PhD degree in Molecular Medicine
curriculum in Molecular Oncology

European School of Molecular Medicine (SEMM)
University of Milan and University of Naples "Federico II"
Settore disciplinare Med/04

HOXB7 in lung cancer: a novel role in stem cell and iPS biology

Simona Monterisi
European Institute of Oncology - IEO, Milan.
Matricola n. R09842

Supervisor:	Prof. **Pier Paolo Di Fiore** IEO - IFOM - University of Milan
Added Supervisor:	Dr. **Fabrizio Bianchi** IEO, Milan
Internal Advisor:	Prof. **Saverio Minucci** IEO - University of Milan
External Advisor:	Prof. **Luis M. Montuenga** CIMA - IDISNA - University of Navarra, Pamplona

Accademic Year 2014 – 2015

Abstract

Current diagnostic tools do not allow prognostic evaluation of patients with early stage lung cancer or selection of patients that might benefit from adjuvant chemotherapy. Therefore, the identification of novel prognostic markers in early-stage lung cancer is paramount. In this scenario, the transcription factor HOXB7, belonging to the homeobox family, has been shown to correlate with poor prognosis in different types of cancer and recently also in stage I lung adenocarcinoma.

To better understand the prognostic implication of alterations in HOXB7 expression in lung cancer, we performed a bioinformatics analysis of multiple lung cancer expression datasets in order to identify gene sets representing cancer-relevant biological functions enriched in high-HOXB7 expressing tumors. We found several gene sets enriched in high-HOXB7 expressing tumors representing molecular mechanisms involved in epithelial to mesenchymal transition, in cancer progression, and, interestingly, in stemness and cellular reprogramming. Based on these results, we hypothesized that HOXB7 may have a role in the expansion of the stem cell compartment in cancer, a mechanism that has been shown to be a hallmark of enhanced tumorigenicity and of increased metastatic potential.

Analysis of the stem-related surface marker CD90 revealed that overexpression of HOXB7 in lung cells increases the $CD90^{high}$ sub population. $CD90^{high}$, but not $CD90^{low}$ cells, are able to form spheroids, which is an hallmark of stemness. Indeed, the sphere forming efficiency of normal lung BEAS-2B cells was 22% and 1.64% in $CD90^{high}$ and $CD90^{low}$ populations, respectively. In addition, we found that silencing of LIN28B coun-

teracts the expansion of the $CD90^{high}$ population. LIN28B was recently described as an oncogene that regulates the cancer stem cell compartment. We found that LIN28B is under the direct transcriptional control of HOXB7. Therefore, we propose a novel molecular mechanism driven by HOXB7 and can increase stem-like properties in lung cells.

We further demonstrated that the HOXB7-LIN28B axis plays an important role in reprogramming of adult cells into induced pluripotent stem cells (iPS). Indeed, HOXB7 may enhance the reprogramming efficiency achieved by the three genes OCT4, KLF4, SOX2 in both mouse embryonic fibroblast and human epithelial BEAS-2B cells by substituting MYC in the transcription factor cocktail of reprogramming factors used by Yamanaka. Of note, LIN28B silencing strongly decreases the number of reprogrammed colonies in high-HOXB7 expressing cells.

These findings suggest that HOXB7, through transcriptional induction of the LIN28B gene, activates a program relevant for stem/iPS cell biology and for tumor progression, possibly opening a new line of research for the development of more effective therapies for metastatic lung cancer patients.

Aknowledgments

Deepest gratitude for the help and support are extended to the following persons that make this study possible.

Prof. Pier Paolo Di Fiore, for the opportunity to follow this project in his laboratory and for his advises and guidance during these years.

Fabrizio Bianchi, for sharing his knowledge and supervising me through this project. Thanks for the help in everyday lab-life and for being a constant source of motivations.

Saverio Minucci and Luis M. Montuenga, for their useful suggestions and for the enthusiasm they showed in following this project.

Wessen Maruwge for her time and effort in critically checking this manuscript.

Giuseppe Testa and Pietro Lo Riso, for introducing me to the field of iPS biology.

This PhD project was supported by a fellowship sponsored by Umberto Veronesi Foundation (FUV) and Università degli Studi di Milano.

Contents

List of Figures

List of Tables

List of acronyms

BEBM Bronchial Epithelia Basal Medium

bFGF basic Fibroblast Grow Factor

BPE Bovine Pituitary Extract

ChIP Chromatine ImmunoPrecipitation

CSC Cancer Stem Cell

Ct Cycle threshold

DMEM Dulbecco's Modified Eagle Medium

EMT Epithelial-to-Mesenchymal Transition

E-cad E-cadherin

EGFR Epidermal Growth Factor Receptor

EGFP Enhanced Green Fluorescent Protein

ESC Embryonic Stem Cell

EV Empty Vector

FACS Fluorescence-activated cell sorting

FFPE Formalin-Fixed Paraffin Embedded

FW Forward primer

FBS Fetal Bovine Serum

GSEA Gene Set Enrichment Analysis

HR Hazard Ratio

Glu L-Glutamine

IF ImmunoFluorescence

IP ImmunoPrecipitation

iPS Induced Pluripotent Stem Cell

LUC Luciferase

MEBM Mammary Epithelia Basal Medium

MEF Mouse Embryonic Fibroblast

N-cad N-cadherin

NSCLC Non-Small Cell Lung Cancer

OKS OCT4, KLF4 and SOX2

ORF Open Reading Frame

PI Propidium Iodide

PE R-phycoerythrin Fluorophore

REV Reverse primer

P/S penicillin and streptomycin

RTqPCR Real Time quantitative Polymerase Chain Reaction

SC Stem Cell

SCLC Small Cell Lung Carcinomas

SCR Scramble

SFE Sphere Forming Efficiency

shRNA short hairpin RNA

TCGA The Cancer Genome Atlas

TF Transcription Factor

TKI Tyrosine Kinase activity Inhibitors

TSS Transcriptional Starting Site

UCSC University of California Santa Cruz

WB Western blot

Chapter 1

Introduction

1.1 HOMEOBOX gene family

The French zoologist Geoffroy Saint-Hilaire developed the unity of plan theory (the unity of anatomical structures), shared by all vertebrates, in the early 19th century. According to this theory, the body plane of various animal phyla is arranged in the same way; the homologous parts (the "body units") remain associated in the same order but may differ in form and size. In the 1940s, the American biologist Edward B. Lewis began to study genes that drive the development of the body units in *Drosophila melanogaster (D. melanogaster)*, a highly specialized insect with two wings and three body segments: head, thorax, and abdomen. Mutations in these genes lead to the absence of a body unit, duplication of a body unit or replacement of a unit with another one (homeotic transformation, from the Greek word Homoiosis, translates literally as a process of making similar). Lewis identified a cluster of genes that, when mutated, caused an extra pair of wings (bithorax complex) and the growth of legs from the head instead of the antenna (antennapedia complex; i.e., Antp gene), and hypothesized thus a common ancestry origin [Lewis, 1978].

In the late 1970s, the German biologists Christiane Nusslein-Volhard and Eric F. Wieschaus identified 15 different homeotic genes that control the development of *D.*

melanogaster [Nusslein-Volhard and Wieschaus, 1980]. They were awarded the Nobel prize in 1995 together with Edward B. Lewis, "for discovering the genetic control of early embryonic development" ([1]). The homeobox is now considered "the Rosetta Stone of developmental biology" [Riddihough, 1992] because it specifies the regional identity along the anterior-posterior axes in *D. melanogaster* and in mammals.

Homeotic genes are characterized by a highly conserved DNA sequence (homeobox), composed of 180 bases that codes for a DNA-binding domain. The consensus 60-polypeptide chain of the homeodomain is highly conserved. In fact, the mouse and the Antp human homeobox b7 (HOXB7) differ only in two amino acids (although the lowest common ancestor dates back to more than 500 millions years ago) [Kornberg, 1993].

The homeobox domain folds into three alpha-helices that are organized in a helix-turn-helix structure. The third helix makes contact with the bases of the major groove of DNA and recognizes the HOX-response elements with a TAA(A)T core sequence [Gilbert, 1996].

The homeobox genes duplicated twice during the evolution of invertebrates into vertebrates (Fig.1.1). Only one HOX cluster is present in *D. melanogaster* and it is composed of 8 genes. In the mouse, instead, there are four clusters (HOX A, B, C, and D) composed of $\sim$10 genes on four different chromosomes. Duplication of genes may provide redundant genetic material. As a consequence, corresponding genes on the separate linkage groups, called paralogs, are more functionally related to each other and work together to give regional identity along the anterior-posterior axes. Combination of mutations in more than one paralogous member gives rise to more severe axial phenotypes than a single mutation does [Wellik, 2007].

As an example, the paralogous mouse genes HOXA-11 and HOXD-11 are both involved in the formation of limb bones. A homozygous mutation in one of these genes leads to malformation in the radius and ulna. In a double-mutant mouse, instead,

[1]The Nobel Prize in Physiology or Medicine 1995; http://www.nobelprize.org/nobel_prizes/medicine/laureates/1995/

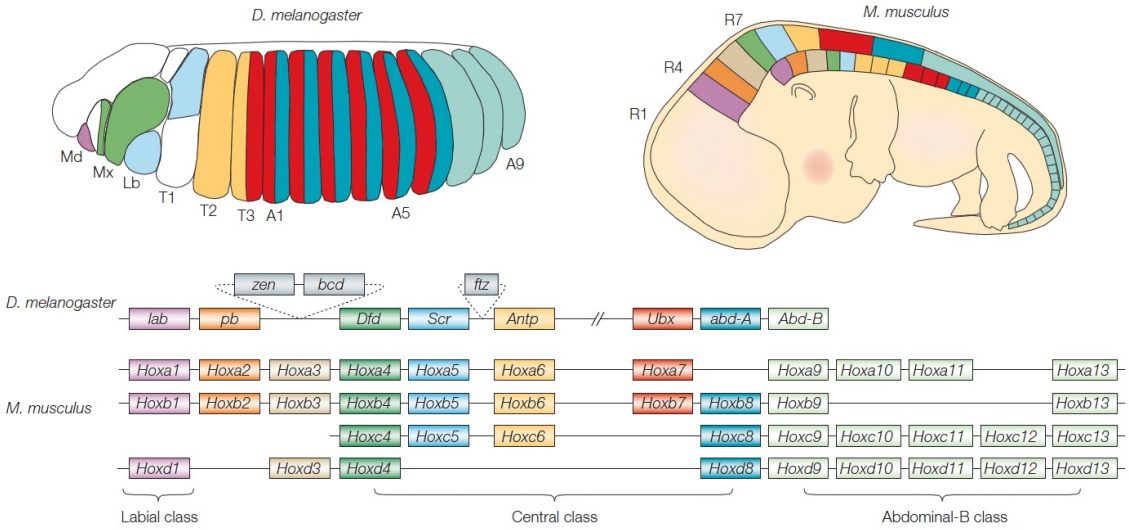


Figure 1.1: **HOX gene clusters.** HOX genes specify the regional identity along the anterior-posterior axes in *D. melanogaster* and in mammals. HOX genes within the cluster are ordered in the 3' to the 5' orientation and paralogous genes are aligned vertically [Pearson et al., 2005]. Homeobox genes duplicated twice during the evolution of invertebrates into vertebrates. Only one cluster is present in *D. melanogaster* and it is composed of eight genes. In the mouse, instead, there are four clusters (HOX A, B, C, and D) composed of ∼10 genes on four different chromosomes. The figure is adapted from Pearson et al., 2005.

these two bones are almost entirely absent [Davis et al., 1995]. There is a partial conservation in HOXA11 gene function in humans. Its involvement in forearm morphogenesis together with HOXA10 has been demonstrated and mutations in HOXA11 have been associated with radio-ulnar synostosis (fusion of the two bones) [Thompson and Nguyen, 2000].

1.1.1 HOX genes and lung development

The expression of HOX genes begins during gastrulation with a spatial and temporal pattern, a phenomena called colinearity [Pearson et al., 2005]. According to this model, the HOX genes are expressed temporally in an order corresponding to their position within the cluster. Thus, genes in the three prime part (3') of each cluster (i.e., HOXA1, HOXB1, HOXA2, HOXB2, etc.) turn on during early gastrulation with a more anterior expression pattern, and genes in the five prime part (5') are expressed later. This temporal delay correlates with progressive generation and growth along the anterior-posterior (AP) axis because 3' genes are generally expressed in anterior tissues and

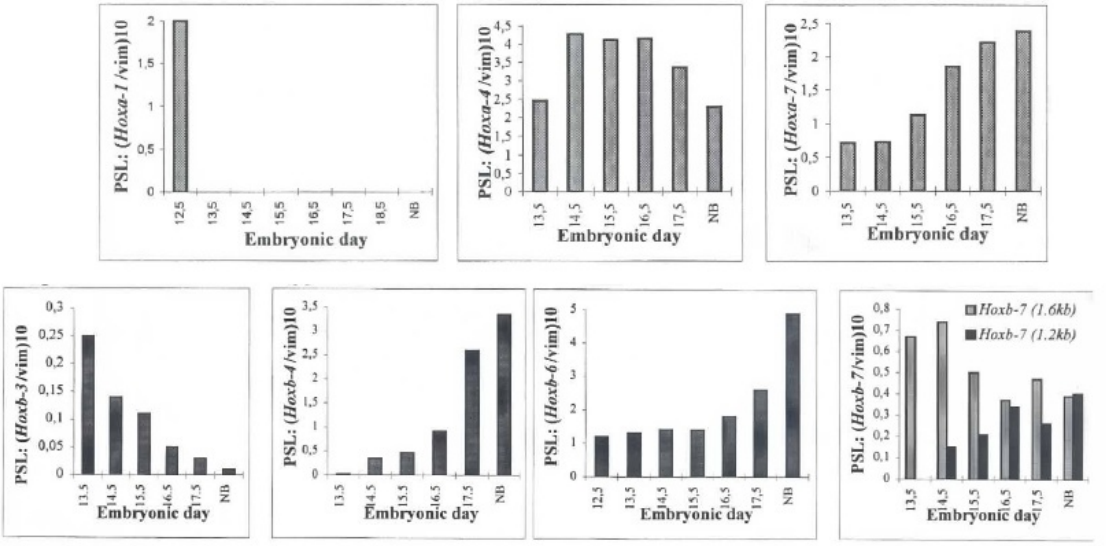


Figure 1.2: **Homeobox gene expression during developing mouse lung.** HOX genes are expressed temporally in an order corresponding to their position within the cluster, so that genes in the 3' part of each cluster (i.e., HOXA1, HOXB3, etc.) are expressed during the first embryonic days, while genes in the 5' part are expressed later. This phenomena is called temporal colinearity [Mollard and Dziadek, 1997].

5' genes are expressed in posterior tissues, after that the anterior somites are formed [Wellik, 2007]. Each body segment is, thus, specified by a combination of functionally active HOX genes, a so called "HOX code". Several studies demonstrate multiple roles for the HOX genes in lung development and maintenance of its functionality in both humans and mice [Grier et al., 2005]. In particular, homeobox genes from clusters A and B are the most highly expressed genes during the branching development of the mouse lung. These genes are characterized by temporal colinearity (Fig.1.2) and differential restrictions in spatial expression domains [Mollard and Dziadek, 1997].

Moreover, expression patterns of HOXB5-9 in the pulmonary mesenchyme of a developing chick have been analyzed [Sakiyama, 2000]. At day four of incubation, before bronchial branching took place, these HOX genes showed a nested pattern of expression, which would correspond to the morphological subdivisions of the lung as morphogenesis proceeded. Indeed, while HOXB5 was expressed throughout the lung and the airways, HOXB6 was expressed in the ventral half. In contrast, HOXB7 and HOXB8 were expressed in the more ventral-distal regions and HOXB9 was first detected

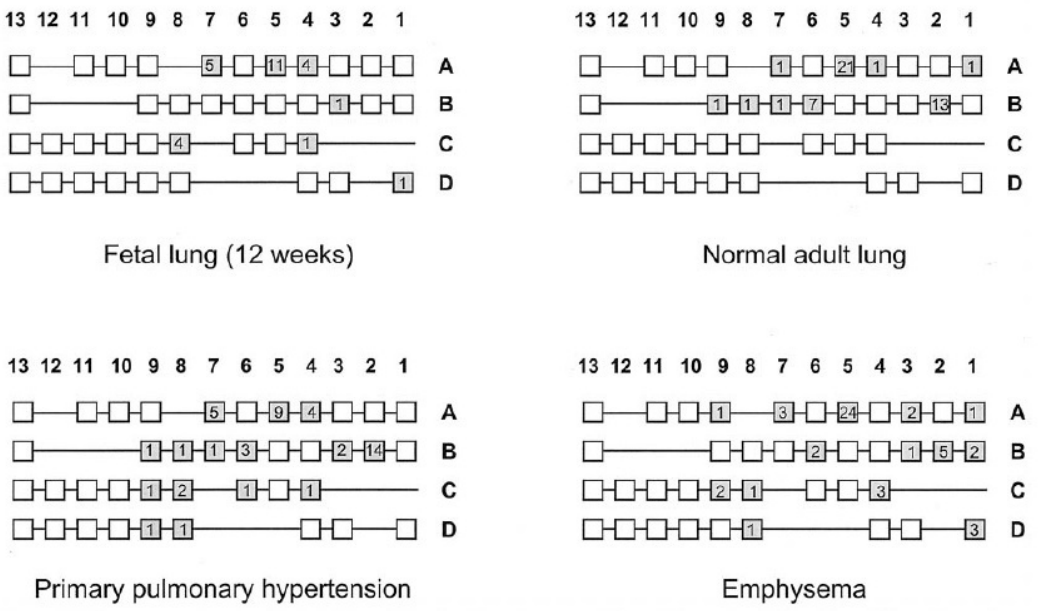


Figure 1.3: **HOX gene expression during developing an adult human lung.** Each box represents a HOX gene: the open boxes represent HOX genes that were not identified by RT-PCR, whereas the shaded boxes represent genes that were detected. The frequency at which positive clones were identified is expressed by the number in the shaded box. The authors suggest that an altered pattern of HOX gene expression may contribute to the development of pulmonary diseases [Golpon et al., 2001].

at day eight in the most distal regions.

The predominantly expressed HOX genes in the human adult lung are those from the 3' end of clusters A and B [Golpon et al., 2001]. Among these, HOXA5 is the most abundant, followed by HOXB2 and HOXB6, while HOXB7 is expressed at low levels. Additional HOX genes from clusters C and D have been found to be expressed during the development of a 12 weeks old fetus, as well as in diseased lung specimens such as emphysema and primary pulmonary hypertension (Fig.1.3).

Due to the role of HOX genes in the development and organogenesis of the embryo and with the breakthrough in the field of stem cells during the last two decades, new efforts have been made to understand HOX genes-driven cell commitment into tissue specific cell lineages.

1.2 HOX genes, stem and cancer stem cells markers

Stem cells (SCs) are defined as cells with the ability to self-renew and to support multilineage differentiation through symmetric and asymmetric cell division. The balance between these two modes of cell division is critical in the maintenance of tissue homeostasis [Alison et al., 2010]. Symmetric cell division gives rise to two daughter cells with SC properties, leading to the expansion of the SC pool. During asymmetric cell division, instead, SCs generate one daughter cell that adopts a SC fate and becomes quiescent and one daughter cell (the progenitor) that enters the transit-amplifying compartment and undergoes rapid proliferation/differentiation, generating mature cells of a particular tissue.

HOX genes were found to be involved in tissue specific stem cell establishment [Shah and Sukumar, 2010]. A detailed study was performed in order to understand the role of HOXB1 in mediating the maintenance and the expansion of posterior neural progenitor cells through the activation of Notch signaling and the JAK/STAT signaling pathways [Gouti and Gavalas, 2008]. The authors suggested that it is possible to drive the differentiation of embryonic stem cells (ESC) toward specific cell fates through the timely expression of specific HOX genes, indicating that HOX genes are important regulators of the very early stages of differentiation. In line with these observations, major studies were performed on hematopoietic stem cell self-renewal. Retrovirus-mediated overexpression of human HOXB4 in mouse bone marrow cells efficiently regenerates the most primitive hematopoietic stem cell compartment, increasing hematopoietic stem cell self-renewal [Sauvageau et al., 1995]. Exogenous HOXB7 in hematopoietic progenitor/stem cells resulted in amplification of the putative hematopoietic stem cell pool and myeloid-restricted progenitor differentiation [Carè et al., 1999]. A growing number of scientific evidence shows that an increment of the SCs compartment is a driving force in tumorigenesis.

1.2.1 Cancer stem cell markers

Cancer stem cells (CSCs) are a rare subpopulation of cells within the tumor, bearing stem cell-like properties. Clinically, the CSC content correlates with disease outcome, indicating that the CSC compartment likely undergoes a greater expansion in tumors with poor-prognosis compared to tumors with a good prognosis [Kim et al., 2005, Pece et al., 2010, Alison et al., 2010]. Therefore, in the last decade, large efforts were made to identify and characterize CSCs in order to better understand their biology and to provide clinically effective tools to improve patient outcome. The first attempt to isolate lung CSCs (using Hoechst 33342 dye exclusion) was reported in 2007 [Ho et al., 2007]. A side population (SP) of several lung cancer cell lines showed an increased tumorigenesis *in vivo* and invasiveness *in vitro*.

Other methods have been proposed for the isolation of human lung CSCs based on: i) the increased activity of ALDH1, already reported to be a SC marker in several human cancers [Jiang et al., 2009]; ii) cell resistance to drugs such as cisplatin, doxorubicin or etoposide [Levina et al., 2008]; and iii) the ability of CSCs to proliferate and form cell clusters (tumor spheres) in absence of adhesion and in serum-free conditions [Eramo et al., 2007]. Isolated lung CSCs cells were subsequently characterized using membrane antigenes expression markers and showed to be positive for several CSC specific biomarkers such as CD166 [Soh et al., 2012], CD133 [Eramo et al., 2007], CD44 and CD90 [Lu et al., 2014, Wang et al., 2013, Yan et al., 2013]. These tumor cells expressed also embryonic markers such as Oct-4, Sox2 and Nanog, confirming thus their undifferentiated phenotype [Eramo et al., 2007].

1.2.2 HOX genes and cancer stem cells

Some HOX genes have been found to be involved in the maintenance of CSC properties. For example, HOXD9 is barely present in normal human brain tissue and astrocytes, but is highly expressed in glioma CSC and in a side population of SK-MG-1 cells (previously identified as enriched stem-like cells) [Tabuse et al., 2011].

In colon, HOXA4 and HOXD10 are expressed in cells at the normal crypt bottom (where stem cells reside), together with the stem markers CD166 and ALDH1. Interestingly, HOXA4 and HOXD10 were found to be overexpressed in colon carcinomas but not in normal colon tissue, suggesting a role for HOX genes in the maintenance of normal stem cells and crypt renewal, contributing to the stem cell overpopulation that drives colon tumorigenesis [Bhatlekar et al., 2014].

Given that HOX genes regulate a variety of cellular processes (including cell adhesion, motility, signaling receptor, differentiation, apoptosis [Morgan, 2006] and, not least, stem cell self-renewal), it is not surprising that an aberrant temporal-spatial expression pattern of HOX genes may drive not only an abnormal development during embryogenesis but also malignancies in adulthood [Pearson et al., 2005].

1.3 HOX genes and cancer

According to the oncogerminative theory of cancer development, HOX genes may drive tumorigenesis by altering the same pathways that they physiologically control during organogenesis [Bhatlekar et al., 2014]. Aberrations in the regulation and expression of HOX genes in cancer were originally described in 1999, when overexpression of HOXA9 was identified by a microarray screening as a marker of poor prognosis in patients with acute myeloid leukaemia [Golub et al., 1999]. Subsequent studies linked the overexpression of HOXA9 to the induction of isulin-like growth factor 1 receptor (IGF1R) and the consequent growth of leukemic cells [Whelan et al., 2008]. From then on, several HOX genes have been linked to cancer processes in a variety of tissues, some of them acting as oncogenes, others as tumorsuppressors [Shah and Sukumar, 2010]. As an example, HOXA5 may act as a tumor suppressor and induce apoptosis by direct transcriptional activation of p53 [Raman et al., 2000], caspase 2 and caspase 8 (involved in the apoptotic signal transduction) [Chen et al., 2004]. Loss of HOXA5 expression through methylation of its promoter has been identified in more than 60%

of breast carcinomas [Raman et al., 2000].

Concerning lung cancer, many members of the HOXC and HOXD clusters such as HOXC4, HOXC8, HOXC9, HOXC13, HOXD8 and HOXD10(not expressed in normal tissue) are strongly up-regulated in primary tumours. In particular, overexpression of HOXC9 and HOXD10 is associated with an increased proliferation rate of the A549 lung cancer cell line [Plowright et al., 2009].

1.3.1 HOXB7 and cancer

HOXB7 was first associated with a higher proliferation rate in melanoma cell lines through the direct transactivation of the basic fibroblast grow factor (bFGF). Indeed, treatment with siRNA against HOXB7 inhibits cell proliferation and expression of bFGF [Caré et al., 1996]. The bFGF promoter has been shown to be a HOXB7 target in the SkBr3 breast carcinoma [Caré et al., 1998] and in IOSE-29 ovarian carcinoma [Naora et al., 2001] cell lines. Increased expression of HOXB7 has been identified in several cancer types: in leukemia [Storti et al., 2011], melanoma [Caré et al., 1996], breast [Caré et al., 1998, Wu et al., 2006, Jin et al., 2012], colorectal [Liao et al., 2011], pancreatic [Nguyen Kovochich et al., 2013] and oral cancer [De Souza Setubal Destro et al., 2010], and, more recently, in lung cancer [Yuan et al., 2014, Bianchi et al., 2007]. Major studies have been performed in breast cancer. Preliminary data showed a 3 and an 18 folds increment in HOXB7 expression in primary and metastatic lesions, respectively, compared with normal tissue [Wu et al., 2006]. Further analyses have demonstrated a potential role for HOXB7 in breast tumorigenesis as a master switch of proangiogenic factors [Carè et al., 2001], and as an epithelial to mesenchymal transition (EMT) promoting factor [Wu et al., 2006]. Transfection of a human normal immortalized breast cell line (MCF10A, negative for HOXB7 expression) with HOXB7 resulted in the activation of the RAS pathway and the acquisition of a phenotype typical of an occurred EMT; i) acquisition of a spindle-like shape; ii) reduction of E-cadherin expression; and iii) expression of Vimentin. EMT is known to correlate in cancer with

an increased malignancy by enhancing proliferation, migration and invasion of cancer cells [Kalluri and Weinberg, 2009]. Alterations in the HOXB7 gene expression may therefore play an important role during tumor progression by conferring, through the induction of EMT, a metastatic potential to cancer cells. Unexpectedly, mammary tumor onset in MMTV-HER-2/neu mice was inhibited when the mice were crossed with MMTV-HOXB7 transgenic mice by ∼6 months. However, after a longer latency the burden of HOXB7 overexpressing tumors was higher and, importantly, cancer cells were more aggressive giving rise to more and larger metastasis [Chen et al., 2008]. Of note, no tumors were detected in HOXB7 overexpressing mice during two years of observation, suggesting a role for HOXB7 in tumor progression rather than in tumorigenesis. Primary mammary cell lines derived from MMTV-HOXB7/HER-2 showed an increased expression of TGFβ2, which was identified as a HOXB7 transcriptional target [Liu et al., 2015]. In two recent works, overexpression of HOXB7 in breast cancer was associated with poorer disease-free survival in estrogen receptor (ER) positive patients treated with tamoxifen. The authors showed how breast cancer cells may acquire tamoxifen resistance through HOXB7-direct activation of the epidermal growth factor receptor (EGFR) [Jin et al., 2012] and HER2 expression [Jin et al., 2015].

1.4 Lung cancer and HOXB7

Lung cancer alone accounts for more than one-quarter of all cancer deaths in both women and men in the Unites States of America [Siegel et al., 2015] and in men in Europe [Ferlay et al., 2013]. The low survival rate (18% at 5 years after diagnosis) is primarily due to the high frequency of late diagnosis when the tumor has become unresectable. Emerging lung cancer screening programs for high-risk individuals (>55years, >30 pack-year) using low-dose Computed Tomography (LDCT) were recently shown to be effective, reducing lung cancer mortality with ∼20% [Med, 2011], underlying the importance of early diagnosis in lung cancer.

1.4.1 Non Small Cell Lung Cancer staging

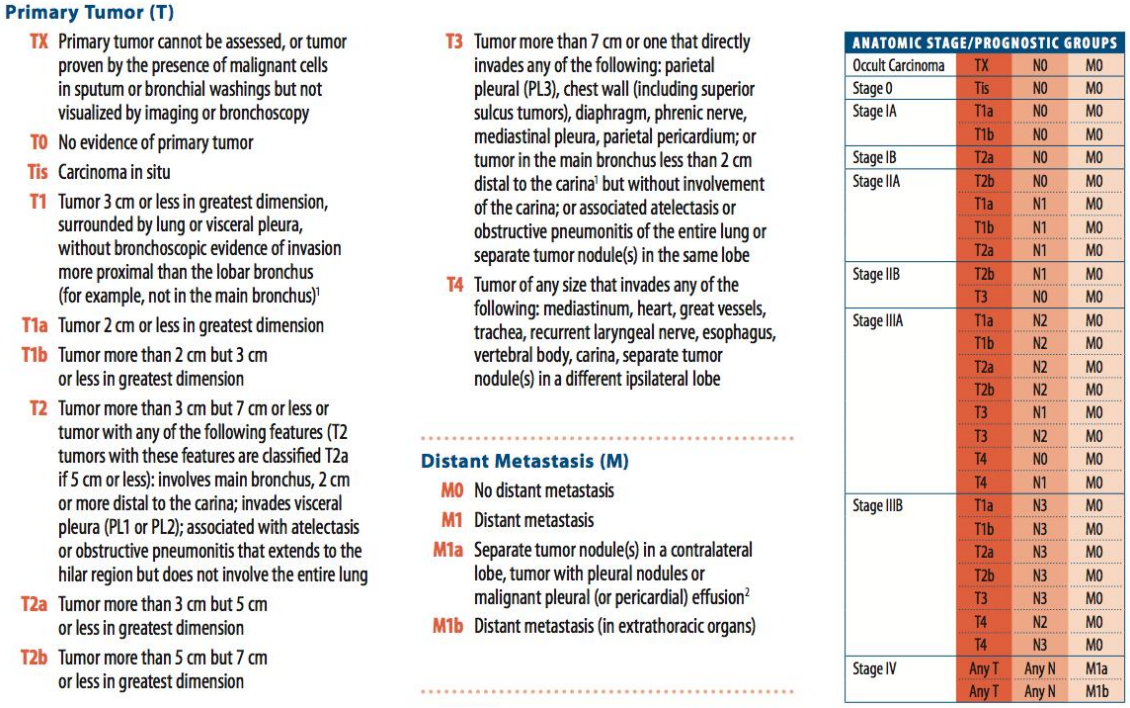
The two main types of lung cancer are small cell lung carcinomas (SCLC), characterized by a scarce cytoplasmic content, and all others, grouped together as Non-Small Cell Lung Cancer (NSCLC) due to their similar, yet different from that of SCLC, response to treatment. SCLC show a good initial response to chemotherapy but most patients relapse within 5 years; surgery treatment in SCLC is rarely used due to frequent disease spread at diagnosis [Jett et al., 2013].

NSCLC have a limited response to chemotherapy treatment but are better characterized genetically. New targeted therapies were recently proposed.

NSCLC accounts for more than 80% of lung cancers. According to the World Health Organization's classification of lung tumors they are divided into different types based on the type of cells found in the tumor: squamous cell carcinoma, large cell carcinoma and adenocarcinoma (the most predominant histological type, accounts for 32% of all lung cancer) [Brambilla et al., 2001]. Adenocarcinomas commonly invade pleura and mediastinal lymph nodes and metastasize to the brain and bones. Patients often have a metastatic disease before the development of symptoms [Hirsch et al., 2008].

To determine the course and spread of lung cancer, a system based on three letters (T, N and M) developed by the American Joint Committee for Cancer Staging and End Results Reporting is used in the clinic. The letter T represents tumor size, N represents regional lymph node involvement, and M represents distant metastases. Numeric subscripts indicate the degree of dissemination. Stage I NSCLC is defined by the clinical stage groupings T1 (or T2a), N0 and M0 which designates a small localized tumor. Stage II NSCLC is defined as T1 (or T2), N1 and M0 which is a primary tumor that has extended to regional nodes, or as T3 (or T2b), N0 and M0 characterized by a tumor mass bigger than 5cm (Fig.1.4).

Surgery is the treatment of choice for patients with stage I to IIIA NSCLC [Crinò et al., 2010]. About 73% of patients with surgical-pathologic stage IA disease are expected to survive more than 5 years following complete resection (58% for stage



Primary Tumor (T)

- **TX** Primary tumor cannot be assessed, or tumor proven by the presence of malignant cells in sputum or bronchial washings but not visualized by imaging or bronchoscopy
- **T0** No evidence of primary tumor
- **Tis** Carcinoma in situ
- **T1** Tumor 3 cm or less in greatest dimension, surrounded by lung or visceral pleura, without bronchoscopic evidence of invasion more proximal than the lobar bronchus (for example, not in the main bronchus)[1]
- **T1a** Tumor 2 cm or less in greatest dimension
- **T1b** Tumor more than 2 cm but 3 cm or less in greatest dimension
- **T2** Tumor more than 3 cm but 7 cm or less or tumor with any of the following features (T2 tumors with these features are classified T2a if 5 cm or less): involves main bronchus, 2 cm or more distal to the carina; invades visceral pleura (PL1 or PL2); associated with atelectasis or obstructive pneumonitis that extends to the hilar region but does not involve the entire lung
- **T2a** Tumor more than 3 cm but 5 cm or less in greatest dimension
- **T2b** Tumor more than 5 cm but 7 cm or less in greatest dimension
- **T3** Tumor more than 7 cm or one that directly invades any of the following: parietal pleural (PL3), chest wall (including superior sulcus tumors), diaphragm, phrenic nerve, mediastinal pleura, parietal pericardium; or tumor in the main bronchus less than 2 cm distal to the carina[1] but without involvement of the carina; or associated atelectasis or obstructive pneumonitis of the entire lung or separate tumor nodule(s) in the same lobe
- **T4** Tumor of any size that invades any of the following: mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, carina, separate tumor nodule(s) in a different ipsilateral lobe

Distant Metastasis (M)

- **M0** No distant metastasis
- **M1** Distant metastasis
- **M1a** Separate tumor nodule(s) in a contralateral lobe, tumor with pleural nodules or malignant pleural (or pericardial) effusion[2]
- **M1b** Distant metastasis (in extrathoracic organs)

ANATOMIC STAGE/PROGNOSTIC GROUPS			
Occult Carcinoma	TX	N0	M0
Stage 0	Tis	N0	M0
Stage IA	T1a	N0	M0
	T1b	N0	M0
Stage IB	T2a	N0	M0
Stage IIA	T2b	N0	M0
	T1a	N1	M0
	T1b	N1	M0
	T2a	N1	M0
Stage IIB	T2b	N1	M0
	T3	N0	M0
Stage IIIA	T1a	N2	M0
	T1b	N2	M0
	T2a	N2	M0
	T2b	N2	M0
	T3	N1	M0
	T3	N2	M0
	T4	N0	M0
	T4	N1	M0
Stage IIIB	T1a	N3	M0
	T1b	N3	M0
	T2a	N3	M0
	T2b	N3	M0
	T3	N3	M0
	T4	N2	M0
	T4	N3	M0
Stage IV	Any T	Any N	M1a
	Any T	Any N	M1b

Figure 1.4: **American Joint Committee on Cancer classification of Lung Cancer Staging.** Figure adapted from AJCC 7th edition staging posters

IB). There is no indication for adjuvant chemotherapy in stage I NSCLC at present [Le Chevalier et al., 2005], even if it would be useful in improving patients prognosis. Stage III NSCLC patients (any T, N3, M0) do not benefit from surgery alone and are best managed by induction chemotherapy (neoadjuvant) with cispaltin-based drugs and adjuvant chemotherapy or radiotherapy depending on the sites of tumor involvement and the performance status of the patient after surgery. Still, more than 45% of patients diagnosed with NSCLC have an advanced disease with distant metastases (i.e., stage IV) [Walters et al., 2013], characterized by very large lesions involving distant sites (Any T, any N, M1). Surgery is not an option and the benefit of cisplatin-based chemotherapy to improve survival, which is used to palliate disease-related symptoms, is dismal.

In recent years, the oncology community recognized NSCLC as a heterogenous disease thanks to the identification of "druggable" target mutations, such as EGFR, KRAS, ALK, MET and ROS-1 (Fig.1.5). These findings have allowed the selection of patients for targeted therapies and have improved the prognosis of patients with

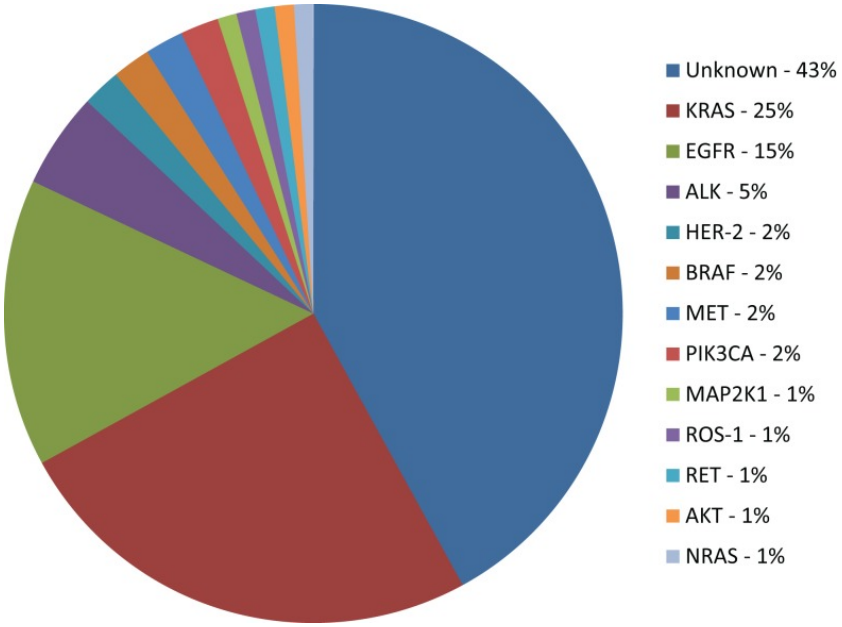


Figure 1.5: **Molecular subsets of lung adenocarcinoma.** The most common mutations in lung adenocarcinoma. Figure adapted from Korpanty et al. [Korpanty et al., 2014].

advanced NSCLC [Korpanty et al., 2014].

1.4.2 Non Small Cell Lung Cancer biomarkers

Preclinical and clinical studies have identified EGFR mutations and amplification as early events in the development of NSCLC, predominantly in adenocarcinoma [Herbst et al., 2008]. The epidermal growth factor receptor (EGFR) is a transmembrane tyrosine kinases receptor. Mutations can occur in its catalytic domain, resulting in a constitutive kinase activity. Since EGFR activates both the PI3K-Akt-mTOR and the RAS-MAPK pathways, these events may lead to an enhancement in proliferation, resistance to apoptosis, invasion and angiogenesis. Mutations in the EGFR in NSCLC are associated with good prognosis as they increase the sensitivity to EGFR tyrosine kinase activity inhibitors (EGFR-TKIs) such as gefitinib and erlotinib. Nevertheless, most patients that show an initial response will eventually relapse. The second-generation EGFR-TKI afatinib demonstrated an increased overall response rate and better progression-free and overall survival in a clinical setting. First-line treatment with EGFR-TKIs is now clinically approved for EGFR mutation-positive NSCLC

[Korpanty et al., 2014].

Another early event in lung cancer are mutations in KRAS, identified already in 1984: 30% of NSCLC are KRAS mutated, especially adenocarcinomas. K-RAS encodes for a small GTPase-protein that can switches from an inactive (GDP-binding state) to an active (GTP-binding) state. When it is active, the K-RAS protein activates a signal transduction cascade, involving MAP Kinases, that leads to cell growth and proliferation. Mutations are usually a single missense mutation at codons 12 (e.g., from glycine to valine G12V) and less frequently at codons 13 and 61, involved in the amino acid exchanges. Mutations prevent the hydrolysis of GTP, intrinsic or catalyzed by the GTPase-activating proteins (GAPs), thereby generating constitutively active and potentially oncogenic RAS molecules.

EGFR and KRAS mutations are mutually exclusive, which can be due to the fact that KRAS-MAPK is an important downstream signaling pathway of the EGFR, explaining why KRAS mutation-positive NSCLC (which were associated with poor prognosis already in the 1990s) are not sensitive to EGFR-TKIs [Suda et al., 2010]. Although mutations in KRAS are the most common mutations in NSCLC, targeted therapies specific for patient presenting mutations in KRAS are not available.

The third most common molecular event occurring in lung adenocarcinoma is ALK translocation and gene fusion (5%) [Korpanty et al., 2014]. The resulting protein shows a constitutive activity and confers sensitivity to TKIs, such as crizotinib and ceritinib.

A target mutation has not been identified in 43% of lung adenocarcinomas [Korpanty et al., 2014]. In conclusion, the fraction of lung cancer patients for whom targeted therapies are available, is still very small. Novel efforts must be dedicated to the identification of new actionable molecular alterations, with significant diagnostic, prognostic, or therapeutic implications, to improve therapy response and prognosis.

1.4.3 Early stage lung cancer biomarkers and HOXB7

Although 73% and 58% of patients with stage IA and IB disease, respectively, are expected to survive more than 5 years following complete resection [Crinò et al., 2010], about 27% develop second primary tumors and/or regional and distant metastases [Martini et al., 1995]. Current diagnostic tools, however, do not allow a precise prognostic evaluation of patients with early stage lung cancer. Selection of patients that might benefit from adjuvant chemotherapy is, thus, not possible. As a consequence, at present there is no indication for adjuvant chemotherapy NSCLC patients with stage I disease [Le Chevalier et al., 2005, Crinò et al., 2010]. The availability of accurate prognostic markers might change this picture by allowing i) the selection of patients at high risk of relapse to be included in clinical trials for new therapeutic strategies and ii) the treatment of high risk patients with stage I disease as patients with more advanced tumors.

