01.

In Appendix B the list of differentially expressed genes of mucinous histotype compared to the other three are reported, while Table 6.5 shows the results of *SPIA* on pathways with genes and miRNAs.

Name of the pathway	Size	NDE	P_{NDE}	t_A	P_{PERT}	P_G	P_{GFDR}	Status
Linoleic acid metabolism	15	6	1.20E-05	-1.5409	0.126	2.18E-05	0.0033	Inhibited
Metabolic pathways	955	32	0.8252	-9.6092	0.000	5.52E-05	0.0042	Inhibited
Amyotrophic lateral sclerosis (ALS)	71	7	0.0194	-17.6675	0.001	0.0002	0.0079	Inhibited
GnRH signaling pathway	103	11	0.0020	-40.6645	0.011	0.0002	0.0079	Inhibited
Pancreatic secretion	72	10	0.0004	-4.3856	0.033	0.0001	0.0079	Inhibited
Dilated cardiomyopathy	77	12	3.63E-05	-0.7160	0.746	0.0003	0.0079	Inhibited
Neurotrophin signaling pathway	171	10	0.1263	-24.0477	0.002	0.0023	0.0381	Inhibited
Colorectal cancer	117	8	0.0833	-12.3675	0.005	0.0036	0.0381	Inhibited
Endocrine and other factor-regulated calcium reabsorption	53	6	0.0159	9.12486	0.029	0.0040	0.0381	Activated
Calcium signaling pathway	133	11	0.0140	-12.675	0.033	0.0040	0.0381	Inhibited
Ether lipid metabolism	28	5	0.0039	-1.5409	0.061	0.0022	0.0381	Inhibited
MAPK signaling pathway	274	20	0.0049	-11.076	0.077	0.0033	0.0381	Inhibited
Focal adhesion	237	18	0.0049	19.400	0.091	0.0039	0.0381	Activated
Small cell lung cancer	140	14	0.0010	9.6926	0.238	0.0023	0.0381	Activated
Pathways in cancer	381	28	0.0008	9.9661	0.443	0.0034	0.0381	Activated
Amoebiasis	96	12	0.0003	0.2660	0.941	0.0027	0.0381	Activated
Ubiquitin mediated proteolysis	152	6	0.5371	-6.6684	0.001	0.0045	0.0385	Inhibited
Spliceosome	130	3	0.8832	-4.0444	0.001	0.0070	0.03856	Inhibited
One carbon pool by folate	20	2	0.1796	-4.0444	0.003	0.0045	0.0385	Inhibited
Folate biosynthesis	14	2	0.0998	-4.0444	0.007	0.0057	0.0385	Inhibited
ErbB signaling pathway	126	9	0.0552	-6.7578	0.016	0.0071	0.0385	Inhibited
alpha-Linolenic acid metabolism	13	3	0.0122	-1.5409	0.068	0.0067	0.0385	Inhibited
Pancreatic cancer	129	11	0.0113	-12.277	0.078	0.0070	0.0385	Inhibited
Hypertrophic cardiomyopathy (HCM)	74	9	0.0021	-1.6577	0.327	0.0057	0.0385	Inhibited
Glioma	117	12	0.0018	-8.3378	0.331	0.0052	0.0385	Inhibited
Neuroactive ligand-receptor interaction	119	12	0.0021	1.5610	0.371	0.0065	0.0385	Activated
p53 signaling pathway	127	13	0.0012	-3.3928	0.459	0.0048	0.0385	Inhibited
Basal cell carcinoma	66	9	0.0009	-0.8041	0.931	0.0069	0.0385	Inhibited
Fat digestion and absorption	28	4	0.0216	-1.5409	0.051	0.0086	0.0443	Inhibited
Vascular smooth muscle contraction	100	11	0.0016	-3.1127	0.687	0.0087	0.0443	Inhibited
Osteoclast differentiation	162	10	0.0974	-14.844	0.012	0.0090	0.0444	Inhibited
Endocytosis	232	12	0.1878	-6.3922	0.007	0.0100	0.0462	Inhibited
Melanoma	113	11	0.0042	-9.7222	0.306	0.0100	0.0462	Inhibited

Table 6.5: *SPIA* results. Topological Pathway Analysis Mucinous EOC vs. other histotypes.

SPIA results although interesting, highlight an awkward behavior. Since miRNA elements represent entry points in each pathway, and given that *SPIA* enhances pathways with deregulated features upstream of their topology, we obtain significant pathways even in case a single miRNA is differentially expressed in a pathway. This uneasy characteristic of *SPIA* leads to a series of significant pathways that could be false positives. For this reason, we consider *SPIA* not suitable for the integrative analysis of miRNA and gene expression using pathways.

6.2.2 Mucinous EOC histotype analyses using *CliPPER*

6.2.2.1 Method

CliPPER implements a topological pathway analysis based on the Gaussian Graphical Models. Assuming to have two classes of samples, *CliPPER* models the data in the two classes with two graphical Gaussian models with the same undirected graph G, but different means and concentration matrix (the inverted matrix of the covariance matrix) which reflects dependencies among variables (genes). Each concentration matrix is a matrix whose element in the $(i,\ j)$ position is the partial correlation measure between the i^{th} and j^{th} variables that are the vectors of gene expression of $gene_i$ and $gene_j$ in a pre-determined biological condition. Partial correlation, that is a measure of the degree of association between two random variables, is measured between those couple of genes defined by the graph connections. After estimating means and concentration matrices, *CliPPER* perform two statistical tests: i) on means and ii) concentration matrices. These tests can be performed at pathway level, to understand which pathway are significantly involved in the biological problem, and at clique level, to understand which portion of the pathway (chain of cliques) is involved. In graph theory, a clique is a set of nodes such that for every two nodes an edge exists (connected component). A tree decomposition strategy is often used in the theory of graphs to propagate the signal. The results of a graph decomposition is a tree of cliques called junction tree. Tests on concentration matrices for each clique within the junction tree is performed and a score, called relevance score, is provided for each chain of cliques (called path). The result is a list of paths ranked by the relevance score. Higher the relevance, higher the association of these paths with the phenotype.

For each pathway, *CliPPER* reports a series of paths, ranked by the relevance with the phenotype under study. In particular, foreach path the following information are reported:

- Start – index of the starting clique of the junction tree path;
- Finish – index of the ending clique of the junction tree path;

- Clique max – index of the clique where the maximum score is reached;
- Length Path – length of the of the junction tree path (the number of cliques considered);
- Max Score – maximum score of the of the junction tree path;
- Ave Score – average score along the of the junction tree path;
- Path Activation – percentage of the junction tree path activation;
- Path Impact – impact of the of the junction tree path on the entire pathway;
- indexes of the involved and signicant cliques;
- indexes of the cliques that forming the junction tree path;
- genes composing the signicant cliques;
- genes composing the junction tree path.

For more details on the methods refer to Martini et al. (2013).

6.2.2.2 Results

The result of the pathway level analysis is reported in Table 6.6. All the pathways showing the p-values of the mean and covariance test less than 0.1 are considered for the subsequent clique-level analyses.

Table 6.6: *CliPPER* pathway level results, Mucinous EOC vs. Other Histotypes.

Pathway Name	Mean test	Covariance test
Amyotrophic lateral sclerosis (ALS)	0	0
Apoptosis	0	0
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0	0
Bacterial invasion of epithelial cells	0	0
Basal cell carcinoma	0	0
Bladder cancer	0	0
Colorectal cancer	0	0
Dilated cardiomyopathy	0	0
Ether lipid metabolism	0	0
Folate biosynthesis	0	0
Glutathione metabolism	0	0
Glycerophospholipid metabolism	0	0
Glycolysis / Gluconeogenesis	0	0
Hedgehog signaling pathway	0	0
Hepatitis C	0	0
Intestinal immune network for IgA production	0	0
Long-term depression	0	0
Maturity onset diabetes of the young	0	0
Melanoma	0	0
Neuroactive ligand-receptor interaction	0	0
Nitrogen metabolism	0	0
Non-small cell lung cancer	0	0
Pancreatic cancer	0	0
Phagosome	0	0
Renal cell carcinoma	0	0
Rheumatoid arthritis	0	0
Small cell lung cancer	0	0
Toxoplasmosis	0	0
Type II diabetes mellitus	0	0
VEGF signaling pathway	0	0
alpha-Linolenic acid metabolism	0	0
p53 signaling pathway	0	0
Arachidonic acid metabolism	0	0.01
B cell receptor signaling pathway	0	0.01
Chronic myeloid leukemia	0	0.01
Cytokine-cytokine receptor interaction	0	0.01
Endocytosis	0	0.01
Fat digestion and absorption	0	0.01
Glioma	0	0.01
Jak-STAT signaling pathway	0	0.01
Pancreatic secretion	0	0.01
Prostate cancer	0	0.01
Adherens junction	0	0.02
Calcium signaling pathway	0	0.02
Fc epsilon RI signaling pathway	0	0.02
Osteoclast differentiation	0	0.02
Wnt signaling pathway	0	0.02
Linoleic acid metabolism	0	0.03
Long-term potentiation	0	0.03
Melanogenesis	0	0.03
Neurotrophin signaling pathway	0	0.03
Oocyte meiosis	0	0.05
Steroid hormone biosynthesis	0	0.05

Table 6.6: continue on the next page.

Pathway Name	Mean test	Covariance test
Vascular smooth muscle contraction	0	0.05
Protein processing in endoplasmic reticulum	0	0.06
Axon guidance	0	0.07
Regulation of actin cytoskeleton	0	0.09
Shigellosis	0	0.09
Arginine and proline metabolism	0	0.1
Taste transduction	0	0.1
RIG-I-like receptor signaling pathway	0.01	0
Vitamin B6 metabolism	0.01	0
Thyroid cancer	0.01	0.07
Tyrosine metabolism	0.01	0.09
Acute myeloid leukemia	0.02	0.08
PPAR signaling pathway	0.03	0.03
Allograft rejection	0.04	0.06
Glycosaminoglycan degradation	0.09	0.02

Table 6.6: *CliPPER* pathway level results, Mucinous EOC vs. Other Histotypes.

A gene can belong to more than one pathway, then, in general, pathways are characterized by a large overlap. A specific signal can start into a pathway and end into another pathway. Then in this case the relevance score will be penalized. To overcome this problem, and to identify more relevant and longer paths, we decide to combine all the the best sub-portions of each pathway, generating a new pathway. Figure 6.1 shows the "union of the best paths", where the colors represent the pathway of origin.

This new pathway has been re-analyzed with *CliPPER*, with the intention to find the chain of genes/miRNAs mostly involved in the separation between the mucinous and the other histotypes.

Here, only an excerpt of the complete table of the results is reported (Table 6.7), while in Figure 6.2, highlighted in red are the genes and the miRNAs belonging to the path with the maximum score: hereafter called the mucinous path (Figure 6.3 for a zoom).

