PhD degree in Molecular Medicine

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Faculty of Medicine

Settore disciplinare: MED/04

CDK12 is a novel oncogene with clinical and pathogenetic relevance in breast cancer

Angelo Taglialatela

IFOM-IEO Campus, Milan

Matricola n. R08924

Supervisor: Prof. Pier Paolo Di Fiore

IFOM-IEO Campus, Milan

Added co-Supervisor: Prof. Salvatore Pece

IFOM-IEO Campus, Milan

Anno accademico 2011-2012

TABLE OF CONTENTS

FIGURES INDEX

Abbreviations List

- STK: serine/threonine kinase
- CDK: cyclin dependent kinase
- CKI: cell cycle kinase inhibitor
- RS: arginine/serine
- ISH: in situ hybridization
- IHC: immunohistochemistry
- TMA: tissue microarray
- HR: hormone receptor
- ER: estrogen receptor
- PR: progesterone receptor
- NGS: next generation sequencing
- DCIS: ductal carcinoma in situ
- IDC: invasive ductal carcinoma
- ILC: invasive lobular carcinoma
- TDLU: terminal ductal lobular unit
- FDA: food and drug administration
- FACS: fluorescence activated cell sorting
- IB: immunoblotting
- IP: immunoprecipitation
- IF: immunofluorescence
- FFPE: formalin-fixed paraffin-embedded
- CTR: control
- Q-PCR: quantitative RT-PCR
- FISH: fluorescence in situ hybridization
- s-dev.: standard deviation
- KD: knock-down
- sh: small-hairpin
- MEC: mammary epithelial cells
- s.e.: short exposure
- l.e.: long exposure
- EV: empty vector

- EMT: epithelial mesenchimal transition
- EGFP: enhanced green fluorescent protein
- FACS: fluorescence activated cell sorting
- DDR: DNA damage response
- OE: overexpression
- GSEA: gene set enrichment analysis
- IPA: ingenuity pathway analysis
- RNAPII: RNA polymerase II
- CTD: carboxy terminal domain

1 ABSTRACT

Breast cancer heterogeneity demands new reliable prognostic markers and therapeutic targets for the personalized management of patients. Over recent years, knowledge on the involvement of different types of kinases in cancer has guided the design of a variety of kinase inhibitors as novel molecularly targeted anti-cancer agents.

In a previous high-throughput screening performed by *in situ* hybridization (ISH) and immunohistochemistry (IHC) on breast cancer tissue microarrays (TMAs), the cyclin-dependent kinase 12 (CDK12) was found to be frequently overexpressed in breast cancer and its overexpression correlated with clinical/pathological parameters of aggressive disease. CDK12 was therefore proposed to be a novel prognostic biomarker in breast cancer.

In the present thesis, we extend these preliminary studies and demonstrate, by IHC analysis on TMAs comprising large cohorts of breast cancer patients, that CDK12 overexpression is significantly associated with clinical/pathological parameters of poor prognosis (high tumor grade, high Ki67 proliferative index, positive ERBB2 status) and with a higher risk of disease recurrence and death. We also show, either in human breast tumors or in breast cancer cell lines, that CDK12 overexpression is due to the amplification of the *CDK12* gene, which may occur either as a single event or in association with *ERBB2* amplification.

Through the use of amenable cell models in diverse *in vitro* and *in vivo* assays, we provide evidence that CDK12 is a *bona fide* oncogene in breast cancer: i) CDK12 overexpression enhances the tumorigenicity of breast cancer cells with normal CDK12 levels; ii) tumor cells harboring CDK12 amplification/overexpression depend

on the continuous presence of this oncogene for the maintenance of their tumor phenotypes. By genome-wide expression analysis, we show that alterations in CDK12 expression are associated with changes in the transcription and splicing of genes involved in cancer-relevant cellular processes, such as DNA damage response, cell cycle and epithelial-to-mesenchymal transition.

In conclusion, we have established that CDK12 is a novel prognostic biomarker in breast cancer and determined that this kinase has an intrinsic oncogenic activity, which most likely involves CDK12 function in the regulation of transcription and splicing of key cancer-related genes. Together these data indicate that CDK12 represents a potential novel molecular target for therapeutic intervention in breast cancer.

2 INTRODUCTION

2.1 BREAST CANCER

Breast cancer is a heterogeneous disease and notwithstanding the significant advances in the comprehension of its pathogenesis, diagnosis and treatment over the last decades, remains one of the most common cancers and a major cause of death. According to the latest worldwide statistics provided by the "International Agency for Research on Cancer", breast cancer is by far the most commonly diagnosed cancer and the most common cause of cancer death in women, with 1.38 million new cases and approximately 460,000 deaths in 2008 [1].

While current diagnostic and therapeutic tools allow us to treat many breast cancer patients at early stages of the disease resulting in good clinical outcomes, a significant proportion of patients relapses and develops metastatic disease. Currently, our therapeutic potential against metastatic breast cancer is still limited, and distant metastases represent the major cause of breast cancer-related death. Thus, the identification of novel prognostic biomarkers that are able to predict the risk of disease progression, as well as of novel molecular targets for the development of anti-cancer therapies that have the potential to prevent and cure metastatic breast cancer, is vital for improving the clinical management of breast cancer patients. Over the last few years, this task has been greatly enhanced by the introduction of a variety of high-throughput technologies that have allowed the identification and clinical validation of molecular signatures and/or individual genes with a potential value as clinical biomarkers and, eventually, targets for rationale therapies. This is, for instance, the case for CDK12, the subject of this thesis work, that is a serine/threonine kinase (STK) initially identified and characterized as a putative novel prognostic biomarker in breast cancer in a high-throughput screening conducted by *in situ* hybridization (ISH) on tissue microarrays (TMA) for the identification of novel kinases involved in breast cancer [2].

Based on this initial evidence, we investigated in this study i) the correlation between CDK12 overexpression and clinical/pathological and prognostic parameters in large cohorts of breast cancer patients, ii) the actual implication of CDK12 as an oncogene in breast carcinogenesis through several complementary functional studies in amenable cell-based models, iii) and the putative molecular mechanisms underlying its function in breast cancer. To introduce this project, I will first provide a brief overview of the state-of-the-art in the clinical management of breast cancer, followed by the analysis of the problem of tumor heterogeneity as the major hurdle to the personalized treatment of breast cancer patients. I will then introduce current knowledge on the molecular pathogenesis of breast cancer. Finally, I will describe the current literature relating to CDK12, including previous evidence implicating CDK12 overexpression in breast cancer.

2.2 The normal mammary gland

In order to understand the pathogenesis of breast cancer it is important to first introduce the structure of the normal mammary gland and its hierarchical tissue organization. The human breast is characterized by a branching network of ducts that end in clusters of small ducts that constitute the terminal ductal lobular units (TDLUs; Figure 1A).

Mammary epithelial cells represent the fundamental functional unit of the gland. The cellular epithelial architecture is composed of a bilayer of luminal cells surrounding an inner lumen and an external layer of myoepithelial and basal cells [3] (Figure 1B). These epithelial cells are surrounded by fibroblasts and adipocytes that compose the stroma of the mammary gland.

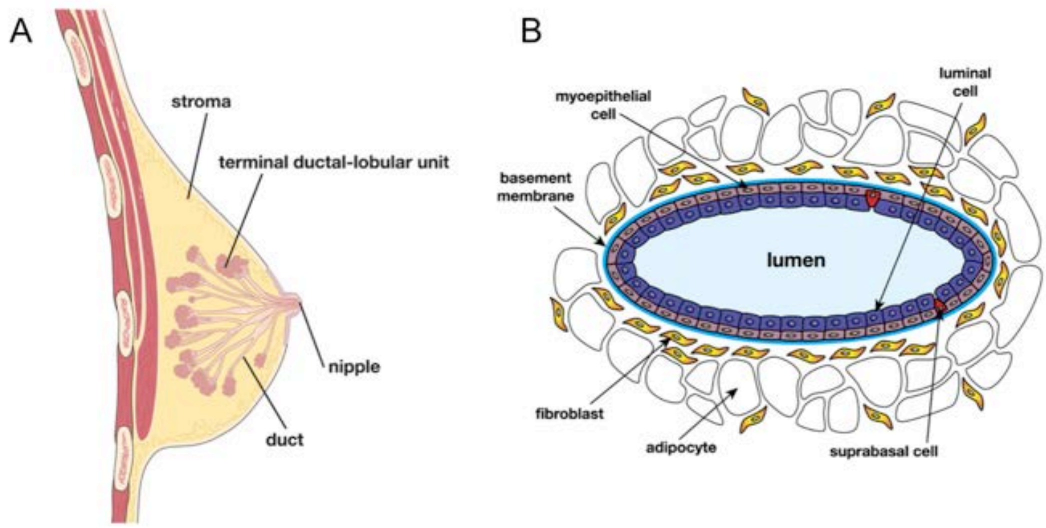


Figure 1. Schematic representation of the mammary gland

(A) Macroscopic structure of the human mammary gland. (B). Cellular composition and architectural organization of a human mammary duct. Figure adapted from [4].

2.3 Clinical management of breast cancer patients

The actual routine clinical management of breast cancer patients relies on prognostic and predictive information acquired by evaluation of clinico-pathological and biological factors. A prognostic factor can be defined as a factor that predicts outcome in the absence of systemic therapy or that predicts an outcome different from that of patients without that factor, despite empiric therapy. A predictive marker is a marker that predicts the differential efficacy (benefit) of a particular therapy based on marker status [5]. An overlap between prognostic and predictive factors exists and a proportion of them exhibit both characteristics.

The classical markers currently used to determine prognosis and response to therapies in breast cancer are:

- Clinical factors (tumor size [6], axillary lymph node status [7], metastases [8]);
- Histological factors (tumor grade and histological type [9]);
- Biological factors (proliferation index [10], *ERBB2* [11] and hormonal receptor (HR) status [12].

These classical markers have been established in several studies as robust prognostic and predictive factors. However, some of them are difficult to determine, and many do not have confirmed independent prognostic value. In order to better refine breast cancer patient stratification into risk groups, many histological and biological factors that determine prognosis are interrelated in multi-parametric systems such as the tumor, node, metastasis (TNM) staging system [13], prognostic algorithms (e.g., Adiuvant!Online [14]), guidelines (e.g., St. Gallen guidelines [15]), and indices (e.g., Nottingham Prognostic Index [16]).

Breast cancer patients undergo local treatments to control local disease and systemic treatment for micrometastatic disease. Local treatments consist of surgery and radiotherapy [17]. Surgery can be partial with excision of the tumor and a part of surrounding normal breast tissue (breast conservative) or total with complete resection of the breast (mastectomy) [17]. Clinical trials have shown insignificant differences in local recurrence and long-time survival between the two approaches, hence, most cases undergo conservative surgery [18,19]. Total removal of the gland is needed only in cases of multicentric invasive carcinomas, extensive intraductal carcinomas and large primary carcinomas that are not reduced in size by pre-operative chemotherapy [20].

In most breast-cancer centers, surgery is followed by postoperative radiotherapy of the whole breast [21]. Systemic adjuvant therapies are given to cure micrometastases, which could potentially be present in all patients with invasive cancer, in order to reduce the risk of relapse. The systemic treatments rely on cytotoxic drugs (chemotherapy) or selective drugs able to target molecules aberrantly expressed in cancer cells (targeted therapy).

Since the discovery of steroid-hormone receptors and their subversion in breast cancer [22], researchers set out to specifically target molecules and networks subverted in cancer cells and the concept of targeted therapy begun to take place.

Targeted therapies, approved and currently used in clinical practice, are endocrine (estrogen receptor antagonists or aromatase inhibitors) and anti-ErbB2 therapies. The status of the biological factors estrogen receptor (ER), progesterone receptor (PR) and *ERBB2* is predictive of the response to these targeted therapies: Tamoxifen, an ER antagonist, is the usual endocrine treatment for hormone receptor (ER and PR) positive breast cancers [23,24]; Trastuzumab, a monoclonal antibody that block the ErbB2 receptor, in association with several cytotoxic compounds, is the treatment of choice for ErbB2-positive cancers. Other ErbB2 small molecule inhibitors are currently under investigation in clinical trials [25].

However, despite the introduction of new and tailored surgical procedures and the development of new targeted therapies based on the increased understanding of the molecular and cell biology processes altered in cancer cells, the clinical management of breast cancer patients remains complicated because of the extensive inter- and intra-tumor heterogeneity that characterizes this disease and that affects response to therapies and guide disease recurrence and metastasis.

2.4 The problem of breast cancer heterogeneity

Breast cancer heterogeneity comprises inter- and intra-tumor heterogeneity. Inter-tumor heterogeneity has long been recognized by histo-pathologists who, based on their microscopic observations, were able to identify and classify 17 different histological subtypes of breast cancer with different clinical behavior [26]. Beyond gross histological differences among tumors, pathologists have also been able to develop a grading system based on the level of differentiation, number of mitoses and nuclear pleomorphism, to classify tumors into different grades with different clinical behavior [9].

An additional level of complexity in the understanding of cancer pathogenesis is the intra-tumor heterogeneity. Intra-tumor heterogeneity has long been observed by histopathologists as areas with different morphology and staining patterns within a tumor sample [27]. However, in a cancer diagnosis, the histopathologists will report the highest grade observed among different regions of the same tumor [28].

Tumor heterogeneity complicates the clinical management of breast cancer patients and represents a major problem in the prediction of prognosis and response to therapy. The molecular understanding of both inter- and intra-tumor heterogeneity has until recently been poor. Now with the advent of new post-genomic technologies, such as microarray techniques and next-generation sequencing (NGS) important advances are being made. This is particularly true for the comprehension of inter-tumor heterogeneity, while our understanding of intra-tumor heterogeneity is still relatively poor.

2.4.1 Histological Classification

Pathologists, on the basis of histological differences between tumors, were the first to perform breast cancer classification. Based on microscopic observations, breast cancer could be classified as non-invasive (referred to as *in situ*) or invasive [26,29]. Non-invasive breast cancers consist of cancer cells restricted within the basement membrane, which covers the underlying connective tissue in the breast. DCIS (Ductal Carcinoma *In Situ*) is the most common type of non-invasive breast cancer [29]. Invasive cancer occurs when cancer cells spread beyond the basement membrane with a higher risk of developing metastasis. These tumors are classified as invasive ductal carcinoma (IDC), comprising between 70 - 80% of all breast cancer cases, and invasive lobular carcinoma (ILC) accounting for 10 - 15% of all breast cancers [26,29,30]. Other histological types of breast cancer, comprising inflammatory, medullary, mucinous, tubular and other carcinoma, are very rare and are all classified as breast cancer of *special types* [26,29].

Despite the fact that classification of breast cancer in ductal or lobular is still largely applied, we now know that this terminology is purely descriptive and does not refer to a particular site or cellular type of cancer initiation. A large-scale histological examination indicates that most breast tumors arise at the junction between the terminal duct and lobule, in the TDLUs [31] (Figure 1). All precursor lesions, including DCIS, are thought to arise in enlarged lobules that have been termed 'atypical lobules' (AL) by Welling *et al.* [32], and 'hyperplastic enlarged lobular units' (HELUs) by Allred and colleagues [33].

