POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Ingegneria Biomedica

A MULTI-OMICS APPROACH TO STUDY

MECHANISMS OF ACTION OF

ALL-TRANS RETINOIC ACID IN BREAST CANCER

Relatore: Prof. Linda Pattini

Correlatori: Dr. Marco Bolis

Dr. Maddalena Fratelli

Tesi di Laurea di:

Arianna Vallerga

Matr.: 814668

Anno Accademico 2017/2018

Et rien n'est tel que le rêve pour engendrer l’avenir.

Utopie aujourd'hui, chair et os demain.

Victor Hugo

Table of Contents

List of Figures

List of Tables

Summary

Breast cancer is the most commonly diagnosed malignancy and the leading cause of cancer death in women worldwide. The highest incidence rates were registered in Western and Northern Europe, Australia/New Zealand and North America; intermediate rates in South America, the Caribbean and Northern Africa; low rates in Sub-Saharan Africa and Asia. Factors that influence this variation in incidence rates are related to differences in reproductive and hormonal status (women in more developed countries have fewer children and a late age at first birth, use oral contraceptives or hormone replacement therapies) or in lifestyle (sedentary lifestyle, higher levels of obesity and higher alcohol consumption). Moreover in developed countries the screening programs have raised the rate of incidence, by permitting diagnoses of cancer that otherwise would have remained undiagnosed. The rate of survival for female breast cancer in developed countries is higher than for most of other types of cancer and this is due both to earlier diagnosis, made possible by screening programs, and to improvement in therapy. Despite the decrease in mortality rate registered during the last years, breast carcinomas still remain the leading cause of cancer death in female worldwide, accounting for about 14% of all cancer deaths.

At present, the broad heterogeneity observed among breast cancer reflects the well-accepted notion that there is not just one disease with disparate variant subtypes, but that breast cancer instead represents a collection of distinct neoplastic diseases of the breast and the cells composing it. Behind this complexity, several systems have been developed to classify this very highly heterogeneous disease and possibly to get information about tumour behaviour and provide more effective therapies. Histological classification categorizes breast cancer either in "*in situ*" , which do not grow into or invade normal tissues within or beyond the breast, or "invasive", which do grow into normal, healthy tissues. However, this classification relies only on histological characteristics, without taking into account molecular markers or morphological features that can be helpful for prognosis or therapy. In addition to this classification, it is crucial to assess the receptor status of a tumour, as it may determine the possibility of using targeted treatments in cancer therapy.

On this basis, various subgroups can be identified, according to their profile of gene expression and positivity to oestrogen receptor (ER), progesterone receptor (PR) and the

tyrosine kinase receptor, HER2. Moreover, breast tumours may be endowed with a different complement of the three above mentioned receptors (ER+/PR+/HER2+; ER+/PR+/HER2-; ER+/PR-/HER2+; ER+/PR-/HER2-; ER-/PR+/HER2-; ER-/PR-/HER2+; ER-/PR-/HER2-). Tumours that lack expression of all three receptors are defined as Triple Negative Breast Cancer (TNBC).

In recent years several studies on gene expression profiles have been conducted using high-throughput technologies, in order to identify molecular subtypes of breast cancer, to allow a better understanding of the complexity of the disease.

Based on hierarchical clustering of gene expression microarrays, six subtypes of breast cancers have been identified: the luminal A breast cancer is the most common subtype accounting for 50 – 60% of all diagnoses; it's characterized by the expression of ER and the absence of HER2 over-expression. The luminal B breast cancer accounts for 10–20% of total, it often expresses HER2 and may express low level of oestrogen receptor. The HER2 positive subtype represents about 15-20% of all breast tumours and is characterized by expression of the HER2 gene, genes associated to its pathway and genes associated to cellular proliferation. The basal-like subtype, which represents 10-20% of total breast cancers, is characterized by expression of genes characteristic of the myo-epithelial (or basal) cells. In general, this subtype does not express ER, PR and HER2 and in clinical practice, it is often referred to as triple negative breast cancer (TNBC). The normal breast subtype is characterized by the expression of genes typical of adipose tissue and accounts for 5-10% of breast cancers; it lacks the expression of hormonal receptors and HER2 and so tumours belonging to this subtype are TNBC, but they are not basal-like cancers since they lack the expression of some genes characteristic of that category. Finally, the claudin-low subtype has been recently identified with a low expression of genes that encode for proteins involved in the formation of tight junctions, including Claudins and E-cadherin. It is rare and characterized by the absence of the oestrogen receptor, progesterone receptor and HER2 expression.

In 2009, Parker and colleagues introduced a new system of analysis, PAM50 (Prediction Analysis of Microarrays), to select a minimum set of genes, whose expression can predict the molecular subtype of a tumour. The PAM50 gene sets allows to obtain a classification similar to the one described above; it is based on a selection of gene sets consisting of a

large number of "intrinsic" genes, and therefore can be used in the clinics to define the molecular phenotype of the tumours.

Molecular classification of breast cancer based on gene expression patterns provides a connection between molecular biology and the behaviour of cancer cells in the corresponding subtypes. However, molecular classification of breast cancer has not yet reached clinical implementation as a routine aspect of patient management.

Although an immunohistochemical staining proxy can be used to stratify and classify breast cancers in a clinical setting, the correspondence between clinical and molecular is not yet remarkable.

The term retinoids refers to a group of compounds comprising metabolites and analogues of vitamin A, both natural and synthetic. The natural retinoids are essential components of diet and physiological regulators of many essential biological processes, such as embryonic development, metabolism and haematopoiesis. In adult mammals, retinoids such as All-Trans Retinoic Acid (ATRA), control homeostasis of different organs and tissues.

All-trans retinoic acid (ATRA) is a small lipophilic molecule and an important regulator of gene expression. The biological action of ATRA and its derivatives is mediated by two classes of nuclear receptors for retinoids called Retinoic Acid Receptor (RAR) and Retinoic X Receptor (RXR). The receptors are ligand-dependent transcription factors that control the activity of several target genes either through a direct or indirect mechanism. Both receptor subtypes exist in three different forms known as alpha, beta and gamma, encoded by different genes (RARA, RARB, RARG/ RXRA, RXRB, RXRG). Each subtype comprises two or more isoforms, which differ in the N-terminal region and are generated by a different promoter or by alternative splicing mechanisms. RAR and RXR receptors form stable hetero-dimers (RAR/RXR) that bind to specific sequences on DNA, called Retinoic Acid Responsive Elements (RAREs), localized within the promoter of target genes. ATRA modulates transcription through different mechanisms: it can directly modulate expression of target genes, through the interaction of the RAR/RXR with a group of co-activators and co-repressors or, in the absence of ligand, the RAR/RXR dimer is bound to the RAREs sequences on DNA and is associated with a complex of co-repressors that inhibit the transcription of target genes.

Tumorigenesis is a multistep process characterized by a series of inherited or acquired genetic changes (mutations, chromosomal rearrangements, epigenetic phenomena), leading

to a disruption of cellular homeostasis and development of the neoplastic process. Several lines of evidence indicate an important role of retinoids in homeostasis.

Mechanisms by which retinoids exert their antitumor activity have not yet been completely clarified, although studies *in vitro* and *in vivo* have shown the ability of retinoids to inhibit proliferation, induce differentiation and apoptosis, making these molecules of therapeutic interest. It is clear that ATRA is able to act through different genomic mechanisms as well as to interact with other intracellular signalling systems, that provide the basis for its pleiotropic action.

Currently, the best example of the anticancer action of retinoids is the use of ATRA in the treatment of patients suffering from acute promyelocytic leukaemia (APL).

The retinoids have been investigated extensively for the prevention and treatment of cancer, predominantly because of their ability to induce cellular differentiation and to arrest proliferation. Systemic retinoids are approved by the U.S. Food and Drug Administration (FDA) also for treatment of cutaneous T-cell lymphoma, other than acute promyelocytic leukemia. The anti-leukemic action of ATRA is not primarily cytotoxic and it is the result of a direct cyto-differentiating action followed by a secondary apoptotic response rendering ATRA the first example of clinically useful cyto-differentiating agent. The use of ATRA in APL is also an example of targeted therapy, as the retinoids' primary target is PML-RARα, the aberrant retinoid receptor expressed into the leukemic cell. To date, more than 85% of patients with APL achieve complete remission following treatment with ATRA in combination with chemotherapy. The unique mechanism of action and the results obtained in APL has raised enthusiasm in generalizing the use of retinoids to other types of cancers, including breast cancer. Pre-clinical data support the idea that ATRA is a promising agent in the treatment and chemoprevention of certain subgroups of breast cancer, with particular reference to ER+ and HER2+ tumours characterized by co-amplification of the retinoic acid receptor alpha gene. There is also a low proportion of triple negative breast cancers which show sensitivity to this unusual anti-tumour agent.

To evaluate the response of breast cancer cell lines to the anti-proliferative effect exerted by retinoic acid, Bolis and colleagues first defined the profile of ATRA-sensitivity in a panel of 48 breast cancer cell lines of the Cancer Cell Lines Encyclopedia (CCLE), well- representing the heterogeneity of the disease. The drug response of each cell line has been quantified by computation of a sensitivity score (ATRA-*score*).

ATRA-score was computed on cell-lines that have been treated with vehicle (DMSO) and 5 logarithmically increasing concentration of ATRA (0.001-10.0 μM) for 9 days and its value is calculated from the relative growth-inhibition (GI) data (ATRA vs. vehicle).

To define the *ATRA-score* sensitivity metric, they fitted growth-inhibition curves relative to DMSO-treated controls and computed the area under the curve (*AUC*) and the maximal inhibitory effect (A_{max}). At this point, ATRA-score values, which are equal to the log2 transformation of AUC x A_{max} , are rescaled in a range between 0 and 1, zero indicating total resistance and one standing for maximum sensitivity.

Moreover, Bolis and colleagues developed a tool capable of predicting ATRA-sensitivity, exploiting the association between this *in vitro* profiling and basal gene-expression data.

In previous studies, a large panel of breast cancer cell lines (>50 lines) representative of the heterogeneity of the disease, have been profiled for their sensitivity to the anti-proliferative action of ATRA. They used a network-guided approach to develop a generalized model based on 21 genes (ATRA-*21*) capable of predicting ATRA-sensitivity across tumour types other than breast cancer.

To identify gene-networks and gene pathways involved in the anti-proliferative action of ATRA, in this study we performed total RNA-sequencing experiments in a panel of 16 sensitive and resistant cell lines, before and after treatment with the retinoid (1.0 μM) for 24 hours. To better represent breast cancer heterogeneity, cell lines have been chosen based on their phenotype (8 luminal, which includes luminal A, luminal B and HER2+, and 8 basal-like) and their widely variable sensitivity to pharmacological treatment with retinoic acid.

Alignment of high-throughput paired-end reads derived from RNA-sequencing experiments to the reference genome has been performed. Genome-generation was performed using the comprehensive gene annotations present in *Gencode*; in particular, the v27 release of the GTF file has been used. As many RNA-sequencing aligners suffer from high mapping error rates, read length limitation or mapping biases, sequence–alignment to reference human genome (hg38) has been performed using STAR (Spliced Transcript Alignment to a Reference) sequence-aligner, which was designed specifically to align non-contiguous sequences directly to the reference genome, using a novel strategy for these spliced alignments.

Differential gene expression analysis was conducted exloiting the R package *DESeq2,* which provides methods to test for differential expression by the use of a negative binomial generalized linear model.

After this first phase of differential expression analysis, we performed a test to verify the correlation between fold changes of differentially expressed genes and the predicted response of cell lines to pharmacological treatment, in terms of ATRA-*score*. To this aim, we computed both Pearson's product moment correlation coefficient and Spearman' rank correlation coefficient. We selected only those genes with a Pearson's coefficient R or a Spearman's coefficient RHO <0.01. To further select only those genes showing a variation across samples that is sufficient to result in a biologically significant action, we computed a variation coefficient, defined as $Sd\{Matrix[i,]/mean(Matrix[i,])\} * 100$.

On the basis of this additionally parameter, a more restrictive selection of genes has been performed, considering only those genes with VC>50%.

Selected genes have been organized in networks based on protein – protein interactions, to a more precise visualization of possible interaction mechanisms induced.

Moreover, gene set enrichment analysis (GSEA) on sequencing data led to the identification of ATRA-dependent pathways and gene-networks with significance for the anti-tumour activity of the retinoid: "interferon-dependent" and immune modulatory pathways are found to be strictly up- regulated after treatment with ATRA. Genes and pathways that are down-regulated specifically in ATRA-sensitive cell lines, are linked to cell proliferation and cell cycle progression which are tightly connected with the antiproliferative effect exerted by retinoic acid, and thus can be considered part of a downstream mechanism of action.

We inspected ChIP-sequencing data from a public database of two forms of RARs transcription factors (RARA, RARG) in one breast cancer cell line treated with retinoic acid (MCF-7): we wanted to evaluate which of the more central genes in our response network were directly perturbated by the binding in the regions of their promoters of the ATRA-activated transcription factors. To this purpose, raw data obtained in "FASTQ" file format have been aligned to the reference genome (hg38), using the Burrows-Wheeler Alignment Tool (BWA), a read alignment algorithm that is based on the backward search with Burrows-Wheeler Transform (BWT). This represents an effective method to align short sequencing reads (50 bp) against a large reference sequence, such as the human genome.

Next, we used a command line tool designed by Zhang and Liu, MACS (Model-based Analysis of ChIP-Sequencing), to analyse pre-processed ChIP-Sequencing data. Given the ChIP-Sequencing data with the correspondent control sample, this peak-finding algorithm can be used to identify transcription factor binding sites (or even histone modifications, if necessary): it uses a dynamic Poisson distribution, which captures local biases in the genome, allowing for robust predictions and giving fewer false positives than the other available methods. Last, we annotated the identified peaks with an Ensembl based annotation package for Homo Sapiens, Ensembl version 86 (*EnsDb.Hsapiens.v86*).

As result, we obtained a list of genes that are part of the above-mentioned interferon signalling, which have been identified as directs targets for RARA or RARG transcription factors; however, some of the most crucial genes involved in such pathways cannot be included in the list. Among the genes identified as RARA direct targets, is of particular interest the presence of "interferon- related" genes, such as DTX3L and PARP9, such as the presence of one of the genes that encodes a protein that is part of the MHC-I complex, HLA-E. In contrast, one of the more important transcription factors involved in the regulation of the interferon signalling, IRF1, is not a direct target, neither for RARA transcription factor, nor for RARG.

Methylation data available for a panel of almost 40 un-treated breast cancer cell lines have been investigated, to find out whether there is a correlation between the basal methylation levels of genes necessary to trigger the mechanism of response to retinoids, and the sensitivity of cell lines to ATRA. All the data were obtained through the HumanMethylation450 BeadChip Array Platform: after a few normalization steps, methylation data have been tested for association with a defined parameter, the ATRA-*score*. To this aim, *dmpFinder* function implemented in the package *minfi* in R environment has been used, which tested each genomic position for association between methylation and a "phenotype", our defined parameter. Given the ATRA-*score* as a continuous parameter, association has been tested with linear regression. Finally, differentially methylated probes have been annotated, exploiting the information stored in the Bioconductor package "*IlluminaHumanMethylation450kanno.ilmn12.hg19.*

Again, a few genes involved in the interferon-related mechanism have found to have a correlation between their methylation levels and the activation of the response to retinoids.

In particular, 73 out of the total 298 identified genes are well interesting, because of their highly interconnected position in the network defined during the differential expression analysis. In particular, among the genes with a higher methylation level we can find TLR3, a crucial part of innate immune response, whose transcription is deeply inducted after treatment with retinoic acid. Moreover, two other genes involved in the innate immune response and inducted by treatment with retinoic acid, HLA-E and PSMB8, have high methylation level in basal condition, correlated with the ATRA-*score.*

In the second part of the study we took again into account the RNA-sequencing data to quantify possible transcription of repetitive elements from retroviral DNA, which are known to be widely distributed in the human genome: we hypostasized that they can be the cause of the above-mentioned interferon-driven immune system reactivation.

To quantify expression of these transposable elements, we retrieved their genomic positions from RepeatMasker database (http://www.repeatmasker.org/). These coordinates were assembled into a customized annotation file (*gene transfer file*, GTF), which was used to determine the abundance of all retroviral-derived transcripts, by using *FeatureCounts*. To avoid detection of false positives, we discarded all transposable elements that show any overlap to known gene-associated exons, according to *Gencode* annotations. Afterwards, viral RNA abundance was normalized for library size and tested for differential expression between ATRA-treated and untreated samples, using the same approach as described in Material and Methods.

A widely distributed up-regulation of these transcripts can be observed: induction (fold change) of the transcriptomic regulation slightly correlates with the sensitivity to ATRA-treatment. Cell lines which are completely resistant to the pharmacological treatment, display no induction; then transcriptomic up-regulation grows with an increasing ATRA-*score.* Despite the presence of a general trend of correlation between the induction of the retroviral elements transcription and the sensitivity to retinoids, of particular interest is the transcriptional effect on a few cell lines (CAMA1, ZR751), which doesn't follow the global behaviour. To better understand the reasons of such tendency, we proceeded with further investigations: it has been shown that this general course tightly correlates with the expression levels of RARA in each cell line.

All things considered, we identified gene-networks whose expression is selectively modulated by ATRA in retinoid-sensitive luminal and triple-negative cell lines as well as

other gene-networks which are commonly regulated in both cell groups. Among the networks stimulated by ATRA, the group of genes involved in interferon- responses is of particular interest, as it indicates that the retinoid exerts a strong and specific immuno-modulatory action in sensitive breast cancer cell lines.

We are evaluating the functional significance of specific elements of these gene-networks for the anti-tumour-metastatic action of ATRA with the use of silencing and over-expression approaches.

The results obtained in our cellular models provide insights into the molecular mechanisms underlying the anti-tumour action of ATRA in breast cancer. In addition, the sequencing data led to the identification of ATRA-dependent pathways and gene-networks with significance for the anti-tumour activity of the retinoid. Finally, the approach provides information as to potential new molecular targets for the design of rational therapeutic combinations based on ATRA for the treatment and secondary chemo-prevention of certain types of breast cancer.

Beside the anti-proliferative effect described above, our data suggest that the pharmacological treatment with ATRA might also have an immunoregulatory effect on these cells. In particular, it has been observed that there is a dramatic up-regulation of the "Antigen-presentation and assembly/loading of class I MHC" pathways, as well as "Inflammatory responses": this may result in an increased antigen presentation mechanism which may activate innate immune response.

All things considered, this study provides a strong rationale for the combination of ATRA with the immune checkpoint inhibitors.

Sommario

Il carcinoma della mammella è il tumore più frequentemente diagnosticato tra le donne nel mondo, sia nei Paesi economicamente più avanzati che in quelli in via di sviluppo. Nel 2018 sono stati stimati 2 milioni di nuovi casi in tutto il mondo.

La più alta incidenza è stata osservata in Nord America, in Europa del Nord e Ovest e in Oceania, dove sono in aumento la prevalenza dei fattori di rischio e le rilevazioni di tumori allo stadio iniziale. Nell'ultimo decennio in questi Paesi si è registrato però anche un aumento della sopravvivenza, grazie alla disponibilità di programmi di prevenzione primaria e diagnosi precoce (mammografia) e di nuove strategie terapeutiche efficaci, che permettono di migliorare la prognosi delle pazienti. Attualmente il tumore della mammella è la quinta causa di morte per tumore (552,000, il 6,4%). Nei Paesi in via di sviluppo, come Sud e Centro America, Africa e Asia, sono invece in aumento sia l'incidenza che la mortalità; questa tendenza riflette il recente cambiamento nello stile di vita (dieta, obesità) insieme a cambiamenti relativi allo stato ormonale e riproduttivo (minor numero di gravidanze e in età più adulta, allattamento di minore durata, utilizzo di contraccettivi), mancanza di programmi di screening efficaci e, in alcuni casi, limitato accesso ai trattamenti.

In Italia, secondo i dati presentati dall'Associazione Italiana di Oncologia (AIOM) e dall'Associazione Italiana dei Registri Tumori (AIRTUM), si registra un aumento dell'incidenza di tumore della mammella nelle donne, con 52,300 nuovi casi stimati nel 2018. Questo aumento può essere in parte ricondotto all'ampliamento della fascia di screening mammografico in alcune Regioni. È il secondo tipo di tumore più frequente nelle donne, ma la sopravvivenza a 5 anni si avvicina al 90%, con percentuali ancora più alte se diagnosticato ad uno stadio precoce. La mortalità è quindi in diminuzione, seppure con differenze tra le Regioni del Nord e Sud Italia, come risultato di una maggiore prevenzione primaria e del miglioramento delle strategie terapeutiche

Il carcinoma della mammella è una malattia eterogenea che comprende entità distinte in termini di istologia, caratteristiche molecolari, prognosi clinica e risposta ai trattamenti. Questa diversità ha reso più complesso lo sviluppo di classificazioni clinicamente utili per determinare il comportamento di un tumore sulla base delle sue caratteristiche biologiche .

La classificazione istopatologica si basa sulle differenti caratteristiche morfologiche dei tumori. Il tumore della mammella può essere innanzitutto classificato come *in situ* o invasivo. *In situ* significa che il tumore rimane confinato all'interno del tessuto epiteliale in cui si sviluppa, mentre è definito invasivo quando invade il tessuto circostante, diffonde nei linfonodi e vasi sanguigni ed eventualmente in altre aree del corpo.

Negli ultimi decenni è stata dimostrata la fondamentale importanza di due recettori di ormoni steroidei, il recettore degli estrogeni (ER) e del progesterone (PR), e del recettore tirosin-chinasico HER2 (*human epidermal growth factor receptor* 2) per l'eziologia, la prognosi e la terapia dei tumori della mammella. Accanto alla classificazione istopatologica, i tumori della mammella si possono distinguere sulla base dell'espressione dei suddetti recettori, valutata mediante analisi immunoistochimica, che permette di rilevare la presenza della proteina. Inoltre, I tumori della mammella possono essere caratterizzati da diverse complementazioni dei tre recettori sopra citati: (ER+/PR+/HER2+; ER+/PR+/HER2-; ER+/PR-/HER2+; ER+/PR-/HER2-; ER-/PR+/HER2-; ER-/PR-/HER2+; ER-/PR-/HER2-). Infine, i tumori della mammella tripli negativi (TNBC, *triple-negative breast cancer*) costituiscono un gruppo di tumori molto eterogenei, caratterizzati dalla mancanza di espressione del recettore degli estrogeni e del progesterone e dalla mancanza di amplificazione/sovra-espressione di HER2.

L'analisi dell'espressione genica, resa possibile dallo sviluppo di tecniche basate su microarray a cDNA, ha permesso di suddividere i tumori in diversi sottotipi molecolari sulla base della somiglianza del profilo di espressione genica. In questo modo è stata definita una classificazione, detta *intrinseca*, che ha individuato sei diversi sottotipi.

Il sottotipo luminale A (50 -60% di tutti i carcinomi della mammella) comprende tumori caratterizzati da alti livelli di espressione del recettore degli estrogeni e dall'assenza di over-espressione del recettore HER2. Il sottotipo luminale B (10-20% di tutti i carcinomi della mammella) comprende tumori

che esprimono alti livelli di HER2 e spesso bassi livelli del recettore per l'estrogeno.

Il sottotipo HER2-arricchito (15-20% di tutti i carcinomi della mammella) è costituito da tumori che sono prevalentemente *HER2*-amplificati, mostrano elevati livelli di espressione di numerosi geni dell'amplicone ERBB2 e geni associati a meccanismi di proliferazione cellulare. I tumori *basal-like* (10-20% di tutti i carcinomi della mammella) sono caratterizzati da alti livelli di espressione di marcatori delle cellule mioepiteliali basali e di geni che

regolano il ciclo cellulare. Sono inoltre caratterizzati dalla mancanza o da bassi livelli di espressione di ER e dei geni correlati ad ER, incluso PR, e dalla frequente assenza della sovra-espressione/amplificazione di HER2. Per questo motive, in pratica clinica, si fa riferimento a questo tipo di tumore come sottotipo triplo negativo.

Il sottotipo *normal breast-like* (5-10% di tutti i carcinomi della mammella) è costituito da tumori caratterizzati da elevati livelli di espressione di geni tipici del tessuto adiposo e di altri tipi cellulari non epiteliali. Sono caratterizzati dalla mancanza di espressione di ER, PR ed HER2, (sottotipo triplo negativo, ma non basal like, in quanto non esprimono geni caratteristici di quella categoria.

Infine, il sottotipo *Claudine-low* (12-14% dei carcinomi della mammella) è costituito per la maggior parte da carcinomi invasivi, caratterizzati da bassi livelli di espressione di geni coinvolti nelle giunzioni e nell'adesione cellula-cellula, come quelli codificanti per le Claudine 3/4/7 e la E-caderina. Inoltre, sono tumori rari e caratterizzati dall'assenza di recettore per l'estrogeno, per il progesterone ed anche HER2 negativi (tripli negativi).

Nel 2009, Parker e colleghi hanno implementato un sistema di analisi, PAM50 (*prediction analysis of microarrays*), per selezionare un set minimo di geni (50 geni) la cui espressione fosse predittiva di uno specifico sottotipo molecolare. Il set di geni PAM50 permette di ottenere una classificazione in accordo con quella precedente, che si basava sulla selezione di set costituiti da un numero maggiore di geni "intrinsechi", e può essere utilizzato in clinica per definire il fenotipo molecolare del tumore.

La classificazione molecolare del tumore della mammella basata sui pattern di espressione genica costituisce un elemento di connessione tra la biologica molecolare del tumore e il conseguente progredire delle cellule tumorali del corrispondente sottotipo. Tuttavia, la classificazione molecolare del tumore del seno non ha ancora raggiunto un'implementazione standardizzata a livello clinico, poiché la corrispondenza tra gli aspetti clinici e quelli molecolari non è ancora completamente definita.

Con il termine retinoidi ci si riferisce a tutte le molecole strutturalmente e funzionalmente analoghe al retinolo (Vitamina A), sia naturali che sintetiche. La vitamina A e i suoi metaboliti biologicamente attivi (acido retinoico tutto-trans, acido 9-cis retinoico e acido 13-cis retinoico) sono molecole essenziali per lo sviluppo embrionale, il meccanismo della visione e l'omeostasi di numerosi tessuti e sistemi, tra cui il sistema nervoso, immunitario e riproduttivo. A livello cellulare, regolano la proliferazione, il differenziamento e l'apoptosi.

L'acido retinoico tutto-trans (ATRA) è una piccolo vitamina liposolubile, importante regolatrice dell'espressione genica. L'acido retinoico e i suoi derivati regolano infatti l'espressione di geni coinvolti nella crescita e nel differenziamento cellulare attraverso specifici recettori nucleari, RAR (*retinoic acid receptor*) e RXR (*retinoid X receptor*). Questi recettori, in forma di omo- o etero-dimeri (RAR-RXR o RXR-RXR), agiscono da fattori di trascrizione. Sia RAR che RXR presentano ciascuno tre sottotipi recettoriali (α, β, γ) codificati da geni distinti. Per ciascun sottotipo esistono più isoforme, generate per splicing alternativo, che possono avere una differente affinità per i vari retinoidi e mediare differenti funzioni biologiche.

L'acido retinoico tutto-trans è un agonista di tutte le isoforme recettoriali RAR, a cui si lega con la stessa affinità, modulando la trascrizione attraverso diversi meccanismi. In assenza di ligando, il dimero RAR-RXR è costitutivamente legato alle sequenze RARE contenute nei promotori dei geni bersaglio ed è associato a co-repressori , che inibiscono la trascrizione genica. Il legame di ATRA a RAR induce modificazioni conformazionali che possono determinare il rilascio di co-repressori e il reclutamento di co-attivatori con attivazione della trascrizione genica (*meccanismo genomico diretto*). Inoltre, i geni regolati direttamente codificano per proteine coinvolte nel trasporto, metabolismo e trasduzione del segnale dell'acido retinoico stesso e per i fattori di crescita, a loro volta, modulano l'espressione di geni coinvolti nella proliferazione cellulare (*meccanismo genomico indiretto*). In questo modo, ATRA inibisce la crescita arrestando il ciclo cellulare e guida la cellula verso un programma di differenziamento.

La carcinogenesi è un processo caratterizzato dal graduale accumulo di alterazioni genetiche ed epigenetiche responsabile della deregolazione dell'omeostasi cellulare. In questo processo, le cellule sane vanno incontro ad una serie di trasformazioni neoplastiche con formazione di lesioni pre-maligne e sviluppo di carcinomi *in situ* e metastatici.

I retinoidi svolgono un ruolo importante nel mantenimento dell'omeostasi: attraverso la regolazione dell'espressione genica, garantiscono il corretto equilibrio tra crescita e differenziamento cellulare. La perdita della loro attività o la diminuzione dei loro livelli intracellulari è associata ad una crescita cellulare aberrante e allo sviluppo di un'ampia varietà di tumori.

I meccanismi con cui i retinoidi esercitano la loro attività antitumorale non sono ancora stati completamente chiariti, sebbene gli studi *in vitro* e *in vivo* abbiano dimostrato la capacità dei retinoidi di inibire la proliferazione, indurre differenziazione e apoptosi, rendendo queste molecole di interesse terapeutico. È chiaro che ATRA sia in grado di agire attraverso diversi

meccanismi genomici e di interagire con altri sistemi di segnalamento intracellulare, che forniscono la base per la sua azione pleiotropica.

L'acido retinoico tutto-trans è stato il primo agente anti-proliferativo e cito-differenziante ad essere utilizzato in clinica nel trattamento di un raro sottotipo di leucemia mieloide acuta (AML, *acute myeloid leukemia*), la leucemia promielocitica acuta (APL, *acute promyelocytic leukemia).* L'azione anti-leucemica di ATRA non è citotossica, ma è il risultato di un'azione diretta anti-proliferativa e cito-differenziante, seguita da una risposta apoptotica secondaria che rende ATRA il primo esempio di agente cito-differenziante correntemente utilizzato in clinica. L'uso di ATRA in APL è inoltre un esempio della cosiddetta "targeted" therapy, poiché l'obiettivo primario del retinoide è la fusione genica PML-RARα, che determina un'alterazione della funzione recettoriale di RARα.

In combinazione con la chemioterapia (antracicline), questa terapia permette a più dell'85% dei pazienti con APL di andare incontro a remissione completa della malattia.

Grazie al successo di ATRA nell'ambito della leucemia promielocitica acuta, l'interesse per il potenziale utilizzo terapeutico dei retinoidi si è esteso anche ad altri tipi di carcinomi, come il tumore della mammella. Ciò ha portato allo sviluppo di numerosi analoghi sintetici dell'acido retinoico, promettenti agenti cito-differenzianti e pro-apoptotici, e alla disponibilità di una serie di dati ottenuti da numerosi studi preclinici. Questi ultimi, tuttavia, si sono tradotti in un numero molto limitato di studi clinici. Tale insuccesso potrebbe essere dovuto al fatto che gli studi sono stati condotti senza tenere in considerazione l'eterogeneità del tumore della mammella e senza una selezione dei sottotipi tumorali.

Si è quindi reso necessario definire i determinanti molecolari della sensibilità e resistenza ad ATRA nei diversi sottotipi di tumore della mammella.

Per valutare la risposta delle linee cellulari di carcinoma mammario all'effetto antiproliferativo esercitato dall'acido retinoico, Bolis e colleghi hanno definito il profilo di sensibilità ad ATRA in un pannello di 48 linee cellulari di carcinoma mammario, che ben rappresentasse l'eterogeneità della malattia. La risposta farmacologica di ciascuna linea cellulare è stata quantificata per il calcolo finale di un punteggio di sensibilità (ATRA-*score*).

Tale punteggio è stato calcolato su linee cellulari che sono state trattate con veicolo (DMSO) e 5 una concentrazione logaritmicamente crescente di ATRA (0,001-10.0 μM) per 9 giorni.

Il puntegio finale è stato calcolato a partire dai dati relativi alle curve di crescita-inibizione (ATRA vs veicolo).

Per definire la metrica di sensibilità dell'ATRA-*score*, è stato eseguito un fitting delle curve di inibizione della crescita e per ciascuna è stata calcolata l'area sottesa alla curva (AUC) e l'effetto inibitorio massimo (A_{max}). A questo punto, i valori di ATRA-*score*, pari alla trasformazione logaritmica del prodotto AUC x A_{max}, sono stati riscalati in un range compreso tra 0 e 1, dove zero indica la resistenza totale e uno la massima sensibilità.

Inoltre, Bolis e colleghi hanno sviluppato uno strumento in grado di predire la sensibilità al trattamento farmacologico con ATRA, sfruttando l'associazione tra questo profilo di sensitività *in vitro* e i dati disponibili di espressione genica basale. A partire da questi dati, utilizzando un approccio "network-guided" è stato sviluppato un modello basato sui dati di espressione basale di 21 geni (ATRA-21), in grado di predire la sensibilità ad ATRA anche in tipi di tumori diversi dal cancro al seno.

