Università degli Studi di Torino

Facoltà di Scienze Matematiche Fisiche e Naturali

Dipartimento di Scienze Oncologiche

Dottorato di Ricerca in
Sistemi Complessi in Biologia Postgenomica

CICLO XX

FUNCTIONAL CLASSIFICATION OF ESTROGEN-RESPONSIVE GENE REGULATORY SEQUENCES IN BREAST CANCER CELLS: TOWARDS THE IDENTIFICATION OF REGULATORY NETWORKS

tesi presentata da:

Dott.ssa Gioia Altobelli

Relatori interni: *Proff. Michele De Bortoli - Michele Caselle*

Relatori esterni: *Dr. A. Benecke, Dr. G. Pavesi, e Prof. S. Subramaniam*

Coordinatore del ciclo: *Prof. F. Bussolino*

Anni Accademici: 2004 - 2007

SETTORE SCIENTIFICO-DISCIPLINARE DI APPARTENENZA:

BIO-11

Be patient, for the world is broad and wide.
Flatland, E.Abbott

Contents

Summary

Understanding regulation of estrogen-responsive genes is central to molecular biology and of great interest in medicine. The transcriptional activation/repression due to estrogen stimuli is gene- and cell- type specific, with a relevant molecular syntax being either unknown or not completely understood. Tissue- and cell type-specificity of the physiological response to estrogen has been addressed in experimental models by employing large-scale approaches, and results suggest that both complexity of transcriptional coregulators and epigenetics of chromatin organization are involved. Existence of several regulatory classes, e.g. early/late up/down- regulated clearly appears in microarrays and ChIP-on-Chip studies, but little is known about the underlying features of the corresponding gene regulatory sequences. Biochemical pathways target different DNA sequence elements and build up the combinatorial control which is a key in the regulatory events. The distribution of these DNA elements in the responsive gene regulatory regions should enable the inference of this regulatory networks.

We collected and compared expression data from genome-wide experiments in breast cancer cell models with a view to characterizing DNA flanking regions of hundreds of estrogen-responsive genes, possibly assessing differences between up- vs. down-regulated classes. In other words, we aimed at identifying the sequence motifs that may describe the differences between genes that are up- and down- regulated by estrogen, suggesting context features and possible control pathways. We set up a bioinformatics pipeline which combines traditional approaches focused on DNA sequence analysis of prox-

imal regions with a method that enables investigation of distal conserved nucleotide blocks of co-regulated estrogen-responsive genes (early responders only). We mainly focused our attention on those motifs identified by all of the tools and/or in different experimental datasets, with a view to inferring both regulatory factors to be tested in laboratory and relevant regulatory networks.

Although chromatin is a major context determinant of gene responsiveness, our pipeline handles the DNA sequences as linear strings. A topographic perspective is achievable to some extent, but we did not attempt it on a large scale. This pipeline also assumes that transcription binding sites tend to be overrepresented and to cluster in modules, and that evolution conservation is a key in their functionality. Despite universality of these assumptions has been recently challenged, we could assess remarkable features of the sequences which may have important biological implications. The upstream regions of early up-regulated genes strongly differ from the ones of early down-regulated genes, suggesting different regulatory mechanisms for the two classes of genes. Significant motifs' localization is provided, along with ontological analysis of gene subsets and transcription factor distributions. An example of co-localization of transcription factor binding sites in the 5'-flanking sequence of cyclin G2, which suggests a direct interaction between estrogen receptor and GATA-3 factor (ER-GATA), is also discussed in detail. This interaction may be important in mammary gland development.

Chapter 1

The complexity of estrogen regulation in health and disease

We provide the background of the project and outline the thesis goals.

1.1 Mechanisms of estrogen receptor signaling

1.1.1 Context-dependent complexity

Estrogen plays a role in a variety of key physiological processes in the human body, as well as in the development and progression of many diseases. The binding of estrogen to its nuclear receptor triggers a chain of molecular events. The transcriptional result (e.g. activation/repression) is determined by the type of co-factor complexes recruited by the activated receptor in a gene- and cell type-specific way. The syntax of these complexes' recruitment is not completely understood, while a comprehension of how the relevant genes are regulated is especially interesting in medicine. In breast cells, estrogen induce cell proliferation and, for this reason, breast cancer is largely treated with antiestrogenic drugs. Antiestrogenic treatments fail in about 40% of cases, since the cellular context becomes progressively capable of converting antiestrogenic action into estrogenic. Similarly, some of these drugs are antiestrogenic in breast cells but estrogenic in other cells (such as osteoblast, for

example). Generally speaking, the cellular context results from the interplays of several factors: chromatin remodeling proteins, RNA polymerase II complexes, activity of co-activators and co-repressors, transcription factors and signal transduction pathways [6]. The cellular context makes the response highly variable in different tissues and gives rise to complex consequences in medicine. *Tamoxifen* – an effective anti-proliferative drug employed in breast cancer, may have opposite effects in endometrium where the ratio of concentration of specific co-repressors/-activators is lower than in breast tissue. Besides, in a type of breast cancer where HER2/neu membrane receptor is over-expressed, resistance to *Tamoxifen* emerges at an early stage of the treatment, due to detrimental synergy between the proliferation generated via estrogen receptor (**ER**) and the signal cascade activated through receptor HER2/neu.

1.1.2 Estrogen receptors as allosteric switches

Over the past 50 years, a host of studies have focused on transcriptional control in vertebrates and, specifically, on the nuclear receptor (NR) superfamily which mediates the response to steroid hormones and to other signaling molecules. Estrogen receptors belong to this superfamily of transcription factors, which are modular proteins composed of three major domains: a *transactivation* domain (N-terminus), which is the most variable among the family members and whose 3D-structure is undetermined; a rigid DNA binding domain (**DBD**) with two zinc-finger motifs common to the entire family; a ligand binding domain (**LBD**), well conserved and structured but able to exhibit remarkable plasticity. The overall structural plasticity enables nuclear receptors to respond to a variety of inputs in the specific tissue, given the specific set of proteins they interact with. In molecular terms, isosteric ligands induce receptor surface changes which favor recruitment of co-activator vs. co-repressor complexes, thus determining transcriptional activation vs. repression. Hence, ligand-induced allosteric alteration of estrogen receptor, ER, is a key molecular event that modulates the response to both estrogen and selective estrogen receptor modulators (**SERM**s). There are two estrogen receptor isoforms, ERα and ERβ, whose structural and functional

organization is very similar [1]. They are specifically distributed in the human body. The concentration ratio of ERα and ERβ is thought to have a functional role [3, 12], with ERα and ERβ often acting in opposite ways in terms of cellular effect [3, 4]. No crystal structure of an entire nuclear receptor is available; but, for example, structures of the DBD domain bound to its responsive element ERE (see Figure 1.1), and of the ligand binding domain bound to several compounds have been obtained for ERα [3, 5]. Estrogen receptors may form both homo- and hetero-dimers, one to another and with other nuclear receptors [1,3]. For example, there is evidence that crosstalk exists between ERα and the estrogen receptor-related receptors (ERRs), which enable estrogen to regulate genes [3].The formation of heterodimers allows for the enhancement of combinatorial effects of ligands. ERs are activated by estrogen and SERMs, but they may be effective even in absence of ligands. There is also evidence for unliganded activation and modulation of ER activities due to post-translational modifications (phosphorylation, ubiquitinylation, acetylation), which can occur in specific sites within transactivation and ligand binding domains (see [6]).

1.1.3 Genomic and non-genomic pathways

Two different models of estrogen action are currently reviewed in literature [1]. In one mechanism – the so-called *genomic pathway*, ERs are located in the nucleus and, after complexation with ligand, bind DNA either directly to the estrogen response element sequences (**ERE**s) or through other proteins already bound to the DNA, such as Sp1 and NFkB. As a result, appropriate co-regulatory proteins are recruited to the promoter and mRNA levels are modulated accordingly. These events produce a physiological response on the hours time scale. This mechanism is often referred to as 'classical' in opposition to the fastest *non-genomic pathway*, which occurs within seconds/minutes after activation of ER located in cytoplasm, near or embedded into the plasma membrane. The non-genomic mechanism involves protein kinase cascades, with rapid changes in the concentration of cellular second-messengers and phosphorylation of several structural proteins and transcription factors. The genomic and non-genomic pathways may converge, con-

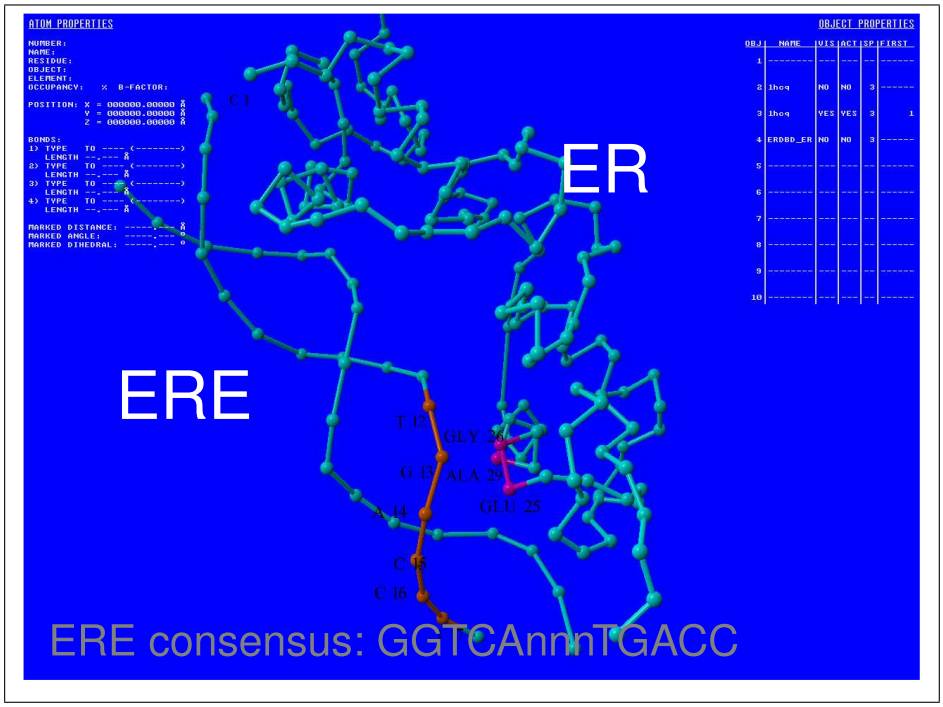

Figure 1.1: Estrogen receptor DNA binding domain-DNA complex: a ball-and-stick model from crystal structure published by Schwabe et al. [2]). Key residues in color. Estrogen receptor element is a palindromic sequence composed of two RGGTCA motifs separated by 3-nt space which is a key in the recognition and binding events. YASARA molecular modeling software package.

tributing the complexity of the estrogen regulation [8, 9]. Phosphorylation and other post-translational modifications mentioned above is the major biochemical events responsible for the achievement of nuclear receptor-mediated and signal transduction pathways' convergence.

Genomic regulation In the genomic pathway, estrogen receptors primarily act as nuclear receptors. In absence of ligand, the receptor is bound to multiproteic complexes which contain heat shock proteins and immunophilins. The binding of estrogen (or other ligands) induces a conformational change that freed the receptor, which is able to form dimers upon binding the DNA either directly or indirectly. When the receptor adopts a different conforma-

tion, a different molecular surface is formed which may interact with different complexes, either co-activators or repressors. Both the receptor transactivation domain (N-terminal) and the ligand binding domain LBD (C-terminal) surfaces contact specific proteins, which in turn mediate interaction with either co-activator or co-repressor complexes. This events result in either up-regulation or down-regulation of the relevant gene. While the interactions involving the N-terminal are not very well understood, the C-terminal interactions have been described in depth. Here do we briefly illustrate a pharmacological example. From the comparison of the crystal structures of LBD bound to estrogen and to raloxifene, an antagonist, a mechanism for explaining agonism/antagonism has been derived [11]. When estradiol occupies the binding pocket, a helix from the LBD, H12, caps the ligand in the cavity and shields the pocket from external environment contributing to the hydrophobic pocket; the resulting molecular surface favors the interaction with co-activators, through the recognition of co-activator LXXLL motif by a hydrophobic groove formed by helices 3,5,12. When instead raloxifene is bound, H12 is displaced by the conformation of ligand and its position prevents the recognition of the nuclear-receptor surface by co-activator LXXLL motif. The receptor now displays increased affinity for the NCoR/SMART co-repressors [6]. Since a pletora of co-activators and co-repressors – each having common and distinct activities– exist in the cell and in different cell types, subtle changes in the LBD structure may determine which set of co-regulators are preferred and, therefore establish the final transcriptional result. Even the DNA sequence may display an allosteric effect on co-activator recruitment, at least in the case of glucocorticoid receptor [7]
As repeatedly emphasized above, estrogen receptor (ER) acts an allosteric switch due to its conformational plasticity, and its conformational changes are induced by several types of inputs: 1) isosteric ligands, 2) ERE sequence variations [10], and 3) phosphorylation which results from the activation of signaling cascades.

Non-genomic regulation The major protein kinase cascades involved in the estrogen response are the following: Src homology and collagen/growth factor receptor binding protein 2/SOS/mitogen-activated protein kinase (MAPK),

phosphatidylinositol 3-kinase/AKT, and cyclic AMP (cAMP)/protein kinase A (PKA) pathways [8]. These pathways phosphorylate and activate several transcription factors among which Jun, for example; and, through phosphorylation, they may change the activity of co-activators and co-repressors and/or determine their nucleo-plasmatic delocalization [6, 8].

1.2 Large-scale approaches to estrogen regulation

Understanding the tissue-specificity of estrogen regulation requires large-scale approaches which enable comparison of transcriptional profiles under different experimental/physiological conditions. Microarrays have been employed in order to address the gene responsiveness to estrogen and/or anti-estrogens in both cellular and clinical models. Expressed genes are mainly linked to proliferation, apoptosis, and development. Results of these experiments display different classes of genes with different kinetics of response, that is genes which respond quickly (1-3 hours) and genes which respond with delay (3-6 or 10-12 hours). In addition to this, some genes respond transiently and others are stably regulated. It was also observed that the number of genes which are repressed by estrogen equals the number of genes that are induced. This was somehow unexpected. Estrogenic regulation was characterized on gene models that turned out to be induced gene, and the theoretical models of estrogenic response in molecular biology textbook only pictures recruitment of transcriptional coactivators as a result of estradiol(E2)/ER binding to DNA. These different groups of co-regulation –e. g. up- vs. down-regulated and early vs. late– clearly point to different mechanisms and pathways of regulation.

As mentioned above, there are several known mechanisms by which estrogen can achieve gene regulation. The first way is by ER binding to responsive regions in target genes, either by direct DNA interaction on ERE [2, 15] or by protein-protein contact with another DNA binding factor, as described in the case of NFkB, Sp1 and AP-1 [15]. The second possibility arises from nongenomic responses through the MAPK pathway: this results in phospho-

rylation, activation or cellular re-location of transcription factors (e.g. Jun) and coactivators or corepressors (e.g. SRC1, SMRT), which finally act on specific genes. In addition to this, even the 'secondary' transcriptional response –i.e. estrogen stimulating or inhibiting the expression of transcription factors (e.g. c-fos, TFAP2c, c-MYB, c-MYC) which in turn bind to and regulate a secondary set of genes– may have a role in estrogen regulation. In principle this modality is limited to delayed responses –it takes some time to a cell to transcribe and to translate significant amounts of new proteins– but examples of this type of regulation are known to be able to influence gene transcription within 2-3 hours.

All of these regulation categories have been well-documented on single genes in the case of positive regulation by estrogen, whereas little or no information is available for gene repression. In order to understand the complexity of genomic responses to estrogen, it is mandatory to distinguish among these possibilities. In pursuing this task, the first essential step is to determine whether a regulated gene is a direct target of estrogen-activated ER, either by direct ER-ERE interaction or by ER-to-factor-DNA interaction. Therefore, experiments were carried out in order to map ER binding sites in the genome and to correlate them with microarray results. The ChIP-on-Chip technique –the immunoprecipitation of chromatin fragments with factor-specific antibodies followed by identification of the associated DNA fragments on promoter or genomic tiled microarrays– has provided a powerful tool for defining transcription factor binding site profiles. ChIP-on-Chip data for ERα were produced essentially by three groups [12, 25, 65], working with different approaches (see Chapter 2 for a list of genes). Data from genome-wide ER mapping have been matched with expression microarray data. From this comparison, it appears that there is not outstanding over-representation of ER-binding in any coregulation class, not even in the early upregulated class, which theoretically represents the paradigm of immediate gene response to a nuclear receptor ligand.

Since the estrogen responsive element (ERE) has been described in some detail, from both mutagenesis and structural data, matrix alignment can be used to map ERE-sequences through bioinformatics approaches. The highest affinity binding sites for ERE in vitro are palindromic elements composed of

two RGGTCA motifs separated by 3 bp (see FIG 1.1). Bourdeau et al. [19] screened for consensus and near-consensus EREs the (-10kb to +5kb) flanking regions of human and mouse genes (refer to Chapter 2 for a list of pairs of orthologs). They identified approximately 70,000 motifs in both genomes and demostrated that near-consensus EREs occur frequently in both genomes, and that far upstream elements can be evolutionarily conserved and bind ERα in vivo. Besides, they found multiple occurences of conserved and non-conserved elements in more than 230 genes that are estrogen-stimulated in microarray experiments. In a more recent yet unpublished work from our group, ERE was mapped by a new algorithm (Cardamone et al.,in preparation) and found to be more or less equally represented in the -2000 to +500 region of both up-regulated and down-regulated genes. The bioinformatics screenings for EREs, anyways, are hampered by the nature of ERE signal itself (see introdution to Chapter 3).

Interestingly, there is evidence that estrogen-bound ER can act as a repressor for certain genes: the cyclin G2 gene, which is robustly downregulated by estrogen treatment, contains a half-ERE that ChIP, mutagenesis and reporter studies demonstrated to be essential for estrogenic effect. Besides, the corepressor NCoR (the same factor which is recruited by antagonist-bound ER) was found at the cyclin G2 promoter [13]. The same was proved for other 2 estrogen-down-regulated genes, among which the bone morphological factor 7 [50]. These results raise the following questions: How can the same transcription factor be an activator of certain genes and a repressor of certain others? How is the E2-ER complex interpreted by other cellular proteins in such a way that it has dramatically divergent effects? We can speculate that the overall capacity of the receptor to interact with other proteins is defined by the context of the gene and associated promoter-specific factors, which together function to establish an activating or a repressing molecular environment. For example, the Sp1 factor is known to interact with ER, and its cognate CG-rich elements are frequently found accompanying the EREs. This is the case of cyclin G2, in which Sp1 plays a pivotal role either in activation or in ER-mediated repression. Of course, the same line of reasoning can be applied with more ease to secondary response, in which the composition of transcription factor binding sites in a gene determine its transcriptional

response to the composition of factors primarily regulated by estrogen.

1.3 Goals of the thesis

The sequence features of the regulatory region of the genes represent the primary environment to receive ER signal and to interpret it. This 'environment' should not be considered as a pure DNA sequence or DNA-protein ensemble, but it has to be thought as a complex interplay of signals that regulate protein-DNA interaction and protein-protein interaction and a complex chromatin dynamics that would or not allow interaction of the proteins with DNA. The key assumption here is that the primary determinant resides in the DNA sequence itself. We reasoned that such a 'sequence environment' may be traced looking at groups of genes that display homogeneous mode of regulation (co-regulons). Thus, among the different gene sets defined by microarray experiments, we concentrated our attention on the more striking co-regulons, i.e. genes that are up- and down- regulated early after estrogen treatment in cell culture, in a time lap when secondary regulation is presumably less important. The 5'-flanking regions of these genes were explored with different bioinformatics tools, in order to identify a number of candidate sequence motifs which could 1) represent target of regulation by transcription factors or other nuclear proteins, 2) significantly associate with the type of transcriptional result observed, and 3) hint at specific pathways and modes of 'control' of the estrogenic signal, independently on the presence or absence of the ER itself.

Therefore, we collected and compared genome-wide data from breast cancer cell models, with a view to functionally characterizing the regulatory regions of estrogen-responsive genes through DNA sequence analyses. We focused on the early-responder genes, setting an upper limit for responsiveness to 4 hours after estrogenic stimulation. We prepared four datasets extracting lists of up/ down-regulated genes from the four homogeneous experiments available in literature, and prepared an additional collection of data from different platforms and cell lines (metaset). We set up a bioinformatics pipeline which combines traditional approaches sequence analysis of proximal DNA flanking regions with a method that allows for exploring distal, conserved nucleotide

blocks. First, we aimed at inferring both regulatory factors to be tested in vitro/in vivo and relevant regulatory networks. Moreover, we pointed to assess difference between up- vs. down-regulated gene 5'-flanking regions, as well as to propose a novel investigation method based upon detection of the combinatorial elements involved in gene regulatory control.
We did not address other regulatory mechanisms, such as epigenetics of chromatin organization –though very relevant to estrogen-mediated regulation– and the possible involvement of miRNAs, recently recognized as important regulators in other biological systems.

Chapter 2

Collection of estrogen-responsive genes

We describe and compare several genome-wide experiments from which we extracted the gene lists for subsequent analyses of 5'-flanking regions. We also define our core-promoters and upstream sequences.

2.1 Dataset1: Metaset

This dataset was obtained by collation of data taken from early microarray experiments, up to the year 2005 [17–20]. We build a list of all the genes which were down-regulated within 4h after estrogen stimulation in several immortalized breast cancer cell lines. In the end, the list size was 135. Each gene has been given a score which reflected the number of independent experiments collected and the experimental assessment of ERE presence. In particular, one point was assigned to a gene for each single experiment (expression, promoter analysis, etc.) in a specific cell line, and 10 points for Chip data. For example, when a gene was down-regulated in two different microarray experiments it was assigned score -2; when chIP data was also available the same gene got -12. A list of the same size for up-regulated genes was also prepared as a control, where each gene was assigned an experimental score (of opposite sign) with the same criterion employed in the case of down-regulated genes. The full list of genes is reported in Tables 2.1-

7. Genes regulated with intermediate (5-6h) and late (>6h) kinetics were also incorporated into the database for future reference, but have not been treated in this work. In this dataset, early genes are identified by Official Symbol, Gene ID and Ensemble! entry codes, and their experimental scores are marked by an A in order to distinguish them from intermediate (B) and late (C).

2.2 Dataset2: M.Brown

The experiment by Carroll ET AL. [22] provides the first map of all estrogen receptor and RNA polymerase II binding sites on a genomic scale in breast cancer cells (CF7), in combination with gene expression data obtained stimulating with Estrada for 0, 3, 6 and 12 h. Expression microarrays were *Aviatrix U133 plus 2.0* (over 47,000 transcripts), and the level of differential expression at each time point was calculated relative to 0h. The microarray employed for the Chip study was *Asymmetric human tiling 1.0* (the entire non-repetitive human genome sequence, NCBI build 35). For our analyses, we extracted the 3-hour dataset (expression data) and the list of genes which presented an ER-binding site (by Chip) in the flanking iatrogenic region from the supplementary files provided by the authors. We sorted the expression data into two lists, up- and down-regulated genes, which respectively contain 362 and 412 genes (RefSeq labels).

2.3 Dataset3: K.Nephew

Fan et al. [23] investigated into the acquired resistance to anti-estrogens tamoxifen and fulvestrant in breast cancer cells, through comparison of gene expression and DNA methylation profiles obtained in three breast cell lines: the MCF7 model and two of its drug-resistant derivatives (MCF7-T and MCF7-F). The microarray platform was *Affimetrix U133 plus 2.0* (over 47,000 transcripts), as in dataset2. We employed the data relevant to the treatment of cell models with estradiol (E2), considering all responsive subsets where the MCF7 wild type was involved (supplementary table S1 of Ref. [23]). Indeed,

many of these genes were expressed accordingly with different fold changes in the drug-resistant models as well. All in all, there are 202 up-regulated genes and 158 down-regulated genes (both RefSeq and Gene Symbol) in this dataset.