2.4.2 Inter-tumor heterogeneity

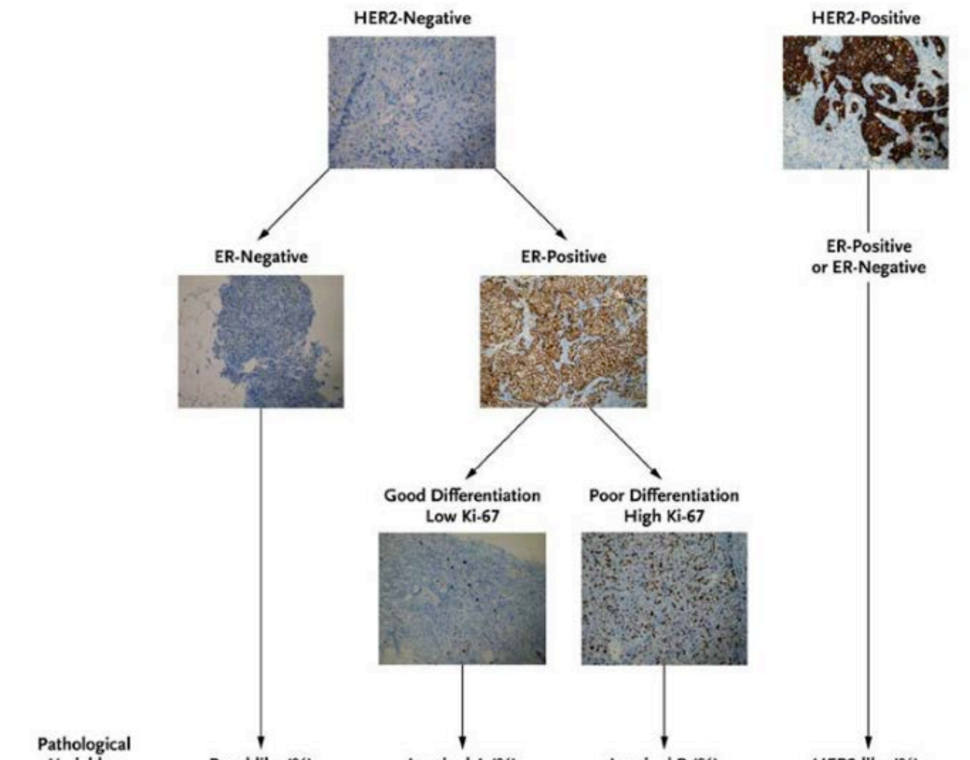
Recent advances in human genome research and high-throughput molecular technologies revealed that the biological and clinical heterogeneity of breast cancer is the result of a concurrent molecular heterogeneity. Since the advent of microarrays for genome-wide expression analyses, two main approaches have been applied to the study of breast cancer:

- molecular-subtype identification studies aimed at establishing different subgroups within a mixed population of patients;
- prognostic/predictive gene signature studies designed to identify cocktails of genes able to predict recurrence and metastasis risk in subgroups of patients and their response to therapies.

2.4.2.1 *Molecular classification*

In a seminal study that analyzed expression profiles of primary breast tumors, Perou *et al.,* showed that breast cancers can be classified according to their patterns of gene expression, which are intrinsically related to tumor biology and behavior [34,35]. By unsupervised hierarchical cluster analysis they were able to identify four main molecular classes ("intrinsic subtypes") of breast cancer that have been validated in several subsequent studies [36-38]. The intrinsic classification divides breast cancer into 4 main groups, namely, Luminal-A, Luminal-B, HER2 and Basal. In addition, other less defined subtypes such as the Normal-like, Apocrine and Claudine-low subtypes have been identified.

Although molecular taxonomy of breast cancer has generated a lot of interest and the expectation that it could result in dramatic improvements in patient management, to date, the practical application of the molecular classification of breast cancers has been limited. The main critical issue is that the molecular classification of breast cancers reflects their classical classification obtained by immunohistochemical (IHC) evaluation of HR/ErbB2 status and proliferation status or histologic grade: Basal-like breast cancers mostly correspond to ER-negative, (PR)–negative, and ErbB2-negative tumors, hence, “triple-negative” tumors; luminal-A cancers are mostly ER-positive and histologically low-grade; luminal-B cancers are also mostly ER-positive, but may express low levels of HRs and are often high-grade; and HER2 cancers show amplification and high expression of the *ERBB2* gene and several other genes of the *ERBB2* amplicon [39] (Figure 2). Nevertheless, the huge knowledge derived from wide molecular dissections of breast cancers can help in identifying new driver molecular alterations and, therefore, new therapeutic targets.



Pathological Variables	Basal-like (%)	Luminal A (%)	Luminal B (%)	HER2-like (%)
HER2-positive (IHC)	10	12	20	100
ER-positive (IHC)	12	96	97	46
Grade III	84	19	53	74
Tumor size >2 cm	75	53	69	74
Node-positive	40	52	65	66

Figure 2. Correspondence between molecular class and clinico-pathological features of breast cancer.

HER2 status, estrogen receptor (ER) status, and proliferation index (Ki-67) were determined by immunohistochemical (IHC) analysis in breast tumor samples. . Figure taken from [39].

CGH (Comparative Genomic Hybridization) is another application of the microarray technique, which permits the assessment of genomic copy number variations among samples. Studies of CGH data showed that breast cancer could be divided into different groups based on their genomic alterations. Three main groups were identified: i) tumors with few rearrangements; ii) tumors with complex alterations; iii) tumors with tightly packed, high-level amplicons. These patterns of alterations can be objectively quantified and can give prognostic information [40].

A very recent study, performed by Curtis *et al.*, presented a genome-wide integrated analysis of copy number and gene expression of a very large cohort of human primary breast tumors composed of 997 discovery and 995 validation cases. By unsupervised analyses of paired DNA-RNA profiles they were able to identify novel molecular subgroups of breast cancer, splitting the whole population in ten clusters with different clinical outcomes [41]. This study provides a novel molecular stratification of breast cancer patients and more importantly permits the identification of novel putative "driver" cancer genes.

2.4.2.1.1 Prognostic/Predictive gene signatures

Predicting outcome or response to therapy using genes differentially expressed in predefined groups of tumors, such as clinical trial cohorts, is an approach first pioneered by Van't Veer *et al.* in 2002 [42]. So far, many gene signature predictors have been generated and validated in specific cohorts of patients, but only a few of them have been subjected to rigorous assay standardization, quality control and clinical validation. The best-known validated assays are Mammaprint (Agendia) and Oncotype DX (Genomic Health).

The Mammaprint prognostic gene signature has been generated from a selected group of 78 patients with node-negative breast cancer who had received no systemic adjuvant therapy [43]. The assay, which measures the expression of 70 genes identified from an initially unselected set of more than 25,000 genes, calculates a prognostic score that classifies patients into poor or good risk groups. The Food and Drug Administration (FDA) has approved this test for use in lymph node-negative breast cancer patients, less than 61 year old and with tumors smaller than 5 cm in size [43]. A comparison of the Mammaprint gene signature with the "Adiuvant!Online" program, which assigns risk according to conventional criteria of tumor size, nodal status, grade, and ER status, showed that 29 percent of patients had discordant results, with the gene signature appearing to be more accurate in these cases [44].

The Oncotype DX assay circumvents the limit of microarray-based tests, such as the Mammaprint assay, that require fresh or frozen tissue and have a limited clinical applicability. Indeed, the Oncotype DX-Assay takes advantage of a real-time RT-PCR method to quantify gene expression in sections of fixed, paraffin embedded tumor tissue. The assay measures the expression of 21 genes (16 cancer-related and 5 control genes) selected from an initial list of 250 candidate genes chosen from published literature and databases of gene expression. A mathematical algorithm calculates a recurrence score (RS), which can be used as a continuous variable to estimate the probability of recurrence at 10 years or to assign patients into low, intermediate or high-risk groups [45]. The Oncotype DX assay has been endorsed by the American Society of Clinical Oncology (ASCO) as a tumor marker and as a clinical tool in decision making about the administration of adjuvant chemotherapy in patients with ER-positive, node-negative breast cancer [46].

Despite the fact that the majority of treatment recommendations are still made without using these tests, in cases where the measure of clinical risk is equivocal (e.g., intermediate expression of the ER and intermediate histologic grade, or high grade, but small tumor size and node-negative cases) these assays could guide clinical decisions. It remains to be seen, however, whether more robust and simpler methods based on IHC will provide comparable information and be more suitable for routine clinical practice.

2.4.3 Intratumor Heterogeneity

Determination of intra-tumor heterogeneity at high resolution is complicated. Using microarray techniques a lot of information regarding physical rearrangements, such as fusion genes or disrupted genomic elements, is lost. Another caveat in the use of microarray methods is that they require large quantities of input DNA and thus the information they provide is limited to the average copy number alterations in a bulk cell population.

Very recent advances in NGS systems have allowed the characterization of the full spectrum of mutations present in a cancer genome and to define a more detailed genomic architectural pattern. The complete sequencing of the human genome was achieved in a period of ten years by using “first-generation” sequencing methods based on the dideoxynucleotide termination chemistry. Today, NGS systems are capable of sequencing a human genome in about one week and provide detailed knowledge of a cancer genome including point mutations, copy number variations and genetic aberrations (deletions, amplifications, inversions or translocations) [47].

NGS cannot resolve the combinations of mutations present in a heterogeneous sample, but it can measure the distribution of allele frequencies it contains. By addressing this feature, in a deep-sequencing study of a basal-like breast cancer, its brain metastasis and a xenograft obtained from the same primary tumor, Ding *et al.*, in a seminal study of potentially great clinical impact, deciphered the clonal relationships between a primary tumor and its synchronous metastasis. The authors showed that few *de novo* mutations appeared in the metastasis compared to the primary tumor, but gross changes in allelic frequencies were observed, and that a significant overlap in the spectra of mutations was present among the metastasis and the xenograft tumor [48]. In a similar study, Shah *et al*. compared mutations of a primary lobular breast cancer with those from its metastasis that developed 12 years later. In this case, the authors identified only 6 mutations in common between the primary tumor and its metastasis, which were present at low frequency in the primary tumor[49].

These studies suggest that minor subpopulations of cells with metastatic potential are pre-existing in the primary tumor and that their spectra of mutations could be predicted by experimental *in vivo* modeling systems. An increasing number of high-resolution/high-throughput studies on cancer genomes are now being performed and much effort is needed to analyze and decipher the huge amounts of data created. It is expected that the enormous amount of information produced will guide the personalized treatment of breast cancer patients by identifying the unique spectrum of mutations that define individual tumors and tumor subpopulations, as well as mutations that are likely to drive metastasis.

2.4.4 Sources of inter-tumor and intra-tumor heterogeneity

Inter- and intra-tumor heterogeneity are relevant to the processes of tumor initiation and progression. Poor knowledge about these events, especially in breast cancer, has generated heated debates in the scientific field and diverging hypotheses.

Inter-tumor heterogeneity can be explained by two different hypotheses. The genetic model proposes that different initiating events in the same cell of origin will lead to different molecular subtypes of cancer. In contrast, the second model, proposes that different cells of origin lead to different subtypes. However, a combination of the two models is also plausible in which different subtypes come from both different cells of origin and different initiating events [50].

Intra-tumor heterogeneity is mainly viewed as a problem of tumor evolution. Also in this case different models have been proposed that are supported by clinical and experimental observations. The classical model of tumor evolution, which was proposed for the first time by Nowell in 1976, is the clonal evolution model. This model supports both a monoclonal and polyclonal evolution of the tumor, in which cells undergo a Darwinian-like selection where one or more cells, respectively, acquire a proliferative advantage and expand to form the tumor mass [51]. A recent study, by Navin *et al*., supports the existence of both monoclonal and polyclonal tumors. In this study the authors developed a system to analyze genomic heterogeneity at single-cell level. The authors isolated single nuclei by FACS and DNA-content profiling from distinct segments of the tumor, and performed single-nucleus sequencing by NGS. This experimental approach revealed the existence of both monogenomic tumors, in which all cells within the tumor share a similar DNA copy number and spectrum of mutations, and polygenomic tumors, in which single cells within the tumor carry different DNA copy numbers and different spectra of mutations. Furthermore, this analysis revealed that single clones could be either topologically restricted to a single sector of the tumor mass or be sparsely disseminated throughout the entire tumor mass [52].

An alternative explanation of intratumor heterogeneity is provided by the cancer stem cell model. This model proposes a hierarchical organization in which a rare population of cancer stem cells proliferates indefinitely, while the majority of tumor cells have limited proliferation. In this model, intratumor heterogeneity is explained by the ability of different precursor cells to give rise to different cell subpopulations within the tumor [53].

Another model explaining intra-tumor heterogeneity is the mutator phenotype. This model proposes that tumors evolve by gradual and random accumulation of mutations, generating the presence of a large diversity of small clones rather than few dominant clonal subpopulations [54] (Figure 3)

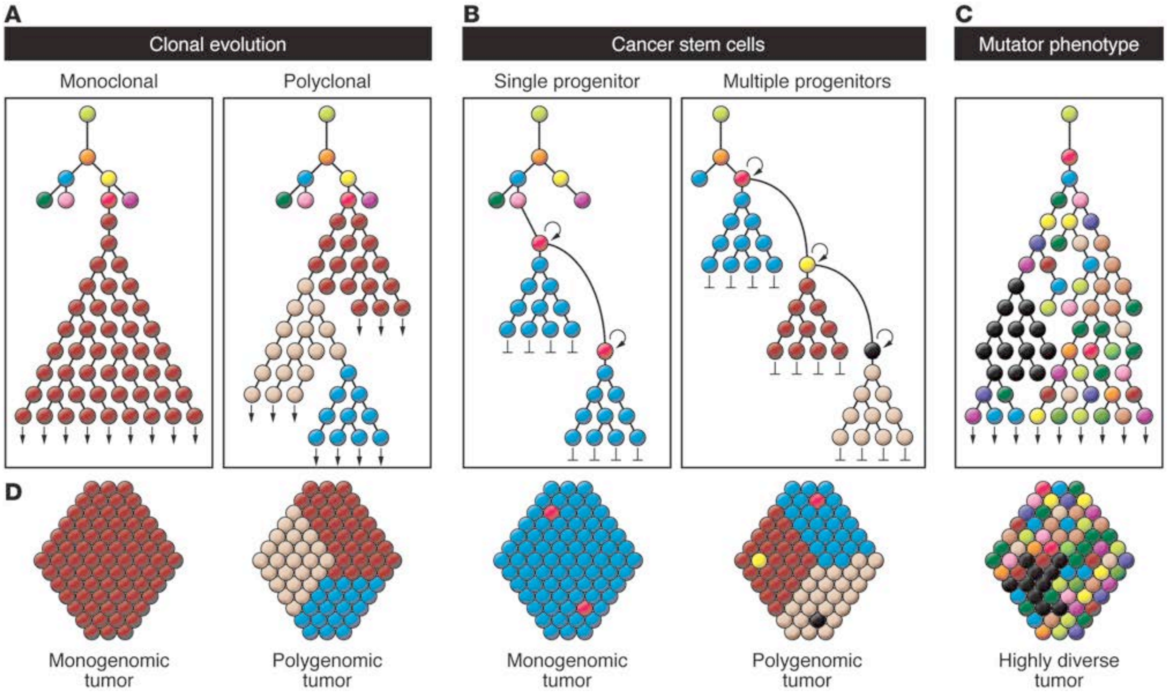


Figure 3. Hypothetical models explaining intra-tumor heterogeneity

Different models of tumor progression can give rise to distinct types of intra-tumor heterogeneity, exemplified here by (A) the clonal evolution, (B) the cancer stem cell, (C) and the mutator phenotype models. (D) The different models can result in distinct spatial distributions of tumor cell subpopulations. Figure taken from [55].