Recently, we showed that *in silico* approaches can be effective for the identification of potentially relevant cancer biomarkers. For example, we identified microRNA-based biomarkers capable to risk stratify breast cancer patients [Monterisi et al., 2015]. The study was based on meta-analysis of publicly available gene expression datasets in order to extract information about the expression of intronic-microRNAs from the profile analysis of the expression of their host genes. The microRNA signature composed by miR-342, miR-483 and miR-1266 was capable to identify sub-types of breast cancer with a high metastatic potential (see Appendix chapter 5).

Concerning lung cancer, previous studies in the laboratory led to the discovery of a 10-gene prognostic signature in stage I lung adenocarcinoma, accurate enough to correctly predict metastatic disease in patients with stage I lung cancer [Bianchi et al., 2007]. Our signature showed an accuracy of 75% in an independent cohort of patients with stage I disease, and outperformed other clinicopathological parameters such as tumor stage (IA vs. IB), grading, age, sex and the presence of K-RAS mutations (Fig.1.6).

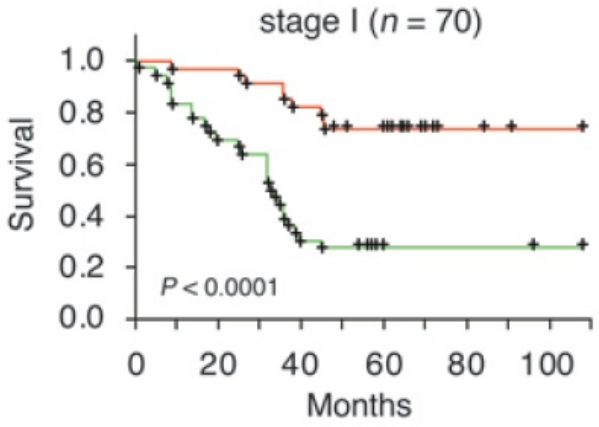


Figure 1.6: **10-gene prognostic signature in stage I lung adenocarcinoma.** The plot reports the 10-gene prognostic model applied to a cohort of stage I patients. Data are shown as the probability of survival (y-axes) over time as a function of a favorable (red line) or unfavorable (green line) signature [Bianchi et al., 2007].

Recently, we performed an extensive validation of the 10-gene signature using formalin-fixed paraffin embedded (FFPE) tumor samples from an additional and independent cohort of 507 lung cancer patients from the IEO hospital, including 351 stage I lung adenocarcinomas. Stage I lung adenocarcinoma patients classified as high-risk patients by the 10-gene signature displayed a 4-fold increase in the risk of death three years post-surgery (univariate analysis: hazard ratio (HR)=4.0, p=0.03; multivariate analysis: HR=4.2, p=0.03,). Of note, the risk model improved the prediction when the "10-gene risk" was added to traditional clinical and pathological parameters (age, sex, smoking status and tumor stage; p=0.01, nested likelihood ratio test) were included (E. Dama et al. *manuscript in preparation*).

Our 10-gene signature includes HOXB7 among other genes (E2F1, E2F4, MCM6, RRM2, SF3B1, NUDCD1, SERPININB5, HSPG2 and SCGB3A1). A recent work using an independent approach (i.e., immunohistochemistry) confirmed, indeed, that HOXB7 is a prognostic biomarkers in lung adenocarcinoma [Yuan et al., 2014]. A precise molecular mechanism through which HOXB7 promotes lung cancer progression and metastatic spreading remains unknown. We believe that understanding the role of HOXB7 in lung cancer could be useful in the identification of actionable targets to develop novel therapeutic strategies for metastatic lung cancer.

Chapter 2

Materials and methods

Cell culture

A549, NCI-H358, HEK-293T and Phoenix-amphotrophic cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 2 mM L-glutamine (Glu), 100 U/ml penicillin and 100 U/ml streptomycin (P/S) at 37°C in a humidified incubator with 5% CO_2. BEAS-2B cells were cultured in Bronchial Epithelia Basal Medium (BEBM) supplemented 0.5 ng/ml EGF, 500 ng/ml hydrocortisone, 0.005 mg/ml insulin, 0.035 mg/ml Bovine Pituitary Extract (BPE), 500 nM ethanolamine, 500 nM phosphoethanolamine, 0.01 mg/ml transferrin, 6.5 ng/ml 3,3',5-triiodothyronine, 500 ng/ml epinephrine and 0.1 ng/ml retinoic acid at 37°C in a humidified incubator with 5% CO_2 in fibronectin coated plates.

Mytomycin-treated Mouse Embryonic Fibroblast (MEF mitotically inactive) and mitotically active Oct4-EGFP-MEFs were cultured in DMEM supplemented with 20% FBS, 100 U/ml P/S and 2 mM Glu at 37°C in a humidified incubator at low oxygen condition (3%). Mytomycin-treated MEFs require 0.1% gelatin coated plates.

Primary cells were established from lung primary tumors from patients operated at the European Institute of Oncology (Milan, Italy) and that had given their informed consent for the research use of human biological materials. Tissues were dissociated mechanically into small pieces with scissors and digested enzymatically with collagenase

type 1 (Invitrogen, 150 u/ml collagenase in DMEM supplemented with P/S and Glu). Cell suspensions were filtered through a 70 μm filter to eliminate cell aggregates. Red blood cells were lysed using ACK Lysing Buffer (Lonza). Epithelial cells were plated in adhesion conditions on collagen coated plates and grown at 37°C in a humidified incubator with 5% CO_2 in EPI LUNG medium (HAM'S F12 and DMEM 1:1, 1% FBS, 1% Glu, 1% P/S, 0.2% amphotericin, 10 μg/ml transferrin, 1 μg/ml hydrocortisone, 1 μg/ml insulin, Hepes pH 7.5 10 mM, 50 μM ascorbic acid, 15 nM sodium selenite, 0.1 mM ethanolamine, 50 ng/ml colera toxin, 10 nM EGF, 35μg/ml BPE and 10 nM T3).

Colony formation assay

A fixed number of cells was seeded in complete medium in 10 cm plates and incubated for 10 days: i) 500 BEAS-2B cells; ii) 5.000 A549 cells; and iii) 1.000 NCI-H358 cells. Colonies were visualized by staining with 1ml/plate of crystal violet (1% w/v in 35% EtOH, Santa Cruz) for 5 minutes at room temperature. The number of colonies was determined using the analyze particles tool of the ImageJ program.

Cell proliferation assay

BEAS-2B cells were seeded in triplicates in 6-well plates (40.000 cells/well). Biorad TC10 automated cell counter was used to count cells every 24 hours for 4 consecutive days.

Soft agar assay

1% and 0.5% low melting agarose was used for the bottom and the top agar, respectively. NCI-H358 cells were seeded in triplicates within the top agar (10.000 cells/ml; 4ml/well in 6-well plates) and allowed to grow for three weeks at 37°C. Colonies were stained with 5mg/ml Thiazolyl Blue Tetrazolium Bromide (MTT, Sigma).

Transwell migration assay

NCI-H358 cells were starved for 24 hours in serum depleted medium, plated on the upper layer of a cell permeable membrane and incubated for 14 hours at 37°C. We used 10% serum medium as chemotactic stimulus and 0% medium as control. Migrated cells were stained with DAPI and images were acquired with a fluorescent microscopy, whereafter cells were counted.

Pneumosphere assay

Cells were plated in low adherent 12 well plates (coated with poly-HEMA, Sigma Aldrich) at a density of 1500 cells/ml in serum-free mammary epithelial medium (MEBM, Lonza) supplemented with 1% glutamine, 1% penicillin-streptomycin antibiotics, 5 μg/ml insulin, 0.5 μg/ml hydrocortisone, 2% B27 (Invitrogen) 20 ng/ml EGF and human b-FGF, 4 μg/ml heparin and 1 mg/ml methylcellulose (Sigma Aldrich).

All images were acquired with a DMI6000 B Leica microscope using a 5x objective lens and LAS-AF image software (Leica). The sphere forming efficiency (SFE) was calculated automatically using a home-made macro developed with Java and run on the Fiji software (ImageJ). Criteria for sphere selection were: area $\geq$ 50 μm; roundness = 0.30-1.00. %SFE = (n° of spheres / n° of plated cells) x 100

Stable gene overexpression and gene silencing

For HOXB7 overexpression, HOXB7 open reading frame (ORF, 674bp) was cloned into the EcoRI site of a standard retroviral pBABE vector (carrying Hygromicin resistance cassette). Primer sequences: FW- GCCAAATTATGAGTTCATTG; REV-TTTCTCCATCCCTCACTCTT.

For LIN28B overexpression, we used a pBABE-hLIN28B carrying puromycin resistance (addgene).

Production of retroviral particles was performed, briefly, through transfection of semi-confluent Phoenix-amphotrophic cells by calcium phosphate precipitation in the

presence of 25 μM chloroquine. Supernatants of transfected cells were collected twice every 12 hours and freshly used to infect target cells.

For HOXB7 dowregulation, two different short hairpin oligos were designed using the web available tool pSICO - oligomaker 1.5 (Table.2.1). Forward and reverse annealed oligos were cloned into the pSICOR vector according to Jacks Lab's protocol (http://web.mit.edu/jacks-lab/protocols/pSico.html). For LIN28B silencing, we designed three different short hairpin RNAs (Table.2.1). Production of lentiviral particles was achieved by plasmids expressing viral proteins GAG, POL, ENV, REV and co-transfection with lentiviral vectors into semi-confluent HEK-293T cells by calcium phosphate precipitation in the presence of 25 μM chloroquine. Supernatants of transfected cells were collected 36 hours post transfection and concentrated by centrifugation. Viral particles were resuspended in PBS and used for infection after freezing at -80°C.

TUNEL assay

A549 cells were harvest 48 hours post infection with pSICOR vectors, fixed with 2% formaldehyde, stained with In Situ Cell Death Detection Kit, Fluorescein (Roche) and propidium iodide (PI, 2.5μg/ml) and acquired on a FACS Calibur instrument (BD Biosciences).

CD90 staining and FACS-sorting

PE hCD90 (BD Pharmingen) staining was performed on lung cancer cell lines and primary cells. 5x10^6 cells were stained with the antibody (pure) for 1 hour at room temperature and subjected to FACS sorting: Influx cell sorted equipped with a 488 nm laser and with a band pass 575/26 nm optical filter for R-phycoerythrin (PE) detection (BD). $CD90^{high}$, $CD90^{low}$ and $CD90^{intermediate}$ cells were collected.

Gene to silence	oligo name	oligo sequence
HOXB7	sh - FW2	TGAAAGAATTTCACTACAATTTCAAGAGAATT GTAGTGAAATTCTTTCTTTTTTC
HOXB7	sh - REV2	TCGAGAAAAAAGAAAGAATTTCACTACAATT CTCTTGAAATTGTAGTGAAATTCTTTCA
HOXB7	sh - FW3	TGACTGTGGGTCTGGACTAATTCAAGAGATT AGTCCAGACCCACAGTCTTTTTTC
HOXB7	sh - REV3	TGACTGTGGGTCTGGACTAATTCAAGAGATT AGTCCAGACCCACAGTCTTTTTTC
LIN28B	sh - FW1	TGCAGAGATCTCAGAACGGTTTCAAGAGAAC CGTTCTGAGATCTCTGCTTTTTTC
LIN28B	sh - Rev1	TCGAGAAAAAAGCAGAGATCTCAGAACGGTT CTCTTGAAACCGTTCTGAGATCTCTGCA
LIN28B	sh - FW2	TGTATAGGGGAACAGTATTTTTCAAGAGAAAA TACTGTTCCCCTATACTTTTTTC
LIN28B	sh - Rev2	TCGAGAAAAAAGTATAGGGGAACAGTATTTTC TCTTGAAAAATACTGTTCCCCTATACA
LIN28B	sh - FW3	TGCAGCTGCACTGACTTTAATTCAAGAGATTA AAGTCAGTGCAGCTGCTTTTTTC
LIN28B	sh - Rev3	TCGAGAAAAAAGCAGCTGCACTGACTTTAATC TCTTGAATTAAAGTCAGTGCAGCTGCA

Table 2.1: **List of short hairpin oligos.** Short hairpin oligos were designed using the web tool pSICO - oligomaker 1.5 in order to be cloned into the pSICOR vector to obtain gene silencing according to Jacks Lab's protocol (http://web.mit.edu/jacks-lab/protocols/pSico.html). FW = forward oligo; REV = reverse oligo.

Dual-Luciferase Reporter Assay

Promoter sequences of LIN28B were cloned upstream of a promoterless firefly luciferase reporter cassette of pGL3 basic vector (Promega). Customized primers were designed in order to amplify the promoter sequences of interest from the BAC clone RP11-633H9 (Table.2.2). For the cloning strategy, we took advantage of the one-step TOPO TA Cloning, passing through an intermediate into pCR2.1 TOPO vector.

Region	Primer sequence	Amplicon
2Kb	FW-TGGTAGTGGACTTTTAAAATGTCAG REV-TTGGTGTACAAATACATCGACTGGA	2239
1Kb	FW-TGGTAGTGGACTTTTAAAATGTCAG REV-CGTGACTTTGTCAATTACATGC	1016
Intron	FW-TACGCTCGAGAAGGAACAGGACAAAAAAGT REV-TACGCTCGAGAAGGAACAGGACAAAAAAGT	1066

Table 2.2: **List of primers used for cloning of LIN28B promoter constructs.** Customized primers were designed in order to amplify the promoter sequences of interest from the BAC clone RP11-633H9: 2Kb, 1Kb and Intron. FW = forward primer; REV = reverse primer. Amplicon length is expressed in base pair.

Phoenix cells were transfected with the pBABE-HOXB7 or pBABE-EV construct in combination with the different pGL3 constructs and the *Renilla* luciferase expression vector under the control of the herpes virus thymidine kinase (HSV-TK). Luciferase assay was performed according to Dual-Glo Luciferase Assay System Protocol (Promega).

Chromatin immunoprecipitation assay

HOXB7 ORF was cloned into a pCDNA3.1 N-term-FLAG frame A vector in EcorI site. A549 cells were transfected with LIPO2000 reagent (Invitrogen) in order to express the HOXB7-FLAG protein. 48 hours post infection, cells were fixed with 1% formaldehyde and chromatine was sonicated with Branson Digital Sonifier. For immunoprecipitation, anti-FLAG M2-Agarose beads (Sigma), Anti-Histone H3 antibody (ab1791 abcam) and protein A Sepharose CL-4B (GE Healthcare) were used. 1% of DNA (obtained through immunoprecipitation after de-crosslinking) was used for PCR amplification with QuantiFast SYBR Green PCR Kit (Qiagen) using specific primers

designed with the Primer3 software (http://primer3.ut.ee/) along the LIN28B promoter sequence (Table.2.3). The Chan2 and Chan3 primers were used by Chang and colleagues [Chang et al., 2009]. the EGFR primers are described in [Jin et al., 2012].

Primer	Primer sequence	In silico PCR / Reference
a	FW-CTATCCTGGCGGCTCCTCT REV-CCCAAGAGCTGGAGGACATA	chr6:105404070+105404147
b	FW-GCCCTATGTCCTCCAGCTC REV-CCCAGTTTTCCAGCCTAACA	chr6:105404124+105404198
c	FW-FWGCTGGAAAACTGGGCTGTTA REV-CACAGGTTTCTCTGCCATCTC	chr6:105404185+105404262
d	FW-ACTGCCATGGAATAGCTGAA REV-GGGAGGGGGTCGTTTAAATA	chr6:105,404,551-105,404,627
Cha2	FW-TGTAATTGACAAAGTCACGTGTGC REV-TCCTCTCTCCAGTTTCTGGCC	Chang et al. 2009
Cha3	FW-GCAAATAACGCTGGATTCAGTG REV-AGAGCTACTAGTTAAGGCACATGGG	Chang et al. 2009
EGFR	FW-CAAGGCCAGCCTCTGAT REV-CCCCTTTCCCTTCTTTTGTT	Jin et al. 2012

Table 2.3: **List of primers used for ChIP.** Customized primers a, b, c and d were designed with Primer3 software along LIN28B promoter sequence. The Chan2 and Chan3 primers are described in [Chang et al., 2009] and the EGFR primer in [Jin et al., 2012].

Cell reprogramming

Mitotically active MEFs were infected with the lentiviral vector expressing OCT4, KLF4 and SOX2 (OKS, was a kind gift from Dr. Naldini) and with the pBABE-HOXB7 or pBABE-EV viral constructs produced as mentioned above. One day after infection, cells were harvested and cultured on mitotically inactivated monolayer feeder cells in ESC medium (DMEM-F12 1:1 supplemented with 20% KnockOut^{TM} Serum Replacement, 1 mM L-glutamine, 100 U/ml penicillin, and 100 U/ml streptomycin, 0,1 mM non-essential amino acids, 1/500 home-made leukaemia inhibitory factor and 0.1 mM 2-β-mercaptoethanol) at 37°C in a humidified incubator at low oxygen conditions (3%).

A slightly different ESC medium was used to reprogram BEAS-2B cells after OKS infection: DMEM-F12 1:1 supplemented with 20% KnockOut^{TM} Serum Replacement,

1 mM L-glutamine, 100 U/ml penicillin, 100 U/ml streptomycin, 0,1 mM non-essential amino acids, 10 ng/ml FGF and 0.05 mM 2-β-mercaptoethanol)

Alkaline phosphatase assay

Alkaline phosphatase staining was performed using the Vector Red Alkaline Phosphatase Substrate kit (Vector Laboratories), following manufacture's instructions, two and three weeks post infection for MEF and BEAS-2B reprogrammed cells, respectively.

Teratoma assay

iPSC were cultured in ESC medium for three passages. 2x10^6 cells were injected subcutaneously into NOD-SCID mice. The mice were sacrificed by cervical dislocation 2-3 weeks after injection and tumors were isolated and fixed in 4% formaldehyde.

Western blot assay

Cells were harvested in Ripa buffer (50 mM Tris HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% Triton, 0.2% SDS, 1% NaDeoxicolate and Protein Inhibitors cocktail CALBIOCHEM). Western blott was performed according to standard procedures: samples were prepared in Laemmli loading buffer (5x: 10% SDS, 50% Glycerol, 300mM TRIS HCl, 0.5 M DTT (Dithiothreitol) and 0.02% Bromophenol blue), loaded into 10% acrylamide gels and transferred to nitrocellulose membranes. Antibodies used against human proteins: HOXB7 (Abcam), LIN28B (Cell signalling), Vimentin (Santa Cruz), E-cadherin (BD), N-cadherin (BD), GAPDH (Abcam), FLAG (Sigma Aldrich), Actin (Sigma Aldrich), Tubulin (home-made) and HA (Babco); and anti-mouse and anti-rabbit (Biorad) as secondary antibodies.

Immunofluorescence assay

Cells were fixed with 4% formaldehyde for 10 minutes at room temperature, permeabilized in 0.1% Triton and incubated with specific antibodies: Vimentin (Santa Cruz),

E-cadherin (BD), N-cadherin (BD). CY3 goat anti-mouse or anti-rabbit were used as secondary antibodies. The nuclei were stained with DAPI. All imagines were acquired with a fluorescence microscopy, excited with ultraviolet light and detected through a blue/cyan filter.

Immunohistochemestry

Teratoma samples were processed in consecutive sections and stained with haematoxylin and eosin. Immunostaining for desmin, protein S-100 and pan-cytokeratin (AE1/AE3) were performed by the Molecular Pathology Division at the IEO hospital.

RNA extraction and RT-qPCR

TRIZOL reagent (Invitrogen), choloroform and isopropyl alcohol were used for total RNA isolation from cells. From FFPE archival lung tumor samples RNAeasy FFPE kit (QIAGEN) was used. RNA was quantified by Nanodrop (Agilent Technologies). Total RNA (1 ug) was retrotranscribed with QuantiTect or miScript Reverse Transcription Kit (Qiagen). mRNA and miRNA expression profiles were obtained by QuantiTec or miScript Primer Assays (Table.2.4-2.5) using QuantiFast or miScript SYBR Green PCR Kit (Qiagen). For HOXB7 mRNA quantification, specific primers were designed (Table.2.6). For LIN28B we used primers described by Viswanathan and colleagues [Viswanathan et al., 2009]. Amplification reactions were performed with LightCycler 480 (ROCHE) using the manufacturer's recommended cycling conditions. Relative expression ratios of miRNAs were obtained using the 2^{ddCT} method: raw data, i.e., cycle threshold (Ct) values, were exported to Excel (Microsoft) and were normalized to the Ct of a housekeeping gene and then to the reference sample. Results are expressed as mean ± standard deviation calculated from at least two technical replicas.

Gene	QuantiTec Primer Assay
ABCG2	Hs_ABCG2_1_SG
ALCAM / CD166	Hs_ALCAM_1_SG
ALDH1A1	Hs_ALDH1A1_1_SG
CD24	Hs_CD24_va.1_SG
CD44	Hs_CD44_1_SG
CD90 / THY1	Hs_THY1_1_SG
E-CADHERIN / CDH1	Hs_CDH1_1_SG
EGFR	Hs_EGFR_1_SG
EPCAM	Hs_EPCAM_1_SG
FGF2 / bFGF	Hs_FGF2_1_SG
GAPDH	Hs_GAPDH_2_SG
GUSB	Hs_GUSB_1_SG
ITGA6	Hs_ITGA6_1_SG
KLF4	Hs_KLF4_1_SG
LIN28	Hs_LIN28_1_SG
MYC	Hs_MYC_1_SG
NANOG	Hs_NANOG_2_SG
N-CADHERIN / CDH2	Hs_CDH2_1_SG
PODXL	Hs_PODXL_va.1_SG
POUF5F1 / OCT4	Hs_POUF5F1_1_SG
PROM1	Hs_PROM1_1_SG
RRN18S	Hs_RRN18S_1_SG
SNAI1	Hs_SNAI1_1_SG
SNAI2	Hs_SNAI2_1_SG
SOX2	Hs_SOX2_1_SG
TWIST	Hs_TWIST1_1_SG

Table 2.4: **List of QuantiTec Primer Assay.**

Gene	miScript Primer Assay
SNORD61	Hs_SNORD61_11
SNORD72	Hs_SNORD72_11
Let-7a	Hs_let-7a_1
Let-7b	Hs_let-7b_1
Let-7c	Hs_let-7c_1
Let-7d	Hs_let-7d_2
Let-7e	Hs_let-7e_2
Let-7f	Hs_let-7f_1
Let-7i	Hs_let-7i_1

Table 2.5: **List of miScript Primer Assay.**

Gene	Customized primer sequence
HOXB7 FW	CTGGATGCGAAGCTCAGG
HOXB7 REV	CAGGTAGCGATTGTAGTGAAATTCT
LIN28B FW	GCCCCTTGGATATTCCAGTC
LIN28B REV	TGACTCAAGGCCTTTGGAAG

Table 2.6: **List of customized RT-qPCR primers.** LIN28B primers were described by Viswanathan and colleagues [Viswanathan et al., 2009]. FW = forward primer; REV = reverse primer.

Gene Set Enrichment Analysis

Gene sets enrichment analysis was performed using GSEA software (Subramanian, Tamayo, et al., (2005, PNAS 102, 15545-15550). Gene Sets (total of 3273 gene sets, C2 category) were downloaded from the MSIGDB database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). A gene set was considered significantly enriched when the nominal p-value was less than 0.05 after 1.000 random shuffling of the experimental labels.

Genomatix

Genomatix is a software tool that utilizes a large library of matrix descriptions for transcription factor binding sites to locate matches in DNA sequences (www.genomatix.de). We used genomatix to search for an enrichment of homeodomain transcription factor binding sites in the 1Kbp sequences upstream and downstream of the LIN28B TSS.

Chapter 3

Results

3.1 Biological relevance of HOXB7 in lung cancer.

As previously discussed, HOXB7 was shown to be overexpressed in leukemia [Storti et al., 2011], melanoma [Caré et al., 1996], breast [Caré et al., 1998, Wu et al., 2006, Jin et al., 2012], colorectal [Liao et al., 2011], pancreatic [Nguyen Kovochich et al., 2013] oral cancer [De Souza Setubal Destro et al., 2010] and more recently in lung cancer [Yuan et al., 2014, Bianchi et al., 2007]. We set out to characterize the molecular mechanisms underlying HOXB7-mediated regulation and pathogenesis of lung cancer.

3.1.1 Analysis of HOXB7 expression in lung cancer expression datasets suggests a role in stemness and reprogramming.

We performed a meta-analysis of high-throughput gene expression datasets of large cohorts of lung cancer patients in order to evaluate the correlation of HOXB7 expression with other genes involved in cancer relevant pathways. We took advantage of three publicly available gene expression datasets comprising a total of ∼1000 lung adenocarcinoma samples:

The LUAD dataset of The Cancer Genome Atlas (TCGA) collection: 499 lung adenocarcinomas profiled by RNA-seq.

Michigan cohort : 442 primary lung adenocarcinomas profiled by Affymetrix (GeneChip Scanner 3000 or 2500), collected by a consortium of four institutions: University of Michigan Cancer Center (UM), Moffitt Cancer Center (HLM), Memorial Sloan-Kettering Cancer Center (MSK) and the Dana-Farber Cancer Institute (DFCI) [Beer, 2008]).

Tokyo Cohort : 226 lung adenocarcinomas subjected to expression profiling by Affymetrix (U133Plus2.0 arrays), consisting of 168 stage I and 58 stage II lung adenocarcinoma samples collated at the National Cancer Center Hospital, Tokyo [Okayama et al., 2012].

We analyzed the mRNA expression profiles generated for thousands of genes from these datasets in order to correlate the gene expression profile of HOXB7 with a collection of gene sets. With gene set, we refer to a group of genes that share a common regulation, biological function or chromosomal location. We used the Gene Set Enrichment Analysis (GSEA), a computational method that permits the assessment of gene sets differentially enriched between two biological states (e.g., high versus low HOXB7 expression) [Subramanian et al., 2005].

Before running the GSEA, lung cancer patients were ranked according to HOXB7 expression level and divided into two classes: HOXB7-high (above the third quartile of HOXB7 expression distribution) and HOXB7-low (below the first quartile of HOXB7 expression distribution). Profiled genes were thus ordered in a ranked list by the software according to their differential expression. We then interrogated the GSEA software in order to determine whether members of a specific gene set tended to occur towards the top (or the bottom) of the ranked list, in which case the gene set was associated with the class distinction. This approach permitted us to determine which gene-sets correlate positively ("enriched") with HOXB7-high lung cancer patients in a total of 3700 samples, representing the entire collection of known molecular pathways.

Among the top scores (nominal p-value <0.05), we identified gene-sets involved in cell proliferation, epithelial-to-mesenchymal transition (EMT), neoplastic transfor-

mation and in tumor progression and metastasis (Fig.3.1). The involvement of these gene-sets was expected due to previous reports describing the involvement of HOXB7 in these pathways [Caré et al., 1996, Wu et al., 2006]. Expression analysis of single genes known as EMT markers, such as N-cadherin and TWIST1, were significantly upregulated in HOXB7-high patients in the three database, SNAI1 in the LUAD and Tokyo cohorts and SNAI2 in LUAD cohort (p-value was calculated by the nonparametric Wilkoxon test; Fig.3.2), further confirming the bona fide of our analysis.

We also analyzed the expression level of FGF2 and EGFR, which are known HOXB7 transcriptional targets [Caré et al., 1998, Jin et al., 2012]. We found that both were, indeed, upregulated in HOXB7-high tumors (Fig.3.2).

All together these results confirmed the potential and reliability of the GSEA and that of our approach to identify molecular pathways possibly modulated by a candidate gene.

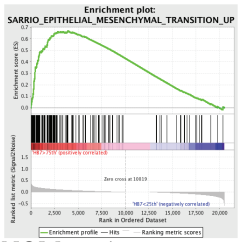
The GSEA also revealed the expression of several enriched gene-sets (p-value <0.05) representing mechanisms involved in stem cell (SC) biology and in induced pluripotent cells (iPS), suggesting a novel role for HOXB7 in modulating these cancer-relevant pathways. To further investigate this important finding, we analyzed in details the gene expression profile of 16 genes that are known or putative SC and IPS markers in these publicly available gene expression datasets. We found that 7 genes (44%) were indeed upregulated upon HOXB7 overexpression: ALDH1A1, BMI-1(COMMD3), CD24, CD49f (ITGA6), CD90 (THY1), LIN28B and SOX2 (Fig.3.2).

In conclusion, the GSEA and gene expression analysis suggest an association for HOXB7 with the activation of a transcriptional program relevant for stem cell biology in cancer, which may support tumor onset and progression. Our findings prompted us to further investigate the possible role of HOXB7 in cancer stem cell biology and lung tumor progression.

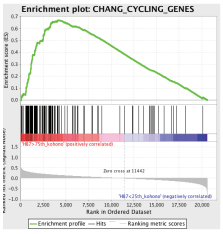
A.

			NES	p-value
EMT	LUAD-TCGA	JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP	1.62	0.018
		SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP	1.58	0.013
	Michigan Cohort	SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP	1.64	0.027
		ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE	1.62	0.042
	Tokyo Cohort	ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE	1.95	<0.001
		SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP	1.94	0.008
		GOTZMANN_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP	1.89	<0.001
		ECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP	1.81	0.01
PROLIFERATION	LUAD-TCGA	CHANG_CYCLING_GENES	1.56	0.021
		REACTOME_SYNTHESIS_OF_DNA	1.55	0.041
	Tokyo Cohort	CHANG_CYCLING_GENES	2.12	0.002
		REACTOME_DNA_REPLICATION	1.89	0.004
		REACTOME_G1_S_TRANSITION	1.87	0.004
CANCER	LUAD-TCGA	WANG_NEOPLASTIC_TRANSFORMATION_BY_CCND1_MYC	1.8	0.004
		BARIS_THYROID_CANCER_UP	1.75	0.002
		KAPOSI_LIVER_CANCER_MET_UP	1.74	0.004
	Michigan Cohort	IIZUKA_LIVER_CANCER_PROGRESSION_L0_L1_UP	1.87	0.002
		LIU_PROSTATE_CANCER_UP	1.76	0.014
		LIAO_METASTASIS	1.84	<0.001
		LI_WILMS_TUMOR_VS_FETAL_KIDNEY_2_UP	1.82	0.002
		CHANDRAN_METASTASIS_UP	1.8	0.001
		RAMASWAMY_METASTASIS_UP	1.75	0.011
	Tokyo Cohort	CROMER_TUMORIGENESIS_UP	2.29	0.001
		VECCHI_GASTRIC_CANCER_EARLY_UP	2.02	0.001
		SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6	2.02	0.006
		JAEGER_METASTASIS_UP	1.97	<0.001
		WINNEPENNINCKX_MELANOMA_METASTASIS_UP	1.93	0.01
		RAMASWAMY_METASTASIS_UP	1.82	0.006
		LEE_LIVER_CANCER_SURVIVAL_DN	1.79	0.016
STEM/iPS	LUAD-TCGA	MIKKELSEN_IPS_WITH_HCP_H3K27ME3	1.92	<0.001
		MIKKELSEN_ES_HCP_WITH_H3K27ME3	1.74	0.013
		BENPORATH_ES_2	1.6	0.019
		CONRAD_STEM_CELL	1.48	0.029
	Michigan Cohort	BENPORATH_ES_1	1.65	0.014
		GUENTHER_GROWTH_SPHERICAL_VS_ADHERENT_UP	1.49	0.032
	Tokyo Cohort	WONG_EMBRYONIC_STEM_CELL_CORE	1.83	0.002
		MEISSNER_NPC_HCP_WITH_H3K27ME3	1.8	0.002

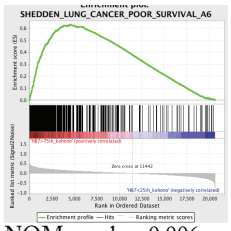
B.



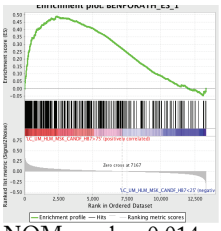
NOM p-value: 0.013



NOM p-value: 0.002



NOM p-value: 0.006



NOM p-value: 0.014

Figure 3.1: **Analysis of HOXB7 expression in lung cancer expression datasets suggests a role in stemness and reprogramming. A.** Analysis of gene expression datasets (LUAD-TGCA, Michigan cohort and Tokyo cohort) of ∼1000 lung adenocarcinomas. Patients were ranked according to their HOXB7 expression and divided into two classes: HOXB7-high (above the 75th percentile) and HOXB7-low (below first quartile). GSEA of more than 3700 gene sets revealed top scoring gene sets in tumor progression, metastasis, EMT, SC homeostasis and iPS, all of which correlate positively with high HOXB7 expression (Nominal p-value: <0.05). **B.** Four enriched plots are reported as examples.

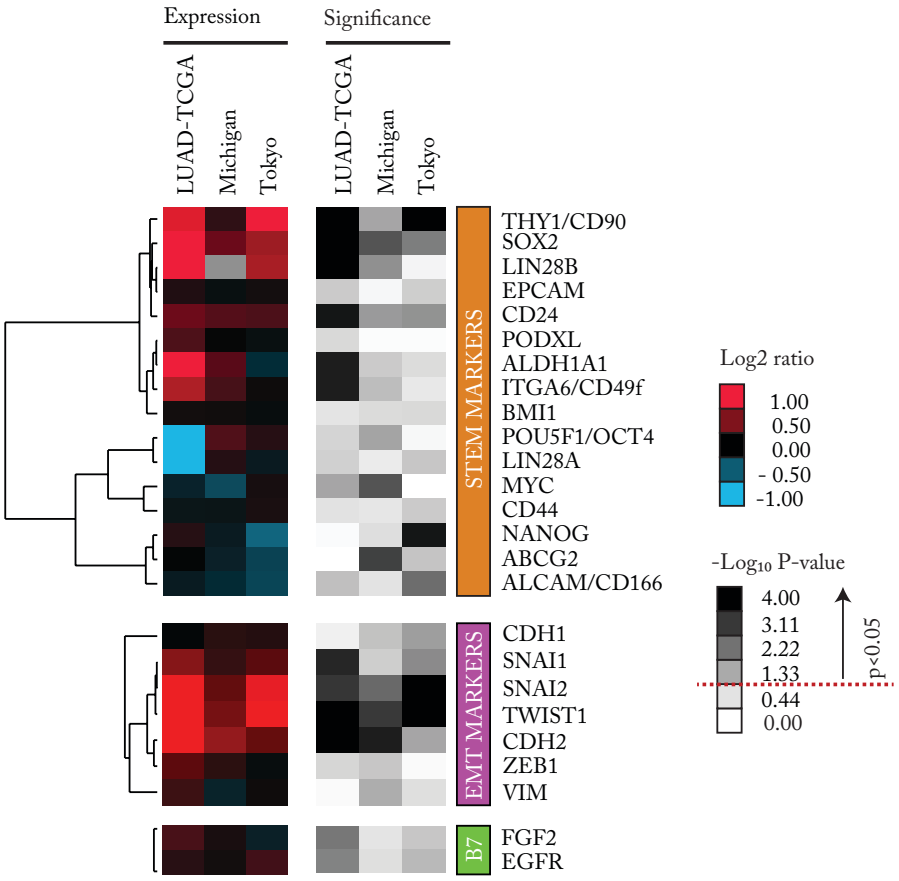


Figure 3.2: **Hierarchical cluster analysis of STEM markers and EMT markers and two known HOXB7 transcriptional targets (EGFR and FGF2).** The markers were collected by literature search and analyzed in three independent cohorts of lung tumors (adenocarcinoma), profiled by RNA-seq (LUAD-TCGA cohort) and Affymetrix microarray (Michigan and Tokyo cohorts). Hierarchical clustering analysis was performed using the log2 of the mean gene expression ratio of HOXB7-high and HOXB7-low expressing tumors. The -log(10)p-value values to shows the significance of the differential expression ($p<0.05$ corresponds to $-LogP>1.3$). P-values were calculated by the non parametric Wilcoxon test.

3.2 HOXB7 overexpression in lung cells augments a sub-population with stem-like properties.

We set out to study the effects of forced HOXB7 expression in a lung cell model system in order to clarify the role of HOXB7 in lung cancer progression. We analyzed the expression of HOXB7 in a panel of lung cancer cell lines and identified a low-HOXB7 expressing normal lung cell line, BEAS-2B (Table.3.1, Fig.3.3 and Supplementary Fig. 5.1). We stably overexpressed HOXB7 (Fig.3.4 A and Fig.3.6 D) in these cells upon retroviral infection with a pBABE-HOXB7 construct (carrying the full length coding sequence of HOXB7) or a pBABE empty vector (EV). Importantly, in the HOXB7 overexpressing cells we scored an increased expression of the previously described HOXB7 target genes, the bFGF (FGF2) and EGFR genes [Caré et al., 1998, Jin et al., 2012], suggesting the functionality of the exogenous protein (Fig.3.4 B).

Cell line	Source	Tissue	Genetics
BEAS-2B	normal lung	normal bronchus	SV40 T antigen, p53 wt
NCI-H1975	lung adenocarcinoma	primary tumor	EGFR and PIK3CA mut, p53 wt
NCI-H358	bronchioalveolar carcinoma	primary tumor	K-Ras mut, p53 negative
NCI-H2228	lung adenocarcinoma	primary tumor	p53 wt
NCI-H23	adenocarcinoma	primary tumor	K-Ras and p53 mut
CALU-3	lung adenocarcinoma	pleura-effusion	K-Ras wt, p53 mut
NCI-H838	lung adenocarcinoma	pleura-effusion	p53 wt
A549	lung adenocarcinoma	primary tumor	K-Ras mut, p53 wt

Table 3.1: **Lung Cell Lines.** Data were collected from the ATCC database and the TP53 web site: http://p53.free.fr/Database/Cancer_cell_lines/NSCLC.html

We then analyzed the gene expression profile of a panel of lung SC/iPS markers in BEAS-2B cells by qRT-PCR (for a total of 14 genes) and found that five of these genes were indeed induced upon HOXB7 overexpression (SOX2, NANOG, CD90/THY1, LIN28 and LIN28B; Fig.3.5 A), confirming the *in silico* analysis by GSEA. Of note, we could not appreciate any induction of CD44, CD166, EPCAM, PODXL, ABCG2 and OCT4 (Fig.3.5 A).

Based on these results, we then assessed whether HOXB7 expression could be func-

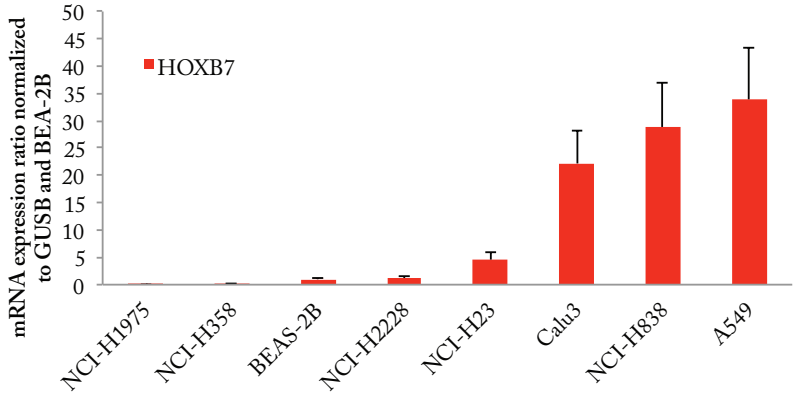


Figure 3.3: **Expression of HOXB7 in NSCLC and BEAS-2B cell lines.** The mRNA expression of HOXB7, (analyzed by RT-qPCR) varied grately between the different cell lines. The normal lung epithelial cell line BEAS-2B showed a low HOXB7 mRNA expression. The NCI-H358 lung bronchoalveolar carcinoma cell line (defined as early stage lung adenocarcinoma [Brambilla et al., 2001]) showed undetectable levels of HOXB7. The Calu-3, NCI-H838 and A549 cell lines displayed the highest expression levels of HOXB7 mRNA (∼30 fold more than BEAS-2B).

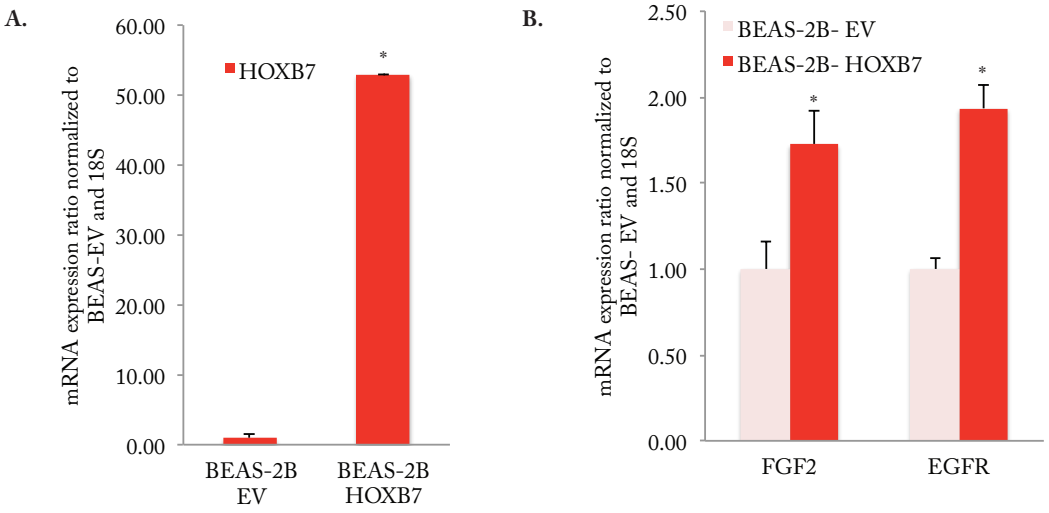


Figure 3.4: **Overexpression of HOXB7 in BEAS-2B increases the expression of known HOXB7 transcriptional targets. A.** Overexpression of HOXB7 by retroviral infection of pBABE-HOXB7 or control empty vector (EV) was verified by RTqPCR. **B.** RT-qPCR analysis of two known HOXB7 transcriptional targets FGF2 (bFGF) and EGFR. *, significant p-value ($p<0.05$; Student's t-test, two tailed).

