Università degli Studi di Trieste

Graduate School in MOLECULAR BIOMEDICINE

Regulatory modules discovery and mesenchymal stem cells characterization from high-throughput cancer genomics data

Yari Ciani

ciclo XXVII – Anno Accademico 2013/2014

Università degli Studi di Trieste

XXVII CICLO DEL DOTTORATO DI RICERCA IN BIOMEDICINA MOLECOLARE

Regulatory modules discovery and mesenchymal stem cells characterization from high-throughput cancer genomics data

Settore scientifico-disciplinare: BIO/13 BIOLOGIA APPLICATA

DOTTORANDO
Yari Ciani

COORDINATORE
Prof. Guidalberto Manfioletti

SUPERVISORI DI TESI
Dott. Silvano Piazza
Prof. Claudio Schneider

ANNO ACCADEMICO 2013/2014

Table of Contents

ABSTRACT

Cancer is a disease characterized by an extreme molecular complexity. Omics approaches, collecting data in public databases for all the genome, transcripts and proteins, attempt to overcome this complexity and find the functional modules that perform the functions involved in tumour related processes. For instance, cancer tissues gene expression profiles are widely used to define genes signatures and test their clinical relevance. I used this kind information in order to characterise interesting genes in breast cancer models. On the other hand, cellular models datasets could provide data that permits to focus on specific molecular mechanisms and probe the effects of molecules in a specific cancer model. One of the most recent omics project is the FANTOM5 project, that has generated a unique resource, the first single molecule sequencing-based expression atlas in mammalian systems. Cap analysis of gene expression (CAGE) was used to measure transcription start sites (TSS) and promoter usage across a wide collection of human samples thereby identifying and measuring levels of the majority of coding and non-coding transcripts in the human genome.

I used this information to characterize a mesenchymal/stromal stem cell line (MSC) derived from high-grade serous ovarian cancer (HG-SOC-MSCs) or derived from normal tissue (N-MSCs) included in the entire FANTOM5 human dataset. I highlighted shared functional programs between HG-SOC-MSCs and N-MSCs suggesting that the global differences between the two cell lines are based on quantitative levels of transcriptional output rather than on qualitative differences. The results suggested that HG-SOC-MSCs are close relatives of mesothelial cells and smooth muscle cells.

Furthermore, we analysed the entire dataset using ScanAll, a newly developed software, to *ab initio* predict the presence of enriched elements in the genomic regions surrounding FANTOM5 promoters. I pinpointed regulatory modules, i.e. groups of enriched motifs co-occurring in co-expressed regions within a fixed distance. These modules are enriched in the co-expressed sequences in each sample respect to random generated sequences. Finally, I created a Compendium of putative expressed and directly interacting transcription factors.

PUBLICATIONS RELEVANT FOR THE THESIS

Forrest AR et al., A promoter-level mammalian expression atlas, Nature. 2014 Mar 27;507(7493):462-70. doi:10.1038/nature13182. PubMed PMID: 24670764.

Verardo R, Piazza S, Klaric E, Ciani Y, Bussadori G, Marzinotto S, Mariuzzi L, Cesselli D, Beltrami AP, Mano M, Itoh M, Kawaji H, Lassmann T, Carninci P, Hayashizaki Y, Forrest AR; Fantom Consortium, Beltrami CA, Schneider C. Specific mesothelial signature marks the heterogeneity of mesenchymal stem cells from high-grade serous ovarian cancer. Stem Cells. 2014 Nov;32(11):2998-3011. doi: 10.1002/stem.1791. PubMed PMID: 25069783.

Emiliano Dalla, .. Yari Ciani et al., *Ab Initio* Prediction of Tissue-Specific Regulatory Modules in the FANTOM5 Project (in preparation)

Pegoraro S, Ros G, Piazza S, Sommaggio R, Ciani Y, Rosato A, Sgarra R, Del Sal G, Manfioletti G. HMGA1 promotes metastatic processes in basal-like breast cancer regulating EMT and stemness. Oncotarget. 2013 Aug;4(8):1293-308. PubMed PMID: 23945276; PubMed Central PMCID: PMC3787158.

Scolz M, Widlund PO, Piazza S, Bublik DR, Reber S, Peche LY, Ciani Y, Hubner N, Isokane M, Monte M, Ellenberg J, Hyman AA, Schneider C, Bird AW. GTSE1 is a microtubule plus-end tracking protein that regulates EB1-dependent cell migration. PLoS One. 2012;7(12):e51259. doi: 10.1371/journal.pone.0051259. Epub 2012 Dec 7. PubMed PMID: 23236459; PubMed Central PMCID: PMC3517537.

AIM OF THE PROJECT

Cancer is a disease characterized by an extreme molecular complexity. Several genes have already been identified as oncogenes or onco-suppressors, but not for all of them has been already identified the specific molecular mechanisms involved in the disease progression. With the use of high-throughput analysis we can now obtain information about the entire genome and transcriptome of a specific tumour or biological model, thus permitting to define the relationships between the different molecular elements that drive oncogenic processes. Omics approaches, collecting data in databases for all the genome, transcripts and proteins, attempt to overcome the complexity of the disease and find the functional modules that perform the functions involved in tumour related processes. One of the most recent omics project is the FANTOM5 project, that has generated a unique resource, the first single molecule sequencing-based atlas of promoters activity in mammalian systems. The FANTOM5 dataset is not strictly a cancer dataset, in fact it does not contain only cancer samples but also healthy tissues and primary cell lines: this feature allowed me to directly compare cell lines coming from the cancer microenvironment against both cancer cell lines and healthy tissues.

During my project I performed different tasks, all of them converging in the final aim of highlight functional and regulatory mechanisms that are relevant for the insurgence and progression of cancer.

1. Characterise mesenchymal stem cells derived from ovarian cancer respect the entire compendium of tissues and cell lines included in the FANTOM5 dataset.
2. Analyse the human promoterome dataset using newly developed software to find cis-regulatory modules that control the expression of the genes.
3. Define the relevance of specific genes (HMGA1 and GTSE1) and their derived gene signatures in cancer processes using gene expression high-throughput data.

All the analyses have been performed starting from publically available cancer datasets or newly generated high-throughput data thus permitting not only to obtain

results relevant for cancer prognosis, but also to make new interesting general observations.

INTRODUCTION

Introduction to Human Cancers

Cancer is a disease in which normal cells progressively acquires a series of characteristics that enable them to become tumorigenic. These acquired features support tumour growth and metastasis through six biological capabilities. In 2000, Hanahan and Weinberg (Hanahan and Weinberg, 2000) described these six biological capabilities and named them as "hallmarks of cancer".

They include six essential steps that allow cells to proliferate, survive and disseminate: sustained proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis and activating invasion and metastasis.

Cancer cells are able to elude the homeostatic control of cell cycle and start to proliferate: this is mostly due to an increased bioavailability of growth factor signalling molecules or circumventing negative feedback that in normal condition control cellular growth. So, cancer cells, in addition to increased proliferation, are unresponsive to suppression of growth factors, and moreover, unlike normal cells, are also able to escape programmed cell death. Apoptosis is usually triggered in response of endogenous stresses (such as DNA damage) in the intrinsic program, or by extracellular death-inducing signal (extrinsic program). Cancer cells adopt a variety of strategies to limit the apoptosis triggering; the most common way is the loss of damage sensor TP53, but also by increasing expression of anti-apoptotic regulators, and by down-regulating pro-apoptotic factors. In addition, normal cells entry in a non-proliferative state called senescence after a limited number of cellular divisions and this process is controlled by the shortening of telomeres. Cancer cells have an increased expression of telomerase, an enzyme that adds DNA sequence repeats to the 3' end of DNA strands in the telomere regions, thus acquiring replicative immortality.

Like normal cells, tumours cell require nutrients and oxygen, they supply these needs by inducing the formation of new vessels. In the advanced stages of the disease, tumour cells start to invade first the surrounding tissue and then distant site forming metastasis. The invasion and metastasis is a multistep process, often termed the invasion-metastasis cascade (Valastyan and Weinberg, 2011). The process starts with

local invasion trough extracellular matrix and stromal cell layer, followed by penetration of cancer cells into blood and lymphatic vessels. The metastatic cells are able to extra-vasate from the lumina of such vessels into the parenchyma of distant tissues, surviving in a foreign microenvironment in order to form metastases (Valastyan and Weinberg, 2011).

To be able to support cell growth and proliferation, cancer cells reprogram their cellular energy metabolism, and cells become more dependent to glucose metabolism, and react by upregulating glucose transport.

A relevant factor for tumour growth and metastasis is the crosstalk between cancer cells and the surrounding microenvironment. Also the the immune system plays a role in this crosstalk by supplying bioactive molecules to the tumour microenvironment, such us growth factors, pro-angiogenic factors and extracellular matrix-modifying enzymes that facilitate angiogenesis, invasion, and metastasis (Grivennikov et al., 2010).

Epithelial Ovarian Cancer

Epithelial ovarian cancer (EOC) is the gynecological disease with the highest death incidence. Despite relatively low morbidity, EOC presents a high mortality and the overall 5-year survival is still less than 30%. The death rate for this disease has not much changed in the last 50 years. This is mainly caused by the fact that the majority of early stage cancers are asymptomatic and over two-thirds of patients are diagnosed with advanced disease (FIGO stage III and IV, see below) (Holschneider and Berek). In advanced in stage the tumour spreads to the upper abdomen (stage III) or beyond (stage IV). The 5-year survival in these cases is only 15 to 20 percent, whereas the 5-year survival rate for stage I disease patients approaches 90 percent and for stage II disease patients approaches 70 percent.

EOC is a morphologically and biologically heterogeneous disease. Despite heterogeneous morphologies all EOC subtypes originate from the single layer of epithelial cells covering the surface of the ovaries (OSE cells), which shares a common embryonic origin with epithelia of Mullerian duct-derived tissues, but is different from the granulosa-thecal cells of the ovary. The evidence that ovarian epithelial cancer arises in the OSE was based on histopathological examination of clinical lesions (Feeley and Wells, 2001). OSE cells go through repeated cycles of proliferation with the growth and rupture of ovarian follicles and this process is

regulated by the interaction between mesenchyme and epithelium (Murdoch and McDonnel, 2002). Deregulation of such micro-environmental regulatory interactions could represent the basis of ovarian cancer progression, whereby molecular pathways, defined by growth factors and hormones, have been suggested to be critically involved. This is supported by the fact that the normal OSE cells retain a high degree of plasticity, being able to acquire mesenchymal or epithelial phenotypes (Auersperg et al., 2001).

From the histopathological point of view EOC are among the most complex of all human malignancies. One of the most particular aspects of ovarian carcinogenesis is the change in differentiation that accompanies neoplastic progression. OSE is a simple, rather primitive epithelium with some stromal features, but as it progresses to malignancy it loses its stromal characteristics and acquires the characteristics of the Mullerian duct-derived epithelia, *i.e.*, the oviduct, endometrium, and uterine cervix. This aberrant differentiation occurs in such a high proportion of ovarian carcinomas that it serves as the basis for the classification of these cancers as serous (fallopian tube-like), endometrioid (endometrium-like), and mucinous (endocervical-like) adenocarcinomas. Tumours in which is impossible to establish any resemblance with Mullerian derived tissues are indicated as undifferentiated tumours. Serous adenocarcinomas comprise approximately 80% of all epithelial ovarian cancers. At the cellular level, Mullerian differentiation is expressed by the appearance of altered cell shapes, E-cadherin, junctional complexes, epithelial membrane antigens, and secretory products including mucins and CA125 antigen. Thus, unlike carcinomas in most other organs in which epithelial cells become less differentiated in the course of neoplastic progression than the epithelium from which they arise, the differentiation of ovarian carcinomas is more complex than that of OSE: EOC, in fact, forms polarised epithelia, papillae, cysts, and glandular structures.

In addition to the histological classification, EOC are also given a grade and a stage after surgery. The grade is based on both architectural and cytological features and is on a scale from 1 to 3. Grade 1 EOC more closely resembles normal tissue and tends to have a better prognosis than Grade 3 EOC. It is noteworthy that there are several other rare histotypes that have been described: Brenner tumours, Mixed Epithelial Tumours, and the Small Cell Carcinomas.

Genome-wide study approaches represent a suitable tool for identification of molecular factors involved in the EOC and in this perspective recent studies have

reported the characterization of gene expression profiles of this type of malignancy and of derivative cell lines (Welsh et al., 2001). These molecular profiling studies have supported the notion that different subtypes likely represent distinct disease entities (TCGA, 2011).

Breast cancer

Breast Cancer is the most common cancer, and the worldwide leading cause of cancer death among women. Despite this, death rates from breast cancer have been declining since about 1989. These decreases are the combined result of earlier detection through screening and improved treatment. Most patients die due to distant metastases that are frequently unresponsive to therapies.

Breast cancer is an heterogeneous disease from a clinical, morphological and, first of all, molecular point of view. The development of high throughput technologies has allowed the classification of breast cancer based on molecular characteristic and not only looking phenotypic parameters (such us tumour size, lymph node involvement, grade, and age) (Perou et al., 2000; Sørlie et al., 2001). Gene expression profiling of breast cancer sample allowed identification of five major molecular subtypes: luminal A, luminal B, HER2 positive, basal-like and claudin-low.

Luminal A is the most common subtype among breast cancers (50-60% of all cases). It is defined by immunohistochemistry as estrogen receptor (ER) positive, and HER2 negative, characterized by low proliferation rate and usually have a good prognosis. Luminal B (10-20% of all BC) tumours have a high proliferation rate, estimated by Ki67 proliferation-related antigen staining. They have a worse prognosis than Luminal A but they respond to chemotherapy. Her2-positive subtype represents 10-15% of breast cancer and highly expresses HER2 gene or genes associated with the HER2 pathway. Commonly they are diagnosed at a high histological grade, and are characterized by poor prognosis.

10-20% of BC corresponds to basal-like subtype: they have poor prognosis, with high mitotic index, high metastatic relapse in visceral organs (lung, central nervous system, lymph nodes). Triple negative breast cancers (TNBC) are the most common sub-group of basal breast: they are negative for ER, PGR and HER2 expression. Treatment for this subtype is very difficult, as well as identification of new

therapeutics targets. The most frequent alterations are p53 mutations and impaired activity of BRCA1 (Bertucci et al., 2012).

Another subtype is the Claudin-low subtype, whose predominant feature is low expression of tight junctions and intracellular adhesion genes (claudin-3, - 4, -7 and E-cadherin) and overexpression of immuno-response related genes, due to high immune-system infiltration (Prat et al., 2010). This sub-group has a poor prognosis.

Normal Breast cancer is a poorly characterized subtype of breast cancer: it occurs only in 5-10% of cases end it express genes typical of adipose tissue. It lacks ER, PGR and HER2 receptor, but it is also missing of CK5 and EGFR basal marker. Usually this kind of BC displays an intermediate prognosis and response to chemotherapy. The classification into subtypes of breast cancer based on molecular features could permit to stratify breast cancer patients and treatments. Nevertheless, processes and pathways in cancer progression and metastasis still need to be investigating in order to better define treatment and prognosis for the various breast cancer subtypes.

Adult Stem cells and cancer

Stem cells, as classically defined, are cells with a capacity to self-renew and to generate daughter cells that can differentiate down several cell lineages to form all of the cell types that are found in the mature tissue. A stem cell might go through an asymmetric cell division to generate one cell that is identical to the stem cell itself and one cell that is distinct and more differentiated. The identical cell provides for self-renewal of the stem cell compartment; the distinct cell goes through a series of cell divisions and differentiation steps to generate the ultimate terminally differentiated cell populations. The cells that form the intermediates between stem cells and terminally differentiated cells are usually referred to as progenitor cells (especially if they give rise to a defined structure or cellular compartment), transit cells or transit amplifying cells. Stem cells could also generate distinct daughter cells by dividing symmetrically into two identical cells, followed by a random decision based on, for example, variation in intensity of cell signalling to establish one daughter cell as a new stem cell and the other as a transit cell.

Stem cells can be divided into two functional classes. First, there are stem cells that are responsible for tissue renewal. Such cells are found, for example, in bone marrow, in the skin and in the intestine, and are responsible for replacing terminally

differentiated cells as they mature and die or are shed from an epithelial surface. These cells are continually active, although at a slow rate. Second, there are stem cells that are inactive until required in response to environmental factors, for example, to repair tissue damage. Satellite cells of muscle might be an example of such a stem cell, as might putative liver stem cells that have been suggested to be responsible for liver regeneration.

Types of Human Adult stem cells

The best characterized types of human adult stem cells in vertebrates are:

HSCs Hematopoietic Stem Cells

NSCs Neural Stem Cells/oligodendrocyte progenitors

IPSCs Induced Pluripotency Stem Cells

MSCs Mesenchymal Stem Cells

CSCs Cancer Stem Cells

HSC Hematopoietic Stem Cells: the stem cells that form blood and immune cells are known as hematopoietic stem cells. They are ultimately responsible for the constant renewal of blood. The first evidence and definition of blood-forming stem cells came from studies of people exposed to lethal doses of radiation in 1945. In the early 1960s, Till and McCulloch began analysing the bone marrow to find out which components were responsible for regenerating blood (Till and McCulloch, 2012) Nowadays hematopoietic stem cell is defined as a cell isolated from the blood or bone marrow that can renew itself, can differentiate to a variety of specialized cells, can mobilize out of the bone marrow into circulating blood, and can undergo programmed cell death (apoptosis).

***NSCs Neural Stem Cells*:** adult NSCs, which are generated from the precursors that build the nervous system during development are maintained into adulthood in at least two niches, the sub-ventricular zone (SVZ) of the lateral ventricles and the sub-granular zone (SGZ) in the hippocampus, although there is lively discussion concerning the possibility that NSCs are more widely scattered throughout the adult brain (Gould, 2007).

IPSC Induced pluripotent stem cells: in 2006, Takahashi and Yamanaka at Kyoto University in Japan identified conditions that would allow specialized adult cells to be genetically "reprogrammed" to assume a stem cell-like state. These adult cells, called induced pluripotent stem cells (iPSCs), were reprogrammed to an embryonic stem cell-like state by introducing genes important for maintaining the essential properties of embryonic stem cells (ESCs) (Takahashi and Yamanaka, 2006). This approach involves taking mature "somatic" cells from an adult and introducing the genes that encode critical transcription factor proteins, which themselves regulate the function of other genes important for early steps in embryonic development. In the initial 2006 study, it was reported that only four transcription factors (Oct4, Sox2, Klf4, and c-Myc) were required to reprogram mouse fibroblasts (cells found in the skin and other connective tissue) to an embryonic stem cell–like state by forcing them to express genes important for maintaining the defining properties of ESCs. These factors were chosen because they were known to be involved in the maintenance of pluripotency, which is the capability to generate all other cell types of the body. The newly-created iPSCs were found to be highly similar to ESCs and could be established after several weeks in culture (Maherali et al., 2007).

MSCs Mesenchymal Stem Cells: more than 30 years ago it has been reported that fibroblast-like cells coming from bone marrow via attachment to tissue culture plastic were inherently osteogenic (Friedenstein et al., 1974). Osteogenic cells were actually capable of differentiating into multiple connective tissue cell types at a clonal level (Dennis et al., 1999), which validated the concept of a mesenchymal stem cell. Nowdays, MSCs are defined as adherent, fibroblastoid-like cells that are able to differentiate to osteoblasts, adipocytes, and chondrocytes in vitro .

In addition to bone marrow, MSCs or MSC-like cells have also been elaborated from skeletal muscle, adipose tissue, umbilical cord, synovium, the circulatory system, dental pulp, and amniotic fluid (Williams et al., 1999)(Zuk et al., 2001)(Erices et al., 2000)(De Bari et al., 2001)(Kuznetsov et al., 2001).

Therefore, it appears that MSCs reside within the connective tissue of most organs. However, it should be noted that these populations are not functionally equivalent with respect to their differentiation potential. Also, clonal studies have shown that plastic adherent populations isolated from bone marrow are functionally heterogeneous and contain undifferentiated stem/progenitors and lineage-restricted

precursors with varying capacities to differentiate into connective tissue cell types (Muraglia et al., 2000). Therefore, characterizing populations as MSC or MSC-like also depends, in part, on the methods used to evaluate their differentiation potential. Finally, because MSCs also generate the stromal component of bone marrow, adherent populations contain cells that express adhesion molecules and cytokines that regulate aspects of hematopoiesis (Deryugina and Müller-Sieburg, 1993).

Despite their functional heterogeneity, MSC populations obtained from most tissues commonly express a number of surface receptors including CD29, CD44, CD49a-f, CD51, CD73, CD105, CD106, CD166, and Stro1 and lack expression of definitive hematopoietic lineage markers including CD11b, CD14, and CD45. MSCs that express the aforementioned surface markers and are capable of differentiating into connective tissue cell types can be enriched from peripheral and umbilical cord blood by selection for CD133 and from bone marrow by selection for stage-specific embryonic antigen (SSEA)-1, SSEA-4, or the nerve growth factor receptor CD271. Other studies have shown that bone marrow-derived MSCs express the pericyte-specific markers CD146 and 3G5 (Shi and Gronthos, 2003) consistent with the fact that specialized vascular pericytes in bone marrow are thought to represent the closest in vivo approximation to MSCs. However, it is important to realize that no single isolation method is regarded as a standard in the field. Therefore, the varied approaches used to culture-expand and select for MSCs make it difficult to directly compare experimental results. Moreover, some isolation schemes introduce epigenetic and genetic changes in cells that may dramatically affect their plasticity and therapeutic utility. Finally, human MSCs exhibit some variation in their pattern of expressed genes among different donor preparations using the same isolation protocols, and larger variations as sparse cultures become confluent and are expanded by serial passage and approach senescence. Many hypotheses on the in situ original cells from which MSCs descend have been raised, but a final answer is to be given.

Cancer stem cells: mutations that deregulate the pathways that control normal stem-cell self-renewal cause a diverse range of cancers (Reya et al., 2001). This indicates that cancer can be considered a disease of unregulated self-renewal in which mutations convert normal stem-cell self- renewal pathways into engines for neoplastic proliferation. Indeed, stem cells are appealing candidates as the 'cell of origin' for

cancer because of their pre-existing capacity for self-renewal and unlimited replication.

When cancer cells of many different types were assayed for their proliferative potential in various *in vitro* or *in vivo* assays, only a small minority of cells were able to proliferate extensively. This gave rise to the idea that malignant tumours are comprised of both cancer stem cells, which have great proliferative potential, as well as more differentiated cancer cells, with limited proliferative potential. So, the growth and progression of many cancers could be driven by a minority population of cancer stem cells.

The discovery of multipotent progenitor cells with the capacity for self-renewal outside the hematopoietic system raises the possibility that cancer stem cells could arise from other tissue stem cells and initiate other cancer types, including solid cancers. Consistent with this possibility, uncultured specimens of human breast cancer cells from patients were separated into fractions that expressed different surface molecules, and then injected into immunodeficient mice. Again, only a small population of the tumour cells was able to induce tumour formation in the mice (Al-Hajj et al., 2003). These findings indicate that, like teratocarcinoma cells and AML cells, breast cancer cells intrinsically differ in their tumorigenic potential.

The tumour niche

Analysis and reconstruction of the hierarchal cellular organization and homeostasis within tumour tissue and along its development represent the most challenging present goals in cancer research. The dissection and characterization of the distinct cellular lineages and their respective progenitors giving rise to the various cell types that form the tumour-tissue could be addressed through the use of defined in vitro growth models.

At present we know that tumour initiation, invasion and progression are determined by molecular and phenotypic alterations arising not only in the tumour cells, but also in their microenvironment. Microenvironment supports malignant growth and the molecular mechanisms and stromal cells may contribute to tumorigenesis. In fact, the heterogeneous nature of the microenvironment confers a large degree of diversity in gene expression and signalling throughout the tumour, resulting in the creation of subpopulations of tumour cells that exhibit phenotypic diversities and differential sensitivity to various forms of treatment. Stromal cells within the tumour, also defined

as CAFs, may derive from various sources including their recruitment from an external Mesenchymal-Stem-Cell compartment. It has been shown that Bone Marrow derived Mesenchymal-Stem-cells respond to chemotactic signals originating from the tumour cells (Karnoub et al., 2007) and acquire the characteristics of CAFs after exposure to tumour-conditioned medium (Mishra et al., 2008). Within the tumour recruited MSC cells could be capable of self-sustaining and expanding as internal colonies thus maintaining their own homeostasis in response to the selection pressure imparted by the tumour cells whose crosstalk should define the basic mechanisms of the reciprocal expansion. MSCs have been shown to increase the kinetics of implantation and facilitate metastasis in a breast cancer xenograft model. (Karnoub et al., 2007). MSC derived from normal tissues have multipotent potential characterized by plastic adhesion properties, differentiation potential and marker expression. Tumour derived stromal cells differ from normal stromal cells in their committed maintenance of an activated state which has been shown, for breast cancer, to most likely depend on epigenetic modifications. This has been recently reinforced also for ovarian serous carcinoma, where evidence has been provided that the tumour stromal compartment does not present clear signs of somatic alterations (Qiu et al., 2008).

DNA microarrays: an invaluable tool for functional genomic exploration

The introduction of automated large-scale sequencing, supported by adequate computational tools and bioinformatics development, has greatly increased our general knowledge on genomic sequences organisation and function. Nowadays, it is possible to analyse thousands of genes in a single assay. Serial analysis of gene expression (SAGE) (Velculescu et al., 1995), oligonucleotide arrays (Lockhart et al., 1996) and cDNA microarrays (Schena et al., 1996) are among the major techniques developed for this type of analysis. In microarray analysis the Northern blotting scheme is reversed: the labelled moiety is obtained from the RNA sample and a certain number of immobilised known sequences are used as probes. The advancements in robotics allowed miniaturising the scale of the reactions and modified microscope slides could be used to deposit thousands of nucleic acid sequences. The same result was also obtained by borrowing photolithography techniques from the semiconductor manufacturing to synthesize oligonucleotides directly onto a solid support. Altogether these progresses leaded, in 1995, to the first

papers in which the term “microarray” was used in its current meaning (Schena et al., 1998).

Principles of microarray analysis

Microarrays are miniaturised hybridisation assays that permit to simultaneously querying thousands of nucleic acid fragments. All microarray systems share the following components:

- The array, which contains the immobilised nucleic acid sequences, known as “probes”;
- One or more labelled samples or “targets” that are hybridised against the microarray;
- A detection system that quantify the hybridisation signals.

Microarrays and DNA-chips

Spotted microarrays consist of a collection of preformed nucleic acid sequences immobilised onto the solid support so that each unique sequence forms a tiny feature called “spot”. These nucleic acids are obtained in numerous ways, and there are different methods for depositing them onto microarray slides (by simple contact, by inkjet technology, or by micro-syringe pumping for instance). In general nucleic acid prepared for deposition on microarrays consist of cDNA clones amplified by polymerase chain reaction (cDNA microarrays), or of synthesised oligonucleotides of various length (oligonucleotide microarrays). The size of the spots differs from one system to another, but it is usually less than two hundred micrometers in diameter. A modified glass slide or glass wafer acts as the solid support onto which up to tens of thousands of spots can be arrayed in a total area of a few square centimetres. On the contrary, DNA-chips are produced by a proprietary technology (GeneChip®, Affymetrix) quite different from the spotted one, as it is based on direct photolithography synthesis of short oligonucleotides (20-25 base pairs) on the solid support.

Target labelling and microarray hybridization

Whatever the kind of microarray used, DNA probes present on the arrays are interrogated by nucleic acid hybridisation with a labelled target. The sample may be mRNA for a gene expression study or genomic DNA for other purposes (promoter usage analysis: CHIP-on-Chip, genomic rearrangements: FISH-on Chip). The sample is converted to a labelled population of nucleic acids, known as the target. These

moieties consist of several thousands of different labelled nucleic acid. Therefore, these hybridisations should be carried out under conditions that do not promote annealing of non-complementary fragments. Fluorescent dyes, and especially the cyanine dyes Cy3 and Cy5, have been widely adopted as the predominant labels in microarray analysis. Fluorescence has the advantage of permitting the detection of two or more different signals in one experiment. This has thus allowed investigators to perform comparative analysis of two or more samples on one microarray.

Then the labelled fragments in the target are expected to form duplexes with their immobilised complementary probes. This phase, called hybridization, requires that the nucleic acids are single-stranded and accessible to each other. The number of duplexes formed reflects the relative number of each specific fragment in the target, as long as the amount of immobilised nucleic acid probe is in excess and not restraining the kinetics of hybridisation. Two or more samples labelled with different fluorescent dyes can be hybridised simultaneously, resulting in simultaneous hybridisation taking place at each spot. By measuring the different fluorescent signals associated with each feature, the relative abundance of specific sequences in each of the samples can be determined.

Scanning and data analysis

Microarray scanners typically contain two different lasers that emit light at wavelengths that are suitable for exciting the fluorescent dyes used as labels. A detector system attached to a confocal microscope records the emitted light from each feature of the array, permitting high-resolution detection of the hybridisation signals. Alternative solutions use CCD-camera devices to detect the fluorescence. Despite their small size, microarrays allow the generation of a large amount of data even from a single hybridisation. For these reasons the use of computerised data processing is necessary in order to handle the amount of generated data and to gain maximum information from the experiment. This is usually achieved by specialised software that extracts primary data from scanned microarray slide images, normalises this data to remove the influence of experimental variation, and finally manipulates the data so that biologically meaningful conclusions can be made.

Applications of microarray analysis

Microarrays represent a high-power approach to perform analyses that were previously time consuming. Due to the availability of millions of data points at once,

microarrays enabled global analysis of fundamental biological processes: gene expression analysis, genome analysis, and drug discovery have been three of the main areas in which microarray analysis has been applied so far.

Gene expression analysis

Gene expression analysis examines the composition of cellular messenger RNA populations. Traditional gene expression analysis was based on techniques such as Northern blotting, RT-PCR and nuclease protection assays, as well as on more advanced methods, such as differential display, subtractive hybridisation, cDNA fragment fingerprinting, and serial analysis of gene expression (SAGE). These techniques were largely used in the past and have enabled the discovery of novel differentially expressed genes. However, the technical challenges of these methods still limit their use to the analysis of just a few samples at a time, while microarray analysis allows the examination of thousands of genes in multiple samples with relative simplicity.

In the simplest scheme a typical microarray gene expression experiment compares the relative expression levels of specific transcripts in two samples. Usually one of the samples is a control while the other is obtained from cells whose response or status is being explored. Each one of the two samples is labelled with a different fluorescent dye, and equal amounts of the labelled samples are combined and hybridised with the microarray. After hybridisation two grey scale images (usually in a 16-bit TIFF format) corresponding to the fluorescent signals of the two dyes are independently obtained by scanning the microarray and fluorescence intensity from each feature is subsequently quantified by a specific software. After normalization, the intensity of the two hybridisation signals can be compared: equal signal from both samples suggests equal expression of the considered genes in both samples, while signals disparity is suggestive of differential expression.

One of the most important remarks that have to be taken into account is that microarray analysis does not give any information about absolute gene expression levels in the samples. This is because the intensity of the fluorescent signals is not only proportional to the number of hybridised fragments, but also to the length of these fragments and the number of fluorescent labels each fragment carries (specific activity of the target or labelling density). These parameters are determined by the unique nucleotide sequence of each transcript, so that they will vary from gene to

gene. If the two samples have been labelled under similar conditions, the length and labelling density of specific transcripts will be similar, allowing the comparison of the relative abundance of the transcripts in the analysed targets.

In cancer research microarrays have been intensively used to find gene expression changes in transformed cells and metastases, to identify diagnostic markers, and to classify tumours based on their gene expression profiles (Alizadeh et al., 2000).

Cancer datasets

Cancer can take hundreds of different forms depending on the tissue of origin and the spectrum of genomic alterations. These differences affect the oncogenesis process and therapeutic response. Although many genomic events with direct phenotypic impact have been identified, much of the complex molecular landscape remains not characterized.

Large numbers of oncogenes had been identified using functional assays on genetic material from tumours in positive-selection systems (Alitalo et al., 1983; Shimizu et al., 1983; Soda et al., 2007), and a subset of tumour suppressor genes was identified by analysing loss of heterozygosis.