2.5 Molecular pathogenesis of breast cancer

Breast carcinogenesis involves a series of progressive changes that accumulate in the stepwise acquisition, by breast epithelial cells, of the so-called “hallmarks of cancer”: i.e., genome instability, sustained proliferative signaling, evasion of growth suppressors, resistance to cell death, replicative immortality, induction of angiogenesis, activation of invasion and metastasis, reprogramming of energy metabolism, and evasion of immune destruction[56](Figure 4). These acquired abilities, which drive and sustain cancer growth and metastasis, reflect the accumulation of genetic changes that are mainly categorized in two classes: 1) loss of function of tumor suppressor genes and 2) gain of function of oncogenes.

An oncogene is a mutated and/or overexpressed gene (defined proto-oncogene in its wild-type form) that alone, or in collaboration with other changes, promotes cellular transformation, growth and invasion. In contrast, a tumor suppressor gene normally counteracts cell growth or other processes that may increase invasive and metastatic potential and whose loss of function promotes malignancy. In addition to protein-coding genes, in the last decade the importance of non-coding RNAs and their involvement in tumorigenesis has been documented, together with their ability to work as oncogenes or tumor suppressor genes [57].

The most frequently activated and best characterized oncogenes in breast cancer include v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (*ERBB2*), phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PI3KCA)*, v-myc avian myelocytomatosis viral oncogene homolog (*MYC*), and cyclin D1 (*CCND1*). In contrast, the tumor suppresor protein p53 gene (*TP53*), the breast cancer susceptibility genes 1 and 2 (*BRCA1 and BRCA2*), the phosphatase and tensin homolog gene (*PTEN*), the E-cadherin gene *(CDH1),* the retinoblastoma gene (*RB1*) and members of the cyclin-dependent kinase inhibitor (CKI) family represent the most frequently altered tumor suppressor genes in breast cancer. Undoubtedly, many more oncogenes and tumor suppressor genes contribute to breast carcinogenesis. Given the heterogeneity of breast cancers, a better understanding of the genetic lesions that drive tumorigenesis in the mammary gland, will lead to improvements in the clinical management of breast cancer patients.

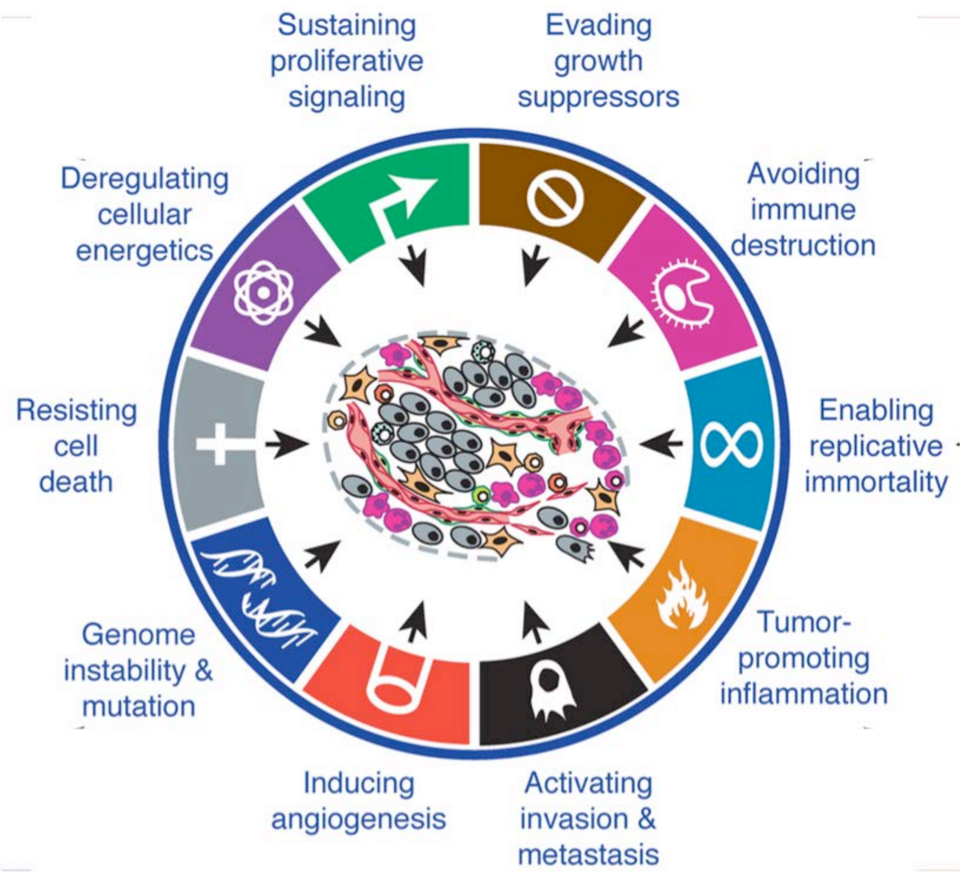


Figure 4. The Hallmarks of Cancer

Schematic diagram of the acquired capabilities necessary for tumor growth and progression. Figure adapted from [56].

2.5.1 ONCOGENES

2.5.1.1 The "case" of the 17q12-q21 amplicon in breast cancer

DNA amplification is a common event in cancer that often involves many genes and large segments of DNA defined as "amplicons". In breast cancer, one of the most frequently amplified chromosomal regions is 17q12-q21, which is particularly rich in genes with known or potential cancer relevance [41,58] (summarized in Table 1). One such gene is the well-characterized oncogene *ERBB2* that is overexpressed in 10 – 35% of human breast cancers (see below, section 2.5.1.2). ErbB2-positive breast cancers, despite being classified as a single class of tumors, remain a heterogeneous disease, as underlined by variable patient outcomes and responses to targeted therapy. This heterogeneity, is likely attributable to differences in the genes that are co-amplified with *ERBB2*. Interestingly, Morrison *et al.* found that the response rate to the ErbB2-targeted therapy, trastuzumab, is inversely correlated with the size of the *ERBB2*-amplicon [59]. This observation suggests that several genes distally co-amplified with *ERBB2* might contribute to the aggressive phenotype of ErbB2-positive tumors.

In an attempt to identify such co-amplified genes that might have diagnostic, prognostic and therapeutic relevance to breast cancer, several studies have sequenced the commonly amplified regions surrounding the *ERBB2* locus in breast cancer cells. Luoh *et al.* in 2002, by Southern Blot analysis, identified the core sequence at the *ERBB2* locus commonly amplified in different breast cancer cell lines [60]. The region identified was rather large, spanning hundreds of kbs in size (>300 kb) from *TRAP220* (*PPARBP*) to *TRAP100* (*THRAP4*) and including 13 genes (Table 1). The authors demonstrated that amplified genes from this region were also overexpressed at the protein level. Among these genes *CRK7* (*CDK12*), a STK located more than 200 kb from *ERBB2,* was found to be amplified and overexpressed in breast cancer cell lines [60].

Kauraniemi *et al.,* by fluorescence *in situ* hybridization (FISH), using both a panel of 16 breast cancer cell lines [61] and a large set of 330 primary breast tumors [62], characterized a minimal common *ERBB2* amplification region of 280 kbs, which is included in the region previously identified by Luoh *et al.* [60]. This region contains ten genes, from *NEUROD2* to *ZNFN1A3* (Table 1). However, this amplicon-mapping study did not identify in the minimal *ERBB2*-amplicon most of the previously reported co-amplified and co-expressed genes, such as *THRA, RARA*, *TOP2A*, *MLN62* (*TRAF4*), *MLN50* (*LASP1*), *PSMB3* and *RPL19* (Table 1). More recently Sircoulomb *et al.,* by comparative genomic hybridization (CGH) microarray analysis of 54 *ERBB2*-amplified breast cancers identified the *ERBB2-C17orf37-GRB7* genomic segment as the minimal common 17q12-q21 amplicon, and *CRK7 (CDK12)* and *ZNFN1A3* (*IKZF3)* as the most frequent centromeric and telomeric amplicon borders, respectively [63]. However, as reported by Kauraniemi *et al.* [64], in most tumors the amplification spans beyond the minimal common region. Therefore, we cannot exclude, in a subset of ErbB2-positive tumors, the functional contribution of *ERBB2* distally co-amplified genes to the aggressive phenotype of these tumors and to trastuzumab therapy resistance.

A few of the *ERBB2* co-amplified genes have been studied for their oncogenic potential or ability to confer drug resistance in breast cancer. Targeted knockdown of the co-amplified genes *GRB7* and *STARD3* decreased proliferation of breast cancer cell lines harboring amplification/overexpression of the protein products of these genes [65]. However, combined targeting of *ERBB2/GRB7* or *ERBB2/STARD3* had no additive effect on reducing proliferation beyond targeting *ERBB2* alone [65]. In addition, *C35* (*C17orf37*) has been shown to function as an oncogene in breast cancer cell lines [66], while *MED1* (*PPARBP*) has been shown to have a role in conferring resistance to tamoxifen treatment of ErbB2-positive tumors [67]. Together, these studies indicate that ErbB2 is not the sole oncogene in the 17q12-q21 amplicon and underline the importance of other *ERBB2* co-amplified genes in exacerbating the complexity and aggressiveness of ErbB2-positive tumors. Of the outmost relevance to this thesis work, it is worth mentioning in this context that CDK12 is among the genes of the 17q12-q21 amplicon that is frequently co-amplified with *ERBB2*, although its function has been poorly characterized and its potential implication in breast carcinogenesis remains elusive. In more general terms, likewise CDK12, the elucidation of the contribution of other genes in the 17q12-q21 amplicon to breast tumorigenesis, alone or in combination, requires additional functional studies.

Table 1. List of genes in the 17q12–q21 region

Symbol	Alias	Description	Start bp position
TRAF4	**MLN62**	TNF receptor-associated factor 4	24095173
TIAF1		TGFB1-induced anti-apoptotic factor 1	24424665
MLLT6		Myeloid/lymphoid or mixed-lineage leukemia; translocated to, 6	34115412
LOC284106		Hypothetical protein LOC284106	34140014
PCGF2	**ZNF144**	Polycomb group ring finger 2	34143676
PSMB3		Proteasome (prosome, macropain) subunit, β type, 3	34162528
PIP5K2B		Phosphatidylinositol-4-phosphate 5-kinase, type II, β	34177324
RPL23		Ribosomal protein L23	34259865
LASP1	**MLN50**	LIM and SH3 protein 1	34279894
FLJ43826		FLJ43826 protein	34439685
PLXDC1		Plexin domain containing 1	34473083
ARL12		ADP-ribosylation factor-like 12	34561546
CACNB1		Calcium channel, voltage-dependent, β1 subunit	34583235
RPL19		Ribosomal protein L19	34610096
STAC2		SH3 and cysteine rich domain 2	34620316
FBXL20		F-box and leucine-rich repeat protein 20	34670367
PPARBP	**TRAP220**	PPAR-binding protein	34816380
CRK7		CDC2-related protein kinase 7	34871818
NEUROD2		Neurogenic differentiation 2	35014575
PPP1R1B	**DARPP-32**	Protein phosphatase 1, regulatory (inhibitor) subunit 1B	35036705
STARD3	**MLN64**	START domain containing 3	35046938
TCAP		Titin-cap (telethonin)	35075125
PNMT		Phenylethanolamine N-methyltransferase	35078033
PERLD1	**MGC9753**	per1-like domain containing 1	35080902
ERBB2		v-erb-b2 erythroblastic leukemia viral oncogene homolog 2	35109922
C17orf37	**MGC14832**	Chromosome 17 open reading frame 37	35138937
GRB7		Growth factor receptor-bound protein 7	35147744
ZNFN1A3		Zinc finger protein, subfamily 1A, 3 (Aiolos)	35174724
ZPBP2		Zona pellucida binding protein 2	35277995
GSDML		Gasdermin-like	35314376
ORMDL3		ORM1-like 3	35330822
GSDM1		Gasdermin 1	35372752
PSMD3		Proteasome (prosome, macropain) 26S subunit, non-ATPase, 3	35390586
CSF3		Colony stimulating factor 3 (granulocyte)	35425214
THRAP4	**TRAP100**	Thyroid hormone receptor associated protein 4	35428877
THRA	**c-erbA**	Thyroid hormone receptor, α	35472589
NR1D1		Nuclear receptor subfamily 1, group D, member 1	35502567
CASC3	**MLN51**	Cancer susceptibility candidate 3	35550100
RAPGEFL1		Rap guanine nucleotide exchange factor (GEF)-like 1	35591394
WIRE		WIRE protein	35666238
CDC6		CDC6 cell division cycle 6 homolog	35697672
RARA		Retinoic acid receptor, α	35740896
GJC1		Gap junction protein, χ 1, 31.9kDa (connexin 31.9)	35770433
TOP2A		Topoisomerase (DNA) II α 170kDa	35798321

Positional information derived from http://www.ncbi.nlm.nih.gov/genome. Dashed lines indicate the minimal common region of amplification defined by Kaurianemi *et al*. Distal genes that have been reported to be co-amplified with *ERBB2* are also reported in the table. Table taken from [64].

2.5.1.2 ERBB2

ErbB2 is a member of the epidermal growth factor receptor (EGFR) family of receptor tyrosine kinase (RTKs). This family comprises four closely related members: EGFR (EGFR1), ErbB2 (Neu, HER-2), ErbB3 (HER-3) and ErbB4 (HER-4) [68]. With the exception of ErbB2 itself, which remains an orphan receptor with no diffusible ErbB2-specific ligand, growth factor receptor activation is initiated by binding to specific extracellular ligands, which induce receptor homo- or hetero-dimerization and autophosphorylation, which in turn leads to activation of multiple transduction cascades including the mitogen-activated protein (MAP) kinase and the PI3K/AKT pathway. Aberrant RTK activity has an oncogenic role in the tumorigenic process being involved in many cellular processes including cell proliferation, angiogenesis, cell-cell interactions, cell motility, metastasis, and resistance to apoptosis [69].

ERBB2 is one of the most intensively studied genes in cancer. The *ERBB2* gene is located on chromosome arm 17q21.1 and genomic amplification of this locus occurs in 10 – 35% of human breast cancers [60,64]. Slamon *et al.*, in 1987 identified *ERBB2* as a biomarker of poor prognosis in breast cancer [70]. This finding has been confirmed and extended in many large-scale studies. Recently in a meta-analysis study summarizing data from 81 studies (27,161 patients), both *ERBB2* amplification and protein overexpression have been confirmed to be strong independent prognostic factors [11]. The majority of these studies also found a strong correlation between *ERBB2* amplification/overexpression and resistance to tamoxifen therapy and sensitivity to anthracycline treatment [11].

The discovery of *ERBB2* gene amplification/overexpression in primary human breast cancer and its association with a more aggressive clinical behavior spurred the development of molecularly targeted therapies. Trastuzumab (Herceptin), a humanized recombinant monoclonal antibody directed against the extracellular portion of the ErbB2 protein [71], is one of the first successful examples of targeted therapy. Nowadays, trastuzumab, in combination with chemotherapy, is the treatment of choice for ErbB2-positive breast cancer patients. However, despite the good response rate, 70% of metastatic breast cancer patients with overexpressed ErbB2 have primary resistance to trastuzumab, and the majority of responders acquire secondary resistance within one or two years [25]. Several mechanisms have been postulated to explain primary and acquired resistance and several alternative drugs (comprising small molecule inhibitors) have been developed and are currently being tested in clinical trials [25].