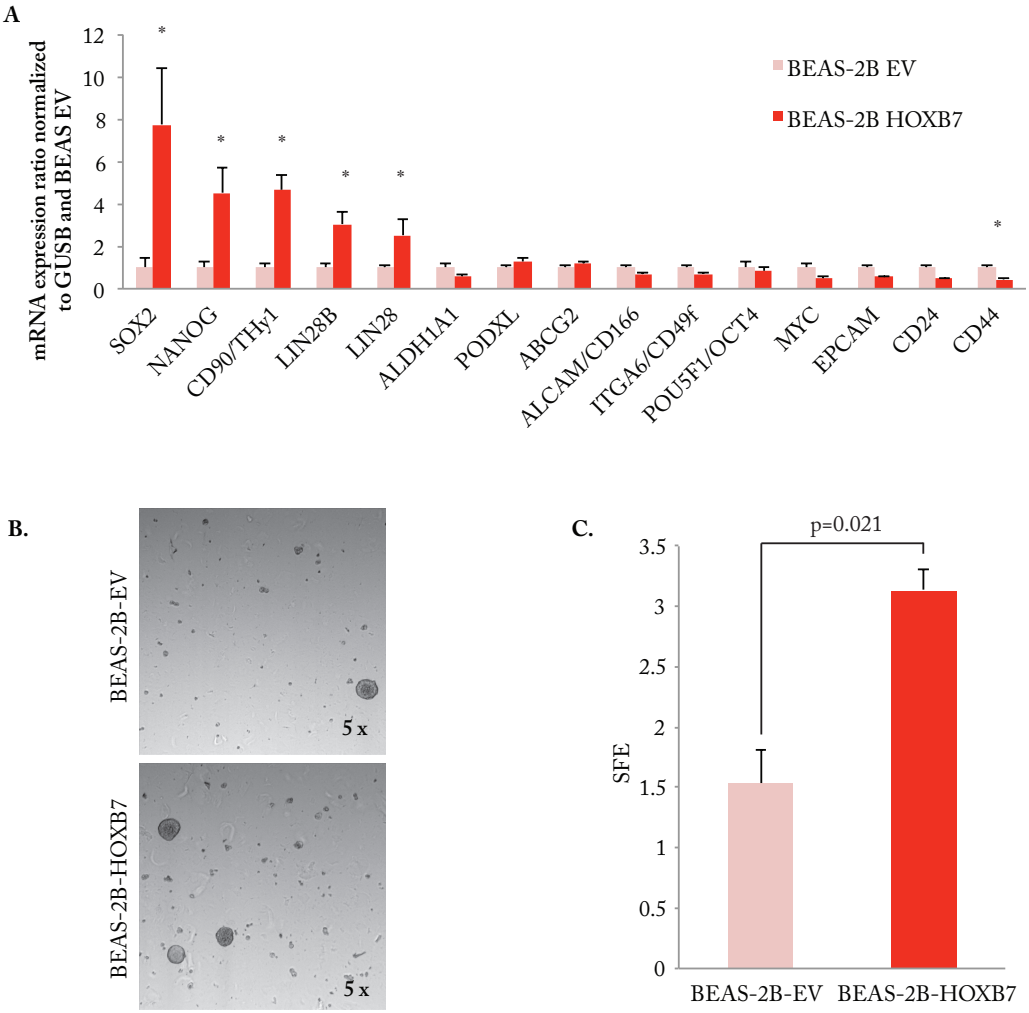


Figure 3.5: **Overexpression of HOXB7 in BEAS-2B lung cells increases stem cell properties. A.** A panel of lung SC markers and genes involved in iPS (for a total of 14 genes) analyzed by RTqPCR in BEAS-HOXB7 and control cells (BEAS-EV). *, significant p-value ($p<0.05$; Student's t-test, two tailed). **B.** Representative images of pneumospheres in cells overexpressing HOXB7 and in control cells. **C.** Quantification of the sphere forming efficiency (SFE) in cells overexpressing HOXB7 and in control cells (n=2).

tionally related with the stem cells compartment. We performed a "pneumosphere assay", following the protocol used for generating mammospheres, as described by Dontu and colleagues [Dontu et al., 2003]. This assay consists in growing epithelial cells *in vitro* under anchorage-independent conditions (we used tissue culture plates that impair the adhesion of cells) in a growing medium that is depleted of serum components in order to preserve the undifferentiated state of the cells. In these conditions, only stem or progenitor cells are able to survive and grow as clonal spheroids, defined as pneumospheres. We can thus assume that the number of formed clonal spheroids is proportional to the number of stem cells present in the initial population [Dontu et al., 2003]. Our results were encouraging: the BEAS-2B-HOXB7 cells showed a 3.1% sphere forming efficiency (SFE) (Fig.3.5 B-C), which was a two-fold increase of the SFE derived from the control cells (1.5%).

This finding shows BEAS-2B-HOXB7 cell are enriched in stem cell properties (i.e., ability to grow as spheres in non adhesion and serum-free conditions), further supported by the observation of an occurred EMT in these cells (Fig.3.6). Indeed, gain of stem cell properties has been recently linked to EMT, revealing a convergence among these two processes [Mani et al., 2008, Kong et al., 2010]. The BEAS-HOXB7 cells appeared scattered and were characterized by a spindle-like shape, whereas control cells maintained their cobblestone-like phenotype, typical of epithelial cells (Fig.3.6 A). We then monitored known EMT markers for 7 weeks post infection. Already in the second week post infection, the BEAS-2B-HOXB7 cells showed an increased expression of the mesenchymal markers N-cadherin and Vimentin (Fig.3.6 C). On the contrary, no significant changes in E-cadherin expression were detected even two months post infection (Fig.3.6 B-D). Previous studies have reported the achievement of EMT without concomitant loss of E-cadherin expression [Hollestelle et al., 2013].

In line with the expected behavior of mesenchymal cells [Kalluri and Weinberg, 2009], cells overexpressing HOXB7 showed an induced proliferative advantage, as shown by previous studies in breast cancer and melanoma cells [Caré et al., 1996, Wu et al.,

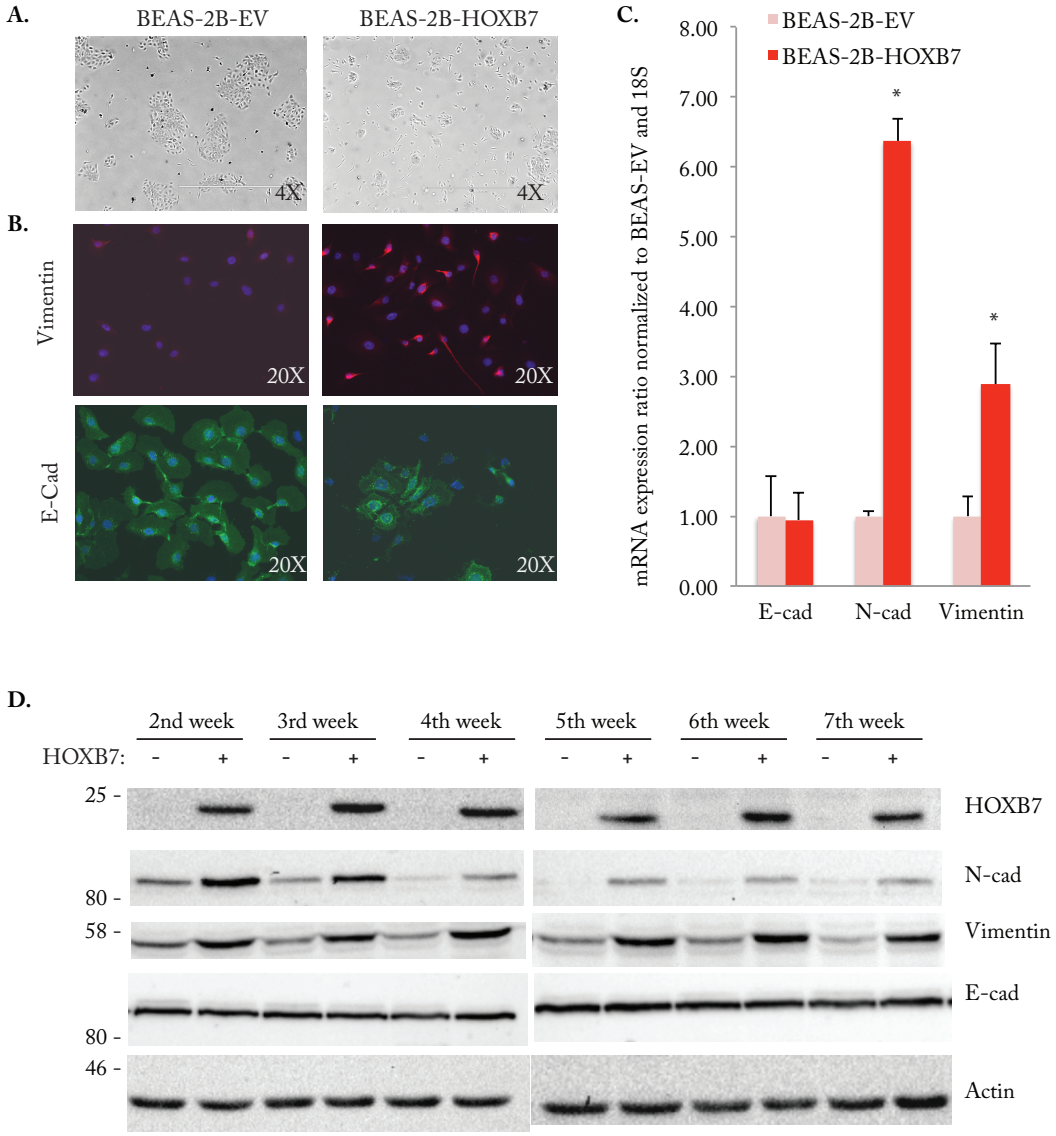


Figure 3.6: **Overexpression of HOXB7 in BEAS-2B cells induces EMT. A.** Light microscopy images of BEAS-2B-HOXB7 cells and empty vector (EV) control cells (4x magnification). **B.** Immunofluorescence staining of cells as in A: Vimentin (red), E-cadherin (green), cell nucleus (blue, stained with DAPI) (20x magnification). **C.** RT-qPCR analysis of EMT marker genes (N-cadherin (N-cad), E-cadherin (E-Cad) and vimentin) in cells as in A. *, significant p-value ($p<0.05$; Student's t-test, two tailed). **D.** Expression of HOXB7 and EMT markers (N-cadherin (N-cad), E-cadherin (E-Cad) and Vimentin) in cells as in A were analyzed by WB analysis; actin, loading control. Whole cell lysates were collected every week post infection for seven weeks.

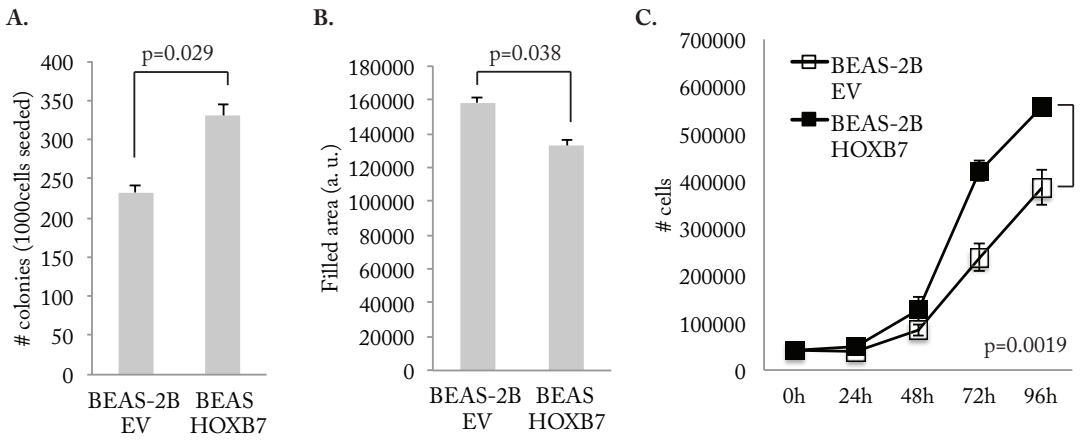


Figure 3.7: **Overexpression of HOXB7 in BEAS-2B cells increases cell proliferation. A-B.** Number and filled area of formed colonies by BEAS-HOXB7 and control cells (wt and EV, n=2). **C.** Cell proliferation assay in BEAS-2B-HOXB7 and control cells (n=2).

2006]. Colony formation assay revealed a 42% increase in colony number in BEAS-2B-HOXB7 cells compared to control, even though control cell colonies occupied a comparable area of the plate surface (Fig.3.7 A-B). Cell proliferation assay confirmed the proliferative advantage in HOXB7-overexpressing cells (Fig.3.7 C).

We then selected the lung cancer cell line NCI-H358, characterized by undetectable levels of HOXB7 (Fig.3.8 A-B), in order to study the effects of overexpressing HOXB7 in a cancer contest. After infection with the pBABE-HOXB7 retroviral construct or the EV control, we found that HOXB7 was indeed able to increase cell proliferation, in line with our results obtained in HOXB7-overexpressing BEAS-2B cells. Colony formation assay in HOXB7-expressing NCI-H358 cells resulted in a 4-fold and 22-fold increase in colony number in normal conditions and in serum depleted condition, respectively, compared to control cells (Fig.3.8 C-D). NCI-H358-HOXB7 cells showed also an increased ability to form colonies in non-adhesion conditions (2.3-fold increase, Fig.3.8 F) and sustained migratory phenotype (10-fold faster in transwell assay compared to control cells, Fig.3.8 D). We also observed an induction of EMT upon HOXB7 overexpression: the expression of Vimentin and N-cadherin was induced and that of E-cadherin was downregulated (Fig.3.9).

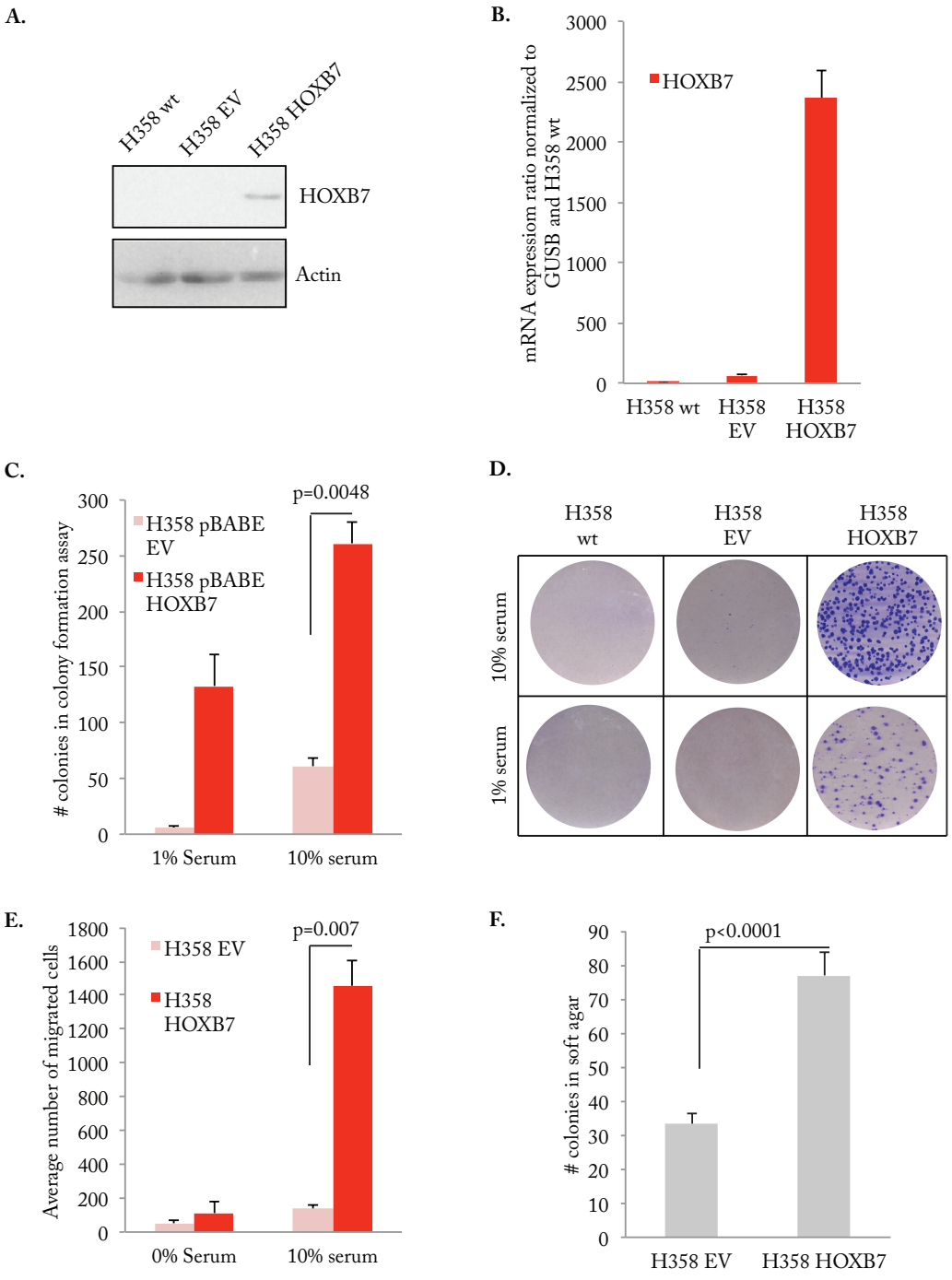


Figure 3.8: **Overexpression of HOXB7 in H358 lung cancer cells increases proliferation and migration. A.** Western blot (WB) analysis of HOXB7 expression in H358-HOXB7 and control cells (wt and EV); Actin: loading control. **B.** RT-qPCR analysis of HOXB7 mRNA expression in cells as in A. **C-D.** Crystal violet staining (D) and quantification (C) of BEAS-HOXB7 and control cell colonies (wt and EV) grown in 1% and 10% serum (n=3). **D.** Transwell migration assay of cells as in A grown in 0% and 10% serum. **F.** Quantification of colonies grown in non-adhesion conditions (soft agar, n=2). P-values were calculated using Student's t-test, two tailed.

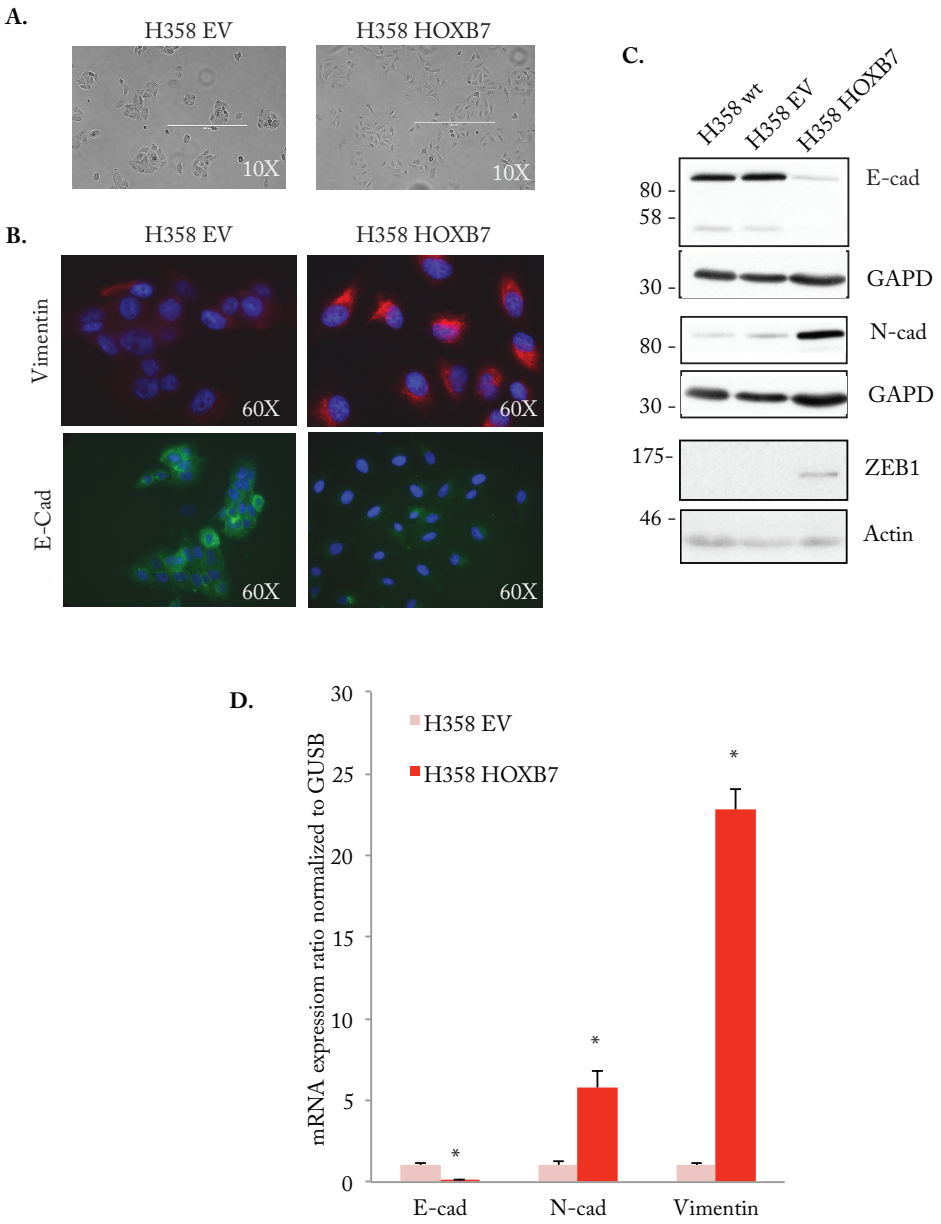


Figure 3.9: **Overexpression of HOXB7 in H358 lung cancer cells induces EMT. A.**. Light microscopy images of H358-HOXB7 and control (EV) cells (10x magnification). **B.** Immunofluorescence staining of vimentin (red), E-cadherin (E-cad, green) and cell nucleus (blue, stained with DAPI) in cells as in A (60x magnification). **C.** WB analysis of EMT markers (N-cadherin (N-cad), E-cadherin (E-cad) and ZEB1) on whole cell extracts. GAPDH and Actin: loading controls.**D.** RT-qPCR analysis of EMT marker genes (E-cadherin (E-cad), N-cadherin (N-cad) and vimentin). *, significant p-value ($p<0.05$; Student's t-test, two tailed).

3.2.1 The CD90^{high} cells form spheres and are enriched upon overexpression of HOXB7.

In order to further characterize stem properties in HOXB7-high expressing cells, we examined the gene expression profile of the panel of the lung SC/iPS markers analyzed by qRT-PCR (Fig.3.5 A). The surface stem marker CD90 showed a 4.5-fold induction and was one of the most upregulated genes in BEAS-HOXB7 cells. CD90, the glicosylphosphatidylinositol-anchored glycoprotein THY1, has recently been proposed and characterized as a lung cancer stem cell marker in A549 and H446 lung cancer cell lines [Yan et al., 2013] and in primary lung cancer cells [Wang et al., 2013]. Furthermore, breast stem-like cells require CD90 to sustain the cancer stem cell compartment [Lu et al., 2014]. Therefore, we used CD90 as a stem marker to further characterize the interplay between HOXB7 expression and the stem cell compartment.

We first analyzed the cell surface expression of CD90 in BEAS-2B cells by FACS analysis using a human CD90 monoclonal antibody. CD90 was heterogeneously expressed in BEAS-2B control cells (Fig.3.10). The bulk population presented a comet shape with a head negative for CD90 (56.9%), a central body with an intermediate expression (21.24%) and a tail of few cells characterized by high CD90 levels (4.43%) (Fig.3.10 A). On the contrary, in BEAS2B-HOXB7 overexpressing cells the bulk population resulted more homogeneously distributed along the x-axis of the plot, with a 5-fold increase of the CD90^{high} population (21.45%) compared to control cells (Fig.3.10 B).

We then performed a SFE assay using BEAS-2B sorted cells and found that the CD90^{high} populations were clearly the ones able to form spheres both in control and in BEAS-HOXB7 cells (22.75% and 14.71% SFE, respectively). Sphere forming efficiency ranged between 1.64% and 00.4% in the CD90^{int} and CD90^{low} populations, respectively (Fig.3.11). Subsequent RTqPCR analysis revealed a 1.5-fold increase of HOXB7 expression in CD90^{high} cells compared to CD90^{low} cells (Fig.3.10 C).

We also quantified the CD90 content in two primary lung cancer cell lines with

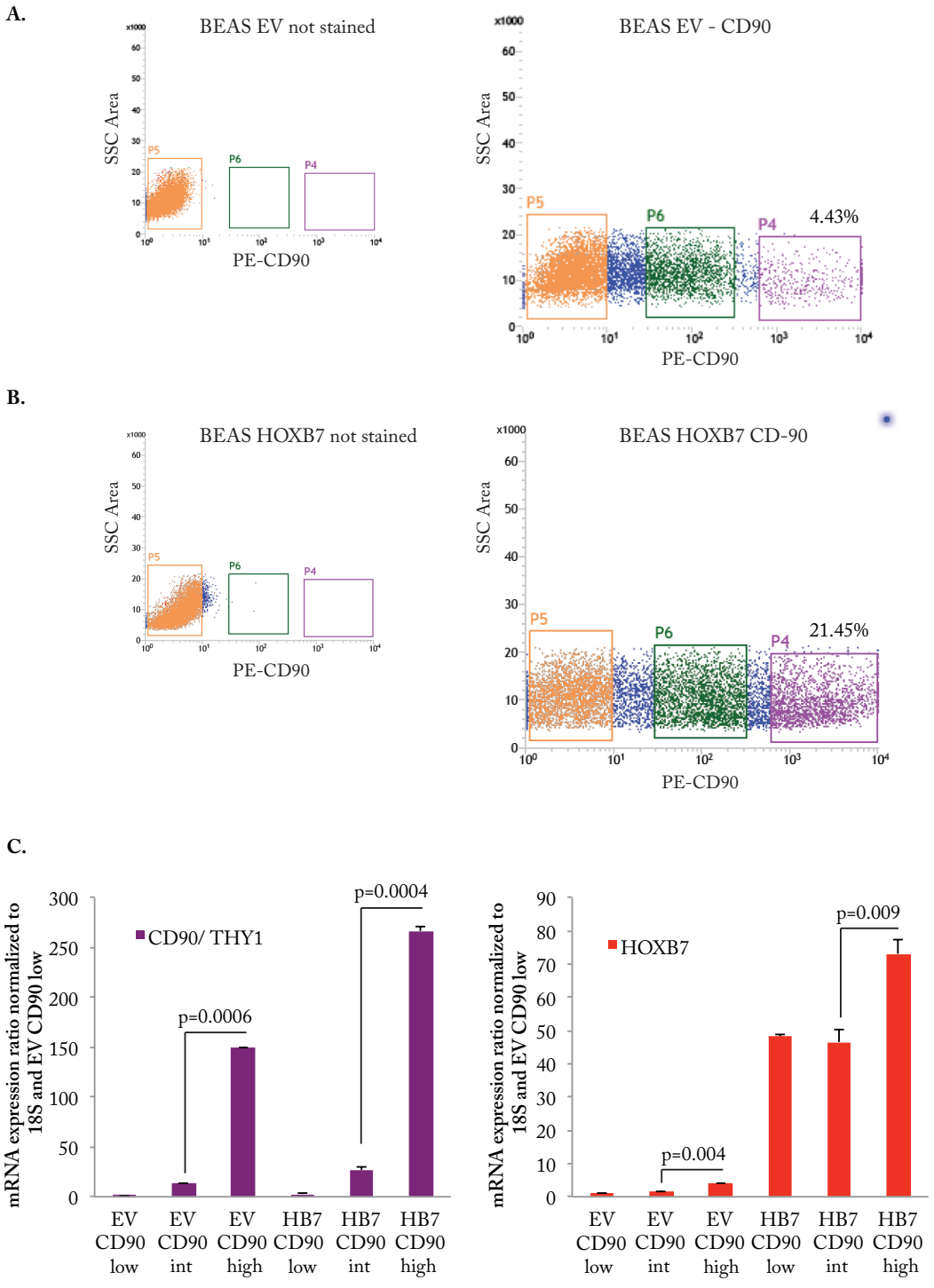


Figure 3.10: **CD90 is enriched upon HOXB7 overexpression in BEAS-2B cells. A-B.** Representative images of two FACS analyses of the CD90 cell surface marker in BEAS-HOXB7 and control (EV and un-stained) cells. Three populations were highlighted: $CD90^{high}$ (purple), $CD90^{int}$ (green) and $CD90^{low}$ (orange). **C.** Expression analysis of the CD90 cell surface marker and HOXB7 by RTqPCR with total RNA extracts from $CD90^{high}$, $CD90^{int}$ and $CD90^{low}$ cells. P-values were calculated using Student's t-test, two tailed).

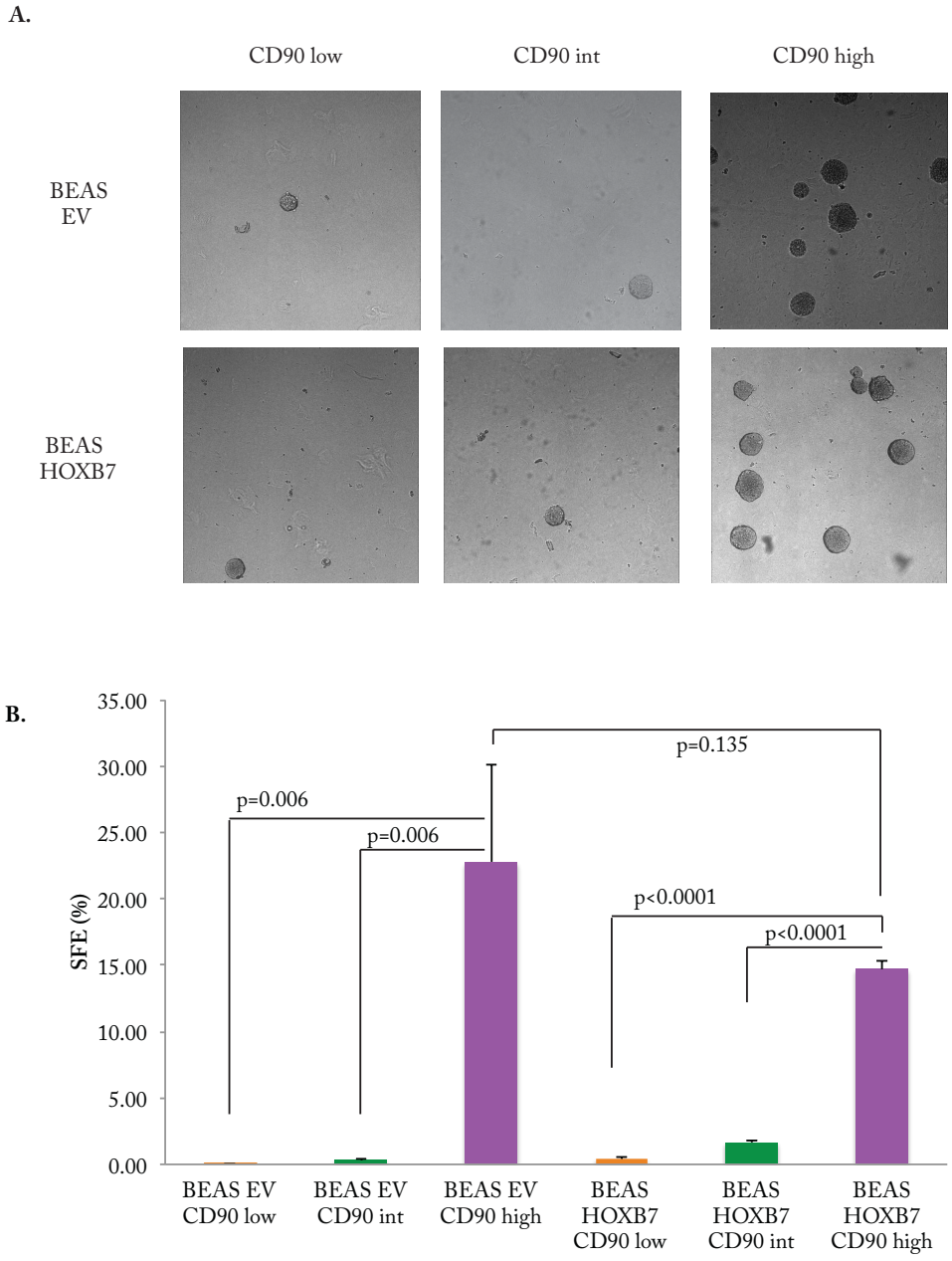


Figure 3.11: **The CD90^{high} cells form more spheres than CD90^{int} and CD90^{low} BEAS-2B cells. A.** Representative images of pneumospheres in cells overexpressing HOXB7 and in control cells (EV) sorted according to CD90 expression. Images were acquired a 5x objective lens. **B.** Quantification of the sphere forming assay performed with CD90-sorted cells; data are reported as SFE=% of grown spheres over the total number of seeded cells (n=1 in technical triplicates. P-values were calculated using Student's t-test, two tailed).

high and low HOXB7 endogenous levels, established from two patients operated at the IEO hospital (Fig.3.12 A): i) a 76-year old patient diagnosed with a Stage IIa adenocarcinoma and negative for HOXB7 (assessed by RTqPCR analysis); and ii) a 78-year old patient diagnosed with a Stage IIa lung adenocarcinoma and positive for HOXB7 expression (RTqPCR data in fig.3.12 A). FACS analysis revealed that CD90 was expressed only in 2.1% of cells from the first patient but in 19.7% of cells from the second patient (positive for HOXB7 expression, Fig.3.12 B).

Taken together, these results confirmed that HOXB7 is capable of expanding a subpopulation of cells with the stem cells characteristic (i.e., ability to grow as spheres) and is characterized by a cell surface marker recently shown to be tightly linked to cancer stem cells [Lu et al., 2014].

3.3 HOXB7 activates the transcription of LIN28B, a gene involved in cancer stem cells biology and tumor progression.

When we analyzed the expression profile of a panel of lung SC markers and genes involved in stem/iPS cells in BEAS-2B cells, we found SOX2, NANOG, CD90, LIN28 and LIN28B to be strongly induced upon HOXB7 overexpression (Fig.3.5).

Expression modulation of both the homologs LIN28 and LIN28B was an important finding for us for two main reasons: i) LIN28/B plays a role in EMT [Liu et al., 2013], in stem cell homeostasis [Newman et al., 2008] and in cancer stem cells [Shyh-Chang and Daley, 2013]; and ii) LIN28/B is involved in lung cancer and other cancer types [Viswanathan et al., 2009].

We, therefore, reasoned that LIN28/B could represent an important effector for HOXB7 signaling in lung cancer and actually represent a link between HOXB7 modulation of the SC-like compartment and lung cancer progression.

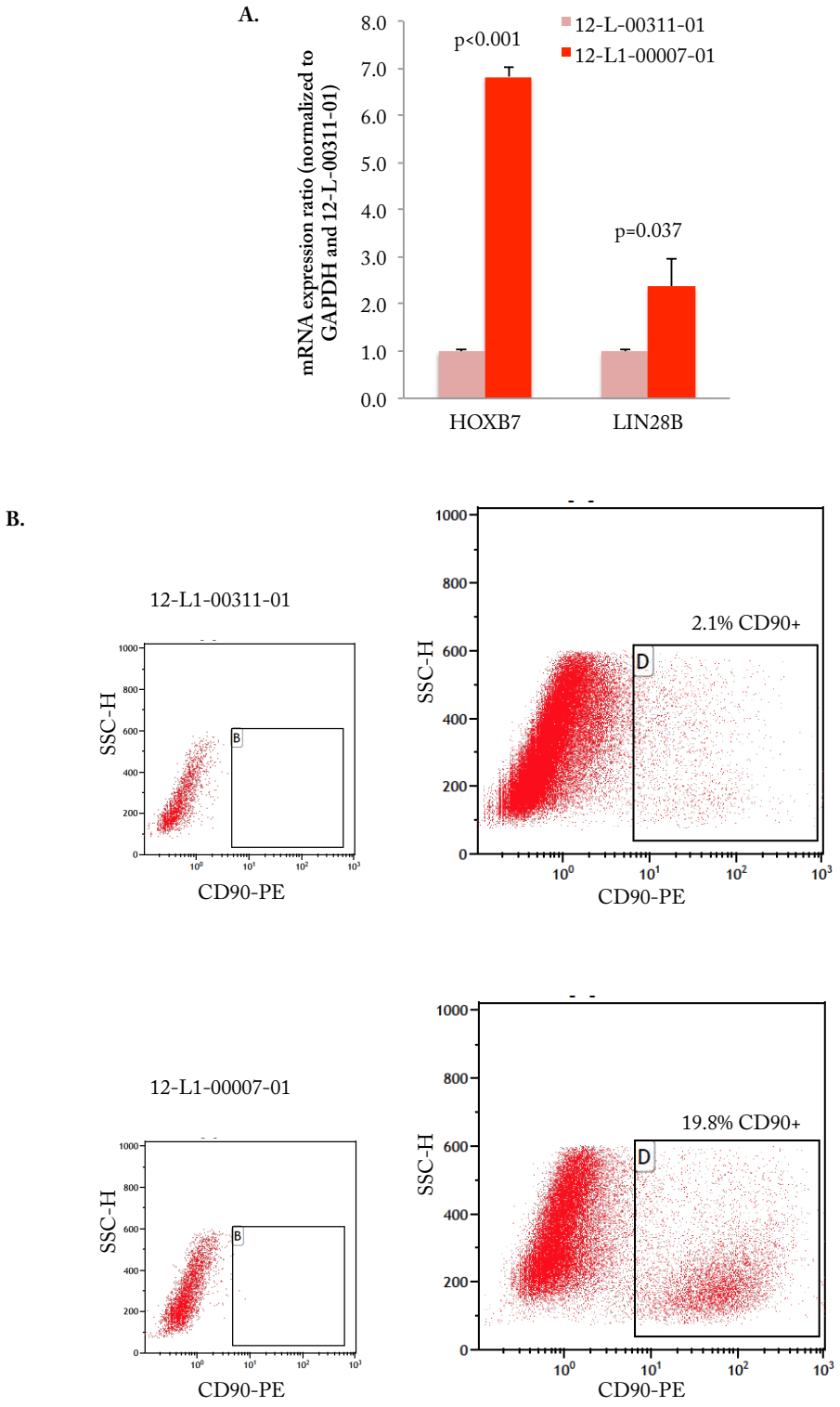


Figure 3.12: **The stem marker CD90 is enriched in a high-HOXB7 expressing lung primary cell line compared to a low-expressing one A.** RT-qPCR analysis of HOXB7 and LIN28B expression. Total RNA was extracted from primary lung cancer cells stabilized from two patients operated at the IEO hospital. P-values were calculated using Student's t-test, two tailed. **B.** Representative images of one FACS analysis of the CD90 cell surface marker in the two primary cell lines and a negative control (not stained).

3.3.1 Correlation of HOXB7 and LIN28B expression.

We initially determined the correlation between HOXB7 and LIN28/B expression in lung cancer patients and in a panel of lung cancer cell lines. Real-Time PCR analysis revealed a strong correlation in 7 out of the 8 cell lines (87.5%) between HOXB7 and LIN28B (Fig.3.13 A) but only in 4 out of 8 cell lines (50%) between HOXB7 and LIN28 (Fig.3.13 B). Indeed, it was more likely to find LIN28B over expressed when HOXB7 was expressed at higher levels and *vice-versa*.

Moreover, as already discussed, gene expression analysis of TCGA lung adenocarcinoma patients showed a significant enrichment of LIN28B in HOXB7-high lung tumors, while LIN28 expression was not enriched (Fig.3.13). Interestingly, both HOXB7 and LIN28B resulted significantly more expressed in cancer samples (468 adenocarcinomas) compared to normal lung tissues (58 samples) from the LUAD dataset (Fig.3.14). Of note, we also found LIN28B (but not LIN28) among the top-ten upregulated genes in a HOXB7-overexpressing myeloma cell line, profiled by expression microarray [Storti et al., 2011].

The coexpression of HOXB7 and LIN28B was also verified in a set of primary lung adenocarcinomas (N=22, R=0.6; p-value = 0.004; Fig.3.13 C) originally screened by qRT-PCR [Bianchi et al., 2007] and in the previously discussed primary lung cancer cell lines established from two lung adenocarcinomas (Fig.3.12 A).

From a molecular point of view, LIN28B gene codes for a RNA binding protein that regulates the maturation of Let-7 microRNA family members, blocking their subsequent maturation [Newman et al., 2008]. In our model system, western blot and immunofluorescence analysis of BEAS-2B cells confirmed the increased expression of LIN28B in HOXB7 overexpressing cells (Fig.3.15 B-C), which was accompanied consistently by a $\sim$40% reduction of six out of eight let-7 miRNA family members (Fig.3.15 D). These finding are similar to the results obtained in NCI-H358 cells overexpressing HOXB7 (Fig. 3.16) and are in line with the massive Let-7s dowregulation observed in BEAS-2B cells upon overexpression of LIN28B (Supplementary fig. 5.2). .