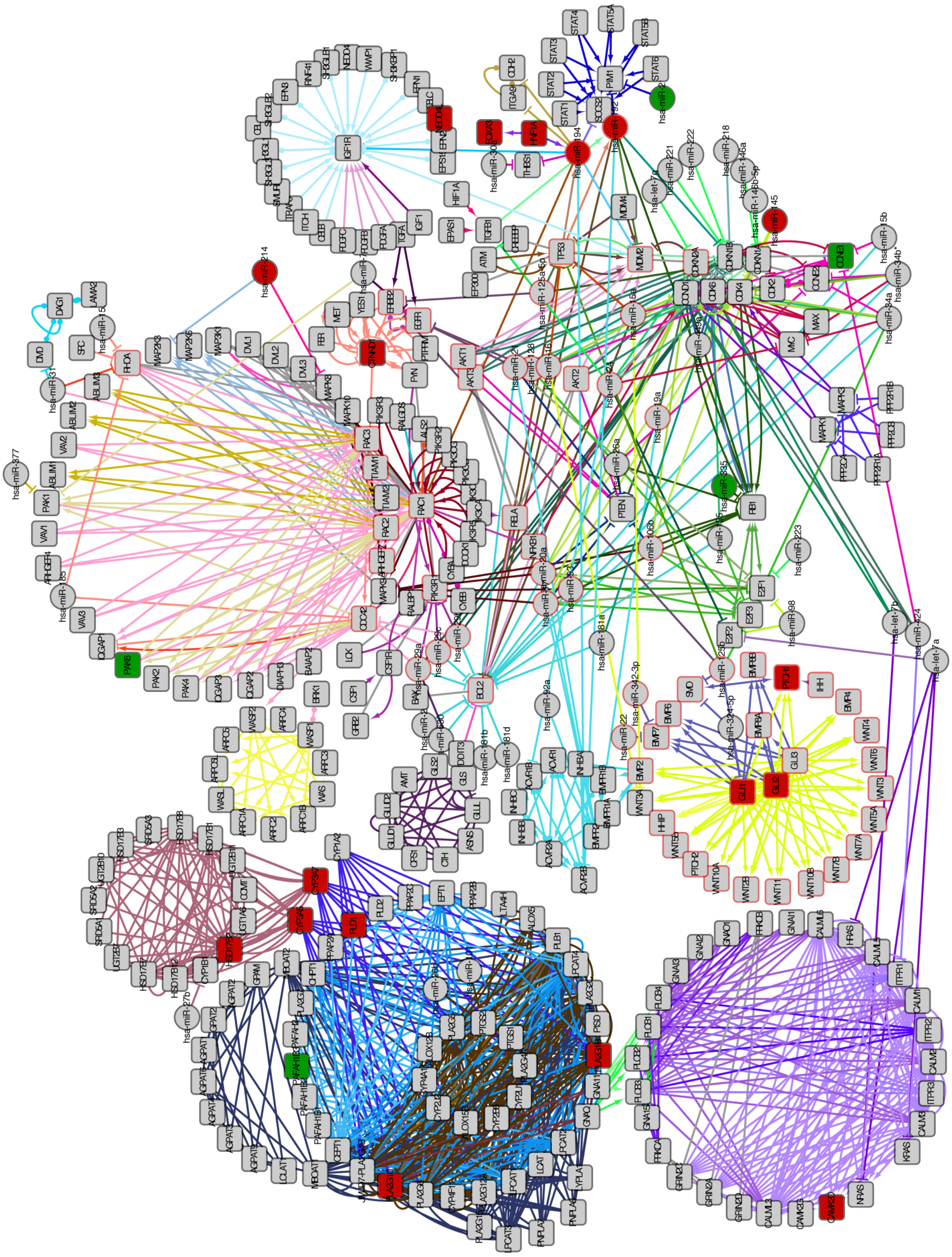


FIGURE 6.1: Union of Best Paths colored by the pathway of origin

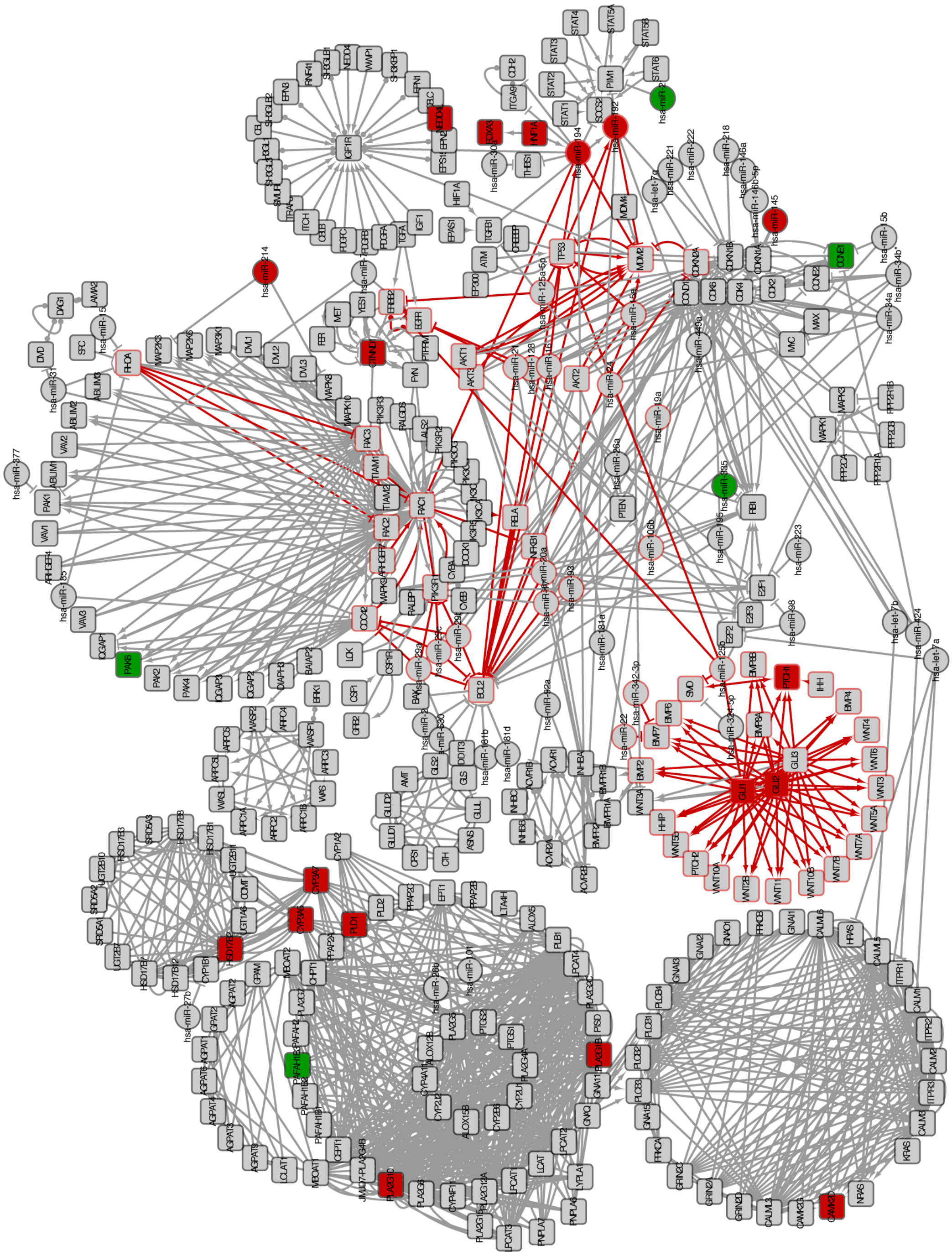


Figure 6.2: Union of Best Paths in red the path with the best score, nodes in red and in green are differentially expressed genes respectively up-regulated and down-regulated.

Start	Finish	Max Clique	Path Length	Max Score	Ave Score	Path Activation	Path Impact
4	98	28	32	101.40	3.62	0.68	0.16
4	134	25	41	67.59	2.70	0.35	0.21
4	95	14	18	49.60	3.54	0.59	0.09
4	137	12	16	41.44	3.45	0.56	0.08
186	190	5	5	19.73	3.94	0.85	0.02
182	185	4	4	18.42	4.60	1	0.02
4	64	6	10	16.57	2.76	0.36	0.05
164	169	3	4	9.84	3.28	0.53	0.02
4	163	4	8	9.21	2.30	0.25	0.04
173	175	2	3	4.94	2.47	0.35	0.01
192	193	1	2	2.30	2.30	0.25	0.01
1	3	1	3	0.99	0.99	0.07	0.01
4	152	1	4	0.74	0.74	0.04	0.02

Table 6.7: *CliPPER* reanalysis of the pathway generated by the union of the best sub-portion of each pathway. Start (index of the starting clique of the junction tree path), Finish (index of the ending clique of the junction tree path), Max Clique (index of the clique where the maximum score is reached), Path Length (length of the of the junction tree path), Max Score (maximum score of the of the junction tree path), Ave Score (average score along the of the junction tree path), Path Activation (percentage of the junction tree path activation), Path Impact (impact of the of the junction tree path on the entire pathway).

6.2.3 The mucinous pathway

A third of the significant pathways, obtained by the *SPIA* and *CliPPER* analyses, are in common, and at least one out of three are cancer pathways or pathways closely related to cancer. Among the most interesting shared by the two algorithms we find Colorectal cancer, Small cell lung cancer, Pancreatic cancer, p53 signaling pathway, Basal cell carcinoma and Melanoma, while *CliPPER* reported Renal cell carcinoma, VEGF signaling pathway, Wnt signaling pathway. All these pathways contain oncogenes that are reported to be associated or involved in EOC at advanced stages, especially worthy of note is the presence of Colorectal cancer. In fact, for several clinical, histopathological and molecular aspects, Colorectal cancer is considered the most similar tumor to Mucinous EOC, even more the ovarian ones (Vaughan et al., 2011; Köbel et al., 2008; Kurman and Shih, 2010; Marabese et al., 2008; Marchini et al., 2011; Prat, 2012b).

The use of pathways to analyze gene and miRNA expression data has the advantage to provide a known structure among elements of the network, but on the other hand has the drawback that pathway annotations can be non-exhaustive. In this perspective we decide to expand the identified mucinous pathway using classical integrative approach (using correlation measure). In case a gene or miRNA of the mutinous pathway will be

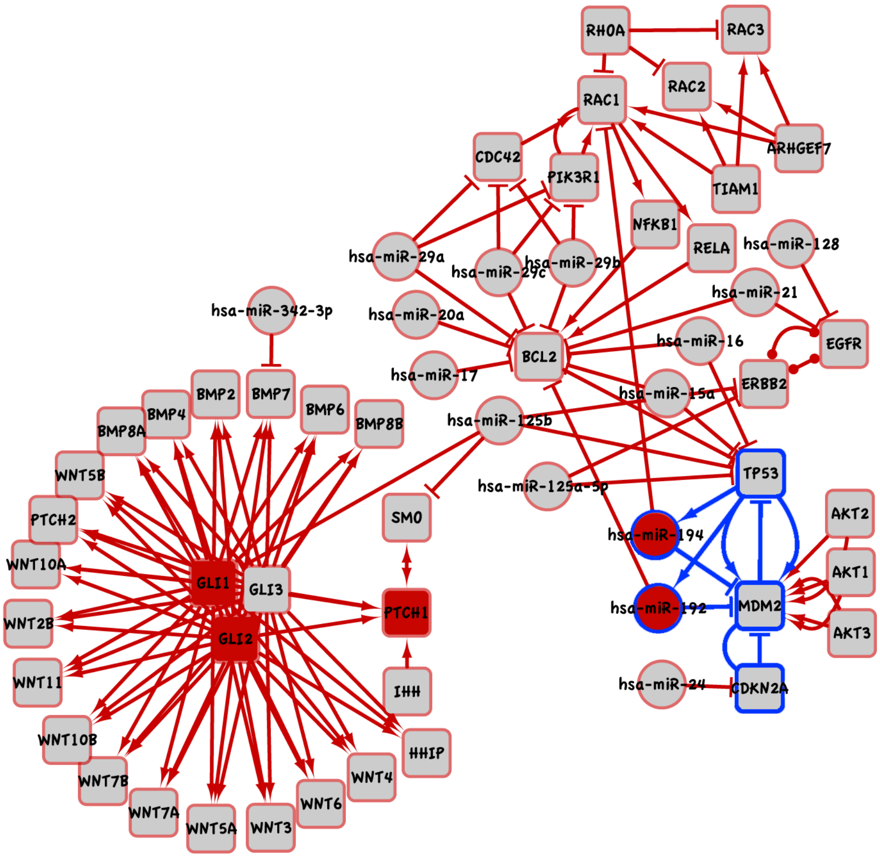


Figure 6.3: The Mucinous EOC Pathway. Highlighted in blu the reactions validated in qRT-PCR.

identified significantly anti correlated with other genes, these genes will be included in the circuit.

Using $MAGIA^2$ web tool (Bisognin et al., 2012), DIANAmicroT as miRNA-target predictor and Pearson correlation, we found miR-192, miR-194 and miR-30a among the top 20 putative interactions, re-confirming the dominant role of these miRNAs in the definition of histotypes. Many putative anti-correlated targets of hsa-miR-192/194 have been identified BMI-1, PSME3 and CUL4A, among others.

The part of the mucinous pathway highlighted in blue in Figure 6.3, involving the mucinous-specific miRNAs and p53 regulation, has been considered as the most interesting part of the pathway, and we decided to focus on it for the validations.

qRT-PCR has been performed on the entire collection of patients for BMI-1, PSME3, CUL4A, miR-192 and miR-194, CDKN2A and MDM2, confirming their differential expression with suitable level of significance ($p.value \leq 0.05$). In particular, BMI-1, PSME3, CUL4A and MDM2 were confirmed as three to four folds down-regulated in the mucinous compared to the other subtypes, while mucinous samples show high expression levels of CDKN2A, miR-192 and miR-194 (Figure 6.4 A).

Specifically, we focus on the portion of the pathway reported in Figure 6.4 B. This circuit regards the regulation of p53, it is known that in EOC p53 play an important role in tumor progression due to its mutation and inactivation at advanced stages. The expression of miR-192/194 cluster is directly controlled by wild type TP53 that, enhancing their transcriptions, is able to down-regulate genes of G1-G2 phase, targets of these two miRNAs, arresting cell cycle (Stehling et al., 2012). Among the targets of miR192/194 is MDM2, a negative regulator of TP53 (Pichiorri et al., 2010). These relationships define a positive feed back loop involving TP53 that, through miR192/194, inhibits its own inhibitor. This positive feed back loop TP53-miR192/194-MDM2 confers the status of tumor suppressor to the mir-192-194 cluster. Recently, the importance of this circuit has been further strengthened by the identification of new oncogenes among the down-regulated targets genes of these miRNAs (Dong et al., 2011; Feng et al., 2011; Yang et al., 2009).

In our analysis, promising miR-192/194 targets are PSME3 and CUL4A because, despite the lacks of these annotations in the pathway, it is known that they physically associate with Mdm2 and are part of the p53 degradation pathway. Specifically, Psme3 is a proteasome activator that promotes the nuclear export of p53 by operating multiple monoubiquitylation enhancing its physical interaction with Mdm2 (Liu et al., 2010), and Cul4A is a Cullin family member that physically associates with Mdm2 and participates as a scaffold in the process of polyubiquitylation of p53 (Nag et al., 2004) and the consequent degradation. In our data, PSME3 and CUL4A have negatively correlated expression profiles with miR192/194 and an *in silico* predicted binding sites, suggesting a functional binding. However, our luciferase binding assay reveals that no direct binding occurs between miR-194 and any of these two genes (data not shown).

Another interesting element linked to this pathway is BMI-1, since the Bmi-1 protein is a repressor of the CDKN2A protein Kim et al. (2004), that prevents the degradation and inactivation of p53 operated by MDM2 (Zhang et al., 1998). Bmi-1 belongs to the polycomb group (PcG) of proteins that form chromatin-modifying complexes commonly deregulated in cancer. BMI-1 is known to be significantly over-expressed in ovarian, endometrial and cervical cancer compared to normal tissue, and its expression is positively correlated with grade and clinical phases of the disease (Honig et al., 2010; Zhang et al., 2008). Notably, miR-194 binding on BMI-1 mRNA was experimentally validated with the luciferase assay in a panel of endometrial cancer cell lines (Dong et al., 2011) and re-confirmed by our assay. In our dataset, BMI-1 mRNA were negatively correlated with expression levels of miR-194, down-regulated in mucinous and up-regulated in other