Per identificare i meccanismi molecolari coinvolti nell'azione anti-proliferative di ATRA, in questo studio sono stati condotti esperimenti di sequenziamento di RNA in un pannello di 16 linee cellulari di carcinoma della mammella, sensibili e resistenti ad ATRA, prima e dopo il trattamento con acido retinoico (1,0 μM) per 24 ore. Per meglio rappresentare l'eterogeneità del cancro al seno, le linee cellulari sono state scelte in base al loro fenotipo (8 di fenotipo luminale, comprendenti luminali A, luminali B e HER2 +, e 8 di fenotipo basale) e la loro variabile sensibilità al trattamento farmacologico con acido retinoico. Le sequenze ottenute dagli esperimenti di sequenziamento di RNA sono state allineate al genoma di riferimento. La generazione del genoma è stata effettuata usando le annotazioni genomiche presenti in GENCODE; in particolare, è stata utilizzata la versione 27 (v27) del file GTF(*gene transfer file*).

Poiché molti allineatori di sequenze di RNA soffrono di alti tassi di errore di mappatura, limitazioni nella lunghezza delle sequenze che possono essere allineate o biases nella mappatura, l'allineamento delle sequenze al genoma di riferimento (hg38) è stato eseguito utilizzando STAR (Spliced Transcripts Alignment to a Reference), progettato in modo specifico per l'allineamento di sequenze non contigue, utilizzando una strategia innovativa per sequenze soggette a splicing.

L'analisi di espressione differenziale è tata poi successivamente condotta in ambiente R, attraverso l'utilizzo del pacchetto DESeq, in grado di testare l'espressione differenziale di

geni tra due diverse condizioni sperimentali (controlli e trattamenti con acido retinoico) attraverso l'utilizzo di un modello lineare (GLM) binomiale negativo.

Dopo questa prima fase di analisi, è stato eseguito un test di correlazione tra l'induzione (fold change) dei geni differenzialmente espressi e la risposta predetta di ciascuna linea cellulare, in termini di ATRA-*score*. Per fare ciò, sono stati calcolati sia il coefficiente di correlazione secondo Pearson (R), sia il coefficiente di correlazione secondo Spearman (RHO). Sono stati selezionati solo quei geni con $R<0.01$ o $RHO<0.01$.

Da ultimo, sono stati ulteriormente selezionati solo quei geni che mostrassero una variabilità nei diversi campioni sufficiente a determinarne un' azione biologicamente significativa. Per fare ciò, è stato calcolato il coefficiente di variazione tra i campioni, definito in termini percentuali. Solo i geni con una variazione superiore al 50% sono stati tenuti in considerazione per analisi successive.

I geni selezionati sono stati organizzati in reti basate sulle interazioni proteina-proteina, per una visualizzazione più precisa dei possibili meccanismi di interazione indotti.

Successivamente, l'analisi di arricchimento di espressione genica (GSEA) sui dati di sequenziamento ha condotto all'identificazione dei pathways molecolari la cui attivazione o repressione sia dipendente dall'attività antitumorale dell'acido retinoico. Le vie dipendenti dall'attivazione di interferone e quelli relative all'attivazione della risposta immunitaria sono fortemente up-regolate dopo il trattamento con ATRA.

Al contrario, le vie connesse con la proliferazione e la progressione del ciclo cellulare, sono fortemente down-regolate: questo meccanismo sembra connesso con l'effetto antiproliferativo dell' acido retinoico, e quindi si ipotizza essere parte di un meccanismo di azione a valle.

Abbiamo ispezionato i dati di sequenziamento di immuno-precipitazione di cromatina (ChIP), depositati in un database pubblico, di due forme di fattori di trascrizione RAR -dipendenti (RARA, RARG) in una linea cellulare di cancro al seno trattata con acido retinoico (MCF-7): abbiamo voluto valutare quale dei geni più interconnessi nella rete di risposta al trattamento, siano stati direttamente perturbati dal legame nelle regioni del loro promotore dei fattori di trascrizione attivati da acido retinoico. A questo scopo, i dati grezzi ottenuti dal sequenzaimento sono stati allineati al genoma di riferimento (hg38), utilizzando lo strumento di allineamento Burrows-Wheeler (BWA), un algoritmo basato sull'utilizzo della trasformata di Burrows-Wheeler (BWT). Si tratta infatti di un metodo efficace per

allineare brevi sequenze (50 paia di basi) su sequenze di riferimento molto lunghe, come appunto il genoma umano.

Successivamente, abbiamo utilizzato un algoritmo implementato da Zhang e Liu, MACS (Model-based Analysis of ChIP-Sequencing), per analizzare i dati di sequenziamento di immuno-precipitazione di cromatina pre-processati. Una volta associato il dato di sequenziamento del trattamento con il corrispondente campione di controllo, questo algoritmo di ricerca può essere utilizzato per identificare i siti di legame dei fattore di trascrizione (o anche le modifiche istoniche, se necessario): utilizzando un modello basato su una distribuzione di Poisson dinamica, permette di ottenere robuste previsioni e un basso numero di falsi positivi, confrontato con altri algoritmi disponibili. Infine, i siti di legame identificati sono stati annotati, con un pacchetto di annotazione che fa riferimento alla release *Ensembl* 86 (https://www.ensembl.org/index.html).

Come risultato, abbiamo ottenuto una lista di geni, alcuni dei quali associabili al signalling di interferone, che sono stati identificati come bersagli diretti per fattori di trascrizione RARA o RARG. Tuttavia, alcuni dei geni cruciali coinvolti nelle vie molecolari interferoniche non possono essere inclusi nella lista. Tra i geni identificati come bersagli diretti di RARA, è di particolare interesse la presenza di geni correlati alla via interferonica come DTX3L e PARP9, insieme ad uno dei geni codificanti una proteina parte del complesso MHC-I, HLA-E. Al contrario, uno dei fattori di trascrizione più importanti coinvolti nella regolazione del segnale dell'interferone, IRF1, non è un bersaglio diretto, né per il fattore di trascrizione RARA, né per RARG.

I dati di metilazione disponibili per un pannello di quasi 40 linee cellulari di carcinoma della mammella in condizioni basali sono stati indagati, per scoprire se esistesse una correlazione tra i livelli di metilazione basale di geni necessari per innescare il meccanismo di risposta ai retinoidi, e la sensibilità delle linee cellulari ad ATRA.

Tutti i dati utilizzati, sono stati ottenuti attraverso la piattaforma HumanMethylation450 BeadChip: dopo alcuni passaggi di normalizzazione, i dati di metilazione sono stati testati per l'associazione con un parametro definito, l'ATRA-*score*. A questo scopo è stata utilizzato l'algoritmo implementato nella funzione *dmpFinder*, implementata nel pacchetto *minfi* in ambiente R, in grado di verificare, per ogni posizione genomica, l'associazione tra la metilazione e un "fenotipo" o parametro definito. Dato l'ATRA-*score* come parametro continuo, l'associazione è stata testata con un modello di regressione lineare.

Infine, le zone genomiche identificate sono state annotate, sfruttando le informazioni contenute nel pacchetto "*IlluminaHumanMethylation450kanno. ilmn12. hg19.*" In *Bioconductor* (https://bioconductor.org/).

Anche in questo caso, per alcuni geni coinvolti nel meccanismo molecolare di azione dipendente dall'attivazione interferonica, è stata trovata una correlazione fra i livelli di metilazione e l'attivazione della risposta all'acido retinoico.

In particolare, tra i geni più rilevanti per il loro ruolo centrale all'interno dei network di interazione identificati, possiamo trovare TLR3, fondamentale nei meccanismi di risposta immunitaria innata, il cui livello di metilazione basale correla con una trascrizione profondamente indotta dopo il trattamento con acido retinoico. Inoltre, altri due geni coinvolti nella risposta immunitaria innata e indotti dal trattamento con acido retinoico, HLA-E e PSMB8, hanno un alto livello di metilazione in condizioni basali, che ben correla con l'ATRA-*score*.

Nella seconda parte di questo studio, abbiamo preso nuovamente in considerazione i dati di sequenziamento di RNA per quantificare la possibile trascrizione di sequenze ripetute da DNA retrovirale, note per essere ampiamente distribuite nel genoma umano: l'ipotesi di partenza dell'analisi, è la possibilità che la perturbazione trascrittomica di queste sequenze possa essere la causa della riattivazione delle vie interferoniche. Per quantificare l'espressione di questi elementi trasponibili, abbiamo ottenuto le loro posizioni genomiche dal database *RepeatMasker* (http://www.RepeatMasker.org/). Queste coordinate sono state assemblate in un file di annotazione personalizzato (*gene transfer file*, GTF), che è stato utilizzato per quantificare l'induzione nella trascrizione di queste sequenze retrovirali, utilizzando l'algoritmo implementato in *FeatureCounts*. Per evitare falsi positivi, tutti gli elementi trasponibili che mostrano la sovrapposizione con esoni noti associati a geni codificanti, sono stati scartati. In seguito, dopo una fase di normalizzazione, la quantificazione dell'espressione differenziale di queste sequenze fra campioni trattati con ATRA e campioni non trattati, è stata condotta usando un approccio analogo a quello usato per i geni codificanti. Ciò che si è potuto osservare è un'induzione ampiamente distribuita di tutte queste sequenze: l'induzione della regolazione trascrittomica è inoltre correlata con la sensibilità all'acido retinoico. Le linee cellulari che sono completamente resistenti al trattamento farmacologico, non mostrano alcuna induzione; linee maggiormente sensibili mostrano un maggiore livello di attivazione. Nonostante la presenza di una tendenza

generale di correlazione tra l'induzione della trascrizione degli elementi retrovirali e la sensibilità ai retinoidi, di particolare interesse è l'effetto trascrizionale su alcune linee cellulari (CAMA1, ZR751), che non segue il comportamento globale. Per meglio comprendere le ragioni di tale tendenza, sono state effettuate ulteriori indagini: è stato dimostrato che questo andamento generale è strettamente correlato con i livelli di espressione basale di RARA in ogni linea cellulare.

I risultati ottenuti in questi modelli cellulari forniscono informazioni sui meccanismi molecolari che sottendono l'azione anti-tumorale di ATRA nel tumore della mammella. Inoltre, i dati di sequenziamento hanno condotto all'identificazione delle vie e delle reti genichedipendenti dall'azione dell'acido retinoico, legate alla sua attività anti-tumorale.

Infine, l'approccio fornisce informazioni sui potenziali nuovi bersagli molecolari per la progettazione di combinazioni terapeutiche razionali basate su ATRA per il trattamento e la prevenzione di una chemioterapia secondaria su alcuni tipi di carcinoma.

Al di là dell'attività antiproliferativa, che potrebbe essere il risultato di una combinazione di fattori diversi, i dati presentati suggeriscono che il trattamento con acido retinoico possa avere un effetto immuno-regulatorio.

Infatti, il possibile aumento nella presentazione di antigeni, unito alla up-regolazione di pathway molecolari legati all'attivazione del sistema immunitario innato, avrebbe effetto significativo sulla rilevazione dei tumori da parte sistema immunitario e la sua conseguente attivazione.

Secondo questa osservazione, questo studio fornisce un forte razionale per lo studio della combinazione di ATRA con farmaci recentemente introdotti nel trattamento immuno-oncologico, cioè gli inibitori del checkpoint immunitario.

Introduction

Breast cancer is the most common malignancy and the leading cause of cancer deaths in women in the Western Hemisphere [1]. This is a very complex and heterogeneous disease and numerous efforts have been made to identify histological and molecular characteristics associated to clinical outcome. On the basis of gene expression data, different subtypes have been identified that show significant differences in incidence, survival and response to drug treatment. The most important determinants of these subtypes are the presence or absence of the estrogen receptor (ER) and the progesterone receptor (PgR), and the over-expression of tyrosine kinase receptor ERBB2 [7]. Given the heterogeneity of the disease, the diversity of the molecular mechanisms activated in different subgroups of this tumor and the developing of resistance to classical therapies, it would be helpful to use combinations of different drugs. All-trans-retinoid acid (ATRA) is the active metabolite of vitamin A and a promising agent in the prevention and treatment of breast cancer. In view of the development of ATRA-based therapeutic strategies aimed at personalized treatment of mammary tumours, a recent study demonstrated that approximately 70% of oestrogen-receptor-positive (ER+) breast cancer cell lines and primary tumours are sensitive to anti-proliferative effects of ATRA [20]. In contrast, only 10-20% of the HER2-positive and triple-negative counterparts respond to the retinoid.

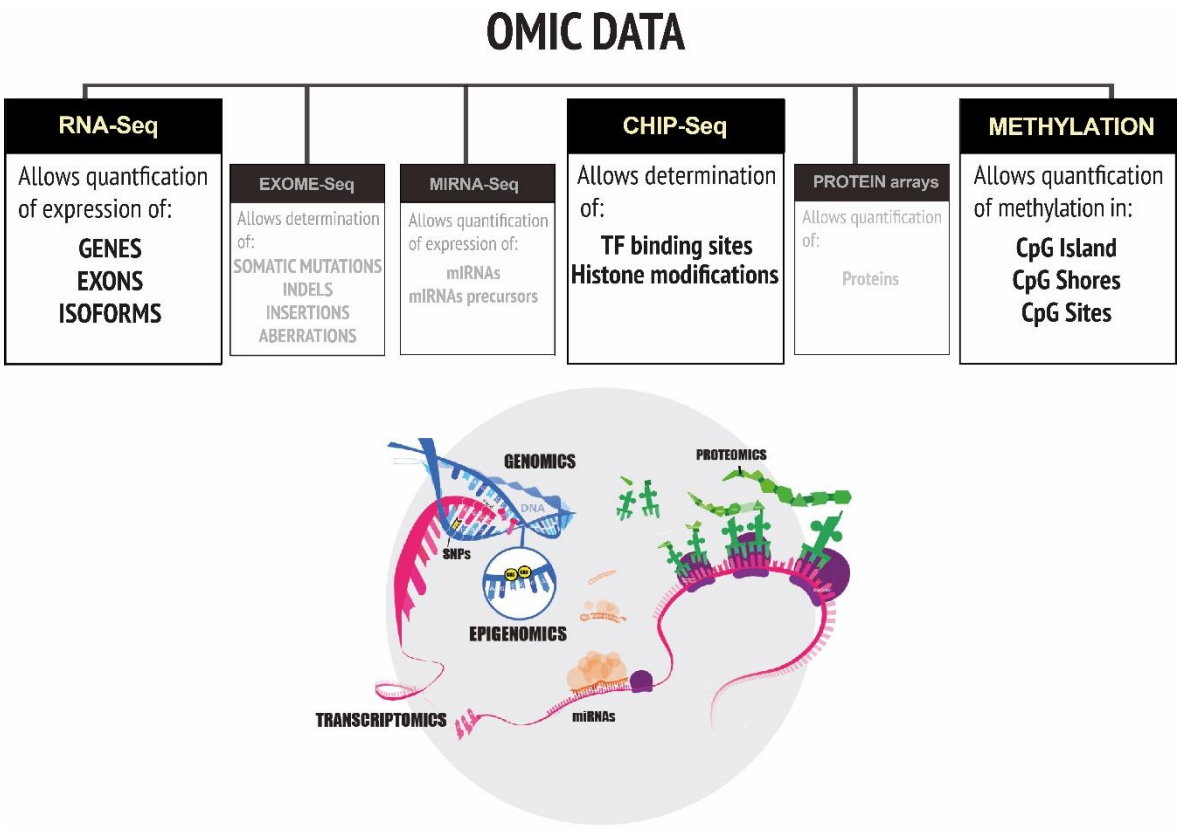
Mechanisms by which retinoids exerts their antitumor activity have not yet been completely clarified, although studies *in vitro* and *in vivo* have shown the ability of retinoids to inhibit proliferation, induce differentiation and apoptosis, making these molecules of therapeutic interest. Therefore, it would be important to decipher retinoids' transcriptionally mechanisms of action, since it could lead to the development of targeted therapeutic strategies, able to implement new drug treatments.

At present, on the basis of the data and the available basal gene-expression profiles of breast cancer cell lines and primary tumors, Bolis and colleagues [38] have developed a model consisting of 21 genes (ATRA-21) which correctly predicts ATRA-sensitivity in the context of breast cancer.

The present study is aimed at getting insights into the molecular mechanisms underlying the anti-tumor action of ATRA in the specific subsets of breast cancer identified. In addition,

we intend to identify specific genes and gene-networks modulated by ATRA which may represent pharmacological targets for the design and development of rational combinations between the retinoid and unrelated therapeutic agents to be used in the personalized treatment of breast cancer agents. A final goal is the identification of potential bio-markers of the anti-tumor response to ATRA and potentially pharmacological targets to be used in the clinics.

To address all these points, we used a multi-omics approach to investigate various aspects of the molecular mechanisms that can be involved in the response of the breast tumour cells to treatment with retinoic acid.



In the first part of the study, we analysed data obtained after performing deep-sequencing experiments on a panel of sixteen cell lines recapitulating the heterogeneity of the breast cancer phenotype and characterized for their anti-proliferative response to ATRA. Each cell line has been exposed to ATRA (1 µM) for 24 hours and total RNA was extracted and subjected to high throughput sequencing. The global gene-expression data were analyzed to

evaluate the transcriptomic profile induced by the pharmacological treatment, with a number of complementary bio-informatic tools.

To complement this analysis, we obtained data from available databases, such as the NCBI GEO (Gene Expression Omnibus), deriving from different techniques, such as Chromatin Immunoprecipitation (ChIP) sequencing and Methylation arrays, performed on a wide number of breast cancer cell lines. We analysed these data and then complemented our results to the data obtained from the first part of the study.

In the second part of the study RNA-sequencing data were analysed to quantify transcription of possible retroviral elements that are part of the human genome, in order to asses if a phenomenon called "viral mimicry" could trigger the mechanism of response to pharmacological treatment with retinoids.

This multi-omics approach gave us a more comprehensive view of the transcriptional and molecular mechanisms that are implicated in the sensitivity or resistance of breast cancers to the treatment with retinoic acid.

It is worth noticing that the huge amount of data to be processed and analysed for this project required a massive computational time and power: for many processing steps we had to exploit CINECA supercomputers.

Background

1 BREAST CANCER

1.1 BREAST CANCER EPIDEMIOLOGY

Breast cancer is the most commonly diagnosed malignancy among females and the leading cause of cancer death in women worldwide, accounting for an estimated 2 million new cancer cases in 2018 [1]. The highest incidence rates were registered in Western and Northern Europe, Australia/New Zealand and North America; intermediate rates in South America, the Caribbean and Northern Africa; low rates in Sub-Saharian Africa and Asia (Figure 1).

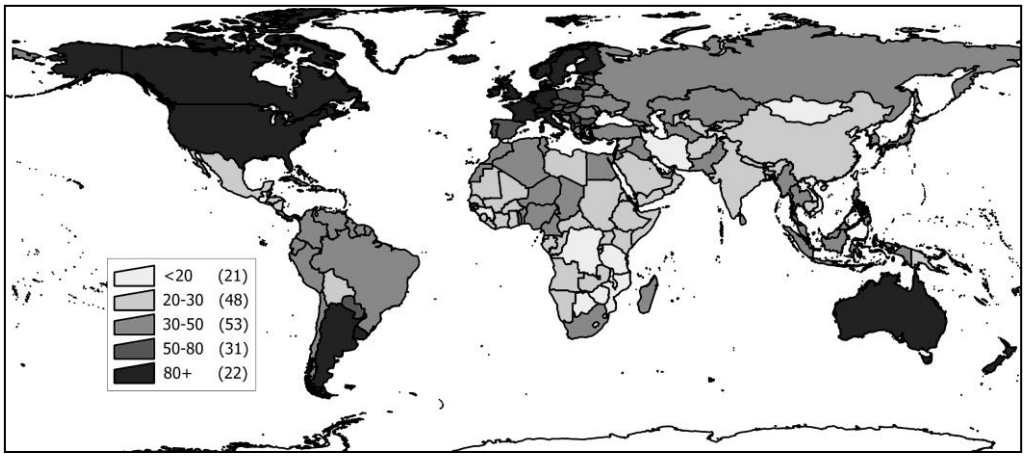


Figure 1 Incidence rates of breast cancer worldwide (adapted from [1])

Despite the increasing number of breast cancer diagnoses over recent decades, the rate of mortality has become stable (or decreasing), reflecting both increased screening programs and improvements in treatment's efficacy.

In Italy breast cancer is the most frequently diagnosed malignancy (excluding non-melanoma skin cancers) in women, with about 52.300 new cases expected in 2018, the 29% of the total, with 12.274 estimated deaths at different stages of life [2].

The main risk factor for developing breast cancer is age, together with female gender and individual hormonal status. Inherited genetic factors, such as hereditary mutations in BRCA1 and BRCA2 tumour suppressor genes, PTEN, p53, CDH1, CHEK2, ATM and a few others

also play an important role, significantly increasing lifetime risk of developing breast cancer [3].

Hormonal and reproductive factors, such as a long menstrual history (early menarche and/or late menopause), use of contraceptives or menopausal hormone replacement therapy, or never having children, are related to an increased risk of developing breast cancer. Lifestyle also affects the risk of developing breast cancer: never breastfed, physical inactivity, overweight/obesity, elevated alcohol consumption, high energy diet and smoking can increase the risk.

1.2 BREAST CANCER CLASSIFICATION

Breast cancer is a heterogeneous disease, characterized by several pathological features, different response to therapeutics, and substantial differences in long-term patient survival. The broad heterogeneity observed among breast cancer reflects the now well-accepted notion that there is not just a one disease with disparate variant subtypes, but that breast cancer instead represents a collection of distinct neoplastic diseases of the breast and the cell composing it [4].

To better understand breast cancer and its heterogeneity it is necessary to briefly explain breast structure.

Human breast is a complex secretory organ, made of two main types of tissue: supporting (stromal) tissue and glandular tissue (mammary gland).

The supporting tissue, or stroma, is composed of collagenous connective tissue and adipose tissue, which sustain the mammary gland with blood and lymphatic vessels associated. The mature mammary gland consists of 15-20 lobes, each of them further divided into several lobules containing the alveoli, the milk producing units, and the ducts, tubes carrying the milk from the alveoli to the nipple.

The alveoli are hollow cavities, delimited by a basement membrane lined by luminal milk secreting cells. Lobules and ducts are formed respectively by alveolar and ductal luminal epithelial cells, which are surrounded by a basal layer of myo-epithelial cells, that contract to excrete the milk through the duct toward the nipple [5]. These cell types form a bi-layer that lies adjacent to the basement membrane and lines both the ducts and cluster of lobules known as *terminal ductal lobules* (TDLUs) (Figure 2).

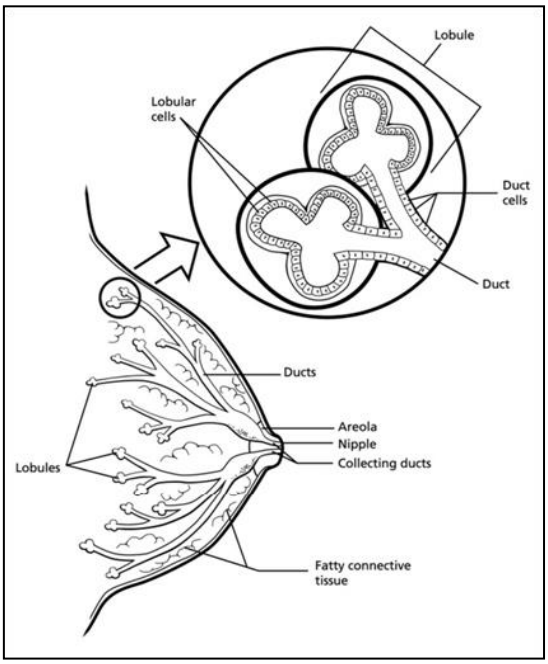


Figure 2 Normal breast structure [www.cancer.org/cancer/breastcancer]

Behind this complexity, several systems have been developed to classify a very highly heterogeneous disease and possibly to get information about tumour behaviour and provide more effective therapies.

Histological classification categorizes breast cancer either in "*in situ*" or "invasive".
The term *in situ* refers to a type of cancer that has developed within the epithelial tissue. It can be divided either in ductal (*ductal carcinoma in situ*, DCIS) or lobular (*lobular carcinoma in situ*, LCIS), depending on the original site of the cancer, ducts or lobules. The lobular *in situ* breast cancer is not considered as a real cancer, but a risk factor for developing it. [6]
The invasive or infiltrating breast cancer has overcome the basement membrane and invaded nearby tissue and vessels and possibly spread to other organs. It can be further sub-classified into multiple sub-groups [6]:

- Invasive Ductal Carcinoma (IDC) (70-80%): it is the most common type of breast cancer, that starts into the duct and then spread and grows into the surrounding tissue. At this point, it may be able to spread (metastasize) to other parts of the body through the lymph system and the bloodstream.

- Invasive Lobular Carcinoma (ILC) (10%): it is the second most common type of breast cancer; it originates in the milk producing lobules and can spread (metastasize) to other parts of the body.

Other sub-types of invasive carcinomas much less common than the breast cancers listed above (less than 5%) are adenocystic carcinoma, medullary carcinoma, mucinous carcinoma, papillary carcinoma and tubular carcinoma.

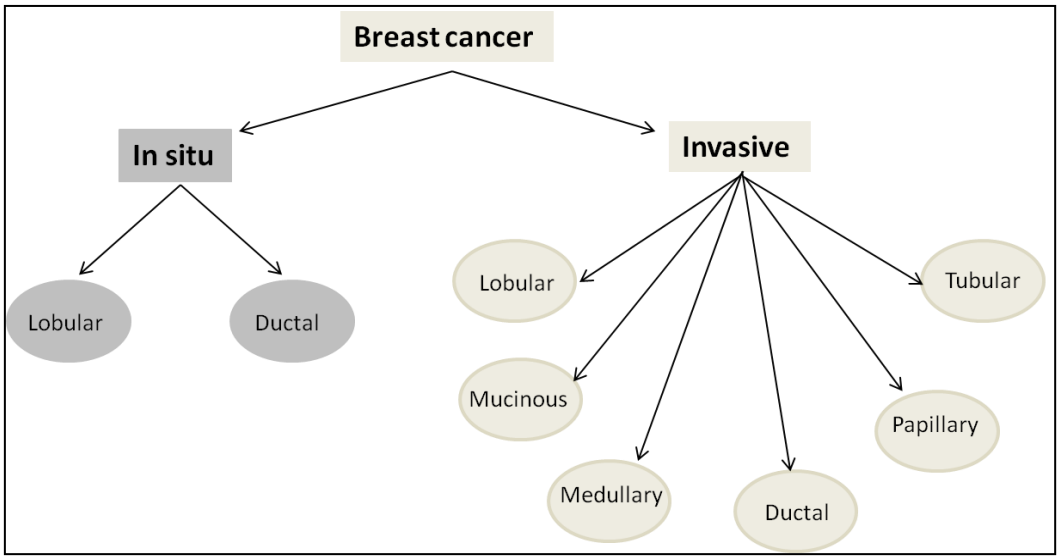


Figure 3 Histological classification of breast cancer (adapted from [6])

This classification relies only on histological characteristics without taking into account molecular markers or morphological features that can be helpful for prognosis or therapy.

In clinical practice breast cancer classification also relies on clinical-histopathological features, on the presence or absence of hormonal receptors and on the Ki67 proliferation index.

In particular, it is critical to assess the receptor status of a tumour, as it determines the possibility of using targeted treatment.

Hormonal receptors for oestrogen (ER) or progesterone (PR) are usually identified through immunohistochemistry (IHC). Cancer cells expressing ER need oestrogen to grow, so ER positive breast cancer can be treated with drugs that reduce the effects of the production of oestrogens.

Human epidermal growth factor receptor 2 (HER2) is a membrane receptor that can be over-expressed in breast cancer. *HER2* positive tumours are characterized by a DNA-amplification of the region containing this tyrosine-kinase receptor which leads to its

overexpression and abnormal functioning. These cancer cells can be treated with the monoclonal antibody Trastuzumab, aimed at blocking HER2 receptor activity, in combination with chemotherapy [7].

The proliferation marker Ki67 is a nuclear protein associated with cellular proliferation, as it is expressed only in proliferating and not in quiescent cells. This factor is associated with histopathological parameters, as it has been shown that poorly differentiated cancers have a high proliferation index [8] and can be considered as an independent prognostic factor for overall survival (OS).

Moreover, breast tumours may be endowed with a different complement of the three above mentioned receptors (ER+/PR+/HER2+; ER+/PR+/HER2-; ER+/PR-/HER2+; ER+/PR-/HER2-; ER-/PR+/HER2-; ER-/PR-/HER2+; ER-/PR-/HER2-). Tumours that lack expression of all three receptors are defined as Triple Negative Breast Cancer (TNBC).

In recent years several studies on gene expression profiles have been conducted using high-throughput technologies, in order to identify molecular subtypes of breast cancer, to allow a better understanding of the complexity of the disease.

Based on hierarchical clustering of gene expression microarray data from 65 tumours and normal breast samples, six subtypes of breast cancers have been identified [9].

The luminal A breast cancer is the most common subtype accounting for 50 – 60% of all diagnoses. It's characterized by the expression of ER, the absence of HER2 over-expression, together with a low Ki67 proliferation rate [10].

The luminal B breast cancer accounts for 10–20% of total. This subtype of cancer often expresses HER2 and may express low level of oestrogen receptor. Moreover, it has a higher proliferative rate (measured by Ki67).

The HER2 positive subtype represents about 15-20% of all breast tumours and is characterized by expression of the HER2 gene, genes associated to its pathway and genes associated to cellular proliferation [11].

The basal-like subtype, which represents 10-20% of total breast cancers, is characterized by expression of genes characteristic of the myo-epithelial (or basal) cells. In general, this subtype does not express ER, PR and HER2 and often, in clinical practice, it is referred to

as triple negative breast cancer, even if the terms are not equivalent since a discordance between the two groups has been observed [12].

The normal breast subtype is characterized by the expression of genes typical of adipose tissue and accounts for 5-10% of total breast cancers [13]. These tumours lack the expression of hormonal receptors and HER2, so they are TNBC, but they are not basal-like cancers since they lack the expression of some genes characteristic of that category. Moreover, this category is still poorly characterized, and its clinical relevance has yet to be determined.

The claudin-low subtype has been recently identified [14] with a low expression of genes that encode for proteins involved in the formation of tight junctions, including Claudins and E-cadherin. It is rare and characterized by the absence of the oestrogen receptor, progesterone receptor and HER2 expression. Furthermore, tumours belonging to this subtype over-express a set of genes related to the Epithelial to Mesenchymal Transition (EMT), that are not present neither in the basal tumours nor in other subtypes.

The molecular subtypes identified show significant differences in incidence, survival and response to treatment (Figure 4), providing a wide range of information that greatly expand the knowledge obtained from the classical pathological markers [15].

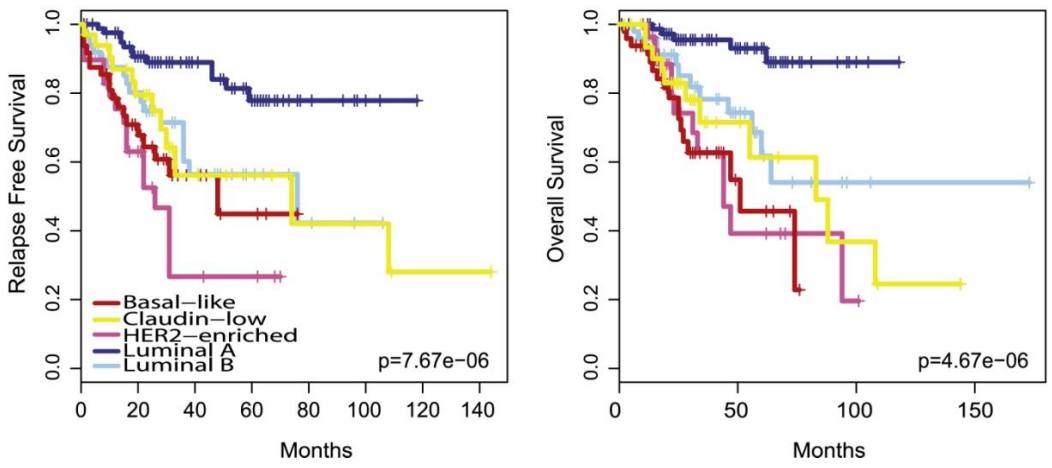


Figure 4. Kaplan-Meier relapse free survival and overall survival of intrinsic subtypes of breast cancer (adapted from [15]).

In 2009, Parker and colleagues[16] introduced a new system of analysis, PAM50 (Prediction Analysis of Microarrays), to select a minimum set of genes, whose expression can predict the molecular subtype of a tumour. The PAM50 gene sets allows to obtain a classification similar to the one described above; it is based on a selection of gene sets consisting of a large number of "intrinsic" genes, and therefore can be used in the clinics [15] to define the molecular phenotype of the tumours.