The advent of microarray allowed collecting gene expression data coming from cancer samples, creating large cohorts of patients of which clinical information is also available. This approach permitted to draw relationships between gene expression patterns and clinical and prognostic features of cancer.

More recently, systematic cancer genomics projects, have applied emerging technologies to the analysis of specific tumour types. This approach has identified novel oncogenic genes and (Davies et al., 2002; Mardis et al., 2009; Tomlins et al., 2005), has established definitions of molecular subtypes (TCGA, 2011, 2012) and has identified new biomarkers. Importantly, some of these biomarkers have important clinical implications (Perou et al., 2000; TCGA, 2013).

The increased number of tumour samples improves the ability to detect and analyse molecular aberrations in cancers. The use of large cohorts has enabled DNA sequencing to uncover a list of recurrent genomic aberrations, both known and novel, as common events across tumour types (Vogelstein et al., 2013). Indeed, a majority of the cancer samples have distinct alterations not shared with other samples. Despite the

apparent uniqueness of each individual tumour, the set of molecular aberrations often integrates into known biological pathways that are shared by sets of tumour samples.
The results of molecular analysis are now revealing that cancers of different organs may share many features, whereas, cancers from the same organ but of different sub-type are often quite distinct.
Important similarities among tumour sub-types from different organs have already been identified. For example, TP53 mutations drive high-grade serous ovarian, serous endometrial and basal-like breast carcinomas, all of which share a global transcriptional signature involving the activation of similar oncogenic pathways (Ciriello et al., 2013; Kandoth et al., 2013; TCGA, 2012).
Shared molecular patterns will enable therapeutic discoveries in one disease to be applied to other types of cancer.

Next-generation sequencing

With the advent of capillary electrophoresis (CE)-based Sanger sequencing, scientists gained the ability to elucidate genetic information from any given biological system.
The concept behind next-generation sequencing (NGS) technology is similar to Sanger sequencing: the bases of a small fragment of DNA are identified from signals emitted as each fragment is re-synthesized from a DNA template strand. NGS extends this process across millions of parallel reactions. This method enables fast sequencing of large sequences of DNA spanning entire genomes, with the newest instruments capable of producing hundreds of gigabases of data in a single run.
Several NGS platforms are commercially available (Metzker, 2010). Most of them are based on sequencing-by-synthesis technology. Roche 454, Illumina, Helicos, and PacBio (Pacific Biosciences) use a DNA polymerase to drive their sequencing reaction, while SOLiD (Life Technologies) and Complete Genomics use a DNA ligase. The sequencing platforms can be single molecule-based (sequencing a single molecule, such as Helicos and PacBio) or ensemble-based (sequencing of multiple identical copies of a DNA molecule, such as Illumina and SOLiD).
Sample preparation and amplification phase change between different platforms, hence the selection of a sequencing platform depends on the experimental goals. For example, the sample preparation protocol for Helicos is relatively simple and might be preferred if the amount of RNA sample is limiting. Helicos avoids the polymerase

chain reaction (PCR) amplification step, giving a direct reflection of RNA expression levels. These characteristics are typical for all single molecule sequencing (SMS) platforms. Generally, non-SMS platforms use amplification steps. Amplification-based protocols can provide relative expression levels for most RNAs but require more controls.

On the other hand single-molecule-based platforms such as Helicos have a high error rate (5%). A higher error rate makes it more difficult to match sequencing reads with a reference genome and lowers the number of usable reads. If a low sequencing error rate is needed, Illumina or SOLiD are often the best choices (< 1). The advantage of low error rates is particularly important for microRNA (miRNA) sequencing. Because of the relatively small sizes of miRNAs (ranging from 15 to 27 nt, with most 20 to 22 nt long on average), high error rates cause many raw reads to be lost at the alignment stage. Whatever platform is chosen, higher sequencing capacity renders lowly expressed transcripts to be detected more easily.

RNA-seq

RNA sequencing (RNA-seq) is the application of any of a variety of next-generation sequencing techniques to study RNA. RNA-seq library preparation usually includes reverse transcription. Data analysis of RNA-seq could give information and quantitation about alternatively spliced transcript and novel transcript.

Transcriptome assembly is necessary to transform individual reads into sequences of entire mRNAs or noncoding transcripts.

Library preparation for RNA-seq consist in converting cellular RNA into molecules that can be easily sequenced. Ribosomal RNAs (rRNA) are abundant in the cell and can comprise up to 80% of total cellular RNA. rRNA may be removed by an enzymatic degradation approach (such as duplex specific nuclease treatment). This helps ensure that rare transcripts are sequenced to adequate depth.

Depending on the chosen sequencing platform, RNA can be fragmented, usually by chemical hydrolysis or enzymatic digestion to a appropriate size. In some cases the RNA species under investigation, such as miRNAs, are small (under 200 bases) and so the fragmentation phase is not required. Other RNAs are long and must be fragmented to smaller sizes, such as *200–250 nt long, to be suitable for sequencing.

Once RNAs of the appropriate size are obtained, they are converted into complementary DNA (cDNA) by a reverse transcriptase using random primers. Adapter oligonucleotides are then ligated to the cDNA to allow amplification and enable sequencing.

For some types of Illumina or SOLiD library preparations the identification of the direction of the original RNA strand is possible thanks to the use of specific adapters ligated to the 3' and 5' ends of RNA. The Helicos Direct RNA Sequencing (DRS) technique replaces the adaptor ligation with a poly-A tailing step that modifies the RNA fragments directly. Because Helicos DRS directly sequences the RNA it directly provides information about strand specificity.

Data Analysis

Dealing with the amount of data generated by RNA-seq experiments time-consuming and challenging.

Data is most often supplied in FASTQ format. This format contains an ID number for each read, the read sequence, and a quality score. There are 2 main stages for sequencing data analysis. First, one must remove sequencing artefacts and errors from the data set that usually are the presence of the ligation adaptors and low-complexity reads. There are publicly available tools that can be used to address these issues (Falgueras et al., 2010; Lassmann et al., 2009). Sequencing errors can be removed or corrected based on the quality score in order to improve the assembly quality. When a large genome is the duty subject, for example the human genome, extremely short reads (< 17 nt) may be filtered out prior to alignment.

At the second stage, the processed data are aligned to a reference genome. The data alignment and analysis approach choice depends on the sequencing platform and particular RNA-seq application. There are some publically available programs that can be freely downloaded and efficiently run by individual laboratories to carry out total RNA-seq data analysis. TopHat, a fast splice junction mapper for RNA-seq reads, is one of the most commonly used programs (Trapnell et al., 2009). Programs such as Cufflinks and Scripture (Guttman et al., 2010) may be used to reconstruct the full transcripts, resolve individual variants, and even quantitate expression levels for each transcript and gene. Further downstream analysis may include differential expression analysis.

CAGE-seq

Cap analysis gene expression (CAGE) is a technique introduced in 2003 as a method to determine transcription start sites on a genome-wide scale. The principle on which the method is based is the isolation and sequencing of short sequence tags originating from the 5′ end of RNA transcripts (Shiraki et al., 2003). Then, mapping the tags to a reference genome identifies all the transcription starting sites from which the transcripts originated.

CAGE is able to capture full length RNAs and at the same time avoid rRNA and tRNA transcripts. First, poly-A terminated RNAs are reverse-transcibed using an oligo-dT primer. Alternatively, a random primer can be used for RNAs without a poly-A tail, which may constitute a consistent part of the transcriptome (Cheng et al., 2005).

The biotinylation and subsequently capture by streptavidin-coated magnetic beads of the 5′ cap structure of RNA/ DNA double-stranded hybrids permits the selection of hybrids that contain mature mRNA. A restriction site about 20 nucleotides downstream the 5′ end of the full-length cDNA is then created by ligation of a linker sequence containing an *Mme*I recognition site. This produces a short CAGE tag starting at the 5′ end of eukaryotic mRNAs (Kodzius et al., 2006).

Due to their size, sequencing CAGE tags is more efficient at detecting transcription start sites than sequencing full-length cDNAs and CAGE tags could also identify new transcription start sites or genes.

In early CAGE experiments many CAGE tags were found only once in a given sample. This was caused by the limited sequencing depth of the sequencers. In the few last years, the next-generation sequencers allowed to generate millions of tags from a low amount of sample thus allowing an accurate estimate of the cellular concentration of the RNA molecule corresponding to each CAGE tag. This method is called Deep-CAGE. The uniqueness of

Deep CAGE

High-throughput gene expression experiments based on microarrays (Schena et al., 1995) and serial analysis of gene expression (SAGE) (Velculescu et al., 1995) give us a snapshot of the RNA concentrations in the cell in a specific condition. Quantitative

real-time PCR (qRT-PCR) (Heid et al., 1996), can provide a valuable standard for validation because of its accuracy and wide dynamic range. The characteristic features of CAGE expression profiling make it particularly suitable for investigating the transcriptional regulatory network that controls the expression of genes.

CAGE tag counts allow us to calculate the cellular amount of the corresponding RNA molecule directly in digital form, without the need of an analog/digital signal conversion phase. Moreover, expression profiling based on tag sequencing is unbiased so one mRNA is not preferentially detected over another. This allows a direct comparison of the expression values of different genes measured in a single deep CAGE experiment.

In contrast, microarray fluorescence levels are affected both by the probe-dependent mRNA affinity and by mRNA: this precludes a direct comparison between genes. In addition, tag counts have a dynamic range that is orders of magnitude larger than microarray expression levels. The accuracy and the dynamic range of CAGE- and SAGE-derived expression levels as well as the sensitivity of detecting lowly expressed transcripts have been improved further by deeper sequencing. Importantly, methodologies based on tag sequencing can also measure the expression of currently unknown transcripts while microarray and qRT-PCR expression profiling need a primer pair in order to study a specific transcript.

CAGE allows us to determine in a single experiment the promoter that regulates the transcription of a specific transcript combined with its expression level.

.Even though the potentiality of this method, the throughput of these CAGE was insufficient to allow genome-wide expression profiling and may not be able to detect lowly expressed transcripts until the advent of next-generation sequencers.

Advances in sequencing chemistry have led to innovations such as pyrosequencing, Solexa sequencing by synthesis (Illumina, Inc., San Diego, CA, USA), SOLiD sequencing by oligo-nucleotide ligation and detection with 2-base encoding (Applied Biosystems, Foster City, CA, USA), and true single molecule sequencing by Helicos (Helicos BioSciences Corp., Cambridge, MA, USA).

These sequencers are particularly suitable for short sequence reads, the kind generated by CAGE. This makes easier the sequencing of the transcriptome while, in contrast, genome sequencing is more difficult because longer sequence reads are preferred. The increased throughput of sequencers enables CAGE sequencing at a much deeper

scale. This implies that it is possible to detect RNA molecules present at very low-copy numbers in the cell.

Deep CAGE also increases the tag count for each particular mRNA: this enables more precise expression profiling also for all expressed transcripts.

The uniqueness of Deep CAGE resides in the capability of detecting both transcription start sites as well as their expression level.

Promoters identification by deep CAGE

A promoter is a region of the DNA that is located near the transcription starting site of a gene, (on the same strand and at the 5' of the gene). It can be 100-1000 nucleotides long and its function is to regulate gene transcription. A promoter contains different regulatory elements:

- the transcription starting site;
- a binding site for RNA polymerase;
- general transcription binding sites;
- specific transcription binding sites.

Due to the inherent variability in transcription start sites, CAGE tags are typically found scattered over short genomic regions. Promoters can be constructed by clustering the 5' ends of the individual CAGE tags based on their position on the genome (Carninci et al., 2005). In deep CAGE experiments, transcription start site clustering may also take into account the similarity in expression profiles. Deep CAGE thus allows us to define the individual promoters and their activity.

Transcription units in mammalian genomes are characterized by alternative transcription start sites and multiple splice forms. These forms could be active in different cellular contexts (Carninci et al., 2006; ENCODE and Consortium, 2004; Ravasi et al., 2006; Zavolan et al., 2003). CAGE tags can correspond to previously identified promoters, or to novel promoters and transcripts that may give rise to novel protein variants. CAGE has also led to the discovery of novel noncoding RNAs.

The analysis of CAGE- defined transcription start sites in human and mouse illustrated that promoters characterized by a TATA-box tend to have a clear, single transcription start site, whereas promoters associated with CpG islands tend to have transcription start sites distributed over a broad area (Carninci et al., 2006). CAGE expression profiling, with its ability to localize the promoter as well as determine its expression level, is ideally suited to study the regulation of bidirectional promoters

(Trinklein et al., 2004), which may be co-regulated by shared transcription factor binding sites. Antisense transcripts, which are transcribed in the opposite direction to the coding strand, may play a role in regulation via RNA interference as well as gene silencing at the chromatin level (Ambros, 2004; Andersen and Panning, 2003; Fukagawa et al., 2004; Imamura et al., 2004). A genome-wide analysis of CAGE transcriptome data revealed frequent concordant regulation of sense/antisense pairs (Katayama et al., 2005).

A major goal of the analysis of deep CAGE is to infer the regulatory network that orchestrates transcription in a cell (Nilsson et al., 2006). This kind of network is qualitatively different from current gene-based networks inferred from microarray or SAGE expression profiling because each of the promoters may be associated with one or more coding or noncoding transcripts.

Transcription factors

Transcription factors (TFs) are proteins that mediate transcriptional regulation. TFs bind to specific short DNA sequences called transcription factor binding sites (TFBSs), which are 5-20 bp in length. These sequences are mainly located in the promoters.

Each TF can regulate multiple genes. A TF can act as an *activator*, increasing the transcription level of the regulated gene, or as a *repressor*, decreasing its transcription level. A TF could act as activator in two main modes: 1) binding to the promoter, the TF interacts with the RNA polymerase. This interaction attracts the RNA polymerase close to the gene promoter, facilitating its binding to the core promoter. 2) When the TF binds the DNA, the structure of the chromatin in the promoter region changes conformation and the binding area of the RNA polymerase becomes accessible.

There are also two main modes of repression: 1) the repressor TF can compete with an activator TF on its BS. Therefore, it decreases the effects of the activator, leading to less efficient binding of RNA polymerase to the promoter. This obviously results in lower expression levels of the gene. 2) The repressor interacts with the same components of RNA polymerase as does an activator. By doing this, the repressor prevents the activator from interacting with the RNA polymerase.

TFs bind to short sequences in the promoter. The features of the binding region (length and base composition) are dictated by the protein structure of the TF. Though the BSs of a TF have a core pattern that is essential for the TF binding, the exact

nucleotide composition may vary. This is because some parts of the sequence are more important for the binding and thus less subject to viable mutations. The BS sequence pattern is called a *motif*. A common way of representing a motif is by using a *motif logo*. The motif logo illustrates the conserved and variable regions of the motif by displaying the information of every position in the motif.

Prediction of conserved regulatory elements

In the last two decades, sequence-based computational methods for describing the DNA-binding specificities of TFs have been developed and the use of microarray- and sequencing-based assays for the high-throughput measurement of protein–DNA binding resulted in a burst of motif discovery methods. Most sequence-based DNA motif discovery methods use position weight matrices (PWMs) to represent the TF–DNA binding specificity (Badis et al., 2009; Bussemaker et al., 2007; Stormo, 2000; Stormo and Zhao, 2010). This type of model can be learned from various data types: from small sets of known binding sites to high-throughput protein–DNA binding data. A different problem arises when looking for motifs (both known and novel, not necessary TFBS), enriched in a set of promoter sequences of co-regulated genes.

The approaches to locate known and novel regulatory elements can be summarized in two categories: comparative genomics (Arnold et al., 2012; Lenhard et al., 2003; Wang and Stormo, 2005; Xie et al., 2005) and single (or statistical) genomic approaches (Bajic et al., 2003; Davuluri et al., 2001; Eskin and Pevzner, 2002; Hughes et al., 2000; Kulakovskiy et al., 2010; Pavesi et al., 2007; Sandelin and Wasserman, 2004; Sinha and Tompa, 2003; Workman and Stormo, 2000). The first strategy, also called “phylogenetic footprinting”, is based on the information coming from conservation between orthologous genes regulatory sequence; single genomic approaches, on the other hand, look for enriched short signals (typically about 5-15nt long) in the midst of a great amount of statistical noise (a typical input being a regulatory region of length 500nt-1.5kb). One of the major points that emerged from distinct assessments (Das and Dai, 2007; Tompa et al., 2005) that exhaustively examined publicly available DNA motif discovery tools, was the difficulty to simultaneously achieve balanced sensitivity and positive predictive value, with the most serious inaccuracy occurring in non-CpG-related promoters (Bajic et al., 2004).

In order to have a more accurate representation of the complex interactions associated with transcriptional regulation, new approaches look for clusters of TFBS called *cis*-

regulatory modules (CRMs) (Bulyk, 2003; Davidson et al., 2002; King et al., 2005; Levine and Davidson, 2005; Yuh et al., 1998). CRMs can represent clusters of two or more closely associated TFBS for transcription factors that physically interact (Ravasi et al., 2010; Sinha et al., 2003). Alternatively, they can describe groups of closely associated binding sites for TFs that form macromolecular complexes but do not directly bind (Muzny et al., 2006) to each other (Blanchette et al., 2006; Ferretti et al., 2007), or that are not sufficiently close to directly interact (Kardassis et al., 2002; Zhou et al., 2010). Finally, they can represent neighbouring enriched genomic elements playing general functional/structural roles (Kolbe et al., 2004; Taylor et al., 2006) like epigenetic markers, chromatin remodelling sites (Yan et al., 2013) or RNA polymerase II stalling-associated sequences (Core and Lis, 2009; Hah et al., 2011).

MATERIAL AND METHODS

Human Samples

Human samples were obtained from the Azienda Ospedaliero-Universitaria of Udine and collected after informed consent in accordance with the declaration of Helsinki and with approval by the Bioethics Committee.

Cell Isolation, Culture and Transfection

N-MSCs were isolated from normal human adult tissues (bone-marrow, heart, and adipose tissues) and cultured under uniform conditions as described in Beltrami et al. (Beltrami et al., 2007). HG-SOC-MSCs were isolated from primary HG-SOC and cultured under uniform conditions as previously described (Bourkoula et al., 2014)(Beltrami et al., 2007) with minor modifications.

Surgical biopsies were freshly collected from patients undergoing surgery, washed several times in PBS solution and then mechanically disaggregated by mincing it with razor blades. Further dissociation was carried out by enzymatic digestion (20µg/ml collagenase IV) for 10 minutes at 37º. Single cell suspensions were obtained by filtering the disaggregated tissue through a nylon mesh with 70µm pores (Cell Strainer, BD Falcon). Recovered cells were cultured in MyeloCult Medium (StemCell Technologies) containing 25% of serum (12.5% horse serum and 12.5% of fetal bovine serum). When the colonies were formed cells were detached with 0.25% trypsin-EDTA (Sigma-Aldrich) and replated at a density of $2.5x10^3/cm^2$ onto fibronectin (Sigma-Aldrich) coated 100mm dishes (Sacco-Falcon), in an expansion medium composed as follows: 60% low glucose DMEM (Invitrogen), 40% MCDB-201, 1mg/mL linoleic acid-BSA, 10^{-9}M dexamethasone, 10^{-4}M ascorbic acid-2 phosphate, 1X insulin-transferrin-sodium selenite (all from Sigma-Aldrich), 2% fetal bovine serum (StemCell Technologies), 10ng/mL hPDGF-BB, 10ng/mL hEGF (both from Peprotech EC). Medium was replaced with fresh one every 3-4 days.

Peripheral blood mononuclear cells (PBMCs) were isolated from HG-SOC-bearing patients through density gradient centrifugation (Biocoll; Biochrom, Berlin, Germany).

MDA-MB-231 and MDA-MB-157 cells were grown in DMEM plus 10% tetracycline-free FBS, MDA- MB-468 cells in RPMI 1640 plus 10% tetracycline-free

FBS. For transfection of siRNAs, all cell lines were transfected with 100 nM siRNAs with Lipofectamine RNAiMAX reagent (Invitrogen). For plasmid, transfection was performed with FuGENE (Roche). For the functional rescue experiment, cells were first transfected with siRNAs and with plasmid 24 hours later.

High-Resolution Genotyping Analysis

Genomic DNA was extracted using QIAamp DNA Mini Kit (Qia- gen, Hilden, Germany) and analysed using HumanCNV370- Quadv3_C SNP platform (Illumina, San Diego, CA). Copy number information was derived normalizing each tumour or each HG- SOC-MSC sample to its matched normal counterpart using B allele frequency (BAF) segmentation and crlmm package (Carvalho et al., 2007).

Real-Time Quantitative RT-PCR

Total RNA was extracted using the TRIZOL Reagent (Life Technologies, Carlsbad, CA) and cDNA was produced using the QuantiTect Reverse Transcription Kit (Qiagen). Real-time quantitative RT-PCR was performed using the SYBR Green Master Mix (Applied Biosystems, Life Technologies, Carlsbad, CA) on a StepOnePlus Real-Time System (Applied Biosystems).

Gene Expression Analysis

Total RNA was extracted using Trizol reagent (Invitrogen) subjected to DNase-I (Invitrogen) treatment and subsequently column-purified with RNeasy kits (QIAGEN). For microarray analysis, four biological mRNA replicates for each group (siCTRL or siA1_3) were hybridized on Affymetrix GeneChip Human Genome U133A 2.0 array. For quantitative RT-PCR, mRNA was transcribed using Superscript II (Invitrogen). Quantitative PCR was performed using SYBR Green PCR master mix (Applied Biosystems) and 7500 Real-Time PCR System (Applied Biosystems).

Low-Level Analysis

For microarray analysis, three biological mRNA replicates for each group (siHMGA1 or siControl) were hybridized on Affymetrix hgu133plus2 chips. Cell intensity values were computed using the Affymetrix Expression Console. Further data processing was performed in the R Computing Environment version 2.14 (http://www.r-project.org/) with BioConductor packages (http://www.bioconductor.org/). Robust Multi- Array Average (RMA) normalization was applied (Irizarry et al., 2003). Statistical analysis for differentially expressed genes was performed with limma (Smyth, 2004). P-values were adjusted for multiple testing using the Benjamini

and Hochberg's method to control the false discovery rate (Hochberg and Benjamini, 1990). Genes with adjusted p-values below 10^{-4} and fold change greater than 2.6 (log 1.4) or lower than -2.6 (-log 1.4) were considered differentially expressed. Gene annotation was obtained from R-Bioconductor metadata packages, and the probesets were converted in Entrez Gene Id and Symbol Id.

Hierarchical Cluster Analysis

Starting from the normalized annotated expression matrix after gene median centering, features that had standard deviation of less than 0.3 were filtered out. Unsupervised hierarchical cluster analysis (average- linkage method) was performed using Cluster software (EisenLab). Cluster results were then visualized using Java TreeView.

For the clustering of functional annotations (average- linkage method), for each sample –log(p-values) of enrichment of functional ontologies have been used: geodesic distances have been calculated and results were visualised using proxy (Meyer and Buchta, 2015) a package for R.

Functional Analysis

Differentially expressed gene lists obtained from low-level procedures were analysed for functional associations.

- Data were analysed through DAVID Bioinformatics Resources v6.7 (Dennis et al., 2003) using the suggested standard parameters.
- Data were analysed through Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com) software. Core analysis was performed, and the top associated networks table was reported.
- Data were analysed through the Oncomine (Rhodes et al., 2007) web tool using suggested standard parameters. Custom concept analysis was performed, and the "Summary view" (adapted) was reported.
- Data were analysed through ClueGo (Bindea et al., 2009), a plug-in for Cytoscape (Shannon et al., 2003) with suggested standard parameters.

To perform Analysis of Functional Annotation (AFA) a combined dataset of the functional analysis results were created, where the significance of each biological functional/transcriptional regulators was transformed as the log(p-value).

Survival analysis

Several published gene expression datasets (breast cancer meta-dataset) were considered and compared with the gene signatures of interest. The raw data were retrieved from the gene expression omnibus (GEO) public gene expression database (GSE1456, GSE4922, GSE5327, GSE6532, GSE7390, GSE11121, GSE12093, GSE2603, GSE16446, GSE19615, GSE20685, GSE21653). Data were normalized in R/Bioconductor environment using the RMA normalization method (affy package), creating a breast cancer meta-dataset. Gene annotation was obtained from brainarray custom CDF metadata packages, and the probesets were converted to Entrez Gene Id and Symbol Id. Each dataset was analysed separately to avoid platform and signal merging problems, and only the results were combined together. To evaluate the correspondence between the signatures expression levels and breast cancer clinical data, we utilized the gene expression-based Outcome for Breast Cancer web tool (GOBO) (Ringnér et al., 2011). To verify the correlation of the gene signatures and breast cancer clinical data, a Mantel-Haenszel test was applied to the normalized meta-dataset (survival R package), and the Kaplan–Meier survival curve of time to distant metastasis (TDM) of breast cancer patients classified according to the expression of the signatures was obtained. With the same meta-dataset, we searched for the distribution of the gene expression intensities of the signatures of interest across different breast cancer subtypes (stats R package).

For the analysis of the mesothelial signature, we performed Kaplan-Meier survival analysis considering progression-free survival (PFS) time of ovarian cancer patients, relapse-free survival time of breast cancer patients, and first progression (FP) time of lung cancer patients. Samples were classified according to the expression of the HG-SOC-MSCs gene signature (HSM-GS) using KMplotter web tool (Gyorffy et al., 2012; Györffy et al., 2010a; Győrffy et al., 2013).

FANTOM5 Genomic sequences

Globally, 889 human CAGE libraries (495 primary cell samples, 259 cancer cell lines and 135 post-mortem tissues) were sequenced at a single-molecule level and analysed. RLE-normalized expression values were calculated using edgeR (Robinson et al. 2010), providing an evaluation of the number of tags per million (TPM) and presenting information on sample-related transcripts relative abundance. A total number of 184,827 robust human Sample-Specific CAGE clusters (SSCs) were

identified as having greater than or equal to 10.0 RLE-TPM and at least ten-fold higher than the median expression of the cluster across all available hCAGE libraries. The genomic sequences for these SSCs clusters were extracted after adding 300nt upstream and 100nt downstream of each cluster and sample-specific regulatory elements were discovered using the software ScanAll.

Sample Ontology

Samples were divided into functional groups using a modified version of the Sample Ontology developed by a member of the FANTOM5 Consortium (Dr. Tom Freeman, University of Edinburgh).

Differential expression analysis of FANTOM5 data

The robust transcriptional (RLE) activity peaks count FANTOM5 data matrices (TPM or counts) were used for the differential transcriptional analysis (a, b, count Matrix), relationship analysis (c, TPM matrix) and correlation analysis (d, TPM matrix). EdgeR(Robinson et al., 2010a) package for R/Bioconductor Environment(Gentleman, 2005) was used for statistical analysis.

a) **Direct comparison between HG-SOC-MSCs and N-MSCs.**

   We considered genes with a differential transcriptional activity with log fold change at least greater/lower than 2/-2 respectively ($p<10^{-5}$, FDR corrected).

b) **HG-SOC-MSCs and N-MSCs vs. other samples.**

   The purpose of this analysis was to obtain specific gene profiles of HG-SOC-MSCs and N-MSCs with respect to all other FANTOM5 project samples. However, it would be totally inappropriate to perform such comparison by pooling all the other samples together (for biological and statistical reasons) since performing such comparison will bias specific groups of samples. To this aim we created 100 randomly selected samples of similar sizes of that of the HG-SOC-MSCs (n=9) and N-MSCs (n=10) subsets. After performing all the contrasts, we considered only the genes showing a differential transcriptional activity ($p<10^{-4}$, FDR corrected) in at least 90% of comparisons between HG-SOC-MSCs or N-MSCs and random samples subsets. The log fold change is the average value of all comparison log fold changes.

c) **Relationship analysis.**

   We compared HG-SOC-MSCs against all other samples divided in biological classes (Class 4_Class1, n=220) with respect to previously described samples

ontology, thus obtaining the lists of differential genes. We then ranked the biological groups on the basis of the number of differential genes with different cutoff of p-values ($p<10^{-10,}$ $p<10^{-15,}$ $p<10^{-20)}$, FDR corrected). The biological groups most similar HG-SOC-MSCs (less than 20 genes) where reorganized into biological categories despite the tissue of origin.

d) correlation analysis.

MSC samples data was extract from the TPM matrix, log tranformed and center normalized. Pearson Correlation matrix among the different MSC groups was computed and graphical rappresented using gplots(Warnes, 2010) package for R/Bioconductor Environment(Gentleman, 2005).

Differential gene peak usage analysis

Starting from the TPM matrix (sample n=889, peaks n=184828), we considered each peak active in a sample if the corresponding TPM was greater than 15. The number of active peaks for each gene in each sample was then calculated. We then performed the following analysis:

a) Direct comparison between of HG-SOC-MSCs and N-MSCs.

Since in this analysis we wanted to select genes with differential global peak activity, we excluded from the analysis the genes with no active peaks in at least 50% of the samples in each group.

Then we compared the total number of peaks activated in each genes between the two the groups using Welch's t test. Finally we considered differential genes with at least pValue less than 0.05.

b) Relationship analysis.

We compared HG-SOC-MSCs against all other samples divided in biological classes (Class 4_Class1, n=220) with respect to previously described samples ontology, using the same approach used in direct comparison between of HG-SOC-MSCs and N-MSCs thus obtaining the lists of global peak activity analysis. We then ranked the biological groups on the basis of the number of differential peaks with $p<10^{-5}$. The biological groups most similar HG-SOC-MSCs (less then 200 genes) where reorganized into biological categories despite the tissue of origin.

ScanAll
ScanAll is a software composed of three main components: ScanPro, which looks for single over-represented motifs in biological sequences; ScanMod, which sub-selects only the motifs belonging to composite regulatory elements; the filtering pipeline, which sorts only the most functionally informative modules.

ScanPro
When starting the analysis, the following parameters were provided:

- a set of *n* sequences (the SSCs to be analysed) of length L (not fixed)
- ℓ, the length of the motif to be found
- *d*, the number of mismatches (or variable nucleotides) allowed in each motif
- *q* (the *quorum*), the minimal number of sequences in which motifs are required to be enriched

By examining PWMs databases, the most common length of motif layouts ranges from 5 to 12nt, with a number of mismatches between 1 and 3. We decided to fix ℓ to 6 and *d* to 1 as they represent the average length/structure of TFBS core sequences (the most conserved and biologically meaningful). Moreover, quorum *q* was fixed to 150 (corresponding to 10% of the average number of genomic sequences in each FANTOM5 sample).
For an extensive description of ScanPro algorithm and Karp-Rabin algorithm (on which ScanPro is based) see Supporting Information.

ScanMod
This phase consisted in finding composite regulatory structures, groups of two or more closely positioned enriched motifs. No parameter exists for fixing the maximum number of modules-composing motifs; hence, very long modules could be potentially obtained. As a side effect, this phase was also very useful in reducing the size of ScanPro output: by sub-selecting only module-composing motifs, in fact, we were also able to define a fraction of ScanPro results that was easier to manage during the downstream analysis phases.
The modules identified were on the form M_1 <d,D> M_2…<d,D> M_x, and the following parameters were provided:

- M_x, a motif belonging to a list (the ScanPro output)
- d, the minimum required distance between left-end and right-end of motifs in the module

- D, the maximum allowed distance between left-end and right-end of motifs in the module
- a set of *n* sequences (the same input received by ScanPro)
- *q* (the new *quorum*), the minimal number of sequences in which a module is required to be enriched

It was also possible to set a filtering value, the *Complexity* (*C*), defining the number of different nucleotides appearing into an instance of an input motif. For example, the "AAC*n*G" motif might have a "complexity" of either three or four, depending on the value of *n*. If the required complexity is set to four, the motif will equate to (AACTG).
To perform the analysis we fixed d to 40, D to 90, q to 60 and *C* to 4.