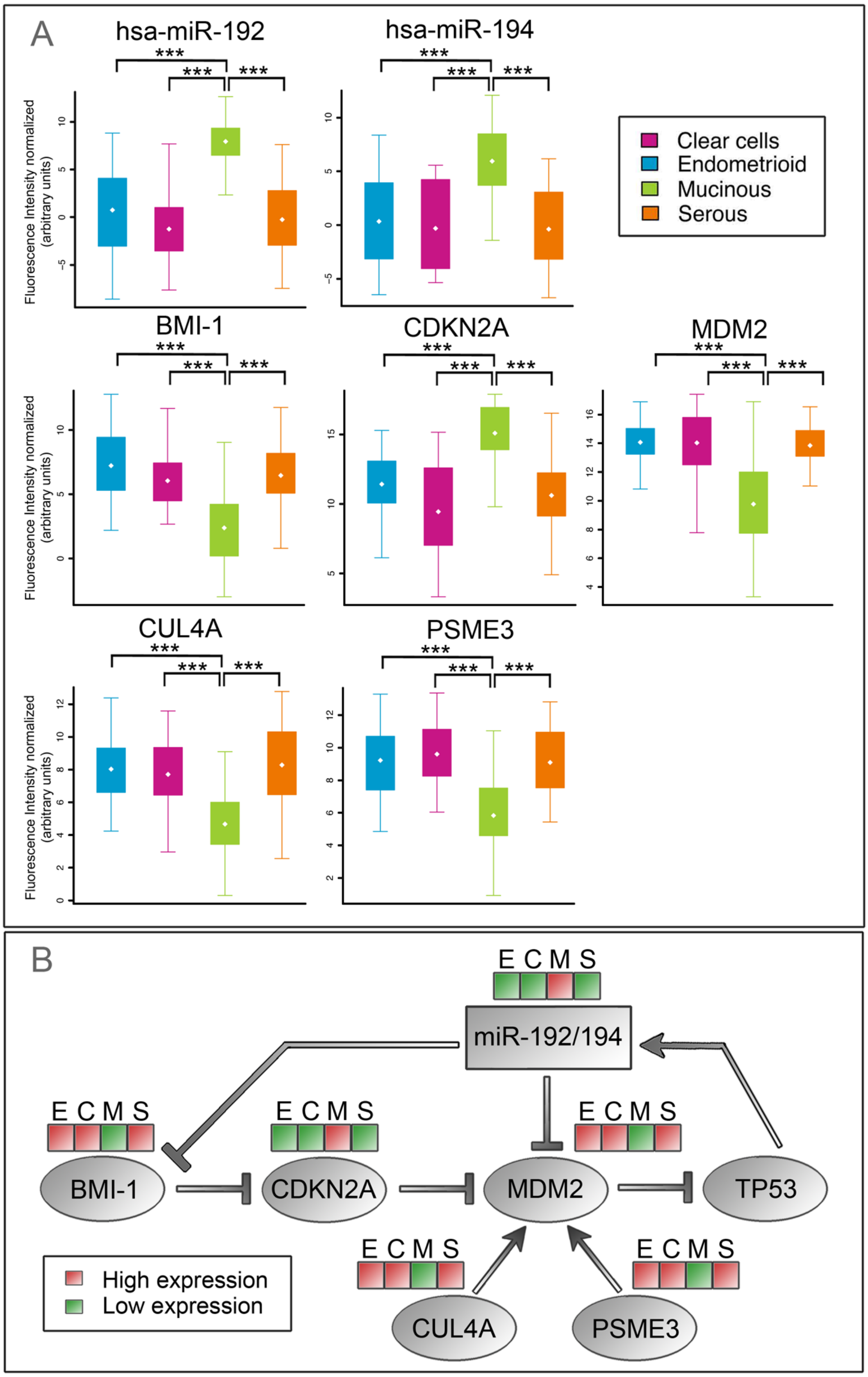


FIGURE 6.4: The Mucinous circuit experimentally validated. Panel A: Real-Time validations of the genes of the circuit in the entire cohort of patients (n=257). Panel B: Schema of the p53 circuit. Bmi-1 is a repressor of Cdkn2A protein (Kim et al., 2004), that prevents the degradation and inactivation of p53 operated by Mdm2 (Zhang et al., 1998). Psme3 and Cul4A, interacting with Mdm2, promote the p53 degradation (Liu et al., 2010; Nag et al., 2004). Moreover, p53 is the transcription factor that controls the expression of miR-192/194 cluster (Pichiorri et al., 2010). Among the targets of this miRNA cluster, miR-194 down regulates BMI-1 and both miR-192 and miR-194 target MDM2 (Dong et al., 2011; Pichiorri et al., 2010). Red and green in color bar represent high expression and low expression respectively in Endometrioid (E), Clear Cell (C), Mucinous (M), Serous (S) histotypes.

histotypes, confirming the differential activation of the signalling circuit in mucinous compared to other EOC subtypes.

Taken together, our results suggest a contrary regulation of p53 circuit in the mucinous subtype as compared to the other ovarian cancer histotypes.

Chapter 7

Conclusions

The aim of my PhD project was the development of tools and methodologies to perform integrated analysis of gene and miRNAs expression data, to provide a better comprehension of circuits of Epithelial Ovarian Cancer histotypes.

Regarding the methodological part of this thesis, we developed *graphite*, an innovative package able to gather and make easily available the contents of the four major pathway databases. In the field of topological analysis *graphite* acts as a provider of biological information by reducing the pathway complexity and considering the biological meaning of the pathway elements. The high number of accesses demonstrates its usefulness. Moreover, this thesis demonstrates that *graphite* can be used as a computational platform for the integration of different sources of information missed in pathway annotations (such as microRNAs), making possible for the first time to run topological pathway analyses on bipartite graphs composed of genes and microRNAs.

I'm strongly convinced that *graphite* enhances and facilitates the development of new tools for network analysis.

Regarding the applicative part of the thesis we obtained important results on the characterization of early stage EOC subtypes. Specifically, there are clear evidences that some ovarian cancers are more similar to certain types of renal, breast and endometrial cancers than other ovarian histologies.

The top priorities of EOC research are i) the identification of possible alternative therapies for different types of ovarian carcinoma and ii) the development of new measures for EOC prevention and early detection (Vaughan et al., 2011).

Considering these two key priorities, we thought that a better characterization of the early stage tumor environment was fundamental to identify biomarkers and new putative targets for subtype specific therapies.

Our results (chapter 5, 6) demonstrate that (i) early stage EOC microRNA pattern is different across subtypes, and that (ii) early stage EOC subtypes seem characterized by specific molecular circuits that differentiate the tumor environment.

In particular, in this study we found that miR-30a and miR-192/194 are key markers of clear cell and mucinous subtypes, respectively, and that in mucinous histotype occurs a different regulation of genes and miRNAs upstream p53.

These finding are important for two reasons. First, mucinous and clear cells histotypes are considered ideal candidates for developing new therapeutic strategies, because of their high mortality and lowest sensitivity to standard chemotherapy (Alexandre et al., 2010). Second, these miRNAs are known to play important roles in other cancer diseases and these findings can be considered as starting point for new treatments of EOC. mir-192 and miR-194 are also markers of colorectal tumors, this connection hints tantalizingly the possibility to treat in the same way mucinous EOC and colon cancer. The strong relation between this two diseases is confirmed, in advanced stage, by many clinical, histologicical, molecular and biochemical evidences, to the point that the Medical Research Council are attempting to treat advanced stage mucinous EOC (mEOC, Clinical Trial Identifier: NCT01081262) with a combination of chemotherapeutic agents commonly used in colon cancer. In this context, the identified miRNAs can be considered as novel therapeutic targets, as well as the identified circuits.

As future perspective, it would be interesting to investigate in III/IV EOC stages the behavior of genes and miRNAs involved in the circuits identified in early stages. This could be important to understand the differences between early and advanced EOCs and to potentially clarify histotype-specific mechanisms of disease progression.

Appendix A

Materials and Methods

In this appendix are collected all the materials and methods about the experimental methodologies included in this thesis. The experiments are performed by researchers of the Dr. Maurizio D'Incalci and Dr. Sergio Marchini Group at the "Mario Negri" Institute in Milano.

A.1 miRNA microarray experiments

Frozen samples (30 mg) were homogenised using TissueLyser LT (Qiagen, Milano Italy) and total RNA enriched in miRNAs fraction was purified using a miRneasy isolation kit according to the manufacturer's instructions (Qiagen). RNA quality control, Cy5-labelling and hybridization were performed as previously published (Marchini et al., 2011). miRNA profiles were generated using commercially available G4470B human miRNA Microarray kit (Agilent Technologies, Palo Alto, CA, USA), which consists of 15K features printed in an 8-plex format (8x15 array), able to detect all known human miRNAs (723 human and 76 human-viral miRNAs) sourced from the Sanger miRBASE public database, release 10.1. The arrays were washed and scanned with a laser confocal scanner (G2565B, Agilent Technologies), according to the manufacturer's instructions. miRNA microarrays underwent standard post hybridisation processing and the intensities of fluorescence were calculated by Feature Extraction software version 11 (Agilent Technologies).

A.2 Gene microarray experiments

Frozen tissues specimens (30 mg) were homogenized in an TissueLyser LT (Qiagen, Milan Italy) and total RNA purified using RNeasy Mini Kit isolation system (Qiagen), following manufacturers protocols. Total RNA concentration and proteins contamination were determined by Nanodrop spectrophotometer (Nanodrop Technologies, Ambion). Only samples with a RIN larger than 6 and a Nanodrop A260:280 ratio between 1.8 and 2.1 were further processed and aliquots stored at -80C until use. Array experiments were performed using standard procedures as previously published by Marchini et al. (2013). Briefly, one hundred ngs of total RNA was reverse transcribed into Cy3-labeled cRNA using LowInput QuickAmp labelling kit (Agilent Technologies, Palo Alto, Ca, US) and hybridized with a RNA labelling and hybridization kit according to the manufacturers instructions (Agilent Technologies). We used the commercially available G4851B human whole GE Microarray kit (SurePrint G3 Human Gene Expression 8x60K v2 Microarray Kit Agilent Technologies) which consists of 60K features printed in an 8-plex format (8x60 array). The arrays were washed and scanned with a laser confocal scanner (G2565B, Agilent Technologies) according to the manufacturers instructions. mRNA microarrays underwent standard post hybridization processing and the intensities of fluorescence were calculated by Feature Extraction software v11 (Agilent Technologies).

A.3 Quantitative reverse transcription real time PCR

Mature miRNA and gene expression levels were examined by quantitative real-time reverse transcription PCR (qRT-PCR) using Sybr Green protocol (Qiagen). miRNA expression analysis was performed using a dedicated set of commercial primers (Qiagen) as previously described by Marchini et al. (2011). Data were normalised using the geometric mean of the four independent housekeeping controls (RNU6B, SNORD61, SNORD72, SNORD68). For gene expression validations, primer pair sequences are listed in Table A.1; data were normalised using the four independent housekeeping genes (ACTB, B2M, PPIA, e HPRT1) as described previously by Marchini et al. (2008). Experiments were run in triplicate, using 384-well reaction plates in an automatic liquid handling station (epMotion 5075LH; Eppendorf, Milano, Italy). Real-time PCR was done on an Applied Biosystems 7900HT (Ambion-ABI). Raw data was generated with SDS Relative Quantification software (version 2.3; Ambion-ABI).

Name	Gene Bank Accession number	Primer	Sequence	Ta (°C)	Ampl Length
BMI1	NM_005180.8	Fw	CTGCTCTTTCCGGGATTTTT	60	88
		Rv	ACACACATCAGGTGGGGATT	60	
CUL4A	NM_003589.2	Fw	AGAATGAGCGGTTCGTCAAC	60	96
		Rv	CCACATGCTTTGCGATCA	60	
PMSE3	NM_005789.3	Fw	ATGCCCAATGGGATGCT	60	110
		Rv	ACCCACATTTTGACCGTGTT	60	
ACTB	NM_001101.3	Fw	CAGAGCCTCGCCTTTGC	60	65
		Rv	TCATCATCCATGGTGAGCTG	60	
B2M	NM_004048.2	Fw	AAGCAGCATCATGGAGGTTT	60	69
		Rv	AGCAAGCAAGCAGAATTTGG	60	
PPIA	NM_021130.3	Fw	GCGTCTCCTTTGAGCTGTTT	60	79
		Rv	CCTTTCTCTCCAGTGCTCAGA	60	
HPRT1	NM_000194.2	Fw	TGAATACTTCAGGGATTTGAATCAT	60	76
		Rv	CTCATCTTAGGCTTTGTATTTTGC	60	
3’ UTR CUL4A	NM_003589.2	Fw	ATGCTAGCTTCCCCTTCATGAAACA	66	1146
		Rv	CAACTCGAGTGCCATGATCAAAATTC	66	
3’ UTR PMSE3	NM_005789.3	Fw	AAAGAGCTCATGATCCTGAAAAATATC	61	1970
		Rv	ATGCTAGCTAACTTTCCCATAATTCAGA	61	
3’ UTR BMI1	NM_005180.8	Fw	CACCTTCATGCCATTACAGCTTTCT	64	1899
		Rv	CTTTCAATGGGCTTTCAAGCAA	64	
CDKN2A	NM_000077.4	Fw	GGGTTTTCGTGGTTCACATC	60	142
		Rv	TCATCATGACCTGGTCTTCTAGG	60	
MDM2	NM_002392.4	Fw	GGAGAGCAATTAGTGAGACAGAAGA	60	213
		Rv	CTGAATGTTCACTTACACCAGCA	60	

Table A.1: Pair sequences, Gene Bank accession number and annealing temperature (T_a) of analyzed genes by RT-qPCR.

Appendix B

Supplementary Information

B.1 Survival analyses across histotypes

Survival analyses were performed using Kaplan-Meier method and significance was assessed with two-sided log-rank statistics (Figure B.1). The Cox proportional hazards model was use to determine the risk ratios and p-value for multivariate analysis (Tables B.1, B.2, B.3, B.4, B.5).

Histotype	coef	exp(coef)	se(coef)	z	p
Endometrioid	-0.256	0.774	0.414	-0.619	0.540
Mucinous	-0.920	0.399	0.463	-1.986	0.047
Serous	-0.164	0.848	0.397	-0.414	0.680

Table B.1: Progression Free Survival Cox model using subtypes as covariate (Clear cells as reference). Likelihood ratio test=5.72 on 3 df, p=0.126 n= 255, number of events= 58

Grade	coef	exp(coef)	se(coef)	z	p
Grade 1	2.04	7.72	1.05	1.94	0.0530
Grade 2	1.92	6.81	1.06	1.81	0.0700
Grade 3	2.77	15.90	1.02	2.70	0.0069

Table B.2: Overall Survival Cox model using grades as covariate (Borderline samples as reference). Likelihood ratio test=17.9 on 3 df, p=0.000462 n= 255, number of events= 39