2.4 Dataset4: M. Rosenfeld

The authors aimed at determining *in vivo* binding profiles of transcription factors, and, employing a new technology specifically developed, they showed "an unprecedented number of the estrogen receptor (ERa) target genes in MCF-7 cells", and that "only a fraction of these ERa direct target genes were highly responsive to estrogen". The expression of this fraction of ER-bound, estrogen-inducible genes was associated with breast cancer progression in humans. In contrast to the others, this experiment [24] –the most recent one in our collection, was performed with Illumina (Human_ WG-6) as regard the expression data (ArrayExpress database, accession no E-MEXP-984). We extracted data at 3h with differential expression set to 50 both for up- and down-regulated genes and cut-off= 0.98. This cutoff –the lowest of the three suggested by the authors: 1, 0.99 and 0.98– represents the chance that the detected spot corresponds to a gene that is expressed, even at low level. Please, note that cut-off choice dramatically affects number of genes which are deemed expressed. In our lists, there are 389 up-regulated genes and 203 down-regulated genes. We added BMP7 to the latter list, which shows a modulation slightly below 0.98. So the size of the down-regulated set was 204 in the end. We also took data relevant to the chIP experiments performed in proximal region (508 genes). From the same laboratory, we also employed ChIP-DSL data which provides a list of 577 gene promoters ([-800,+200]) where ER is bound to [25]. The ChIP experiment was performed with a microarray containing probes to 22K human gene promoters in RefSeq. Only 54 genes out of 575 were also found to be regulated by estradiol in the same experiment [25].

2.5 Dataset5: M.Rae

Creighton et al. [26] generated gene expression profiles from three different estrogen receptor *a* -positive breast cancer cell lines (MCF7, T47D, and BT474) stimulated by estradiol *in vitro* over time (4h, 8h, 24h, 48h) and compared these profiles to the ones obtained from MCF7 cells grown as xenograph. They also compared their data to published clinical data [26], showing good overlap among *in vitro* and *in vivo* data, as well as published clinical results. They also indicated enrichment for transcriptional targets of the *myc* oncogene in the estrogen-responsive genes, and correlation with MYC expression in human tumors. They employed *Affymetrix U133A*, which contains almost 45,000 probe sets representing more than 39,000 transcripts derived from approximately 33,000 well-substantiated human genes (UniGene database; build 133, April 20, 2001); and *Affimetrix U133 plus 2.0* (over 47,000 transcripts). Then, they took the 22,283 probe sets shared between the two platforms in their analysis. They used a p-value < 0.01. For our analyses, we extracted from their supplementary excel file clusters A and B, e.g. the early (4h) up-regulated, and clusters E and F, e.g. the early (4h) down-regulated. Clusters A and E contains the genes that after early induction/repression return to the baseline within 24h, while the genes in clusters B and F showed sustained induction or repression through 24h. The authors examined each of the gene clusters for significantly enriched Gene Ontology annotation terms [16] and found that 1) there were no significant GO terms in the early up-regulated genes that return to baseline before 24 hours (cluster A; dataset5a) and in the early genes with sustained down-regulation (cluster F; dataset5b); 2) there were significant GO terms which include 'ribosomal function' and 'RNA and protein processing' in early sustained up-regulated genes (cluster B; dataset5b); 3) for the early down-regulated genes that return to the baseline before 24 hours (cluster F; dataset5a), significant GO terms involve 'transcription factor activity', 'development', and 'cell adhesion'. In the remaining gene clusters relevant to 8,12,24 hours, the author found enrichment of cell cycle-related terms, which is consistent with the observation that breast cancer cells stimulated with estrogen begin to divide and proliferate by 24 hours.

2.6 Summary

The panel below (Figure 2.1) summarizes key features of the datasets in our database. Dataset1 –a collation of data extracted from a variety of experiments published till the 2005, is collected in Tables 2.1-7. In Table 2.8, size of each dataset before sequence analyses.

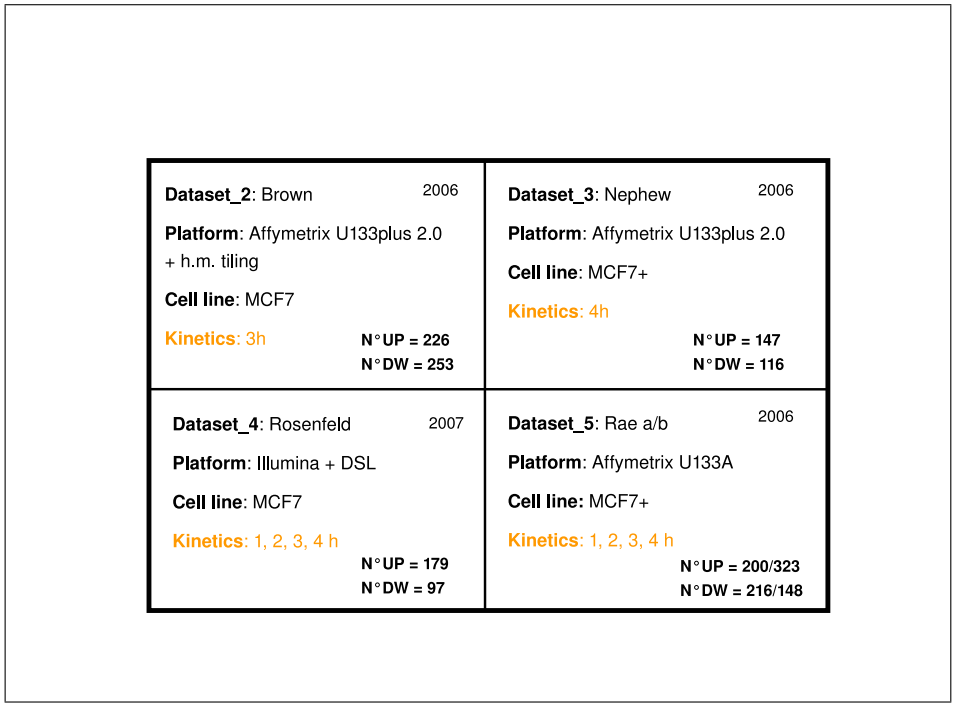

Figure 2.1: Features of datasets 2,3,4,5. MCF7+ indicates responsive genes in more than one cell lines/xenograph

Intersections: Down-regulated class We were interested in the most robust genes in the down-regulated class of the database. The relevant 9 pairwise-dataset intersections are collected in Table 2.9 below (metalist not directly included).

All datasets share the following three genes: **BMP7**, bone morphogenesis factor 7; **CCNG2**, cyclin G2; and **GTF2IRD1**, GTF2I repeat domain containing 1. The genes BMP7 and CCNG2 are also contained in the metalist (dataset1) where they represent the highest score entries. It is worth noticing

that dataset5a is complemental to dataset5b, and it does not contain BMP7 and CCNG2 by construction. None are found in the ER-bound promoter list of dataset4 [25].
In addition to the gene triad mentioned above, for dataset2 ∩ dataset3 ∩ dataset4 we observed the following subsets {EFNA1, FZD2, MXD4, ID3}; for dataset3 ∩ dataset4 ∩ dataset5b {BTG2, MARCKS, PPFIBP2, SOCS2, BIK}; for dataset2 ∩ dataset4 ∩ dataset5b {LMNA, RAB26, RXRA}. Among these genes, the following ones are in the metalist as well: BIK (-1A), BTG2 (-1A), EFNA1 (-1A), and RXRA (-2A). None of these genes are in the ER-bound promoter list of dataset4 (ChIP-DSL data) [25]. MXD4 (Max dimerization protein), ID3 (inhibitor of DNA binding), and RAB26 (member Ras oncogene family) are instead found in the list of ER-bound genes in distal region (dataset2) [22]. GTF2IRD1, GTF2I repeat domain containing 1, is also in this list. In Table 2.12, we gathered the down-regulated genes of our collection which are found in the ChIP-DSL data [25]; and the genes found in the collection of 660 pairs of orthologs which bear high affinity EREs within 2kb from TSS by Bourdeau et al. [19] are listed in Table 2.13.

Intersections: Up-regulated class The relevant 9 pairwise-dataset intersections are collected in Table 2.11-12 below (metalist not directly included). At least 9 up-regulated genes are in all datasets. They are as follows (HUGO IDs): AMD1, ASB13, CCND1, CXCL12, DDX21, IGFBP4, NRIP1, NP and RLN2. Surprisingly, TFF1 or pS2 – a gene whose induction by estradiol is employed as a positive control, is not contained in dataset3. None of these genes are in the ER-bound promoter list of dataset4 [25]; however, all but two genes (NP and RLN2) are found in the the list of ER-bound genes in distal region (dataset2) [22]. In Table 2.14, we gathered the up-regulated genes of our collection which are found in the ChIP-DSL data [25]; and the genes found in the collection of 660 pairs of orthologs which bear high affinity EREs within 2kb by Bourdeau et al. [19] are listed in Table 2.15.

2.7 Discussion

The four experiments, e.g. dataset2,3,4,5, overlap to a very limited extent, especially in the case of down-regulated class where only three genes appear in all lists. Even in the class of the more populated up-regulated class, we find less than 10 genes in all lists. Comparison between datasets from different genome-wide experiments are complicated by several facts inherent to the experimental techniques, as discussed in Ref. [27] –a review of strengths and weakness of genome-wide approaches for identification of nuclear receptor target genes. There is no standard way to interpret microarray data, and variability is linked to experimental protocols, intrinsic stochastic noise in gene expression, as well as experimental model. However, this should not suggest that microarrays data is not valuable. Indeed, even though genes may be different, the functions they carry out are consistent. The limited overlap of gene lists may be partially due to the fact that we chose to take the data as they were published; an alternative procedure could have started with the re-normalization of raw data before comparison and sequence analyses, but the shortcomings mentioned above would not be avoided even in this case. Several genes in both regulatory classes, anyway, appear in more than one datasets and certainly are robust entries.

The overlap between, on the one hand, the gene lists from the five datasets and, on the other hand, both ChIP data by Kwon et al. [25] and list of orthologs which bear putative EREs according to Bourdeau et al. [19] is limited to about a tenth of genes per dataset. The intersection between ChIP [25] and in-silico data [19] contains 3 genes from the down-regulated class (ANKRD2, CITED2, NR1D1), and 6 genes from the up-regulated class (CASP7, CYP1B1, FOXC1, HNRPDL, IFRD1, TPBG). These 9 genes are indicated in bold in Table 2.13 and Table 2.15. This suggests that the genes in our database –a subpopulation of the estrogen-responsive genes in breast cancer cells– might be influenced by indirect mechanisms (e.g. E2-ER interacting with DNA-bound proteins, non-genomic pathways and secondary response). It is worth noticing that the bioinformatics screening by Bourdeau et al. was performed on a old version of human and mouse genomes (h.s. NCBI33, m.m. NCBI30), while the experiments refer to new versions (h.s.

NCBI35). Moreover, given the limited number of genes (respectively 577 and 660) in the ChIP-DSL and in-silico data lists, primary targets directly regulated by ER through ERE may be more than the few ones found.

2.8 A database called EREGLON

All of the lists are stored in a database, EREGLON, which is installed on a Linux SunV20Z server and managed through MySQL open software. Clinical and model data relevant to responsiveness to anti-estrogens, for example, can be added to the database and analyzed accordingly. A comparison between clinical and model data in diverse conditions shall provide valuable insight into breast cancer pathogenesis. EREGLON should become an integrated tool for storage, analysis and experimental design in the study of estrogen regulation. In addition to primary data such as gene lists/sequences relevant to diverse samples/experiments, it will contain secondary information obtained from sequence analyses. The fully integrated database may become a publicly accessible resource, useful for the identification of tumor markers and the design of new experiments. Sequence analyses and relevant pathway inference –the dynamic core of EREGLON, is performed by a bioinformatics pipeline described in Chapter 3.

GOS	GID	ensgID	experiments	scr
ACTN1	87	ENSG00000072110	ZR75(ma)	+1A
ADCY9	115	ENSG00000162104	ma	+1A
AKAP1	8165	ENSG00000121057	ZR75(ma)	+11A
AREG	374	ENSG00000109321	MCF7(ma)++	+1A
ARL3	403	ENSG00000138175	ZR75 (ma), MCF7 (ma)	+2A
ARMCX6	54470	ENSG00000198960	ZR75 (ma)	+1A
ASB13	79754	ENSG00000196372	MCF7 (ma),T47D, MDA-MB-436	+3A
ASS	445	ENSG00000130707	MCF7 (SAGE, North; ma)	+1A
B4GALT1	2683	ENSG00000086062	MCF7 (ma)	+1A
BCAT1	586	ENSG00000060982	T47D (ma)	+1A
BCL2	596	ENSG00000171791	MCF7 (ma)	+1A
BIRC5	332	ENSG00000089685	ZR75 (ma)	+1A
CALCR	799	ENSG00000004948	MCF7 (ma, PCR)	+1A
CAV1	857	ENSG00000105974	MCF7 (SAGE, North)	+1A
CBFA2T3	863	ENSG00000129993	ma	+1A
CCND1	595	ENSG00000110092	MCF7 (SAGE, North), ZR75 (ma), ZR75-1 (North, West,+)	+3A
CD44	960	ENSG00000026508	ZR75 (ma)	+1A
CDC6	990	ENSG00000094804	MCF7 (ma, PCR)	+1A
COL4A6	1288	ENSG00000197565	MCF7 (ma), T47D (ma)	+2A
CTSD	1509	ENSG00000117984	MCF7 (ma) ++; MCF7(EMSA)	+2A
CXCL12	6387	ENSG00000107562	ZR75 (ma), MCF7 (ma,PCR)	+2A
CYR61	3491	ENSG00000142871	MCF7 (in situ hyb, North, West) ++	+2A
E2IG2	51287	ENSG00000181924	MCF7 (SAGE, North)	+1A
E2IG5 (C3orf28)	26355	ENSG00000114023	MCF7 (SAGE, North)	+1A
EDG2	1902	ENSG00000198121	T47D (ma)	+1A
EGR1	1958	ENSG00000120738	ma	+1A
EGR3	1960	ENSG00000179388	MCF7 (ma, PCR)	+1A
EIF3S9	8662	ENSG00000106263	ZR75 (ma)	+1A
EIF5A	1984	ENSG00000132507	MCF7 (SAGE, North)	+1A
ELL2	22936	ENSG00000118985	MCF7 (PCR,ma), T47D (ma)	+2A
F12	2161	ENSG00000131187	ZR75 (ma)	+1A
FKBP4	2288	ENSG00000004478	MCF7 (SAGE, North), ZR75 (ma)	+2A
FLJ13611	80006	ENSG00000113597	T47D (ma)	+1A
FLJ36166	349152	ENSG00000170629	MCF7(ma), ZR75 (ma)	+2A
FOS	2353	ENSG00000170345	ZR75 (ma), MCF7 (ma, PCR)	+2A
FOXC1	2296	ENSG00000054598	ZR75 (ma)	+11A
FOXP1	27086	ENSG00000114861	MCF7 (ma), T47D (ma)	+2A
G6PD	2539	ENSG00000160211	ZR75 (ma) +	+1A
GADD45B	4616	ENSG00000099860	MCF7 (ma, PCR)	+1A
GAPDH	2597	ENSG00000111640	T47D (ma); MCF7 (EMSA)	+2A

Table 2.1: Metalists from metaset; up-/down-regulated +/- score; A= early. cont'd

GOS	GID	ensgID	experiments	scr
GLRB	2743	ENSG00000109738	MCF7 (ma)	+1A
GREB1	9687	ENSG00000196208	MCF7 (ma); MCF7 (EMSA)	+2A
H2AFZ	3015	ENSG00000164032	MCF7 (SAGE, North)	+1A
H3F3A	3020	ENSG00000163041	MCF7 (SAGE, North)	+1A
HBG2	3048	ENSG00000196565	T47D (ma)	+1A
HIP12 (HIP1R)	9026	ENSG00000130787	ZR75 (ma)	+1A
HMGB1	3146	ENSG00000189403	MCF7 (SAGE), ZR75 (ma)	+2A
HOXC5	3222	ENSG00000172789	ma	+1A
HOXC6	3223	ENSG00000197757	MCF7 (ma, PCR)	+1A
HS3ST3A1	9955	ENSG00000153976	MCF7 (ma),T47D, MDA-MB-436	+3A
HSPA1A	3303	ENSG00000204389	ZR75 (ma)	+1A
HSPA5	3309	ENSG00000044574	ZR75 (ma)	+1A
HSPA8	3312	ENSG00000109971	MCF7 (SAGE, North)	+1A
HSPCB	3326	ENSG00000096384	MCF7 (SAGE, North), ZR75 (ma)	+2A
HSPD1	3329	ENSG00000144381	MCF7 (SAGE, North), ZR75 (ma)	+2A
IGFBP4	3487	ENSG00000141753	ZR75 (ma), MCF7 (ma, PCR) +; MCF7(EMSA)	+3A
INHBB	3625	ENSG00000163083	MCF7 (SAGE, North)	+1A
ISG20	3669	ENSG00000172183	MCF7(ma) +	+1A
JUN	3725	ENSG00000177606	ZR75 (ma)	+1A
KCNAB1	7881	ENSG00000169282	MCF7 (ma), T47D (ma)	+2A
CRKL	1399	ENSG00000099942	MCF7(ChIP)	+10A
KLF10	7071	ENSG00000155090	ZR75 (ma)	+1A
LAMA3	3909	ENSG00000053747	ZR75 (ma)	+1A
LDHA	3939	ENSG00000134333	MCF7 (ma), T47D (ma)	+2A
LOC92017	92017	no entry	MCF7 (SAGE, North), ZR75 (ma)	+2A
LRRC49	54839	ENSG00000137821	MCF7 (ma), T47D (ma)	+2A
LTF	4057	ENSG00000012223	MCF7 (ma), T47D (ma)	+2A
MGC16121	84848	ENSG00000165705	MCF7(ma)	+1A
MYB	4602	ENSG00000118513	ZR75 (ma), MCF7 (nucl run-on trans anal., North, West)	+2A
MYBL2	4605	ENSG00000101057	ZR75 (ma)	+1A
MYC	4609	ENSG00000136997	ZR75 (ma)	+1A
MYH11	4629	ENSG00000133392	ZR75 (ma)	+1A
NME1	4830	ENSG00000011052	ZR75 (ma), MCF7 (SAGE, ?)	+2A
NOLC1	9221	ENSG00000166197	ZR75 (ma)	+1A
NRP1	8829	ENSG00000099250	ZR75 (ma)	+1A
NRP2	8828	ENSG00000118257	T47D (ma)	+1A
NTNG1	22854	ENSG00000162631	MCF7 (ma), T47D (ma)	+2A
OLFM1	10439	ENSG00000130558	MCF7(ma), ZR75 (ma)	+2A
OSTF1	26578	ENSG00000134996	MCF7 (ma), MCF7 (ma, PCR)	+2A
PA2G4	5036	ENSG00000170515	MCF7 (SAGE, North)	+1A

Table 2.2: Metalists from metaset; up-/down-regulated +/- score; A= early. cont'd

GOS	GID	ensgID	experiments	scr
PCNA	5111	ENSG00000132646	MCF7(ma)	+1A
PPP2CA	5515	ENSG00000113575	T47D (ma)	+1A
PTGER3	5733	ENSG00000050628	MCF7(SAGE,North)	+1A
PTMA	5757	ENSG00000187514	MCF7(el.phgelassay, north,west), MDA-MB-231(idem)	+12A
RAB31	11031	ENSG00000168461	MCF7(ma), MCF7 (ma)	+2A
RAN	5901	ENSG00000132341	MCF7(SAGE,North), ZR75	+2A
RARA	5914	ENSG00000131759	MCF7(ma,PCR), MCF7 (ma),T47D, MDA-MB-436	+3A
RASGRP1	10125	ENSG00000172575	ma	+1A
RBBP7	5931	ENSG00000102054	MCF7 (ma, PCR)	+1A
RERG	85004	ENSG00000134533	ZR75 (ma)	+1A
RET	5979	ENSG00000165731	ZR75 (ma), MCF7 (ma,PCR)	+2A
RFC5	5985	ENSG00000111445	ZR75 (ma)	+1A
RPL14	9045	ENSG00000188846	ZR75 (ma)	+1A
RPS3	6188	ENSG00000149273	ZR75 (ma)	+1A
RPSA	3921	ENSG00000168028	ZR75 (ma)	+1A
RRM1	6240	ENSG00000167325	ZR75 (ma)	+1A
RSAD2	91543	ENSG00000134321	ZR75 (ma)	+1A
RUSC1	23623	ENSG00000160753	ZR75 (ma)	+1A
SEPT2	4735	ENSG00000168385	ZR75 (ma)	+1A
SERPINB6	5269	ENSG00000124570	ZR75 (ma)	+1A
SFRS1	6426	ENSG00000136450	ZR75 (ma)	+11A
SFRS7	6432	ENSG00000115875	ZR75 (ma)	+1A
SKB1	10419	ENSG00000100462	ZR75 (ma)	+1A
SLC16A6	9120	ENSG00000108932	ZR75 (ma)	+1A
SLC1A2	6506	ENSG00000110436	MCF7 (ma), T47D (ma)	+2A
NRIP1	8204	ENSG00000180530	MCF7 (ma, PCR); MCF7(EMSA); MCF7(ChIP, PCR)	+12A
SLC26A2	1836	ENSG00000155850	ZR75 (ma)	+2A
SLC7A5	8140	ENSG00000103257	ZR75 (ma), MCF7 (ma, PCR)	+2A
SLC9A3R1	9368	ENSG00000109062	ZR75 (ma), MCF7 (ma)	+2A
SLK	9748	ENSG00000065613	ma	+1A
SNRPA	6626	ENSG00000077312	ZR75 (ma) +	+1A
SOCS3	9021	ENSG00000184557	ZR75 (ma)	+1A
SPRED1	161742	ENSG00000166068	MCF7 (ma), T47D (ma)	+2A
STAR	6770	ENSG00000147465	MCF7 (ma), T47D (ma)	+2A
STC2	8614	ENSG00000113739	MCF7 (ma), ZR75 (ma)	+12A
STK6	6790	ENSG00000087586	ZR75 (ma)	+1A
TFF1	7031	ENSG00000160182	ZR75(ma), MCF7 (SAGE, North) ++; MCF7(ChIP,PCR); MCF7(EMSA)	+13A
TGIF2	60436	ENSG00000137801	MCF7 (ma)	+1A
THBS1	7057	ENSG00000118707	MCF7 (ma)	+1A
TMF1	7110	ENSG00000144747	MCF7 (ma, PCR)	+1A

Table 2.3: Metalists from metaset; up-/down-regulated +/- score; A= early. cont'd