Filtering
Modules found by ScanMod were filtered based on a Z-score. The Z-score uses the normal approximation to the binomial distribution to compare the rate of occurrence of a TFBS in the target set of genes to the expected rate estimated from the pre-computed background set. For a given TFBS, the random variable *X* denotes the number of predicted binding site nucleotides in the dataset. Let *B* be the number of predicted binding site nucleotides in the background dataset.
Using a binomial model with *n* events, where *n* is the total number of nucleotides examined (i.e. the total number of nucleotides in the conserved non-coding regions) from the co-expressed genes, and *N* is the total number of nucleotides examined from the background genes, then the expected value of *X* is $u = B * C$, where $C = n / N$ (i.e. *C* is the ratio of sample sizes). Then taking $p = B / N$ as the probability of success, the standard deviation is given by $s = \mathrm{sqrt}(n * p * (1 - p))$.
Now, let *x* be the observed number of binding site nucleotides in the dataset. By applying the Central Limit Theorem and using the normal approximation to the binomial distribution with a continuity correction, the z-score is calculated as $z = (x - u - 0.5) / s$. Background sequences are generated by shuffling the di-nucleotides of the original sequences. Then, for perfectly overlapping modules, that means they are different layouts of the same conserved sequence, only the module with the best z-score is retained.

Motifs Annotation
We proceeded to the annotation of the module-composing motifs using TOMTOM (http://meme.nbcr.net/meme/cgi-bin/tomtom.cgi). We queried the TRANSFAC, SwissRegulon, JASPAR and UniPROBE human public databases, along with the ENCODE *cis*-regulatory lexicon and the HOMER and HOCOMOCO collections. We considered as positive matches all the hits to any of these databases within an E-value < 0.5, corresponding to less than 1 hit being expected at random, the remaining results constituting our list of putative novel motifs.

FANTOM 5 Cluster analysis
The TPM FANTOM5 data matrix (samples n=889, peaks n=184827) was used for cluster analysis. After gene median centering, features whose standard deviation was less than 0.55 were filtered out, thus obtaining the TPM filtered matrix (sample n=889, peaks n= 26266). Unsupervised hierarchical cluster analysis (average-linkage method) was performed using the Cluster software (Eisen et al., 1998) and cluster results were visualized using Java TreeView (Saldanha, 2004).

Expressed TFs
The list of expressed TFs for each biological group was obtained starting from the TPM matrix: a TF was considered expressed in a biological class if at least one peak associated with that TF had a TPM value >2.5

Protein-protein interactions
Interaction data generation from TFs lists was performed using iRefR (Mora and Donaldson, 2011), a R package based on the iRefIndex protein-protein interaction database (Razick et al., 2008). We considered only Human protein-protein interactions data.

Motifs Mapping to Transposable Elements (TE)
The ScanMod output was converted in standard BED-6 format. The BEDTools utility (Quinlan and Hall, 2010) was used to compare .bed files containing ScanMod results to a file with all the annotated human transposable elements (retrieved from the TranspoGene project, http://transpogene.tau.ac.il/index.html). For each module-composing motif identified in the 38 selected FANTOM5 samples, or in the four analysed ENCODE cell lines (HeLaS3, HepG2, GM12878 and K562), we counted the number of instances mapping to TEs and to the promoter regions not overlapping TEs. The resulting ratios were then compared to the sample-specific ratios obtained

by dividing the length of the promoter regions overlapping TE regions by the promoter regions not overlapping TE regions, and we considered as over/under-represented those motifs whose folds were greater than two or smaller than minus two, respectively.

Coding and software implementation

ScanPro and ScanMod were implemented using C++ language. The R/Bioconductor environment (R Development Core Team, 2011) was used to handle filtering, functional validation, text manipulation, statistical analysis, graphical representation and all the automatic annotation steps. The following packages were used: ade4 (Dray and Dufour, 2007), Biostrings (Pages et al., 2013), coin (Hothorn et al., 2006), edgeR, ggplot2 (Wickham, 2009), gplots (Warnes et al., 2013), pROC (Robin et al., 2011), ROCR (Sing et al., 2005), seqLogo (Bembom, 2013) and Vennerable (Swinton, 2011).

Data Access

The accession number in the Gene Expression Omnibus public database for the MDA-MB-231 expression array experiment is GSE35525.

The full set of sample-specific analysed FANTOM5 promoters regions can be accessed as indicated in the FANTOM5 main paper(Forrest et al., 2014).

RESULTS

Using high-throughput data to characterise cell lines isolated from healthy and cancer tissue

Cellular models experiments could provide data that permits to focus on specific molecular mechanisms and probe the effects of molecules in a specific cancer model. It is known that the observations obtained in cellular models can't always be reproduced in the tissue of interest. The reasons of these differences are many: on one hand in vivo and in vitro environments interact in different ways with the biological process that take place in the cells. In particular, in vitro experiments do not take into account the supporting role of tumour microenvironment. To complicate matters, cancer cell lines can't always be assigned univocally to a specific tissue of origin. There are different causes of this fact: 1) the intrinsic plasticity and tendency to accumulate mutations of the cancer cell lines; 2) the difficulty of comparing cancer cell lines genotypes in respect to healthy tissues genotypes; 3) the intrinsic heterogeneity of the cancer tissue; 4) the fact that, for a lot of cancer types, the tissue of origin of the cancer cells is not known yet; 5) also observing the same type of cancer, different molecular subtypes can potentially originate from different healthy tissues. In the next section I describe a publicly available high-throughput dataset that includes cancer cell lines, primary cell lines and human tissues, thus making possible a direct comparison between cancers derived cell lines and human tissues.

The FANTOM5 promoterome atlas

Although the genome information is the same in almost all cells of an individual, each distinct cell type has its own regulatory repertoire of active and inactive genes. Every single cell responds to external stimuli and alterations in environment with changes in gene expression. Cells interact with other cells in order to perform specific functions in the organism. The majority of genes have more than one TSS, and the regulatory inputs that determine TSS choice and activity are diverse and complex. Specific sets of transcription factors are induced or repressed in each cell line. These transcription factors provide regulatory inputs that are integrated at transcription start sites (TSSs) to control the process of transcription. Annotation of the regulation, expression and function of all the mammalian genes requires systematic analysis of the distinct mammalian cell types. It is also necessary to identify the set of TSS in the mammalian genome and the transcription factors that regulate their utilization.

The FANTOM5 project has the aim of mapping the sets of transcripts, transcription factors, promoters and enhancers active in the majority of mammalian primary cell types.

During the FANTOM5 project 889 human CAGE libraries were sequenced, providing a collection made of 495 human primary cell samples (~150 cell types from 3 donors), 259 cancer cell lines (representing 154 distinct cancer subtypes) and 135 human post-mortem tissues. For each sample, several millions of mapped tags (also referred to as “peaks”) were generated, clustered based on proximity and subsequently subdivided according to their different expression profiles using a method called Decomposition-based Peak Identification (DPI; (Forrest et al., 2014)). This allowed defining “robust” and “permissive” clusters, based on the number of independent single molecule observations (> 10 or > 2, respectively) occurring at least in one nucleotide position per library. These peaks account for the majority of known 5’ tags, possibly represent transcription start sites and, more generally, provide evidence of promoter utilization. In order to distinguish true transcripts from post-transcriptional artefacts, robust peaks found in the selected libraries were normalized using the RLE method (Anders and Huber, 2010) implemented in edgeR (Robinson et al., 2010b), obtaining an evaluation of the number of tags per million (TPM) and providing information on sample-related transcripts relative abundance.

The FANTOM5 dataset has been critical in different projects in which I am involved. One of the main features of the dataset is that it includes samples coming from human tissues, primary cell lines and transformed cell lines thus allowing the comparison of the cell lines of interest, not only respect to all the other cell lines, but also respect to the tissues of the human body. In the next sections of the results the dataset is utilized in order to characterize specific cell types that are included in the FANTOM5 dataset.

Mesenchymal stem cells characterization

Mesenchymal Stem/Stromal cells (MSCs) reside in almost all types of tissues and are believed to play a central role in wound repair, tissue regeneration, and maintenance of tissue homeostasis (da Silva Meirelles et al., 2006; Uccelli et al., 2008). The interactions between mesenchymal and epithelial parenchyma are essential for organogenesis and also play a critical role in cancer progression (Hanahan and Weinberg, 2011). MSCs are in fact present also in the tumour microenvironment. They could be recruited via the blood stream from the bone marrow thus exerting multiple and complex roles. In my laboratory it has been previously reported a protocol for the derivation of multipotent stromal cells from normal human adult bone marrow, heart, liver and adipose tissues (hereafter referred to as N-MSCs) that maintain the ability to differentiate along various lineages (Beltrami et al., 2007; Zeppieri et al., 2013). The protocol has been adapted in order to obtain mesenchymal stem/stromal cells (MSCs) from primary high-grade serous ovarian cancers (HG-SOC-MSCs) (see Supporting Material Table S1). In breast and ovarian cancers, the stromal tissues supporting the tumour mass are composed of cell populations with no evident genomic alterations (Qiu et al., 2008), thus originating from normal adult tissue stem cells. The genomic status of the HG-SOC-MSCs had been assessed by high-resolution genotyping analysis using the Illumina Infinium SNP BeadChips. The investigated data set was composed of matching HG-SOC primary tissues, HG-SOC-MSC cells and peripheral blood mononuclear cells (PBMCs), each triplet from three patients. As clearly shown in Figure 1 A and B, no differences were detected between PBMC and HG-SOC-MSC genomes, while genome-wide aberrations were clearly found between matched PBMC and HG-SOC genomes ($p<0.03$, paired t-test). In order to test for tumourigenicity, cultured HG-SOC-MSCs were injected into NOD-

SCID mice in parallel to SKOV3. $1x10^6$ HG-SOC-MSCs from three separate patients were injected subcutaneously in duplicate into mice and followed for 12 weeks. These cells were incapable of forming tumours. In contrast, $1x10^6$ SKOV3 cells (in five replicas) were capable of forming subcutaneous tumours in 2 weeks (see Supporting Material Table S2).

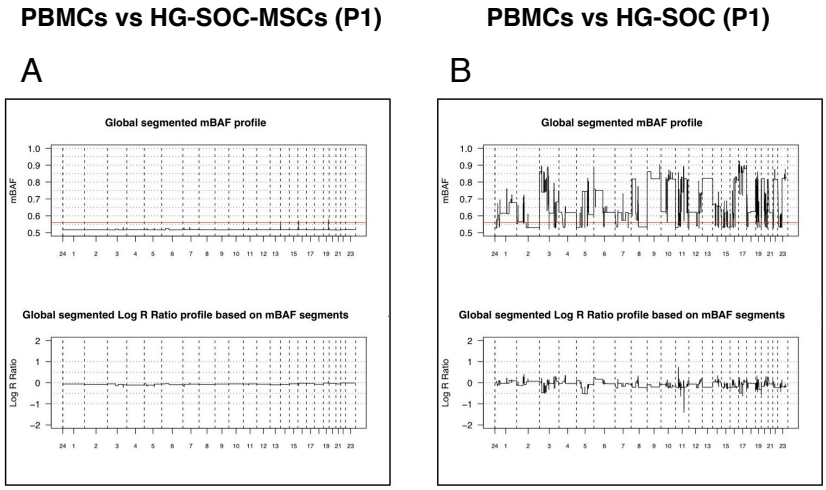

Figure 1: Transformation of B allele frequency data for matched samples. Application of the segmentation strategy (BAF estimates) to the matched three samples belonging to the same patient (SOC-43-01): Peripheral Blood Mononuclear Cells (PBMCs), HG-SOC whole tumour sample and HG-SOC-MSCs derived from the tumour sample independently hybridized on Affymetrix 450k SNP arrays. Top panel shows BAF estimates and the lower panel copy number estimates for PBMCs versus HG-SOC-derived-MSCs (A) and PBMCs versus HG-SOC (B).

Promoter activity analysis of MSCs

The transcriptional identity of HG-SOC-MSCs was analysed by comparing their profiles to the large collection of samples in the FANTOM5 Project. Ten normal tissue-derived MSC populations (N-MSCs) obtained from ten different patients (three from adipose tissue, three from bone-marrow, and four from heart tissue) and nine HG-SOC-MSCs (obtained from four different patients) (Table 1) were profiled by deep-CAGE technology (Kanamori-Katayama et al., 2011a) as part of the FANTOM5 project. In order to analyse the promoter activity peaks in terms of deep-CAGE-derived expression profiles, I performed differential gene expression analysis using a read count-based statistics approach (Robinson et al., 2010a). This kind of approach considers each peak as representative of the expression of the related gene. This permitted me to identify a list of differentially expressed genes: 625 genes that were more expressed in HG- SOC-MSCs and 450 genes more expressed in N-MSCs were identified (p-value $< 10^{-25}$, false discovery rate (FDR) corrected) as shown in Figure 2A set of differentially expressed genes has been experimentally validated by qRT-PCR analysis, in N-MSCs with respect to HG-SOC-MSCs (in Figure 2B). The resulting HeatMap in in Figure 2C shows the agreement between the expression analyses based on CAGE-seq peak activity and the qRT-PCR results. These results confirm that FANTOM5 peak activity data are representative of gene expression.

Table 1: MSC samples. N-MSCs and HG-SOC-MSCs samples used as mRNA source for the deep-CAGE sequencing analysis.

MSC line	Sex	Age	Tissue
HSLIM-16	F	71	Adipose
HSLIM-26	F	46	Adipose
HSLIM-85	F	50	Adipose
BM-21	M	56	Bone Marrow
BM-29	F	56	Bone Marrow
BM-37	M	70	Bone Marrow
HH-337-AD	M	27	Heart tissue
HH-354-AD	M	70	Heart tissue
HH-394-AD	M	57	Heart tissue
HH-421-AD	M	54	Heart tissue
SOC-19-01	F	47	Serous Ovarian Carcinoma - Left Ovary
SOC-19-02	F	47	Serous Ovarian Carcinoma - Right Ovary
SOC-41-01	F	39	Serous Ovarian Carcinoma - Left Ovary
SOC-41-02	F	39	Serous Ovarian Carcinoma - Right Ovary
SOC-43-01	F	83	Serous Ovarian Carcinoma - Left Ovary
SOC-43-02	F	83	Serous Ovarian Carcinoma - Right Ovary
SOC-57-01	F	63	Serous Ovarian Carcinoma - Left Ovary
SOC-57-02	F	63	Serous Ovarian Carcinoma - Right Ovary
SOC-57-03	F	63	Serous Ovarian Carcinoma - Right Ovary

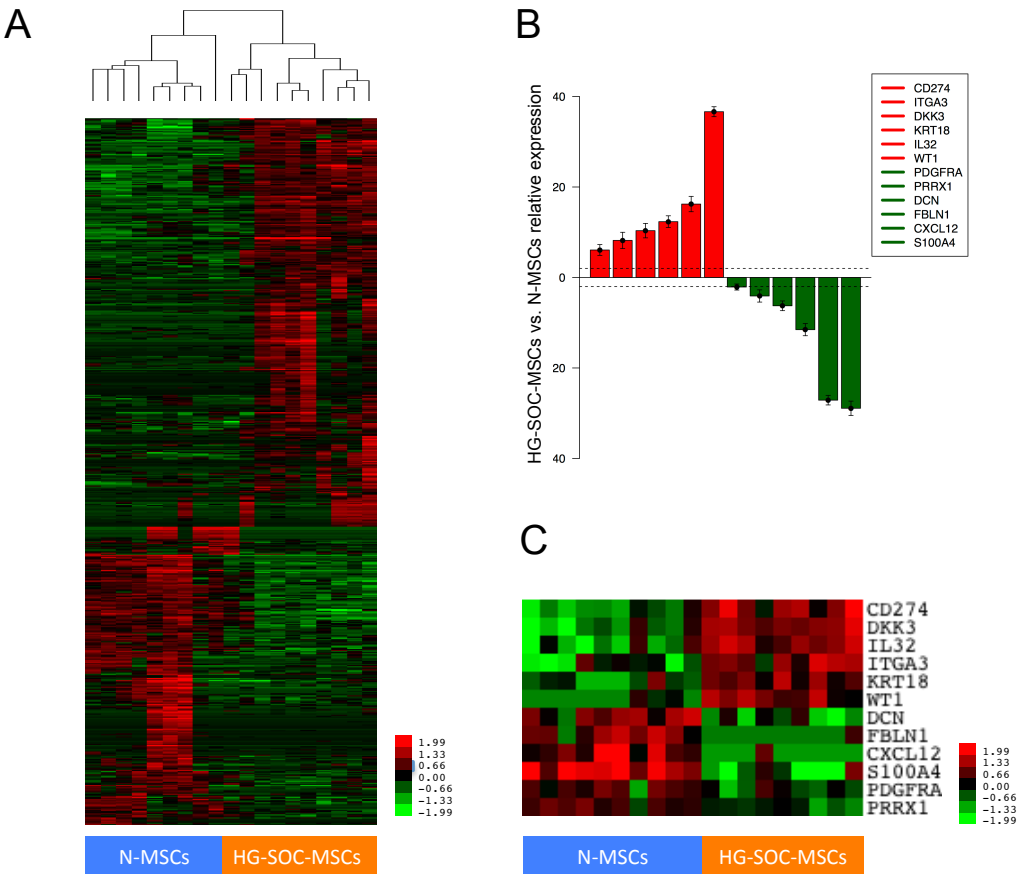

Figure 2: Transcriptional signatures of N-MSCs and HG-SOC-MSCs. A) Hierarchical clustering of the peaks data generated by deep-CAGE sequencing of HG-SOC-MSCs and N-MSCs for the differentially expressed genes (single peak analysis). Red and green bars represent, respectively, up-regulated or down-regulated genes with respect to the average of all samples. B) qRT-PCR validation of selected genes differentially expressed between HG-SOC-MSCs and N-MSCs. C) HeatMap of qRT-PCR confirmed genes from TPM matrix.

Functional characterization of MSCs

The next step was to compare the HG-SOC-MSCs and N-MSCs deep-CAGE profiles with the entire FANTOM5 phase1 dataset, which includes 889 distinct samples. For this analysis I used an approach based on the comparison of the cell lines of interest against 100 randomly generated datasets (see Materials and Methods). The obtained results were subjected to analysis of functional annotation (AFA) (Figure 3). The analysis revealed (left panel for functions, right panel for transcriptional regulators) that N-MSCs and HG-SOC-MSCs share the activation of common core biological functions and transcriptional regulators, the differences residing in the different level of their activation. Among the most up-regulated biological themes in HG-SOC-MSCs versus N-MSCs, the following emerged as significant: cancer-related and developmental functional terms (red bar and blue bar, respectively), invasion, vasculogenesis, and cell motility (Table 2). The transcriptional regulators analysis (including the activation of TGFB, VEGF, and HGF) led to the same conclusion, suggesting that globally the differences between HG-SOC-MSCs and N-MSCs are more quantitative than qualitative. In the two groups of cell lines similar core functions and the same transcriptional regulators are activated, however at different levels. The transcriptional identity of HG-SOC-MSCs was also investigated with respect to the expression of genes previously reported to be associated with ovarian cancer-derived MSCs (Lis et al., 2011; McLean et al., 2011; Spaeth et al., 2009), showing an overall gene expression concordance of 75%. Specifically, a significant correlation was found for BMP2 and BMP4 (McLean et al., 2011) and secreted factors involved in fibrovascular organization (Spaeth et al., 2009), thus enforcing their critical functions in the homeostasis of the HG-SOC microenvironment.

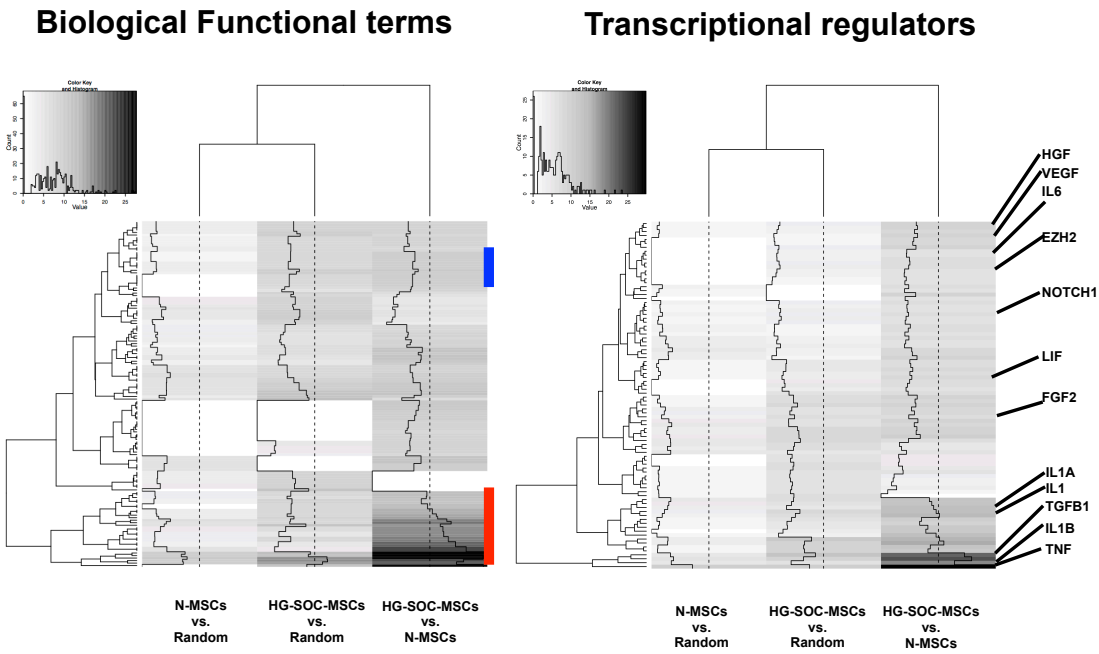

Figure 3: functional analysis of MSCs. Heat-map showing the results of the annotation of functional analysis (AFA) Starting from the genes up-regulated in HG- SOC-MSCs vs. N-MSCs, and in HG-SOC-MSCs or N-MSCs vs. 100 random- selected samples datasets, the biological functions and associated transcriptional regulators were obtained using the Ingenuity Pathway (IPA) tool and then clustered. Left panel: each block represents a single functional theme. Colored vertical bars represent the functional themes that are overrepresented in both the N-MSCs and HG- SOC-MSCs with respect to the random gene lists, but that are also globally overrepresented in the HG-SOC-MSCs with respect to the N-MSCs. Among them, a vast group is composed of terms that are predominantly cancer-related (red bar). Another group of functional terms is represented by themes related to cellular, tissue and organismal development (blue bar). Right panel: the same analysis was performed for the IPA associated transcriptional regulators. Each block represents the single transcription factor activation status, predicted by the differential expression levels of its known targets. The bar intensity reflects the statistical significance (corrected for multiple testing) of the analysis and the actual value is also plotted in trace line.

Category	Functions Annotation	p-Value	Predicted Activation State	Activation z-score
Cellular Movement	cell movement	7.45E-04	Increased	3.036
Cellular Movement	invasion of cells	6.70E-04	Increased	2.212
Gene Expression	transactivation	3.81E-04	Increased	2.067
Cellular Movement	invasion of tumour cell lines	8.87E-03	Increased	2.193
Cancer	Cancer	2.26E-02	Increased	2.2
Infectious Disease	replication of virus	1.93E-02	Increased	2.054
Organismal Development	vasculogenesis	1.41E-02	Increased	2.26
Cardiovascular System Development and Function	vasculogenesis	1.41E-02	Increased	2.26
Cellular Movement	migration of tumour cell lines	1.35E-02	Increased	2.292
Organismal Development	development of blood vessel	1.04E-02	Increased	2.262
Cardiovascular System Development and Function	development of blood vessel	1.04E-02	Increased	2.262
Cellular Movement	chemotaxis of tumour cell lines	1.01E-02	Increased	2

Table 2: Functional annotation of the HG-SOC-MSCs vs. N-MSCs differential genes. List of functional annotation terms generated by Ingenuity Pathway Analysis tool associated to genes up-regulated in HG-SOC-MSCs with respect to N-MSCs. For each term p-Value and Activation z-score is reported.

HG-SOC-MSCs Are Close Relatives of Mesothelial Cells

The FANTOM5 project dataset represents to date the broadest transcription start site-based atlas obtained from the majority of the mammalian cells and tissues (Forrest et al., 2014). Using this comprehensive dataset I had been able to define the relationships between the HG-SOC-derived MSCs and the entire set of primary cells, cell lines, and tissues (n=889), in search of the closest cell/tissue in terms of gene expression. Starting from the TPM matrix (see Materials and Methods), and excluding genes with low variance across samples (normalized SD > 0.5), I performed hierarchical cluster analysis (Figure 4A). All the samples in the dataset aggregated into four major groups: cluster 1 is mainly composed of blood and immune-related primary cells; cluster 2 is composed of adult tissues, especially of neuronal origin; cluster 3 is mainly composed of mesenchymal tissue/primary cells; cluster 4 is composed of cancer tissues/cell lines. HG-SOC-MSCs and N-MSCs are both included in cluster 3 (Mesenchymal cluster). Interestingly, comparing the different MSCs populations, we noticed that the MSCs derived from heart were the most similar to HG-SOC-MSCs (Figure 4B). To find the closest transcriptional relatives of HG-SOC-MSCs I used a hierarchical sample ontology (modified version of the Freeman ontology, hereafter referred to as sample ontology, see Materials and Methods) to divide the FANTOM5 samples into biological-related groups (n=220). I then performed a pairwise differential expression analysis between the HG-SOC-MSCs group and all the other biological groups. The analysis showed that, HG-SOC-MSCs are one of the most similar to primary mesothelial cells and several cell types (smooth muscle cells and fibroblasts) (Table 3A) hypothesized to derive from mesothelial precursors (Rinkevich et al., 2012a).

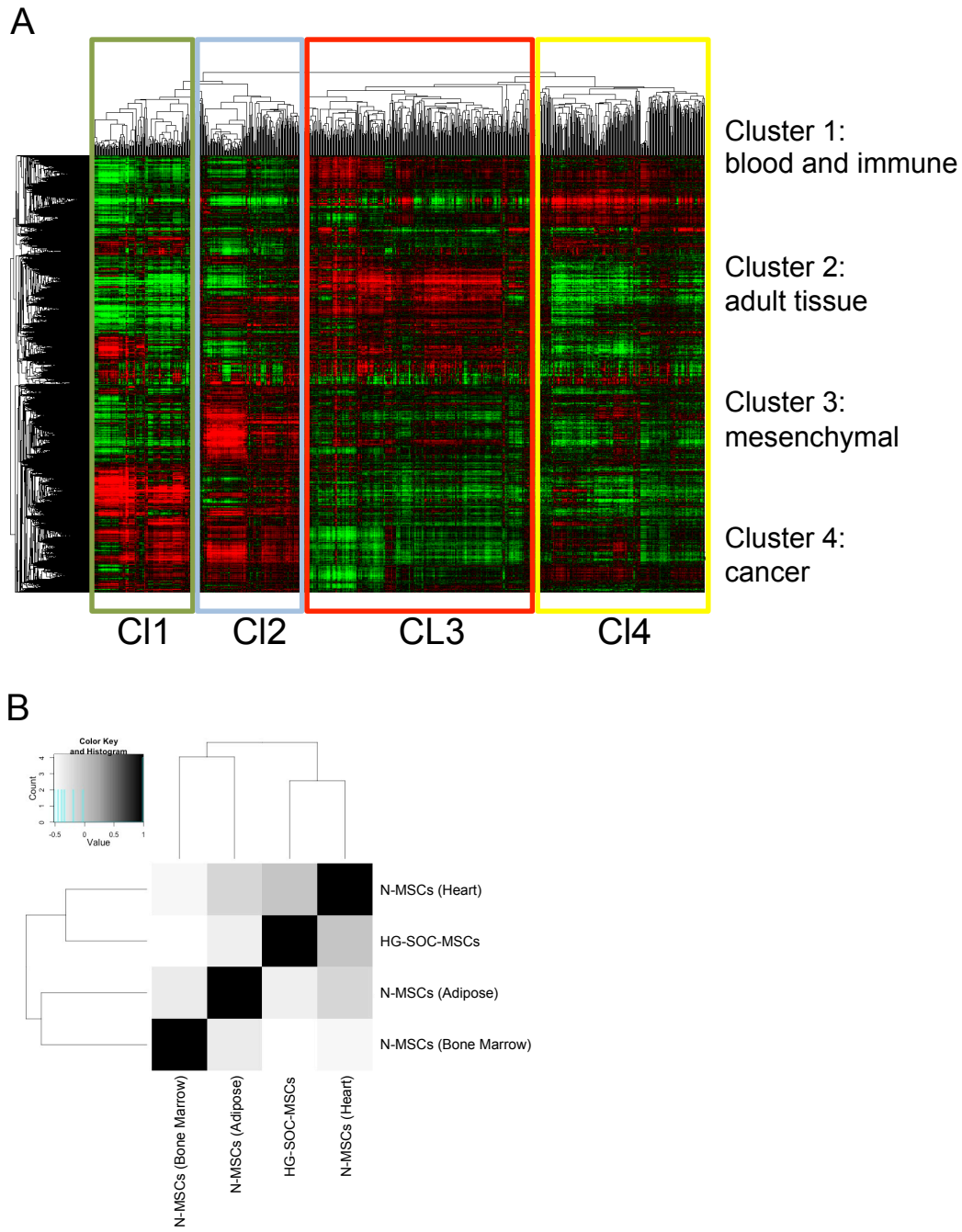

Figure 4: Hierarchical clustering of the core FANTOM5 dataset. A) Hierarchical clustering of the peaks data generated by deep-CAGE sequencing of the primary cells and tissues of the FANTOM5 project belonging to the phase1 sample dataset (n=889). Red and green bars represent peaks up-regulated or down- regulated with respect to the average value of all samples. Colored boxes highlight the four main identified clusters. B) Correlation among MSC groups. TPM Data for the different MSC groups (adipose tissue-derived N-MSCs; bone marrow-derived N-MSCs; heart tissue-derived N-MSCs; High-Grade Serious Ovarian Cancer-derived HG-SOC-MSCs) was extracted from the global FANTOM5 TPM matrix. The computed Pearson correlation matrix for all the normalized and log fold tranforform TMP peaks was then clustered.

A	B
mesothelium: mesothelial cells	**mesothelium:** mesothelial cells
smooth muscle: coronary , carotid, artery brachiocephalic, artery pulmonary, artery umbilical, artery subclavian, prostate	**smooth muscle:** aorta, artery brachiocephalic, colon, umbilical vein, artery subclavian, internal thoracic artery
fibroblasts: lymph node, pulmonary, heart, choroid plexu, conjunctiva	**fibroblasts:** periodontal ligament, heart, aortic, gingival
mesenchymal cells: Whartons Jelly, amniotic membrane, N-MSCs, hepatic, adipose	**mesenchymal cells:** bone marrow, adipose precursors, N-MASCs
stem cells: hair follicle stem cells, cord blood	**stem cells:** embryonic
tumour cell lines: serous adenocarcinoma **mesenchymal tumour cells**: lipocyte (liposarcoma) **teratoma:** peritoneum, sacrococcyx	**tumour cell lines:** carcinomas (endometrial, mucinous, squamous basal, bile duct, hepatocellular, serous, cervical), glioblastoma, neuroectodermal tumour.
	skeletal muscle: skeletal muscle

Table 3: HG-SOC-MSCs close relatives in FANTOM5 dataset. The most similar samples to HG-SOC-MSCs were reorganized into biological categories. A) The list of biological groups that resulted similar to HG-SOC-MSCs using the single peak activity data. B) The list of biological groups that resulted similar to HG-SOC-MSCs using the global peak activity data.

