TEL AVIV UNIVERSITY

SACKLER SCHOOL OF MEDICINE

DEPARTMENT OF MOLECULAR GENETICS AND BIOCHEMISTRY

# Development of Bioinformatics Methods for Analysis of Functional Genomics Data and Their Application to the Study of DNA Damage Responses

THESIS SUBMITTED FOR THE DEGREE
"DOCTOR OF PHYLOSOPHY" BY
RAN ELKON

SUBMITTED TO THE SENATE OF TEL AVIV UNIVERSITY
FEBRUARY 2006

This work was carried out under the supervision of

| | |
|---|---|
| Prof. Yosef Shiloh<br>Department of Molecular Genetics and Biochemistry<br>Sackler School of Medicine | Prof. Ron Shamir<br>School of Computer Science |

## Acknowledgements

I wish to express my profound gratitude to my two supervisors, Prof. Yossi Shiloh and Prof. Ron Shamir. If any significant results were obtained during my doctoral research, it was mainly due to the fruitful and unique collaboration between these two exceptional labs, a collaboration in which I was privileged to take part. I had the great fortune to work with Chaim Linhart, a dear friend and a brilliant research partner. I wish to thank Sharon Rashi-Elkeles who faithfully ran all the microarray experiments performed during my research. I thank Prof. Ari Barziali and his lab for always reacting so positively to our ideas and for the productive collaboration. And I would like to express my appreciation and gratitude to Dr. Josef Sassoon who so generously donated my doctoral fellowship.

## Abstract

Background: Life sciences and biomedical research have undergone a revolutionary change in the last few years, with the emergence of a new paradigm, termed *systems biology*, which aims at gaining systems-level understanding of biological networks. This approach has become feasible thanks to the combination of three indispensable factors: the completion of the sequencing of genomes of various organisms, which provides us with entire blueprints of the 'program of life' in these species; the maturation of novel high-throughput biotechnologies for large-scale analysis of cellular constituents that yields comprehensive views of life in a cell; and the development of powerful computational algorithms and data analysis tools. The large-scale sequencing projects and the novel high-throughput biotechnologies have transformed biology into an information-rich science. Mining meaningful biological knowledge out of the huge volume of accumulated data is critically dependent on the availability of supporting bioinformatics tools.

Prominent among the novel high-throughput biotechnologies are gene expression microarrays that allow parallel recording of expression levels of thousands of genes in a single assay, providing genome-wide snapshots of the cellular transcriptome under the examined biological conditions. They have proven to be very powerful tools for molecular characterization of pathological conditions, and for global delineation of transcriptional programs induced by various stimuli or of programs associated with physiological processes such as differentiation, cell cycle, aging and neoplastic transformation.

DNA damage poses one of the greatest threats to the function and life of the cell and the organism and therefore cells acquired intricate mechanisms to sense and handle such challenges. The efficiency and quality of cellular responses to DNA

damage determine whether this insult will be repaired with no lasting effect on cellular life, or divert the cell from normal growth to programmed cell death (apoptosis), or end up in neoplastic transformation. Understanding of DNA damage responses has broad implications for basic life processes such as cell cycle control, aging, tissue development and degeneration. It is highly relevant for human health, primarily to coping with environmental hazards, cancer formation, and many neurodegenerative disorders.

Cellular responses to DNA damage have long been viewed mainly in terms of the concerted activation of DNA repair mechanisms and cell cycle checkpoints. However, studies that applied functional genomics approaches demonstrate that the damage-invoked network is much broader than DNA repair and cell cycle control. These recent studies showed that DNA damage sets off a wide array of signaling pathways that cover most aspects of cellular physiology, ranging from metabolic pathways to changes in protein turnover, cellular trafficking and cell-to-cell signaling. The biological mechanisms and the significance of most parts of this network are barely understood.

Our lab is interested in understanding cellular responses to DNA damage, and in particular the role of the ATM protein kinase in modulating the response to DNA double strand breaks (DSBs). ATM is positioned at the center of a physiological junction from which the cell activates a vast array of pathways in response to a DSB. To date, more than 20 direct substrates of ATM have been identified, including p53, CHEK2, and BRCA1.

Goals: The major goal of my research was to develop bioinformatics approaches for the analysis of gene expression microarray data and to apply them, as well as existing, state-of-the-art computational techniques, to the study of transcriptional networks

induced by DNA damage, with special emphasis on the role of ATM. Specific goals were to identify ATM-dependent components in the transcriptional network induced by DSBs, identify by computational means transcription factors that control the transcriptional response induced by DNA damage, and dissect the damage response network into arms mediated by these regulators.

Methods: A large-scale, gene expression microarray project that was carried out in our lab yielded an enormous amount of data. Mining meaningful biological insights from the raw data poses a major bioinformatic challenge. To meet this challenge we adopted an integrative approach for the analysis of the data that starts with the initial preprocessing steps of signal extraction, normalization and filtering, and continues through partition analysis (clustering or biclustering) to high-level statistical analyses that seek enriched functional categories and cis-regulatory promoter elements in the clusters/biclusters. This approach is implemented in the EXPANDER package, that serves as the central platform for the integration of all the microarray data analysis algorithms developed in Shamir's lab.

Results: The results of six projects are presented in my thesis. In the first project, we developed the PRIMA (PRomoter Integration in Microarray Analysis) tool for integrating computational promoter analysis in the analysis of gene expression datasets. Microarray experiments provide genome-wide snapshots of the cellular transcriptome under the examined biological conditions. They do not, however, directly reveal the transcription factors (TFs) that mobilize the observed modulation of the transcriptional program. Computational promoter analysis can potentially shed light on this hidden layer using a '*reverse engineering*' approach, in which sets of genes that show similar expression patterns are first identified (usually, by applying cluster analysis), and then the promoters of these co-expressed genes are scanned for

over-represented sequence motifs, which presumably reflect the common regulatory elements through which these promoters are co-regulated. PRIMA applies such approach. In short, given a target set and a background set of promoters, PRIMA performs statistical tests aimed at identifying TFs whose binding site signatures are significantly more prevalent in the target set than in the background set. First, we demonstrated the power of PRIMA in delineating transcriptional networks in human cells. At present, in addition to human data, we have also extracted genome-wide promoter sets for twelve organisms, including worms, insects, fish, chicken, rodents, and dog, and successfully applied PRIMA to gene expression data obtained from these species.

In the second project we established the SHARP (SHowcase of ATM Related Pathways) knowledge base for signaling pathways. Our motivation for developing this bioinformatic tool was two-fold. First, the overwhelming complexity of the cellular responses to DNA damage turns the assimilation and interpretation of extant data no less acute a problem than lack of data. We therefore realized that a computational environment for storing, visualizing and analyzing this signaling web had to be developed. Second, we envisioned SHARP as a pivotal component of our computational arsenal for analyzing functional genomics datasets. Using SHARP, we superimpose gene expression data on signaling maps to elucidate biological endpoints mediated by various pathways that are induced in response to genotoxic stress. SHARP is being developed in our labs and includes two main software components: A database for biological interactions and a visualization package that allows graphic viewing of the biological interactions stored in the database, dynamic layout and navigation through the networks, and superposition of DNA microarray data on the interaction maps.

In the third project we focused on transcriptional mechanisms that control cell cycle progression in human cells. Our computational analyses revealed eight transcription factors whose binding sites are significantly over-represented in promoters of genes whose expression is cell cycle-dependent. The enrichment of some of these factors was specific to certain phases of the cell cycle. In addition, several pairs of these transcription factors show a significant co-occurrence rate in cell cycle-regulated promoters. Each such pair suggests functional cooperation between its members in regulating the transcriptional program associated with cell cycle progression. In this project we demonstrated for the first time that the *reverse engineering* approach, which infers regulatory mechanisms from gene expression patterns, can reveal transcriptional networks in human cells. Before this study, such methodologies were successfully demonstrated only in prokaryotes and low eukaryotes.

In the forth project, we further demonstrated how functional genomics data can be utilized to discover novel functional links between transcription factors based on significant co-occurrence of their binding site signatures on common target promoters. Focusing on the oncoprotein c-Myc, we identified nine transcription factors whose binding site signatures are highly over-represented in a promoter set of c-Myc targets, which points to possible functional links between these transcription factors and c-Myc. We showed that the binding sites of most of these transcription factors were also enriched on the set of mouse homolog promoters, suggesting functional conservation.

In the fifth project we assessed the ability to precisely dissect transcriptional networks by the combination of the RNA interference (RNAi) and gene expression techniques. Analyzing a DNA damage-induced transcriptional network, we recorded

expression profiles in human cells exposed to a radiomimetic drug that induces DNA double strand breaks (DSBs). Profiles were measured in control cells and in cells knocked-down for the Rel-A subunit of NF-kB and for p53, two pivotal stress-induced transcription factors, and for ATM, a major transducer of the cellular responses to DSBs. We observed that NF-kB and p53 mediated most of the damage-induced gene activation; that they controlled the activation of largely disjoint sets of genes; and that ATM was required for the activation of both pathways. Applying computational promoter analysis, we demonstrated that the dissection of the network into ATM/NF-kB- and ATM/p53-mediated arms was highly accurate.

In the sixth project we aimed at obtaining global dissection of the transcriptional response to ionizing radiation in murine lymphoid tissue using gene expression microarrays. Probing Atm-knockout and wild-type mice, we identified a prominent cluster containing dozens of genes whose response to irradiation was Atm-dependent. Computational analysis identified significant enrichment of the binding site signatures of NF-κB and p53 among promoters of these genes, pointing to the major role of these two transcription factors in mediating the Atm-dependent transcriptional response in the irradiated lymphoid tissue. Examination of the response showed that pro- and anti-apoptotic signals were simultaneously induced, with the pro-apoptotic pathway mediated by p53 targets, and the pro-survival pathway by NF-κB targets.

Discussion: Functional genomics is changing the way biological research is done. For the first time it is possible to study biological systems as a whole and to obtain large-scale snapshots of cellular transcriptome and proteome. In our studies, we developed and applied functional genomics approaches to dissect transcriptional programs that are associated with cell cycle progression and responses to DNA damage in human

and mouse model systems. Our results elucidated novel regulatory links within these intricate signaling networks. Applying genome-wide computational promoter analyses, we pointed to novel regulators of the transcriptional program associated with cell cycle progression, and we have shed more light on the mechanisms by which the c-Myc oncogene promotes cell growth and transformation.

Fine dissection of complex transcriptional responses has posed a long-standing challenge in the signal transduction field. External and internal stimuli may activate complex networks whose analysis by traditional biochemistry can be daunting. The DNA damage response is an example of such a complex network. The combination of gene expression microarrays, manipulation of genes activity using siRNAs, and powerful computational tools holds promise for systematic and rapid dissection of such networks. The study in which we dissected the transcriptional network induced by DSBs into two major arms, the ATM/NF-κB- and the ATM/p53-dependent arms, provided a proof-of-principle for the power of this combined experimental approach, despite possible nonspecific effects of RNAi, which can be neutralized by controlled experimental design and computational analysis of the data.

Our findings on the lymph nodes dataset further elucidate the molecular network induced by IR, and might have implications for cancer management. They raised a model in which pro- and anti-apoptotic signals are induced in parallel, where the former is mediated by p53 and the latter – by NF-κB. This model suggests that restoring the p53-mediated apoptotic arm while blocking the NF-κB-mediated pro-survival arm could effectively increase the radiosensitivity of lymphoid tumors.

Our results demonstrate that the new paradigm of systems biology provides global delineation of complex cellular networks. Although systems biology is in its infancy, it is already a vital part of modern biomedical research. Its potential benefits are

enormous in both scientific and practical terms. It is expected to impact on clinical medicine as well as on pharmaceutical industries. This emerging field will eventually provide us with detailed mechanistic models for the etiology of diseases, pointing the way to novel strategies for rational intervention in pathological conditions and the design of improved personalized drugs.

# 1. Introduction

## *1.1. Systems biology and functional genomics*

Life sciences and biomedical research have undergone a revolutionary change in the last few years, with the emergence of a new paradigm, termed *systems biology*, that aims at a systems-level understanding of biological networks [1-7]. Biological research has traditionally applied reductionist experimental approaches whereby cellular systems are deconstructed into their elementary components (genes, proteins) and particular, isolated parts of the system are characterized. The transition to a new experimental paradigm in biology, often called the '*post-genome era*', was triggered by the rapid advance in the human genome project and in large-scale sequencing projects in other model organisms.

The availability of sequences of complete genomes allows us, in principle, to identify all the genes in an organism (and thereby also the entire collection of encoded proteins) — analogous to listing all the parts of a mechanical system. While such a catalog of individual components is invaluable for studying the system, it is not sufficient by itself for understanding the system's function. In the case of the living organism, we need to decipher how the components dynamically interact and regulate each other to form highly intricate physiological systems. This is the goal of the new field called *functional genomics*. More than merely assigning genes into functional categories, functional genomics aims at a comprehensive understanding of genetic networks: how gene products interact and regulate each other to produce coherent and coordinated physiological processes during the organism's development and in response to homeostatic challenges [8].

In contrast to the reductionist approach, functional genomics takes a holistic approach in which the cellular system is analyzed as a whole [9]. This systems-level

approach has become feasible in biomedical research thanks to the combination of three indispensable factors. First, as noted above, the completion of the sequencing of genomes of various organisms providing us with entire blueprints of the 'program of life' in these species. Second, the maturation of novel high-throughput biotechnologies for large-scale analysis of cellular constituents that yield comprehensive views of life in a cell, a tissue, and ultimately the whole organism. Third, the development of powerful computational algorithms and data analysis tools. The large-scale sequencing projects and the novel high-throughput biotechnologies have transformed biology into an information-rich science. Experimental biological data are being generated at an unprecedented pace. Mining meaningful biological knowledge out of the huge volume of accumulated data is critically dependent on the availability of supporting bioinformatics tools. Therefore, this novel research paradigm is multidisciplinary and necessitates intimate collaboration between biologists and computational scientists.

## *1.2. Functional genomics technologies*

The novel high-throughput functional genomics technologies analyze cellular constituents at various layers: from the level of the DNA sequence (the *genome* tier), through the expressed RNA molecules (the cellular *transcriptome* tier) to the level of the proteins (the *proteome* tier).

Prominent functional genomics tools for the study of cells at the DNA level are SNP (Single Nucleotide Polymorphisms) genotyping and CGH (Comparative Genomic Hybridization) microarrays. SNP arrays enable scientists to conduct genome-wide linkage and association studies in order to discover genetic variations underlying complex human traits [10-12]. For example, the new generation of SNP-genotyping high density array manufactured by Affymetrix, the GeneChip Mapping

100K Set, allows the genotyping of more than 100,000 distinct human SNPs in a single assay [13]. CGH microarrays have significantly increased the resolution of conventional CGH in detection of DNA copy number aberrations; this greatly improves the ability to characterize the chromosomal imbalances resulting in gain and/or loss of genomic material that are recurrent in human cancers [14, 15].

Major functional genomics technologies for analysis of the cellular transcriptome are gene expression microarrays (see detailed description in Section 1.3 below) and SAGE (Serial Analysis of Gene Expression). Gene expression microarrays rely on hybridization between RNA molecules extracted from the examined cells (or cDNA molecules derived from them), and complementary probes deposited or synthesized on the array [16-18]. SAGE is based on the isolation of unique sequence tags (typically 10-14 bp in length) from defined locations in mRNA molecules, and concatenation of these tags in a serial way into long DNA molecules for a lump-sum sequencing indicating the expression level of the corresponding RNA molecules [19]. These technologies allow parallel recording of expression levels of thousands of genes in a single assay, providing genome-wide snapshots of the cellular transcriptome under the examined biological conditions. They have proven to be very powerful tools for molecular characterization of pathological conditions, andfor global delineation of transcriptional programs induced by various stimuli or programs associated with physiological processes such as differentiation, cell cycle, aging and neoplastic transformation [20-23].

Another functional genomics approach that greatly enhances the study of transcriptional networks combines chromatin immunoprecipitation (ChIP) and promoter microarrays . This technique, also termed ‘ChIP-on-chip’, enables genome-scale identification of promoters that are bound by specific transcription factors (TFs)

under certain conditions, in a single experimental assay [24, 25]. ChIP-on-chip was recently applied in a seminal study by Lee et al. [24] to map all TF-promoter binding relationships in yeast under standard growth conditions. The microarrays used in this study contained probes corresponding to the promoters of all known and predicted genes in *S. cerevisiae*. These arrays were reacted with all known TFs in this organism, yielding a comprehensive map of the transcriptional network controlling yeast life under standard growth conditions. The approach was also applied to mammalian cells to identify genome-wide direct targets of many TFs, including E2F, c-Myc and NF-κB [26-28].

Mass spectrometric analysis for protein identification is a major technology in systems analysis of the cellular proteome (*proteomics*) [29]. It has been applied in recent years to identify protein-protein interactions on a proteome-wide scale (the *interactome*) in organisms ranging from yeast to human [30-32]. In addition, mass spectrometric techniques have been established for the analysis of post-translational modifications, such as phosphorylation and glycosylation [29, 33]. At the same time, protein microarrays are being developed. They can be divided into two categories according to use. In the first category are chips for profiling protein levels. Dozens of antibodies with high specificity are deposited on the chip to enable comparison of the expressions of the respective protein antigens from different samples. This technology was recently applied to identify proteins that are upregulated in specific cancers [34]. Furthermore, antibodies that recognize specific modified states of their target proteins (e.g., recognizing a target that is phosphorylated on a certain site) are used to measure the level of various post-translational modifications. In the second category are chips for biochemical characterization of the function of proteins. The proteins themselves are deposited on the chip and assayed in parallel for a specific biochemical reaction.

For example, in a recent study [35], a protein chip on which all *S. cerevisiae* proteins were spotted was assayed for interactions with specific substrates to identify, among others, all calmodulin-binding proteins.

Another major advance in functional genomics came with the discovery of RNA interference (RNAi) and its harnessing as a research tool [36, 37]. RNAi has dramatically expanded the scope and versatility of cell culture systems for the analysis of gene function and involvement in biological processes. Prior to this discovery, obtaining cell lines deficient for a specific protein depended largely on the availability of cells from human patients affected with genetic disorders, or from knockout mice. Methods for silencing the expression of specific genes using antisense oligos allowed only transient silencing and their efficiency was limited. The advent of RNAi technology changed this situation. The introduction of short (21 nt) double-stranded RNA (or an RNA oligo that acquires a secondary small hairpin structure that imitates double-stranded RNA) into cells results in the degradation of complementary cellular mRNA molecules via activation of the RISC complex (RNA-induced silencing complex) [36]. Furthermore, it is now possible to express ectopically such small hairpin RNA (shRNA) to generate cell lines that are stably knocked-down for the target gene [38, 39]. Several labs have undertaken the task of systematically constructing cell lines that collectively will be knocked-down for most genes in the human genome and in other model organisms [36, 37, 40, 41]. Initial progress towards this goal was recently reported when a large-scale RNAi screen was carried out in human cells to identify new components of the p53 pathway [42].

## 1.3. Gene expression microarrays: platforms

During its first few years gene expression microarray technology suffered from 'infancy maladies' [43]. Now that it has reached the stage where it yields accurate and reproducible results, it has become a standard research tool in molecular biology labs.

Gene expression microarrays come in three platforms that differ in the nature of the probe molecules used to detect RNA levels, and the method of placing them on the array surface. The three platforms are called cDNA microarrays, high-density oligonucleotide arrays, and long-oligonucleotide arrays.

cDNA microarrays use as probes PCR products (typically several hundred bps in length) of cDNA libraries. These probes are mechanically deposited ('spotted') on the array surface by a robotic arm carrying a pin head [16, 17, 44] in a simple, widely used procedure. The problems of irregular spot size and shape and probe concentration are addressed by co-hybridization to the same chip of two samples, a test sample and a reference sample, each labeled with a different florescent dye that allows relative measurements of gene expression levels.

The high-density oligonucleotide array technology was developed by Affymetrix (Santa Clara, CA). In this platform, called GeneChip, 25-mer oligonucleotide probes are synthesized directly on the array using photolithography and photosensitive oligonucleotide synthesis chemistry [8, 18, 45]. Each target gene in these arrays is represented by a probe set of 11-20 perfect match (PM) oligos complementary to different regions along the respective mRNA molecule. A parallel set of mismatch (MM) oligo probes that differ from the PM probes by a single nucleotide at the central position serves as a control that improves the discrimination between specific and nonspecific hybridization signals. Recent studies questioned the utility of the MM probes as negative controls [46, 47]. The design of the Affymetrix GeneChip requires

knowledge of the sequence of the target genes. Affymetrix offers whole- or near-whole genome GeneChip expression arrays for many organisms whose genome sequencing is near completion, including bacteria, yeast, worm, fly, chicken, rat, mouse and human.

The third gene expression platform uses as probes in-situ synthesized long oligonucleotide (60-70 mers), which are spotted on the arrays. In a study of the dependence of specificity and sensitivity of hybridization signals on probe length, Hughes et al [48] found that 60-mer probes yield optimal results. Long oligonucleotide arrays,e commercially manufactured by Agilent Technologies (Palo Alto, CA), improved probe deposition by an ink-jet printing process. Mechanically spotted long oligonucleotide arrays were produced by some academic groups [49].

Several studies compared the performance of different platforms [50, 51]. cDNA microarrays were found to be more prone to cross-hybridizations but are cheap and affordable to academic labs. Oligonucleotide-based arrays have higher flexibility in probe selection, which improves specificity. The 60-mer oligonucleotide arrays are highly sensitive (and were shown to detect genes expressed at levels as low as one copy per cell [48]). The GeneChip records are robust thanks to multiple probes used to measure each target. Extensive efforts are underway to understand the factors affecting inter-platform and inter-lab variability and to set standards for the gene expression technology [52-54].

## *1.4. Gene expression microarrays: applications*

To date, the most notable achievements of systems biology have been in global delineation of transcriptional regulatory networks that control various biological processes. These achievements were made possible by the maturation of genome-

wide gene expression microarrays and the ChIP-on-chip technique, both of which specifically shed light on the transcriptome layer of cellular systems [55-58]. Gene expression microarrays are now used in all fields of biomedical research. Some of the major applications of this technology are summarized below.

- Identification of transcriptional programs activated in response to perturbations of cellular life. Regulation of transcription is a key component of physiological networks and is the endpoint of many signal transduction pathways triggered by either extracellular or intracellular stimuli. Therefore, delineation of transcriptional programs activated during normal development or in response to homeostatic challenges is one of the important tasks of functional genomics. Even the simplest implementation of gene expression microarrays – comparison of expression profiles of two biological samples, one exposed to a certain manipulation and the other a control – has demonstrated its tremendous power to identify new connections between genes and cellular processes. In one of the early implementations of this technology, Jelinsky and Samson [59] identified several hundred genes whose expression was modified in the budding yeast following exposure to the alkylating carcinogen MMS. This single study increased by about 10-fold what is known about the transcriptional program activated in response to this stress. More advanced applications record expression profiles over multiple perturbations, time points, genetic manipulations, etc. For example, microarrays were used to disclose global transcriptional programs associated with cell cycle progression in yeast [60] and humans [21], cellular responses to various stresses in yeast [61] and human [62], aging [22], and development [63].
- Functional characterization of unknown genes. Although most of the human genome has been sequenced, a significant proportion of genes remain functionally

uncharacterized; they reside in databases as ESTs (expressed sequence tags) or predicted genes. Complementary to insights that can be derived from sequence analysis, microarrays provide a systematic and comprehensive means to obtain putative functional annotations for such genes. Many studies demonstrated that genes that share a common function or participate in a common pathway tend to be expressed together in many different biological conditions. Therefore, an uncharacterized gene with an expression pattern similar to that of a group of genes that function in a certain biological process can be tentatively assigned to this functional category. This approach is called 'guilt-by-association' [64].

- Definition of diagnostic molecular signature. Another major application of gene expression microarrays is the identification of "molecular signatures" associated with pathological conditions. Prominent successes have been reported in cancer research. Comparison of expression profiles between patients and healthy controls and among sets of patients with different prognoses has led to identification of molecular signatures that are predictive of survival and treatment outcome in several cancers [65-67]. The predictive power of molecular signature defined by microarrays outperformed conventional prognostic markers for breast cancer, and led to the first use of this technology in clinics [68]. Microarrays have served not only for identifying classification signatures of known cancers, but also for discovering novel subtypes of cancers associated with different survival rates [23, 69]. Here, microarrays hold immediate promise for disease diagnosis and personalized medicine.
- Drug development and toxicogenomics. Gene expression microarrays are widely used by drug companies: for selection of target candidates, identification of drug targets, and early identification of toxic side-effects [70, 71]. The integration of

this functional genomics technology into drug development holds promise for significant reduction in development time and in the proportion of drug candidates that fail at the late stage of in-vivo testing due to toxic effects on the animal. A closely related novel research field termed *toxicogenomics* is building a comprehensive database for expression profiles recorded after exposure of cells and animal tissues to known toxins. With such a collection of profiles, the molecular signatures ('fingerprints') that characterize toxin families sharing common mode of toxic operation can be defined. Comparing the expression profile that results from exposure of cells to a new candidate drug with the molecular signatures of known toxins will flag its toxic side effects at very early trial phases, thereby boosting the efficiency of drug development [72].

## *1.5. Reverse Engineering of transcriptional networks*

Microarray experiments provide genome-wide snapshots of the cellular transcriptome under the examined biological conditions. Comparisons of gene expression profiles under normal and pathological conditions and in response to various perturbations elucidate the corresponding alteration in the cellular transcriptional programs. Microarray measurements do not, however, directly reveal the regulatory networks that underlie the observed transcriptional modulation: the layer of the transcription factors (TFs) that mobilize the observed program is to a large extent hidden from the microarray measurements. This is because microarrays record gene expression levels, while the activity of many TFs is regulated at the protein level. In many cases, TFs are controlled by post-translational modifications such as phosphorylation and ubiquitination, which affect their activity, stability and

subcellular localization. Combining computational promoter analysis with microarray results can potentially shed light on this hidden layer.

A computational approach designed to infer transcriptional regulators from the observed expression data is termed '*reverse engineering*' of transcriptional network. This approach has two main steps: sets of genes that show similar expression patterns are identified (usually, by applying cluster analysis), and the promoters of these co-expressed genes are scanned for over-represented sequence motifs, which presumably reflect the common regulatory elements through which these promoters are co-regulated. When the approach was first being developed it successfully deciphered transcriptional networks only in lower organisms, including *E. coli* and *S. cerevisiae* [55, 58, 73].

Several computational methods implement this approach. All of them need to deal with the fact that DNA elements recognized and bound by TFs are very short (typically, 7-10 bps) and highly flexible. The binding site of most TFs contains only a very short core, typically of 3-4 bps, in which the constraint for specific nucleotides is rigid. Therefore, genome-wide computational scans of promoters for putative binding sites inevitably yield many false positives. Tools for reverse-engineering of transcriptional networks differ in several features: Some restrict the analysis to TFs whose binding site (BS) signatures are characterized and utilize pre-compiled models for these signatures (e.g., COMET [74], Toucan [75]). Others do not rely on known BS motifs but try to recover novel motifs *ab initio* (e.g., MEME [76] which applies Expectation-Maximization algorithm, AlignAce [77] which applies Gibbs sampling, MITRA [78] which performs efficient string enumeration, and Ann-Spec [79] which applies neural networks). These tools also differ in the method they use to model the BS of the TFs. Common methods are *degenerate patterns*, which specify the allowed

set of nucleotides in each position of the BS (patterns are conventionally represented using the IUPAC nomenclature); and *position weight matrices* (PWMs), which specify the probability for observing each nucleotide at each position of the BS, based on a set of empirically validated BSs of the respective TF. More complicated models that account for inter-dependencies between different positions in the BS were also suggested [80], but training them requires large sets of validated BSs and this information is currently available only for very few TFs.

The availability of sequences of many genomes in addition to the human genome greatly boosts the specificity of *in-silico* identification of regulatory elements embedded in the genome [81]. Because higher selective pressure imposed on functional elements makes them more conserved than their surrounding non-functional DNA, scanning for evolutionarily conserved elements, an approach called *phylogenetic footprinting*, markedly reduces false-positive hit rates [82-84].

Transcriptional regulation in eukaryotes is combinatorial in essence. That is, the conditions under which a gene is transcribed are determined by an intricate interplay of multiple positive and negative transcriptional regulators that recognize and bind to cis-regulatory elements within and beyond the gene's promoter region. Thus, a major task in deciphering transcriptional regulation networks is to identify combinations of TFs that cooperate in the regulation of multiple genes; that is, to identify combinations of TFs whose binding site signatures co-occur in promoters and form recurrent regulatory motifs, termed *regulatory modules*. Recent studies successfully undertook a computational approach for genome-wide mapping of such transcriptional regulation modules in *S. cerevisiae* [55, 85] and Drosophila [86-88]. Transcriptional modules in mammalian cells were defined and identified by several

pioneering studies [89-91]. Computational tools that try to define such modules include CRÈME [92] and COMET [74].

## *1.6. DNA damage response networks*

Cell life is governed by a highly structured network of biochemical pathways that evolved to maintain its metabolism, and in higher organisms also to allow it to carry out specific functions according to tissue context. This carefully laid-out plan of operation may be perturbed by numerous physical and chemical environmental agents that damage cellular constituents. Notable among them are agents that damage the DNA, posing one of the greatest threats to the function and life of the cell and the organism. DNA damage stems from several sources. It inevitably occurs during normal DNA replication (e.g., via replication errors); it is constantly induced by intermediates of normal cellular metabolism, usually by reactive oxygen species formed during cellular respiration or inflammation; and it is inflicted by exposure to environmental physical and chemical agents that induce a large variety of chemical modifications in DNA components or strand breaks.

Cells possess intricate mechanisms to sense and handle the challenge posed by DNA damage. The essentiality of these mechanisms for cell life is reflected by the conservation of their core throughout evolution. Elements in the DNA damage response network can be generally divided into a three-layered hierarchy. At the top of the network are specialized DNA surveillance *sensors* that scan the genome for abnormalities. Once the sensors detect damage, they lead to the activation of *transducers* which amplify and convey the alarm message throughout the cell by modulating the activity of downstream *effectors* that in turn affect the biological endpoints of the damage response [93]. The efficiency and quality of cellular

responses to DNA damage determine whether this insult will be repaired with no lasting effect on cellular life, or divert the cell from normal growth to programmed cell death (apoptosis), or end up in neoplastic transformation [94]. Understanding of DNA damage responses has broad implications for basic life processes such as cell cycle control, aging, tissue development and degeneration. It is highly relevant for human health, primarily to coping with environmental hazards, cancer formation [95, 96], and many neurodegenerative disorders. Strong evidence for this is provided by genetic disorders caused by defects in cellular responses to DNA damage. Patients with such disorders exhibit acute predisposition to cancer, degenerative changes in specific tissues, premature aging, and body malformations (e.g., Bloom syndrome [97], xeroderma pigmentosum (XP) [98], hereditary non-polyposis colorectal carcinoma (HNPCC) [99, 100], ataxia-telangiectasia (A-T) [101], and Nijmegen breakage syndrome [102]). The central nervous system (CNS) seems to be especially sensitive to defects in DNA damage response [103-105], possibly due to its high oxidative stress and lack of cellular turnover.

Cellular responses to DNA damage have long been viewed mainly in terms of the concerted activation of DNA repair mechanisms and cell cycle checkpoints that are activated in order to prevent cell death during DNA replication, or fixation of genetic alterations at the damage sites, or transmission of unbalanced genetic content to daughter cells [106]. However, studies that applied functional genomics approaches demonstrate that the damage-invoked network is much broader than DNA repair and cell cycle control [59, 73, 107-109]. These recent studies show that DNA damage sets off a wide array of signaling pathways that cover most aspects of cellular physiology, ranging from metabolic pathways to changes in protein turnover, cellular trafficking

and cell-to-cell signaling. The biological mechanisms and the significance of most parts of this network are barely understood.

It is also becoming apparent that different tissues induce distinct damage responses, adding to the complexity of the DNA damage response network. In general, tissue sensitivity to DNA damage may be linked to its proliferation rate: terminally differentiated, post-mitotic cells tend to be more resistant (e.g., adult brain, muscle), while tissues with high cellular turnover are more sensitive (e.g., epithelia, bone marrow, spermatogonia and hair follicles) [110-114]. This model fits well with the general radiosensitivity of tumors, which are made up of actively proliferating cells; however, it fails to explain the radiosensitivity of some of the most radiosensitive tissues – spleen and thymus - which in adults consist mainly of non-dividing cells, as well as the high radiosensitivity of bone-marrow haematopoietic stem cells, which are predominantly quiescent [110].