GOS	GID	ensgID	experiments	scr
TP53	7157	ENSG00000141510	ma	+1A
TPBG	7162	ENSG00000146242	ZR75(ma)	+11A
TPD52L1	7164	ENSG00000111907	ZR75 (ma), MCF7 (ma, PCR)	+12A
TSPAN5	10098	ENSG00000168785	MCF7 (ma), MCF7 (ma, PCR)	+2A
TXNIP	10628	ENSG00000117289	ZR75 (ma)	+1A
ZNF703	80139	ENSG00000183779	MCF7 (ma), ZR75 (ma)	+2A
ZNF9	7555	ENSG00000169714	MCF7(SAGE,North)	+1A
CYP1B1	1545	ENSG00000138061	T47D (ma)	+11A
FOXA1	3169	ENSG00000129514	MCF7(ChIP)	+11A
PTGES	9536	ENSG00000148344	MCF7(ma)	+1A
STS	412	ENSG00000101846	MCF7(ChIP)	+1A
NR0B2	8431	ENSG00000131910	MCF7(ChIP)	+1A
FEM1A	55527	ENSG00000141965	MCF7(ChIP)	+11A
CYP4F11	57834	ENSG00000171903	MCF7(ChIP)	+11A
RPS6KL1	83694	ENSG00000198208	MCF7(ChIP)	+1A
ABCC5	10057	ENSG00000114770	MCF7 (ChIP)+	-10A
ABCG2	9429	ENSG00000118777	MCF7 (ChIP)	-10A
ACHE	43	ENSG00000087085	ZR-75 (ma)	-1A
ACPL2	92370	ENSG00000155893	T-47D (ma)	-1A
ALOX12B	242	ENSG00000179477	T-47D (ma)	-1A
ANGPTL4	51129	ENSG00000167772	MCF7 (ma)	-1A
ANXA2	302	ENSG00000182718	ZR-75 (ma)	-1A
ARF4L	379	ENSG00000113966	MCF7 (ma)+	-1A
ARID5B	84159	ENSG00000150347	T-47D (ma)	-1A
ATP2A3	489	ENSG00000074370	MCF-7 (ma)+	-1A
ATP9A	10079	ENSG00000054793	ZR-75 (ma)	-1A
BAK1	578	ENSG00000030110	MCF-7 (ma)+	-1A
BCL3	602	ENSG00000069399	ChIP+	-10A
BIK	638	ENSG00000100290	MCF-7 (ma)++	-1A
BIRC4BP	54739	ENSG00000132530	ZR-75 (ma)	-1A
BLNK	29760	ENSG00000095585	MCF-7 (ma,RT-PCR)	-1A
BMP7	655	ENSG00000101144	ChIP; T-47D(ma)	-11A
BTG2	7832	ENSG00000159388	MCF-7 (ma,RT-PCR)++	-1A
C10ORF110	55853	no ensg	MCF-7 (ma)	-1A
C10ORF45 (FAM107B)	83641	no ensg	MCF-7 (ma)+	-1A
C13ORF10 (RBM26)	64062	ENSG00000139746	MCF-7 (ma)	-1A
C17ORF37	84299	ENSG00000141741	MCF-7 (ma)	-1A
CBLB	868	ENSG00000114423	MCF-7 (ma)+	-1A
CCNG2	901	ENSG00000138764	MCF-7 (ma)++; T-47D(ma); ZR75(ma); ChIP	-13A
CD7	924	ENSG00000173762	T-47D (ma)	-1A

Table 2.4: Metalists from metaset; up-/down-regulated +/- score; A= early. cont'd

GOS	GID	ensgID	experiments	scr
CDC42EP3	10602	ENSG00000163171	T-47D (ma)	-1A
CDC42EP4	23580	ENSG00000179604	T-47D(ma)	-1A
CDK6 (MGC59692)	1021	ENSG00000105810	MCF-7 (ma);T-47D (ma)	-2A
CDKN1A	1026	ENSG00000124762	MCF-7 (ma,RT-PCR)+	-1A
CEACAM6	4680	ENSG00000086548	MCF-7 (ma)	-1A
CENTG1	116986	ENSG00000135439	T-47D (ma)	-1A
CLDN4	1364	ENSG00000189143	ZR-75 (ma)+;+MCF7	-1A
COPA	1314	ENSG00000122218	MCF-7 (ma); ZR-75-1 (SAGE)	-2A
CPS1	1373	ENSG00000021826	MCF-7 (ma)	-1A
CRABP2	1382	ENSG00000143320	T-47D (ma)	-1A
CREBBP	1387	ENSG00000005339	MCF-7 (ma)	-1A
CRISP3	10321	ENSG00000096006	T-47D (ma)	-1A
CTBS	1486	ENSG00000117151	T-47D (ma)	-1A
CXCR4	7852	ENSG00000121966	MCF-7 (ma, RT-PCR)	-1A
DDIT4	54541	ENSG00000168209	MCF-7 (ma)++, ChIP+	-11A
DNER	92737	ENSG00000187957	T-47D (ma)	-1A
DUSP4	1846	ENSG00000120875	MCF-7 (ma, RT-PCR)	-1A
EEF1G	1937	ENSG00000186676	MCF-7 (ma)	-1A
EFEMP1	2202	ENSG00000115380	MCF-7 (ma)+	-1A
EFNA1	1942	ENSG00000169242	MCF-7 (ma)+	-1A
EGFL4	1954	ENSG00000105429	ZR-75 (ma)	-1A
ENC1	8507	ENSG00000171617	MCF-7 (ma)+	-1A
EPB41L5	57669	ENSG00000115109	MCF-7 (ma)	-1A
EPHA4	2043	ENSG00000116106	T-47D (ma); MCF-7 (ma)+	-2A
EPLIN	51474	ENSG00000050405	MCF-7 (ma)+	-1A
EPOR	2057	ENSG00000187266	MCF-7 (ma, RT-PCR)	-1A
ERBB2	2064	ENSG00000141736	MCF-7 (ma, PCR)+	-1A
ERBB3	2065	ENSG00000065361	MCF-7 (ma)+; ZR-75-1 (SAGE)	-2A
ERBP	30836	ENSG00000067334	ZR-75 (ma)	-1A
ERF	2077	ENSG00000120705	MCF-7 (ma)	-1A
F10	2159	ENSG00000126218	T-47D(ma)	-1A
FBN1	2200	ENSG00000166147	T-47D (ma)	-1A
FLJ11336	55346	ENSG00000176148	MCF-7 (ma)	-1A
FLJ14201	81539	ENSG00000111371	MCF-7 (ma)	-1A
FLJ14213	79899	ENSG00000135362	T-47D (ma)	-1A
FLJ21963	79611	ENSG00000111058	MCF-7 (ma)++	-1A
FVT1	2531	ENSG00000119537	MCF-7 (ma)	-1A
GRB10	2887	ENSG00000106070	MCF-7 (ma)	-1A
GTF2H2	2966	ENSG00000145736	ZR-75 (ma)	-1A
HBG1	3047	ENSG00000019655	T-47D (ma)	-1A

Table 2.5: Metalists from metaset; up-/down-regulated +/- score; A= early. cont'd

GOS	GID	ensgID	experiments	scr
HBP1	26959	ENSG00000105856	MCF-7 (ma), ChIP+	-11A
HBQ1	3049	ENSG00000086506	T-47D (ma)	-1A
HEATR1	55127	ENSG00000119285	MCF-7 (ma)	-1A
HGS	9146	ENSG00000185359	ZR-75 (ma)	-1A
HIG2	29923	ENSG00000135245	T-47D(ma)	-1A
HUWE1	10075	ENSG00000086758	MCF-7 (ma)	-1A
ID2	3398	ENSG00000115738	MCF-7 (ma)	-1A
IDI1	3422	ENSG00000067064	ZR-75 (ma)	-1A
IDS	3423	ENSG00000010404	ZR-75 (ma)	-1A
IFNB1	3456	ENSG00000171855	T-47D(ma)	-1A
IFT122 (WDR10)	55764	ENSG00000163913	MCF-7 (SAGE,)+	-1A
IGFBP5	3488	ENSG00000115461	ZR-75 (ma)+	-1A
IKIP	121457	ENSG00000166130	ZR-75 (ma)	-1A
IL1R1	3554	ENSG00000115594	MCF-7 (ma, RT-PCR)	-1A
IL1RL2	8808	ENSG00000115598	ZR-75 (ma)	-1A
IL4	3565	ENSG00000113520	MCF-7 (ma)	-1A
INHBA	3624	ENSG00000122641	T-47D(ma)	-1A
IRX5	10265	ENSG00000176842	T-47D(ma)	-1A
ISGF3G	10379	ENSG00000092098	ZR-75 (ma);++	-1A
KCNG1	3755	ENSG00000026559	T-47D (ma)	-1A
KIAA0492	57238	no entry	ZR-75 (ma)	-1A
KIFAP3 (SMAP)	22920	ENSG00000075945	T-47D (ma);EMSA	-2A
KLF6	1316	ENSG00000067082	MCF-7 (ma)	-1A
KRT23	25984	ENSG00000108244	T-47D (ma)	-1A
KYNU	8942	ENSG00000115919	MCF-7 (ma)+	-1A
LIM (PDLIM5)	10611	ENSG00000163110	MCF-7 (ma)	-1A
LIMK2	3985	ENSG00000182541	MCF-7 (ma)	-1A
LMCD1	29995	ENSG00000071282	T-47D (ma)	-1A
LOC440281	440281	no entry	ZR-75 (ma)	-1A
LOC441027 (FLJ12993)	441027	no entry	MCF-7 (ma)	-1A
MAP2K6	5608	ENSG00000108984	MCF-7 (ma,RT-PCR)	-1A
MFSD7	84179	ENSG00000169026	T-47D (ma)	-1A
MGC10500 (YPEL3)	83719	ENSG00000090238	MCF-7 (ma)+	-1A
MGC12335	84830	ENSG00000111863	ZR-75 (ma)	-1A
MKNK2	2872	ENSG00000099875	T-47D (ma)	-1A
MYO1B	4430	ENSG00000128641	MCF-7 (ma)+	-1A
N4BP3	23138	ENSG00000145911	MCF-7 (SAGE,); ZR-75-1 (SAGE)	-2A
NCOA2	10499	ENSG00000140396	MCF-7 (ma,RT-PCR)	-1A
NDRG1	10397	ENSG00000104419	T-47D (ma)+	-1A
NMA (BAMBI)	25805	ENSG00000095739	T-47D (ma)	-1A

Table 2.6: Metalists from metaset; up-/down-regulated +/- score; A= early. cont'd

GOS	GID	ensgID	experiments	scr
NMSE1	84419	ENSG00000166920	T-47D (ma)	-1A
NR2C1	7181	ENSG00000120798	MCF-7 (ma)	-1A
OPHN1	4983	ENSG00000079482	MCF-7 (ma)	-1A
PAFAH1B1	5048	ENSG00000007168	T-47D (ma)+	-1A
PCM1	5108	ENSG00000078674	MCF-7 (ma), ChIP+	-11A
PIK3R3	8503	ENSG00000117461	MCF-7 (ma)	-1A
PLEKHF2	79666	ENSG00000175895	MCF-7 (ma)++	-1A
PRKCBP1	23613	ENSG00000101040	MCF-7 (ma)+	-1A
PTPN12	5782	ENSG00000127947	MCF-7 (ma)	-1A
RAB9B	51209	ENSG00000123570	ZR-75 (ma)	-1A
REA (PHB2)	11331	no ensg	T-47D (ma)	-1A
RERE	473	ENSG00000142599	MCF-7 (ma)	-1A
RFPL2	10739	ENSG00000128253	T-47D (ma)	-1A
RHOB	388	ENSG00000143878	T-47D (ma)	-1A
RXRA	6256	ENSG00000186350	MCF-7 (ma)+; ZR-75-1 (PCR)	-2A
SERPINE1	5054	ENSG00000106366	T-47D(ma)	-1A
SIAT1 (STGGAL1)	6480	ENSG00000073849	T-47D (ma)	-1A
SLC12A2	6558	ENSG00000064651	MCF-7 (ma)+	-1A
SMAD6	4091	ENSG00000137834	MCF-7 (ma)+	-1A
SNK	10769	ENSG00000145632	T-47D (ma); +	-1A
SREC (SCARF1)	8578	ENSG00000074660	ZR-75 (ma)	-1A
SYTL2	54843	ENSG00000137501	MCF-7 (ma)+	-1A
TAPBP (TAPBPL)	55080	ENSG00000139192	MCF-7 (ma); ZR-75-1 (SAGE)	-2A
TNFSF10	8743	ENSG00000121858	MCF-7 (ma)	-1A
TRIM24 (TIF1)	8805	ENSG00000122779	MCF-7 (ma,RT-PCR)+	-1A
TRPS1	7227	ENSG00000104447	MCF-7 (ma)	-1A
TTC3	7267	ENSG00000182670	MCF-7 (ma)	-1A
ZFN217	7764	ENSG00000171940	MCF-7 (ma)+	-1A
ZFYVE26	23503	ENSG00000072121	ZR-75 (ma)	-1A
ZNF33A	7581	ENSG00000189180	MCF-7 (ma)	-1A

Table 2.7: Metalists from metaset; up(+) and down(-) regulated gene score; A=early. cont'd

dataset	repressed	activated
1	97	125
2	392	332
3	131	174
4	132	234
5a	298	270
5b	226	473

Table 2.8: Number of genes in each dataset, annotated in *Ensemble40!* (h.g. NCBI36)

2 ∩ 3	**4 ∩ 2**	**2 ∩ 5a**	**2 ∩ 5b**	**3 ∩ 5a**	**3 ∩ 5b**	**4 ∩ 3**	**4 ∩ 5a**	**4 ∩ 5b**
AMACR	BAK1	CDC42EP4	ABCG1	BTG2	ARL4A	BIK	BTG2	ACAA2
BMP7	BCL9L	GATA2	BAMBI	RAB4B	BIK	**BMP7**	PCDH1	ATP6V0A4
CCNG2	**BMP7**	OASL	**BMP7**	SOCS3	BLNK	BTG2	SRPK2	BIK
EFNA1	BTG1		CASP9	WNT6	**BMP7**	**CCNG2**		**BMP7**
FZD2	**CCNG2**		**CCNG2**		BTG2	CDKN2B		BTG2
GTF2IRD1	CGN		CHD3		**CCNG2**	CTGF		**CCNG2**
ID1	CTDSP2		CNP		GTF2IRD1	CYP1A1		DKK1
ID3	EFNA1		COL9A2		IL1R1	EFNA1		GTF2IRD1
MXD4	FNBP1		DDIT4		KCNJ3	ENC1		LMNA
PIK3R3	FZD2		ELF3		MARCKS	EVA1		MARCKS
SLC2A10	GPR30		ENTPD2		PIK3R3	FZD2		PLEKHF2
ZNF467	GTF2IRD1		GAS2L1		PNRC1	GTF2IRD1		PPFIBP2
	ID3		GRB7		PPFIBP2	ID2		RAB26
	JUB		GTF2IRD1		PSD3	ID3		RAB9A
	LMNA		GUSB		SOCS2	KYNU		RXRA
	MC1R		ITGA3		SPDEF	MARCKS		SOCS2
	MXD4		KIAA0247			MLLT3		TLE1
	RAB26		LIMK2			MXD4		
	RAB3D		LMNA			OXTR		
	RXRA		MB			PMP22		
	SELENBP1		MKNK2			PPFIBP2		
	SLC17A5		PHF1			RBMS1		
			PIAS3			REL		
			PIK3R3			SALL4		
			RAB26			SLC6A14		
			RPRM			SOCS2		
			RXRA			SYTL2		
			SCN1B			TACC1		
			SH3BP4			TNFRSF11B		
			SHC2			TP53INP1		
			SREBF1					
			SSH3					
			TBX2					
			TIMP3					
			TNFAIP1					
			ZDHHC7					
			ZNF580					

Table 2.9: Dataset intersections: down-regulated genes, hugo codes. BMP7 and CCNG2 are also contained in dataset1; GTF2IRD1, present in all homogeneous sets, is not contained in dataset1

2 ∩ 3	**4 ∩ 2**	**2 ∩ 5a**	**2 ∩ 5b**	**3 ∩ 5a**	**3 ∩ 5b**	**4 ∩ 3**	**4 ∩ 5a**	**4 ∩ 5b**
AKAP1	ABCE1	C10orf22	ABCE1	ADCY9	CBFA2T3	ADCY9	ABCA3	ABCE1
AMD1	AHSA1	CEBPZ	ABHD2	AMD1	FUT4	AKAP1	CALM1	ADCY9
ASB13	AK3	DOCK5	AK3	ANXA9	KLF4	AMD1	CALM1	AK3
CA12	AMD1	EEF1E1	AKAP1	ASB13	MICAL2	AP1G1	CALM1	AMD1
CCND1	ASB13	EIF1AX	AMD1	B4GALT1	MITF	ASB13	CAMKK2	AREG
CDCA7	CA12	GTPBP4	ASB13	C1QTNF6	PDLIM3	CA12	CBFA2T3	AREG
CTPS	CCND1	HNRPDL	BYSL	CA12	RAB31	CCND1	CDR2	ASB13
CXCL12	CXCL12	KLF4	CA12	CBFA2T3	RET	CSPG5	CYP1B1	C6orf66
DDX21	DDX21	MTR	CCND1	CCND1	SLC22A5	CTPS	DICER1	CA12
FOS	DDX55	RET	COPS8	CELSR2	TRAF3	CXCL12	DNAJB9	CAMKK2
FZD7	DKC1	RHOBTB3	CTPS	CHPT1		DDX10	E2F5	CCND1
HCK	DNAJB6	RNF138	CUEDC1	CXCL12		DDX21	GFOD1	CXCL12
HEY2	ELF1	SH3BP5	CXCL12	DDX10		EGR3	GGA2	DDX10
IGFBP4	FAM8A1	SLC20A1	CYCS	DDX21		FHL2	GTPBP4	DDX21
KLF4	FER1L3	SLC22A5	DDX21	FKBP4		FLNB	KLF4	DICER1
NP	FOS	SLC25A24	DEPDC6	FLNB		FUT4	MBD4	DKC1
NRIP1	GEMIN5	STK17A	DKC1	FOS		GADD45B	PEX11A	FER1L3
PKIB	GTPBP4	SYNCRIP	EEF1E1	FOXC1		HEY2	RAB31	FLNB
PMAIP1	HIF1A	TGFA	EIF1AX	GAB2		HLA-DRB1	RFP —	GADD45B
RET	IGFBP4	TNPO1	FER1L3	GADD45B		HSPB8	SFRS7	GEMIN4
RLN2	KLF4	TOMM20	GTPBP4	IGFBP4		IGFBP4	SH3BP5	GTPBP4
SGK	KLK11	ZNF161	HEY2	KCNK5		KCNK5	SLC20A1	IGFBP4
SIAH2	MKI67IP	ZNF239	HNRPAB	KCNK6		LRIG1	SLC22A5	KCNK5
SLC22A5	MYC		HPRT1	KLF4		MYBL1	TFB2M	KIAA0133
SLC26A2	NCOA4		IFRD1	NCOR2		NP	TRIM8	KIAA0182
SLC7A2	NOLC1		IGFBP4	NP		NRIP1	ZNF239	KIAA0690
SLC9A3R1	NP		LRP8	NRIP1		OLFM1		METTL1
TIPARP	NPY1R		LRPPRC	OLFM1		PADI3		MYB
TPD52L1	NR4A2		LRRFIP2	OPN3		PDLIM3		MYBBP1A
	NRIP1		MRPL39	PODXL		PMAIP1		MYC
	PAICS		MYC	PPP2R2C		PODXL		NOLC1
	PCP4		NCKAP1	RAB31		PRSS23		NP
	PHLDA1		NOLC1	RASGRP1		PTGES		NPY1R
	PLOD2		NP	RLN2		RAPGEFL1		NRIP1
	POLR1B		NPY1R	SIAH2		RASGRP1		NXT1

Table 2.10: Dataset intersections: most robust up-regulated genes, hugo codes.Cont'd.

2 ∩ 3	**4 ∩ 2**	**2 ∩ 5a**	**2 ∩ 5b**	**3 ∩ 5a**	**3 ∩ 5b**	**4 ∩ 3**	**4 ∩ 5a**	**4 ∩ 5b**
	PPAT		NRIP1	SLC22A5		RLN2		OLFM1
	PUS1		OXR1	SLC26A2		SIAH2		PAK1IP1
	RARA		PEO1	SLC3A2		SLC1A4		PLOD2
	RERG		PFDN2	SLC7A5		SLC26A2		PODXL
	RLN2		PLOD2	SLC9A3R1		SLC7A5		POGK
	RRS1		PMAIP1	SMOX		SLC9A3R1		POLG2
	SH3BP5		PPAT	SNX24		SNX24		POLR1C
	SIAH2		PPIF	SULT2B1		SVIL		POLS
	SLC20A1		PTK9	SVIL		THBS1		PPAT
	SLC22A5		PUS1	TPD52L1		TIAM1		PUS1
	SLC26A2		RARA	UNC119		TIPARP		RARA
	SLC9A3R1		RB1	WFS1		TMPRSS3		RASGRP1
	SRM		RHOBTB3	ZNF185		TPD52L1		RFP
	STC2		RLN2					RLN2
	TFF1		RRS1					RRS1
	TPD52L1		SEH1L					RUNX1
	UGCG		SFRS2					SDCCAG3
	XBP1		SIAH2					SFRS7
	ZNF239		SLC26A2					SIAH2
			SLC7A1					SLC26A2
			SLC9A3R1					SLC2A1
			STC2					SLC7A5
			STS					SLC9A3R1
			TFF1					SNAPC4
			TFRC					SNX24
			TIPARP					STC1
			TNPO1					STC2
			TOMM20					SVIL
			TPBG					TAF4B
			TPD52L1					TFF1
			UCHL5					TPD52L1
			UCK2					UGCG
			UGCG					WDR12
			WDR3					XBP1
			XBP1					ZNF259
			XPOT					

Table 2.11: Cont'd. Dataset intersections: most robust up-regulated genes, hugo codes.

dataset1	dataset2	dataset3	dataset4	dataset5a	dataset5b
ABCC5	ABCC5	GLRX	CRIP2	ANKRD2	BCL3
BCL3	ANXA6	IGSF3	ESR1	COPE	CITED2
HBP1	ANXA9	PTPN13	ZNF444	NR1D1	ELF3
PCM1	BAD	YPEL3		VRK3	FAM13A1
	ELF3			WDTC1	HBP1
	PPP1R13L				LTA4H
	TSC22D3				SNX27
	WBP2				TTC9
	ZFYVE1				USP31
					WDTC1

Table 2.12: Down-regulated genes which are found in the ChIP-DSL data [25]

dataset1	dataset2	dataset3	dataset4	dataset5a	dataset5b
CD7	BMF	CYP1A1(2)	ATP1B1	ADCY9	CASP9
CDC42EP4	CASP9	PDK4	AXUD1	**ANKRD2**	**CITED2**
	CDC42EP4	RAB27B	CYP1A1(2)	CDC42EP4	DMPK
	CDKN2C(2)	TGFB2	LMNA	CHRNE	LMNA
	CSNK1D		MATN2	CRHR1	MB
	DDT(2)		RARG(2)	DMPK	NUDT2
	DLG3			DRD4	SOX13(2)
	DPP7(2)			ECM1	SREBF1
	FADS3			FN3K	VAT1
	GALE			GRAP	
	GRHPR			HCK	
	HMGCL			HEYL	
	IKBKG(2)			IL18BP(2)	
	LCN2			KCNE1	
	LMNA			KCNIP2	
	MB			MSI1(3)	
	MVP(2)			NOS3	
	NDRG2			**NR1D1**(2)	
	PTK6			PLD2	
	SLC4A2			POU5F1(3)	
	SOX9			RGS11(2)	
	SPAG4			SH3BP1	
	SREBF1			SOX13(2)	
	UPK1A			TFR2	
				TP53	
				ZAP70	

Table 2.13: Down-regulated genes which are found in supplemental data by Bourdeau et al. [19]. Number between parenthesis indicates multiple EREs were found.

dataset1	dataset2	dataset3	dataset4	dataset5a	dataset5b
AKAP1	AGR2	ADORA1	ABCA3	ABCA3	AKAP1
CYP1B1	AHSA1	AKAP1	AHSA1	BICD2	DICER1
CYP4F11	AKAP1	ANXA9	ANXA9	C10orf22	HNRPA1
FEM1A	ARL6IP2	BRI3BP	BLVRB	COX11	IFRD1
FOXA1	C10orf22	C1QTNF6	C1QTNF6	CYP1B1	KIAA0664
FOXC1	CBX3	CHPT1	CASP7	DICER1	MAX
PTMA	HNRPDL	FOXC1	CHPT1	DNAJB9	MXI1
SFRS1	IFRD1	MAP3K4	CRKL	GARNL4	NMT1
STC2	PKIB	MCM6	CYP1B1	HNRPDL	PHB
TFF1	PUS1	MITF	DICER1	ITPR1	PUS1
TPBG	SLC7A2	PDZK1	DLG5	MAX	QTRTD1
TPD52L1	STC2	PKIB	DNAJB9	MITF	SFRS1
	TFF1	SLC25A25	ELF3	PEX1	STC2
	TGFA	SLC7A2	FOXC1	RBBP5	STCH
	TIPARP	TIPARP	MORF4L2	RBL1	TFF1
	TPBG	TPD52L1	PUS1	TGFA	TIPARP
	TPD52L1	USP31	RAB30	ZNF571	TPBG
			SEMA3B		TPD52L1
			SLC25A19		ZNF331
			STC2		
			TFF1		
			TPD52L1		
			WISP2		

Table 2.14: Up-regulated genes which are found in the ChIP-DSL data [25]

dataset1	**dataset2**	**dataset3**	**dataset4**	**dataset5a**	**dataset5b**
ADCY9	CHML	ADCY9	ADCY9	CBFA2T3	ADCY9
CAV1	DNAJB6	CBFA2T3	BIRC3	CDR2	ATP1A1(2)
CBFA2T3	GMFB	CDT1	**CASP7**(18)	CYP1A1(2)	DHODH(2)
CTSD	HCK	CTSD	CBFA2T3	CYP1B1(2)	HNRPAB
CYP1B1(2)	HNRPAB	FKBP4	CDR2	**HNRPD**L	IFRD1
EDG2(2)	HNRPDL	**FOXC1**	CYP1B1(2)	LPIN1	IRS1
FKBP4	HSPA8(2)	HCK	DIRAS1(2)	NDST2	KARS
FOXC1	**IFRD1**	KCNK5(2)	DNAJB6	NETO2	KCNK5(2)
G6PD(4)	MRPL15	LOXL4(2)	FKBP4	PSEN2(2)	KIAA0409
HSPA8(2)	NRIP1	NRIP1	FOXC1	RFP(3)	MTMR4
KCNAB1(2)	RAB18	OPN3	KCNK5(2)	SLC1A1	NRIP1
KCNIP2	SFRS2	P4HA2	NRIP1	SOX9	RFP(3)
LTF(2)	TPBG	TRAF3	OPN3	TRAF3	RPA1
TP53		UNC119	RFP(3)	USP14	SDF2L1(2)
TPBG			SLC12A4	ZNF142(2)	SFRS2
			TST		SLC25A12(2)
			UNC119		TFDP1
					TPBG
					ZNF142(2)

Table 2.15: Up-regulated genes which are found in supplemental data by Bourdeau et al. [19]. Number between parenthesis indicates multiple EREs were found.