Differential Gene Peaks Usage Analysis Supports the Relationship Between HG-SOC-MSC and Mesothelial Lineage

To explore the full potential of FANTOM5 CAGE-seq data, we investigated the distribution of the peaks for all genes across their entire length, highlighting that, on average, nine peaks are associated to each gene. In 15% of the genes more than one peak could be considered active (TPM > 15). The presence of multiple active peaks could be an indication of open chromatin status leading to multiple transcription start sites. This finding hints to an epigenetic mechanisms of transcriptional regulation (Onder et al., 2012). I then investigated the differences of multiple active peaks between N-MSCs and HG-SOC-MSCs in order to elucidate if this kind of regulation could be important to define the particular transcriptional identity of the MSC derived from healthy and cancer tissues. In spite of the presence of a dominant peak, there were significant differences when considering the presence of multiple peaks along the entire gene length, As shown for two representative genes (Figure 5A and B). I identified 203 genes with differential number of active peaks between N-MSCs and HG-SOC-MSCs (Supporting Material Table S3A and B). Only a fraction of the resulting genes (22%) were highlighted in the previous differential expression analysis, unveiling differences not otherwise detected. Finally, differential gene peaks

usage data were used to search the cell/tissue similar to HG-SOC-MSCs (see Materials and Methods): as shown in Table 3B, this analysis further confirmed the evidence that HG-SOC-MSCs are close relatives of mesothelial cells.

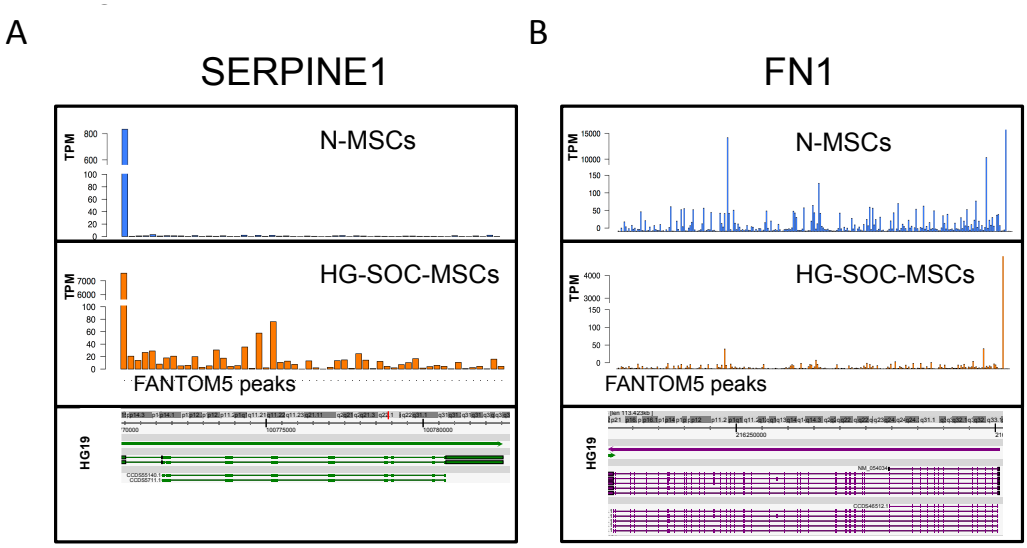

Figure 5: Example of two genes with differential global peak activity. Starting from the TPM matrix, the TPM values of the peaks were plotted across the entire length of each gene. A) a representative gene (SERPINE1) having a higher number of active peaks in HG-SOC-MSCs with respect to N-MSC; B) a representative gene (FN1) having a lower number of active peaks in HG-SOC-MSCs with respect to N-MSC. Bottom panels show the graphical representation of the genomic position and the structure of the genes from UCSC refseq and CCDS HG19 tracks.

A Specific HG-SOC-MSCs Mesothelial-Related Gene Signature Correlating With Serous Ovarian Cancer Prognosis

Starting from the lists of genes found to be differentially expressed between HG-SOC-MSCs and N-MSCs, derived from both the differential gene expression analysis and the differential gene peaks usage analysis, I performed an exhaustive literature search using automated tools (Ingenuity Pathway Analysis tool or molecular signatures database (Liberzon et al., 2011)) and manual tools (PubMed, Protein Atlas (Pontén et al., 2008)) to search for mesothelial-related genes. Among them, I selected six genes previously reported as markers of the mesothelial lineage (CALB2, SPP1, KRT7, MSLN, CNN1, and LRRN4) and four mesothelial/mesothelioma-related genes (TLR2, FN1, MCAM/CD146, and KRT8) (Barberis et al.; Bidlingmaier et al., 2009; Connell and Rheinwald, 1983; Kachali et al., 2006; Kanamori-Katayama et al., 2011b; Ksiazek et al., 2009; LaRocca and Rheinwald, 1984; Park et al., 2007; Sato et al., 2010; Taniguchi et al., 2001; Tigrani and Weydert, 2007) and I generated a HSM-GS composed of nine genes (CALB2, SPP1, KRT7, MSLN, CNN1, FN1, MCAM/CD146, TLR2, and KRT8). The expression of all these genes has been confirmed by qRT-PCR (Figure 6A).

In order to study the specificity of the identified transcriptional signature with respect to other types of cancers, I analysed high-throughput gene expression datasets (GSE23066 for non- small cell lung cancer, GSE29270 for breast cancer, and GSE36474 for myeloma) comparing MSCs derived either from cancerous tissue or from the healthy tissue counterpart. As shown in Table 4, the HSM-GS signature is unique to MSCs derived from HG-SOC and not other cancers.

	GSE23066		GSE29270		GSE36474		HG-SOC-MSCs vs N-MSCs	
	Fold change	**pValue**	**Fold change**	**pValue**	**Fold change**	**pValue**	**Fold change**	**pValue**
FN1	2.19	0.298	NA	0.445	-2.73	0.354	-26.12*	0.029
SPP1	1.00	0.589	1.11	0.205	1.28	0.622	5.56	1.43E-11
CNN1	1.00	0.294	NA	NA	-1.03	0.085	12.03	1.74E-19
MCAM	1.01	0.319	1.06	0.597	1.03	0.036	16.70	1.53E-22
KRT7	1.29	0.971	NA	0.040	-1.98	0.675	39.14	2.25E-10
TLR2	-1.43	0.370	1.05	0.002	1.09	0.003	90.84	7.35E-15
KRT8	1.70	0.799	1.02	0.391	1.26	0.169	100.62	7.75E-35
CALB2	1.03	0.067	NA	NA	-1.22	0.031	133.94	2.47E-08
MSLN	-1.00	0.205	1.06	0.303	1.19	0.617	346.65	1.86E-13

* data from differential peaks usage analysis

Table 4: HG-SOC-MSCs gene signature (HSM-GS) specificity. Differential gene expression analyses between MSCs isolated from cancerous (lung, breast and myeloid cancers) and the corresponding healthy tissues (GSE23066, GSE29270 and GSE36474 datasets). In the table we report for each dataset the fold change and the corresponding p-values of the HSM-GS genes. Among the non-ovarian tissue-derived MSCs, only very few HSM-GS genes were differentially expressed between the MSCs derived from the cancerous and the corresponding normal tissue. Data from differential expression analysis obtained comparing HG-SOC-MSCs and N-MSCs are also shown in the last column of the Table.

I was then interested to see if the HSM-GS derived from HG-SOC-MSCs might correlate with ovarian cancer prognosis. I analysed several microarray datasets of serous ovarian cancer, to my knowledge encompassing all publicly available SOC profiles based on the Affymetrix platform at that time, collectively consisting of more than 900 patients (Gyorffy et al., 2012). Kaplan-Meier survival analysis of the combined datasets showed that SOC patients with higher levels of the selected genes displayed shorter PFS time ($p<2.5^{-05}$), as shown in Figure 6B. To evaluate the HG-SOC-MSCs gene signature specificity I performed the same analysis on several breast cancer microarray datasets comprising more than 2,500 patients (Györffy et al., 2010b) and several lung cancer microarray datasets comprising more than 700 patients (Győrffy et al., 2013). Interestingly I observed that breast cancer patients expressing higher levels of the HG-SOC- MSCs gene signature displayed longer PFS time ($p<7.1^{-05}$ in Figure 6C), while lung cancer patients did not show significant correlation between the HSM-GS gene signature and FP survival time ($p=0.06$ in Figure 6D). Different performances of the HG-SOC-MSCs signature in different kind of tumours suggest that the genes included in the signature have a specific role in serous ovarian cancer. This specificity could reflect a specific interplay between cancer cells and the tumour micro-environment.

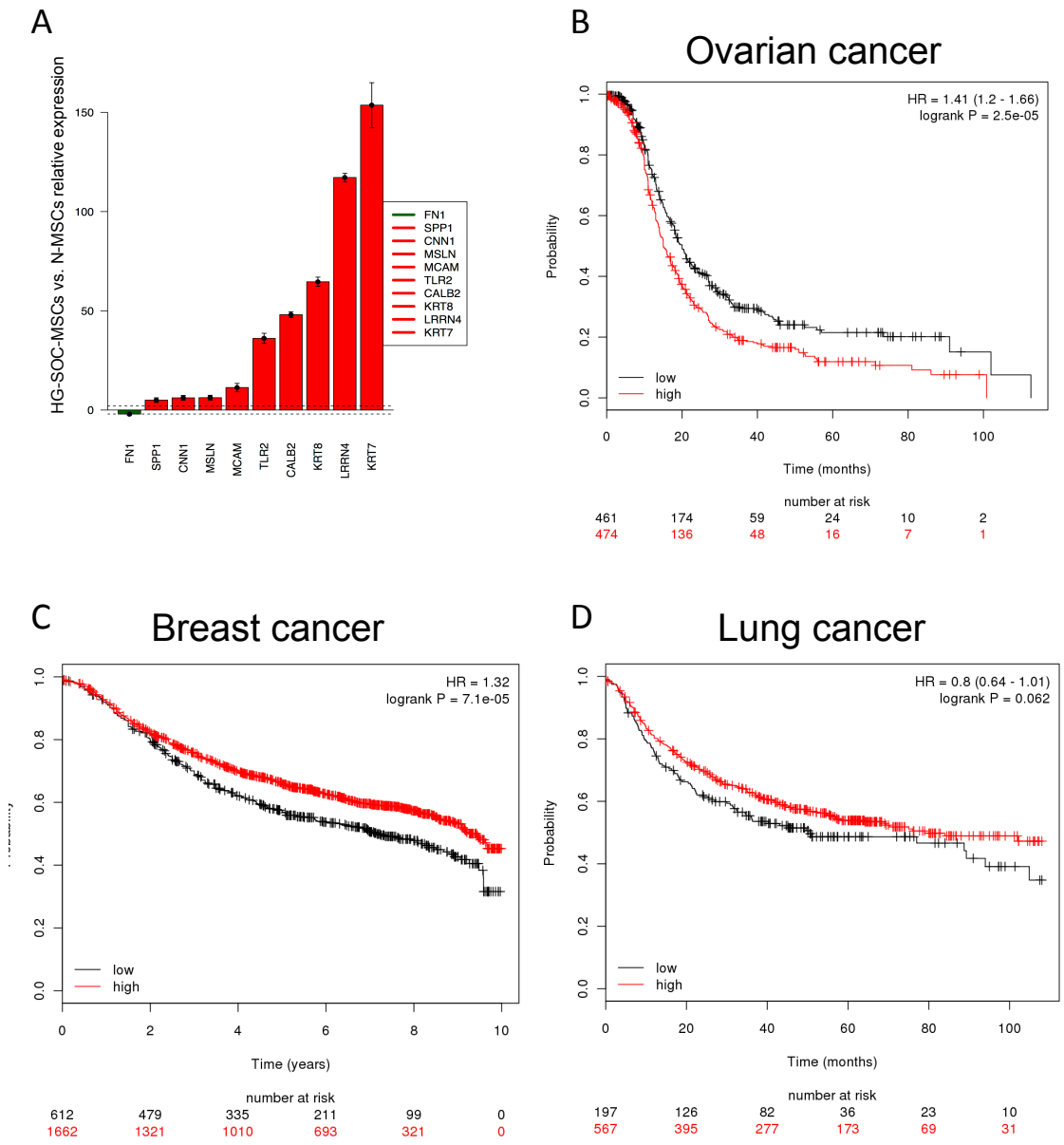

Figure 6: A) qRT-PCR analysis of HG-SOC-MSCs mesothelial marker genes and selected mesothelial-related genes. Relative expression levels in HG-SOC-MSCs with respect to N-MSCs. All the genes are significantly over-expressed (at least $p<0.05$) in HG-SOC-MSCs with respect to N-MSCs. B) Survival curves are shown for serous ovarian cancer patients with high (red) and low (black) expression of the selected gene signature. PFS time (PFS< 96 months) has been selected as clinical outcome. Hazard ratio and significance are also reported. SOC patients expressing higher levels of the selected genes in tumours displayed shorter progression free survival time ($p=2.5^{-05}$). C) Survival curves are shown for breast cancer patients with high (red) and low (black) expression of the selected genes. PFS time (PFS< 96 months) has been selected as clinical outcome. Hazard ratio and significance are also shown. Patients expressing higher levels of the selected genes in tumours displayed longer progression free survival time ($p=7.1^{-05}$). D) Survival curves are shown for lung cancer patients with high (red) and low (black) expression of the selected genes. FP time (FP< 96 months) has been selected as clinical outcome. Hazard ratio and significance are also shown. Patients expressing higher levels of the selected genes in tumours displayed no significant change in first progression survival time ($p=0.06$).

Mesenchymal stem cells characterization final remarks

I analysed the identity of HG-SOC-MSCs respect to the vast panel of tissues and cell lines included in the FANTOM5 dataset. Taking advantage of deep-CAGE sequencing I have been able to show that HG-SOC-MSCs possess distinct and specific CAGE-peak promoter activities with respect to N- MSCs. Among all the FANTOM5 samples, both gene expression and gene peak usage analysis results suggested that HG-SOC- MSCs were close relatives of mesothelial cells, and their derived mesothelial-related signature is correlated with bad prognosis in ovarian cancer. All the results have been published (Verardo et al., 2014).

Create new tools to discover regulatory elements in the human genome

The FANTOM5 dataset includes information about the transcription starting sites in a wide range of human tissues and cell lines. This is very useful not only to evaluate the expression of the genes, but also to specifically study the activity of their promoters and the fine regulatory network that drives gene expression. Every experiment represents a set of genes that are co-expressed in a specific tissue or cell line and it is very likely that co-expressed genes share common transcriptional regulators. Because of the fact that we know the position of each transcription starting site (TSS), we are able to define a set of sequences that are included in the promoters of a set of co-expressed genes. In order to find the transcriptional regulators that are able to bind such promoter sequences and regulate genes, me and my research group developed a de novo motif discovery software: ScanAll. The main feature of this software is that it is able to find modules (combinations of motifs) that are enriched in a set of sequences. In the following sections I will show the results obtained by analysing the FANTOM5 dataset using ScanAll.

ScanAll working pipeline

We used our newly developed software ScanAll on the genomic regions surrounding the peaks of the FANTOM5 dataset to discover sample-specific motifs and composite regulatory modules (Figure 7). ScanAll is a software composed of three main components: ScanPro, which looks for single over-represented motifs in biological sequences; ScanMod, which sub-selects only the motifs belonging to composite regulatory elements; a final Filtering phase, which retains only the most significant modules.

During the *ab initio* motif discovery phase we fixed ScanPro motif length ℓ=6 and the number of mismatches d=1, in order to account for the most common structure/length of the core sequence of a TFBS (see Discussion and Methods for more details). This variation is prevented from occurring in the first position of the element and the number of different nucleotides appearing into a motif (*Complexity, C*) is set to four. ScanMod then processed these motifs, generating modules that are combinations of motifs. There is no limit to the number of motifs that can compose a module, so potentially, ScanMod can find modules composed by any number of motifs. Despite this absence of limitations ScanMod has not found modules composed by more than

three motifs in any sample. The next step is the statistical validation and filtering of the modules. To do so we compared our results respect to the results obtained analysing ten randomly generated datasets. The random datasets have been obtained shuffling the di-nucleotides of the original dataset. This approach permits us to obtain random datasets that, respect to the original FANTOM5 dataset, have the following features: 1) the datasets have the same number of samples respect to the original dataset 2) each specific sample has the same number of sequences respect to the original sample 3) each sequence has the same length respect to the original sequence 4) each sequence has the same di-nucleotide composition respect to the original sequence. This approach not only permits to have a reliable background, but also prevents the creation of low complexity sequences in the random datasets. Even though ScanAll does not directly filter out low complexity sequences, we set the minimum complexity of motifs to four.

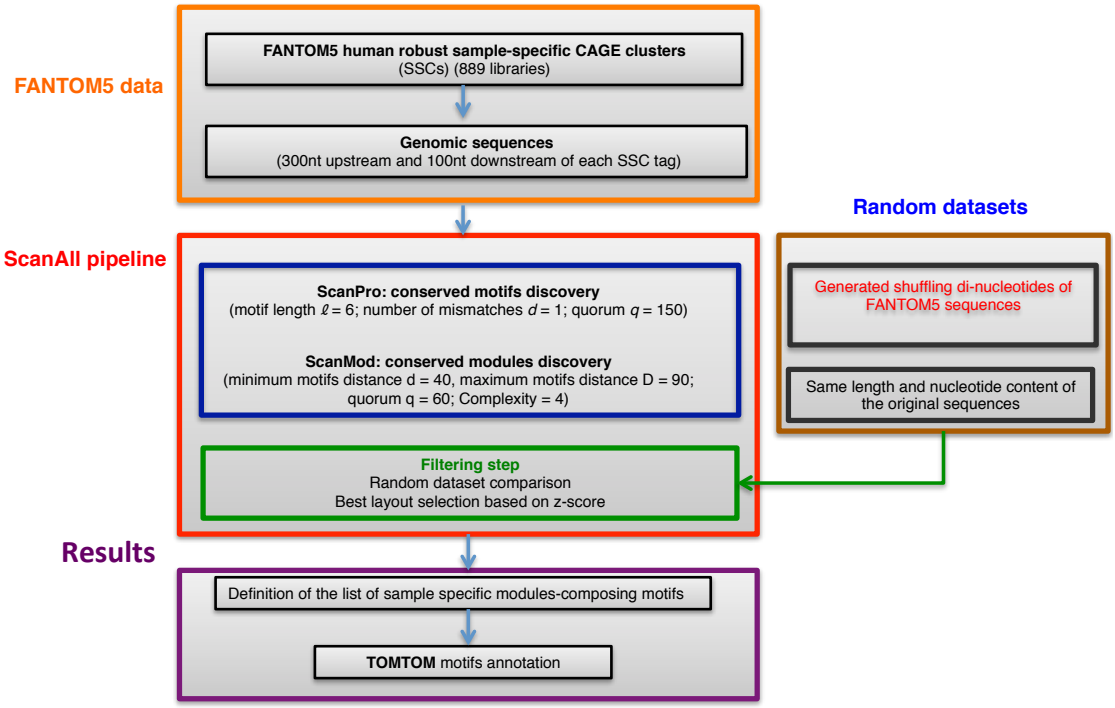

Figure 7: ScanAll pipeline. Block diagram describing the different steps of the ScanAll pipeline. Starting from the FANTOM5 genomic sequences derived from Sample-Specific CAGE clusters, the ScanAll pipeline proceeds through 3 phases: ScanPro, the motif-finding step; ScanMod, the module-finding step; the filtering step.

Motifs and *cis*-regulatory modules discovery

Globally, ScanAll identified 580773 unique modules, with more than two thirds of the results displaying combinations of two motifs, the remaining composed of three motifs and none made of a higher number of motifs. Looking into more detail to the composition of these enriched modules, they were generated by the association of 2822 non-redundant motifs. The identified enriched elements localise only to specific, well-defined sub-regions of the examined promoters. Notably, the positions of the modules on the promoter regions are not completely random respect to the transcription starting sites (Figure 8). The frequency of modules is low in the portion of the promoter that is close to the transcription-starting site. This could be possibly due to the presence of low complexity sequences, such as TATA box.

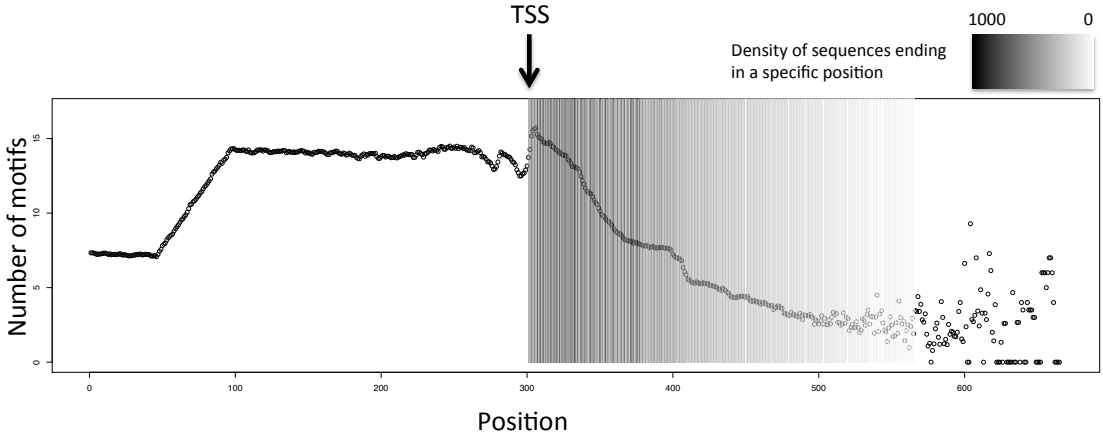

Figure 8: figure showing the position of the modules derived motifs found by ScanAll respect to the transcription starting site. The number of motifs is divided by the number of sequences that cover that position. The grey scale shows the number of sequences ending in that specific position. The low number of motifs in the first hundred positions is due to the fact that modules are composed by motifs with a distance constrain (40-90 nucleotides); because of this in the beginning and in the end of sequences the probability to find modules is lower.

Considering the distribution across samples of the number of motifs and motif instances (Figure 9A and B), we found that the majority of samples displayed less than two hundreds of different motifs. We then considered the distribution of modules across the samples (Figure 9C and D). In this case we can observe that a high number of modules are found in only one sample, indicating that our modules are possibly regulatory elements with high specificity for a particular cell line or tissue.

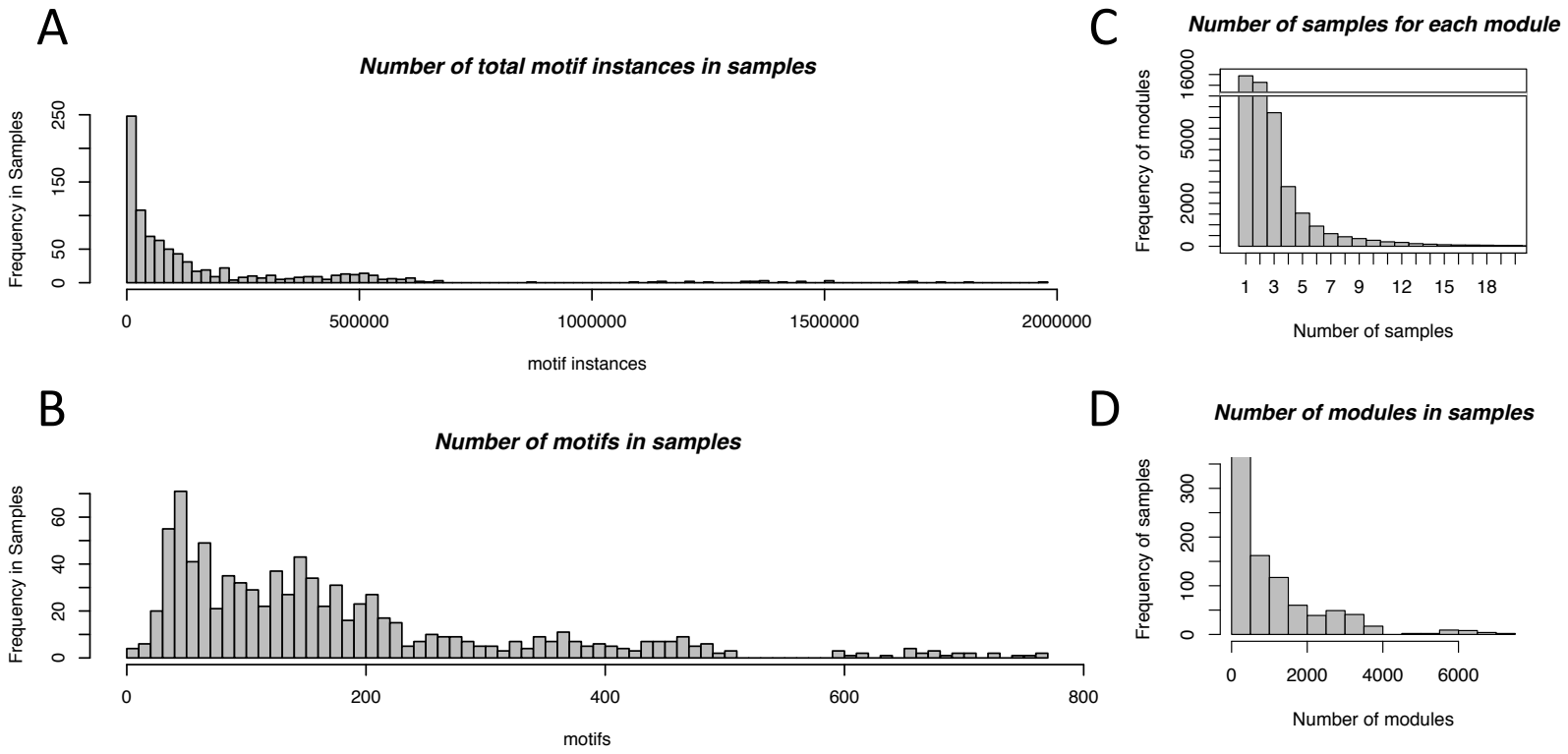

Figure 9: Distribution of the number of motifs and modules. A) Distribution of motif instances across samples. B) Distribution of motifs in samples. C) Distribution of the number of modules in which the same module has been found. D) Distribution of the number of modules in each sample.

Annotation of cis-regulatory modules

To define the TFs putatively able to bind these sequences, the module-composing motifs were independently annotated with the TOMTOM tool (Gupta et al., 2007). Hits were retained within an E-value < 0.5 to human known TFBS in the TRANSFAC (Matys et al., 2006), SwissRegulon (Pachkov et al., 2007), JASPAR (Bryne et al., 2008) and UniPROBE (Newburger and Bulyk, 2009) public databases. Data from the ENCODE *cis*-regulatory lexicon (Neph et al., 2012), the ChIP-seq derived HOMER motifs (Heinz et al., 2010) and the HOCOMOCO collection (Kulakovskiy et al., 2013) was also used. The outcome resulted in 2117 motifs (75% of the total number of module-composing motifs) that were associated with a known TFBS, the remaining 705 (25%) being without any correspondence. There can be several reasons for this result but the most likely are on the one hand the possibility that part of the motifs composing these modules are highly enriched structural elements but not TFBS, and on the other hand that they represent uncharacterized binding sites whose function still needs to be defined. In fact, although these motifs were not annotated using TOMTOM, MACRO-APE (http://autosome.ru/macroape/) allowed to co-cluster 70% of them with novel motifs predicted by other FANTOM5 methods (DMF (Marchand et al., 2011), HOMER and ChIPMunk) or identified by the ENCODE Consortium. Moreover, the remaining 30% co-clustered with other ScanAll annotated motifs, this possibly indicating their status of biologically relevant sequences and, more specifically, suggesting a role as novel TFBS. The transcription factors binding sites identified by our method and the integration with the other motif findings methods used during the FANTOM5 project, had been recently published on Nature (Figure 10) (Forrest et al., 2014). Interestingly, we observed that the totally unknown modules were significantly enriched in samples obtained from tissues with respect to cell lines and primary cell specimens (P-value < 10^{-11}, Chi-squared test), with central nervous system and gastrointestinal tract samples being among the major contributors to this category.

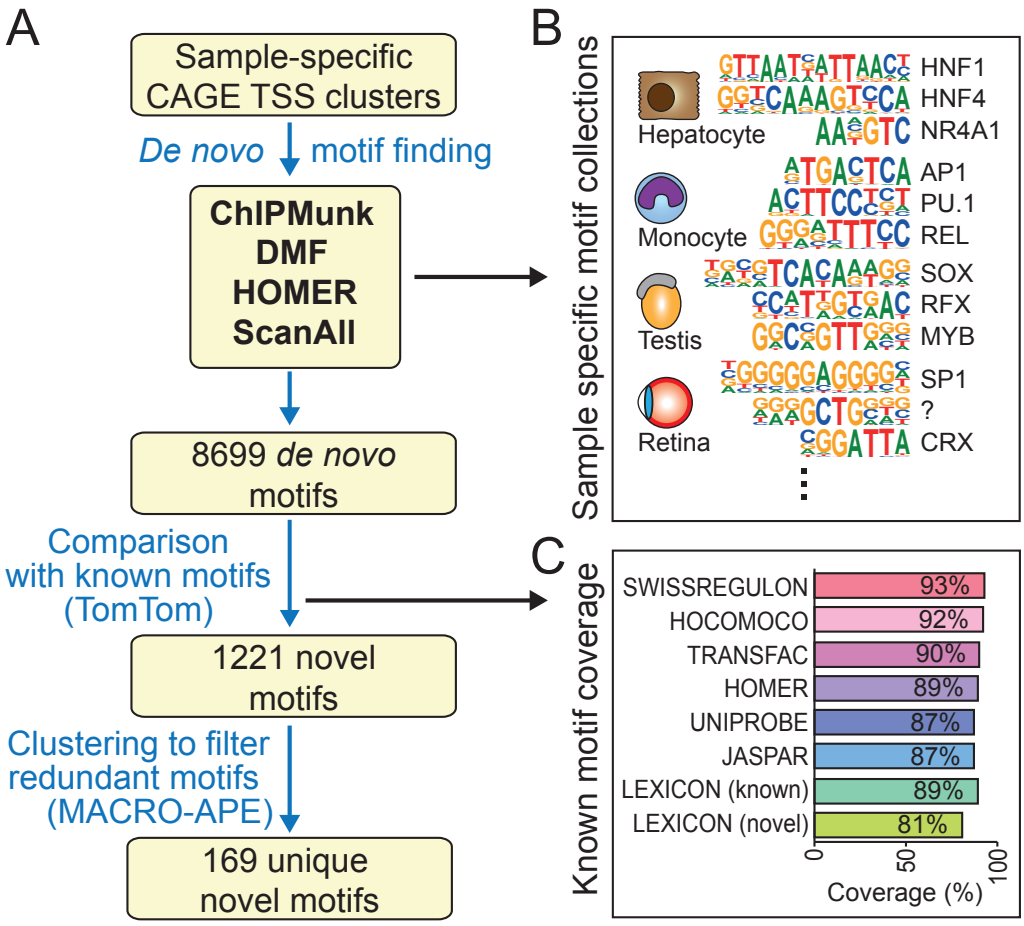

Figure 10: De novo derived, cell-state-specific motif signatures. The de novo motif discovery tools DMF, HOMER, ChIPMunk and ScanAll were applied to detect sequence motifs enriched in the vicinity of sample-specific peaks; a), yielding 8,699 de novo motifs; b). The coverage of known motif space by the de novo motifs was evaluated by comparing them to the SWISSREGULON, HOCOMOCO, TRANSFAC, HOMER, JASPAR, and ENCODE LEXICON motif collections; c)The remaining 1,221 de novo motifs that were not similar to known motifs were then clustered using MACRO-APE, resulting in 169 unique novel motifs.

Regulatory Modules Compendium

Afterwards, we analysed the distribution of modules across samples (Figure 9C) and the number of different modules found in each sample (Figure 9D) and we found that they originated mainly from small, well-defined groups of samples.

Only a small number of modules (1155 modules, representing 0.2% of the total number of modules) are present in more than a hundred samples: the more common module is AGCTGN, AGCTGN, found in 431 samples. Interestingly, this module is composed by two copies of the same motif, that is associated to TFAP4, an helix-loop-helix protein that can form homodimers and contains multiple protein-protein interfaces (Hu et al., 1990).