1.3 CLINICAL AND MOLECULAR CLASSIFICATION OF BREAST CANCER

Molecular classification of breast cancer based on gene expression patterns provides a connection between molecular biology and the behaviour of cancer cells in the corresponding subtypes [16]. However, molecular classification of breast cancer has not yet reached clinical implementation as a routine aspect of patient management. Although an immunohistochemical staining proxy can be used to stratify and classify breast cancers in a clinical setting, the correspondence between clinical (i.e., immunohistochemical) and molecular (i.e., gene expression) classification is not remarkable. This situation is illustrated in Figure 5 for 381 breast cancers for which both immunohistochemical staining data and molecular classification were available [17].

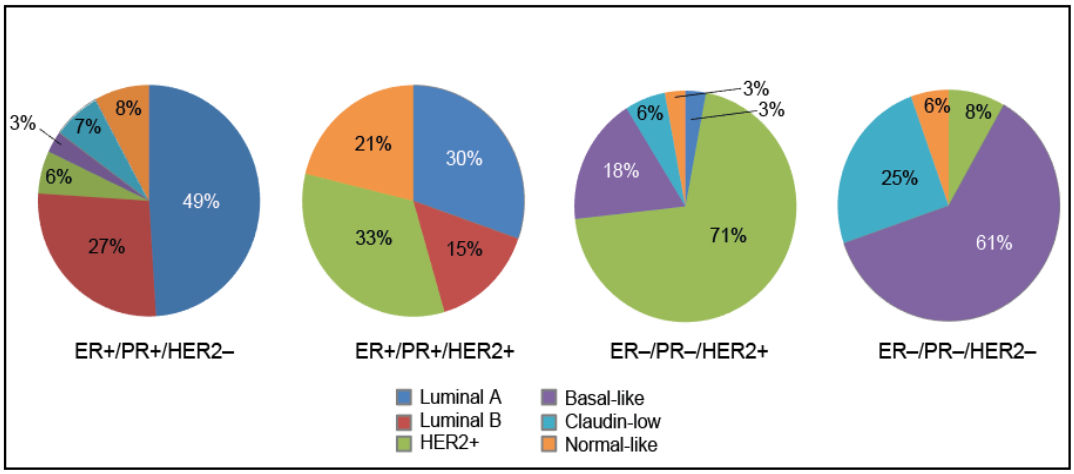


Figure 5. - Correspondence between PAM50 and traditional classification of breast-cancer (adapted from [17]).

Correspondence between the immuno-histochemical and the molecular/transcriptomic classification of breast cancer. Mammary-tumors are divided in four groups according to the traditional classification based on the immuno-histochemical determination of the ERα, PR and HER2 molecular markers. For each of the four groups, the percentage of cases showing the six indicated transcriptomic phenotypes determined on the basis of a modification of the PAM50 fingerprint is illustrated.

2. RETINOIDS

The term retinoids refers to a group of compounds comprising metabolites and analogues of vitamin A, both natural and synthetic. The natural retinoids are essential components of diet and physiological regulators of many essential biological processes, such as embryonic development, metabolism and haematopoiesis. In adult mammals, retinoids such as All-*Trans* Retinoic Acid (ATRA), control homeostasis of different organs and tissues [18].

All-*Trans* Retinoic Acid (ATRA) is a small lipophilic molecule and an important regulator of gene expression. ATRA, synthesized intracellularly from circulating retinol or diffusing from an adjacent cell, binds to cellular retinoic acid binding proteins I and II (CRABPI and CRABPII). Binding to these two proteins have opposite effects, while binding to CRABPII promotes ATRA activity, by stimulating the transfer to the nucleus and the activation of transcription, binding to CRABPI reduces its activity, by promoting its degradation [18].

In addition to degradation, another balancing mechanism of ATRA concentration, to avoid excessive stimulation of the cells, is isomerization in less-active and more water-soluble isoform 9-*cis* Retinoic Acid (9-cisRA) and 13-*cis* Retinoic Acid (13-cisRA), thus the plasma concentration of retinoic acid in humans is very low, between 5-10nM [19].

The biological action of ATRA and its derivatives is mediated by two classes of nuclear receptors for retinoids called Retinoic Acid Receptor (RAR) and Retinoic X Receptor (RXR). RAR and RXR receptors have different affinity towards specific ligands: 9-cisRA binds both receptors, while ATRA binds only RARs [19].

The receptors are ligand-dependent transcription factors that control the activity of several target genes either through a direct or indirect mechanism. Both receptor subtypes exist in three different forms known as alpha, beta and gamma, encoded by different genes (RARA, RARB, RARG/ RXRA, RXRB, RXRG). Each subtype comprises two or more isoforms, which differ in the N-terminal region and are generated by a different promoter or by alternative splicing mechanisms [20]. Each RAR and RXR isoform encoded by a distinct gene and transcribed into different splicing-variants is represented in Figure 6 [21].

RAR and RXR receptors form stable hetero-dimers (RAR/RXR) that bind to specific sequences on DNA, called Retinoic Acid Responsive Elements (RAREs), localized within the promoters of target genes.

RXR receptors form hetero-dimers with other nuclear receptors such as the Vitamin D Receptor (VDR), the Thyroid Hormone Receptor (TR), the Peroxisomal Proliferation Receptor (PPAR) and other orphan nuclear receptors [22], allowing modulation of several signalling pathways [23]. Moreover, the possibility of forming different combinations between the retinoids receptors contributes to different cellular responses in different cell types.

All-*trans* retinoic acid modulates transcription through different mechanisms. It can directly modulate expression of target genes, through the interaction of the RAR/RXR with a group of co-activators and co-repressors. In the absence of ligand, or in presence of antagonists, the RAR/RXR dimer is bound to the RAREs sequences on DNA and is associated with a complex of co-repressors that inhibit the transcription of target genes.

There are also indirect target genes, which do not contain RARE sequences in their promoters, but are regulated by genes that are direct target of ATRA. Among these genes, there are several growth factors and growth factors receptors, such as Tumour Necrosis Factor (TNF), transforming growth factor β and Epidermal Growth Factor (EGF) [24].

Moreover, ATRA can exert its activity in a non-genomic way. Although still largely to characterize, this relies on the ability of retinoids to activate specific kinases, suggesting an atypical, nongenomic activation. In line with this new concept, though classically thought to reside in the nucleus, RARs have been recently reported to be present in the cytosol or in membranes [25].

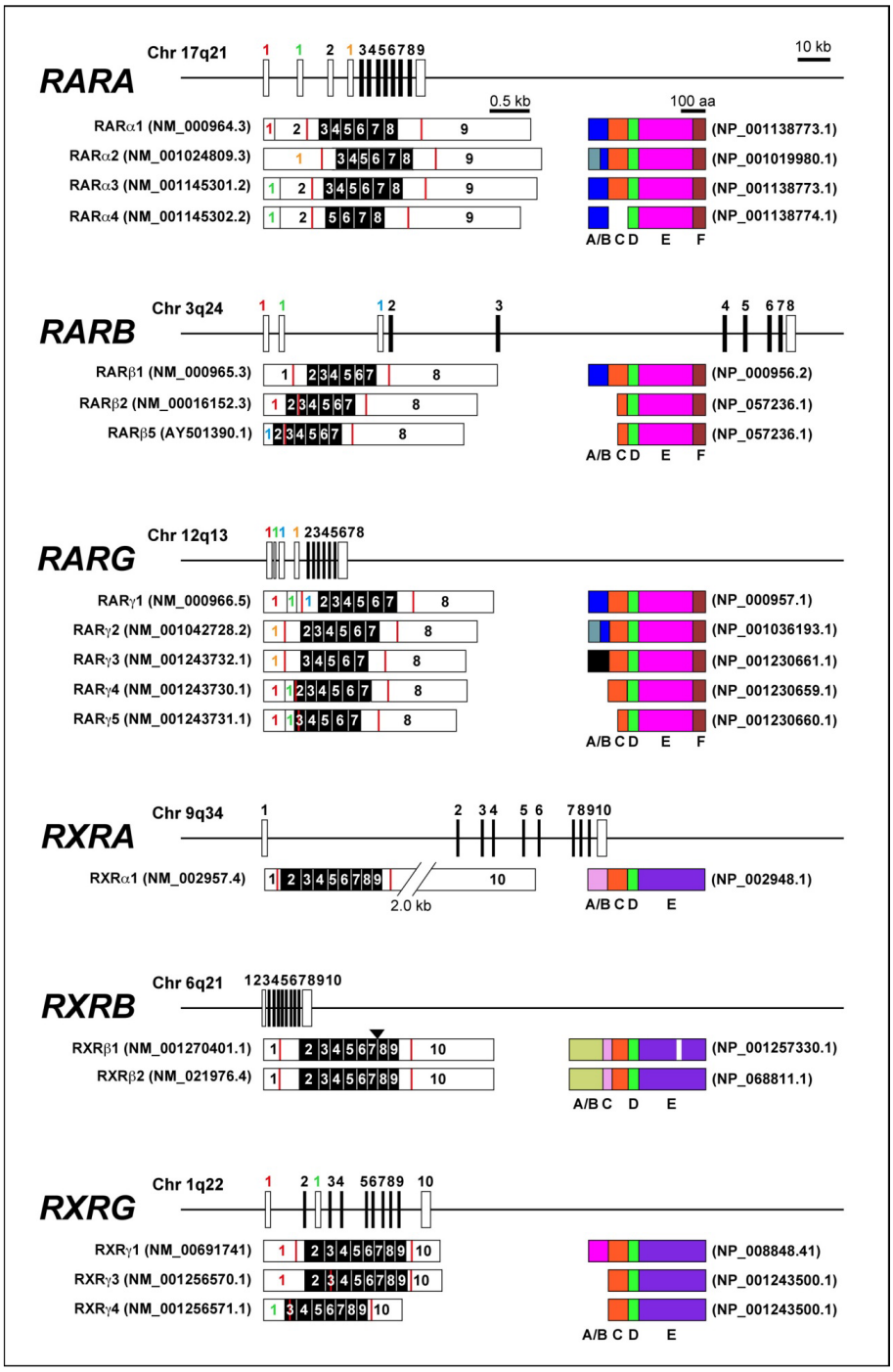


Figure 6. Structure of the RAR/RXR genes and relative mRNA/protein products.

The exonic structure of the genes coding for the human RARA, RARB, RARG, RXRA, RXRB and RXRG genes along with the corresponding chromosomal location are indicated. Below each gene the structure of the corresponding transcript variants is shown on the left side. The structure of the encoded proteins from the NH- to the COOH terminus (left to right) are indicated on the right. Boxes drawn in different colours represent the known structural domains of the various receptors [21].

2.1 RETINOIDS IN ONCOLOGY

Tumorigenesis is a multistep process characterized by a series of inherited or acquired genetic changes (mutations, chromosomal rearrangements, epigenetic phenomena) leading to a disruption of cellular homeostasis and development of the neoplastic process.

Several lines of evidence indicate an important role of retinoids in homeostasis. Moreover, retinoids are essential components of the diet and regulate embryological development. Low levels of vitamin A during development can cause malformations, while high concentrations are teratogenic. In adult humans, vitamin A regulates the fertility, the visual function, prevents tumour growth and the development of neurodegenerative diseases [26].

In particular, *in vivo* studies showed the existence of a correlation between vitamin A deficiency and the onset of cancer. Moreover, changes in the bioavailability of retinoids, due to alterations in their metabolism, are associated with tumorigenesis. In many cancers, including breast cancer, it seems that the gene coding for RARβ is silenced, through a deletion, a mutation, or a hyper-methylation of his promoter; the loss of expression of RARβ receptor is associated with tumour progression [27].

In contrast, the addition of retinol to the diet reduces the risk of hyperplasia and tumour incidence in animal models and in organ cultures and changes in the levels of functional expression or activity of retinoid receptors (RAR and RXR) have been observed in some cancers [28].

Mechanisms by which retinoids exert their antitumor activity have not yet been completely clarified, although studies *in vitro* and *in vivo* have shown the ability of retinoids to inhibit proliferation, induce differentiation and apoptosis, making these molecules of therapeutic interest.

It is clear that ATRA is able to act through different genomic mechanisms as well as to interact with other intracellular signalling systems, that provides the basis for its pleiotropic action.

Currently, the best example of the anticancer action of retinoids is the use of ATRA in the treatment of patients suffering from Acute Promyelocytic Leukaemia (APL). More than 85% of patients with APL achieve complete remission following treatment with ATRA in combination with anthracyclines. Retinoids also show efficacy in the treatment of three precancerous lesions: leucoplakia, actinic keratosis and cervical dysplasia, and are also able

to slow the development of melanoma in patients suffering from *Xeroderma Pigmentosum* [29].

2.2 RETINOIDS IN BREAST CANCER

Retinoids are of therapeutic interest in breast cancer as they are able to prevent tumours induced by carcinogenic agents in murine models [30] and have anti-proliferative effect on breast cancer cell lines *in vitro*. Despite promising preclinical studies, however, retinoids have not yet found effective application in clinical practice.

Interest in the use of retinoids in breast cancer is reflected by the vast scientific literature (>1300 articles) available, which concentrates on ATRA. On the other hand, the large number of pre-clinical studies has been translated into few clinical trials (Table 1).

No chemo-preventive trials involving ATRA or classic retinoids have yet been published. Except for one report conducted with bexarotene (RXR agonist), all the known chemo-preventive trials are based on fenretinide, which is not a *bona fide* functional retinoid despite its chemical structure, as it can be considered a retinoid, but has a different mechanism of action. In fact, the mechanism underlying its antitumor activity is not linked to the classical mode of activation of nuclear receptors for ATRA. *In vitro* fenretinide inhibits the growth of breast cancer cell lines by acting through a mechanism independent from the retinoid receptors. Indeed, fenretinide was proven to be a selective RARG agonist and not a pan-RAR agonist like ATRA. On the basis of this study on fenretinide, retinoids are not recommended outside of a clinical trial.

In invasive breast cancer, three major therapeutic trials on retinoids, used as single agents, have been reported. The only trial involving ATRA is a phase-II study in pre-treated patients which failed to achieve the primary end-point [31].

A few clinical trials using retinoid-based combinations are available:

- ATRA + tamoxifen is the object of a dose-escalation phase I/II study conducted in patients with ER+ hormone-refractory tumours. Objective responses or stable disease (SD) were observed in 9 patients [32].
- A second pre-operative study in locally advanced breast cancer was conducted to determine both the biologic effects and the minimal effective dose of ATRA with or

without tamoxifen and interferon alpha-2 (IFNα2). Neither ATRA + tamoxifen nor ATRA + tamoxifen + IFNα2 potentiated the ATRA-induced effects [33].

- A small pilot study was conducted to evaluate the efficacy of combinations between ATRA and paclitaxel in pre-treated metastatic breast cancer. In 17 evaluable patients, 3 showed partial remission (PR) and 10 presented with SD. Despite the small cohort of patients analysed, the data suggest that this well-tolerated combination induces a modest frequency of PR but relatively high rates of stable disease [34].
- Retinyl-palmitate, 9-cis-RA and 13-cis-RA [35] were also tested in combination with other agents. Combinations of 13-cisRA and tamoxifen or interferon alpha-2 were investigated in post-menopausal pre-treated metastatic breast cancer. No significant difference in the overall response rate and overall survival was observed in the 3 treatment arms.

The data available stress the paucity and the generally disappointing nature of the clinical results obtained with retinoids.

The molecular determinants responsible for resistance or sensitivity of cancer cells to retinoids are still poorly understood but it is reasonable to suppose that they include the expression of RARs and RXRs and factors related to them [36]: from this point of view, breast cancer is a classic example of the heterogeneous response to retinoids. For that reason, although they have shown a potential for therapeutic use in breast cancer, up to now retinoids haven't been approved for clinical use. The variability in the response is yet the major problem that limits their use in clinics.

The starting hypothesis here is that the negativity of the results is predominantly due to the design of the clinical trials, which did not take into account the heterogeneity of breast cancer and were conducted on cohorts of patients without selection for any particular sub-type of tumour.

Given the heterogeneity of breast cancer, a clinical use of retinoids requires the identification of a subset of patient sensitive to their effect.

To challenge this goal, it would be important to understand the molecular mechanisms and the determinants of retinoid sensitivity or resistance.

Compound	End point	Trials (No.)	Pts (No.)	Clinical Phase	Reference
Fenretinide	BC prevention	5	6521	Ph II/III	-Veronesi U et al, (1999) *J NatlCancerInst***91,** 1847-56 -Veronesi U et al, (2006) *AnnOncol***17,** 1065-71
	BC treatment	2	441	PhI/III	-Cobleigh MA et al, (1993) *J ClinOncol***11**, 474-7 -Rao RD et al, (2011) *Med Oncol***28**, 1:S39-47
ATRA	BC prevention	0	0		
	BC treatment	4	73	Ph I/II	-Sutton LM et al, (1997) *Cancer Chemotherapy and Pharmacology***40**, 335-341 -Budd GT et al, (1998) *Clin Cancer Res* **4**, 635-42 -Toma S et al, (2000) *Int J Oncol* **17**, 991-1000
9-*cis*-RA	BC prevention	0	0		
	BC treatment	2	34	Ph I	-Kurie JM et al, (1996) *Clin Cancer Res* **2,**287-93 -Lawrence JA et al, (2001) *J ClinOncol***19**, 2754-63
13-*cis*-RA	BC prevention	0	0		
		1	94	Ph II	-Chiesa MD et al, (2007) *Acta Biomed***78**, 204-9

	BC treatment				
Bexarotene	BC prevention	1	87	Ph I	-Brown P et al, (2008) *Cancer Prev Res* **1,** CN04-04
	BC treatment	1	148	Ph II	-Esteva FJ et al, (2003) *J ClinOncol***21**, 999-1006
Retinyl palmitate	BC prevention	0	0		
	BC treatment	1	65	Ph II	-Recchia F et al, (2009) *Oncol Rep* **21**, 1011-6

Table 1. Selected clinical trials of retinoids in breast cancer

The table lists the published clinical trials on retinoids in breast cancer along with the corresponding references.

3 ATRA SENSITIVITY

3.1 ATRA – *SCORE*

As mentioned above, all-*trans* retinoic acid and its derivatives have shown a potential for therapeutic and preventive use in breast cancer because of their ability to modulate cell growth and differentiation [37].

To evaluate the response of breast cancer cell lines to the anti-proliferative effect exerted by retinoic acid, Bolis and colleagues [38] first defined the profile of ATRA-sensitivity in a panel of 48 breast cancer cell lines of the Cancer Cell Lines Encyclopedia (CCLE), well-representing the heterogeneity of the disease. The drug response of each cell line has been quantified by computation of a sensitivity score (ATRA-*score*).

ATRA-score is computed on cell lines that were treated with vehicle (DMSO) and 5 logarithmically increasing concentrations of ATRA (0.001-10.0 μM) for 9 days; its value has been finally calculated from the relative growth-inhibition (GI) data (ATRA vs. vehicle). To define the *ATRA-score* sensitivity metric (as shown in Figure 7), they fitted growth-inhibition curves relative to DMSO-treated controls and computed the area under the curve (*AUC*) and the maximal inhibitory effect (A_{max}). At this point, ATRA-*score* values, which are equal to the log2 transformation of AUC x A_{max} , are rescaled in a range between 0 and 1, zero indicating total resistance and one standing for maximum sensitivity.

Relative to the sole *AUC* value, the *ATRA-score* gives more weight to the maximal growth inhibitory effect. This choice was taken in order to benefit in particular those cell lines that reached striking levels of maximal growth inhibition.

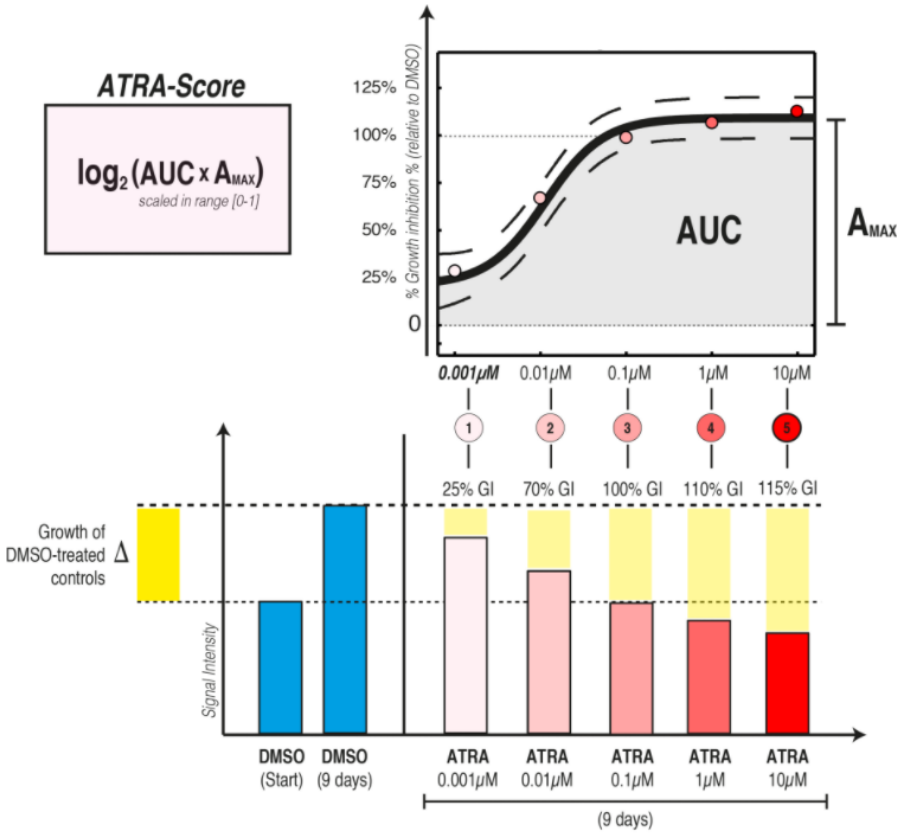


Figure 7. ATRA-score metrics (adapted from[38]).

The figure illustrates an example of a fitted growth-inhibition curve (top) and provides insights on how ATRA-dependent growth-inhibition (GI) was determined relative to DMSO treated controls (bottom). Cell growth was determined with the sulforhodamine assay.

Moreover, the aim of the study of Bolis and colleagues [38] was to develop a tool capable of predicting ATRA-sensitivity, exploiting the association between this *in vitro* profiling and basal gene-expression data.

Starting from the generation of a predictive model based on basal gene expression profile of 139 genes (ATRA-*139*), they used a network-guided approach to develop a generalized model based on a selection of 21 out of the original 139 genes (ATRA-*21*) capable of predicting ATRA-sensitivity across tumour types other than breast cancer.

Materials and methods

1. EXPERIMENTAL SETUP

A panel of 16 breast cancer cell lines has been subjected to a total RNA-sequencing procedure, before and after treatment with retinoic acid (1.0 μM) for 24 hours. Each experiment has been conducted in triplicate, with a total of 96 samples. To better represent breast cancer heterogeneity, cell lines have been chosen based on their phenotype (luminal, which includes luminal A, luminal B and HER2 [10,11], or basal [12]) and their widely variable sensitivity to pharmacological treatment with retinoic acid. Table 2 reports cell lines phenotype, their receptors status and their ATRA-*score*.

CELL-LINE	PHENOTYPE	ER (IHC)	HER2(IHC)	PR(IHC)	ATRA-SCORE
CAL851	BASAL	-	-	-	0.00
CAMA1	LUMINAL	+	-	+	0.66
HCC1187	BASAL	-	-	-	0.00
HCC1419	LUMINAL	-	+	-	0.09
HCC1500	LUMINAL	+	-	+	0.66
HCC1599	BASAL	-	-	-	1.00
HCC202	LUMINAL	-	+	-	0.24
Hs578T	BASAL	-	-	-	0.19
MB157	BASAL	-	-	-	0.28
MDAMB157	BASAL	-	-	-	0.25
MDAMB175VII	LUMINAL	+	-	-	0.19
MDAMB231	BASAL	-	-	-	0.01
MDAMB361	LUMINAL	+	+	-	0.58
MDAMB436	BASAL	-	-	-	0.00
SKBR3	LUMINAL	-	+	-	0.99
ZR751	LUMINAL	+	-	-	0.14

Table 2. Panel of 16 sensitive and resistant cell lines.

For each cell line is reported the phenotype, the receptor status and the ATRA-*score.*

2. RNA-SEQUENCING DATA ANALYSIS

2.1 LOW LEVEL PROCESSING

Alignment of high-throughput paired-end reads derived from RNA-sequencing experiments to the reference genome has been performed. Genome-generation was performed using the comprehensive gene annotations present in *Gencode* [39]; in particular, the v27 release of the *Gene Transfer File* (GTF) file has been used.

High-throughput sequencing (HTS) experiments generate hundreds of millions of sequences that present the unique challenge of detection and characterization of spliced transcript, dealing with reads that contain both mismatches, insertions, deletions of genomic regions and sequencing errors. The reads sequenced are small fragments (150 bp) compared with the median gene size in homo sapiens (24 kbp) [40] and therefore they may ambiguously align to multiple genomic regions.

As many RNA-sequencing aligners [41, 42,43] suffer from high mapping error rates, read length limitation or mapping biases, sequence–alignment to reference human genome (hg38) has been performed using STAR (Spliced Transcript Alignment to a Reference) [44] sequence-aligner, which was designed specifically to align non-contiguous sequences directly to the reference genome, using a novel strategy for these spliced alignments.

STAR algorithm involves two major steps: a first seed searching step and a second clustering/stitching/scoring step.

The central idea of the STAR seed finding phase is the sequential search for a Maximal Mappable Prefix (MMP) in uncompressed suffix arrays (SA) [45]. To find the best alignment of a given RNA-sequencing read R, a read location *i* and a reference genome sequence *G*, the MMP (R,i,G) has to be identified, as the longest substring (Ri, Ri + 1, … , Ri + MML − 1) that matches exactly one or more substrings of G, where MML is the maximum mappable length (Figure 8).

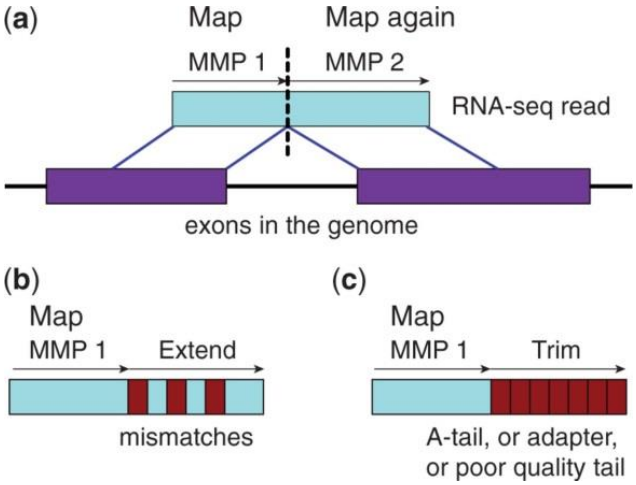


Figure 8. STAR alignment algorithm

Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (a) splice junctions, (b) mismatches and (c) tails [44].

As a first step, the algorithm finds the MMP starting from the first base of the read (Figure 8.A). If the read comprises a splice junction, it cannot be mapped contiguously to the genome, and thus the first seed will be mapped to a donor splice site. Moreover, in the presence of mismatches, the MMP will serve as an anchor in the genome that can be extended, allowing for alignment even in presence of divergencies (Figure 8.B).

As a second step, the MMP search is repeated for the unmapped portion of the read, which, in this case, will be mapped to an acceptor splice site.

The splice junctions are detected in a single alignment pass without any *a priori* knowledge of splice junctions' loci or properties, and without a preliminary contiguous alignment pass needed by the junction database approaches.

The MMP search is performed in both forward and reverse direction of the read sequence.

In a second phase, STAR algorithm builds alignment of the entire read sequence by stitching together all the seed that were aligned to the genome in the previous phase.

First, the seeds are clustered together by proximity to a selected set of "anchor" seeds. The size of the user-defined genomic windows determines the maximum intron size for the spliced alignments. STAR algorithm allows for any number of mismatches, still for only one insertion or deletion (gap).

Then, a stitching process is guided by a local alignment scoring scheme, with defined scores (and penalties) for matches, mismatches, insertions, deletions and splice junctions gaps, leading to a subsequently quantitative assessment of the alignment qualities and ranks.

Notably, the seeds from the mates of paired-end RNA-sequencing reads are clustered and stitched concurrently, with each paired-end read represented as a single sequence, allowing for a possible genomic gap or overlap between the inner ends of the mates. This is a fundamental way to use the paired-end information, underlying the nature of the paired-end reads, reflecting the fact that the mates are pieces (ends) of the same sequence. This type of approach increases the sensitivity of the algorithm, as only one correct anchor from one of the mates is sufficient to accurately align the entire read.

2.2 DIFFERENTIAL GENE EXPRESSION ANALYSIS

All the analyses and the processing of the RNA-sequencing data were performed using R 46, 47], a free software environment for statistical computing and graphics.

The R package *DESeq2* [48,49] was used to detect differentially expressed genes between ATRA-treated and untreated samples.

As first step, the input for the *Deseq2* package has been defined, as un-normalized counts of sequencing reads obtained from the RNA-sequencing experiment, in the form of a matrix of integer values. The value in the i-th row and the j-th column of the matrix defines the number of reads that can be assigned to gene i in sample j. Moreover, as a further step before starting the analysis, a pre-filter of low count genes has been carried out, by removing rows in which all the read counts all equal to 0, to both reduce the memory size of the data objects and increasing the speed of the transformation and testing functions.

The phases of the differential expression analysis are designed into a single function, *DESeq*. The analysis that it performs is based on the Negative Binomial (Gamma-Poisson) distribution and goes through three different steps: estimation of the size factors, estimation of the dispersion coefficients and a Negative Binomial Generalized Linear Model (GLM) fitting.

The generalized linear model used in differential expression analysis is on the form:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_j \beta_i$$

where counts K_{ij} for gene i, sample j, are modelled using a Negative Binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean is composed of a sample-specific size factor s_j and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j. The coefficients β_i give the log2 fold changes for gene i for each column x_j of the model matrix X.

To sum up, the first step performed by the function is the estimation of size factor s_j, followed by the estimation of dispersion coefficient α_i, and the negative binomial GLM fitting for β_i, coupled with a Wald statistic.

The estimation estimate of size factors is implemented in the function *estimateSizeFactors*, that uses the "mean ration method" described by Anders and Huber [50], to obtain a normalized count matrix.

EstimateDispersions is then the function used in the second step to obtain dispersion estimates for such negative binomial distributed data. The dispersion parameter α_i defines the relationship between the variance of the counts observed and its mean value. Specifically, how far the observed count is expected to be from the mean value, which depends both on size factors s_j and the covariate dependent part q_{ij}, as defined above.

The fitting method proceeds as follows: for each gene is found an estimate of the dispersion that maximize the Cox- Reid-adjusted profile likelihood [51]. Then, a trend line capturing the dispersion mean relationship is fitted to the maximum likelihood estimates.

A normal prior is determined for the log dispersion estimates centred in the predicted value from the trended fit with variance equal to the difference between the observed variance of the log dispersion estimates and the expected sampling variance (Equation 1) . Finally, maximum a posteriori dispersion estimates are returned.

$$Var(K_{ij}) = E\left[\left(K_{ij} - \mu_{ij}\right)^2\right] = \mu_{ij} + \alpha_i {\mu^2}_{ij} \qquad (1)$$

This final dispersion parameter is used in subsequent tests.

Using the previously calculated size factors and dispersion estimates, the function *nbinomWaldTest* tests for significance of coefficients in a negative binomial generalized linear model (GLM).

First, standard maximum likelihood estimates for the generalized linear model coefficients (β_i, or log2fold changes) are calculated. To obtain the Wald test p-values, the coefficients are scaled by their standard errors and then compared to a standard normal distribution.

Notably, the *DESeq2* package performs independent filtering, in order to filter out from the procedure those tests that have no, or little chance of showing significant evidence, before without even looking at their statistics.
As underlined in the study of Bourgon and Gentlemen [52], this permits to increase detection power at the same experiment-wide type error I, using a two-stage approach that filters variable by a criterion independent from the statistics, and then testing only those variables that passes the filter.
Independent filtering is performed here using the mean of normalized counts as a filter. Wald test p-values of the subset of genes that have passed the filtering phase are adjusted using the Benjamini and Hochberg False Discovery Rate [53] procedure. At the end, the adjusted p-values for the genes which do not pass the filter threshold are set to NA.
The filter threshold value and the number of rejections at each quantile of the filter statistic are available as metadata of the object returned as result.

2.2.1 CORRELATION ANALYSIS

After a first phase of differential expression analysis, a test to verify the correlation between fold changes of identified differentially expressed genes and cell lines predicted response to pharmacological treatment (ATRA-*score* [38]), has been carried out.
After consulting the study of Hauke and Kossowski [54], we decided to use both Spearman's and Pearson's correlation coefficients, in order to have a more inclusive result, independent from the procedure of association. In fact, Pearson's correlation coefficient is a measure of the strength of the linear relationship between two variables, whether Spearman's rank correlation coefficient is a nonparametric (distribution-free) rank statistic used in this case as a measure of the strength for the same comparison. More in details, it is a measure of a monotone association, that can be used when the distribution of data may make Pearson's correlation coefficient someway misleading. Unlike Pearson's product-moment correlation coefficient, Spearman's correlation coefficient does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level. In principle, it can be considered a special case of Pearson's product-moment coefficient in which the data are converted to ranks before calculating the coefficient.