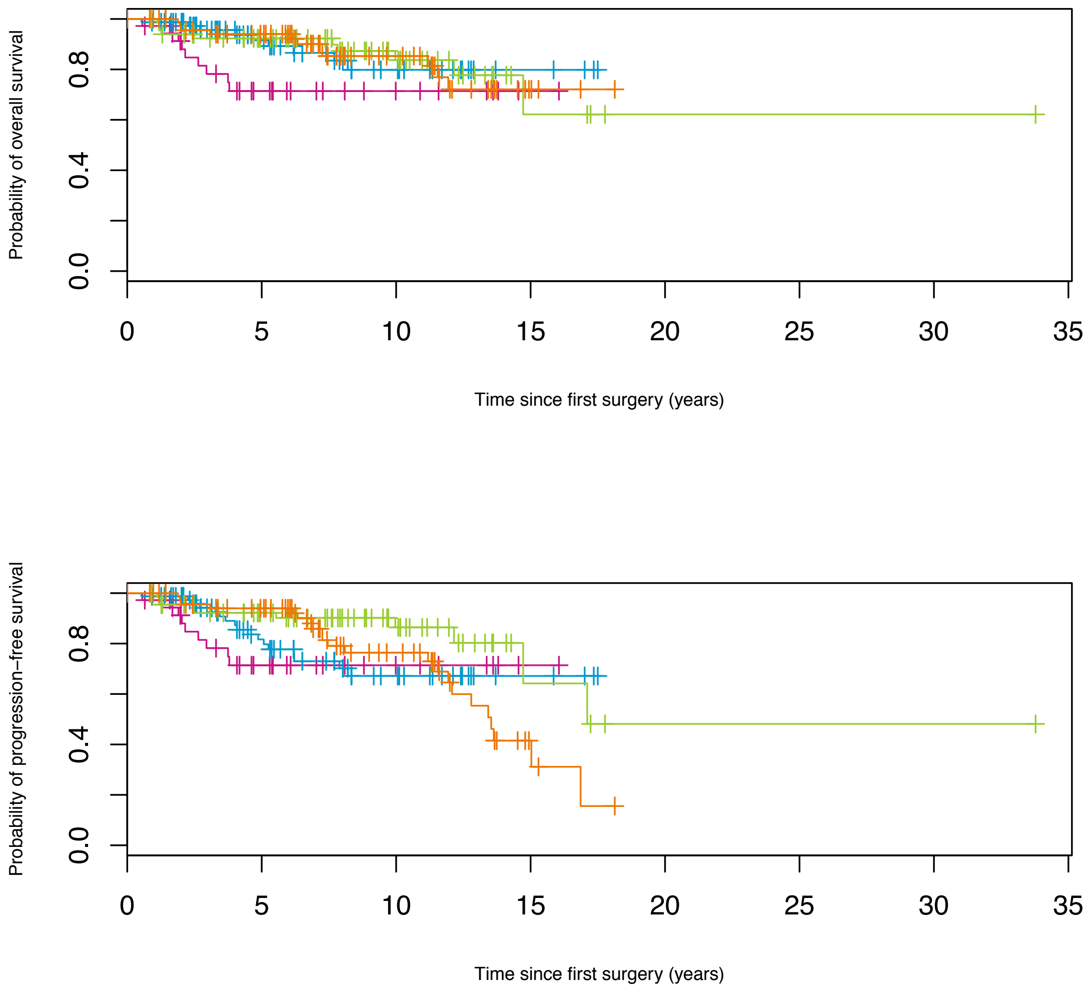


Figure B.1: Kaplan-Meier curves. Panel A: Overall Survival (n=255, p-value=0.16). Panel B: Progression Free Survival (n=255, p-value=0.15). Pink for Clear cells, Orange for Serous, Green for Mucinous, Blue for Endometrioid histotypes.

Grade	coef	exp(coef)	se(coef)	z	p
Grade 1	0.445	1.56	0.477	0.934	0.3500
Grade 2	0.338	1.40	0.485	0.698	0.4900
Grade 3	1.177	3.25	0.424	2.774	0.0055

Table B.3: Progression free Survival Cox model using grades as covariate (Borderline samples as reference). Likelihood ratio test=11.9 on 3 df, p=0.0078 n= 255, number of events= 58

miRNA	coef	exp(coef)	se(coef)	z	p
miR192	0.0883	1.092	0.110	0.803	0.42
miR194	-0.1769	0.838	0.137	-1.296	0.20
miR30a	-0.0600	0.942	0.163	-0.367	0.71
miR30a*	0.1464	1.158	0.128	1.140	0.25

Table B.4: Overall Survival Cox model using miRNA expression as covariate. Likelihood ratio test=3.82 on 4 df, p=0.43 n= 254, number of events= 39

miRNA	coef	exp(coef)	se(coef)	z	p
miR192	0.0310	1.031	0.0863	0.359	0.72
miR194	-0.1032	0.902	0.1080	-0.956	0.34
miR30a	-0.0699	0.933	0.1382	-0.506	0.61
miR30a*	0.1688	1.184	0.1079	1.564	0.12

Table B.5: Progression free Cox model using miRNA expression as covariate. Likelihood ratio test=6.41 on 4 df, p=0.17 n= 254, number of events= 58

B.2 Complete List of Differentially expressed miRNAs

MicroRNAs differentially expressed in all possible couples of comparisons:

- Clear cell vs. Endometrioid
- Clear Cell vs. Mucinous
- Clear Cell vs. Serous
- Endometrioid vs. Mucinous
- Endometrioid vs. Serous
- Mucinous vs. Serous

For each miRNA of each comparison were provided:

- log_2(Fold Change);
- p-value adjusted for multiple tests (adj.P.Val) obtained using Limma using all the 183 samples;
- the max number of Simulated Resampling Score (SRS);
- the Resampling Score (RS).

The results provided are selected for $adj.P.Val \leq 0.01$ and are ordered by Resampling Score (RS).

microRNAs	log_2(FC)	adj.P.Val	Max SRS	RS
hsa-miR-30a	2.178221964	6.93932E-13	112	500
hsa-miR-30a*	1.626517454	4.67144E-07	99	494
hsa-miR-181a	-1.360369063	1.0446E-05	112	486
hsa-miR-181b	-1.076253543	2.40595E-05	80	475
hsa-miR-205	-3.00045046	2.92244E-05	26	464
hsa-miR-193b	1.150661602	7.866E-05	83	429
hsa-miR-30c	1.187338837	0.000203527	111	368
hsa-miR-181c	-1.101136057	0.000758317	24	327
hsa-miR-92a	-0.754437111	0.001341581	32	297
hsa-let-7i	1.173828572	0.001389269	71	291
hsa-miR-200b*	-0.879910948	0.003103202	22	228
hsa-miR-148a	-1.170526396	0.003567933	43	226
hsa-miR-185	0.660247135	0.004622196	39	212

Table B.6: Clear Cell vs. Endometrioid histotype differentially expressed miRNAs

microRNAs	log_2(FC)	adj.P.Val	Max SRS	RS
hsa-miR-192	-4.576489119	2.3573E-19	28	500
hsa-miR-30a	2.476917077	1.27358E-16	35	500
hsa-miR-194	-4.488328243	1.27358E-16	15	500
hsa-miR-30a*	1.908687702	8.94055E-10	178	500
hsa-miR-30c	1.421447032	2.64924E-06	33	483
hsa-miR-1274a	1.617274552	2.64924E-06	76	477
hsa-miR-222	-1.658988055	3.25091E-06	32	482
hsa-miR-338-3p	-1.746691771	8.34732E-06	37	490
hsa-miR-145	-1.392430686	0.000223802	58	402
hsa-miR-214	-1.535040388	0.000538999	60	366
hsa-miR-365	0.732745265	0.000912231	13	339
hsa-miR-30d	0.866090808	0.001509118	37	315
hsa-miR-221	-0.998459893	0.001611332	34	309
hsa-miR-193b	0.934407313	0.001861248	41	288
hsa-miR-377	-1.075660458	0.003029451	137	261
hsa-miR-200b*	-0.866616497	0.003071745	54	245
hsa-miR-532-5p	0.795612694	0.00321089	16	243
hsa-miR-376c	-1.052506251	0.003635647	25	258
hsa-let-7i	1.035743742	0.00439688	101	250
hsa-miR-185	0.630775276	0.004430261	84	219
hsa-miR-30e*	0.765833391	0.005344596	27	198
hsa-miR-1274b	0.999145742	0.00546896	60	207
hsa-miR-30b	1.306220711	0.006775498	72	178
hsa-miR-181a	-0.862589598	0.006963722	106	181
hsa-miR-10a	1.433787263	0.009769118	21	170

Table B.7: Clear Cell vs. Mucinous histotype differentially expressed miRNAs

microRNAs	log_2(FC)	adj.P.Val	Max SRS	RS
hsa-miR-30a	1.98939242	8.08689e-11	11	500
hsa-miR-92a	-1.0549930	3.19074e-06	14	490
hsa-miR-34b*	-2.280838	2.49918e-05	18	484
hsa-miR-30a*	1.2858667	8.72349e-05	19	408
hsa-miR-222	-1.277072	0.0016498999	13	304
hsa-miR-17*	-1.0391151	0.0024902687	10	269
hsa-miR-19b	-0.875396	0.0030416892	17	245
hsa-miR-193a-3p	-0.985749	0.0056551662	92	203
hsa-miR-30c-2*	0.9197421	0.0078809350	40	186
hsa-miR-29b	-1.2552117	0.0095792861	63	184
hsa-miR-1308	1.3057282	0.0095792861	132	176
hsa-miR-193b	0.8551035	0.0095792861	58	175
hsa-miR-575	1.4909159	0.0095792861	53	165
hsa-miR-29c	-0.938144	0.0095792861	53	159

Table B.8: Clear Cell vs. Serous histotype differentially expressed miRNAs

microRNAs	log_2(FC)	adj.P.Val	Max SRS	RS
hsa-miR-192	-4.456901911	1.80092E-21	29	500
hsa-miR-194	-4.012306899	6.09125E-17	55	500
hsa-miR-338-3p	-1.791611501	1.1649E-06	13	491
hsa-miR-96	1.61909503	4.9799E-05	24	442
hsa-miR-20b	1.249263862	0.000136048	93	404
hsa-miR-183	1.298797021	0.001194581	142	308
hsa-miR-222	-1.063356031	0.001194581	17	303
hsa-miR-141	1.158775513	0.001194581	124	296
hsa-miR-497	-1.074432848	0.003045108	102	234
hsa-miR-93	0.878133669	0.004003256	241	225
hsa-miR-181c	0.748650383	0.00503931	49	199
hsa-miR-30d	0.706895756	0.00503931	33	178
hsa-miR-376c	-0.918312032	0.007016886	20	170
hsa-miR-1274a	0.919275326	0.007193528	101	168

Table B.9: Endometrioid vs. Mucinous histotype differentially expressed miRNAs

microRNAs	log_2(FC)	adj.P.Val	Max SRS	RS
hsa-miR-146b-5p	-1.386438884	0.000198779	65	392
hsa-miR-29b	-1.521002838	0.000286949	112	364
hsa-miR-29c	-1.080214183	0.00070553	153	341
hsa-miR-484	0.649515566	0.00260455	42	254
hsa-miR-101	-0.745487733	0.00260455	34	251
hsa-miR-29c*	-0.651810471	0.00260455	23	250
hsa-miR-1225-3p	0.979074041	0.00260455	126	248
hsa-miR-575	1.445641549	0.003786505	71	235
hsa-miR-1234	0.879150849	0.006049702	93	213
hsa-miR-150*	1.264372168	0.006349042	93	209
hsa-miR-514	-1.622195144	0.006583467	162	202
hsa-miR-1225-5p	1.168814698	0.006583467	92	198
hsa-miR-572	1.240903836	0.006583467	102	194
hsa-miR-557	1.272441414	0.006583467	203	193
hsa-miR-34b*	-1.345890203	0.006583467	29	189
hsa-miR-513c	-1.131672064	0.006583467	68	185
hsa-miR-877*	0.845435947	0.006583467	57	171
hsa-miR-513a-5p	-1.093407519	0.006607725	73	175

Table B.10: Endometrioid vs. Serous histotype differentially expressed miRNAs

microRNAs	log_2(FC)	adj.P.Val	Max SRS	RS
hsa-miR-192	4.288388557	1.93547E-20	26	500
hsa-miR-194	4.147014794	2.75664E-18	122	500
hsa-miR-1225-5p	1.793013009	8.81332E-06	89	491
hsa-miR-20a	-1.39515799	1.48515E-05	53	494
hsa-miR-20b	-1.344149297	1.69514E-05	77	493
hsa-miR-34b*	-1.915035209	2.99683E-05	52	482
hsa-miR-193a-3p	-1.110131825	5.81468E-05	25	477
hsa-miR-17	-1.272144444	8.39811E-05	36	462
hsa-miR-188-5p	1.41523542	0.0002525	101	442
hsa-miR-22	0.778625655	0.000279902	8	459
hsa-miR-145	1.198337052	0.000345076	45	428
hsa-miR-1207-5p	1.444466298	0.000345076	66	427
hsa-miR-92a	-0.697720894	0.000356708	28	440
hsa-miR-141	-1.204247989	0.000356708	81	424
hsa-miR-338-3p	1.269349245	0.000384408	12	422
hsa-miR-497	1.148237194	0.000521258	39	422
hsa-miR-29b	-1.312050202	0.000521258	137	404
hsa-miR-575	1.554934939	0.000521258	79	400
hsa-miR-146b-5p	-1.140320799	0.000526351	137	400
hsa-miR-93	-0.94801489	0.000552286	112	393
hsa-miR-96	-1.286492834	0.000552286	37	384
hsa-miR-99b	-0.843618434	0.000552286	63	380
hsa-miR-150*	1.349673535	0.000842048	79	379
hsa-miR-134	1.257441784	0.001097997	69	359
hsa-miR-1246	1.486772311	0.001146725	33	355
hsa-miR-1249	1.168186792	0.001274883	77	346
hsa-miR-200c	-1.247868741	0.001449652	88	351
hsa-miR-125a-5p	-0.907679032	0.001525345	46	336
hsa-miR-324-5p	-0.690790352	0.001558047	38	336
hsa-miR-296-5p	1.064346369	0.001677927	79	337
hsa-miR-1275	1.231039729	0.001739698	53	333
hsa-miR-574-5p	0.854702462	0.001758872	64	332
hsa-miR-557	1.260014191	0.001759154	67	324
hsa-miR-19b	-0.71764471	0.001769056	70	333
hsa-miR-30c	-0.836727999	0.001769056	70	319
hsa-miR-362-3p	-0.72465274	0.001769056	42	301
hsa-miR-135b	-1.243556395	0.002126579	36	302
hsa-miR-1202	1.059862896	0.002459214	71	313
hsa-miR-638	1.143638189	0.002533228	73	310
hsa-miR-532-5p	-0.677178833	0.002579142	30	280
hsa-miR-155	-0.821102463	0.002652694	24	304
hsa-miR-142-3p	-1.385247887	0.003008615	33	310
hsa-miR-106b	-0.851954056	0.003462208	65	284
hsa-miR-1268	1.077024082	0.003462208	89	280
hsa-miR-1238	0.64661252	0.003462208	66	270
hsa-miR-30e*	-0.657704694	0.003555385	38	260
hsa-miR-532-3p	-0.713174598	0.003873134	43	246
hsa-miR-572	1.156191361	0.00483514	87	260
hsa-miR-1234	0.794140304	0.005210659	95	239
hsa-miR-1915	1.068602197	0.005284211	88	258
hsa-miR-720	-0.933296394	0.006013686	128	227
hsa-miR-183	-0.981239607	0.006782023	29	240
hsa-miR-1228	0.84097685	0.006782023	87	236
hsa-miR-1274a	-0.817022386	0.006782023	95	224
hsa-miR-17*	-0.723808576	0.006782023	50	206
hsa-miR-320c	0.681622511	0.006782023	46	204
hsa-miR-630	1.039829562	0.0068716	111	219
hsa-let-7e	-0.971560456	0.007690934	49	209