## *1.7. ATM and A-T*

The nuclear protein kinase ATM is positioned at the center of a physiological junction from which the cell activates a vast array of pathways in response to a specific DNA lesion, the double strand break (DSB). When DSBs crop up in the DNA, the cell activates an intricate web of pathways that includes DNA repair mechanisms, cell cycle checkpoints, and numerous other stress responses. ATM plays a pivotal role in the activation of all these branches of the damage response [96] by phosphorylating key players in each of the pathways. To date, more than 20 direct substrates of ATM have been identified that control the signaling cascades that execute the endpoint physiological processes (Fig 1.7.1).

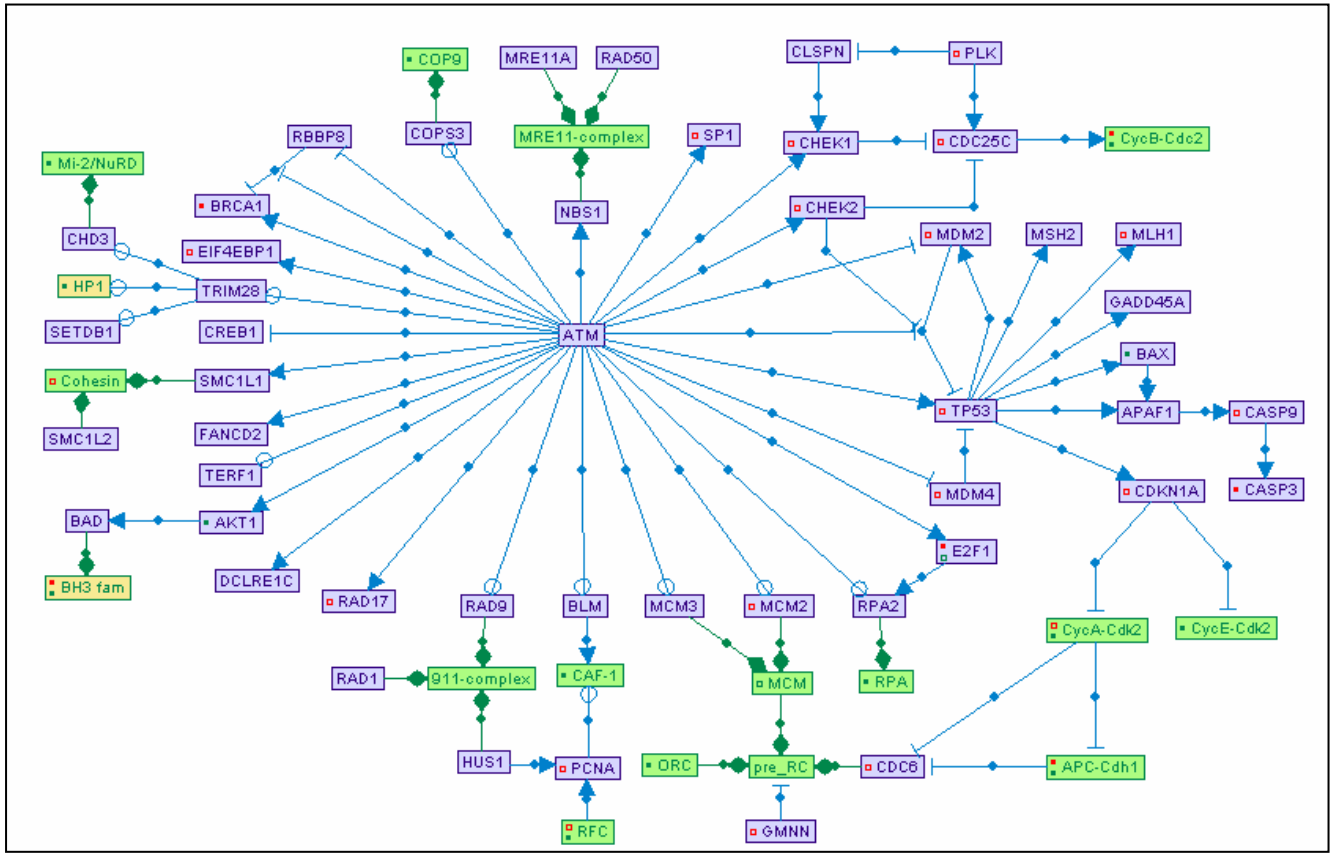



**Figure 1.7.1. ATM-regulated network.** ATM is a master regulator of an intricate web of cellular responses induced by DNA double strand breaks. In the presence of such lesions in the DNA, ATM sets off a wide array of signaling pathways by directly phosphorylating numerous substrates. Among the processes modulated by ATM are cell-cycle checkpoints, apoptosis, DNA repair, gene transcription, and protein degradation. The interactions shown in the map are selective: the entire network of all ATM interacting proteins plus their own documented interactions contains more than a hundred proteins with dozens of interconnections. This figure was generated using our SHARP tool which is described in detail in Section 4.2. Briefly, violet nodes correspond to proteins; green nodes to protein complexes, and yellow nodes to protein families. Blue edges represent regulations: arrows correspond to activation; T shape edges – to inhibition; and open circles denote regulations whose effect is still not clear. Green edges represent association between nodes (e.g., association between a protein complex and its components). Red and green dots within a node indicate that not all the regulations and associations stored in SHARP database for the node are displayed in the map.

Notable examples for ATM substrates are p53 that mediates the activation of G1/S checkpoint, DNA repair and apoptosis; the phosphorylation of the cell cycle checkpoints CHK1 and CHK2; and BRCA1 (for a recent review see [115]). A recurrent mode of operation in the tactic taken by ATM in modulating downstream pathways is its parallel regulation of several players within a target pathway. For example, ATM stimulates p53 activity by its phosphorylation on Ser 15 and augments this activation by directly phosphorylating MDM2 and MDMX [116] which interfere with their inhibitory effect on p53. In an analogous manner, ATM phosphorylates both BRCA1 and its inhibitor CtIP to achieve robust activation of this response arm [117, 118].

We are only just beginning to understand the very early events that lead to ATM activation in response to DSB. Recent evidence suggests that the complex containing the DNA repair proteins Mre11, Rad50 and Nbs1 (MRN complex) acts as sensor of DSB and is responsible for the recruitment and activation of ATM at DNA damage sites [119, 120]. In a proposed model, Mre11 and Rad50 form structural bridges between free DNA ends at DSB sites via the coiled-coil arms of Rad50 dimers, and NBS1 facilitates the recruitment of ATM and other downstream effectors by protein-protein interactions [121]. The nuclease activity of the MRN complex exerted by MRE11, which resects broken DNA ends to produce single-strand ends, was shown to be required for the activation of ATM.

The gene that encodes ATM is mutated in patients with the autosomal recessive disorder ataxia-telangiectasia (A-T). A-T is a devastating multifaceted disorder characterized by progressive degeneration of the cerebellum leading to severe neuromotor dysfunction; immunodeficiency that stems from compromised

functioning of both the B- and T-cell arms of the immune system; premature aging; growth retardation; and extreme predisposition to cancer, mainly of lymphoid origin [101]. Significantly, A-T patients show acute sensitivity to ionizing radiation and other radiomimetic chemicals and therefore cannot be treated effectively for cancer using radiotherapy and commonly used chemotherapies.

The pleiotropic nature of A-T points to the high complexity of the DNA damage network and its essential role in the maintenance of proper functioning cells and tissues. The connection between compromised DNA damage response and cerebellar degeneration in A-T was intensely debated for many years. It was even suggested that in contrast to its role as a nuclear DNA damage protein in proliferating tissues, ATM plays other roles as a cytoplasmic protein in post-mitotic neurons of the CNS [122, 123]. Recent findings brought this debate to an end. One finding is associated with the discovery of the genetic disorder ATLD (AT-like disease), which shares many common features with A-T, including the cerebellar degeneration [124]. The responsible gene encodes the MRE11 DNA nuclease protein. This similarity between the two diseases suggested that the MRN complex is required for ATM activation by DNA damage. This was experimentally confirmed by Uziel et al. [120], strongly suggesting that the cerebellar degeneration in both A-T and ATLD results from the defective DNA damage response. Moreover, Frappart et al. [125] recently showed that while inactivation of the murine *Nbs1* gene, which encodes the Nbs1 subunit of the MRN-complex, is embryonic lethal, its conditional inactivation in neural tissues results in a combination of neurological anomalies, including cerebellar defects and ataxia.

## 1.8. Signaling pathway databases

The complexity of the networks that regulate cellular physiology is growing commensurate with the enormous growth in biological knowledge. It is now clear that proteins that carry out physiological processes do not function in isolation, but rather as units of large multi-protein complexes. And, biological pathways that govern cellular development and responses to environmental challenges are not linear, parallel, and independent, but form an intricate web of interlocking processes tightly controlled by various logics of positive and negative feedback loops [126, 127].

Given this high degree of complexity, it is essential to develop computational means for processing, presenting and analyzing cellular signaling networks. However, at present, most biological knowledge resides as free text in archives of scientific journals. Before it can be processed by computers, it has to be transformed into symbolic form using a highly structured language. This process is complicated by the fact that the same protein is often referred to by different aliases in different publications, and the same designation is sometimes assigned to several different proteins. Such ambiguities can be avoided by basing the symbolic representation of biological knowledge on standardized vocabularies and controlled nomenclature, where possible.

The need to represent biological knowledge in formal language within electronic knowledgebases (KBs) is well recognized and several have been established in recent years. Most of them (e.g., EcoCyc [128], KEGG [129], WIT [130]) focus on metabolic pathways in lower organisms, which at present are the most characterized pathways. KBs for signal transduction pathways in higher eukaryotes are also coming on the scene (e.g., CSNDB [131], TRANSPATH [132], aMAZE [133, 134], BIND [135] and Reactome [136]).

In addition to their contribution as central repositories for data on protein-protein interactions, biochemical reactions and signaling pathways, and making these data available in a form amenable to computational analysis, these KBs are becoming an essential part of the analysis of data obtained by high-throughput functional genomics technologies. When examining the effect of a certain perturbation on the cellular transcriptome, hundreds of genes are typically found to respond. A major challenge is to understand the biological meaning of the observed modulation of the transcriptional program. Microarray publications usually provide long lists of genes that were found to respond to certain stimuli, but making global sense of the biological endpoints from such lists is very difficult. One way to tackle this task is to systematically integrate these results with current biological knowledge. One of the first tools for the integration of microarray data with extant biological knowledge was GeneMapp [137]. This tool provides a software environment for drawing pathways by the users, and once a pathway is drawn and submitted, it is posted on the web and can be utilized by the research community. Maps for canonical processes such as cell cycle, apoptosis and metabolic cycles were contributed by various researchers and are available from the GeneMapp website (http://www.genmapp.org/default.asp). Using GeneMapp, expression data can be linked to the drawings and used for coloring the maps. The user can navigate among the different maps to seek out those that are densely colored by the data, pointing to the activation/repression of the respective pathway in the analyzed dataset. The main drawback of this approach is that representing the knowledge in such a drawing, rather than storing it in a strictly structured DB, precludes its algorithmic processing by computers. In addition, the maps are not connected to each other, and are not updated on a regular basis.

Cytoscape [138] is another software environment for integrating high-throughput data (on cellular transcriptome, proteome, or interactome) with current knowledge on biomolecular interaction networks. In addition to providing network visualization utilities, Cytoscape implements a powerful computational approach that helps elucidate pathways and processes that are responsive in the analyzed dataset. It superimposes the user's high-throughput input dataset on the entire genome-wide network and, with no bias by the researcher's assumption about the responding pathways, searches this global map for local regions that are significantly enriched for responding genes [139].

## 2. Research goals and specific aims

The major goal of my research was to develop bioinformatics approaches for the analysis of gene expression microarray data and to apply them to the study of DNA damage response networks, with special emphasis on the role of ATM.

Specific aims:

- Develop new methods for the analysis of gene expression data.
- Apply these methods as well as existing, state-of-the-art computational techniques to experimental datasets in order to delineate transcriptional responses to DSBs in mouse tissues and human cells.
- Use these methods to identify ATM-dependent components in the transcriptional network induced by DSBs.
- Identify by computational means transcription factors that control the transcriptional response induced by DNA damage, and dissect the damage response network into arms mediated by these regulators.
- Elucidate biological endpoints of the transcriptional program induced in response to DNA damage.

# 3. Methods

## *3.1. Integrative approach for analysis of gene expression data*

A large-scale, gene expression microarray project that was carried out in our lab yielded an enormous amount of data. Mining meaningful biological insights from the raw data poses a major bioinformatic challenge. We adopted an integrative approach for the analysis of the data that includes the following steps:

- Intensity signal extraction.
- Chip normalization.
- Gene clustering and biclustering based on the chip data.
- Enrichment analysis of functional categories and cis-regulatory promoter elements.
- Superposition of the microarray data on signaling maps that reflect current knowledge on cellular signal transduction pathways.

The methods and algorithms applied at each step are described below. The analysis flow is presented in Fig 3.1.1. This integrative approach was designed and developed in Prof. Ron Shamir's group by Amos Tanay, Chaim Linhart, Roded Sharan and myself. Each application was first implemented and tested separately, and then integrated into a general analysis platform by Adi Maron and Israel Steinfeld. This integrated package, called EXPANDER, serves as the central platform for the integration of all the microarray data analysis algorithms developed in our lab (http://www.cs.tau.ac.il/~rshamir/expander/) (see Appendix A for a manuscript we recently submitted that describes EXPANDER in detail).

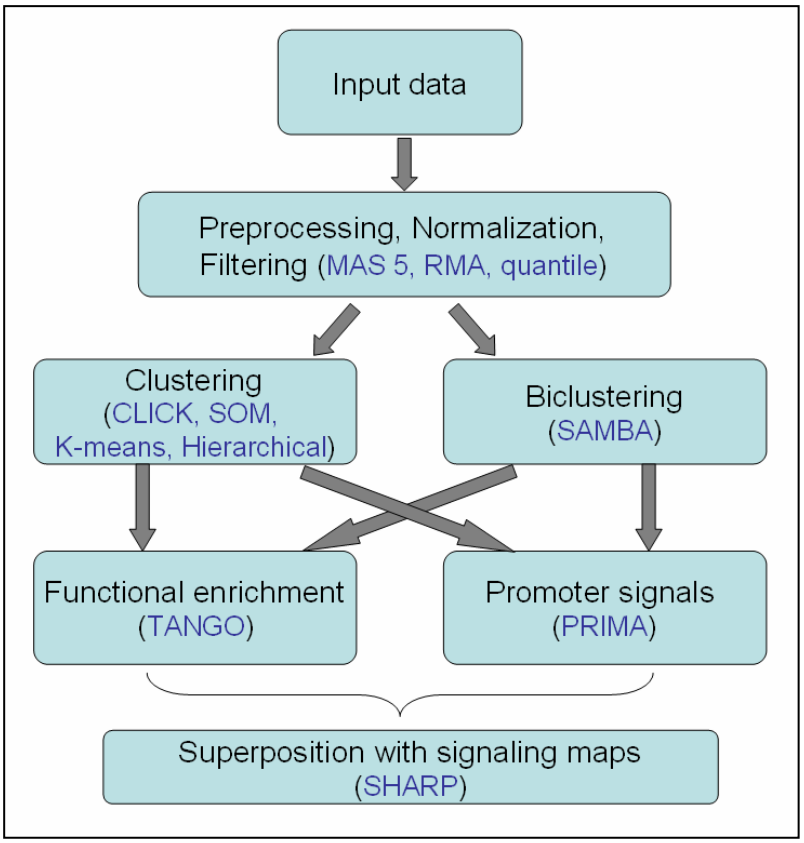



**Figure 3.1.1. Analysis flow of gene expression datasets.** A flow chart illustrating our analysis of gene expression datasets. This integrated approach starts with the initial preprocessing steps of signal extraction, normalization and filtering, and continues through partition analysis (clustering or biclustering) to high-level statistical analyses that seek enriched functional categories and cis-regulatory promoter elements in the clusters/biclusters. In the last step, we superimpose the results on signaling maps that reflect current biological knowledge. Algorithms and tools applied in each step are indicated. This approach is implemented in the EXPANDER package developed in our lab.

Extraction of expression signals. Our microarray project uses the Affymetrix GeneChip technology. In these arrays each target is probed by a set of 11-20 pairs of perfect-match (PM) and mismatch (MM) probes, each of which is 25 bp long and is complementary to a different region along the respective mRNA molecule. Arrays were scanned using an Affymetrix supplied scanner, and images were analyzed using Affymetrix image analysis software to yield 'CEL files' that assign an intensity level for each probe ('cell') in the chip. Several methods are used for summarizing the intensity levels obtained for probes in the same probe set into a representative signal that is indicative of the expression level of the respective target gene [47].

The most naïve approach was applied by Affymetrix Microarray Analysis Suite version 4 (MAS 4). The following model was assumed to describe the relationship

between the intensity of signals measured by the probes in a specific probe set, and the expression level of their target gene (i.e., the concentration of the gene's RNA):

$$PM_{ij} - MM_{ij} = \theta_i + \varepsilon_{ij}$$

Where:

$i=1,..,I$ is the index of the chip (in an experiment that includes I chips)

$j=1,\ldots,J$ is the index of the probe within a probeset.

$PM_{ij}$ is the signal measured by PM probe j of the probeset in chip i.

$MM_{ij}$ is the signal measured by MM probe j of the probeset in chip i.

$\theta_i$ is the expression level of the target gene in the sample probed by chip i.

$\varepsilon_{ij}$ is a term that reflects random error assumed to have an identical distribution over all probes in the probe set and over all conditions.

This model assumes a constant additive error term. Therefore, an appropriate statistic of the probeset signals is an arithmetic average:

$$Ei = \Sigma(PMij\text{-}MMij)/N^{*}$$

In practice, outlier probe pairs (i.e., their difference deviates from the mean difference by more than 3 standard deviations) are excluded from the sum.

Affymetrix Microarray Analysis Suite version 5 (MAS 5) refined the model after observing that the measurement error is generally proportional to the probe signal, and therefore introduced a multiplicative error term (see Affymetrix technical report, http://www.affymetrix.com/products/software/index.affx):

$$PM_{ij}\text{-}MM'_{ij} = \varepsilon_{ij} * \theta_i$$

The measurement error is therefore: $(\varepsilon ij - 1) * \theta i$, and we can write equivalently:

$$\log(PM_{ij}\text{-}MM'_{ij}) = \log(\theta_i) + \log(\varepsilon_{ij});$$

Here it is assumed that the error terms $\varepsilon_{ij}$ follow log-normal distribution. *MM′* stands for mismatch signals that are manipulated in cases where they are above perfect-match values to prevent negative values in the log transformation.

The summary $E_i$ of the probeset signals that is appropriate to this model is a (weighted) average over logarithm of the signals.

$$\mathrm{Log}(E_i) = \Sigma w_j * \mathrm{Log}(PM_{ij} - MM_{ij})$$

The $w_j$ weights are computed using Tukey biweight's function, in which the distance of each data point from the median determines how each value is weighted. Outliers far from the median contribute little to the average making the summary resistant to them.

Li and Wong [140] observed that the variability among signals measured by different probes within a probeset is very large. Therefore they introduced into their model a parameter that captures this probe affinity effect. Denote by $\alpha_j$ the affinity effect of probe j, the model assumes the following relationship:

$$PM_{ij} - MM_{ij} = \alpha_j \theta_i + \varepsilon_{ij}.$$

Multiple arrays are required in order to fit a model and to obtain good estimates for the α parameters. Ei, the estimator of θi, is obtained using linear least square procedure. This method for computation of probe signals from Affymetrix chips is implemented in the dChip tool [140].

Recently, the Robust Multi-array Average (RMA) method was introduced and was demonstrated by several studies to outperform the other methods [47, 141]. Notably, these studies questioned the utility of the MM probes as negative controls and recommended ignoring them in the computation of intensity signals (the MM signals are still utilized in estimation of global background signals, which we do not discuss here). RMA assumes the following model:

$$\mathrm{Log}(PM'_{ij}) = \log(\alpha_j \theta_i) + \varepsilon_{ij}$$

where PM' denotes PM values after background correction and normalization; αj is the affinity effect of probe j; and the error terms $\varepsilon_{ij}$ are assumed to follow a normal distribution.

RMA estimates θi using a robust linear fitting procedure.

In the early stages of our project, we used the MAS 5.0 method for extraction of intensity signals. After RMA was published we utilized it, in its implementation provided by the BioConductor project [142].

Arrays normalization. The comparison of expression levels measured under different conditions should be preceded by removal of systematic biases between arrays. The process of removing such biases is called normalization, and several methods were developed for this task. The normalization scheme implemented by Affymetrix analysis software is a basic one that computes a single scaling factor per chip. Multiplying all intensity signals in an array by the scaling factor brings the average signal to a fixed predefined level.

Several studies pointed out that systematic variation between chips is often intensity- or spatially-dependent [141, 143]. Such non-linear biases cannot be removed by global scaling and necessitate more advanced approaches. Yang et al. [143] introduced the lowess normalization scheme that computes a normalization function using local regression. Using lowess, the mean signal is equalized among the arrays over the entire range of intensities (neutralizing non-linear, intensity-dependent biases), and between all spatial sectors on the chip. An even more stringent normalization scheme, quantile normalization, was introduced by Bolstad et al. [141]. This scheme forces the distribution of signals in all analyzed chips to be identical. A recent comparative study reported that quantile normalization outperforms the other methods in removing systematic biases while retaining true biological variation. In the first phases of our project, chips were normalized using global scaling. Later, we adopted quantile normalization.

Filtering. After normalizing the arrays to a common scale, we focus our high-level analyses on the set of *responding genes*; that is, we filter out genes that are either not

expressed in the probed conditions or do not respond to the examined perturbations. First, we filter out probesets that get 'Absent' calls by Affymetrix software across all the arrays in the dataset. Typically, signals of such probesets are at the lower tail of the intensity distribution, close to the level of background noise. Then we scan the data for genes that show a variation above a predefined threshold across the probed conditions.

The number of replicates used in our experiments (2-3 repeats) usually does not allow sufficient statistical power to robustly detect differentially expressed genes; rigorous statistical tests at this stage usually pass too few genes for subsequent analyses. Therefore, at this initial filtering step, we usually apply a naive fold-change (FC) filtering criterion. Statistical tests with controlled rate of false positive discoveries are applied in our downstream analyses in order to detect global phenomena within the set of responding genes. As a default we use FC threshold of 1.75. Affymetrix reported that such a threshold corresponds to an average of 5% false positives in a single differential experiment without replicates [18]. Before applying this filtering scheme we set a floor intensity level: all intensities below a certain floor level are set to this level to reduce false calls in the low intensity range. We set the floor level to the $75^{th}$ percentile intensity of the distribution of the 'Absent' probesets.

Identification of major expression patterns in the dataset. In the next step, we subject the set of responding genes to cluster analysis. Clustering algorithms that are applied to gene expression data partition the genes into distinct groups according to their expression patterns over a set of experimental conditions. Such partition should assign genes with similar expression patterns to the same cluster (keeping the *homogeneity* merit of the clustering solution) while retaining the distinct expression pattern of each cluster (ensuring the *separation* merit of the solution). Cluster analysis eases the

interpretation of the data by reducing its complexity and revealing the major underlying expression patterns. We used several clustering algorithms, including SOM [20], K-means [58], hierarchical clustering [144], and *CLICK* [145]. The last one is a graph theory-based algorithm developed in Shamir's group by Roded Sharan. *CLICK* was demonstrated to outperform other algorithms according to several objective figures of merit [146].

As expression data accumulate and profiles over hundreds of different biological conditions are readily available, clustering becomes too restrictive. Clustering algorithms globally partition the genes into disjoint sets according to the overall similarity in their expression patterns; i.e., they search for genes that exhibit similar expression levels over all the measured conditions. Such a requirement is appropriate when small to medium size datasets from one or a few related experiments are analyzed, as it provides statistical robustness and produces results that are easily visualized and comprehended. But when larger datasets are analyzed, a more flexible approach is needed. A *bicluster* is defined as a set of genes that exhibit significant similarity over a subset of the conditions. A biclustering algorithm can dissect a large gene expression dataset into a collection of biclusters, where genes or conditions can take part in more than one bicluster. A set of biclusters can thus characterize a combined, multifaceted gene expression dataset [147]. For this analysis we utilize *SAMBA* (**S**tatistical-**A**lgorithmic **M**ethod for **B**icluster **A**nalysis), an algorithm that was developed in Shamir's group by Amos Tanay and Roded Sharan [148].

Functional enrichment analysis. After identifying the main co-expressed gene groups in the data (either by clustering or biclustering), one of the major challenges is to ascribe to them a biological significance. To this end, we applied statistical analysis

that seeks specific functional categories that are significantly over-represented in the analyzed gene groups with respect to a given background set of genes. This analysis utilizes functional annotation files that associate genes with functional categories using the standard vocabulary defined by the Gene Ontology (GO) consortium ((http://www.geneontology.org/), [149]). The statistical significance of the enrichment of a specific cluster for genes from a particular functional category is determined by computing the upper tail of the hypergeometric distribution (see, for example, [58]), taking into account the number of genes represented on the chip that are associated with this functional category. While certain genes are represented by several probe sets, to avoid biases, each gene is counted only once.

The functional enrichment module integrated in EXPANDER currently supports six organisms: yeast (*S. cerevisiae*), worm (*C. elegans*), fly (*D. melanogaster*), rat (*R. norvegicus*), mouse (*M. musculus*) and human. We compiled annotation files for these organisms based on data provided by the GO consortium and by the central databases for these organisms.

An improved module for functional enrichment analysis was recently integrated in EXPANDER. The *TANGO* (**T**ool for **AN**alysis of **GO** enrichments) algorithm developed by Amos Tanay (manuscript in preparation) accounts better for extensive multiple testing typically done in such analysis (hundreds of categories are typically tested for enrichment). While standard methods for accounting for multiple testing assume independent tests (e.g., Bonferroni, False Discovery Rate), the hierarchical tree-like structure of the GO ontology induces strong dependencies among the categories. TANGO accounts for these dependencies, thus yielding more reliable p-value estimations.

Cis-regulatory element analysis. Microarray measurements provide snapshots of cellular transcriptional programs that take place in the examined biological conditions. These measurements do not, however, directly reveal the regulatory networks that underlie the observed transcriptional activity, i.e., the transcription factors (TFs) that control the expression of the responding genes. Computational promoter analysis can shed light on the regulators layer of the network (see Section 1.5). We developed the PRIMA tool for this task (PRIMA is described in Section 4.1) and routinely apply it to gene expression datasets to discover TFs that control the observed alteration in the cellular transcriptome.

Superposition of gene expression data on signaling maps. Knowledgebases for signaling pathways are becoming highly instrumental in the analysis of high-throughput data in general, and gene expression in particular. A simple way to integrate gene expression data with extant biological knowledge is to present microarray results on signaling maps. This can be done, for example, by coloring genes in the maps according to their expression levels. Such coloring points to sub-regions in the network that are turned on or shut down in response to the examined perturbations. Such sub-regions will correspond to subgraphs densely populated with genes that are induced or repressed in the dataset. We use our SHARP tool for this task (SHARP is described in Section 4.2).

Applying this integrative approach to analysis of gene expression datasets allows us to systematically identify major expression patterns in the data (by applying cluster or bicluster analysis), assign clusters with putative functional roles (based on the enriched functional categories), and reveal transcription factors that underlie the transcriptional response exhibited by the clusters (using PRIMA). Superposition of the data on signaling maps using SHARP helps us to identify active pathways and to

generate hypothetical mechanistic models for cellular networks that respond to various stresses. Gene expression projects in which we applied this flow of analysis are described in Sections 4.5 and 4.6.

# 4. Results

The results of six projects are presented in this thesis. Section 4.1 describes our promoter analysis tool PRIMA (**PR**omoter **I**ntegration in **M**icroarray **A**nalysis) and Section 4.2 presents the signaling knowledgebase SHARP (**SH**owcase of **A**TM **R**elated **P**athways). Section 4.3 describes a project in which we demonstrate that the reverse engineering approach can successfully delineate transcriptional networks in human biological systems. As a test case we applied this approach to data on cell cycle progression in human cell lines. In Section 4.4 we further demonstrated how computational promoter analysis can be utilized in the analysis of ChIP-on-chip datasets. The c-Myc TF served as our test case. Section 4.5 presents a study in which we combined the microarray and siRNA technologies to dissect transcriptional network induced by DNA damage in human cell line. Finally, in Section 4.6 we applied our experimental and computational strategy to delineate transcriptional responses to ionizing radiation (IR) under physiological conditions: murine tissues from wild-type and Atm-deficient animals.

## *4.1. PRIMA*

At the time we began this project, reverse engineering of transcriptional networks had been successfully applied only to lower organisms, such as *E. coli* and *S. cerevisiae* [55, 58, 73]. We were motivated to demonstrate that, notwithstanding their much higher complexity, networks that govern transcriptional networks in human cells can be elucidated by this approach as well. For this purpose Chaim Linhart of Ron Shamirs' group and I developed PRIMA (Promoter Integration in Microarray Analysis) for integrating computational promoter analysis in the analysis of gene

expression datasets. PRIMA is described in [150] (see Appendix B and http://www.cs.tau.ac.il/~rshamir/prima). Based on the assumption that genes that exhibit similar transcriptional expression patterns across multiple conditions share cis-regulatory elements in their promoters, PRIMA computationally scans promoters for TF binding sites in search of these common sequence elements. In short, given a target set and a background set of promoters, PRIMA performs statistical tests aimed at identifying TFs whose binding site signatures are significantly more prevalent in the target set than in the background set. Technical details on PRIMA's principles of operation are given in this section. Projects in which we utilized PRIMA to delineate transcriptional networks in humans and mouse are described in Sections 4.3-4.6.

A first requirement for PRIMA is the availability of genome-wide promoter sequences. Thus, we initially downloaded the entire human genome data assembled into genomic contigs by the NCBI Reference Sequence project [151] (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens, release of June 2001). The version in which human repetitive sequences are masked was used (mfa files). From these genomic contigs, we extracted putative promoter sequences of known human genes based on their start annotations provided by NCBI (gbs files provided at the same url). Next, we had to determine the promoter region around the putative transcription start site (TSS) in which to search for transcriptional regulatory elements in human networks. We examined the location distribution of 1075 empirically validated TF binding sites in human promoters (data obtained from TRANSFAC database [152]). We found that some 80% of these elements were located within 1,200 bases upstream of the genes' TSS, and therefore confined our analyses to this region. Clearly, current knowledge is biased towards binding sites at short distances from the TSS. Certain regulatory elements are known to act over very large distances, up to several kilobases

from TSS; yet it is clear that ample information resides in sequences in close proximity to the TSS. At the time we began this project our promoter set contained sequences for putative promoter regions of 12,981 known human genes, each 1,200 bp in length. We refer to this promoter set as the '13K set' (which can be downloaded from http://www.cs.tau.ac.il/~rshamir/prima/PRIMA.htm). Since the vast majority of the genes' TSSs in the human genome are not experimentally validated, it was important to estimate the quality of this promoter set by benchmarking it with experimentally validated human promoters extracted from the EPD database [153]. EPD contained validated promoter sequences for 247 distinct human genes. The 13K set contained promoter sequences for 180 of these genes. When the pairs of putative and validated promoters were aligned, the distance between the putative and true TSS was within 200 bp in 70% of cases.

As part of the maintenance of PRIMA, we routinely update the promoter set with every major update of the human genome assembly. At present, in addition to human data, we have also extracted genome-wide promoter sets for all organisms supported by the Ensembl project [154]), using a Perl script based on the application programming interface provided by Ensembl. We have already downloaded putative promoter sequences for twelve organisms, including worms, insects, fish, chicken, rodents, dog and human. Ensembl TSS annotations are derived from alignment of cDNAs, ESTs and proteins sequences to the genomic sequence (for detailed description of Ensembl's method for gene building see [155]). Best TSS annotations are obtained for organisms for which large collections of full-length cDNAs are available (e.g., the FANTOM2 set for mouse [156] and the comprehensive human full-length cDNA libraries [157, 158]). In order to ensure that the quality of the annotated TSS in all organisms suffices for the detection of cis-regulatory elements,

we verified that the TATA-box signal detected by PRIMA peaks at the correct location, in the very proximity of the TSS, in each of the tested species (Fig. 4.1.1).