For that reason, a first correlation test using Pearson's method has been carried out, between each row of the fold change matrix obtained for differentially expressed genes (i.e. the fold change computed for each i-th gene in each j-th sample) and the ATRA-*score* vector. When the statistic test is based on Pearson's product moment correlation coefficient, it follows a *t*-distribution with *length(x)-2* degrees of freedom. The asymptotic confidence interval is given based on Fisher's Z transform. Secondly, when Spearman's method has been used to estimate the rank-based measure of association, p-values are computed via the asymptotic *t* approximation.

2.2.2 VARIATION COEFFICIENT

To further select only those genes showing a variation across samples that is sufficient to result in a biologically significant action, we decided to compute a coefficient of variation, as defined in Equation 2:

$$VC = Sd\{Matrix[i,]/mean(Matrix[i,])\} * 100 \quad (2)$$

For each row of the normalized count matrix, i.e. for each i-th gene, we determined the ratio between the value of its expression in each sample and the mean value across all the samples. Finally, the normalized standard deviation of this value has been represented in percentage. On the basis of this additionally parameter, a more restrictive selection of genes has been performed, considering only those genes with VC>50%.

2.3 GENE SET ENRICHMENT ANALYSIS

Gene set enrichment analysis has been performed using the package *Limma* in R environment [55].

The collection of annotated gene sets here used (more than 10.000) is provided by the Molecular Signatures Database MSigDB (http://software.broadinstitute.org/gsea/msigdb). In this compendium, gene set collections are organized in eight major categories, as summarize in Table 3.

To our aim, we decided to focus on the Hallmark, C2 (KEGG and REACTOME) and C5 (Gene Ontology Biological Process, Molecular Function and Cellular Compartment) collections.

Name of the collection	Number of gene sets	Description
HALLMARK gene sets	50	Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression.
C1: Positional gene sets	326	Gene sets corresponding to each human chromosome and each cytogenetic band that has at least one gene.
C2: Curated gene sets,	4762	Gene sets curated from various sources such as online pathway databases, the biomedical literature, and knowledge of domain experts.
C3: Motif gene sets,	836	Gene sets representing potential targets of regulation by transcription factors or microRNAs. The sets consist of genes grouped by short sequence motifs they share in their non-protein coding regions. The motifs represent known or likely cis-regulatory elements in promoters and 3'-UTRs.
C4: Computational gene sets	858	Computational gene sets defined by mining large collections of cancer-oriented microarray data.
C5: Gene Ontology (GO) gene sets	5917	Gene sets that contain genes annotated by the same GO term. The C5 collection is divided into three sub-collections based on GO ontologies: Biological Process, Cellular Compartment, and Molecular Function.
C6: Oncogenic signatures	189	Gene sets that represent signatures of cellular pathways which are often dis-regulated in cancer. The majority of signatures were generated directly from data from NCBI GEO .
C7: Immunologic signatures,	4872	Gene sets that represent cell states and perturbations within the immune system. The signatures were generated by manual curation of published studies in human and mouse immunology.

Table 3. 3 MSigDB collections (adapted from (http://software.broadinstitute.org/gsea/msigdb))

Competitive gene set tests are widely used in molecular pathways analyses: they are useful to test for enrichment of a particular gene annotation category among the differential expression results from RNA-sequencing analysis, to display statistically significant and concordant differences between two biological states. Briefly, they are similar to differential expression analyses in which as set of genes is considered as a unit and therefore associated with a p-value.

Many traditional competitive tests assume independence of genes [56], evaluating p-values by permutation of gene labels, because they rely on parametric approximations that are asymptotically equivalent to a gene permutation [57]. Among them, the very popular Gene Set Enrichment Analysis (GSEA) [58] procedure, uses sample permutation to test the significance of a competitive gene test statistics, but this may result in a hybrid test for which the null and alternative hypothesis are difficult to characterize in terms of population parameters.

To avoid these critical issues, we decided to perform this kind of analysis using a new gene set test procedure presented by Wu and Gordon [59], *CAMERA*, that is based on the idea of estimating inter-gene correlation from the data and then use them to adjust the statistics of the test.

Starting from a given collection of gene set tests, CAMERA tests whether the genes in the set are highly ranked in terms of differential expression relative to genes not in the set.

This algorithm can be used for any RNA-sequencing experiment resulting in log-expression values y_{gi} for genes i = 1, … ,G and RNA samples i =1, …, n. Assuming a linear model for the expected value of each expression value given the experimental design,

$$E(y_{gi}) = \mu_{gi} = \sum_{j=1}^{p} \alpha_{gj} x_{ij} \qquad (3)$$

where x_{ij} are the design variables specifying which treatment condition is associated with each RNA sample, whether α_{gj} are unknown regression coefficient representing expression fold changes between the two experimental condition in the experimental set up.

Each gene is supposed to have its own variance $var(y_{gi}) = \sigma_g^2$.

CAMERA estimates p-values after adjusting the variance of statistic tests by an estimated variance inflation factor (VIF).

Let's consider a set of m genewise statistics z_1, .., z_m. The variance of the mean of the statistics is defined as

$$\text{var}\,\bar{z} = \frac{1}{m^2}\left(\sum_{i=1}^{m}\tau_i^2\right) + \sum_{i<j}\rho_{ij}\tau_i\tau_j \quad (4)$$

where τ_i is the standard deviation of z_i and the $\rho_{ij}$are the pairwise correlations. The second term here represents the increase in the variance of the mean that derives from the correlation between the genes. In case all the τ_i are equal to τ

$$\text{var}\,\bar{z} = \frac{\tau^2}{m}\,VIF \quad (5)$$

where the variance inflation factor is equal to c, being $\bar{\rho}$ the average of ρ_{ij}.

Notably, the inflation factor depends on estimated genewise correlation and the number of genes in the gene set.

To estimate the inter-gene correlation $\bar{\rho}$, let's consider Y = $\{y_{gi}\}$ for the *m* x *n* matrix of genes in the test set. Here, rows correspond to genes and columns to RNA samples. We assume that the expression values can be represented by genewise linear models with a *n* x *p* design matrix X = $\{x_{ij}\}$. The rows of the design matrix correspond to RNA samples and the columns to coefficients of the linear model.

To estimate the average pairwise correlation, the first step is the computation of *d* independent residual for each gene. Let's now consider X = QR for the QR-decomposition of the design matrix, where Q is *n* x *n* and R is *n* x *p*. Here, R is upper-triangular and Q satisfies $Q^TQ = 1$. The *n* x *d* matrix of independent residual is obtained by $U = YQ_2$, Q_2 representing the trailing *d* columns of Q.

At this point, correlation matrix for the *m* genes can be obtained as $C = UU^T$. As m is large, the column means u_k of U are computed, so that the estimate of VIF became

$$\widehat{\text{VIF}} = \frac{m}{d}\sum_{k=1}^{d}u_k^{-2} \quad (5)$$

In this case, the estimate of the average correlation equal to the average of all pairwise correlations in the matrix C, can be estimated by solving $VIF = 1 + (m-1)\bar{\rho}$ for $\bar{\rho}$.

The estimate of the mean pair-wise correlation within each set of genes is implemented in the function `interGeneCorrelation`:

- if `interGeneCorrelation = NA`, the algorithm will estimate the inter-gene correlation for each set. In this way, it gives a rigorous error rate control for all sample sizes and all gene sets.
- If `interGeneCorrelation = 0.01`, CAMERA will rank biologically interpretable sets more highly. This gives a useful compromise between strict error rate control and interpretable gene set rankings.

For our analysis, we decided to use a slightly modified version of the presented algorithm, called *cameraPR*: this is a "pre-ranked" version of CAMERA where the genes are pre-ranked according to a pre-computed statistic. In this case, the statistical values given to the function arises the significance (p-value) of the statistical correlation parameter calculated through Pearson's procedure presented in the previous paragraph.
As final result, CAMERA return a matrix with a row for each gene set tested and a column for each of the following parameters: number of genes in the set, direction of change ("up" or "down"), a two-tailed p-value and the Benjamini and Hochberg adjusted p-value (FDR).

3. CHIP-SEQUENCING DATA ANALYSIS

CHIP-sequencing raw data have been obtained from the NCBI GEO repository (www.ncbi.nlm.nih.gov/geo/) series GSE60272 and are publicly available. In particular, a subset composed by GSM1469981, GSM1469982, GSM1469983, GSM1469984, GSM1469985, GSM1469986, GSM1469987, GSM1469988, GSM1469989, GSM1469990, GSM1469991 has been chosen for our analysis. We took into consideration for higher level analysis only those experiments where biotin ChIP-sequencing have been performed, for wild-type (WT) forms of two RAR transcription factors (RARα/γ) in MCF-7 breast cancer cells upon retinoic acid (RA) and oestrogen (E2) stimulation.

3.1 LOW-LEVEL PROCESSING

Raw data in "FASTQ" file format have been aligned to the reference genome (hg38), using the Burrows-Wheeler Alignment Tool (BWA), a read alignment package that is based on backward search with Burrows-Wheeler Transform (BWT) [60]. This represent an effective

method to align short sequencing reads (50 bp) against a large reference sequence, such as the human genome.

Actually, using backward search [61] with BWT, the algorithm is able to effectively mimic the top down traversal on the prefix trie of the genome with relatively small memory footprint [62] and to count the number of exact hits of a string (read) of a generic length *m* in a time independent of the size of the genome.

For inexact search, BWA usually sample from the implicit prefix trie the different substrings that are less than *k* edit distance away from the query read. Because exact repeats are collapsed on one path on the prefix trie, we don't need to align the reads against each copy of the repeat. This is the main reason why BWT-based algorithms are efficient and BWA has been chosen among different category of software available [63, 64, 65].

The prefix tie for a generic string *X (*in our case the human genome) is a tree where each edge is labelled with a symbol and the string concatenation of the edge symbols on the path from the leaf to the root gives the unique prefix of *X*. Hence, the string concatenation of the edge symbols from a node to the root gives a unique substring of *X*, called the string represented by the node.

Notably, the prefix trie of X is the exact copy of the suffix trie of reverse X and therefore suffix trie theory can also be applied to prefix trie.

Once defined the prefix trie, testing whether a generic query *W* is an exact substring of *X* is equivalent to finding the node that represents *W*, which can be done by matching each symbol in *W* to an edge, starting from the root.

To allow mismatches, we can completely traverse the trie and match *W* to each possible path. It is also possible to accelerate this search by using prefix information of *W*.

To compute the Burrows–Wheeler transform of the human genome, let's consider Σ being an alphabet. Consider now a symbol \$, that is not present in Σ and is lexicographically smaller than all the symbols in Σ. A string $X=a_0a_1 \ldots a_{n-1}$ is always ended with symbol \$ (i.e. a_{n-1}=\$) and this symbol only appears in the end.

Let $X[i]=a_i$, $i=0,1,\ldots,n-1$, be the *i*-th symbol of *X*, $X[i,j]=a_i \ldots a_j$ a substring and $Xi=X[i,n-1]$ a suffix of *X*.

Suffix array (SA) *S* of *X* is a permutation of the integers $0,\ldots,n-1$ such that $S(i)$ is the start position of the *i*-th smallest suffix.

Now, we can define the BWT of *X* as $B[i]$=\$ when $S(i)=0$ and $B[i]=X[S(i)-1]$ otherwise.

In practice, the algorithm usually construct the suffix array first, and then generate the BWT of the reference genome.

To describe the final sequences alignment procedure, we must first consider the suffix array intervals. If string W is a substring of X, the position of each occurrence of W in X will result in an interval in the suffix array. This is because all the suffixes that have W as prefix are sorted together. Based on this observation, we can define:

$$\underline{R}(W) = \min\{k : W \text{ is the prefix of } XS(k)\} \quad (7)$$

$$\bar{R}(W) = \max\{k : W \text{ is the prefix of } XS(k)\} \quad (8)$$

In particular, if W is an empty string, $R(W)=1$ and $R(W)=n-1$. The interval $[R(W),R(W)]$ is called the *SA interval* of W and the set of positions of all occurrences of W in X is

$$\{S(k) : R(W) \leq k \leq R(W)\}. \quad (9)$$

Knowing the intervals in suffix array we can define the positions. Hence, sequence alignment is equivalent to searching for the suffix array intervals of substrings of X that match the query. For the exact matching problem, we can find only one such interval.

The *backward search* procedure, can be now explained as follow. Let $C(a)$ be the number of symbols in $X[0,n-2]$ that are lexicographically smaller than $a \in \Sigma$ and $O(a,i)$ the number of occurrences of a in $B[0,i]$. Ferragina and Manzini [61] proved that if W is a substring of X:

$$\underline{R}(aW) = C(a)+O(a,R(W)-1)+1 \quad (10)$$

$$\bar{R}(aW) = C(a)+O(a,R(W)) \quad (11)$$

and that $R(aW) \leq R(aW)$ if and only if aW is a substring of X. This result makes it possible to test whether W is a substring of X and to count the occurrences of W in $O(|W|)$ time by iteratively calculating $\underline{R}$ and $\bar{R}$ from the end of W. This procedure is called *backward search*.

Notably, equations (10) and (11) realize the topdown traversal on the prefix trie of X; given that we can calculate the SA interval of a child node in constant time if we know the interval of its parent. In these sense, backward search is equivalent to exact string matching on the prefix trie, but without explicitly putting the trie in the memory.

The default output alignment format is SAM (*Sequence Alignment Map* format). Once obtained, we used SAMtools (http:// http://samtools.sourceforge.net/) to convert it to its binary format (BAM file, *Binary Alignment Map* format) [66].

At this point, a BED (*Browser Exstensible Data*) file format has been obtained using *bedtools* (https://bfastqedtools.readthedocs.io/en/latest/), which was the required input for the further steps of analysis.

3.2 PEAK CALLING

A command line tool designed by Zhang and Liu [67], MACS (Model-based Analysis of ChIP-Sequencing), has been used to analyse pre-processed ChIP-Sequencing data.

Combining both Chromatin immunoprecipitation (ChIP) and high throughput sequencing (Seq), this popular technique permits to study the cistrome of transcription factors (TFs) [68], the genome-wide set of *in vivo cis*-elements bound by TFs, necessary to determine which genes are directly regulated by those transcription factors.

To identify TFs binding sites, the information previously obtained of mapped genomic locations for sequencing are needed: the input of the analysis was composed by mapped reads from CHIP-sequencing experiments, in the "BED" (*Browser Extensible Data*) file format, together with their control data, i.e. "input" DNA, that has been cross-linked and sonicated but not immuno-precipitated. Input DNA has also been aligned to the reference genome and obtained in the BED file format.

The final output of MACS is represented by the *narrowPeak* file format (a type of BED file) which contains locations of peaks and some measurements of their statistical significance.

Given the ChIP-Sequencing data with the correspondent control sample, MACS can be used to identify transcription factor binding sites, using a two-step strategy: modelling the reads shift size, and then peak calling [69].

On the basis of the reads distribution, MACS analytically models the shift size of ChIP-sequencing reads. As ChIP-DNA fragments are equally likely to be sequenced from both ends, the reads density around a real TFs binding site is likely to show a bimodal enrichment pattern, with forward strand reads enriched upstream of binding and reverse strand reads enriched downstream (Figure 9).

Given two user-dependent parameters, *bandwidth* (300 bp as default, maintained in our study) and *mfold* (a high-confidence fold-enrichment interval), MACS scans 2*bandwidth windows across the genome to find regions with certain reads enrichment relative to the expectation (larger than 10 fold and smaller than 30 fold, as default). MACS selects only

these high-quality peaks, separates their forward and reverse reads, and then aligns them by the midpoint.

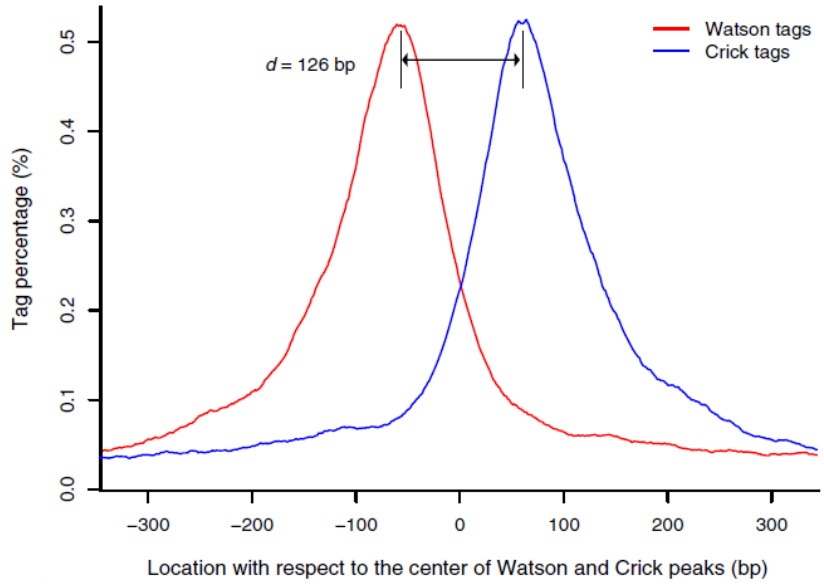


Figure 9. MACS model for ChIP-Seq (adapted from [69]).

The 5' ends of strand-separated tags from a random sample of 1,000 model peaks, aligned by the centre of their forward strand reads (Watson) and reverse strand reads (Cricks).

The distance between the modes of the forward and reverse peaks in the alignment is defined as *d*, and MACS shifts all the reads by *d/2* toward the 3' end to better locate the precise binding sites. It is worth observing that the parameter *fold* is used in the procedure only in the first step, where a suitable *mfold* parameter will lead to several thousand paired peaks from ChIP-Sequencing data for model building.

Within the genome coverage of ChIP-Sequencing experiments, reads distribution along the genome could be modelled by a Poisson distribution [70], which can express the probability of a number of events (lambda) happening in a fixed period (in this case distance along the genome). It takes this single parameter, lambda, to define the expected number of instances that occur in the given region.

As the background level varies across the genome, during the phase of peak calling, instead of using a uniform λ_{BG} (of the entire background) estimated from the whole genome, MACS uses a dynamic parameter, λ_{local}, defined for each reads enriched region, that can be defined as

$$\lambda_{local} = max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda v_{10k}) \quad (12)$$

where $\lambda 1k$, $\lambda 5k$ and $\lambda 10k$ are λ estimated from the 1 kb, 5 kb and 10 kb window centred at the peak location in the control sample.

Using λ_{local} is useful to capture influence of local biases, making the value robust against occasional low read counts at small local regions.

MACS applies λ_{local} to calculate the p-values for each read enriched region, and only those regions with p-values below a user-defined threshold (10^{-5}) are reported as identified peaks, being the ratio between the ChIP-Sequencing read counts and λ_{local} the *"fold enrichment"*.

For each peak identified, the detailed output information includes the chromosome number, the start position, the end position, the length of peak region, the summit location related to the peak start position, the number of reads in peak region, the fold enrichment for this region (compared to the expectation from Poisson distribution with local lambda) and finally the False Discovery Rate (FDR).

3.3 PEAK ANNOTATION

To perform batch annotation of enriched peaks identified from CHIP-sequencing data, we used a package available in *Bioconductor* [71], an open source and open development software project specialized in biological data analysis and integration, within the statistical programming environment R.

The package *ChIPpeakAnno* [72] in *Bioconductor* uses the *IRanges* package and represents the peak list as *RangedData* to accurately find the nearest or overlapping gene, exon, 5'UTR, 3' UTR, microRNA (miRNA) or transcription factor binding sites.

In the previous phase of analysis, as final output of the peak calling algorithm (MACS2), we obtained a file containing a list of chromosome coordinates in a BED (*Browser Extendible Data*) file format, that is all *ChIPpeakAnno* package needs.

Moreover, even if the genome annotations are update periodically, upon the pre-built annotation data packages (*TSS.human.NCBI36, TSS.human.GRCh37, ExonPlusUtr.human.GRCh37*), the user has the possibility to customize the annotation data, following his specific purpose. To our aim, we loaded in the environment an Ensembl-based annotation package for Homo Sapiens, Ensembl version 86 (*EnsDb.Hsapiens.v86*). One of the main advantages of this package is the fact that it generates "versioned" annotation packages, i.e. annotation packages that are built for a specific Ensembl release (version 86 in our study) and are also named according to that. This ensures reproducibility of our

analysis, as it allows to load annotations from a specific Ensembl release also if newer versions of annotation packages became available.

To annotate peaks, we used the function *annoPeaks* implemented in the *ChIPpeakAnno* package. Once loaded the annotation data, in the form of a *GRanges* object, the criteria to associate peaks with annotations must be specified. In our case, we set up parameters to obtain peaks within 2kb upstream and up to 2kb downstream from Transcription Start Site (TSS) within the gene bodies.

The output of this phase of analysis, is a *GRanges* object of the annotated peaks.

4. METHYLATION ARRAYS DATA ANALYSIS

The DNA-methylation data were provided by the Broad Institute Cancer Cell Line Encyclopaedia (CCLE) and downloaded from the cBioPortal for Cancer Genomics online archive. DNA-methylation profiles have been collected from the NCBI GEO repository (www.ncbi.nlm.nih.gov/geo/) series GSE68379.

All this data were obtained through the HumanMethylation450 BeadChip Array Platform, an Illumina scanners which give as output binary two-colour .IDAT files (pair of files with names ending in *_Red.idat* or *_Grn.idat*).

All the analyses and the processing of the DNA-methylation data were performed using R environment [46,47].

The methylation array experiments were read through the function *read.metharray.exp* available in the Bioconductor package *minfi* [73] in R: the function finds all IDAT files in the directory and returns an unique object of class *RGChannelSet.*

4.1 LOW LEVEL PROCESSING

The sequential processing of samples can give rise to array-to-array variation in background fluorescence, which contribute to an additive error to the measured signal: it can arise from many sources, such as non-specific binding or spatial heterogeneity across the array. To overcome this critical issue, we applied a background correction method in order to estimate the true signal from the observed foreground, modelled as the sum of true signal and ambient

signal [74]. For all background-correction methods, probes are pooled and then corrected within each single colour channel.

As suggested in the study of Liu and Siegmund [75], we decided to implement in R a normal exponential convolution method (*Noob*) using the function *preprocessNoob*, which returns an object of class *MethylSet*, implemented in the Bioconductor package *minfi* [73].

The analysis proceeded as follows. Let's consider the background signal normally distributed $X_B \sim N(\mu,\sigma^2)$ and the true signal following an exponential distribution $X_S \sim Exp(\gamma)$, X being either the Green or Red channel, and the observed foreground intensity as their sum $X_f = X_S + X_B$.

For each channel separately, the parameters are estimated from the background distribution using the small number of control probes (n= 614 for the HumanMethylation450, designed to not match any genomic regions and thus measure background fluorescence), while the signal parameter γ is obtained by subtracting the background mean from the observed foreground intensities ($\gamma = X_f - \mu$). The conditional expectation of the signal, given the observed foreground and background, is computed by:

$$E = [X_S|X_f] = \mu_{sf} + \sigma^2 \frac{\phi(0; \mu_{sf}, \sigma^2)}{1-\Phi(0; \mu_{sf}, \sigma^2)} \qquad (12)$$

where $\mu_{sf} = X_f - \mu - \sigma^2/\gamma$, $\phi(\cdot)$ is the standard normal density and Φ is the cumulative normal distribution. The conditional expectation allows the estimation of the background corrected intensities and is used to smoothly interpolate probes with intensities near the background level [76].

4.2 PROBE DESIGN BIAS CORRECTION

A critical statistical issue when dealing the Illumina 450k BeadChip is the bias introduced by the two different types of assay chemistry technologies used, that lead to widely different distributions of the methylation values derived from these two probe designs (that we will refer to as type 1 and type 2 probe) [77].

For that reason, type 2 probe values must be normalized into type 1 values distributions, enabling them to be comparable and thus reducing the bias.

Using again as reference the study of Liu and Siegmund [75] in which it was found that the within-array combination of *Noob* + *BMIQ* always improved signal sensitivity, we implemented in R the Beta-Mixture Quantile dilation normalization strategy (*BMIQ*)

exploiting the function *bmiq.mc,* available in the Bioconductor package *ENmix,* that returns a data matrix of methylation Beta-values [78].

4.3 CORRELATION ANALYSYS

Methylation data have been subsequently tested for association with a defined parameter, the ATRA-*score*. To this aim, *dmpFinder* function implemented in the package *minfi* in R environment has been used. This function tests each genomic position for association between methylation and a "phenotype" or defined parameter. Given the ATRA-*score* as a continuous parameter, association has been tested with linear regression.

4.4 PROBE ANNOTATION

Essential analysis of 450k DNA-methylation data depends on annotating probes with their genomic location: in our study, we exploited the annotation information stored in the Bioconductor package "*IlluminaHumanMethylation450kanno.ilmn12.hg19*" [79] and the array design stored in the Bioconductor package "*IlluminaHumanMethylation450kmanifest*".

5. RETROVIRAL TRANSCRIPTS QUANTIFICATION

Nearly half of the human genome is constituted of repetitive elements that are tightly regulated to protect the host genome from destructive consequences associated to their inappropriate reactivation [80]. Both full length and fragmented copies of these viral genomes have propagated through host genomes to produce repeating instances of their sequences [81].

Major families of repeat elements are shown in Figure 10 and include autonomous and nonautonomous retrotransposons as well as DNA transposons. Aberrant reactivation of transposable elements has been shown to activate cell autonomous anti-viral response [82].

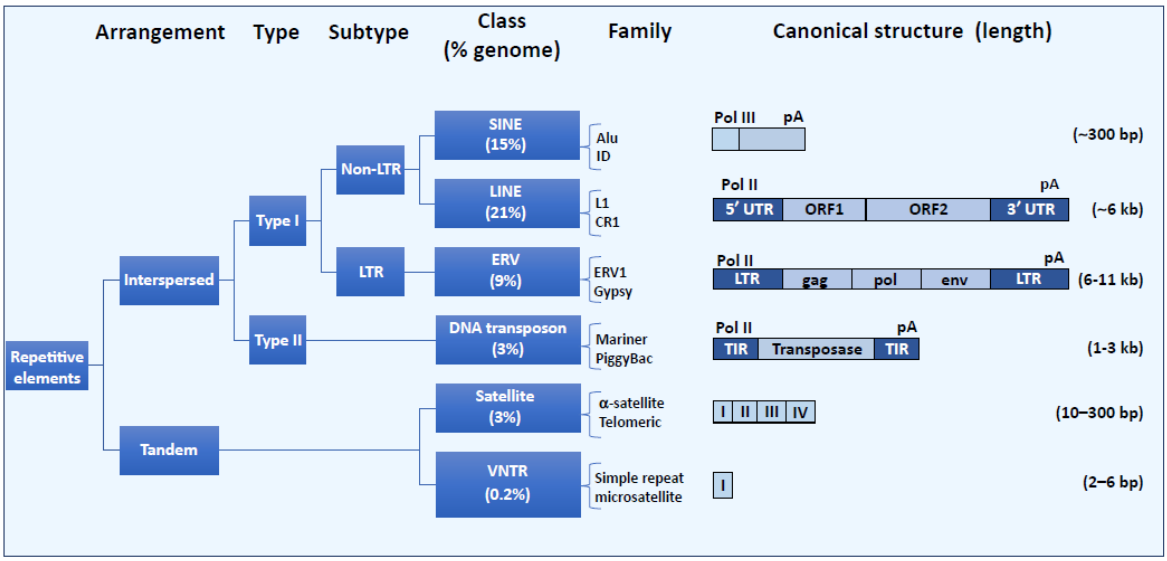


Figure 10. Classification and Organization of Repetitive Elements in the Human Genome [81].

Two examples of families are shown per class, and families are further stratified into subfamilies. Listed abundances are reported in brackets.

To quantify expression of these transposable elements, we retrieved their genomic positions from *RepeatMasker* database (http://www.repeatmasker.org/). These coordinates were assembled into a customized annotation file (*gene transfer file*, GTF), which was used to determine the abundance of all retroviral-derived transcripts, by using *FeatureCounts* [83]. To avoid detection of false positives, we discarded all transposable elements that show any overlap to known gene-associated exons, according to *Gencode* [39] annotations. Afterwards, viral RNA abundance was normalized for library size and tested for differential expression between ATRA-treated and untreated samples, using the same approach as described in 2.2. To this purpose, we stratified all the identified transposable elements according to their characteristic family (LINE/SINE/LTR/ALU) and normalized them together with the coding genes raw counts.

6. NETWORK GENERATION

The protein-protein interaction network was generated using the *stringApp* implemented in the Cytoscape Network Inference Toolbox, an open-source software environment for the large scale integration of integrating biomolecular interaction networks with high-throughput expression data [84].

On the base of the results obtained with the differential expression analysis, we entered a list of gene symbols in the STRING protein query [85] to import the matching network. STRING is a database of known and predicted protein-protein interactions which aims at collecting and integrate information about functional interactions between the expressed proteins, by consolidating known and predicted protein–protein association data for a large number of organisms.

The network was furtherly analysed using the MCODE algorithm implemented in *Cytoscape*, which allowed the identification of sub-networks. MCODE, a plugin developed to perform network module identification specifically in biology, weights nodes by local neighbourhood density, then performs an outward traversal from a locally dense seed protein node to isolate larger dense regions, and finally graphically displays extracted modules and associated information.

Results

1. DATA QUALITY ASSESSMENT

Principal components analysis was conducted on raw counts normalised on library size, averaged for sample replicates. It was performed to assess if data quality was consistent with respect to the experimental conditions and to identify possible outliers that might influence further analysis.

As shown in Figure 11, cell lines are grouped on the basis of their phenotype (luminal or basal), with a further division between basal cell lines which have or not gone through the Epithelial to Mesenchymal Transition (EMT) [86]. Moreover, this kind of analysis separates treated from un-treated samples, in a manner which appears to depend on the amplitude of their sensitivity to the pharmacological treatment.

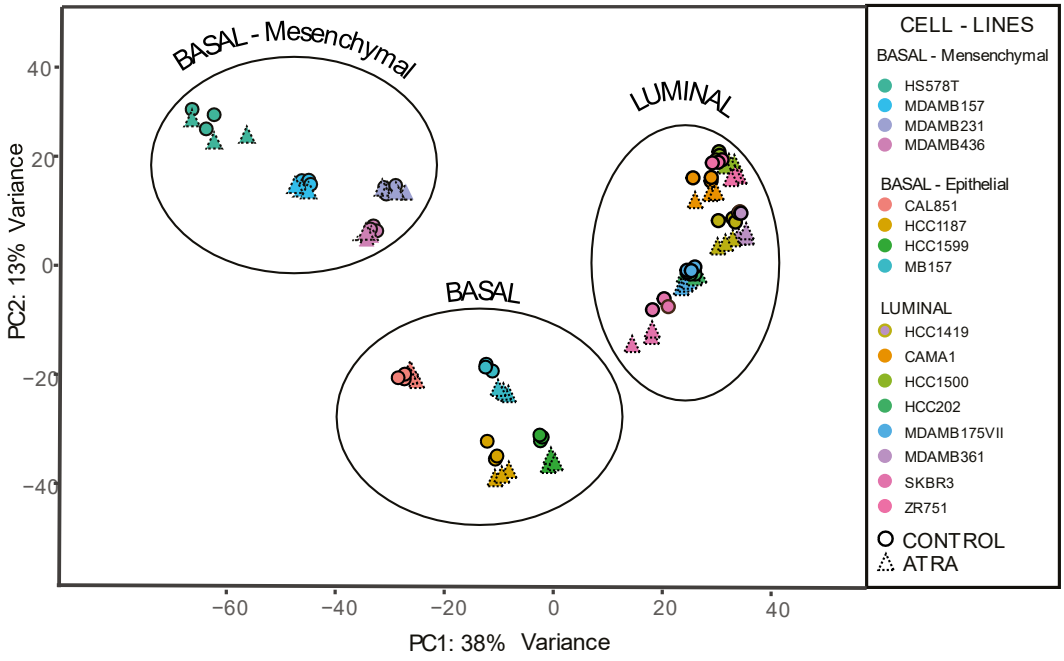


Figure 11. Principal components Analysis

The figure illustrates a PCA plot representing gene expression data averaged for each sample (cell line), summarized at the two first principal components coordinates. Each sample is represented with a different colour.

Unsupervised hierarchical clustering is represented in Figure 12. It groups the cell lines into two principal clusters on the basis of their phenotypes, with a perfect match between every

couple of treated and un-treated sample. Again, the value of the normalized counts for each sample is obtained as the average of the three replicates.