We decided to collect all our findings and create a compendium that could be useful in order to easily access information about regulator modules in each specific tissue or cell line. We started considering all the 580773 ScanAll identified modules: we collapsed all the modules holding the same TOMTOM annotation, obtaining a list of 86390 group-specific TF modules. They were subsequently filtered to retain only those expressed (>2.5 TPM) in the specific functional groups, generating an Expressed TFs module Compendium containing 26741 elements (22136 for >5 TPM). Finally, since Expressed TFs modules describe the particular condition in which two TFs are expressed and putatively able to bind to nearby regions in the same promoter, we identified couples of TFs that could establish protein-protein interactions (PPI), obtaining what we called the PPI Expressed Compendium. Figure 11A and B show respectively an example of the Expressed TFs module Compendium and PPI Expressed Compendium for a thyroid tissue sample. The composition across samples of the Compendium TFs is variable, accounting for the different roles played by widely expressed (possibly housekeeping) or group-specific factors. However, what impressively characterizes our results is the combinatorial effect of Compendium TFs, as it allows underlining very specific regulative modules that would be lost considering each TF separately.

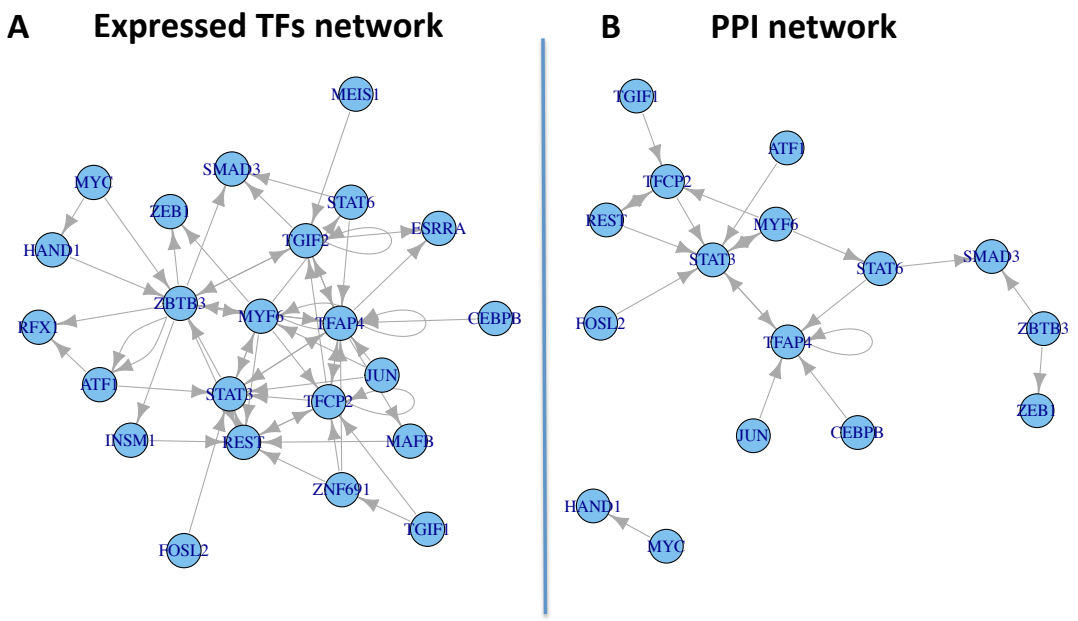

Figure 11: Example of the Expressed TFs module Compendium A) and PPI Expressed Compendium; B) for a thyroid tissue sample. TFs are represented as nodes. In A edges indicate the presence of a module between two TFs expressed in that specific sample. In B edges indicates the presence of a module between two expressed TFs that can interact at protein level.

Functional analysis

The Compendium can be used to describe factors that could perform housekeeping functions, or that play a specific role in gene regulation in every cell line or tissue. In order to delve deeper into the biological functions that are regulated by the TFs included in our compendium we performed functional analysis starting from the list of TFs that form modules in each biological group, using ClueGo, a function enrichment tool for Cytoscape.

We then tried to understand if the functional enrichment analysis results were consistent when comparing the functions performed by every specific tissue. To do so, using the functional enrichment information, we performed unsupervised hierarchical clustering of every biological class. As shown in Figure 12, groups of samples that are biologically related, such as blood and brain samples, are clustered together. Despite the good results obtained for this specific groups of tissues and cell lines, it has not been possible to cluster all the tissues and cell lines in a satisfactory manner. The reasons of this are probably many: 1) Functional annotation of TFs is partially biased by the nature itself of these genes. Each TFs is involved in different biological functions and probably it plays more important roles not yet well defined or discovered. This observation could be more important in the light of our results. Our

analysis revealed that the regulatory specificity in each tissue is controlled by the specific combination of TFs. So, the functional annotation of single TFs probably can't give us enough information about the biological functions performed by the gene. 2) Cancer cell lines seem to be totally different from their healthy tissue of origin. The dataset contains healthy tissues, primary cell lines and transformed cell lines. This analysis clearly evidences that cell lines cluster together with other cell lines, despite of their tissue of origin. We can suppose that the differences between different tissues are not conserved in the derived cell lines. 3) It's difficult to define the active functions in a specific tissue or cell line. We do not have enough information to define which biological functions should be active in a specific tissue or cell line. From a semantic point of view, the terms included in the gene ontology could be biased by the type of tissue or cell line in which a certain gene or process had been studied.

Taking into account all these results, we can say that the combination of TFs that can bind our modules in every specific sample could be linked to specific biological functions. These functions are shared between samples coming from the same tissue of origin in the case of Brain and blood samples. Despite these particular cases, the cell lines share biological functions between them and probably lose specific functional traits that are characteristic of their tissue of origin.

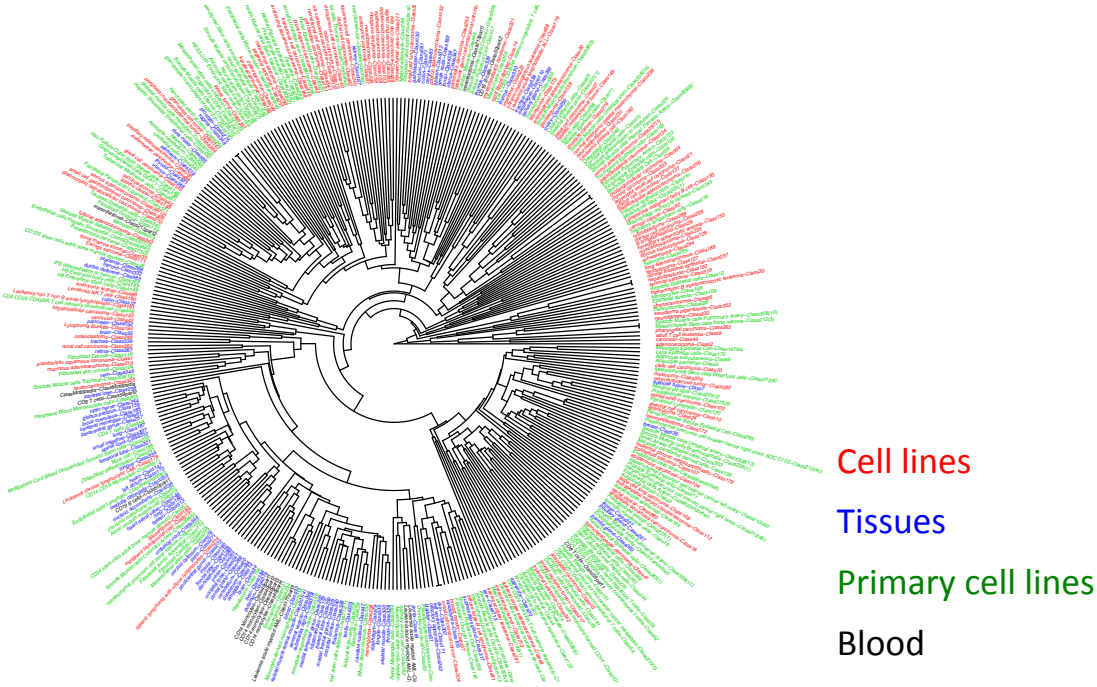

Figure 12: hierarchical clustering of functional annotation. All the samples included in the FANTOM5 dataset have been clustered based on the functional annotation of the modules TFs found by ScanAll. Brain and blood samples cluster together while cancer cell lines do not cluster close to their tissue of origin.

Association with Transposable Elements

There is increasing evidence (Brosius, 2003; Johnson et al., 2006; Laperriere et al., 2007) associating the presence of binding regions to transposable elements. Given the fact that a portion of ScanAll motifs could not be coupled to any known TFBS, we decided to verify the degree of correlation existing between the modules-composing motifs identified in the subset of samples also analysed in the ENCODE project (Wang et al., 2012) and the genomic Transposable Elements (TEs). We investigated on the one hand if transposition events could have spread specific binding sites across the mammalian regulatory regions and on the other hand if motifs themselves could be enriched structural elements enhancing transposition.

As a general feature, we observed that only a small fraction of promoters overlapped known TE regions. This, however, was somehow expected considering that in the analysed human transposable elements database (retrieved from the TranspoGene project, (Levy et al., 2008)) only 9% of the genes (1772/20114) had TEs associated with their promoting regions. Nevertheless, we observed that in the samples in which more than two different promoters overlapped TE regions, the motifs found in TE

regions were over- or under-represented with respect to non-TE regions (P-value < 0.05, permuted rank tests, multiple correction applied).

We decided to focus firstly on the HeLaS3, HepG2, GM12878 and K562 cell line samples profiled during the FANTOM5 project, as the ENCODE Consortium already analysed them to find binding sites enriched in TEs. Our results showed there were indeed significant correlations (Figure 13), finding one and three motifs being respectively over- and under-represented in TEs in at least two cell lines. Moreover, six other motifs displayed opposite behaviour depending on the cellular context, while 65 had a lineage-specific occurrence, eight being found only in HeLaS3 (three over- and five under-represented), four only in HepG2 (three over and one under), 23 only in GM12878 (12 over and 11 under) and 30 more only in K562 (14 over- and 16 under-represented).

Interestingly, the ENCODE Consortium results provided supporting evidence regarding the spatial association we found in HepG2 TEs between MAFB and NFE2L2 binding sites, as well as for the presence of JUN sites; besides, we also confirmed the mapping of ESR1 sites to transposable elements, as previously seen in MCF-7 cells (Bourque et al., 2008; Polak and Domany, 2006). On the other hand we could not verify the association between CEBPB, STAT1 and JUN sites, found by the ENCODE Consortium in HeLaS3 and HepG2 cells, as we co-mapped them only in GM12878 cells, where they displayed an opposite behaviour being under-represented in TEs.

Motifs	Annotation	HeLaS3			HepG2			GM12878			K562		
		CNhs12325	CNhs12326	CNhs12327	CNhs12328	CNhs12329	CNhs12330	CNhs12331	CNhs12332	CNhs12333	CNhs12334	CNhs12335	CNhs12336
GCTCAN	ZNF691	4.44	13.45										
TTNCAG	UW.Motif.0494	2.55	8.28	2.62									
TGGCNA	RFX1		2.76	4.66									
CAGTGN	ZBTB3	4.71		2.33							2.16	2.04	
ANCCTG	UW.Motif.0363	3.08		2.35								-2.69	-4.34
AGCTGN	TFAP4	2.53		2.37				-3.26	-3.55	-2.06			
AGNCCT	UW.Motif.0435											-2.34	-3.53
CTGNGA	UW.Motif.0480		-2.69	-2.01									
CTNGGA	STAT4	-2.72		-2.10									
TNCCAG	HAND1	-3.05		-2.75	-3.66	-2.99	-6.64					-2.71	-2.18
CTGGNA	HAND1	-3.73		-2.95									
CTGNAG	UW.Motif.0627		-4.12	-2.83				-2.75	-2.88		-4.48	-3.74	-2.09
AGNCTG	UW.Motif.0012		-3.09	-5.51									
TGCAGN	UW.Motif.0515		-5.01	-3.81									
TCAGGN	ESR1				9.59		10.35	-15.25	-4.78				
CTNAGG	STAT6				3.85	4.41	4.17					-3.58	-7.12
GCTGAN	MAFB				2.94		3.04						
GCTNAG	NFE2L2				2.57		2.04						
CAGNCT	UW.Motif.0012					2.03	2.07						
GAGCTN	nomatch				-3.64	-2.06							
CTGGAN	SMAD3				-8.07	-11.48	-6.94				-3.28	-2.06	-2.80
CAGNGT	UW.Motif.0235							4.44	7.89	7.08			
AGNTGC	NHLH1							4.65		3.76	-2.07		-3.47
CTCANG	FOSL2							2.92		5.16			
GANGCT	nomatch							3.65		3.65			
TGANCC	RXRA							2.08		4.45			
AGACTN	nomatch								3.02	3.50			
TGAGCN	nomatch							2.11	4.77	2.40			
GNCTCA	JUN								2.93	3.10			
TGNCCA	UW.Motif.0045							2.33	3.72	2.54			
CACTGN	ZNF691							2.13		3.09			
TNCTGA	REST							2.16	2.19	2.62			
GNCTGA	PAX5								2.35	2.23			
CNTCAG	MAFB							2.10	2.03				
CANGTG	TFE3							-2.54	-2.05	-2.25	2.98	3.03	
TGNAGC	nomatch								-2.58	-2.06			
CTNAGA	STAT5B								-2.13	-2.55			
TCAGNC	PAX5							-2.19	-2.03	-3.09			
GCANCT	NHLH1								-3.04	-2.08			
TCTGAN	nomatch							-2.24		-3.49			
TGNGCA	CEBPB							-2.67		-3.19			
TCAGNA	REST							-3.54	-2.65				
TNTCAG	UW.Motif.0118							-3.99	-2.49	-3.29			
GANCTG	UW.Motif.0229							-5.14		-2.11			
CCTNAG	STAT6							-5.42	-7.03	-13.43			
CCTGAN	THA								-7.20	-10.61			
CNAGTG	UW.Motif.0221										7.88	10.33	7.15
GACNTC	JUN											6.54	5.19
GNCAGT	MYBL1										6.30	3.61	4.24
CNAGGT	RORA										3.42	3.90	4.64
TGGCAN	TLX1										4.32		3.12
TCNGGA	nomatch										6.30	2.73	2.01
CCANGT	ATF6											2.73	3.63
GACCTN	RARA										2.27	3.03	3.42
CTGACN	MAFB										2.10	3.40	2.75
TGGNCA	UW.Motif.0045										2.54	2.42	
AGANCT	UW.Motif.0229										2.07	2.63	2.48
CNAGCT	NR2E3										2.23	2.48	
CCAGNT	TFCP2											2.26	2.26
CTGANT	UW.Motif.0625										2.30	2.05	
CAGCTN	TGIF2											-2.61	-2.06
CACNTG	TFE3											-2.05	-2.85
AGCTNC	UW.Motif.0218											-3.06	-2.22
CTGAGN	PAX5											-2.32	-3.36
GNTGCA	CEBPB											-2.48	-3.34
CAGTCN	nomatch											-3.53	-3.41
TNAGGC	TFAP2A											-2.83	-4.12
TCNGAG	nomatch										-4.16	-2.66	-3.64
AGCCTN	UW.Motif.0205											-2.68	-4.34
TCANGG	RARG											-2.15	-5.52
TCCAGN	nomatch										-3.36	-5.60	-5.25
CTCAGN	PAX5										-7.32	-3.56	-4.92
TGNCAG	UW.Motif.0028										-7.06	-4.23	
GAGNCT	nomatch										-4.86	-8.06	
CTGANG	MAFB											-4.38	-9.48

Figure 13: Motifs overlapping TE regions in the twelve ENCODE cell line samples. The table represents the Z-score of the number of motifs overlapping TE regions with respect to the total motifs. For each sample, values were normalized for the proportion of promoters covered by TE regions (in base pairs) with respect to promoters not covered by TE regions. Values are separately showed for each motif and for each sample.

Ongoing research: The complete module atlas of the human genome

During my PhD project I found the regulatory modules using the information coming from the FANTOM5 dataset. In particular, the analysis focused on finding modules that were shared between the promoters that were active in a specific sample. The dataset included different tissues and cell lines, and only rare and really specific tissues (such as embryonic tissues) are not included. We can really think that the promoters found in this work are the majority of all the promoters in our genome. The next step is to use ScanAll to analyse the entire promoterome, without taking into account of which promoters are active in each sample. This approach will permit to find genomic elements that are conserved in the promoters of human genome. These elements are not necessary linked to regulation of transcription through TFs binding, in fact some of them could represent structural elements or have not yet characterised functions.

To do so we collapsed all the human samples in a single FASTA file that virtually contains all the human promoters found in the FANTOM5 project. We ran ScanAll with the same parameters as the previous analysis obtaining the list of conserved modules.

We then wanted to define the relationship between the found modules and the peaks. To do so we performed unsupervised hierarchical clustering of the modules. This analysis is computationally intensive: the matrix is composed by ~420000 rows (modules) and ~65000 columns (peaks). We are actually trying different approaches in order to overcome this problem.

The first approach is based on the possibility to simply dividing and parallelising the process, although any tools seem to not solve our particular problem.

The second approach is based on the possibility to filter the data in order to reduce the size of the matrix and consequently the computational weight. The filtering could be done for both modules and peaks. In the case of modules, we can filter out all the modules with a low variance for the number of instances in each peak. Filtering the modules could be prone to bias because of two reasons: 1) We can't establish a priori which level of variance should be considered as a cut-off; 2) This could lead to the

exclusion of housekeeping modules (if they exist); 3) This could lead to the exclusion of very specific modules that we found only in few peaks. Despite these limitations, taking into account the fact that we want to found the most important modules that define a specific group of peaks, this approach is acceptable.

On the other hand, filtering the peaks based on the variance of their expression could be a successful approach. We know the expression level of the peaks in each sample, so excluding the peaks that do not change between samples is licit. This will also lead to the exclusion of all the modules that are found only in those specific peaks. It could also be interesting to analyse separately all the peaks that have really low change in expression levels in different samples, because of the fact that we can consider them as housekeeping transcription starting sites.
We think that the filtering approach will permit us to perform the clustering and associate groups of samples directly to the peaks and so, also to the samples.

Ongoing research: Build regulatory interaction networks for every specific human tissue

The possibility to assign regulatory modules to specific genes is even more intriguing in the prospective of creating tissue specific regulatory networks. It is possible to use directed graphs to represent regulatory interactions (Arda et al., 2010). In gene regulatory networks (GNR) regulators and targets are represented as nodes and the interaction between them as edges. These networks could be useful to explain the dynamics and organization principles of gene expression.

Our compendium already includes a section representing the interaction networks between transcription factors, where an edge indicates the presence of a module composed by the two linked TFs or Unknown motifs (represented as nodes). The direction of the edges represents the order of the TFs or motif in modules.
From preliminary data we observed that in a not negligible number of samples the hubs with higher degree are specific Unknown motifs. These motifs are likely to be not yet characterised TFBS or structural elements that play a role in gene expression regulation.
We are still working on the representation of networks including also the regulated genes.

Taking into account all the possible regulatory interactions between TFs and genes, network analysis could help to identify specific regulation circuits that are shared between samples or that are specific for certain tissues.

Ongoing research: Investigate cancer and cancer related cell lines in FANTOM5 dataset

The FANTOM5 dataset is not a cancer samples dataset; in fact, it contains also tissues and primary cell lines. This feature makes it particularly suitable for comparing cancer cells against healthy tissues. I already performed an intensive analysis comparing MSCs coming from HG-SOCs against MSCs coming from healthy tissues; besides, the dataset includes also MSCs coming from HG-SOC metastasis. I performed differential expression analysis and enlighten the presence of genes that are differentially expressed between MSCs from HG-SOC and from metastasis. Functional annotation performed using IPA indicates that the genes differentially expressed are linked to Inflammation and Immune system (Table 5).

Furthermore, upstream regulation prediction by IPA indicates progesterone as possible responsible for the specific expression pattern seen in metastasis MSCs, while inflammatory cytokines seem to revert the observed transcriptional program (Table 6).

The next step will be to integrate specific comparison between cancer related cell lines and healthy tissues with regulatory modules information coming from our Compendium, thus making possible to enlighten the regulatory component of specific expression patterns relevant for cancer processes.

Table 5

Metastasis MSCs vs HG-SOC MSCs

Canonical Pathways	-log(p-value)	Activation z-score
Granulocyte Adhesion and Diapedesis	15.80	
Agranulocyte Adhesion and Diapedesis	14.10	
Hepatic Fibrosis / Hepatic Stellate Cell Activation	11.20	
Inhibition of Matrix Metalloproteases	6.85	
Role of IL-17A in Arthritis	6.38	
Atherosclerosis Signaling	6.14	
Role of IL-17A in Psoriasis	6.02	
Cholecystokinin/Gastrin-mediated Signaling	5.79	-0.54
Role of IL-17F in Allergic Inflammatory Airway Diseases	5.33	
HMGB1 Signaling	4.89	-0.54
Role of Cytokines in Mediating Communication between Immune Cells	4.44	
TREM1 Signaling	4.15	-1.27
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	4.07	
Leukocyte Extravasation Signaling	4.05	1.29
LXR/RXR Activation	3.60	2.11
Acute Phase Response Signaling	3.28	0.30
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	3.07	
Bladder Cancer Signaling	2.96	
IL-17 Signaling	2.89	
Airway Pathology in Chronic Obstructive Pulmonary Disease	2.89	
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	2.85	-0.45
Regulation of Cellular Mechanics by Calpain Protease	2.85	2.00
Communication between Innate and Adaptive Immune Cells	2.83	
Prostanoid Biosynthesis	2.72	
IL-6 Signaling	2.63	-1.00

Table 6

Metastasis MSCs vs HG-SOC MSCs

Upstream Regulator	Log Ratio	Predicted Activation State	Activation z-score	p-value of overlap
HOXB3	3.33	Activated	2.00	3.19E-06
FBXO32	2.00	Activated	2.16	7.76E-10
NR3C1	1.96	Activated	2.15	7.57E-10
PTGS2	-2.03	Inhibited	-2.03	1.42E-08
CSF2	-2.22	Inhibited	-2.58	1.31E-09
HGF	-3.45	Inhibited	-2.21	7.05E-18
IL1B	-4.92	Inhibited	-2.37	1.08E-22
IL17F		Inhibited	-3.08	2.71E-15
progesterone		Activated	2.22	5.20E-12
peptidoglycan		Inhibited	-3.39	2.51E-11
TLR5		Inhibited	-3.08	3.48E-08
TLR3		Inhibited	-2.98	4.47E-08
TLR4		Inhibited	-2.61	1.90E-07

Using gene expression cancer datasets to define the role of genes in cancer

A variety of gene expression cancer datasets are publically available and can be investigated in order to find new biomarkers, extracting relevant genomic information and validating biological hypotheses. For large cohorts of patients clinical information is also available thus making possible to draw relationships between gene expression patterns and clinical and prognostic features of cancer. Integrated bioinformatics approaches permit to identify tumour related signatures, molecular subtypes of cancer, and biomarkers. Relevant findings can then be further investigated in cellular models in order to define precise molecular mechanisms.

In our laboratory we focus particularly on the study of serous ovarian cancer and breast cancer. During the first part of my doctorate I concentrated on the study of two genes: GTSE1 and HMGA1. Both genes (and their coded proteins) are well characterised from a molecular point of view but their role in cancer is still not clear. Using gene expression data coming from cancer patients, I tried to define the relevance of these genes for cancer related processes, for the prognosis of the disease and for their use as a molecular marker.

HMGA1 expression in primary breast tumours

The HMGA gene family plays important roles in proliferation, differentiation and stem cell self-renewal (Shah and Resar, 2012). Generally, the expression of HMGA genes is restricted to embryogenesis and, with few exceptions, is very low in normal adult cells (Sgarra et al., 2004). However, in transformed cells, HMGA genes are expressed at high levels, representing a possible feature of human malignancies. Several studies have reported that HMGA1 expression is high in a variety of human cancers, including carcinomas derived from prostate, colon and breast tissues (Fusco and Fedele, 2007). In recent years, studies have demonstrated a causal role of the HMGA1 protein in promoting a transformed phenotype (Berlingieri et al., 2002; Dolde et al., 2002; Liau et al., 2006; Reeves et al., 2001; Takaha et al., 2004; Wood et al., 2000) and the presence of the HMGA1 protein has been correlated with a higher grade in mammary epithelial cancer (Chiappetta et al., 2004; Ram et al., 1993). These results suggest that HMGA1 may be a key player in cancer related processes, in particular in breast cancer. It has already been shown that the HMGA1 gene is overexpressed in 60% of sporadic ductal carcinomas (Chiappetta et al., 2004), but it is

still unclear whether HMGA1 is enriched in a particular molecular subtype.

To enlighten the importance of HMGA1 in breast cancer, I compared the abundance of HMGA1 mRNA with clinical variables such as tumour subtype and grade, which are important indicators of breast cancer prognosis (Perou et al., 2000; Sørlie et al., 2001), by analysing a public primary breast cancer microarray meta-dataset (1881 samples). The analysis revealed that HMGA1 mRNA levels are higher in the basal-like and HER2+ subtypes respect to the luminal A, luminal B and normal-like subtypes (Figure 14A). I also found a strong association between HMGA1 expression and the absence of the oestrogen receptor (Figure 14B). In fact, both the basal-like and HER2+ subtypes are oestrogen receptor-negative breast cancer subtypes and, comparing directly tumours positive for the receptor against tumours negative for the receptor, the HMGA1 gene is significantly more expressed in the latter group. Finally, HMGA1 mRNA expression was associated with tumour grade (Figure 14C); tumours with a higher HMGA1 expression exhibit a higher grade.

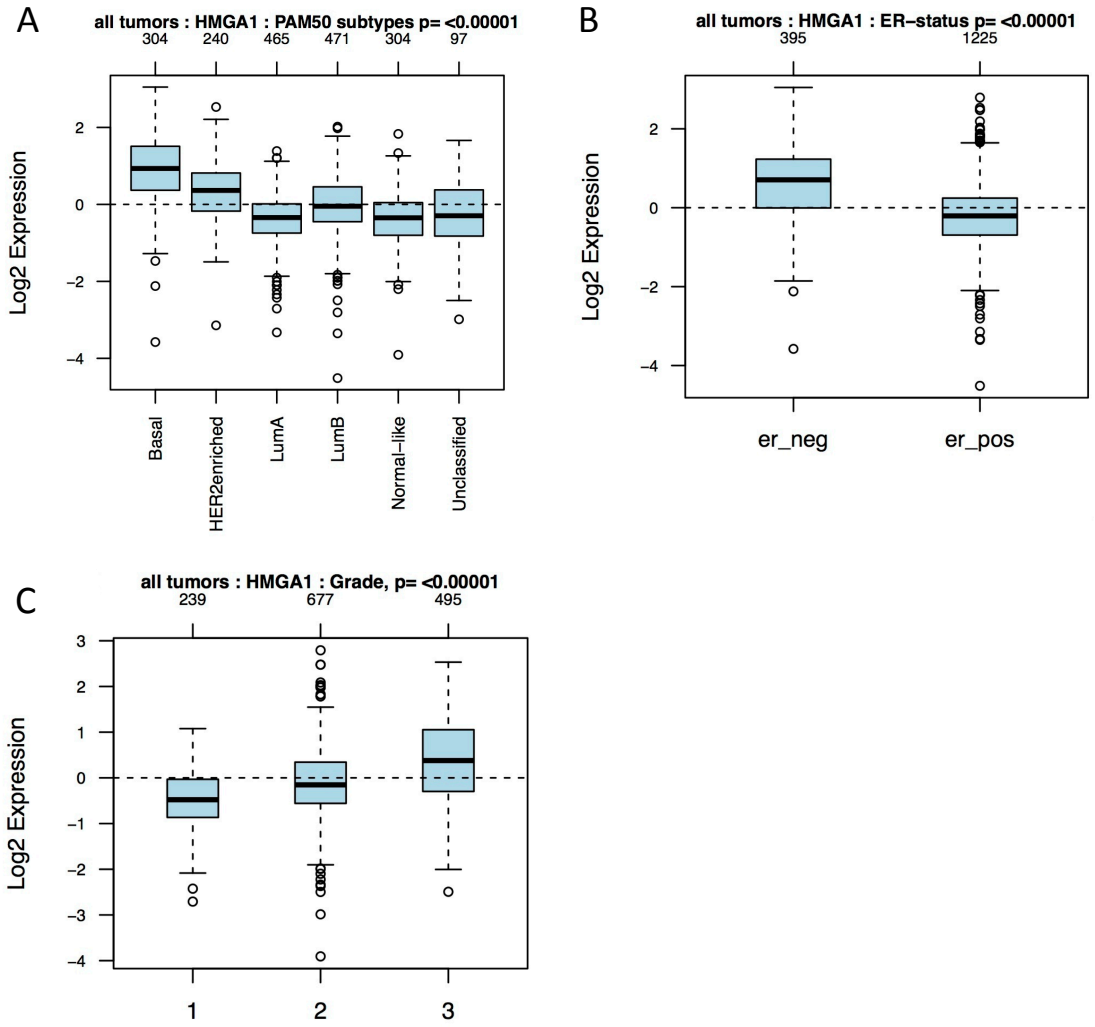

Figure 14: HMGA1 expression in breast cancer. The level of expression of HMGA1 gene is shown across breast cancer subtypes (A), ER-status (B) and grades (C).

The HMGA1 gene signature is an independent predictor of poor clinical outcome

In order to understand the functional involvement of HMGA1 in breast cancer malignancy I investigated how HMGA1 alters the transcriptional program in breast cancer cell lines by analysing the transcriptional profile of breast cancer cells in the presence and absence of HMGA1. For this analysis a cellular model of an inducible cellular system for HMGA1 silencing based on siRNA in the oestrogen receptor-negative basal-like human breast cancer cell line MDA-MB-231 was used.

Performing unsupervised hierarchical clustering of gene expression data, I identified two gene cluster of genes differentially expressed after HMGA1 silencing (Figure 15): the first cluster contains the genes that were most up-regulated (siHMGA1 UP

genes, n=38); the second cluster is larger and contains the genes that were most down-regulated (siHMGA1 DW genes, n=130) (see Supporting Material Table S4 and S5) for complete lists of differential expressed genes). This finding is consistent with notion that HMGA1 acts predominantly as a transcription activator. The effects of HMGA1 depletion on the expression of selected genes were confirmed by qRT- PCR (see Supporting Material Figure S1). Functional annotation analysis of the genes regulated by HMGA1 silencing, performed with the DAVID/EASE and Ingenuity Pathway Analysis (IPA) tools, led to the conclusion that HMGA1 silencing affects genes involved in the regulation of the cell cycle, cellular movement, growth, proliferation, metabolism and cancer (Table 7).

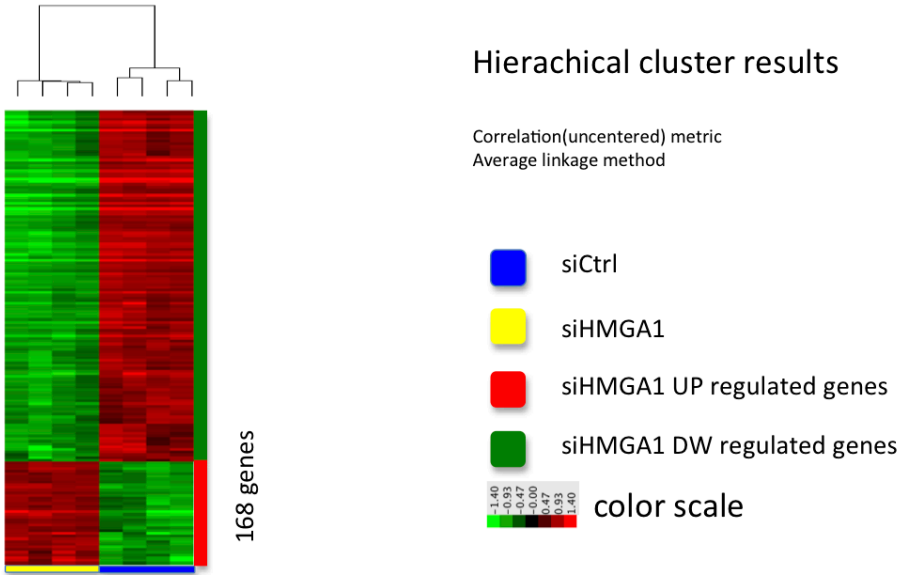

Figure 15: Microarray analysis of MDA-MB-231 breast cancer cell after HMGA1 depletion. Using the Affymetrix microarray technology, I characterized the gene expression profiles of human MDA-MB-231 cell lines with HMGA1-depleted and control cells. The results of the cluster analysis for the genes obtained from differential genes analysis are shown.