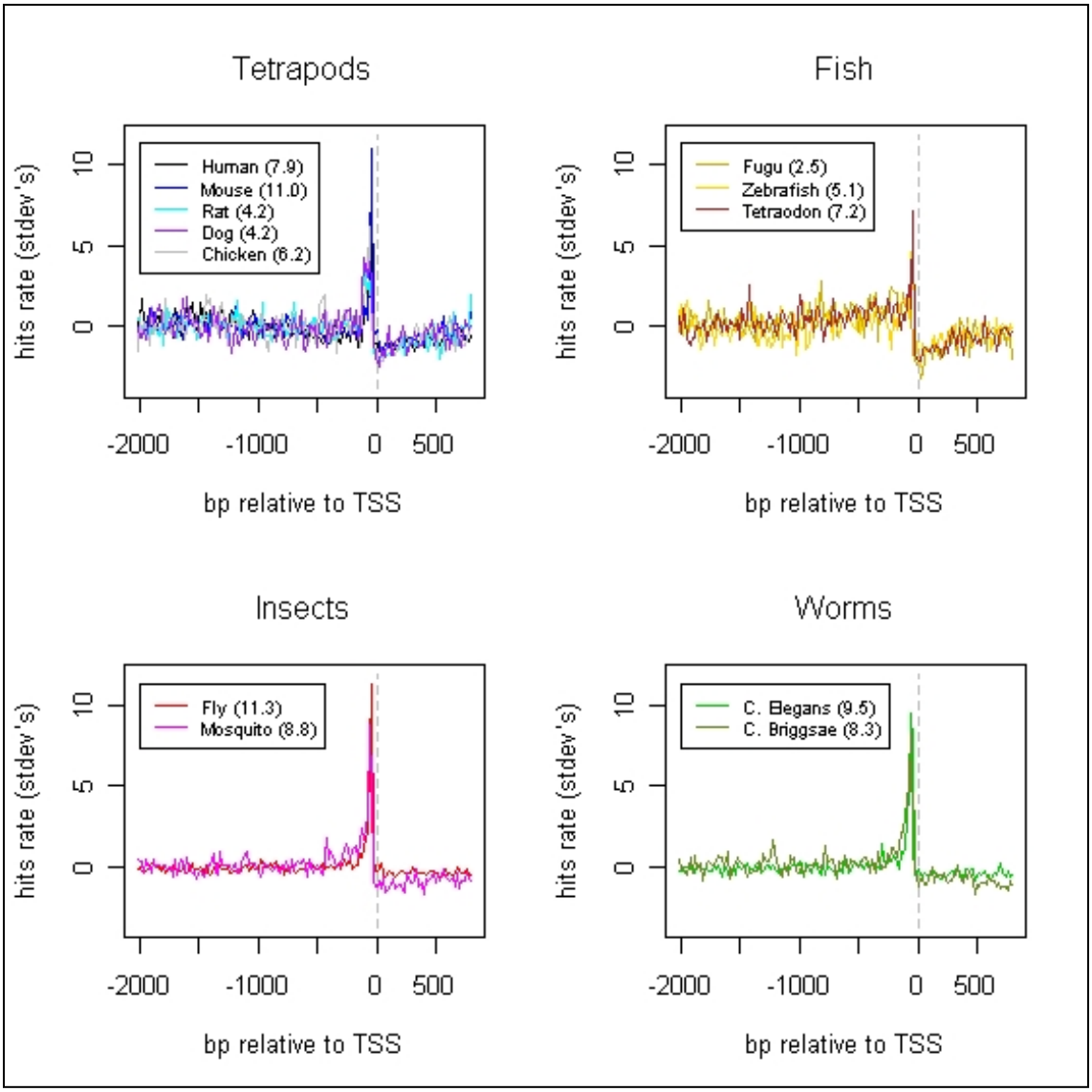



**Figure 4.1.1. Location distribution of TATA-box signal in 12 species**. Promoter sequences spanning 3000 bps were scanned for hits of TATA-box using PRIMA. For each species, the promoters were divided into bins of 20 bps length, and the total number of TATA hits in all promoters was recorded for each bin. Y-axis represents the normalized counts in each bin. A very sharp peak was observed at the expected position (-30 relative to the TSS) in all organisms except Fugu, confirming the high quality of the TSS annotation. The peak values (in standard deviations) are shown in the legends.

In scanning promoters for putative TFBSs, PRIMA relies on extant information on binding site signatures of known TFs. PRIMA uses *position weight matrices* (*PWMs*) as models for TFBSs, and we obtained PWMs for known mammalian TFs from the TRANSFAC database [152] which has the largest PWM collection currently available.

PRIMA was primarily designed to elucidate TFs that underlie the transcriptional response in gene expression datasets. In this context, PRIMA gets sets of co-expressed genes (identified by cluster analysis), scans their promoters for TFBSs of known TFs, and identifies PWMs that are significantly over-represented in each examined set compared to some background set of promoters. Typically, the set of genes represented on the microarray or the set of genes that were expressed in the examined conditions serves as the background set in PRIMA's tests.

Formally, given a PWM $P$ of length $l$, both strands of each promoter are scanned by sliding a window of length $l$ along the promoter. At each position of the window, a similarity score is computed between $P$ and the corresponding subsequence of the promoter. Denote by $p(i,j)$ the frequency of base $i$ at position $j$ in the PWM $P$. Given a promoter subsequence $s_1 s_2 \ldots s_l$, we originally defined its similarity to $P$ as follows:

$$sim(P, s_1 s_2 \ldots s_l) = \prod_{j=1}^{l} p(s_j, j)$$

At later stages, we added an option to weigh each position by its information content, which gives further importance to matches in the BS core positions. The value of the information content at each position is between 0 – for positions with uniform distribution of the nucleotides ('low information') and 2 – for positions with absolute requirement for one specific nucleotide ('maximal information') [159]. In order to identify putative binding sites, or *hits*, of a TF, a threshold $T(P)$ for the

similarity score of the TF's PWM *P* is determined. Subsequences with a similarity score above *T(P)* are regarded as hits of *P*. The threshold *T(P)* is controlled by two parameters, α and *β*. The first parameter controls the rate of hits of *P* in random sequences as follows: A set of 1,000 random promoters of the same length as the real promoters is generated by an order-2 Markov model learnt from the background promoters. A threshold $T_1$ is computed, such that α percent of the random promoters contain one or more sites whose similarity score to *P* is above $T_1$. For several PWMs, we observed that typical values of α (e.g., 5%-10%) yielded very low hit rates on the real promoters, which makes it difficult to reveal statistical enrichment for the corresponding TFs. Therefore, we introduced a second parameter, *β,* which controls the rate of hits of *P* in a background set of promoters. A threshold $T_2$ is computed, such that a fraction of *β* background promoters contain one or more sites whose similarity score to *P* is above $T_2$. The threshold *T(P)* is set as the minimum of $T_1$ and $T_2$. Default parameter values were empirically set to α=10%, and *β*=10%. Although the choice of these particular parameter values is somewhat arbitrary, we observed that significant TF BS enrichments are not sensitive to a large range of settings for these parameters.

Once a similarity score threshold is set, the PWM *P* is used to scan the promoters. Given a set *B* of *n* background promoters, and a subset *T* of *m* target promoters, we compute an analytical score for the observed enrichment of PWM *P* in *T* with respect to its abundance in *B*. Suppose there are *h* hits of *P* in *T*, where three hits at most are counted per promoter. Let $n_1$, $n_2$ and $n_3$ denote the number of background promoters containing one, two, or at least three hits, respectively. Assuming that *T* is randomly chosen out of B, the analytical score for the probability of observing at least *h* hits in *T* is:

$$p = \frac{\sum_{i+2j+3k \geq h} \binom{n_1}{i}\binom{n_2}{j}\binom{n_3}{k}\binom{n-n_1-n_2-n_3}{m-i-j-k}}{\binom{n}{m}}$$

We examined the accuracy of our analytical scores by an empirical statistical procedure. We tested how often each of the PWMs with analytical score $p$<0.001 received at least $h$ hits on 10,000 random sets of promoters (analytical score < 0.001 indicates that on average there should be less than 10 sets with at least h hits out of 10,000 random sets). Each set was generated by randomly choosing a subset of $m$ promoters from B. The empirical validations indicated that the analytical scores are reliable (after accounting for multiple testing).

Transcriptional regulation in eukaryotes is to a large extent combinatorial. Therefore, in addition to the identification of TFs whose BSs are enriched in given target sets, PRIMA also applies a statistical test to identify pairs of TFs whose BSs tend to appear on common promoters much more frequently than would be expected by chance alone. Significant co-occurrence of TF signatures can point to regulatory interplays maintained among the respective TFs. Formally, given a set of $m$ promoters, and a pair of PWMs, $P_a$ and $P_b$, denote by $f_a$, $f_b$ the number of promoters that contain a hit for $P_a$, $P_b$, respectively. Let $f_{ab}$ be the number of promoters with a hit for both $P_a$ and $P_b$. The p-value for observing $f_{ab}$ or more promoters containing hits for both PWMs is:

$$p = \sum_{h=f_{ab}}^{\min\{f_a, f_b\}} \frac{\binom{f_a}{h}\binom{m-f_a}{f_b-h}}{\binom{m}{f_b}}$$

Overlapping hits of $P_a$ and $P_b$ are omitted from counting.

In its stand-alone version, an execution of *PRIMA* typically takes several hours to complete. To facilitate the execution of *PRIMA* from within *EXPANDER*, we added a preprocessing phase, which decreased the running time to only a few minutes on a standard PC. The preprocessing phase is run by us on occasions of major updates of genome sequence assemblies of the supported organisms (yeast, worm, fly, rat, mouse, and human). This step generates promoter-fingerprints file for each organism. These fingerprints files, which are supplied with EXPANDER, map computationally-identified, high scoring putative binding sites ('*hits*') of all TFs to the entire set of promoters in the organisms. In the version integrated in *EXPANDER*, *PRIMA* loads the hits data from the fingerprints files rather than scanning promoter sequences de-novo on each run, thereby enormously reducing the running time. This improvement greatly enhanced the flexibility of *PRIMA*, enabling its execution in an iterative way, in which results obtained by different clustering solutions can be routinely compared. The following sample PRIMA run demonstrates the drastic improvement in running time due to the use of fingerprints. A PRIMA analysis of target and background sets of 500 and 22,240 genes, respectively, that tested for enrichment of 430 PWMs in the region from -1000 to +200 relative to the TSS took 4 hrs and 48 min (some 40 sec per PWM) on a standard PC (Pentium 4 2.4 Ghz, 768 MB of RAM). The same run took less than a minute using the fingerprints file.

## *4.2. SHARP – a DNA damage signaling knowledgebase*

We initiated and designed the SHARP (**SH**owcase of **A**TM **R**elated **P**athways) tool as a signaling pathway knowledgebase focused on ATM- and DNA-damage

related networks. SHARP, based on the PIVOT tool developed by Nir Orlev [160], is currently developed by Giora Strenberg, Ran Blekhman and Jackie Assa of our group. SHARP is expected to be officially released by September 2005. The tool is already operational and utilized in our lab and in several other labs serving as beta sites.

Our motivation for developing this bioinformatic tool was two-fold. First, the complexity of the cellular responses to DNA damage is much higher than originally thought. The network regulated by the ATM protein exemplifies this high degree of complexity. Although our present understanding of ATM functions is only partial, the signaling network that involves all currently known ATM-interacting proteins, ATM substrates, and downstream effectors is already known to contain hundreds of proteins with dozens of interconnections (see Fig 1.7.1 in the Introduction). Given this complexity, the assimilation and interpretation of data become no less acute a problem than lack of data. We therefore realized that a computational environment for storing, visualizing and analyzing this signaling web had to be developed. Second, we envisioned SHARP as a pivotal component of our computational arsenal for analyzing the high volume of data expected from our gene expression microarray project. Using SHARP, we superimpose gene expression data on signaling maps to elucidate biological endpoints mediated by various ATM-dependent pathways.

In this thesis I will describe in detail only the data model of SHARP, since this was the module in which I was most deeply involved. Other modules of SHARP software were designed and implemented by others and therefore are presented here only in general principles.

**SHARP's data model.** We have devised a formal data model in which information on signaling interactions is summarized in a format amenable to computerized manipulation and analysis. Two fundamental requirements instructed our model:

First, the formal language should be of sufficient expressive power to capture information on most aspects of regulatory pathways. Second, it should be kept as simple as possible for easy entry of novel data by all SHARP users. Too complicated a model would hamper our goal that SHARP DB be populated primarily by the user community. With these considerations in mind, we designed a data model that is based on two basic types of objects: "*biological entities*" and "*regulatory interactions*".

1. ***Biological entities***: SHARP data model comprises the following four biological entities:

a. *Genes/Proteins*. Genes (and the proteins they encode) are the atomic elements of our model. To avoid ambiguities about the identity of genes, only characterized human genes that were assigned a formal designation by the Human Genome Nomenclature Committee (HGNC) are included in SHARP gene/protein space. At present, over 20,000 human genes have been assigned such unique symbols.

b. *Protein families*. These families are groups of isoforms (encoded by distinct loci) that share most of their biological activities. Well known examples are the JNK family which includes JNK1, JNK2 and JNK3 proteins (whose official names are MAPK8, MAPK9 and MAPK10), and the p38 family that is comprised of four p38 isoforms officially designated as MAPK11-14.

c. *Protein complexes*. These complexes are groups of proteins (or protein families) that carry out a specific function only when associated with their complex mates. For example, the DNA-damage sensor MRN complex is composed of three proteins — Mre11, RAD50 and NBS1; the NF-κB transcription factor is a hetrodimeric complex that is composed of two subunits, each a family of proteins (the first subunit is a

member of the Rel family that contains the Rel, RelA and RelB proteins; the second subunit is either NF-κB1 (whose common alias is p105) or NF-κB2 (p49/p100)).

d. *Small molecules*. This entity enables the model to include descriptions of interactions involving or modulated by small signaling molecules such as GTP, cAMP, $Ca^{++}$, etc.

2. ***Regulatory interactions***. The second object in our data model is the *regulatory interaction*. A regulatory interaction has the following structured form:

[*Source*] <*regulatory mode*> [*Target*]

(e.g., ATM activates p53, p21 inhibits CycE-CDK2 complex).

A regulatory interaction can be defined between any two biological entities or between a biological entity and another regulatory interaction. A regulatory interaction can have one of three modes: "promote/activate", "inhibit/repress", or "unknown".

A regulatory mode can be achieved by different biochemical mechanisms (phosphorylation, transcriptional regulation etc.). Each regulatory interaction (defined by its source, target and mode/effect) is associated with several attributes: the biochemical mechanism by which the regulation is driven, at least one supporting reference, its submitter, and reliability and status flags for quality control (see below).

Allowing the definition of one regulation as the target of another regulation enhances the flexibility and expressivity of the SHARP data model. First, it enables our model to describe regulatory interactions that affect only a subset of downstream interactions emanating from its target, as illustrated in Fig 4.2.1a. In addition, it improves the specificity of the description, as demonstrated by the comparison between Fig 4.2.1b and Fig 4.2.1c. For regulations that act on other regulations, an

additional attribute is added: the *physical target*, which specifies the physical entity on which the biochemical interaction is exerted.

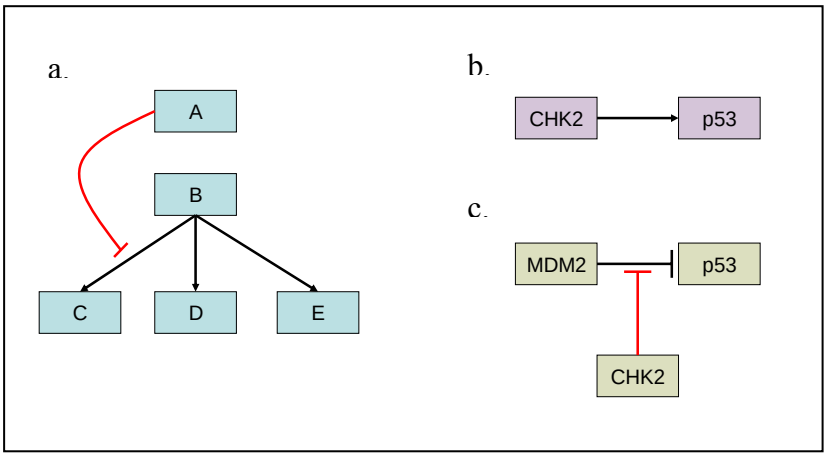



**Figure 4.2.1. SHARP data model**. (a). The utility of adding the SHARP data model with the option to define regulation between a biological entity and another regulation. This type of regulation is helpful in cases where the effect of regulator A on target B is transmitted to some but not all targets of B. In this schematic example, A specifically inhibits the B-mediated activation of C. (b). A schematic representation of p53 activation by Chk2. (c). Introducing a regulation between a biological entity and another regulation allows our model to be more elaborative. Here, we present the information that the activation of p53 by Chk2 is achieved by interfering with the inhibition of p53 by Mdm2 (See Hiaro A. et al. [161]). Therefore, the target of this regulation is the inhibition of Mdm2 on p53, and its physical target is p53.

**SHARP components**. SHARP includes three main software modules: A database (DB) for signaling interactions, a visualization package for presenting pathway maps, and an algorithmic engine for analyzing the networks.

SHARP DB for biological interactions. SHARP data model is implemented in a relational database using MySQL as the DB management system (DBMS). SHARP contains data from two sources: 1) interactions inserted individually by SHARP users; and 2) data imported en masse from general signaling pathway DBs (we have already imported data from KEGG [129] and plan to import data also from the Reactome [136] and BIND [162] DBs). Data uploaded by SHARP users will focus on pathways intimately related to ATM, such as cell cycle checkpoints, DNA repair, and apoptosis. Data loaded en masse will cover many other aspects of cellular signaling, which will facilitate novel discoveries on functional relationship between ATM and various physiological processes (such as energy metabolism, transcriptional regulation,

protein turnover, RNA processing). Having these two sources combined should produce a high volume of data in the SHARP DB, which will be especially comprehensive on pathways closely related to ATM. At present, we have already populated SHARP DB with several hundred interactions pertinent to ATM signaling. We decided that SHARP DB will contain only direct regulatory interactions although the model supports both direct and indirect interactions.

The issue of data quality is tackled by several means. Each submission is attached with reliability and status flags. The reliability flag is set by the user who uploads the data and reflects the confidence the user ascribes to it; in general, interactions derived from highly focused biochemical studies are assigned high reliability, while those derived from high-throughput experiments are assigned low reliability. The status flag is set to 'Draft' at the time of submission. SHARP curators will review all new submissions, decide whether to accept, edit or reject the submitted data, and set the status flag accordingly. SHARP visualization module allows the filtering of displayed data according to their reliability and status values.

Standardization of metadata is important in the context of signaling DBs, but there are no standard widely-accepted ontologies for summarizing biological experimental details. A collective effort for their definition is required by the biomedical research community (similar to the MIAME standards for microarray experiments' metadata). Once such standardization is defined, SHARP will be adjusted to support it. Currently, each regulation is linked to at least one supporting reference from which details on experimental conditions can be retrieved. Standard nomenclature is both available and critical for genes' identifiers, and therefore is strictly enforced by SHARP – genes are uniquely referred to by their official HUGO symbol. In addition, in order to impose basic uniformity, fields that describe the regulation are provided as

closed vocabularies (e.g., regulation effect is defined as 'promote', 'inhibit', 'unknown'; regulation mechanism is selected from a closed list of biochemical processes: 'phosphorylation', 'transcriptional regulation', 'ubiquitination' etc).

SHARP visualization package. SHARP pathway visualization module, based on the PIVOT protein visualization tool that was developed in our lab by Nir Orlev [160], allows dynamic presentation of the biological interactions stored in the DB and gradual navigation through the networks. Researchers can manipulate the graph interactively, control its layout, expand or collapse selected segments, and retrieve further information on both biological entities and displayed interactions. Another utility supported by SHARP visualization module is the superposition of functional genomics data onto the signaling map to color it according to supplied experimental results (Fig 4.2.2).

SHARP algorithmic inference engine. This module is not yet implemented. We plan to include in it basic graph analysis utilities such as: the path-finding algorithm that will enable the user to find the shortest path(s) of interactions connecting two selected entities in the network; and more sophisticated tasks such as identification of 'hot-spots' in the entire network given high-throughput results. Examples of the latter are identification of sub-regions that are significantly dense for genes/proteins that show some property of interest and thereby point to active pathways in the analyzed dataset, and identification of apparent discrepancies between interactions in the network and experimental data.

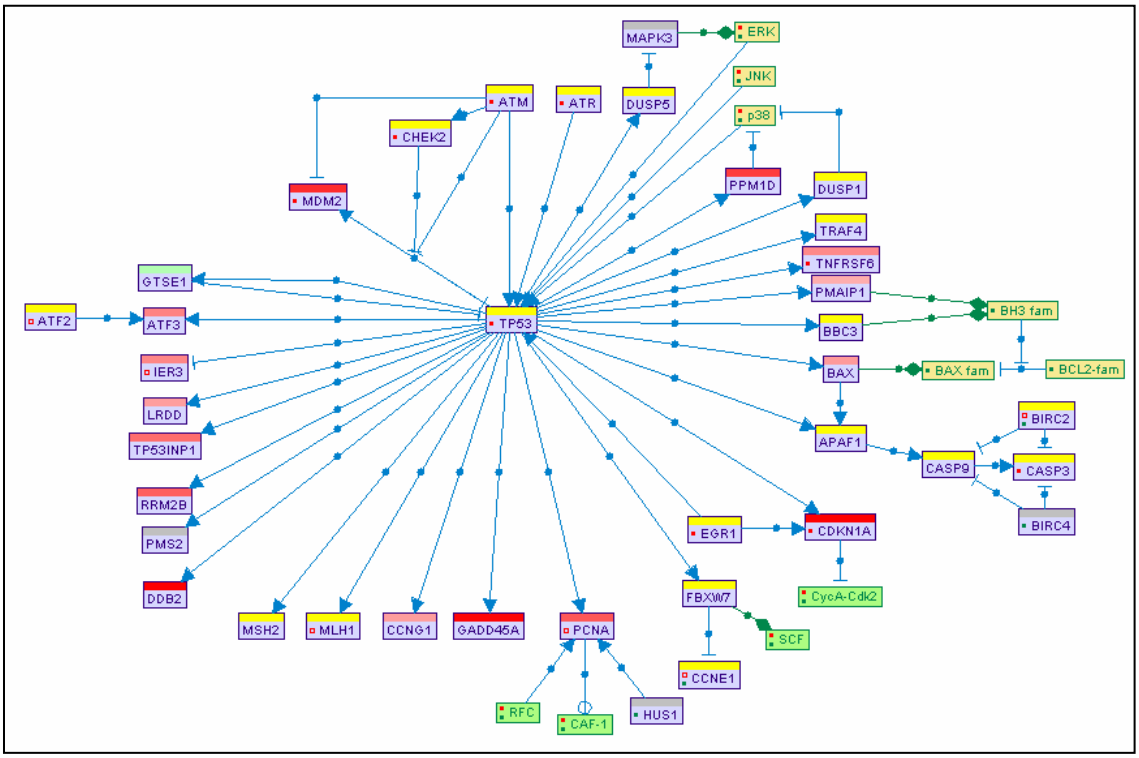



**Figure 4.2.2. Superposition of gene expression data on signaling maps using SHARP.** SHARP enables the user to superimpose functional genomics data on pathway maps, enhancing the interpretation of such data. In this example, the map of the p53-mediated transcriptional network was superimposed with gene expression data measured 4 hrs after exposure of human lymphocyte cells to ionizing radiation. The color of the bar above gene nodes indicates its response in the dataset. Upregulated genes are reddish in color: the darker the read the greater the fold of induction. Downregulated genes are colored green. Genes whose expression was not changed are yellow, and genes for which data are not available (e.g., genes not present on the microarray) are grey. Superimposing the data on this map clearly shows that the p53 network was activated by the examined stress. SHARP also allows superimposition of clustering data: any division of the genes into groups (according to GO annotation, cellular compartment, clustering algorithms, or user-specific definitions) can be viewed on the pathway map (not shown). (Explanation of the different types of nodes and edges displayed in SHARP maps is detailed in the legend of Fig 1.7.1).

**SHARP architecture**. SHARP currently operates as a stand-alone package, and is installed and used by seven research labs serving as beta-sites. The first official release of SHARP 1.0, planned for November 2005, will be based on a decentralized database architecture, in which a local copy of SHARP database is installed in each research lab, and the databases are periodically synchronized. Data exchange between

the local DBs will be based on the SHARP XML scheme that was developed for this and other interoperability goals. SHARP 2.0 will be upgraded to a web-based system, allowing the worldwide ATM and DNA damage research communities to benefit from one central database.

**SHARP – Limitations.** The data model we have devised is intentionally simple, facilitating easy uploading of data by many end-users and rapid implementation of the project. Despite its simplicity, the model can describe most aspects of signal transduction pathways. However, at this stage it does not account for the following elements:

A. Metabolic reactions. SHARP is focused on signaling pathways rather than on metabolic ones. Its model, which is based on the structure "*Source regulates Targe*t", is adequate for describing signal transduction pathways, but not for metabolic cycles where the structure "reactants – enzyme – products" is more suitable.

B. Splice variants. Since there is still no accepted nomenclature for different splice variants originating from the same gene, and since details on the specific variants that participate in reported interactions are usually missing in current publications, we feel it is premature to cover this aspect at this point. It is known, however, that some interactions are carried out only by certain splice variants of a gene and not by others.

C. Cell-type specificity. Currently, SHARP represents a "generic" cell in which all regulations potentially take place. Yet, some interactions may be specific for certain cell types. Again, we felt that current knowledge is not sufficient to add a cell-type layer to the data.

D. Quantitative aspects. Our model is a qualitative one. Any quantitative aspect that might affect the interactions is not modeled (for example, certain interactions are reported only when a physiological trigger rises above a certain threshold).

E. Single organism. SHARP currently supports regulatory data obtained in human cellular systems.

Despite these limitations, SHARP design is general enough to allow its adjustment to support many of the above features in the future, as biological knowledge expands and standard ontologies and nomenclatures are become widely accepted.

## *4.3. Reverse-engineering of transcriptional networks in human cells – analysis of cell cycle regulation as a test case*

This project was carried out in collaboration with Chaim Linhart and Roded Sharan from Prof. Shamir's group. Its results were published in [150], which is attached as Appendix B of this thesis.

Reverse engineering infers regulatory mechanisms from gene expression patterns. Several studies have applied this approach to discover novel transcriptional networks in yeast (see Section 1.5 and [55, 73]). In this study we utilized human genomic sequences, models for binding sites of known transcription factors, and gene expression data to demonstrate for the first time that reverse engineering can disclose transcriptional networks in human cells despite their greater complexity compared to yeast.

In this project we introduced the PRIMA tool for computational promoter analysis (see Section 4.1). Our test case was the transcriptional mechanisms that control cell cycle progression in human cells. Our first task was to use in-silico analysis to identify TFs that cooperate with any particular TF of interest. The scheme of the analysis is as follows: A set of promoters of genes that are directly regulated by the TF of interest (termed 'targets' of this TF) is constructed and scanned for over-represented binding sites corresponding to other TFs. Such over-representations may

point to a functional link between the over-represented TFs and the TF of interest. We used this scheme to ferret out TFs that cooperate with E2F. Since robust statistics require the largest possible set of E2F targets, we used recent results published by Ren et al. [26], who combined ChIP (chromatin immunoprecipitation) and microarray technologies to identify 124 genes whose promoters bind either E2F1 or E2F4 in-vivo. Our 13K set of human promoters (see Section 4.1) contained promoter sequences for 103 of these genes. This set of E2F target promoters was scanned with experimentally-derived position weight matrices (PWMs) for 107 human TFs (PWMs were taken from TRANSFAC database [152]). The occurrence frequencies of each PWM in the E2F target set and in the 13K set, which served as a background set, were compared, and an analytical score computed for the significance of its observed abundance in the E2F target set (see Section 4.1 for details on statistical computations). For PWMs that achieved a highly significant analytical score, we applied an additional empirical test vs. random promoter sets. We determined the occurrence frequency of those high-scoring PWMs on 10,000 subsets of promoters that were randomly chosen from the 13K set and with the same size as the target set (103 promoters). We report only PWMs whose abundance on the E2F target set was significantly higher than on the random sets. The screening criterion that we applied corresponded to $p<0.05$ after accounting for multiple testing. We identified four significantly enriched PWMs in the E2F target set (Table 4.3.1). As expected, the PWM of E2F itself is highly enriched in this set. Since E2F is a true positive in this set, the identification of its PWM demonstrates the sensitivity of our approach for detecting true signals. PWMs of three TFs — NF-Y, CREB and NRF-1 — were also significantly enriched, pointing to possible functional links between these TFs and E2F.

| TF | Number of promoters with hits | Number of hits | Analytical p-val | Rank relative to abundance in random sets |
|---|---|---|---|---|
| E2F (M00516) | 28 | 35 | $1.9 \times 10^{-10}$ | 1 |
| NF-Y (M00185) | 44 | 64 | $1.7 \times 10^{-14}$ | 1 |
| CREB (M00113) | 28 | 41 | $2.5 \times 10^{-5}$ | 1 |
| NRF-1 (M00652) | 32 | 77 | $3.1 \times 10^{-4}$ | 3 |

**Table 4.3.1. Enriched TF PWMs in promoters of E2F target genes**. A set of 103 promoters corresponding to E2F target genes reported by Ren et al. [26] was scanned for over-represented binding sites corresponding to 107 human TF PWMs. Four significantly enriched PWMs were found. Indicated for each one are: the number of promoters with hits of the PWM and the total number of hits of the PWM (some promoters have multiple hits of a PWM), the analytical score for observing such enrichment, and the rank of the PWM's abundance in the E2F target set relative to its abundance in 10,000 sets of randomly selected promoters of the same size as that of the E2F target set. Similarity score thresholds for declaring hits were stringently determined in order to enable identification of real enrichments in the examined set. Therefore, the number of promoters having E2F binding sites in this E2F target set is underestimated. Nevertheless, the observed occurrence rate of E2F is highly significant. Notably, the enrichment of the NF-Y PWM in this set is even more significant than the enrichment of the E2F PWM.

Next, we utilized functional annotation data to delineate regulatory mechanisms. Hughes et al. [77] demonstrated that groups of functionally related genes in *S. cerevisiae* often share common cis-regulatory elements in their promoters. Hence, analyzing promoters of genes with common function could reveal regulatory elements characteristic of specific functional categories. We examined whether this approach could be applied to human promoters, using the functional categorization of human genes provided by the LocusLink DB [151], which adopts the standard Gene Ontology vocabulary for description of biological processes [149]. We focused on four cell cycle-related categories: cell cycle control, mitotic cell cycle, DNA metabolism, and M phase (some genes are assigned to several functional categories, hence the groups are not mutually exclusive). The methodology described above was

applied to each category, again using the 13K set as the background set and scanning with all 107 PWMs. Significantly enriched PWMs were disclosed in all functional categories (Table 4.3.2). The E2F PWM is enriched in all categories, reflecting its central role in regulating these processes. Notably, it is enriched in promoters of genes known to function in the M phase of the cell cycle. This is in accordance with recent studies [163, 164] showing that E2F's role in controlling the cell cycle goes beyond its previously documented control of the entry into the S phase. NF-Y and NRF-1 PWMs are enriched in three out of the four categories, Sp1 PWM is enriched in the cell cycle control and DNA metabolism categories, and ETF and ATF PWMs are enriched in the cell cycle control and the M phase categories, respectively.

| Biological process category | Number of genes | TF | Analytical p-val | Rank relative to abundance in random sets |
|---|---|---|---|---|
| Cell cycle control (GO:000074) | 223 | ETF (M00695) | $1.5x10^{-7}$ | 1 |
| | | E2F (M00516) | $1.5x10^{-6}$ | 1 |
| | | NRF1 (M00652) | $2.5x10^{-5}$ | 1 |
| | | Sp1 (M00196) | $2.5x10^{-4}$ | 4 (2) |
| Mitotic cell cycle (GO:0000278) | 175 | E2F (M00516) | $1.4x10^{-9}$ | 1 |
| | | NF-Y (M00185) | $1.3x10^{-4}$ | 1 (2) |
| | | NRF1 (M00652) | $1.6x10^{-4}$ | 1 |
| DNA metabolism (GO:0006259) | 240 | E2F (M00516) | $6.7x10^{-5}$ | 1 |
| | | NF-Y (M00185) | $4.6x10^{-4}$ | 4 (2) |
| | | Sp1 (M00196) | $6.8x10^{-4}$ | 5 (5) |
| M phase (GO:0000279) | 100 | NRF1 (M00652) | $5.9x10^{-6}$ | 1 |
| | | NF-Y (M00185) | $2.5x10^{-4}$ | 2 (2) |
| | | ATF (M00338) | $3.4x10^{-4}$ | 4 (5) |
| | | E2F (M00516) | $3.8x10^{-4}$ | 1 |

**Table 4.3.2. Enriched TF PWMs in promoters of genes that function in the cell cycle.** Four categories related to cell cycle, containing a total of 672 distinct genes, were analyzed (certain genes are assigned to several categories; hence the categories are not mutually exclusive). The number of promoters and the TF PWMs significantly enriched in each category are indicated. Indicated for each over-represented PWM are the analytical score for observing such enrichment, and the rank of the PWM's abundance in the functional category relative to its abundance in 10,000 sets of randomly selected promoters of the same size as that of the functional category set. Numbers in parentheses represent the number of random sets in which the PWM was equally abundant as in the functional category set.