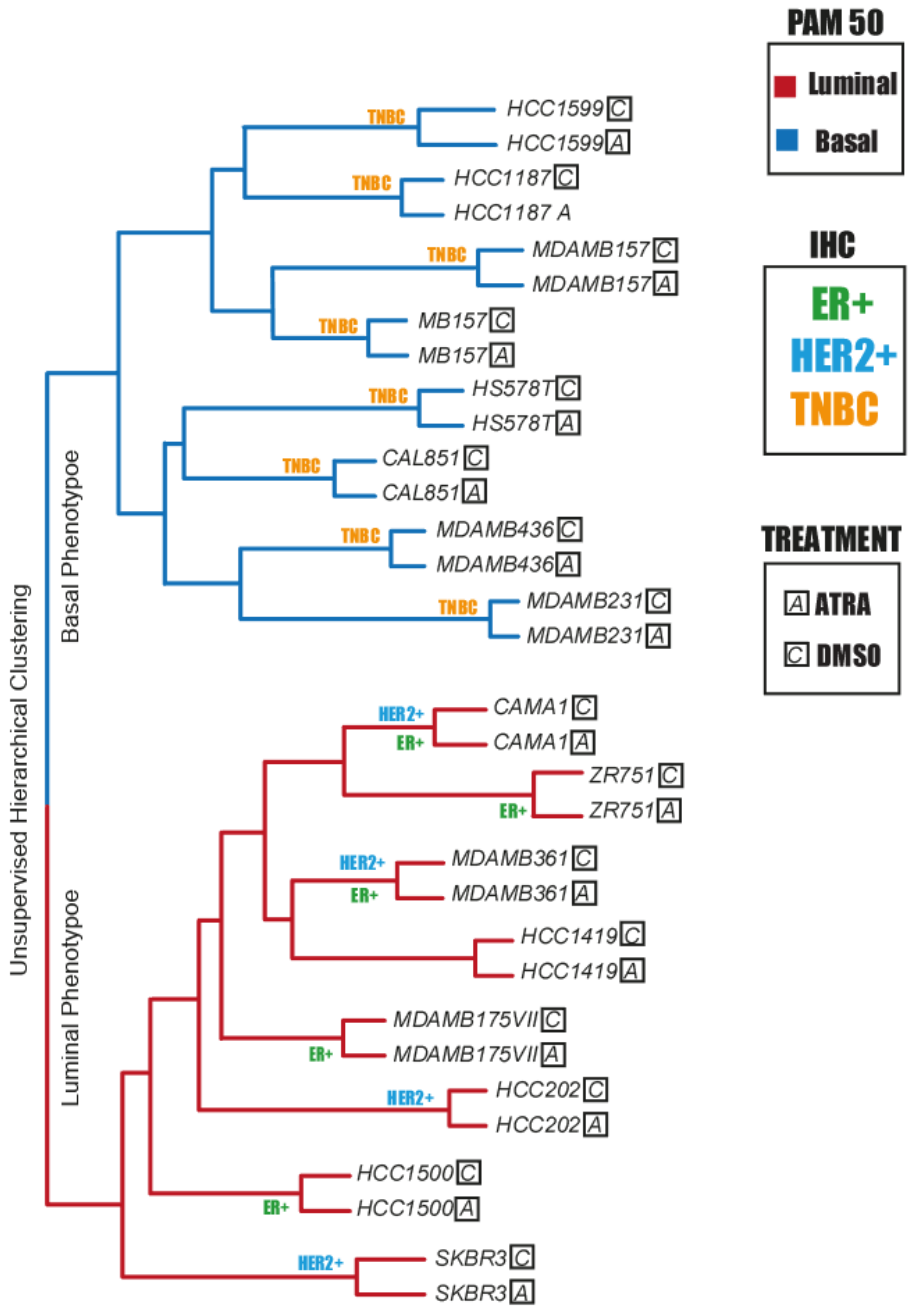


Figure 12. Unsupervised hierarchical clustering.

Cell lines are clustered based on their phenotype. Additional information about PAM50 and histological classification are reported for each cell line.

2. RETINOIC ACID - INDUCED TRANSCRIPTIONAL PERTURBATIONS

The overall amount of transcriptional perturbations induced by the pharmacological treatment with retinoic acid is directly proportional to the associated ATRA-*score*.

As shown in Table 13, the number of genes differentially expressed in each cell line decreases directly with the sensitivity-*score*, with a non-significant difference in the number of up-regulated and down-regulated genes.

CELL LINES	TOTAL	UP	DOWN	ATRA-SCORE
HCC1599	*6852*	3314	3538	1
SKBR3	*4919*	2480	2439	0.99
HCC1500	*1921*	798	1123	0.66
CAMA1	*1587*	996	591	0.66
MDAMB361	*2713*	1348	1365	0.58
MB157	*1509*	763	746	0.28
MDAMB157	*1107*	474	633	0.25
HCC202	*1168*	601	567	0.24
MDAMB175VII	*1803*	822	981	0.19
HS578T	*1515*	647	868	0.19
ZR751	*2232*	1300	932	0.14
HCC1419	*1998*	896	1102	0.09
MDAMB231	*330*	203	127	0.01
CAL851	*145*	85	60	0
HCC1187	*1255*	687	568	0
MDAMB436	*692*	454	238	0

Table 4. Number of differentially expressed genes in ATRA-treated cell lines.

Coloured conditional formatting highlights the correlation between the number of differentially expressed genes and the ATRA-*score*. The number of differentially expressed genes is computed considering only genes having a p-value < 0.05 after multiple test correction.

Indeed, the correlation between the value of the ATRA-*score* associated to each cell line and the number of its differentially expressed genes is equal to 0.833. To identify genes whose up/down regulation by the retinoid is directly correlated to ATRA-sensitivity, we determined the correlation between drug-induced fold changes of each individual gene and the cell lines predicted sensitivity to ATRA (ATRA-*score*). This step has been carried out, both calculating Pearson's product moment coefficient (R) and Spearman's coefficient (RHO). As a result, we obtained for each gene the correspondent correlation coefficient and p-value which represents the significance of the correlation.

We took into consideration for downstream analysis, only those gens with a Pearson's coefficient associated p-value (pR) < 0.01 or a Spearman's coefficient p-value (pRHO) < 0.01. This step resulted in a restricted list of 1776 genes.

Subsequently, we calculated the variation coefficient for each of the resulting genes, to identify those genes whose variation across samples is enough to result in a biologically significant action. On the basis of this parameter, we selected a group of 754 genes, which showed a variation higher than the 50% among all the samples.

Overall, our analysis identified 414 genes as negatively correlated with the sensitivity to retinoids and 340 as positively correlated. Genes with a positive correlation are represented in a protein - protein interaction network to a more effective visualization [Figure 13]. Genes with a negative correlation are similarly represented in Figure 14.

Most meaningful genes are selected based on the number of gene-neighbours (degree) they are connected to. These genes are highlighted in [Figure 13-14].

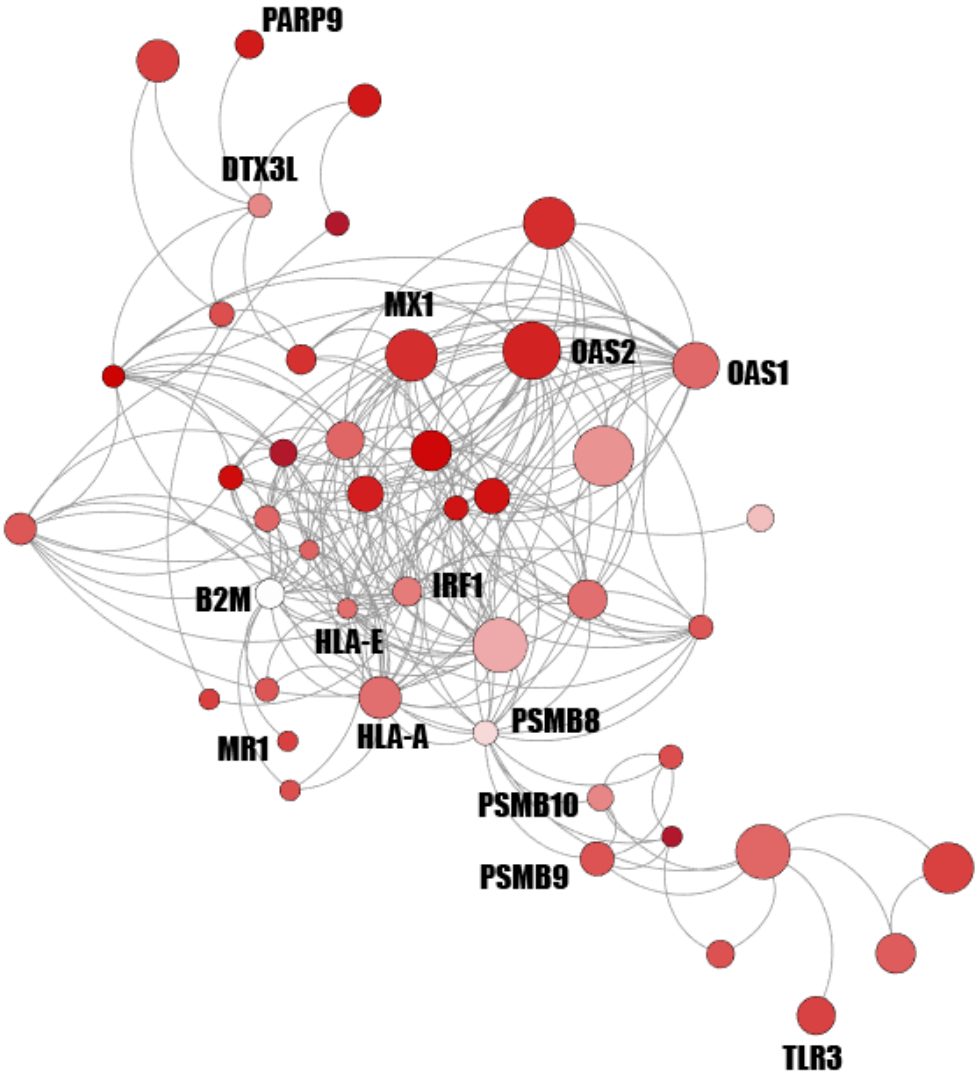


Figure 13. Network of protein-protein interaction based on positive-correlated genes.

OAS1 and OAS2 are genes which encode two members of the 2-5A synthetase family, essential proteins involved in the innate immune response to viral infection; HLA-A is a major histocompatibility complex (MHC) antigen specific to humans and one of three major types of human MHC class I cell surface receptors; HLA-E is a non-classical MHC class I molecule that is characterized by a limited polymorphism and a lower cell surface exposure and more important, it has a very specialized role in cell recognition by natural killer cells (NK cells); B2M is a gene that encode β_2 microglobulin, a component of MHC class I molecules; PSMB8, PSMB9 and PSMB10 genes encode members of the proteasome B-type family, that are a 20S core beta subunit. They are located in the class II region of the MHC complex; expression of these genes is induced by gamma interferon and their product

replaces a subunit of the immunoproteasome. MR1 encode the major histocompatibility complex class I-related protein, which is an antigen-presenting molecule, involved in the development and expansion of a small population of T-cells expressing an invariant T-cell receptor alpha chain. IRF1, Interferon regulatory factor 1, encode a member of the interferon regulatory transcription factor (IRF) family. It is shown to function as a transcriptional activator or repressor of a variety of target genes; it regulates expression of such genes by binding to an interferon stimulated response element (ISRE) in their promoters. MX1 encode the Interferon-induced GTP-binding protein MX1, which has an antiviral activity against a wide range of RNA viruses and some DNA viruses, targeting viral negative-stranded RNA; TLR3 gene encode a protein member of the toll-like receptor family of pattern recognition receptors, in the innate immune system. This proteins recognize pathogen-associated molecular patterns expressed on infectious agents, and mediate the production of cytokines necessary for the development of effective immunity.

Finally, we highlight DTX3L ubiquitin-protein ligase which, in association with ADP-ribosyltransferase PARP9, plays a role in DNA damage repair and in interferon-mediated antiviral responses.

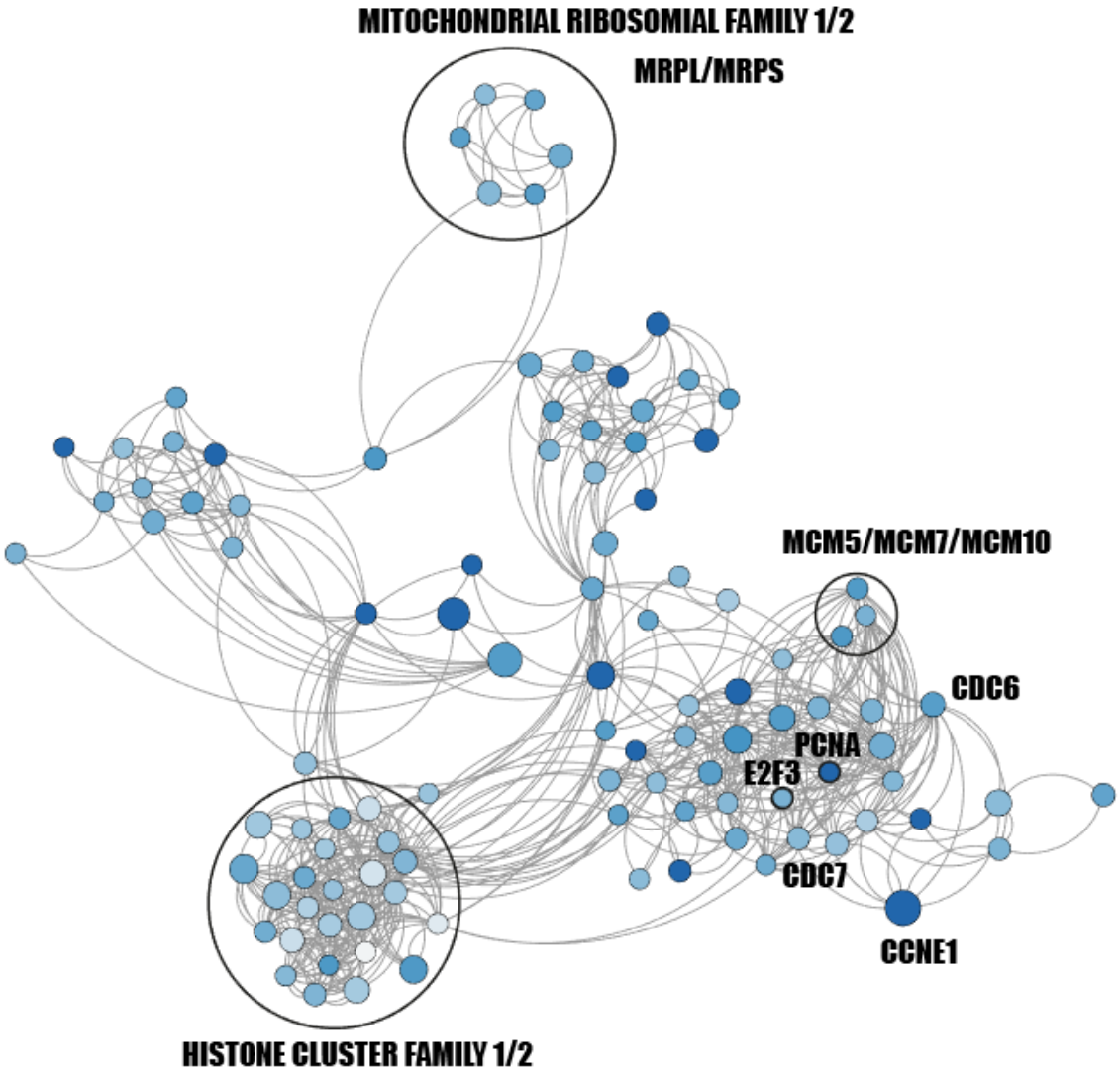


Figure 14. Network of protein-protein interaction based on negative-correlated genes.

E2F3 encodes a protein that is a member of the E2F family of transcription factors, which plays a crucial role in the control of cell cycle and synthesis of DNA; PCNA encode for the Proliferating Cell Nuclear Antigen, that is found in the nucleus and acts as a cofactor of DNA polymerase delta. It also operate increasing the processivity of leading strand synthesis during DNA replication. CDC6 and CDC7 encode the cell division cycle 6/7-related protein kinase, involved in the regulation of the cell cycle at the point of chromosomal DNA replication. They are also required for loading minichromosome maintenance (MCM) proteins onto the DNA, as an essential step in the initiation of DNA synthesis. CCNE1 belongs to the cyclin family, whose members are characterized by a periodicity in protein abundance through the cell cycle. Cyclins function as regulators of the Cyclin-dependent

kinases, which are involved in regulating transcription, mRNA processing and cell differentiation.

Moreover, the MCM5, MCM7, MCM10 genes encode three proteins of the MCM complex, DNA helicases essential for genomic DNA replication. They are critical proteins for cell division and the complex is also the target of various checkpoint pathways, such as the S-phase entry and S-phase arrest checkpoints. Both the loading and activation of MCM helicase are strictly regulated and are coupled to cell growth cycles. MRPL and MRPS family are nuclear genes which encode mitochondrial ribosomal proteins that help in protein synthesis within the mitochondrion.

Finally, a wide down-regulation of genes belonging to the Histone Cluster Family 1 and 2 can be underline: they encode nuclear proteins involved in the maintenance of the nucleosome structure of the chromosomal fibres.

For the genes which show more than 5 edges in the protein- protein interaction networks above, information about their variation coefficient, their correlation (R, RHO) with the ATRA-*score* and the statistics of the correlation (p-values) are reported in Appendix I.

3. RETINOIC ACID-INDUCED PATHWAYS PERTURBATIONS

To analyse the functional enrichment of differentially expressed genes, gene set enrichment analysis was performed, considering a various collection of gene sets, from MSigDB collection. We considered as significant, enrichments with an FDR adjusted P-values below a cut-off threshold of 0.1. We determined pathway enrichments using a pre-ranked order of genes, namely those showing a significant direct or indirect correlation between fold induction and ATRA-sensitivity (ATRA-*score*).

Figure 15 shows the significantly enriched gene sets of the Hallmark collections: the major negatively-correlated gene sets are related to cell cycle progression (E2F and G2M) and to oxidative phosphorylation, with FDR adjusted p-values < 0.001. The more significant gene sets that exhibit a positive correlation with the *score* can be all associated to cellular response to interferon alpha and gamma activation (FDR adjusted p-value < 0.001) and with the inflammatory response (FDR adjusted p-value < 0.01), as shown in Figure 15.

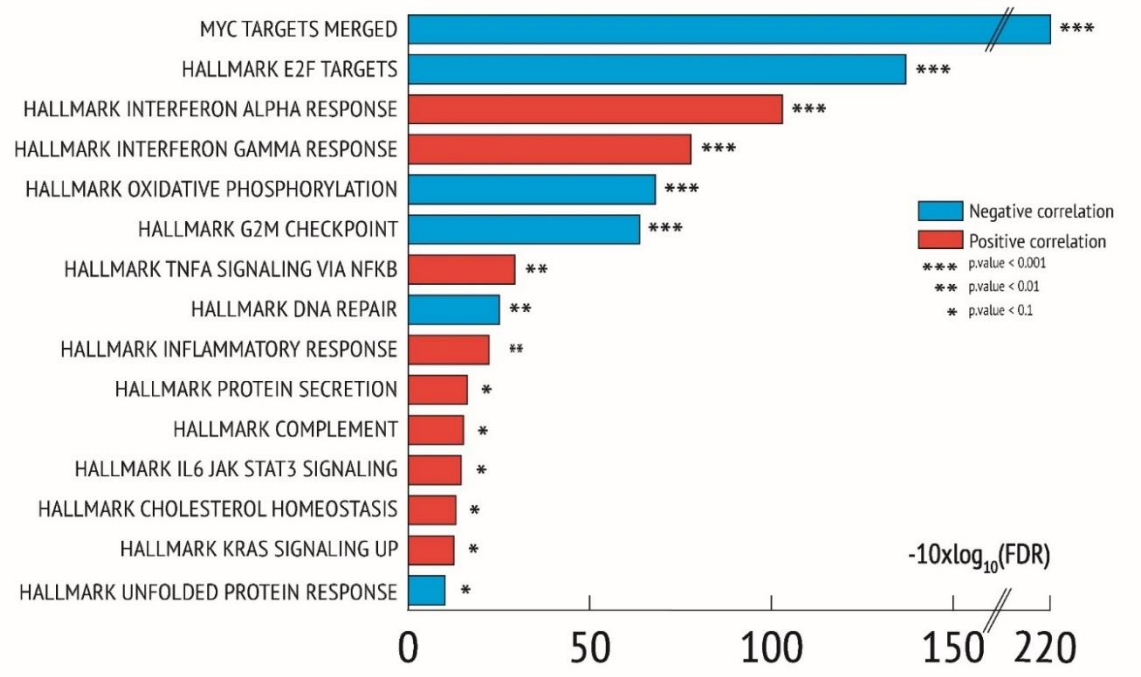


Figure 15.Hallmark collection.

The figure illustrates the result of the gene set enrichment analysis using the Hallmark collection

GO term enrichment analysis showed that negatively-correlated genes were significantly enriched in 231 biological processes, while positive-correlated genes in 156 biological processes (FDR adjusted P-value <0.1). The main GO biological process terms for negatively-correlated genes showed a wide variety of functional processes ranging from RNA splicing and progression, DNA repair, cellular developmental process, cell cycle progression, cell differentiation, cell development and regulation of system processes (FDR adjusted p-value $<10^{-5}$). On the other side, the primary GO terms for positively-correlated genes are related to interferon alpha, beta and gamma response, innate immune response and vesicular- mediated transport (FDR adjusted p-value < 0.001).

GO term enrichment analysis based on cellular compartments and molecular function relate negatively-correlated genes to the helicase and the endonuclease activity in the nucleus, whereas positively-correlated genes show enrichment for transport mediated by vesicles in the Golgi apparatus.

A table that provides an overview of the significantly enriched GO terms is provided in Appendix II. The table reports the name of the collection, the number of genes it involves, the p-value and the FDR adjusted P-value of the enrichment analysis.

Gene set enrichment analysis for KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways has also been performed using the same significance threshold of 0.1.

Figure 16 shows all significantly enriched gene sets. The main KEGG pathways for negatively-correlated genes are involved in all the major DNA repair mechanisms and in the cell cycle progression; positively-correlated genes exhibit an enrichment for those gene set which represent various type of response to viral infections.

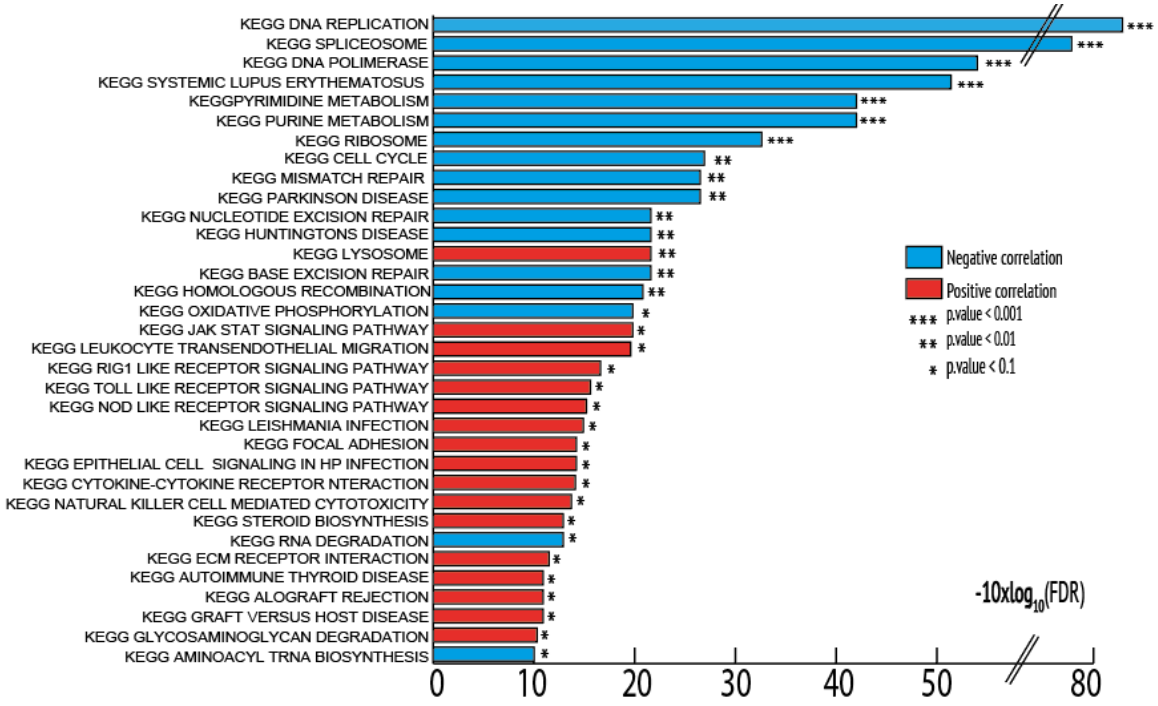


Figure 16. KEGG collection.

The figure illustrates the result of the gene set enrichment analysis using the KEGG gene sets collection.

REACTOME database of reactions, pathways and biological processes has also been investigated. The main Reactome pathways related to negatively-correlated genes are significantly enriched in 132 biological processes, while pathways explored with positive-correlated genes in 48 (FDR adjusted P-value <0.1). As reported in Appendix III, significantly enriched terms for negative-correlated genes are related to mRNA processing, Chromosome maintenance, DNA elongation and, more in general, cell cycle development (FDR adjusted p-value $<10^{-7}$). Positively-correlated genes, show enrichment for those biological processes related to interferon alpha and beta signalling, antigen presentation and peptide loading of class I MHC and innate immune system activation (FDR adjusted p-value <0.001).

4. IDENTIFICATION OF DIRECT TARGETS THROUGH CHIP-SEQUENCING DATA ANALYSIS

ChIP-sequencing data set for two forms of the RAR transcription factors (RARA and RARG) obtained from the Gene Expression Omnibus (GEO) were used in this study to complement the previous analysis.

Prior to associating features of interest with peaks, we made a comparison between the data obtained from the peak calling phase in the two data sets of interest to evaluate the difference in the number of genomic regions that can be bind specifically by the two different transcription factors. Also, it is biologically interesting to obtain overlapping peaks from different ChIP-sequencing experiments to evaluate the potential formation of transcription factor complexes. To this aim, we obtained a Venn Diagram (Figure 17) showing the number of overlapping enriched genomic regions. In addition, the significance of overlap was determined with hypergeometric testing and an associated p-value attributed.

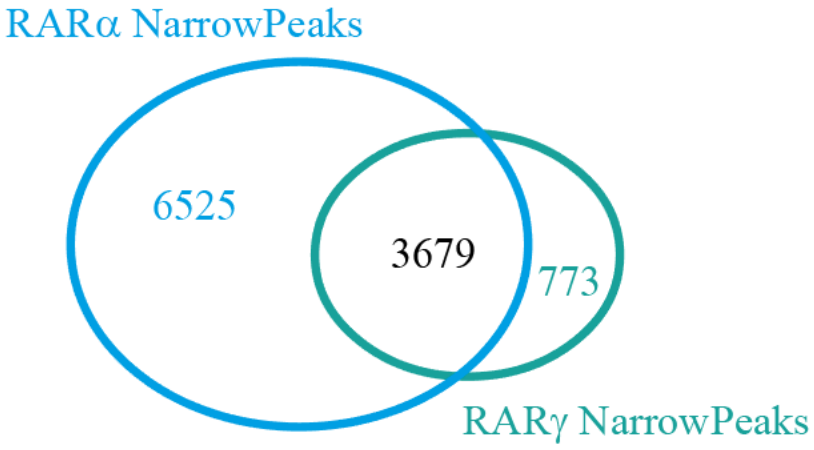


Figure 17. Venn diagram of overlapping peaks between RARA and RARG.

Next, we annotated peaks, to more precisely identify RARA and RARG binding sites. Exploiting the annotation tool described in chapter 3.3 of Material and Methods, we identify 3092 unique direct target genes for RARA and 1116 unique direct target genes for RARG. The size of the binding region considered is between 2 kb downstream and 2 kb upstream

the defined binding site, because the binding can take place both upstream and inside the gene body.

This additional information was then linked to the result of the differential expression analysis, to find out for which of the selected genes fold induction is related to direct binding of the transcription factor, whether it is RARA or RARG. 98 out of the 754 genes correlated with the ATRA-*score* are potentially direct targets of RARA; out of these, 30 genes are also potentially direct targets of RARG.

Interestingly, among the genes identified as RARA direct targets, it is of particular interest the presence of "interferon- related" genes, such as DTX3L and PARP9, and of one of the genes that encode a protein that is part of the MHC-I complex, HLA-E. In contrast, one of the more important transcription factors being involved in the regulation of the interferon signalling, IRF1, is not a direct target, neither for RARA transcription factor, nor for RARG. Appendix I reports for each gene which transcription factor (RARA, RARG or both) has been found to be able to directly bind its promoter.

5. CORRELATION BETWEEN METHYLATION LEVELS AND ATRA-*SCORE*

Data from the HumanMethylation450 BeadChip Array Platform have been analysed, in order to identify genomic regions that are differentially methylated with respect to the associated ATRA-*score*.

We retrieved methylation data for 52 breast cancer cell lines from GEO repository, and performed correlation analysis using 39 of these samples as 13 of these cell-lines were not part of our panel of 48 cell-lines previously profiled for retinoid sensitivity.

After the application of the *dmpFinder* algorithm, each of the 458k probes is associated with a p-value and a q-value (FDR adjusted p-value) based on the correlation between its methylation level and the associated ATRA-*score* in each cell line.

A significance-threshold of 0.01 results in the selection of 16648 probes out of the initial 458k. Each probe has been subsequently annotated to the nearest Transcription Start Site (TSS). As multiple probes fall within the same genes, when summarizing at the gene level, we obtained a final list of 7459 gene, which cointains features whose DNA methylation levels correlate positively/negatively with the sensitivity to the retinoid (p-value < 0.01). To

complement these results with the previous steps of the analysis, we identified genes, whose fold-induction was previously determined to be associated to ATRA-sensitivity, that also resulted to be differentially methylated: 298 out of these 754 genes have been found to have strong correlation between the ATRA-*score* and their methylation level.

Moreover, each of them can be characterized in term of number of methylated sites identified: in fact, each probe is able to identify only one methylated site, but each gene may have more than one methylated site all over its length. Genes that show multiple sites being associated to ATRA-sensitivity are likely to be more reliable, as it's highly unlikely that they represent false positives. Among the genes which show more than 1 differentially methylated sites associated to ATRA-sensitivity (143 genes), we can find a group of genes involved in the Interleukin-2 / STAT5 signalling: DHRS3, PHLDA1, ODC1, PLEC; CDCP1 and SPRY4 are part of a critical signalling pathway which entrains regulatory T cell differentiation and affects regulatory T cell function.

Furthermore, 73 out of the 298 genes are highly interconnected in the network defined during the initial step of differential expression analysis. In particular, among the genes with a higher level of methylation we can find TLR3, a crucial part of innate immune response, whose transcription is deeply inducted after treatment with retinoic acid. Moreover, two other genes involved in the innate immune response and induced by treatment with retinoic acid, HLA-E and PSMB8, show high methylation level in basal condition, which correlateswith the ATRA-*sensitivity.*

Appendix I reports information about their methylation status: if their promoters are differentially methylated and how many methylation counts can be reported.

6. GENOME-WIDE REACTIVATION OF ENDOGENOUS RETROVIRUSES

We quantified transcription of endogenous retroviruses through a differential expression approach. Starting from 5 million different sequences, we stratified all these identified transposable elements, according to their characteristic belonging family, (LINE/SINE/LTR/ALU), into 60 final subgroups. Appendix 3 shows how many sequences belongs to each identified family. The transcriptomic analysis performed (see chapter 6 of Material an Methods) indicates that ATRA induces a potent up-regulation of the transcription of endogenous retroviruses (retrotransposons) that mimics viral infection.