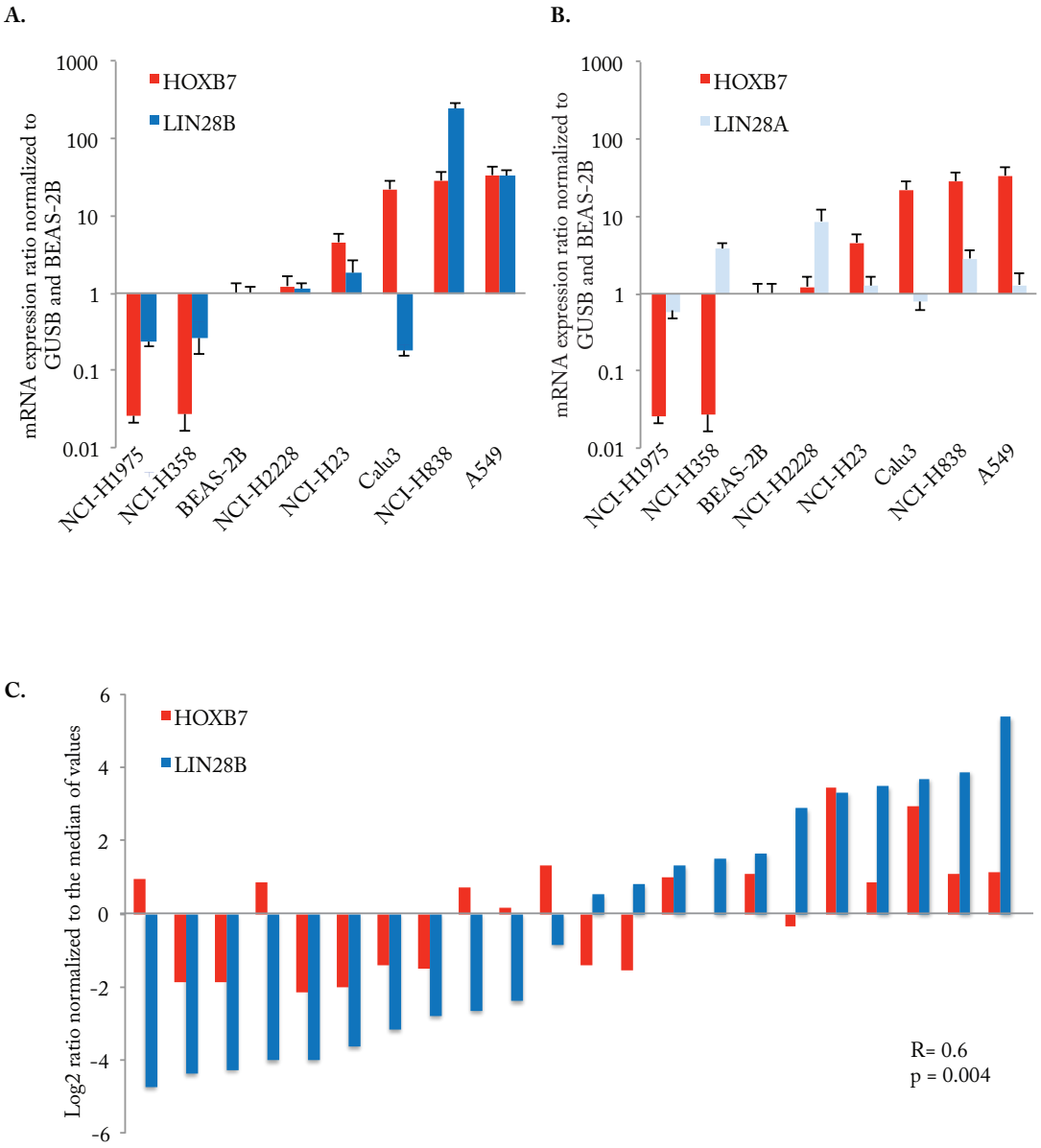


Figure 3.13: **Correlation of HOXB7 and LIN28B expressions in lung cancer cells. A.** RT-qPCR analysis of HOXB7 and LIN28B expression performed on RNA extracts from NSCLC cell lines. Data are normalized to GUSB and BEAS-2B expression. **B.** HOXB7 and LIN28A expression profile in cells as in A. **C.** RNA expression profile of 22 paraffin embedded, primary lung adenocarcinomas.

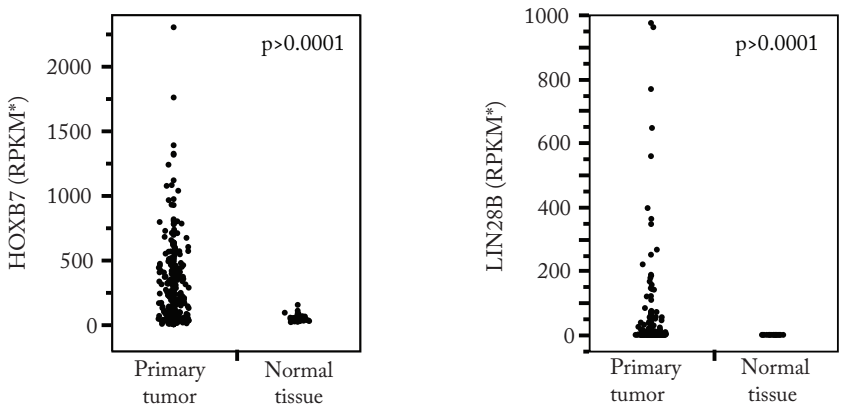


Figure 3.14: **HOXB7 and LIN28B are more expressed in cancer samples compared to normal lung tissues.** *Reads Per Kilobase per Million (RPKM) mapped reads of HOXB7 and LIN28B in primary tumors (468 adenocarcinomas) compared to normal lung tissues (58 samples) from the LUAD dataset. P-values were calculated by the non parametric Wilcoxon test.

Conversely, downregulation of HOXB7 in BEAS-2B cells upon infection with a lentiviral vector carrying short hairpin RNAs against HOXB7 (Supplementary fig.5.3) produced a strong down-modulation of LIN28B expression, both at mRNA and protein levels (Fig.3.17 A-B), accompanied by a 3-6 fold increment in the expression of Let-7 microRNA family members compared to BEAS-2B-scr control (Fig.3.17 C).

These observations further support the existence of a transcriptional network regulated by the transcription factor (TF) HOXB7 and the RNA binding protein LIN28B and we decided to further investigate this possibility.

3.3.2 Promoter analysis of the LIN28B gene.

To elucidate whether LIN28B could be a transcriptional target of HOXB7, we analyzed LIN28B promoter region looking for possible HOXB7 binding sites. We analyzed a ”2Kb” region, starting 1kb upstream of the LIN28B Transcriptional Starting Site (TSS) and extended to 1kb downstream of the TSS. We took advantage of the Genomatix suite software in order to identify the HOX predicted TF binding site (Fig.3.18 and Fig. 3.19 A) and the UCSC Genome Browser (UCSC, University of California Santa Cruz) to calculate the sequence conservation of the predicted TF binding sites

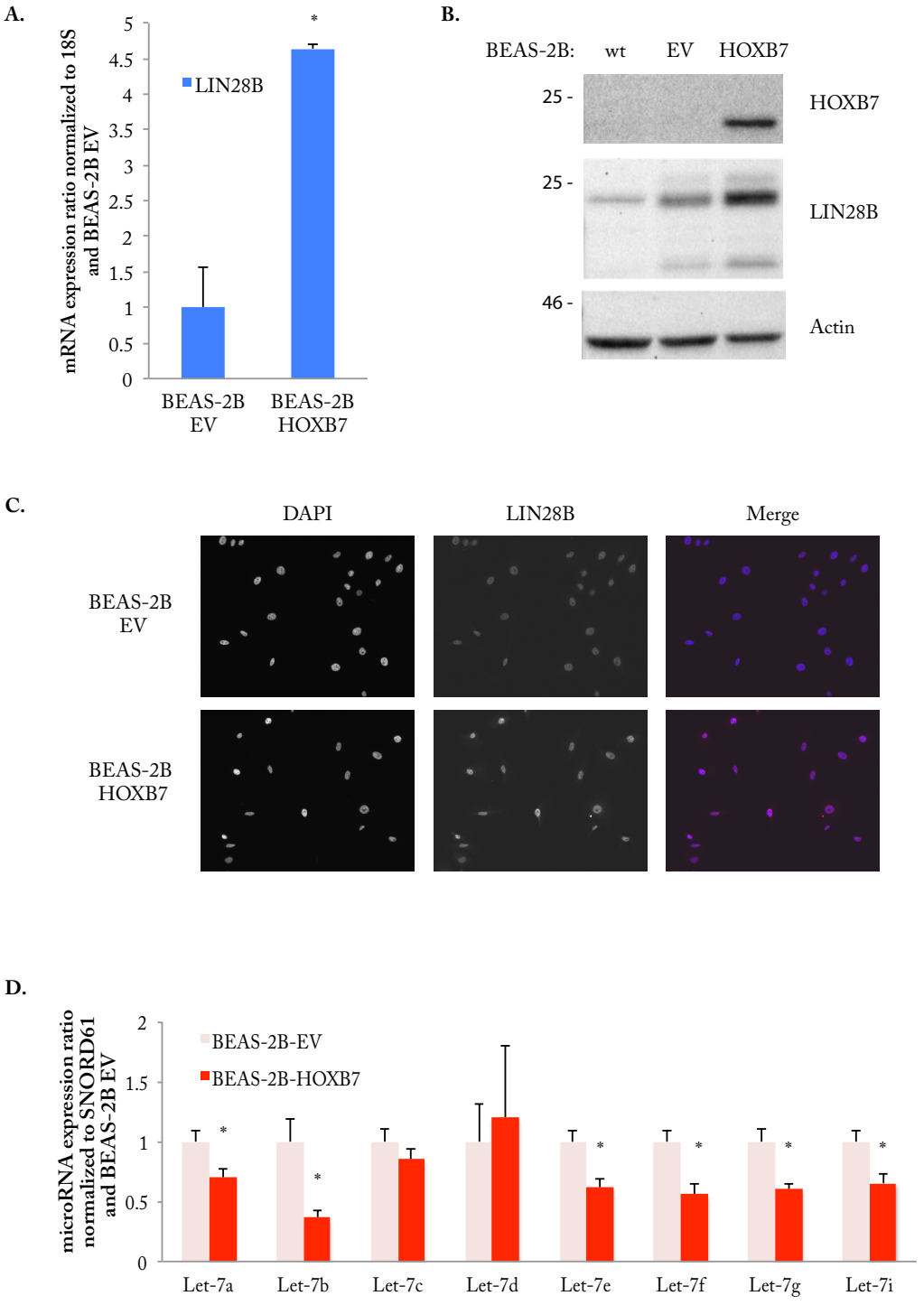


Figure 3.15: **BEAS-2B-HOXB7 cells show increased LIN28B expression and concomitant Let-7s downregulation. A-B.** RTqPCR and WB analysis of LIN28B expression performed on the whole cell extracts from BEAS-2B cells infected with pBABE-Hygro-EV or pBABE-Hygro-HOXB7. **C.** Immunofluorescence of LIN28B expression (blu=DAPI, red=LIN28B). **D.** Expression ratio of Let-7s microRNA family members, relative to BEAS-2B-EV cells by RT-qPCR analysis of cells as in A. *, significant p-value ($p<0.05$; Student's t-test, two tailed).

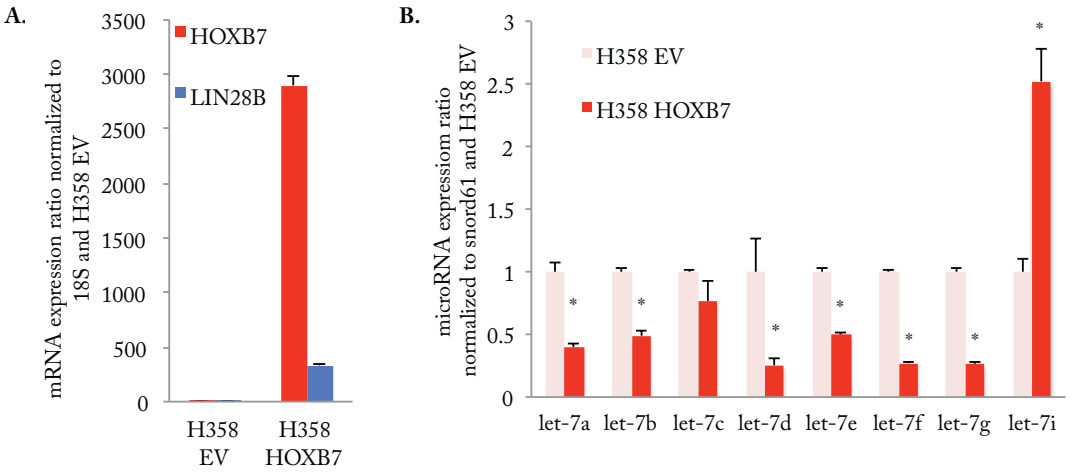


Figure 3.16: **LIN28B is up-regulated in H358 cells overexpressing HOXB7, resulting in Let-7s microRNA down-modulation. A.** RTqPCR analysis of HOXB7 and LIN28B expression performed on RNA extracts from H358 cells infected with pBABE-Hygro-EV or pBABE-Hygro-HOXB7. **B.** Expression ratio of Let-7s microRNA family members, relative to H358-EV cells by RT-qPCR analysis of cells as in A. *, significant p-value ($p<0.05$; Student's t-test, two tailed).

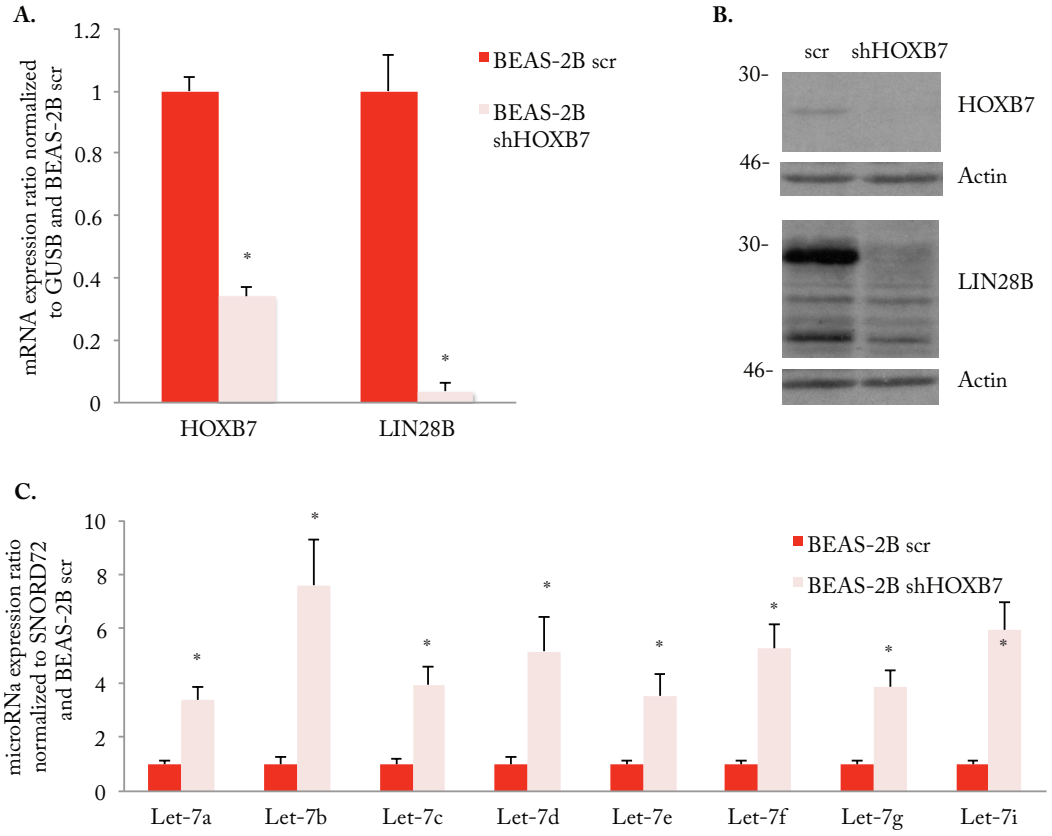


Figure 3.17: **Silencing of HOXB7 in BEAS-2B determines the down-regulation of LIN28B and the up-regulation of Let-7s.** Silencing of HOXB7 in BEAS-2B cells infected with the lentivector construct pSICOR-shLIN28B, BEAS shHOXB7. Negative control: pSICOR-scr). **A-B.** RT-qPCR and WB analysis of HOXB7 and LIN28B expression in BEAS shHOXB7 and control cells. **C.** Expression ratio of Let-7s microRNA family members by RT-qPCR analysis of cells as in A. *, significant p-value ($p<0.05$; Student's t-test, two tailed).

among 46 vertebrates (Fig.3.19 B). In the 1Kb region upstream of the TSS region (i.e. 1Kb region) we found an enrichment of predicted HOX TF binding sites with highly conserved sequences (Fig.3.19 B). Additional predicted TF sites were present within the first intron of LIN28B (i.e., Intron region, [Chang et al., 2009, Iliopoulos et al., 2009]). In order to verify the HOXB7 transactivation activity on the LIN28B promoter, we decided to use a luciferase reporter assay and to perform a Chromatin Immunoprecipitation (ChIP) analysis.

3.3.3 HOXB7 is involved in transcriptional activation of LIN28B: Luciferase assay

Dual-Luciferase assay setup

Dual-Luciferase reporter assay system consists in the expression and measurement of two individual reporter enzymes within a single system. In our case, the expression of the firefly luciferase under the control of different LIN28B promoter constructs was normalized to the expression of the *Renilla reinforces* (known as sea pansy) luciferase, which was under the control of the herpes virus thymidine kinase (HSV-TK) constitutive promoter. The TK promoter provided a moderate but constant expression of renilla luciferase to give an internal control of transfection efficiency. In order to reduce the signal crosstalk during the assay, we performed a co-transfection experiment to optimize the ratio of the co-reporter vectors added to the transfection mixture.

We co-transfected cells with a pGL3-SV40 promoter vector (constitutively active) and a pRL-TK vector at different ratios. From this set-up experiment, we concluded that 1:0.05 is the best ratio to suppress the occurrence of trans effect between promoter elements and provides a low but constant level of luciferase control activity (Fig.3.18 A).

We then cloned three different LIN28B promoter constructs upstream the luciferase reporter gene into a pGL3 vector, according to the promoter scheme shown in figure

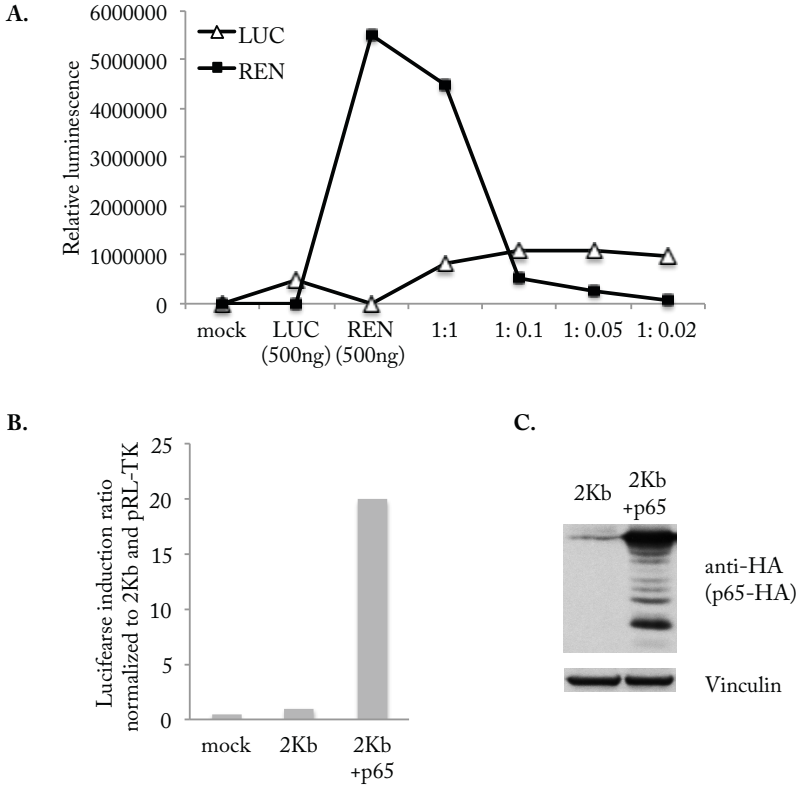


Figure 3.18: **Dual-Luciferase reporter assay setup A.** Luminescence produced by cells transfected with a pGL3-SV40 promoter vector and a pRL-TK vector, at different ratios (1:1, 1:0.1, 1:0.05, 1:0.02, SV40 and TK). **B.** Activity of the 2Kb promoter reporter construct in Phoenix cells with or without p65-HA expression normalized to *Renilla* luciferase expression; not transfected cells as control (mock). **C.** WB analysis of p65-HA expression on whole cell extracts from Phoenix cells (anti-HA WB); Vinculin: loading control.

3.19: pGL3-2Kb, pGL3-1Kb and pGL3-Intron.

In order to test the promoter functionality of our constructs, we co-transfected Phoenix Amphotropic cells with: i) a vector carrying the HA-tagged p65 subunit of NFkB, know to transactivate LIN28B expression through binding to the first intron region [Iliopoulos et al., 2009]; ii) a pGL3-2Kb vector (containing all promoter sequence); and iii) a pRL-TK vector as normalizer. Reporter activity increased almost by 20 folds when p65-HA was overexpressed compared to controls (overexpression of p65-HA was assessed by western blot using a commercial monoclonal antibody anti-HA, Fig.3.18B).

Luciferase assay revealed HOXB7-induced transcriptional activation of LIN28B

By luciferase assay, we found that HOXB7 was indeed able to induce up to 22-fold higher reporter activity with the 2Kbp-construct compared to control cells not overexpressing HOXB7. Similarly, luciferase induction was obtained using the 1Kb region upstream of the TSS of LIN28B. Conversely, induction of the reporter gene was impaired when the Intron sequence was used as promoter (Fig.3.19 D-E).

These experiments confirmed that HOXB7 is involved in LIN28B transcriptional activation and that the preferential HOXB7 binding region lies within the 1000 bp region upstream of the LIN28B TSS.

3.3.4 HOXB7 is involved in LIN28B transcriptional activation: ChIP assay

Setting-up ChIP assay conditions

In order to determine whether HOXB7 is associated with the promoter region of LIN28B, we decided to perform a Chromatin Immunoprecipitation assay (ChIP). This technique requires four main steps: i) Crosslink of DNA and associated proteins (chromatin); ii) Chromatin sonication to obtain 200-300bp DNA fragments; iii) Immunoprecipitation of the cross-linked DNA-protein fragments with a specific antibody; and iv) De-crosslinking and enrichment of specific DNA sequences by RTqPCR with specific primers. We set up and optimized these steps as follows:

Antibody setup.

We initially tested a HOXB7 monoclonal antibody by IP but it failed to efficiently immunoprecipitate the protein (Fig. 3.20 D). Thus, we decided to take advantage

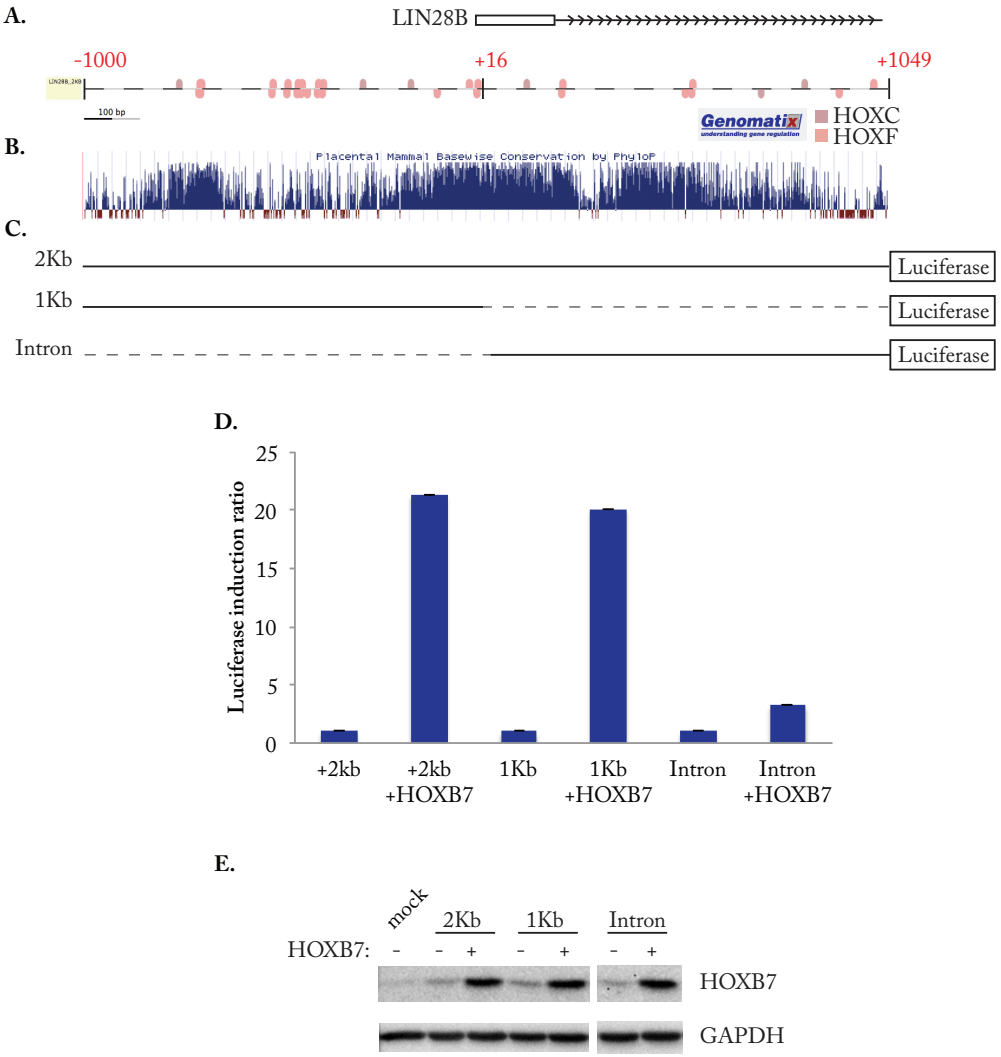


Figure 3.19: **HOXB7 induces LIN28B transcriptional activation. A.** Schematic representation of the genomic region extending from -1000bp upstream of the transcriptional starting site (TSS) of LIN28B to +1049bp downstream of the TSS (including the sequence for the first exon, blue bar, and the first intron). In pink, the predicted homeodomain transcription factor binding sites (predicted using Genomatix software). The qPCRamplicon is represented by the black line. **B**. The phyloP plot depicts evolutionary conservation between 46 vertebrate species. Predicted conserved sites are assigned positive scores (blue), while sites predicted to be fast-evolving are assigned negative scores (red). **C**. Schematic representation of the LIN28B promoter sequences cloned in pGL3 basic vector. **D**. The activity of promoter reporter constructs in Phoenix cells with high HOXB7 expression (normalized to low HOXB7 expression). Values represent relative firefly luciferase activity normalized to *Renilla* luciferase activity. **E**. WB analysis of HOXB7 expression on whole cell extracts from Phoenix cells transfected with pBABE-HOXB7 (+) or pBABE empty vector (-) and the specific pGL3 construct; control: not transfected cells (mock).

of a pCDNA-FLAG-HOXB7 construct in order to overexpress a FLAG-protein (an analogous strategy was adopted by Jin and colleagues [Jin et al., 2012]) in our cells. Protein overexpression was achieved through transfection of plasmid DNA in Phoenix Amphotropic cells and A549 lung cancer cells. Immunoprecipitation of FLAG-HOXB7 with an anti-FLAG antibody or anti-FLAG beads increased the IP efficiency (Fig. 3.20 E-F).

For the ChIP experiment, we used an antibody directed against Histone H3 as a positive control for the IP step and a generic IgG directed against an unrelated cytosolic protein (e. g. anti-DPDC1B) as a negative control, defining thus the background of the assay.

Primer setup.

We designed four couples of primers in order to obtain amplicons along the 1Kb promoter region, as shown in fig. 3.21. Using the PCR tool of Genome Browser PCR tool (https://genome.ucsc.edu/cgi-bin/hgPcr), we aligned amplicons to the genome to control for their target specificity (human genome assembly: Feb 2009 GRCh37/hg19, Table.2.3). The chromatin to be used as template for RTqPCR was then sonicated: the ratio between template quantity and cycle detection appeared to be linear, a prerequisite for a successful ChiP (Fig. 3.20A).

To test for the presence of HOXB7 on the Intron region, we took advantage of one couple of primers used by Chang and colleagues when performed the ChIP against MYC on the LIN28B promoter (Chan3). We also used the primer couple "Chan2", designed to anneal close to the TTS of LIN28B (Table.2.3) [Chang et al., 2009]. As positive control, we used primers that anneal to the EGFR-promoter region used by Jin and colleagues to demonstrate HOXB7 and EGFR-promoter direct interaction [Jin et al., 2012].

Chromatin sonication setup.

We optimized the setting for the chromatin sonication in order to obtain 200-300

bp long DNA fragments, the optimal suggested size for ChIP experiments. Ten cycles resulted enough to obtain the right size of DNA fragments (as visualized in the agarose gel of fig. 3.20 C) and did not affect DNA integrity (RTqPCR cycle detection remained unchanged even after several cycles of sonication; Fig. 3.20 B).

ChIP assay revealed a direct involvement of HOXB7 in LIN28B transcriptional activation.

ChIP analysis revealed up to a 15.4 fold increment in HOXB7 binding to the LIN28B promoter region in HOXB7 expressing cells compared to EV control (primer d, fig. 3.21 A). The fold increment in HOXB7 binding decreased to 2.7 and 4.4 folds moving along the 2kb region toward the 3' or 5'-end respectively (primer d, fig. 3.21 A).

To increase the resolution of this experiment, we tested anti-FLAG beads instead of the monoclonal antibody. We found that the HOXB7 LIN28B-promoter binding was increased up to 30 folds in the 1kb region and 6 folds in the first intron (Fig.3.21 B). As control, we observed a ∼10 fold increase in HOXB7 binding to the EGFR promoter regions, in line with a previous study [Jin et al., 2012].

We further validated these data by immunoprecipitation of the FLAG-HOXB7 protein transiently transfected into A549 lung adenocarcinoma cells. A sustained HOXB7 binding to LIN28B 1kb region was obtained (up to 18-folds with primers "d") while the signal was dramatically reduced when a FLAG peptide competed for the binding with the beads, demonstrating the specificity of our ChIP experiment (Fig.3.21 C).

In summary, we have identified a region upstream of the LIN28B TSS sufficient for HOXB7-dependent transcriptional activation in Phoenix and A549 cells. The HOXB7-mediated induction of LIN28B may suggest a possible mechanism of action of HOXB7 relevant for lung cancer and for stem/tumor-initiating-cell biology.

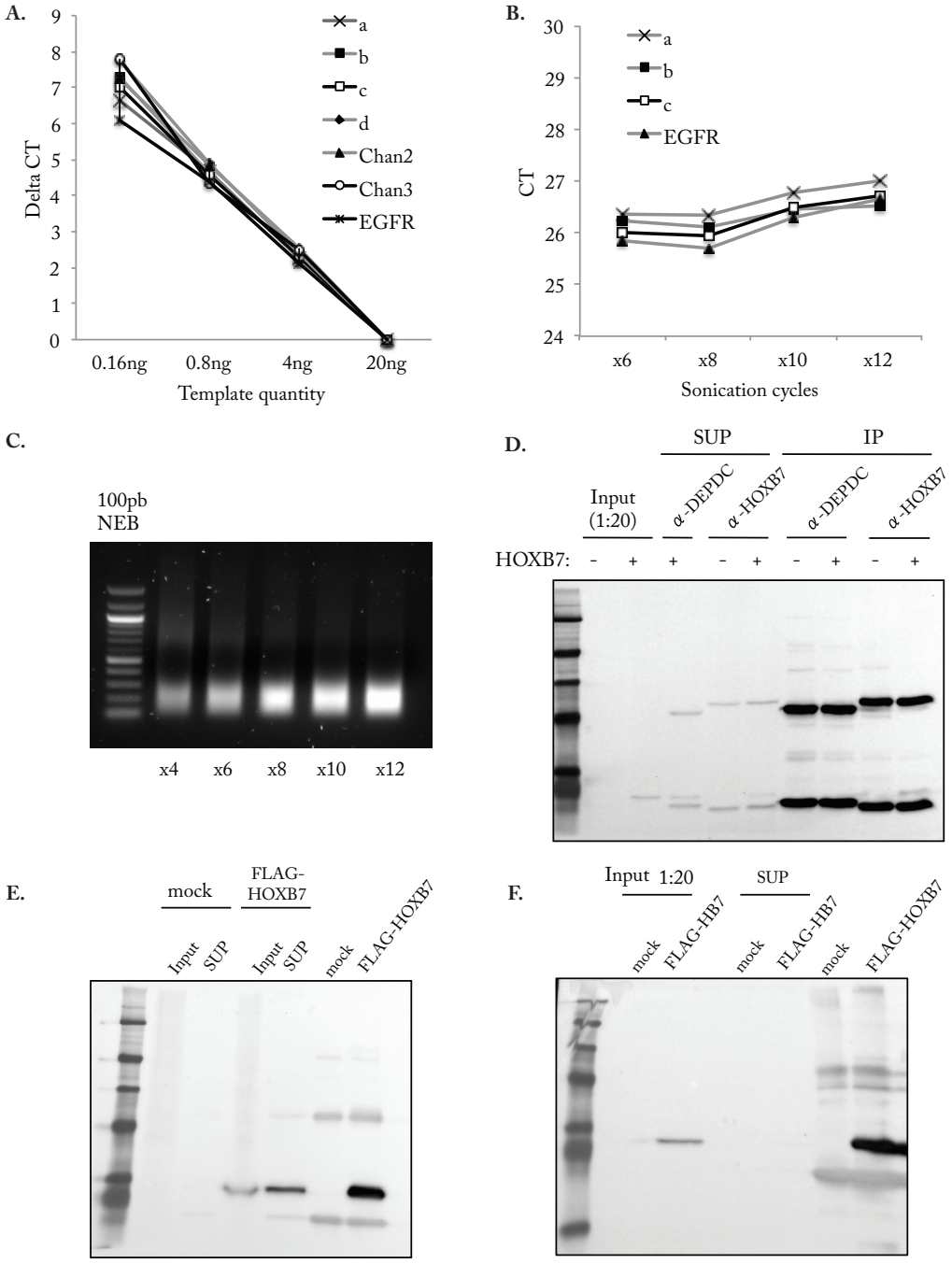


Figure 3.20: **ChIP assay conditions setup. A.** Primer setup: the plot shows the relative ratio between template quantity and cycle detection. X-axis: quantity of sonicated chromatin used as template for PCR; Y-axis: difference in cycle detection (Delta CT) relative to detection with 20ng template. **B.**Sonication setup: cycle detection (Y-axis) at increased number of chromatin sonication cycles (from 4 to 12 cycles). 5ng of the template were loaded in the PCR mixture. **C.**Sonicated chromatin was loaded into a 1% agarose gel in order to visualize the magnitude of the DNA fragments. DNA marker: 100pb NEB. **D.** Antibody setup: HOXB7 immunoprecipitation (IP) with anti-HOXB7 (4ug) from 1mg of NCI-H358 cell lysate overexpressing (+) or not (-) HOXB7. Input: 50ug cell lysate. SUP: supernatant. A generic IgG direct to an unrelated cytosolic protein (e. g., anti-DPDC1B) was used as negative control. The samples were loaded into a 12% acrylamide gel and WB with anti-HOXB7. **E-F.** IP with anti-FLAG (E) or anti-FLAG beads (F) from 1mg Phoenix cell lysate expressing the N-FLAG-HOXB7 construct (mock: lysate from cells transfect with pCDNA3.1 FLAG EV). Input: 50ug cell lysate. SUP: supernatant. The samples were loaded into a 12% acrylamide gel and WB with anti-HOXB7.

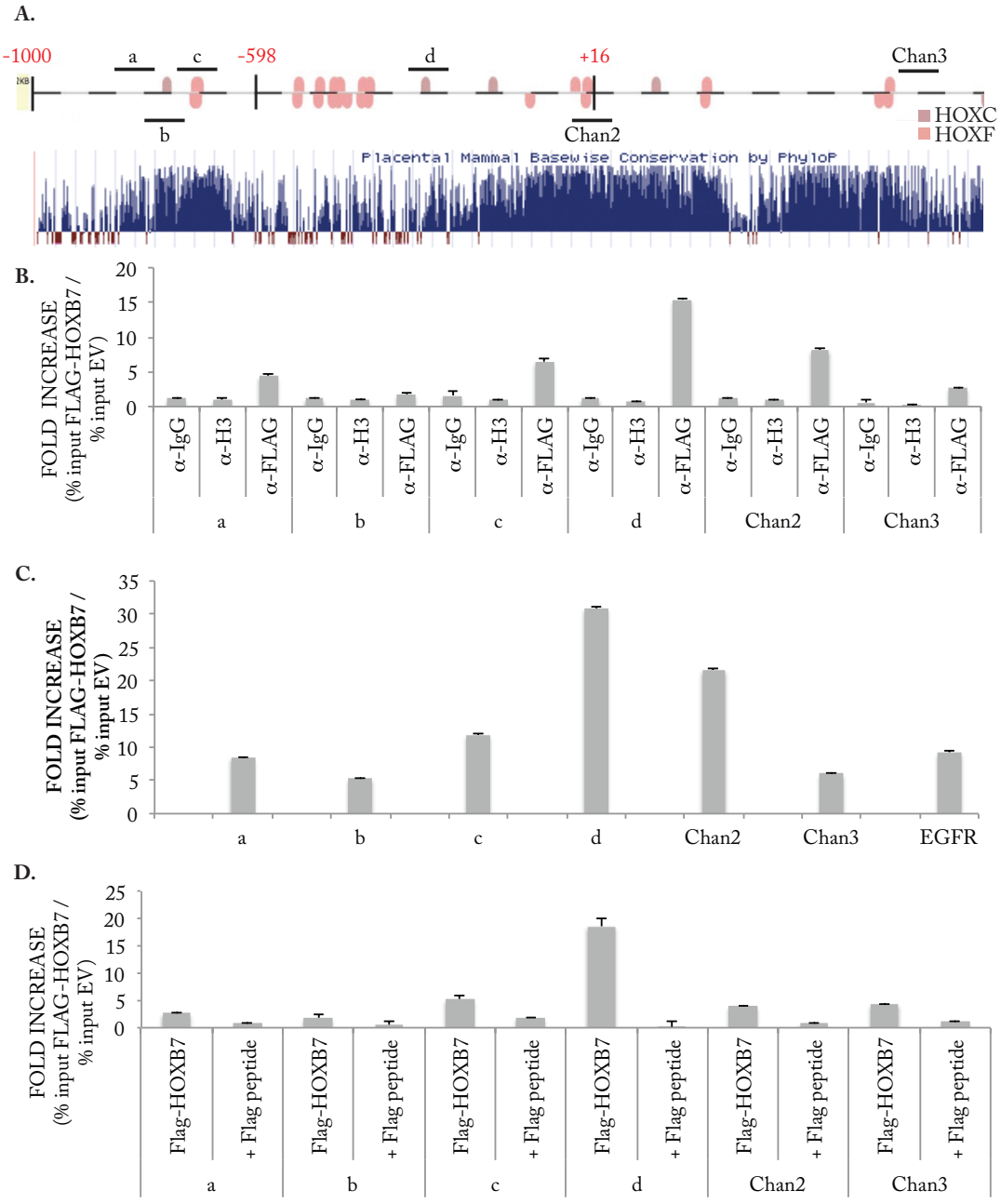


Figure 3.21: **ChIP assay revealed the presence of HOXB7 on LIN28B promoter region.** **A.** Schematic representation of LIN28B promoter region highlighted with predicted homeodomain transcription factor binding sites (pink) and phyloP plot. **B.** ChIP with anti flag. Fold increase of % input in cells overexpressing FLAG-HOXB7 Phoenix cells and control cells (EV); input: 1% of starting chromatin; IP controls: anti-IgG or anti-H3. PCR amplification was performed with six different couples of primers (a, b, c, d, Chan2 and Chan3). **C.** ChIP with anti-FLAG beads. Fold increase of % input in cells overexpressing FLAG-HOXB7 cells and control cells. **D.** ChIP was performed with beads anti-FLAG in A549 cells overexpressing FLAG-HOXB7 (or FLAG-EV). A flag peptide was added to the IP mixture to compete for the binding.

3.4 Silencing of LIN28B prevents HOXB7-mediated increase of stem markers in BEAS-2B cells.

To explore the interaction of HOXB7 and LIN28B in the biology of cancer cells, we knocked down the LIN28B gene in BEAS-2B cells using lentiviral infection with a pSICOR vector carrying short hairpin RNAs specific for LIN28B (we used shRNA oligo3, the best performing oligo in A549 cells: 85% of LIN28B silencing; fig.5.4 A). After selecting infected cells with puromycin, a 60% LIN28B silencing efficiency was scored in BEAS-2B, which was, however, biologically sufficient to allow 2.5 fold increase in the expression of mature Let-7 microRNAs, compared to BEAS-scr control cells (Fig.3.22 C).

We then overexpressed HOXB7 in BEAS-2B cells and assessed whether subsequent LIN28B induction by HOXB7 was abrogated by the short hairpin RNA construct. Indeed, we scored a 2-folds induction of LIN28B in BEAS-scr cells (compared to cells infected with pSICOR-scr and pBABE-EV), while the overexpression of HOXB7 in BEAS-shLIN28B led to a LIN28B expression comparable to control cells, BEAS-scr-EV (Fig.3.23 A-B).

In order to understand whether and which of the HOXB7-dependent phenotypes may require LIN28B activation to occur, we went back to analyze the expression of the panel of EMT and Stem marker genes by RTqPCR.