Table B.11: Mucinous vs. Sierous histotype differentially expressed miRNAs

B.3 Survival Analyses on Grades within Mucinous histotype

B.3.1 Overall survival

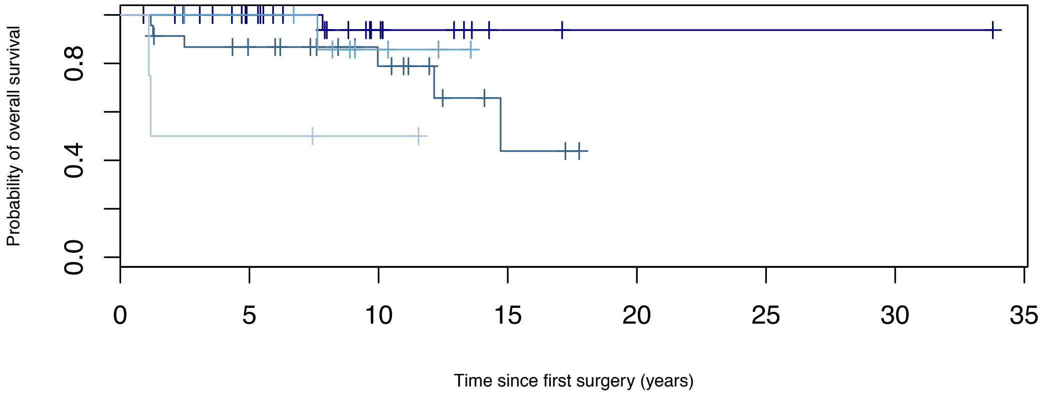


Figure B.2: Kaplan Meyer curves for Overall Survival in the Mucinous subtypes according to their grade (p-values=0.004)

B.3.2 Progression Free survival

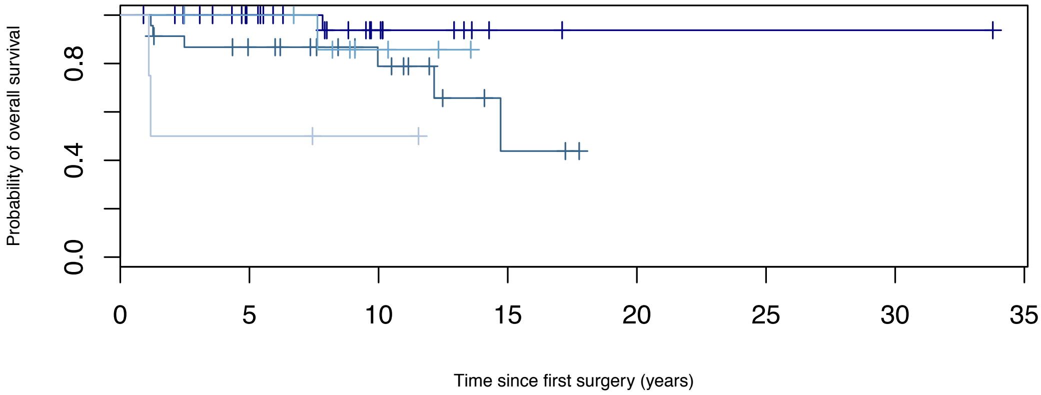


Figure B.3: Kaplan Meyer curves for Progression Free in the Mucinous Samples divided by grades (p-values=0.008)

B.4 Differentially expressed genes and miRNAs between Mucinous and other histotypes used for *SPIA* analyses

B.4.1 miRNAs

miRNA	logFC	adj.P.Val
hsa-miR-192	4.044470612	2.07879E-11
hsa-miR-194	4.02387039	9.87869E-10
hsa-miR-200c	-1.118990711	0.010642539
hsa-miR-145	1.252030074	0.010642539
hsa-miR-338-3p	1.540952688	0.01663384
hsa-miR-96	-1.179182871	0.036711442
hsa-miR-301a	-1.109432673	0.041937882
hsa-miR-20b	-0.984266574	0.066515973
hsa-miR-497	1.007091155	0.066515973
hsa-miR-362-5p	-0.734909312	0.066515973
hsa-miR-210	-0.766896196	0.066515973
hsa-miR-141	-0.854919178	0.081436696
hsa-miR-335	-0.864608152	0.083912645
hsa-miR-214	1.03674175	0.083912645
hsa-miR-183	-0.989627457	0.083912645
hsa-miR-30d	-0.792547238	0.08424842
hsa-miR-107	-0.535256818	0.08424842

Table B.12: Mucinous vs. Other histotypes differentially expressed miRNAs

B.4.2 Genes

Here for editorial reasons, gene list has been cut at a lower threshold ($p - value \leq 0.01$) than those used in the analyses.

Table B.13: Mucinous vs. Other histotypes differentially expressed genes

GeneSymbol	logFC	adj.P.Val
TFF1	4.434130557	3.63438E-06
LGALS4	4.701214179	5.21281E-06
LRRC66	2.479798915	1.57372E-05
SPINK1	4.84668279	1.57372E-05
MYZAP	1.50738676	1.57372E-05
RAB27B	2.112706856	1.63424E-05

Table B.13: continue on the next page.

GeneSymbol	logFC	adj.P.Val
PLA2G10	2.326301545	2.03161E-05
MYO1A	3.19173963	7.91367E-05
CEACAM6	4.152909092	7.91367E-05
SLC39A5	3.152411174	9.26524E-05
ANG	1.922497805	0.000123874
PRSS3	2.238985055	0.000148356
IL22RA1	2.413954232	0.000148356
CCL15	1.608149241	0.000172193
NR5A2	1.566734587	0.000207831
SYT13	2.877758176	0.000321438
SLC22A18AS	1.355754715	0.000528924
PLAC8	2.61284333	0.000576194
SLC9B2	1.12697286	0.000576194
CTSE	3.621933541	0.000628096
FMO5	2.781182353	0.000628096
ERN2	2.663659039	0.00068929
CP	-2.673705775	0.00068929
NPC1L1	1.641892743	0.000833552
BCAS1	2.799324436	0.00087236
SMPDL3A	1.699290232	0.000966416
FOLR1	-2.581747267	0.00096865
FBXO16	-1.272146672	0.001000462
CLDN3	-1.733869151	0.001000462
FOXS1	2.154517309	0.001097864
SIX4	-1.560785311	0.001154362
CLEC3B	1.769693701	0.001154362
FNIP2	1.05383133	0.001257983
UNC5CL	1.246777081	0.001399352
CDHR2	1.732772926	0.002438209
C1orf186	-2.17557003	0.002571349
AZGP1	1.92914555	0.002710447
SH3BGRL2	1.219752963	0.002710447
MARCH3	1.165645356	0.002772439
ITLN1	2.666644842	0.002772439
MAPRE2	1.102123098	0.003073387
RNASE4	1.289442666	0.003073387
SOX17	-1.977324862	0.003073387
CYP3A5	2.150437967	0.003234718
SH2B2	-1.445007767	0.003834804
FOXA3	1.175496573	0.004395511
SAMD5	1.695987	0.004395511
VSIG2	2.475823296	0.004395511
RGL3	-1.539464794	0.004577705
PLEK2	1.416682145	0.004866037
SDCBP2	1.580566917	0.00545028
TPM1	1.034590979	0.005506123
AIF1L	-1.352583467	0.005506123
C10orf35	-1.030234118	0.005506123
LFNG	1.120519797	0.005506123
DHTKD1	-0.881526837	0.00551
PAG1	1.038164152	0.005735513
MXD1	1.092633826	0.005815497
SAMD13	1.211635167	0.006193405
CYP3A7	2.740875149	0.007019751
TRIM54	1.752507941	0.007028482
VILL	1.65729073	0.007042674

Table B.13: continue on the next page.

GeneSymbol	logFC	adj.P.Val
TSPO2	1.260070156	0.007175694
OLFML2A	1.052032108	0.007294847
EML4	0.635294515	0.007399134
PDIA4	-0.723848997	0.007548295
CCDC157	-0.921496923	0.007767063
MYOZ3	1.406476336	0.007767063
TESC	2.091074629	0.007767063
HNF1A	2.276781298	0.008357846
CIDEC	0.956013123	0.008564873
KLK8	-3.114099597	0.008564873
FBP1	1.561509671	0.008628993
EPS8L3	1.840871526	0.008804382
TNF	-1.550469627	0.008804382
WFDC2	-2.25641872	0.00911212
UNC13B	0.625459656	0.00911212
E2F5	-0.910277267	0.009308925
DNAL1	-0.838387221	0.009508292
C14orf176	1.312300807	0.010009152
ITPKA	1.547992181	0.010009152
SLPI	-1.693099295	0.010009152
CES2	0.844237483	0.010009152
MIA2	1.777636117	0.010009152
TNNT1	-2.545081503	0.010009152
CAMK2D	0.727178811	0.010009152
PTCH1	0.694491215	0.010009152
IL17RE	0.947958013	0.010009152
RGS5	1.548379717	0.010009152
PLD1	1.079465237	0.010031078
COL17A1	2.161497443	0.011053277
NPTN	0.555750368	0.011258286
INSL3	2.284999627	0.011258286
CEACAM7	2.651567331	0.011365777
EDA	-1.119250772	0.011413606
SMPD3	1.891344602	0.011636491
SPTBN1	1.056120938	0.011636491
MTMR11	1.20301869	0.011846393
ATP2A3	1.196774725	0.011849605
AHCYL2	1.014735314	0.012083973
NFE2	-1.839054994	0.012302538
EDEM3	0.726696341	0.012302538
TFF2	2.942947002	0.012302538
TRIM15	1.88665857	0.012302538
FOXP1	0.617310189	0.012302538
VTCN1	-2.343983669	0.012302538
HSD17B2	2.471808862	0.012302538
FCN3	0.778889591	0.012659651
TSPAN8	2.5119881	0.012997127
MRPS26	-0.556876087	0.013072475
C11orf82	-1.027920065	0.013496372
SPAG4	-1.19408042	0.013955254
GNG12	0.96927124	0.014251991
FHL2	0.997513108	0.014334561
SLC36A1	0.857367734	0.014334561
DQX1	2.240277453	0.014334561
SLC1A7	1.510187587	0.014334561
ZDHHC3	0.57180131	0.014336132

Table B.13: continue on the next page.

GeneSymbol	logFC	adj.P.Val
HEYL	1.565926287	0.0144661
ARTN	-0.752589736	0.0144661
PLS1	1.383494537	0.015091066
AGR2	2.668231685	0.015310712
VPS13A	0.673413888	0.015501889
TBC1D8B	0.632538657	0.015667325
REEP3	0.75299725	0.016075394
LYST	0.767777478	0.016075394
UBL3	0.730106855	0.016206247
STARD10	1.127436	0.016696594
C4orf19	1.477091712	0.016696594
DACT3	1.3875729	0.016788126
CIDEB	0.862242663	0.016950565
CAPN9	1.556220096	0.018695143
PAX8	-1.79456781	0.018695143
GPX2	2.359366854	0.019216276
PLCL2	0.791803439	0.019328773
FEM1B	0.672782769	0.019328773
PRKACB	0.855320566	0.019328773
C9orf85	0.57970109	0.019523946
ACTG2	2.069573223	0.019523946
LGALS2	1.828102135	0.019523946
NDNF	1.184303325	0.019523946
MARCKS	0.970589419	0.019523946
ZNF664-FAM101A	0.914831161	0.019523946
CAMK2N1	1.635252	0.019523946
ADRA2C	-1.716065349	0.019523946
SPEF2	-0.863605671	0.019523946
PRSS2	1.965044418	0.019523946
ADH1C	1.581301176	0.019523946
RHOJ	1.045420813	0.019611792

Publications during the Ph.D.

- **graphite - a Bioconductor package to convert pathway topology to gene network.**

  Sales G*, Calura E*, Cavalieri D, Romualdi C.

  BMC Bioinformatics. 2012 Jan 31;13:20. doi: 10.1186/1471-2105-13-20. PubMed PMID: 22292714; PubMed Central PMCID: PMC3296647.

  * Equally Contribution

- **A systems biology approach to characterize the regulatory networks leading to trabectedin resistance in an in vitro model of myxoid liposarcoma.**

  Uboldi S, Calura E, Beltrame L, Fuso Nerini I, Marchini S, Cavalieri D, Erba E, Chiorino G, Ostano P, D'Angelo D, D'Incalci M, Romualdi C.

  PLoS One. 2012;7(4):e35423. doi: 10.1371/journal.pone.0035423. Epub 2012 Apr 16. PubMed PMID: 22523595; PubMed Central PMCID: PMC3327679.

- **The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways.**

  Beltrame L*, Calura E*, Popovici RR, Rizzetto L, Guedez DR, Donato M, Romualdi C, Draghici S, Cavalieri D.

  Bioinformatics. 2011 Aug 1;27(15):2127-33. doi:10.1093/bioinformatics/btr339. Epub 2011 Jun 7. PubMed PMID:21653523; PubMed Central PMCID: PMC3137220.

  * Equally Contribution

- **Association between miR-200c and the survival of patients with stage I epithelial ovarian cancer: a retrospective study of two independent tumour tissue collections.**

  Marchini S, Cavalieri D, Fruscio R, Calura E, Garavaglia D, Nerini IF, Mangioni C, Cattoretti G, Clivio L, Beltrame L, Katsaros D, Scarampi L, Menato G, Perego P, Chiorino G, Buda A, Romualdi C, D'Incalci M.

  Lancet Oncol. 2011 Mar;12(3):273-85. doi: 10.1016/S1470-2045(11)70012-2. Epub 2011 Feb 21. PubMed PMID: 21345725.

- **DC-ATLAS: a systems biology resource to dissect receptor specific signal transduction in dendritic cells.**

  Cavalieri D, Rivero D, Beltrame L, Buschow SI, Calura E, Rizzetto L, Gessani S, Gauzzi MC, Reith W, Baur A, Bonaiuti R, Brandizi M, De Filippo C, D'Oro

U, Draghici S, Dunand-Sauthier I, Gatti E, Granucci F, Gundel M, Kramer M, Kuka M, Lanyi A, Melief CJ, van Montfoort N, Ostuni R, Pierre P, Popovici R, Rajnavolgyi E, Schierer S, Schuler G, Soumelis V, Splendiani A, Stefanini I, Torcia MG, Zanoni I, Zollinger R, Figdor CG, Austyn JM.