Our next step was to apply the reverse engineering approach to infer transcriptional regulatory mechanisms from gene expression data. We analyzed a human cell cycle dataset published recently by Whitfield et al. [21]. Their study recorded genome-wide gene expression levels over multiple time points during the progression of the cell cycle in HeLa human cell line; 874 genes showed periodic expression patterns over several cell cycles. Our 13K promoters set contained putative promoter sequences for 568 of these genes. Whitfield et al. [21] partitioned the cell cycle-regulated genes according to their expression periodicity patterns into five clusters, corresponding to cell cycle phases G1/S, S, G2, G2/M and M/G1. We analyzed clusters of 103, 105, 122, 145 and 93 promoters, respectively.

We searched for significantly enriched PWMs in the entire set of 568 cell cycle-regulated promoters using the 13K set as the background set. Six out of the 107 PWMs, corresponding to E2F, NF-Y, NRF-1, Sp1, ATF and CREB TFs, were significantly over-represented in this target set (Table 4.3.3*a*). We then searched for PWMs enriched only in specific phase clusters and found that Arnt and YY1 PWMs were specifically enriched in the G1/S and the M/G1 clusters, respectively (Table 4.3.3*b*). Caution must be exercised when examining whether PWMs that were enriched in the entire set favor any specific phase cluster. Given their significant over-representation in the entire set, random partitions of the dataset are also expected to yield clusters where these PWMs are enriched with respect to their genomic prevalence. So, what should be tested is whether these PWMs favor any specific phase cluster given their prevalence in this dataset, rather than their genomic background prevalence. Hence, in this examination, the set of 568 cell cycle-regulated promoters was used as the background set. E2F PWM was found to be significantly over-represented in the G1/S and S phases ($p=3.2*10^{-7}$ for the observed prevalence in

these 2 clusters together) and under-represented in the M/G1 cluster (p=0.015); NF-Y PWM was over-represented in the G2 and G2/M phases (p=0.0096 for the observed prevalence in these 2 clusters together); and Sp1 PWM slightly favored the G1/S cluster (p=0.02). NRF-1, ATF and CREB PWMs were more uniformly distributed and showed no bias for any particular phase (Fig. 4.3.1).

***a***

| TF | Number of promoters with hits | Number of hits | Analytical p-val | Rank relative to abundance in random sets |
|---|---|---|---|---|
| NF-Y (M00185) | 152 | 203 | $1.2x10^{-11}$ | 1 |
| E2F (M00516) | 78 | 92 | $1.2x10^{-8}$ | 1 |
| NRF1 (M00652) | 127 | 234 | $3.3x10^{-6}$ | 1 |
| Sp1 (M00196) | 223 | 365 | $1.3x10^{-4}$ | 1 |
| ATF (M00338) | 113 | 162 | $5.3x10^{-4}$ | 2 |
| CREB (M00113) | 91 | 117 | $9.3x10^{-4}$ | 2 (1) |

***b***

| TF | Number of promoters with hits | Number of hits | Cell cycle phase | Analytical p-val | Rank relative to abundance in random sets |
|---|---|---|---|---|---|
| Arnt (M00236) | 33 | 37 | G1/S | $5.1x10^{-4}$ | 5 (4) |
| YY1 (M00069) | 20 | 25 | M/G1 | $8.1x10^{-4}$ | 5 (3) |

**Table 4.3.3. Enriched TF PWMs in promoters of cell cycle regulated genes.** *a*. A set of 568 promoters of cell cycle-regulated genes scanned for over-represented TF PWMs, disclosing six significantly enriched PWMs. Information for each PWM is the same as in Table 4.3.1. *b.* Whitfield et al. [21] partitioned the cell cycle-regulated genes according to their expression periodicity patterns into five clusters corresponding to different phases of the cell cycle. When the promoter sequences of these clusters were scanned for enriched PWMs, two PWMs were enriched in a specific phase cluster, but not in the 568 set as a whole.

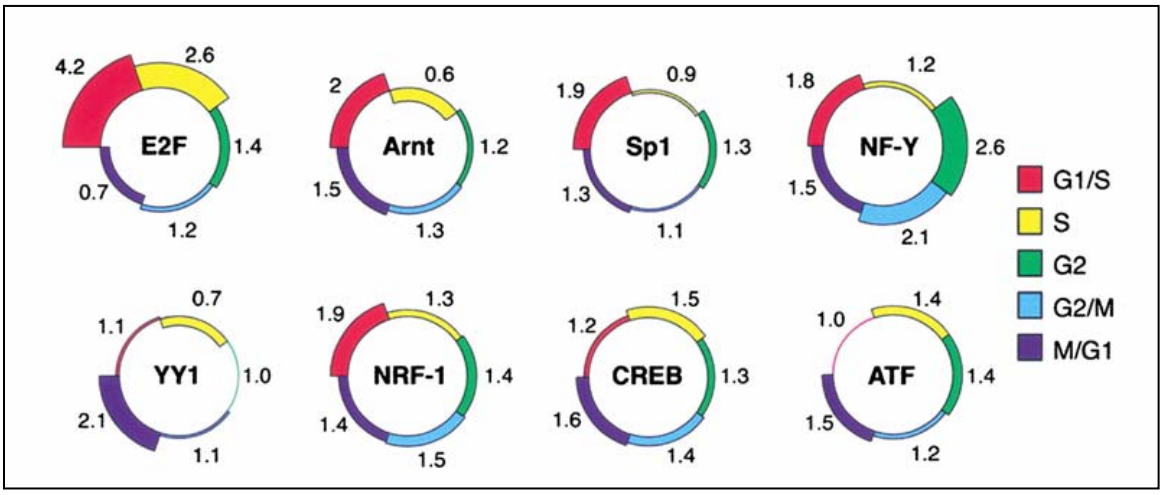
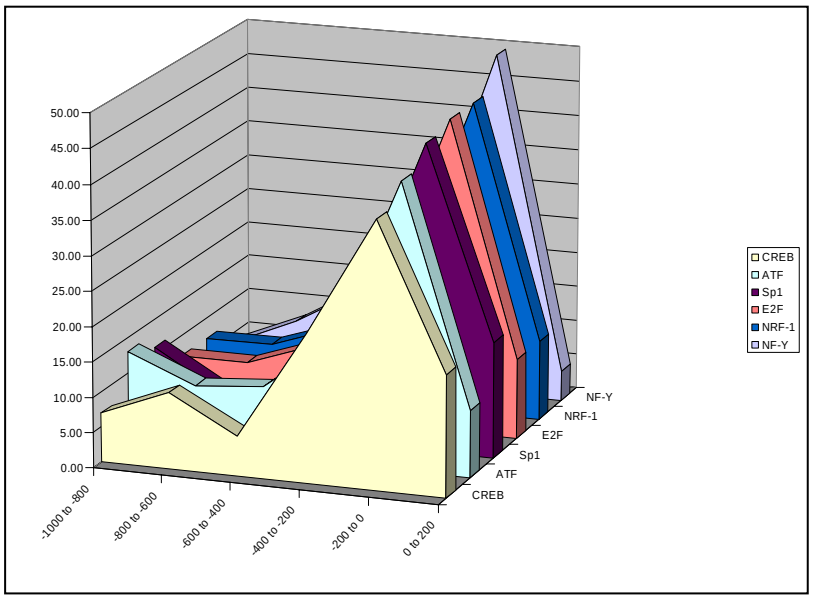



**Figure 4.3.1. TFs whose binding site signatures are highly enriched on promoters of cell-cycle regulated genes.** The circles correspond to the eight PWMs that are highly enriched in promoters of cell-cycle-regulated genes (Table 4.3.3). Each circle is divided into 5 zones, corresponding to the phase clusters. The number adjacent to the zone represents the ratio of the TF hits prevalence in promoters contained in each of the cell-cycle phase clusters to its prevalence in the set of 13K background promoters (e.g., E2F hits were 4.2-fold more prevalent on promoters of genes that peak at G1/S phase). Note that several TFs show a tendency towards specific cell-cycle phases: e.g., over-representation of the E2F PWM in promoters of the G1/S and S clusters, and its under-representation in promoters of the M/G1 cluster.

Next, we examined the location distribution of the computationally identified putative binding sites of the enriched PWMs. The putative binding sites for E2F, NF-Y, NRF-1, Sp1, ATF and CREB tend to concentrate in the proximity of the TSS (Fig 4.3.2). This observation is in agreement with experimental data on the locations of *in-vivo* binding sites of E2F [165] and NF-Y [166]. In addition to the fact that the positions of the computationally identified hits are not uniformly distributed, but rather concentrate near the TSSs, we also observed that their occurrence rate declines sharply downstream of the putative TSSs. These observations provide an additional indication of both the high quality of the putative human promoters that we used and of PRIMA's specificity in detecting TF hits (a high false positive rate would result in a uniform hit distribution as expected for random hits).

**Figure. 4.3.2. Location distributions of the computationally identified hits peak at TSS.** The figure presents the distribution of locations of TFs putative binding sites found by PRIMA in the 568 cell-cycle-regulated promoters. Promoters were divided into six intervals, 200 bp each. For each of the PWMs listed in Table 4.3.3, the number of times its computationally identified binding sites appeared in each interval was counted (after accounting for the actual number of bps scanned in each interval; this number changes as the masked sequences are not uniformly distributed among the six intervals), and normalized by the total number of hits. The Y-axis represents the percentage of hits in each interval for each PWM. Locations of NRF-1, CREB, NF-Y, Sp1, ATF and E2F binding sites were significantly concentrated in the vicinity of the TSSs (chi-square test, $p<0.01$). The sharp peaks of these distributions point to both the specificity of PRIMA-identified TF hits and the high quality of TSS annotations in the human genome.




The approaches applied thus far identified TF PWMs that were enriched in target sets of promoters, with the tests performed separately on each PWM. Finding several enriched PWMs on the same target set may indirectly point to functional links between the corresponding TFs. We sought a direct method to test the associations between distinct PWMs. In an effort to identify pairs of PWMs that exhibit a significant tendency to appear together in the same promoters, we examined whether the prevalence of promoters containing hits for two PWMs was significantly higher than would be expected if the PWMs occurred independently. This analysis was applied to the set of 568 promoters of cell cycle-regulated genes. We examined all possible pairs formed by the 9 PWMs found to be enriched in any of the analyses reported above. Eight pairs showed a significant tendency to co-occur in this promoter set. Each such pair constitutes a hypothetical regulatory module, or a part thereof.

Figure 4.3.3 suggests that NRF-1, Sp1, ETF and E2F may constitute transcriptional modules of higher orders, i.e., recurrent motifs of three or four TFs.

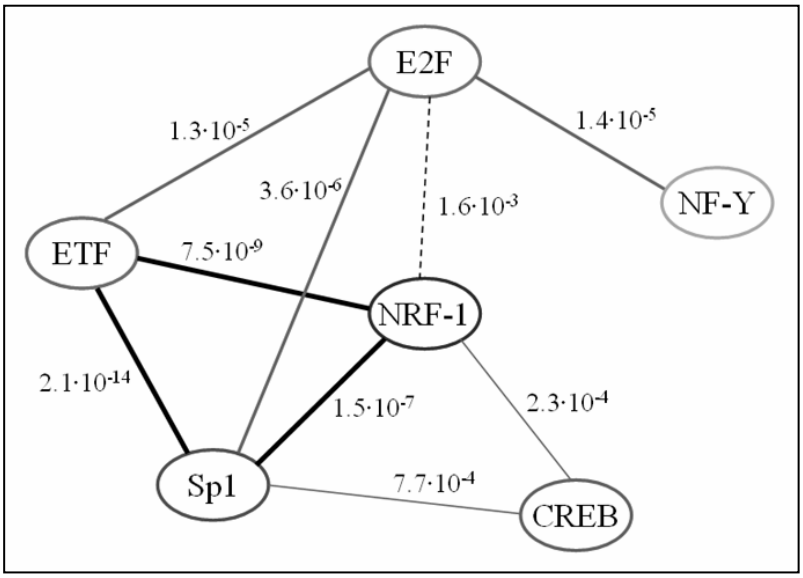



**Figure. 4.3.3. Pairs of PWMs that co-occur significantly in promoters of genes regulated in a cell cycle-dependent manner**. We examined whether the nine PWMs reported in Tables 4.1.1-3 can be organized into regulatory modules. Eight significant pairs were identified, each connected by an edge. The corresponding p-value is indicated next to the edge. The edge connecting the E2F-NRF1 pair is dashed to indicate that its borderline significance

Our results elucidate some novel aspects of the transcriptional regulation of cell cycle progression. They were derived by computational means and as such should be regarded as hypotheses on regulatory relationship between the enriched TFs and the progression of the cell cycle that requires empirical validation. However, there is evidence strongly supporting most of our findings, and is discussed in Section 5. The methods we presented in this test case study are general and can be applied to the analysis of transcriptional networks controlling any biological process.

Since the publication this project [150] in 2003, two other works analyzed the same human cell cycle dataset, aiming to identify TFs that control cell cycle progression. Sharan et al. [92] sought to identify human-mouse conserved cis-regulatory modules that are over-represented on cell-cycle regulated promoters. They report on seven such modules, together composed of ten different PWMs. Only the strongest signals (i.e., enrichment of E2F and NF-Y binding signatures) were reported by both Sharan and our study. ZF5 (PWM M00716), which appears in four out of the

seven modules reported by Sharan et al., was added to TRANSFAC DB only after our analysis was carried out. Testing ZF5 using PRIMA indeed finds that this TF is highly enriched on cell cycle-regulated promoters ($p<10^{-11}$). USF2 (M00726) passes Sharan's filtering threshold (p=0.05) but not ours (p=0.001), yet it gets by PRIMA an enrichment score of 0.02. DELTAEF1 was defined using chicken binding sites and therefore was not included in our analysis. The other five PWMs reported by Sharan et al. are not enriched in PRIMA tests.

Several reasons can account for the different findings. Sharan et al. applied a very different methodology, aimed at identification of conserved modules. Therefore, they sought for 1) conservation of TF hits between human and mouse promoters and 2) significant co-occurrence of hits at spatially restricted intervals. The requirement for human-mouse conservation was demonstrated to reduce false positive discoveries in the identification of regulatory elements [81, 82, 167]. However, low conservation is observed in many known regulatory regions [168-171]. In such cases, imposing the requirement for evolutionary conservation can lead to missing functional sites. Indeed, only 336 cell-cycle regulated promoters that contained human-mouse conserved segments of at least 80 bp were included in Sharan's analysis, compared to 568 promoters that were included in our analysis. The higher number of promoters gives our analysis higher statistical power in detecting real over-representation of signatures.

Dieterich et al. [172] applied another methodology, seeking for TF signatures that are enriched on cell-cycle regulated promoters and that are conserved between human and mouse. There was no overlap between the 22 enriched PWMs reported in this study and ours, and only one overlap with Sharan's results. I believe that the major reason for these disparate findings is the different promoter region analyzed by

Dieterich et al.: In this study, up to 12,000 bp were scanned in each promoter, compared to 1,200 bp in [92] and in our analysis. This increase in the search area by an order of magnitude greatly reduces the signal-to-noise ratio. It should be noted that E2F and NF-Y signals, which are well-known master regulatory signals of cell cycle progression, were not detected by Dietrich et al., again probably due to the large promoter region analyzed in this study (E2F and NF-Y signals are strongly biased to the close proximity of the TSS). In a recent study, we demonstrated that E2F signal is conserved in promoters of G1/S genes in all organisms from worm to human, and that NF-Y signal is conserved in promoters of G2/M regulated genes in vertebrates [173].

## *4.4. Computational identification of transcriptional modules – c-Myc as a test case*

This project was carried out in collaboration with Chaim Linhart of Prof. Shamir's group and Karen Zeller of Prof. C.V. Dang's lab at the Cancer Center, Johns Hopkins University School of Medicine. Its results were published in [174], which is attached as Appendix C of this thesis.

In this study we further demonstrated how ChIP-on-chip data can be computationally utilized to discover novel functional links between transcription factors based on significant co-occurrence of their binding site signatures on common target promoters. We focused on the oncoprotein c-Myc and aimed at identifying TFs that form recurrent cis-regulatory modules with it. To this end, we analyzed the data reported by Li et al. [28]. This study applied the ChIP-on-chip technique and identified 776 human genes in Burkitt's lymphoma cells whose promoters are bound by c-Myc and its heterodimer partner Max.

c-Myc regulates cell cycle proliferation, apoptosis and differentiation. Overexpression of c-Myc is one of the most common alterations in human cancer, yet it is not clear how it promotes malignant transformation [175-178]. It is widely accepted that transcriptional regulation activity of c-Myc is critical for development of malignancy associated with it, but the target genes that mediate this process remain elusive. Identification of TFs that are functionally linked to c-Myc can provide novel clues to the mechanisms by which its oncogenic form drives cells into unbalanced proliferation.

First, we extracted genome-wide promoter sequences from the human and mouse genomes (version 13 released by Ensembl in Dec. 2002). For each gene flagged as a 'known gene', a genomic sequence was extracted that spanned 1,000 bp upstream to 200 bp downstream of the gene's putative TSS. The extracted sequences were masked for repetitive elements. A total of 19,351 and 18,748 putative promoters were extracted from the human and mouse genome, respectively. To avoid biases due to highly similar promoters, we constructed for each organism a non-redundant set of promoters by running all-against-all BLAST comparisons. For every promoter pair with BLAST E-score $< 10^{-50}$, we excluded one of the members from the non-redundant set. The non-redundant human and mouse promoter sets contain 17,390 and 15,521 promoters, respectively.

In order to identify enriched TF signatures in promoters bound by c-Myc, we applied PRIMA to a dataset recently published by Li et al., who used ChIP-on-Chip to identify 876 human promoters that were bound by c-Myc, 931 promoters that were bound by Max, and 776 promoters that were bound by both TFs in Burkitt's lymphoma cells. In order to reduce false positives among the reported bound promoters, we focused on the set of promoters that were reported to be bound by both

c-Myc and Max, hereafter referred to as the c-Myc/Max target set. Our collection of human promoter sequences contains a total of 19,351 promoters, of which a subset of 17,390 promoters was defined by us as non-redundant. The full and non-redundant promoter sets include sequence data for 615 and 519 genes, respectively, out of the 776 genes in the Myc/Max target set. We applied PRIMA to scan these 519 non-redundant promoters of the Myc/Max target set for over-represented TF binding site signatures. This means that we searched for TFs whose binding site signatures are significantly more abundant on this promoter set than expected by chance, given their abundance on the entire collection of non-redundant human promoters. Such over-representation suggests the existence of functional links between the over-abundant TFs and c-Myc/Max. Three hundred PWMs that represent mammalian TF binding sites (obtained from the TRANSFAC database) were tested by PRIMA, and nine of them were significantly enriched in the c-Myc/Max target set ($p\_value<10^{-5}$, and after conservative adjustment for multiple testing $p<0.05$, Table 4.4.1A). We observed that PWMs that were enriched on both the human and its mouse homolog sets tend to show similar distributions of hit location (i.e., those PWMs that show picked distribution in human, also show such pattern on mouse promoters). However, in general, the location of specific hits in homolog promoters was not conserved

Previously, we used PRIMA to analyze human genes whose expression is cell cycle dependent (Section 4.3, [150]). Interestingly, most of the TFs whose signatures were enriched in the cell cycle-dependent promoter set were also enriched in the c-Myc/Max target set (E2F, NF-Y, Sp1, NRF1, ETF, CREB and AhR/Arnt). We therefore checked the overlap between the cell cycle and the c-Myc/Max target sets. The cell cycle set contains 568 genes, only 30 of which are common to the c-Myc/Max set. When we analyzed the c-Myc/Max set after deleting these 30 genes, the

over-representation of all the TFs reported in Table 4.4.1 remained significant. Thus, the overlap between the results obtained on the two datasets is not explained by common genes, and is probably due to a general role of these regulators in cell cycle progression, where they control different sets of genes in different cell types. The mitogen- and stress-induced ELK-1 and EGR-1 TFs were also enriched on the Myc/Max target set.

Li et al. reported, somewhat surprisingly, that only 25% of the promoters bound by c-Myc/Max contain a core E-box motif (CACGTG), which is directly recognized and bound by the c-Myc/Max heterodimer. In accordance with this observation, PRIMA found that PWMs with canonical E-box core motifs are only slightly enriched on the *Myc/Max target set*, and they did not pass our significance threshold. However, another PWM that models c-Myc/Max binding sites, whose core motif is a variant of

**Table 4.4.1. TF binding site signatures enriched in c-Myc/Max target sets**

| TF [PWM accession ID in TRANSFAC DB] | A. p-value for PWM hits' abundance on Myc/max target set | B. p-value for PWM hits' abundance on mouse ortholog set | C. p-value for PWM hits' abundance on mode-1 subset |
|---|---|---|---|
| ETF [M00695] | $1.2x10^{-15}$ | $6.8x10^{-13}$ | $3.2x10^{-7}$ |
| Sp1 [M00196] | $1.7x10^{-14}$ | $8.9x10^{-12}$ | $6.3x10^{-10}$ |
| Nrf-1 [M00652] | $6.5x10^{-14}$ | $7.9x10^{-11}$ | $3.2x10^{-7}$ |
| NF-Y [M00185] | $3.2x10^{-12}$ | Not enriched | $1.5x10^{-5}$ |
| CREB [M00177] | $4.7x10^{-8}$ | $1.4x10^{-7}$ | Not enriched |
| c-Myc/Max [M00322] | $1.5x10^{-7}$ | $2.8x10^{-6}$ | $5.7x10^{-62}$* |
| Egr-1 [M00243] | $2.4x10^{-7}$ | $6.7x10^{-5}$ | $3.4x10^{-8}$ |
| Elk-1 [M00025] | $3.9x10^{-7}$ | $5.5x10^{-11}$ | Not enriched |
| E2F [M00516] | $5.8x10^{-7}$ | $6.6x10^{-5}$ | Not enriched |
| AhR/Arnt [M00237] | $6.4x10^{-7}$ | Not enriched | Not enriched |
| E-Box | $8.4x10^{-4}$ | $4.1x10^{-5}$ | $6.5x10^{-7}$ |

*c-Myc/Max signature is markedly enriched on 'mode1' subset, since by definition this set comprises the genes whose promoters were found to have a hit for cMyc/Max. Not enriched indicates p-value > 0.0001 (p-values are before correction for multiple testing).

the E-box (CAYGYG, Y=[C or T]), was significantly enriched on this target set (PWM M00322 in TRANSFAC DB) (Table 4.4.1A). It cannot be ruled out that we failed to detect E-box over-representation because a fraction of these regulatory elements are located outside the analyzed promoter region.

The enrichment of binding site signatures of specific TFs in the c-Myc/Max target set raises the possibility that c-Myc/Max and these TFs maintain functional relationships and together form recurrent transcriptional regulation modules that control the expression of numerous genes. In order to strengthen this *in-silico*-derived hypothesis, we repeated the tests for TF binding site enrichment, this time on the set of promoters comprising the mouse orthologs of the human c-Myc/Max target set. Extracting promoter sequences for all known mouse genes, we collected 18,478 mouse promoter sequences, of which 15,521 were non-redundant. The non-redundant set includes sequence data for 407 mouse orthologs of the human Myc/Max target set. Applying PRIMA to this set, we found that seven out of the nine PWMs that were enriched in the human set were also enriched in the mouse set (Table 4.4.1B). The enrichment of the same TF binding sites on the ortholog sets suggests that the functional relationships between these TFs are conserved between mice and humans.

As noted above, a canonical E-box was found in only about 25% of the promoters identified by Li et al. as direct targets of c-Myc/Max. Therefore, it was suggested that the c-Myc/Max heterodimer controls its target by two different modes: in the first one it directly binds its target promoters through its classical E-box element or a variant thereof; in the second mode it participates in the regulation of target promoters without binding directly to the DNA, but by physically interacting with other sequence specific TFs or with components of the general transcription machinery.

We attempted to identify TFs that form cis-regulatory modules with c-Myc/Max via the first mode. To this end we identified a subset of the c-Myc/Max target set, comprising genes whose promoters contain high scoring putative binding sites (*hits*) for c-Myc/Max. As the c-Myc/Max variant binding element represented by PWM M00322 was more enriched than the canonical E-box, we scanned the c-Myc/Max target set for promoters with hits for this PWM. Out of 615 promoters, we identified hits for M00322 in 134, making them good candidates for being regulated by c-Myc/Max through its direct DNA binding. We refer to this subset as 'c-Myc/Max mode1 subset'. Since it is smaller than the complete c-Myc/Max target set, we expected statistical phenomena associated with it to be less significant. However, we observed that one TF, EGR-1, was even more significantly enriched in the mode1 subset than in the complete set (Table 4.4.1C). This makes EGR-1 a strong candidate for being c-Myc/Max's partner in regulation of mode1 target promoters. To determine whether the increased abundance of EGR-1 hits in the mode1 subset compared to the entire c-Myc/Max target is statistically significant, we ran PRIMA using mode1 subset and the c-Myc/Max set as target and background sets, respectively. The over-representation of EGR-1 hits in the mode1 subset was significant ($p=0.0034$).

Of note, E2F, a pivotal regulator of the transcriptional program associated with cell cycle progression, was not enriched in the mode1 subset, suggesting that the c-Myc/Max cooperation with E2F is maintained mainly through mode2.

As noted above, the EGR-1 binding site signature was more enriched in the c-Myc/Max mode1 subset than in the complete c-Myc/Max set. Inspection of EGR-1 PWM (M00243) showed that it has a high GC content with a GCGTGGG core. This led us to compare the GC-content of the mode1 subset, the c-Myc/Max target set and the full collection of non-redundant human promoters (Table 4.4.2). We observed that

the c-Myc/Max target set is more GC-rich than the full set of human promoters (57% vs 53%, respectively, z-score=9.4), and the c-Myc/Max mode1 subset is even more GC-rich (60.1%, z-score=5.8 when compared to GC content of the c-Myc/Max target set). This raised the question whether the higher abundance of hits for EGR-1 in the mode1 subset could be explained merely by its high GC-content. To address this concern, we generated random PWMs based on the EGR-1 PWM in a way that preserved its GC content. We generated the random PWMs by permuting the columns of the original PWM, and randomly interchanging A with T and G with C. We then compared the enrichment of five permuted and the original EGR-1 PWM on the c-Myc/Max mode1 subset. Significantly, the original EGR-1 PWM yielded an enrichment score that was far more significant than the scores obtained by the random PWMs generated based on it (Table 4.4.3). For further examination, we sorted all TRANSFAC mammalian PWMs according to their GC-content and recorded their enrichment in the mode1 subset. Table 4.4.4 lists the results for the top 25 GC-rich PWMs. This list shows that the over-representation of EGR-1 in the mode1 subset is not merely a reflection of its high GC-content, as there are many PWMs at least as GC-rich as EGR-1 that are not at all enriched on this subset.

**Table 4.4.2. High GC content of the c-Myc/max mode-1 subset**

| | Number of promoters | %A | %C | %G | %T |
|---|---|---|---|---|---|
| Non-redundant human promoters | 17,390 | 23.1 | 26.4 | 27.0 | 23.5 |
| c-Myc/Max target set | 615 | 21.5 | 28.3 | 28.7 | 21.6 |
| c-Myc/Max mode-1 subset | 134 | 19.5 | 30.4 | 30.5 | 19.7 |

**Table 4.4.3. Enrichment of the original and permuted EGR-1 PWM on the c-Myc/Max mode-1 subset**

| PWM | EGR-1 | EGR-1 Rand1 | EGR-1 Rand2 | EGR-1 Rand3 | EGR-1 Rand4 | EGR-1 Rand5 |
|---|---|---|---|---|---|---|
| p-val | $4.5x10^{-7}$ | 0.17 | 0.28 | 0.25 | 0.088 | 0.0034 |

**Table 4.4.4. Enrichment of top 25 GC-rich PWMs on the c-Myc/Max mode-1 subset**

| TF | PWM ACCNUM (TRANSFAC DB) | PWM GC content | Enrichment on mode-1 subset |
|---|---|---|---|
| ETF | M00695 | 0.87 | 6.92E-09 |
| MAZ | M00649 | 0.82 | 8.66E-02 |
| AP-2gamma | M00470 | 0.82 | 1.13E-04 |
| AP-2alpha | M00469 | 0.80 | 2.81E-02 |
| AP-2 | M00189 | 0.79 | 1.80E-04 |
| Sp1 | M00196 | 0.79 | 6.29E-10 |
| Nrf-1 | M00652 | 0.76 | 3.21E-07 |
| c-Myc/Max | M00322 | 0.75 | 5.74E-62 |
| ----- | M00051 | 0.74 | 1.47E-02 |
| LBP-1 | M00644 | 0.74 | 9.65E-02 |
| E2F-1 | M00428 | 0.74 | 9.12E-01 |
| USF2 | M00726 | 0.72 | 6.52E-07 |
| Sp1 | M00008 | 0.72 | 1.59E-03 |
| Egr-3 | M00245 | 0.72 | 1.86E-04 |
| HEB | M00698 | 0.70 | 3.92E-01 |
| E2F-1 | M00431 | 0.70 | 3.37E-02 |
| MTF-1 | M00650 | 0.70 | 6.80E-01 |
| Egr-2 | M00246 | 0.69 | 2.14E-02 |
| MZF1 | M00083 | 0.68 | 7.21E-01 |
| **Egr-1** | **M00243** | **0.68** | **3.36E-08** |
| LF-A1 | M00646 | 0.67 | 7.45E-01 |
| cMyc-E-Box | M99996 | 0.67 | 6.52E-07 |
| Sp3 | M00665 | 0.67 | 4.06E-02 |
| NF-muE1 | M00651 | 0.67 | 2.11E-01 |
| GABP | M00341 | 0.66 | 1.61E-04 |

High scoring hits for c-Myc/Max (M00322) and EGR-1 (M00243) were found in 134 and 167 promoters, respectively, out of a total of 615 promoters in the c-Myc/Max target set. Fifty-four promoters contained strong hits for both c-Myc/Max and EGR-1. Finally, we examined the location distribution of c-Myc/Max and EGR-1 hits on the promoters of the c-Myc/Max target set. For both TFs, the computationally-

identified binding sites are significantly concentrated in the proximity of the TSS and their density drops downstream of it (Fig 4.4.1). In contrast, the distribution of hits identified for a random PWM generated by permuting the EGR-1 PWM was quite uniform between -650 bp to +200 bp with respect to the TSS, as expected for random hits. The fact that the hit distributions for cMyc/Max and EGR-1 show a prominent peak in the anticipated position is an additional indication of the quality of the human genome TSS annotations (such peaks would have not been obtained if significant deviations had existed between the locations of the annotated and real TSS in a large proportion of the genes), and of the specificity of the hits identified by PRIMA (high false positive rates for hits identified by PRIMA would have obscured the peak of true hits).

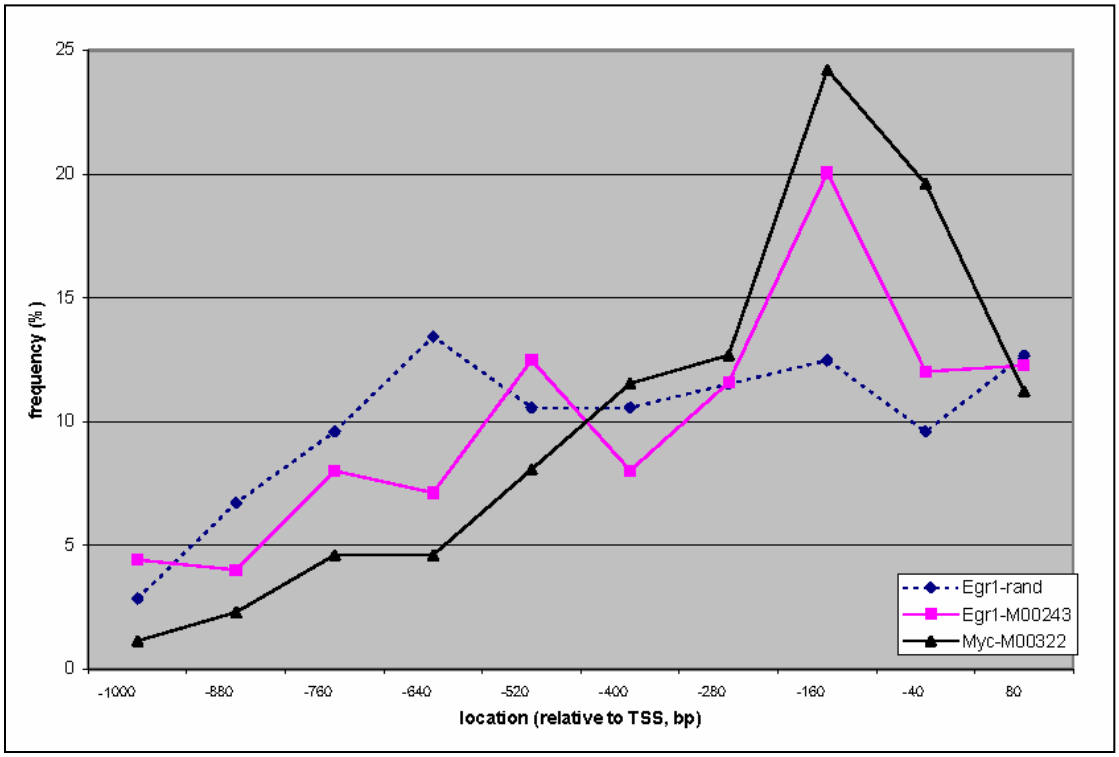



**Figure 4.4.1. Location distribution of hits for c-Myc/Max and EGR-1 on promoters of the cMyc-Max target set.** The promoter region spanning 1000 bps upstream to 200 bps downstream of the TSS was divided into 10 bins of 120 bps. The graph represents the relative frequency of hits over the bins for cMyc/Max (M00322), EGR-1 (M00243), and for a random PWM derived from the EGR-1 PWM as explained in the text. The number of hits in each bin was normalized by the effective sequence length scanned in the bin (effective lengths can be different in different bins due to masking of repetitive elements in promoters).