In BEAS-scr-HOXB7 cells we obtained a significant induction ($p < 0.05$, Student's t-test, two-tailed) of N-cadherin and Vimentin (12.6 and 7.8-fold induction respectively compared to BEAS scr-EV cells), which is in line with an epithelial to mesenchymal transition (Fig.3.23 C). Interestingly, the lack of LIN28B induction upon HOXB7 over-expression in the BEAS shLIN28B + HOXB7 did not impair the induction of these mesenchymal markers (11.7 fold induction for N-cadherin and 7.1-fold Vimentin), which suggests that HOXB7 drives EMT gene regulation independently of LIN28B activation, possibly through the direct induction of TGFb as recently published [Liu et al., 2015].

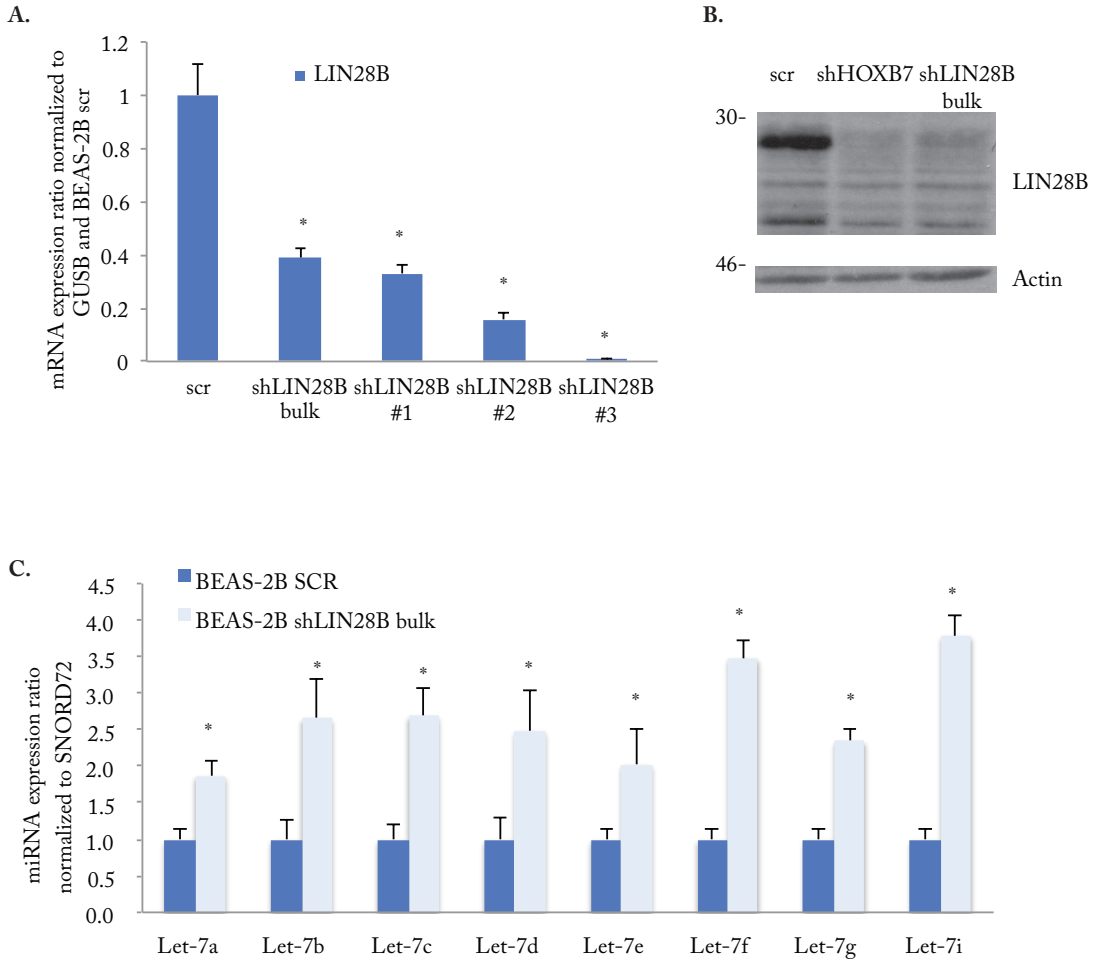


Figure 3.22: **Silencing of LIN28B in BEAS-2B determines Let-7s upregulation.** Down-regulation of LIN28B in BEAS-2B cells by infection with the lentivector construct pSICOR-shLIN28B (negative control: pSICOR-scr). Three single clone populations of BEAS shLIN28B were selected. **A** Expression analysis of LIN28B by RTqPCR in BEAS-shLIN28B cells (bulk and clone populations) normalized to LIN28B expression in BEAS-scr cells. **B.** WB analysis of LIN28B in BEAS-scr BEAS-shHOXB7 and BEAS-shLIN28B bulk cells. **C.** Expression ratio of Let-7s microRNA family members by RT-qPCR analysis of cells as in B. *, significant p-value ($p<0.05$; Student's t-test, two tailed).

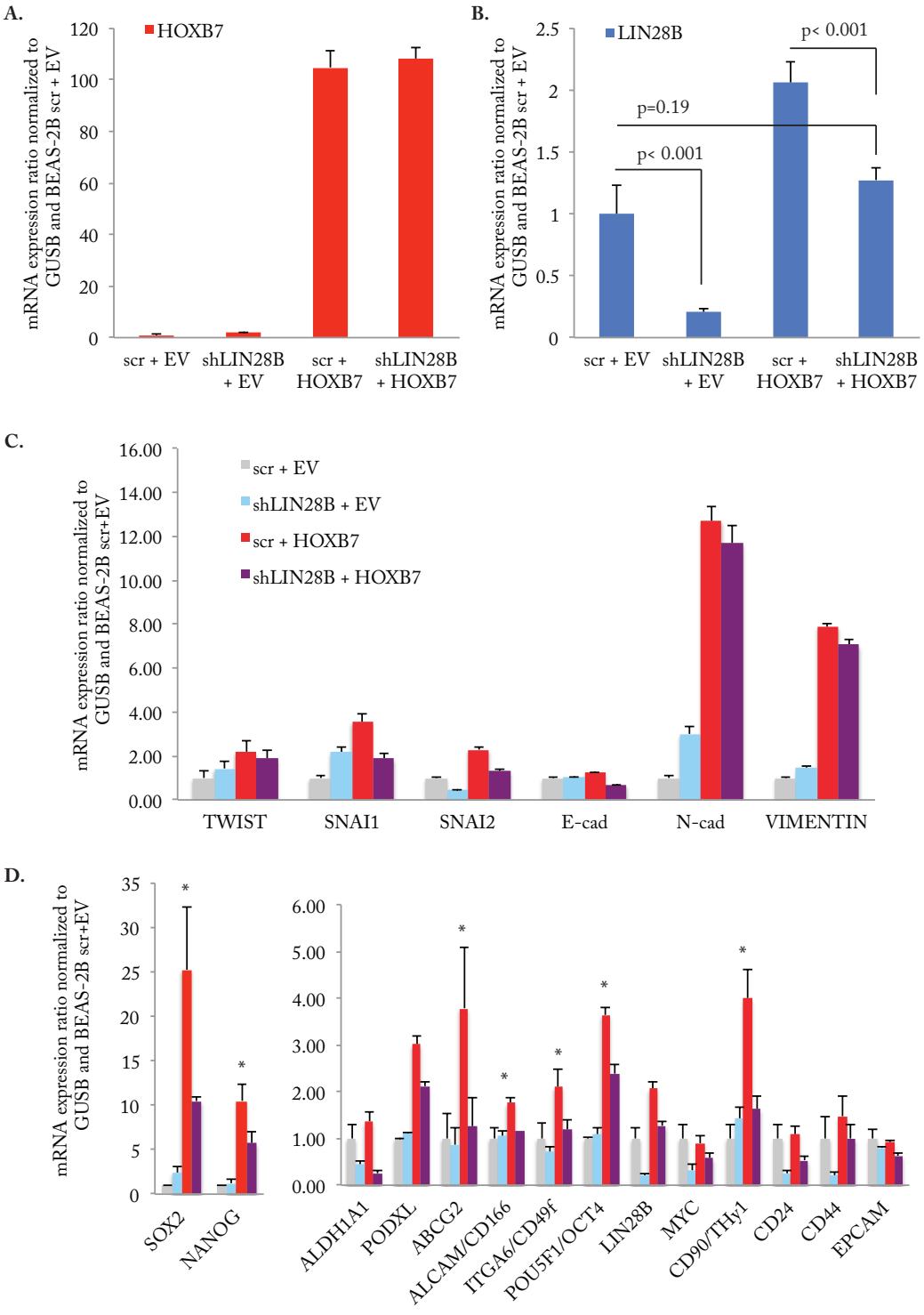


Figure 3.23: **Silencing of LIN28B in BEAS-2B cells prevents HOXB7-dependent induction of stem marker genes.** BEAS-2B cells were first infected with pSICOR shLIN28B construct (or pSICOR-scr as control) and then with pBABE-HOXB7 construct (or pBABE-EV as control). The shLIN28B prevents the induction of LIN28B when HOXB7 is overexpressed. Gene expression analysis of **A.** HOXB7, **B.** LIN28B, **C.** EMT and **D.** stem/iPS markers was performed by RTqPCR. *****, genes significantly enriched in scr +HOXB7 cells respect to scr +EV cells and significantly dowregulated in shLIN28B +HOXB7 cells respect to scr +HOXB7, when there is not any significant difference between gene expression in shLIN28B +EV and scr +EV (p-values were calculated using Student's t-test, two tailed).

Conversely, the up regulation of seven stem marker genes (SOX2, NANOG, ABCG2, ALCAM/CD166, ITGA/CD49f, POUF5F1/OCT4 and CD90) observed in HOXB7 overexpressing cells was prevented upon LIN28B silencing. In some cases the expression level was comparable to that of control cells, scr+EV (Fig.3.23).

We then analyzed the surface stem marker CD90 by FACS analysis. $CD90^{high}$ population, which positively correlates with sphere forming capacity, resulted more than halved in shLIN28B + HOXB7 cells compared to control cells, scr + HOXB7 (11.47% and 25.5% of cells, respectively; Fig.3.24 A). Interestingly, also in shLIN28B + HOXB7 cells the $CD90^{high}$ population was the highest HOXB7 expressing cell population, followed by $CD90^{int}$ and, lastly, by $CD90^{low}$ population (Fig.3.24 B). Indeed, stem marker gene analysis revealed an axis between HOXB7 and LIN28B with a possible role in the regulation of genes involved in stem cell biology.

New emerging evidences are also highlighting a novel important role for LIN28B in cell reprogramming processes [Chien et al., 2015]. This finding in a cancer related contest is intriguing because it is known that aggressive tumor cells may loose characteristics of the tissue of origin and acquire a more undifferentiated phenotype in order to drive metastatisation[Kim et al., 2005, Pece et al., 2010, Alison et al., 2010]. Cell reprogramming allows, indeed, terminally differentiated cells to revert to an undifferentiated and stem-like condition [Takahashi and Yamanaka, 2006]. We thus wondered whether HOXB7, through the induction of LIN28B, may participate to this cancer relevant feature, as already suggested by our GSEA analysis (Fig.3.1).

3.5 HOXB7 enhances the efficiency of cell reprogramming.

According to Yamanaka and colleagues, only four TFs are sufficient to convert mouse embryonic fibroblast (MEFs) into iPS cells: SOX2, OCT4, KLF4 and c-MYC [Takahashi and Yamanaka, 2006]. All four TFs are not essential, as two of them

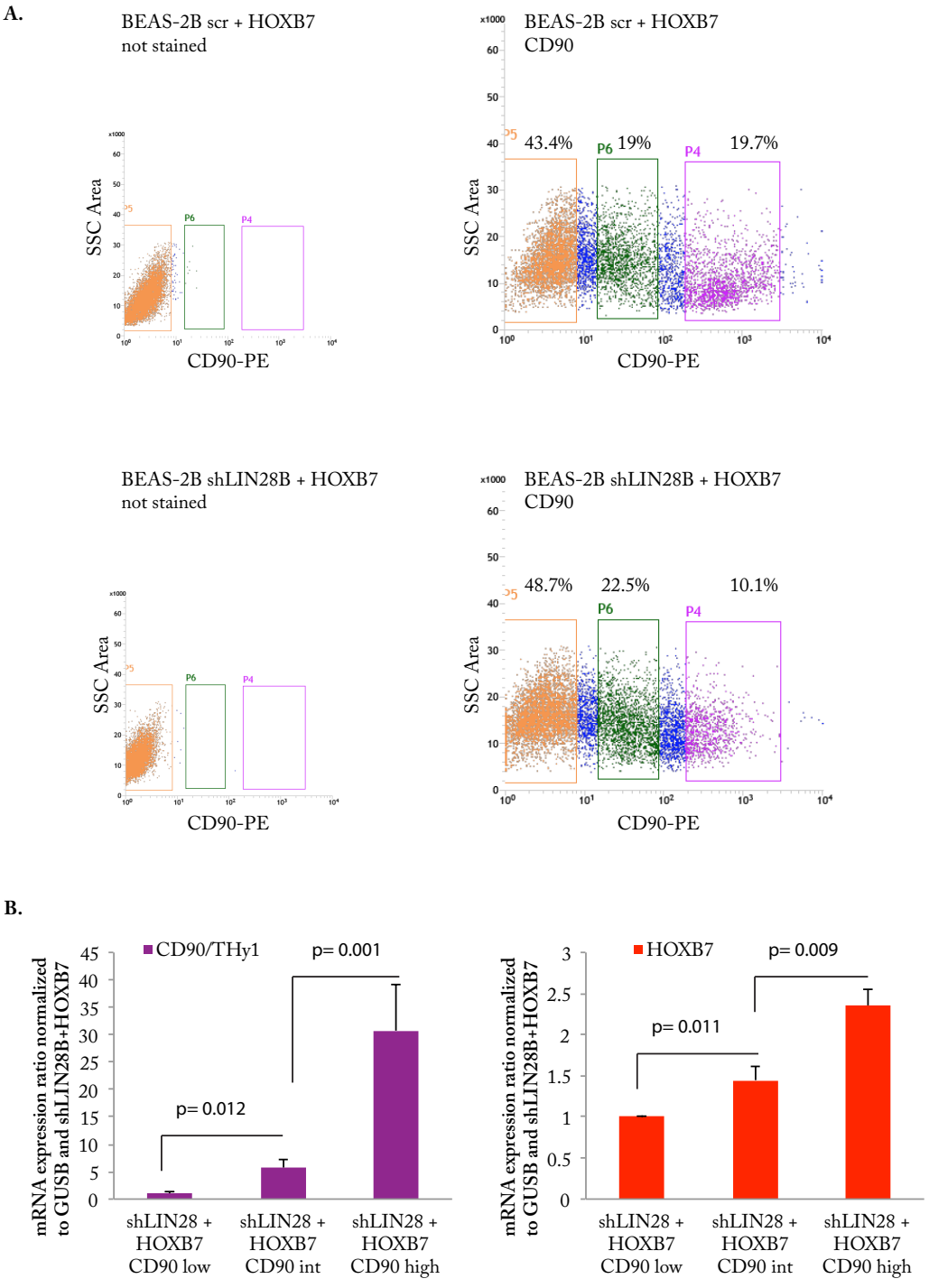


Figure 3.24: **Silencing of LIN28B in BEAS-2B cells prevents HOXB7-dependent induction of the stem marker gene CD90. A.** FACS analysis of the CD90 cell surface marker in BEAS shLIN28B+HOXB7 cells and control (scr+HOXB7 and un-stained) cells. Three populations are highlighted according to the expression of CD90 (high, low or intermediate): $CD90^{high}$ in purple, $CD90^{int}$ in green, $CD90^{low}$ in orange. **B.** Expression analysis of of CD90 and HOXB7 by RTqPCR with total RNA extracts from the $CD90^{high}$, $CD90^{int}$ and $CD90^{low}$ populations of BEAS shLIN28B+HOXB7 cells. P-values were calculated using Student's t-test, two tailed.

can be substituted with other genes and efficient cell reprogramming will still occur. Indeed, it has been shown that the LIN28B gene can stand in for c-MYC to form iPS (Yamanaka and Tanabe, 2011; Patent number: 8993329), supported by the fact that c-MYC activates the transcription of LIN28B [Chang et al., 2009]. The LIN28B/Let-7 axis has been shown to play a critical role in the reprogramming of normal human oral terminally differentiated cells (i.e., keratocyte) into iPS together with OCT4, SOX2 and NANOG [Chien et al., 2015].

Thus, we reasoned that HOXB7 might take part in cell reprogramming processes by substituting c-MYC in Yamanaka's transcription factor cocktail similar to LIN28B.

3.5.1 HOXB7 positively contributes to reprogramming of mouse embryonic fibroblast.

In order to verify this hypothesis, we used MEFs derived from a mouse carrying a knock-in of EGFP coding sequence in the 3'-UTR of endogenous Oct4 to expressed GFP concomitantly with Oct4 expression in iPS cells [Unternaehrer et al., 2014]. We infected these fibroblasts with a lentiviral vector expressing human OCT4, KLF4 and SOX2 (OKS) together with a pBABE-HOXB7 or a pBABE-EV construct. The reprogrammed (iPS) colonies were then visualized by alkaline-phosphatase assay: two weeks post infection a ∼30 fold increase of iPS colonies was visible when HOXB7 was overexpressed compared to control (Fig.3.25 A-C). Indeed the reprogramming efficiency was 0.21% and 0.43% in MEF-OKS-HOXB7 cells and 0.006% and 0.014% in control cells after two and three weeks, respectively.

Finally, to verify that full reprogramming was achieved, we isolated two iPS clones obtained by OKS expression alone or with HOXB7 (both resulted positive for GFP expression; Fig. 3.25 B). We tested their ability to differentiate along the three embryonic lineages through the formation of teratomas. Differentiation was assessed by combining hematoxylin/eosin with immunohistochemical stainings for lineage-specific markers (Desmin and S-100 for mesodermal, and Cytokeratin for ectodermal lineages)

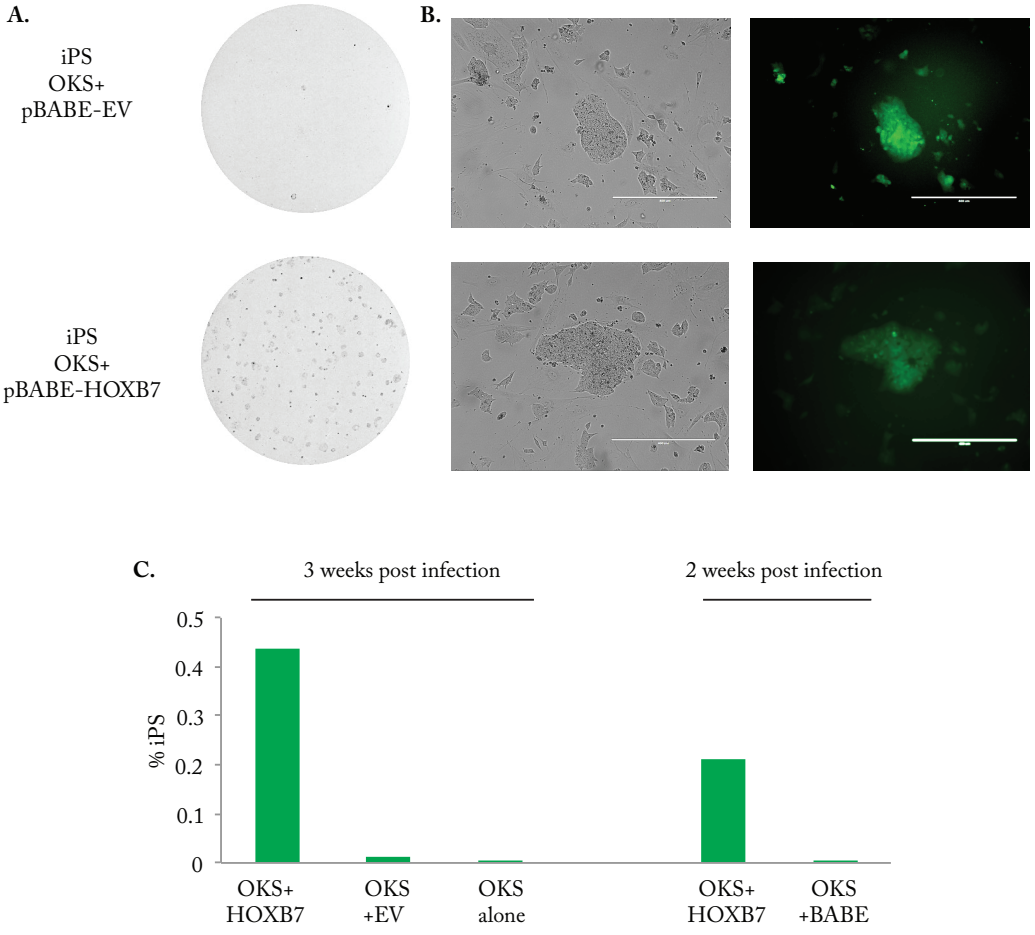


Figure 3.25: **HOXB7 positively contributes to MEF reprogramming. A.** Alkaline phosphatase staining of control MEFs (lower row) and MEFs overexpressing HOXB7 (upper row) two weeks post infection with a lentiviral vector expressing OCT4, KLF4 and SOX2 (OKS). **B.** Light microscopy (left) and Oct4-driven GFP fluorescence (right) images of iPS clones isolated from cells as in A. **C.** Quantification of iPS colonies obtained two or three weeks post lentiviral transduction with the indicated constructs (n=1).

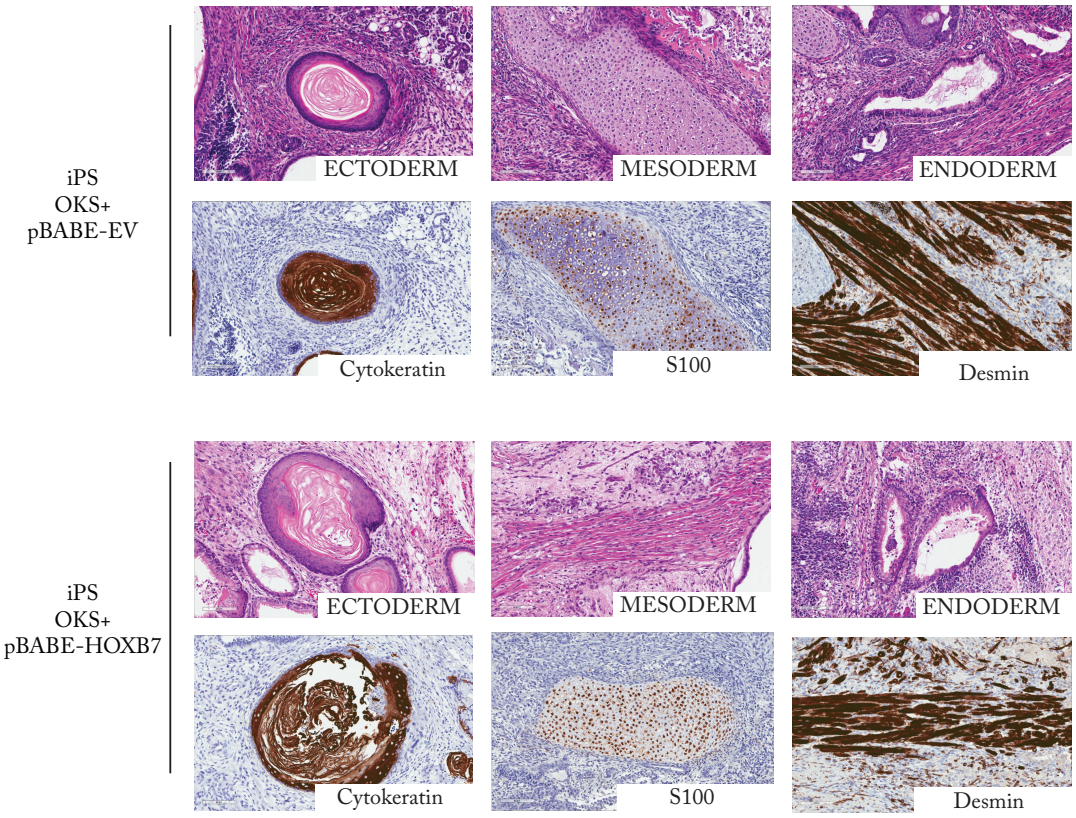


Figure 3.26: **iPS clones can give rise to a complete teratoma.** Hematoxylin and eosin staining and immunohistochemical (IHC) analysis of representative sections of teratomas generated from OKS control or OKS+HOXB7 iPSC cells. Stainings for the ectoderm marker cytokeratin (on the left) and the mesoderm markers S100 (in the middle) and desmin (on the right) are displayed.

(Fig.3.26). We found that both iPS clones had the potential to give rise to differentiated tissue derived from all three primary germ layers: epithelial annexes (ectoderm), cartilage and muscles (mesoderm) and gut epithelium (endoderm). In conclusion, we were able to achieve full reprogramming of mouse fibroblasts with OCT4, KLF4 and SOX2. The reprogramming was ∼30 fold more efficient in combination with HOXB7 overexpression.

In order to quantify the strength of HOXB7 in affecting reprogramming processes, we will next compare reprogramming efficiency using OKS+HOXB7 constructs or the canonical four-TFs protocol (i.e., OKS+MYC).

3.5.2 HOXB7 positively contributes to reprogramming of human lung BEAS-2B cells by activating LIN28B.

Cell reprogramming may play a critical role in lung cancer progression, giving rise to stem-like cells resistance to chemotherapy and cause tumor relapse [Suvà et al., 2013]. Thus, studying this process in our human epithelial BEAS-2B cell line model system was challenging.

We forced the expression of OCT4, KLF4 and SOX2 (OKS) in BEAS-2B-EV and BEAS-2B-HOXB7 cells and waited three weeks in stem conditions before evaluating their pluripotent status by alkaline phosphatase staining. In line with the reprogramming experiment in MEFs, HOXB7 overexpression positively contributed to generating iPS colonies from BEAS-2B cells. Reprogramming efficiency was more than doubled: we obtained 0.29% of iPS colonies from BEAS-HOXB7-OKS cells and 0.11% from control cells (p=0.02; fig.3.27 A-B). When we induced cell reprogramming with the canonical set of four genes (e.g., OKS+MYC), we were not able to stabilize any iPS colony due to over-proliferation of epithelial cells induced by the MYC oncogene, as already observed in other studies [Xu et al., 2013].

In order to understand whether HOXB7-induced reprogramming efficiency requires LIN28B increased expression, we forced the expression of OCT4, KLF4 and SOX2 (OKS) in BEAS-scr +HOXB7 and BEAS-shLIN28B +HOXB7 cells and assessed their capability to form iPS colonies. Pluripotent colonies were visualized by alkaline phosphatase assay. We scored a 1.4-fold reduction in BEAS-shLIN28B cells compared to scr cells (p=0.08, fig.3.27 C). We analyzed also a BEAS-shLIN28B clone characterized by 90% of gene silencing (BEAS shLIN28B clone #3, Fig.3.22 A). With this clone, overexpression of HOXB7 was impaired in positively contributing to reprogramming, scoring a 0.01% of reprogramming efficiency (p=0.0001, fig.3.27 C).

These results pointed to a role for HOXB7 in the activation of a transcription program, involving the LIN28B pathway, relevant for stem/iPS cell biology. Since it is known that the enhancement of stem cell compartment increases tumor aggressiveness,

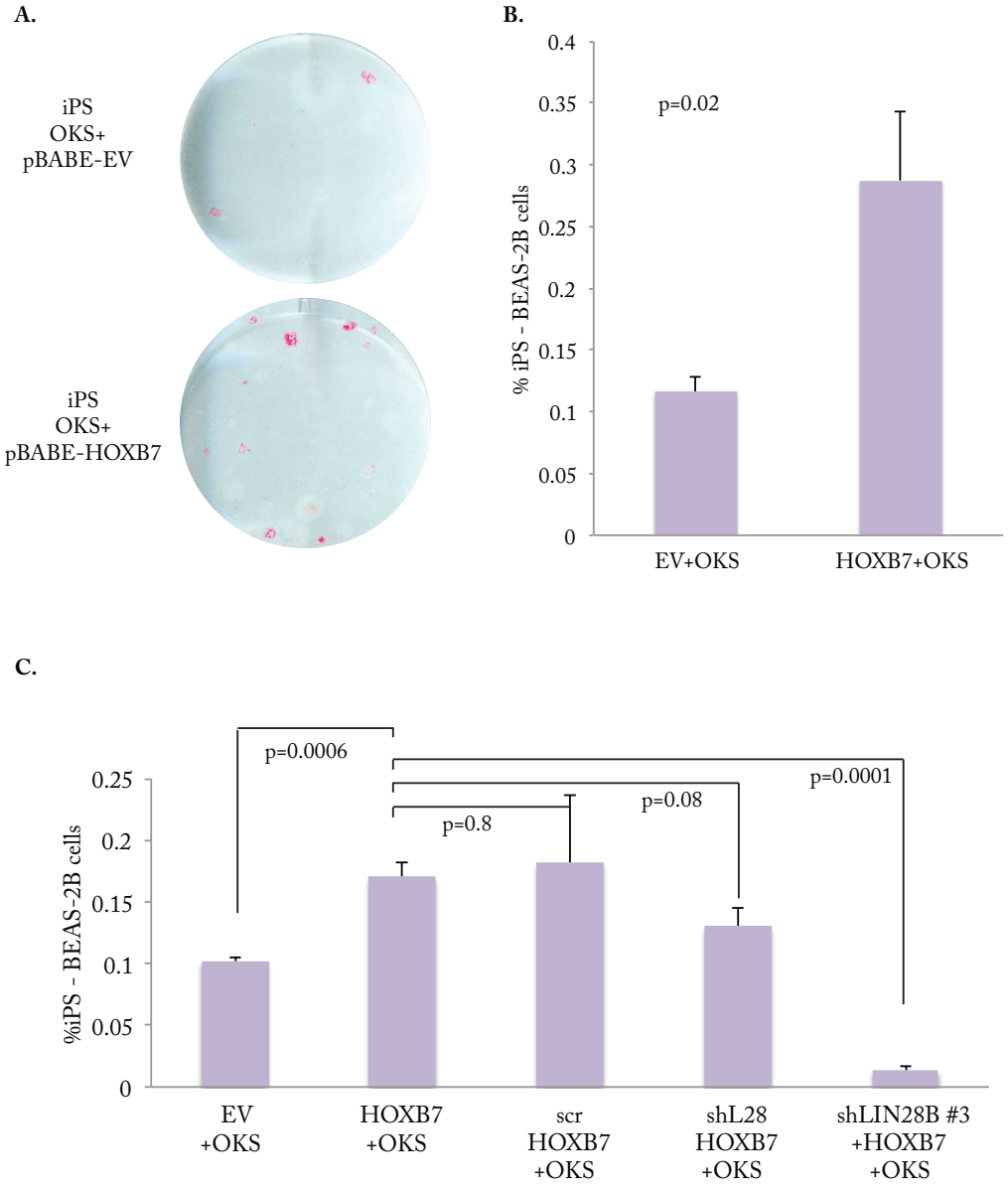


Figure 3.27: **HOXB7 increases the efficiency of cell reprogramming in BEAS-2B cells.** **A.** Alkaline phosphatase staining of BEAS-2B overexpressing HOXB7 and control cells two weeks post lentiviral transduction of the construct for the expression of OCT4, KLF4 and SOX2 (OKS). **B.** Quantification of the alkaline phosphatase assays (n=2, p-values were calculated using Student's t-test, two tailed.). **C.** Quantification of the alkaline phosphatase assays, including analysis of the reprogramming efficiency in BEAS-2B cells carrying shRNAs for silencing LIN28B and scrambled control (scr, n=2).

the mechanism we proposed can, at least in part, explain the role of HOXB7 in lung cancer progression. Further analysis is required to elucidate whether HOXB7 may have an oncogenic role in lung cancer and whether it might be involved in chemotherapy resistance, which would suggest it as a candidate for novel molecularly targeted therapies.

Chapter 4

Discussion

4.0.3 HOXB7 expression is prognostic in patients with early-stage of lung adenocarcinomas.

We have previously shown that a increased expression of HOXB7 is a hallmark of adverse prognosis in lung cancer patients [Bianchi et al., 2007]. Our study integrated the gene expression profile of three cohorts of lung cancer patients to identify genes correlating with poor prognosis, and a gene signature derived from a cancer cell model system (E1A infected terminal differentiated mouse myotubes forced to reenter the cell cycle). Recently, HOXB7 was confirmed to be a prognostic biomarker in lung adenocarcinoma patients using an independent approach (i.e., immunohistochemistry) [Yuan et al., 2014].

In line with these reports, we found that HOXB7 is significantly more expressed in cancer samples by comparing gene expression analysis of 468 lung adenocarcinomas and 58 lung normal tissue samples from the LUAD-TCGA dataset. However, the precise molecular mechanism through which HOXB7 promotes lung cancer progression and metastatic spreading is still unknown.

Previous studies in breast cancer have highlighted a role for HOXB7 in the induction of EMT [Wu et al., 2006], a process that plays an important role during tumor progression [Kalluri and Weinberg, 2009]: excessive epithelial cell proliferation and ac-

quisition of motility and invasiveness, typical characteristics of mesenchymal cells, are thought to herald metastatic dissemination. Of note, we demonstrated that HOXB7 overexpression drives EMT in the normal lung cell line BEAS-2B and in the lung cancer cell line NCI-H358, suggesting that the mechanism is shared among physiological and pathological condition. Indeed, EMT has been described as a cell-biological program that allows remodeling of cells primarily during normal embryonal development and wound healing repair. More recently, EMT has been associated with tumor metastasis, enabling cancer cell dissemination (Thiery, 2003).

During the formation of metastasis, cancer cells seem to require the ability to reach the secondary tissue sites, proliferate and participate in the reconstruction of a new tissue (i.e., a macroscopic metastases), exhibiting stem-like capabilities. This observation raised a still ongoing debate about the possibility that the EMT may generate cells with properties of stem cells, revealing a convergence between these two processes [Mani et al., 2008, Kong et al., 2010]. We thus wondered whether HOXB7 may be involved in lung cancer stem cells processes.

4.0.4 A novel role for HOXB7 in acquisition of cancer stem-like properties.

To better understand the effect of alterations in HOXB7 expression in the contest of lung cancer, we used an *in silico* approach (GSEA analysis) to try to identify biological functions specifically altered in high-HOXB7 expressing lung cancers. Strikingly, in addition to gene-sets representing molecular mechanisms involved in EMT and cancer progression, we scored a significant enrichment of genes/pathways involved in stemness and cell reprogramming, upregulated in high-HOXB7 lung tumors. We have shown in an experimental model that HOXB7 was able to expand a subpopulation of cells with stem cells characteristics (i.e., anoikis resistance) and characterized by a cell surface marker that was recently shown to be tightly linked with cancer stem cells: the CD90 [Lu et al., 2014].

CD90 is a glicosylphosphatidylinositol-anchored glycoprotein (THY1) expressed by mesenchymal cells. It was recently shown to be required for the physical interaction between breast cancer stem cells and cells of the stem niche [Lu et al., 2014]. CD90 expression was characterized in lung cancer cell lines (A549 and H446 cells [Yan et al., 2013]) and in primary lung cancer cells [Wang et al., 2013]. CD90 has also been proposed to be a lung cancer stem cell marker. We have showed that the $CD90^{high}$ population in BEAS-2B cells was ∼15-fold more proficient to form spheroids in a pneumosphere assay. Thus, $CD90^{high}$ cells were those that retained the stem cell property of anoikis resistance.

Importantly, the $CD90^{high}$ population was overrepresented in HOXB7-high expressing primary lung tumors and in HOXB7 overexpressing BEAS2B cells, further pointing to a role for HOXB7 in the expansion of cancer stem cells/tumor-propagating cells. Enlargement of the stem cell compartment in cancer has been shown to be a hallmark of enhanced tumorigenicity and increased metastatic potential [Kim et al., 2005]. Therefore, our hypothesis is that HOXB7 may enhance tumor aggressiveness by inducing EMT and expanding the cancer stem cell compartment.

4.0.5 HOXB7 transcriptionally activates a gene involved in stem cell features sustainment and tumor progression: the LIN28B.

Among the stem marker genes found enriched in high-HOXB7 expressing lung tumors and cell lines, we identified LIN28B as a HOXB7 transcriptional target.

Analysis of the promoter region revealed an enrichment of homeodomain transcription factor motives in a 1kb region upstream of the LIN28B TSS. Cloning of this region upstream of a promoterless firefly luciferase reporter cassette was sufficient to drive HOXB7-dependent reporter activity, while MYC and NF-kB directly activate LIN28B expression through a binding in the first intron [Chang et al., 2009] [Iliopoulos

et al., 2009].

Accordingly, with HOXB7 transcriptional control of LIN28B expression (verified by chromatin immunoprecipitation), we found the two genes concomitantly expressed in lung cell lines and primary tumors; moreover LIN28B expression levels were modulated upon HOXB7 overexpression or downregulation in BEAS-2B and NCI-H358 cells.

The identification of LIN28B as a HOXB7 target was an important finding for us since it could explain the mechanism by which HOXB7 induces the expansion of a cell subpopulation with stem cell characteristics. Indeed, the LIN28B gene was recently described as "an emerging oncogenic driver in cancer stem cells" [Zhou et al., 2013].

LIN28B is a RNA binding protein with two homologs in mammalian, LIN28 and LIN28B, that share more than 70% of protein sequence identity. Both genes are reported to be aberrantly expressed in about 15% of human cancer cell lines and in primary tumors and have been associated with poor clinical prognosis [Viswanathan et al., 2009]. When overexpressed in NIH-3T3 cells, LIN28/B drive tumorigenesis *in vivo* characterized by delayed latency and evidence of local invasion [Viswanathan et al., 2009]. Physiologically, LIN28/B are highly expressed in stem and progenitors cells but are silenced during cell differentiation, inversely proportional to Let-7 microRNAs, a known family of oncosupressor genes [Viswanathan and Daley, 2010]. The best characterized course of action consists in direct binding of LIN28/B to Let-7 microRNA precursors (pre-Let-7s), recruitment of a non-canonical poly (A) polymerase, TUTase4 (TUT4), and the subsequent uridylylation of pre-Let-7s [Heo et al., 2009], which blocks Dicer processing and further maturation of microRNAs [Newman et al., 2008]. These events lead to the subsequent derepression of Let-7s canonical targets, such as RAS, MYC, HMGA2 and LIN28/B itself, further sustaining cell proliferation and transformation. Interestingly, there is a positive loop involving MYC and LIN28B as MYC was found on LIN28B promoter, resulting in its transcriptional activation [Chang et al., 2009]. This last observation may explain why it was possible to substitute MYC with LIN28B within the original reprogramming group of transcription factors (Yamanaka

and Tanabe, 2011; Patent number: 8993329), and inspired us to verify whether HOXB7 might take in part in cell reprogramming by substituting c-MYC in Yamanaka's transcription factor cocktail.

In our experimental model we found that HOXB7 requires, at least in part, the action of LIN28B to sustain the expansion and viability of subpopulation of cells with stem cell characteristics. The $CD90^{high}$ population, which positively correlates with sphere forming capacity, resulted more than halved upon LIN28B silencing in HOXB7 overexpressing cells compared to control cells.

This line of investigation was also supported by the GSEA predictions and the identification of four pivotal stem marker genes (correlating with HOXB7 overexpression) in cell reprogramming and differentiation into pluripotent stem cells (iPS): SOX2, KLF4, NANOG and LIN28B [Takahashi and Yamanaka, 2006, Yu et al., 2007, Chien et al., 2015].

4.1 HOXB7 positively contributes to cell reprogramming by activating LIN28B.

New studies have recently underlined the parallelism between oncogenic transformation and cell reprogramming from terminally differentiated cells into iPS [Suvà et al., 2013]. Indeed, during oncogenic transformation, cancer cells may acquire SC-like properties, such as unlimited self-renewal and pluripotent potential. Several studies were thus focused on the understanding of the mechanisms shared between cellular reprogramming and oncogenic transformation. Historically, cell reprogramming was first achieved by somatic cell nuclear transfer (SCNT) into an enucleated oocyte (1952). The system was efficient but laborious. In 2006 Takahashi and Yamanaka developed a more feasible method to obtain pluripotent cells: starting from a combination of 24 genes, they identified four transcription factor (SOX2, KLF4, OCT4 and c-MYC) as the minimal "cocktail" able to convert mouse embryonic fibroblast (MEFs) into ES-

like cells, the so called induced pluripotent stem cells (iPS) [Takahashi and Yamanaka, 2006]. Importantly, those reprogramming factors are bona fide oncogenes.

In order to fully reprogram differentiated cells, only the core factors OCT4 and SOX2 are strictly required while the other TFs may enhance reprogramming efficiency and can be substitute by other genes, such as NANOG and LIN28 [Yu et al., 2007]. Interestingly, both genes are shown to be involved in carcinogenesis as well [Jeter et al., 2015, Viswanathan et al., 2009].

Further evidences underline the importance of LIN28/B during reprogramming. In a very recent work, LIN28B was silenced during reprogramming of a normal human oral keratocyte to reprogramm using the Yamanaka 4-gene protocol: silencing of LIN28B efficiently impaired the reprogramming [Chien et al., 2015]. Conversely, only in the presence of LIN28B the three genes OCT4-SOX2-NANOG were able to give rise to iPS cells.

We have shown that HOXB7 enhanced cell reprogramming efficiency through the three genes OCT4, KLF4, SOX2 in both mouse embryonic fibroblast and in human epithelial BEAS-2B cells. However, LIN28B silencing counteracted the expression of many stem/iPS marker genes in BEAS-2B-HOXB7 cells and strongly decreased the number of iPS colonies obtained upon HOXB7 overexpression. These observations suggest that the HOXB7 transcriptional activation of LIN28B is fundamental for the HOXB7 activation of a transcription program relevant for stem/iPS cell biology.

Further studies are required for the assessment of whether the mechanism we have described might have an impact on tumor initiation (e.g. upon limiting dilution and serial *in vivo* xenotransplantation assays). The identification of tumor initiating cells may ultimately lead to more effective prognostic tools and therapies for metastatic lung cancer.