Chapter 3

The pipeline of DNA sequence analyses

We describe our method and discuss its advantages and its limitations.

3.1 Introduction

As described in Chapter 1, ChIP-on-Chip technique is a powerful tool for defining transcription factor binding site profiles that has helped shed light into the ERE-mapping issue. Generally speaking, ChIP-on-chip data is noisy and incomplete due to the inadequate resolution of whole genomic arrays and small number of transcription factors employed, especially in mammalian studies (see [27] for a review of strengths and weakness of genome-wide approaches for identification of nuclear receptor target genes). It requires to be complemented by bioinformatics methods in order to identifying cis-regulatory elements on genome-wide scale. There are two major bioinformatics ways to address transcription factor profiling. One relies on determining evolutionary conserved motif sequences through phylogenetic footprinting (see [28]); the other one points to determining shared or similar *motifs* between co-expressed genes, assuming that these genes would be co-regulated as well. The motifs searched can be either previously characterized, as in the case of matrix-based methods; or entirely novel, as in the case of enumeration algorithms. Bioinformatics approaches are flawed by several limitations as

much as the experimental genome-wide counterparts. For a general account of these problems, as well as of the state-of-the art, see references [28–35].
No significant difference in the number of estrogen responsive elements (EREs) between up- and down-regulated genes emerged in a pilot study of ours, and other computational studies proved ineffective as well (Pavesi's and Katznellenbogen's personal communications). This reflects the underlying complexity of the regulation in breast cancer cells, and, more in general, of the gene regulation in eukaryotes. As a matter of fact, transcription factor binding sites (TFBSs) may be highly degenerated motifs, e.g. signal which are very difficult to detect with the computational tools currently available. As soon as algorithm sensitivity is increased, the number of false positives explodes (see *futility theorem* in Ref. [28]). No surprise, thus, if it is so difficult to characterize DNA sequences using matrix-based algorithms when dealing with detection of highly degenerated, long TFBSs such as the EREs. Besides, transcription factor binding sites (TFBSs), are masked by chromatin –a factor that is not usually modeled by bioinformatics tool. It has been estimated that only a small fraction ($< 0.001\%$) of the whole number of putative TFBSs for retinoic acid receptor transcription factors in human genome ($50\text{x}10^6$) are actually functional, due to the impact of the chromatin dynamics on accessibility of transcription factor binding sites (see Ref. [37] for a detailed discussion).
We set up a bioinformatics pipeline which employs the two major approaches (enumeration/matrix-based algorithms and phylogenetic conservation) in order to explore the 5'-flanking regions of sets of genes that display homogeneous mode of regulation (co-regulons) by estrogen. The bioinformatics tools provide patterns of motifs and hint at putative transcription factors which may be involved in estrogenic regulation in conjuction with estrogen receptor. The *motifs* are nucleotide strings of various length ('words') which are statistically over-represented (see ahead) –e.g. whose actual occurrences are compared to to a background frequency. Denoting by $U(g)$ the length of the upstream region considered for a gene g and by $n(m, g)$ the number of occurrences of the motif m in such region and by $b(m)$ the length of m, the

background frequency of a motif m, $f(m)$, is defined as follows:

$$f(m) = \frac{\sum_g n(m,g)}{\sum_g u(g,m)} \tag{3.1}$$

where both sums are taken over all the specific genome under consideration, and where

$$u(g,m) \equiv U(g) - b(m) + 1 \tag{3.2}$$

is the number of words of length $b(m)$ that can be read in the upstream region of g.

3.2 Pipeline architecture

Our pipeline combines traditional approaches focused on sequence analysis of proximal regions with a method that allows for exploring distal conserved nucleotide blocks of sequences upstream of the gene. We investigated 5'-flanking regions of different sizes: (1) up to 15 Kbp from ATG (hg NCBI 36; Ensembl release 40); (2) [-800;+200] (hg NCBI 36; Ensembl release 40); (3) [-100,+100] (from RIKEN experiments); (4) [-450,+50] (hg NCBI 36; mg NCBI 37). Proximal regions (2), (3) and (4) are referred to as *core-promoters* in the following.

For core-promoter analyses, both a matrix algorithm and exhaustive enumeration tool were employed. The matrix-based algorithm directly identified and localized transcription factor binding sites in the core-promoters of co-regulated gene sets. Substrings of nucleotide sequences of different sizes, the so-called **motifs**, were obtained with the other tools both for proximal and distal regions. Each motif was associated with both a gene subset –that is the genes that contain it in their regulatory region, and a chromosomal position. Output motifs could match transcription factor binding sites contained in repository databases –in particular, TRANSFAC Professional [46]. Validation of best match candidates is being performed in laboratory.

In detail, we employed three different procedures, and crossed all of the output motif patterns afterwards. In order to explore the distal regulatory regions, we exploited one method based on Ab initio identification of DNA motifs by comparative genomics [60]. To investigate into proximal regions, we

employed both the software package Weeder [43] –an exhaustive enumeration algorithm; and, from the class of matrix-based algorithms, a novel tool [44]. The pipeline architecture is illustrated in Figure 3.1.

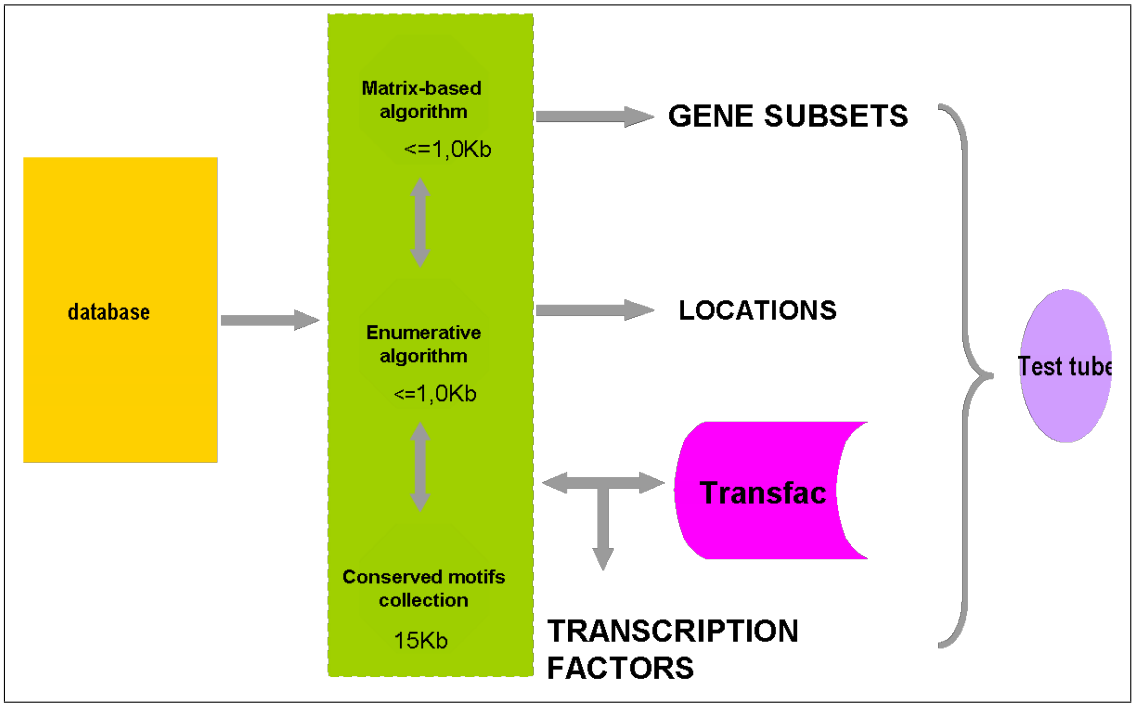

Figure 3.1: Pipeline architecture. Input: gene lists/regulatory regions. Output: motifs associated with gene subsets; motif chromosomal location. Motifs may match transcription factor binding sites contained in TRANSFAC database: validation of best matches is performed in laboratory

3.3 Input

Input of our pipeline comprised both gene lists and gene core-promoters of different sizes. The genes taken from the genome-wide experiments lists (see Chapter 2) were assigned ENSEMBL ID labels (ENSGIDs), in order to prepare them for extraction of conserved motifs. Not all genes, however, could be assigned one ENSGID, so that this procedure resulted in a partial resizing of the experimental lists. The 1000nt core-promoters ([-800,+200]) were extracted from ENSEMBL database [56] using the ENS-

GIDs. The 200nt core-promoters were extracted from the RIKEN database, using repeated BLAST alignments of small probes (Bajic promoters [38]) to extended sequences from the (entire) human genome, retaining the highest score matches only. As a result, some gene in the lists could not be assigned any [-100,+100] RIKEN core-promoter, while others obtained more than one 200nt core-promoter. Mammal genes, indeed, exhibit multiple Transcriptional Start Sites (TSSs) [39]. The matrix-based algorithm directly acts on 500nt promoters from the NCBI DNA sequence database (RefSeq IDs required).

3.4 Collection of conserved motifs

Our conserved motifs collection was extracted from a larger collection obtained with method by Corà et al. [55, 59, 60] with update to *Ensembl!40* human-mouse genome version. This method clusters genes from an input list, based on statistical over-representation of motifs from aligned 5'-flanking regions defined as follows: upstream sequence that extends up to 15Kbp (upper bound) from the start of translation of the longest transcript. So the over-represented motifs come from upstream sequence contained in conserved nucleotide blocks –the conserved non-coding sequence blocks, CNBs. The upstream regions could be shorter than 15Kbp, depending on the size of the intergenic region. In this procedure, some of the genes from the input lists have been filtered out due to 1) lack of ENSGID label, as already mentioned above; and 2) lack of relevant mouse ortholog gene. The database of conserved motifs is based on a universe of 12,381 genes.

Statistical over-representation of motifs in the distal conserved blocks was defined with respect to the background frequency to which the number of occurrences was compared. A motif m which occurs n times in the upstream region of a gene is considered to be over-represented if the probability for the motif to occur n or more times by chance is less than 0.01. This probability is computed with the right tail of a binomial distribution, assuming that motifs are randomly distributed, each with its background frequency [59].

The minimum score for an over-represented motif in output is 2.0 (score = 2.0 corresponds to p-value = 10^{-2}), and higher scores correspond to lower

p-values. All of the motifs obtained in output (a few hundreds depending on dataset and on regulatory class) have been crossed with core-promoter motifs. However, in the tables of Chapter 4, we only reported the motifs whose scores are higher or equal to a threshold which we have called *high-score motifs*: they were obtained arbitrarily setting a score threshold to 2.75 (a few tenths depending on dataset and on regulatory class).
Major limits of this method are as follows: 1) it only handles fixed motifs; 2) the analyzed sequence starts from ATG of the longest gene transcript, which occasionally does not guarantee it really extends in the upstream region of the gene; 3) it does not provide motif localization straightforwardly; 4) it is difficult to update when new genome releases become available; 5) when multiple transcripts are available, it only takes into account the longest one; 6) the region that is actually explored is reduced with respect to the original 15Kbp, due to repeats masking and alignment to mouse. Despite of these limitations –which determine a large number of false negatives, this method proved effective by studies both in yeast and human [59, 60]. Besides, to our knowledge is one of the very few available methods which enable distal exploration of DNA regulatory regions.

3.5 Collection of core-promoter motifs

Weeder [43] is the best performer in the class of the exhaustive enumeration approaches [35]. Weeder is a consensus-based method which enumerates exhaustively all of the motifs up to a maximum length and collects their occurrences with substitutions from input sequences in fasta format. Number of motif mutations allowed by the algorithm in each run increases with length of the motifs (1 for 6-mers, 2 for 8-mers, 2 for 10-mers and 3 for 12-mers). In addition to enumerating motifs, the algorithm provides the location in the specific sequences where these motifs were found.
Each motif is evaluated by Weeder according to the number of sequences in which it appears and on how well conserved it is in each sequence –in comparison to the expected values derived from the analysis of all of the upstream sequences of the same organism. A weight matrix, which is built using the discovered motifs, selects best instances of each motif. The top-

scoring motifs of each run are analyzed and compared in order to detect which one could be more likely to be a transcription factor binding site.
We applied Weeder to both strands of 1000nt core-promoters in all datasets, and also of 200nt core-promoters of dataset1. We retrieved the first 30 motifs for each of the 4 category in all datasets. This has produced a collection of 120 (30x4) motifs for the 1000nt core-promoters, and a collection of 80 (20x4) motifs for the 200nt core-promoters of dataset1. Best advice motifs for each dataset and regulatory class are reported in Chapter 4. We did not perform a similar study on mouse orthologs.

3.6 Transcription factor profiling

The advantage of using *Pscan* [44] relies on the fact that – in contrast to other matrix-based algorithms, it does not require the definition of a matching threshold and even the use of homologous sequences. As other matrix-based algorithms, it provides a straightforward way to identify putative transcription factors involved in the regulation of a gene set in input. The algorithm implements the following idea: the set of promoter sequences from co-regulated genes (co-regulons) is a sample of a larger population –e.g. the collection of all of the promoters from the genome of the same species; and the 'random' model is constituted by the whole set of promoters available for the species.
The highest scoring binding value of a matrix M on sequence Si in the sample, B(M,i), is a random variable whose mean value is compared to the one of the whole population through a z-test. The z-scores for matrix M given the sample P and the population S are instances of a random variable which follows a normal distribution; the significance of the difference between sample and population –that is, the associated p-values– is estimated by employing the normal cumulative distribution function.
Matrix-based algorithm are known to be affected by large number of false positives; after several tests, the authors suggest the outcomes should be considered reliable when their p-values stayed below 10^{-4}. When the up- and down- regulated classes were compared one to another directly, as for datasets 2,3,4, the p-value threshold was set to 0.01 (t-test). Pscan acts on core pro-

moters of size 500nt, and employed matrix profiles based on JASPAR [45] and TRANSFAC [46] databases. TRANSFAC profile contains many redundant matrices, while JASPAR contains a limited number of matrices.

3.7 Output

Motifs. The size of the sequences extracted is variable. The conserved motifs size always range between 5 to 8 nucleotides (nt); the core-promoter motifs may reach 12 nt, depending on parameter T; for example, with parameter T set to 30 they have length of 6 and 10 nt. Each motif was associated with both a gene subset and a chromosomal position. **Gene subsets**. Gene subset size varies from a few genes up to the entire collection in input, depending on both the algorithm employed and specific motif. Typical output size of gene subsets for conserved motifs is 3-7 –rarely above ten. Considering the procedure applied to core-promoters, number of associated sequences where a motif is found is higher –usually, a large amount that can be almost the size of the input set.**Transcription factor binding sites matching**. Output motifs could match transcription factor binding sites contained in repository databases –in particular, TRANSFAC Professional [46]. We mostly performed the matching on anecdotic basis, employing the program PATCH available within TRANSFAC database [46]. We also exploited the consensus tables published by Xie et al., 2005 [57].**Motif localizations**. The matrix-based algorithm, PSCAN, directly identified and localized transcription factor binding sites in core-promoters of co-regulated gene sets, and so does Weeder with its unknown motifs. Conserved motifs must be localized through an external procedure, such as for example *fuzznuc* within EMBOSS package [58]. Validation of the best matches is being performed in laboratory.

3.8 Assumptions and limits of the pipeline

As common bioinformatics tools, our pipeline handles the DNA sequences as linear strings where transcription binding sites are *signals* to be extracted from a *background noise* (see [30] for a brief historical perspective). There

are several levels of organization for the genetic material, from nucleosomes as beads on a string to more compact forms resulting from the regular organization of several nucleosomes together. Packaging of eukaryotic genomes must have a strong impact on gene regulation, and two studies have recently addressed – with somehow discordant results– the issue of nucleosome positioning determination by DNA sequence analysis [40, 41]. We have not attempted the dimensional perspective on a large scale for the moment.

Since co-expression does not always imply co-regulation, the set of sequences we employed in our study are 'noisy' by default –due to experimental artifacts for example. This may affect the identification of the transcription factor binding sites and, hence, of the relevant transcription factors which supposedly regulate the input genes. Our pipeline also assumes that transcription factor binding sites tend to be over-represented and to cluster in modules, with evolutionary conservation being the key to their functionality. The universality of the first two assumptions has been recently challenged [42], while the fact that site conservation does not guarantee functionality – at least in the case of estrogen responsive elements, has been often reported in literature (see [14] for example). Transcriptional cis-regulatory modules (CRMs) are more complex than what previously thought, with binding sites scattered over a large DNA segment, concentrated in one dense cluster, or even arranged in a composite way (for an orderly outlook, see fig.1 in ref. [42]). Our approach cannot obviously handle "composite signals", e.g. the ones that result from the combination of two distinct, distant signals –each statistically non over-represented.

These assumptions, though arguable, do not severely limit our analyses. We estimate that this study should provide a reliable description of regulatory regions of estrogen-responsive genes on first approximation, which is what is achievable at the moment (see final discussion in Chapter 5). With the progression of experimental techniques targeting chromosome territories, it shall become possible to include important information in the characterization of regulatory regions of co-regulated genes, such as their chromosomal localization. It should be also possible to take into account chromatin dynamics and long range interactions.

Chapter 4

DNA sequence analyses

We compare motif patterns within each dataset/regulatory class and between regulatory classes. We discuss a few instances of transcription factor binding sites identification/localization in detail. See Chapter 3 for terminology.

4.1 Dataset1: Metaset

4.1.1 Conserved motifs

We found 200 conserved motifs in the 15kbp upstream sequences of up-regulated genes and 177 motifs in the corresponding sequences of the down-regulated ones. The intersection of these two DNA motif sets comprises four words, as follows: AAGGTAGA, AGGGTG, ATAAG, and ATGCG motifs (and reverse). In the up (dw)-regulated list, 24 (22) motifs out of 200 (177) are statistically very significant (score threshold = 2.75; an arbitrary value: see Chapter 3 for an explanation). The high-score motifs for both regulatory classes are collected in Table 4.1. In both classes, 4 motifs are associated with scores higher than 4.00, while all the others have score comprised between 2.78 and 3.91 (up-regulated), and between 2.79 and 3.96 (down-regulated). Total number of these high-score motifs in the 5'-flanking regions of the up - and down-regulated genes is 45 and 42 respectively. No high-score motif is shared by the two lists.
Some of the output motifs differ by 2 or 3 mismatches; and, in principle, one would think it possible to extend the corresponding gene subsets, considering

these motifs as some variants of the same binding site. In order to illustrate an example of this, in the up-regulated class, the two following instances are provided: 1) motif GGCAAGGA, for genes ADCY9, NRP1, SPRED1, ZNF9, score 3.91; 2) motif GGGAACAA, for genes CRKL, NRP2, ELL2, score 3.08; where genes NRP1 and NRP2 are members of the same family, the neurophilin family of receptor proteins involved in versatile roles –among which angiogenesis, tumorigenesis and invasion. (almost all genes related to signal transduction–AMPcyclic and MAPras). In the down-regulated class, we observed, for example, the following associations: 1) motif GGTCGAAA, for genes ALOX12B, EFEMP1, PCM1, score 2.79; 2) motif CCTCGGAG, for ACHE, C6orf105, KCNG1, score 2.79. The safest motifs composition should be performed when the relevant gene subsets overlap, at least partially, that is when the same genes are present in subsets corresponding to the variant motifs. In this dataset, nonetheless, there are no such instances.

4.1.2 Core-promoter motifs

The exhaustive enumeration algorithm was applied to both 1000nt and 200nt core-promoters (see Chapter 3 for details). In this way, we obtained 120 motifs in each output (30x4). The entire set of motifs identified in both core-promoter sets is not shown.

The best advice motifs in 1000nt core-promoters are as follows: in down-regulated gene class, TCGTCGGG and reverse are highest ranking motifs (2 redundant motifs: GGTCGTCGGG - GTTCGTCGGG) in all but 19 sequences (data not shown); in up-regulated gene class, highest ranking motifs TCGCGCGT and reverse (9 redundant motifs: ACGCGCGAAC - CGACGCGC - CGCGCGTA - TCGAGCGG - CGCGCGTT - GCGCGTTA - TTCGCGCG - ACGCGCGT - GCGCGT) in all but 25 sequences (data not shown). The core of this motif, CGCGCG, may represent either a Sp1 binding site or a portion of CpG islands.

The best advice motifs in 200nt core-promoters are as follows: in the down-regulated gene class, highest-ranking GCGGAG and reverse (7 redundant motifs: GGGCGGAG - CTCCGCGC - GCTCCGCG - GGCGGA - GCGGTG - CGGAGC - TCCGCG) in all of the 60 sequences but 1 which cor-

responds to gene HEATR1; highest ranking are also TCGGCGGA and reverse (5 redundant motifs: TGCGCGGA - CCGCCGAG - TCGGCGGT - GGCGGA - CGCCGA). In up-regulated class, CGAGCG and reverse are highest ranking motifs (5 redundant ones: GCGAGCGG - GCCGCTCG - CGCGAGCG - AGCGCG - GCGAGC) in all of the 81 sequences but 5 (HSPA5, GADD45B, FLJ13611, RSADD2, and FOS).

4.1.3 Intersections of motifs

Crossing the motif patterns of up- and down-regulated gene classes, we obtained the collection reported in Table 4.2. We also crossed the motifs obtained from 1000nt and 200nt promoters within the same regulatory class; the result is collected in Table 4.3, where intersection of core-promoter and conserved motifs (partially discussed above) is reported as well.

As an example of shared motif analysis, we report the following: in the up-regulated class, the core-promoter motif *CGCGTT* is shared by all but eight 1000nt-gene flanking regions. NRIP1 and GREB1 –a gene found in the recent list of ER-bound promoter genes of dataset4 [25], are two of the eight up-regulated genes which do not hold this motif; their experimental scores are +1A and +2A respectively. This motif, *CGCGTT*, is also found in the conserved motif collection, as reverse complement AACGCG (score 2.1); it is found in the regulatory regions of three genes: **MYC**, HS3ST3A1, and CRKL, whose experimental scores are respectively +1A, +3A and +10A (see table Table 2.1). CRKL is also contained in the ChIP data list of dataset4 [25]. Additional interesting intersection motifs in the up-regulated class are as follows: *CTCGCG* (all but 5 sequences) and CC*CTCGCG* a conserved motif for 3 genes which are all in the larger set of sequences identified by the enumeration algorithm; GCGTCG (all but 5 1000nt-sequences) and GCGTCG-AA (conserved), 2 genes, **EGR1** and **B4GALT1**, also found within the larger set. In the down-regulated class, we observed a partial overlap between the *GAACGG* (1000nt core-promoters) and GAACGG-TA (conserved motifs, score 2.94).