Fifty-four promoters of the cMyc/Max mode1 subset contained strong hits for EGR-1. To experimentally test our *in-silico* derived hypothesis that these two TFs together form a recurrent transcriptional cis-regulatory module, our collaborator K. Zeller in the laboratory of Prof. C.V. Dang (Cancer Center, Johns Hopkins University School of Medicine) selected six of these genes and examined the binding of both c-Myc and EGR-1 to their promoters using chromatin immunoprecipitation assays (ChIP). For all 6 genes examined (*RAP2B*, *KHSRP*, *PolH*, *PTPN1*, *PP* and *KPNA3*), the signals obtained for both c-Myc and EGR-1 were above the background level (p value < 0.025, one tail t-test, Fig 4.4.2). For a negative control, we chose from the c-Myc/Max target set the *MCCC2* gene, whose promoter did not contain any hit for EGR-1. The signal obtained by the ChIP assay for EGR-1 binding to this promoter was very close to background level. These results demonstrate that c-Myc and EGR-1 co-binding occurs on multiple promoters. Further experiments are required to establish a direct functional link between these two transcriptional regulators.

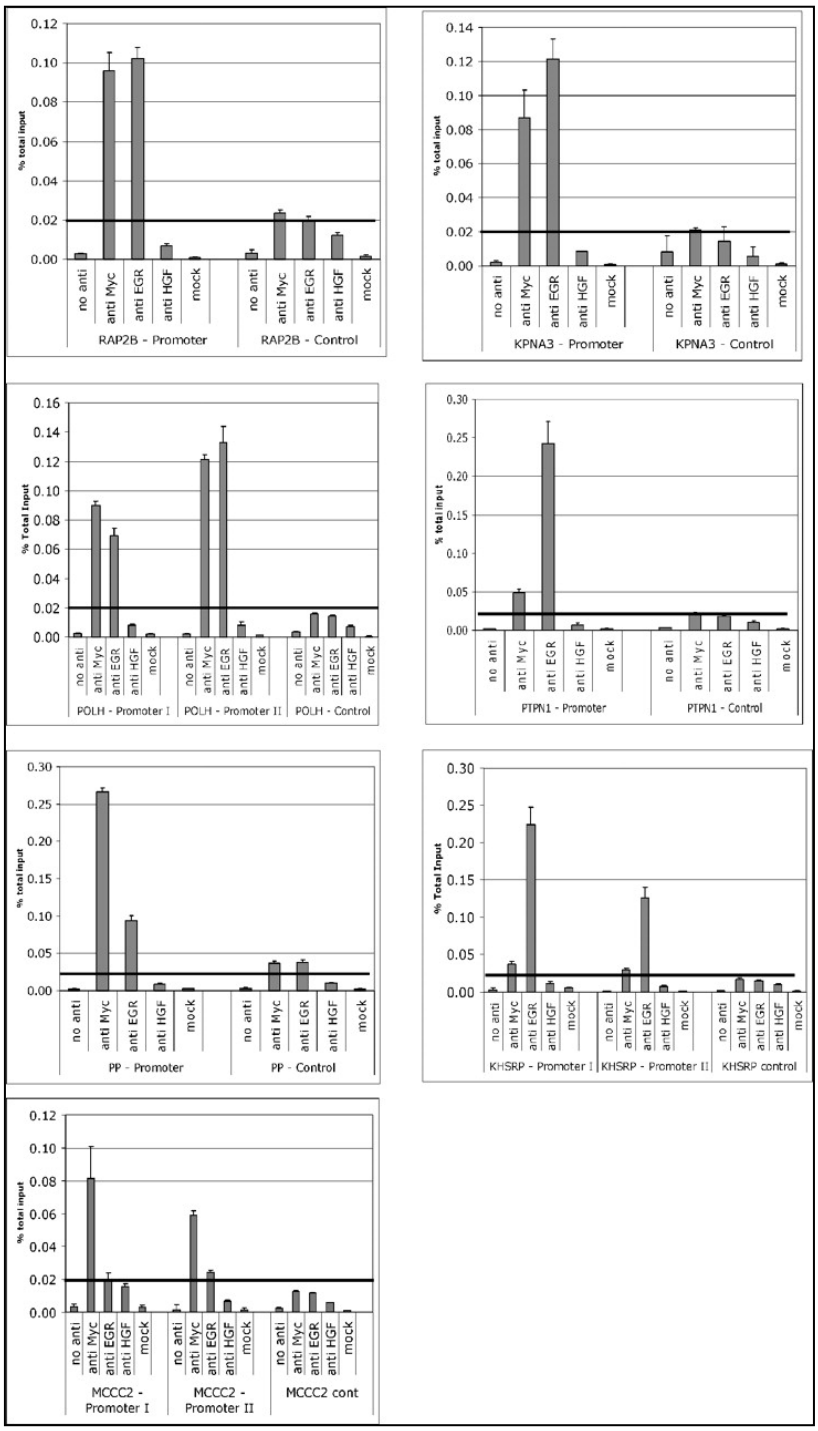



**Figure 4.4.2. Chromatin immunoprecipitation of c-Myc and EGR-1 targets.** Each graph represents real-time PCR amplification of the promoter and control regions of each gene using anti-Myc, anti-EGR-1, anti-HGF, and no antibody-precipitated chromatin as template. Bars represent the percentage of total input DNA for each ChIP sample averaged over 3 PCR reactions. Error bars represent one standard deviation. The solid horizontal line represents 0.02% total input DNA, the background signal for this assay. The signals obtained for the binding of c-Myc and EGR-1 to the promoter regions of all 6 genes examined (*RAP2B*, *KHSRP*, *PolH*, *PTPN1*, *PP* and *KPNA3*), but not for the negative control, MCCC2, were above the background level (p value < 0.025, one tail t-test). (Experiments carried out by K. Zeller in the laboratory of Prof. C.V. Dang (Cancer Center, Johns Hopkins University School of Medicine).

## 4.5. Dissection of a DNA damage-induced transcriptional network using a combination of microarrays and RNAi

This project was carried out in collaboration with Sharon Rash-Elkeles, Yaniv Lerenthal, and Tamar Tenne of Prof. Shiloh's group, Chaim Linhart of Prof. Shamir's group, and Dr. Ninette Amariglio and Prof. Gideon Rechavi of the Functional Genomics Unit at the Sheba Medical Center. Its results were published in [179], which is attached as Appendix D of this thesis.

In this test-case study, we applied two of the most prominent functional genomics technologies, gene expression microarrays and RNA interference (RNAi), to demonstrate that this combined experimental approach can yield accurate dissections of transcriptional networks induced in a human cellular system. In this experiment, cellular systems knocked-down for key regulators of the DNA damage induced network were established by Yaniv Lerenthal, and microarray hybridizations and RT-PCR validations were carried out by Sharon Rashi-Elkeles from our lab.

Soon after the discovery of the RNAi phenomenon and its utilization as an experimental tool to modulate gene activity, it was realized that the combination of global gene expression profiling and RNAi-mediated silencing of key regulatory genes offers a powerful method for systematic dissection of transcriptional networks in mammalian systems. However, recent studies pointed out that applying RNAi to mammalian cells triggers some nonspecific pathways [180-182] and affects an unpredicted number of off-targets [183] in addition to knocking-down the target of interest. This raised the concern that nonspecific responses to short interfering RNA (siRNA) might obscure the consequences of silencing the target of interest.

In this work where we focused on a DNA damage-induced transcriptional network as a test case, we established human cellular systems stably knocked-down for the ATM protein kinase, for the Rel-A subunit of NF-κB, and for p53. Stable knock-down of the proteins was obtained by infecting HEK293 cells with retroviral vectors expressing the corresponding short hairpin RNAs (shRNAs). Efficient reduction of protein levels was confirmed using western blotting analysis (Fig 4.5.1). Controls for our experiments were uninfected cells and cells infected with a vector carrying siRNA against LacZ, which has no significant homology to any human gene. Using Affymetrix Human Focus GeneChip arrays, which represent some 8,500 well-annotated genes, we recorded gene expression profiles in these cellular systems prior to and 4 hrs after exposure to neocarzinostatin (NCS), an enediyne antitumor antibiotic that intercalates into the DNA and induces DSBs [184]. All samples were probed in independent triplicates.

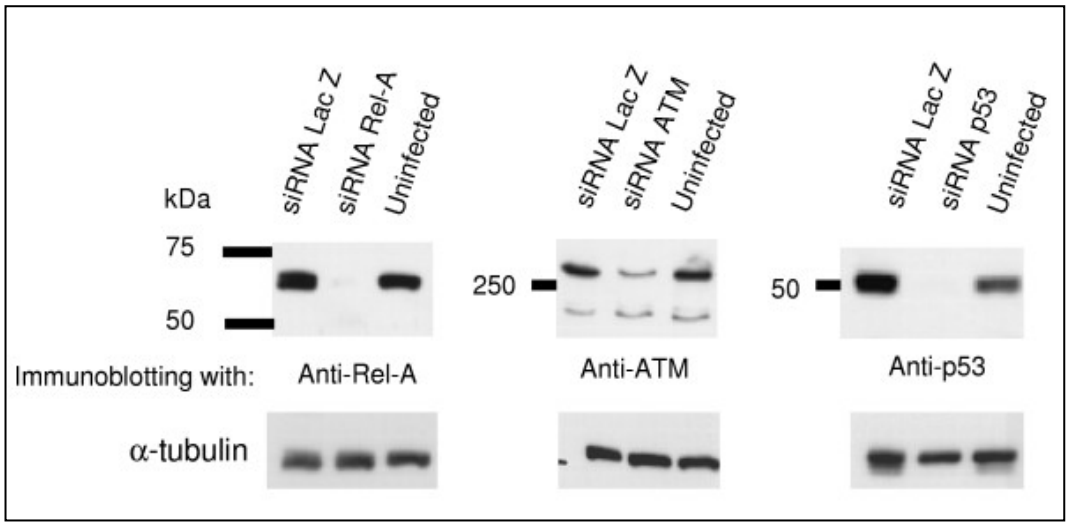



**Figure 4.5.1. Western blotting analysis demonstrating the reduction in protein levels encoded by mRNAs that were targeted by siRNAs**. Tubulin was used as a loading control. (Experiments carried out by Tamar Tenne).

Expression levels were computed using the RMA method [47] that was run from the BioConductor package (http://www.bioconductor.org/). We preferred to use RMA over Affymetrix' MAS5 for two reasons: First, several studies indicated that the mismatch signals are correlated with the mRNA concentration of their corresponding gene; i.e., they themselves contain information on the expression level of the genes, hence subtracting their signals from the perfect-match ones, as MAS5 does, may add noise to the measurement and therefore be counterproductive [47]. RMA ignores the mismatch probes and computes expression levels based only on perfect match signals. When we examined the mismatch probe signals for several genes activated by the NCS treatment, we found that these signals indeed increased, in a manner correlated with the increase exhibited by their corresponding perfect-match signals (Fig. 4.5.2). Second, while MAS5 uses global scaling to normalize between arrays, RMA applies the quantile normalization that was demonstrated to perform better [141]. Comparison of expression levels computed by MAS5 and RMA showed that RMA reduced noise between replicates (Fig 4.5.3), as well as the range of fold-changes in gene expression after the treatment (Table 4.5.1).

Probesets that received 'Absent' calls in all chips were filtered out, leaving 6002 probesets for subsequent analysis. Averaging expression levels over replicates, our dataset contained measurements for ten conditions: five cellular systems (uninfected and the LacZ control cells and cells knocked-down for Rel-A, p53 and ATM), each probed at two time points — without treatment and 4 hrs after exposure to NCS.

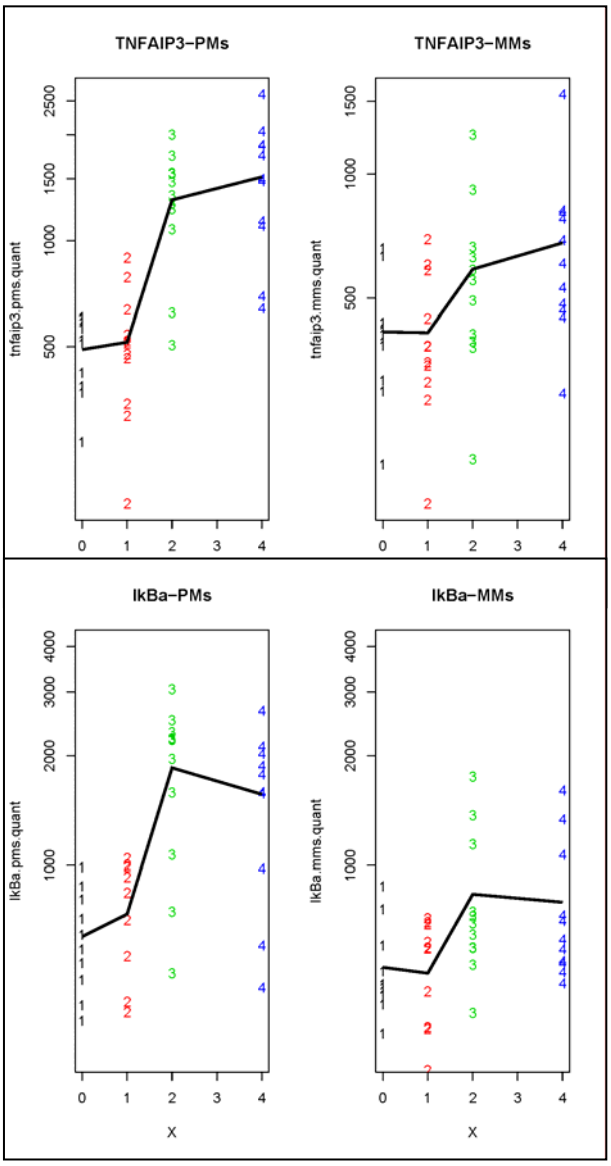



**Figure 4.5.2. Perfect-match (PM) and mismatch (MM) probe signals measured prior to treatment and at 1, 2, and 4 hrs after treatment with NCS.** These signals are shown for two sample genes (TNFAIP3 and IKBα) that were induced by the NCS treatment. Mismatch signals were increased as well, in strong correlation with their PM counterparts, demonstrating that MM probes too contain information on expression level of their target genes. This correlation questions the role of MM probes as negative controls and the utility of subtraction of these signals from the ones measured by PM probes.

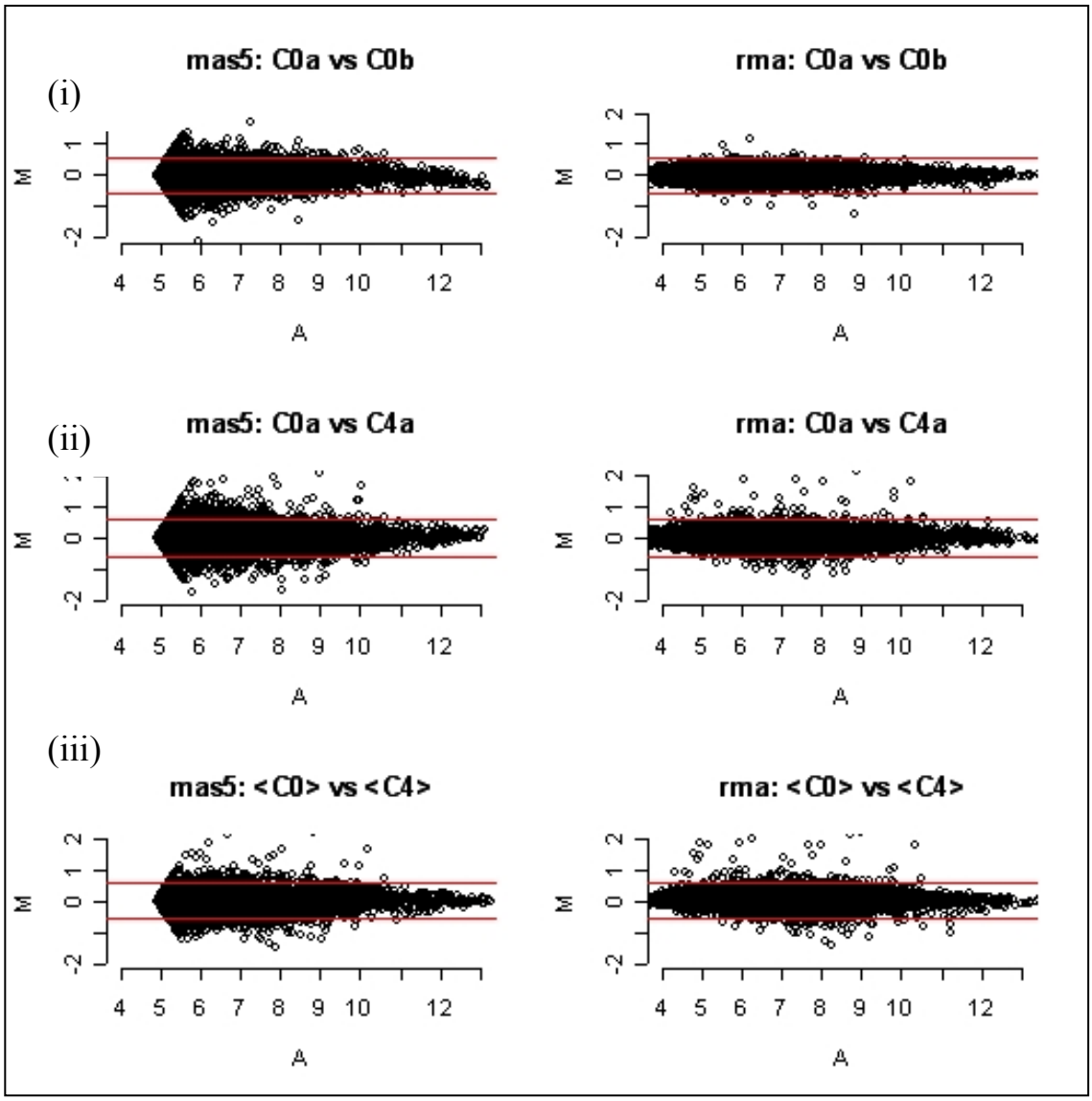



**Figure 4.5.3. Comparison between RMA and MAS 5 computed signals.** M (log of the difference between probe set signals in the compared chips) vs. A (log of the mean probe set signals in the compared chips) plots, as introduced by Speed's lab (http://stat-www.berkeley.edu/users/terry/zarray/Html/normspie.html), based on expression levels that were computed by MAS5 or RMA for comparison between: (i) two replicated chips (C0a vs. C0b), (ii) post-treatment vs. pre-treatment chips (C0a vs. C4a), and (iii) same as (ii) but expression levels were averaged on triplicate chips at both time points. In all comparisons, the fold induction distributions (represented by the Y-axis) were markedly narrower when expression levels were computed by RMA. Distributions based on MAS5 were especially noisy in the low intensity genes.

| | | (I) C_0a vs C_0b | (II) C_0a vs C_4a | (III) <C_0> vs <C_4> |
|---|---|---|---|---|
| **MAS5** | **Up** | 278 | 342 | 148 |
| | **Down** | 202 | 318 | 251 |
| **RMA** | **Up** | 11 | 74 | 95 |
| | **Down** | 7 | 58 | 68 |

**Table 4.5.1. Comparison of expression levels computed by MAS5 and RMA**. Number of genes whose expression level was increased (Up) or decreased (Down) by at least 1.5-fold in the following comparisons: (I) two replicates of control cells are compared at time 0; (II) single chip 4 hrs after NCS treatment is compared to a single chip ; and (III) triplicate averaged chips of samples prior to and after NCS treatment. Note the 25-fold decrease in the noise between two control replicates measured by the two methods (see also Fig 4.5.2).

As a first step in our data analysis we searched for nonspecific responses to siRNA expression. We scanned the dataset for genes that were either consistently up- or down-regulated in cells expressing all five siRNAs compared to their basal level in the uninfected control, all before exposure to NCS. We observed a subtle but statistically significant response to viral infection/siRNA expression. Very few genes were consistently responsive when a cut-off of 1.5 fold-change was set, but lowering the threshold to 1.3 resulted in 20 consistently up-regulated and 75 consistently down-regulated genes in the infected cells. The threshold is low, but the number of genes that showed consistent response is significantly higher than expected at random (in 1,000 datasets with randomly permutated entries for each gene, an average of 0.1 and 0.2 consistently up- and down-regulated genes, respectively, were found). The set of consistently up-regulated genes contained mainly genes involved in different aspects of cellular metabolism. The consistently down-regulated genes included metabolic genes and genes that function in control of cell growth, signal transduction and stress

responses. In contrast to some reports [180, 182], we did not observe induction of the interferon pathway following the introduction of siRNA to the cells.

Next, we searched the dataset for genes that responded to the NCS treatment in the control uninfected cells and whose response was not disturbed by the introduction of siRNA to the cells: namely, genes that responded to the treatment in a coherent manner in the uninfected and the LacZ control cells. We defined the *damage-responding gene set* as all genes whose expression levels changed by at least 1.5-fold in one control (either the uninfected or the LacZ-infected cells), and at least 1.4-fold in the same direction in the other control. A total of 112 genes that were induced in both controls met this criterion. We chose thresholds of 1.5 and 1.4 — lower than those usually used in microarray analysis — because the RMA method significantly narrows the distribution of expression levels and of the fold changes compared to Affymetrix' MAS5 package (Table 4.5.1 and Fig 4.5.3). Although the thresholds are low, the expected false positive rate in our damage-induced gene set is low: not one single gene passed this criterion when applied to expression levels measured 30 min after exposure of the cells to NCS. In addition, this number is significantly higher than expected at random: in 1,000 datasets with randomly permutated entries for each gene, the average number of genes that met this criterion was 14.1. Only 7 genes met an analogue criterion for repression in response to NCS treatment; six of them are related to mitosis, presumably reflecting the activation of cell cycle checkpoints in response to DNA damage.

We divided the expression level of each of the 112 damage-induced genes at the 4-hr time point by its level in untreated cells in the same cellular system, yielding a 112x4 data matrix, with rows corresponding to genes. We standardized each row to mean=0 and SD=1, and subjected the standardized matrix to average-linkage

hierarchical clustering using our EXPANDER package for microarray data analysis. The damage-induced gene set was found to fall into four major response patterns (Fig. 4.5.4): Cluster 1 contained 26 damage-induced genes whose response was strongly reduced in the absence of ATM and Rel-A, and only partially affected by the absence of p53. Cluster 2 contained 11 genes whose response was abolished in the absence of ATM and p53, but augmented in the absence of Rel-A, suggesting some negative regulatory effect for NF-κB on their expression. Cluster 3 contained 46 genes whose response was markedly attenuated in the absence of ATM and p53, and not substantially affected by the absence of Rel-A. Cluster 4 contained 12 genes whose induction was strongly reduced in the absence of p53, partially affected by the absence of ATM, and not affected by the absence of Rel-A. This analysis shows the following. (1) The transcriptional network induced upon exposure to NCS in these cells is almost completely mediated by NF-κB and p53, and these two TFs induced nearly disjoint sets of genes: the former controls the induction of cluster 1 genes, the latter controls the induction of the genes in clusters 2-4. (2) ATM is required for the activation of a major part of the damage-induced transcriptional program, comprising both the NF-κB and p53 response arms (the activation of clusters 1-3 genes is ATM-dependent). (3) There is some cross-talk between the NF-κB and p53 pathways: the absence of p53 partially reduces the induction of the NF-κB arm (cluster 1), suggesting a positive effect of p53 on the induction of the NF-κB mediated response; and the absence of Rel-A leads to increased activation of a subset of the p53-mediated arm (cluster 2), pointing to a negative regulatory role for NF-κB in the induction of these genes.

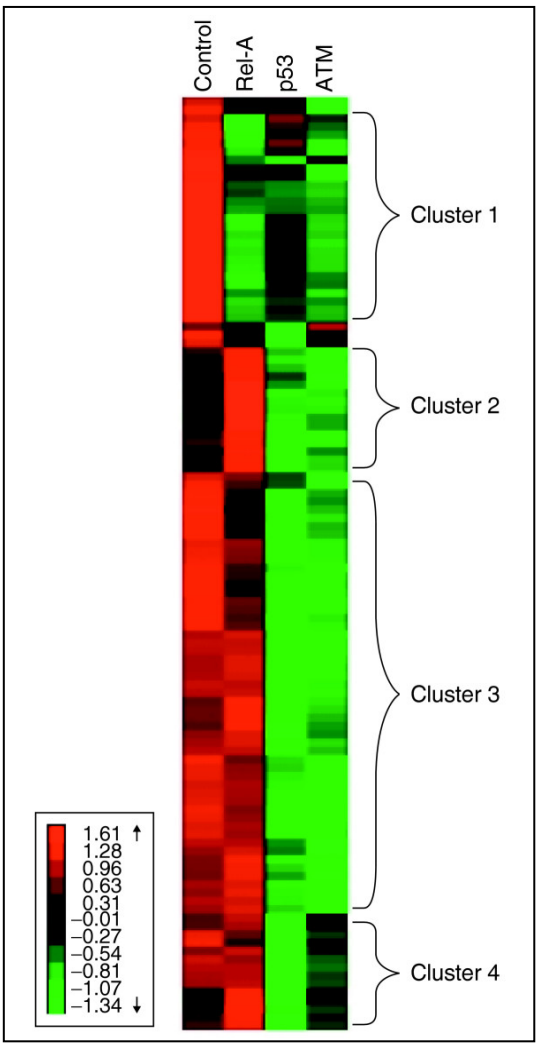



**Figure 4.5.4. Four major expression patterns in the damage-induced gene set revealed by cluster analysis.** For each of the 112 damage-induced genes, induction fold of expression level 4 hrs after NCS treatment was computed in uninfected cells and in the cells knocked-down for Rel-A, p53 and ATM, yielding a 112x4 data matrix, with the rows corresponding to genes. The matrix rows were subjected to hierarchical clustering after normalizing the rows to have mean=0 and SD=1. The heat map visually represents the normalized matrix after clustering. Red, green and black entries represent above-, below- and near-average fold of induction, respectively. Four prominent expression patterns are evident: Cluster 1 represents genes whose induction is strongly attenuated in cells knocked-down for Rel-A and ATM (compared to the response in the control uninfected cells), and only partially attenuated in cells knocked-down for p53. Cluster 2 represents genes whose response is attenuated in cells knocked-down for p53 and ATM, but increased in cells knocked-down for Rel-A. Cluster 3 represents genes whose response is attenuated in cells knocked-down for p53 and ATM, but not affected by knocking-down Rel-A. Cluster 4 represents genes whose response is markedly attenuated in cells knocked-down for p53, and only partially attenuated in cells knocked-down for ATM.

Cluster analysis identified both ATM/NF-κB- and ATM/p53-mediated transcriptional responses. We sought to demonstrate that this dissection of the ATM-mediated transcriptional network induced by DNA damage is precise and cannot be ascribed to some nonspecific or off-targets effects. To this end, we examined the effect of knocking-down Rel-A and p53 on several of their respective known direct targets that were included in the damage-induced genes set. Table 4.5.2A shows that knocking-down Rel-A and ATM significantly blocked the induction of known NF-κB target genes, whereas knocking-down p53 had a much milder effect on their

induction. Table 4.5.2B shows that knocking-down p53 and ATM specifically blocked the induction of known p53 target genes, whereas knocking-down Rel-A did not disrupt their induction (and even augmented it for some genes). Results of quantitative real-time RT-PCR, performed to validate the microarray results for these genes, were in good agreement with the microarray data in most cases; the magnitudes of induction differed between the two experimental systems, but the dependency of transcriptional induction on the various regulators was similar for 10 out of 13 examined genes.

To further substantiate the accuracy of the network dissection obtained by our experimental setup, we applied the PRIMA tool to the dataset in order to identify TFs whose binding site signatures are significantly more prevalent in a given set of promoters than expected at random. In this analysis, the four gene clusters were used as target sets, and the entire collection of genes present on the chip (after filtering out those that got 'Absent' calls under all conditions) served as the background set. Putative promoter sequences corresponding to all known human genes were extracted from the human genome (Ensembl, version 19, Feb 2004). PRIMA tests were confined to 800 bp upstream to the putative genes' transcription start site. Repetitive elements were masked out. Both strands were scanned.

Markedly, promoters of genes assigned to cluster 1, which represents an ATM-NF-κB dependent response, were specifically and highly significantly enriched for the binding site signature of NF-κB (Table 4.5.3), while p53-dependent clusters 3 and 4 were specifically enriched for the binding site of ATF2. ATF2 regulates transcription after heterodimerization with either ATF3 or c-Jun [185]. Of note, in our dataset the induction of both *ATF3* and *c-Jun* was p53-dependent (Table 4.5.2B), hence the enrichment for this signature probably reflects a second wave of

**Table 4.5.2.** Fold induction of genes' expression after 4 hrs of exposure to 200 ng/ml of NCS as measured by microarrays and by quantitative real-time RT-PCR.

**A. Known direct targets of NF-κB.**

| Gene | Affy_ID | Fold induction - Microarray | | | | | | Fold induction - RT-PCR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C | LacZ | Rel-A (NFκB) | p53 | ATM | | C | Rel-A (NFκB) | p53 | ATM |
| TNFAIP3 | 202644_s_at | 8.28 | 5.34 | 1.15 | 3.02 | 1.19 | | 9.5 | 1.1 | 9.5 | 0.9 |
| RELB | 205205_at | 3.7 | 2.89 | 0.82 | 2.95 | 0.91 | | 15.7 | 6.0 | 21.3 | 2.5 |
| TNFRSF9 | 207536_s_at | 4.01 | 3.5 | 1.1 | 2.08 | 1.21 | | 14.3 | 3.5 | 11.0 | 1.4 |
| NFKBIA | 201502_s_at | 4.61 | 5.4 | 1.26 | 2.67 | 1.02 | | 4.2 | 1.7 | 4.5 | 1.2 |
| CD83 | 204440_at | 3.46 | 2.99 | 1.0 | 1.73 | 1.06 | | 6.5 | 1.0 | 5.7 | 1.3 |
| IER3 | 201631_s_at | 4.44 | 5.12 | 1.43 | 2.35 | 1.44 | | 6.6 | 1.8 | 3.4 | 1.8 |

**Table 4.5.2 (cont.)**
**B. Known direct targets of p53.**

| Gene | Affy_ID | Fold induction - Microarray | | | | | | Fold induction - RT-PCR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C | LacZ | Rel-A NF-κB | p53 | ATM | | C | Rel-A NF-κB | p53 | ATM |
| ATF3 | 202672_s_at | 3.44 | 3.74 | 7.03 | 1.54 | 1.47 | | 5.2 | 5.9 | 1.6 | 1.6 |
| EGR1 | 201694_s_at | 2.78 | 1.77 | 6.77 | 1.04 | 1.02 | | 4.4 | 13.4 | 0.7 | 2.4 |
| JUN* | 213281_at | 2.01 | 1.45 | 2.71 | 1.36 | 1.25 | | 6.6 | 3.9 | 0.64 | 2.5 |
| FOS | 209189_at | 1.72 | 1.42 | 2.22 | 1.07 | 1.22 | | 3.4 | 13.1 | 3.4 | 1.9 |
| ETR101* | 202081_at | 1.97 | 2 | 2.6 | 1.06 | 1.13 | | 2.0 | 3.0 | 1.4 | 1.4 |
| GADD45A | 203725_at | 2.36 | 2.07 | 2.00 | 1.07 | 1.22 | | 1.8 | 2.3 | 1.8 | 1.3 |
| DUSP1 | 201041_s_at | 2.06 | 2.57 | 3.45 | 1.11 | 1.22 | | 2.2 | 4.5 | 2.0 | 1.9 |

*These genes are not reported as direct targets of p53 but are known to be functionally related to p53.

transcriptional regulation controlled by these TFs, whose induction is mediated by p53. This is in agreement with other studies that reported a p53-dependent activation of ATF3 in response to DNA damage [186, 187]. PRIMA did not identify enrichment for the p53 binding site signature in the p53-dependent clusters. It is possible that PRIMA is not sensitive enough to detect p53 enrichments due to the complex nature of p53's binding sites [188] or the relatively long distance of these binding sites from the transcription start sites (many experimentally validated p53 binding sites are located outside the promoter region included in PRIMA analysis). However, using the same parameters, PRIMA did identify significant enrichment for p53 binding signature in several other microarray datasets that we analyzed (e.g., results presented in Section 4.6). We therefore believe that p53 signature is not over-represented in these clusters, suggesting that p53 in the cells we used exerts its direct effect on a limited number of target genes, which are then further expanded into a wider network of transcriptional response mediated mainly by the ATF/Jun.