Immunome Res. 2010 Nov 19;6:10. doi: 10.1186/1745-7580-6-10. PubMed PMID: 21092113; PubMed Central PMCID: PMC3000836.

Submitted papers

- **Graphite Web: web tool for gene set analysis exploiting pathway topology**

  Sales G, Calura E, Martini P, Romualdi C.

  Submitted to NAR - web server issue

- **miRNA landscape in Stage I Epithelial Ovarian Cancer defines the histotype specificities**

  E. Calura, R. Fruscio, L. Paracchini, E. Bignotti, A. Ravaggi, P. Martini, G. Sales, L. Beltrame, L. Clivio, L. Ceppi, M. Di Marino, I. Fuso Nerini, L. Zanotti, D. Cavalieri, G. Cattoretti, P. Perego, R. Milani, D. Katsaros, G. Tognon, E. Sartori, S. Pecorelli, C. Mangioni, M. DIncalci, C. Romualdi, S. Marchini

  Submitted to Clinical Cancer Research

Papers in preparation

- **miRNA expression in Grades of Stage I Epithelial Ovarian Cancer**

  E. Calura, R. Fruscio, L. Paracchini, E. Bignotti, A. Ravaggi, P. Martini, G. Sales, L. Beltrame, L. Clivio, L. Ceppi, M. Di Marino, I. Fuso Nerini, L. Zanotti, D. Cavalieri, G. Cattoretti, P. Perego, R. Milani, D. Katsaros, G. Tognon, E. Sartori, S. Pecorelli, C. Mangioni, M. DIncalci, C. Romualdi, S. Marchini

Bibliography

Robert J Kurman and Ie-Ming Shih. The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory. *The American journal of surgical pathology*, 34 (3):433–443, March 2010. ISSN 1532-0979. doi: 10.1097/PAS.0b013e3181cf3d79. URL http://www.ncbi.nlm.nih.gov/pubmed/20154587. PMID: 20154587.

Sebastian Vaughan, Jermaine I Coward, Jr Bast, Robert C, Andy Berchuck, Jonathan S Berek, James D Brenton, George Coukos, Christopher C Crum, Ronny Drapkin, Dariush Etemadmoghadam, Michael Friedlander, Hani Gabra, Stan B Kaye, Chris J Lord, Ernst Lengyel, Douglas A Levine, Iain A McNeish, Usha Menon, Gordon B Mills, Kenneth P Nephew, Amit M Oza, Anil K Sood, Euan A Stronach, Henning Walczak, David D Bowtell, and Frances R Balkwill. Rethinking ovarian cancer: recommendations for improving outcomes. *Nature Reviews. Cancer*, 11(10): 719–725, October 2011. ISSN 1474-1768. doi: 10.1038/nrc3144. URL http://www.ncbi.nlm.nih.gov/pubmed/21941283. PMID: 21941283.

C Blake Gilks. Molecular abnormalities in ovarian cancer subtypes other than high-grade serous carcinoma. *J Oncol*, 2010:740968, 2010. doi: 10.1155/2010/740968.

Danijela Jelovac and Deborah K Armstrong. Recent progress in the diagnosis and treatment of ovarian cancer. *CA Cancer J Clin*, 61(3):183–203, 2011. doi: 10.3322/caac.20113.

J Prat. New insights into ovarian cancer pathology. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, 23 Suppl 10:x111–x117, September 2012a. ISSN 1569-8041. doi: 10.1093/annonc/mds300. PMID: 22987944.

Martin Köbel, Steve E Kalloger, Niki Boyd, Steven McKinney, Erika Mehl, Chana Palmer, Samuel Leung, Nathan J Bowen, Diana N Ionescu, Ashish Rajput, Leah M Prentice, Dianne Miller, Jennifer Santos, Kenneth Swenerton, C. Blake Gilks, and David Huntsman. Ovarian carcinoma subtypes are different diseases: Implications for biomarker studies. *PLoS Med*, 5(12):e232, December 2008. doi: 10.1371/journal.pmed.0050232. URL http://dx.doi.org/10.1371/journal.pmed.0050232.

Linda E Kelemen and Martin Köbel. Mucinous carcinomas of the ovary and colorectum: different organ, same dilemma. *The Lancet Oncology*, 12(11):1071–1080, October 2011. ISSN 14702045. doi: 10.1016/S1470-2045(11)70058-4. URL `http://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(11)70058-4/abstract`.

Anais Malpica, Michael T Deavers, Karen Lu, Diane C Bodurka, Edward N Atkinson, David M Gershenson, and Elvio G Silva. Grading ovarian serous carcinoma using a two-tier system. *Am J Surg Pathol*, 28(4):496–504, Apr 2004.

Deborah K Armstrong. Relapsed ovarian cancer: challenges and management strategies for a chronic disease. *Oncologist*, 7 Suppl 5:20–8, 2002.

Sergio Marchini, Pietro Mariani, Giovanna Chiorino, Eleonora Marrazzo, Riccardo Bonomi, Robert Fruscio, Luca Clivio, Annalisa Garbi, Valter Torri, Michela Cinquini, Tiziana Dell'Anna, Giovanni Apolone, Massimo Broggini, and Maurizio D'Incalci. Analysis of gene expression in early-stage ovarian cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 14(23):7850–7860, December 2008. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-08-0523. URL `http://www.ncbi.nlm.nih.gov/pubmed/19047114`. PMID: 19047114.

Stephen A Cannistra. Cancer of the ovary. *N Engl J Med*, 351(24):2519–29, Dec 2004. doi: 10.1056/NEJMra041842.

David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 116 (2):281–97, Jan 2004.

David P Bartel. Micrornas: target recognition and regulatory functions. *Cell*, 136(2): 215–33, Jan 2009. doi: 10.1016/j.cell.2009.01.002.

Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. mirbase: tools for microrna genomics. *Nucleic Acids Res*, 36(Database issue):D154–8, Jan 2008. doi: 10.1093/nar/gkm952.

Sam Griffiths-Jones. mirbase: microrna sequences and annotation. *Curr Protoc Bioinformatics*, Chapter 12:Unit 12.9.1–10, Mar 2010. doi: 10.1002/0471250953.bi1209s29.

Isaac Bentwich, Amir Avniel, Yael Karov, Ranit Aharonov, Shlomit Gilad, Omer Barad, Adi Barzilai, Paz Einat, Uri Einav, Eti Meiri, Eilon Sharon, Yael Spector, and Zvi Bentwich. Identification of hundreds of conserved and nonconserved human micrornas. *Nat Genet*, 37(7):766–70, Jul 2005. doi: 10.1038/ng1590.

Eugene Berezikov, Victor Guryev, José van de Belt, Erno Wienholds, Ronald H A Plasterk, and Edwin Cuppen. Phylogenetic shadowing and computational identification of human microrna genes. *Cell*, 120(1):21–4, Jan 2005. doi: 10.1016/j.cell.2004.12.031.

Ulf Andersson Ørom and Anders H Lund. Experimental identification of microrna targets. *Gene*, 451(1-2):1–5, Feb 2010. doi: 10.1016/j.gene.2009.11.008.

Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–10, Jan 2011. doi: 10.1093/nar/gkq1184.

Helen Parkinson, Ugis Sarkans, Nikolay Kolesnikov, Niran Abeygunawardena, Tony Burdett, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Ele Holloway, Natalja Kurbatova, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Gabriella Rustici, Anjan Sharma, Eleanor Williams, Tomasz Adamusiak, Marco Brandizi, Nataliya Sklyar, and Alvis Brazma. Arrayexpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, 39(Database issue):D1002–4, Jan 2011. doi: 10.1093/nar/gkq1040.

Andreas Ruepp, Andreas Kowarsch, Daniel Schmidl, Felix Buggenthin, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, and Fabian J Theis. Phenomir: a knowledgebase for microrna expression in diseases and biological processes. *Genome Biol*, 11(1):R6, 2010. doi: 10.1186/gb-2010-11-1-r6.

Aly A Khan, Doron Betel, Martin L Miller, Chris Sander, Christina S Leslie, and Debora S Marks. Transfection of small rnas globally perturbs gene regulation by endogenous micrornas. *Nat Biotechnol*, 27(6):549–55, Jun 2009. doi: 10.1038/nbt.1543.

Marcus E Peter. Let-7 and mir-200 micrornas: guardians against pluripotency and cancer progression. *Cell Cycle*, 8(6):843–52, Mar 2009.

M Maragkakis, M Reczko, V A Simossis, P Alexiou, G L Papadopoulos, T Dalamagas, G Giannopoulos, G Goumas, E Koukis, K Kourtis, T Vergoulis, N Koziris, T Sellis, P Tsanakas, and A G Hatzigeorgiou. Diana-microt web server: elucidating microrna functions through target prediction. *Nucleic Acids Res*, 37(Web Server issue):W273–6, Jul 2009. doi: 10.1093/nar/gkp292.

Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of mirna targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69, 2007. doi: 10.1186/1471-2105-8-69.

Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome Biol*, 11(8):R90, 2010. doi: 10.1186/gb-2010-11-8-r90.

Sabbi Lall, Dominic Grün, Azra Krek, Kevin Chen, Yi-Lu Wang, Colin N Dewey, Pranidhi Sood, Teresa Colombo, Nicolas Bray, Philip Macmenamin, Huey-Ling Kao, Kristin C Gunsalus, Lior Pachter, Fabio Piano, and Nikolaus Rajewsky. A genome-wide map of conserved microrna targets in c. elegans. *Curr Biol*, 16(5):460–71, Mar 2006. doi: 10.1016/j.cub.2006.01.050.

Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microrna target recognition. *Nat Genet*, 39(10):1278–84, Oct 2007. doi: 10.1038/ng2135.

Kevin C Miranda, Tien Huynh, Yvonne Tay, Yen-Sin Ang, Wai-Leong Tam, Andrew M Thomson, Bing Lim, and Isidore Rigoutsos. A pattern-based method for the identification of microrna binding sites and their corresponding heteroduplexes. *Cell*, 126 (6):1203–17, Sep 2006. doi: 10.1016/j.cell.2006.07.031.

Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mrnas are conserved targets of micrornas. *Genome Res*, 19(1):92–105, Jan 2009. doi: 10.1101/gr.082701.108.

Panagiotis Alexiou, Manolis Maragkakis, Giorgos L Papadopoulos, Martin Reczko, and Artemis G Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics (Oxford, England)*, 25(23):3049–3055, December 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp565. URL `http://www.ncbi.nlm.nih.gov/pubmed/19789267`. PMID: 19789267.

Ming Lu, Qipeng Zhang, Min Deng, Jing Miao, Yanhong Guo, Wei Gao, and Qinghua Cui. An analysis of human microrna and disease associations. *PLoS One*, 3(10):e3420, 2008. doi: 10.1371/journal.pone.0003420.

Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Res*, 37(Database issue):D98–104, Jan 2009. doi: 10.1093/nar/gkn714.

Seungyoon Nam, Bumjin Kim, Seokmin Shin, and Sanghyuk Lee. mirgator: an integrated system for functional annotation of micrornas. *Nucleic Acids Res*, 36(Database issue):D159–64, Jan 2008. doi: 10.1093/nar/gkm829.

Molly Megraw, Praveen Sethupathy, Benoit Corda, and Artemis G Hatzigeorgiou. mirgen: a database for the study of animal microrna genomic organization and function. *Nucleic Acids Res*, 35(Database issue):D149–55, Jan 2007. doi: 10.1093/nar/gkl904.

Priyanka Shahi, Serguei Loukianiouk, Andreas Bohne-Lang, Marc Kenzelmann, Stefan Küffer, Sabine Maertens, Roland Eils, Herrmann-Josef Gröne, Norbert Gretz, and

Benedikt Brors. Argonaute–a database for gene regulation by mammalian micrornas. *Nucleic Acids Res*, 34(Database issue):D115–8, Jan 2006. doi: 10.1093/nar/gkj093.

Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. mirecords: an integrated resource for microrna-target interactions. *Nucleic Acids Res*, 37(Database issue):D105–10, Jan 2009. doi: 10.1093/nar/gkn851.

Thanasis Vergoulis, Ioannis S Vlachos, Panagiotis Alexiou, George Georgakilas, Manolis Maragkakis, Martin Reczko, Stefanos Gerangelos, Nectarios Koziris, Theodore Dalamagas, and Artemis G Hatzigeorgiou. Tarbase 6.0: capturing the exponential growth of mirna targets with experimental support. *Nucleic Acids Res*, 40(Database issue): D222–9, Jan 2012. doi: 10.1093/nar/gkr1161.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, Mar 2011. doi: 10.1016/j.cell.2011.02.013.

Yong Sun Lee and Anindya Dutta. Micrornas in cancer. *Annu Rev Pathol*, 4:199–227, 2009. doi: 10.1146/annurev.pathol.4.110807.092222.

Marilena V. Iorio, Rosa Visone, Gianpiero Di Leva, Valentina Donati, Fabio Petrocca, Patrizia Casalini, Cristian Taccioli, Stefano Volinia, Chang-Gong Liu, Hansjuerg Alder, George A. Calin, Sylvie Ménard, and Carlo M. Croce. MicroRNA signatures in human ovarian cancer. *Cancer Research*, 67(18):8699–8707, September 2007. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-07-1936. URL `http://cancerres.aacrjournals.org/content/67/18/8699`.

Andrea Ventura and Tyler Jacks. MicroRNAs and cancer: short RNAs go a long way. *Cell*, 136(4):586–591, February 2009. ISSN 1097-4172. doi: 10.1016/j.cell.2009.02.005. URL `http://www.ncbi.nlm.nih.gov/pubmed/19239879`. PMID: 19239879.

Georgios Pampalakis, Eleftherios P Diamandis, Dionyssios Katsaros, and Georgia Sotiropoulou. Down-regulation of dicer expression in ovarian cancer tissues. *Clinical Biochemistry*, 43(3):324–327, February 2010. ISSN 1873-2933. doi: 10.1016/j.clinbiochem.2009.09.014. URL `http://www.ncbi.nlm.nih.gov/pubmed/19782670`. PMID: 19782670.

Sergio Marchini, Duccio Cavalieri, Robert Fruscio, Enrica Calura, Daniela Garavaglia, Ilaria Fuso Nerini, Costantino Mangioni, Giorgio Cattoretti, Luca Clivio, Luca Beltrame, Dionyssios Katsaros, Luca Scarampi, Guido Menato, Patrizia Perego, Giovanna Chiorino, Alessandro Buda, Chiara Romualdi, and Maurizio D'Incalci. Association between miR-200c and the survival of patients with stage i epithelial ovarian cancer: a retrospective study of two independent tumour tissue collections. *The Lancet*

Oncology, 12(3):273–285, March 2011. ISSN 1474-5488. doi: 10.1016/S1470-2045(11)70012-2. URL `http://www.ncbi.nlm.nih.gov/pubmed/21345725`. PMID: 21345725.