In the up- and down-regulated gene sequences, both in the core promoter and conserved motif sets, different motif patterns are observed. This dataset

is the most heterogeneous in our database, which likely explains why the up-down intersection is more populated than in other datasets.

4.1.4 Transcription factor profiling

The study on human promoter of the genes in this dataset, where a direct comparison between up- and down-regulated classes was performed, no difference in transcription factor profiles emerged (data not shown). The p-values obtained could not guarantee the reported occurrences were not false positives; this is likely due to the fact that this set is highly inhomogeneous by construction.

With the matrix-based algorithm and the same two different sets of matrices applied to mouse ortholog core-promoters, we obtained the results summarized in table Table 4.4 (down-regulated class) and Table 4.5 (up-regulated lass). In the [-450,+50] core-promoter region of the dw-regulated genes, the two best matrices (TRANSFAC), p53 and REBP1, have p-values that do not guarantee absence of false positives (p-value more than 10^{-4}). The former is also identified by the corresponding JASPAR matrix, although with even less significant p-value. Nonetheless, p53 is a nuclear protein which regulates cell cycle; it is responsible for DNA repair and eventually for cell apoptosis. In the up-regulated class, p-values of some matrices are reasonably good (p-value less or equal to 10^{-4}). ZNF42 is the best score with JASPAR matrices; TAXCREB, MZF1, SPZ1, MAZR with TRANSFAC profile. SPZ1, MZF1, MAZ are reported in literature as transcription factors with binding sites enriched in 1,234 ChIP-PET clusters [12]. We underline that the matrix corresponding to the ERE binding site (ESR1, JASPAR) exhibited p-values 0.065 and 0.0033, respectively in the down/up-regulated classes.

4.2 Dataset2: M.Brown

4.2.1 Conserved motifs

There are 247 motifs in the up-regulated class and 254 the down-regulated one. The 50 high-score motifs found in the up-regulated gene upstream

regions and the 54 high-score motifs found in the down-regulated counterpart are collected in Table 4.6 and Table 4.7 (threshold score=2.75), where 3 motifs are associated to score higher than 4.00 in both regulatory classes. All of the others score between 2.76 and 3.95 (up-regulated) and 2.79 and 3.9 (down-regulated). Intersection of motif patterns from the two regulatory classes comprises one motif: ACGGCCCA. In the down-regulated gene class, the motif GCCCCCCCC, score: 3.57, is found in the regulatory regions of 5 genes; the motif GCCCCACC, score: 3.50, is found in 7 sequences. These two gene subsets do not overlap. We found the same result when we considered the gene subset (6 genes) associated with motif ACCGCCCC. Thus, it might be not plausible that these three motifs are variations of the same motif (by 1 and 2 mismatches). On the other hand, when considering the motif CGCCCCCA (score: 3.39) we found two genes, in a gene subset of 6, which also are in the gene subset associated with ACCGCCCC. These genes are as follows: CASP9 and OTTHUMT00000077092(ENSG00000162755).

4.2.2 Core-promoter motifs

The enumeration algorithm was applied to 1000nt core-promoters (349 down-regulated and 328 up-regulated), with parameter T set to 30. The 120 (30x4) motifs identified in both classes of genes are not shown; all of them have been crossed with the conserved motifs (see ahead). In this section we only report the best advice motifs. These are as follows: down-regulated class, (sense) GGTCCG and (antisense) CGGACC (10 redundant motifs: CTGGTCCG - ACGGACCC - GGGTCCGG - CGGGTCCG - GGTCGG - CCGGAC - CCCGGA - GTCCGC - GGACCG - GGACCG); (sense) CGCCATCG and (antisense) CGATGGCG (1 redundant motifs: CGACGTCG); up-regulated class, (sense) GTCGCG and (antisense) CGCGAC (21 redundant motifs: GTCGCGAG - ACGTCGCG - GTCGCGTC - CGTCGCGA - TCGTCGCG - CGTCGCGT - CGCGAT - CGCGGT - GCGACC - GCGCGA - CGCGAG - CGCGTA - TCCGCG - GACGCG - GACGCG - TCGCGG - CGTCGC - ACGCGA - TTCGCG - TTCGCG - AACGCG); (sense) CGACCGTTCG and (antisense) CGAACGGTCG (8 redundant motifs: CGTCGATTCG - CGGCGATTCG - CGACCTGTCG - CGTCGCTTCG - CGCACGTTCG -

CGACTGGTCG - CGAACGATCG - AACGGTCG).

4.2.3 Intersections of motifs

Intersecting the motif patterns of up-regulated and down-regulated classes, we obtained the results summarized in Table 4.8. There are no shared motifs between conserved and 1000nt core-promoter patterns in the down-regulated class; and only one motif, CGCACG, in the up-regulated one. This dataset contains more genes than other ones in our database and also exhibits abundance of motif patterns possibly presenting high redundancy.

4.2.4 Transcription factor profiling

The study on human core-promoters (same size:[-450,+50]), considering p-value less than 10^{-4} which may guarantee no false positives, highlighted the following:

down-regulated. With JASPAR matrix set, TFAP2A, NFKB1, ZNF42.1-4, SP1, ESR1, Pax5, and NFkB are the best scores; the overlap with mouse is good. Considering the TRANSFAC matrix set, SP1, ZNF219, MAZ, WT1, ZIC2, AHRARNT, ZIC1, MFZ1, VDR, ZIC3, SP3, and HES1. The overlap with mouse concerns SP1 only (see ahead).

up-regulated. With JASPAR matrix set, E2F1, CREB1, ELK4, ELK1, Arnt, GABPA. The overlap with mouse is good. With TRANSFAC matrix set, AHR, SP1, E4F1, ARNT, YY1, AHRARNT, STAT1, SP3, and WT1, and the overlap with mouse concerns ARNT. This matrix corresponds to the binding site of factor the aryl hydrocarbon (Ah) receptor.
Indeed, there is evidence that ARNT forms a complex with Sp1 and ER, which is required for positive regulation of cathepsin D [54]. Cathepsin D codes for a very important protease in estrogenic response and it is a tumor marker. Arnt matrix (JASPAR M0004; TRANSFAC ARNT.01) is overrepresented when the analysis is performed up vs. down with both matrix sets (but not when down vs. up was performed).
With the matrix-based algorithm and two different sets of matrices applied

to mouse orthologs (see Chapter 3), we obtained the results summarized in Table 4.9 (down-regulated), and Table 4.10 (up-regulated). In the 500nt core-promoter region of the dw-regulated genes, TFAP2A, Roaz, NFKB1, SP1 (JASPAR) have good p-values, while ESR1, the estrogen responsive element is borderline along with NFkB. This finding partially overlaps with the results obtained with the TRANSFAC set, which see SP1, AP2GAMMA, AP2ALPHA, EGR3, MAZR (a repressor) highly overrepresented. In the class of up-regulated, several matrices obtained good p-value: CREB1, Arnt, Arnt-Ahr, TFAP2A, Mycn, USF1, ELK4, GABPA, ELK1,NHLH1, MYC-MAX, MAX, E2F1 (JASPAR profile). Many matrices from the TRANSFAC set display very significant p-values, with very good overlap with former set (the MYC-related, AP2-related, CREB-related, E2F, etc.). Good statistics may be due to the high number of genes in each list of this dataset.

4.3 Dataset3: K.Nephew

4.3.1 Conserved motifs

Total number of motifs is 272 for the up-regulated genes and 205 for the down-regulated ones. As previously mentioned, the minimum score for a motif to be reported in the pipeline output is 2.0, while there is no score upper limit. The high-score motifs for both regulatory classes are collected in Table 4.11 and Table 4.12, where three motifs are associated to a score higher than 4 and all the others have score comprised between 2.76 and 3.83 (down) or 3.96 (up). Total number of these high-score motifs in the 5'-flanking regions of the up-regulated genes is 53. Total number of these high-score motifs in the 5'-flanking regions of the up-regulated genes is 43. No high-score motif is shared by the two lists. Actually, no overlap of conserved motifs for up- and down-regulated classes is found.

4.3.2 Core-promoter motifs

The enumeration algorithm was applied to 1000nt core-promoters, with parameter T set to 30. The 120 (30x4) motifs identified in both classes of genes

are not all shown here. Nonetheless, they have been all crossed with the conserved motifs. (best advice motifs)

4.3.3 Intersections of motifs

The intersection of the conserved motifs from the up-regulated and down-regulated is empty, while 8 motifs from the 1000nt promoters are shared between the two regulatory classes (Table 4.13). There are no motifs at the intersection between conserved and core-promoter motif patterns within each regulatory class.

4.3.4 Transcription factor profiling

In the study on human core-promoters, where the p-values are less than 10^{-4}, TFAP2 was overrepresented in the up-regulated class along with NHLH1, ZNF42, and Pax5 using JASPAR set; with TRANSFAC profile, ZNF219,SP1, WT1, HES1, USF2 and MAZ. In the down-regulated, JASPAR profile found: En1, Fos, Foxq1, FOXD1, Prrx2, SOX9, FOXF2, Sox17, and HLF; and TRANSFAC profile: SP1, ZNF219, MAZ, WT1, SP1, VDR, MZF1, SP3, AHRARNT, ZIC1, ZIC2, and HES1.
WT1 binding site was also found in dataset2; WT1 protein might directly interact with ER in repressing IGF-I receptor (IGF-IR), which has an important role in breast cancer development and progression [47]. However, this matrix is found overrepresented in both up and down- regulated lists. Indeed, comparing up- vs. down- regulated class, two matrices emerged: USF2 and YY1; while comparing dw- vs. up-regulated class, HLF and TITF1 (p-values less than 0.01, t-test). Anyways, WT1 is a tumor suppressor and an oncogene [48], and plays multiple roles in several types of cancer [49]. With JASPAR, direct comparison provided the follows: Arnt and Mycn (up- vs. down- regulated class) and Fos, En1, Foxq1, FoxD1, SOX9 (down- vs. up-regulated).
With the matrix-based algorithm and two different sets of matrices applied to mouse ortholog 500nt core-promoters (see Chapter 3), we obtained the results summarized in the following tables, Table 4.14 (up) and Table 4.15 (down). In the class of dw-regulated, ZNF42 and TFAP2A are best scores

with JASPAR set; with TRANSFAC profile, MZF1, AP2alpha and SP1. In the class of up-regulated, overrepresented JASPAR matrices are TFAP2A and RREB1; the list of TRANSFAC matrices is much longer: SP1, AP2, MAZR, SPZ1, RREB1, MZF1, AP2ALPHA, AP2GAMMA, EGR2, EGR3, and more.

4.4 Dataset4: M.Rosenfeld

4.4.1 Conserved motifs

Total number of motifs is 233 for the up-regulated genes and 214 for the down-regulated ones. As previously mentioned, the minimum score for a motif to be reported in the pipeline output is 2.0, while there is no score upper limit. The high-score motifs for the up-regulated class are collected in Table 4.16, where three motifs are associated to a score higher than 4 and all the others have score comprised between 2.76 and 3.87. Total number of these high-score motifs in the 5'-flanking regions of the up-regulated genes is 34. In the same table, Table 4.16, the high-score motifs for the down-regulated class are collected as well; three motifs are associated to a score higher than 4 and all the others have score comprised between 2.79 and 3.61. Total number of these high-score motifs in the 5'-flanking regions of the up-regulated genes is 49. No high-score motif is shared by the two lists.

4.4.2 Core-promoter motifs

The enumeration algorithm was applied to 1000nt core-promoters, with parameter T set to 30. The 120 (30x4) motifs identified in both classes of genes are not all shown here. Nonetheless, they have been all crossed with the conserved motifs (see ahead). (best advice motifs)

4.4.3 Intersections of motifs

The intersection of motifs across the two regulatory classes is empty; and the overlap of conserved and core-promoters motifs is minimal within each regulatory classes. In particular, three conserved motifs (**CAGCCC**, CCCAG,

CCCAGAC) were found in both regulatory classes. The first of the three has high score in the down-regulated class (score=3.48) and a medium score in the up-regulated class (score=2.53). In the former class, 7 genes are associated to the motif (IDH1, ENC1,TNFRSF11B, HOXC13, HEIS2, ZFP64), and in the latter one there are 8 genes (BCL6, KLF4, BATF, CBFA2T3, CEBPB, ZNF185, CXXC5, HSPB8). Mapping to consensus motifs in the table of Xie et al. 2005 [57](see Chapter 3), this motif could match either ER binding site (ERE) or CAC-BP (Sp1 and related). Localization of these conserved motifs in each of the 15 genes' flanking regions is the way to assess the nature of each of them. Localization can also guide experiments when the case is considered. Due to the total number of motifs, the size of the intersection between up-regulated and down-regulated classes of genes is essentially empty.
The intersection between the conserved and the 1000nt core-promoter motif patterns within each regulatory class is as follows: for upregulated, AACGCG; for downregulated, CCCGCA.

4.4.4 Transcription factor profiling

In the human 500nt core-promoters of up-regulated genes, MAX is the best matrix, followed by Arnt (3.09 10^{-10}), TFAP2A, Arnt-Ahr, Mycn, MYC-MAX, USF1 (JASPAR profile); with TRANSFAC profile: best matrices are SP1, AHRARNT, USF2; good matrices are HES1, AHR, ARNT, ZNF219, WT1, MAZ, and MYC. In the human 500nt core-promoters of down-regulated genes, we found: (JASPAR) TFAP2A, NHLH1, Pax5, SP1, E2F1, Arnt, NFKB1; (TRANSFAC) ARNT, SP1,ZNF219, MAZ, WT1. When we compared up- vs. down- regulated class, and viceversa, no matrix emerged as a distintive one.
With the matrix-based algorithm and two different sets of matrices applied to mouse orthologs, we obtained the results summarized in the following tables, Table 4.17 (down) and Table 4.18 (up). In the 500nt core-promoter region of the dw-regulated genes, JASPAR profile identified NFKB1, SP1; TRANSFAC finds SP1 and AP2. In the class of up-regulated, Arnt turns out with both JASPAR and TRANSFAC profiles, as in dataset2.

As pointed out above, this matrix (Arnt) corresponds to the binding site of factor the aryl hydrocarbon (Ah) receptor. Indeed, there is evidence that ARNT forms a complex with Sp1 and ER, which is required for positive regulation of cathepsin D [54]. Cathepsin D codes for an important protease in estrogenic response which is also a tumor marker.

4.5 Dataset5: M.Rae

4.5.1 Conserved motifs

This dataset is composed by two subsets: dataset5a which contains the early responders which do return to the baseline within 24h; and dataset5b which contains the early genes whose repression/induction is sustained throughout the 24h. Thus, dataset5a and dataset5b do not share genes by construction (in principle!).

dataset5a

Number of output motifs is 191 for the up-regulated genes and 257 for the down-regulated ones. Motifs with score higher than 2.75, for both regulatory classes are collected in Table 4.19 and Table 4.20. There are 5 and 7 motifs associated with score higher than 4.00 respectively in up- and down-regulated classes, while other scores range between 2.76 and 3.97 (down-regulated) and 2.90 and 3.78 (up-regulated). Total number of these high-score motifs in the 5'-flanking regions of the up-regulated genes is 41; and total number of these high-score motifs in the 5'-flanking regions of the down-regulated genes is 64. No motif is shared by the two lists.

dataset5b

Number of output motifs is 230 for the up-regulated genes and 249 for the down-regulated ones. The high-score motifs for the up- and down- regulated class are collected in Table 4.21, where 5 motifs in down-regulated class and 2 motifs in up-regulated class are associated with score higher than 4. Other motifs in the table are associated with score comprised between 2.76 and

3.96 for the up-regulated class, and 2.77 and 3.97 for the down-regulated one. Total number of these high-score motifs in the 5'-flanking regions of the genes in both regulatory classes is 50. Two motifs are shared by the two lists, ACACTCAG and CCCGGCCG –none of them are contained in table 4.21.

4.5.2 Core-promoter motifs

The best advice motifs (T=30) are as follows: from the down-regulated promoters in dataset5a, ATCGGGGG and CCCCCGAT (5 redundant motifs: GAATCGGGGG - TCGGGGGA - ATCCCCCG - AACGGGGG - CCCCGA); from the up-regulated promoters in dataset5a, (sense) GTCGCG and (antisense) CGCGAC (15 redundant motifs: ACGCGACG - TCGTCGCG - AACGCGAC - CGGTCGCG - TCGCGT - TCGCGC - CCGCGA - CGCGGT - CGCGTC - GCGACC - CGCGTT - CGTCGC - TTCGCG - TTCGCG - TACGCG). In all of 223 down-regulated sequences but 9 in dataset5b, the algorithm adviced the following motifs: (sense) CGTCCG and (antisense) CGGACG (6 redundant motifs: CGTCCGAG - GCGGACGA - TCCGCT - CCGTCC - GCGGAC - CGACCG).

4.5.3 Intersections of motifs

The 1000nt core-promoter motifs and the conserved ones from dataset5a and dataset5b have been crossed within each regulation class. Intersections of conserved motifs and of 1000nt core-promoter motifs in dataset5a were found empty. In dataset5b, instead, we found two motifs, CGCTTC and CGTCGCGC, in the up-regulated class; no motifs in the down-regulated class. In the same dataset, crossing 1000nt core promoter motif patterns of regulatory classes produced the following intersection: CGGCGT and CGACCG. (no table available).

4.5.4 Transcription factor profiling

With the matrix-based algorithm and two different sets of matrices applied to mouse orthologs (see Chapter 3), we obtained the results summarized in

Table 4.22. (study on human)

4.6 Meta-analyses

We looked for motifs shared by more than one dataset within each regulatory class, analyzing gene subsets where the motif was identified, and providing its chromosomal locations. We performed our meta-analyses in anecdotic way, although the pipeline provided us with more than one example worth investigating into (see tables Table 4.1-23).

4.6.1 Down-regulated class

GATA motif

There are 177 and 205 significant motifs in the down-regulated 15Kbp sequences of dataset1 and dataset3 respectively. Dataset1 and dataset3 share 13 genes (BIK, BLNK, BMP7, BTG2, CCNG2, CDC42EP3, EFNA1, ENC1, EPHA4, ID2, IL1R1, KYNU, PIK3R3, SMAD6, SYTL2, TRPS1). Some of the conserved motifs are common to both datasets and hold remarkably high statistical/biological significance. One of these motifs, AGATAAAA, strikingly resembles a binding site for GATA transcription factor (consensus WGATAR, Xie et al. 2005 [57]). Genes associated with AGATAAAA in the dataset1 are BMP7, CCNG2, TRPS1 and DDIT4; and in dataset3, BMP7, CCNG2, TRPS1, GPC4, and LIN7A. Three genes are contained in both subsets: BMP7, CCNG2 and TRPS1. Among them, BMP7 and CCNG2 possess very high experimental scores and are shared by all of the datasets (see Chapter 2). CCNG2 is a negative regulator of the cell cycle: when it is repressed the cells divides more rapidly. Same conclusion holds for the repression of BMP7 –a bone morphogenic factor that normally activates transduction pathways involving SMAD proteins, which in turn are inhibitory signals of proliferation. All of this is consistent with the effect of estrogen on breast cancer cells. TRPS1 is also known as atypical GATA protein GC79; TRPS1 overexpression is correlated with breast cancer. Alteration of TRPS1 is associated with a craniofacial malformation (tricho-rhino-phalangeal syndrome) and mild skeletal dysplasia. It is worth noticing that GATA binding motifs

have been found in both immediate upstream of TSS and distant enhancers of various genes crucially involved in development, cell proliferation and differentiation. None of these genes appear in the list of promoter regions where ERs are bound (ChIP-DSL data [25]). As regard the other genes associated with GATA motif, GPC4–a cell surface proteoglycan, may plays a role in the control of cell division and growth regulation, while LIN7A encodes a protein implicated in junction of the epithelial cells. DDIT4, DNA damage-induced, is involved in cell cycle and apoptosis. This gene is also found in the list of promoters where ER is bound to [25].
Common motifs shared by genes in dataset1 and dataset3 are collected in Table 4.23. A few genes in the table appear in more than one set. It is the case of ID2, which bears three different (non-redundant) motifs in its conserved upstream region; and of SMAD6, CCNG2 and EPHA4, which hold two different (non-redundant) motifs each. EPHA4 –a tyrosine kinase, ephrin A, which is a negative regulator of tumorigenesis in glioma (signal transduction); SMAD6 is involved in growth control, and is one component of the signal pathways controlled by the bone morphological factors (see above); so it makes sense that it results down-regulated here, being BMP7 down-regulated. ID2 is required for lobulo-alveolar differentiation of mammary gland.

Localization of GATA motifs We localized the motif AGATAAAA (chromosomal coordinates) in the conserved blocks of the upstream sequence of the relevant genes, by applying *fuzznuc* –a module of EMBOSS package [58]. Relevant data are collected in Table 4.24, where the genomic location, along with the one related to TSS (Ensembl!40), is provided for all of these sites in BMP7, CCNG2 and TRPS1 flanking regions. In the upstream region of CCNG2, we found three motif instances –all of them localized upstream more or less in a range of 5,000 bp from the TSS. In the flanking regions of BMP7 gene, we identified two GATA motifs in the same conserved block. In the case of TRPS1, GATA motifs are found in four different conserved blocks located in the second intron. This is an unusual outcome, for the ATG of TRPS1 is located in the 2nd exon and 1st and 2nd introns of this gene are extremely wide. A correction based on RIKEN data to these locations, in

the case of CCNG2, should be taken into account: the TSS shift is as follows: TSS bajc=TSS ens40 -130nt (which is the TSS in the current Ensemble version, by the way).
In the upstream sequence of DDIT4 gene, there were two (forward) motifs in two different short conserved blocks. In GPC4, there were two (reverse) motifs in two different large blocks; and in LIN7A, one (forward) motif in a short block. Localization of these motifs is not shown.

Co-localization ERE-GATA Using a consensus definition by Xie et al., 2005 [57], R-nnn-TGACCT, we could find a putative transcription factor binding site for estrogen receptor, or estrogen receptor element ERE, near to one of the GATA sites previously localized and reported in Table 4.24 for CCNG2 gene. This motif is the reverse of A-CCA-TGACCT (extended motif found: TT. AGGTCA-TGG-T.T), as shown in Figure 4.1. The GATA position related to TSS is -5443; the absolute distance between the two putative transcription factor binding sites, approximately 150nt, is compatible with a hypothesis of direct interaction between the ER and GATA transcription factors. A putative estrogen responsive element could be also found downstream one of GATA motif in DDIT4 flanking regions. These two motifs, GATA and ERE, are separated by 4 nucleotides, more or less the space for half helix turn (not shown). Our findings are being tested in laboratory (see conclusion and perspective also for a relevant biological hypothesis about mammary gland development). It is important to note that GATA binding sites have been repeatedly found in the surrounding of ERE sites in ChIP-on-Chip experiments [12, 22].