2.5.1.3 PI3KCA

An important RTK downstream signaling pathway that is often deregulated in breast cancer is the PI3K/AKT/mTOR pathway. This pathway is activated in response to external stimuli and regulates several cellular functions, such as cell growth, metabolism, survival and proliferation [72]. Uncontrolled activation of this pathway by alteration of any of its components is capable of driving cell transformation and malignant progression and occurs very frequently in cancer. Indeed, mutations in *PIK3CA*, the catalytic subunit of PI3K, occur in ~36% of all breast cancers and its locus is also frequently amplified [73]. Loss of PTEN, a tumor suppressor that antagonizes PI3K activity, is also very frequent in breast cancer [74] and leads to hyperactivation of the PI3K pathway.

AKT is a STK that operates as a downstream effector in the PI3K pathway, mediating cell proliferation, survival and metabolism [72]. Amplification or activating mutations of the *AKT* gene have been rarely observed in breast cancer, however, it is often hyper-activated due to activating mutations in upstream components of the pathway. Another important downstream molecule in the PI3K pathway is mTOR, a STK, which functions as a master regulator of protein translation. Hyper-activation of mTOR is a frequent event in breast cancer and leads to the accumulation of cell cycle regulatory proteins, such as cyclin D1, and enhances AKT activation through a positive feedback loop [75,76].

Since aberrant activation of the PI3K pathway is very frequent in breast cancer and plays a crucial role in tumor progression and drug resistance, much effort has been placed on the development of specific drugs targeting components of this pathway. Inhibitors targeting PIK3CA, such as wortmannin and LY294002, have been tested in preclinical studies, but poor solubility, instability and high toxicity have limited their clinical applications [25]. Perifosine, an AKT inhibitor that prevents its recruitment to the membrane and subsequent activation, has been tested in clinical studies, however, low response rates were observed [25]. Rapamycin, a mTOR inhibitor, was shown to have potent anti-proliferative activity in cancer cell lines, but was ineffective *in vivo* due to its low solubility [25]. Rapamycin analogues with increased solubility, such as Temsirolimus (Torisel), have however been developed and are now approved by the Food and Drug Administration (FDA) for the treatment of metastatic renal cell carcinoma [25]. mTOR inhibitors are currently under investigation in clinical trials for the treatment of breast cancer [25].

2.5.1.4 CCND1

CCND1 encodes the cyclin D1 cell cycle protein, which forms an active complex with CDK4 and is responsible for the Rb phosphorylation and cell cycle progression through G1-S phases. Its overexpression increases proliferation of cancer cells. Cyclin D1 has been found to be overexpressed in 40 – 50% of human breast cancers and its gene, which is located on chromosome 11q13, is amplified in 10 – 20% of breast cancer cases [77].

2.5.1.5 MYC

Myc is a basic helix-loop-helix zipper (bHLHZ) motif–containing transcription factor, whose activity is tightly regulated by its direct binding to another bHLHZ protein Max that works as a Myc coactivator. Myc activity is precisely controlled by the activity of multiple competing repressive Max binding partners, such as Mad. Myc activation leads to transcriptional activation or repression of specific genes. The global transcriptional influence of Myc has effects on multiple pathways including those involved in cell growth, cell proliferation, metabolism, microRNA regulation, cell death, and cell survival [78,79]. The *MYC* gene located on chromosome 8q24 is amplified and overexpressed in 15 – 25% of breast tumors[80].

2.5.2 TUMOR SUPPRESSOR GENES IN BREAST CANCER

2.5.2.1 TP53

The tumor suppressor gene *TP53*, located on 17p13.1, is the most frequently mutated tumor suppressor in human tumors. More than 50% of all cancer cases carry *TP53* mutations. This frequency is slightly lower in breast cancer with 15 – 34% of cases harboring *TP53* mutations [81].

TP53 has been defined as the "guardian of the genome" [82]. When normal cells are damaged by ionizing radiation or mutagens, the protein encoded by this gene, p53, is activated by phosphorylation and accumulates in the nucleus where it binds as a tetramer to specific sequences to activate transcription of target genes. These target genes mediate the pleiotropic effects of p53 on cellular homeostasis, including cell cycle checkpoint activation, DNA repair, cell migration, cell metabolism, cellular senescence, apoptosis and autophagy [55].

One of the target genes induced by p53 is p21, an inhibitor of cyclin-dependent kinases (CDKs) that causes cell cycle arrest. GADD45, another p53 target gene, is in charge of repairing damaged DNA. When repair is successful p53 is degraded by the action of the ubiquitin-ligase MDM2 and the cell cycle restarts. When GADD45 cannot repair the genome because of excessive DNA damage, p53 trans-activates the pro-apoptotic gene BAX, which induces apoptosis [83]. p53 dysfunction/inactivation allows cells to survive DNA damaging insults, which may result in the accumulation of activating mutations in proto-oncogenes or inactivating mutations in tumor suppressor genes and, consequently, to malignant transformation.

2.5.2.2 BRCA1 and BRCA2

BRCA1 and BRCA2 are known as the "caretakers" of genome stability maintenance [84]. These proteins share some functional similarities and play an important role in the repair of DNA double strand breaks (DSBs), particularly in the process of homologous recombination. Several studies showed that BRCA1 and BRCA2 interact with and regulate the activity of Rad51, a key enzyme in the homologous recombination process [85-87].

In contrast to BRCA2, BRCA1 may also be involved in additional DNA repair mechanisms such as non-homologous end joining (NHEJ) and nucleotide excision repair (NER) [25]. BRCA1 also participates in the regulation of other cellular processes, such as cell cycle control, gene transcription regulation and apoptosis regulation [84,88].

Genome instability is an early event in the process of cell transformation that is often acquired, in the case of breast cancer, by inherited or somatic mutations of *BRCA1* and *BRCA2* that result in a greater potential to accumulate genetic alterations [89]. Hundreds of different mutations have been detected in *BRCA1* and *BRCA2* genes in breast cancer with most of them causing premature truncation of the proteins [90]. Although *BRCA1* and *BRCA2* mutations appear to be quite rare in sporadic breast cancer cases, germ-line mutations of these genes predispose to breast cancer and appear to be present in over 80% of familial breast cancer cases [84].

2.5.2.3 E-cadherin

E-cadherin is a calcium dependent cell-adhesion molecule with a well-defined role in cell-cell adhesion, signal transduction and epithelial differentiation [91]. Loss of function of E-cadherin seems to facilitate the malignant invasive potential of breast cancer cells. The E-cadherin gene, *CDH1*, is located on 16q22.1, a region frequently deleted in breast tumors [92]. *CDH1* has been found to be commonly mutated in breast cancers of the lobular histological subtype resulting in decreased expression of this gene at both the mRNA and protein levels [73].

2.5.2.4 Retinoblastoma

Retinoblastoma (*RB1*) was the first tumor suppressor gene to be identified. The Rb protein protects against tumorigenesis by regulating cell cycle progression, cellular senescence, differentiation, apoptosis and chromosomal integrity [93]. The main role of Rb is the regulation of the G1-S transition by interaction and repression of the E2F transcription factors that mediate expression of genes required for cell cycle progression [94]. Specific cyclin D1-CDK4 and cyclin E-CDK2/6 complexes phosphorylate Rb, leading to dissociation and release of E2F, which in turn activates the expression of target genes responsible for cell cycle progression.

The *RB1* gene was originally identified in retinoblastoma [95]. Later studies showed that changes of this gene occur in over half of human malignancies. In breast cancer, mutation or loss of *RB1* is present in up to 30% of cases [73,96]. Moreover, deregulation of the Rb pathway may occur by dysfunction of Rb itself or by deregulation of other components of the pathway, such as loss of function of CDK inhibitors or amplification and overexpression of cyclin D1 [97]

2.5.2.5 CKIs

CKIs are negative regulators of CDKs and cell cycle progression, and can be categorized as the tumor suppressor genes. Some members of the family have also been associated with other cellular functions including apoptosis, senescence, transcription or cell migration [98]. CKIs are divided into two families including the INK4 family (p16, P15, p18, p14), which inhibit CDK4 and CDK6 to prevent G1-S transition and the Cip/Kip family (p21, p27, p57), whose members are able to inhibit all CDKs.

p27, a member of the Cip/Kip family, has been characterized as a tumor suppressor. p27 is capable of binding to several cyclin/CDK complexes thereby inhibiting their activity and typically causing a cell cycle block in the G1 phase. Many roles, in addition to cell cycle regulation, have been proposed for p27, including modulation of drug resistance, cell differentiation and protection from inflammation [90]. Its expression has been found to be downregulated in a wide range of human cancers including breast cancer. However, mutations of the p27 gene, *CDKN1B*, seem to be an uncommon event in breast cancer (1% of cases) [73] suggesting that its downregulation is mainly due to epigenetic or other upstream genetic alterations.

2.5.2.6 PTEN

The *PTEN* gene codifies for a protein and lipid phosphatase that serves as a negative regulator of AKT. Loss of PTEN results in reduced dephosphorylation of phosphatidylinositol (3,4,5)-triphosphate (PIP3), which allows PI3K to phosphorylate phosphatidylinositol (4,5)-biphosphate (PIP2) and further enhance levels of PIP3. PIP3 induction causes increased cell proliferation and cell migration, cell survival and cell size through activation of downstream proteins such as Akt [99]. PTEN mutations are found predominantly in advanced glial and prostate tumors. However, mutations of PTEN have been detected in breast cancer with a low frequency (3%) [73] and germline mutations have been shown to increase the risk of breast and ovarian cancer[100].

2.6 CDK12

As mentioned above, one of the genes contained within the 17q12-q21 amplicon that is frequently co-amplified with *ERBB2* is *CDK12* (formerly known as CRKRS or CRK7).

2.6.1 Classification and structure

CDK12 (alias CRKRS, CRK7, CRKR, KIAA0904) was isolated and originally identified as a Cdc2-related STK possessing an arginine/serine (RS)-rich domain, which was closely related to the family of CDKs [101]. CDK12 was later confirmed to be a CDK after Chen *et al.* reported its interaction with cyclins L1 and L2 (CycL) [102]. However, in this study the CycL-CDK12 interaction was only observed with exogenously expressed CycL. Recent studies have now identified cyclin K (CycK), but not CycL, as the endogenous binding CDK12 partner, both in *Drosophila melanogaster* and in humans[103,104].

CDK12 is highly conserved throughout evolution. The closest orthologues to the human protein are the gene products of *CG7597* in *Drosophila melanogaster* and *B0285* in *Caenorhabditis elegans,* which display an overall identity of 41% and 53%, respectively. The closest homologue in humans is CDK13, which displays 89% identity in the central 421 amino acids containing the cdc2-kinase domain, but is completely unrelated outside this sequence [101].

The human CDK12 gene is located on chromosome 17q12. It is composed of 14 exons and codifies two alternatively spliced isoforms that differ only in the last exon. The short isoform (CDK12s) finishes with an in-frame stop codon soon after the skipped 5' splice site of exon 13, and encodes a protein containing 1481 amino acids; the long isoform (CDK12l) splices directly from exon 13 to 14, and encodes a protein comprised of 1490 amino acids (Figure 5A). Both isoforms contain the same functional domains (Figure 5B): a central conserved catalytic serine-threonine kinase domain closely related to the cdc2 family of protein kinases; an amino-terminal RS-rich domain, which is typically found in splicing regulators and, together with the RNA recognition motif (RRM), defines the family of serine/arginine-rich splicing factors (SRSF)[105,106]; two proline-rich motifs (PRMs) that are known to mediate protein-protein interactions by binding to domains, such as SH3, WW or profilin [107,108]; seven potential PEST regions that are associated with protein degradation [109,110]; four bipartite nuclear localization signals (NLS) responsible for the nuclear localization of the protein [101] (Figure 5B).

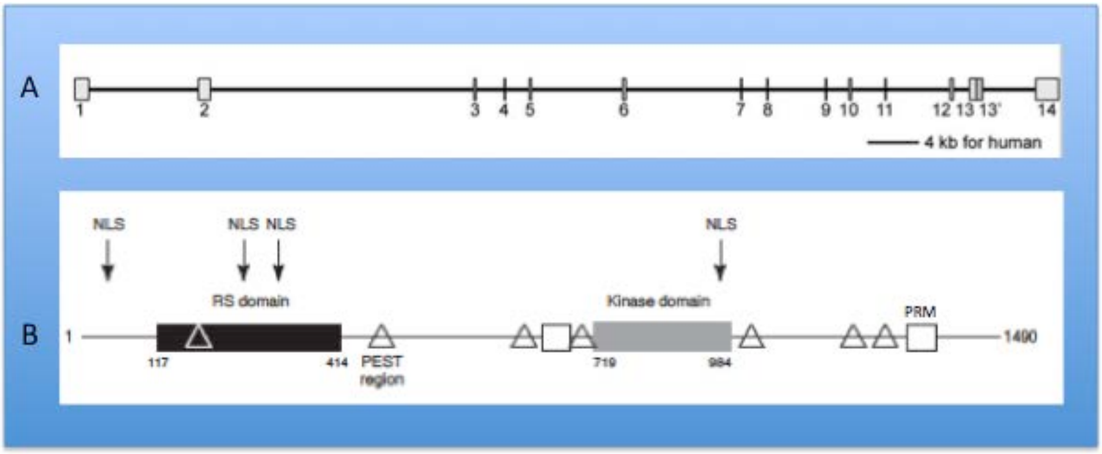


Figure 5. Schematic representation of CDK12 gene and protein structure

A) CDK12 intron/exon genomic organization. (B) CDK12 protein domain structure. NLS: nuclear localization signal; RS domain: arginine/serine-rich domain; PEST region: peptide sequence rich in proline; kinase domain: serine-threonine kinase domain; PRM: proline-rich motif. Figure adapted from [101] and [102].

2.6.2 CDK12 expression, localization and *in vitro* kinase activity

Northern blot analysis revealed that CDK12 is ubiquitously expressed in human tissues. CDK12 localizes exclusively in the nucleus in well-defined speckled structures that usually define the localization of components of the splicing machinery [111].

Consistent with this observation, CDK12 colocalizes with the splicing factor SC35 in immunofluorescence experiments [101](Figure 6).

CDK12 expression does not vary significantly during cell cycle, while differences in its phosphorylation status have been observed between interphase and mitosis [101]. CDK12 has a kinase activity as demonstrated by *in vitro* assays showing the ability of the immunoprecipitated protein to phosphorylate itself and exogenous substrates, including the alternative splicing factor (ASF), the myelin basic protein, and a GST-fusion protein of the RNA polymerase II carboxy-terminal domain (CTD) [101].

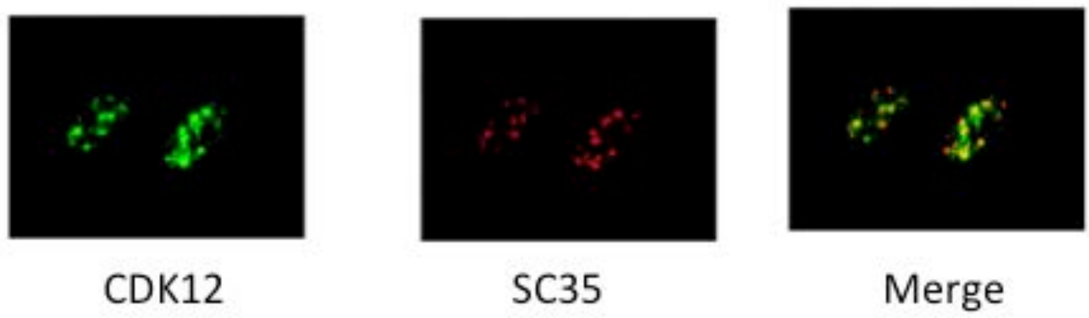


Figure 6 Cellular localization of CDK12

Immunofluorescence images showing the colocalization of CDK12 (CrkRS) with the splicing factor SC35 in nuclear speckles, also defined as "splicing factories". Figure taken from [101]

2.6.3 **Physiological role of CDK12**

CDKs are proteins involved in critical cellular processes. Some CDKs, such as CDK1, CDK2, CDK4 and CDK6, have a role in regulating the cell cycle, while others, such as Cdk7 and Cdk9, are implicated in the regulation of transcription and post-transcriptional mRNA processing [112].