Chapter 5

Appendix

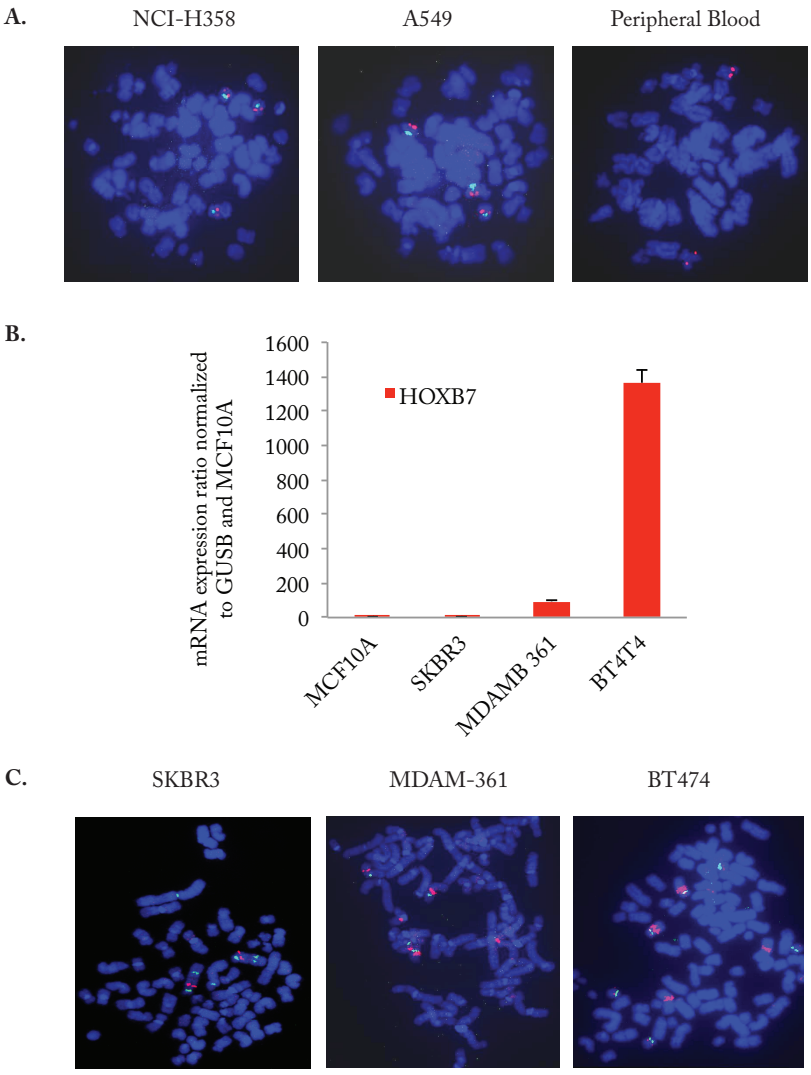


Figure 5.1: **FISH analysis of HOXB7 gene locus in lung and breast cancer cell lines. A.** FISH analysis of two lung cancer cell lines (NCI-H358 and A549) and peripheral blood cells as control: HOXB locus (17q21.32) in red; chromosome 17 centromere in green. Only one copy gain involving the entire chromosome was detected in both cell lines (also the 17 centromere specific probe gave one signal). **B.** RTqPCR analysis of HOXB7 expression in three breast cancer cell lines. **C.** HOXB7 expression and FISH analysis in three breast cell lines revealed a strong correlation of HOXB7 expression with the copy number.

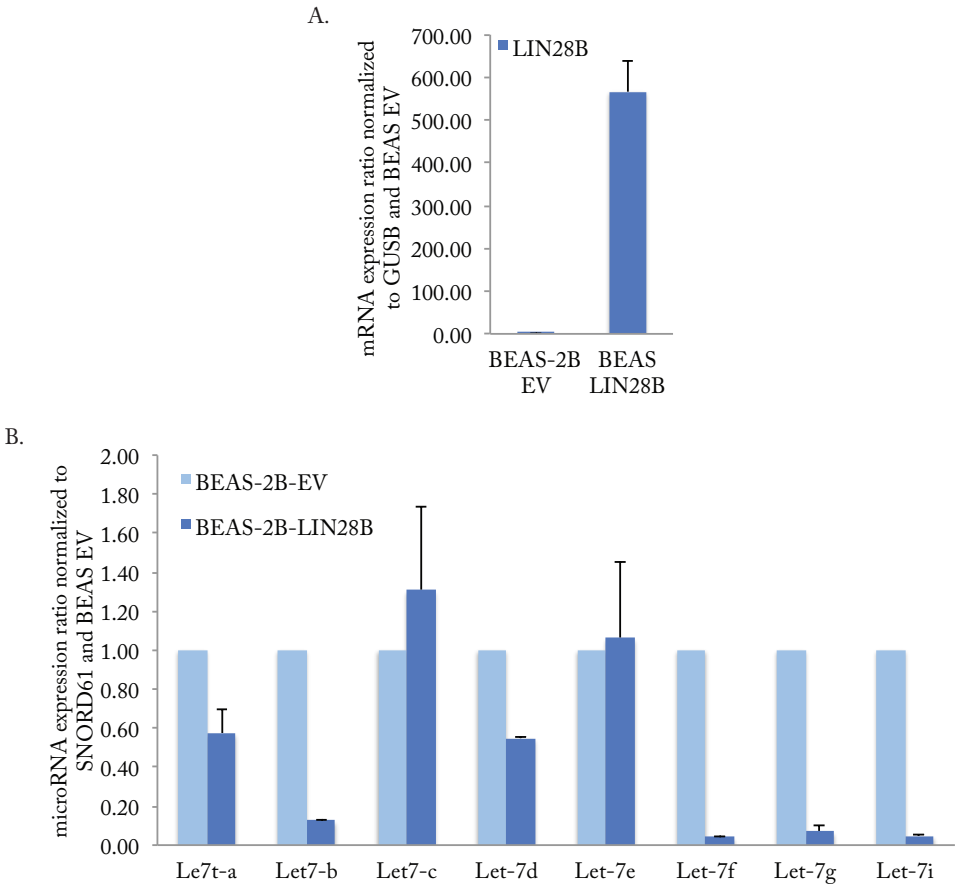


Figure 5.2: **Overexpression LIN28B in BEAS-2B cells resulted in the down-regulation of Let-7 microRNAs. A.** Overexpression of LIN28B by infection with the retroviral pBABE-hLIN28B construct (addgene) or control empty vector (EV) and was verified by RTqPCR analysis. **B.** Expression ratio of Let-7s microRNA family members by RT-qPCR analysis of cells as in A (relative to BEAS-2B-EV cells).

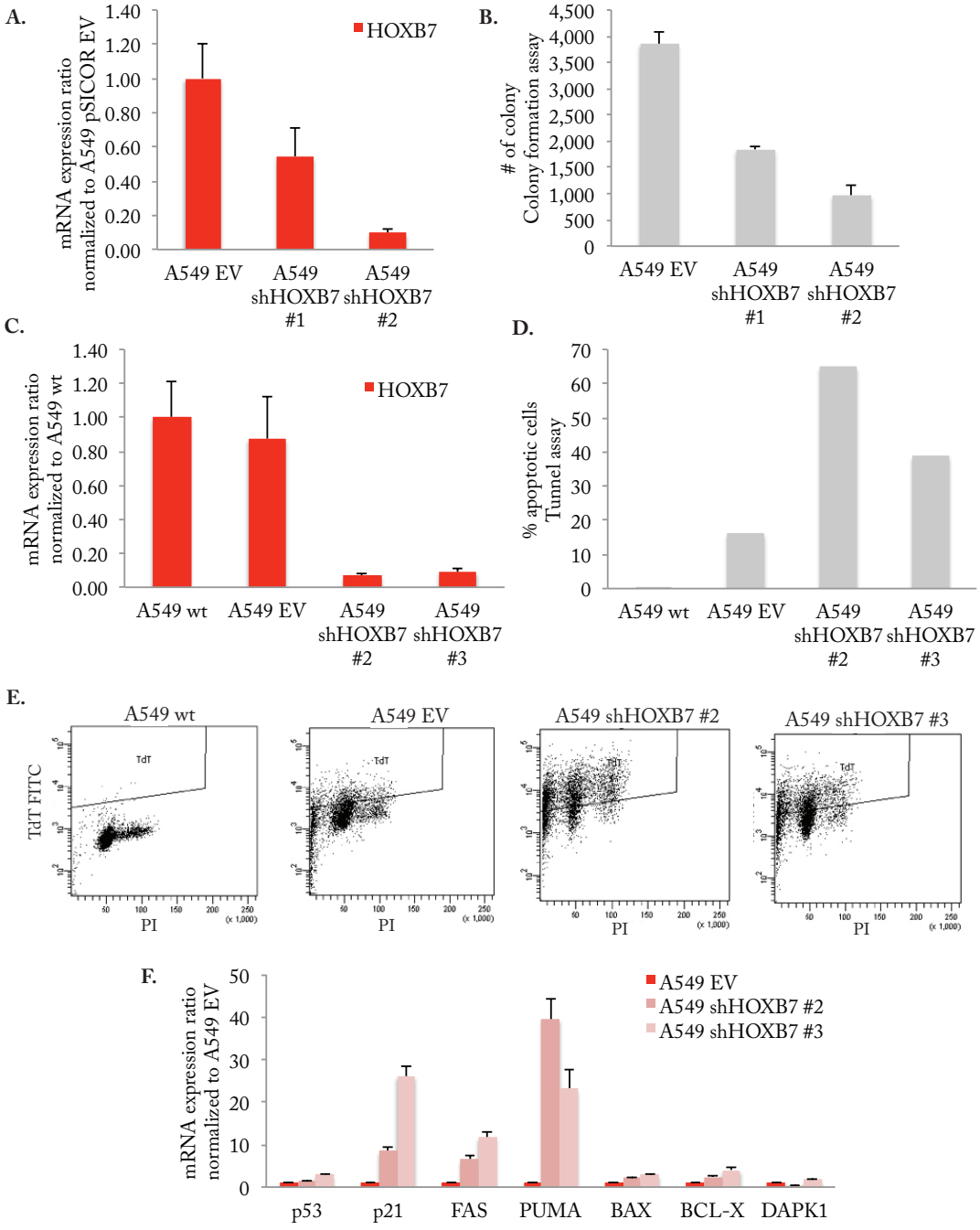


Figure 5.3: **Silencing of HOXB7 in A549 cells induces apoptosis.** HOXB7 was silenced in A549 lung cancer cells by infection with lentiviral vectors carrying shRNA against HOXB7. Control: pSiCOR empty vector (EV). **A.** RT-qPCR analysis of HOXB7 expression revealed a 50% and 90% gene silencing with oligo1 and oligo2, respectively. **B.** Quantification of colony formation assay highlighted an impairment in cell proliferation upon down-regulation of HOXB7 (50% and 70% reduction with oligo1 and oligo2, respectively) compared to wild type cells, consistent with results recently reported by Yuan and colleagues [Yuan et al., 2014]. **C-D.** Cell cycle profile and apoptotic events analysis by TUNEL assays, propidium iodide (PI) and FITC staining. DNA fragmentation, resulting from apoptotic signaling cascades, increased by 57% and 25% with oligo2 and oligo3, respectively, compared to control EV cells. **E.** Silencing efficiency of oligo3 was evaluated by RTqPCR analysis. **F.** RT-qPCR analysis of pro-apoptotic genes (p53, PUMA, BAX, FAS, BCL-X) and the cyclin dependent kinase inhibitor p21, involved in G1-S cell cycle transition. Increased expression of these genes correlated in inverse proportion with the expression levels of HOXB7.

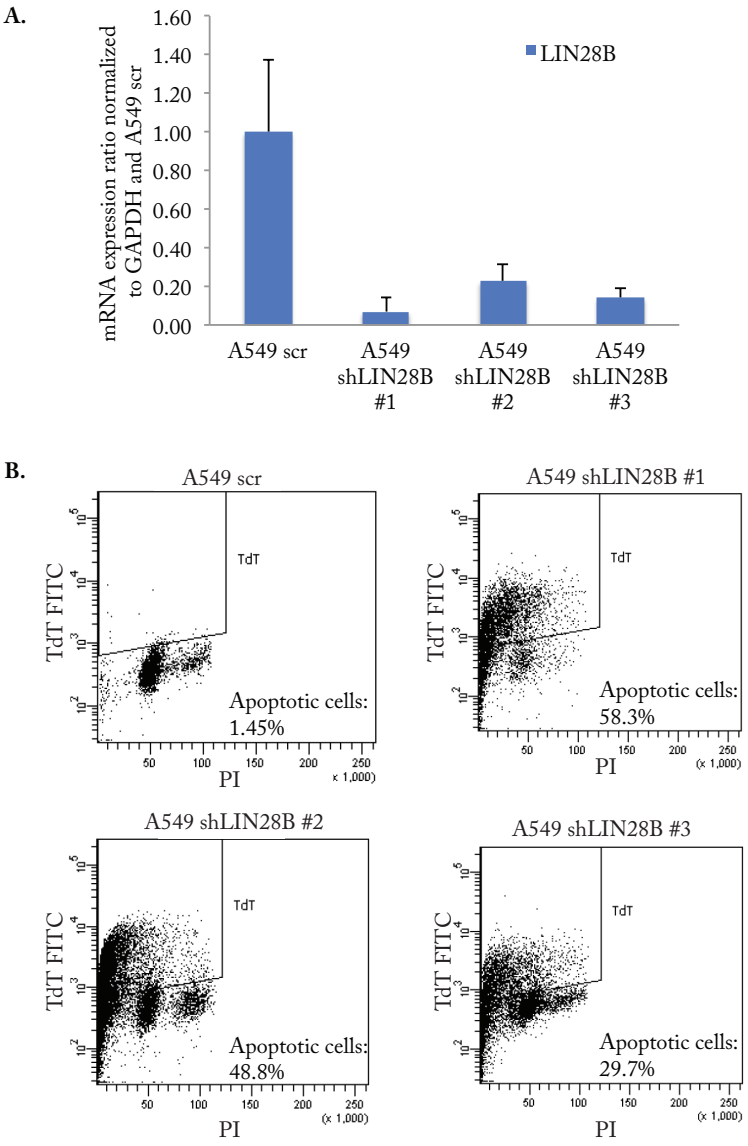


Figure 5.4: **Silencing of LIN28B in A549 cells line leads to apoptosis. A.** RT-qPCR analysis of LIN28B in A549 cells (with high HOXB7 endogenous levels) was infected with a pSICOR vector encoding three short hairpin RNAs targeting LIN28B or scrambled vector as control. Data are plotted as relative ratio to LIN28B expression in A549 scr cells. **B**. FACS analysis of A549 cells stained with PI and FITC (TUNEL assay) revealed an increased apoptotic rate upon silencing of LIN28B.

Gene	TCGA	Michigan	Tokyo	Class
ABCG2	0.01	-0.18	-0.36	STEM
ALCAM/CD166	-0.15	-0.23	-0.37	STEM
ALDH1A1	1.55	0.36	-0.24	STEM
BMI1	0.07	0.06	-0.03	STEM
CD24	0.43	0.34	0.31	STEM
CD44	-0.09	-0.08	0.09	STEM
EPCAM	0.13	-0.04	0.07	STEM
ITGA6/CD49f	0.68	0.30	0.05	STEM
LIN28A	-2.27	0.15	-0.14	STEM
LIN28B	1.10	NA	0.66	STEM
MYC	-0.19	-0.41	0.08	STEM
NANOG	0.16	-0.14	-0.55	STEM
PODXL	0.32	0.02	-0.05	STEM
POU5F1/OCT4	-1.11	0.33	0.15	STEM
SOX2	1.43	0.42	0.62	STEM
THY1/CD90	0.88	0.21	1.11	STEM
CDH1	-0.01	0.17	0.14	EMT
CDH2	1.68	0.58	0.41	EMT
SNAI1	0.53	0.23	0.37	EMT
SNAI2	1.16	0.40	0.91	EMT
TWIST1	1.64	0.47	1.91	EMT
VIM	0.26	-0.19	0.05	EMT
ZEB1	0.37	0.17	-0.03	EMT
FGF2	0.30	0.09	-0.18	HOXB7 TARGET
EGFR	0.16	0.07	0.28	HOXB7 TARGET

Table 5.1: **$\mathbf{Log_2}$ ratio of the mean expression of the markers in HOXB7-high expressing tumors vs. HOXB7-low expressing tumors.** The markers were collected by literature searching and analyzed in three independent cohorts of lung tumors (adenocarcinoma), profiled by RNA-seq (TCGA cohort) and Affymetrix microarray (Michigan and Tokyo cohorts).

Gene	p-value			-LOG10 p-value			Class
	TCGA	Michigan	Tokyo	TCGA	Michigan	Tokyo	
ABCG2	0.9700	0.0010	0.1200	0.013	3.000	0.921	STEM
ALCAM/CD166	0.1000	0.3600	0.0050	1.000	0.444	2.301	STEM
ALDH1A1	0.0003	0.1500	0.2800	3.523	0.824	0.553	STEM
BMI-1	0.3700	0.2700	0.2500	0.432	0.569	0.602	STEM
CD24	0.0002	0.0270	0.0200	3.699	1.569	1.699	STEM
CD44	0.3500	0.4100	0.1500	0.456	0.387	0.824	STEM
EPCAM	0.1500	0.7500	0.1700	0.824	0.125	0.770	STEM
ITGA6/CD49	0.0003	0.0900	0.4400	3.523	1.046	0.357	STEM
LIN28	0.1800	0.4800	0.1300	0.745	0.319	0.886	STEM
LIN28B	<0.0001		0.6800	4.000		0.167	STEM
MYC	0.0400	0.0020	0.9600	1.398	2.699	0.018	STEM
NANOG	0.8500	0.3000	0.0002	0.071	0.523	3.699	STEM
POU5F1/OCT4	0.2100	0.0370	0.7700	0.678	1.432	0.114	STEM
PODXL	0.2500	0.9100	0.9000	0.602	0.041	0.046	STEM
SOX2	<0.0001	0.0020	0.0090	4.000	2.699	2.046	STEM
THY1/CD90	<0.0001	0.0400	<0.0001	4.000	1.398	4.000	STEM
CDH1	0.57	0.11	0.03	0.244	0.959	1.523	EMT
CDH2	<0.0001	0.0003	0.04	4.000	3.523	1.398	EMT
SNAI1	0.0004	0.17	0.015	3.398	0.770	1.824	EMT
SNAI2	0.0007	0.0044	<0.0001	3.155	2.357	4.000	EMT
TWIST1	<0.0001	0.0008	<0.0001	4.000	3.097	4.000	EMT
VIM	0.82	0.052	0.31	0.086	1.284	0.509	EMT
ZEB1	0.23	0.13	0.84	0.638	0.886	0.076	EMT
FGF2	0.0072	0.37	0.13	2.143	0.432	0.886	HOXB7 TARGET
EGFR	0.0112	0.31	0.08	1.951	0.509	1.097	HOXB7 TARGET

Table 5.2: **P-values of differential expression of the markers in HOXB7-high expressing tumors vs. HOXB7-low expressing tumors.**. -Log(10)p-value values were used to show significance of the differential expression ($p<0.05$ corresponds to -LogP>1.3). P-values were calculated by the non parametric Wilcoxon test.

MOLECULAR ONCOLOGY 9 (2015) 473–487

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/molonc

Technical Note

Mining cancer gene expression databases for latent information on intronic microRNAs

Simona Monterisi[a], Giovanni D'Ario[a,1], Elisa Dama[a,b], Nicole Rotmensz[b], Stefano Confalonieri[a,c], Chiara Tordonato[a], Flavia Troglio[a], Giovanni Bertalot[a], Patrick Maisonneuve[b], Giuseppe Viale[d,e], Francesco Nicassio[a,c,f], Manuela Vecchi[a,c], Pier Paolo Di Fiore[a,c,g,**,2], Fabrizio Bianchi[a,*,2]

[a]Molecular Medicine Program, Department of Experimental Oncology, European Institute of Oncology, Milan, Italy
[b]Division of Epidemiology and Biostatistics, European Institute of Oncology, Milan, Italy
[c]IFOM, The FIRC Institute for Molecular Oncology Foundation, Milan, Italy
[d]Division of Pathology, European Institute of Oncology, Milan, Italy
[e]School of Medicine, University of Milan, Milan, Italy
[f]Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Milan, Italy
[g]Department of Scienze della Salute, University of Milan, Milan, Italy

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form
23 September 2014
Accepted 2 October 2014
Available online 15 October 2014

Keywords:
MicroRNA
Cancer
Gene expression
Breast cancer

ABSTRACT

Around 50% of all human microRNAs reside within introns of coding genes and are usually co-transcribed. Gene expression datasets, therefore, should contain a wealth of miRNA-relevant latent information, exploitable for many basic and translational research aims. The present study was undertaken to investigate this possibility. We developed an *in silico* approach to identify intronic-miRNAs relevant to breast cancer, using public gene expression datasets. This led to the identification of a miRNA signature for aggressive breast cancer, and to the characterization of novel roles of selected miRNAs in cancer-related biological phenotypes. Unexpectedly, in a number of cases, expression regulation of the intronic-miRNA was more relevant than the expression of their host gene. These results provide a proof of principle for the validity of our intronic miRNA mining strategy, which we envision can be applied not only to cancer research, but also to other biological and biomedical fields.

* *Corresponding author*. European Institute of Oncology, Department of Experimental Oncology, Via Ripamonti, 435, 20141 Milan, Italy. Tel.: +39 02 94375173; fax: +39 02 94379305.
** *Corresponding author*. European Institute of Oncology, Department of Experimental Oncology, Via Ripamonti, 435, 20141 Milan, Italy. Tel.: +39 02 57489832; fax: +39 02 94375991.
E-mail addresses: pierpaolo.difiore@ieo.eu (P.P. Di Fiore), fabrizio.bianchi@ieo.eu (F. Bianchi).
[1] Present address: Swiss Institute of Bioinformatics, Lausanne, Switzerland.
[2] These two authors are equal last authors.
http://dx.doi.org/10.1016/j.molonc.2014.10.001

Abbreviations

FFPE	formalin-fixed paraffin-embedded
PCR	polymerase chain reaction
FDR	false discovery rate
qRT-PCR	quantitative reverse transcriptase PCR
GEO	gene expression omnibus
ER	estrogen receptor
HER2 (ErbB2)	human epidermal growth factor receptor 2
IHC	immunohistochemistry

1. Introduction

MicroRNAs are small, non-coding RNA molecules (18–22 nucleotides in length) that function as endogenous triggers of the mRNA interference pathway and are involved in the regulation of pleiotropic biological functions (Krol et al., 2010; Yendamuri and Kratzke, 2011). Almost 50% of all human miRNA genes are located within introns of host genes, with which they usually share transcriptional regulation (Baskerville and Bartel, 2005; Griffiths-Jones, 2007; He et al., 2012; Monteys et al., 2010; Ozsolak et al., 2008; Rainer et al., 2009; Rodriguez et al., 2004; Saini et al., 2007). In principle, this property could be exploited to predict the expression of intronic miRNAs (int-miRNAs) through the analysis of the expression of their host genes (miRNA host genes, miR-HG). Similar approaches have already been successfully employed to identify miRNA target genes, to predict miRNA tissue expression, and to characterize miRNA/miR-HG feedback loops (Lutter et al., 2010; Radfar et al., 2011; Wang et al., 2009).

The major potential stemming from the mode of regulation of int-miRNAs is, however, untapped. In recent years, enormous effort has been dedicated to the profiling of various physiological and pathological conditions at the transcriptomic (mRNA) level. While almost every field of biology and biomedicine has been explored through this approach, cancer biology is arguably the field in which the highest investment has been made, with the dual objective of: i) obtaining a global view of cancer processes by systems-based analysis (Basso et al., 2005; Minn et al., 2005; Sweet-Cordero et al., 2005); ii) identifying biomarkers for improved management of cancer patients (Ivshina et al., 2006; Sorlie et al., 2001; Sotiriou et al., 2003; van 't Veer et al., 2002). As a result, thousands of human tumors have been profiled and the datasets made publicly available, frequently associated with high quality clinical information. In our view, these datasets are amenable to mining "latent" information on int-miRNA expression.

There is growing interest in miRNAs, both as potential cancer determinants and biomarkers (Calin and Croce, 2006). From a general perspective, miRNA profiling might be advantageous over mRNA profiling, since the complexity of miRNome is at least 20-fold lower than that of a reference transcriptome (if one makes the somewhat rough comparison of ~1000 miRNAs *vs.* ~ 20,000 genes). This means that sufficient statistical power can be reached with a much lower number of analyzed samples. This is particular relevant to studies, such as those involving human pathological samples, in which genetic variability represents a relevant confounding factor.

Thus, the explicit goal of this study was to exploit cancer datasets, in particular, breast cancer datasets, to provide a proof of principle that meta-analysis of miR-HG expression profiles can accurately identify int-miRNAs that are relevant to cancer, both in terms of their potential utility as biomarkers and their role in breast cancer cell biology.

2. Material and methods

2.1. Patient selection criteria

Written informed consent for research use of biological samples was obtained from all patients. Patients underwent surgery at the European Institute of Oncology between 1998 and 2010. Only tumor samples with a cellularity >70% were included in the study.

2.2. Affymetrix microarray analysis

Retroviral infection of the MCF10A cell line with the SV40-large T antigen was performed using a pBABE-neo retroviral vector. After 48 h of infection, cells were collected and total RNA extracted using the RNeasy Mini Kit (QIAGEN). RNA quality was controlled using the 2100 Bioanalyzer (Agilent). Total RNA (5 μg) was then retrotranscribed into double stranded cDNA using SuperScript® Double-Stranded cDNA Synthesis Kit (Invitrogen).

In vitro anti-sense RNA transcription was performed through an Eberwine's modified *in vitro* transcription reaction (MEGAscript, Ambion) using labeled rNTP (Enzo® BioArray™ HighYield™ RNA Transcript Labeling Kit, ENZO Biolabs). Briefly, we added 14.5 μl of rNTPs mix, 2 μl of T7 polymerase and 2 μl of reaction buffer to 1.5 μl of purified cDNA, and incubated the reaction mix at 37 °C for 6 h. Labeled cRNA was then fragmented (30-200 base fragments), checked by agarose gel, and hybridized on Human Genome U133A 2.0 Arrays in duplicate for each condition (i.e., MCF10A SV40-large T, and MCF10A pBABE-empty).

Data were normalized using the Robust Multi-array Average (RMA) method. Information on human int-miRNAs, associated host genes and mature miRNA sequences was retrieved from miRBase v13.0. The Entrez IDs of the miRNA host genes was extracted from the "*org.Hs.eg.db*" Bioconductor annotation package. Probe sets mapping to miRNA host genes were identified using the Bioconductor *hgu133a2.db* annotation package. Differentially expressed probe sets were identified using the limma Bioconductor package. *P*-values were adjusted using the Benjamini–Hochberg correction.

2.3. Bioinformatics analysis of external Affymetrix datasets

Breast cancer microarray datasets and associated clinical information were downloaded from the Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/geo/). The accession numbers of the datasets used are GSE1456, GSE2990, GSE7390, and GSE4922. All datasets were based on

the GeneChip® Human Genome U133A to avoid batch bias effects during the analysis.

We applied a quality control procedure on CEL files to identify flawed arrays using Relative Log Expression (RLE) values and Normalized Unscaled Standard Error (NUSE) values (Bolstad et al., 2004; Gentleman et al., 2004). For each array, we computed the median value and the interquartile range (IQR) of both the NUSE and RLE statistics. We then calculated the IQRs across the arrays for each dataset. Arrays were rejected if IQR values were $>q3 + 1.5IQRs$ or $< q1 - 1.5IQRs$, where q1 and q3 are the first and third quartile, respectively.

This filtering resulted in the exclusion of 47 arrays in the GSE4922 dataset, 9 arrays in the GSE1456 dataset, 9 arrays in the GSE7390 dataset, and 5 arrays in the GSE2990 dataset. In addition, in the GSE2990 dataset we considered only the "Uppsala" cohort of patients because of the low signal intensity distribution of several arrays of the "Oxford" cohort compared to the "Uppsala" cohort, which determined a batch bias effect (Figure S1). Data were normalized using the RMA method.

Information relative to human int-miRNAs and associated host genes was retrieved from miRBase (www.mirbase.org, release 13.0). Probe sets were filtered for signal intensity using the Bioconductor genefilter package. Only probe sets that had a normalized signal greater than 150 (7.2 on the log2 scale) in at least 10% of the samples were retained for further analysis. Differentially expressed probe sets were identified using the limma Bioconductor package. All *P*-values were adjusted using the Benjamini–Hochberg correction.

Monte Carlo simulation was performed for each dataset in the following manner: 1) all miRNA-associated probe sets were excluded from the dataset; 2) we randomly selected *n* probe sets, where *n* is the number of miRNA-associated probe sets ($n = 422$); 3) the number of probe sets significantly regulated between G3 *vs*. G1 and/or ER + *vs*. ER-tumors were annotated; 4) we repeated steps '2' and '3' 999 times. An empirical *P*-value was calculated as the fraction of simulations yielding a larger list of significantly regulated probe sets than the list obtained in the original analysis. Expression dataset Breast subtype analysis was performed using the TCGA breast cancer (October 2012 release, 599 patients) downloaded from the TCGA web data portal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). Data were gene centered on relative medians and log2 transformed before clustering analysis.

Pathway analysis was performed using the online available webtool Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com/). Predicted and experimentally validated miRNA target gene sets were obtained using miRWalk database (Dweep et al., 2011). MicroRNA target prediction was performed using four different methods: miRanda, miRDB, miRWalk and Targetscan. Genes predicted in 4 out of 4 methods were retained for subsequent IPA analysis.

2.4. RNA isolation and RT-PCR

Total RNA was extracted from cell lines using the TRIZOL reagent (Invitrogen) or from FFPE archival breast tumor samples (with a tumor cellularity >70%) using the RNAeasy FFPE kit (QIAGEN). RNA was quantified by Nanodrop (Agilent Technologies).

miRNA expression profiles of MCF10A cells were obtained using the TaqMan® Low Density Array microRNA Signature Panel (v1.0; Applied Biosystems) and reactions were carried out on an Applied Biosystems 7900HT thermocycler using the manufacturer's recommended cycling conditions.

miRNA expression profiles of FFPE archival breast tumor samples or of MDA-MB-231 and MDA-MB-361 cells were obtained using miScript Primer Assays and the miScript SYBR Green PCR Kit (Qiagen). Total RNA (400 ng) was reverse transcribed using the miScript Reverse Transcriptase kit (Qiagen) according to manufacturer's instructions. Briefly, the two-step protocol involves reverse transcription of miRNA to cDNA using miRNA-specific primers followed by qRT-PCR. Reactions were run in duplicate using 5 ng of cDNA as template in 20 µl final reaction volume. All probes were normalized to U5A for FFPE archival breast tumor samples or to *SNORD25* for breast cancer cell lines, as an internal control. Amplification reactions were performed with LightCycler 480 (ROCHE) using the manufacturer's recommended cycling conditions. Relative expression ratios of miRNAs were obtained using the 2^{-ddCT} method.

2.5. Cell lines and infection

The MDA-MB-231 and MDA-MB-361 breast cancer cell lines were grown in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100 U/ml penicillin, and 100 U/ml streptomycin at 37 °C in a humidified incubator with 5% CO_2. Mature miRNAs were overexpressed using GFP Lenti-miR™ vectors obtained from Systems Biosciences (SBI, Mountain View, CA, USA). The pCDH-CMV-MCS-EF1-copGFP was used as a control vector (SBI, Mountain View, CA, USA). GFP positive cells were sorted with a BD Influx™ cell sorter (BD Bioscience, San Jose, CA, USA). Host genes overexpression was achieved by Lipofectamine® 2000 transfection (LifeTechnologies) of pCMV-Sport6-*MYO5C*, pCMV-Sport6-*EVL* and pDEST26-*IGF2* vectors (derived from the original pENTR221-*IGF2* vector) provided by Life Science (Source BioScience, Nottingham, UK). Gene expression was verified by qRT-PCR using QuantiTect Primer Assay (Qiagen), or, for *MYO5C*, using custom primers: *MYO5C* – Forward, GAATCTCTGCCTCCACTTCG; *MYO5C* – Reverse, GATAGCTGAGAGCCGTGAGG. miR-HG expression was normalized to GUSB expression as an internal control.

2.6. Cell proliferation and colony forming assays

For proliferation assays, MDA-MB-231 and MDA-MB-361 cells were seeded in triplicate in 6-well plates (BD Falcon™ 353046) at 4×10^4 and 25×10^4 cells/well, respectively. Bio-Rad TC10 automated cell counter was used to count cells every 24 h for 4 days. The first measurement was taken at 24 h for MDA-MB-231 or at 72 h for MDA-MB-361 after seeding cells. Colony forming assays were performed by seeding 5000 cells/type in 10-cm plates (BD Falcon™ 353003) and then incubating plates for 10 days. Colony formation was visualized by staining for 5 min at RT with crystal violet (1% w/v in 35% EtOH, Santa Cruz).

2.7. Wound-healing assay

Time-lapse video microscopy was performed as described previously (Palamidessi et al., 2008; White et al., 2007) with slight modifications. Confluent monolayers of MDA-MB-231 or MDA-MB-361 cells in 12-well plates (BD Falcon™ 353043) were wounded with a plastic pipette tip to induce migration into the wound. Cells were placed on the stage of an inverted motorized microscope (Leica AF600) in a cage incubator (Okolab) at 37 °C and 0% CO_2 for time-lapse video microscopy. Phase-contrast images were collected every 20 min over a 12-h period. Videos were generated using the ImageJ software for image analysis. Cell trajectories were determined using the MTrackJ plugin of ImageJ (Meijering et al., 2012). The distance covered by each cell and the migration speed were extracted from the track plots. Fifteen cells from 3 independent experiments were analyzed for each condition, and data are expressed in micrometers as mean ± σ.

2.8. Statistical analysis

The significance of the overlap between gene lists was based on the hypergeometric distribution (Fury et al., 2006). The extension to the case of more than two overlapping lists was based on marginalization of joint probability and chain rule of probability.

RT-PCR gene expression analysis and relative statistical analyses were performed using JMP 10.0 64-bit edition (SAS Institute Inc.). Statistical analyses were performed on log2 data using parametric tests (t-test, ANOVA). Cluster 3.0 for Mac OS X (http://bonsai.hgc.jp/~mdehoon/software/cluster/) and Java Treeview (http://jtreeview.sourceforge.net) were used for the hierarchical clustering analysis. Uncentered correlation and centroid clustering methods were used on log2 median centered data. The multivariate model to predict risk of having an aggressive tumor subtype was built using diagonal linear discriminant analysis (DLDA) with BRB ArrayTools (http://linus.nci.nih.gov/BRB-ArrayTools.html). Briefly, the model assigned a risk index to every patient and classified them as high- or low-risk of having an aggressive tumor subtype based on linear combination of gene expression values weighted by coefficients calculated during training of the classifier. The critical cutoff value to predict high-/low-risk was identified by the receiver operating characteristics curve analysis (ROC) using JMP 10.0 software, and it was set at −1.23. In the training set, twenty-eight out of the 29 LuA tumors (~97%) were predicted as 'low risk', which is consistent with the reported low metastatic behavior of LuA breast tumors (Voduc et al., 2010). In contrast, 43 out of the 66 patients with the more aggressive LuB tumor subtype (~65%) were predicted as 'high risk' ($P < 0.0001$). Overall, the test displayed an accuracy, sensitivity and specificity of ~75%, ~65% and ~97%, respectively, in the training set, and it was independent from other risk factors such as nodal status, tumor size (pT), and HER2 positivity (Table S1).

Multivariate nominal logistic regression of miRNA risk class was performed using JMP 10.0 software. The Odds ratio of miRNA high-risk class was adjusted for ErbB2 (HER2), nodal status and tumor size (pT).

Table 1 – Overlapping miR-HGs in the different analyses. Overlapping miR-HGs (highlighted in grey scale) in G3 *vs.* G1 and ER+ *vs.* ER-analyses ($P < 0.05$, Benjamini–Hochberg correction). Dataset, Gene Expression Omnibus accession numbers of the Affymetrix datasets. N, total number of unique miR-HGs found regulated in the relative dataset. %, percentage of miR-HGs regulated out of the total miR-HGs mapped on the array ($N = 243$). N overlapping, number of miR-HGs commonly found regulated in highlighted datasets (light/dark grey areas). *P*, significance of the overlaps.

G3/G1

Dataset	N	% (N/243)	N overlapping		
GSE4922	110	45	53 P<0.0001	47 P<0.0001	25 P<0.0001
GSE1456	56	23			
GSE2990	78	32			
GSE7390	35	14			

ER+/ER-

Dataset	N	% (N/243)	N overlapping	
GSE4922	97	40	27 P<0.0001	17 P<0.0001
GSE2990	29	12		
GSE7390	43	18		

3. Results

3.1. An in silico approach to extract information on the regulation of int-miRNAs from microarray gene expression (mRNA) datasets

Several studies have indicated that there is a good correlation between the expression of miR-HGs and their respective int-miRNAs (Baskerville and Bartel, 2005; Wang et al., 2009; Wang and Li, 2009). Therefore, the first step in our strategy was the development of an *in silico* approach to predict int-miRNA expression by means of microarray-based analysis of miR-HG expression profiles. To this end, we used a controllable and syngeneic model system, the non-transformed breast cell line MCF10A infected with the Simian virus 40 (SV40) large T antigen, which causes cell transformation and alters the expression of numerous genes (Carbone et al., 1997; De Luca et al., 1997; Girardi et al., 1962). We compared the mRNA and miRNA expression profiles of these MCF10A-SV40 cells to those of mock-infected MCF10A cells (see Methods). miR-HGs were identified by mapping the genomic position of 706 known human miRNA precursors (pri-miRNAs) to the genomic coordinates of the entire human genome. This resulted in the identification of 317 pri-miRNAs located within the introns of 269 unique miR-HGs (Table S2). Affymetrix gene expression analysis revealed 43 miR-HGs differentially expressed in MCF10A-SV40 *vs.* control cells (FDR<10%, Table S3), which contain 51 pri-miRNAs in their introns, corresponding to 84 mature int-miRNAs. We validated the analysis, by directly measuring the expression of 47 of these miRNAs, in the same samples, by qRT-PCR (using available TaqMan assays). Of the 47 int-miRNAs, 38 were detectable by qRT-PCR, and of these 31 were congruently regulated with their miR-

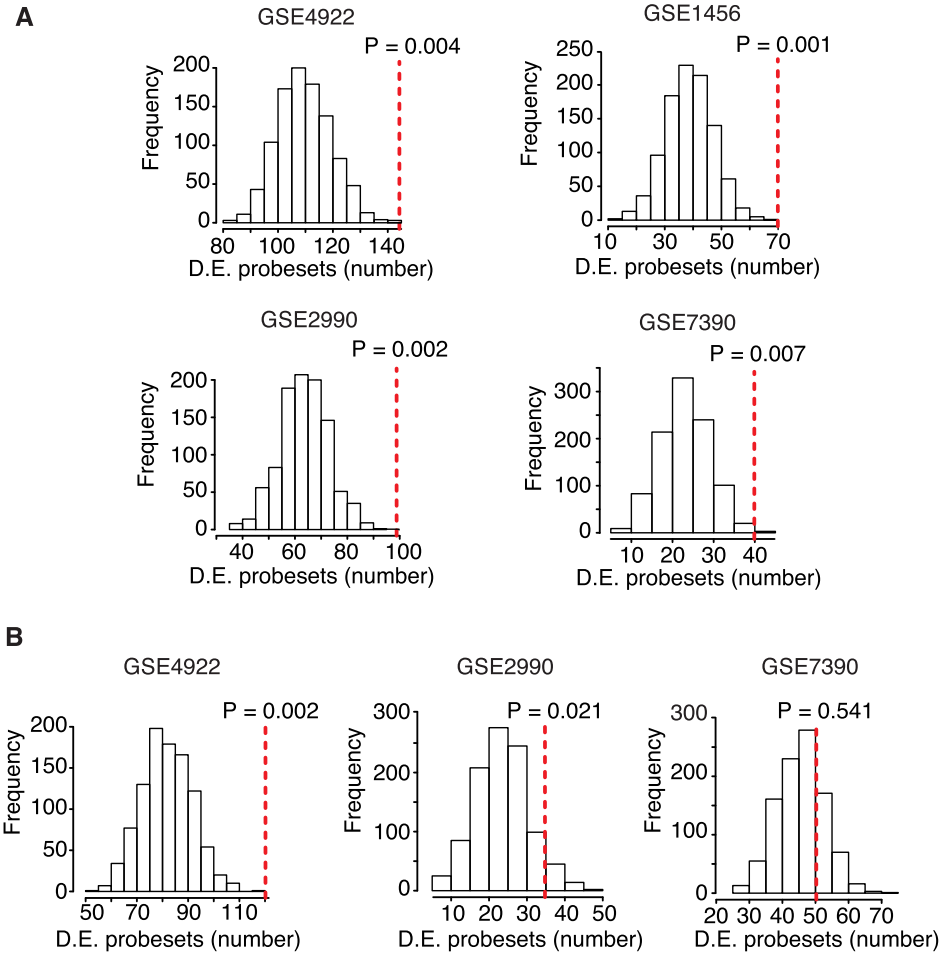


Figure 1 – Global gene expression profiles of miR-HGs in breast cancer. A. Results of the Monte Carlo simulation for the G3 *vs.* G1 breast cancer Affymetrix meta-analysis. For this simulation, we used 1000 lists of randomly selected genes that were of the same size as the original list of miR-HGs, and which contained genes that, to date, have not been associated with any int-miRNA ("non-host genes"). Histograms represent the distribution of the random lists (1000 random lists of 422 probe sets) in the indicated GEO datasets. *X*-axes: number of significantly regulated probe sets (D.E., differentially expressed probe sets) found in the random lists ($P < 0.05$; Benjamini–Hochberg correction). *Y*-axes: frequency of random lists. Red dashed lines indicate the position, within the distributions, of the miR-HG list (422 probe sets) identified through our analysis. *P*-values indicate the fraction of random lists having an equal or larger number of significantly regulated probe sets compared to the miR-HG list. B. Results of the Monte Carlo simulation for the ER+ *vs.* ER- breast cancer Affymetrix analysis. Results are presented as in (A).