Diana M Cittelly, Irina Dimitrova, Erin N Howe, Dawn R Cochrane, Annie Jean, Nicole S Spoelstra, Miriam D Post, Xian Lu, Russell R Broaddus, Monique A Spillman, and Jennifer K Richer. Restoration of mir-200c to ovarian cancer reduces tumor burden and increases sensitivity to paclitaxel. *Mol Cancer Ther*, 11(12):2556–65, Dec 2012. doi: 10.1158/1535-7163.MCT-12-0463.

Marcos Malumbres. mirnas versus oncogenes: the power of social networking. *Mol Syst Biol*, 8:569, 2012. doi: 10.1038/msb.2012.2.

Azra Krek, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microrna target predictions. *Nat Genet*, 37 (5):495–500, May 2005. doi: 10.1038/ng1536.

Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005. doi: 10.1016/j.cell.2004.12.035.

M E Peter. Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene*, 29(15):2161–2164, April 2010. ISSN 1476-5594. doi: 10.1038/onc.2010.59. URL `http://www.ncbi.nlm.nih.gov/pubmed/20190803`. PMID: 20190803.

Daehyun Baek, Judit Villén, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. The impact of micrornas on protein output. *Nature*, 455(7209):64–71, Sep 2008. doi: 10.1038/nature07242.

Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. A coding-independent function of gene and pseudogene mrnas regulates tumour biology. *Nature*, 465(7301):1033–8, Jun 2010. doi: 10.1038/nature09144.

Yvonne Tay, Lev Kats, Leonardo Salmena, Dror Weiss, Shen Mynn Tan, Ugo Ala, Florian Karreth, Laura Poliseno, Paolo Provero, Ferdinando Di Cunto, Judy Lieberman, Isidore Rigoutsos, and Pier Paolo Pandolfi. Coding-independent regulation of the tumor suppressor pten by competing endogenous mrnas. *Cell*, 147(2):344–57, Oct 2011. doi: 10.1016/j.cell.2011.09.029.

Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A cerna hypothesis: the rosetta stone of a hidden rna language? *Cell*, 146(3):353–8, Aug 2011. doi: 10.1016/j.cell.2011.07.014.

Masafumi Inui, Graziano Martello, and Stefano Piccolo. Microrna control of signal transduction. *Nat Rev Mol Cell Biol*, 11(4):252–63, Apr 2010. doi: 10.1038/nrm2868.

Héctor Herranz and Stephen M Cohen. Micrornas and gene regulatory networks: managing the impact of noise in biological systems. *Genes Dev*, 24(13):1339–44, Jul 2010. doi: 10.1101/gad.1937010.

Yuri Lazebnik. Can a biologist fix a radio?–or, what i learned while studying apoptosis. *Cancer Cell*, 2(3):179–82, Sep 2002.

Luca Beltrame, Enrica Calura, Razvan R Popovici, Lisa Rizzetto, Damariz Rivero Guedez, Michele Donato, Chiara Romualdi, Sorin Draghici, and Duccio Cavalieri. The biological connection markup language: a sbgn-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics*, 27(15):2127–33, Aug 2011. doi: 10.1093/bioinformatics/btr339.

Deyun Pan, Ning Sun, Kei-Hoi Cheung, Zhong Guan, Ligeng Ma, Matthew Holford, Xingwang Deng, and Hongyu Zhao. Pathmapa: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for arabidopsis. *BMC Bioinformatics*, 4:56, Nov 2003. doi: 10.1186/1471-2105-4-56.

Nicolas Le Novère, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, Emek Demir, Katja Wegner, Mirit I Aladjem, Sarala M Wimalaratne, Frank T Bergman, Ralph Gauges, Peter Ghazal, Hideya Kawaji, Lu Li, Yukiko Matsuoka, Alice Villéger, Sarah E Boyd, Laurence Calzone, Melanie Courtot, Ugur Dogrusoz, Tom C Freeman, Akira Funahashi, Samik Ghosh, Akiya Jouraku, Sohyoung Kim, Fedor Kolpakov, Augustin Luna, Sven Sahle, Esther Schmidt, Steven Watterson, Guanming Wu, Igor Goryanin, Douglas B Kell, Chris Sander, Herbert Sauro, Jacky L Snoep, Kurt Kohn, and Hiroaki Kitano. The systems biology graphical notation. *Nat Biotechnol*, 27(8):735–41, Aug 2009. doi: 10.1038/nbt.1558.

Tobias Czauderna, Christian Klukas, and Falk Schreiber. Editing, validating and translating of sbgn maps. *Bioinformatics*, 26(18):2340–1, Sep 2010. doi: 10.1093/bioinformatics/btq407.

M Hucka, A Finney, H M Sauro, H Bolouri, J C Doyle, H Kitano, A P Arkin, B J Bornstein, D Bray, A Cornish-Bowden, A A Cuellar, S Dronov, E D Gilles, M Ginkel, V Gor, I I Goryanin, W J Hedley, T C Hodgman, J-H Hofmeyr, P J Hunter, N S Juty, J L Kasberger, A Kremling, U Kummer, N Le Novère, L M Loew, D Lucio, P Mendes, E Minch, E D Mjolsness, Y Nakayama, M R Nelson, P F Nielsen, T Sakurada, J C Schaff, B E Shapiro, T S Shimizu, H D Spence, J Stelling, K Takahashi, M Tomita, J Wagner, J Wang, and SBML Forum. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–31, Mar 2003.

Joanne S Luciano. Pax of mind for pathway researchers. *Drug Discov Today*, 10(13): 937–42, Jul 2005. doi: 10.1016/S1359-6446(05)03501-4.

M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. ISSN 0305-1048. URL `http://www.ncbi.nlm.nih.gov/pubmed/10592173`. PMID: 10592173.

Anna Bauer-Mehren, Laura I Furlong, and Ferran Sanz. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol Syst Biol*, 5: 290, 2009. doi: 10.1038/msb.2009.47.

Gary D Bader, Michael P Cary, and Chris Sander. Pathguide: a pathway resource list. *Nucleic Acids Res*, 34(Database issue):D504–6, Jan 2006. doi: 10.1093/nar/gkj126.

Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Ozgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39(Database issue): D685–90, Jan 2011. doi: 10.1093/nar/gkq1039.

Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl Schaefer, Joanne Luciano, Frank Schacherer, Irma Martinez-Flores, Zhenjun Hu, Veronica Jimenez-Jacinto, Geeta Joshi-Tope, Kumaran Kandasamy, Alejandra C Lopez-Fuentes, Huaiyu Mi, Elgar Pichler, Igor Rodchenkov, Andrea Splendiani, Sasha Tkachev, Jeremy Zucker, Gopal Gopinath, Harsha Rajasimha, Ranjani Ramakrishnan, Imran Shah, Mustafa Syed, Nadia Anwar, Ozgün Babur, Michael Blinov, Erik Brauner, Dan Corwin, Sylva Donaldson, Frank Gibbons, Robert Goldberg, Peter Hornbeck, Augustin Luna, Peter Murray-Rust, Eric Neumann, Oliver Ruebenacker, Oliver Reubenacker, Matthias Samwald, Martijn van Iersel, Sarala Wimalaratne, Keith Allen, Burk Braun, Michelle Whirl-Carrillo, Kei-Hoi Cheung, Kam Dahlquist, Andrew Finney, Marc Gillespie, Elizabeth Glass, Li Gong, Robin Haw, Michael Honig, Olivier Hubaut, David Kane, Shiva Krupa, Martina Kutmon, Julie Leonard, Debbie Marks, David Merberg, Victoria Petri, Alex Pico, Dean Ravenscroft, Liya Ren, Nigam Shah, Margot Sunshine, Rebecca Tang, Ryan Whaley, Stan Letovksy, Kenneth H Buetow, Andrey Rzhetsky, Vincent Schachter, Bruno S Sobral, Ugur Dogrusoz, Shannon McWeeney, Mirit Aladjem, Ewan Birney, Julio Collado-Vides, Susumu Goto, Michael Hucka, Nicolas Le Novère, Natalia Maltsev, Akhilesh Pandey, Paul Thomas, Edgar Wingender, Peter D Karp, Chris Sander, and Gary D Bader. The biopax community standard for pathway data sharing. *Nat Biotechnol*, 28(9):935–42, Sep 2010. doi: 10.1038/nbt.1666.

Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:47, 2009. doi: 10.1186/1471-2105-10-47.

Jelle J Goeman and Ulrich Mansmann. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24(4):537–44, Feb 2008. doi: 10.1093/bioinformatics/btm628.

Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Brief Bioinform*, 9(3):189–97, May 2008. doi: 10.1093/bib/bbn001.

Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran, and Yutaka Yasui. Gene-set analysis and reduction. *Brief Bioinform*, 10(1):24–34, Jan 2009. doi: 10.1093/bib/bbn042.

Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. A systems biology approach for pathway level analysis. *Genome Res*, 17(10):1537–45, Oct 2007. doi: 10.1101/gr.6202607.

Maria Sofia Massa, Monica Chiogna, and Chiara Romualdi. Gene set analysis exploiting the topology of a pathway. *BMC Syst Biol*, 4:121, 2010. doi: 10.1186/1752-0509-4-121.

Senol Isci, Cengizhan Ozturk, Jon Jones, and Hasan H Otu. Pathway analysis of high-throughput biological data within a bayesian network framework. *Bioinformatics*, 27 (12):1667–74, Jun 2011. doi: 10.1093/bioinformatics/btr269.

P. Neuvial L. Jacob and S. Dudoit. Gains in power from structured two-sample tests of means on graphs. Technical Report arXiv:q-bio/1009.5173v1, arXiv, Berkeley University, 2010.

Paolo Martini, Gabriele Sales, M Sofia Massa, Monica Chiogna, and Chiara Romualdi. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res*, 41(1):e19, Jan 2013. doi: 10.1093/nar/gks866.

Rui Alves, Fernando Antunes, and Armindo Salvador. Tools for kinetic modeling of biochemical networks. *Nat Biotechnol*, 24(6):667–72, Jun 2006. doi: 10.1038/nbt0606-667.

Jitao David Zhang and Stefan Wiemann. Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, 25(11):1470–1, Jun 2009. doi: 10.1093/bioinformatics/btp167.

Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Geeta Joshi-Tope, Gopal Gopinath, David Croft, Bernard de Bono, Marc Gillespie, Bijay Jassal, Suzanna Lewis, Lisa Matthews, Guanming Wu, Ewan Birney, and Lincoln Stein. Reactome: a knowledge base of biologic pathways and processes. *Genome biology*, 8(3):R39, 2007. ISSN 1465-6914. doi: 10.1186/gb-2007-8-3-r39. URL `http://www.ncbi.nlm.nih.gov/pubmed/17367534`. PMID: 17367534.

Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic Acids Res*, 37(Database issue):D674–9, Jan 2009. doi: 10.1093/nar/gkn653.

Robert Gentleman. *Bioinformatics and computational biology solutions using R and Bioconductor*. Birkhäuser, August 2005. ISBN 9780387251462.

Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-Sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, Jan 2009. doi: 10.1093/bioinformatics/btn577.

Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–2, Feb 2011. doi: 10.1093/bioinformatics/btq675.

Lisa M Ooms, Kristy A Horan, Parvin Rahman, Gillian Seaton, Rajendra Gurung, Dharini S Kethesparan, and Christina A Mitchell. The role of the inositol polyphosphate 5-phosphatases in cellular function and human disease. *Biochem J*, 419(1): 29–49, Apr 2009. doi: 10.1042/BJ20081673.

Davide Ruggero and Nahum Sonenberg. The akt of translational control. *Oncogene*, 24 (50):7426–34, Nov 2005. doi: 10.1038/sj.onc.1209098.

T Kitamura, Y Kitamura, S Kuroda, Y Hino, M Ando, K Kotani, H Konishi, H Matsuzaki, U Kikkawa, W Ogawa, and M Kasuga. Insulin-induced phosphorylation and activation of cyclic nucleotide phosphodiesterase 3b by the serine-threonine kinase akt. *Mol Cell Biol*, 19(9):6286–96, Sep 1999.

M Hołlysz, N Derebecka-Hołysz, and W H Trzeciak. Transcription of lipe gene encoding hormone-sensitive lipase/cholesteryl esterase is regulated by sf-1 in human adrenocortical cells: involvement of protein kinase a signal transduction pathway. *J Mol Endocrinol*, 46(1):29–36, Feb 2011. doi: 10.1677/JME-10-0035.

Sabina Chiaretti, Xiaochun Li, Robert Gentleman, Antonella Vitale, Kathy S Wang, Franco Mandelli, Robin Foà, and Jerome Ritz. Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *Clin Cancer Res*, 11(20):7209–19, Oct 2005. doi: 10.1158/1078-0432.CCR-04-2165.

Manhong Dai, Pinglang Wang, Andrew D Boyd, Georgi Kostov, Brian Athey, Edward G Jones, William E Bunney, Richard M Myers, Terry P Speed, Huda Akil, Stanley J Watson, and Fan Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res*, 33(20):e175, 2005. doi: 10.1093/nar/gni179.

Fangxin Hong, Rainer Breitling, Connor W McEntee, Ben S Wittner, Jennifer L Nemhauser, and Joanne Chory. Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–7, Nov 2006. doi: 10.1093/bioinformatics/btl476.

Gabriele Sales, Enrica Calura, Duccio Cavalieri, and Chiara Romualdi. graphite - a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13:20, 2012. doi: 10.1186/1471-2105-13-20.

Harvey A. Risch, Loraine D. Marrett, Meera Jain, and Geoffrey R. Howe. Differences in risk factors for epithelial ovarian cancer by histologic type results of a case-control study. *American Journal of Epidemiology*, 144(4):363–372, August 1996. ISSN 0002-9262, 1476-6256. URL `http://aje.oxfordjournals.org/content/144/4/363`.

Jason Madore, Fengge Ren, Ali Filali-Mouhim, Lilia Sanchez, Martin Köbel, Patricia N Tonin, David Huntsman, Diane M Provencher, and Anne-Marie Mes-Masson. Characterization of the molecular differences between ovarian endometrioid carcinoma and ovarian serous carcinoma. *The Journal of pathology*, 220(3):392–400, February 2010. ISSN 1096-9896. doi: 10.1002/path.2659. PMID: 19967725.