As shown in Figure 18, a widely distributed up-regulation of transcripts can be observed: induction (fold change) of the transcriptomic regulation slightly correlates with the sensitivity to ATRA-treatment. Cell lines which are completely resistant to the pharmacological treatment, display no induction (Appendix 3); then, transcriptomic up-regulation grows with an increasing ATRA-*score.*

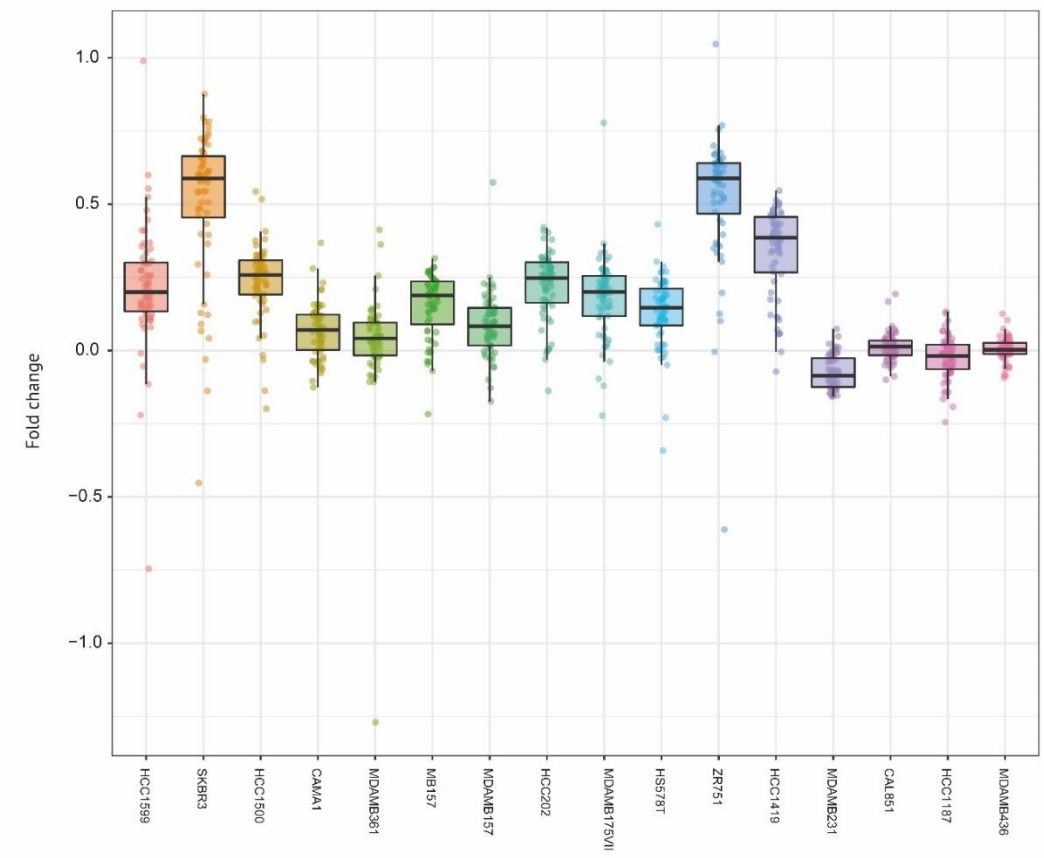


Figure 18 Induction of endogenous retroviruses.

Despite the presence of a general trend of correlation between the induction of the retroviral elements transcription and the sensitivity to retinoids, the transcriptional effect on a few cell lines (CAMA1, ZR751) doesn't follow the global behaviour.

To better understand the reasons of such tendency, we proceeded with further investigations. It has been shown that this general course tightly correlates with the expression levels of RARA in each cell line (Figure 19).

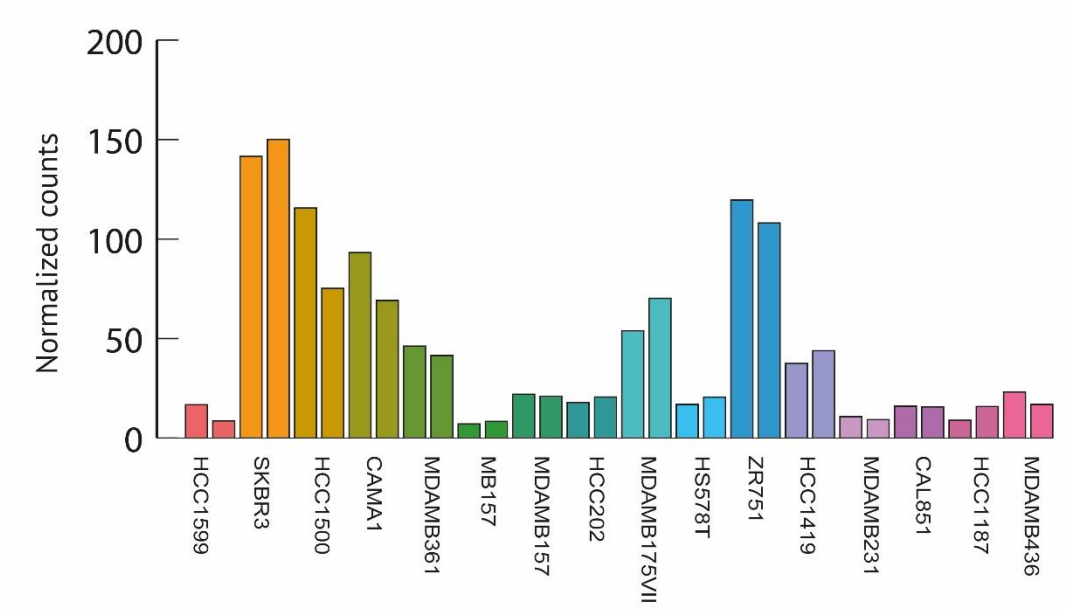


Figure 19 Basal Gene expression of RARA

For each cell lines we reported two bars: left bar represents expression level after ATRA treatment, right bar represents expression levels in untreated samples (DMSO). The 3 replicates for each condition are averaged.

To a more analytical analysis, we computed the correlation between the expression levels in basal condition (DMSO) of RARA and the fold induction of the retroviral elements (fold changes) in the same samples. The scatterplot of such correlation analysis is represented in Figure 20.

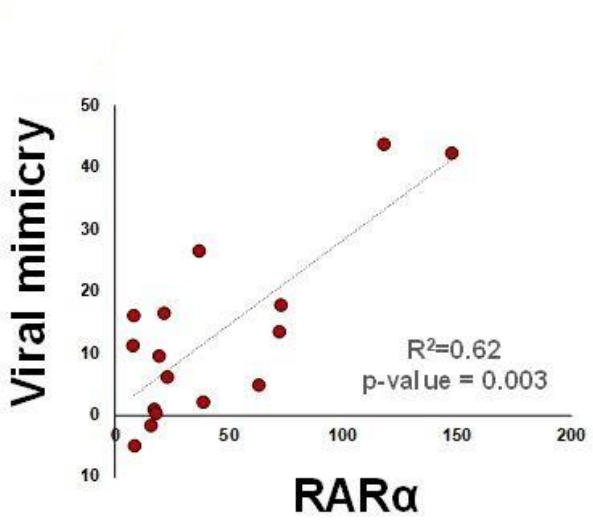


Figure 20. Correlation between induction of retroviral elements and basal expression of RARA.

The figure demonstrate that the ATRA-dependent process of viral mimicry is correlated (R^2=0.62) with the expression levels of RARA in our panel of 16 breast cancer cell lines.

7. PROTEIN-PROTEIN INTERACTION NETWORK

We imported from the STRING database a protein-protein interaction network matching the 754 genes that emerged from the various phases of the differential expression analysis.
The resulting network with 342 nodes is shown in [Figure 20]: any non-interacting gene or with less than two connections was excluded from the network. The size of each node is proportional to its variation coefficient; the colour of each node is proportional to its Pearson's product moment coefficient (R): nodes along the blue scale are negatively correlated with the ATRA-*score*, while nodes along the red scale are correlated positively.
Moreover, information from the ChIP-sequencing data analysis and the Methylation arrays are also included to constitute a multi-layer and multi-omics network. Nodes which represent genes identified as direct target of RARA are represented as triangles, while genes which appears to be direct targets of both RARA and RARG are represented as rhombus. The width of the border of each node is related to its methylation level: indeed, for those genes whose

methylation levels are strictly correlated with the ATRA-*score*, borders are drawn thicker; decreasing correlation is associated with decreasing thickness.

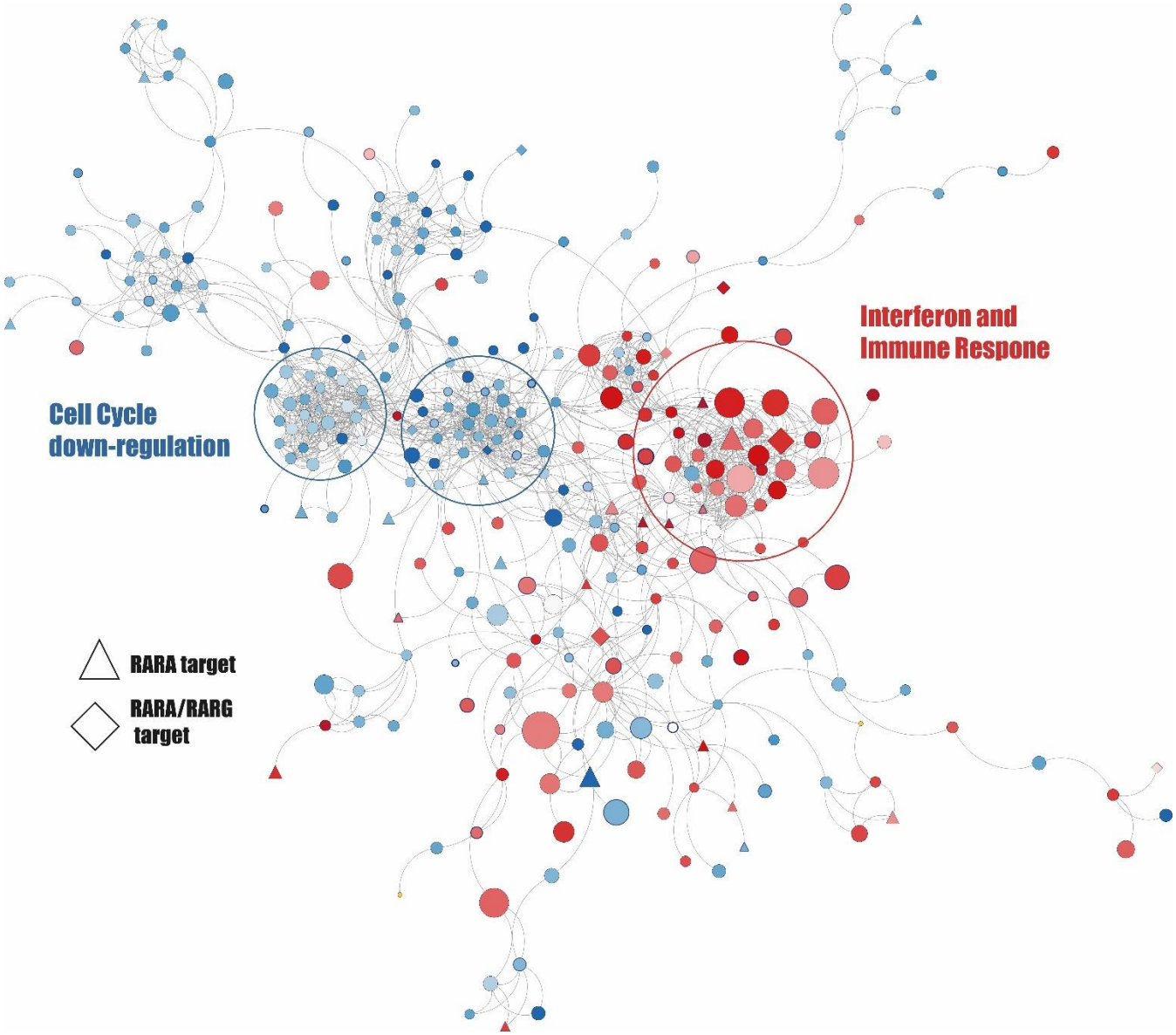


Figure 21. Protein – protein interaction network of genes perturbed by ATRA-treatment

Protein-protein interaction network, imported from STRING database, matching the top-ranked genes that emerged from the sequencing data analysis.

Discussion

The retinoids have been investigated extensively for the prevention and treatment of cancer, predominantly because of their ability to induce cellular differentiation and to arrest proliferation. Systemic retinoids are approved by the U.S. Food and Drug Administration (FDA) for the treatment of cutaneous T-cell lymphoma and acute promyelocytic leukemia (APL) [87]. The anti-leukemic action of ATRA is not primarily cytotoxic and it is the result of a direct anti-proliferative and cyto-differentiating action followed by a secondary apoptotic response rendering ATRA the first example of clinically useful cyto-differentiating agent. More than 85% of patients with APL achieve complete remission following treatment with ATRA in combination with chemotherapy [88]. The unique mechanism of action and the results obtained in APL has raised enthusiasm in generalizing the use of retinoids to other types of cancers.

Retinoids are of therapeutic interest in breast cancer, that is the subject of the present thesis work, because of its anti-tumor activity in *in vitro* cell lines models and *in vivo* mouse models. The majority of studies refer to ATRA since it is considering the prototype of the family and one of the most powerful retinoids available.

It is surprising to notice that the huge amount of information obtained at the pre-clinical level has translated into a very limited number of clinical trials. Indeed, 18 published clinical trials are available, seven of which report on data obtained with Fenretinide and only four refer to the use of ATRA. To date, retinoids are not recommended for the clinical practice on the basis of chemo-preventive clinical trials performed. Overall, ATRA trials showed no activity even if some benefits have been observed in combination with other anti-cancer drugs [89]. We surmise that these disappointing results are predominantly the consequence of the design of the clinical trials which consist of a small cohort of patients and did not take into account the intrinsic heterogeneity of breast cancer: they were conducted on cohorts of patients recruited without prior selection for any particular sub-type of breast tumor. There is a general agreement in that breast cancer is a large collection of different diseases and we don’t know yet, why there are specific subtypes of this tumor which seem to be particularly sensitive or refractory to the anti-tumor action of retinoids. Thus, it is important to define whether common or distinct molecular determinants and mechanisms are active in different

types of breast cancer, or eventually if it possible to identify a novel ATRA-sensitive subgroup.

By treating 16 breast cancer cell-lines characterized by different sensitivity to the retinoid, and by analyzing their associated transcriptomic perturbations, we collected evidence of a common mechanisms of action.

Of the identified 754 genes, whose perturbations by ATRA are quantitatively correlated with the *ATRA-score*, of particular interest are those being up-regulated by the retinoid (n=414), as they are more likely to be the main effectors of ATRA-mediated anti-tumour activity. Indeed, genes and pathways that are down-regulated specifically in ATRA-sensitive cell lines, are linked to cell proliferation and cell cycle progression which are tightly connected with the antiproliferative effect exerted by retinoic acid, and thus can be considered part of a downstream mechanism of action.

In contrast, the observed strong induction of the interferon signalling on treated cell lines may be directly involved in cell-cycle arrest and apoptosis.

Interferon itself has been proven to exert a strong anti-tumour and anti-proliferative effect in large amount of cancer types. Indeed, interferon is administered in some chemotherapeutics protocols [90], but its efficacy is limited to tumours that constitutively express interferon receptors. Thus, it is not effective in activating its own pathway in case there is a block at the receptor level. At this point, we investigated how ATRA could activate the interferon pathway in these cells, as we determined that retinoid-associated up-regulation of the interferon-dependent genes is not the consequence of an increase in the levels of any of the type I, type II or type III interferons, whose transcripts are undetectable in all cell lines regardless of ATRA exposure. Therefore, there must be a mechanism of activation independent of the levels of IFN; this is of a particular importance, as the system does not depend on the presence of the receptor to activate the pathways. In this scenario, we can speculate that ATRA can activate the interferon response in a receptor-independent fashion. To evaluate whether direct target of RARs (activated by ATRA) can induce the pathway themselves, we wanted to evaluate whether ATRA could increase the transcription of endogenous retroviruses, as interferon is mainly a mechanism of innate immune response to viral infection events. If this is to be the case, ATRA would be capable of inducing a phenomenon known as ‘Viral-mimicry’. At this moment, we still await experimental data to determine what is the main mechanism of interferon activation.

We need to discriminate if the pathway induction is due to RAR-mediated overexpression of DTX3L/PARP9, which would suggest a direct effect, or if it is mediated by increased expression of viral double-stranded RNAs, which would favour and indirect effect (Viral mimicry). To confirm these two alternative hypothesis, we are performing wet-lab experiments aimed at silencing DTX3L/PARP in breast cancer cell-lines or by artificially activating viral mimicry using polyriboinosinic:polyribocytidylic acid (poly(I:C)) .

Beside the anti-proliferative effect described above, our data suggest that the pharmacological treatment with ATRA might also have an immunoregulatory effect on these cells. In particular, it has been observed that there is a dramatic up-regulation of the "Antigen-presentation and assembly/loading of class I MHC" pathways, as well as "Inflammatory responses": this may result in an increased antigen presentation mechanism which may activate innate immune response. From an immunologic point of view, this is of a particular interest. Indeed, it has been observed that ATRA induces the interferon pathway selectively in those cell lines that have low levels of antigens presentation and therefore an inactivated interferon signalling. Hence, those tumours have a low immunogenicity. Immunologically quiet tumours usually progress without arousing attention of immune system. Therefore, a strong reactivation of the interferon signalling in these tumours and the consequent increased exposure of antigens, provides the rational to hypothesize that ATRA could favour recognition by the immune system.

Moreover, of great relevance is the correlation between endogenous retrovirus reactivation and the RARA basal expression in each cell line. According to this, RARA expression levels could be used as simple biomarker of induction of viral mimicry and associated increased antigen presentation. Following this idea, ideal target for treatment with retinoic acid would be tumours with high levels of RARA and low levels of basal interferon activity (immunologically quiet). This provides a strong rationale for the combination of ATRA with the immune checkpoint inhibitors.

Finally, we took into consideration the predictive model mentioned before (ATRA-21) [38]. Notably, it defined a low level in interferon activation (basal condition) to be a sensitivity marker of ATRA-treatment. This suggests that other tumours predicted to be sensitive to ATRA may show a similar behaviour, e.g. induction of interferon signalling, antigen presentation and endogenous retroviruses transcription.

This preliminary evidence led to the planification of experiments involving ATRA in gastric carcinoma where a good percentage of tumour (10%) has an amplification of the oncogene ERBB2 together with RARA [91].

Concluding remarks

ATRA and its derivatives have shown a potential for therapeutic and preventive use in breast cancer because of their ability to modulate cell growth and differentiation .

To date, a huge amount of studies, proving the antitumor activity of retinoids in *in vitro* and *in vivo* models of breast cancer, have translated into a very limited number of clinical trials, with disappointing results. There is therefore a need to go define the cellular and molecular determinants of retinoid sensitivity in breast cancer.

To this aim, human breast cancer cell lines have been used, as useful pre-clinical cancer models that reflect the heterogeneity of human cancers, thus representing useful tools to define the molecular determinants and the mechanisms, underlying the pathogenesis and the progression of the disease, and to evaluate the sensitivity to pharmacological treatments.

The results obtained in the present thesis project provide insights into the molecular mechanisms underlying the anti-tumour action of ATRA in breast cancer.

In the first part of the study, we analysed data from high throughput technologies (RNA-sequencing and ChIP-sequencing) and Methylation arrays, to go insight the gene expression profile inducted by the pharmacological treatment with retinoic acid.

First, RNA-sequencing data from a panel of 16 breast cancer cell lines treated with retinoic acid (1 μM, 24 hours) have been analysed to identify the transcriptomic perturbation inducted by the pharmacological treatment. Selected genes have been organized in networks based on protein – protein interactions, to more a precise visualization of possible interaction mechanisms induced. The sequencing data led to the identification of ATRA-dependent pathways and gene-networks with significance for the anti-tumour activity of the retinoid: "interferon-dependent" and immune modulatory pathways are found to be strictly up-regulated after treatment with ATRA. On the contrary, pathways associated with cell proliferation and cell cycle progression, are down-regulated, dealing with the idea that this effect is tightly connected with the antiproliferative effect of retinoic acid, and thus can be considered part of a downstream mechanism of action.

Then, we inspected ChIP-sequencing data from a public database of two forms of RARs transcription factors (RARA, RARG) in one breast cancer cell line treated with retinoic acid: we evaluated which of the more central genes in our response network were directly

perturbated by the binding in the regions of their promoter of the ATRA-activated transcription factors. As results, we obtained a list of genes that are part of the above-mentioned interferon signalling, which have been identified as directs targets for RARA or RARG transcription factors; however, some of the most crucial genes involved in such pathways cannot be included in the list.

Last, Methylation data available for a panel of almost 40 un-treated breast cancer cell lines have been investigate, to find out whether there is a correlation between the basal methylation levels of genes necessary to trigger the mechanism of response to retinoids, and the sensitivity of cell lines to ATRA. Again, a few genes involved in the interferon-related mechanism have found to have a correlation between their methylation levels and the activation of the response to retinoids.

In the second part of the study we took again into account the RNA-sequencing data to quantify possible transcription of repetitive elements from retroviral DNA, which are known to be widely distributed in the human genome: we hypostasized that they can be the cause of the above-mentioned interferon-driven immune response.

Our data support the idea that up-regulation of the interferon-dependent genes is mediated by the induction of non-coding RNAs transcribed from endogenous retroviral DNA. Indeed, the transcriptomic analysis conducted indicates that ATRA induces a potent up-regulation of the transcription of these endogenous retroviruses (retrotransposons) that mimics viral infection. This mechanism, known as viral mimicry, leads to a strong activation of cell-autonomous interferon response, which markedly results in increased transcription of interferon-related genes and MHC class-I components. Finally, the approach provides information as to potential new molecular targets for the design of rational therapeutic combinations based on ATRA for the treatment and secondary chemo- prevention of certain types of breast cancer. In fact, these last results are consistent with the idea that ATRA exerts a strong immune-modulatory action in breast cancer cells: all things considered, it represents proof of principle for the evaluation of combination between the retinoid and cancer immunotherapeutic in the treatment of ATRA-sensitive breast cancer subtypes.

Bibliography

1. Bray F, Ferlay J, Soerjomataram Iet al, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians.
2. AIOM AIRTUM. I numeri del cancro in Italia 2018.
3. Jemal A, Bray F, Center MM, et al. Global cancer statistics. CA Cancer J Clin 2011;61(2):69–90.
4. Polyak, K. (2007). Breast cancer: origins and evolution. The Journal of clinical investigation, 117(11), 3155-3163.
5. Visvader JE. Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. Genes dev. 2009;23(22):2563-77.
6. Malhotra GK, Zhao X, Band H et al. Histological, molecular and functional subtypes of breast cancers. Cancer Biol Ther. 2010(10):955-60.
7. Perez EA, Romond EH, Suman VJ et al. Four-year follow-up of trastuzumab plus adjuvant chemotherapy for operable human epidermal growth factor receptor 2-positive breast cancer: joint analysis of data from NCCTG N9831 and NSABP B-31. J Clin Oncol. 2011;29(25):3366-73
8. Urruticoechea A, Smith IE, Dowsett M Proliferation marker Ki-67 in early breast cancer. J Clin Oncol. 2005;23:7212–7220
9. Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., ... & Fluge, Ø. (2000). Molecular portraits of human breast tumours. nature, 406(6797), 747.
10. Kennecke H, Yerushalmi R, Woods R, et al. Metastatic behavior of breast cancer subtypes. J Clin Oncol. 2010;28(20):3271–7
11. Gianni L, Dafni U, Gelber RD, et al. Treatment with trastuzumab for 1 year after adjuvant chemotherapy in patients with HER2-positive early breast cancer: a 4-year follow-up of a randomised controlled trial. Lancet Oncol. 2011;12(3):236–44.
12. Kreike B, van Kouwenhove M, Horlings H, et al. Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. Breast Cancer Res. 2007;9(5):R65.
13. Eroles P, Bosch A, Pérez Fidalgo AJ et al. Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. Cancer Treat Rev. 2012;38:698–707.

14. Prat A, Parker JS, Karginova O, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12:R68
15. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol oncol. 2011;5: 5-23.
16. Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., ... & Quackenbush, J. F. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, *27*(8), 1160.
17. Rivenbark, A. G., O'Connor, S. M., & Coleman, W. B. (2013). Molecular and cellular heterogeneity in breast cancer: challenges for personalized medicine. The American journal of pathology, 183(4), 1113-1124.
18. Napoli JL. Retinoic acid biosynthesis and metabolism. FASEB J. 1996;10(9):993-1001
19. Altucci L, Gronemeyer H. The promise of retinoids to fight against cancer. Nat Rev Cancer. 2001;1(3):181-93.
20. Germain PP, Chambon P, Eichele G et al. International Union of Pharmacology. LX. Retinoic acid receptors. Pharmacol Rev. 2006;58(4):712-25
21. Centritto, F., et al., Cellular and molecular determinants of all-trans retinoic acid sensitivity in breast cancer: Luminal phenotype and RARalpha expression. EMBO Mol Med, 2015. **7**(7): p. 950-72.
22. Morgan NP, Gudas LJ. Diverse actions of retinoid receptors in cancer prevention and treatment. Differentiation. 2007;75(9):853-70
23. Minucci S, Pelicci PG. Retinoid receptors in health and disease: co-regulators and the chromatin connection. Semin Cell Dev Biol. 1999;10:215-25
24. Balmer JE, Blomhoff R. Gene expression regulation by retinoic acid. J Lipid Res. 2002;43(11):1773-808.
25. Dey N, De PK, Wang M, et al. CSK controls retinoic acid receptor (RAR) signaling: a RAR-c-SRC signaling axis is required for neuritogenic differentiation. Mol Cell Biol. 2007 Jun;27(11):4179-97.
26. Niederreither K, Dolle P Retinoic acid in development: towards an integrated view. Nat Rev Genet. 2008;9(7):541-53
27. Radaeva S. Alcohol, Retinoic Acid, and Cancer. In: Alcohol and Cancer. S. V. Zakhari, Vasilis; Guo, Q. Max (Eds.), Springer: 2011. 127-153

28. Soprano DR, Qin P, Soprano KJ. Retinoic acid receptors and cancers. Annu Rev Nutr. 2004;24:201-21.
29. Altucci L, Leibowitz MD, Ogilvie KM, et al. RAR and RXR modulation in cancer and metabolic disease. Nat Rev Drug Discov. 2007;6(10):793-810
30. Anzano MA, Byers SW, Smith JM, et al. Prevention of breast cancer in the rat with 9-cis-retinoic acid as a single agent and in combination with tamoxifen. Cancer Res. 1994;54(17):4614-7.
31. Sutton, L.M., et al., Pharmacokinetics and clinical impact of all-trans retinoic acid in metastatic breast cancer: a phase II trial. Cancer Chemother Pharmacol, 1997. **40**(4): p. 335-41.
32. Budd, G.T., et al., Phase I/II trial of all-trans retinoic acid and tamoxifen in patients with advanced breast cancer. Clin Cancer Res, 1998. 4(3): p. 635-42.
33. Toma, S., et al., Biological activity of all-trans-retinoic acid with and without tamoxifen and alpha-interferon 2a in breast cancer patients. Int J Oncol, 2000. 17(5): p. 991-1000.
34. Bryan, M., et al., A pilot phase II trial of all-trans retinoic acid (Vesanoid) and paclitaxel (Taxol) in patients with recurrent or metastatic breast cancer. Invest New Drugs, 2011. 29(6): p. 1482-7.
35. Chiesa, M.D., et al., Tamoxifen vs Tamoxifen plus 13-cis-retinoic acid vs Tamoxifen plus Interferon alpha-2a as first-line endocrine treatments in advanced breast cancer: updated results of a phase II, prospective, randomised multicentre trial. Acta Biomed, 2007. 78(3): p. 204-9.
36. Rochette-Egly C, Germain P. Dynamic and combinatorial control of gene expression by nuclear retinoic acid receptors (RARs). Nucl Recept Signal. 2009;7:e005.
37. Anzano MA, Byers SW, Smith JM, et al. Prevention of breast cancer in the rat with 9-cis-retinoic acid as a single agent and in combination with tamoxifen. Cancer Res. 1994;54(17):4614-7.
38. Bolis, M., Garattini, E., Paroni, G., Zanetti, A., Kurosaki, M., Castrignanò, T., ... & Terao, M. (2016). Network-guided modeling allows tumor-type independent prediction of sensitivity to all-trans-retinoic acid. Annals of Oncology, 28(3), 611-621.
39. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., ... & Barnes, I. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research, 22(9), 1760-1774.

40. Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I., & Oren, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome biology, 15(5), R69.
41. PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq dataZhang Y, Lameijer EW, 't Hoen PA, Ning Z, Slagboom PE, Ye KBioinformatics. 2012 Feb 15; 28(4):479-86.
42. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Au KF, Jiang H, Lin L, Xing Y, Wong WH Nucleic Acids Res. 2010 Aug; 38(14):4570-8.
43. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Wu TD, Nacu S Bioinformatics. 2010 Apr 1; 26(7):873-81.
44. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21.
45. Manber, U., & Myers, G. (1993). Suffix arrays: a new method for on-line string searches. siam Journal on Computing, 22(5), 935-948.
46. RCore Team, R: A language and environment for statistical computing. 2017.
47. RStudio Team, RStudio: Integrated Development Environment for R. 2016.
48. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 550.
49. Love, M., Anders, S., & Huber, W. (2014). Differential analysis of count data–the DESeq2 package. Genome Biol, 15(550), 10-1186.
50. Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology, 11(10), R106.
51. McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research, 40(10), 4288-4297.
52. Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent _ltering increases detection power for high-throughput experiments. PNAS, 107(21):9546{9551, 2010.
53. Benjamini, Hohberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing.

54. Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones geographicae, 30(2), 87-93.
55. Smyth, G. K. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and bioconductor. Edited by: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. 2005.
56. Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics, 23(8), 980-987.
57. Kim, S. Y., & Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. BMC bioinformatics, 6(1), 144.
58. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550.
59. Wu, D., & Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic acids research, 40(17), e133-e133.
60. Burrows, M., & Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm.
61. Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. In Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on (pp. 390-398). IEEE.Lam, T. W., Sung, W. K., Tam, S. L., Wong, C. K., & Yiu, S. M. (2008). Compressed indexing and local alignment of DNA. Bioinformatics, 24(6), 791-797.
62. Lippert, R. A., Mobarry, C. M., & Walenz, B. P. (2005). A space-efficient construction of the Burrows–Wheeler transform for genomic data. *Journal of Computational Biology*, *12*(7), 943-951.
63. Campagna, D., Albiero, A., Bilardi, A., Caniato, E., Forcato, C., Manavski, S., ... & Valle, G. (2009). PASS: a program to align short sequences. Bioinformatics, 25(7), 967-968.
64. Eaves, H. L., & Gao, Y. (2009). MOM: maximum oligonucleotide mapping. Bioinformatics, 25(7), 969-970.

65. Lin, H., Zhang, Z., Zhang, M. Q., Ma, B., & Li, M. (2008). ZOOM! Zillions of oligos mapped. Bioinformatics, 24(21), 2431-2437.
66. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.
67. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology, 9(9), R137.
68. Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science, 316(5830), 1497-1502.
69. Feng, J., Liu, T., & Zhang, Y. (2011). Using MACS to identify peaks from ChIP-Seq data. Current protocols in bioinformatics, 34(1), 2-14.
70. Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., ... & Lee, W. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 448(7153), 553.
71. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... & Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology, 5(10), R80
72. Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., & Green, M. R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC bioinformatics, 11(1), 237.
73. Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics, 30(10), 1363-1369.
74. Ritchie, M.E., et al., A comparison of background correction methods for two-colour microarrays. Bioinformatics, 2007. **23**(20): p. 2700-7.
75. Liu, J., & Siegmund, K. D. (2016). An evaluation of processing methods for HumanMethylation450 BeadChip data. BMC genomics, 17(1), 469.
76. Triche Jr, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., & Siegmund, K. D. (2013). Low-level processing of Illumina Infinium DNA methylation beadarrays. *Nucleic acids research*, *41*(7), e90-e90.

77. Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., & Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. Epigenomics, 3(6), 771-784.
78. Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., & Beck, S. (2012). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics, 29(2), 189-196.
79. Hansen, K.D., IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays, in Bioconductor. 2016.
80. Ishak, C. A., Classon, M., & De Carvalho, D. D. (2018). Deregulation of Retroelements as an Emerging Therapeutic Opportunity in Cancer. Trends in cancer.
81. Mager, D.L. and Stoye, J.P. (2015) Mammalian endogenous retroviruses. Microbiol. Spectr. 3, Mdna3-0009–2014.
82. Cañadas, I., Thummalapalli, R., Kim, J. W., Kitajima, S., Jenkins, R. W., Christensen, C. L., ... & Zhang, G. (2018). Tumor innate immunity primed by specific interferon-stimulated endogenous retroviruses. Nature medicine, 1.
83. Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923-93.
84. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504.
85. Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., ... & Jensen, L. J. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research.
86. Singh, A., & Settleman, J. E. M. T. (2010). EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. *Oncogene*, *29*(34), 4741.
87. Garattini E, Gianni M, Terao M. Cytodifferentiation by retinoids, a novel therapeutic option in oncology: rational combinations with other therapeutic agents. Vitam Horm 2007;75:301-54.

88. Tallman, M. S., Andersen, J. W., Schiffer, C. A., Appelbaum, F. R., Feusner, J. H., Ogden, A., ... & Wiernik, P. H. (1997). All-trans-retinoic acid in acute promyelocytic leukemia. *New England Journal of Medicine*, *337*(15), 1021-1028.
89. Connolly RM, Nguyen NK, Sukumar S. Molecular pathways: current role and future directions of the retinoic acid pathway in cancer prevention and treatment. Clin Cancer Res.2013;19(7):1651-9.
90. Yeo, W., Mok, T. S., Zee, B., Leung, T. W., Lai, P. B., Lau, W. Y., ... & Hui, P. (2005). A randomized phase III study of doxorubicin versus cisplatin/interferon α-2b/doxorubicin/fluorouracil (PIAF) combination chemotherapy for unresectable hepatocellular carcinoma. *Journal of the National Cancer Institute*, *97*(20), 1532-1538.
91. Xiang, Z., Huang, X., Wang, J., Zhang, J., Ji, J., Yan, R., ... & Yu, Y. (2018). Cross-database analysis reveals sensitive biomarkers for combined therapy for ERBB2+ gastric cancer. *Frontiers in pharmacology*, *9*.