BMP7 and CCNG2: collection of conserved motifs and related gene subsets

BMP7 and CCNG2 genes are down-regulated by estradiol in all of the experimental datasets. Their functions are widely investigated, and many studies report their role in cancer/inflammation. In addition to this, they share at least one conserved motif, AGATAAAA, as outlined above. Here do we provide the entire collection of the conserved motifs extracted from their 5'-flanking regions, along with the gene subsets associated with them. In Table

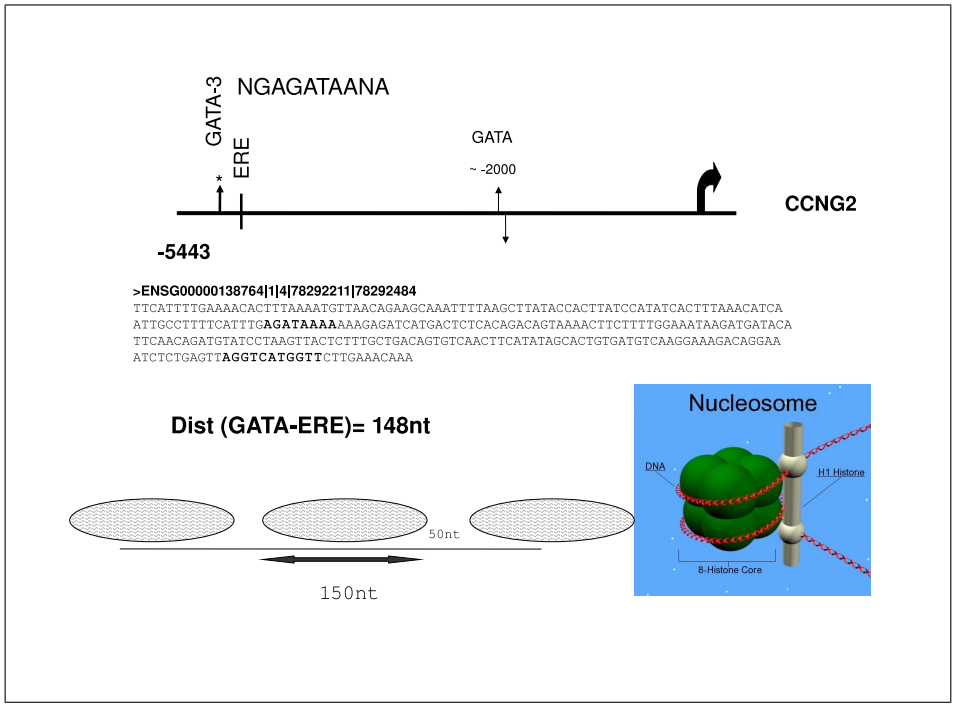

Figure 4.1: Localization of GATA motifs and ERE site in the regulatory region of CCNG2, cyclin G2

4.25 we gathered all motifs for BMP7. In Tables 4.30-34, we collected the motifs for CCNG2, dataset- by- dataset. Genes that are found in the ChIP data [25] are indicated in bold. These tables highlight associations with several interesting genes, among which GATA transcription factors. Specifically, there is GATA2 in dataset2 (CCGGC, score=3.62; GCCCC, score=2.18); and GATA3 in dataset4 (CGCCC, score=2.49). There is evidence CCNG2 is down-regulated by GATA-3 [61], and that a positive cross-regulatory loop links GATA3 and ERα and regulates their own expression in breast cancer cells [62].

We also noticed the presence of motifs shared by more than one dataset. For example, CCGAGC in both dataset1 (score=2.44) and dataset3 (score=2.24). In addition to CCNG2, the first gene subset (dataset1) comprises ACHE, IFT122, EPH4 (tyrosine kinase, ephrin A; differentiation), PLK2 (Polo-like kinase 2, Drosophila); and the second gene subsets (dataset3) EPHA4, VAV3 (Vav 3 oncogene), PLK2, PPAP2B (Phosphatidic acid phosphatase type 2B;

activates among others, PKC). VAV3 exchanges GTP/GDP for small-G proteins (especially Rho) of cytoskeleton in shape rearrangement and motility, and replication. PLK2 is a kinase serum-inducible which regulates the proliferative response. Thus EPHA4, PLK2 and CCNG2 form the subset core associated with CCGAGC motif which seems to be associated with signal transduction phosphorylative response. It might be possible that estrogen down-regulates these genes in order to maintain the ratio of co-activator/co-repressor concentration. An additional instance is taken from Table 4.27, where GATGTTTA motif (score 2.63) is associated with JDP2, CRTC2 and ZNF784, as well as CCNG2. JDP2, Jun dimerization protein, inhibits cell transformation– it's a cell survival protein; CRTC2 is a regulator of CREB (cAMP response) and it regulates recruitment of certain co-activators to some gene promoters in response to cAMP; and ZNF784 is a putative transcription factor.

CACCC-binding motif We investigated into some other motifs collected in Table 4.23, employing PATCH, an algorithm available within TRANSFAC Professional [46]. We set parameters as follows: minimum sequence length= 5, max number of mismatches 1, mismatch penalty 98, lower score boundary 87.5. CACCCTC could be associated with the CACCCC-binding factor site through the CACCC portion. Besides, CACCCC-binding factor is a transcription repressor. Typically, CACCC is a binding site for Sp1-4, and for KLF (Kruppel-like factors) which are proteins involved in growth control, apoptosis, and angiogenesis. One of them, KLF9, inhibits the estradiol transactivation in estrogen-responsive genes such as PR (progesteron receptor), whose promoter contains a half-ERE surrounded by CG-rich sites (Sp1) [63]
However, this motif may also match other matrices (for ex. Sp1, EKLF, FKLF, GLI2alpha; EGR-1 and EGR-2, important factors in growth control through reverse complement GGGTG site). The localization of this motif in the specific conserved sequence block of the relevant genes could help assess the nature of the putative binding site. The gene subset associated with CACCCTC in dataset1 comprises: ATP2A3, NDRG1 and IL1R1; the corresponding gene subset in dataset3 is composed by AFAP1, GAB1, EMP1, as

well as IL1R1.

Interleukin 1 receptor type I, **IL1R1**, encodes a cytokine receptor that belongs to the interleukin 1 receptor family. It is an important mediator involved in many cytokine-induced immune and inflammatory responses (inflammatory cytokine). Although IL1R1 is not in the list of the genes that have ERs bound in the proximal regulatory region (ChIP-DSL) [25], the response to IL-1 may affect the genomic response to estrogen. In particular, it controls the re-localization of NCoR co-repressors from nucleus to cytoplasm, in such a way that when IL1R1 is down-regulated the persistance of NCoR in the nucleus helps select those genes that must be up-regulated vs. those that must be down-regulated in response to estradiol [50]. Besides, IL1R1 is one of the genes whose expression is altered due to promoter methylation in MCF7-tamoxifen/raloxifene resistant cell lines [23]. In the regulatory region of interleukin 1 receptor type I, there are 2 instances of CACCCTC in 2 different, short conserved blocks. The ATG of IL1R1 is unusually located in the 3rd exon, and the first intron contains a bit less than 11,000 base pairs. As a result, the first motif (chromosomal location: 102,136,840) is located in the 2nd (untranslated) exon and the second one (chromosomal location: 102,137,335) is instead found within the 2nd intron. The TSS location of IL1R1 is 102,125,678 on chromosome 2.

GAB1, GRB2-associated protein, and AFAP1, actin filament associated protein 1, both contain binding sites for SH2 and SH3; the latter can effectively interact with SH2 and SH3, as well as with other non-receptor tyrosine kinases (signal transduction). ATP2A3, ATPase Ca++ transporting, is a membrane spanning protein found in sarco/endoplasmic reticulum. The ATPase refills Ca++-stores in endoplasmic reticulum and it might have a critical role here, for the cellular responses produced by the activation of some signalling pathways is production of IP3 which in turn mobilizes Ca++ from the endoplasmic stores. NDRG1, N-myc downstream regulated gene 1, is supposed to have a role in growth arrest and cell differentiation as a signaling protein shuttling between cytoplasm and nucleus.

Transcription factor profiling

With PSCAN, we investigated into DNA sequences of the genes which are down-regulated in all of the homogeneous datasets, and of the gene subsets associated with the GATA motif in both dataset1 and dataset3. Despite of the fact that the number of genes does not guarantee significant statistics for analyses of their core-promoters, we could observed that at least one matrix was overrepresented in both situations.

Set of most robust genes The three genes found in all of the homogeneous datasets (BMP7, CCNG2, GTF2IRD1) share the CDP CR1 protein transcription factor (Transfac matrix M00246), cut homeodomain protein, also known as cux or cutl. The matrix-based algorithm identified CDP factor with p-value equals to 0.005 as best output. It is interesting that the human cut homeodomain protein, CDP or CCAAT displacement protein, represses transcription from the c-myc promoter [51] and it is also associated with repression in H60 myeloid leukemia cells [52]. Besides, CDP expression is inversely correlated with survival in breast cancer [53].

Set of GATA-motif associated genes The seven genes associated with AGATAAAA motif share EGR-2 (M00246), early growth response 2, as best output.

4.6.2 Up-regulated class

CGAAACAC motif

This conserved motif is shared by up-regulated class of dataset1 and dataset3. In dataset1 (score 2.70), it is found in the flanking regions of 3 genes: HSPA5, SLC9A3R1 and CDC6; in dataset3 (score 2.10), the motif is detected in SLC9A3R1 and CDC6 flanking regions as well as in the one of PRKAG2. All of these genes are not in the list of ER-bound core-promoters [25]. The genes common to both subsets are as follows: SLC9A3R1 (experimental score +2A) is a solute carrier (family 9, isoform 3 regulator1), and CDC6 (experimental score +1A) is CDC6 cell division cycle 6 homolog (S. cerevisiae).

The other ones are the following: PRKAG2 (dataset3) is an AMP-activated protein kinase (gamma2 non-catalytic subunit); and HSPA5 (dataset1) is the heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa).
Allowing one mismatch, CGAAACAC motif could overlap the sequence CAAAACA, which corresponds to the insuline responsive element, IRS. Binding factors provided by TRANSFAC are several –the most striking seems to be either HNF-3alpha or FOXA1. The former, HNF-3, has consensus **TRTT-TRY**TYW according to Xie et al. [57]. The latter, FOXA1, breast cancer cell specific and capable of interaction with BRCA1, has been reported as a key factor by all authors of the large-scale study [12, 22]. FOXA1 is required for *nuclear clustering* of genes in response to estradiol [20, 65]. However, the Forkhead motif slogo found by Carroll et al. [22] is Aa*GxAAAcA*a, and the consensus according to Xie et al. [57] is W**AAA**A**YAAACA**ATM. Localization of the motif in the gene flanking regions can help filter the best match and better describe the hit.

Transcription factor profiling

Less than a tenth of up-regulated genes are shared by all of the datasets – once more a number not sufficient for good statistics employing Pscan algorithm. The genes are as follows: AMD1, ASB13, CCND1, CXCL12, DDX21, IGFBP4, NRIP1, NP, RLN2; and the matrix-based algorithm provided the following results (TRANSFAC profile): RP58 or ZNF238 (p-value=0.004), and HNF4 (pvalue=0.009).
We remind here of the matrix Arnt/ARNT, which turned out to be over-represented in up-regulated class of both dataset2 and dataset4. This matrix corresponds to the binding site of the aryl hydrocarbon (Ah) receptor (see FIG 4.4). ARNT forms a complex with Sp1 and ER, which is required for positive regulation of the tumor marker cathepsin D [54]. An additional piece of evidence–which suggests experiments to be performed on our genes, is that ligands of aryl hydrocarbon receptor may increase the occupancy of ER on certain promoters [64].

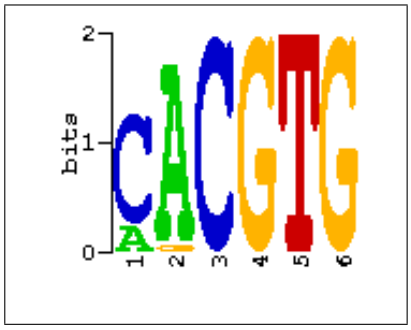

Figure 4.2: Slogo of matrix Arnt/ARNT, aryl hydrocarbon receptor.

4.7 Discussion

Presence of a distinctive conserved motif in the flanking regions of genes belonging to the same subset may indicate that these genes are co-regulated through the same biochemical pathway, and that they share some function/ontology. This seems to be reasonable, based on the instances in which the motifs did not look like CpG islands. However, the size of the gene subsets did not enable us to carry out a systematic ontology study. A way to handle this drawback may take advantage of clustering algorithms, which help in generating motif subgroups based on a similarity measure. Thus, each motif subgroup shall be associated with a consensus and a function to be validated in laboratory. This seems to be mandatory in order to exploit all information extracted from the sequences, for the conserved motifs might also represent matrix attachment domains (MARs) or other biological signals as important as transcription factor binding sites (TFBSs).

The overlap between conserved and core-promoter motifs depends upon the dataset taken under consideration, but generally speaking is not high. This was somehow expected due to the fact that the regions analyzed may not always overlap and that fixed motifs were employed. Besides, the core-promoter analysis was only performed on human promoters. The motifs so identified did not apparently match any of the TFBSs in TRANSFAC database. On the other hand, this inference generally proved very difficult.

The matrix-based algorithm helped identify a few interesting TBFSs, which suggested working hypothesis to be tested in laboratory. Similarity of outcomes in different datasets is not always cogent, likely due to the fact that

the overlap between datasets is not very high (see Chapter 2). These results seem to depend strongly on gene sets provided in input and on matrix profiles employed–which by the way contain redundant matrices at the moment. It could be interesting to apply the algorithm to restricted sets of genes, for example to the ones which are found in ChIP-DSL data [25], provided that the size of subset is sufficiently high to produce sensible statistics. This is certainly possible in the case of the up-regulated class (see Chapter 2, Table 2.10 and Table 2.13).

As seen in Chapter 2, very few genes (less than 10% in each list) in our database seem to bear estrogen responsive elements (EREs) in their flanking regions, according to both ChIP-DSL [25] and in-silico data [19] (Tables 2.12-15). Accordingly, there are few of these genes in the Tables 4.25-30, which contain all of the motifs and relevant associated gene subsets containing CCNG2 and BMP7 genes. As argued in Chapter 2, primary targets directly regulated by ER-ERE might be far more than the few ones we found crossing this published data –CCNG2, for example, which is not in ChIP data list, is directly down-regulated by ER-(half)ERE interaction [13]. Nevertheless, genes in our database –a subpopulation of the estrogen-responsive genes in breast cancer cells– seem to be strongly influenced by indirect mechanisms (e.g. E2-ER interacting with DNA-bound proteins, non-genomic pathways and secondary response).

motifs dw	SSS	score	motifs up	SSS	score
ATGAACCG	6	6.47	ATATGG	5	4.35
CGATCTAC	4	4.36	AAGTTCTG	4	4.23
CACACGC	5	4.14	ACGCGG	6	4.09
TAGGGCCA	4	4.02	CTGCGCTC	4	4.01
CCAAACAC	4	3.96	GGCAAGGA	4	3.91
CGTCGAAA	3	3.92	AGGGTG	6	3.86
ATCTG	6	3.92	ATGCGTCA	4	3.86
CAACAA	5	3.71	AAATGAC	4	3.77
CATTCGC	4	3.70	GGGAA	6	3.54
ATCCCGTG	4	3.70	ATGACTG	4	3.45
GACGAGTA	3	3.53	GAGCGCA	4	3.45
ACGTCGAA	3	3.44	ACGCGCG	4	3.45
CCGGAC	5	3.41	CTGGTTTA	3	3.36
AGTCCGG	4	3.31	AGTTCTG	4	3.35
AGATAAAA	4	3.31	CTTATC	4	3.28
TAATGCTA	3	3.30	CTGAATAA	3	3.19
CAAAGTA	4	3.27	GAGATTCA	3	3.13
AGACCG	4	3.23	ATCCA	5	3.10
CCCGGAC	4	3.20	ATCTGGCG	3	3.08
GTCGCCTA	4	3.20	CCTATGCG	3	3.08
GTATGCGA	3	3.16	GGGAACAA	3	3.08
GCGATCTA	3	3.04	AAAGTTGC	3	3.08
CCCGG	12	2.98	GCAACC	4	3.08
TGAGCAAA	3	2.94	CACGCG	5	3.01
GAACGGTA	3	2.94	GATGAGAA	3	2.99
ACCGTAAT	3	2.94	AAGCCCTT	3	2.99
CGCCATAC	3	2.89	ACGCG	7	2.97
ATACACTC	3	2.89	CGCCCAG	4	2.97
ACCTGCC	4	2.88	ACAATACG	3	2.94
CCCTAGGC	3	2.84	ATGGTGAA	3	2.94
AACCCTGC	3	2.84	ACTTCTGC	3	2.94
AGTCACTG	3	2.84	ACGCTTAT	3	2.94
CGGACTAA	3	2.84	ACACGGC	3	2.90
CTCGCA	4	2.82	GTACGTCA	3	2.90
CTATA	5	2.82	CTTACCTC	3	2.86
AACAGAGT	3	2.79	GATTCAAA	3	2.86
ATTACG	3	2.79	GTTCTGCA	3	2.86
GGTCGAAA	3	2.79	GCTACAGC	3	2.81
CCTCGGAG	3	2.79	AGTCGGTC	3	2.81
ACGAAAGT	3	2.79	AATGGTCG	3	2.81
GCTGGTCA	3	2.79	CTGTAAAA	3	2.81
GCTGAAAC	3	2.79	AATCTCTG	3	2.78
			GTGTTAA	3	2.78
			ATATCTGA	3	2.78
			ATCACTTC	3	2.78

Table 4.1: High-score conserved motifs from the 5'-flanking regions of genes in metaset. Gene set size: dw=104, up=89; SSS=subset size

region type	DW-UP motif
1000nt core	CGCGAA
	CGAACG
	ACTCGC
	TACGCG
	TCGCGCGT
	CGAACGGC
	CGCTCGCG
	GTCGAGCG
200nt core	TCGCCG
	CGCGGC
	CGGCGG
conserved	ATAAG
	ATGCG
	AGGGTG
	AAGGTAGA

Table 4.2: Motifs from all flanking region sizes which are shared by both regulatory classes in dataset1

type of intersection	motif dw	motif up
1000nt-200nt	GGCGGT	TCGCGG
	AGCGCC	CGACGC
	CGGAGACCGC	CGAGCG
	GCGGAGACCGCG	GCGCGT
		ACGCCG
		CGCGAACG
		GCGGCGTA
		GCTCGCGG
		CGACGCGC
		CGCTCGTCCG
		GCGAACGGGCGG
conserved-1000nt	CTCGCA	none
conserved-200nt	none	CGCCGCCGC

Table 4.3: Intersection motifs from regulatory regions of different sizes within each regultory class in dataset1

matrix	p-value
JASPAR	
TP53	0.00721277
ZNF42.5-13	0.00861142
TRANSFAC	
P53.01	0.00276853
SREBP1.02	0.00365697
SP1.Q6	0.00756064

Table 4.4: Best p-value matrices in dw-regulated class of dataset1 (mouse orthologs)

matrix	p-value
JASPAR	
ZNF42.5-13	0.000446069
ZNF42.1-4	0.00134352
CREB1	0.0129926
TRANSFAC	
TAXCREB.02	5.25778e-05
MZF1.01	7.42287e-05
SPZ1.01	8.98066e-05
MAZR.01	0.000577514
MZF1.02	0.000626964

Table 4.5: Best p-value matrices in up-regulated class of dataset1 (mouse orthologs)

motifs dw	SSS	score	motifs up	SSS	score
GGCCC	20	4.63	AACTCTAG	6	4.85
CCAGG	23	4.50	CAGTCTAA	5	4.21
CAGACAGG	6	4.03	ACCACGT	6	4.10
CGGGTATC	5	3.98	ATCATACA	5	3.95
CCCAAAG	6	3.96	CGAACC	6	3.70
AGTCAC	8	3.89	AAGCAA	8	3.70
AAGGAGG	7	3.82	ACCGCGGG	6	3.60
CCCAG	22	3.67	CGGTCTA	5	3.58
CTGGGCC	9	3.63	CTAACGAA	5	3.52
CCGGC	21	3.62	CACGTGCA	5	3.46
AGGGGCA	6	3.60	CGCCTAC	5	3.46
AGTTCGAC	4	3.59	CAAAAAC	5	3.46
ATAGGCCG	5	3.57	CGCACG	8	3.38
GCCCCCCC	5	3.57	CGCTCCTA	5	3.34
GCCCCACC	7	3.50	CGTGCGGC	5	3.34
ACCGCCCC	6	3.44	CGGGCTCA	5	3.34
GGTGTCAA	5	3.42	AGTCGCG	5	3.28
AACCATAC	5	3.42	GGCGGTAA	5	3.28
CGCCCCCA	6	3.39	GAGCGTCC	5	3.28
GTCTCTGA	5	3.36	GAGTCCGA	5	3.28
CGCCGTAA	5	3.36	CGGAACAG	5	3.17
CCCGAACC	5	3.36	CTCTTTAC	4	3.14
ATATCGC	5	3.29	TAACCCA	4	3.14
CGGCCC	13	3.25	ACGCCGCT	5	3.12
GCCCGGA	6	3.24	AACTTTA	6	3.08
ATTGAGGC	5	3.23	CAATTCAC	4	3.06
CCGTGA	6	3.20	CGAGGAGA	4	3.06
ACCTAACC	4	3.20	ATGCGATC	4	3.06
CCCGGC	14	3.17	ATAAATTC	5	3.02
CCGCCCC	12	3.12	TAGACCGA	4	2.99
AGGTCCC	6	3.02	TCAACGAA	4	2.99
TGGCAA	6	2.98	ACGTTGTA	4	2.99
CAGGACA	6	2.98	AGCAATCA	4	2.99

Table 4.6: High-score conserved motifs from the 5'-flanking regions of genes in dataset2. Gene set size: dw=226, up=253; SSS=subset size. cont'd on the following table.

motifs dw	SSS	score	motifs up	SSS	score
CGGGTCGA	4	2.95	GCGGAGAC	5	2.97
CCCTCACG	5	2.95	ACTATTGA	4	2.93
GCCTCAGC	5	2.95	AATTTTGA	5	2.93
ATTAAG	7	2.95	AACAGTA	5	2.93
GCGGGCC	7	2.95	GATTGA	5	2.88
GGAGGGA	8	2.94	ACTTAAAA	5	2.88
AGGCAG	9	2.91	CGGTAAA	4	2.87
CCGCCTCA	5	2.90	GTCAACGA	4	2.87
AGTCATAC	4	2.88	CCTCCGAA	4	2.87
CAAAGGGG	4	2.88	AACAAAAA	6	2.86
CCTGCGCG	6	2.86	CGGACTCC	5	2.84
GTGGCAA	5	2.85	ACGATCC	4	2.80
CCTGGCC	7	2.82	GATGAGCC	4	2.80
AGGACGA	4	2.81	GGCAAGA	5	2.80
AAACTGTG	4	2.81	CCGCACCG	5	2.80
AGTGGATC	4	2.81	CCCGACGC	5	2.76
CTTCCCAA	4	2.81	AACCCCGC	5	2.76
CGGCC	20	2.81			
TCCTGAAA	5	2.80			
CCACCGCC	5	2.80			
AGGGTC	6	2.79			

Table 4.7: cont'd from the previous table. High-score conserved motifs from the 5'-flanking regions of genes in dataset2. Gene set size: dw=226, up=253; SSS=subset size.

region type	DW-UP motif
1000nt core	CCGACG
	GACCGC
	CGCGTATCCG
conserved	ACGGCCCA

Table 4.8: Motifs from all flanking region sizes which are shared by both regulatory classes in dataset2.

matrix	p-value
JASPAR	
TFAP2A	7.61312e-06
Roaz	0.000147698
NFKB1	0.000800458
SP1	0.000999516
TRANSFAC	
SP1.Q6	8.85719e-08
AP2GAMMA.01	9.53851e-07
MAZR.01	5.42933e-06
AP2ALPHA.01	7.61312e-06
SP1.01	7.74592e-06
EGR3.01	0.000166892
AP2.Q6	0.000215851

Table 4.9: Best p-value matrices in dw-regulated class of dataset2 (mouse orthologs).

matrix	p-value
JASPAR	
CREB1	3.38008e-08
Arnt	1.18166e-06
Arnt-Ahr	6.67902e-06
TFAP2A	4.77396e-05
Mycn	9.29683e-05
USF1	0.000104534
ELK4	0.000320515
GABPA	0.000428103
ELK1	0.000496579
TRANSFAC	
ATF.01	1.29886e-08
E2F.03	2.40471e-08
AP2.Q6	3.59769e-08
PAX4.01	5.44881e-07
AHRARNT.02	2.48205e-06
NMYC.01	3.23477e-06
SP1.Q6	7.21341e-06
AP2GAMMA.01	2.28992e-05
EGR3.01	3.54873e-05
AP2ALPHA.01	4.77396e-05
CREB.02	7.40373e-05
EGR1.01	8.38691e-05
SPZ1.01	0.000145442
MAZR.01	0.000158069
TAXCREB.01	0.000234645
PAX5.01	0.00025087
MYCMAX.01	0.000362072
NRF2.01	0.000428103
CREB.Q2	0.000533569
CREBP1.Q2	0.000556606
AHRARNT.01	0.000806279
ATF6.01	0.000936263
ARNT.02	0.000937415
SP1.01	0.000956468