Hiroaki Itamochi, Junzo Kigawa, Ryoji Akeshima, Shinya Sato, Shunji Kamazawa, Masakuni Takahashi, Yasunobu Kanamori, Mitsuaki Suzuki, Michitaka Ohwada, and Naoki Terakawa. Mechanisms of cisplatin resistance in clear cell carcinoma of the ovary. *Oncology*, 62(4):349–353, 2002. ISSN 0030-2414. doi: 65067. PMID: 12138243.

Gianpiero Polverino, Fabio Parazzini, Giovanna Stellato, Giovanna Scarfone, Sonia Cipriani, and Giorgio Bolis. Survival and prognostic factors of women with advanced ovarian cancer and complete response after a carboplatin-paclitaxel chemotherapy. *Gynecologic oncology*, 99(2):343–347, November 2005. ISSN 0090-8258. doi: 10.1016/j.ygyno.2005.06.008. PMID: 16051334.

J Alexandre, I Ray-Coquard, F Selle, A Floquet, P Cottu, B Weber, C Falandry, D Lebrun, and E Pujade-Lauraine. Mucinous advanced epithelial ovarian carcinoma:

clinical presentation and sensitivity to platinum-paclitaxel-based chemotherapy, the GINECO experience. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, 21(12):2377–2381, December 2010. ISSN 1569-8041. doi: 10.1093/annonc/mdq257. PMID: 20494964.

Rebecca T Marquez, Keith A Baggerly, Andrea P Patterson, Jinsong Liu, Russell Broaddus, Michael Frumovitz, Edward N Atkinson, David I Smith, Lynn Hartmann, David Fishman, Andrew Berchuck, Regina Whitaker, David M Gershenson, Gordon B Mills, Jr Bast, Robert C, and Karen H Lu. Patterns of gene expression in different histotypes of epithelial ovarian cancer correlate with those in normal fallopian tube, endometrium, and colon. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 11(17):6116–6126, September 2005. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-04-2509. PMID: 16144910.

J Baptist Trimbos, Mahesh Parmar, Ignace Vergote, David Guthrie, Giorgio Bolis, Nicoletta Colombo, Jan B Vermorken, Valter Torri, Constantino Mangioni, Sergio Pecorelli, Andrea Lissoni, and Ann Marie Swart. International collaborative ovarian neoplasm trial 1 and adjuvant ChemoTherapy in ovarian neoplasm trial: two parallel randomized phase III trials of adjuvant chemotherapy in patients with early-stage ovarian carcinoma. *Journal of the National Cancer Institute*, 95(2):105–112, January 2003. ISSN 0027-8874. URL `http://www.ncbi.nlm.nih.gov/pubmed/12529343`. PMID: 12529343.

B.m. Bolstad, R.a Irizarry, M. Åstrand, and T.p. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185 –193, January 2003. doi: 10.1093/bioinformatics/19.2.185. URL `http://bioinformatics.oxfordjournals.org/content/19/2/185.abstract`.

Wynn L Walker, Isaac H Liao, Donald L Gilbert, Brenda Wong, Katherine S Pollard, Charles E McCulloch, Lisa Lit, and Frank R Sharp. Empirical bayes accomodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from duchenne muscular dystrophy patients. *BMC Genomics*, 9(1):494, October 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-494. URL `http://www.biomedcentral.com/1471-2164/9/494`.

Gordon K Smyth. *Limma: linear models for microarray data*. Springer, New York, 2005.

Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–75, Feb 2003.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Number Book, Whole. 2010.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004. ISSN 1465-6914. doi: 10.1186/gb-2004-5-10-r80. URL `http://www.ncbi.nlm.nih.gov/pubmed/15461798`. PMID: 15461798.

Hua Zhu, Hao Wu, Xiuping Liu, Biao Li, Yun Chen, Xingcong Ren, Chang-Gong Liu, and Jin-Ming Yang. Regulation of autophagy by a beclin 1-targeted microRNA, miR-30a, in cancer cells. *Autophagy*, 5(6):816–823, August 2009. ISSN 1554-8635. URL `http://www.ncbi.nlm.nih.gov/pubmed/19535919`. PMID: 19535919.

Peter Kreuzaler and Christine J Watson. Killing a cancer: what are the alternatives? *Nature reviews. Cancer*, 12(6):411–424, June 2012. ISSN 1474-1768. doi: 10.1038/nrc3264. URL `http://www.ncbi.nlm.nih.gov/pubmed/22576162`. PMID: 22576162.

Christian J Braun, Xin Zhang, Irina Savelyeva, Sonja Wolff, Ute M Moll, Troels Schepeler, Torben F Ørntoft, Claus L Andersen, and Matthias Dobbelstein. p53-responsive MicroRNAs 192 and 215 are capable of inducing cell cycle arrest. *Cancer Research*, 68(24):10094–10104, December 2008. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-08-1569. URL `http://cancerres.aacrjournals.org/content/68/24/10094.full`.

Kimihiro Hino, Kiichiro Tsuchiya, Taro Fukao, Kotaro Kiga, Ryuichi Okamoto, Takanori Kanai, and Mamoru Watanabe. Inducible expression of microRNA-194 is regulated by HNF-1alpha during intestinal epithelial cell differentiation. *RNA (New York, N.Y.)*, 14(7):1433–1442, July 2008. ISSN 1469-9001. doi: 10.1261/rna.810208. URL `http://www.ncbi.nlm.nih.gov/pubmed/18492795`. PMID: 18492795.

Lina Ma, Yanyan Huang, Wangyu Zhu, Shiquan Zhou, Jihang Zhou, Fang Zeng, Xiaoguang Liu, Yongkui Zhang, and Jun Yu. An integrated analysis of miRNA and mRNA expressions in non-small cell lung cancers. *PloS One*, 6(10):e26502, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0026502. URL `http://www.ncbi.nlm.nih.gov/pubmed/22046296`. PMID: 22046296.

Zhipeng Meng, Xianghui Fu, Xiaosong Chen, Samuel Zeng, Yan Tian, Richard Jove, Rongzhen Xu, and Wendong Huang. miR-194 is a marker of hepatic epithelial cells and suppresses metastasis of liver cancer cells in mice. *Hepatology*, 52(6):2148–2157, December 2010. ISSN 02709139. doi: 10.1002/hep.23915. URL `http://doi.wiley.com/10.1002/hep.23915`.

Bo Song, Yuan Wang, Kenji Kudo, Elaine J. Gavin, Yaguang Xi, and Jingfang Ju. miR-192 regulates dihydrofolate reductase and cellular proliferation through the p53-miRNA circuit. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(24):8080–8086, December 2008. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-08-1422. PMID: 19088023 PMCID: 2653201.

Chao Cheng, Xuping Fu, Pedro Alves, and Mark Gerstein. mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. *Genome Biology*, 10(9):R90, 2009. ISSN 1465-6906. doi: 10.1186/gb-2009-10-9-r90. PMID: 19723326 PMCID: 2768979.

Gabriele Sales, Alessandro Coppe, Andrea Bisognin, Marta Biasiolo, Stefania Bortoluzzi, and Chiara Romualdi. MAGIA, a web-based tool for miRNA and genes integrated analysis. *Nucleic Acids Research*, 38(Web Server issue):W352–359, July 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq423. URL `http://www.ncbi.nlm.nih.gov/pubmed/20484379`. PMID: 20484379.

Andrea Bisognin, Gabriele Sales, Alessandro Coppe, Stefania Bortoluzzi, and Chiara Romualdi. MAGIA2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic acids research*, 40(Web Server issue):W13–21, July 2012. ISSN 1362-4962. doi: 10.1093/nar/gks460. URL `http://www.ncbi.nlm.nih.gov/pubmed/22618880`. PMID: 22618880.

Mirko Marabese, Sergio Marchini, Eleonora Marrazzo, Pietro Mariani, Dario Cattaneo, Roldano Fossati, Anna Compagnoni, Mauro Signorelli, Ute M Moll, A Maria Codegoni, and Massimo Broggini. Expression levels of p53 and p73 isoforms in stage i and stage iii ovarian cancer. *Eur J Cancer*, 44(1):131–41, Jan 2008. doi: 10.1016/j.ejca.2007.10.011.

J Prat. Ovarian carcinomas: five distinct diseases with different origins, genetic alterations, and clinicopathological features. *Virchows Archiv*, 460(3):237–249, 2012b. ISSN 0945-6317. doi: 10.1007/s00428-012-1203-5. URL `http://www.springerlink.com/content/l57131633804g001/abstract/`.

Joo Heon Kim, Sun Young Yoon, Chang-Nam Kim, Joung Hyuck Joo, Sang Kyoung Moon, In Seong Choe, Yong-Kyung Choe, and Jae Wha Kim. The bmi-1 oncoprotein is overexpressed in human colorectal cancer and correlates with the reduced p16INK4a/p14ARF proteins. *Cancer Letters*, 203(2):217–224, January 2004. ISSN 0304-3835. doi: 10.1016/j.canlet.2003.07.009. URL `http://www.sciencedirect.com/science/article/pii/S030438350300692X`.

Y Zhang, Y Xiong, and W G Yarbrough. ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the rb and p53 tumor suppression pathways. *Cell*, 92(6):725–734, March 1998. ISSN 0092-8674. URL `http://www.ncbi.nlm.nih.gov/pubmed/9529249`. PMID: 9529249.

Jian Liu, Guowu Yu, Yanyan Zhao, Dengpan Zhao, Ying Wang, Lu Wang, Jiang Liu, Lei Li, Yu Zeng, Yongyan Dang, Chuangui Wang, Guang Gao, Weiwen Long, David M. Lonard, Shanlou Qiao, Ming-Jer Tsai, Bianhong Zhang, Honglin Luo, and Xiaotao Li. REG modulates p53 activity by regulating its cellular localization. *Journal of Cell Science*, 123(23):4076–4084, December 2010. ISSN 0021-9533. doi: 10.1242/jcs.067405. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2987440/`. PMID: 21084564 PMCID: PMC2987440.

Alo Nag, Srilata Bagchi, and Pradip Raychaudhuri. Cul4A physically associates with MDM2 and participates in the proteolysis of p53. *Cancer research*, 64(22):8152–8155, November 2004. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-04-2598. URL `http://www.ncbi.nlm.nih.gov/pubmed/15548678`. PMID: 15548678.

Flavia Pichiorri, Sung-Suk Suh, Alberto Rocci, Luciana De Luca, Cristian Taccioli, Ramasamy Santhanam, Wenchao Zhou, Jr Benson, Don M, Craig Hofmainster, Hansjuerg Alder, Michela Garofalo, Gianpiero Di Leva, Stefano Volinia, Huey-Jen Lin, Danilo Perrotti, Michael Kuehl, Rami I Aqeilan, Antonio Palumbo, and Carlo M Croce. Downregulation of p53-inducible microRNAs 192, 194, and 215 impairs the p53/MDM2 autoregulatory loop in multiple myeloma development. *Cancer Cell*, 18 (4):367–381, October 2010. ISSN 1878-3686. doi: 10.1016/j.ccr.2010.09.005. URL `http://www.ncbi.nlm.nih.gov/pubmed/20951946`. PMID: 20951946.

Peixin Dong, Masanori Kaneuchi, Hidemichi Watari, Junichi Hamada, Satoko Sudo, Jingfang Ju, and Noriaki Sakuragi. MicroRNA-194 inhibits epithelial to mesenchymal transition of endometrial cancer cells by targeting oncogene BMI-1. *Molecular Cancer*, 10:99, August 2011. ISSN 1476-4598. doi: 10.1186/1476-4598-10-99. PMID: 21851624 PMCID: 3173388.

Oliver Stehling, Ajay A Vashisht, Judita Mascarenhas, Zophonias O Jonsson, Tanu Sharma, Daili J A Netz, Antonio J Pierik, James A Wohlschlegel, and Roland Lill. MMS19 assembles iron-sulfur proteins required for DNA metabolism and genomic integrity. *Science (New York, N.Y.)*, 337(6091):195–199, July 2012. ISSN 1095-9203. doi: 10.1126/science.1219723. URL `http://www.ncbi.nlm.nih.gov/pubmed/22678362`. PMID: 22678362.

Shipeng Feng, Shujie Cong, Xin Zhang, Xichen Bao, Wei Wang, Huiping Li, Zhe Wang, Guoxin Wang, Jianzhen Xu, Bowen Du, Dezhong Qu, Wei Xiong, Menghui Yin, Xiaoshuai Ren, Feifei Wang, Jianxing He, and Biliang Zhang. MicroRNA-192 targeting retinoblastoma 1 inhibits cell proliferation and induces cell apoptosis in lung cancer cells. *Nucleic Acids Research*, 39(15):6669–6678, August 2011. ISSN 0305-1048. doi: 10.1093/nar/gkr232. PMID: 21511813 PMCID: 3159440.

Shuang Yang, Jun Du, Zhaoqi Wang, Jidong Yan, Wei Yuan, Jie Zhang, and Tianhui Zhu. Dual mechanism of deltaEF1 expression regulated by bone morphogenetic protein-6 in breast cancer. *The International Journal of Biochemistry & Cell Biology*, 41(4):853–861, April 2009. ISSN 1878-5875. doi: 10.1016/j.biocel.2008.08.030. URL `http://www.ncbi.nlm.nih.gov/pubmed/18805502`. PMID: 18805502.

A Honig, C Weidler, S Häusler, M Krockenberger, S Buchholz, F Köster, S E Segerer, J Dietl, and J B Engel. Overexpression of polycomb protein BMI-1 in human specimens of breast, ovarian, endometrial and cervical cancer. *Anticancer research*, 30 (5):1559–1564, May 2010. ISSN 1791-7530. URL `http://www.ncbi.nlm.nih.gov/pubmed/20592341`. PMID: 20592341.

F Zhang, L Sui, and Tao Xin. Correlations of BMI-1 expression and telomerase activity in ovarian cancer tissues. *Experimental oncology*, 30(1):70–74, March 2008. ISSN 1812-9269. URL `http://www.ncbi.nlm.nih.gov/pubmed/18438345`. PMID: 18438345.

Sergio Marchini, Robert Fruscio, Luca Clivio, Luca Beltrame, Luca Porcu, Ilaria Fuso Nerini, Duccio Cavalieri, Giovanna Chiorino, Giorgio Cattoretti, Costantino Mangioni, Rodolfo Milani, Valter Torri, Chiara Romualdi, Alberto Zambelli, Michela Romano, Mauro Signorelli, Silvana di Giandomenico, and Maurizio D'Incalci. Resistance to platinum-based chemotherapy is associated with epithelial to mesenchymal transition in epithelial ovarian cancer. *Eur J Cancer*, 49(2):520–30, Jan 2013. doi: 10.1016/j.ejca.2012.06.026.