HGs (Table S4). The positive correlation between the expression of miR-HGs and int-miRNAs was significant when analyzed both qualitatively (congruent direction of regulation; $P = 0.02$, Figure S2A) and quantitatively (expression ratios; $P = 0.0003$, $R = 0.4$, Figure S2B). In contrast, the 74 int-miRNAs (for which TaqMan assays were available) within introns of unregulated miR-HGs (FDR>10%), did not show significant co-regulation with their cognate miR-HGs ($P = 0.2$ and $P = 0.5$; Figure S2C and D), possibly due to a different post-transcriptional regulation of mRNA and miRNA cognate transcripts.

In summary, these data indicate that our bioinformatics approach can be used to infer int-miRNA regulation through host gene expression profile analysis.

3.2. In silico *prediction of miR-HGs differentially regulated in breast cancer*

We next analyzed the expression of miR-HGs in breast cancer through a meta-analysis of expression datasets of 666 patients, from four independent studies with clinical and pathological information, and raw data available through the GEO database (Table S5). We mapped 243 unique miR-HGs (containing 264 int-miRNAs) whose expression data were present in the datasets (Table S6). Through the comparison of tumors with different clinical and pathological parameters, we identified a significant number of differentially expressed miR-HGs ($P < 0.05$, Benjamini–Hochberg correction; Table S7A and B), especially in the comparisons between poorly differentiated (G3) and well-differentiated (G1) tumors, or between estrogen receptor positive (ER+) and estrogen receptor negative (ER−) tumors. In the first instance (G3 *vs.* G1), 14–45% of all miR-HGs (depending on the considered dataset, Table 1) were differentially regulated; in the second case (ER + *vs.* ER−) the differential expression of miR-HGs was 12–40% (Table 1).

It has been reported that the differences in the transcriptomic profiles of different types of breast cancer (i.e. G3 *vs.* G1, or ER−*vs.* ER+) are so vast that regulation of a set of genes of interest might simply reflect large-scale transcriptional changes (Ivshina et al., 2006; Sotiriou et al., 2003, 2006), and even that a significant number of randomly chosen "signatures" may have prognostic value (Venet et al., 2011). Thus, to determine whether the large fraction of miR-HGs differentially regulated in breast cancer was not the mere result of large-scale transcriptional changes, we performed a Monte Carlo simulation using 1000 random lists of non-host genes (which represent more than 95% of the entire genome). In G3 *vs.* G1 tumors, there was a significant enrichment of differentially regulated miR-HGs, with respect to non-host genes ($P < 0.01$ in all datasets; Figure 1A). A similar enrichment was also observed, although to a lesser extent, when ER+ tumors were compared with ER− (Figure 1B).

The above result argues that miR-HGs are preferentially and selectively regulated among different types of breast cancers. While this concept will be further discussed later, it is of note that there is a high degree of overlap between the G3/G1 and the ER+/ER− lists of miR-HGs regulated with same trend, in independent datasets (Table 1 and Tables S7C and D).

Table 2 — Significantly regulated miR-HGs in G3 versus G1 breast tumors and in the different tumor subtypes.

Probeset	miRBase	miRNA	Gene symbol	GEO-1 FC	GEO-2 FC	GEO-3 FC	GEO-4 FC	GEO-1 P	GEO-2 P	GEO-3 P	GEO-4 P	TCGA subtype P (ANOVA)	
201664_at	MI0000115/MI0000438	16-2/15b	*SMC4*	1.94	1.57	1.91	1.74	<0.001	<0.001	<0.001	0.001	–	
201663_s_at	MI0000115/MI0000438	16-2/15b	*SMC4*	1.71	1.60	1.56	1.88	<0.001	<0.001	0.002	<0.001	<0.001	
210983_s_at	MI0000095	106b/25/93	*MCM7*	1.57	1.68	1.52	1.57	<0.001	<0.001	0.001	<0.001	<0.001	
201852_x_at	MI0006380	1245	*COL3A1*	−1.40	−1.53	−1.70	−1.56	0.005	0.012	0.001	0.016	<0.001	
209897_s_at	MI0000294	218-1	*SLIT2*	−1.49	−1.51	−1.57	−1.57	<0.001	0.001	<0.001	0.003	<0.001	
218966_at	MI0006403	1266	*MYO5C*	−1.61	−1.37	−1.63	−1.54	<0.001	0.007	<0.001	0.004	<0.001	
217838_s_at	MI0000805	342	*EVL*	−2.45	−1.88	−2.25	−1.95	<0.001	<0.001	<0.001	<0.001	<0.001	
202409_at	MI0002467	483	*IGF2*	−2.50	−1.72	−2.65	−2.31	<0.001	0.012	<0.001	<0.001	<0.001	
214053_at	MI0006375	548f-2	*ERBB4*	−2.79	−2.43	−2.66	−2.38	<0.001	<0.001	<0.001	<0.001	<0.001	

The miR-HGs significantly regulated (fold-change > 1.5 or < −1.5, in at least 3 out of the 4 datasets) between G3 and G1 breast tumors in the four Affymetrix GEO datasets are shown. The significance by ANOVA analysis of gene expression regulation among different tumor subtypes is also reported in a fifth independent breast tumor dataset (TCGA dataset (2012)), in which molecular subtyping using the PAM50-model was available (Agilent array). miRNA genes are indicated together with their relative miRBase accession number. Fold change (FC) and P-values (P) are reported. GEO-1, GSE4922 dataset; GEO-2, GSE1456 dataset; GEO-3, GSE2034 dataset; GEO-4, GSE7390 dataset.

3.3. *From miR-HGs to int-miRNAs: in silico predictions of relevance to cancer*

We next attempted to translate the information on miR-HGs into the corresponding information concerning their hosted int-miRNAs. Since this step was preparatory to the actual validation and analysis of biological relevance in "real" tumors, we concentrated on the miR-HGs differentially expressed between G3 and G1 tumors, which displayed the most consistent and significant regulation among datasets (Figure 1). To eliminate candidates whose fluctuations might be due to technical or biological variability, we applied a stringent threshold and selected only those miR-HGs that displayed a fold-change of at least 1.5 (positive or negative) in at least 3 out of the 4 datasets. This yielded a list of 8 candidate miR-HGs, of which 2 were upregulated and 6 downregulated (Table 2).

The two upregulated miR-HGs were *SMC4* and *MCM7* that are involved in DNA synthesis, mitosis and DNA repair (Hagstrom and Meyer, 2003; Lei and Tye, 2001), and contain two miRNA clusters in their introns: the *miR-15b~16-2* and the *miR-25~106b* cluster, respectively. Both miRNA families have been described as being relevant to cancer (Bonci et al., 2008; Poliseno et al., 2010). Conversely, the six downregulated miR-HGs contain nine miRNAs whose relevance to cancer has not been investigated in detail: *miR-548f-2*, *miR-1245*, *miR-218/218**, *miR-342-3p/-5p*, *miR-483-3p/-5p* and *miR-1266* (Grady et al., 2008; Song et al., 2012; Soon et al., 2009; Tie et al., 2010; Veronese et al., 2010).

To gain initial insights into the potential relevance of these int-miRNAs to breast cancer, we performed pathway analyses of their predicted and validated target genes (see Methods). We observed a statistically significant enrichment in cancer-relevant genes among the predicted targets, which was confirmed also among the experimentally validated target genes (Figure 2A and B).

3.4. *Validation of predicted breast cancer-regulated int-miRNAs by qRT-PCR*

To validate the results of the *in silico* analysis, we analyzed the eleven identified int-miRNAs by qRT-PCR in a cohort of 36 FFPE archival G1 and G3 breast cancers (Table S8A). *miR548f-2* and *miR-1245* were undetectable in all samples and were thus excluded from further analyses. The hierarchical clustering analysis of the tumor samples, based on the expression of the remaining nine int-miRNAs, displayed a clear separation at the first tree branching between G3 and G1 tumors (70% and 75% of G3 and G1 tumors, respectively, clustering at the first branch; $P = 0.006$, likelihood-ratio test; Figure 3A). This result confirmed a distinct pattern of expression of the int-miRNA signature in high-*vs.* low-grade breast tumors, as predicted by the miR-HG expression analysis. Individually, *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-*

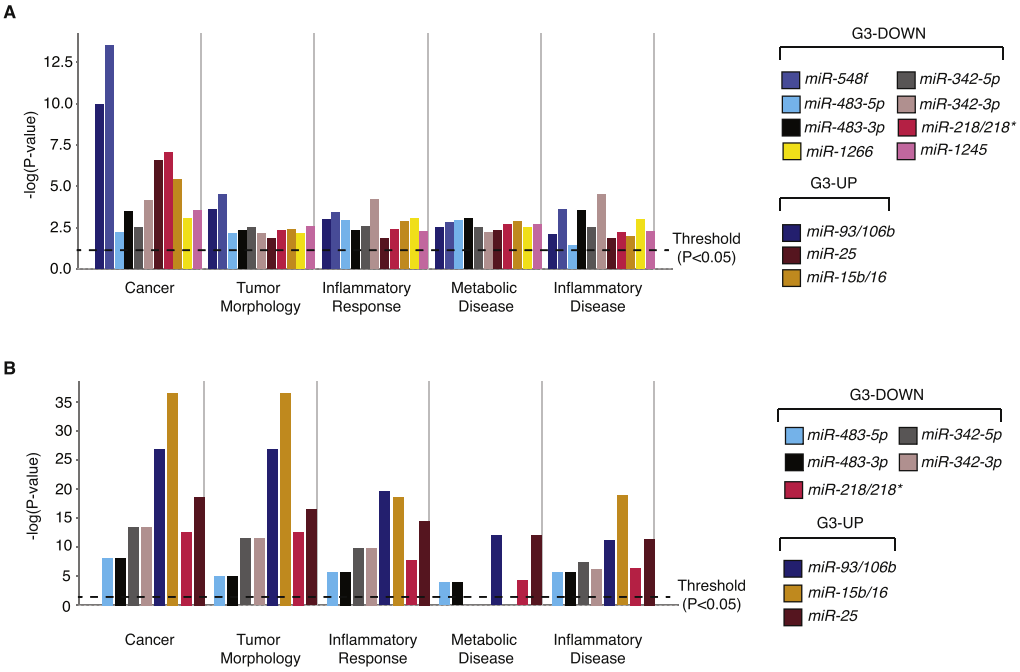
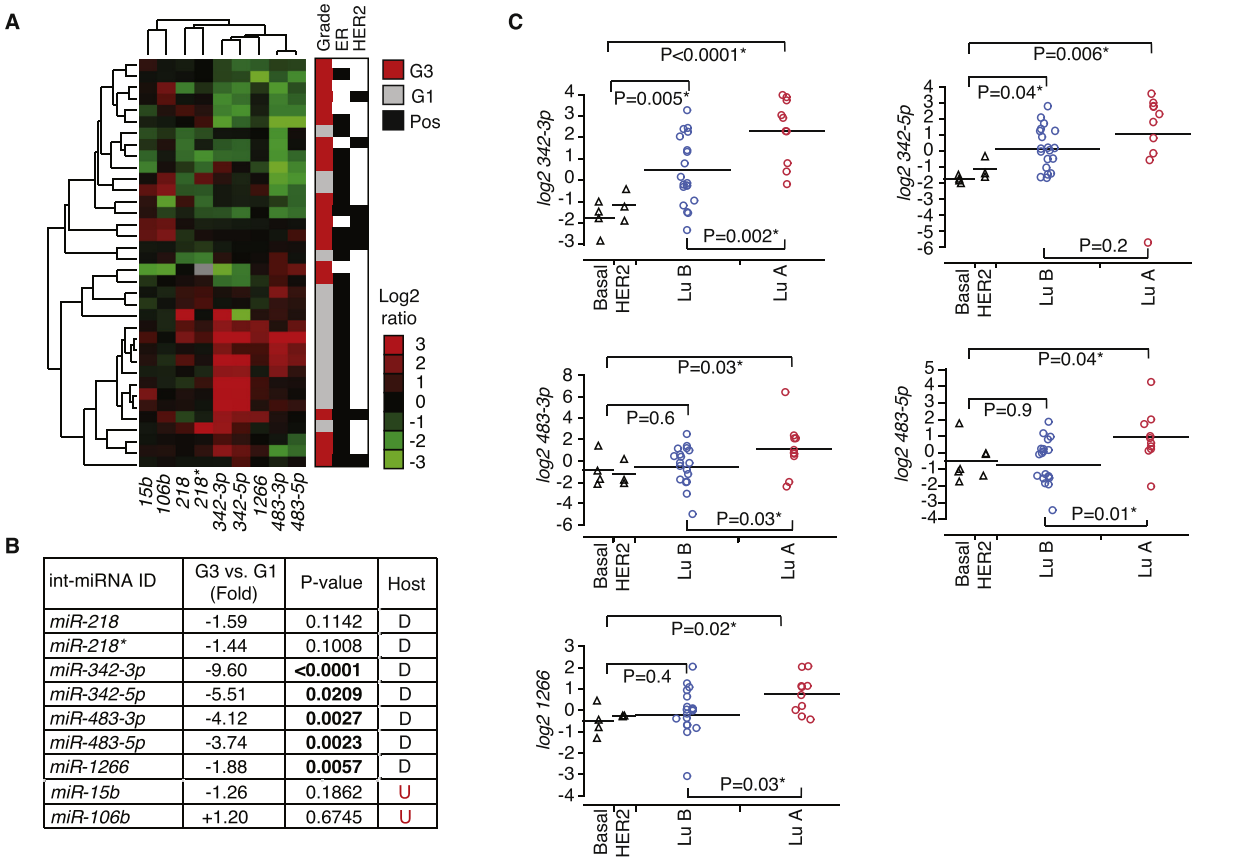


Figure 2 – Ingenuity Pathway Analysis of predicted and experimentally validated int-miRNA target genes. A. Bio-functions analysis of predicted target genes. B. Bio-functions analysis of experimentally validated target genes (see also Table S12). No validated targets were available for *miR-548f*, *miR-1245* and *miR-1266*. *Y*-axis, *P*-value (-log10; Fisher's Exact test) of the enrichment of int-miRNA targets in the biomolecular functions (displayed on the *X*-axis) annotated in the Ingenuity Bio-functions database. G3-DOWN, target genes of int-miRNAs located in host genes that are downregulated in G3 breast tumors. G3-UP, target genes of int-miRNAs located in host genes that are upregulated in G3 breast tumors. *P*-value (Fisher's exact test) cutoff was set at 0.05 (Threshold).



B

int-miRNA ID	G3 vs. G1 (Fold)	P-value	Host
miR-218	-1.59	0.1142	D
*miR-218**	-1.44	0.1008	D
miR-342-3p	-9.60	**<0.0001**	D
miR-342-5p	-5.51	**0.0209**	D
miR-483-3p	-4.12	**0.0027**	D
miR-483-5p	-3.74	**0.0023**	D
miR-1266	-1.88	**0.0057**	D
miR-15b	-1.26	0.1862	U
miR-106b	+1.20	0.6745	U

Figure 3 – Validation of the int-miRNA signature in the 36-patient cohort of G1 and G3 breast tumors by qRT-PCR. A. Hierarchical clustering of tumors based on the expression of selected int-miRNAs. *miR-15b* and *miR-106b* were used as representatives of their respective co-transcribed miRNA clusters, while *miR-548f-2* and *miR-1245* were undetectable by qRT-PCR. Columns represent log2 ratios of expression of each miRNA (median centered); rows represent tumor samples. Colored bars indicate the class of each patient. The color code, on the right, shows the characteristics of each patient: red, grade 3 tumor (G3); grey, grade 1 tumors (G1); ER, estrogen receptor (black = positive, white = negative); HER2, ErbB2 receptor (black = positive, white = negative). B. Differences in int-miRNA expression between G3 and G1 tumors. G3 *vs.* G1 (Fold): fold-change difference in expression; *P*-value calculated by Student's *t*-test; Host: Affymetrix host gene expression change in G3 *vs.* G1 tumors (D: downregulated, U: upregulated). C. Expression ratios (Log2) of int-miRNAs in breast tumor subtypes defined by ER/PgR, HER2 and Ki67 status. The tumor subtypes, shown on the *x*-axes, were identified as follows: basal = ER−, PgR- and HER2-; HER2 = ER−, PgR- and HER2+; LuB = ER+ and/or PgR+, Ki67 ≥ 14%; LuA = ER+ and/or PgR+, Ki67 < 14%. *P*-values were calculated using the Student's *t*-test. Asterisks, statistically significant *P*-values.

1266 were strongly downregulated in G3 breast tumors ($P < 0.05$; Figure 3B). *miR-218/218** were also downregulated, but not significantly ($P > 0.05$; Figure 3B). Similarly, the regulation of *miR-15b* and *miR-106b* (used as representatives of the *miR-15b~16-2* and the *miR-25~106b* naturally co-transcribed clusters) was not significant ($P > 0.05$; Figure 3B).

Since the expression of the miR-HGs was also differentially regulated in different breast cancer molecular subtypes (Table 2, Figure S3), we compared the expression levels of the significantly regulated int-miRNAs in the breast tumor subtypes in the 36-patient cohort. The different molecular subtypes – luminal A (LuA), Luminal B (LuB), basal, and HER2 – were identified using ER, progesterone receptor (PgR), HER2 and Ki67 immunohistochemistry markers (Blows et al., 2010; Cheang et al., 2009; Nielsen et al., 2004). Overall, the int-miRNAs showed significant downregulation in the most aggressive molecular subtypes (LuB, basal, and HER2) with respect to the less aggressive LuA subtype (Figure 3C). Similar results were obtained using an external dataset (Enerly et al., 2011) with matched miR-HG and int-miRNA expression profile (Figure S4). Importantly, we also observed a general downregulation of these int-miRNAs, and of their relative miR-HGs, in breast tumors when compared to normal breast epithelium (Figure S5). This latter finding suggests that the loss of expression of the analyzed int-miRNAs/miR-HGs might be directly involved in the transformation process.

3.5. *Detection of aggressive G2 breast cancers using the int-miRNA signature*

The sum of the previous results suggested an association between the downregulation of four int-miRNAs – *miR-342-3p*, *miR-483-3p*, *miR-483-5p* and *miR-1266* – and features of aggressiveness in breast cancer. We directly tested this possibility by taking advantage of a category of breast tumors with a moderate degree of differentiation (G2 tumors). The reason for doing so was dual. On the one hand, G2 tumors were not considered in our previous analyses, thus circumventing the risk of overfitting the data because of the selection of candidate int-miRNAs. On the other, G2 tumors represent a heterogeneous category, composed of tumors with varying degrees of aggressiveness (Gnant et al., 2011; Ivshina et al., 2006; Rakha et al., 2010; Sotiriou et al., 2006).

An independent cohort of 95 G2 tumors was profiled for *miR-483-3p/5p*, *miR-342-3p* and *miR-1266* expression (Table S8B). Hierarchical clustering analysis of int-miRNA expression profiles of this cohort, alone or together with those of the previously described 36-tumor G1/G3 cohort, revealed two main clusters, characterized by opposite regulation of the four int-miRNAs (Figure 4A). The cohort of G2 breast cancer patients was almost equally distributed between the two clusters. Interestingly, luminal A (less aggressive) and luminal B (more aggressive) G2 cancers co-segregated significantly with G1 and G3 tumors, respectively (Figure 4A), as also confirmed by contingency analysis ($P < 0.0001$; Figure 4B).

Three of the four int-miRNAs, *miR-483-3p*, *miR-483-5p* and *miR-1266*, were significantly downregulated in G2-LuB *vs.* G2-LuA tumors ($P < 0.0001$; Figure 4C), suggesting that they might identify more aggressive subtypes (i.e. Luminal B) even in G2 tumors. To investigate this possibility, we built a multivariate model based on these three miRNAs using the 95-patient cohort as the training set (see Materials and Methods). The model was then validated in an additional independent cohort

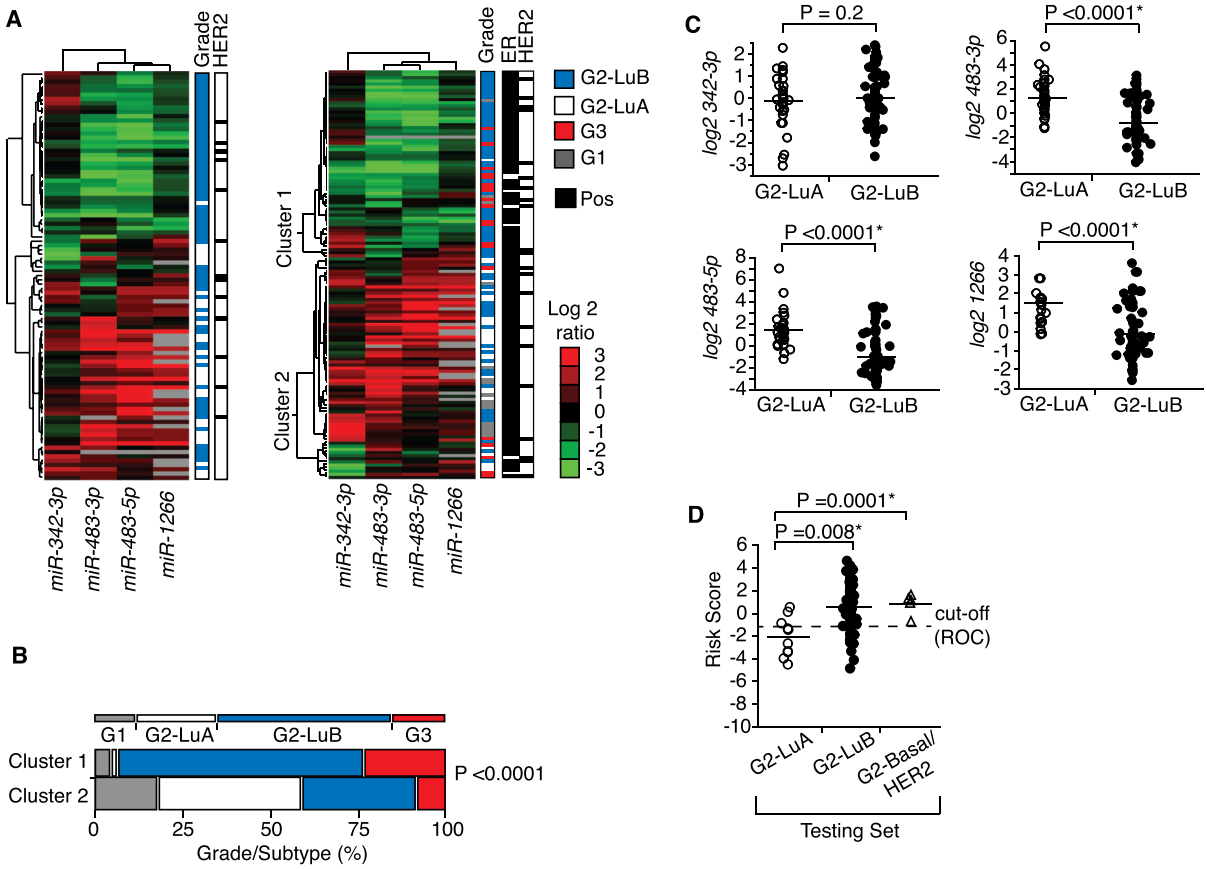


Figure 4 – Molecularly "aggressive" G2 breast cancers are identified by an int-miRNA signature. A. The cohort of 95 G2 tumors was analyzed by qRT-PCR for the expression of *miR-342-3p*, *miR-483* and *miR-1266*. Results, alone or together with those obtained in G3:G1 cohort (36 patients), were subjected to hierarchical clustering analysis. Columns represent log2 ratios of expression of int-miRNAs (median centered); rows represent tumor samples. The color code, on the right, indicates tumor grade/subtype: G3, red; G1, grey; G2-LuB, light blue; G2-LuA, white; ER, (black = positive); HER2, (black = positive). B. Contingency analysis of tumor distribution in Clusters 1 and 2. Color codes as in (A). *P*-value, likelihood-ratio test. C. Expression analysis of int-miRNAs in G2-luminal A (G2-LuA) and G2-luminal B (G2-LuB) tumors. *P*-values, Student's *t*-test. D. Performance of the int-miRNA signature composed of *miR-483-3p*, *miR-483-5p* and *miR-1266* in the additional independent cohort (Testing Set) of 90 G2 breast tumors. *Y*-axes: risk scores of the model. *X*-axes: breast tumor subtypes. Dashed line: decision cutoff used to classify patients in the high-*vs*. low-risk category, determined by nominal logistic regression and ROC analysis. *P*-values were calculated by the Student's *t*-test. Asterisks, statistically significant *P*-values.

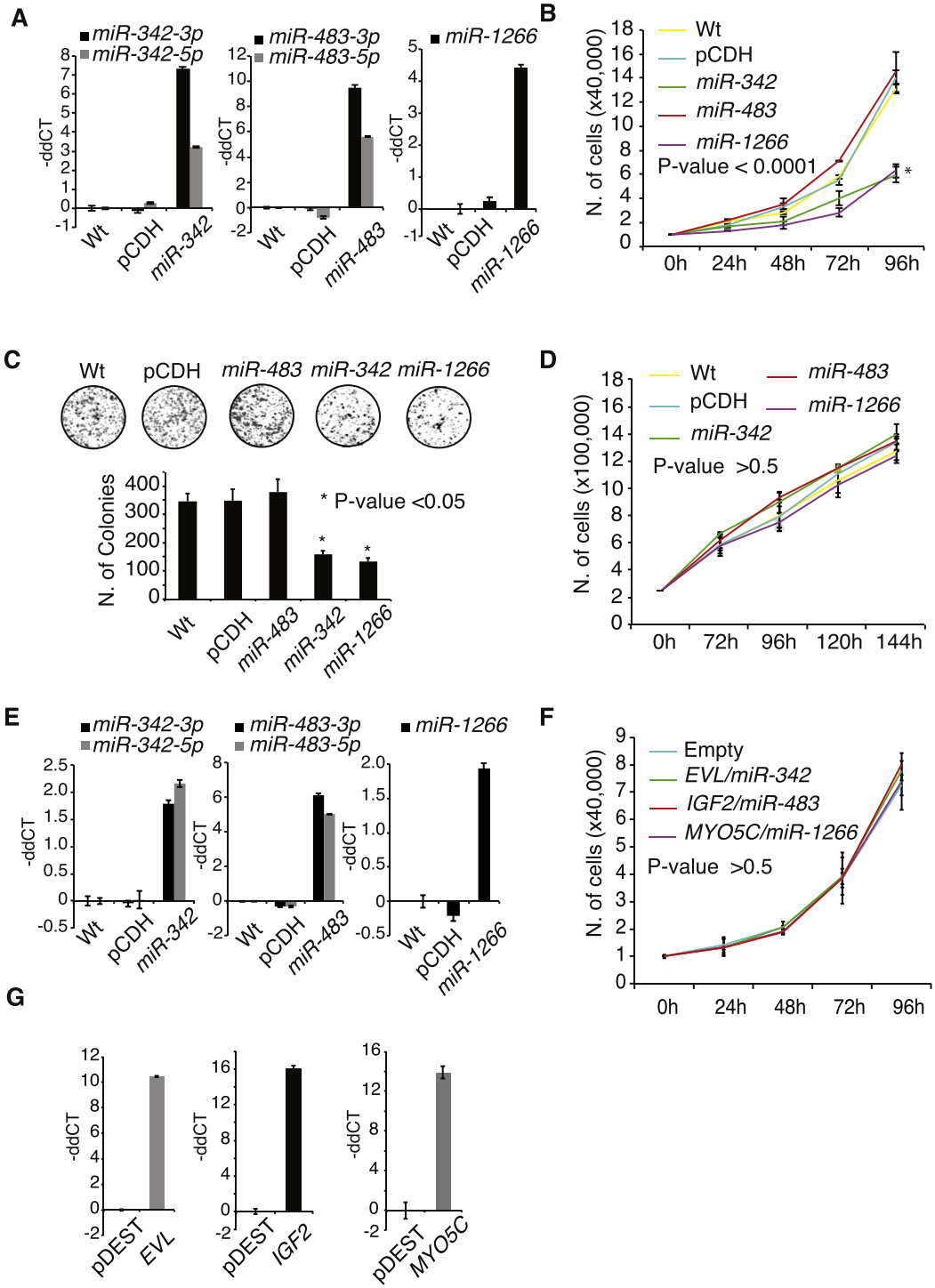


Figure 5 – Effects of overexpression of *miR-1266*, *miR-342-3p/5p* and *miR-483-3p/5p* and of relative host genes in breast cancer cell lines. Cells were infected with lentiviral vectors expressing precursors of *miR-342*, *miR-483* and *miR-1266*, or transfected with vectors expressing full-length *EVL*, *IGF2* and *MYO5C* miR-HGs. A. qRT-PCR analysis of int-miRNA expression in MDA-MB-231 cells. miRNA levels are reported as the Log2 normalized ratio of expression (-ddCT) relative to wild-type (non-infected) cells. pCDH, the empty vector pCDH-CMV-MCS-EF1-GFP was used as control. B. Cell proliferation assay with MDA-MB-231 cells. Data represent the mean ± σ from three independent experiments ($n = 3$). *P*-value, two-way ANOVA test relative to control cells (i.e., WT or pCDH). The asterisk indicates statistical significance. C. Colony forming assay with MDA-MB-231 cells. Images (top) represent colonies formed ten days after seeding. The bar graph (bottom) displays the mean ± σ from two independent experiments ($n = 2$). *, statistically significant *P*-value ($P < 0.05$) relative to control cells (WT or pCDH). D. Cell proliferation assay with MDA-MB-361 cells. Data represent the mean ± σ from three independent experiments. No significant differences

of 90 patients with G2 node-negative breast cancer, in which all subtypes were represented (Table S9). In this cohort, the model correctly identified all 5 basal/HER2 tumors and 60 of 75 LuB tumors (80%) as 'high risk', and 7 of 10 LuA tumors (70%) as 'low-risk' (Figure 4D). Thus, the risk model detected aggressive tumor subtypes with an accuracy, sensitivity and specificity of 80%, 81% and 70%, respectively.

In conclusion, we have provided a proof of principle that the detection of int-miRNAs, through data mining of miR-HGs in published cancer expression datasets, represents a viable strategy for the identification *in silico* of miRNAs of potential cancer relevance.

3.6. *Increased expression of miR-342 and miR-1266, but not of their host genes, impairs breast cancer cell proliferation and migration*

We investigated whether the regulated expression of the loci corresponding to *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-1266* has an impact on the biology of breast cancer cells. Hierarchical clustering of the *miR-342*, *miR-483* and *miR-1266* miR-HG expression profiles (i.e. *EVL*, *IGF2* and *MYO5C*, respectively) in a panel of 51 breast cancer cell lines, for which expression data are publicly available [Table S10 (Neve et al., 2006)], revealed two main clusters, enriched in cell lines of the basal (Cluster 1, 20 out of 22, $P < 0.0001$) or luminal (Cluster 2, 23 out of 29, $P < 0.0001$) subtypes (Figures S6A and B). We selected as a model system, the MDA-MB-231 cell line that displayed the lowest median expression levels of the *IGF2*, *EVL2* and *MYO5C* miR-HGs (Figure S6C). As a control, we selected the MDA-MB-361 cell line, which displayed an opposite, and quantitatively comparable, regulation of the same miR-HGs (Figure S6C). By qRT-PCR analysis, we confirmed that the expression of *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-1266* was indeed congruent with the expression of their host genes in both of the selected cell lines (Figure S6D).

We selected two relevant cancer phenotypes, proliferation and migration, to test the impact of restoration of high levels of expression of *miR-342*, *miR-483* and *miR-1266* in MDA-MB-231 cells, using MDA-MB-361 as a specificity control. We also tested the effects of overexpression of the corresponding miR-HGs (*IGF2*, *EVL2* and *MYO5C*). The lentiviral-mediated expression of *miR-342* and *miR-1266* (Figure 5A), caused a significant reduction in the proliferation rate and colony forming ability of MDA-MB-231, but not MDA-MB-361, cells (Figure 5B–E). Similarly, *miR-1266* expression significantly impaired cell migration of MDA-MB-231, but not MDA-MB-361, cells (Figure 6A–C). Conversely, overexpression of the miR-HGs did not affect proliferation or migration of MDA-MB-231 cells (Figure 5F–G, and Figure 6D). Finally, since negative self-regulation of miR-HGs has been reported in the literature (Bosia et al., 2012), we analyzed the expression of miR-HGs upon overexpression of their cognate int-miRNAs: no significant changes were observed (Figure S7).

4. Discussion

Here, we describe an approach to exploit an intrinsic characteristic of miRNAs, i.e. that ~50% of their genes reside within introns of protein-coding genes and share their regulation (He et al., 2012; Monteys et al., 2010; Ozsolak et al., 2008; Rodriguez et al., 2004). We reasoned that the wealth of publicly available microarray (mRNA) expression datasets might contain "encrypted" miRNA-related information that could be exploited for the discovery of biologically relevant miRNAs simply through meta-analysis.

In designing a proof-of-principle validation, we concentrated on breast cancer for which several high-quality, independent, transcriptome datasets are publicly available. We did so with multiple intents: i) to verify whether we could identify differentially expressed int-miRNAs simply by extracting the latent information in published gene expression (mRNA) datasets, ii) to verify whether int-miRNA signatures can be identified that would allow patient stratification, iii) to identify int-miRNAs whose involvement in biological processes, most notably cancer, was not previously known, iv) to investigate whether, at least in some cases, the dysregulations emerging from gene expression profiling might be more informative if viewed from the point of view of the hosted int-miRNA, rather than of the hosting gene.

Our efforts were successful on all accounts. Firstly, we were able to identify several miR-HGs, and their corresponding int-miRNAs, that are differentially expressed in various breast cancer subtypes. Importantly, several of the int-miRNAs that we identified have recently been found to be regulated in high-throughput miRNA expression profilings of independent cohorts of breast cancer patients [Table S11; (Aure et al., 2013; Blenkiron et al., 2007; Dvinge et al., 2013; Volinia et al., 2012)]. These data further support the effectiveness of our *in silico* approach to predict cancer-regulated int-miRNAs.

Secondly, we were able to identify an int-miRNA cancer signature – composed of *miR-342*, *miR-483* and *miR-1266* – that successfully stratified G2 cancers according to their molecular subtype, and therefore, according to their aggressiveness. We are not claiming a direct, even prospective, clinical utility of the identified signature or of the related risk model. Clearly, further studies are needed in this direction, aimed at, for example, comparing our int-miRNA cancer signature with existing stratification tools, such as molecular subtyping. However, our data demonstrate that the int-miRNA-related information "hidden" in transcriptomic profiles can be

were found by two-way ANOVA test relative to control cells ($P > 0.5$; WT or pCDH). E. qRT-PCR analysis of int-miRNA expression in MDA-MB-361 cells. miRNA levels are reported as the Log2 normalized ratio of expression (-ddCT) relative to wild-type (non-infected) cells. pCDH, the empty vector pCDH-CMV-MCS-EF1-GFP was used as control. F. Cell proliferation assay in MDA-MB-231 cells overexpressing the indicated miR-HGs. Data represent the mean ± σ from three independent experiments ($n = 3$). No significant differences were found by two-way ANOVA test ($P > 0.5$) relative to control cells (transfected with a pDEST26 empty vector). G. qRT-PCR analysis of miR-HG expression in MDA-MB-231 cells. *EVL*, *IGF2* or *MYO5C* host gene expression is reported as the Log2 normalized ratio of expression (-ddCT) relative to control (empty vector transfected) cells. Empty, the empty vector pDEST26 was used as a control.

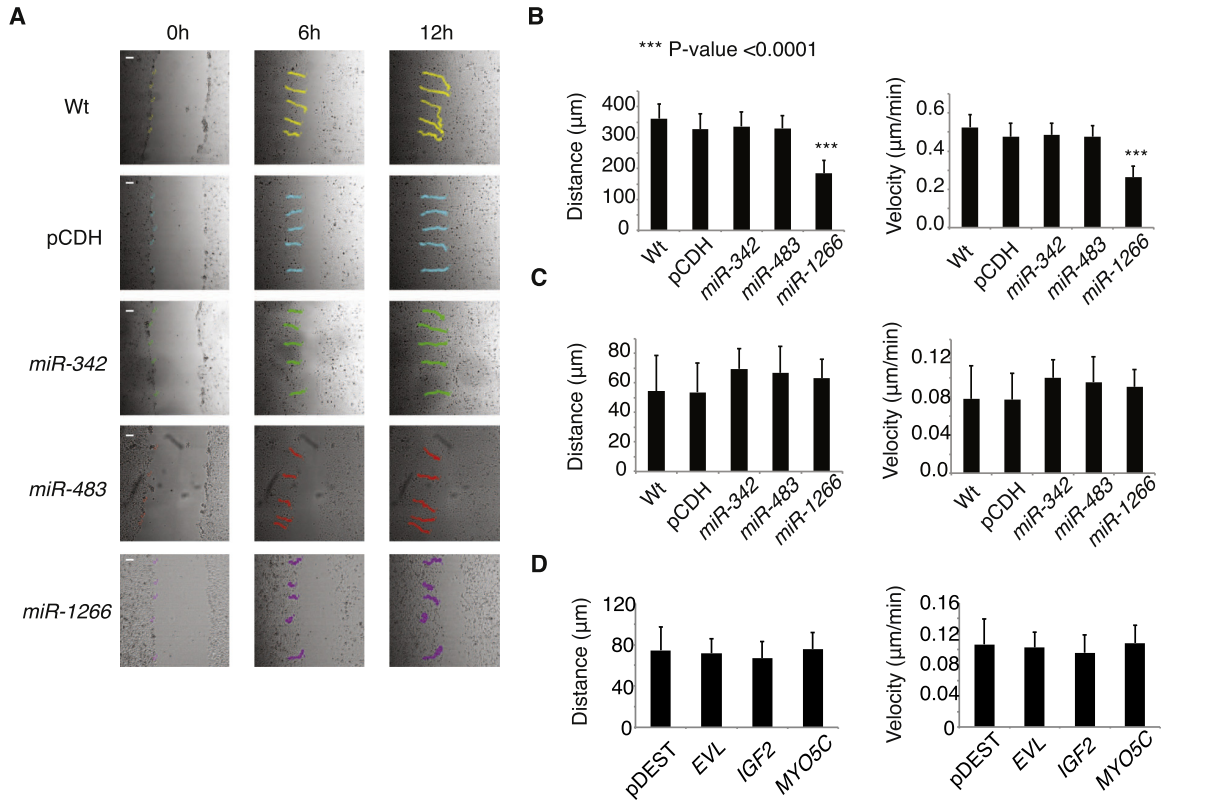


Figure 6 – Effects of overexpression of *miR-1266*, *miR-342-3p/5p* and *miR-483-3p/5p* on the migration of breast cancer cell lines. A. Monolayers of infected or wild-type (WT) MDA-MB-231 cells were scratch-wounded, as shown in the images on the left, and monitored by time-lapse video microscopy. Representative images were taken from movies at 0, 6 and 12 h. Colored lines show tracks of 5 representative cells. White Bar, 30 µm. B. Quantitation of the experiment shown in (A). Mean distance covered (left) and velocity (right) are shown. Data represent the mean ± σ from 15 individually tracked cells from 3 independent experiments. *P*-values were calculated using Welch's *t*-test analysis. ***, $P < 0.0001$ relative to control cells (Wt or pCDH). C. Quantitation of migration of MDA-MB-361 cells overexpressing *miR-1266*, *miR-342-3p/5p* and *miR-483-3p/5p*, relative to control cells (Wt or pCDH). Mean distance covered (left) and velocity (right) are shown. Data represent the mean ± σ from 15 individually tracked cells from 3 independent experiments. *P*-values were calculated using Welch's *t*-test analysis and were not significant ($P > 0.5$). D. Quantitation of migration of MDA-MB-231 cells overexpressing *EVL*, *IGF* and *MYO5C*, relative to control cells (pDEST). Mean distance covered (left) and velocity (right) are shown. Data represent the mean ± σ from 15 individually tracked cells from 3 independent experiments. *P*-values were calculated using Welch's *t*-test analysis and were not significant ($P > 0.5$).

successfully extracted *in silico* and used to identify miRNAs of potential clinical utility.

From a biological viewpoint, our approach led to the identification of two miRNAs, *miR-342* and *miR-1266*, that are involved in cancer-related phenotypes. The fact that overexpression of these miRNAs inhibited cancer-relevant phenotypes specifically in cells that display low expression of these miRNAs, but not in cells that express normal levels, argues that the downmodulation of *miR-342* and *miR-1266* might have a causal role in determining the aggressiveness of some breast cancers. This latter notion is supported by our findings that *miR-342* and *miR-1266* are differentially expressed between different breast cancer subtypes, and also underexpressed in some tumor tissues with respect to the normal breast epithelium. Of note, while there is some evidence in the literature indicating an involvement of *miR-342* in cancer (Dvinge et al., 2013; Veronese et al., 2010), there have been no reports, prior to this study, of a role of *miR-1266* in cancer (Ichihara et al., 2012).

Lastly, in the case of the loci encoding *miR-342/EVL* and *miR1266/MYO5C*, we report the intriguing observation that restoration of the expression of the int-miRNA, but not that of the host gene, inhibits cancer-relevant phenotypes. Thus, adding the "int-miRNA perspective" to the analysis and validation of gene expression studies is likely to increase the likelihood of identifying significant biological mechanisms involved in cancer.