**Table 4.5.3.** Significantly enriched transcription factor (TF) binding site signatures in promoters of co-clustered genes.

| Cluster | Number of genes* | Dependency of genes' induction on** | | | Enrichment for binding sites of*** | |
|---|---|---|---|---|---|---|
| | | ATM | Rel-A (NF-κB) | p53 | NF-κB (M00054) | ATF2 (M00179) |
| 1 | 26 | ++ | ++ | + | 9.7 ($6.0x10^{-12}$) | ---- |
| 3 | 46 | ++ | - | ++ | ---- | 2.9 ($2.7x10^{-5}$) |
| 4 | 12 | + | - | ++ | ---- | 6.6 ($3.6x10^{-6}$) |

* Number of genes with promoter sequence data.

** Strong attenuation in induction of the cluster's genes in the respective cells is denoted by '++', partial attenuation is denoted by '+', and no attenuation by '-'.

*** The ratio between TF hits prevalence in the cluster and in the background sets of promoters, and its p-value (Accession numbers for TF binding site models are of TRANSFAC DB).

## *4.6. Functional genomics delineation of Atm-dependent transcriptional responses induced by ionizing radiation in murine lymphoid tissues*

This project was carried out in collaboration with Sharon Rash-Elkeles of Prof. Shiloh's group; Nir Weizman of Prof. Barzilai's group (Faculty of Life Sciences, Tel Aviv University); Chaim Linhart of Prof. Shamir's group; and Dr. Ninette Amariglio and Prof. Gideon Rechavi of the Functional Genomics Unit at the Sheba Medical Center.

In this study we applied gene expression microarrays combined with our computational battery to delineate transcriptional responses induced by ionizing radiation (IR) in murine lymphoid tissues and components of this network whose activation is Atm-dependent.

The critical cytotoxic DNA lesion inflicted by IR is the DNA double strand break (DSB), and ATM is the master regulator of the cellular response to this DNA lesion [93]. Two phenomena prompted us to examine responses to IR in lymphoid tissues. First, A-T patients show severe immunodeficiency stemming from aberrant development of both the B- and T- lymphoid arms, and they are highly prone to cancer, mainly of lymphoid origin. Second, *ATM* is frequently mutated in sporadic cancers of lymphoid origin [189], among them B-cell chronic lymphocytic leukemias (B-CLLs), which are the most common leukemias in western countries. B-CLL tumors that carry mutations in either *ATM* or in *TP53* are associated with poor clinical course, with *ATM*-mutated B-CLL tumors less aggressive than the *TP53*-mutant ones [190, 191]. Studying expression profiles in untreated and irradiated Atm-deficient and control lymphoid tissues was expected to elucidate molecular factors that promote

malignancies, and molecular determinants that affect sensitivity or resistance to IR and other forms of chemotherapies used to treat cancers.

Global transcriptional responses were recorded in wild type and Atm-deficient lymph node tissues of unirradiated mice, and 30 and 120 min after exposure to whole body irradiation with 15 Gy of IR. All mice were 5-7-week old males. Affymetrix GeneChip MGU74Av2 arrays were used. Each sample represented a pool of tissues from 3 animals. Samples from untreated mice were probed in independent hybridization triplicates, and samples from irradiated mice were probed in independent hybridization duplicates. Mice were handled by Nir Weizman. Samples and chip hybridizations were prepared by Sharon Rashi-Elkeles.

Signal intensities were computed using Affymetrix MAS 5.0 software. All chips were scaled to an average signal intensity of 150. Probe sets that registered 'Absent' flags by MAS 5.0 in all measured conditions were excluded. To reduce false positive calls of differential genes, which is especially frequent at the low range of intensities, signal intensities below 40 were set to 40. This dataset was analyzed before publication of the RMA method and we repeated the analysis reported here using RMA and quantile normalization. All the results reported below remained solid. A representative expression level for each probe set in each of the six tested conditions (two genotypes, three time points) was computed by averaging the probe set's signal intensities in the replicate arrays. As a filtering step, we defined the set of 'responding genes', consisting of genes whose expression level was changed by at least 1.75 fold across the tested conditions. Some 10% of the probe sets present in the array, 1206 out of 12488, met this criterion.

After these preprocessing steps, we analyzed the data using the EXPANDER package. The analysis included the following steps:

Identification of major expression patterns in the dataset. We subjected the set of responding genes to CLICK, a clustering algorithm that yields an optimal balance of intra-cluster homogeneity and inter-cluster separation [145]. Prior to clustering, expression levels of each gene were standardized to have mean equal to zero and variance equal to one; hence, genes clustered together share expression patterns across the tested conditions, but might differ in the magnitude of their response. The six major clusters identified by CLICK are shown in Fig. 4.6.1. Clusters 1 and 2 represent Atm-dependent expression patterns: that is, they contain genes that were induced by IR in the Atm+/+ tissue, while their activation in the Atm-deficient tissue was significantly abrogated. Cluster 1 represents 'early responders' that were already transcriptionally activated at the early time point of 30 min post IR, and whose activation was Atm-dependent. Cluster 2 represents a later wave of Atm-dependent response. Clusters 3 and 4 contain genes that were either activated (cluster 3) or repressed (cluster 4) in both genotypes. Clusters 5 and 6 contain genes that responded only in the Atm-deficient tissue.

Functional categories within gene clusters. Examination of the genes that responded to IR indicated that the network activated following IR spans many biological processes covering most aspects of the cellular physiology. In an attempt to systematically characterize this network, we applied tests aimed at identifying functional categories that are statistically enriched in the clusters. We utilized functional annotations of mouse genes provided by the Mouse Genome Informatics

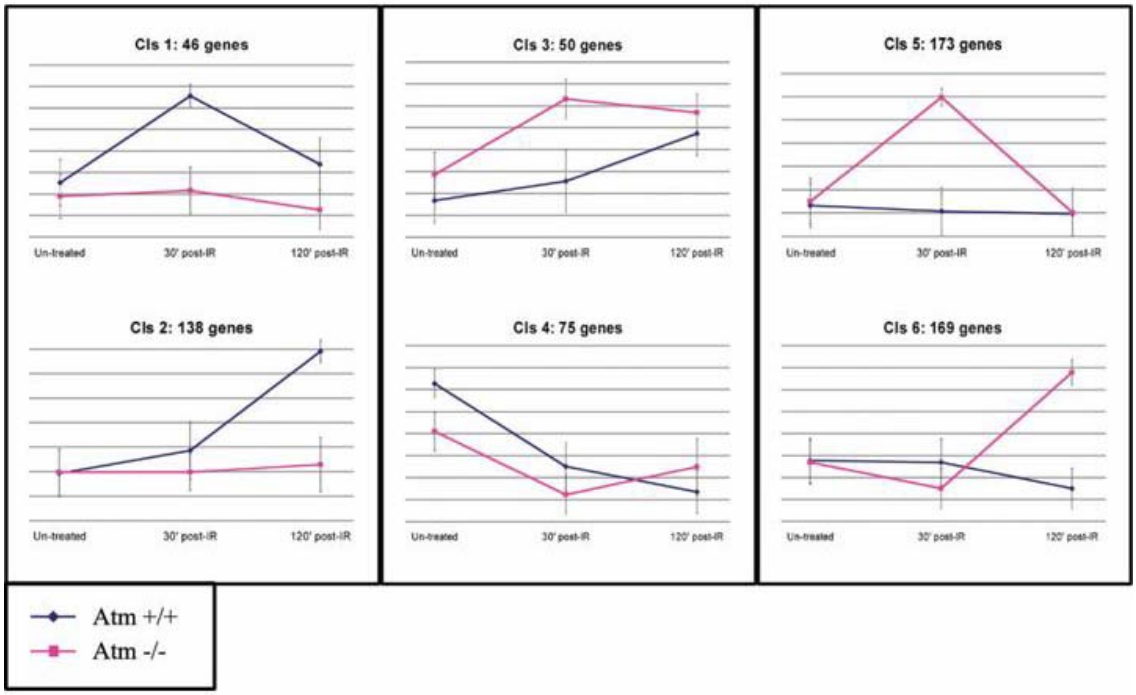



**Figure 4.6.1. Major clusters identified by CLICK in the set of 1206 responding genes**. Each cluster represents a set of genes with a similar expression pattern. Prior to clustering, the expression levels of each gene were standardized to have a mean value of 0, and variance of 1. The Y axis represents these standardized values. The X axis corresponds to the tested conditions: unirradiated animals, and 30 and 120 min post irradiation. Shown for each cluster is the mean expression pattern calculated over all the genes contained in it, in Atm+/+ (blue) and Atm-deficient tissues (red). Error bars represent +/- one S.D. The total number of genes in each cluster is indicated. Clusters 1 and 2 contain early and late Atm-dependent responders. Clusters 3 and 4 represent genes that exhibited similar response patterns in both genotypes. Clusters 5 and 6 represent early and late responding genes that were activated in Atm-deficient but not in the Atm+/+ tissue.

(MGI), which uses the standard vocabulary introduced by the Gene Ontology (GO) consortium [149]. Enriched functional categories ($p<0.01$) were identified in four of the clusters (Table 4.6.1). Importantly, regulation of cell cycle and apoptosis were among the categories enriched in the Atm-dependent clusters, pointing to their defective activation in the Atm-deficient tissue.

| Cluster | Number of genes with functional annotations | Functional category | GO ID | Number of Genes associated with the category |
|---|---|---|---|---|
| 1 | 25 | Cell cycle | GO:0007049 | 7 |
| 2 | 63 | Regulation of cell cycle | GO:0000074 | 7 |
| | | Cytokine activity | GO:0005125 | 5 |
| | | Apoptosis | GO:0006915 | 5 |
| 5 | 83 | Electron transport | GO:0006118 | 8 |
| 6 | 54 | Response to pathogen | GO:0009613 | 13 |
| | | Inflammatory response | GO:0006954 | 6 |

**Table 4.6.1. Functional categories enriched in the clusters** ($p<0.01$).

Computational search for mediating transcriptional regulators. Next, we sought to identify transcriptional regulators that control the observed modulation in gene expression following IR. We were particularly interested in regulators whose activation is compromised in Atm-deficient tissues. To this end, we applied PRIMA. Each gene cluster was considered a target set, and the entire collection of putative murine promoters corresponding to genes present on the chips and expressed in the lymph node was the background set. Mouse and human promoter sequences used here were downloaded from Ensembl project (v13, May 2003 release) [192]. Analysis was done on the region from 1,000 bp upstream to 200 bp downstream to genes' putative TSS.

PRIMA identified several TFs whose binding site signatures were significantly over-represented in cluster 2, which corresponds to the later wave of Atm-dependent response (Table 4.6.2). The highest enrichment was observed for NF-κB and p53,

| Transcription factor | Accession number in TRANSFAC DB | Enrichment factor* | p-value |
|---|---|---|---|
| NF-kappaB | M00054 | 4.3 | $8.0x10^{-9}$ |
| p53 | M00034 | 4.2 | $1.5x10^{-5}$ |
| Sp1 | M00196 | 1.6 | $3.1x10^{-5}$ |
| STAT1 | M00496 | 2.9 | $4.1x10^{-4}$ |

**Table 4.6.2. TF binding site signatures enriched in the later wave of Atm-dependent response (cluster #2).**
*The ratio between the prevalence of transcription factor hits found by PRIMA in promoters of the genes contained in cluster #2 and in promoters of the background set of all mouse promoters.

which are both well-established stress-induced transcriptional regulators. The incidence of the NF-κB binding signature was more than 4-fold higher among the promoters of cluster 2 than in the background set. This enrichment was robust, remaining solid over a large range of threshold values. PRIMA identified 19 promoters in this cluster that contained at least one high-scoring putative NF-κB binding site (the total number of such 'hits' was 27, as several promoters contained more than one putative NF-κB binding site). We believe that the number of genes whose response is controlled by NF-κB in response to IR is higher than we report; some were probably not picked up because of the stringent threshold used when scanning for putative NF-κB hits. Several of the genes in which we identified strong NF-κB binding site signature were previously reported to be under direct control of NF-κB, while the others are novel putative NF-κB targets and require experimental validation. Of note, this set of genes contained those that encode subunits of the NF-κB heterodimer itself (Relb, Nfkb2), as well as two of its direct inhibitors, Nfkbia (IκBα) and Nfkbib (IκBβ). This pattern of parallel activation of positive and negative regulators, which probably represents positive and negative feedback loops, appears to be a recurrent theme in the logic of cellular signaling networks. This theme appears

in the p53-mediated arm as well, where p53 induces its inhibitor MDM2. PRIMA identified high enrichment for p53 binding site signatures in promoters of this cluster - 4-fold higher than expected according to the prevalence of p53 hits in the background set (Table 4.6.2). p53 binds to a consensus DNA sequence consisting of two conserved decamers separated by a spacer varying in length from 0-13 base pairs [193]. The PWM used by PRIMA to represent p53 binding sites does not model the flexibility in the length of the spacer between the decamer repeats, and therefore may have missed possible p53 binding sites in the promoters of this cluster that contain a spacer between the decamers.

To validate the results obtained by the microarrays, Sharon Rashi-Elkeles from our lab performed quantitative real-time RT-PCR analysis of the expression of eleven genes that responded in an Atm-dependent manner. This analysis was focused on the putative NF-κB-mediated arm, as the p53-mediated arm is well-documented. We selected for validation genes whose promoters were found to contain a strong NF-κB binding signature. To reduce false positive rate, we required that a strong NF-κB binding signature appear also in the promoters of the human ortholog genes. We found good agreement between the microarray and RT-PCR results; the magnitudes of induction differed for some genes but the dependency of their activation on functional Atm was validated for all eleven examined genes (Fig. 4.6.2).

A combination of microarray-based and computational analysis pointed out the major involvement of NF-κB in Atm-mediated gene regulation in the lymphoid cells

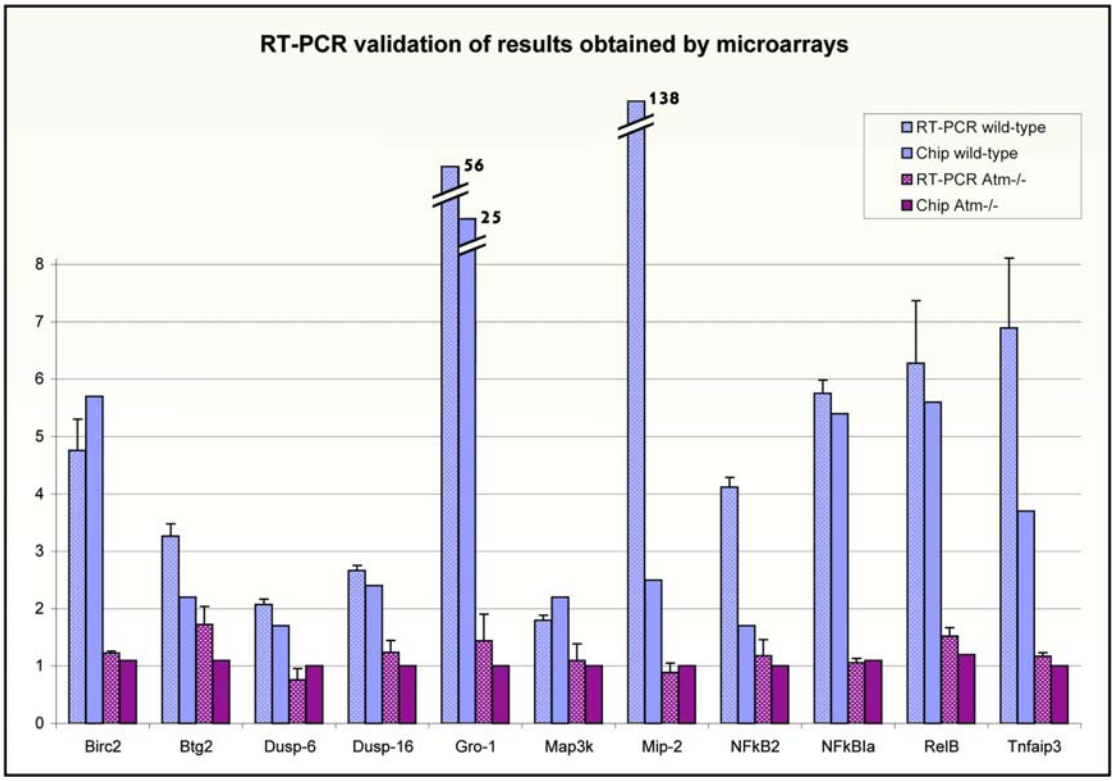



**Figure 4.6.2. Real-time RT-PCR validations.** Comparison of gene expression (induction fold) obtained from microarray data and real-time RT-PCR analysis for 11 genes selected from the set of Atm-dependent, putatively NF-κB-mediated responding genes. Four bars are shown for each gene: the blue and red bars represent fold induction of that gene 120 min post IR in the Atm+/+ and Atm-deficient tissues, respectively. Within each genotype, the dotted bar denotes the result obtained by RT-PCR (averaged over three independent measurements, error bars represent one S.D.), and the solid bar denotes the value obtained from microarray measurements. Note that although the magnitude of induction of some genes differs in the real-time RT-PCR and microarray measurements, all eleven genes show agreement between the two technologies on the dependence of the induction on Atm. (RT-PCR experiments were carried out by Sharon Rashi-Elkeles.)

following IR. We sought to confirm this phenomenon by direct biochemical demonstration of the dependence of the IR-induced NF-κB activation on Atm. Electro-mobility shift assays (EMSA) performed by Nir Weizman on nuclear extracts from lymph node tissues of untreated and irradiated Atm+/+ and Atm-deficient mice showed that while NF-κB binding activity in the Atm+/+ tissue was induced by 2.5-fold following IR, the irradiated Atm-deficient tissue failed to induce NF-κB binding

activity (Fig. 4.6.3). These data demonstrate that the induction of NF-κB in lymphocytic cells in response to IR is Atm-dependent.

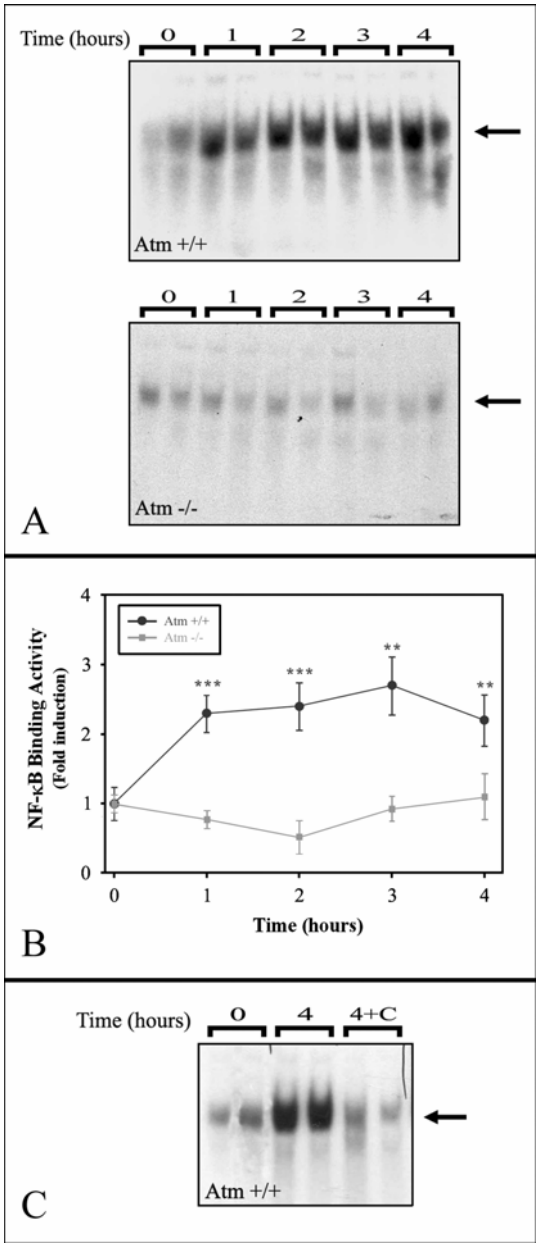



**Figure 4.6.3. Reflecting NF-κB activation using EMSA.** NF-κB binding activity in wild-type and Atm-/- lymphoid tissues following exposure to 20 Gy X-rays. (A) EMSA results obtained from two animals are shown for each indicated time point. The arrows indicate the position of the NF-κB-DNA complexes. (B) Quantitative analysis of NF-κB binding activity fold induction (n=4 for each time point). ***$p<0.01$; **$p<0.025$. Error bars represent ± S.D. Statistical analyses were performed using two-tailed Student's t-test. (C) To demonstrate the NF-κB specificity of the shifted bands, nuclear proteins isolated from irradiated Atm+/+ tissues were exposed to 100-fold excess of unlabeled oligonucleotides representing NF-κB-binding consensus sequence, and then incubated with radiolabeled probe (4+C). Band intensity was significantly reduced under this condition. (Experiments carried out by Nir Weizman.)

In order to identify biological endpoints of the Atm-dependent gene regulation mediated by NF-κB and p53, we applied our SHARP tool (described in Section 4.2) to this dataset. Figure 4.6.4, generated using SHARP, indicates that while the p53-regulated arm induced by IR included two major pro-apoptotic regulators (Apaf1 and

Bax, which are both direct targets of p53 [194-196]), the NF-κB-mediated arm contained three pivotal anti-apoptotic genes (Birc2, Birc3 and TNFaip3, all are direct targets of NF-κB [196, 197]). Birc2 and Birc3 (cIAP1, cIAP2), are members of the Inhibitor of Apoptosis (IAP) family of proteins that inhibit apoptosis probably by directly interfering with activation of several caspases, including Casp3 and Casp9 [196, 198]. Tnfaip3 (A20) was reported to inhibit TNFα-induced apoptosis by disrupting the recruitment of TRADD and RIP to the complex that assembles at the TNF receptor shortly after it is bound by its ligand [197]. Our results show that in response to IR, pro- and anti-apoptotic signals are induced in parallel, and the induction of both arms is Atm-dependent. The pro- and anti-apoptotic signals were conveyed via direct targets of p53 and NF-κB.

In a recent study, Stankovic et al. [199] recorded gene expression profiles in ATM-deficient, p53-deficient and ATM/p53 proficient B-CLL cancer samples. Similar to the results obtained in our dataset, these investigators reported that the ATM-dependent transcriptional response is composed of two major arms: one is p53-dependent and contains many pro-apoptotic genes, while the other, controlled by an unknown transcription factor, is enriched with pro-survival genes. In an attempt to reveal the regulator of the ATM-dependent, p53-indepndent response observed in that study, we applied PRIMA to the cluster of 61 genes that were reported to respond to IR in the wild type and *TP53*-mutant but not in the *ATM*-mutant B-CLL tumors. In full concordance with the result obtained on our dataset, we found that the ATM-dependent, p53-independent, pro-survival cluster reported by Stankovic et al. was significantly enriched for the NF-κB binding site signature. The prevalence of the NF-κB binding site signature on genes' promoters of this cluster was more than 4-fold higher than in the background set comprised of all human known-genes' promoters

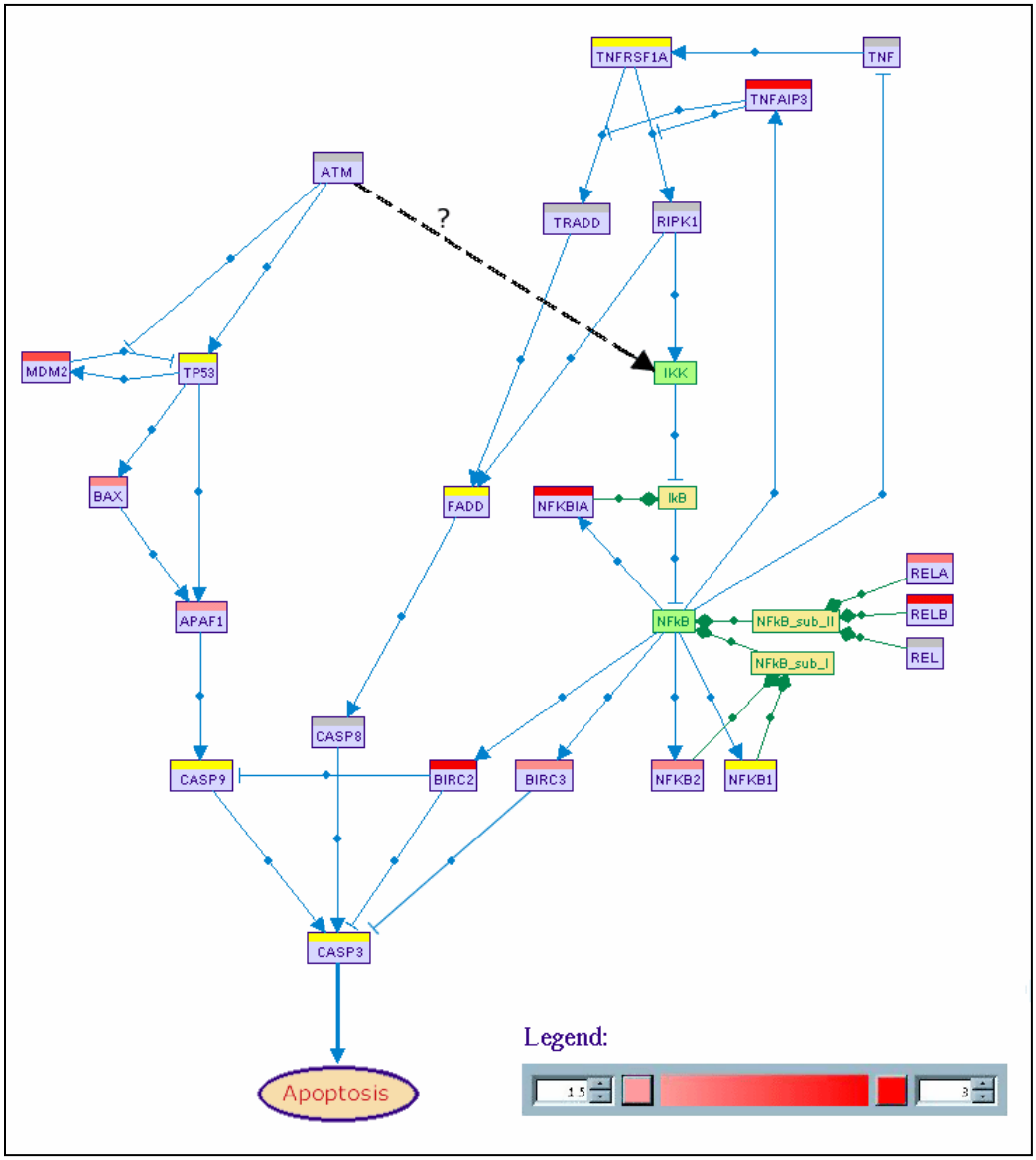



**Figure 4.6.4. Parallel, Atm-dependent induction of pro- and anti-apoptotic signals in response to IR in murine lymphoid tissue.** Superposition of microarray results on gene-interaction map using SHARP pointed to parallel induction of pro- and anti-apoptotic signals, with the pro-apoptotic pathway mediated by p53 (through its induction of Apaf1 and Bax), and the anti-apoptotic arm mediated by NF-κB (through its induction of Birc2, Birc3 and Tnfaip3). The activation of both arms was dependent on Atm in our dataset. p53 is a direct substrate of ATM. The mechanism by which ATM activates the NF-κB arm remains to be determined. The figure was created using SHARP. The interaction map contains 3 types of nodes and 2 types of edges. gGrey nodes represent single proteins (denoted by the official Human Genome Nomenclature Committee (HGNC) symbol of their encoding gene), yellow nodes represent protein families (e.g., IκB), and green nodes represent protein complexes (e.g., IKK). The first type of edge denotes regulation relations (→ for 'activation', ⊣ for 'inhibition'), and the second type denotes containment relations (green arrows) among nodes (e.g., RelA is contained in the NF-κB_subII family). Protein nodes were colored by SHARP according to the fold-induction exhibited by their encoding genes in the Atm+/+ tissue 120 min post-IR: red- induction, green- repression, yellow- no change, gray- data not available (gene was either not present on the microarray or its expression level was below detection limit).

(p-value of this over-representation is $4.7x10^{-6}$). Similar to findings on our data, the cluster contains several subunits of NF-κB itself (NFKB1, NFKB2, RelB), further supporting a major role for this TF in the induction of the pro-survival transcriptional response. This observation suggests an underlying molecular model for the phenotypic difference between *ATM*- and *TP53*-mutant B-CLL tumors: In contrast to ATM-deficient tumors, in which the induction of both pro- and anti-apoptotic signals is compromised, in p53-deficient B-CLLS, the induction of the apoptotic arm is abolished while that of pro-survival signals mediated by ATM and NF-κB is intact, making these tumors more aggressive and resistant to chemotherapies.

In Sections 4.3 - 4.5 we demonstrated that *in-silico* dissection of transcriptional networks is feasible in the analysis of gene expression data obtained from a homogenous population of culture cells. Here, we demonstrate that this functional genomics approach is sensitive enough to dissect transcriptional networks in a more physiological relevant situation: a mixture of cells in the tissue of an irradiated animal. The model that emerges has implications for rational therapeutic strategies for managing cancers of lymphoid origin. It suggests that restoring the p53-mediated apoptotic arm while blocking the NF-κB-mediated pro-survival arm could conceivably increase the radiosensitivity of lymphoid tumors.

# 5. Discussion

Functional genomics is changing the way biological research is done. For the first time it is possible to study biological systems as a whole and to obtain large-scale snapshots of cellular transcriptome and proteome. The maturation of novel high-throughput biotechnologies has turned biology into an information-rich science. Yet, the field is still in its infancy in terms of mining biological knowledge out of the vast volume of collected data. Indeed, the development of novel techniques for analysis of functional genomics data is one of the major challenges in bioinformatics. Notable successes in systems biology have been achieved in global delineation of transcriptional networks in various organisms ranging from primitive prokaryotes to human. In our studies, we developed and applied functional genomics approaches to dissect transcriptional programs that are associated with cell cycle progression and responses to DNA damage in human and mouse model systems. Our results elucidated novel regulatory links within these intricate signaling networks.

Usage of computational tools. Section 4.1 presents our PRIMA tool for promoter analysis. In addition to its utilization in the analysis of the microarray datasets collected in our lab, PRIMA was installed in more than ten labs from various countries, and there are already several publications that reported biological findings that were obtained using this tool [200, 201]. In addition to the direct installation of PRIMA, many more labs use this module as part of our EXPANDER package, which is now the preferred way for running PRIMA given the great improvement in running time achived by using the pre-complied promoter fingerprint files provided with EXPANDER for six organisms. Over a thousand labs installed EXPANDER to date.

The SHARP KB described in Section 4.2 is not officially released yet, but it is installed in several beta sites in addition to our own lab. These sites include the labs of Prof. Danny Michaelson and Prof. Yoel Kloog from the Department of Neurobiochemistry, Faculty of Life Sciences, Tel Aviv University and three research groups the Curie Institute, Paris, France. We have also established a SHARP version for fly (Drosophila) installed in the lab of Dr. Danny Chamovitz, Dept. Plant Sciences, Faculty of Life Sciences, Tel Aviv University. We expect that with its official release and publication SHARP will become popular among the DNA damage research community.

Elucidation of key regulators of the human cell cycle transcriptional program. In the study presented in Section 4.3, we demonstrated for the first time that the reverse-engineering approach, which infers transcriptional mechanisms from measured gene expression data, can accurately reveal transcription factors that control the observed modulation in the human cellular transcriptome. Employing genome-wide *in-silico* promoter analysis, we revealed eight transcription factors (E2F, NF-Y, Sp1, ATF, NRF-1, CREB, Arnt and YY1) whose binding site signatures are significantly over-represented in promoters of genes whose expression is cell cycle dependent. The enrichment of some of these factors was specific to particular phases of the cell cycle. In addition, we found that several pairs of these TFs show a significant co-occurrence rate in cell cycle-regulated promoters. Most of our computationally-derived findings are strongly supported by experimental evidence.