Based on the domain structure of CDK12 and its colocalization with splicing machinery in nuclear speckles, Ko *et al*. proposed that CDK12 could play a role in the regulation of transcription and alternative splicing rather than cell cycle progression[101]. The authors hypothesized that CDK12 could be a novel RNA polymerase II (RNAPII) kinase that might directly link transcription with the splicing machinery.

Chen *et al*. in 2006, provided the first experimental indication of the possible role of CDK12 in the regulation of alternative splicing. These authors found that overexpression or knockdown of CDK12 in human cells altered, in an opposite manner, the splicing pattern of a synthetic E1A minigene [102]. However, these studies were not performed using the physiological cyclin partner of CDK12 (CycK) and further studies are required to clearly define the involvement of CDK12 in the splicing mechanism.

Very recently, two studies by Bartkowiak *et al*. and Blazek *et al*., showed that *Drosophila* and human CDK12 are able to phosphorylate RNAPII in its carboxy-terminal domain (CTD), both *in vitro* and *in vivo* [103,104]. The RNAPII-CTD contains several repeats of the heptapeptide YSPTSPS that can be phosphorylated in each of its serine, tyrosine and threonine residues. The different possible patterns of CTD phosphorylation correlate with the dynamics of RNAPII during transcription: to allow transcription initiation the "formatted" CTD begins to be phosphorylated by CDK7 specifically on Ser5 of the heptapeptides; during transcription elongation, Ser5 residues are gradually dephosphorylated, while Ser2 phosphorylation accumulates; at the end of transcription (termination), Ser2 phosphorylation diminishes facilitating the release of RNAPII, which can then be recruited for a new transcription cycle [113,114]. The positive transcription elongation factor b (P-TEFb) complex, composed of cycT/cycK-CDK9, has long been considered to be the only kinase responsible for CTD-Ser2 phosphorylation [115,116]. However, using a knockdown approach, Bartkowiak *et al*. and Blazek *et al*., have established that CDK12 is another important elongation-associated CTD-Ser2 kinase, which probably acts in complex with CycK [103,104].

Bartkowiak *et al.*, by immunofluorescence and chromatin immunoprecipitation (ChIP) experiments, also showed that CDK12, in *Drosophila melanogaster* cells, mainly localized on active genes and predominantly in the middle and at the end of the transcriptional units [103].

Confirming the involvement of CDK12 in transcriptional regulation, Blazek *et al.*, also found, by gene expression microarray analysis, that CDK12 depletion in human cells, downregulates the expression of mainly long and complex genes [104]. This observation is consistent with the shown kinase activity of CDK12 and suggests a role in stabilizing the RNAPII-DNA interaction, which in turn allows the transcription of long genes. However, more studies are needed to verify this hypothesis and to elucidate the role of CDK12 as a transcriptional regulator.

A very recent publication has confirmed the role of CDK12 as a splicing regulator. In their study, Rodrigues *et al.* studied the regulation of glial-specific splicing of *NeurexinIV*. This gene transcribes two isoforms generated by alternative splicing of two mutually exclusive exons. They found that the sequence-specific RNA binding protein HOW is required for the splicing regulation of *NeurexinIV* and that CDK12 is necessary for the splicing activity mediated by HOW [117]. Based on their study and the literature, Rodrigues *et al.* proposed a model in which CDK12 kinase activity is a determinant for the recruitment of splicing factors and the correct assembly of the splicing machinery on specific pre-mRNA sites defined by HOW (Figure 7).

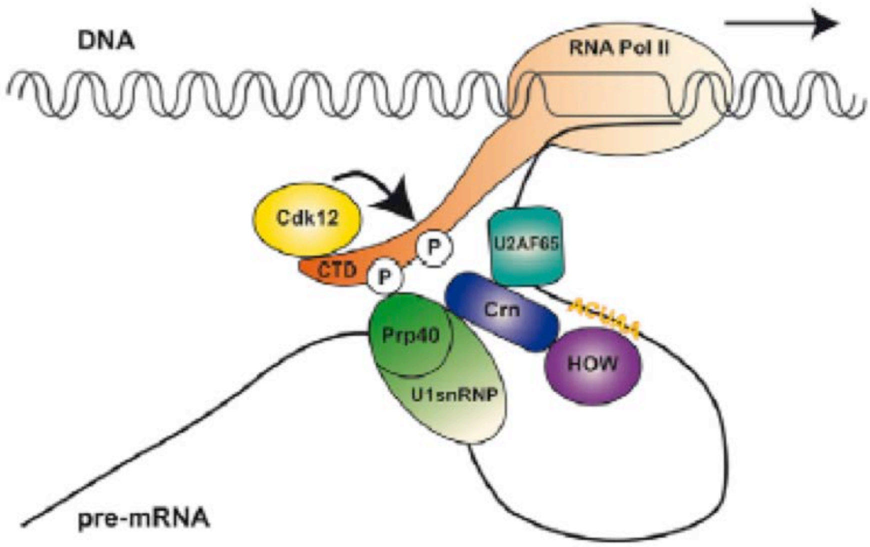


Figure 7. Model of CDK12/HOW dependent splicing

CDK12 binds to and phosphorylates the CTD of RNAPII [103,104] that permits the recruitment of the U1snRNP complex by direct interaction with its protein component Prp40 [118]. Prp40 in turn, interacts with Crn/Clf1, which also binds U2AF65 and the sequence-specific RNA binding protein HOW [119]. Figure taken from [117].

Beyond the mechanistic mode of action of CDK12, the cellular processes in which CDK12 is physiologically or pathologically involved remain another important issue to be addressed. Blazek *et al.* in their microarray analysis, found that CDK12 or CycK depletion affected the expression of the same subset of genes involved in the maintenance of genomic stability [104]. The authors also demonstrated that CycK/CDK12 depletion increases the number of cells in the G2-M phase of the cell cycle and sensitizes cells to the apoptotic effects of DNA damaging agents [104]. This study indicates an indirect role of CDK12 in the cellular process of DNA damage response (DDR) and maintenance of genomic stability by modulating the expression of DDR genes. Nevertheless, further studies are needed to completely elucidate the role of CDK12 in the cellular homeostasis.

2.6.4 CDK12 as a new candidate biomarker in breast cancer

In a recent high-throughput analysis to assess the expression of 125 STKs in different types of human cancers using ISH on TMAs, CDK12 was found to be overexpressed in breast cancer [2]. In this study, in a cohort of 92 breast IDC patients, 20% of cases displayed high CDK12 expression levels compared to normal breast [2]. Moreover, CDK12 overexpression correlated, both at the transcriptional and protein level, with clinico-pathological parameters of aggressive breast cancer disease, such as high tumor grade, high proliferative index, negative HR status and positive HER2 status, arguing for a putative role of CDK12 in breast carcinogenesis.

These preliminary findings indicated that CDK12 might represent a novel prognostic biomarker in breast cancer, and provided the basis for further studies investigating the oncogenic potential of CDK12 and its possible use as therapeutic target in breast cancer.

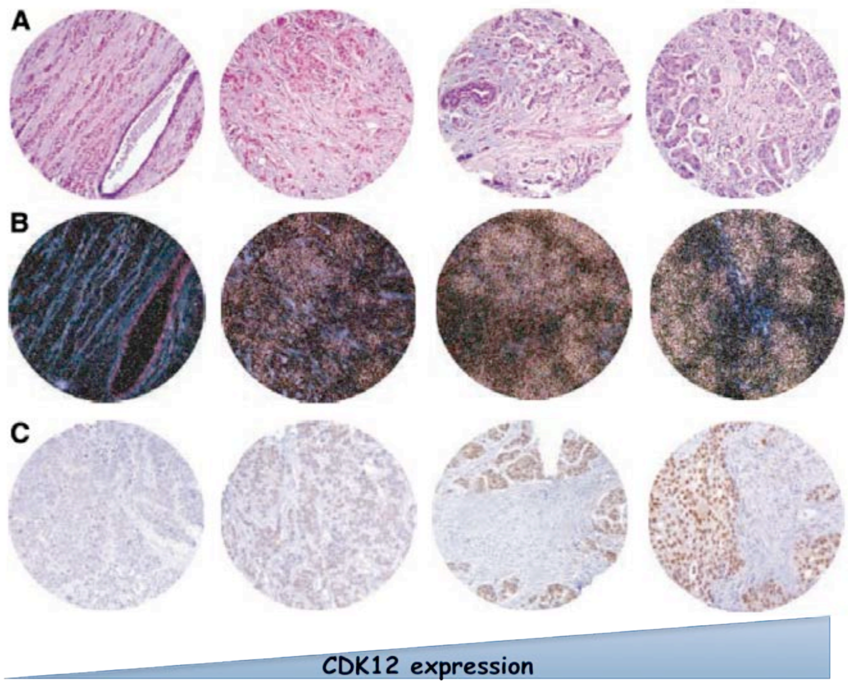


Figure 8. CDK12 expression in human breast tumor samples

Four breast carcinomas are shown, displaying increasing expression levels of CDK12, from left to right. *(A)* H&E stain*; (B) in situ* hybridization *(ISH); (C)* immunohistochemistry (*IHC*). Figure adapted from [2].

3 PRELIMINARY UNPUBLISHED DATA

3.1 CDK12 overexpression and amplification in breast cancer

From a large-scale survey of 125 STK by ISH on TMA, CDK12 was found to be overexpressed in 20% of breast tumor samples. Moreover, CDK12 overexpression significantly correlated with prognostic clinical indicators of aggressive disease, including high tumor grade, high proliferative index (Ki67), and positive HER2 [2].

In order to extend this analysis to larger case collections of breast cancer samples and to gain insights into a possible role of CDK12 as a prognostic marker in breast cancer, we decided to perform a preliminary analysis of CDK12 expression at mRNA level by ISH on TMA. We used a training set, the 'case-control' group, comprised of a representative number of patients who developed an event (loco-regional relapse, distant metastasis or controlateral breast cancer) within seven years from the first breast cancer surgery, compared to patients alive and free of event after seven years of follow-up. We also used a test set, the 'validation' group, composed of consecutive tumors.

We analyzed the association of CDK12 expression by ISH with different clinico-pathological parameters in the 'case-control' and the 'validation' cohorts. In the two study groups, CDK12 mRNA expression (ISH-CDK12) directly correlated with grade of differentiation ('case-control', $p<0.001$; 'validation', $p<0.001$), hormone-receptor status (ER 'case-control', $p=0.002$; ER 'validation', $p=0.002$; PgR 'case-control', $p=0.001$; PgR 'validation', $p=0.002$), Ki67 ('case-control', $p=0.020$; 'validation', $p<0.001$) and HER2 expression ('case-control', $p<0.001$; 'validation', $p<0.001$). Overall, the correlation analysis indicated that CDK12 expression, both in the case-control and validation study group, is associated with traditional prognostic markers of aggressive disease confirming our previous results [2].

To establish whether CDK12 is predictive of breast cancer recurrence and whether it is a reliable independent prognostic factor, we analyzed its expression, by ISH on TMA, both in the case-control and validation cohort. Univariate and multivariate logistic regression analyses of disease-free survival (any event and distant-type of event) of CDK12 expression (ISH-CDK12) indicated that CDK12 is a novel prognostic marker strongly associated with a higher risk of disease recurrence (Table 2). Of note, in the multivariate analysis in which all events were evaluated in relation to relevant prognostic factors, such as tumor grade, estrogen receptor, nodal status, Ki67 and ErbB2 expression, CDK12 overexpression remains significantly associated with a higher risk of recurrence and distant metastasis (Table 2). To validate the prognostic potential of CDK12 observed in the 'case-control' study, we performed the same analyses for the validation dataset. Univariate and multivariate analyses confirmed that CDK12 was a prognostic marker strongly associated with a higher risk of disease recurrence (Table 3). Remarkably, in multivariate analysis, in addition to CDK12 overexpression, only positive nodal status was associated with a higher risk of any and distant-type event suggesting that CDK12 is a novel predictive marker. Kaplan-Meier estimates of disease-free survival according to CDK12 expression are shown in Figure 9. Patients were grouped according to ISH-CDK12 expression levels and associations with disease-free survival (ISH-CDK12 log-rank breast-related events and distant metastasis, $p<0.001$) were maintained (Figure 9A-B). At 5 years, the risk rate of any breast-related event and of a distant-type event was 3.74% and 2.41%, respectively, for the ISH-CDK12 negative group and 20.83% and 12.50%, respectively, for the ISH-CDK12 positive group (Figure 9A,B). Kaplan-Meier estimates of disease-free survival according to ISH-CDK12 expression were also calculated in the subpopulation of HER2 negative patients. Associations with disease-free survival were significantly maintained (HR=6.811, $p<0.001$ for any breast-related event and HR=6.622, $p<0.001$ for distant-type events) (Figure 9C-D). Therefore, CDK12 retained its predictive power of any breast-related event and distant-type event in HER2 negative tumors. Importantly, the finding that CDK12 is prognostic in the HER2-negative tumors suggests that it may have a biological role in tumorigenesis independently of HER2.

Table 2. Univariate and multivariate analyses for breast-related events and distant metastasis in the case-control dataset.

	BREAST-RELATED EVENTS		DISTANT METASTASIS	
	UNIVARIATE			
	*OR (95% CI)	Pvalue	*OR (95% CI)	Pvalue
Age (≥50 vs <50)	1.060 (0.73-1.54)	0.759	1.379 (0.88-2.15)	0.157
Histotype(Lobular*vs*Ductal)	0.905 (0.54-1.51)	0.702	0.660 (0.34-1.27)	0.215
pT (2-3-4 vs 1)	2.478 (1.67-3.68)	<0.001	3.083 (1.95-4.88)	<0.001
Nodal Status (Pos vs Neg)	2.516 (1.71-3.69)	<0.001	3.785 (2.38-6.01)	<0.001
Grade (G3 vs G1-G2)	3.099 (2.00-4.81)	<0.001	3.410 (2.07-5.62)	<0.001
ER (≥10% vs <10%)	0.688 (0.45-1.04)	0.077	0.576 (0.36-0.93)	0.023
PgR (≥10% vs <10%)	0.519 (0.35-0.77)	0.001	0.393 (0.25-0.63)	<0.001
Ki-67 (≥16 vs <16)	2.455 (1.65-3.66)	<0.001	3.207 (1.97-5.23)	<0.001
HER2 (High vs Low)	2.732 (1.52-4.92)	<0.001	3.281 (1.71-6.29)	<0.001
ISH-CDK12 (≥1.5 vs <1.5)	5.069 (2.86-8.98)	<**0.001**	7.033 (3.77-13.1)	<**0.001**
	MULTIVARIATE			
	**OR (95% CI)	Pvalue	**OR (95% CI)	Pvalue
Age (≥50 vs <50)	1.253 (0.72-2.19)	0.428	1.772 (0.88-3.55)	0.107
pT (2-3-4 vs 1)	1.364 (0.76-2.45)	0.300	1.561 (0.78-3.14)	0.211
Nodal Status (Pos vs Neg)	2.606 (1.45-4.69)	0.001	3.840 (1.84-8.00)	0.000
Grade (G3 vs G1-G2)	1.469 (0.79-2.75)	0.229	1.545 (0.74-3.22)	0.245
ER (≥10% vs <10%)	1.857 (0.88-3.93)	0.106	1.518 (0.64-3.61)	0.346
PgR (≥10% vs <10%)	0.443 (0.23-0.87)	0.018	0.398 (0.18-0.87)	0.022
Ki-67 (≥16 vs <16)	2.263 (1.22-4.2)	0.010	2.987 (1.39-6.43)	0.005
HER2 (High vs Low)	1.928 (0.80-4.62)	0.141	1.662 (0.58-4.74)	0.343
ISH-CDK12 (≥1.5 vs <1.5)	2.533 (1.22-5.26)	**0.013**	3.169 (1.38-7.28)	**0.007**

*Odds ratio (OR) and 95% Confidence Intervals (CI) obtained from logistic regression models. **Odds ratio (OR) and 95% Confidence Intervals (CI) adjusted for age, pathological stage, tumor grade, hormone-receptor status, nodal status, ki67 and HER2 expression.