Appendices

Appendix 1

Differentially expressed genes: variation coefficient and correlation of their induction with the ATRA-*score.*

SYMBOL	VC	R	p-value R	RHO	p-value RHO	TFs	Is methylated	Meth. Counts
H6PD	60.94175347	0.839774	4.71E-05	0.716817	0.001779434		No	
DHX58	72.49401231	0.836593	5.35E-05	0.793514	2.44E-04		No	
TRIM5	86.75056845	0.820271	9.95E-05	0.781714	3.48E-04		No	
REL	56.72582029	0.814124	1.24E-04	0.764015	5.70E-04	RARA	No	
STAT6	71.41966133	0.807511	1.55E-04	0.722717	0.001562577	RARA	Yes	1
TRIM21	65.25023602	0.796459	2.22E-04	0.764015	5.70E-04		No	
IFIT1	147.1123157	0.789463	2.76E-04	0.746316	8.98E-04		No	
PHF21A	58.99307046	0.788204	2.87E-04	0.643071	0.007205587		Yes	3
VAMP4	60.25004003	0.786491	3.02E-04	0.778764	3.79E-04	RARA	No	
SP100	74.89353251	0.783581	3.29E-04	0.710918	0.002020147		Yes	1
CHMP1B	51.91950675	0.779332	3.73E-04	0.619472	0.010491653		No	
FBXL20	157.2738491	0.775223	4.19E-04	0.722717	0.001562577		No	
CEBPD	116.3071425	0.77476	4.24E-04	0.584073	0.01751624		Yes	4
IFIT3	123.9657048	0.774538	4.27E-04	0.734516	0.001193198		No	
IFI35	73.97797279	0.770773	4.74E-04	0.823012	9.01E-05		Yes	1
PARP14	111.9287226	0.763872	5.72E-04	0.687319	0.003261473		No	
PARP9	92.28707989	0.762018	6.01E-04	0.766965	5.27E-04	RARA/RARG	Yes	1
TRIM56	56.39516038	0.756677	6.92E-04	0.707968	0.002150048	RARA	Yes	1
RNF217	95.38647039	0.756419	6.96E-04	0.563424	0.023040837		No	
GBP2	126.0116956	0.755296	7.17E-04	0.536875	0.032008547		No	
FOXO3	64.2687726	0.755151	7.19E-04	0.675519	0.004079798		Yes	1
ITGAV	73.48032	0.753089	7.58E-04	0.775815	4.12E-04		No	
OAS2	224.5471714	0.749776	0.001288	0.738356	0.001669669		No	
IFI44	196.0620033	0.732537	0.00125	0.766965	5.27E-04		No	
TMEM173	115.3133257	0.73253	0.00125	0.746316	8.98E-04		Yes	2
MX1	197.4743712	0.731221	0.001288	0.814163	1.24E-04	RARA/RARG	No	
DECR1	79.98076475	0.727365	0.001407	0.607672	0.012526367	RARA	No	
UBE2L6	97.80328957	0.725762	0.001459	0.710918	0.002020147		Yes	1
CHMP2B	54.80507037	0.720656	0.001636	0.601773	0.013653454	RARA	No	
ARHGAP31	146.4794684	0.720573	0.002442	0.630828	0.011683494		No	
PJA1	54.29377928	0.719862	0.001665	0.271388	0.30927721		No	
WSB1	55.18716376	0.716844	0.001778	0.607672	0.012526367		No	
SAT2	71.95157473	0.716271	0.001801	0.545725	0.028767026		No	
CYLD	63.61743602	0.715297	0.001839	0.66667	0.004795669		No	

ARFGAP3	52.40818634	0.712676	0.001946	0.551625	0.026749489		No	
PPP2R5A	55.39272816	0.711565	0.001992	0.702068	0.002430184	RARA	No	
GGT7	66.59309427	0.708348	0.002133	0.483778	0.057610502		No	
DDX60	121.5133868	0.708235	0.002138	0.66667	0.004795669		Yes	1
UBA7	155.0033355	0.705763	0.002251	0.802363	1.84E-04		No	
LRP10	52.6080542	0.705239	0.002276	0.495577	0.050929762	RARA	No	
STAT3	56.54389923	0.703652	0.002352	0.707968	0.002150048		No	
CTSS	192.1318228	0.703219	0.002373	0.769915	4.86E-04		Yes	1
MR1	55.10949866	0.701691	0.002449	0.634221	0.008324656		No	
TLR3	136.1325925	0.699501	0.003701	0.645165	0.009397773		Yes	10
CPD	108.4572875	0.696527	0.00272	0.678469	0.003861255		No	
SORT1	58.05819441	0.696262	0.002734	0.469029	0.066843812		No	
TAF13	74.33701323	0.693362	0.002897	0.64897	0.006528495		No	
USP9X	61.08572028	0.691436	0.003009	0.749266	8.35E-04		No	
ETS1	126.29724	0.688481	0.006478	0.536429	0.047976914		Yes	3
ACTR2	50.15503768	0.687795	0.003231	0.598823	0.014245964		No	
DPYD	185.9830248	0.686134	0.003337	0.64897	0.006528495		No	
NFKB2	71.56651194	0.681366	0.003656	0.581123	0.018235058		No	
CAPZA1	59.76247417	0.681087	0.003675	0.637171	0.007937331		No	
PIF1	72.61642493	0.680535	0.003714	0.43658	0.090887581		No	
TAPBP	56.63414783	0.67961	0.003779	0.65192	0.006209506		No	
RNF114	76.06746664	0.679093	0.003816	0.557525	0.024842079		Yes	1
FAM149B1	50.63626371	0.677004	0.003969	0.542775	0.029818337		No	
BCL10	89.00244794	0.674879	0.004128	0.404132	0.120544092		No	
PIK3CD	117.2191648	0.673986	0.004197	0.575224	0.019741194		Yes	2
PIK3R1	130.8116123	0.672853	0.004286	0.528026	0.035520369	RARA/RARG	No	
PSMB9	117.1985144	0.672569	0.004308	0.545725	0.028767026		No	
SEC24D	70.4744044	0.669254	0.004577	0.598823	0.014245964		Yes	4
MAGED1	54.79883695	0.669149	0.004586	0.43953	0.08847836		No	
PDGFC	117.2401562	0.668497	0.00464	0.66667	0.004795669		No	
GABARAPL1	101.5481131	0.668493	0.004641	0.672569	0.004308166	RARA	No	
RNASEL	73.34261867	0.667969	0.004685	0.616522	0.01097398		No	
GBP3	107.2333649	0.667694	0.004708	0.681419	0.003652225		No	
GAA	63.96525771	0.666497	0.004811	0.572274	0.020529464		No	
PTK2B	93.64966155	0.664909	0.004949	0.230089	0.391284989		No	
BCL6	105.8418162	0.660281	0.005372	0.477878	0.06118325		Yes	3
CTSB	142.902409	0.65925	0.00547	0.737466	0.001113013		Yes	1
PRKCE	70.51891553	0.659161	0.005479	0.539825	0.030898677		No	
NMI	69.40897344	0.658342	0.005558	0.713867	0.00189669		No	
KLHL5	94.65731052	0.658134	0.005578	0.563424	0.023040837		No	
CAMK2D	64.43584067	0.65466	0.005924	0.598823	0.014245964		No	
GPX8	117.5636691	0.654019	0.00599	0.560474	0.023928433		No	
SAMD9L	195.2084606	0.652703	0.006127	0.646021	0.006860399		No	
LRP1	219.0958765	0.65121	0.006285	0.256638	0.33731508		No	
TRIM8	51.7168892	0.648072	0.006628	0.474928	0.063029429		No	

ALG10B	78.06886449	0.645856	0.006879	0.566374	0.022178802		Yes	1
STAP2	79.31546166	0.64452	0.007034	0.637171	0.007937331		No	
TRIM38	79.16832978	0.643944	0.007102	0.769915	4.86E-04		No	
GBP1	135.3183917	0.643715	0.007129	0.643071	0.007205587		No	
BIRC3	210.035201	0.642886	0.007228	0.530976	0.034318889		Yes	1
OAS1	173.6373292	0.640288	0.007544	0.728617	0.001367745	RARA	No	
DDX60L	101.216574	0.639412	0.007653	0.755166	7.19E-04		Yes	1
PLD3	53.2249134	0.63865	0.007749	0.362833	0.167211533		Yes	3
UBQLN2	71.41797839	0.635852	0.008109	0.507377	0.044842307		No	
TBC1D1	61.09039106	0.635286	0.008183	0.707968	0.002150048	RARA	Yes	2
TM7SF2	98.78930363	0.634318	0.008312	0.563424	0.023040837		No	
PARVA	85.52423114	0.633106	0.008475	0.410031	0.11470988		Yes	3
HLA-E	54.4122551	0.631957	0.008632	0.327435	0.215724174	RARA	Yes	3
SH3GLB1	70.57890845	0.630661	0.008812	0.44838	0.081525044		No	
GCC2	52.6950163	0.630422	0.008845	0.589973	0.016144758	RARA	No	
IFIT2	141.1213865	0.629397	0.00899	0.530976	0.034318889		No	
HLA-A	154.390396	0.629285	0.009006	0.581123	0.018235058		No	
NT5C2	131.1761346	0.625686	0.00953	0.66372	0.005055461		No	
MAT2B	52.15417155	0.624545	0.009701	0.637171	0.007937331		No	
KLC4	55.36984564	0.623356	0.009882	0.687319	0.003261473		Yes	1
SCCPDH	141.0380828	0.622807	0.009967	0.584073	0.01751624		No	
AKT3	130.2088722	0.620035	0.010402	0.672569	0.004308166		Yes	2
PTPN6	89.97881731	0.615842	0.011088	0.634221	0.008324656		No	
EGF	144.0878064	0.613537	0.01148	0.690268	0.003079151		No	
IRF1	95.03340854	0.610545	0.012005	0.65192	0.006209506		No	
FN1	289.5496453	0.609023	0.012279	0.65487	0.005903065		No	
QPRT	89.49323555	0.60721	0.016366	0.713137	0.002838928		No	
CD47	64.21211875	0.604526	0.013118	0.799414	2.02E-04		Yes	1
PSMB10	81.83847743	0.594812	0.015084	0.746316	8.98E-04	RARA	No	
DTX3L	70.21533034	0.590457	0.016036	0.684369	0.003452398	RARA/RARG	No	
TBC1D8B	75.74191337	0.581471	0.018149	0.634221	0.008324656	RARA	No	
IFI44L	235.3077592	0.573849	0.020106	0.684369	0.003452398		No	
RAB20	90.10132244	0.550672	0.027068	0.631271	0.008726804		Yes	1
OASL	210.3639243	0.537854	0.031637	0.66372	0.005055461		No	
DCP1B	73.72295952	0.513794	0.041769	0.646021	0.006860399		Yes	1
PARP12	84.66898049	0.508029	0.044523	0.678469	0.003861255		No	
AGPAT4	100.0523765	0.488098	0.055094	0.66667	0.004795669		No	
PSMB8	76.32974227	0.46619	0.068738	0.637171	0.007937331		Yes	2
CSAD	60.67941634	0.466158	0.06876	0.628321	0.009144172	RARA/RARG	No	
SYNJ2	80.64393729	0.308142	0.245596	0.65487	0.005903065		Yes	7
B2M	96.23471692	0.235222	0.380506	0.746316	8.98E-04		No	
PPP2R2C	133.3280625	-0.29096	0.29276	-0.67739	0.005529364		No	
HIST1H4J	54.39620506	-0.41948	0.105783	-0.74927	8.35E-04		No	
RBBP7	55.31424891	-0.45716	0.075026	-0.65782	0.005608813		No	
HIST1H2BJ	78.67151307	-0.48651	0.056011	-0.62832	0.009144172		No	

HIST1H2AJ	67.34486596	-0.50361	0.046724	-0.66667	0.004795669		No	
HIST1H2AL	69.57856943	-0.50586	0.045593	-0.71092	0.002020147		No	
HIST1H1D	64.24294239	-0.54591	0.028703	-0.64602	0.006860399		No	
SNF8	69.21772618	-0.55478	0.025715	-0.62537	0.009577164		No	
SDC3	93.08924551	-0.55698	0.025014	-0.63422	0.008324656		No	
HIST1H3B	57.75595754	-0.56144	0.023635	-0.74042	0.001037287		No	
HIST1H1E	54.02657795	-0.56681	0.022053	-0.65487	0.005903065		No	
RAD54L	53.79388482	-0.57295	0.020346	-0.67257	0.004308166		No	
ORC1	59.99050377	-0.57595	0.01955	-0.70797	0.002150048		No	
CACNG4	148.5394804	-0.57859	0.023841	-0.64701	0.009131655		No	
HIST1H3F	68.29939339	-0.58139	0.018169	-0.72272	0.001562577		No	
LRR1	60.19079718	-0.58268	0.017852	-0.63127	0.008726804		No	
HIST1H4D	77.22111264	-0.58593	0.017076	-0.80826	1.51E-04		No	
HIST1H2BO	64.46132892	-0.58809	0.016574	-0.72567	0.001462519		No	
HIST1H3J	86.24685507	-0.59223	0.015643	-0.75812	6.66E-04		No	
HIST1H2AG	53.1838641	-0.59318	0.015435	-0.72272	0.001562577		No	
HIST2H3D	81.49684387	-0.59689	0.014645	-0.66077	0.005326394		No	
HIST1H1B	66.99400065	-0.60012	0.013983	-0.74632	8.98E-04		No	
HIST2H2AA4	52.5297785	-0.60427	0.013168	-0.62537	0.009577164		No	
HPRT1	53.17952859	-0.60512	0.013005	-0.64012	0.007564437		No	
PKMYT1	62.26426496	-0.61018	0.01207	-0.73157	0.001278041		Yes	1
ACD	52.02289113	-0.61256	0.01165	-0.78466	3.19E-04	RARA	No	
E2F8	54.35523778	-0.61453	0.01131	-0.69027	0.003079151		No	
HIST1H2AI	50.02718898	-0.61824	0.010691	-0.73157	0.001278041	RARA	No	
HIST1H2BH	60.10763616	-0.62005	0.010399	-0.70502	0.002286641		No	
PGAM1	66.05234413	-0.62152	0.010166	-0.68142	0.003652225		No	
ORC6	64.3450466	-0.62274	0.009978	-0.70502	0.002286641		No	
ZNRD1	61.57365314	-0.62343	0.00987	-0.71682	0.001779434		No	
ATP5I	51.89121045	-0.62406	0.009774	-0.38643	0.139275348		Yes	1
RMI2	53.56802979	-0.62496	0.00964	-0.67257	0.004308166		Yes	1
GNL2	55.71712393	-0.62594	0.009493	-0.45723	0.074973989		No	
ESRP2	81.83626608	-0.62668	0.009383	-0.70502	0.002286641		No	
FKBP9	98.88592146	-0.62694	0.009346	-0.36578	0.163530133		No	
GAPDH	57.28685646	-0.62735	0.009285	-0.54573	0.028767026		No	
PEMT	57.99112679	-0.62922	0.009016	-0.48378	0.057610502		Yes	4
GIT1	69.52545101	-0.62953	0.008972	-0.47493	0.063029429		No	
DUT	50.13344149	-0.63206	0.008618	-0.67257	0.004308166		No	
BLM	54.88528427	-0.63769	0.007871	-0.60767	0.012526367		Yes	1
LYAR	84.60244905	-0.63786	0.007848	-0.46313	0.070824076		No	
UBFD1	59.65601893	-0.63811	0.007817	-0.69617	0.002739165		No	
HIST1H3A	81.9952244	-0.64022	0.007552	-0.73452	0.001193198		No	
GPHN	56.05347973	-0.64148	0.007397	-0.51623	0.040646801		No	
POP1	54.19748575	-0.64306	0.007207	-0.33628	0.202837527		No	
MRPL41	54.06892486	-0.64311	0.007201	-0.43953	0.08847836	RARA/RARG	No	
KNTC1	55.22091445	-0.64321	0.007189	-0.62537	0.009577164		Yes	1

PSMD8	77.22398073	-0.64486	0.006994	-0.57227	0.020529464		No	
FANCB	52.71941906	-0.64563	0.006905	-0.64897	0.006528495		No	
CDC7	60.8547125	-0.64939	0.006482	-0.54573	0.028767026		Yes	3
PRPF3	58.48974004	-0.64995	0.006421	-0.66077	0.005326394		No	
KCTD7	53.7977105	-0.65277	0.00612	-0.77286	4.48E-04		Yes	1
FGFR4	162.9031651	-0.65363	0.00603	-0.66372	0.005055461		Yes	1
EXO1	54.87667589	-0.65448	0.005943	-0.71387	0.00189669		No	
MDN1	56.83201351	-0.65481	0.00591	-0.28319	0.287866173		No	
SRPK1	53.0849639	-0.65499	0.005891	-0.56047	0.023928433		No	
NDUFAF4	59.07661539	-0.65626	0.005763	-0.31564	0.233702896		No	
MRPS12	67.878305	-0.65715	0.005675	-0.35693	0.174739123	RARA	No	
HIST1H4C	52.85708628	-0.6574	0.00565	-0.75222	7.75E-04		No	
AGPAT5	62.67157205	-0.6581	0.005582	-0.46018	0.072877593		No	
KIF1C	51.42387272	-0.65828	0.005564	-0.65192	0.006209506		No	
NIP7	55.99108558	-0.65936	0.00546	-0.51033	0.043409397		No	
DGKE	77.58314549	-0.65939	0.005457	-0.51918	0.039316106		No	
ERLIN1	59.68311761	-0.65941	0.005455	-0.63422	0.008324656		No	
CHMP7	54.97746056	-0.66095	0.00531	-0.59587	0.014858383		Yes	1
ARF6	59.8108576	-0.66144	0.005264	-0.71682	0.001779434		No	
LBR	53.75215116	-0.66176	0.005235	-0.73747	0.001113013		Yes	1
GPC1	90.57087609	-0.66178	0.005232	-0.48378	0.057610502		Yes	1
PODXL2	84.33865043	-0.66194	0.005218	-0.40118	0.12353713		No	
NOP16	65.87474774	-0.66249	0.005167	-0.39528	0.129676812		No	
RNASEH1	62.0226095	-0.66268	0.005149	-0.22714	0.397553714	RARA	No	
CHEK1	52.79467489	-0.6637	0.005057	-0.67552	0.004079798		No	
GNA12	56.21548408	-0.66456	0.00498	-0.69322	0.002905144		Yes	9
HIST1H2BL	66.79241717	-0.66552	0.004896	-0.71977	0.00166814	RARA	No	
DCLRE1B	68.12739941	-0.66673	0.00479	-0.64012	0.007564437	RARA	No	
AP1M1	71.93679501	-0.66822	0.004663	-0.71977	0.00166814		No	
GLO1	63.55249546	-0.66887	0.004609	-0.31564	0.233702896		No	
ARHGAP39	92.85192612	-0.67027	0.004493	-0.50738	0.044842307		No	
HIST1H4B	61.88344854	-0.67083	0.004447	-0.79056	2.67E-04		No	
NFRKB	62.65696931	-0.67231	0.004329	-0.58112	0.018235058		Yes	1
ATP5O	52.63184997	-0.67266	0.004301	-0.46313	0.070824076		No	
VMA21	85.78324868	-0.67334	0.004248	-0.35988	0.170947788		No	
RAD51C	59.83916819	-0.67365	0.004223	-0.61062	0.011990908	RARA/RARG	No	
CASP2	56.04472864	-0.67428	0.004175	-0.61947	0.010491653		No	
TPI1	66.4116892	-0.67632	0.004019	-0.63127	0.008726804		No	
PCBP4	77.68308694	-0.67635	0.004017	-0.54867	0.027744243	RARA	No	
PSMA5	59.37822583	-0.67708	0.003963	-0.34808	0.186445717		Yes	1
RRM2	61.47607086	-0.67816	0.003884	-0.75517	7.19E-04		No	
GINS2	66.77518297	-0.6782	0.003881	-0.73157	0.001278041		Yes	10
RPF2	55.27304801	-0.67838	0.003868	-0.31859	0.229122599	RARA	No	
RSL1D1	50.17783153	-0.67868	0.003846	-0.29204	0.272406977		Yes	1
DNAJC8	55.42726331	-0.68048	0.003718	-0.64897	0.006528495		No	

PCGF6	60.62322647	-0.68054	0.003714	-0.49853	0.049353523		No	
VIM	201.7136611	-0.6806	0.003709	-0.54573	0.028767026		Yes	2
RFC2	57.16571366	-0.68248	0.003579	-0.70207	0.002430184		Yes	4
E2F3	53.14583486	-0.68289	0.003551	-0.46608	0.068812972	RARA	No	
MCM5	51.95924846	-0.68376	0.003493	-0.71682	0.001779434		No	
INF2	54.93943449	-0.6852	0.003398	-0.55457	0.025782266	RARA	Yes	1
RPP40	53.53013937	-0.68533	0.00339	-0.41003	0.11470988		Yes	1
MAK16	54.85352188	-0.68599	0.003346	-0.60472	0.013080403		No	
C1QBP	57.99478034	-0.69189	0.002983	-0.46608	0.068812972		No	
IMPDH2	61.66499149	-0.6926	0.002941	-0.63717	0.007937331		No	
CLSPN	75.90371259	-0.69891	0.002592	-0.70207	0.002430184		No	
SCML2	74.86253463	-0.69979	0.002546	-0.67847	0.003861255	RARA	No	
PDGFA	114.132698	-0.69991	0.00254	-0.50443	0.046310296		No	
HIST1H2AK	59.11138147	-0.70054	0.002507	-0.72567	0.001462519		No	
UTP18	52.08766381	-0.70102	0.002483	-0.43658	0.090887581		No	
WDR3	68.4642988	-0.70194	0.002436	-0.37758	0.149348246		Yes	1
NDUFS3	61.65275593	-0.70201	0.002433	-0.69027	0.003079151		No	
TRIM3	96.83055657	-0.70319	0.002375	-0.73157	0.001278041		No	
IRAK1	86.66139978	-0.70405	0.002333	-0.58407	0.01751624		No	
CCDC86	60.48926259	-0.70417	0.002327	-0.51033	0.043409397		No	
SOCS7	64.80194522	-0.70752	0.00217	-0.56637	0.022178802		No	
LPCAT1	62.18059753	-0.70775	0.00216	-0.63422	0.008324656		Yes	3
AMD1	57.52313675	-0.70795	0.002151	-0.25074	0.348922716		No	
POLA1	52.8900045	-0.70815	0.002142	-0.55752	0.024842079		No	
MRPL27	70.60405779	-0.70836	0.002132	-0.59587	0.014858383		No	
WDR77	72.10688748	-0.71139	0.002	-0.52508	0.036753395		No	
U2AF2	56.52985192	-0.71146	0.001997	-0.73157	0.001278041		No	
HIST1H3H	57.99186648	-0.7122	0.001966	-0.84071	4.53E-05		No	
AP2S1	50.96588179	-0.71235	0.001959	-0.74042	0.001037287		No	
LPAR2	67.81001834	-0.71409	0.001888	-0.63127	0.008726804		No	
DNAJC9	55.5112475	-0.7154	0.001835	-0.70502	0.002286641		No	
MSH2	60.24145858	-0.71584	0.001818	-0.55457	0.025782266		No	
PAFAH1B3	55.15414636	-0.7176	0.001749	-0.69617	0.002739165		Yes	2
POLR3K	50.08094597	-0.71824	0.001725	-0.48083	0.059377103		No	
LDHA	61.57555965	-0.71915	0.001691	-0.51918	0.039316106		No	
MAGOHB	67.1916996	-0.71946	0.001679	-0.61062	0.011990908		Yes	2
TMEM199	57.62444475	-0.71975	0.001669	-0.53393	0.033148449		No	
SEH1L	62.15944479	-0.72077	0.001632	-0.62537	0.009577164	RARA/RARG	No	
HIST1H2BI	89.03273127	-0.72298	0.001553	-0.71682	0.001779434		No	
HIST1H2AH	57.29095843	-0.7237	0.001529	-0.72862	0.001367745		No	
SNRPF	63.53337178	-0.72385	0.001524	-0.51328	0.042011063		No	
NOLC1	56.08191106	-0.72414	0.001514	-0.46903	0.066843812		No	
UBA2	68.06670727	-0.72623	0.001444	-0.56047	0.023928433		No	
PHB	109.3898129	-0.72626	0.001443	-0.59587	0.014858383		No	
POLR2I	61.88623484	-0.72659	0.001432	-0.50148	0.047813868		No	

KLHL11	51.17025469	-0.72796	0.001388	-0.57227	0.020529464		No	
EFHD2	81.95876872	-0.72796	0.001388	-0.72272	0.001562577		No	
RNGTT	54.51578383	-0.72928	0.001347	-0.38348	0.142579872		Yes	1
PVR	69.80292995	-0.7304	0.001313	-0.61947	0.010491653		No	
GEMIN5	53.08909086	-0.73088	0.001299	-0.56047	0.023928433		No	
ENO2	133.4603408	-0.73361	0.001219	-0.75517	7.19E-04		No	
B3GALT6	52.10357821	-0.73431	0.001199	-0.58112	0.018235058		No	
MRPS34	51.77613396	-0.73442	0.001196	-0.60767	0.012526367		No	
LAS1L	54.3905378	-0.73646	0.00114	-0.53983	0.030898677		No	
ADRM1	62.90570232	-0.73677	0.001131	-0.81711	1.11E-04		Yes	1
MSH6	50.84270801	-0.73684	0.00113	-0.71682	0.001779434		No	
CYC1	58.14144396	-0.73711	0.001123	-0.62242	0.010026187		No	
STX2	94.52072469	-0.74161	0.001008	-0.67257	0.004308166		Yes	1
GGCT	84.12594735	-0.74279	9.80E-04	-0.55162	0.026749489		No	
NOP2	61.41021053	-0.74299	9.75E-04	-0.51918	0.039316106		No	
PHB2	71.25572153	-0.7444	9.42E-04	-0.55457	0.025782266		No	
PRKDC	51.34901379	-0.7458	9.10E-04	-0.58407	0.01751624		No	
LSM2	52.22170135	-0.74636	8.97E-04	-0.55752	0.024842079		No	
YWHAH	51.41918998	-0.74882	8.44E-04	-0.79646	2.22E-04		No	
PTCD3	53.51276487	-0.75022	8.15E-04	-0.57817	0.018976551		No	
ALG8	51.95061995	-0.7506	8.07E-04	-0.66077	0.005326394		No	
POLE4	67.45276516	-0.75421	7.37E-04	-0.65487	0.005903065		No	
MRPS7	52.99987224	-0.75797	6.69E-04	-0.66077	0.005326394		No	
CDC6	70.70200926	-0.7587	6.56E-04	-0.74632	8.98E-04		No	
CMSS1	56.89945553	-0.75902	6.51E-04	-0.46313	0.070824076		Yes	1
EIF3K	96.85919636	-0.75954	6.42E-04	-0.71682	0.001779434		No	
HAGHL	127.668463	-0.76182	6.05E-04	-0.59882	0.014245964		Yes	1
SKI	59.78896263	-0.76221	5.98E-04	-0.45723	0.074973989		Yes	8
UBE2N	51.12128222	-0.76297	5.86E-04	-0.61357	0.01147359		No	
FBL	110.4442636	-0.76583	5.43E-04	-0.77286	4.48E-04		No	
FEN1	76.46373416	-0.76734	5.21E-04	-0.76696	5.27E-04		No	
LSM6	54.36525066	-0.76997	4.85E-04	-0.62832	0.009144172		No	
MCM7	57.08722298	-0.77031	4.80E-04	-0.76402	5.70E-04		No	
SRM	61.79732046	-0.7704	4.79E-04	-0.69027	0.003079151		Yes	1
CBX2	82.78459116	-0.77605	4.09E-04	-0.78761	2.92E-04		No	
MCM10	55.62910555	-0.77688	4.00E-04	-0.77286	4.48E-04		Yes	2
TIMM50	52.21353022	-0.77892	3.77E-04	-0.66667	0.004795669	RARA	No	
HIST2H4A	50.01936937	-0.77908	3.75E-04	-0.76107	6.17E-04		No	
TSEN54	53.60470376	-0.77962	3.69E-04	-0.61357	0.01147359		No	
PAICS	56.0976597	-0.78095	3.56E-04	-0.59292	0.015491163		No	
RPL26L1	62.05832438	-0.78148	3.50E-04	-0.60767	0.012526367		No	
VDAC3	54.54245403	-0.78686	2.99E-04	-0.64602	0.006860399		Yes	1
GEMIN4	51.05957881	-0.78765	2.92E-04	-0.60767	0.012526367		No	
SUPT16H	53.72549155	-0.78895	2.80E-04	-0.62537	0.009577164		No	
RPS6KA4	57.55477795	-0.79124	2.61E-04	-0.76107	6.17E-04		No	

UQCRH	56.61520213	-0.79303	2.47E-04	-0.69322	0.002905144		No	
POLE2	85.96918987	-0.79689	2.19E-04	-0.56047	0.023928433		No	
SNRPD1	56.62664824	-0.79931	2.03E-04	-0.68437	0.003452398		No	
NME1	83.40783663	-0.80367	1.76E-04	-0.69912	0.002580936		No	
ADCY7	103.3032281	-0.80456	1.71E-04	-0.78171	3.48E-04		No	
DDX20	67.44243655	-0.80534	1.67E-04	-0.62832	0.009144172		No	
TRIP13	52.83887149	-0.80848	1.50E-04	-0.78761	2.92E-04		Yes	1
NLE1	51.46785694	-0.80911	1.47E-04	-0.65782	0.005608813		No	
POLR1E	56.9283612	-0.81211	1.33E-04	-0.78761	2.92E-04		No	
BOP1	60.90441092	-0.81249	1.31E-04	-0.65192	0.006209506		No	
NUP155	64.84569835	-0.81435	1.23E-04	-0.75222	7.75E-04		No	
TOP3A	52.2503718	-0.81492	1.20E-04	-0.76107	6.17E-04		No	
GPD1L	71.29646826	-0.81564	1.17E-04	-0.64897	0.006528495	RARA/RARG	No	
CCNE1	116.5425842	-0.81608	1.16E-04	-0.75812	6.66E-04		No	
LRP8	85.75922348	-0.81915	1.04E-04	-0.77876	3.79E-04		No	
PCNA	50.8483213	-0.81929	1.03E-04	-0.75517	7.19E-04	RARA/RARG	No	
AEN	60.09964241	-0.82092	9.72E-05	-0.60472	0.013080403		No	
PFAS	63.20774704	-0.8236	8.82E-05	-0.56047	0.023928433		No	
UNG	50.71275291	-0.82392	8.71E-05	-0.74927	8.35E-04		No	
VAV2	58.65829591	-0.82623	8.00E-05	-0.76696	5.27E-04		Yes	1
GAS6	145.1435784	-0.83223	1.19E-04	-0.76345	9.26E-04	RARA	No	
FAM57A	76.49933212	-0.83957	4.75E-05	-0.73157	0.001278041		No	
SLBP	54.92094584	-0.84416	3.92E-05	-0.71387	0.00189669		No	
CDC25A	54.58656107	-0.84461	3.85E-05	-0.82596	8.08E-05		No	
SNRNP25	56.34875997	-0.84595	3.64E-05	-0.79941	2.02E-04		Yes	1
PLXND1	125.3312603	-0.84925	3.15E-05	-0.71682	0.001779434		No	
CTPS1	50.80919704	-0.86021	1.92E-05	-0.79646	2.22E-04		Yes	1
CCT5	53.58234413	-0.8644	1.57E-05	-0.71092	0.002020147		No	
RAC3	65.81446059	-0.8667	1.40E-05	-0.77876	3.79E-04		No	
F12	105.5270626	-0.87903	7.34E-06	-0.90266	1.71E-06		No	
COPS3	61.53124145	-0.88163	6.35E-06	-0.77581	4.12E-04		No	
ALYREF	56.39342391	-0.89449	2.94E-06	-0.85546	2.39E-05		No	
TIPIN	74.67344359	-0.90057	1.97E-06	-0.76696	5.27E-04		Yes	1
DDX39A	52.13601059	-0.90342	1.62E-06	-0.91151	8.99E-07		Yes	2
C9orf142	59.11652356	-0.94468	3.67E-08	-0.85251	2.73E-05		No	