The E2F family is well documented as a prime regulator of the mammalian cell cycle. Pathways that modulate the activity of E2F are frequently disrupted in human cancers, leading to misregulated cellular proliferation [202]. The E2F PWM obtained highly significant enrichment scores in all our analyses, demonstrating the sensitivity of our methods to reveal true signals. The role of this family of TFs in the cell cycle was underscored by several recent studies showing that E2F regulates not only genes

that function in the G1/S and S phases, but also many M phase genes [163, 164]. Our analysis indicates that the E2F PWM is indeed enriched in promoters of genes that are expressed in G2, although its enrichment in promoters of genes expressed in G1/S and S phases is much more prominent (Fig. 4.3.1).

Published experimental data support our findings on most of the other TFs as well. NF-Y and Sp1 PWMs obtained highly significant enrichment scores. Though involved in many different aspects of cellular life, both TFs have an established role in the regulation of the cell cycle. NF-Y was demonstrated to control the expression of several key regulators of the G2/M phases of the cell cycle [203-206], in line with our observation of its significant enrichment in these phases. The transcriptional activity of Sp1 is modulated in a cell cycle-dependent manner through its phosphorylation by Cyclin A-CDK complexes [207]. In addition, several cell cycle regulators were reported to be controlled by Sp1 [208-211].

Our findings that E2F and NF-Y binding sites, as well as E2F and Sp1 binding sites, significantly co-occur in promoters of cell cycle-regulated genes suggest functional cooperation between these TFs in the regulation of cell cycle progression. Experimental evidence supports the existence of such relations. Physical interactions were demonstrated between members of the E2F and Sp1 families [212], and functional cooperation between E2F and Sp1 was reported in several cell cycle-related promoters [212-216]. As for E2F and NF-Y, co-occurrence of functional binding sites for both TFs was reported in several promoters, including Cdc2, TK, POLA, Cyclin A, and several histone genes [217]. Functional synergism between E2F and NF-Y was demonstrated in the regulation of the E2F-1 promoter [218]. Our findings substantially expand the generality of these functional links, pointing to possible synergism between these TFs on dozens of cell cycle-regulated promoters.

Other TFs that were significantly over-represented in cell cycle-related promoters in our analyses have not been established as prominent regulators of the cell cycle, but data suggest they are involved in regulation of cellular proliferation. ATF/CREB is a family of over a dozen TFs that bind a common regulatory element, the ATF/CRE (cAMP Response Element) motif. One member of the family, CREB, undergoes cell cycle-regulated phosphorylation [219], and was recently reported to control the expression of multiple cell cycle regulatory genes [220]. Over-expression of another family member, ATF2, inhibits the G1/S phase transition in human cancer cell line [221], and is directly involved in the regulation of cyclin A [222] and cyclin D1 [223].

YY1 was reported to control several S-phase-induced genes [224, 225]. Over-expression of YY1 was reported to induce DNA synthesis [226], and a cell cycle-regulated physical interaction between YY1 and pRb was reported in the same study. These findings link YY1 to induction of the S phase. In contrast, we found the YY1 PWM to be under-represented in the S phase, but significantly enriched in the M/G1 cluster.

Arnt forms a dimeric TF with the aryl hydrocarbon receptor (AhR). It is implicated in developmental processes and tissue homeostasis, and several studies have linked the AhR-Arnt dimer to cell cycle regulation. Activation of AhR was reported to induce G1 arrest [227, 228]. Recently, this negative regulation was shown to depend on physical interaction between AhR and pRb [229]. In agreement, we find the enrichment of the Arnt PWM in the G1/S cluster.

Transition of cells from quiescence to proliferation increases the cell demand for energy. One way of responding to the increased demand for ATP is to modulate the activity of the respiratory chain components. NRF-1 regulates the expression of many genes required for mitochondrial respiratory function [230]. The hypothesis that we

raised in this study for the first time of a functional relationship between NRF-1 and E2F recently received strong support from a study by Cam et al. [231].

In addition, we found that several pairs of these TFs show a significant co-occurrence rate; that is, some combinations of their binding site signatures form recurrent cis-regulatory modules that are embedded in multiple cell cycle-regulated promoters. We expect our findings will provide guidelines for experimental dissection of regulatory mechanisms that control cell cycle in mammalian cells. Moreover, the methods demonstrated in this study are general and can be applied to the analysis of transcriptional networks controlling any biological process.

Computational identification of TFs associated with c-Myc. In the study presented in Section 4.4, we further demonstrated how computational promoter analysis can be utilized in the analysis of data obtained by the ChIP-on-chip technique. As a test case, we focused on promoters bound by c-Myc. We identified nine TFs whose binding site signatures were significantly over-represented on promoters of c-Myc target genes. We showed that the binding site signatures of most of these TFs were also enriched on the set of mouse homolog promoters, suggesting functional conservation of their putative association with c-Myc.

Our computational analysis sheds more light on the mechanisms by which c-Myc promotes cell growth and transformation. Among the TFs we found enriched in the c-Myc/Max target promoter set were the pivotal regulators of the transcriptional program associated with cell cycle progression, E2F and NF-Y. Functional links between c-Myc and E2F are well documented [232-235]. Myc promotes cell cycle progression by coordinated activation of cell cycle driving genes (e.g., *Cdc25A*, *Cdk4*, and *Cyclins D2*, *E* and *A*), and by suppression of cell cycle arrest genes (such as *p15*,

*p21*, *p27* and *GADD*s) [233]. Some of the cell cycle promoting genes are common targets of c-Myc and members of the E2F family TFs [236]. As for the link between c-Myc and NF-Y, physical interaction between c-Myc and the NF-YB and NF-YC subunits of the NF-Y trimer has been demonstrated [237, 238].

Another TF whose binding signature was highly enriched in the c-Myc target promoters is EGR-1, which is rapidly activated by many types of stress, including hypoxia, DNA damage and vascular injury, and has a central role in angiogenesis — the formation of new blood vessels from pre-existing vasculature [239, 240]. The possible functional links between c-Myc and EGR-1 is intriguing because of the pivotal role of EGR-1 in angiogenesis. Uncontrolled angiogenesis plays an important role in tumor growth, and the sprouting of new blood vessels into tumors suggests that angiogenesis is necessary for the progression of malignancy. Recent reports underscored the critical roles of both EGR-1 and c-Myc in angiogenesis. Fahmy et al. [239] reported that inhibition of EGR-1 expression repressed neovascularization and blocked angiogenesis and tumor growth in mouse and rat models. Baudino et al. [241] reported that c-Myc is also required for the proper expression of several major angiogenic factors, and that c-Myc(-/-) ES cells are dramatically impaired in their ability to form tumors in immune-compromised mice, and the small tumors that do develop are poorly vascularized. Here, we proposed a possible synergism between c-Myc and EGR-1 in transcriptional regulation of target genes, and experimentally demonstrated the binding of both c-Myc and EGR-1 to several target promoters.

Computational identification of enriched TF hits on a set of co-expressed genes points to a role of the respective TF in the regulation of these genes. However, such an *in-silico* approach alone cannot precisely and uniquely decipher the regulatory effect of the TF. It is possible that the enriched TF acts as an activator and hence its

binding to the regulatory elements is necessary for the induction of the analyzed genes. On the other hand, it cannot be ruled out that the enriched TF acts as a suppressor, and the removal of its binding to the gene promoters is required for their induction. For example, we identified a significant enrichment for NF-Y binding site signature on promoters of genes whose expression peaks in the G2 phase of the cell cycle. This observation alone does not tell us whether NF-Y is an activator or repressor of these promoters at this cell cycle phase. Additional experiments are needed to understand the nature of the regulatory effect of this TF. The expression pattern of a TF may point to its role in cases where its activity is regulated at the transcriptional level. In the above example, if the expression of NF-Y itself was to peak at G2 as well, it would strongly suggest that it acts as an activator, and vice versa. Similarly, regulatory modules that are identified computationally based on significant co-occurrence of TF hits cannot by themselves reveal the regulatory effect of each member of the module, nor the nature of the interplay between them. For example, we identified significant co-occurrence of c-Myc and EGR1 binding site signatures, which suggests the involvement of both TFs in the regulation of a large set of common targets. But we cannot know whether these two elements are synergistic or antagonistic in the induction of their targets. Nor can computational means tell us whether these two TFs bind their common targets simultaneously, or their binding is mutually exclusive. It is also possible that these two TFs control a common set of genes but each regulator is activated in response to different triggers. Again, correlations between the expression of these regulators and that of the genes putatively regulated by this module can elucidate the regulation logic.

Dissection of DNA damage responses using a combination of microarrays and RNAi. The results in Section 4.5 represent another test-case study. Fine dissection of

complex transcriptional responses has posed a long-standing challenge in the signal transduction field. External and internal stimuli may activate complex networks whose analysis by traditional biochemistry can be daunting. The DNA damage response is an example of such a complex network. This highly branched signaling web spans numerous aspects of cellular metabolism and involves a vigorous wave of gene transcription across the genome. The combination of gene expression microarrays, manipulation of genes activity using siRNAs, and powerful computational tools holds promise for systematic and rapid dissection of such networks.

Our analysis provides a proof-of-principle for the power of this combined experimental approach, despite possible nonspecific effects of RNAi [180-183], which can be neutralized by controlled experimental design and computational analysis of the data. One way to filter out off-target effects is to use several different siRNA sequences against the same target on the assumption that completely different siRNAs will not induce the same off-target effects [37, 183]. Following this logic, dissecting a signaling pathway that is mediated by several regulators using independent targeting of these regulators should also boost confidence. In this case, overlapping sets of genes whose expression is attenuated by knocking down different regulators are unlikely to be a result of off-target effects. It is also important to show that the observed effects are not a general consequence of the expression of siRNAs in the cells.

In our study we focused on two arms in the DNA damage-induced network that are mediated by the ATM/NF-κB and the ATM/p53 regulators. First, we identified a set of genes whose induction in response to DNA damage was abrogated in cells knocked-down for two different components of the damage-induced signaling

pathway, ATM and the Rel-A subunit of NF-κB. Importantly, the induction of these genes was not disrupted in cells expressing siRNA against LacZ and was only mildly attenuated in cells knocked-down for p53, indicating that the loss of induction was not a general nonspecific consequence of siRNA expression. Moreover, computational promoter analysis showed that the set of promoters of these genes was highly and specifically enriched for the binding site signature of NF-κB, providing independent evidence of the accuracy of this analysis. We then identified a set of genes whose induction in response to DNA damage was significantly abrogated in cells knocked-down for ATM and p53, but not in cells knocked-down for the Rel-A subunit of NF-κB, or in the LacZ control. Again, it is unlikely this dissection of the ATM/p53-mediated arm can be ascribed to nonspecific or off-targets effects. According to computational promoter analysis, this set was highly enriched for the binding signature of ATF2/ATF3/Jun, a secondary-transcriptional pathway whose induction was indeed p53-dependent in our data. This observation is in agreement with several studies that reported p53-dependent activation of this transcriptional pathway in response to DNA damage [186, 187]. However, evidence suggests that p53-dependence of the induction of the ATF2/ATF3/Jun pathway depends on the cellular context, the type of DNA lesion or the extent of damage, as p53-independent induction of this pathway was observed in other studies [242, 243].

In summary, we have dissected the network into two major arms, the ATM/NF-κB- and the ATM/p53-dependent arms. Statistical tests coupled with computational promoter analysis demonstrated that this dissection was highly accurate. Given the success of this pilot study, we are pursuing this strategy to obtain finer dissection of the cellular responses to DNA damage. Cellular systems knocked-down for all the TFs that are known to be involved in DNA damage responses are being established in

our lab. Analyzing these systems at several timepoints after exposure to damaging agents will allow us to delineate kinetic waves in the induced network. The results we obtained indicated that, in the cells we used, a large component of the p53-dependent transcriptional response was mediated by activation of a second wave of transcriptional induction controlled by ATF/Jun TFs. Examination of the damage response in cells knocked-down for these regulators will allow us to test this model. Our results also suggested that most of the first wave of the transcriptional response is transmitted by the ATM-dependent activation of p53 and NF-κB. It will be interesting to probe the transcriptional response induced by DNA damage in cells knocked-down for both p53 and NF-κB. In our follow-up studies we will construct systems knocked-down for combinations (pairs, triplets) of regulators. This should allow further elucidation of interlinks among key players in this network.

Delineation of transcriptional responses to ionizing radiation in murine lymphoid tissues. In the study described in Section 4.6, we reversed-engineered components of the transcriptional network induced by IR in murine lymphoid tissue. Using the microarray technology, we first identified a prominent cluster of genes whose activation by IR was Atm-dependent. Then, using a computational method, we searched the promoters of these genes for over-represented cis-regulatory elements. Without any bias of prior knowledge, PRIMA revealed highly significant enrichment for NF-κB and p53 binding site signatures, suggesting that these two transcription factors are the major transcriptional regulators in the Atm-dependent response to IR in lymphoid tissues. These results are in agreement with previous studies that reported compromised IR-induced activation of both NF-κB and p53 in murine Atm-deficient tissues and in cell lines derived from A-T patients [244-247]. Focusing on the putative NF-κB-mediated arm, we biochemically validated the Atm dependence of IR-induced

enhancement in DNA binding activity of NF-κB in lymphoid cells. Our approach could thus dissect two major arms in the Atm-mediated transcriptional response, according to the transcriptional regulators that function downstream of Atm.

The mechanisms by which ATM activates and stabilizes p53 are well-established. ATM directly phosphorylates p53 as well as its inhibitor and E3 ubiquitin ligase, Mdm2, and the checkpoint kinase Chk2, which in turn phosphorylates p53 on yet another site (reviewed by Shiloh, 2003 [93]). p53's response to DNA damage also depends on Mdm2-dependent proteolysis of Mdmx, a homologue of Mdm2 that represses p53's transactivation function [248]. Recently, Yaron Pereg of our lab showed that efficient damage-induced degradation of human Hdmx depends on its direct phosphorylation by ATM in response to DSBs [116]. All these ATM-dependent modifications contribute to the stabilization and rapid accumulation of p53 in response to IR-induced DNA damage. In contrast, the mechanisms by which ATM activates the NF-κB pathway remain elusive. Li et al. [245] demonstrated that ATM is required for DSB-induced activation of the NF-κB pathway, including activation of the IKK complex that phosphorylates IκBα, NF-κB's inhibitor. Recently, Hur et al. [249] reported the death domain kinase RIP to be an essential component upstream to IKK in the activation of NF-κB by DNA damage. Importantly, RIP was demonstrated to physically interact with IKK upon exposure of cells to DNA damaging agents, and this interaction was ATM-dependent. The direct substrate(s) of ATM in this signaling pathway and how the alarm signal is propagated from the nucleus to the cytoplasm remain to be determined. A recent study pointed to the IKBKG (NEMO/IKKγ) subunit of the IKK complex as a key player, which shuttles between the nucleus and the cytoplasm, in the transmission of the signal [250].

Importantly, we observed that the Atm-dependent response contained several p53-direct targets with major pro-apoptotic role, as well as several NF-κB-direct targets with anti-apoptotic function. Taken together, our findings suggest a model in which, in the response of lymphoid cells to IR, pro- and anti-apoptotic signals are induced in parallel, the former being mediated by NF-κB and the latter by p53, while the activation of both is Atm-dependent. This model is in strong agreement with the results recently reported by Stankovic et al. [199], who examined IR-responses in three groups of B-CLL tumor cells: *ATM/TP53*-proficient, *ATM*-mutant/*TP53*-w.t. and *ATM*-w.t./*TP53*-mutant B-CLL tumors. In line with our results, Satnkovic et al. observed parallel induction of pro-apoptotic ATM- and p53-dependent transcriptional response, and ATM-dependent, p53-independent pro-survival response. Applying PRIMA to this dataset pointed to NF-κB as the missing piece in the puzzle, i.e., the major regulator downstream to ATM that mediates the anti-apoptotic arm in lymphoid cells. Indeed, Weston et al. [251] very recently showed that increased NF-κB signaling confers acute lymphoblastic leukemia tumors with resistance to IR-induced apoptosis. However, a model that depicts the ATM-p53 and ATM-NF-κB pathways as parallel, linear and independent is probably over-simplified. Accumulating data derived from several cell types suggest that the p53- and NF-κB-mediated arms maintain multiple cross-talks [252-255], which indicate that the logic of the IR-induced response network and the balance between apoptotic and survival signals are highly intricate and depend on the cellular context.

Taken together, our findings on the lymph nodes and B-CLL datasets further elucidate the molecular network induced by IR, and might have implications for cancer management. They suggest that restoring the p53-mediated apoptotic arm while blocking the NF-κB-mediated pro-survival arm could effectively increase the

radiosensitivity of lymphoid tumors. This may point the way to the development of new, effective therapeutic protocols for leukemias of lymphoid origin.

**Future prospects.**

Phylogenetic footprinting. A major difficulty in computational promoter analysis stems from the fact that cis-regulatory elements recognized and bound by TFs are typically very short and highly flexible. Therefore, genome-wide computational scans for putative TF binding sites inevitably yield many false positive hits. We have demonstrated that this problem does not prevent successful reverse engineering of transcriptional networks in human cells. However, because of the significant false positive rate, our analyses were focused on the identification of global statistical phenomena in the studied target set of promoters, and less on yielding high confidence lists of putative TF targets. The availability of the genome sequences of multiple species in addition to the human genome is expected to greatly boost the specificity of *in-silico* identification of regulatory elements embedded in the genome [81-83, 256, 257]. The higher selective pressure imposed on functional elements makes them more conserved than their surrounding non-functional DNA. Several studies demonstrated the utility of computational identification of evolutionary conserved elements, an approach called *phylogenetic footprinting*, in drastically reducing false-positive hit rates [81, 82, 257]. Thus, we downloaded genome-wide promoter data for twelve organisms and searched for conservation of various motifs across species. In a preliminary analysis, we identified core promoter elements that are conserved throughout evolution from yeast to humans, as well as elements specific to certain species groups. We intend to integrate multi-organism data in our promoter analysis tool. This should allow us not only to reveal the major regulators

that control observed transcriptional programs, but also to point out, by computational means alone and with very high specificity, target genes that are controlled by each transcription factor.

Future generation gene expression microarrays. Probes deposited on most current gene expression microarrays do not allow the distinction between splice variants of the target genes. Therefore, this important layer of modulation of gene activity is left completely uncovered by most expression profiling studies. Yet, it is becoming clear that alternative splicing occurs in a large proportion of mammalian genes and is a central contributor to increasing the diversity of the mammalian proteome [258]. A major development in this field is the design of microarrays deposited with exon-specific or exon-junction probes [259, 260]. Using such chips will shed new light on roles of alternative splice variants in different developmental stages, and on the involvement of specific variants in pathological conditions. The use of such exon-chips should also boost our ability to decipher regulatory signals that control alternative splicing and determine which variant is expressed in which spatial-temporal conditions.

Another major advance in the field is the development of genome-tiling microarrays with probes that will eventually cover the entire genome. These arrays will enable measuring transcription from various regions of the genome without bias towards location of known genes. Several pioneering studies with such chips turned up evidence of large amounts of transcription outside the boundaries of known genes [261-264], suggesting that the universe of gene expression is much broader than currently thought (these newly identified entities in the cellular transcriptome were recently referred to as the 'dark matter' in the genome [265]). Of special note is the constantly growing family of genes encoding for regulatory micro-RNAs (miRNAs)

[266, 267]. miRNAs are endogenous, ~ 22 nt RNAs that are assumed to play important regulatory roles in animal development by controlling gene activity through targeting mRNAs for degradation or translational repression. In human, as of 2004 some 170 unique miRNAs have been identified and validated [266]. The total number of genes that encode for miRNA precursors is probably much higher. Elucidation of these genes and their functions will be significantly enhanced with the maturation of genome-tiling microarrays.

The mechanism by which miRNAs exert their regulatory role is believed to share many features with RNAi mechanisms [266]. According to current models, pre-miRNAs that acquire hairpin double-stranded secondary structure are processed by the Dicer complex, resulting in short single-stranded miRNAs that are loaded onto the RNA-induced silencing complex (RISC). The loaded miRNAs direct the RISC apparatus to downregulate the expression of their target genes either by mRNA cleavge or repression of mRNA translation into proteins. Both processes are assumed to be mediated by various degrees of complementarity between miRNAs and their mRNA targets; high homology is believed to favor degradation while more modest homology between miRNA and its target mRNA is believed to favor translational repression. Regulatory elements through which many of the miRNA discovered to date are embedded in the 3'-UTR region of their mRNA targets. Current knowledge on such regulatory elements is scant. Adopting a strategy that combines gene expression profiling with genome-tiling microarrays and phylogenetic footprinting focused on 3'-UTR regions is expected to disclose many new functional elements that control gene expression by means of miRNA-directed suppression. A promising indication for the potential of the comparative genomics approach in deciphering 3'-UTR regulatory elements was recently demonstrated by Xie et al. [268], who

computationally identified more than a hundred 3'-UTR motifs likely involved in post-transcriptional regulation.

Proteomic technologies. Measurement of protein levels and post-translational modifications is much more challenging than nucleic acid measurements. It is not surprising, therefore, that technologies for large-scale profiling of the cellular proteome lag behind those for profiling the transcriptome. Nevertheless, one can expect considerable progress in the proteomics technologies in the coming years. Availability of robust proteomic chips, in addition to the genome-tiling ones, will allow simultaneous profiling of TF-DNA interactions, gene expression and protein levels and modifications. Reverse engineering of cellular regulatory networks (not limited to the tier of transcriptional regulation) from such multi-layer data will necessitate the development of novel algorithms, and pose one of the greatest bioinformatics challenges in the years to come. Having such experimental and computational tools will allow generating detailed mechanistic models for the cellular function. This will eventually pave the way to models that are detailed and accurate enough to implement computational simulations of the living cell [269, 270]. The impact of accurate simulations of the functions of human cellular systems on biomedical research cannot be overestimated.

We do, however, anticipate intriguing findings that may throw into question some basic assumptions. One such surprising finding that still averts scientific attention relates to our interpretation of gene expression data. The interpretation of expression data for analysis of genes function rests on several assumptions. First, it is assumed that evolutionary selection was tight enough to ensure that genes are expressed only under conditions in which their products are needed for the proper functioning of the cell/tissue. Second, it assumes that regulation of transcription is the pivotal factor for

regulation of gene activity. These assumptions are shaken by experiments documenting low correlation between changes in mRNA and protein levels in cells [271, 272], and poor overlap between genes induced in response to perturbations and genes whose deletion compromises fitness of the cells to the same perturbations [273, 274]. It is conceivable that some fraction of gene expression has no significant functional role. Future studies will have to address this issue.

In this work, we developed several novel functional genomics approaches and applied them to the study of cellular responses to DNA damage. Our results demonstrated that the new paradigm of systems biology provides global delineation of complex cellular networks. Although systems biology is in its infancy, it is already a vital part of modern biomedical research. Its potential benefits are enormous in both scientific and practical terms. Advances in the field will enable us to construct mechanistic models for the operation of the cellular systems, test and refine them using experimental approaches, and gradually witness the emergence of robust, dynamic, adapting, and developing systems from the information encoded in the genomes. Gaining such understanding will elucidate causal relationships between defective components (e.g., mutated genes) and compromised biological systems (i.e., abnormal phenotypes of organisms). Thus, systems biology is expected to impact on clinical medicine as well as on pharmaceutical industries. This emerging field will eventually provide us with detailed mechanistic models for the etiology of diseases, pointing the way to novel strategies for rational intervention in pathological conditions and the design of improved personalized drugs.

# References


1. **Aggarwal K, Lee KH: Functional genomics and proteomics as a foundation for systems biology. *Brief Funct Genomic Proteomic* 2003, 2:175-184.**
2. **Ehrenberg M, Elf J, Aurell E, Sandberg R, Tegner J: Systems biology is taking off. *Genome Res* 2003, 13:2377-2380.**
3. **Ideker T: Systems biology 101--what you need to know. *Nat Biotechnol* 2004, 22:473-475.**
4. **Ideker T, Galitski T, Hood L: A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2001, 2:343-372.**
5. **Kitano H: Computational systems biology. *Nature* 2002, 420:206-210.**
6. **Kitano H: Systems biology: a brief overview. *Science* 2002, 295:1662-1664.**
7. **Hood L, Heath JR, Phelps ME, Lin B: Systems biology and new technologies enable predictive and preventative medicine. *Science* 2004, 306:640-643.**
8. **Lockhart DJ, Winzeler EA: Genomics, gene expression and DNA arrays. *Nature* 2000, 405:827-836.**
9. **Lander ES, Weinberg RA: Genomics: journey to the center of biology. *Science* 2000, 287:1777-1782.**
10. **Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, et al: Large-scale genotyping of complex DNA. *Nat Biotechnol* 2003, 21:1233-1237.**
11. **Middleton FA, Pato MT, Gentile KL, Morley CP, Zhao X, Eisener AF, Brown A, Petryshen TL, Kirby AN, Medeiros H, et al: Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am J Hum Genet* 2004, 74:886-897.**
12. **Mitra N, Ye TZ, Smith A, Chuai S, Kirchhoff T, Peterlongo P, Nafa K, Phillips MS, Offit K, Ellis NA: Localization of cancer susceptibility genes by genome-wide single-nucleotide polymorphism linkage-disequilibrium mapping. *Cancer Res* 2004, 64:8116-8125.**
13. **Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, et al: Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 2004, 1:109-111.**
14. **Mantripragada KK, Buckley PG, de Stahl TD, Dumanski JP: Genomic microarrays in the spotlight. *Trends Genet* 2004, 20:87-94.**
15. **Snijders AM, Pinkel D, Albertson DG: Current status and future prospects of array-based comparative genomic hybridisation. *Brief Funct Genomic Proteomic* 2003, 2:37-45.**
16. **Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995, 270:467-470.**
17. **Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. *Nat Genet* 1999, 21:10-14.**
18. **Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: High density synthetic oligonucleotide arrays. *Nat Genet* 1999, 21:20-24.**

19. **Velculescu VE, Vogelstein B, Kinzler KW: Analysing uncharted transcriptomes with SAGE. *Trends Genet* 2000, 16:423-425.**
20. **Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 1999, 96:2907-2912.**
21. **Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D: Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* 2002, 13:1977-2000.**
22. **Lee CK, Klopp RG, Weindruch R, Prolla TA: Gene expression profile of aging and its retardation by caloric restriction. *Science* 1999, 285:1390-1393.**
23. **Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000, 403:503-511.**
24. **Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al: Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002, 298:799-804.**
25. **Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al: Genome-wide location and function of DNA binding proteins. *Science* 2000, 290:2306-2309.**
26. **Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD: E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* 2002, 16:245-256.**
27. **Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, et al: Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A* 2003, 100:12247-12252.**
28. **Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B: A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci U S A* 2003, 100:8164-8169.**
29. **Aebersold R, Mann M: Mass spectrometry-based proteomics. *Nature* 2003, 422:198-207.**
30. **Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, et al: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002, 415:180-183.**
31. **Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, et al: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, 415:141-147.**
32. **Forler D, Kocher T, Rode M, Gentzel M, Izaurralde E, Wilm M: An efficient protein complex purification method for functional proteomics in higher eukaryotes. *Nat Biotechnol* 2003, 21:89-92.**

**33. Mann M, Ong SE, Gronborg M, Steen H, Jensen ON, Pandey A: Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol* 2002, 20:261-268.**
**34. Sreekumar A, Nyati MK, Varambally S, Barrette TR, Ghosh D, Lawrence TS, Chinnaiyan AM: Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. *Cancer Res* 2001, 61:7585-7593.**
**35. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, et al: Global analysis of protein activities using proteome chips. *Science* 2001, 293:2101-2105.**
**36. Hannon GJ: RNA interference. *Nature* 2002, 418:244-251.**
**37. Hannon GJ, Rossi JJ: Unlocking the potential of the human genome with RNA interference. *Nature* 2004, 431:371-378.**
**38. Brummelkamp TR, Bernards R, Agami R: A system for stable expression of short interfering RNAs in mammalian cells. *Science* 2002, 296:550-553.**
**39. Brummelkamp TR, Bernards R, Agami R: Stable suppression of tumorigenicity by virus-mediated RNA interference. *Cancer Cell* 2002, 2:243-247.**
**40. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, et al: Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* 2003, 421:231-237.**
**41. Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K, et al: A resource for large-scale RNA-interference-based screens in mammals. *Nature* 2004, 428:427-431.**
**42. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, et al: A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* 2004, 428:431-437.**
**43. Knight J: When the chips are down. *Nature* 2001, 410:860-861.**
**44. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G: Making and reading microarrays. *Nat Genet* 1999, 21:15-19.**
**45. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ: Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotechnol* 1997, 15:1359-1367.**
**46. Naef F, Lim DA, Patil N, Magnasco M: DNA hybridization to mismatched templates: a chip study. *Phys Rev E Stat Nonlin Soft Matter Phys* 2002, 65:040902.**
**47. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003, 4:249-264.**
**48. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, et al: Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001, 19:342-347.**
**49. Barczak A, Rodriguez MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP, Erle DJ: Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res* 2003, 13:1775-1785.**

**50.** **Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, Churchill G: A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J Biomol Tech* 2004, 15:276-284.**

**51.** **Park PJ, Cao YA, Lee SY, Kim JW, Chang MS, Hart R, Choi S: Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J Biotechnol* 2004, 112:225-245.**

**52.** **Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, et al: Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2005, 2:351-356.**

**53.** **Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, et al: Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2005, 2:345-350.**

**54.** **Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: Independence and reproducibility across microarray platforms. *Nat Methods* 2005, 2:337-344.**

**55.** **Pilpel Y, Sudarsanam P, Church GM: Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* 2001, 29:153-159.**

**56.** **Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003, 34:166-176.**

**57.** **Segal E, Yelensky R, Koller D: Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* 2003, 19 Suppl 1:i273-282.**

**58.** **Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. *Nat Genet* 1999, 22:281-285.**

**59.** **Jelinsky SA, Samson LD: Global response of Saccharomyces cerevisiae to an alkylating agent. *Proc Natl Acad Sci U S A* 1999, 96:1486-1491.**

**60.** **Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998, 9:3273-3297.**

**61.** **Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 2000, 11:4241-4257.**

**62.** **Murray JI, Whitfield ML, Trinklein ND, Myers RM, Brown PO, Botstein D: Diverse and specific gene expression responses to stresses in cultured human cells. *Mol Biol Cell* 2004, 15:2361-2374.**

**63.** **Furlong EE, Andersen EC, Null B, White KP, Scott MP: Patterns of gene expression during Drosophila mesoderm development. *Science* 2001, 293:1629-1633.**

**64.** **Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nat Genet* 2002, 31:255-265.**

**65.** **van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al: A gene-**

**expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002, 347:1999-2009.**

**66. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, 415:530-536.**

**67. Bohen SP, Troyanskaya OG, Alter O, Warnke R, Botstein D, Brown PO, Levy R: Variation in gene expression patterns in follicular lymphoma and the response to rituximab. *Proc Natl Acad Sci U S A* 2003, 100:1926-1930.**

**68. Schubert CM: Microarray to be used as routine clinical screen. *Nat Med* 2003, 9:9.**

**69. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999, 286:531-537.**

**70. Debouck C, Goodfellow PN: DNA microarrays in drug discovery and development. *Nat Genet* 1999, 21:48-50.**

**71. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al: Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* 1998, 4:1293-1301.**

**72. Hamadeh HK, Amin RP, Paules RS, Afshari CA: An overview of toxicogenomics. *Curr Issues Mol Biol* 2002, 4:45-56.**

**73. Jelinsky SA, Estep P, Church GM, Samson LD: Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. *Mol Cell Biol* 2000, 20:8157-8167.**

**74. Frith MC, Spouge JL, Hansen U, Weng Z: Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res* 2002, 30:3214-3224.**

**75. Aerts S, Van Loo P, Thijs G, Mayer H, de Martin R, Moreau Y, De Moor B: TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res* 2005, 33:W393-396.**

**76. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994, 2:28-36.**

**77. Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000, 296:1205-1214.**

**78. Eskin E, Pevzner PA: Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 2002, 18 Suppl 1:S354-363.**

**79. Workman CT, Stormo GD: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput* 2000:467-478.**

**80. Barash Y, Elidan G, Kaplan T, Friedman N: CIS: compound importance sampling method for protein-DNA binding site p-value estimation. *Bioinformatics* 2005, 21:596-600.**

**81. Duret L, Bucher P: Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* 1997, 7:399-406.**

**82.** **Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: Identification of conserved regulatory elements by comparative genome analysis.** ***J Biol*** **2003, 2:13.**

**83.** **Sandelin A, Wasserman WW, Lenhard B: ConSite: web-based prediction of regulatory elements using cross-species comparison.** ***Nucleic Acids Res*** **2004, 32:W249-252.**

**84.** **Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints.** ***J Mol Biol*** **1988, 203:439-455.**

**85.** **Sudarsanam P, Pilpel Y, Church GM: Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae.** ***Genome Res*** **2002, 12:1723-1731.**

**86.** **Halfon MS, Grad Y, Church GM, Michelson AM: Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** ***Genome Res*** **2002, 12:1019-1028.**