DOWN regulated genes upon siHMGA1 Functional Analysis	
DAVID	p-value
cell cycle	8.49E-12
M phase	6.40E-09
condensed chromosome	1.57E-06
nuclear lumen	2.10E-06
nuclear division	1.23E-05
mitosis	1.23E-05
Ingenuity Pathway Analysis	p-value
Cancer	4.56E-08 - 3.84E-02
Reproductive System Disease	2.25E-04 - 3.84E-02
Cell Cycle	3.48E-13 - 3.68E-02
DNA replication and Repair	1.45E-08 - 3.30E-02
Embryonic Development	2.25E-05 - 3.84E-02
Hematological System Development	3.04E-04 - 3.30E-02

UP regulated genes upon siHMGA1 Functional Analysis	
DAVID	p-value
regulation of kinase activity	0.031
domain:Cadherin 7	0.034
regulation of transferase activity	0.034
FAD-dependent pyridine nucleotidedisulphide oxidoreductase	0.038
EGF	0.043
Ingenuity Pathway Analysis	p-value
Cancer	3.92E-06 - 1.73E-02
Gastrointestinal Disease	3.92E-06 - 1.73E-02
Cellular Movement	3.32E-05 - 1.73E-02
Cellular Growth and Proliferation	4.88E-05 - 1.73E-02
Lipid Metabolism	1.71E-04 - 1.73E-02
Cell-Mediated Immune Response	3.32E-05 - 1.16E-02

Table 7: Functional analysis of genes regulated by HMGA1. Starting from the genes that were expressed in the silenced HMGA1 cells that had greater than a 1.4 log-fold change or lower than a 1.4 log-fold change with respect to the control cells, I used the publicly accessible software DAVID/EASE and Ingenuity Pathway Analysis. The most over-represented terms ($p<10^{-5}$) in the down- regulated gene cluster in the silenced HMGA1 cells were related to the mitotic cell cycle and mitosis (upper panel), while the up-regulated gene cluster was characterised by GO and was related to metabolism, movement and proliferation (lower panel).

In order to address the hypothesis that the transcriptional programme induced by HMGA1 is important for tumour aggressiveness, I investigated the expression of the HMGA1 gene signature in different cancer datasets using the Oncomine web tool (www.oncomine.org)(Rhodes and Chinnaiyan, 2004; Rhodes et al., 2007). Interestingly, the results of this analysis clearly revealed higher expression in tumour tissues respect to the normal tissue (ratio 89/6) of the genes that were down-regulated after HMGA1 silencing (i.e., genes induced by HMGA1). In particular, these genes are highly expressed also comparing bad vs. good clinical outcome (ratio 55/4), especially for breast cancer (ratio 31/0) (Figure 16). Therefore, to further evaluate this clinical correlation, I analysed several breast cancer microarray datasets, which collectively consisted of more than 2000 patients. A Kaplan-Meier survival analysis showed that the expression of these genes was significantly correlated with clinical outcome. In particular, patients expressing high levels of these genes displayed a shorter time to distant metastasis (TDM) (Figure 17A). In addition, a higher HMGA1 gene signature expression was associated with basal-like subtype and high-grade (G3) breast cancers (Figure 17B and C). Notably, the signature was also correlated with HMGA1 mRNA expression (Figure 18) supporting the idea that HMGA1 is able to control this specific transcriptional program both in cancer cell lines and *in vivo*.

Cancer Type	Cancer vs. Normal		Clinical Outcome	
	Over-expression or Copy Gain	Under-expression or Copy Loss	Over-expression or Copy Gain	Under-expression or Copy Loss
Bladder Cancer	2		1	
Brain and CNS Cancer	5	1	4	
Breast Cancer	8	1	31	
Cervical Cancer	2			
Colorectal Cancer	14		2	1
Esophageal Cancer	3			
Gastric Cancer	4			
Head and Neck Cancer	10			
Kidney Cancer			2	
Leukemia	1	3		
Liver Cancer	4			
Lung Cancer	13		4	
Lymphoma			3	3
Melanoma	1			
Myeloma			5	
Other Cancer	5	1	2	
Ovarian Cancer	4			
Pancreatic Cancer	2			
Prostate Cancer	2		1	
Sarcoma	10			
Significant Unique Concepts	89	6	55	4
Ratio	14.83 (p<0.0007)		13.75(p<0.009)	

Figure 16: Oncomine analysis. Starting with the genes expressed in the silenced HMGA1 cells that displayed less than a 1.4 log-fold change with respect to the control cells (HMGA1 signature), I used the Oncomine web tool to determine if there were any associations between the gene expression profiles of the different cancer types present in the database. This table displays the number of significant results, coloured red or blue for over- or under-expression, respectively, across all cancer types, with an analysis of the correlation with clinical outcomes. P-values were calculated using a two-sample paired Wilcoxon test.

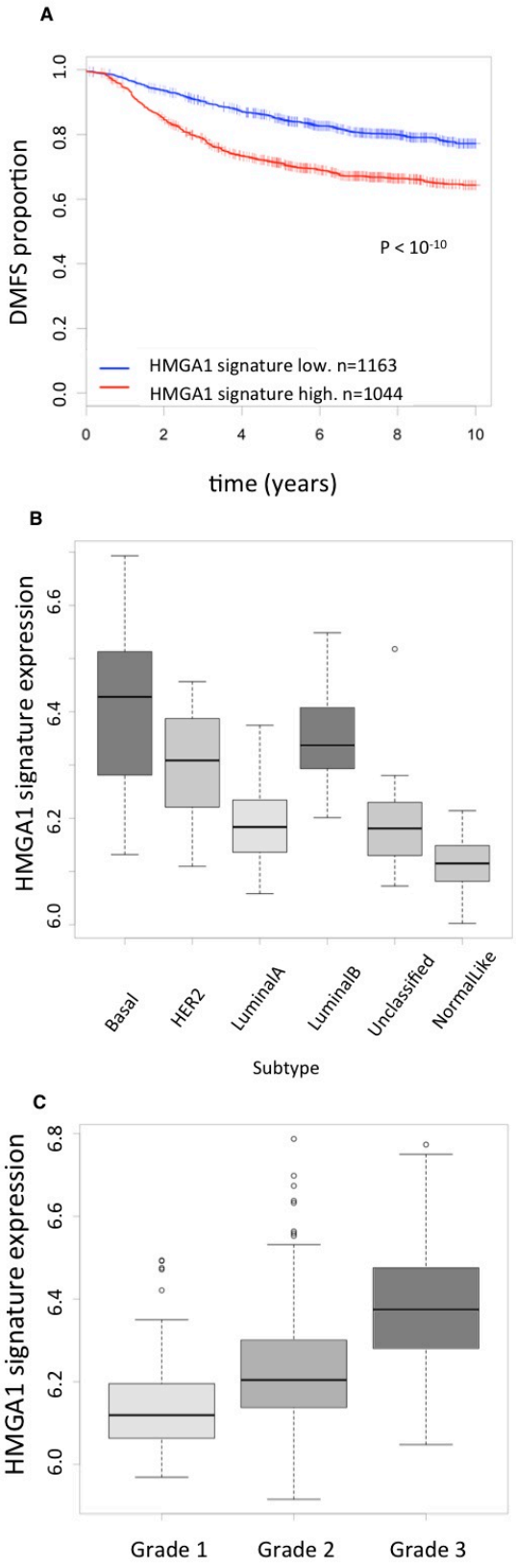

Figure 17: HMGA1 signature in breast cancer. (A) Kaplan–Meier survival curve of time to distant metastasis (TDM) for breast cancer patients who were classified according to HMGA1 expression. Red line: cases with high HMGA1 expression; blue line: cases with low HMGA1 expression. (B-C) Boxplots of the distribution of the gene expression intensities of the HMGA1 gene signature across different breast cancer subtypes (intrinsic subtypes or Grades 1, 2 or 3).

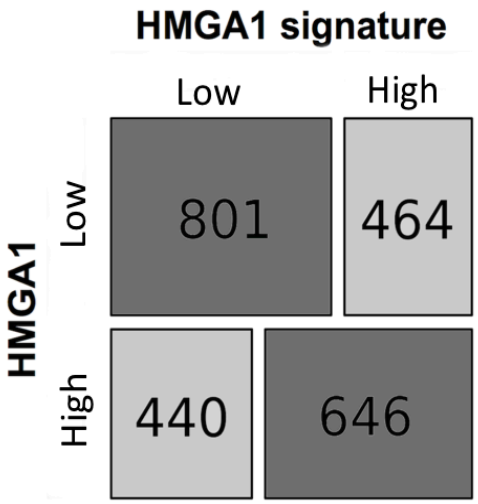

Figure 18: Correlation between HMGA1 and HMGA1-gene signature expression. We classified all of the breast cancer samples using the expressions of HMGA1 or corresponding signature obtained by our microarray experiments to evaluate their association. The mosaic plot shows the proportion of the four possible groups along with the number of samples in each group. Statistical analysis (using the Pearson's Chi-squared test) showed a significant correlation ($p< 2.2e^{-16}$).

I then assessed whether this signature could be an independent predictor of clinical outcome. Cox multivariate analysis revealed that the HMGA1 gene signature behaves as a significant ($p<0.05$) independent prognostic factor (Table 8). Then I repeated the analysis in a cohort of 115 patients in which many clinical variables were available. Also in this case I confirmed that the HMGA1 signature yields prognostic information (Table 9). Hence, the combined expression of the genes up-regulated by HMGA1 has prognostic significance and may be considered as a marker of breast cancer malignancy.

Clinical variable	HR	se(coef)	z	p-value
ER status (negative)	1.457	0.215	1.749	0.015
Linfonode (negative)	0.624	0.197	-2.394	0.0011
Size	3.447	0.25	4.945	2^{-09}
Age	1.034	0.104	0.322	0.57
HMGA1 signature	1.574	0.195	2.326	0.019

n= 586

Table 8: Multivariate analysis of risk of death.

Clinical variable	HR	se(coef)	z	p
ER	4.56	0.646	2.34912	0.019
PR	0.89	0.674	-0.17372	0.86
HER2	0.999	0.534	-0.00196	1
Grade	2.363	0.544	1.57972	0.11
Age	1.04	0.329	0.11829	0.91
LN	0.489	0.632	-1.13249	0.26
Size	1.516	0.225	1.85226	0.064
HMGA1 signature	0.295	0.606	-2.01266	0.044

n= 115

Table 9: Multivariate analysis of risk of death.

HMGA1 final remarks

The bioinformatics analysis was integrated with biological experiments in order to demonstrate that HMGA1 plays a pivotal role in regulating invasive processes and determining poor prognostic outcomes in breast cancer by sustaining the mesenchymal phenotype and stemness. Taken together, these data suggest that HMGA1 plays a key role in breast cancer malignancy and the progression of metastatic disease acting as an activator of a specific gene network. All results have already been published (Pegoraro et al., 2013).

GTSE1 Expression in Breast Cancers Correlates with Time to Metastasis, Invasiveness and Clinical Outcome

The protein GTSE1 (G-2 and S-phase expressed 1) is a negative regulator of p53 that can shuttle between the cytoplasm and nucleus. After DNA damage, GTSE1 accumulates in the nucleus, where it interacts with p53 and shuttles it out of the nucleus to promote its down-regulation and recovery from the p53-induced G2 DNA damage checkpoint (Liu et al., 2010; Monte et al., 2003, 2004). In the absence of DNA damage, GTSE1 localizes to interphase microtubules networks (Collavin et al., 2000; Monte et al., 2000), and has also been found associated with clathrin-containing complexes (Borner et al., 2012; Hubner et al., 2010). Interestingly GTSE1 resulted regulated by HMGA1 and, in our laboratory, it has already been observed that GTSE1 is overexpressed in several transformed cell lines with respect to non-transformed cell lines (see Supporting Material Figure S2): these observations led to the hypothesis that GTSE1 could act as an effector, performing specific molecular cancer-related functions regulated by HMGA1 transcriptional program.

In order to identify potential clinical or cancer-related correlations with GTSE1 expression, I looked for associations with any type of tumour using the Oncomine cancer microarray database (www.oncomine.org) (Rhodes et al., 2007) (Figure 19). This analysis identified 61 unique datasets where GTSE1 had significantly higher expression in tumour tissues as compared to normal tissues. Notably, worse clinical outcome was associated with increased GTSE1 expression in 12 cases, 10 of which were in breast cancer.

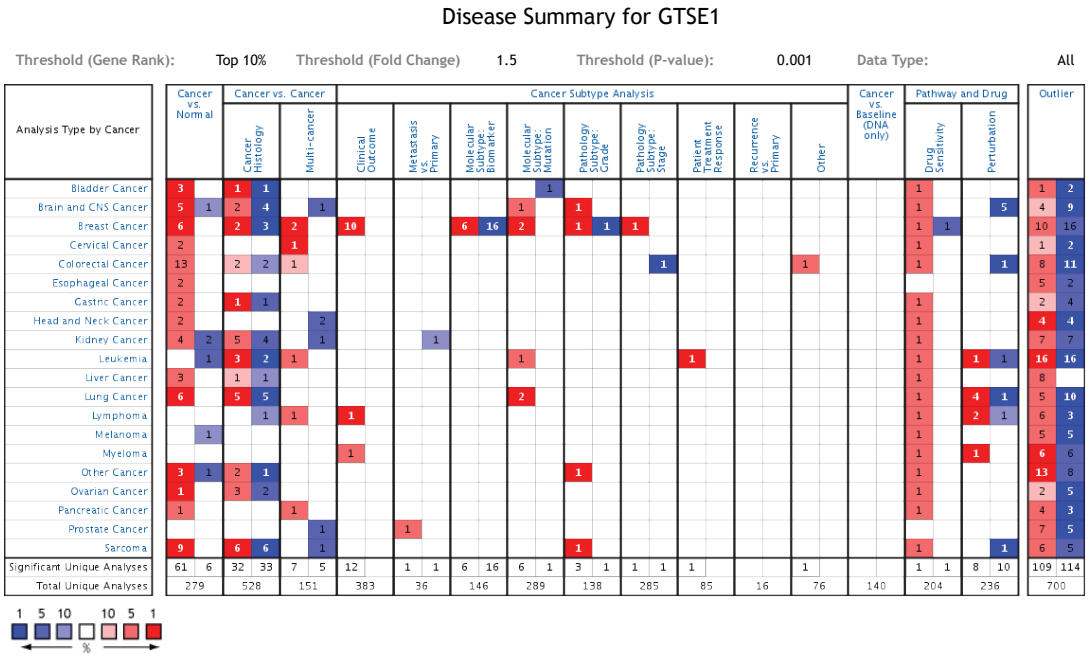

Figure 19: Oncomine analysis for GTSE1 Disease Summary: this view displays the number of significant results colored in red or blue for over- or under-expression, respectively, across all cancer types and analysis types in Oncomine.

I then analysed several microarray data sets of breast cancer, collectively consisting of more than 2000 patients in order to delve deeper into clinical correlations with GTSE1 expression. Kaplan-Meier survival analysis of the combined data sets showed that breast cancer patients expressing higher GTSE1 levels in tumours displayed shorter time to distant metastasis ($p<10^{-15}$; Figure 20A). Notably, the correlation between GTSE1 expression and the grade of breast cancers is significant, with the most invasive and aggressive cancers (Grade 3) showing highest expression of GTSE1 (Figure 20B). Taking all together, these data show a correlation between the deregulation and overexpression of GTSE1 and tumour invasiveness and cancer prognosis.

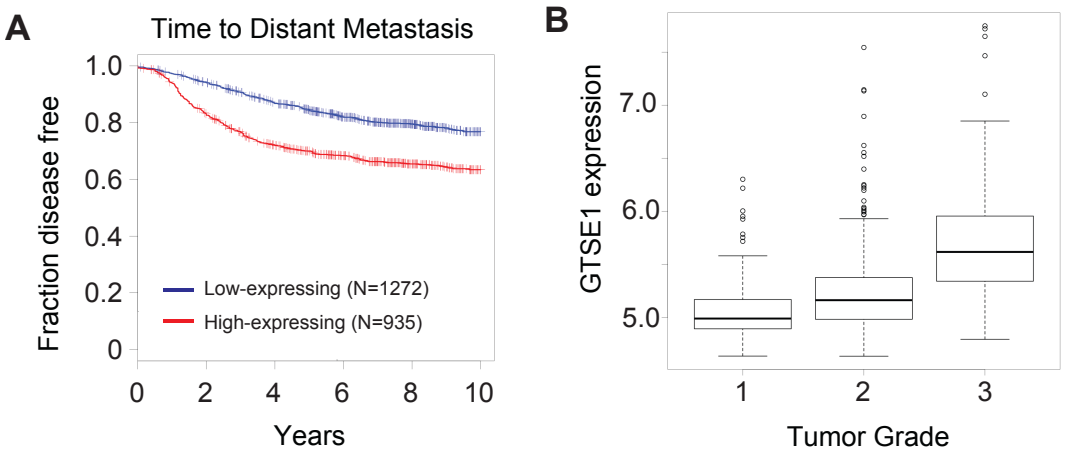

Figure 20: GTSE1 expression in breast cancer tumours and cells correlates with time to metastasis and invasiveness. (A) Kaplan–Meier survival curve of time to distant metastasis of breast cancer patients classified according to the expression of GTSE1. Red line: cases with high expression of GTSE1, blue line: cases with low expression of GTSE1. ($p<10^{-15}$) (B) Boxplots of the distribution of gene expression intensities of GTSE1 across different breast cancer subtypes (Grade 1, 2 or 3; $p<10^{-5}$; linear regression analysis).

GTSE1 final remarks

In order to delve deeper the functional role of GTSE1, in our laboratory we performed gene expression profiling silencing GTSE1 in TNBC-cell lines. Functional analysis confirmed that the perturbed functions include cell adhesion, cell junction and endocytosis biological themes. This analysis also led to the selection of potential molecules that can target specific genes modulated by GTSE1 silencing.

The bioinformatics analysis was integrated with biological experiments demonstrating that the molecular activity of GTSE1 leading to stimulation of cell migration and loss of focal adhesions is EB1-dependent microtubule plus-end tracking, providing an intriguing link between microtubule plus-end functions and metastasis. All results have been published (Scolz et al., 2012).

DISCUSSION

Nowadays, different cancer gene expression datasets are publically available and can be interrogated with the aim of generating new biological hypothesis or validating the existing ones. Usually genes are characterized as onco-genes or onco-supressors. Such definitions are restricting and do not give information about how the gene performs its pro- or anti- tumour activity. Even through the mechanisms and pathways exploited by cancer cells in order to elude apoptosis and cell cycle control are the same, each tumour subtype relies on a different set of alterations. Because of this, often, gene signatures have clinical relevance only in some tumour types or subtypes but not in others. In the case of HMGA1, it is known that the HMGA gene family plays important roles in proliferation, differentiation and stem cell self-renewal and that in transformed cells, HMGA genes are expressed at high levels, thus indicating a possible role in cancer. Using bioinformatics approaches I demonstrated that HMGA1 is highly expressed in breast cancers with poor prognosis and with the tendency to metastasize. Interestingly, the HMGA1 signature overlaps with other signatures that identify patients with poor prognosis; in fact, some genes in the signature (CENPF, CENPA, CCNE2, BUB1 and PSMD2) are part of the 70-gene prognosis profile (van 't Veer et al., 2002) and Pin1/mutant p53 signature (Girardini et al., 2011).

The microarray analysis indicates that HMGA1 is able to regulate genes implicated in microtubules dynamics, and among them, GTSE1 has emerged as a microtubule-associated protein. My bioinformatics analyses indicate both GTSE1 and HMGA1 as genes implicated in cancer related processes and having relevance for prognosis; the former acting as regulator of a general transcription program, the latter as effector performing specific molecular functions.

Importantly, the molecular activity of GTSE1 leading to stimulation of cell migration and loss of focal adhesions is EB1-dependent microtubule plus-end tracking: this result provides an intriguing link between microtubule plus-end functions and metastasis. We can suppose that this mechanism is not specific of breast cancer, but shared between cancer types with tendency to metastasize. In fact, upregulation of GTSE1 expression was identified as a potential marker for metastasis in oral tongue squamous cell carcinoma (Zhou et al., 2006) and more recently, GTSE1 was identified as one of three cell cycle regulatory genes (along with CDKN3 and Cyclin

B1) whose upregulation in gastro-enteropancreatic neuroendocrine tumours correlate with metastasis (Lee et al., 2012). These results are consistent with the observation that GTSE1 mRNA expression levels correlate with time to metastasis and tumour grade in breast cancer.
The question regarding how HMGA1 can regulate GTSE1 is still unresolved, but we are actually collecting multiple evidences that GTSE1 is regulated by TEAD, a component in the Hippo pathway, playing essential roles in mediating biological functions of YAP (Zhao et al., 2008). Intriguingly, HMGA1 is able to regulate YAP localization (Pegoraro et al., in preparation) thus permitting to hypothesize a novel axis capable of modulate breast cancer aggressiveness. The Integration of different bioinformatics and molecular approaches lead to the characterization of this axis, identifying HMGA1 as the main regulator of a complex transcriptional program and GTSE1 as one of the down-stream effectors.

Even though cancer tissues gene expression profiles are widely used to define genes signatures and test their clinical relevance, they do not take into account to the fact that tumours are composed by heterogeneous tissue; cancer cells interact with the cancer microenvironment in order to orchestrate tumour growth, drug resistance and metastasis. Focusing on the cancer supporting cell compartment could enlighten specific molecular mechanisms and biomarkers, also relevant for prognosis and clinical outcome; actually, the dissection and characterization of the distinct cellular lineages and their respective progenitors giving rise to the various cell types that form the tumour-tissue could be addressed only through in vitro models.
One of the cellular lineages that compose both normal and cancer tissue microenvironments are Mesenchymal Stem/Stromal Cells (MSCs). In our laboratory, MSCs from High-Grade Serous Ovarian Cancers (HG-SOCs) and various normal tissues (N-MSCs) have been isolated and their transcriptional activity have been analysed with respect to the large comprehensive FANTOM5 sample dataset.

The FANTOM5 project used single molecule CAGE sequencing to generate a promoter-level expression atlas displaying the transcriptional regulatory networks that define the majority of mammalian cells and tissues.
Taking advantage of the deep-CAGE-seq data it has been possible to show that HG-SOC-MSCs possess distinct and specific CAGE-peak promoter activities with respect

to N-MSCs. Moreover, looking at the distribution of all CAGE-peaks across the entire gene length, the differential analysis suggested the possible existence of a specific epigenetic-based mechanism explaining such global CAGE-peak patterns. Altogether, the highlighted shared functional programs between HG-SOC-MSCs and N-MSCs suggest that the global differences are based on quantitative levels of transcriptional output rather than on qualitative differences in the expressed genes. Taking into account the fact that HG-SOC-MSCs are not transformed cells, this results enlighten a particular plasticity of MSCs from a transcriptional point of view, possibly in response to external stimuli coming from tumour microenvironment. The existence of specific identities in MSCs has been recently described in murine MSC-like cells derived from different tissues (bone-marrow, heart and kidney)(Pelekanos et al., 2012). It has been observed that bone marrow-derived MSC-like cells cannot replace MSC-like cells of pro-epicardial origin in myocardial infarction (Chong et al., 2011) and MSCs originating from different sources show a different healing performance in cardiac regeneration (Gaebel et al., 2011), so such specific heterogeneity seems to be critical at the functional level.

Both differential gene expression and differential gene peak usage analysis results suggested that HG-SOC-MSCs are close relatives of mesothelial cells, smooth muscle cells and fibroblasts. This result could be related to the district of origin of these cells. It has been recently shown (Rinkevich et al., 2012b) that the mesothelium derived from various adult mouse organs contains the precursors of Fibroblasts and Smooth Muscle Cells (FSMCs). A possible hypothesis is that HG-SOC-MSCs could derive from the ovarian mesothelium or other local mesothelia. Similarly, all the accumulated evidences (Auersperg, 2013; Berek et al., 2012; Karst et al., 2011; Kim et al., 2012; Levanon et al., 2008) support the mesothelia of ovarian surface epithelium (OSE) and fallopian tube fimbriae as the site of origin of the cancer cell compartment of high-grade serous ovarian, peritoneal, and fallopian tube cancers. An alternative to the above-proposed hypothesis, the HG-SOC microenvironment could instruct, via epigenetic mechanisms, the conditioning of recruited bone marrow MSCs to acquire the overall mesothelial signatures common to their specific origin. Future transcriptomics analysis of mesothelial-derived cells obtained from different tissue districts will uncover their complex biology and heterogeneity, shedding light on the potential role they might play in the microenvironment of HG-SOC and other aggressive tumours.

Noticeably we found CAGE-peak differences also between the various normal tissue-derived MSCs: heart-derived MSCs showed a similar overall mesothelial signature as found in HG-SOC-MSCs with respect to adipose and bone marrow-derived N-MSCs. In order to address a potential role of HG-SOC-MSCs in serous ovarian cancer, I derived a gene signature composed of nine genes. These genes consist of five established mesothelial markers and four mesothelial-related genes. This signature showed a statistically significant correlation with bad prognosis when used to interrogate a large SOC meta-dataset. Given that its performance was either statistically non-significant in the case of lung cancer or correlated with good prognosis in the case of breast cancer we can hypothesise that such correlation is specific for HG-SOC. This observation supports the hypothesis that high-grade serous ovarian cancer derives from mesothelium of ovarian surface epithelium (OSE) and fallopian tube fimbriae.

For the characterization of HG-SOC-MSCs it had been essential the use of the FANTOM5 dataset: in fact, one of the main features of the dataset is that it includes samples coming from human tissues, primary cell lines and transformed cell lines thus allowing the comparison of the cell lines of interest, not only respect to all the other cell lines, but also respect to the tissues of the human body. Furthermore, since the FANTOM5 dataset is focused on promoter utilization, it represents an excellent source of information for the study of gene expression regulation mechanisms.
I and my group wondered if the identification of enriched composite elements in the genomic regions surrounding FANTOM5 CAGE peaks could enhance the ability to understand the regulatory networks in different tissues.
Our new software, ScanAll, tries to get rid of the limits associated with phylogenetic and statistical approaches (Simcha et al., 2012) (such as computational time, input sequences size limitations, and use of inappropriate background), some of which depending on the fact of relying on existing knowledge.
In fact, available collections of transcription factor binding sites account for some hundreds of non-redundant profiles, with respect to an estimated number of ~1700 different human TFs (Vaquerizas et al., 2009; Wingender et al., 2013). A single DNA-binding domain shared by multiple TFs could allow the interaction of several proteins to very similar DNA sequences; however, splicing variants of a single TF could have very different binding specificities (Giguère et al., 1995). It is hence likely that

existing catalogues of binding profiles provide an underrepresentation of the total number of binding sites in the genome.

Nowadays, an increasing number of high quality binding site profiles are added on a regular basis to TFBS collections. However, methods based on PWMs lead to the risk of underestimating the importance of less studied transcription factors (Pavesi et al., 2004; Sandve et al., 2007; Tompa et al., 2005). Moreover, even for methods not using PWMs, the risk of inaccurate predictions (Pavesi et al., 2004) is high when the significance of an analysis involves motif frequencies calculated by only taking into account known regulatory regions, as a background. Finally, there are methods that only predict a single enriched motif per set of sequences (Pavesi et al., 2004; Sandve et al., 2007; Tompa et al., 2005), that account for up to one single occurrence of any enriched motif per sequence, or that do not permit to analyse subsets of the input datasets.

The possibility provided by ScanAll to look for structured enriched modules, instead of single motifs, permits to getting rid of the aforementioned restrictions. The introduction of a dedicated filtering pipeline based on the comparison of our results against random sequences greatly reduced the number of false positive results.

The choice of motif length ℓ (set to six) reflected the average length of the conserved TFBS core sequences; in combination with the number of mismatches d (set to one) they constitute one of the motif layouts commonly searched by existing methods (Narasimhan et al., 2003; Pavesi et al., 2007). By looking at the promoter coverage obtained by mapping the module-composing enriched elements, we could find that instead of providing a completely random painting of the analysed region, they specifically localized at precise spots, thus confirming the topological specificity of our findings in identifying binding regions. These results are obtained by introducing two concepts: the notion of *complexity* and that of *quorum*. Complexity is the number of different nucleotides required to appear in every motif and it was fixed to four in order to maximize the structural diversity of enriched sequences (and hence the probability to identify biologically relevant elements). The choice of quorum during motifs discovery was arranged in order to correspond (on average) to 10% of the estimated number of genomic sequences (i.e. active promoters) in each FANTOM5 sample. The values we used for these parameters, although not arbitrary, could sometimes lead to evident drawbacks, like the loss of known TFBS with simple

structures (i.e. sites for SP1, SP3, NHLH or ZFX) or the absence of seldom-represented modules occurring only in very few promoters.

Spacing between motifs forming CRMs varies depending on the involved TFs and on the face of the DNA double helix where the interactions occur. Recent evidence suggest that the lower bound is generally between 25 and 50 nucleotides (Huang et al., 2012), a value frequently used also by composite motifs discovery methods (Frith et al., 2001, 2003; Kel et al., 2006; Zhou and Wong, 2004). Conversely, the upper bound still preserving the possibility for two closely positioned TFs to interact synergistically is estimated to be around 100 nucleotides (Courey, 2008; Girgis and Ovcharenko, 2012). For these reasons, we set our pipeline to discover enriched modules made of motifs separated by 40-90 nucleotides. We wanted them to be close enough to ensure co-binding between putative interacting TFs, while also allowing a sufficient interval enabling the combinations of TFBS for regulatory proteins not directly interacting with each other, as happening in large transcriptional complexes.

All these parameters were chosen based on specific considerations, but they are however all completely user-definable and can hence be adapted to different analytical conditions.

In order to get a broader representation of the full biological potential of these modules, we evaluated 1) the expression status of all the TFs that putatively bound the discovered enriched elements and 2) their ability to establish direct protein-protein interactions (PPI). This led us to the creation of the "Expressed module Compendium" and "PPI Expressed Compendium". These collections allowed us to have direct evidence of the sample-specific switches taking place between the expressed TFs: depending on the cellular context, in fact, they selectively associated with different binding partners.

Based on these results, our Compendium represents a useful source of information in order to associate the presence of enriched composite regulatory modules, the expression of TFs to human tissues and known direct protein interactions between TFs to specific human tissues.

This information could be of particular relevance is cancer, which has been described as a disease of disrupted gene regulation: it is therefore unsurprising that non-coding variations are linked to tumorigenesis. A possible challenge is to predict whether the alteration in a specific *cis*-regulatory element is likely to have a functional consequence. Existing predictive approaches can involve multiple lines of evidence:

TF–DNA interactions arise from the interplay between DNA sequence motifs, chromatin accessibility, epigenetic marks, and interactions with cofactors. In order to prioritize functional regulatory elements, multiple aspects of TF–DNA interactions have to be taken into account and the integration of different types of information is needed (Mathelier et al., 2015).

This integrative approach makes also feasible the perspective of creating tissue specific regulatory networks and analyse them in order to find specific or shared regulatory circuits. Recently, network analysis has been proposed in drug discovery studies (Huang et al., 2014; Ryall and Tan, 2015; Tang et al., 2013) where, because of the failure of single targets to successfully translate into clinical practice and the problem of development of drug resistance with single target cancer therapies, the interest in discovery of effective drug combinations has increased. The combination therapies may dramatically improve efficacy of cancer therapies and systems biology approaches are needed to prioritize combinations for experimental testing.

BIBLIOGRAPHY

Van ’t Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536.

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., and Clarke, M.F. (2003). Prospective identification of tumorigenic breast cancer cells. Proc. Natl. Acad. Sci. U. S. A. *100*, 3983–3988.