Table 4.10: Best p-value matrices in up-regulated class of dataset2 (mouse orthologs). cont'd on the following table

motifs dw	SSS	score	motifs up	SSS	score
AGATAAAA	6	5.50	AGGGTG	9	5.21
AGCCAGAA	5	4.59	CTGGACGA	5	4.47
ATACGATC	4	4.25	GCGCATAA	4	4.01
CCCTGGGA	5	3.96	ACATTCC	5	3.83
TGCTCAAA	4	3.90	GCGCCAAA	5	3.69
CAGAACAC	4	3.84	CATTTA	7	3.63
AGATCG	4	3.67	CCGCG	17	3.54
TCCTTCA	4	3.61	ATTCCATG	4	3.50
ATGTTGAC	4	3.56	CGCCAAA	5	3.40
ATTAAGGG	4	3.56	AAGCAAGC	4	3.33
ACGATCTC	4	3.56	GGGTAAC	4	3.33
ACAAGAGA	4	3.51	ATTCCACA	4	3.33
AGCGTTAA	4	3.47	CGGGGGA	5	3.29
GACTTCGC	4	3.42	ACGTACTC	4	3.28
ACATGTC	4	3.37	AGCAAGCC	4	3.28
ATACG	4	3.37	CCCCAA	6	3.18
ACACGCG	4	3.25	AATTATG	5	3.18
ACATGTCA	3	3.23	CCGTTAGG	4	3.17
ACAGGCCT	4	3.21	GTCATTTA	3	3.17
ATAACAT	4	3.13	CGCCAA	5	3.15
CAAAGTA	4	3.09	AATTC	7	3.12
AAAATT	9	3.07	AGCCTGCA	4	3.12
GTATGCGA	3	3.03	ATGCA	6	3.11
AACGTTCG	3	3.03	AGATTGA	4	3.08
CTAGGCGA	4	3.02	GAGCATC	4	3.08
GATAATGC	3	2.97	AAATC	7	3.06
GTTCCATA	3	2.91	CAATGAC	4	3.03
ACGGACTA	3	2.91	AATTAAAG	4	3.03
CAGGCCTG	4	2.89	CGACGTCC	4	3.03

Table 4.11: High-score conserved motifs from the 5'-flanking regions of genes in dataset3. Gene set size: dw = 116, up = 147 ; SSS=subset size. cont'd on the following table

motifs dw	SSS	score	motifs up	SSS	score
CACCCTC	4	2.85	GCGCA	13	3.01
GATCAA	4	2.85	CGCGCATA	3	3.01
AAAACCA	4	2.85	CATCCATG	3	3.01
ATCAACCG	3	2.85	GAGTGTCA	4	2.99
CGCCTATA	3	2.85	AACCCGC	4	2.99
CTCTACCC	3	2.80	CCCAG	14	2.91
ACGGCATA	3	2.80	TAGCAAA	4	2.90
ACACCTGA	3	2.80	GGCCCAA	4	2.90
ACCTCCC	4	2.79	AGAACCTG	4	2.90
ATGCAAA	4	2.79	CGCCGCAA	4	2.86
AATGC	5	2.79	ACGACCTA	3	2.86
AAATTC	5	2.79	AGGTCCTC	3	2.86
CGAGC	8	2.79	CATAATTA	3	2.86
GATAAAA	4	2.76	CCGCGC	10	2.85
			CACCCTG	5	2.84
			GCTGCAGA	4	2.82
			GCTGAAC	4	2.82
			CCGGTATA	3	2.79
			CGAAC	6	2.79
			AAGGCCAG	4	2.78
			AGCGTGA	4	2.78
			CACATTCC	4	2.78
			ATGGAAT	4	2.78
			CTGCAGC	5	2.76

Table 4.12: cont'd. High-score conserved motifs from the 5'-flanking regions of genes in dataset3. Gene set size: dw = 116, up = 147 ; SSS=subset size

region type	DW-UP motif
1000nt core	CGCTCG
	CGTTCG
	GCGGAC
	ACGCGCGT
	CGAACGGG
	CCGATCGC
	CGGACGACCG
	CGCTCGTCCG
conserved	none

Table 4.13: Motifs identified in both regulatory classes in dataset3.

matrix	p-value
JASPAR	
ZNF42.5-13	0.000687222
TRANSFAC	
MZF1.01	4.71939e-05
MZF1.02	0.000533595

Table 4.14: Best p-value matrices in down-regulated class of dataset3 (mouse orthologs)

matrix	p-value
JASPAR	
TFAP2A	0.000153662
RREB1	0.000207429
Roaz	0.000528005
SP1	0.000551663
TRANSFAC	
SP1.Q6	8.02992e-07
AP2.Q6	1.6064e-06
MAZR.01	2.83945e-06
SPZ1.01	3.7037e-05
RREB1.01	0.000108802
MZF1.02	0.000138269
AP2ALPHA.01	0.000153662
SP1.01	0.000268412
AP2GAMMA.01	0.000968987

Table 4.15: Best p-value matrices in up-regulated class of dataset3 (mouse orthologs)

motifs dw	SSS	score	motifs up	SSS	score
TGTAAACA	5	5.34	CCCAGAG	7	4.23
CCGGGGC	7	4.45	AAATCTCA	5	4.20
AACAACTA	4	4.33	ACTACG	6	4.10
GCCCC	12	3.61	CCCAG	18	3.87
AATAGT	5	3.55	ACTACGG	5	3.53
CAGCCC	7	3.48	CTAGACAC	4	3.45
CTTAAAAA	4	3.42	TGCCA	8	3.44
GCGCAAA	4	3.42	GTCATGGA	4	3.38
ACTTGCTA	3	3.32	TACAGCAA	4	3.31
GTAAAAAC	3	3.25	ACGTGCGC	5	3.30
CCCAGCC	6	3.21	CCGCGAGG	5	3.25
AGGAG	7	3.18	GCTCACAC	4	3.18
CGAGTTTA	3	3.13	TCTTGGCA	4	3.18
CTCCTA	4	3.11	CGGAAGGA	5	3.17
CCCCGCA	4	3.11	CCGGAAC	5	3.13
CCCAGC	7	3.10	ACCTGTGC	4	3.12
GTAAACA	4	3.08	TAAATTAA	5	3.09
AGTATCGA	3	3.08	GGCAAAAA	4	3.06
ACGAAAT	3	3.08	CGCAGACA	4	3.01
TGAAA	7	3.08	ACGCGCCC	4	3.01
AGGAGG	6	3.07	AAACCTGG	4	3.01
GGTACCGA	3	3.02	CAAGTCCC	4	3.01
AGCCCC	6	3.02	CAGAACCC	4	3.01
AGGGAG	6	2.99	ACTTCAAG	4	3.01
CAAATAA	4	2.94	AGGGTG	7	2.97
CCCGCA	5	2.93	AGAAACCT	4	2.96
ACAACTAA	3	2.92	CGCTTTAA	4	2.96
CGAACATA	3	2.92	GGATCTAA	4	2.90
GCTAGCA	3	2.92	ACCTCG	5	2.90
AGGGTACG	3	2.88	CACGTGGG	5	2.87
ATAATGTG	3	2.88	GAAAGTTC	4	2.86
GTACGTAA	3	2.88	AAGTCACC	4	2.86
AGCAAGCA	3	2.88	ACTACGA	4	2.76
AATCGTCC	3	2.88	CAAGCTAA	3	2.76
AGCCCCGG	3	2.88			
CGGATATA	3	2.88			
GTAAACAC	3	2.83			
GCCACTTA	3	2.83			
AATTAGGT	3	2.83			
AATAAGTC	3	2.83			
TACGTAA	3	2.83			
CCCTTTGA	3	2.83			
ATACACAT	3	2.83			
TACGTACA	3	2.83			
AATTGGGG	3	2.83			
AAAAG	7	2.83			
GTTTTA	5	2.82			
AAGTTGG	3	2.79			
ATCGCCTA	3	2.79			

Table 4.16: High-score conserved motifs from the 5'-flanking regions genes of dataset4. Gene set size: dw = 97, up = 179 ; SSS=subset size

matrix	p-value
JASPAR	
NFKB1	2.16947e-05
SP1	0.000263987
TRANSFAC	
SP1.Q6	7.9656e-06
AP2.Q6	2.37965e-05
SP1.01	3.03465e-05
E2F.03	0.000298879

Table 4.17: Best p-value matrices in dw-regulated gene list of dataset4 (mouse orthologs).

matrix	p-value
JASPAR	
TFAP2A	9.71496e-06
Arnt-Ahr	1.04919e-05
SP1	6.64661e-05
Arnt	0.000131393
TRANSFAC	
AP2ALPHA.01	9.71496e-06
AP2GAMMA.01	3.22254e-05
USF.Q6	3.2342e-05
SP1.01	3.90012e-05
MAZR.01	4.6936e-05
AP2.Q6	7.55904e-05
E2F.03	0.00016181
SP1.Q6	0.000352901
NMYC.01	0.000485396
AHR.01	0.000997355

Table 4.18: Best p-value matrices in up-regulated class of dataset4

motifs dw	SSS	score	motifs up	SSS	score
GGGGA	18	5.86	CACTTTC	7	5.03
AGGGG	17	5.32	ACCGCGA	7	4.84
CCCCCG	13	4.80	GGCCGTAA	6	4.74
GATGCAA	6	4.21	AAAAATCG	5	4.28
CCCCACCC	9	4.18	ACCGCGAG	6	4.10
CCCCGG	13	4.08	CGGTGTC	5	3.90
CCCCCCAC	6	4.03	ACCCCTAA	4	3.77
CCCCCGG	7	3.97	CTGCTTCA	5	3.76
CCCCA	19	3.93	CCAATCGC	6	3.65
CCATCCCC	6	3.92	AGATTGAT	5	3.58
AGCTCCCC	5	3.81	GCCGTCCC	5	3.47
ACAGCCC	6	3.81	CGAGCACG	5	3.47
GCCCCC	12	3.80	GAAAGTCC	4	3.41
GAGGATTA	4	3.75	CCGCC	26	3.37
CCCCC	17	3.70	CGGCAGGC	5	3.31
CACACACG	5	3.67	GATCCGTA	4	3.26
CCCCCACC	7	3.56	CCTTACTA	4	3.26
CACACGC	6	3.56	CGCCGACC	5	3.26
CCCTTTCC	5	3.55	CGCAGCCA	5	3.17
ACACACGC	5	3.55	GTTGACGA	4	3.13
CACCCC	11	3.52	ACCTTGTA	4	3.13
CTTTATAC	5	3.49	CAAGTTA	5	3.12
CACCCCC	8	3.45	ACCACTCT	3	3.09
GCAGGGAC	5	3.43	TACAAAGA	5	3.08
CCACTGCC	5	3.37	AAAGCTAA	4	3.06
CCCCGGGG	5	3.37	TCAAGAGA	4	3.00
GCCCC	19	3.34	ATCGTCAA	4	3.00
CAGAGC	8	3.34	CCACGGGA	4	3.00
GCAGAGC	6	3.26	CAATCGC	5	2.99
ATTCACC	5	3.21	CAGATGCA	4	2.89
AGGAGGGA	6	3.18	CAAAGATC	4	2.89
AGGGGC	9	3.18	AAGTGTAT	4	2.89
CAATGTA	5	3.16	GCCGCCGA	5	2.87
GAGCACA	5	3.11	CCGCCGA	5	2.87
AGGGGTCA	5	3.11	CGCCGCGA	5	2.87
GACAGGC	5	3.11	GAGCGGCA	4	2.83
GTTTGGA	5	3.11	CGTGAGAA	4	2.78
CGATCCTA	4	3.07	CGGGTGTC	4	2.78
GAACTGCC	4	3.07	ATACAAAG	4	2.78
CACAGCCC	5	3.06	AGGGCGAT	4	2.78
ACCCC	12	3.05	AGGGAATG	4	2.78
GGGACCC	6	3.04			

Table 4.19: High-score conserved motifs from the 5'-flanking regions of genes in dataset5a. Gene set size: dw = 216, up = 200; SSS=subset size. cont'd.

motifs dw	SSS	score
GTTTGCA	5	3.02
ACACACG	5	3.02
GGGACC	7	3.01
CTGTGCTC	4	2.94
CGATGTGC	4	2.94
ATTAGTTA	4	2.94
CCCCTC	10	2.89
GCTGAAA	5	2.88
GACTGATA	4	2.88
AGGTATGG	4	2.88
ACCCCCAC	5	2.84
GTGAGCCA	4	2.82
ATAAATCG	4	2.82
ACCTACTG	4	2.82
ACCACAAG	4	2.82
CAGGGAC	5	2.80
GGACCC	7	2.79
GGGAA	8	2.79
TCTGGGAA	5	2.76
CCCTGGCC	5	2.76
CGGGAACA	4	2.76
CAGCAAGC	4	2.76

Table 4.20: cont'd. High-score conserved motifs from the 5'-flanking regions of down-regulated genes in dataset5a. In this dataset, there are more motifs in the down-regulated class than in the up-regulated one. Gene set size: dw = 216, up = 200; SSS=subset size.

motifs dw	SSS	score	motifs up	SSS	score
TCCGATAA	5	4.75	CGCTCCA	8	4.66
CCCGG	18	4.50	AGGAGCA	8	4.26
CCGATGAA	5	4.38	CCGTTCGC	7	3.96
CCCCG	18	4.16	GTTCACA	6	3.81
CAGTCGGA	5	4.08	AAGCGTTC	6	3.74
CGGCACCA	5	3.97	CGCTTC	10	3.67
CGGCC	16	3.82	CCGTAGC	6	3.66
GCCCCCC	6	3.73	ATTTAGCG	6	3.66
CCGGCC	10	3.66	CGCGGGAA	8	3.66
CGCCCCCA	5	3.59	AAGTCGCC	6	3.52
AAAACAC	5	3.50	ACGTAC	6	3.52
AGAGGGCC	4	3.49	CGAACGGA	6	3.52
CCCCGC	12	3.49	CACTTGG	6	3.45
CGGCCGCC	6	3.48	AGCAGGGC	6	3.39
AGCCGATG	4	3.43	AAGTAGTA	6	3.39
GCCCC	15	3.39	GCGTCCGA	6	3.39
ACGGGATG	4	3.37	ACCGCAC	6	3.39
ATCGGGGG	4	3.37	AGGAAAGC	5	3.23
CCCGC	19	3.36	CGCTTCCG	9	3.19
GCCCGGAC	4	3.32	GTACTGTA	5	3.15
ACATGGAA	4	3.32	AGGACG	7	3.09
CCGGGAG	6	3.30	ATACGGAG	5	3.08
TATTTGAA	4	3.26	CCGAATAG	5	3.08
CCGATAAA	4	3.26	ATTTGCG	5	3.08
CCGCCCC	9	3.21	ACCTCTGA	5	3.08
CCCGGCC	7	3.21	CGCCAAG	5	3.08
ATGGAGTA	4	3.16	CTCCCGCA	5	3.01
ATCCCGTG	4	3.11	AGTCGC	6	2.98
CCATCAA	4	3.11	CACAAGAC	5	2.94
ATCGCCG	4	3.11	CCGCGAAA	5	2.94
AGCGTTAA	4	3.07	ACTCTGGA	5	2.94
GGGGCGGA	5	3.04	CGTCCTCC	5	2.94
GTAAGCA	4	3.02	AGTAGTAC	5	2.94
TAAGCAAA	4	3.02	CGGGAGTC	5	2.94
CCGCCC	12	2.96	TCACGGA	5	2.94
AAGGAAG	5	2.95	CCGACCG	6	2.93
ATCCGCGC	4	2.93	AGTCCCGC	7	2.91
AAGCGGAC	4	2.89	TCTGTGCA	5	2.88
GGAGGGA	6	2.86	CTCGCGAA	5	2.88
ATACGAGC	3	2.85	CTCGCTAA	5	2.88
CGAGAGTC	4	2.85	TAATGCA	5	2.88
TATTTCAA	4	2.81	CTAAGAC	5	2.88
GCCGATGA	4	2.81	ACCGTA	5	2.88
CGGGGC	10	2.79	AGGGCGAT	5	2.88
CTATATTC	3	2.78	GCGGGATA	5	2.82
CCAGCGGG	4	2.77	CCTTGGCG	5	2.82
CCGGACTC	4	2.77	AACACCAC	5	2.82
CCGGACCG	4	2.77	AAAACACC	5	2.82
CCGGGA	7	2.77	AAGTGGTA	4	2.80
ACCCCGCT	4	2.77	AAATCTCT	5	2.76

Table 4.21: High-score conserved motifs from the 5'-flanking regions of genes in dataset5b. Gene set size: dw = 148, up = 323; SSS=subset size.

profile	**dw5a**	**up5a**	**dw5b**	**up5b**
JASPAR	ZNF42	TFAP2A	ZNF42 (0.0033)	E2F1
		Arnt		ELK4
		Mycn		Arnt
				TFAP2A
	p300	AP2ALPHA	SP1	E2F
TRANSFAC	MAZR	AP2GAMMA	NFKAPPAB50	AHRRARNT
		AP2	AP2	AP2
		NMYC		SP1
		TAXCREB		NMYC
		SP1		EGR3
		NGFIC		AP2GAMMA
				AP2ALPHA

Table 4.22: Best matrices in dataset5a and dataset5b (mouse orthologs)

motifs	**genes**
CTCGC	SMAD6
AGACCG	EPHA4
CCGAGC	CCNG2,SNK
CAAAGTA	ID2, KYNU,EPHA4
CACCCTC	IL1R1
AGATAAAA	BMP7,TRPS1,CCNG2
GACTTCGC	none
GTATGCGA	ID2,ARID5B
GTCGTCGA	ID2,SMAD6

Table 4.23: Conserved motifs shared by down-regulated genes in dataset1 and dataset3.

gene	motif	C	T
BMP7 (-; chr20)	TTTTATCT	55,286,958	-11,867
	TTTTATCT	55,287,015	-11,924
CCNG2 (+; chr4)	AGATAAAA	78,292,308	-5,443
	AGATAAAA	78,295,722	-2,029
	TTTTATCT	78,295,784	-1,967
TRPS1 (-; chr8)	TTTTATCT	116,704,152	+184,787
	TTTTATCT	116,704,365	+184,574
	AGATAAAA	116,706,201	+184,738
	TTTTATCT	116,707,986	+184,953

Table 4.24: Chromosomal positions of conserved motif AGATAAAA. C: location in absolute coordinates of first nucleotide. T: location relative to TSSensemble.

dataset	motif	score	gene subset
Dataset1	AGATAAAA	3.31	BMP7,DDIT4, TRPS1,CCNG2
Dataset2	ACGTGAC	2.07	ERP29,BMP7,UBXD1, MTHFR,CCNG2
	AACCATAC	3.42	C190RF21,BMP7,OR10B1P, POR,PKIG
	ACCATACC	2.41	C190RF21,BMP7, OR10B1P,POR
	CAGAGGAA	2.27	BMP7,ST14, FAM83H,C7orf20
	CTGGGTGC	2.27	XRCC1,BMP7,JUNB, PHLDA3
Dataset3	AAATTC	2.79	BMP7,OXTR,LIN7A,TMEM45B, PSD3
	ATAAAA	2.11	BMP7,TRPS1,LIN7A,SPINK5,LYN,CCDC68, SLITRK6
	AAAACCA	2.85	BMP7,LGALS8,TM4SF1,MLLT3
	CTGTAAA	2.49	BMP7, WNT6,PPFIBP2
	GATAAAA	2.76	BMP7,ALAD,TMED4, GRB14
	ACAGGCCT	3.21	BMP7,ID1, CDK9,PPFIBP2
	ACCTGGCA	2.34	BMP7,NRP1,CTGF
	AGATAAAA	5.5	GPC4,BMP7,TMED4,TRPS1, LIN7A,CCNG2
	CAGGCCTG	2.89	BMP7,ID1,BLNK,TMEM45B
Dataset4	CAAAA	2.38	BMP7, LHX2, ID2, CN004. HUMAN, CT11. HUMAN,TMEM49
	GTTTTA	2.82	BMP7,LHX2,PQLC1,OXTR,SNX7
	AAAACCA	2.11	BMP7, TRAFD1,MLLT3
	ATAAAAC	2.73	BMP7,LHX2,PQLC1,SNX7
Dataset5b	AATGGCC	2.31	BMP7, CNP,RXRA
	AAAGTCGT	2.05	ACHE,BMP7,CD226
	ACCTGGCA	2.05	BMP7, MMP15,MYH14
	CAGGCCTG	2.5	BMP7,PHF1,BLNK,TCF2
	CGACACAA	2.12	BMP7, PHF1,CASKIN2
	TGTGAAAA	2.15	BMP7,ANXA5,TLE1

Table 4.25: Gene subsets associated with all conserved motifs found in the 5'-flanking region of gene BMP7

motif	score	gene subset
CGTCGAAA	3.92	CCNG2,RBM26,PDLIM5
AGATAAAA	3.31	BMP7,DDIT4,TRPS1,CCNG2
GTCGCCTA	3.2	KIFAP3,BAMBI,CCNG2,IFNB1
CGGACTAA	2.84	ARL6,PIK3R3,CCNG2
ATTACG	2.79	BAMBI,CCNG2,RERE
ATGCCGGA	2.62	ACHE,KRT23,CCNG2
CGGAC	2.53	ACHE,MKNK2,KLF6,ZFYVE26,PIK3R3,CCNG2
CCGAGC	2.44	ACHE,IFT122,EPHA4,CCNG2,PLK2
CGGGGCTC	2.25	BIK, **HBP1**,CCNG2
GGTAGAA	2.23	HUWE1,SLC38A1,CCNG2
CCCTCTGA	2.23	IL1R1,MYO1B,CCNG2
CCCGTCC	2.12	CREBBP,CBLB,CCNG2
CGGGGC	2.09	CREBBP,**BCL3**,MKNK2,KLF6,BAMBI,EPHA4,CCNG2
ATAAG	2.05	FLJ14213,PAFAH1B1,CPS1,CCNG2

Table 4.26: Gene subsets associated with all conserved motifs found in the 5́-flanking region of gene CCNG2: dataset1

motif	score	gene subset
CCGGC	3.62	NCAPD2,TLE2,RPRC1,TBX2,MXD4,PVRL2,BHLHB2,CSNK1D, N4BP3,TLCD1,SLC43A2,TMEM142C,GATA2,MAF1,RGAG4,LOC727800, BAMBI,PIK3R3,CCNG2,FAM84A,RPRM
CGCCGTAA	3.36	CTNNBL1,TBC1D2,BAMBI,LRP16,CCNG2
GCCTCAGC	2.95	TLE2,RASAL1,GALE,CCNG2,RNASEH2C
GGAGGGA	2.94	TMEM132A,GTF2IRD1,TLE2,TMEPAI,BHLHB2,COL9A2,CCNG2,FAM53B
GGGGCAC	2.63	POR,GRB7,SLC4A2,BCL9L,CCNG2
GATGTTTA	2.63	JDP2,CRTC2,ZNF784,CCNG2
CGGAC	2.61	PARP12,MKNK2,SYNGR3,CSNK1D,SLC43A2,GATA2,MAF1,PIK3R3, CCNG2,FNBP1
TCTAAGCA	2.51	LSR,SOX9,DPP7,CCNG2
CCCGACAA	2.27	CCNG2,**ZFYVE1**,FAM53B,NMB
CAGTCGCG	2.18	XRCC1,IER2,BAP1,CCNG2
GCCCC	2.18	ITPKC,ERP29,C19orf21,MKNK2,HSPF2,MCG18,DNAJC4,PHF2, PHF1, GLIS2,CASP9,N4BP3,INPPL1,JUNB,TMEM142C,GATA2,ADSSL1,BCL9L, ANKRD35,RHOV,CCNG2
CGCCCC	2.16	TP53INP2,ERP29,MKNK2,PSCD2,GALE,MXD4,CASP9, CSNK1D,INPPL1,FBXW4,BTG1,CCNG2,SLC35C1
GCTCTGCA	2.1	HIG2,C1orf210,CCNG2,FAM53B
ACGTGAC	2.07	ERP29,BMP7,UBXD1,MTHFR,CCNG2
AGAACCTG	2.06	PARP12,ATAD4,CCNG2,TMEM140