ISH-CDK12 expression is missing for 88 patients. The number of scored cases is lower than the total number of cases since: i) in some cases, individual cores detached from the slides during the manipulations; ii) clinical information was not available for all patients.

Table 3. Univariate and multivariate analyses for breast-related events and distant metastasis in the validation dataset.

	BREAST-RELATED EVENTS		**DISTANT METASTASIS**	
	UNIVARIATE			
	*HR (95% CI)	Pvalue	*HR (95% CI)	Pvalue
Age (≥50 vs <50)	1.024 (0.46-2.27)	0.950	1.045 (0.38-2.85)	0.930
Histotype(Lobular*vs*Ductal)	0.949 (0.23-3.97)	0.940	0 (0-Inf)	1.000
pT (2 vs 1)	1.578 (0.65-3.82)	0.310	1.124 (0.41-3.07)	0.820
Nodal Status (Pos vs Neg)	5.017 (2.33-10.8)	<0.001	5.995 (2.20-16.37)	<0.001
Grade (G3 vs G1-G2)	3.027 (1.52-6.04)	0.002	3.740 (1.59-8.81)	0.003
ER (≥10% vs <10%)	0.237 (0.11-0.51)	<0.001	0.222 (0.09-0.57)	0.002
PgR (≥10% vs <10%)	0.466 (0.24-0.92)	0.028	0.586 (0.25-1.39)	0.230
Ki-67 (≥16 vs <16)	3.040 (1.41-6.54)	0.005	3.655 (1.34-9.98)	0.011
HER2 (High vs Low)	2.042 (0.78-5.32)	0.140	2.002 (0.58-6.87)	0.270
ISH-CDK12 (≥1.5 vs <1.5)	5.96(2.88-12.36)	<**0.001**	5.536 (2.20-13.95)	<**0.001**
	MULTIVARIATE – ISH			
	**HR (95% CI)	Pvalue	**HR (95% CI)	Pvalue
Age (≥50 vs <50)	0.761 (0.33-1.76)	0.520	0.715 (0.24-2.10)	0.540
pT (2 vs 1)	1.204 (0.44-3.3)	0.720	0.837 (0.25-2.76)	0.770
Nodal Status (Pos vs Neg)	6.85(2.81-16.72)	0.000	8.557 (2.58-28.34)	0.000
Grade (G3 vs G1-G2)	1.175 (0.42-3.28)	0.760	1.661 (0.43-6.40)	0.460
ER (≥10% vs <10%)	0.434 (0.16-1.21)	0.110	0.384 (0.10-1.49)	0.170
PgR (≥10% vs <10%)	0.665 (0.27-1.66)	0.380	0.987 (0.28-3.45)	0.980
Ki-67 (≥16 vs <16)	0.981 (0.34-2.82)	0.970	0.989 (0.24-4.06)	0.990
HER2 (High vs Low)	0.334 (0.11-1.01)	0.052	0.393 (0.09-1.64)	0.200
ISH-CDK12 (≥1.5 vs <1.5)	12.15(4.3334.12)	**0.000**	10.153(2.6439.00)	**0.001**

*Hazard ratio (HR) and 95% Confidence intervals (CI) obtained from Cox regression models.

**Hazard ratio (HR) and 95% Confidence intervals (CI) adjusted for age, tumor size, nodal status, grade, hormone-receptor status, Ki67 and HER2 expression.

ISH-CDK12 expression is missing for 70 patients. The number of scored cases is lower than the total number of cases since: i) in some cases, individual cores detached from the slides during the manipulations; ii) clinical information was not available for all patients.

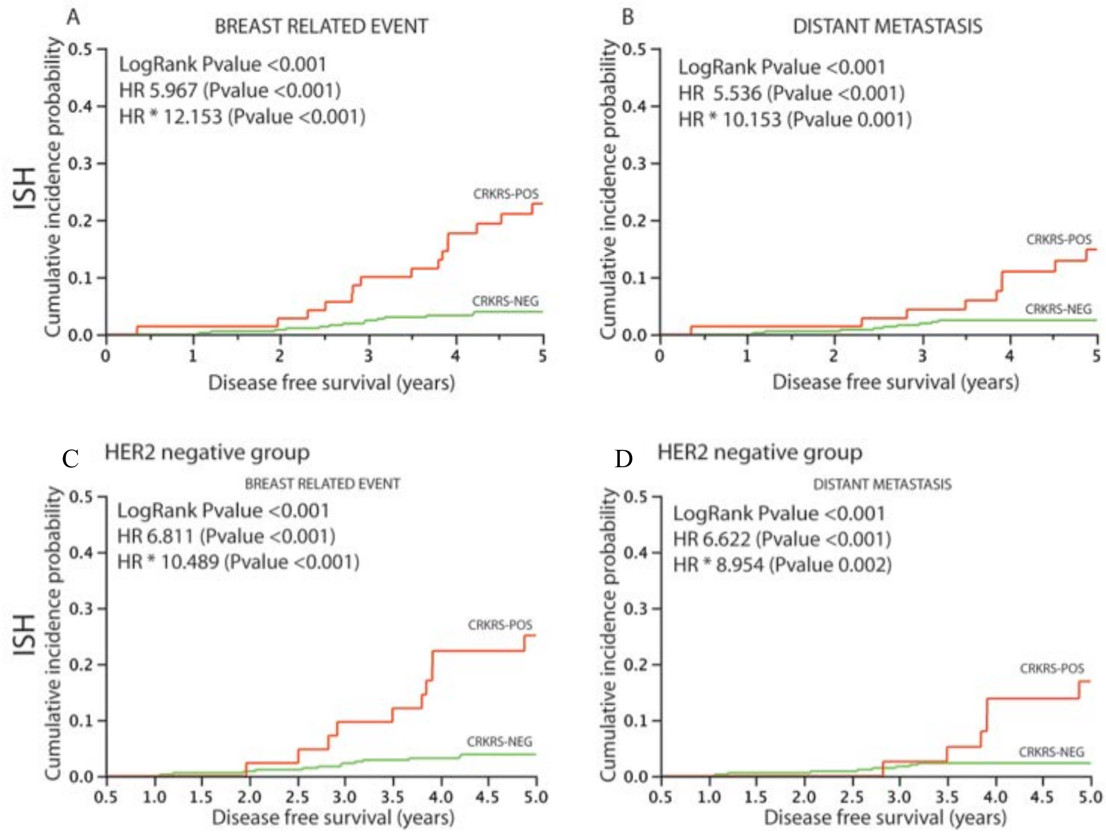


Figure 9. Cumulative incidence of breast-related events and distant metastasis in the 'validation' dataset.

Kaplan-Meier plots of the 5-year cumulative incidence of breast-related events (A and C) and distant metastasis (B and D) according to CDK12 positive (CDK12-POS, red line) and negative (CDK12-NEG, green line) expression measured by *in situ* hybridization in all patients of the validation dataset (A and B) and in the subgroup of HER2-negative patients (C and D).

Log-rank P values and Hazard ratios of univariate (HR) and multivariate (HR*) analyses derived from Table 3 are indicated.

Gene amplification is one of the molecular mechanisms leading to oncogene activation [120]. Importantly, the identification of amplified oncogenes may have both diagnostic and therapeutic implications [121-124]. Indeed, fluorescence *in situ* hybridization (FISH) analysis, widely used to study gene amplification in breast cancer, has revealed a number of gene copy number alterations, including regions with high-level amplification that were associated with poor clinical outcome [121,122,124].

CDK12 is located on chromosome 17 (17q12 region). The 17q12-q21 region is frequently amplified in breast cancer. Historically this region has been described as the ERBB2-amplicon and many studies focused on the role of this RTK in breast cancer both as a prognostic marker and as a therapeutic target.

Previous studies, reported CDK12 as a frequent ERBB2 co-amplified gene [60 63]. However, so far nobody addressed in a systematic way specifically the amplification of CDK12 and its correlation with ERBB2 amplification. Therefore, we performed a dual color interphase FISH analysis on breast TMAs, using a centromeric probe for chromosome 17 and a BAC clone encompassing the *CDK12* gene. We also performed on the same cases a FISH analysis to determine ERBB2 amplification. In total, we analyzed ~350 breast tumor samples, of which 132 gave evaluable results for both the *CDK12* and *HER2* genes. The *CDK12* gene was considered to be amplified when the cut-off ratio of the *CDK12* signal *vs.* the centromeric signal) was higher than 2.25.

Surprisingly, in the analyzed patient cohort CDK12 amplification was an event more frequent than ERBB2 amplification and as a consequence it can occur as an event independent from ERBB2 amplification. In particular, *CDK12* was amplified in 32% (42 out of 132) of primary breast tumors, whereas *ERBB2* was amplified in approximately 13% of cases (17 out of 132) (Figure 10). However, among the cases with *CDK12* amplification (42 in total) only 26% (n = 11) displayed *ERBB2* co-amplification. In contrast, most of tumors harboring *HER2* amplification (65%, 11 out of 17 in total) displayed concomitant *CDK12* amplification (Figure 10).

Remarkably, these findings, together with the observation that CDK12 overexpression has a potential prognostic role for breast cancer patients that remains significant in the subgroup of ErbB2-negative patients, suggest a causative involvement of CDK12 in breast cancer progression, both in cooperation and independent of *ERBB2* amplification.

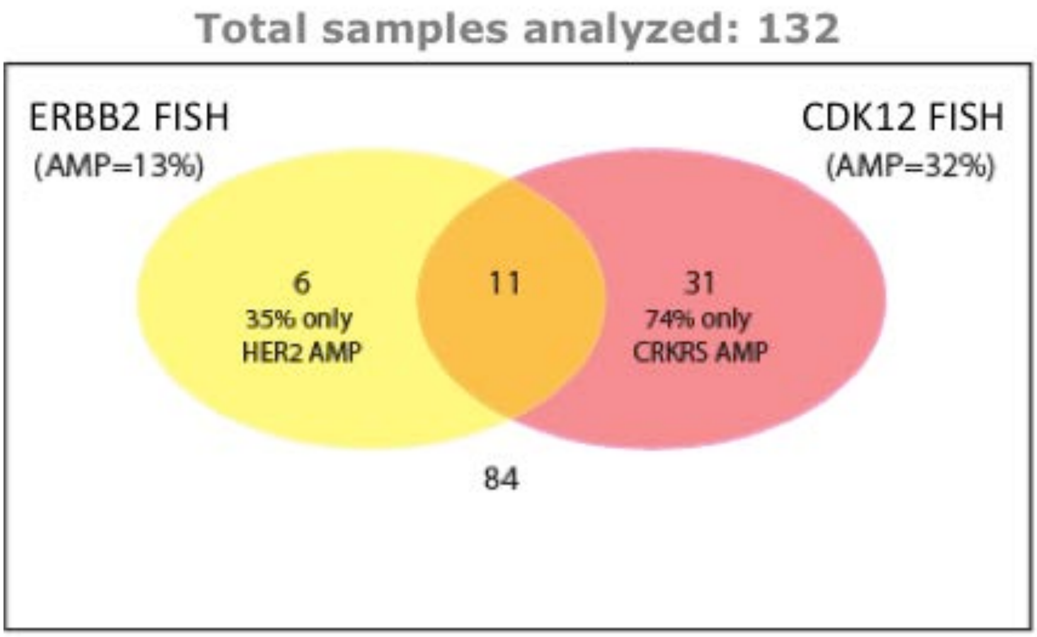


Figure 10. Correlation analysis between *CDK12* and *ERBB2* gene amplification.

Venn diagram showing the number of breast tumors with *CDK12* and/or *ERBB2* gene amplifications. Both *CDK12* and *ERBB2* gene amplification, measured by FISH on breast TMAs (FISH-CDK12 and FISH-ERBB2, respectively), were considered negative or amplified when the ratio between each specific gene and the chromosome 17 centromere copy number was <2.25 or ≥ 2.25, respectively. For a total of 132 samples, it was possible to determine both *CDK12* and *ERBB2* gene copy numbers: 84 samples were not amplified for either *CDK12* or *ERBB2*, whereas 48 samples were amplified for at least one gene. The indicated percentages were calculated based on the total number of analyzed samples, i.e., 132.

4 AIMS AND RATIONALE OF THE STUDY

The identification of novel biomarkers and new therapeutic targets for more refined prognostic stratification and clinical treatment of breast cancer patients remains a major unmet clinical need. Protein kinases constitute a large family of regulatory enzymes involved in cellular homeostasis by virtually controlling every cellular process. It is therefore not surprising that functional subversion of several kinases, due to different possible mechanisms, is frequently implicated in malignant transformation. For these reasons, kinases hold great promise both as cancer biomarkers and as molecular targets for the development of new therapies. This is witnessed by the increasing number of kinase inhibitors that have already found application in cancer therapy, such as Imatinib (Gleevec) for the treatment of chronic myeloid leukemia, Gefininib (Iressa) and Erlotinib (Tarceva) for the treatment of non-small cell lung cancer[125], Lapatinib (Tyverb) for the treatment of HER-2 positive breast cancer[25], while others such as Seliciclib (Roscovitine), Sorafenib (BAY 43-9006) and Vatalanib (PTK787) are currently under evaluation in clinical trials[126]. Therapeutic targeting of kinases can also be achieved by the development of inhibitory monoclonal antibodies, as epitomized by the monoclonal anti-HER2 antibody Trastuzumab (Herceptin) for the treatment of HER2-positive breast cancer patients[25 126].

In line with this rationale, a recent high-throughput analysis was conducted in our lab to analyze the expression of 125 STKs in different types of human cancers using ISH on TMA: among the positive hits of this study CDK12 was found to be overexpressed in 20% of breast cancers. Of note, in this study, CDK12 expression correlated, both at the transcriptional and protein level, with clinical/pathological parameters of aggressive breast cancer disease, such as high tumor grade, high proliferative index, negative hormonal status and positive ERBB2 status, in a cohort of 92 breast invasive ductal carcinoma (IDC) patients [2], arguing for a putative functional role of CDK12 in breast carcinogenesis.