Appendix 2

GO Biological Process: 30 more significantly up-regulated pathways

Gene Set	NGenes	PValue	FDR
GO_RESPONSE TO TYPE I INTERFERON	42	1.86E-10	7.50E-08
GO_INTERFERON GAMMA MEDIATED SIGNALING PATHWAY	38	1.88E-08	4.15E-06
GO_CELLULAR RESPONSE TO INTERFERON_GAMMA	49	6.53E-07	7.61E-05
GO_RESPONSE TO INTERFERON GAMMA	60	1.96E-06	0.000181
GO_VACUOLE ORGANIZATION	136	3.07E-05	0.001579
GO_PROTEIN LOCALIZATION TO GOLGI APPARATUS	25	9.57E-05	0.003925
GO_GOLGI ORGANIZATION	77	0.000108	0.00434
GO_IMMUNE EFFECTOR PROCESS	227	0.000121	0.004755
GO_DEFENSE RESPONSE TO VIRUS	96	0.000121	0.004755
GO_POSITIVE REGULATION OF RESPONSE TO EXTERNAL STIMULUS	120	0.000147	0.005529
GO_ENDOSOME ORGANIZATION	54	0.000191	0.006786
GO_GOLGI TO PLASMA MEMBRANE PROTEIN TRANSPORT	21	0.00022	0.007629
GO_RETROGRADE TRANSPORT VESICLE RECYCLING WITHIN GOLGI	16	0.000256	0.008667
GO_I KAPPAB KINASE NF KAPPAB SIGNALING	48	0.000342	0.010893
GO_REGULATION OF I KAPPAB KINASE NF KAPPAB SIGNALING	153	0.000421	0.01331
GO_REGULATION_OF VACUOLE ORGANIZATION	37	0.000426	0.013379
GO_UTERUS_DEVELOPMENT	8	0.000497	0.015074
GO_INNATE IMMUNE RESPONSE	240	0.000516	0.015374
GO_POST GOLGI VESICLE MEDIATED TRANSPORT	74	0.000516	0.015374
GO_REGULATION OF VESICLE MEDIATED TRANSPORT	271	0.000517	0.015374
GO_VESICLE MEDIATED TRANSPORT	787	0.000537	0.01585
GO_SERTOLI CELL DIFFERENTIATION	11	0.000716	0.020638
GO_ESTABLISHMENT OF PROTEIN LOCALIZATION TO GOLGI	13	0.000717	0.020638
GO_IMMUNE RESPONSE	390	0.000735	0.020892
GO_VACUOLAR TRANSPORT	214	0.000736	0.020892
GO_REGULATION OF RHO PROTEIN SIGNAL TRANSDUCTION	78	0.000744	0.021007
GO_CYTOKINE MEDIATED SIGNALING PATHWAY	191	0.000792	0.021947
GO_REGULATION OF ACTIN FILAMENT_LENGTH	99	0.000793	0.021947
GO_PATTERN RECOGNITION RECEPTOR SIGNALING PATHWAY	67	0.000815	0.022412
GO_RESPONSE TO INTERFERON ALPHA	15	0.000835	0.022561

GO Biological Process: 30 more significantly down-regulated pathways

Gene Set	NGenes	PValue	FDR
GO_RIBOSOME BIOGENESIS	285	2.30E-14	1.02E-10
GO_NCRNA_PROCESSING	349	1.31E-13	2.91E-10
GO_RIBONUCLEOPROTEIN COMPLEX BIOGENESIS	404	3.08E-13	3.83E-10
GO_TRANSLATIONAL TERMINATION	84	3.45E-13	3.83E-10
GO_MITOCHONDRIAL TRANSLATION	97	1.28E-12	1.13E-09
GO_rRNA METABOLIC PROCESS	237	1.63E-12	1.20E-09
GO_tRNA PROCESSING	95	5.90E-12	3.73E-09
GO_ncRNA METABOLIC PROCESS	468	1.47E-11	8.16E-09
GO_TRANSLATIONAL ELONGATION	96	4.85E-11	2.39E-08
GO_TRNA METABOLIC PROCESS	153	1.25E-10	5.53E-08
GO_RNA PROCESSING	724	2.40E-10	8.87E-08
GO_SPLICEOSOMAL SNRNP ASSEMBLY	35	8.24E-10	2.69E-07
GO_RIBOSOMAL LARGE SUBUNIT BIOGENESIS	46	8.51E-10	2.69E-07
GO_CELLULAR PROTEIN COMPLEX DISASSEMBLY	108	1.81E-09	5.33E-07
GO_DNA DEPENDENT DNA REPLICATION	93	4.13E-09	1.14E-06
GO_tRNA MODIFICATION	49	7.26E-09	1.89E-06
GO_RIBONUCLEOPROTEIN COMPLEX LOCALIZATION	106	9.46E-09	2.33E-06
GO_DNA REPLICATION INITIATION	27	1.36E-08	3.16E-06
GO_RNA_SPLICING_VIA_TRANSESTERIFICATION_REACTIONS	244	2.19E-08	4.61E-06
GO_RIBOSOMAL SMALL SUBUNIT BIOGENESIS	55	2.91E-08	5.86E-06
GO_AMIDE BIOSYNTHETIC PROCESS	402	3.09E-08	5.95E-06
GO_REGULATION OF TELOMERASE RNA LOCALIZATION TO CAJAL BODY	15	5.94E-08	1.10E-05
GO_DNA TEMPLATED TRANSCRIPTION TERMINATION	90	7.00E-08	1.24E-05
GO_RNA MODIFICATION	98	1.51E-07	2.57E-05
GO_RIBONUCLEOPROTEIN COMPLEX SUBUNIT ORGANIZATION	176	1.90E-07	3.12E-05
GO_tRNA TRANSPORT	33	2.14E-07	3.39E-05
GO_DNA REPLICATION	188	2.28E-07	3.48E-05
GO_RNA SPLICING	318	2.73E-07	4.03E-05
GO_RIBOSOME ASSEMBLY	46	2.91E-07	4.16E-05
GO_MULTI ORGANISM METABOLIC PROCESS	132	4.56E-07	6.31E-05

GO Cellular Compartments: 30 more significantly up (UP)– and down (DN)-regulated pathways

Gene set	NGenes	Direction	PValue	FDR
GO_NUCLEOLAR PART	56	DN	1.50E-12	8.64E-10
GO_ORGANELLAR RIBOSOME	64	DN	2.95E-11	8.53E-09
GO_RIBOSOME	194	DN	7.49E-11	1.44E-08
GO_MITOCHONDRIAL PROTEIN COMPLEX	117	DN	2.82E-10	3.90E-08
GO_INNER MITOCHONDRIAL MEMBRANE PROTEIN COMPLEX	91	DN	3.38E-10	3.90E-08

GO_RIBOSOMAL_SUBUNIT	145	DN	4.82E-10	4.54E-08
GO_PRERIBOSOME	57	DN	5.50E-10	4.54E-08
GO_MITOCHONDRIAL MEMBRANE PART	144	DN	5.62E-09	4.06E-07
GO_ORGANELLE INNER MEMBRANE	407	DN	1.04E-08	6.66E-07
GO_RIBONUCLEOPROTEIN COMPLEX	613	DN	1.58E-08	9.15E-07
GO_LARGE RIBOSOMAL SUBUNIT	87	DN	5.22E-08	2.75E-06
GO_SMALL SUBUNIT PROCESSOME	31	DN	5.95E-08	2.86E-06
GO_SMN - SM PROTEIN COMPLEX	15	DN	7.21E-08	3.04E-06
GO_SMALL NUCLEOLAR RIBONUCLEOPROTEIN COMPLEX	17	DN	7.37E-08	3.04E-06
GO_DNA PACKAGING COMPLEX	79	DN	9.22E-08	3.55E-06
GO_SPLICEOSOMAL TRI SNRNP COMPLEX	24	DN	1.98E-07	7.16E-06
GO_90S PRERIBOSOME	23	DN	2.62E-07	8.92E-06
GO_METHYLOSOME	12	DN	3.94E-07	1.27E-05
GO_MITOCHONDRIAL ENVELOPE	539	DN	7.49E-07	2.28E-05
GO_SMALL RIBOSOMAL SUBUNIT	58	DN	1.01E-06	2.91E-05
GO_CYTOSOLIC RIBOSOME	97	DN	1.16E-06	3.18E-05
GO_RESPIRATORY CHAIN	69	DN	2.38E-06	6.26E-05
GO_SMALL_NUCLEAR RIBONUCLEOPROTEIN COMPLEX	56	DN	2.76E-06	6.89E-05
GO_NUCLEAR PORE	66	DN	2.86E-06	6.89E-05
GO_PROTEIN DNA COMPLEX	130	DN	3.16E-06	7.31E-05
GO_MITOCHONDRIAL PART	749	DN	3.46E-06	7.69E-05
GO_MITOCHONDRIAL MATRIX	337	DN	4.26E-06	9.11E-05
GO_CATALYTIC STEP 2 SPLICEOSOME	85	DN	4.42E-06	9.11E-05
GO_ORGANELLAR LARGE RIBOSOMAL SUBUNIT	30	DN	4.59E-06	9.14E-05
GO_SPLICEOSOMAL COMPLEX	159	DN	5.08E-06	9.78E-05

GO Molecular Functions: 30 more significantly up (UP)– and down (DN)-regulated pathways.

Gene Sets	NGenes	Direction	PValue	FDR
GO_STRUCTURAL CONSTITUENT OF RIBOSOME	176	DN	3.80E-11	3.37E-08
GO_RNA POLYMERASE ACTIVITY	37	DN	5.79E-08	2.57E-05
GO_RAN GTPASE BINDING	28	DN	2.67E-07	7.93E-05
GO_POLY A RNA BINDING	1064	DN	3.75E-07	8.33E-05
GO_SNORNA BINDING	23	DN	1.08E-06	0.000192
GO_RNA BINDING	1357	DN	1.87E-06	0.000278
GO_STRUCTURAL CONSTITUENT OF NUCLEAR PORE	13	DN	1.05E-05	0.00134
GO_PURINE NTP DEPENDENT HELICASE ACTIVITY	81	DN	2.09E-05	0.002232
GO_RIBONUCLEASE ACTIVITY	73	DN	2.50E-05	0.002232
GO_ENDONUCLEASE ACTIVITY ACTIVE WITH EITHER RIBO OR DEOXYRIBONUCLEIC ACIDS AND PRODUCING 5' PHOSPHOMONOESTERS	30	DN	2.51E-05	0.002232
GO_ENDORIBONUCLEASE ACTIVITY PRODUCING 5' PHOSPHOMONOESTERS	24	DN	3.02E-05	0.002319

GO_TRNA SPECIFIC RIBONUCLEASE ACTIVITY	14	DN	3.13E-05	0.002319
GO_RNA HELICASE ACTIVITY	54	DN	3.75E-05	0.002497
GO_NUCLEASE ACTIVITY	151	DN	3.93E-05	0.002497
GO_RRNA BINDING	51	DN	5.94E-05	0.00347
GO_SINGLE STRANDED DNA BINDING	76	DN	6.24E-05	0.00347
GO_TRANSLATION FACTOR ACTIVITY RNA BINDING	72	DN	8.79E-05	0.004438
GO_TRANSLATION INITIATION FACTOR ACTIVITY	46	DN	8.99E-05	0.004438
GO_METAL CLUSTER BINDING	51	DN	0.00012	0.005593
GO_PSEUDOURIDINE SYNTHASE ACTIVITY	12	DN	0.000154	0.006839
GO_NUCLEOCYTOPLASMIC TRANSPORTER ACTIVITY	23	DN	0.000171	0.007234
GO_ENDONUCLEASE ACTIVITY	87	DN	0.000231	0.009324
GO_ENDORIBONUCLEASE_ACTIVITY	38	DN	0.000241	0.00933
GO_RAS GUANYL NUCLEOTIDE EXCHANGE FACTOR ACTIVITY	122	UP	0.000287	0.010213
GO_EXONUCLEASE ACTIVITY	66	DN	0.000287	0.010213
GO_DNA SECONDARY STRUCTURE BINDING	20	DN	0.000341	0.011675
GO_HELICASE ACTIVITY	133	DN	0.000444	0.014626
GO_DNA HELICASE ACTIVITY	48	DN	0.000468	0.014861
GO_RNA METHYLTRANSFERASE ACTIVITY	38	DN	0.000502	0.015386
GO_FOUR WAY JUNCTION DNA BINDING	13	DN	0.000578	0.017131
GO_GUANYL NUCLEOTIDE EXCHANGE FACTOR ACTIVITY	174	UP	0.000604	0.017311
GO_ARF GUANYL NUCLEOTIDE EXCHANGE FACTOR ACTIVITY	19	UP	0.000664	0.018446
GO_NUCLEOTIDYLTRANSFERASE ACTIVITY	106	DN	0.000735	0.019792
GO_TRANSFERASE ACTIVITY TRANSFERRING ONE CARBON GROUPS	160	DN	0.000815	0.021319
GO_PEPTIDE ANTIGEN BINDING	13	UP	0.000953	0.024217
GO_SH3 DOMAIN BINDING	74	UP	0.001192	0.028864
GO_4 IRON 4 SULFUR CLUSTER BINDING	38	DN	0.001201	0.028864
GO_RIBONUCLEASE P ACTIVITY	9	DN	0.001522	0.035246
GO_DNA DEPENDENT ATPASE ACTIVITY	70	DN	0.001546	0.035246
GO_ENZYMEACTIVATOR ACTIVITY	308	UP	0.001752	0.038546
GO_PHOSPHATIDYLINOSITOL BINDING	139	UP	0.001778	0.038546
GO_STRUCTURAL_MOLECULE_ACTIVITY	402	DN	0.001828	0.038687
GO_RHO GUANYL NUCLEOTIDE EXCHANGE FACTOR ACTIVITY	56	UP	0.001915	0.039596
GO_PHOSPHATIDYLINOSITOL PHOSPHATE BINDING	78	UP	0.002005	0.040506
GO_RIBONUCLEOPROTEIN COMPLEX BINDING	84	DN	0.002163	0.041377
GO_PROTEIN CHANNEL ACTIVITY	9	DN	0.002181	0.041377
GO_TRANSLATION INITIATION FACTOR BINDING	23	DN	0.002189	0.041377
GO_OXIDOREDUCTASE ACTIVITY ACTING ON NAD P H QUINONE OR SIMILAR COMPOUND AS ACCEPTOR	44	DN	0.002234	0.041377
GO_CADHERIN BINDING	19	UP	0.002613	0.046818

GO_TRANSLATION REPRESSOR ACTIVITY	13	UP	0.002633	0.046818
GO_1-PHOSPHATIDYLINOSITOL BINDING	13	UP	0.002787	0.048582

REACTOME: 30 more significantly up -regulated pathways.

Gene Sets	NGenes	PValue	FDR
REACTOME_INTERFERON ALPHA-BETA SIGNALING	38	1.19E-08	6.03E-07
REACTOME_INTERFERON GAMMA SIGNALING	33	2.94E-06	4.49E-05
REACTOME_ANTIGEN PRESENTATION FOLDING ASSEMBLY AND PEPTIDE LOADING OF CLASS I MHC	17	0.000145	0.001136
REACTOME_GRB2 SOS PROVIDES LINKAGE TO MAPK SIGNALING FOR INTERGRINS	8	0.000321	0.002318
REACTOME_SIGNAL REGULATORY PROTEIN SIRP FAMILY INTERACTIONS	7	0.000471	0.003298
REACTOME_MEMBRANE TRAFFICKING	102	0.000563	0.003779
REACTOME IL RECEPTOR_SHC SIGNALING	14	0.000677	0.0045
REACTOME_RIG 1 MDA5 MEDIATED INDUCTION OF IFN ALPHA-BETA PATHWAYS	49	0.000759	0.004846
REACTOME_CELL CELL COMMUNICATION	70	0.000905	0.005672
REACTOME_INNATE IMMUNE SYSTEM	138	0.002596	0.014776
REACTOME_IL 6 SIGNALING	8	0.002605	0.014776
REACTOME_CELL SURFACE INTERACTIONS AT THE VASCULAR WALL	40	0.002672	0.015029
REACTOME_SYNTHESIS OF PIPS AT THE PLASMA MEMBRANE	23	0.002777	0.015357
REACTOME_SIGNALING BY ILS	72	0.003303	0.017961
REACTOME_P130CAS LINKAGE TO MAPK SIGNALING FOR INTEGRINS	8	0.003616	0.019342
REACTOME_IL 3 5 AND GM CSF SIGNALING	26	0.003934	0.020711
REACTOME_TRANS GOLGI NETWORK VESICLE BUDDING	52	0.004405	0.023003
REACTOME_REGULATION OF IFNA SIGNALING	10	0.00567	0.028922
REACTOME_BIOLOGICAL OXIDATIONS	41	0.006274	0.031515
REACTOME_APOPTOTIC CLEAVAGE OF CELLULAR PROTEINS	30	0.00783	0.037882
REACTOME_PI METABOLISM	38	0.008375	0.040083
REACTOME_NEPHRIN INTERACTIONS	13	0.010027	0.047129
REACTOME_SIGNALING BY HIPPO	20	0.010505	0.048257
REACTOME_ENDOSOMAL VACUOLAR PATHWAY	6	0.010561	0.048257
REACTOME_REGULATION OF KIT SIGNALING	12	0.011207	0.050858
REACTOME_CYTOKINE SIGNALING IN IMMUNE SYSTEM	178	0.011382	0.051299
REACTOME_METABOLISM OF LIPIDS AND LIPOPROTEINS	322	0.012209	0.054652
REACTOME_TIE2 SIGNALING	10	0.0128	0.056527
REACTOME_TRANSPORT TO THE GOLGI AND SUBSEQUENT MODIFICATION	26	0.013687	0.06004
REACTOME_SIGNALING BY BMP	15	0.014538	0.06335

REACTOME: 30 more significantly down -regulated pathways.

Gene set	NGenes	PValue	FDR
REACTOME_PROCESSING OF CAPPED INTRON CONTAINING PRE MRNA	133	4.65E-12	3.06E-09
REACTOME_MRNA PROCESSING	147	1.86E-11	4.72E-09
REACTOME_MRNA SPLICING MINOR PATHWAY	40	2.15E-11	4.72E-09
REACTOME_METABOLISM OF NON CODING RNA	46	3.58E-11	5.21E-09
REACTOME_TELOMERE MAINTENANCE	68	3.96E-11	5.21E-09
REACTOME MRNA SPLICING	104	4.62E-10	5.07E-08
REACTOME_TRANSCRIPTION	184	5.79E-10	5.45E-08
REACTOME_DNA STRAND ELONGATION	29	1.87E-09	1.54E-07
REACTOME_CHROMOSOME MAINTENANCE	104	4.10E-09	3.00E-07
REACTOME_MEIOTIC RECOMBINATION	66	7.83E-09	5.15E-07
REACTOME_ACTIVATION OF THE PRE REPLICATIVE_COMPLEX	29	9.56E-09	5.72E-07
REACTOME_RNA POL I RNA POL III AND MITOCHONDRIAL TRANSCRIPTION	103	1.05E-08	5.78E-07
REACTOME_RNA POL I PROMOTER OPENING	52	1.57E-08	7.37E-07
REACTOME_TRANSPORT OF MATURE TRANSCRIPT TO CYTOPLASM	52	3.04E-08	1.26E-06
REACTOME_INFLUENZA LIFE CYCLE	126	3.22E-08	1.26E-06
REACTOME_EXTENSION OF TELOMERES	25	3.26E-08	1.26E-06
REACTOME_RNA POL I TRANSCRIPTION	74	3.46E-08	1.26E-06
REACTOME_S PHASE	100	6.22E-08	2.15E-06
REACTOME_TRANSPORTOF MATURE MRNA DERIVED FROM AN INTRONLESS TRANSCRIPT	33	9.98E-08	3.19E-06
REACTOME_DEPOSITION OF NEW CENPA CONTAINING NUCLEOSOMES AT THE CENTROMERE	58	1.02E-07	3.19E-06
REACTOME_MITOCHONDRIAL PROTEIN IMPORT	44	1.07E-07	3.21E-06
REACTOME_ACTIVATION OF ATR IN RESPONSE TO REPLICATION STRESS	34	1.24E-07	3.55E-06
REACTOME_METABOLISM OF RNA	239	2.88E-07	7.90E-06
REACTOME_MITOTIC G1/G1/S PHASES	121	3.29E-07	8.67E-06
REACTOME_RNA POL II TRANSCRIPTION	93	3.78E-07	9.50E-06
REACTOME_CELL CYCLE	366	4.01E-07	9.50E-06
REACTOME MEIOSIS	87	4.04E-07	9.50E-06
REACTOME_PACKAGING OF TELOMERE ENDS	43	5.25E-07	1.19E-05
REACTOME_HIV LIFE CYCLE	103	5.64E-07	1.20E-05
REACTOME_G1/S TRANSITION	100	5.65E-07	1.20E-05

Appendix 3

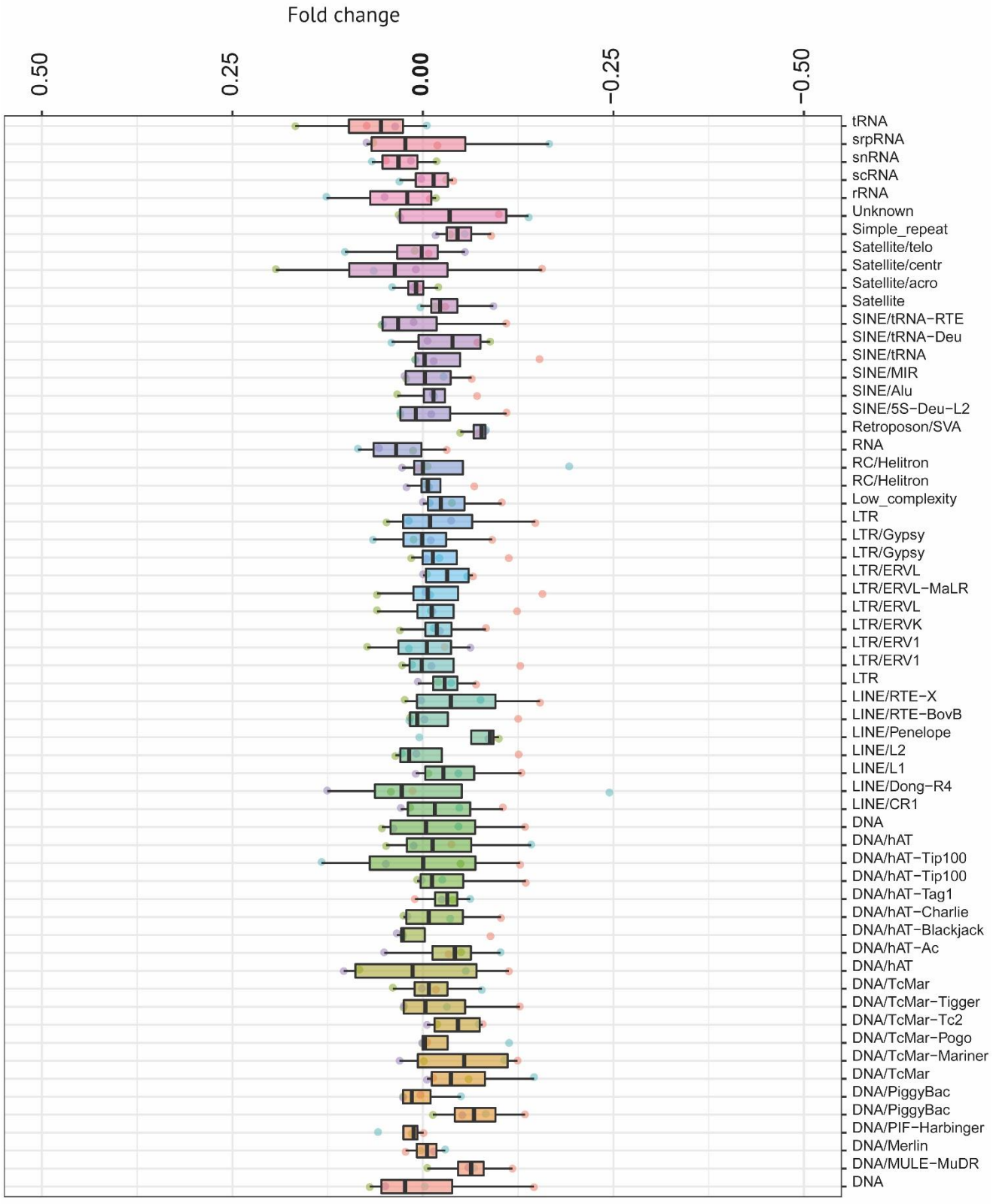
Number of sequences belonging to each family of endogenous retrotransposon

Family	Number of sequnces
SINE/Alu	1205213
LINE/L1	983004
Simple_repeat	671061
SINE/MIR	581929
LINE/L2	461251
LTR/ERVL-MaLR	348738
DNA/hAT-Charlie	256286
LTR/ERV1	175743
LTR/ERVL	163045
DNA/TcMar-Tigger	116242
Low_complexity	96306
LINE/CR1	65860
DNA/hAT-Tip100	45158
DNA/hAT-Blackjack	19212
LTR/Gypsy	16631
DNA/TcMar-Mariner	16012
LINE/RTE-X	15215
LTR/ERVK	10923
DNA/hAT	8631
LINE/RTE-BovB	8609
DNA/TcMar-Tc2	8019
LTR/Gypsy	7303
Retroposon/SVA	5529
LTR	5523
Unknown	5433
SINE/tRNA-RTE	5382
Satellite	4270
DNA/hAT-Ac	4108
DNA	3167
LTR	3143
Satellite/centr	2782
snRNA	2512
SINE/5S-Deu-L2	2390
DNA	2200
LTR/ERVL	2147
DNA/PiggyBac	2146
SINE/tRNA	2102
DNA/hAT-Tip100	1979
DNA/MULE-MuDR	1953
tRNA	1932

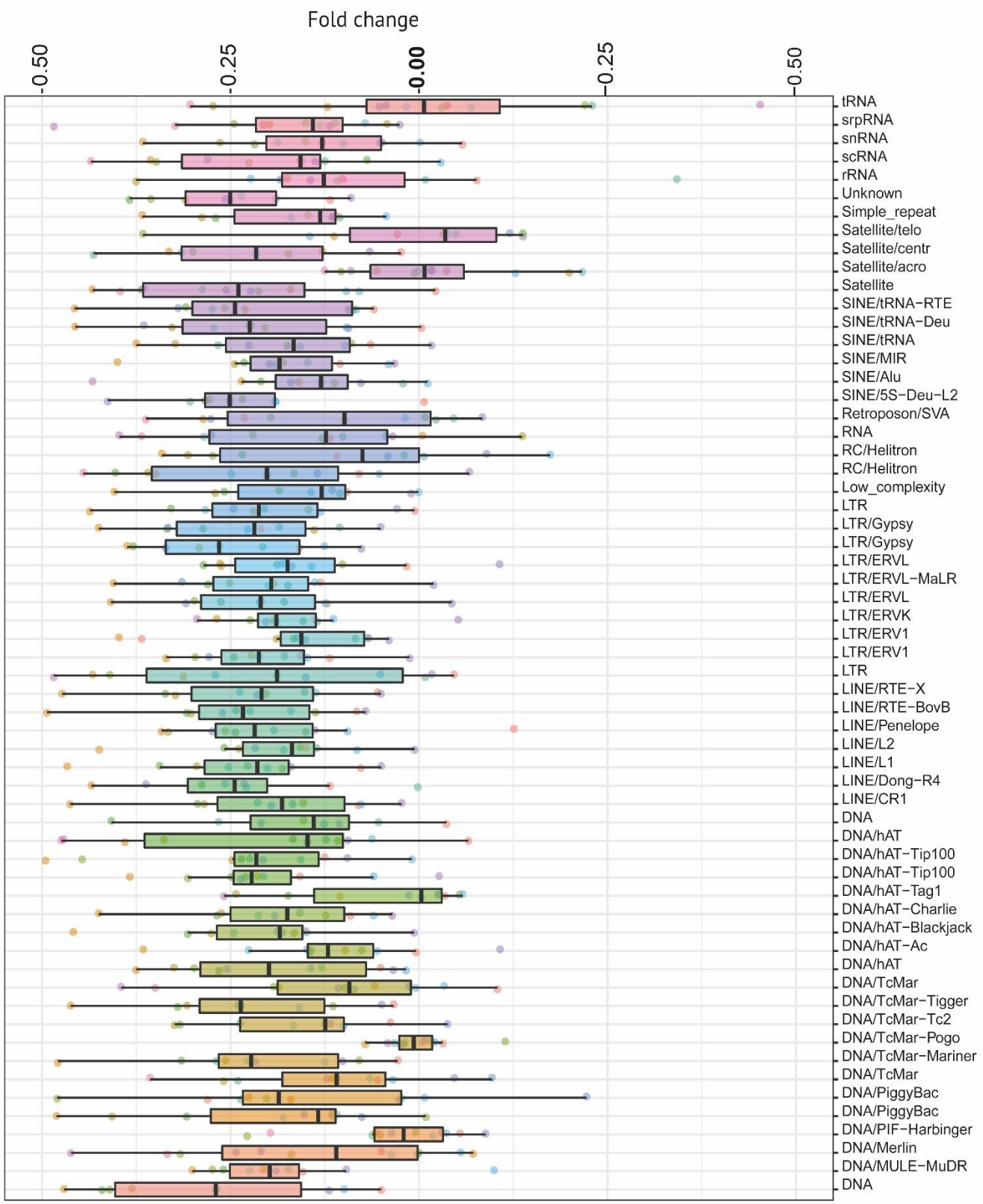
Family	Number of sequnces
RC/Helitron	1739
DNA/hAT	1501
rRNA	1246
LTR/ERV1	1235
LINE/Penelope	1040
srpRNA	982
scRNA	601
SINE/tRNA-Deu	600
LINE/Dong-R4	518
RNA	399
RC?/Helitron	395
Satellite/telo	367
DNA/TcMar	345
DNA/PiggyBac	212
DNA/TcMar	169
DNA/hAT-Tag1	145
Satellite/acro	85
DNA/Merlin	56
DNA/TcMar-Pogo	35
SINE	33
DNA/PIF-Harbinger	30
SINE/tRNA	5

Induction of Endogenous retroviruses in ATRA totally-resistant (ATRA-*score* = 0, part A) and ATRA-sensitive (part B) cell lines.

Part A



Part B


Fold change
-0.50
-0.25
0.00
0.25
0.50
tRNA
srpRNA
snRNA
scRNA
rRNA
Unknown
Simple_repeat
Satellite/telo
Satellite/centr
Satellite/acro
Satellite
SINE/tRNA-RTE
SINE/tRNA-Deu
SINE/tRNA
SINE/MIR
SINE/Alu
SINE/5S-Deu-L2
Retroposon/SVA
RNA
RC/Helitron
RC/Helitron
Low_complexity
LTR
LTR/Gypsy
LTR/Gypsy
LTR/ERVL
LTR/ERVL-MaLR
LTR/ERVL
LTR/ERVK
LTR/ERV1
LTR/ERV1
LTR
LINE/RTE-X
LINE/RTE-BovB
LINE/Penelope
LINE/L2
LINE/L1
LINE/Dong-R4
LINE/CR1
DNA
DNA/hAT
DNA/hAT-Tip100
DNA/hAT-Tip100
DNA/hAT-Tag1
DNA/hAT-Charlie
DNA/hAT-Blackjack
DNA/hAT-Ac
DNA/hAT
DNA/TcMar
DNA/TcMar-Tigger
DNA/TcMar-Tc2
DNA/TcMar-Pogo
DNA/TcMar-Mariner
DNA/TcMar
DNA/PiggyBac
DNA/PiggyBac
DNA/PIF-Harbinger
DNA/Merlin
DNA/MULE-MuDR
DNA

Ringraziamenti

Questo progetto di tesi è stato sviluppato durate il mio tirocinio presso l'Istituto di Ricerche Farmacologiche Mario Negri IRCCS.
Un ringraziamento particolare va alla Prof. Pattini, mia relatrice, che oltre ad avermi guidato nella stesura di questo lavoro, mi ha offerto l'opportunità di lavorare in un ambiente stimolante e ricco di stimoli per il mio futuro.
Vorrei ringraziare tutti i membri del Laboratorio di Biologia Molecolare dell'Istituto Mario Negri, in particolare il Dr. Enrico Garattini e la Dr. Maddalena Fratelli, che mi hanno accolta e supportata nello sviluppo del mio progetto, spronandomi a dare sempre il meglio di me, senza porre limiti agli obiettivi che avrei potuto raggiungere.
Un ringraziamento particolare al Dr. Marco Bolis, che oltre ad avermi ispirato con la sua inesauribile passione per questo lavoro, mi ha accompagnato nel mio percorso di formazione, e, forse per primo, ha creduto nelle mie capacità, incoraggiandomi a fare altrettanto.

Un ringraziamento speciale alla mia famiglia, a mamma e papà: è grazie a loro sostegno incondizionato se oggi sono riuscita a raggiungere questo traguardo.

Grazie a chi c'è stato quando non credevo di potercela fare, a te che mi hai fatto alzare lo sguardo verso i traguardi che non credevo più di poter avvicinare.

Grazie ai miei amici, alle amiche di sempre e a chi invece c'è da poco, ma è come se ci fosse sempre stata. Grazie agli amici che hanno condiviso con me gli anni al Politecnico, e a quelli che sono ancora accanto a me dopo tutti questi anni.

Infine, il mio grazie va tutti coloro che dedicano la loro vita alla ricerca scientifica nel nostro Paese, e nonostante le difficoltà che questo lavoro comporta, vi si dedicano ogni giorno con passione, contribuendo a mantenere alto il nome della ricerca italiana nel mondo.