**87.** **Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** ***Proc Natl Acad Sci U S A*** **2002, 99:757-762.**

**88.** **Markstein M, Markstein P, Markstein V, Levine MS: Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** ***Proc Natl Acad Sci U S A*** **2002, 99:763-768.**

**89.** **Kel A, Kel-Margoulis O, Babenko V, Wingender E: Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells.** ***J Mol Biol*** **1999, 288:353-376.**

**90.** **Wasserman WW, Fickett JW: Identification of regulatory regions which confer muscle-specific gene expression.** ***J Mol Biol*** **1998, 278:167-181.**

**91.** **Frech K, Quandt K, Werner T: Muscle actin genes: a first step towards computational classification of tissue specific promoters.** ***In Silico Biol*** **1998, 1:29-38.**

**92.** **Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** ***Bioinformatics*** **2003, 19 Suppl 1:i283-291.**

**93.** **Shiloh Y: ATM and related protein kinases: safeguarding genome integrity.** ***Nat Rev Cancer*** **2003, 3:155-168.**

**94.** **Norbury CJ, Hickson ID: Cellular responses to DNA damage.** ***Annu Rev Pharmacol Toxicol*** **2001, 41:367-401.**

**95.** **Coultas L, Strasser A: The molecular control of DNA damage-induced cell death.** ***Apoptosis*** **2000, 5:491-507.**

**96.** **Shiloh Y, Kastan MB: ATM: genome stability, neuronal development, and cancer cross paths.** ***Adv Cancer Res*** **2001, 83:209-254.**

**97.** **Kaneko H, Kondo N: Clinical features of Bloom syndrome and function of the causative gene, BLM helicase.** ***Expert Rev Mol Diagn*** **2004, 4:393-401.**

**98.** **Magnaldo T, Sarasin A: Xeroderma pigmentosum: from symptoms and genetics to gene-based skin therapy. *Cells Tissues Organs* 2004, 177:189-198.**
**99.** **Charames GS, Bapat B: Genomic instability and cancer. *Curr Mol Med* 2003, 3:589-596.**
**100.** **Jo WS, Chung DC: Genetics of hereditary colorectal cancer. *Semin Oncol* 2005, 32:11-23.**
**101.** **Chun HH, Gatti RA: Ataxia-telangiectasia, an evolving phenotype. *DNA Repair (Amst)* 2004, 3:1187-1196.**
**102.** **Digweed M, Sperling K: Nijmegen breakage syndrome: clinical manifestation of defective response to DNA double-strand breaks. *DNA Repair (Amst)* 2004, 3:1207-1217.**
**103.** **Abner CW, McKinnon PJ: The DNA double-strand break response in the nervous system. *DNA Repair (Amst)* 2004, 3:1141-1147.**
**104.** **Caldecott KW: DNA single-strand breaks and neurodegeneration. *DNA Repair (Amst)* 2004, 3:875-882.**
**105.** **Uberti D, Ferrari Toninelli G, Memo M: Involvement of DNA damage and repair systems in neurodegenerative process. *Toxicol Lett* 2003, 139:99-105.**
**106.** **Laiho M, Latonen L: Cell cycle control, DNA damage checkpoints and cancer. *Ann Med* 2003, 35:391-397.**
**107.** **Fornace AJ, Jr., Amundson SA, Bittner M, Myers TG, Meltzer P, Weinsten JN, Trent J: The complexity of radiation stress responses: analysis by informatics and functional genomics approaches. *Gene Expr* 1999, 7:387-400.**
**108.** **Koch-Paiz CA, Amundson SA, Bittner ML, Meltzer PS, Fornace AJ, Jr.: Functional genomics of UV radiation responses in human cells. *Mutat Res* 2004, 549:65-78.**
**109.** **Amundson SA, Bittner M, Fornace AJ, Jr.: Functional genomics as a window on radiation stress signaling. *Oncogene* 2003, 22:5828-5833.**
**110.** **Gudkov AV, Komarova EA: The role of p53 in determining sensitivity to radiotherapy. *Nat Rev Cancer* 2003, 3:117-129.**
**111.** **Komarova EA, Diatchenko L, Rokhlin OW, Hill JE, Wang ZJ, Krivokrysenko VI, Feinstein E, Gudkov AV: Stress-induced secretion of growth inhibitors: a novel tumor suppressor function of p53. *Oncogene* 1998, 17:1089-1096.**
**112.** **Burns TF, Bernhard EJ, El-Deiry WS: Tissue specific expression of p53 target genes suggests a key role for KILLER/DR5 in p53-dependent apoptosis in vivo. *Oncogene* 2001, 20:4601-4612.**
**113.** **Rosen EM, Fan S, Goldberg ID, Rockwell S: Biological basis of radiation sensitivity. Part 2: Cellular and molecular determinants of radiosensitivity. *Oncology (Huntingt)* 2000, 14:741-757; discussion 757-748, 761-746.**
**114.** **Rosen EM, Fan S, Goldberg ID, Rockwell S: Biological basis of radiation sensitivity. Part 1: Factors governing radiation tolerance. *Oncology (Huntingt)* 2000, 14:543-550.**
**115.** **Kurz EU, Lees-Miller SP: DNA damage-induced activation of ATM and ATM-dependent signaling pathways. *DNA Repair (Amst)* 2004, 3:889-900.**
**116.** **Pereg Y, Shkedy D, de Graaf P, Meulmeester E, Edelson-Averbukh M, Salek M, Biton S, Teunisse AF, Lehmann WD, Jochemsen AG, Shiloh Y:**

**Phosphorylation of Hdmx mediates its Hdm2- and ATM-dependent degradation in response to DNA damage. *Proc Natl Acad Sci U S A* 2005, 102:5056-5061.**

**117. Cortez D, Wang Y, Qin J, Elledge SJ: Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks. *Science* 1999, 286:1162-1166.**

**118. Li S, Ting NS, Zheng L, Chen PL, Ziv Y, Shiloh Y, Lee EY, Lee WH: Functional link of BRCA1 and ataxia telangiectasia gene product in DNA damage response. *Nature* 2000, 406:210-215.**

**119. Lee JH, Paull TT: ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science* 2005, 308:551-554.**

**120. Uziel T, Lerenthal Y, Moyal L, Andegeko Y, Mittelman L, Shiloh Y: Requirement of the MRN complex for ATM activation by DNA damage. *Embo J* 2003, 22:5612-5621.**

**121. Abraham RT, Tibbetts RS: Cell biology. Guiding ATM to broken DNA. *Science* 2005, 308:510-511.**

**122. Oka A, Takashima S: Expression of the ataxia-telangiectasia gene (ATM) product in human cerebellar neurons during development. *Neurosci Lett* 1998, 252:195-198.**

**123. Barlow C, Ribaut-Barassin C, Zwingman TA, Pope AJ, Brown KD, Owens JW, Larson D, Harrington EA, Haeberle AM, Mariani J, et al: ATM is a cytoplasmic protein in mouse brain required to prevent lysosomal accumulation. *Proc Natl Acad Sci U S A* 2000, 97:871-876.**

**124. Stewart GS, Maser RS, Stankovic T, Bressan DA, Kaplan MI, Jaspers NG, Raams A, Byrd PJ, Petrini JH, Taylor AM: The DNA double-strand break repair gene hMRE11 is mutated in individuals with an ataxia-telangiectasia-like disorder. *Cell* 1999, 99:577-587.**

**125. Frappart PO, Tong WM, Demuth I, Radovanovic I, Herceg Z, Aguzzi A, Digweed M, Wang ZQ: An essential function for NBS1 in the prevention of ataxia and cerebellar defects. *Nat Med* 2005, 11:538-544.**

**126. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: Network motifs: simple building blocks of complex networks. *Science* 2002, 298:824-827.**

**127. Shen-Orr SS, Milo R, Mangan S, Alon U: Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* 2002, 31:64-68.**

**128. Karp PD: Pathway databases: a case study in computational symbolic theories. *Science* 2001, 293:2040-2044.**

**129. Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, 28:27-30.**

**130. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Jr., Kyrpides N, Fonstein M, Maltsev N, Selkov E: WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* 2000, 28:123-125.**

**131. Takai-Igarashi T, Nadaoka Y, Kaminuma T: A database for cell signaling networks. *J Comput Biol* 1998, 5:747-754.**

**132. Schacherer F, Choi C, Gotze U, Krull M, Pistor S, Wingender E: The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics* 2001, 17:1053-1057.**

**133. van Helden J, Naim A, Lemer C, Mancuso R, Eldridge M, Wodak SJ: From molecular activities and processes to biological function. *Brief Bioinform* 2001, 2:81-93.**
**134. van Helden J, Naim A, Mancuso R, Eldridge M, Wernisch L, Gilbert D, Wodak SJ: Representing and analysing molecular and cellular function using the computer. *Biol Chem* 2000, 381:921-935.**
**135. Bader GD, Betel D, Hogue CW: BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 2003, 31:248-250.**
**136. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, et al: Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005, 33:D428-432.**
**137. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 2002, 31:19-20.**
**138. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, 13:2498-2504.**
**139. Ideker T, Ozier O, Schwikowski B, Siegel AF: Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002, 18 Suppl 1:S233-240.**
**140. Li C, Wong WH: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 2001, 98:31-36.**
**141. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003, 19:185-193.**
**142. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004, 5:R80.**
**143. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002, 30:e15.**
**144. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998, 95:14863-14868.**
**145. Sharan R, Shamir R: CLICK: a clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol* 2000, 8:307-316.**
**146. Sharan R, Elkon R, Shamir R: Cluster analysis and its applications to gene expression data. *Ernst Schering Res Found Workshop* 2002:83-108.**
**147. Tanay A, Sharan R, Kupiec M, Shamir R: Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* 2004, 101:2981-2986.**

148. Tanay A, Sharan R, Shamir R: Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 2002, 18 Suppl 1:S136-144.
149. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25:25-29.
150. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res* 2003, 13:773-780.
151. Maglott DR, Katz KS, Sicotte H, Pruitt KD: NCBI's LocusLink and RefSeq. *Nucleic Acids Res* 2000, 28:126-128.
152. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 2000, 28:316-319.
153. Praz V, Perier R, Bonnard C, Bucher P: The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res* 2002, 30:322-324.
154. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, et al: An overview of Ensembl. *Genome Res* 2004, 14:925-928.
155. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: The Ensembl automatic gene annotation system. *Genome Res* 2004, 14:942-950.
156. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al: Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 2002, 420:563-573.
157. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, et al: The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* 2004, 14:2121-2127.
158. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, et al: Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* 2004, 2:e162.
159. Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, 16:16-23.
160. Orlev N, Shamir R, Shiloh Y: PIVOT: protein interacions visualizatiOn tool. *Bioinformatics* 2004, 20:424-425.
161. Hirao A, Kong YY, Matsuoka S, Wakeham A, Ruland J, Yoshida H, Liu D, Elledge SJ, Mak TW: DNA damage-induced activation of p53 by the checkpoint kinase Chk2. *Science* 2000, 287:1824-1827.
162. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, et al: The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 2005, 33:D418-424.
163. Ishida S, Huang E, Zuzan H, Spang R, Leone G, West M, Nevins JR: Role for E2F in control of both DNA replication and mitotic functions as

**revealed from DNA microarray analysis. *Mol Cell Biol* 2001, 21:4684-4699.**

**164. Polager S, Kalma Y, Berkovich E, Ginsberg D: E2Fs up-regulate expression of genes involved in DNA replication, DNA repair and mitosis. *Oncogene* 2002, 21:437-446.**

**165. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ: Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J Mol Biol* 2001, 309:99-120.**

**166. Mantovani R: A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res* 1998, 26:1135-1143.**

**167. Blanchette M, Schwikowski B, Tompa M: Algorithms for phylogenetic footprinting. *J Comput Biol* 2002, 9:211-223.**

**168. Vavouri T, Elgar G: Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. *Curr Opin Genet Dev* 2005, 15:395-402.**

**169. Romano LA, Wray GA: Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development* 2003, 130:4187-4199.**

**170. Erives A, Levine M: Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci U S A* 2004, 101:3851-3856.**

**171. Simpson P: Evolution of development in closely related species of flies and worms. *Nat Rev Genet* 2002, 3:907-917.**

**172. Dieterich C, Rahmann S, Vingron M: Functional inference from non-random distributions of conserved predicted transcription factor binding sites. *Bioinformatics* 2004, 20 Suppl 1:I109-I115.**

**173. Linhart C, Elkon R, Shiloh Y, Shamir R: Deciphering transcriptional regulatory elements that encode specific cell cycle phasing by comparative genomics analysis. *Cell Cycle* 2005, 4:1788-1797.**

**174. Elkon R, Zeller KI, Linhart C, Dang CV, Shamir R, Shiloh Y: In silico identification of transcriptional regulators associated with c-Myc. *Nucleic Acids Res* 2004, 32:4955-4961.**

**175. Nilsson JA, Cleveland JL: Myc pathways provoking cell suicide and cancer. *Oncogene* 2003, 22:9007-9021.**

**176. Pelengaris S, Khan M, Evan G: c-MYC: more than just a matter of life and death. *Nat Rev Cancer* 2002, 2:764-776.**

**177. Cole MD, McMahon SB: The Myc oncoprotein: a critical evaluation of transactivation and target gene regulation. *Oncogene* 1999, 18:2916-2924.**

**178. Oster SK, Ho CS, Soucie EL, Penn LZ: The myc oncogene: MarvelouslY Complex. *Adv Cancer Res* 2002, 84:81-154.**

**179. Elkon R, Rashi-Elkeles S, Lerenthal Y, Linhart C, Tenne T, Amariglio N, Rechavi G, Shamir R, Shiloh Y: Dissection of a DNA-damage-induced transcriptional network using a combination of microarrays, RNA interference and computational promoter analysis. *Genome Biol* 2005, 6:R43.**

**180. Bridge AJ, Pebernard S, Ducraux A, Nicoulaz AL, Iggo R: Induction of an interferon response by RNAi vectors in mammalian cells. *Nat Genet* 2003, 34:263-264.**

**181. Persengiev SP, Zhu X, Green MR: Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). *Rna* 2004, 10:12-18.**
**182. Sledz CA, Holko M, de Veer MJ, Silverman RH, Williams BR: Activation of the interferon system by short-interfering RNAs. *Nat Cell Biol* 2003, 5:834-839.**
**183. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS: Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* 2003, 21:635-637.**
**184. Povirk LF: DNA damage and mutagenesis by radiomimetic DNA-cleaving agents: bleomycin, neocarzinostatin and other enediynes. *Mutat Res* 1996, 355:71-89.**
**185. van Dam H, Castellazzi M: Distinct roles of Jun: Fos and Jun: ATF dimers in oncogenesis. *Oncogene* 2001, 20:2453-2464.**
**186. Fan F, Jin S, Amundson SA, Tong T, Fan W, Zhao H, Zhu X, Mazzacurati L, Li X, Petrik KL, et al: ATF3 induction following DNA damage is regulated by distinct signaling pathways and over-expression of ATF3 protein suppresses cells growth. *Oncogene* 2002, 21:7488-7496.**
**187. Zhang C, Gao C, Kawauchi J, Hashimoto Y, Tsuchida N, Kitajima S: Transcriptional activation of the human stress-inducible transcriptional repressor ATF3 gene promoter by p53. *Biochem Biophys Res Commun* 2002, 297:1302-1310.**
**188. Hoh J, Jin S, Parrado T, Edington J, Levine AJ, Ott J: The p53MH algorithm and its application in detecting p53-responsive genes. *Proc Natl Acad Sci U S A* 2002, 99:8467-8472.**
**189. Stankovic T, Stewart GS, Byrd P, Fegan C, Moss PA, Taylor AM: ATM mutations in sporadic lymphoid tumours. *Leuk Lymphoma* 2002, 43:1563-1571.**
**190. Pettitt AR, Sherrington PD, Stewart G, Cawley JC, Taylor AM, Stankovic T: p53 dysfunction in B-cell chronic lymphocytic leukemia: inactivation of ATM as an alternative to TP53 mutation. *Blood* 2001, 98:814-822.**
**191. Stankovic T, Stewart GS, Fegan C, Biggs P, Last J, Byrd PJ, Keenan RD, Moss PA, Taylor AM: Ataxia telangiectasia mutated-deficient B-cell chronic lymphocytic leukemia occurs in pregerminal center cells and results in defective damage response and unrepaired chromosome damage. *Blood* 2002, 99:300-309.**
**192. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, et al: Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res* 2003, 31:38-42.**
**193. el-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B: Definition of a consensus binding site for p53. *Nat Genet* 1992, 1:45-49.**
**194. Robles AI, Bemmels NA, Foraker AB, Harris CC: APAF-1 is a transcriptional target of p53 in DNA damage-induced apoptosis. *Cancer Res* 2001, 61:6660-6664.**
**195. Miyashita T, Reed JC: Tumor suppressor p53 is a direct transcriptional activator of the human bax gene. *Cell* 1995, 80:293-299.**
**196. Wang CY, Mayo MW, Korneluk RG, Goeddel DV, Baldwin AS, Jr.: NF-kappaB antiapoptosis: induction of TRAF1 and TRAF2 and c-IAP1 and c-IAP2 to suppress caspase-8 activation. *Science* 1998, 281:1680-1683.**

**197. He KL, Ting AT: A20 inhibits tumor necrosis factor (TNF) alpha-induced apoptosis by disrupting recruitment of TRADD and RIP to the TNF receptor 1 complex in Jurkat T cells. *Mol Cell Biol* 2002, 22:6034-6045.**
**198. Deveraux QL, Reed JC: IAP family proteins--suppressors of apoptosis. *Genes Dev* 1999, 13:239-252.**
**199. Stankovic T, Hubank M, Cronin D, Stewart GS, Fletcher D, Bignell CR, Alvi AJ, Austen B, Weston VJ, Fegan C, et al: Microarray analysis reveals that TP53- and ATM-mutant B-CLLs share a defect in activating proapoptotic responses after DNA damage but are distinguished by major differences in activating prosurvival responses. *Blood* 2004, 103:291-300.**
**200. Ophir G, Amariglio N, Jacob-Hirsch J, Elkon R, Rechavi G, Michaelson DM: Apolipoprotein E4 enhances brain inflammation by modulation of the NF-kappaB signaling cascade. *Neurobiol Dis* 2005, 20:709-718.**
**201. De Leon DD, Farzad C, Crutchlow MF, Brestelli J, Tobias J, Kaestner KH, Stoffers DA: Identification of transcriptional targets during pancreatic growth after partial pancreatectomy and exendin-4 treatment. *Physiol Genomics* 2006, 24:133-143.**
**202. Nevins JR: The Rb/E2F pathway and cancer. *Hum Mol Genet* 2001, 10:699-703.**
**203. Manni I, Mazzaro G, Gurtner A, Mantovani R, Haugwitz U, Krause K, Engeland K, Sacchi A, Soddu S, Piaggio G: NF-Y mediates the transcriptional inhibition of the cyclin B1, cyclin B2, and cdc25C promoters upon induced G2 arrest. *J Biol Chem* 2001, 276:5570-5576.**
**204. Jung MS, Yun J, Chae HD, Kim JM, Kim SC, Choi TS, Shin DY: p53 and its homologues, p63 and p73, induce a replicative senescence through inactivation of NF-Y transcription factor. *Oncogene* 2001, 20:5818-5825.**
**205. Yun J, Chae HD, Choy HE, Chung J, Yoo HS, Han MH, Shin DY: p53 negatively regulates cdc2 transcription via the CCAAT-binding NF-Y transcription factor. *J Biol Chem* 1999, 274:29677-29682.**
**206. Imbriano C, Gurtner A, Cocchiarella F, Di Agostino S, Basile V, Gostissa M, Dobbelstein M, Del Sal G, Piaggio G, Mantovani R: Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters. *Mol Cell Biol* 2005, 25:3737-3751.**
**207. Fojas de Borja P, Collins NK, Du P, Azizkhan-Clifford J, Mudryj M: Cyclin A-CDK phosphorylates Sp1 and enhances Sp1-mediated transcription. *Embo J* 2001, 20:5737-5747.**
**208. Eto I: Molecular cloning and sequence analysis of the promoter region of mouse cyclin D1 gene: implication in phorbol ester-induced tumour promotion. *Cell Prolif* 2000, 33:167-187.**
**209. Martino A, Holmes JHt, Lord JD, Moon JJ, Nelson BH: Stat5 and Sp1 regulate transcription of the cyclin D2 gene in response to IL-2. *J Immunol* 2001, 166:1723-1729.**
**210. Cram EJ, Liu BD, Bjeldanes LF, Firestone GL: Indole-3-carbinol inhibits CDK6 expression in human MCF-7 breast cancer cells by disrupting Sp1 transcription factor interactions with a composite element in the CDK6 gene promoter. *J Biol Chem* 2001, 276:22332-22340.**
**211. Paskind M, Johnston C, Epstein PM, Timm J, Wickramasinghe D, Belanger E, Rodman L, Magada D, Voss J: Structure and promoter activity of the mouse CDC25A gene. *Mamm Genome* 2000, 11:1063-1069.**

**212. Rotheneder H, Geymayer S, Haidweger E: Transcription factors of the Sp1 family: interaction with E2F and regulation of the murine thymidine kinase promoter. *J Mol Biol* 1999, 293:1005-1015.**
**213. Chang YC, Illenye S, Heintz NH: Cooperation of E2F-p130 and Sp1-pRb complexes in repression of the Chinese hamster dhfr gene. *Mol Cell Biol* 2001, 21:1121-1131.**
**214. Huang D, Jokela M, Tuusa J, Skog S, Poikonen K, Syvaoja JE: E2F mediates induction of the Sp1-controlled promoter of the human DNA polymerase epsilon B-subunit gene POLE2. *Nucleic Acids Res* 2001, 29:2810-2821.**
**215. Nishikawa N, Izumi M, Yokoi M, Miyazawa H, Hanaoka F: E2F regulates growth-dependent transcription of genes encoding both catalytic and regulatory subunits of mouse primase. *Genes Cells* 2001, 6:57-70.**
**216. Parisi T, Pollice A, Di Cristofano A, Calabro V, La Mantia G: Transcriptional regulation of the human tumor suppressor p14(ARF) by E2F1, E2F2, E2F3, and Sp1-like factors. *Biochem Biophys Res Commun* 2002, 291:1138-1145.**
**217. Matuoka K, Yu Chen K: Nuclear factor Y (NF-Y) and cellular senescence. *Exp Cell Res* 1999, 253:365-371.**
**218. van Ginkel PR, Hsiao KM, Schjerven H, Farnham PJ: E2F-mediated growth regulation requires transcription factor cooperation. *J Biol Chem* 1997, 272:18367-18374.**
**219. Saeki K, Yuo A, Takaku F: Cell-cycle-regulated phosphorylation of cAMP response element-binding protein: identification of novel phosphorylation sites. *Biochem J* 1999, 338 (Pt 1):49-54.**
**220. Klemm DJ, Watson PA, Frid MG, Dempsey EC, Schaack J, Colton LA, Nesterova A, Stenmark KR, Reusch JE: cAMP response element-binding protein content is a molecular determinant of smooth muscle cell proliferation and migration. *J Biol Chem* 2001, 276:46132-46141.**
**221. Crowe DL, Shemirani B: The transcription factor ATF-2 inhibits extracellular signal regulated kinase expression and proliferation of human cancer cells. *Anticancer Res* 2000, 20:2945-2949.**
**222. Djaborkhel R, Tvrdik D, Eckschlager T, Raska I, Muller J: Cyclin A down-regulation in TGFbeta1-arrested follicular lymphoma cells. *Exp Cell Res* 2000, 261:250-259.**
**223. Recio JA, Merlino G: Hepatocyte growth factor/scatter factor activates proliferation in melanoma cells through p38 MAPK, ATF-2 and cyclin D1. *Oncogene* 2002, 21:1000-1008.**
**224. Johansson E, Hjortsberg K, Thelander L: Two YY-1-binding proximal elements regulate the promoter strength of the TATA-less mouse ribonucleotide reductase R1 gene. *J Biol Chem* 1998, 273:29816-29821.**
**225. Wu F, Lee AS: YY1 as a regulator of replication-dependent hamster histone H3.2 promoter and an interactive partner of AP-2. *J Biol Chem* 2001, 276:28-34.**
**226. Petkova V, Romanowski MJ, Sulijoadikusumo I, Rohne D, Kang P, Shenk T, Usheva A: Interaction between YY1 and the retinoblastoma protein. Regulation of cell cycle progression in differentiated cells. *J Biol Chem* 2001, 276:7932-7936.**
**227. Puga A, Barnes SJ, Dalton TP, Chang C, Knudsen ES, Maier MA: Aromatic hydrocarbon receptor interaction with the retinoblastoma**

protein potentiates repression of E2F-dependent transcription and cell cycle arrest. *J Biol Chem* 2000, 275:2943-2950.
228. Weiss C, Kolluri SK, Kiefer F, Gottlicher M: Complementation of Ah receptor deficiency in hepatoma cells: negative feedback regulation and cell cycle control by the Ah receptor. *Exp Cell Res* 1996, 226:154-163.
229. Elferink CJ, Ge NL, Levine A: Maximal aryl hydrocarbon receptor activity depends on an interaction with the retinoblastoma protein. *Mol Pharmacol* 2001, 59:664-673.
230. Evans MJ, Scarpulla RC: NRF-1: a trans-activator of nuclear-encoded respiratory genes in animal cells. *Genes Dev* 1990, 4:1023-1034.
231. Cam H, Balciunaite E, Blais A, Spektor A, Scarpulla RC, Young R, Kluger Y, Dynlacht BD: A common set of gene regulatory networks links metabolism and growth inhibition. *Mol Cell* 2004, 16:399-411.
232. Leone G, Sears R, Huang E, Rempel R, Nuckolls F, Park CH, Giangrande P, Wu L, Saavedra HI, Field SJ, et al: Myc requires distinct E2F activities to induce S phase and apoptosis. *Mol Cell* 2001, 8:105-113.
233. Gartel AL, Shchors K: Mechanisms of c-myc-mediated transcriptional repression of growth arrest genes. *Exp Cell Res* 2003, 283:17-21.
234. Santoni-Rugiu E, Duro D, Farkas T, Mathiasen IS, Jaattela M, Bartek J, Lukas J: E2F activity is essential for survival of Myc-overexpressing human cancer cells. *Oncogene* 2002, 21:6498-6509.
235. Baudino TA, Maclean KH, Brennan J, Parganas E, Yang C, Aslanian A, Lees JA, Sherr CJ, Roussel MF, Cleveland JL: Myc-mediated proliferation and lymphomagenesis, but not apoptosis, are compromised by E2f1 loss. *Mol Cell* 2003, 11:905-914.
236. Stevens C, La Thangue NB: E2F and cell cycle control: a double-edged sword. *Arch Biochem Biophys* 2003, 412:157-169.
237. Izumi H, Molander C, Penn LZ, Ishisaki A, Kohno K, Funa K: Mechanism for the transcriptional repression by c-Myc on PDGF beta-receptor. *J Cell Sci* 2001, 114:1533-1544.
238. Mao DY, Barsyte-Lovejoy D, Ho CS, Watson JD, Stojanova A, Penn LZ: Promoter-binding and repression of PDGFRB by c-Myc are separable activities. *Nucleic Acids Res* 2004, 32:3462-3468.
239. Fahmy RG, Dass CR, Sun LQ, Chesterman CN, Khachigian LM: Transcription factor Egr-1 supports FGF-dependent angiogenesis during neovascularization and tumor growth. *Nat Med* 2003, 9:1026-1032.
240. Thiel G, Cibelli G: Regulation of life and death by the zinc finger transcription factor Egr-1. *J Cell Physiol* 2002, 193:287-292.
241. Baudino TA, McKay C, Pendeville-Samain H, Nilsson JA, Maclean KH, White EL, Davis AC, Ihle JN, Cleveland JL: c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes Dev* 2002, 16:2530-2543.
242. Hayakawa J, Depatie C, Ohmichi M, Mercola D: The activation of c-Jun NH2-terminal kinase (JNK) by DNA-damaging agents serves to promote drug resistance via activating transcription factor 2 (ATF2)-dependent enhanced DNA repair. *J Biol Chem* 2003, 278:20582-20592.
243. Kool J, Hamdi M, Cornelissen-Steijger P, van der Eb AJ, Terleth C, van Dam H: Induction of ATF3 by ionizing radiation is mediated via a signaling pathway that includes ATM, Nibrin1, stress-induced MAPkinases and ATF-2. *Oncogene* 2003, 22:4235-4242.

**244. Piret B, Schoonbroodt S, Piette J: The ATM protein is required for sustained activation of NF-kappaB following DNA damage. *Oncogene* 1999, 18:2261-2271.**
**245. Li N, Banin S, Ouyang H, Li GC, Courtois G, Shiloh Y, Karin M, Rotman G: ATM is required for IkappaB kinase (IKKk) activation in response to DNA double strand breaks. *J Biol Chem* 2001, 276:8898-8903.**
**246. Banin S, Moyal L, Shieh S, Taya Y, Anderson CW, Chessa L, Smorodinsky NI, Prives C, Reiss Y, Shiloh Y, Ziv Y: Enhanced phosphorylation of p53 by ATM in response to DNA damage. *Science* 1998, 281:1674-1677.**
**247. Saito S, Goodarzi AA, Higashimoto Y, Noda Y, Lees-Miller SP, Appella E, Anderson CW: ATM mediates phosphorylation at multiple p53 sites, including Ser(46), in response to ionizing radiation. *J Biol Chem* 2002, 277:12491-12494.**
**248. Marine JC, Jochemsen AG: Mdmx as an essential regulator of p53 activity. *Biochem Biophys Res Commun* 2005, 331:750-760.**
**249. Hur GM, Lewis J, Yang Q, Lin Y, Nakano H, Nedospasov S, Liu ZG: The death domain kinase RIP has an essential role in DNA damage-induced NF-kappa B activation. *Genes Dev* 2003, 17:873-882.**
**250. Huang TT, Wuerzberger-Davis SM, Wu ZH, Miyamoto S: Sequential modification of NEMO/IKKgamma by SUMO-1 and ubiquitin mediates NF-kappaB activation by genotoxic stress. *Cell* 2003, 115:565-576.**
**251. Weston VJ, Austen B, Wei W, Marston E, Alvi A, Lawson S, Darbyshire PJ, Griffiths M, Hill F, Mann JR, et al: Apoptotic resistance to ionizing radiation in pediatric B-precursor acute lymphoblastic leukemia frequently involves increased NF-kappaB survival pathway signaling. *Blood* 2004, 104:1465-1473.**
**252. Ding GR, Honda N, Nakahara T, Tian F, Yoshida M, Hirose H, Miyakoshi J: Radiosensitization by inhibition of IkappaB-alpha phosphorylation in human glioma cells. *Radiat Res* 2003, 160:232-237.**
**253. Ryan KM, Ernst MK, Rice NR, Vousden KH: Role of NF-kappaB in p53-mediated programmed cell death. *Nature* 2000, 404:892-897.**
**254. Tergaonkar V, Pando M, Vafa O, Wahl G, Verma I: p53 stabilization is decreased upon NFkappaB activation: a role for NFkappaB in acquisition of resistance to chemotherapy. *Cancer Cell* 2002, 1:493-503.**
**255. Wu H, Lozano G: NF-kappa B activation of p53. A potential mechanism for suppressing cell growth in response to stress. *J Biol Chem* 1994, 269:20067-20074.**
**256. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004, 306:636-640.**
**257. Loots GG, Ovcharenko I: rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res* 2004, 32:W217-221.**
**258. Matlin AJ, Clark F, Smith CW: Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* 2005, 6:386-398.**
**259. Fehlbaum P, Guihal C, Bracco L, Cochet O: A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res* 2005, 33:e47.**
**260. Le K, Mitsouras K, Roy M, Wang Q, Xu Q, Nelson SF, Lee C: Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res* 2004, 32:e180.**

**261.** **Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, et al: Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 2004, 116:499-509.**

**262.** **Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR: Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002, 296:916-919.**

**263.** **Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, et al: Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 2004, 14:331-342.**

**264.** **Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, et al: The transcriptional activity of human Chromosome 22. *Genes Dev* 2003, 17:529-540.**

**265.** **Johnson JM, Edwards S, Shoemaker D, Schadt EE: Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* 2005, 21:93-102.**

**266.** **Bartel DP: MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004, 116:281-297.**

**267.** **Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 2005, 433:769-773.**

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

**269.** **Tomita M: Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol* 2001, 19:205-210.**

**270.** **Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, Hutchison CA, 3rd: E-CELL: software environment for whole-cell simulation. *Bioinformatics* 1999, 15:72-84.**

**271.** **Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR, 3rd: Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 2003, 100:3107-3112.**

**272.** **Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R: Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics* 2002, 1:323-333.**

**273.** **Begley TJ, Rosenbach AS, Ideker T, Samson LD: Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping. *Mol Cancer Res* 2002, 1:103-112.**

**274.** **Begley TJ, Rosenbach AS, Ideker T, Samson LD: Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping. *Mol Cell* 2004, 16:117-125.**