Based on this initial study, we hypothesized that CDK12 deregulation could also have a pathogenetical role in breast carcinogenesis and therefore, once functionally validated, constitute a further option for targeted therapy in the broader perspective of personalized clinical management of breast cancer patients. With this idea in mind, we embarked on an in-depth study of CDK12 implication in breast cancer through the integration of molecular pathology, post-genomics and classical protein function studies, a comprehensive approach that unpredictably also shed light on some novel aspects concerning the role of CDK12 in physiology and pathology.

In particular, in the present thesis work, the first goal was to corroborate initial data on the clinical value of CDK12 as a predictive prognostic biomarker in breast cancer by performing an extensive immunohistochemical analysis of CDK12 expression on a TMA derived from a large cohort of breast cancer patients with complete clinical follow-up (see sections 3.1 and 5.2). For the purpose of this analysis, a new monoclonal anti-CDK12 antibody was generated, which was instrumental to assess the correlation of CDK12 overexpression with clinical/pathological markers of aggressive breast cancer disease and with the ability to predict the likelihood of breast cancer recurrence and patient survival.

The second objective was to functionally implicate CDK12 dysfunction as a driver pathogenetic lesion in breast cancer (see Section 5.4 and 5.5 of Results). To this aim, we used well-established *in vitro* and *in vivo* assays to study cell transformation and acquisition of tumorigenic potential to compare the biological effect of CDK12 overexpression or ablation in normal and tumorigenic backgrounds, in order to unequivocally understand whether breast cancer cells are functionally dependent on CDK12 expression/overexpression.

Finally, to provide mechanistic insights on how CDK12 deregulation might eventually mediate tumorigenesis, based on recent evidences supporting a role of CDK12 in the regulation of both transcription [127] and splicing [117], we performed a genome-wide analysis of the transcription/splicing alterations linked to CDK12 perturbation in cell-based model systems, using the high-throughput Affymetrix Exon Array technology (see Section 5.6 of Results).

5 RESULTS

5.1 Generation of a specific antibody against human CDK12

To perform extensive CDK12 protein expression analyses we produced, in collaboration with the Antibody Facility at IFOM, an anti-CDK12 monoclonal antibody, as well as the corresponding hybridoma, thus allowing an unlimited source of antibody to perform various analyses without reagent limitation. An advantage of using a monoclonal antibody rather than a polyclonal antibody is that it avoids the variability in antigen recognition linked to polyclonal antibodies, which are produced *ex novo,* starting each time from the animal immunization. The monoclonal antibody would be useful to perform analyses, such as immunoblotting, immunoprecipitation, immunofluorescence and immunohistochemistry. Most importantly, the availability of a monoclonal antibody, once validated as a clinical grade reagent, is of paramount importance for routine patient stratification based on the targeted biomarker.

Using the bioinformatic tool GlobPlot, we selected a specific globular region within the CDK12 moiety, outside of the kinase domain, in order to avoid recognition of regions conserved in other kinases. An antigenic peptide corresponding to a sequence unique to the CDK12 protein (amino acid residues 400–524) was used as GST-fusion protein to immunize mice. The "AQ19" monoclonal antibody was selected, affinity-purified (final concentration: 250 μg/ml), and tested for its ability to specifically detect CDK12 using different techniques, namely immunoblotting (IB), immunoprecipitation (IP), and immunofluorescence (IF) (Figure 11).

In the IB analysis, the AQ19 antibody recognized a major band of approximately 190 kDa in MCF10A total cell lysates, corresponding to the predicted molecular weight (MW) of the CDK12 protein (Figure 11A). This band disappeared in CDK12-silenced MCF10A cells, demonstrating that the antibody specifically recognizes CDK12 (Figure 11A).

We performed IP experiments with the "AQ19" using MCF10A total cell lysates. IB detection of CDK12 showed that the CDK12 protein was enriched in the IP samples compared to the control input, and it was immunodepleted in the corresponding supernatant (Figure 11B). CDK12 was not immunoprecipitated by the HA-antibody used as a control. These results demonstrated that the "AQ19" antibody is able to specifically pull-down CDK12.

In the IF experiment, staining with the AQ19 antibody resulted in a nuclear-specific signal, which is consistent with the expected subcellular localization of CDK12 [101], in control MCF10A cells, that was significantly reduced in CDK12-silenced cells (Figure 11C). This result indicates that the AQ19 antibody specifically recognizes CDK12 in IF experiments.

We also tested the suitability of the antibody for immunohistochemistry (IHC) analysis of formalin-fixed paraffin-embedded (FFPE) samples. To this aim, we prepared FFPE samples of control and CDK12-silenced MCF10A breast cells. IHC staining of these paired samples with the AQ19 antibody yielded a nuclear signal in control MCF10A cells that disappeared in CDK12-silenced cells, indicating that the antibody specifically and efficiently recognizes CDK12 in IHC experiments (Figure 12).

Therefore, we successfully generated a CDK12-specific monoclonal antibody that is suitable for use in a wide range of techniques including IB, IP, IF and IHC.

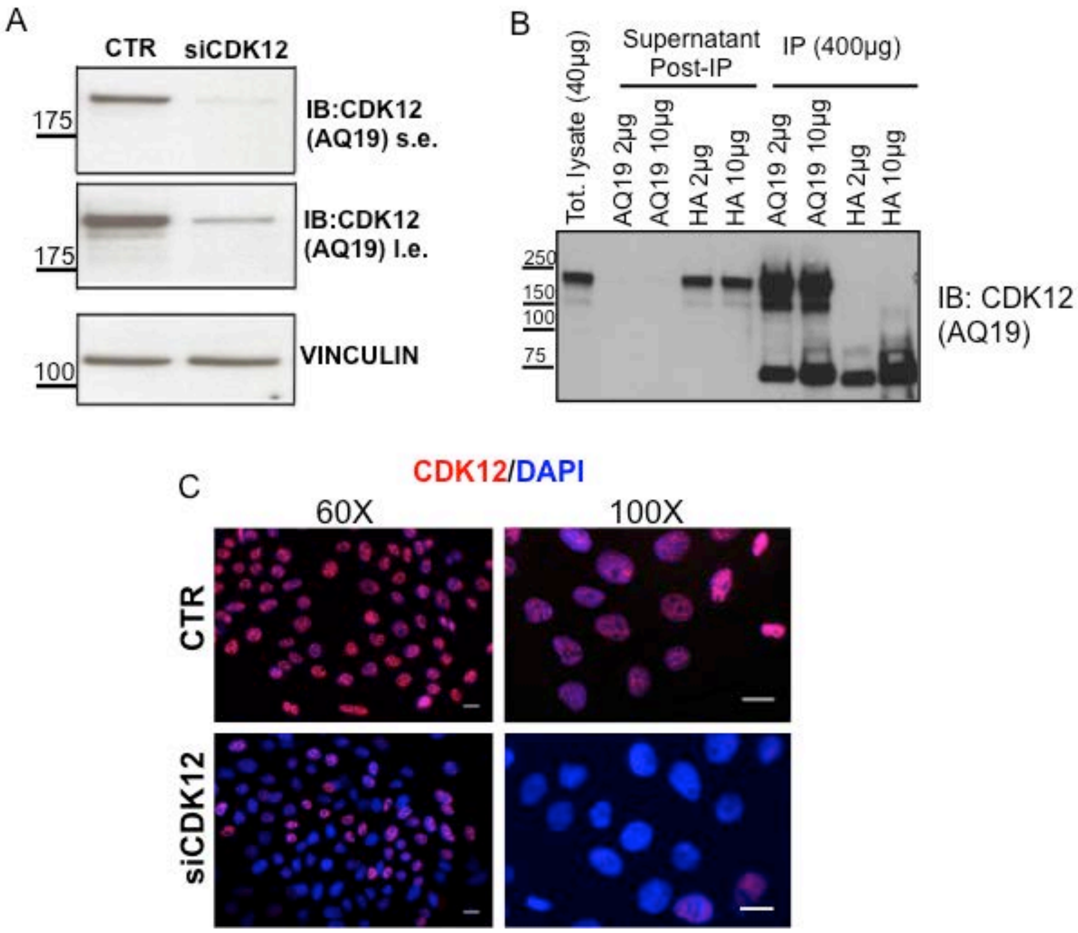


Figure 11. Characterization of the AQ19 monoclonal antibody

A) Immunoblotting (IB) with AQ19. Total cell lysates (40μg) from control (CTR) or CDK12-silenced (siCDK12) MCF10A cells were immunoblotted with the AQ19 monoclonal antibody. Both short and long exposure blots are shown (s.e. and l.e., respectively). Vinculin was detected as a loading control. MW markers are reported to the left of the blots.

(B) Immunoprecipitation (IP) with AQ19. The AQ19 antibody (2 or 10μg) was used to immunoprecipitate CDK12 from MCF10A total cell lysates (400 μg). An anti-HA antibody (2 or 10μg) was used to control for non-specific binding in the IP reaction. Blot shows input of the IP reaction (40μg), supernatants of the IP to demonstrate the efficiency of CDK12 immunodepletion (40μg), and the anti-CKD12 and anti-HA immunoprecipitates (400μg). MW markers are reported to the left of the blot.

(C) Immunofluorescence (IF) with AQ19. Control (CTR) or CDK12-silenced (siCDK12) MCF10A cells were fixed and stained with the AQ19 antibody (red). Nuclei were counterstained with DAPI (blue). Representative overlaid images at two different magnification are shown. Scale bars: 10 μm. Blots and images are representative of 3 repeats.

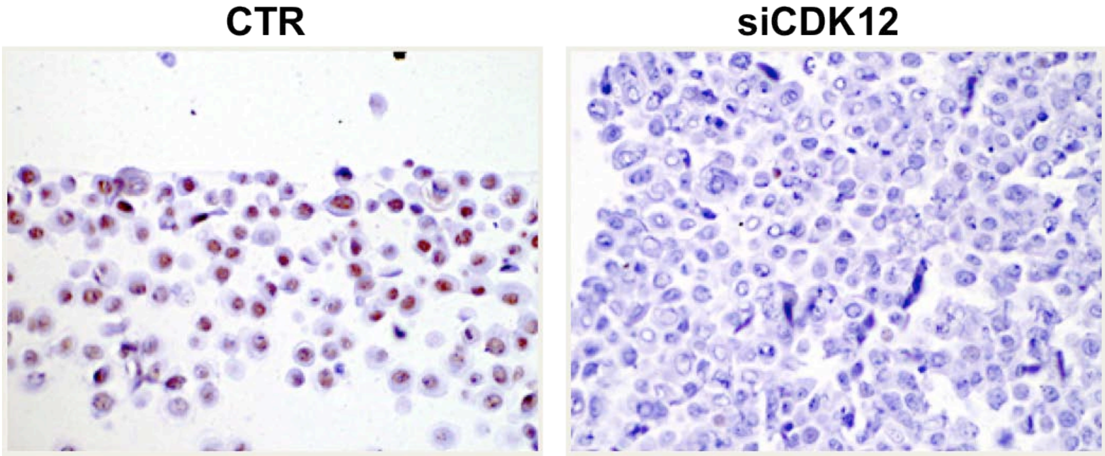


Figure 12. The AQ19 CDK12 monoclonal antibody specifically recognizes CDK12 protein in FFPE samples by IHC analysis

IHC analysis of FFPE samples of control (CTR) and CDK12-silenced (siCDK12) MCF10A cells. Representative images are shown. Magnification 40X.

5.2 Immunohistochemical analysis of CDK12 expression in breast cancer patients

5.2.1 Correlation of CDK12 expression and clinical/pathological parameters in invasive breast carcinomas

To assess the clinical value of CDK12 protein expression in breast cancer, we decided to investigate the correlation between intrinsic CDK12 status of patients with standard clinical/pathological parameters and with disease outcome. To this aim, we took advantage of two independent case collections of breast cancer specimens available on TMAs, which were analyzed using the AQ19 anti-CDK12 monoclonal antibody produced in house (see Results Section 5.1)

The IHC analysis was initially conducted in a case-control study group of 349 cases and results obtained from this patient cohort were subsequently validated in an independent collection of 970 consecutive cases (Table 4 and Table 5). In parallel CDK12 expression was measured in a group of normal breast tissue samples (N=38). In normal mammary tissues, CDK12 was expressed at low level (median score =1, average score =0.8). Tumors with high CDK12 expression were thus defined as those displaying an expression score > 1. On this basis, we were able to classify tumors in the case-control and validation cohorts as CDK12-high and CDK12-low according to the score assigned (Table 6 and Table 7).

We analyzed the association of CDK12 expression by IHC with different clinico-pathological parameters in the 'case-control' and in the 'validation' cohorts. In the two study groups, CDK12 expression (IHC-CDK12) directly correlated with high grade, i.e. with a poor degree of differentiation ('case-control', p=0.0055; 'validation set', p<0.0001), negative estrogen (ER) and progesterone (PgR) receptor status (ER: 'case-control', p=0.0322; 'validation set', p=0.011; PgR: 'case-control', p=0.0012; PgR ‘validation’, p=0.002), high Ki67 proliferation index (‘case-control’, p=0.0052; ‘validation set’, p<0.0001) and high ErbB2 expression status (‘case-control’, p<0.001; ‘validation set’, p<0.0001) (Table 6 and Table 7).

Overall, these IHC analyses indicated that CDK12 expression is associated with traditional prognostic markers of aggressive disease, which extends and further corroborates initial evidences on the clinical value of CDK12 detection in breast cancer [2].

Furthermore, we found a significant association between high CDK12 status and increased risk of disease relapse within 15 years from the first breast cancer surgery both in the case-control (p=0.0050) and in the validation group (p=0.025). In the same patient cohorts, high CDK12 status also significantly correlated with survival status (p =0.0102 in case-control; 0.0146 in validation set) (Table 6 and Table 7).

Table 4. Clinical and pathological information of the case-control cohort of breast cancer patients (N=349)

PARAMETER	GROUP	CASE-CONTROL DATASET (N = 349)	
		N	**%**
Histotype	*DUCTAL*	290	83.09
	LOBULAR	50	14.33
	OTHER	9	2.58
pT	*1*	178	51.44
	2	142	41.04
	3	18	5.2
	4	8	2.3
Nodal Status	*NEG*	153	43.84
	POS	196	56.16
GRADE	*G1*	56	18.42
	G2	132	43.42
	G3	116	38.16
ER	*NEG*	106	31.45
	POS	231	68.55
PgR	*NEG*	140	41.54
	POS	197	58.46
Ki-67	*NEG*	116	34.42
	POS	221	65.58
ErbB2	*NEG*	227	88.22
	POS	37	11.78
NPI	*GPG*	84	28
	MPG	133	44.33
	PPG	83	27.67
ANY RELAPSE	*NO*	165	47.28
	YES	184	52.72
DISTANT RELAPSE	*NO*	245	70.2
	YES	104	29.8
STATUS	*ALIVE*	233	66.76
	DEAD	116	33.24

The clinical and pathological information of the case-control study group of breast cancer patients operated at the European Institute of Oncology (IEO) between 1994 and 1997 is reported. Disease recurrence (any relapse and distant relapse) was within 18 years (median 9.2 years and 10.6 years, respectively). For some patients not all information was available. Nottingham Prognostic Index (NPI) combines nodal status, tumor size and histological grade. According to NPI's score, patients can be divided into 3 classes: Good Prognosis Group (GPG), Moderate Prognosis Group (MPG) and Poor Prognosis Group (PPG).