Table 4.27: Gene subsets associated with the conserved motifs found in the 5́-flanking region of gene CCNG2: dataset2

motif	score	gene subset
AGATAAAA	5.5	GPC4,BMP7,TMED4,TRPS1,LIN7A,CCNG2
ACGATCTC	3.56	LYPD1,OXTR,KCNJ8,CCNG2
CGCCTATA	2.85	SOCS3,NR3C1,CCNG2
AGGGTACG	2.65	BIK,CYP1A1,CCNG2
CAATTCGG	2.49	CCNG2,**IGSF3**,LOC649698,FAM84A
CCGAGC	2.24	EPHA4,VAV3,CCNG2,PLK2,PPAP2B
AATCAGA	2.04	CCNG2,AMIGO2,CDKN2B
GTCGCCTA	2.02	LYPD1,SPINK5,CCNG2
CAATCAGA	2.02	PDK4,ST8SIA4,CCNG2

Table 4.28: Gene subsets associated with all conserved motifs found in the 5́-flanking region of gene CCNG2: dataset3

motif	score	gene subset
GCCCC	3.61	C19orf21,ABTB1,JUP,TMEM142C,BCL9L,HOXC13,MEIS2, CCNG2,CG018,EVA1,TP53INP1
AGCCCC	3.02	CYP1A1,BCL9L,EDN1,MEIS2,CCNG2
AGCCCCGG	2.88	TMEM142C,FAM20C,CCNG2
AGGGTACG	2.88	BIK,CYP1A1,CCNG2
TACGTAA	2.83	JUB,C20orf111,CCNG2
AATAAGTC	2.83	SLC6A14,MEIS2,CCNG2
ATTGCGCA	2.67	C20orf111,ACAA2 LOC648603,CCNG2
ACCGTGCC	2.56	C19orf21,LHX2,CCNG2
AGCCCCG	2.51	ABTB1,TMEM142C,CCNG2,CG018
AGAACCTG	2.43	BIK,ATAD4,CCNG2
GGGGCAC	2.26	IRF1,BCL9L,CCNG2
GGAGGGA	2.08	GTF2IRD1,HOXC13,MEIS2,CCNG2
CAGCCCCG	2.06	ABTB1,BCL9L,CCNG2
CCCCG	2.04	ABTB1,DNAJB1,TMEM142C,FAM20C,EPOR,CCNG2,TP53INP1, ENC1 LOC730775,ENPP1
GCGTCGAA	2.02	IRF1,CCNG2

Table 4.29: Gene subsets associated with the conserved motifs found in the 5́-flanking region of gene CCNG2: dataset4

motif	**score**	**gene subset**
CCCCG	4.16	**BCL3**,POLB,ACHE,MKNK2,MMP15,DYRK1B,MYH14,SHC2,CASP9,FLOT1, PDLIM7,UBE2A,BAMBI,CCNG2,ADAM17,MAOA,SPTAN1
CCCCGC	3.49	ACHE,MKNK2,DYRK1B,IQCE,CASP9, PDLIM7,BAMBI,CCNG2, TGFBR2,MAOA,SPTAN1
GCCCC	3.39	**BCL3**,POLB,MKNK2,ABCD1 hCG 2042779,DYRK1B,GNG13,SHC2,CASP9, CASKIN2,PDLIM7,CCNG2,*CITED2*,CA5B,MAOA
CCCGC	3.36	**BCL3**,MKNK2,MMP15,DYRK1B,IQCE,VAT1,TBX2,CASP9,FLOT1,CASKIN2, PDLIM7,BAMBI,CCNG2,PNRC1,TGFBR2,*CITED2*,TLE1,SPTAN1
GGAGGGA	2.86	GTF2IRD1,**BCL3**,TGFB3,COL9A2,CCNG2
CGGGGC	2.79	**BCL3**,POLB,MKNK2,MYH14,PQLC1,PDLIM7,BAMBI,CCNG2,SPTAN1
CCGGACCG	2.77	SMARCD1,TBX2,TNK1,CCNG2
CGCCCC	2.59	MKNK2,IQCE,PQLC1,CASP9,CASKIN2,CCNG2,CA5B,MAOA,SPTAN1
CGCCC	2.49	SMARCD1,POLB,MKNK2,IQCE,GATA3,SHC2,CASP9,FLOT1,CASKIN2, PDLIM7,CCNG2,PNRC1,*CITED2*,MAOA,SPTAN1
CCCCC	2.39	**BCL3**,PHF1,CASP9,CNP,CASKIN2,**HBP1**,CCNG2,ADAM17,*CITED2*,CHD3
AGGGTACG	2.36	BIK,RNF103,CCNG2
ATTACG	2.36	BAMBI,CCNG2,RERE
GGACC	2.35	TBX2,TCF2,PIK3R3,CCNG2,*CITED2*
ACGTACCC	2.27	BIK,CCNG2,*CITED2*
ACCCGCAT	2.12	CNP,VAMP2,CCNG2
CTTAACAA	2.08	SMAD9,CCNG2,NUDT2

Table 4.30: Gene subsets associated with the conserved motifs found in the 5́-flanking region of gene CCNG2: dataset5b. In italics, a gene that is found in both ChIP data [25] and in the in-silico screening by Bourdeau et al. [19] (see Chap2, discussion section).

Chapter 5

Conclusion and perspectives

The remarkable difference between motif patterns found in the DNA upstream regions of up- and down-regulated estrogen-responsive genes suggest that regulative mechanisms for the two classes of early responsive genes differ accordingly, involving several proteins in addition to the estrogen receptor. We have identified and illustrated a few of them. In the down-regulated class, we identified two factors, GATA and CACCC-binding factor, which are consistent with down-regulation by estrogen according to data in literature. In particular, we focused our attention on GATA transcription factor (GATA3), which is involved in differentiation, in the expression of estrogen receptor [62], and which is prognostic of invasive tumor. We localized the relevant conserved motif in the upstream sequence of cyclin G2, a robust gene which is down-regulated by GATA3 [61]. We also localized an estrogen responsive element in the vicinity of this GATA binding site. We proposed a direct interaction between estrogen receptor and GATA3 factor is feasible, and may contribute the down-regulation of CCNG2 by estrogen. This hypothesis is being tested in laboratory by ChIP and siRNA techniques. In the up-regulated class, we identified ARNT factor as an important positive regulator in response to estrogen and we plan to perform appropriate experiments.

We may also raise a developmental hypothesis based upon our findings of conserved GATA binding sites in the class of down-regulated genes. A very relevant role for the GATA-3 transcription factor in mammary gland devel-

opment has been recently evidenced [66]. The mammary gland is formed by branched ducts encased in muoepithelial cells on the exterior and lined with ductal luminal cells along their length. The genetic abrogation of GATA-3 results in loss of the luminal component of the gland and inability to lactate. Furthermore, epithelial progenitors do not divide any longer. Luminal cells are also unique to express ER alpha and other markers of epithelial differentiation, while myoepithelial components are ER*alpha* negative.

The presence of GATA-3, therefore, seems necessary for estrogen to evoke the transcriptional response required for luminal cells to divide and establish the tissue. Thus, it is possible that co-presence of ERalpha and GATA factors on the same gene regulatory sequences may cooperate specifically in order to obtain the desired transcriptional response. In the example of the CCNG2 gene illustrated here, it may be hypothesized that ER alone, in the absence of GATA-3, is not able to shut-down the transcription of this cell cycle inhibitor (CCNG2). The same may be true for other genes such as BMP7. Interestingly,there is evidence that ER and GATA-3 sustain their reciprocal expression and also that these two protein interact physically, making the hypothesis of a direct involvement in the same gene contexts realistic.

Our approach proved effective in identifying meaningful motifs/factors whose validation was found in literature before than in laboratory. This approach can be applied to other sets of experimentally-defined and/or clinically relevant co-regulated genes. In order to infer pathways –an additional goal of the thesis– we plan to employ clustering algorithms in order to create motif subgroups to be associated with a consensus/function. This will also enable us to address systematically the combinatorial layer of estrogen regulation through co-localization of multiple motifs in the gene flanking regions. A topographic inspection performed on a large scale–employing algorithms for nucleosome positioning prediction, shall complete our analyses.

Presentations and publications

- ⋆ **SysBioHealth** Symposium 2007, *Characterization of regulatory sequences of estrogen-responsive genes in breast cancer cells*, poster, Milan, Italy
- ⋆ **FISV**, Italian Society of Molecular Biology and Biophysics annual meeting 2007, *Functional classification of estrogen-responsive genes through DNA sequence analyses*, poster, Riva del Garda, Italy
- ⋆ **EMBO** conference 2007, Nuclear Receptors: Structure & Function in Health and Disease, *Identifying networks of estrogen responsive genes in breast cancer cells (III)*, short talk, Gardone Riviera, Italy
- ⋆ **BITS**, Italian Bioinformatics Society Annual Meeting 2007, *Identifying networks of estrogen responsive genes in breast cancer cells (II)*, poster presentation, Naples, Italy
- ⋆ **EMBL** 3rd Biennial Symposium 2006: From Functional Genomics to Systems Biology, *Identifying networks of estrogen responsive genes in breast cancer cells (I)*, poster, Heidelberg, Germany
- ⋆ Sismondi P, Biglia N, Ponzone R, Fuso L, Scafoglio C, Cicatiello L, Ravo M, Weisz A, Cimino D, Altobelli G, Friard O, De Bortoli M.
  Influence of estrogens and antiestrogens on the expression of selected hormone-responsive genes.
  Maturitas. 2007 May 20;57(1):50-5. Epub 2007 Mar 28.
- ⋆ **Manuscript in preparation**

Acknowledgments

I am delighted to thank the many who contributed to the outline and development of this project: in the first place, my scientific advisors, *Proff. Michele De Bortoli* and *Michele Caselle*, who brilliantly supported the entire development of my thesis. I especially thank MDB for prompt, thorough explanations relevant to the biological topics and for hosting me in his laboratory; and MC for sharing methodological and theoretical perspectives, as well as for strongly encouraging me to pursue this degree.
I warmly thank *Dr. Davide Corà* for help with extraction of conserved motifs and RIKEN core promoters; *Mr. Olivier Friard* for technical support and everyday sympathy; finally, for fruitful discussions, *all of the faculties and colleagues* from the International School of Advanced Studies at University of Turin (I.S.A.S.U.T.), Ph.D program in Complex Systems in Post-genomic Biology.
I am grateful to the external referees, *Dr. Arndt Benecke*, *Dr. Giulio Pavesi* and *Prof. Shankar Subramaniam*, who kindly reviewed the final manuscript. Last but not least, I wish to thank *Prof. Federico Bussolino*, Director of our Ph.D program, for taking care of this unique school.

This work was supported by grants from the CRT (Cassa di Risparmio Torinese), Regione Piemonte, and Ministero della Salute.

... point n'est besoin d'espérer pour entreprende, ni de réussir pour pérséver
Guglielmo d'Orange

Bibliography

[1] B. J. Deroo, K. S. Korach
Estrogen receptors and human disease
J. Clin. Invest. Mar;116(3):561-70. Review. (2006).

[2] J. Schwabe, L. Chapman, J. T. Finch, D. Rhodes
The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptor discriminate between their response elements
Cell,Nov 5;75:567-578. (1993).

[3] H. Gronemeyer, J.A. Gustaffson, V. Laudet
Principle for modulation of the nuclear receptor superfamily
Nature; Nov; 3: 950-964. Review. (2004).

[4] J.A. Gustaffson
What pharmacologist can learn from recent advances in estrogen signaling
Trends Pharmacol. Sci.; 24: 479-485. (2003).

[5] Ruff M., Gangloff M., Wurtz J.M., Moras D.
Estrogen-receptor transcription and transactivation: Structure-function relashionship in DNA-and ligand-binding domains of estrogen receptors
Breast Cancer Research;2:353-359. Review. (2000).

[6] V. Perissi, M.G. Rosenfeld
Controlling nuclear receptors: the circular logic of cofactor cycles
Nature; 6: 542-554. (2005).

[7] JA Lefstin, KR Yamamoto
Allosteric effects of DNA on transcriptional regulation.
Nature; 392: 885-888. (1998).

[8] L. Björnström, M. Sjöberg
Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes
Molecular Endocr.; Jun; 19(4): 833-842. Minireview. (2006).

[9] B.O'Malley
A life-long search for the molecular pathways of steroid hormone action
Molecular Endocr.; Aug; 19(6): 1402-1411. Perspective (2005).

[10] C.M. Klinge
Estrogen receptor interaction with estrogen response elements
Nucleic Acids Research;; 29(14): 2905-2919.(2001).

[11] AM Brzozowski, AC Pike, Z Dauter, RE Hubbard et al.
Molecular basis of agonism and antagonism in the oestrogen receptor
Nature; 389:753-758. (1997).

[12] Chin-Yo Lin et al.
Whole-Genome cartography of estrogen receptor alpha binding sites
PLoS Genetics; Jun; 3(6): e87. (2007).

[13] Stossi F. Likhite V.S., Katzenellenbogen J., Katzenellenbogen B.
Estrogen-occupied estrogen receptor represses cyclin G2 gene expression and recruits a repressor complex at the cyclin G2 promoter.
J. Biol. Chem.;Jun. 16; 281(24);16272-16278. (2006)

[14] O Lone et al.
Genome targets of nuclear estrogen receptors
Mol. Endocr. 18(8);1859-1875 (2004).

[15] R. Sanchez, D Nguyen, W Rocha, JH White, S Mader.
Diversity in the mechanisms of gene regulation by estrogen receptors
BioEssays;24;244-254. (2002)

[16] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.
Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
Nat Genet. May;25(1):25-9 (2000).

[17] Cicatiello L, Scafoglio C, Altucci L, Cancemi M, Natoli G, Facciano A, Mazzetti G, Calogero R, Biglia N, De Bortoli M, Sfiligoi C, Sismondi P, Bresciani F, A Weisz
A genomic view of estrogen actions in human breast cancer cells by expression profilino of the hormone-responsive transcriptome
J. of Mol. Endocr.; 32, 719-775 (2004).

[18] Finlin BS, Gau CL, Murphy GA, Shao H, Kimel T, Seitz RS, Chiu YF, Botstein D, Brown PO, Der CJ, Tamanoi F, Andrei DA, CM Perou
RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer
J. of Biol. Chem.; : 42259-67 (2001).

[19] Bourdeau V, Deschênes J, Métivier R, Nagai Y, Nguyen D, Breschneider N, Gannon F, White JH, S. Mader
Genome-wide identification of high-affinity estrogen response elements in human and mouse
Mol Endocr.;18(6),1411-27 (2004).

[20] Laganière J, Deblois G, Lefebvre C, Bataille AR, Robert F, and V Guiguère
Location analysis of estrogen receptor a target promoters reveals that FOXA1 defines a domain of estrogen response
PNAS; Oct; 10(10): 102,33,11651-56. (2005).

[21] Carroll JS, Liu SX, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, M Brown

Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1
Cell; 122,33-43. (2006).

[22] Carroll JS et al.
Genome-wide analysis of estrogen receptor binding sites
Nature genetics; Oct; doi:10.1038/ng1901. (2006).

[23] Fan M et al.
Diverse gene expression and DNA methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens Tamoxifen and Fulvestrant
Cancer Research; Dec.; 66(24);11954-11966 (2006).

[24] Basset IG et al.
Histone methylation-dependent mechanisms impose ligand dependency for gene activation by nuclear receptors
Cell; Feb.;128; 505-518 (2007).

[25] Kwon Y-S. et al.
Sensitive ChIP-DSL technology reveals an extensive estrogen receptora-binding program on human gene promoters
PNAS; Mar.; 104(12); 4852-4857 (2007).

[26] Creighton C.J. et al.
Sensitive ChIP-DSL technology reveals an extensive estrogen receptora-binding program on human gene promoters
PNAS; Mar.; 104(12); 4852-4857 (2007).

[27] Tavera-Mendoza LE, Mader S, White JH.
Genome-wide approaches for identification of nuclear receptor target genes
Nucl Recept Signal.;4:e018. Jul 7; Epub (2006).

[28] Wasserman WW and Sandelin A.
Applied bioinformatics for the identification of regulatory elements
Nature; 276(5); Apr.; 276-287 (2006).

[29] Pavesi G, Mauri G, Pesole G.
In silico representation and discovery of transcription factor binding sites.
Brief Bioinform. Sep;5(3):217-36 (2004).

[30] Stormo GD
DNA binding sites: representation and discovery
Bioinformatics; 16(1):16-23 (2000).

[31] Jones SJM
Prediction of genomic functional elements
Annu.Rev.Genom.Human Genet.; 7:315-338 (2006).

[32] Blanchette M, Bataille AR, Chen X, Poitras C, Laganière J, Lefebvre C, Deblois G, Giguère V, Ferretti V, Bergeron D, Coulombe B, Robert F.
Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression
Genome Res. May;16(5):656-68. Epub 2006 Apr 10 (2006).

[33] Sandve GK, Drablos F.
A survey of motif discovery methods in an integrated framework
Biol Direct. Apr 6;1:11 (2006).

[34] Elnitski L, Jin VX, Farnham PJ, Jones SJ.
Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques
Genome Res.; Oct; 19 (2006)

[35] Tompa M, Li N, Bailey TL, Church GM, et al.
Assessing computational tools for the discovery of transcription factor binding sites.
Nat Biotechnol. Jan;23(1):137-44 (2005).

[36] Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J.
Genome-wide prediction of mammalian enhancers based on analysis of

transcription-factor binding affinity.
Cell. Jan 13;124(1):47-59 (2006).

[37] Benecke A.
Chromatin code, local non-equilibrium dynamics, and the emergence of transcription regulatory programs
Eur. Phys. J. E;19: 353-66 (2006).

[38] Bajic VB, Tan SL, Christoffels A, Schönbach C et al.
Mice and men: their promoter properties.
PLoS Genet.; Apr 28;2(4):e54. Epub(2006).

[39] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group).
The transcriptional landscape of the mammalian genome
Science. Sep 2;309(5740):1559-63 (2005).

[40] Segal E.
A genomic code for nucleosome positioning
Nature; Aug.; 44(17): 772-778 (2006).

[41] Ioshikhes I.P. et al.
Nucleosome positions predicted through comparative genomics
Nature Genetics; Oct.; 38(10):1210-1215 (2006).

[42] Halfon M.S.
(Re)modeling the transcriptional enhancer
Nature Genetics; Oct.; 38(10):1102-1103 (2006).

[43] Pavesi G, Mereghetti P, Mauri G, Pesole G
Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes
Nucleic Acids Res.,32: W199-W203(2004).

[44] Giulio Pavesi, Federico Zambelli
Prediction of Over Represented Transcription Factor Binding Sites in Co-Regulated Genes Using Whole Genome Matching Statistics.

Lecture Notes in Computer Science - Applications of Fuzzy Sets Theory. Volume 4578/2007; 651-658 (2007).

[45] Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B.
A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.
Nucleic Acids Res. Jan 1;34(Database issue):D95-7 (2006).

[46] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, et al.
TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.
Nucleic Acids Res. Jan 1;34(Database issue):D108-10 (2006).

[47] Reizner N, Maor S, Sarfstein R, Abramovitch S, Welshons WV, Curran EM, Lee AV, Werner H.
The WT1 Wilms' tumor suppressor gene product interacts with estrogen receptor-alpha and regulates IGF-I receptor gene transcription in breast cancer cells.
J Mol Endocrinol.,Aug;35(1):135-44 (2005)

[48] Yang L, Han Y, Suarez Saiz F, Minden MD.
A tumor suppressor and oncogene: the WT1 story.
Leukemia. May;21(5):868-76. (2007) Epub 2007 Mar 15.

[49] Davies R, Moore A, Schedl A, Bratt E, Miyahawa K, Ladomery M, Miles C, Menke A, van Heyningen V, Hastie N.
Multiple roles for the Wilms' tumor suppressor, WT1.
Cancer Res., Apr 1;59(7 Suppl):1747s-1750s; discussion 1751s. (1999)

[50] Zhu P, Baek SH, Bourk EM, Ohgi KA, Garcia-Bassets I, Sanjo H, Akira S, Kotol PF, Glass CK, Rosenfeld MG, Rose DW.
Macrophage/cancer cell interactions mediate hormone resistance by a nuclear receptor derepression pathway.
Cell.; Feb 10;124(3):615-29 (2006)

[51] Dufort D, Nepveu A.
The human cut homeodomain protein represses transcription from the

c-myc promoter.
Mol Cell Biol. Jun;14(6):4251-7 (1994)

[52] Lievens PM, Donady JJ, Tufarelli C, Neufeld EJ.
Repressor activity of CCAAT displacement protein in HL-60 myeloid leukemia cells.
J Biol Chem.; May 26;270(21):12745-50 (1995)

[53] Michl P, Ramjaun AR, Pardo OE, Warne PH, Wagner M, Poulsom R, D'Arrigo C, Ryder K, Menke A, Gress T, Downward J.
CUTL1 is a target of TGF(beta) signaling that enhances cancer cell motility and invasiveness.
Cancer Cell., Jun;7(6):521-32 (2005)

[54] Wang F., Hoivik D, Pollenz R., Safe S.
Functional and physical interactions between the estrogen receptor Sp1 and nuclear aryl hydrocarbon receptor complexes.
NAR;26(12):3044-3052 (1998).

[55] Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M.
CORG: a database for Comparative Regulatory Genomics.
Nucleic Acid Res, 31:55-57 (2003).

[56] Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G,Cox T, Cunningham F, Curwen V, Cutts T et al.
Ensembl 2006.
Nucleic Acids Res. 34 D556-561 (2006).

[57] Xie X., Lu J., Kulbokas EJ., Golub TR., Mootha V., Lindblad-Toh K.,Lander ES. and Kellis M.
Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.
Nature 434, 338 -345. (2005) .

[58] Rice P, Longden I, Bleasby A.
EMBOSS: the European Molecular Biology Open Software Suite.
Trends Genet.;Jun;16(6):276-7.(2000).

[59] Corà D., Di Cunto F., Provero P., Silengo L. and Caselle M.
Computational identification of transcription factor binding sites by functional analysis of set of genes sharing overrepresented uptream motifs.
BMC Bioinformatics; May; 11;5(1):57 (2004).

[60] Corà D., Herrmann C., Dieterich C., Di Cunto F., Provero P. and Caselle M.
Ab initio identification of putative human transcription factor binding sites by comparative genomics.
BMC Bioinformatics; May; 2;6(1):110 (2005).

[61] D.S. Oh, M. A. Troester, J. Usary et al.
Estrogen-Regulated Genes Predict Survival in Hormone ReceptorPositive Breast Cancers.
J Clin Oncol 24:1656-1664.(2006)

[62] Eeckhoute J. et al.
Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer.
Cancer Res.; Jul 1; 67(13):6477-6483. (2007)

[63] Velarde MC, Zeng Z, McQuown JR, Simmen FA, Simmen RC.
Kruppel-like factor 9 is a negative regulator of ligand-dependent estrogen receptor alpha signaling in Ishikawa endometrial adenocarcinoma cells.
Mol Endocrinol. 2007 Aug 23; [Epub ahead of print]

[64] J. Matthews et al.
Co-planar 3,3',4,4',5-pentachlorinated biphenyl and non co-planar 2,2',4,6,6'-pentachlorinated biphenyl differentially induce recruitment of oestrogen receptor alpha to aryl hydrocarbon receptor target genes.
Biochem. J.; Sep 1; 406(2):343-353. (2007)

[65] Carrol J.S. et al.
Chromosome-wide mapping of estrogen receptor binding reveals long-

range regulation requiring the forhead protein FoxA1.
Cell;122;33-43,(2005)

[66] Tlsty T.
Luminal cell GATA have it
Nature Cell Biol; Feb.; 9(2); 133-134(2007)

WEB address of the Ensembl database
`http://www.ensembl.org`

WEB address of the UCSC database
`http://genome.ucsc.edu/`

WEB address of the NCBI - NIH database
`http://www.ncbi.nlm.nih.gov/`

WEB address of the CORG database
`http://corg.molgen.mpg.de`