Alitalo, K., Ramsay, G., Bishop, J.M., Pfeifer, S.O., Colby, W.W., and Levinson, A.D. (1983). Identification of nuclear proteins encoded by viral and cellular myc oncogenes. Nature *306*, 274–277.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Andersen, A.A., and Panning, B. (2003). Epigenetic gene regulation by noncoding RNAs. Curr. Opin. Cell Biol. *15*, 281–289.

Arda, H.E., Taubert, S., MacNeil, L.T., Conine, C.C., Tsuda, B., Van Gilst, M., Sequerra, R., Doucette-Stamm, L., Yamamoto, K.R., and Walhout, A.J.M. (2010). Functional modularity of nuclear hormone receptors in a Caenorhabditis elegans metabolic gene regulatory network. Mol. Syst. Biol. *6*, 367.

Arnold, P., Erb, I., Pachkov, M., Molina, N., and van Nimwegen, E. (2012). MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics *28*, 487–494.

Auersperg, N. (2013). The origin of ovarian cancers--hypotheses and controversies. Front. Biosci. (Schol. Ed). *5*, 709–719.

Auersperg, N., Wong, A.S., Choi, K.C., Kang, S.K., and Leung, P.C. (2001). Ovarian surface epithelium: biology, endocrinology, and pathology. Endocr. Rev. *22*, 255–288.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Bajic, V.B., Seah, S.H., Chong, A., Krishnan, S.P.T., Koh, J.L.Y., and Brusic, V. (2003). Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. J Mol Graph Model *21*, 323–332.

Bajic, V.B., Tan, S.L., Suzuki, Y., and Sugano, S. (2004). Promoter prediction analysis on the whole human genome. Nat Biotechnol *22*, 1467–1473.

Barberis, M.C., Faleri, M., Veronese, S., Casadio, C., and Viale, G. Calretinin. A selective marker of normal and neoplastic mesothelial cells in serous effusions. Acta Cytol. *41*, 1757–1761.

De Bari, C., Dell'Accio, F., Tylzanowski, P., and Luyten, F.P. (2001). Multipotent mesenchymal stem cells from adult human synovial membrane. Arthritis Rheum. *44*, 1928–1942.

Beltrami, A.P., Cesselli, D., Bergamin, N., Marcon, P., Rigo, S., Puppato, E., D'Aurizio, F., Verardo, R., Piazza, S., Pignatelli, A., et al. (2007). Multipotent cells can be generated in vitro from several adult human organs (heart, liver, and bone marrow). Blood *110*, 3438–3446.

Bembom, O. (2013). Sequence logos for DNA sequence alignments. R Packag. Version 1.16.0 1–5.

Berek, J.S., Crum, C., and Friedlander, M. (2012). Cancer of the ovary, fallopian tube, and peritoneum. Int. J. Gynaecol. Obstet. *119 Suppl* , S118–S129.

Berlingieri, M.T., Pierantoni, G.M., Giancotti, V., Santoro, M., and Fusco, A. (2002). Thyroid cell transformation requires the expression of the HMGA1 proteins. Oncogene *21*, 2971–2980.

Bertucci, F., Finetti, P., and Birnbaum, D. (2012). Basal breast cancer: a complex and deadly molecular subtype. Curr. Mol. Med. *12*, 96–110.

Bidlingmaier, S., He, J., Wang, Y., An, F., Feng, J., Barbone, D., Gao, D., Franc, B., Broaddus, V.C., and Liu, B. (2009). Identification of MCAM/CD146 as the target antigen of a human monoclonal antibody that recognizes both epithelioid and sarcomatoid types of mesothelioma. Cancer Res. *69*, 1570–1577.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093.

Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganière, J., Lefèbvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., et al. (2006). Genome-wide

computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res *16*, 656–668.

Borner, G.H.H., Antrobus, R., Hirst, J., Bhumbra, G.S., Kozik, P., Jackson, L.P., Sahlender, D.A., and Robinson, M.S. (2012). Multivariate proteomic profiling identifies novel accessory proteins of coated vesicles. J. Cell Biol. *197*, 141–160.

Bourkoula, E., Mangoni, D., Ius, T., Pucer, A., Isola, M., Musiello, D., Marzinotto, S., Toffoletto, B., Sorrentino, M., Palma, A., et al. (2014). Glioma-associated stem cells: a novel class of tumor-supporting cells able to predict prognosis of human low-grade gliomas. Stem Cells *32*, 1239–1253.

Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.H., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res *18*, 1752–1762.

Brosius, J. (2003). The contribution of RNAs and retroposition to evolutionary novelties. Genetica *118*, 99–116.

Bryne, J.C., Valen, E., Tang, M.-H.E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res *36*, D102–D106.

Bulyk, M.L. (2003). Computational prediction of transcription-factor binding site locations. Genome Biol *5*, 201.

Bussemaker, H.J., Foat, B.C., and Ward, L.D. (2007). Predictive modeling of genome-wide mRNA expression: from modules to molecules. Annu. Rev. Biophys. Biomol. Struct. *36*, 329–347.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626–635.

Carvalho, B., Bengtsson, H., Speed, T.P., and Irizarry, R.A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics *8*, 485–499.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149–1154.

Chiappetta, G., Botti, G., Monaco, M., Pasquinelli, R., Pentimalli, F., Di Bonito, M., D'Aiuto, G., Fedele, M., Iuliano, R., Palmieri, E.A., et al. (2004). HMGA1 protein overexpression in human breast carcinomas: correlation with ErbB2 expression. Clin. Cancer Res. *10*, 7637–7644.

Chong, J.J.H., Chandrakanthan, V., Xaymardan, M., Asli, N.S., Li, J., Ahmed, I., Heffernan, C., Menon, M.K., Scarlett, C.J., Rashidianfar, A., et al. (2011). Adult Cardiac-Resident MSC-like Stem Cells with a Proepicardial Origin. Cell Stem Cell *9*, 527–540.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. *45*, 1127–1133.

Collavin, L., Monte, M., Verardo, R., Pfleger, C., and Schneider, C. (2000). Cell-cycle regulation of the p53-inducible gene B99. FEBS Lett. *481*, 57–62.

Connell, N.D., and Rheinwald, J.G. (1983). Regulation of the cytoskeleton in mesothelial cells: reversible loss of keratin and increase in vimentin during rapid growth in culture. Cell *34*, 245–253.

Core, L.J., and Lis, J.T. (2009). Paused Pol II captures enhancer activity and acts as a potent insulator. Genes Dev *23*, 1606–1612.

Courey, A.J. (2008). Mechanisms in Transcriptional Regulation. (Wiley-Blackwell).

Das, M.K., and Dai, H.-K. (2007). A survey of DNA motif finding algorithms. BMC Bioinformatics *8*, S21.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002). A Genomic Regulatory Network for Development. Science (80-.). *295*, 1669–1678.

Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M.J., Bottomley, W., et al. (2002). Mutations of the BRAF gene in human cancer. Nature *417*, 949–954.

Davuluri, R. V, Grosse, I., and Zhang, M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat Genet *29*, 412–417.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, P3.

Dennis, J.E., Merriam, A., Awadallah, A., Yoo, J.U., Johnstone, B., and Caplan, A.I. (1999). A quadripotential mesenchymal progenitor cell isolated from the marrow of an adult mouse. J. Bone Miner. Res. *14*, 700–709.

Deryugina, E.I., and Müller-Sieburg, C.E. (1993). Stromal cells in long-term cultures: keys to the elucidation of hematopoietic development? Crit. Rev. Immunol. *13*, 115–150.

Dolde, C.E., Mukherjee, M., Cho, C., and Resar, L.M.S. (2002). HMG-I/Y in human breast cancer cell lines. Breast Cancer Res. Treat. *71*, 181–191.

Dray, S., and Dufour, A.B. (2007). The ade4 package: implementing the duality diagram for ecologists. J Stat Soft *22*, 1–20.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. *95*, 14863–14868.

ENCODE, and Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Erices, A., Conget, P., and Minguell, J.J. (2000). Mesenchymal progenitor cells in human umbilical cord blood. Br. J. Haematol. *109*, 235–242.

Eskin, E., and Pevzner, P.A. (2002). Finding composite regulatory patterns in DNA sequences. Bioinformatics *18*, S354–S363.

Falgueras, J., Lara, A.J., Fernández-Pozo, N., Cantón, F.R., Pérez-Trabado, G., and Claros, M.G. (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics *11*, 38.

Feeley, K.M., and Wells, M. (2001). Precursor lesions of ovarian epithelial malignancy. Histopathology *38*, 87–95.

Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., and Blanchette, M. (2007). PReMod: a database of genome-wide mammalian cis-regulatory module predictions. Nucleic Acids Res *35*, D122–D126.

Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassman, T., Kulakovskiy, I. V, Lizio, M., Itoh, M., et al. (2014). A promoter level mammalian expression atlas. Nature.

Friedenstein, A.J., Deriglasova, U.F., Kulagina, N.N., Panasuk, A.F., Rudakowa, S.F., Luriá, E.A., and Ruadkow, I.A. (1974). Precursors for fibroblasts in different populations of hematopoietic cells as detected by the in vitro colony assay method. Exp. Hematol. *2*, 83–92.

Frith, M.C., Hansen, U., and Weng, Z. (2001). Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics *17*, 878–889.

Frith, M.C., Li, M.C., and Weng, Z. (2003). Cluster-Buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Res *31*, 3666–3668.

Fukagawa, T., Nogami, M., Yoshikawa, M., Ikeno, M., Okazaki, T., Takami, Y., Nakayama, T., and Oshimura, M. (2004). Dicer is essential for formation of the heterochromatin structure in vertebrate cells. Nat. Cell Biol. *6*, 784–791.

Fusco, A., and Fedele, M. (2007). Roles of HMGA proteins in cancer. Nat. Rev. Cancer *7*, 899–910.

Gaebel, R., Furlani, D., Sorg, H., Polchow, B., Frank, J., Bieback, K., Wang, W., Klopsch, C., Ong, L.-L., Li, W., et al. (2011). Cell origin of human mesenchymal stem cells determines a different healing performance in cardiac regeneration. PLoS One *6*, e15652.

Gentleman, R. (2005). The Bioconductor Project : Open-source Statistical Software for Bioinformatics and Computational Biology.

Giguère, V., McBroom, L.D., and Flock, G. (1995). Determinants of Target Gene Specificity for ROR alpha 1: Monomeric DNA Binding by an Orphan Nuclear Receptor. Mol Cell Biol *15*, 2517–2526.

Girardini, J.E., Napoli, M., Piazza, S., Rustighi, A., Marotta, C., Radaelli, E., Capaci, V., Jordan, L., Quinlan, P., Thompson, A., et al. (2011). A Pin1/mutant p53 axis promotes aggressiveness in breast cancer. Cancer Cell *20*, 79–91.

Girgis, H.Z., and Ovcharenko, I. (2012). Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. BMC Bioinformatics *13*, 25.

Gould, E. (2007). How widespread is adult neurogenesis in mammals? Nat. Rev. Neurosci. *8*, 481–488.

Grivennikov, S.I., Greten, F.R., and Karin, M. (2010). Immunity, inflammation, and cancer. Cell *140*, 883–899.

Gupta, S., Stamatoyannopoulos, J. a, Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol *8*, R24.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. *28*, 503–510.

Gyorffy, B., Lánczky, A., and Szállási, Z. (2012). Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. Endocr. Relat. Cancer *19*, 197–208.

Györffy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q., and Szallasi, Z. (2010a). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res. Treat. *123*, 725–731.

Győrffy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q., and Szallasi, Z. (2010b). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res. Treat. *123*, 725–731.

Győrffy, B., Surowiak, P., Budczies, J., and Lánczky, A. (2013). Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. PLoS One *8*, e82241.

Hah, N., Danko, C.G., Core, L., Waterfall, J.J., Siepel, A., Lis, J.T., and Kraus, W.L. (2011). A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. Cell *145*, 622–634.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Heid, C.A., Stevens, J., Livak, K.J., and Williams, P.M. (1996). Real time quantitative PCR. Genome Res. *6*, 986–994.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576–589.

Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. Stat. Med. *9*, 811–818.

Holschneider, C.H., and Berek, J.S. Ovarian cancer: epidemiology, biology, and prognostic factors. Semin. Surg. Oncol. *19*, 3–10.

Hothorn, T., Hornik, K., van de Wiel, M.A., and Zeileis, A. (2006). A Lego System for Conditional Inference. Am Stat *60*, 257–263.

Hu, Y.F., Lüscher, B., Admon, A., Mermod, N., and Tjian, R. (1990). Transcription factor AP-4 contains multiple dimerization domains that regulate dimer specificity. Genes Dev. *4*, 1741–1752.

Huang, L., Li, F., Sheng, J., Xia, X., Ma, J., Zhan, M., and Wong, S.T.C. (2014). DrugComboRanker: drug combination discovery based on target network analysis. Bioinformatics *30*, i228–i236.

Huang, Q., Gong, C., Li, J., Zhuo, Z., Chen, Y., Wang, J., and Hua, Z.-C. (2012). Distance and Helical Phase Dependence of Synergistic Transcription Activation in cis -Regulatory Module. PLoS One *7*, e31198.

Hubner, N.C., Bird, A.W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., Hyman, A., and Mann, M. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J. Cell Biol. *189*, 739–754.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000). Computational Identification of Cis-regulatory Elements Associated with Groups of Functionally Related Genes in Saccharomyces cerevisiae. J Mol Biol *296*, 1205–1214.

Imamura, T., Yamamoto, S., Ohgane, J., Hattori, N., Tanaka, S., and Shiota, K. (2004). Non-coding RNA directed DNA demethylation of Sphk1 CpG island. Biochem. Biophys. Res. Commun. *322*, 593–600.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. *31*, e15.

Johnson, R., Gamblin, R.J., Ooi, L., Bruce, A.W., Donaldson, I.J., Westhead, D.R., Wood, I.C., Jackson, R.M., and Buckley, N.J. (2006). Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. Nucleic Acids Res *34*, 3862–3877.

Kachali, C., Eltoum, I., Horton, D., and Chhieng, D.C. (2006). Use of mesothelin as a marker for mesothelial cells in cytologic specimens. Semin. Diagn. Pathol. *23*, 20–24.

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Bertin, N., Kaiho, A., Ninomiya, N., Daub, C.O., et al. (2011a). Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res. *21*, 1150–1159.

Kanamori-Katayama, M., Kaiho, A., Ishizu, Y., Okamura-Oho, Y., Hino, O., Abe, M., Kishimoto, T., Sekihara, H., Nakamura, Y., Suzuki, H., et al. (2011b). LRRN4 and UPK3B are markers of primary mesothelial cells. PLoS One *6*, e25391.

Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73.

Kardassis, D., Falvey, E., Tsantili, P., Hadzopoulou-Cladaras, M., and Zannis, V. (2002). Direct Physical Interactions between HNF-4 and Sp1 Mediate Synergistic Transactivation of the Apolipoprotein CIII Promoter. Biochemistry *41*, 1217–1228.

Karnoub, A.E., Dash, A.B., Vo, A.P., Sullivan, A., Brooks, M.W., Bell, G.W., Richardson, A.L., Polyak, K., Tubo, R., and Weinberg, R.A. (2007). Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. Nature *449*, 557–563.

Karst, A.M., Levanon, K., and Drapkin, R. (2011). Modeling high-grade serous ovarian carcinogenesis from the fallopian tube. Proc. Natl. Acad. Sci. U. S. A. *108*, 7547–7552.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564–1566.

Kel, A., Konovalova, T., Waleev, T., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E. (2006). Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. Bioinformatics *22*, 1190–1197.

Kim, J., Coffey, D.M., Creighton, C.J., Yu, Z., Hawkins, S.M., and Matzuk, M.M. (2012). High-grade serous ovarian cancer arises from fallopian tube in a mouse model. Proc. Natl. Acad. Sci. U. S. A. *109*, 3921–3926.

King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R.C. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res *15*, 1051–1060.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. (2006). CAGE: cap analysis of gene expression. Nat. Methods *3*, 211–222.

Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., and Chiaromonte, F. (2004). Regulatory Potential Scores From Genome-Wide Three-Way Alignments of Human, Mouse, and Rat. Genome Res *14*, 700–707.

Ksiazek, K., Mikula-Pietrasik, J., Korybalska, K., Dworacki, G., Jörres, A., and Witowski, J. (2009). Senescent peritoneal mesothelial cells promote ovarian cancer cell adhesion: the role of oxidative stress-induced fibronectin. Am. J. Pathol. *174*, 1230–1240.

Kulakovskiy, I. V, Boeva, V.A., Favorov, A. V, and Makeev, V.J. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. Bioinformatics *26*, 2622–2623.

Kulakovskiy, I. V, Medvedeva, Y.A., Schaefer, U., Kasianov, A.S., Vorontsov, I.E., Bajic, V.B., and Makeev, V.J. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res *41*, D195–D202.

Kuznetsov, S.A., Mankani, M.H., Gronthos, S., Satomura, K., Bianco, P., and Robey, P.G. (2001). Circulating skeletal stem cells. J. Cell Biol. *153*, 1133–1140.

Laperriere, D., Wang, T.-T., White, J.H., and Mader, S. (2007). Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. BMC Genomics *8*, 23.

LaRocca, P.J., and Rheinwald, J.G. (1984). Coexpression of simple epithelial keratins and vimentin by human mesothelium and mesothelioma in vivo and in culture. Cancer Res. *44*, 2991–2999.

Lassmann, T., Hayashizaki, Y., and Daub, C.O. (2009). TagDust--a program to eliminate artifacts from next generation sequencing data. Bioinformatics *25*, 2839–2840.

Lee, J., Sung, C.O., Lee, E.J., Do, I.-G., Kim, H.-C., Yoon, S.H., Lee, W.Y., Chun, H.K., Kim, K.-M., and Park, Y.S. (2012). Metastasis of neuroendocrine tumors are characterized by increased cell proliferation and reduced expression of the ATM gene. PLoS One *7*, e34456.

Lenhard, B., Sandelin, A., Mendoza, L., Engström, P., Jareborg, N., and Wasserman, W.W. (2003). Identification of conserved regulatory elements by comparative genome analysis. J Biol *2*, 13.

Levanon, K., Crum, C., and Drapkin, R. (2008). New insights into the pathogenesis of serous ovarian cancer and its clinical impact. J. Clin. Oncol. *26*, 5284–5293.

Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. Proc Natl Acad Sci *102*, 4936–4942.

Levy, A., Sela, N., and Ast, G. (2008). TranspoGene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. Nucleic Acids Res *36*, D47–D52.

Liau, S.-S., Jazag, A., and Whang, E.E. (2006). HMGA1 is a determinant of cellular invasiveness and in vivo metastatic potential in pancreatic adenocarcinoma. Cancer Res. *66*, 11613–11622.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Lis, R., Touboul, C., Mirshahi, P., Ali, F., Mathew, S., Nolan, D.J., Maleki, M., Abdalla, S.A., Raynaud, C.M., Querleu, D., et al. (2011). Tumor associated mesenchymal stem cells protects ovarian cancer cells from hyperthermia through CXCL12. Int. J. Cancer *128*, 715–725.

Liu, X.S., Li, H., Song, B., and Liu, X. (2010). Polo-like kinase 1 phosphorylation of G2 and S-phase-expressed 1 protein is essential for p53 inactivation during G2 checkpoint recovery. EMBO Rep. *11*, 626–632.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M. V, Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. *14*, 1675–1680.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., et al. (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell *1*, 55–70.

Marchand, B., Bajic, V.B., and Kaushik, D.K. (2011). Highly scalable ab initio genomic motif identification. In Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis on - SC '11, (New York, New York, USA: ACM Press), p. 1.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. N. Engl. J. Med. *361*, 1058–1066.

Mathelier, A., Shi, W., and Wasserman, W.W. (2015). Identification of altered cis-regulatory elements in human disease. Trends Genet. *31*, 67–76.

Matys, V., Kel-Margoulis, O. V, Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108–D110.

McLean, K., Gong, Y., Choi, Y., Deng, N., Yang, K., Bai, S., Cabrera, L., Keller, E., McCauley, L., Cho, K.R., et al. (2011). Human ovarian carcinoma–associated mesenchymal stem cells regulate cancer stem cells and tumorigenesis via altered BMP production. J. Clin. Invest. *121*, 3206–3219.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31–46.

Meyer, D., and Buchta, C. (2015). proxy: Distance and Similarity Measures.

Mishra, P.J., Mishra, P.J., Humeniuk, R., Medina, D.J., Alexe, G., Mesirov, J.P., Ganesan, S., Glod, J.W., and Banerjee, D. (2008). Carcinoma-associated fibroblast-like differentiation of human mesenchymal stem cells. Cancer Res. *68*, 4331–4339.

Monte, M., Collavin, L., Lazarevic, D., Utrera, R., Dragani, T.A., and Schneider, C. (2000). Cloning, chromosome mapping and functional characterization of a human homologue of murine gtse-1 (B99) gene. Gene *254*, 229–236.

Monte, M., Benetti, R., Buscemi, G., Sandy, P., Del Sal, G., and Schneider, C. (2003). The cell cycle-regulated protein human GTSE-1 controls DNA damage-induced apoptosis by affecting p53 function. J. Biol. Chem. *278*, 30356–30364.

Monte, M., Benetti, R., Collavin, L., Marchionni, L., Del Sal, G., and Schneider, C. (2004). hGTSE-1 expression stimulates cytoplasmic localization of p53. J. Biol. Chem. *279*, 11744–11752.

Mora, A., and Donaldson, I.M. (2011). iRefR: an R package to manipulate the iRefIndex consolidated protein interaction database. BMC Bioinformatics *12*, 455.

Muraglia, A., Cancedda, R., and Quarto, R. (2000). Clonal mesenchymal progenitors from human bone marrow differentiate in vitro according to a hierarchical model. J. Cell Sci. *113 (Pt 7*, 1161–1166.

Murdoch, W.J., and McDonnel, A.C. (2002). Roles of the ovarian surface epithelium in ovulation and carcinogenesis. Reproduction *123*, 743–750.

Muzny, D.M., Scherer, S.E., Kaul, R., Wang, J., Yu, J., Sudbrak, R., Buhay, C.J., Chen, R., Cree, A., Ding, Y., et al. (2006). The DNA sequence, annotation and analysis of human chromosome 3. Nature *440*, 1194–1198.

Narasimhan, C., LoCascio, P., and Uberbacher, E. (2003). Background rareness-based iterative multiple sequence alignment algorithm for regulatory element detection. Bioinformatics *19*, 1952–1963.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res *37*, D77–D82.

Nilsson, R., Bajic, V.B., Suzuki, H., di Bernardo, D., Björkegren, J., Katayama, S., Reid, J.F., Sweet, M.J., Gariboldi, M., Carninci, P., et al. (2006). Transcriptional network dynamics in macrophage activation. Genomics *88*, 133–142.

Onder, T.T., Kara, N., Cherry, A., Sinha, A.U., Zhu, N., Bernt, K.M., Cahan, P., Marcarci, B.O., Unternaehrer, J., Gupta, P.B., et al. (2012). Chromatin-modifying enzymes as modulators of reprogramming. Nature *483*, 598–602.

Pachkov, M., Erb, I., Molina, N., and van Nimwegen, E. (2007). SwissRegulon: a database of genome-wide annotations of regulatory sites. Nucleic Acids Res *35*, D127–D131.

Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2013). Biostrings: String objects representing biological sequences, and matching algorithms. Bioconductor.

Park, J.-H., Kim, Y.-G., Shaw, M., Kanneganti, T.-D., Fujimoto, Y., Fukase, K., Inohara, N., and Núñez, G. (2007). Nod1/RICK and TLR signaling regulate chemokine and antimicrobial innate immune responses in mesothelial cells. J. Immunol. *179*, 514–521.

Pavesi, G., Mauri, G., and Pesole, G. (2004). In silico representation and discovery of transcription factor binding sites. Br. Bioinform *5*, 217–236.

Pavesi, G., Zambelli, F., and Pesole, G. (2007). WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. BMC Bioinformatics *8*, 46.

Pegoraro, S., Ros, G., Piazza, S., Sommaggio, R., Ciani, Y., Rosato, A., Sgarra, R., Del Sal, G., and Manfioletti, G. (2013). HMGA1 promotes metastatic processes in basal-like breast cancer regulating EMT and stemness. Oncotarget *4*, 1293–1308.

Pelekanos, R. a, Li, J., Gongora, M., Chandrakanthan, V., Scown, J., Suhaimi, N., Brooke, G., Christensen, M.E., Doan, T., Rice, A.M., et al. (2012). Comprehensive

transcriptome and immunophenotype analysis of renal and cardiac MSC-like populations supports strong congruence with bone marrow MSC despite maintenance of distinct identities. Stem Cell Res. *8*, 58–73.

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Polak, P., and Domany, E. (2006). Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics *7*, 133.

Pontén, F., Jirström, K., and Uhlen, M. (2008). The Human Protein Atlas--a tool for pathology. J. Pathol. *216*, 387–393.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. *12*, R68.

Qiu, W., Hu, M., Sridhar, A., Opeskin, K., Fox, S., Shipitsin, M., Trivett, M., Thompson, E.R., Ramakrishna, M., Gorringe, K.L., et al. (2008). No evidence of clonal somatic genetic alterations in cancer-associated fibroblasts from human breast and ovarian carcinomas. Nat. Genet. *40*, 650–655.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. *1*, 1–1731.

Ram, T.G., Reeves, R., and Hosick, H.L. (1993). Elevated high mobility group-I(Y) gene expression is associated with progressive transformation of mouse mammary epithelial cells. Cancer Res. *53*, 2655–2660.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. *16*, 11–19.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell *140*, 744–752.

Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.

Reeves, R., Edberg, D.D., and Li, Y. (2001). Architectural transcription factor HMGI(Y) promotes tumor progression and mesenchymal transition of human epithelial cells. Mol. Cell. Biol. *21*, 575–594.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

Rhodes, D.R., and Chinnaiyan, A.M. (2004). Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. Ann. N. Y. Acad. Sci. *1020*, 32–40.

Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B.B., Barrette, T.R., Anstet, M.J., Kincead-Beal, C., Kulkarni, P., et al. (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia *9*, 166–180.

Ringnér, M., Fredlund, E., Häkkinen, J., Borg, Å., and Staaf, J. (2011). GOBO: gene expression-based outcome for breast cancer online. PLoS One *6*, e17911.

Rinkevich, Y., Mori, T., Sahoo, D., Xu, P.-X., Bermingham, J.R., and Weissman, I.L. (2012a). Identification and prospective isolation of a mesothelial precursor lineage giving rise to smooth muscle cells and fibroblasts for mammalian internal organs, and their vasculature. Nat. Cell Biol. *14*, 1251–1260.

Rinkevich, Y., Mori, T., Sahoo, D., Xu, P.-X., Bermingham, J.R., and Weissman, I.L. (2012b). Identification and prospective isolation of a mesothelial precursor lineage giving rise to smooth muscle cells and fibroblasts for mammalian internal organs, and their vasculature. Nat. Cell Biol. *14*, 1251–1260.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010a). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010b). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Ryall, K.A., and Tan, A.C. (2015). Systems biology approaches for advancing the discovery of effective drug combinations. J. Cheminform. *7*, 7.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Sandelin, A., and Wasserman, W.W. (2004). Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics. J Mol Biol *338*, 207–215.

Sandve, G.K., Abul, O., Walseng, V., and Drabløs, F. (2007). Improved benchmarks for computational motif discovery. BMC Bioinformatics *8*, 193.

Sato, A., Torii, I., Okamura, Y., Yamamoto, T., Nishigami, T., Kataoka, T.R., Song, M., Hasegawa, S., Nakano, T., Kamei, T., et al. (2010). Immunocytochemistry of CD146 is useful to discriminate between malignant pleural mesothelioma and reactive mesothelium. Mod. Pathol. *23*, 1458–1466.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. U. S. A. *93*, 10614–10619.

Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E., and Davis, R.W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol. *16*, 301–306.

Scolz, M., Widlund, P.O., Piazza, S., Bublik, D.R., Reber, S., Peche, L.Y., Ciani, Y., Hubner, N., Isokane, M., Monte, M., et al. (2012). GTSE1 is a microtubule plus-end tracking protein that regulates EB1-dependent cell migration. PLoS One *7*, e51259.

Sgarra, R., Rustighi, A., Tessari, M.A., Di Bernardo, J., Altamura, S., Fusco, A., Manfioletti, G., and Giancotti, V. (2004). Nuclear phosphoproteins HMGA and their relationship with chromatin structure and cancer. FEBS Lett. *574*, 1–8.

Shah, S.N., and Resar, L.M.S. (2012). High mobility group A1 and cancer: potential biomarker and therapeutic target. Histol. Histopathol. *27*, 567–579.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res *13*, 2498–2504.

Shi, S., and Gronthos, S. (2003). Perivascular niche of postnatal mesenchymal stem cells in human bone marrow and dental pulp. J. Bone Miner. Res. *18*, 696–704.

Shimizu, K., Goldfarb, M., Suard, Y., Perucho, M., Li, Y., Kamata, T., Feramisco, J., Stavnezer, E., Fogh, J., and Wigler, M.H. (1983). Three human transforming genes are related to the viral ras oncogenes. Proc. Natl. Acad. Sci. U. S. A. *80*, 2112–2116.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. U. S. A. *100*, 15776–15781.

Da Silva Meirelles, L., Chagastelles, P.C., and Nardi, N.B. (2006). Mesenchymal stem cells reside in virtually all post-natal organs and tissues. J. Cell Sci. *119*, 2204–2213.

Simcha, D., Price, N.D., and Geman, D. (2012). The Limits of De Novo DNA Motif Discovery. PLoS One *7*, e47836.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.

Sinha, S., and Tompa, M. (2003). YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res *31*, 3586–3588.

Sinha, S., van Nimwegen, E., and Siggia, E.D. (2003). A probabilistic method to detect regulatory modules. Bioinformatics *19 Suppl 1*, i292–i301.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561–566.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. U. S. A. *98*, 10869–10874.

Spaeth, E.L., Dembinski, J.L., Sasser, A.K., Watson, K., Klopp, A., Hall, B., Andreeff, M., and Marini, F. (2009). Mesenchymal stem cell transition to tumor-associated fibroblasts contributes to fibrovascular network expansion and tumor progression. PLoS One *4*, e4992.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16–23.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. *11*, 751–760.

Swinton, J. (2011). Vennerable: Venn and Euler area-proportional diagrams. R package version 2.11.0. 1–33.

Takaha, N., Resar, L.M.S., Vindivich, D., and Coffey, D.S. (2004). High mobility group protein HMGI(Y) enhances tumor cell growth, invasion, and matrix metalloproteinase-2 expression in prostate cancer cells. Prostate *60*, 160–167.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Tang, J., Karhinen, L., Xu, T., Szwajda, A., Yadav, B., Wennerberg, K., and Aittokallio, T. (2013). Target inhibition networks: predicting selective combinations

of druggable targets to block cancer survival pathways. PLoS Comput. Biol. *9*, e1003226.

Taniguchi, S., Takeoka, M., Ehara, T., Hashimoto, S., Shibuki, H., Yoshimura, N., Shigematsu, H., Takahashi, K., and Katsuki, M. (2001). Structural Fragility of Blood Vessels and Peritoneum in Calponin h1-deficient Mice, Resulting in an Increase in Hematogenous Metastasis and Peritoneal Dissemination of Malignant Tumor Cells. Cancer Res. *61*, 7627–7634.

Taylor, J., Tyekucheva, S., King, D.C., Hardison, R.C., Miller, W., and Chiaromonte, F. (2006). ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res *16*, 1596–1604.

TCGA (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

TCGA (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

TCGA (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43–49.

Tigrani, D.-Y., and Weydert, J.A. (2007). Immunohistochemical expression of osteopontin in epithelioid mesotheliomas and reactive mesothelial proliferations. Am. J. Clin. Pathol. *127*, 580–584.

Till, J.E., and McCulloch, E.A. (2012). A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. 1961. Radiat. Res. *178*, AV3–AV7.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science *310*, 644–648.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A. V, Frith, M.C., Fu, Y., Kent, W.J., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. *23*, 137–144.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res. *14*, 62–66.

Uccelli, A., Moretta, L., and Pistoia, V. (2008). Mesenchymal stem cells in health and disease. Nat. Rev. Immunol. *8*, 726–736.