5. Conclusions

In summary, we have developed an approach for the identification of biologically relevant int-miRNAs that has the potential to accelerate miRNA research by bypassing the lengthy and costly phase of initial screenings, and substituting them with meta-analysis of miR-HGs in publicly available expression datasets. The approach apparently performs well both in the holistic-oriented field of signature identification, which

finds primary application in projects of clinical interest, and in the more traditional field of "gene hunting" to guide high-resolution studies. While we have applied the methodology to the breast cancer setting to obtain a proof of principle of its utility, we envision applications in several fields of biology and medicine for which high quality gene expression datasets are available.

Acknowledgments

We are indebted with the Molecular Pathology Unit and the Clinical Biomarker Lab (Francesca De Santis, Stefania Pirroni) and the Imaging Facility at the Molecular Medicine Program of IEO; Pascale R. Romano and Rosalind Gunby for critically editing the manuscript. We also thank the IEO Biobank and Biomolecular Resource Infrastructure (IBBRI) and the Division of Pathology at IEO. This work was supported by grants to PPDF from Associazione Italiana per la Ricerca sul Cancro (AIRC, IG 10349 and 14404 and MCO 10.000 to PPDF), MIUR (the Italian Ministry of University and Scientific Research), the Italian Ministry of Health; from European Research Council (Mammastem Project), and from the Monzino Foundation.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.molonc.2014.10.001.

REFERENCES

Aure, M.R., Leivonen, S.K., Fleischer, T., Zhu, Q., Overgaard, J., Alsner, J., Tramm, T., Louhimo, R., Alnaes, G.I., Perala, M., Busato, F., Touleimat, N., Tost, J., Borresen-Dale, A.L., Hautaniemi, S., Troyanskaya, O.G., Lingjaerde, O.C., Sahlberg, K.K., Kristensen, V.N., 2013. Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. Genome Biol. 14, R126.

Baskerville, S., Bartel, D.P., 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA 11, 241–247.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A., 2005. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37, 382–390.

Blenkiron, C., Goldstein, L.D., Thorne, N.P., Spiteri, I., Chin, S.F., Dunning, M.J., Barbosa-Morais, N.L., Teschendorff, A.E., Green, A.R., Ellis, I.O., Tavare, S., Caldas, C., Miska, E.A., 2007. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol. 8, R214.

Blows, F.M., Driver, K.E., Schmidt, M.K., Broeks, A., van Leeuwen, F.E., Wesseling, J., Cheang, M.C., Gelmon, K., Nielsen, T.O., Blomqvist, C., Heikkila, P., Heikkinen, T., Nevanlinna, H., Akslen, L.A., Begin, L.R., Foulkes, W.D., Couch, F.J., Wang, X., Cafourek, V., Olson, J.E., Baglietto, L., Giles, G.G., Severi, G., McLean, C.A., Southey, M.C., Rakha, E., Green, A.R., Ellis, I.O., Sherman, M.E., Lissowska, J., Anderson, W.F., Cox, A., Cross, S.S., Reed, M.W., Provenzano, E., Dawson, S.J., Dunning, A.M., Humphreys, M., Easton, D.F., Garcia-Closas, M., Caldas, C., Pharoah, P.D., Huntsman, D., 2010. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. Plos Med. 7, e1000279.

Bolstad, B.M., Collin, F., Simpson, K.M., Irizarry, R.A., Speed, T.P., 2004. Experimental design and low-level analysis of microarray data. Int. Rev. Neurobiol. 60, 25–58.

Bonci, D., Coppola, V., Musumeci, M., Addario, A., Giuffrida, R., Memeo, L., D'Urso, L., Pagliuca, A., Biffoni, M., Labbaye, C., Bartucci, M., Muto, G., Peschle, C., De Maria, R., 2008. The miR-15a-miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. Nat. Med. 14, 1271–1277.

Bosia, C., Osella, M., Baroudi, M.E., Cora, D., Caselle, M., 2012. Gene autoregulation via intronic microRNAs and its functions. BMC Syst. Biol. 6, 131.

Calin, G.A., Croce, C.M., 2006. MicroRNA signatures in human cancers. Nat. Rev. Cancer 6, 857–866.

Carbone, M., Rizzo, P., Grimley, P.M., Procopio, A., Mew, D.J., Shridhar, V., de Bartolomeis, A., Esposito, V., Giuliano, M.T., Steinberg, S.M., Levine, A.S., Giordano, A., Pass, H.I., 1997. Simian virus-40 large-T antigen binds p53 in human mesotheliomas. Nat. Med. 3, 908–912.

Cheang, M.C., Chia, S.K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P.S., Parker, J.S., Perou, C.M., Ellis, M.J., Nielsen, T.O., 2009. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J. Natl. Cancer Inst. 101, 736–750.

Chen, D.T., Nasir, A., Culhane, A., Venkataramu, C., Fulp, W., Rubio, R., Wang, T., Agrawal, D., McCarthy, S.M., Gruidl, M., Bloom, G., Anderson, T., White, J., Quackenbush, J., Yeatman, T., 2010. Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. Breast Cancer Res. Treat. 119, 335–346.

De Luca, A., Baldi, A., Esposito, V., Howard, C.M., Bagella, L., Rizzo, P., Caputi, M., Pass, H.I., Giordano, G.G., Baldi, F., Carbone, M., Giordano, A., 1997. The retinoblastoma gene family pRb/p105, p107, pRb2/p130 and simian virus-40 large T-antigen in human mesotheliomas. Nat. Med. 3, 913–916.

Dvinge, H., Git, A., Graf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E.A., Chin, S.F., Provenzano, E., Turashvili, G., Green, A., Ellis, I., Aparicio, S., Caldas, C., 2013. The shaping and functional consequences of the microRNA landscape in breast cancer. Nature 497, 378–382.

Dweep, H., Sticht, C., Pandey, P., Gretz, N., 2011. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J. Biomed. Inform. 44, 839–847.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.K., Aure, M.R., Russnes, H.G., Ronneberg, J.A., Johnsen, H., Navon, R., Rodland, E., Makela, R., Naume, B., Perala, M., Kallioniemi, O., Kristensen, V.N., Yakhini, Z., Borresen-Dale, A.L., 2011. miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. PLoS One 6, e16915.

Fury, W., Batliwalla, F., Gregersen, P.K., Li, W., 2006. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. Conf. Proc. IEEE Eng. Med. Biol. Soc. 1, 5531–5534.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Girardi, A.J., Sweet, B.H., Slotnick, V.B., Hilleman, M.R., 1962. Development of tumors in hamsters inoculated in the neonatal period with vacuolating virus, SV-40. Proc. Soc. Exp. Biol. Med. 109, 649–660.
Gnant, M., Harbeck, N., Thomssen, C., 2011. St. Gallen 2011: summary of the consensus discussion. Breast Care (Basel) 6, 136–141.
Grady, W.M., Parkin, R.K., Mitchell, P.S., Lee, J.H., Kim, Y.H., Tsuchiya, K.D., Washington, M.K., Paraskeva, C., Willson, J.K., Kaz, A.M., Kroh, E.M., Allen, A., Fritz, B.R., Markowitz, S.D., Tewari, M., 2008. Epigenetic silencing of the intronic microRNA hsa-miR-342 and its host gene EVL in colorectal cancer. Oncogene 27, 3880–3888.
Griffiths-Jones, S., 2007. Annotating noncoding RNA genes. Annu. Rev. Genomics Hum. Genet., 279–298.
Hagstrom, K.A., Meyer, B.J., 2003. Condensin and cohesin: more than chromosome compactor and glue. Nat. Rev. Genet. 4, 520–534.
He, C., Li, Z., Chen, P., Huang, H., Hurst, L.D., Chen, J., 2012. Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene coevolution. Nucleic Acids Res. 40, 4002–4012.
Ichihara, A., Jinnin, M., Oyama, R., Yamane, K., Fujisawa, A., Sakai, K., Masuguchi, S., Fukushima, S., Maruo, K., Ihn, H., 2012. Increased serum levels of miR-1266 in patients with psoriasis vulgaris. Eur. J. Dermatol. 22, 68–71.
Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J.E., Liu, E.T., Bergh, J., Kuznetsov, V.A., Miller, L.D., 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 66, 10292–10301.
Krol, J., Loedige, I., Filipowicz, W., 2010. The widespread regulation of microRNA biogenesis, function and decay. Nat. Rev. Genet. 11, 597–610.
Lei, M., Tye, B.K., 2001. Initiating DNA synthesis: from recruiting to activating the MCM complex. J. Cell Sci. 114, 1447–1454.
Lutter, D., Marr, C., Krumsiek, J., Lang, E.W., Theis, F.J., 2010. Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. BMC Genomics 11.
Meijering, E., Dzyubachyk, O., Smal, I., 2012. Methods for cell and particle tracking. Methods Enzymol. 504, 183–200.
Minn, A.J., Kang, Y., Serganova, I., Gupta, G.P., Giri, D.D., Doubrovin, M., Ponomarev, V., Gerald, W.L., Blasberg, R., Massague, J., 2005. Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. J. Clin. Invest. 115, 44–55.
Monteys, A.M., Spengler, R.M., Wan, J., Tecedor, L., Lennox, K.A., Xing, Y., Davidson, B.L., 2010. Structure and activity of putative intronic miRNA promoters. RNA 16, 495–505.
Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F., Speed, T., Spellman, P.T., DeVries, S., Lapuk, A., Wang, N.J., Kuo, W.L., Stilwell, J.L., Pinkel, D., Albertson, D.G., Waldman, F.M., McCormick, F., Dickson, R.B., Johnson, M.D., Lippman, M., Ethier, S., Gazdar, A., Gray, J.W., 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 10, 515–527.
Nielsen, T.O., Hsu, F.D., Jensen, K., Cheang, M., Karaca, G., Hu, Z., Hernandez-Boussard, T., Livasy, C., Cowan, D., Dressler, L., Akslen, L.A., Ragaz, J., Gown, A.M., Gilks, C.B., van de Rijn, M., Perou, C.M., 2004. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin. Cancer Res. 10, 5367–5374.
Ozsolak, F., Poling, L.L., Wang, Z., Liu, H., Liu, X.S., Roeder, R.G., Zhang, X., Song, J.S., Fisher, D.E., 2008. Chromatin structure analyses identify miRNA promoters. Genes Dev. 22, 3172–3183.
Palamidessi, A., Frittoli, E., Garre, M., Faretta, M., Mione, M., Testa, I., Diaspro, A., Lanzetti, L., Scita, G., Di Fiore, P.P., 2008. Endocytic trafficking of Rac is required for the spatial restriction of signaling in cell migration. Cell 134, 135–147.
Poliseno, L., Salmena, L., Riccardi, L., Fornari, A., Song, M.S., Hobbs, R.M., Sportoletti, P., Varmeh, S., Egia, A., Fedele, G., Rameh, L., Loda, M., Pandolfi, P.P., 2010. Identification of the miR-106b~25 microRNA cluster as a proto-oncogenic PTEN-targeting intron that cooperates with its host gene MCM7 in transformation. Sci. Signal. 3, ra29.
Radfar, M.H., Wong, W., Morris, Q., 2011. Computational prediction of intronic microRNA targets using host gene expression reveals novel regulatory mechanisms. PLoS One 6.
Rainer, J., Ploner, C., Jesacher, S., Ploner, A., Eduardoff, M., Mansha, M., Wasim, M., Panzer-Grumayer, R., Trajanoski, Z., Niederegger, H., Kofler, R., 2009. Glucocorticoid-regulated microRNAs and mirtrons in acute lymphoblastic leukemia. Leukemia 23, 746–752.
Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., Palacios, J., Richardson, A.L., Schnitt, S.J., Schmitt, F.C., Tan, P.H., Tse, G.M., Badve, S., Ellis, I.O., 2010. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 12, 207.
Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., Bradley, A., 2004. Identification of mammalian microRNA host genes and transcription units. Genome Res. 14, 1902–1910.
Saini, H.K., Griffiths-Jones, S., Enright, A.J., 2007. Genomic analysis of human microRNA transcripts. Proc. Natl. Acad. Sci. USA, 17719–17724.
Song, L., Dai, T., Xie, Y., Wang, C., Lin, C., Wu, Z., Ying, Z., Wu, J., Li, M., Li, J., 2012. Up-regulation of miR-1245 by c-myc targets BRCA2 and impairs DNA repair. J. Mol. Cell Biol. 4, 108–117.
Soon, P.S., Tacon, L.J., Gill, A.J., Bambach, C.P., Sywak, M.S., Campbell, P.R., Yeh, M.W., Wong, S.G., Clifton-Bligh, R.J., Robinson, B.G., Sidhu, S.B., 2009. miR-195 and miR-483-5p identified as predictors of poor prognosis in adrenocortical cancer. Clin. Cancer Res. 15, 7684–7692.
Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Lonning, P.E., Borresen-Dale, A.L., 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA 98, 10869–10874.
Sotiriou, C., Neo, S.Y., McShane, L.M., Korn, E.L., Long, P.M., Jazaeri, A., Martiat, P., Fox, S.B., Harris, A.L., Liu, E.T., 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc. Natl. Acad. Sci. USA 100, 10393–10398.
Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., Delorenzi, M., 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. 98, 262–272.
Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J.J., Ladd-Acosta, C., Mesirov, J., Golub, T.R., Jacks, T., 2005. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat. Genet. 37, 48–55.
Tie, J., Pan, Y., Zhao, L., Wu, K., Liu, J., Sun, S., Guo, X., Wang, B., Gang, Y., Zhang, Y., Li, Q., Qiao, T., Zhao, Q., Nie, Y., Fan, D., 2010. MiR-218 inhibits invasion and metastasis of gastric cancer by targeting the Robo1 receptor. Plos Genet. 6, e1000879.
van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J.,

Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Venet, D., Dumont, J.E., Detours, V., 2011. Most random gene expression signatures are significantly associated with breast cancer outcome. Plos Comput. Biol. 7, e1002240.

Veronese, A., Lupini, L., Consiglio, J., Visone, R., Ferracin, M., Fornari, F., Zanesi, N., Alder, H., D'Elia, G., Gramantieri, L., Bolondi, L., Lanza, G., Querzoli, P., Angioni, A., Croce, C.M., Negrini, M., 2010. Oncogenic role of miR-483-3p at the IGF2/483 locus. Cancer Res. 70, 3140–3149.

Voduc, K.D., Cheang, M.C., Tyldesley, S., Gelmon, K., Nielsen, T.O., Kennecke, H., 2010. Breast cancer subtypes and the risk of local and regional relapse. J. Clin. Oncol. 28, 1684–1691.

Volinia, S., Galasso, M., Sana, M.E., Wise, T.F., Palatini, J., Huebner, K., Croce, C.M., 2012. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc. Natl. Acad. Sci. USA 109, 3024–3029.

Wang, D., Lu, M., Miao, J., Li, T.T., Wang, E., Cui, Q.H., 2009. Cepred: predicting the co-expression patterns of the human intronic microRNAs with their host genes. PLoS One 4.

Wang, Y.P., Li, K.B., 2009. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. BMC Genomics 10, 218.

White, D.P., Caswell, P.T., Norman, J.C., 2007. Alpha v beta3 and alpha5beta1 integrin recycling pathways dictate downstream Rho kinase signaling to regulate persistent cell migration. J Cell Biol. 177, 515–525.

Yendamuri, S., Kratzke, R., 2011. MicroRNA biomarkers in lung cancer: MiRacle or quagMiRe? Transl. Res. 157, 209–215.

Bibliography

[Alison et al., 2010] Alison, M. R., Islam, S., and Wright, N. A. (2010). Stem cells in cancer: instigators and propagators? *Journal of Cell Science*, 123(14):2357–2368.

[Beer, 2008] Beer, D. G. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827.

[Bhatlekar et al., 2014] Bhatlekar, S., Addya, S., Salunek, M., Orr, C. R., Surrey, S., McKenzie, S., Fields, J. Z., and Boman, B. M. (2014). Identification of a developmental gene expression signature, including HOX genes, for the normal human colonic crypt stem cell niche: overexpression of the signature parallels stem cell overpopulation during colon tumorigenesis. *Stem cells and development*, 23(2):167–179.

[Bianchi et al., 2007] Bianchi, F., Nuciforo, P., Vecchi, M., Bernard, L., Tizzoni, L., Marchetti, A., Buttitta, F., Felicioni, L., Nicassio, F., and Di Fiore, P. P. P. (2007). Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. *The Journal of clinical investigation*, 117(11):3436–3444.

[Brambilla et al., 2001] Brambilla, E., Travis, W. D., Colby, T. V., Corrin, B., and Shimosato, Y. (2001). The new World Health Organization classification of lung tumours. *The European respiratory journal*, 18(6):1059–1068.

[Carè et al., 2001] Carè, A., Felicetti, F., Meccia, E., Bottero, L., Parenza, M., Stoppacciaro, A., Peschle, C., and Colombo, M. P. (2001). HOXB7: A Key Factor for Tumor-associated Angiogenic Switch. *Cancer Research*, 61(17):6532–6539.

[Caré et al., 1998] Caré, A., Silvani, A., Meccia, E., Mattia, G., Peschle, C., and Colombo, M. P. (1998). Transduction of the SkBr3 breast carcinoma cell line with the HOXB7 gene induces bFGF expression, increases cell proliferation and reduces growth factor dependence. *Oncogene*, 16(25):3285–3289.

[Caré et al., 1996] Caré, A., Silvani, A., Meccia, E., Mattia, G., Stoppacciaro, A., Parmiani, G., Peschle, C., and Colombo, M. P. (1996). HOXB7 constitutively activates basic fibroblast growth factor in melanomas. *Molecular and cellular biology*, 16(9):4842–4851.

[Carè et al., 1999] Carè, A., Valtieri, M., Mattia, G., Meccia, E., Masella, B., Luchetti, L., Felicetti, F., Colombo, M. P., and Peschle, C. (1999). Enforced expression of HOXB7 promotes hematopoietic stem cell proliferation and myeloid-restricted progenitor differentiation. *Oncogene*, 18(11):1993–2001.

[Chang et al., 2009] Chang, T.-C. C., Zeitels, L. R., Hwang, H.-W. W., Chivukula, R. R., Wentzel, E. A., Dews, M., Jung, J., Gao, P., Dang, C. V., Beer, M. A., Thomas-Tikhonenko, A., and Mendell, J. T. (2009). Lin-28B transactivation is necessary for Myc-mediated let-7 repression and proliferation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(9):3384–3389.

[Chen et al., 2004] Chen, H., Chung, S., and Sukumar, S. (2004). HOXA5-induced apoptosis in breast cancer cells is mediated by caspases 2 and 8. *Molecular and cellular biology*, 24:924–935.

[Chen et al., 2008] Chen, H., Lee, J. S. S., Liang, X., Zhang, H., Zhu, T., Zhang, Z., Taylor, M. E., Zahnow, C., Feigenbaum, L., Rein, A., and Sukumar, S. (2008). Hoxb7 inhibits transgenic HER-2/neu-induced mouse mammary tumor onset but promotes progression and lung metastasis. *Cancer research*, 68(10):3637–3644.

[Chien et al., 2015] Chien, C.-S. S., Wang, M.-L. L., Chu, P.-Y. Y., Chang, Y.-L. L., Liu, W.-H. H., Yu, C.-C. C., Lan, Y.-T. T., Huang, P.-I. I., Lee, Y.-Y. Y., Chen, Y.-

W. W., Lo, W.-L. L., and Chiou, S.-H. H. (2015). Lin28B/let-7 regulates expression of oct4 and sox2 and reprograms oral squamous cell carcinoma cells to a stem-like state. *Cancer research*, 75(12):2553–2565.

[Crinò et al., 2010] Crinò, L., Weder, W., van Meerbeeck, J., Felip, E., and behalf of the ESMO Guidelines Working Group, O. (2010). Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 21(suppl 5):v103–v115.

[Davis et al., 1995] Davis, A., Peter, Witte, D. P., Hsieh-Li, H. M., Potter, S., Steven, and Capecchi, M. R. (1995). Absence of radius and ulna in mice lacking hoxa-11 and hoxd-11. *Nature*, 375(6534):791–795.

[De Souza Setubal Destro et al., 2010] De Souza Setubal Destro, M. F. F., Bitu, C. C. C., Zecchin, K. G., Graner, E., Lopes, M. A., Kowalski, L. P. P., and Coletta, R. D. (2010). Overexpression of HOXB7 homeobox gene in oral cancer induces cellular proliferation and is associated with poor prognosis. *International journal of oncology*, 36(1):141–149.

[Dontu et al., 2003] Dontu, G., Abdallah, W. M., Foley, J. M., Jackson, K. W., Clarke, M. F., Kawamura, M. J., and Wicha, M. S. (2003). In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes & development*, 17(10):1253–1270.

[Eramo et al., 2007] Eramo, A., Lotti, F., Sette, G., Pilozzi, E., Biffoni, M., Di Virgilio, A., Conticello, C., Ruco, L., Peschle, C., and De Maria, R. (2007). Identification and expansion of the tumorigenic lung cancer stem cell population. *Cell Death & Differentiation*, 15(3):504–514.

[Ferlay et al., 2013] Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence

and mortality patterns in europe: estimates for 40 countries in 2012. *European journal of cancer (Oxford, England : 1990)*, 49(6):1374–1403.

[Gilbert, 1996] Gilbert, S. F. (1996). *Biologia dello sviluppo*. Zanichelli, 2nd italian edition.

[Golpon et al., 2001] Golpon, H. A., Geraci, M. W., Moore, M. D., Miller, H. L., Miller, G. J., Tuder, R. M., and Voelkel, N. F. (2001). HOX genes in human lung: altered expression in primary pulmonary hypertension and emphysema. *The American journal of pathology*, 158:955–966.

[Golub et al., 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

[Gouti and Gavalas, 2008] Gouti, M. and Gavalas, A. (2008). Hoxb1 controls cell fate specification and proliferative capacity of neural stem and progenitor cells. *Stem cells (Dayton, Ohio)*, 26(8):1985–1997.

[Grier et al., 2005] Grier, D. G., Thompson, A., Kwasniewska, A., McGonigle, G. J., Halliday, H. L., and Lappin, T. R. (2005). The pathophysiology of HOX genes and their role in cancer. *Journal of Pathology*, 205:154–171.

[Heo et al., 2009] Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., Yeom, K.-H., Han, J., and Kim, V. N. (2009). TUT4 in concert with lin28 suppresses MicroRNA biogenesis through Pre-MicroRNA uridylation. *Cell*, 138(4):696–708.

[Herbst et al., 2008] Herbst, R. S., Heymach, J. V., and Lippman, S. M. (2008). Lung cancer. *The New England journal of medicine*, 359(13):1367–1380.

[Hirsch et al., 2008] Hirsch, F. R., Spreafico, A., Novello, S., Wood, M. D. D., Simms, L., and Papotti, M. (2008). The prognostic and predictive role of histology in ad-

vanced non-small cell lung cancer: a literature review. *Journal of thoracic oncology*, 3(12):1468–1481.

[Ho et al., 2007] Ho, M. M., Ng, A. V., Lam, S., and Hung, J. Y. (2007). Side population in human lung cancer cell lines and tumors is enriched with stem-like cancer cells. *Cancer Research*, 67(10):4827–4833.

[Hollestelle et al., 2013] Hollestelle, A., Peeters, J. K., Smid, M., Timmermans, M., Verhoog, L. C., Westenend, P. J., Heine, A. A., Chan, A., Sieuwerts, A. M., Wiemer, E. A., Klijn, J. G., van der Spek, P. J., Foekens, J. A., Schutte, M., den Bakker, M. A., and Martens, J. W. (2013). Loss of e-cadherin is not a necessity for epithelial to mesenchymal transition in human breast cancer. *Breast cancer research and treatment*, 138(1):47–57.

[Iliopoulos et al., 2009] Iliopoulos, D., Hirsch, H. A., and Struhl, K. (2009). An epigenetic switch involving NF-kappaB, lin28, let-7 MicroRNA, and IL6 links inflammation to cell transformation. *Cell*, 139(4):693–706.

[Jeter et al., 2015] Jeter, C. R., Yang, T., Wang, J., Chao, H.-P. P., and Tang, D. G. (2015). Concise review: NANOG in cancer stem cells and tumor development: An update and outstanding questions. *Stem cells (Dayton, Ohio)*, 33(8):2381–2390.

[Jett et al., 2013] Jett, J. R., Schild, S. E., Kesler, K. A., and Kalemkerian, G. P. (2013). Treatment of small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143(5 Suppl).

[Jiang et al., 2009] Jiang, F., Qiu, Q., Khanna, A., Todd, N. W., Deepak, J., Xing, L., Wang, H., Liu, Z., Su, Y., Stass, S. A., and Katz, R. L. (2009). Aldehyde dehydrogenase 1 is a tumor stem cell-associated marker in lung cancer. *Molecular cancer research : MCR*, 7(3):330–338.

[Jin et al., 2012] Jin, K., Kong, X., Shah, T., Penet, M.-F., Wildes, F., Sgroi, D. C., Ma, X.-J., Huang, Y., Kallioniemi, A., Landberg, G., Bieche, I., Wu, X., Lobie, P. E., Davidson, N. E., Bhujwalla, Z. M., Zhu, T., and Sukumar, S. (2012). The HOXB7 protein renders breast cancer cells resistant to tamoxifen through activation of the EGFR pathway. *Proceedings of the National Academy of Sciences*, 109(8):2736–2741.

[Jin et al., 2015] Jin, K., Park, S., Teo, W. W., Korangath, P., Cho, S. S., Yoshida, T., Gyorffy, B., Goswami, C. P., Nakshatri, H., Cruz, L.-A., Zhou, W., Ji, H., Su, Y., Ekram, M., Wu, Z., Zhu, T., Polyak, K., and Sukumar, S. (2015). HOXB7 is an ERα cofactor in the activation of HER2 and multiple ER target genes leading to endocrine resistance. *Cancer Discovery*, pages CD–15–0090.

[Kalluri and Weinberg, 2009] Kalluri, R. and Weinberg, R. A. (2009). The basics of epithelial-mesenchymal transition. *The Journal of clinical investigation*, 119:1420–1428.

[Kim et al., 2005] Kim, C. B. F., Jackson, E. L., Woolfenden, A. E., Lawrence, S., Babar, I., Vogel, S., Crowley, D., Bronson, R. T., and Jacks, T. (2005). Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell*, 121(6):823–835.

[Kong et al., 2010] Kong, D., Banerjee, S., Ahmad, A., Li, Y., Wang, Z., Sethi, S., and Sarkar, F. H. (2010). Epithelial to mesenchymal transition is mechanistically linked with stem cell signatures in prostate cancer cells. *PloS one*, 5(8).

[Kornberg, 1993] Kornberg, T. B. (1993). Understanding the homeodomain. *Journal of Biological Chemistry*, 268(36):26813–26816.

[Korpanty et al., 2014] Korpanty, G. J., Graham, D. M., Vincent, M. D., and Leighl, N. B. (2014). Biomarkers that currently affect clinical practice in lung cancer: EGFR, ALK, MET, ROS-1, and KRAS. *Frontiers in oncology*, 4.

[Le Chevalier et al., 2005] Le Chevalier, T., Arriagada, R., Pignon, J.-P. P., and Scagliotti, G. V. (2005). Should adjuvant chemotherapy become standard treat-

ment in all patients with resected non-small-cell lung cancer? *The lancet oncology*, 6:182–184.

[Levina et al., 2008] Levina, V., Marrangoni, A. M., DeMarco, R., Gorelik, E., and Lokshin, A. E. (2008). Drug-selected human lung cancer stem cells: cytokine network, tumorigenic and metastatic properties. *PloS one*, 3(8):e3077.

[Lewis, 1978] Lewis, E. B. (1978). A gene complex controlling segmentation in drosophila. *Nature*, 276(5688):565–570.

[Liao et al., 2011] Liao, W.-T. T., Jiang, D., Yuan, J., Cui, Y.-M. M., Shi, X.-W. W., Chen, C.-M. M., Bian, X.-W. W., Deng, Y.-J. J., and Ding, Y.-Q. Q. (2011). HOXB7 as a prognostic factor and mediator of colorectal cancer progression. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 17(11):3569–3578.

[Liu et al., 2015] Liu, S., Jin, K., Hui, Y., Fu, J., Jie, C., Feng, S., Reisman, D., Wang, Q., Fan, D., Sukumar, S., and Chen, H. (2015). HOXB7 promotes malignant progression by activating the TGFb signaling pathway. *Cancer research*, 75(4):709–719.

[Liu et al., 2013] Liu, Y., Li, H., Feng, J., Cui, X., Huang, W., Li, Y., Su, F., Liu, Q., Zhu, J., Lv, X., Chen, J., Huang, D., and Yu, F. (2013). Lin28 induces Epithelial-to-Mesenchymal transition and stemness via downregulation of let-7a in breast cancer cells. *PLoS ONE*, 8(12):e83083.

[Lu et al., 2014] Lu, H., Clauser, K. R., Tam, W. L., Fröse, J., Ye, X., Eaton, E. N., Reinhardt, F., Donnenberg, V. S., Bhargava, R., Carr, S. A., and Weinberg, R. A. (2014). A breast cancer stem cell niche supported by juxtacrine signalling from monocytes and macrophages. *Nat Cell Biol*, 16(11):1105–1117.

[Mani et al., 2008] Mani, S. A., Guo, W., Liao, M.-J., Eaton, E. N., Ayyanan, A., Zhou, A. Y., Brooks, M., Reinhard, F., Zhang, C. C., Shipitsin, M., Camp-

bell, L. L., Polyak, K., Brisken, C., Yang, J., and Weinberg, R. A. (2008). The Epithelial-Mesenchymal transition generates cells with properties of stem cells. *Cell*, 133(4):704–715.

[Martini et al., 1995] Martini, N., Bains, M. S., Burt, M. E., Zakowski, M. F., McCormack, P., Rusch, V. W., and Ginsberg, R. J. (1995). Incidence of local recurrence and second primary tumors in resected stage I lung cancer. *J Thorac Cardiovasc Surg*, 109(1):120–129.

[Med, 2011] Med, N. E. J. (2011). Reduced Lung-Cancer mortality with Low-Dose computed tomographic screening. *N Engl J Med*, 365(5):395–409.

[Mollard and Dziadek, 1997] Mollard, R. and Dziadek, M. (1997). Homeobox genes from clusters A and B demonstrate characteristics of temporal colinearity and differential restrictions in spatial expression domains in the branching mouse lung. *The International journal of developmental biology*, 41:655–666.

[Monterisi et al., 2015] Monterisi, S., D'Ario, G., Dama, E., Rotmensz, N., Confalonieri, S., Tordonato, C., Troglio, F., Bertalot, G., Maisonneuve, P., Viale, G., Nicassio, F., Vecchi, M., Di Fiore, P. P. P., and Bianchi, F. (2015). Mining cancer gene expression databases for latent information on intronic microRNAs. *Molecular oncology*, 9(2):473–487.

[Morgan, 2006] Morgan, R. (2006). Hox genes: a continuation of embryonic patterning? *Trends in genetics : TIG*, 22(2):67–69.

[Naora et al., 2001] Naora, H., Yang, Y. Q., Montz, F. J., Seidman, J. D., Kurman, R. J., and Roden, R. B. (2001). A serologically identified tumor antigen encoded by a homeobox gene promotes growth of ovarian epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America*, 98(7):4060–4065.

[Newman et al., 2008] Newman, M. A., Thomson, J. M., and Hammond, S. M. (2008). Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. *RNA (New York, N.Y.)*, 14(8):1539–1549.

[Nguyen Kovochich et al., 2013] Nguyen Kovochich, A., Arensman, M., Lay, A. R., Rao, N. P., Donahue, T., Li, X., French, S. W., and Dawson, D. W. (2013). HOXB7 promotes invasion and predicts survival in pancreatic adenocarcinoma. *Cancer*, 119(3):529–539.

[Nusslein-Volhard and Wieschaus, 1980] Nusslein-Volhard, C. and Wieschaus, E. (1980). Mutations affecting segment number and polarity in drosophila. *Nature*, 287(5785):795–801.

[Okayama et al., 2012] Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., Furuta, K., Tsuta, K., Shibata, T., Yamamoto, S., Watanabe, S.-i., Sakamoto, H., Kumamoto, K., Takenoshita, S., Gotoh, N., Mizuno, H., Sarai, A., Kawano, S., Yamaguchi, R., Miyano, S., and Yokota, J. (2012). Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer research*, 72(1):100–111.

[Pearson et al., 2005] Pearson, J. C., Lemons, D., and McGinnis, W. (2005). Modulating hox gene functions during animal body patterning. *Nat Rev Genet*, 6(12):893–904.

[Pece et al., 2010] Pece, S., Tosoni, D., Confalonieri, S., Mazzarol, G., Vecchi, M., Ronzoni, S., Bernard, L., Viale, G., Pelicci, P. G. G., and Di Fiore, P. P. P. (2010). Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. *Cell*, 140(1):62–73.

[Plowright et al., 2009] Plowright, L., Harrington, K. J., Pandha, H. S., and Morgan, R. (2009). HOX transcription factors are potential therapeutic targets in non-small-

cell lung cancer (targeting HOX genes in lung cancer). *British journal of cancer*, 100:470–475.

[Raman et al., 2000] Raman, V., Martensen, S. A., Reisman, D., Evron, E., Odenwald, W. F., Jaffee, E., Marks, J., and Sukumar, S. (2000). Compromised HOXA5 function can limit p53 expression in human breast tumours . *Nature*, 405:974–978.

[Riddihough, 1992] Riddihough, G. (1992). Homing in on the homeobox. *Nature*, 357(6380):643–644.

[Sakiyama, 2000] Sakiyama, J. (2000). Coordinated Expression of Hoxb Genes and Signaling Molecules during Development of the Chick Respiratory Tract. *Developmental Biology*, 227:12–27.

[Sauvageau et al., 1995] Sauvageau, G., Thorsteinsdottir, U., Eaves, C. J., Lawrence, H. J., Largman, C., Lansdorp, P. M., and Humphries, R. K. (1995). Overexpression of HOXB4 in hematopoietic cells causes the selective expansion of more primitive populations in vitro and in vivo. *Genes & development*, 9(14):1753–1765.

[Shah and Sukumar, 2010] Shah, N. and Sukumar, S. (2010). The Hox genes and their roles in oncogenesis. *Nature reviews. Cancer*, 10(5):361–371.

[Shyh-Chang and Daley, 2013] Shyh-Chang, N. and Daley, G. Q. (2013). Lin28: Primal regulator of growth and metabolism in stem cells. *Cell Stem Cell*, 12(4):395–406.

[Siegel et al., 2015] Siegel, R. L., Miller, K. D., and Jemal, A. (2015). Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29.

[Soh et al., 2012] Soh, B. S. S., Zheng, D., Li Yeo, J. S. S., Yang, H. H. H., Ng, S. Y. Y., Wong, L. H. H., Zhang, W., Li, P., Nichane, M., Asmat, A., Wong, P. S. S., Wong, P. C. C., Su, L. L. L., Mantalaris, S. A., Lu, J., Xian, W., McKeon, F., Chen, J., Lim, E. H. H., and Lim, B. (2012). CD166(pos) subpopulation from differentiated

human ES and iPS cells support repair of acute lung injury. *Molecular therapy : the journal of the American Society of Gene Therapy*, 20(12):2335–2346.

[Storti et al., 2011] Storti, P., Donofrio, G., Colla, S., Airoldi, I., Bolzoni, M., Agnelli, L., Abeltino, M., Todoerti, K., Lazzaretti, M., Mancini, C., Ribatti, D., Bonomini, S., Franceschi, V., Pistoia, V., Lisignoli, G., Pedrazzini, A., Cavicchi, O., Neri, A., Rizzoli, V., and Giuliani, N. (2011). HOXB7 expression by myeloma cells regulates their pro-angiogenic properties in multiple myeloma patients. *Leukemia*, 25(3):527–537.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[Suda et al., 2010] Suda, K., Tomizawa, K., and Mitsudomi, T. (2010). Biological and clinical significance of KRAS mutations in lung cancer: an oncogenic driver that contrasts with EGFR mutation. *Cancer and Metastasis Reviews*, 29(1):49–60.

[Suvà et al., 2013] Suvà, M. L., Riggi, N., and Bernstein, B. E. (2013). Epigenetic reprogramming in cancer. *Science*, 339(6127):1567–1570.

[Tabuse et al., 2011] Tabuse, M., Ohta, S., Ohashi, Y., Fukaya, R., Misawa, A., Yoshida, K., Kawase, T., Saya, H., Thirant, C. A., Chneiweiss, H. A., Matsuzaki, Y., Okano, H., Kawakami, Y., and Toda, M. (2011). Functional analysis of HOXD9 in human gliomas and glioma cancer stem cells. *Molecular Cancer*, 10(1):10–60.

[Takahashi and Yamanaka, 2006] Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676.

[Thompson and Nguyen, 2000] Thompson, A. A. and Nguyen, L. T. (2000). Amegakaryocytic thrombocytopenia and radio-ulnar synostosis are associated with HOXA11 mutation. *Nature Genetics*, 26(4):397–398.

[Unternaehrer et al., 2014] Unternaehrer, J. J., Zhao, R., Kim, K., Cesana, M., Powers, J. T., Ratanasirintrawoot, S., Onder, T., Shibue, T., Weinberg, R. A., and Daley, G. Q. (2014). The Epithelial-Mesenchymal transition factor SNAIL paradoxically enhances reprogramming. *Stem Cell Reports*, 3(5):691–698.

[Viswanathan and Daley, 2010] Viswanathan, S. R. and Daley, G. Q. (2010). Lin28: A microRNA regulator with a macro role. *Cell*, 140(4):445–449.

[Viswanathan et al., 2009] Viswanathan, S. R., Powers, J. T., Einhorn, W., Hoshida, Y., Ng, T. L., Toffanin, S., O'Sullivan, M., Lu, J., Phillips, L. A., Lockhart, V. L., Shah, S. P., Tanwar, P. S., Mermel, C. H., Beroukhim, R., Azam, M., Teixeira, J., Meyerson, M., Hughes, T. P., Llovet, J. M., Radich, J., Mullighan, C. G., Golub, T. R., Sorensen, P. H., and Daley, G. Q. (2009). Lin28 promotes transformation and is associated with advanced human malignancies. *Nature Genetics*, 41(7):843–848.

[Walters et al., 2013] Walters, S., Maringe, C., Coleman, M. P., Peake, M. D., Butler, J., Young, N., Bergström, S., Hanna, L., Jakobsen, E., Kölbeck, K., Sundstrøm, S., Engholm, G., Gavin, A., Gjerstorff, M. L., Hatcher, J., Johannesen, T. B., Linklater, K. M., McGahan, C. E., Steward, J., Tracey, E., Turner, D., Richards, M. A., Rachet, B., Brostrøm, S., Bryant, H., Currow, D., Gavin, A., Gunnarsson, G., Hanson, J., Harper, T., Kaasa, S., Richards, M. A., Sherar, M., Thomas, B., Adolfsson, J., Andersen, O., Bryant, H., Coldman, A., Dhaliwal, D., Engholm, G., Gavin, A., Forman, D., Gjerstorff, M. L., Hatcher, J., Hosbond, C., Johannesen, T. B., Lambe, M., Linklater, K. M., Marrett, L., McGahan, C. E., McLaughlin, J., Meechan, D., Middleton, R., Milnes, K., Nishri, D., Quin, N., Rabenek, L., Russell, C., Shin, J., Steward, J., Tracey, E., Turner, D., Bergström, S., Boyer, M., Butler, J., Evans, W., Hanna, L., Jakobsen, E., Kölbeck, K., Lester, J., McAleese, J., Peake, M. D.,

Sundstrøm, S., Young, N., Williamson, I., Walters, S., Maringe, C., Coleman, M. P., and Bernard, R. (2013). Lung cancer survival and stage at diagnosis in australia, canada, denmark, norway, sweden and the UK: a population-based study, 2004-2007. *Thorax*, pages thoraxjnl–2012–202297.

[Wang et al., 2013] Wang, P., Gao, Q., Suo, Z., Munthe, E., Solberg, S., Ma, L., Wang, M., Westerdaal, N. A. C. A., Kvalheim, G., and Gaudernack, G. (2013). Identification and characterization of cells with cancer stem cell properties in human primary lung cancer cell lines. *PloS one*, 8(3).

[Wellik, 2007] Wellik, D. M. (2007). Hox patterning of the vertebrate axial skeleton. *Dev. Dyn.*, 236(9):2454–2463.

[Whelan et al., 2008] Whelan, J. T., Ludwig, D. L., and Bertrand, F. E. (2008). HoxA9 induces insulin-like growth factor-1 receptor expression in b-lineage acute lymphoblastic leukemia. *Leukemia*, 22(6):1161–1169.

[Wu et al., 2006] Wu, X., Chen, H., Parker, B., Rubin, E., Zhu, T., Lee, J. S., Argani, P., and Sukumar, S. (2006). HOXB7, a Homeodomain Protein, Is Overexpressed in Breast Cancer and Confers Epithelial-Mesenchymal Transition. *Cancer Research*, 66(19):9527–9534.

[Xu et al., 2013] Xu, Y., Wei, X., Wang, M., Zhang, R., Fu, Y., Xing, M., Hua, Q., and Xie, X. (2013). Proliferation rate of somatic cells affects reprogramming efficiency. *Journal of Biological Chemistry*, 288(14):9767–9778.

[Yan et al., 2013] Yan, X., Luo, H., Zhou, X., Zhu, B., Wang, Y., and Bian, X. (2013). Identification of CD90 as a marker for lung cancer stem cells in a549 and h446 cell lines. *Oncology reports*, 30(6):2733–2740.

[Yu et al., 2007] Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin,

I. I., and Thomson, J. A. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920.

[Yuan et al., 2014] Yuan, W., Zhang, X., Xu, Y., Li, S., Hu, Y., and Wu, S. (2014). Role of HOXB7 in regulation of progression and metastasis of human lung adenocarcinoma. *Molecular carcinogenesis*, 53(1):49–57.

[Zhou et al., 2013] Zhou, J., Ng, S.-B. B., and Chng, W.-J. J. (2013). LIN28/LIN28B: an emerging oncogenic driver in cancer stem cells. *The international journal of biochemistry & cell biology*, 45(5):973–978.