Valastyan, S., and Weinberg, R.A. (2011). Tumor metastasis: molecular insights and evolving paradigms. Cell *147*, 275–292.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S. a, and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat Rev Genet *10*, 252–263.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484–487.

Verardo, R., Piazza, S., Klaric, E., Ciani, Y., Bussadori, G., Marzinotto, S., Mariuzzi, L., Cesselli, D., Beltrami, A.P., Mano, M., et al. (2014). Specific mesothelial signature marks the heterogeneity of mesenchymal stem cells from high-grade serous ovarian cancer. Stem Cells *32*, 2998–3011.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Wang, T., and Stormo, G.D. (2005). Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci *102*, 17400–17405.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res *22*, 1798–1812.

Warnes, G.R. (2010). gplots: Various R programming tools for plotting data. Text *2*.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2013). gplots: Various R programming tools for plotting data. R package version 2.11.0. 1–61.

Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A., and Hampton, G.M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. U. S. A. *98*, 1176–1181.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. R package version 2.11.0. 1–195.

Williams, J.T., Southerland, S.S., Souza, J., Calcutt, A.F., and Cartledge, R.G. (1999). Cells isolated from adult human skeletal muscle capable of differentiating into multiple mesodermal phenotypes. Am. Surg. *65*, 22–26.

Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. Nucleic Acids Res *41*, D165–D170.

Wood, L.J., Maher, J.F., Bunton, T.E., and Resar, L.M. (2000). The oncogenic properties of the HMG-I gene family. Cancer Res. *60*, 4256–4261.

Workman, C.T., and Stormo, G.D. (2000). ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac Symp Biocomput 467–478.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338–345.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., et al. (2013). Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. Cell *154*, 801–813.

Yuh, C.H., Bolouri, H., and Davidson, E.H. (1998). Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene. Science (80-.). *279*, 1896–1902.

Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D.A., Hayashizaki, Y., and Gaasterland, T. (2003). Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res. *13*, 1290–1300.

Zeppieri, M., Salvetat, M.L., Beltrami, A.P., Cesselli, D., Bergamin, N., Russo, R., Cavaliere, F., Varano, G.P., Alcalde, I., Merayo, J., et al. (2013). Human adipose-derived stem cells for the treatment of chemically burned rat cornea: preliminary results. Curr. Eye Res. *38*, 451–463.

Zhao, B., Ye, X., Yu, J., Li, L., Li, W., Li, S., Yu, J., Lin, J.D., Wang, C.-Y., Chinnaiyan, A.M., et al. (2008). TEAD mediates YAP-dependent gene induction and growth control. Genes Dev. *22*, 1962–1971.

Zhou, Q., and Wong, W.H. (2004). CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci *101*, 12114–12119.

Zhou, X., Temam, S., Oh, M., Pungpravat, N., Huang, B.-L., Mao, L., and Wong, D.T. (2006). Global expression-based classification of lymph node metastasis and extracapsular spread of oral tongue squamous cell carcinoma. Neoplasia *8*, 925–932.

Zhou, Z., Li, X., Deng, C., Ney, P.A., Huang, S., and Bungert, J. (2010). USF and NF-E2 Cooperate to Regulate the Recruitment and Activity of RNA Polymerase II in the beta-Globin Gene Locus. J Biol Chem *285*, 15894–15905.

Zuk, P.A., Zhu, M., Mizuno, H., Huang, J., Futrell, J.W., Katz, A.J., Benhaim, P., Lorenz, H.P., and Hedrick, M.H. (2001). Multilineage cells from human adipose tissue: implications for cell-based therapies. Tissue Eng. *7*, 211–228.

SUPPORTING MATERIAL

Table S1: Clinical-pathological characteristics of the patients considered in the study. Characteristics of each tumor sample from which HG-SOC-MSCs used in cell-based assays were isolated.

SAMPLE	AGE	DIAGNOSIS	MITOSIS (SCORE)	GRADING (S)	pTNM	FIGO ST
sample 12	56	HIGH-GRADE SEROUS ADENOCARCINOMA	79/10HPF(3)	3 (8)	pT3b,Nx,Mx	IIIB
sample 17	75	HIGH-GRADE SEROUS ADENOCARCINOMA	48/10HPF (3)	3 (8)	pT3c,Nx,Mx	IIIC
sample 19	47	HIGH-GRADE SEROUS ADENOCARCINOMA	46/10HPF (3)	3 (9)	pT3c,Nx,Mx	IIIC
sample 26	64	HIGH-GRADE SEROUS ADENOCARCINOMA	37/10HPF (3)	3 (8)	pT3c,Nx,Mx	IIIC
sample 31	23	HIGH-GRADE SEROUS ADENOCARCINOMA	17/10HPF (2)	2(6)	pT3c,N1,M1(cute)	IV
sample 33	70	HIGH-GRADE SEROUS ADENOCARCINOMA	46/10HPF (3)	3 (9)	pT3c,Nx,Mx	IIIC
sample 34	51	HIGH-GRADE SEROUS ADENOCARCINOMA	60/10HPF (3)	3 (9)		
sample 36	52	HIGH-GRADE SEROUS ADENOCARCINOMA	48/10HPF (3)	3 (9)	pT2c,Nx,Mx	IIC
sample 38	68	HIGH-GRADE SEROUS ADENOCARCINOMA	68/10HPF (3)	3 (9)	pT3c,N0,Mx	IIIC
sample 39	30	BORDER SEROUS ADENOCARCINOMA	0/10HPF(1)	1(5)	pT1C,N0,Mx	IC
sample 40	55	HIGH-GRADE SEROUS ADENOCARCINOMA	28/10HPF (3)	3 (9)	pT2b,Nx,Mx	IIB
sample 41	39	HIGH-GRADE SEROUS ADENOCARCINOMA	66/10HPF (3)	3 (9)		
sample 43	83	HIGH-GRADE SEROUS ADENOCARCINOMA	50/10HPF (3)	3 (9)	pT3b,Nx,Mx	IIIB
sample 48	47	BORDER SEROUS ADENOCARCINOMA	12/10HPF (2)	2(6)	pT1A,N0,Mx	IA
sample 49	56	HIGH-GRADE SEROUS ADENOCARCINOMA	166/10HPF (3)	3 (8)	pT3a,Nx,Mx	IIIA
sample 57	63	HIGH-GRADE SEROUS ADENOCARCINOMA	36/10HPF (3)	3 (9)	pT3c,Nx,Mx	IIIc

Table S2

	INJECTED MICE (n)	TUMOURS (n)
HG-SOC-MSCs (19-02)	**2**	**0**
HG-SOC-MSCs (41-01)	**2**	**0**
HG-SOC-MSCs (43-01)	**2**	**0**
TOT	**6**	**0**

	INJECTED MICE (n)	TUMOURS (n)
SKOV-3	**5**	**5**

Table S3 A

GENE SYMBOL	P.VALUES	DIFFERENTIAL GLOBAL PEAK VALUE (HG-SOC-MSCs_vs_N-MSCs)
SERPINE1	0.032597848	4.24
ERO1L	1.86E-05	2.22
RAC2	8.54E-06	1.80
GNG11	0.01561648	1.73
ITGB1	0.014141219	1.67
F2RL1	0.000630888	1.63
FRMD4A	0.000415651	1.51
SULF1	0.040224159	1.51
MIR663B	0.020494662	1.46
TRIO	0.027126448	1.41
KPNA4	0.000745618	1.30
DCBLD2	0.018337521	1.23
DPYSL3	0.036048586	1.22
UBE2D3	0.005904836	1.20
ITGA3	0.000456793	1.18
ANTXR2	0.000708471	1.14
HMGA2	0.026622099	1.13
C1QTNF1	0.033070151	1.11
EZR	0.008053422	1.09
MAP2K1	0.001698256	1.09
NF1	0.002856186	1.08
LRRFIP1	0.004071572	1.07
ECE1	0.021515579	1.04
CORO2B	0.044379945	1.01
MYADM	0.001833474	1.01
SPON2	0.038437698	1.01
DKK3	0.00846815	1.00
CD82	0.012641189	0.99
PGAM1	0.012641189	0.99
STC1	0.002154894	0.99
ACTN1	0.001163193	0.98
SEMA3C	0.017761272	0.98
ITGA2	0.013781792	0.97
PGK1	0.013609481	0.97
IL33	0.005434391	0.96
HUWE1	0.028007808	0.96
PLS3	0.006032561	0.94
FNDC3A	0.005058224	0.91
SGK1	0.020603751	0.91
BNIP3	0.000198602	0.89
PHLDB1	0.000202043	0.88

FAM177A1	0.012631333	0.87
TNIP1	0.009971161	0.86
ABLIM3	0.008416762	0.84
BAZ2A	0.011638538	0.84
ANTXR1	0.04870766	0.80
CDV3	0.017432394	0.79
INHBA	0.020159711	0.79
SPTBN1	0.009496772	0.79
CREB3	0.001725101	0.78
HDAC1	0.031940382	0.77
NBEAL1	0.031940382	0.77
PDLIM4	0.006101891	0.77
ZMAT3	0.031940382	0.77
CDCP1	0.044024973	0.74
CTPS	0.005479976	0.74
LACTB	0.034427427	0.74
SH3BP4	0.005479976	0.74
TMEM45A	0.004733273	0.72
YWHAG	0.036916396	0.71
DYNLT3	0.001322951	0.70
FGFR1OP2	0.009534613	0.70
VMP1	0.025909045	0.70
CTNNB1	0.001016734	0.69
YAP1	0.027023452	0.64
NUTF2	0.012450834	0.63
FAR1	0.011519521	0.62
STT3B	0.007912131	0.61
RALA	0.005121073	0.60
TRIP12	0.005121073	0.60
ACTA2	0.006628844	0.59
ARPC5	0.006628844	0.59
TMEM123	0.009907153	0.58
TOR1AIP1	0.009907153	0.58
TSPAN9	0.0493146	0.57
CASK	0.013349063	0.56
OSBPL9	0.013349063	0.56
RNF181	0.013349063	0.56
SUN2	0.013349063	0.56
ASAP1	0.023910764	0.53
PLXNB2	0.014956364	0.50
SERPINB6	0.048949735	0.50
RHOA	0.037702915	0.48
CAPNS1	0.044214858	0.47
VDAC1	0.044214858	0.47
GBE1	0.042385344	0.46

TCEB1	0.043653925	0.43
ARHGAP18	0.043653925	0.43
H6PD	0.048680346	0.42
STK17B	0.048680346	0.42
TAF9	0.044839993	0.41
ARFIP1	0.036787498	0.40
ATP1A1	0.036787498	0.40
C9orf25	0.036787498	0.40
CHMP7	0.036787498	0.40
COX19	0.036787498	0.40
CPPED1	0.036787498	0.40
EPT1	0.036787498	0.40
F3	0.036787498	0.40
GOLIM4	0.036787498	0.40
GPATCH2	0.036787498	0.40
GPR180	0.036787498	0.40
GYS1	0.036787498	0.40
ICAM1	0.036787498	0.40
IDE	0.036787498	0.40
IFNAR2	0.036787498	0.40
JMJD1C	0.036787498	0.40
KCTD20	0.036787498	0.40
MACF1	0.036787498	0.40
MLLT11	0.036787498	0.40
OSGIN2	0.036787498	0.40
PLD3	0.036787498	0.40
PPFIA1	0.036787498	0.40
RBM6	0.036787498	0.40
TBX3	0.036787498	0.40
TERF2IP	0.036787498	0.40
TES	0.036787498	0.40
TNFAIP3	0.036787498	0.40
TP53RK	0.036787498	0.40
VGLL4	0.036787498	0.40

Table S3 B

GENE SYMBOL	P.VALUES	DIFFERENTIAL GLOBAL PEAK VALUE (HG-SOC-MSCs_vs_N-MSCs)
FN1	0.029767291	-26.12
COL1A2	2.65E-02	-2.77
MFAP5	1.18E-02	-2.27
COL6A1	0.038126203	-2.21
ISLR	0.00318969	-2.09
CCDC80	0.027140069	-1.94
PRRX1	0.006529897	-1.69
DCN	0.001511675	-1.67
PMP22	0.001823658	-1.63
PCOLCE	0.003538774	-1.26
PDGFRB	0.037153657	-1.21
TMEM43	0.03742545	-1.19
MYOF	0.014424954	-1.14
EIF6	0.001108348	-1.13
ANGPTL2	0.008413587	-1.12
PTRF	0.013349063	-1.11
41526	0.032150582	-1.08
HSPB6	0.005320076	-1.06
AKT1S1	0.007972306	-1.03
PDE4DIP	0.046313095	-0.98
BRP44L	0.000937456	-0.93
RPL34	0.001961531	-0.91
ACAA2	0.014774056	-0.90
SH3D19	0.00044674	-0.90
NME4	0.006210589	-0.88
NFIX	0.021786337	-0.87
IL1R1	0.008416762	-0.84
CD248	0.032156376	-0.83
CTSA	0.020217864	-0.83
BAG2	0.028149726	-0.80
JTB	0.000202499	-0.80
NFIC	0.009396685	-0.79
MRPS33	0.021259985	-0.74
FAM125A	0.036916396	-0.71
EFEMP1	0.001016734	-0.69
TGFBR2	0.00939512	-0.69
EMP3	0.001820564	-0.68
IFI16	0.009663676	-0.68
MLF2	0.003949773	-0.67
PDGFRA	0.039080981	-0.66
C16orf45	0.032271821	-0.63

SERTAD2	0.039335268	-0.62
LOXL1	0.023856385	-0.60
MYO1B	0.005121073	-0.60
TXNDC5	0.006628844	-0.59
UGP2	0.006628844	-0.59
C6orf72	0.009907153	-0.58
GLT8D1	0.043653925	-0.58
TGFB1	0.009907153	-0.58
GCLM	0.014956364	-0.50
LTA4H	0.014956364	-0.50
CNBP	0.025277358	-0.49
PLIN3	0.025277358	-0.49
RPL39	0.025277358	-0.49
HNRNPUL1	0.037702915	-0.48
PCBP1	0.037702915	-0.48
FKBP1A	0.042385344	-0.46
NDUFAF3	0.042385344	-0.46
VAV3	0.042385344	-0.46
ALAS1	0.035265203	-0.44
ALG10	0.035265203	-0.44
C6orf108	0.035265203	-0.44
CCDC12	0.035265203	-0.44
DGCR6L	0.035265203	-0.44
FAM165B	0.035265203	-0.44
GPR108	0.035265203	-0.44
METRNL	0.035265203	-0.44
NADK	0.035265203	-0.44
RNF24	0.035265203	-0.44
SIGIRR	0.035265203	-0.44
SLFN11	0.035265203	-0.44
TBC1D5	0.035265203	-0.44
TERC	0.035265203	-0.44
CYBRD1	0.035265203	-0.44
JAGN1	0.043653925	-0.43
AK1	0.048680346	-0.42
PAMR1	0.048680346	-0.42
ACOT13	0.044839993	-0.41
ACAT1	0.036787498	-0.40
ARL2-SNX15	0.036787498	-0.40
FKBP9	0.036787498	-0.40
PHLDA3	0.036787498	-0.40
PRUNE2	0.036787498	-0.40

Table S4

Genes up-regulated after HMGA1 silencing

SymbolID_coef1	ID	logFC	P.Value	adj.P.Val
ERBB4	214053_at	2.38	1.06E-08	9.92E-07
NAALAD2	1554507_at	2.04	1.08E-07	4.31E-06
PTGS1	238669_at	1.98	1.64E-07	5.85E-06
CAPN8	229030_at	1.94	2.06E-07	6.68E-06
LGR5	213880_at	1.90	1.72E-08	1.29E-06
COL3A1	215076_s_at	1.87	1.91E-10	1.09E-07
SH2D3A	219513_s_at	1.84	8.35E-10	2.17E-07
FABP4	203980_at	1.81	1.03E-09	2.39E-07
TMEM156	241844_x_at	1.81	5.72E-10	1.91E-07
DNAH12	243802_at	1.76	1.15E-10	8.68E-08
BCCIP	227896_at	1.74	1.99E-10	1.09E-07
GDPD1	238681_at	1.70	1.36E-08	1.15E-06
KBTBD8	239835_at	1.68	1.16E-09	2.56E-07
DPYD	1554536_at	1.67	4.23E-10	1.55E-07
ABCG2	209735_at	1.66	1.38E-11	3.91E-08
GKN2	238222_at	1.64	1.01E-08	9.66E-07
AGXT2L1	221008_s_at	1.62	1.32E-07	4.95E-06
AKR1B10	206561_s_at	1.60	1.09E-09	2.48E-07
PFKFB2	209992_at	1.58	4.55E-09	6.01E-07
C9orf95	219147_s_at	1.56	5.19E-08	2.66E-06
LOC100505989	238103_at	1.55	1.28E-09	2.64E-07
CHMP4C	226803_at	1.53	6.88E-09	7.61E-07
TXNRD3	59631_at	1.53	4.21E-08	2.31E-06
CBLB	227900_at	1.50	1.68E-09	3.19E-07
MAP7	202890_at	1.49	4.41E-08	2.38E-06
SAMD5	228653_at	1.49	7.42E-09	7.94E-07
MAL2	224650_at	1.48	3.85E-08	2.19E-06
RELN	205923_at	1.47	2.65E-08	1.67E-06
FAT3	236029_at	1.46	6.49E-10	1.97E-07
PGAP1	244321_at	1.45	1.38E-08	1.15E-06
ESRP2	229223_at	1.44	2.03E-07	6.61E-06
PPM1H	212686_at	1.42	3.95E-09	5.48E-07
FAM83H	226129_at	1.42	2.44E-10	1.14E-07
LOC100506713	229833_at	1.42	5.82E-09	6.77E-07
PCDH7	205535_s_at	1.40	2.87E-09	4.51E-07
ST20	217104_at	1.40	2.41E-07	7.51E-06
TMEM45A	219410_at	1.40	4.33E-08	2.35E-06
PPP2R1B	222351_at	1.40	2.16E-09	3.79E-07

Table S5

Genes up-regulated after HMGA1 silencing

SymbolID_coef1	ID	logFC	P.Value	adj.P.Val
ERBB4	214053_at	2.38	1.06E-08	9.92E-07
NAALAD2	1554507_at	2.04	1.08E-07	4.31E-06
PTGS1	238669_at	1.98	1.64E-07	5.85E-06
CAPN8	229030_at	1.94	2.06E-07	6.68E-06
LGR5	213880_at	1.90	1.72E-08	1.29E-06
COL3A1	215076_s_at	1.87	1.91E-10	1.09E-07
SH2D3A	219513_s_at	1.84	8.35E-10	2.17E-07
FABP4	203980_at	1.81	1.03E-09	2.39E-07
TMEM156	241844_x_at	1.81	5.72E-10	1.91E-07
DNAH12	243802_at	1.76	1.15E-10	8.68E-08
BCCIP	227896_at	1.74	1.99E-10	1.09E-07
GDPD1	238681_at	1.70	1.36E-08	1.15E-06
KBTBD8	239835_at	1.68	1.16E-09	2.56E-07
DPYD	1554536_at	1.67	4.23E-10	1.55E-07
ABCG2	209735_at	1.66	1.38E-11	3.91E-08
GKN2	238222_at	1.64	1.01E-08	9.66E-07
AGXT2L1	221008_s_at	1.62	1.32E-07	4.95E-06
AKR1B10	206561_s_at	1.60	1.09E-09	2.48E-07
PFKFB2	209992_at	1.58	4.55E-09	6.01E-07
C9orf95	219147_s_at	1.56	5.19E-08	2.66E-06
LOC100505989	238103_at	1.55	1.28E-09	2.64E-07
CHMP4C	226803_at	1.53	6.88E-09	7.61E-07
TXNRD3	59631_at	1.53	4.21E-08	2.31E-06
CBLB	227900_at	1.50	1.68E-09	3.19E-07
MAP7	202890_at	1.49	4.41E-08	2.38E-06
SAMD5	228653_at	1.49	7.42E-09	7.94E-07
MAL2	224650_at	1.48	3.85E-08	2.19E-06
RELN	205923_at	1.47	2.65E-08	1.67E-06
FAT3	236029_at	1.46	6.49E-10	1.97E-07
PGAP1	244321_at	1.45	1.38E-08	1.15E-06
ESRP2	229223_at	1.44	2.03E-07	6.61E-06
PPM1H	212686_at	1.42	3.95E-09	5.48E-07
FAM83H	226129_at	1.42	2.44E-10	1.14E-07
LOC100506713	229833_at	1.42	5.82E-09	6.77E-07
PCDH7	205535_s_at	1.40	2.87E-09	4.51E-07
ST20	217104_at	1.40	2.41E-07	7.51E-06
TMEM45A	219410_at	1.40	4.33E-08	2.35E-06
PPP2R1B	222351_at	1.40	2.16E-09	3.79E-07

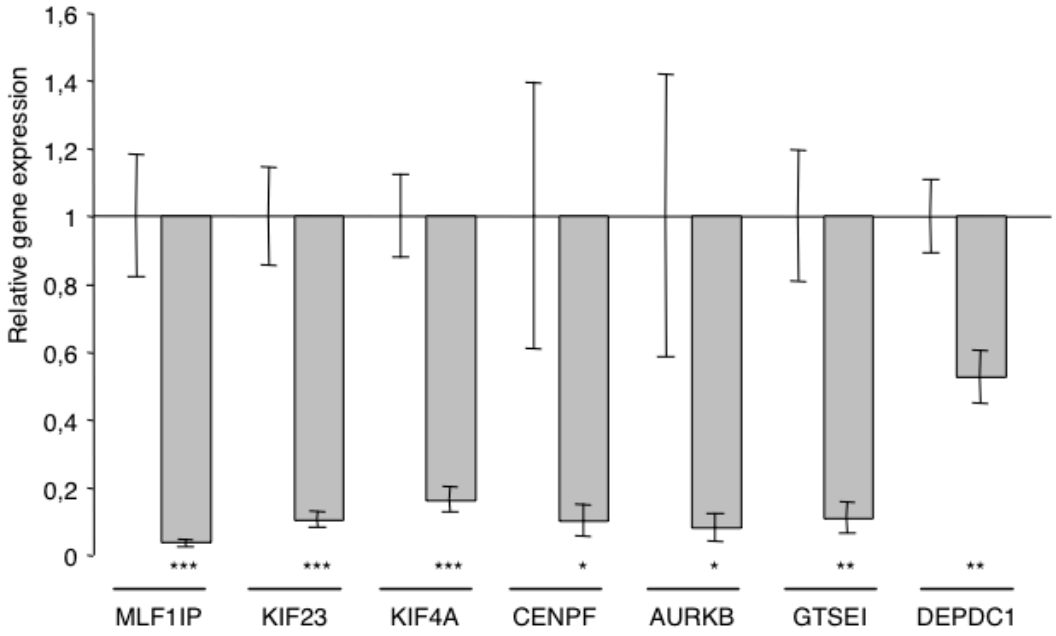

Figure S1: Gene expression analysis after HMGA1 silencing in MDA-MB-231 cells. Downregulation of selected genes after HMGA1 silencing (gray bar) was measured by qRTPCR. Expression was normalized to the level in MDA-MB-231 cells transfected with siCTRL; GAPDH was used as an internal control. Data are presented as the mean ± SD (n=3) (***p<0.001; **p<0.01; *p<0.05).

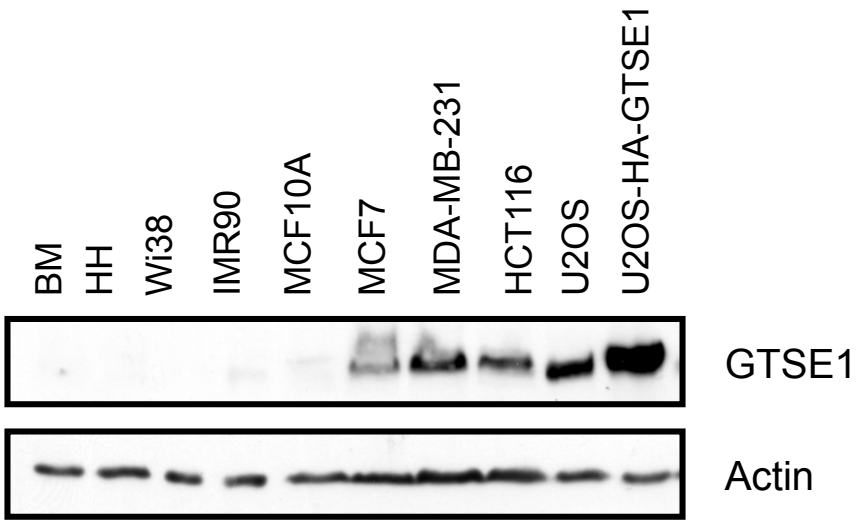

Figure S2: Western blot showing GTSE1 protein levels in transformed and non-transformed cell lines. Non-transformed cell lines are BM (Human bone marrow-derived multipotent adult stem cells), HH (Human heart-derived multipotent adult stem cells), Wi38, and IMR-90. Transformed cell lines are MCF10A, MCF7, MDA-MB-231, HCT116, and U2OS.

Karp-Rabin algorithm

The classic approach to DNA motif discovery is to start by solving a *local multiple alignment* problem, using specific assumptions and statistical hypothesis. In a more general setting, after finding a set of similar—multiply aligned—substrings of the same length as the strings in the input, there is the necessity to compute a barycentric sequence: the "most representative string", according to some criteria, for the found set. This problem is, in general, NP-hard, therefore some constraints must be introduced to cut the search space to a feasible size. In order to find a manageable solution, the consensus problem is then reduced to the *fixed-layout consensus problem*, thereby obtaining the sought computational cut on the search space.

The general strategy used to tackle the problem is based on the introduction of a data structure encoding '*à la* Karp-Rabin'. The Karp-Rabin algorithm (Karp and Rabin 1987) solves the pattern discovery problem on the exact string matching background. This algorithm (reasonably) assumes that it is possible to efficiently shift a vector of bits and efficiently perform arithmetical operations on integers as strings of bits. To take advantage of these assumptions, a string should be seen like an integer, mapping each character of the underlying alphabet Σ in a digit using a function f_m. For example, in the DNA context, $\Sigma = \{A, C, G, T\}$ can be mapped into $\Sigma' = \{0, 1, 2, 3\}$. For a given text string T, let T_r denote the ℓ−length substring of T starting at character r. It is then possible to define the following function:

$$H(T_1) = \sum_{i=1}^{i=\ell} |\Sigma|^{\ell-i} \cdot f_m(T[i]) \quad (1)$$

$$H(T_r) = |\Sigma| \cdot \left(H(T_{r-1}) - |\Sigma|^{\ell-1} \cdot f_m(T[r-1])\right) + f_m(T[r+\ell-1]) \quad (2)$$

The following statement holds: *there is an occurrence of a pattern P starting at position r of T if and only if H(P) = H(T_r).*

Karp and Rabin introduced a method called the *randomized fingerprint* method, which preserves the spirit of the above numerical approach, but allows dealing with larger numbers in an extremely efficient way. It is a *randomized* method because it introduces a probability of error, but the probability that a false match occurs can be bound as stated: *the Karp-Rabin randomized algorithm for pattern matching requires* $\vartheta(n+m)$ *time and has a probability of error* $\vartheta(1/n)$.

The fingerprint function is especially useful when dealing with long patterns. It allows obtaining reasonably small and usable encoding numbers, to compare against each other. In this work, however, we use only the first part of the idea since we are

interested in finding short substrings, a reasonable hypothesis from a biological point of view.

The problem can then be specified as follows: let $\mathcal{F} = \{s_1, \ldots, s_m\}$ be the set of sequences (not necessarily of the same length), *d* the number of variations allowed in comparisons, and *q* a parameter denoting the minimum size of the set of $\mathcal{F}$-elements containing a closest substring (motif) of length ℓ with at most *d* variations. Without loss of generality, we consider input sequences of the same length *n*.

ScanPro algorithm

For a string *s* over Σ, we introduce the following notation:

- $|s|$ denotes the length of s
- $s[i]$ is the *i-th* character of the string s
- $s[i \ldots i + \ell - 1]$ is the substring of ℓ characters starting from $s[i]$
- $s \trianglelefteq t$ denotes that s is a substring of t

The Hamming distance d_H between two strings of the same length is the number of symbols (nucleotides) that are different.

The algorithm aims at solving the *fixed-layout (ℓ, d, q)-consensus problem*, and then extends the results to obtain an approximate solution for the (ℓ, d, q)-consensus problem solution. The difference with solving immediately the former problem lies in the generation step of the consensus sequence and in the final complexity. Furthermore, thanks to the fixed-layout approach, it becomes possible to propose versions of the algorithm that exploit new biological constraints on protein/DNA 3D binding (Sagot et al. 1995).

We classify error layouts into two classes: *basic* and *shifted*. The algorithm exploits the relation between these classes to perform a cost-effective encoding of substrings according to all possible layouts. A generic error layout $f_l = \{i_1, \ldots, i_d\}$, such that $1 \leq i_1 < \ldots < i_d \leq \ell$, is the set of the *d* positions where an error may occur during a comparison between strings. Without loss of generality, we assume that i_1 is strictly greater than 1, that is no error can be in the first position; in this way there are $tfl = \binom{\ell-1}{d}$ error layouts called $FL = \{fl_1, \ldots, fl_{tfl}\}$. *Basic layouts*, bfl, are characterized by having $i_d = \ell$ and are $\binom{\ell-2}{d-1}$ overall. *Shifted layouts* relative to a given basic layout bfl_x are denoted by $sfl_{x,j}$ and look like $\{i_1 - j, \ldots, i_d - j\}$, with $i_1 - j \geq 2$.

For a given error layout, the function fl_code: $\left(\Sigma^{\ell} \times FL \to \mathbb{N}\right)$ gives the encoding of a string of $\ell - d$ characters. For each string $s \in \mathcal{F}$, with $|s| = n$, and each basic layout $bfl_x = \{i_1, \dots, i_d\}$, we have to encode all possible $n - \ell + 1$ substrings of length ℓ in *s*, thus perform $\vartheta(n \cdot \ell)$ operations (for the base case). Considering a shifted layout $sfl_{x,j} = \{i_1 - j, \dots, i_d - j\}$, with $j = 1, \dots, i_1 - 2$, we obtain the encoding of the substring $s[i, \dots, i + \ell - 1]$ for the given layout by taking the encoding of $s[i - 1, \dots, i + \ell - 2]$ for the error layout $sfl_{x,j-1}$, performing a left shift and adding the value relative to $s[i + \ell - 1]$, making $\vartheta(n)$ operations only. [In a left-to-right traveling of the string a right-to-left shift of the error-layout (in the string) takes places. Hence, a recursive computation with base case corresponding to the basic layouts can be -and in fact is- implemented, allowing for occurrences of substrings containing *any* possible error layout to be searched by ScanPro.]

Since $|\Sigma| = 4$, we map the alphabet on $\mathbb{Z}_4$ using only 2 bits for each character. Hence, an ℓ-characters-long string is mapped into a 2ℓ bits integer number, so any shift involves only 2 bits.

The implementation of the algorithm exploits this compact binary representation to improve performances. This idea was suggested by the *Shift-Or* algorithm (Baeza-Yates and Gonnet 1992).

The function fl_code returns a value used as index of an array: in each position *c* of this array there is a pointer to a matrix M_c of dimension $tfl \times (m + 1)$, with $m = |\mathcal{F}|$, whose elements are lists of positions from column 1 to m, while column $m + 1$ contains the "rank", that is the number of different strings $s \in \mathcal{F}$ in which we can find that specific motif. $M_k(i, j) \neq NULL$ if in s_j there exists a substring w such that $|w| = \ell$ and $fl\_code(w, fl_i) = k$. A rank is incremented when an element is put in a empty position in the relative row. If this last value is greater than the quorum, the motif is a solution for the *fixed-layout problem* and we retrieve all positions of the occurrences using a generalized suffix tree. Using these occurrences we can then generate a consensus c_k, according to a consensus string model (i.e. the majority string). Finally, thanks to the previous data structure, it is possible to have in constant time all the positions of the substrings s', such that $d_H(c_k, s') \leq d$.