Table 5. Clinical and pathological information of the consecutive cohort of breast cancer patients (N=970)

PARAMETER	GROUP	Consecutive Cohort (N = 970)	
		N	**%**
Histotype	*DUCTAL*	775	79.90
	LOBULAR	104	10.72
	OTHER	91	9.38
pT	*1*	86	8.87
	2	802	82.68
	3	67	6.90
	4	15	1.55
Nodal Status	*NEG*	320	33.47
	POS	636	66.53
GRADE	*G1*	72	7.83
	G2	359	39.02
	G3	489	53.15
ER	*NEG*	207	21.34
	POS	763	78.66
PgR	*NEG*	319	32.89
	POS	651	67.11
Ki-67	*LOW*	192	19.81
	HIGH	777	80.19
ErbB2	*NEG*	273	28.14
	POS	106	10.93
	N.A.	591	60.93
ANY RELAPSE	*NO*	575	59.46
	YES	392	40.54
DISTANT RELAPSE	*NO*	752	78.17
	YES	210	21.83
STATUS	*ALIVE*	719	74.12
	DEAD	251	25.88
Subtype	*Luminal A*	80	8.25
	Luminal B	160	16.49
	Luminal B, Her2 Pos.	67	9.91
	ErbB2 Pos.	39	4.02
	Triple Negative	33	3.40
	ErbB2 status Unknown	591	60.92

The clinical and pathological information of the breast cancer patients included in the consecutive cohort and operated at the European Institute of Oncology (IEO) between 1997 and 2000 is reported. For some patients not all information was available. ErbB2 status before year 2000 was not routinely assessed. Subtypes are defined according to the clinico-pathologic criteria defined in the St. Gallen guidelines 2011.

Table 6. Correlation of CDK12 expression and clinico-pathological parameters in the case-control cohort of 349 breast cancer patients.

		CASE-CONTROL COHORT (N = 323*)		
		CDK12 LOW	**CDK12 HIGH**	**χ2 P-value**
All Patients		234	89 (26 %)	
ER	NEG	64	34 (**34.69 %**)	
	POS	166	50 (23.15 %)	**0.0322**
PgR	NEG	82	47 (**36.43 %**)	
	POS	148	37 (20 %)	**0.0012**
Ki67	NEG	88	18 (16.98 %)	
	POS	142	66 (**31.73 %**)	**0.0052**
ErbB2	NEG	211	54 (20.8 %)	
	POS	6	28 (**82.35 %**)	**< 0.001**
pT	1	124	42 (25.30 %)	
	2-3-4	109	45 (29.22%)	0.4311
GRADE	G1	43	8 (15.69 %)	
	G2	92	30 (24.59 %)	
	G3	69	43 (**38.39** %)	**0.0055**
GRADE	G1-G2	135	38 (21.97 %)	
	G3	69	43 (**38.39** %)	**0.0027**
Node	NEG	97	45 (31.69 %)	
	POS	137	44 (34.31 %)	0.1406
NPI	GPG	59	17 (22.37 %)	
	MPG	92	35 (27.56 %)	
	PPG	51	27 (34.62 %)	0.2355
Any Relapse	NO	125	32 (20.38 %)	
	YES	109	57 (**34.34 %**)	**0.0050**
Distant Relapse	NO	174	59 (25.32 %)	
	YES	60	30 (33.33 %)	0.1485
Status	Alive	169	51 (23.18 %)	
	Dead	65	38 (**36.89 %**)	**0.0102**

CDK12 expression was measured by IHC on TMA using the AQ19 antibody. Tumors with a high CDK12 status (CDK12 High) were defined as those tumors displaying an expression score > 1, while low CDK12 status (CDK12 low) tumors were those with an expression score ≤ 1. P - values were determined by the Pearson's chi-squared test. Note that the number of scored cases (N=323) is lower that the total number of patients composing the case-control study group (N=349) because of a number of possible reasons: i) in some cases, individual cores detached from the slides during the manipulations; ii) clinical information was not available for some patients. In tumor tissues, IHC signals were associated with the tumor cell component and not with the adjacent or infiltrating stroma. *Only for 323 out of 349 samples expression data for CDK12 were available.

Table 7. Correlation of CDK12 expression and clinico-pathological parameters in the consecutive cohort (N=970).

		Consecutive Cohort (N = 801*)		
		CDK12 LOW	**CDK12 HIGH**	**χ2 P-Value**
All Patient		472	329 (41.07%)	
Nodal Status	Negative	158	106 (40.15 %)	0.734
	Positive	307	217 (41.41 %)	
SubType	Luminal A	53	3 (5.36 %)	**<0.0001**
	Luminal B	95	40 (29.63 %)	
	Luminal B Her2 Positive	15	40 (**72.73 %**)	
	HER2 Positive	6	22 (**78.57 %**)	
	Triple Negative	20	8 (28.57 %)	
ER	Negative	84	83 (**49.70 %**)	**0.011**
	Positive	388	246 (38.80 %)	
Grade	G1	45	9 (16.67 &)	**<0.0001**
	G2	207	85 (29.11 %)	
	G3	196	221 (**53.00%**)	
PgR	Negative	137	130 (48.69 %)	**0.002**
	Positive	335	199 (37.27 %)	
Ki-67	LOW	121	23 (15.97 %)	**<0.0001**
	HIGH	350	306 (**46.65 %**)	
ERBB2 Status	Negative	168	51 (23.29 %)	**<0.0001**
	Positive	21	62 (**74.70 %**)	
Status	Alive	361	226 (38.50 %)	**0.0146**
	Dead	111	103 (**48.13 %**)	
Any Event	NO	293	179 (37.92 %)	**0.025**
	YES	177	150 (**45.87 %**)	
Distant Event	NO	371	246 (39.87 %)	0.2268
	YES	98	80 (44.94 %)	

CDK12 expression was measured by IHC on TMA using the AQ19 antibody. Tumors with a high CDK12 status (CDK12 High) were defined as those tumors displaying an expression score > 1, while low CDK12 status (CDK12 low) tumors were those with an expression score ≤ 1. P - values were determined by the Pearson's chi-squared test. Note that the number of scored cases is lower than the total number of cases since: i) in some cases, individual cores detached from the slides during the manipulations; ii) clinical information was not available for some patients. In tumor tissues, the IHC signals were associated with the tumor cell component and not with the adjacent or infiltrating stroma. *Only for 801 of 970 samples expression data for CDK12 were available.

5.2.2 Analysis of the association of CDK12 expression with overall survival and disease free-survival in breast cancer patients

As CDK12 overexpression correlates with indicators of aggressive disease in breast cancer, we investigated whether CDK12 status can be considered as a new prognostic marker able to predict breast cancer recurrence and overall survival.

Univariate logistic regression analysis performed in the case-control cohort indicated that CDK12 overexpression is strongly associated with a higher risk of disease recurrence (Odds Ratio=2.042; p=0.0048) and death (Odds Ratio=1.937; p=0.0113).

To validate the prognostic potential of CDK12 observed in the case-control cohort, we performed Kaplan-Meier analysis on the consecutive cohort of 970 patients stratified on the basis of their IHC-CDK12 expression levels in high (score>1) and low (≤1) CDK12 status. Kaplan-Meier curves of disease recurrence and overall survival showed inverse correlations of CDK12 status with disease-free survival (log-rank=0.0058) and overall survival (log-rank=0.0098) (Figure 13). To refine this analysis, we stratified patients into three classes according to IHC-CDK12 levels, comprising a low (score ≤1), moderate (1 < score ≤ 2), and high (2 < score ≤ 3) CDK12 subgroup, in which significant differences in terms of disease-free survival (log-rank=0.0162) and overall survival (log-rank=0.0095) were maintained. Moreover, in these three IHC-CDK12 subgroups of patients, Kaplan-Meier estimates showed a proportionally higher risk of disease relapse and death associated with increasing levels of CDK12 expression (Figure 14).

We also analyzed the association between IHC-CDK12 expression and disease-free survival in the subpopulation of ErbB2-negative patients present in the consecutive cohort. Also in this case, correlation between high CDK12 status and increased risk of disease recurrence (log-rank=0.0162) *vis a vis* decreased overall survival (log-rang=0.0095) were significantly maintained (Figure 15). We therefore concluded that CDK12 overexpression retains its predictive value as a poor prognostic marker in ErbB2-negative tumors.

Altogether, our findings indicate that CDK12 overexpression is a new prognostic marker in breast cancer significantly associated with a high risk of recurrence and fatal outcome. Importantly, the finding that CDK12 overexpression retains its prognostic value in the ErbB2-negative subset of patients suggests that it may play a functional role in breast tumorigenesis independent of ErbB2. This hypothesis is consistent with previous observations from our lab demonstrating that, whereas CDK12 is frequently co-amplified with ERBB2, in a sizable number of cases CDK12 amplification occurs independently of ERBB2. These results further prompted us to embark on a series of functional studies to prove the functional implication of aberrant CDK12 activity in breast cancer.

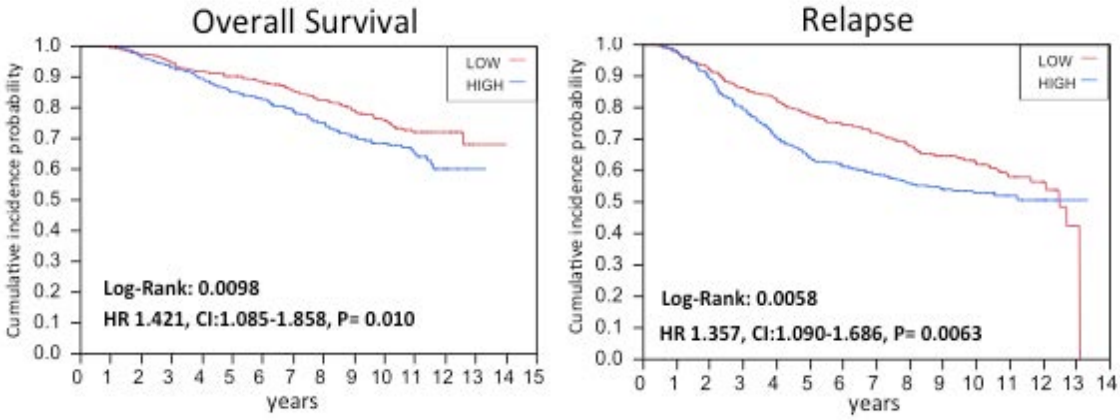
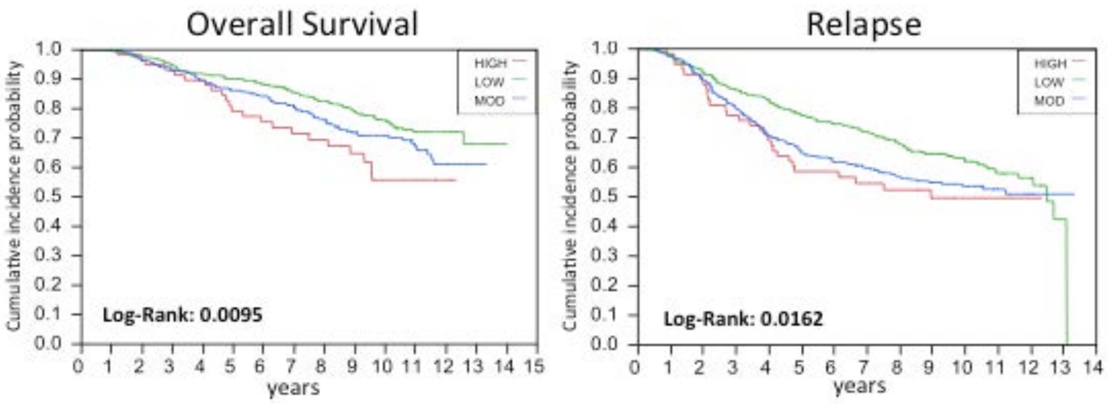


Figure 13. Cumulative incidence probability of overall survival and breast cancer relapse in the consecutive cohort.

Kaplan-Meier plots of the 15-year cumulative incidence probability of overall survival and breast cancer relapse according to LOW- ((score ≤ 1, red line) and HIGH- (score > 1, blue line) CDK12 expression measured by IHC in the consecutive cohort of 970 breast cancer patients. Log-rank values, Hazard ratios (HR) and P-values determined by the Cox-model are indicated.



Overall Survival

	Risk Ratio	Lower 95%	Upper 95%	P-value
MODvsLOW	**1.331**	0.997	1.769	**0.0527**
HIGHvsLOW	**1.896**	1.169	2.934	**0.0109**
HIGHvsMOD	1.424	0.868	2.240	0.1557

Relapse

	Risk Ratio	Lower 95%	Upper 95%	P-value
MODvsLOW	**1.321**	1.047	1.662	**0.0193**
HIGHvsLOW	**1.539**	1.011	2.254	**0.0444**
HIGHvsMOD	1.165	0.757	1.730	0.4739

Figure 14. Cumulative incidence probability of overall survival and breast cancer relapse in the consecutive cohort.

Kaplan-Meier plots of the 15-year cumulative incidence probability of overall survival and breast cancer relapse according to low- (score ≤ 1, green line), moderate- (1 < score ≤ 2, blue line) and high- (2 < score ≤ 3, red line) CDK12 expression measured by IHC in the consecutive cohort of 970 breast cancer patients. Log-rank values are indicated in the figures. Hazard ratios (HR) and P–values, determined by the Cox-model, between groups relative to overall survival and relapse, are reported in tables.

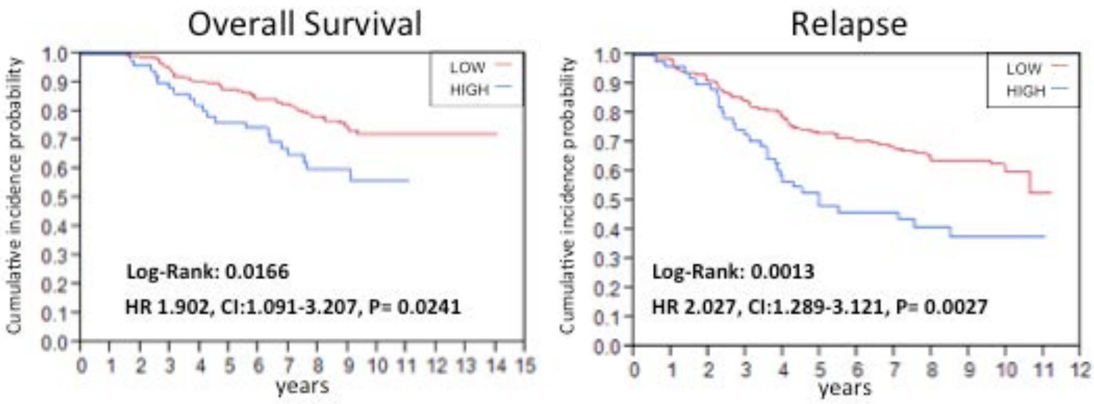


Figure 15. Cumulative incidence probability of overall survival and breast cancer relapse in the ErbB2-negative patients of the consecutive cohort

Kaplan-Meier plots of the 15-year cumulative incidence probability of overall survival and breast cancer relapse according to low- (score ≤ 1, LOW, red line) and high (score > 1, HIGH, blue line) CDK12 expression measured by IHC on TMA in the ErbB2-negative patients of the consecutive cohort. Log-rank values, Hazard ratios (HR) and P–values, determined by the Cox-model, are indicated.

5.3 Analysis of CDK12 expression and gene amplification in breast cell lines.

We have shown that CDK12 is overexpressed in human breast tumors due to amplification of the *CDK12* gene (see section 3). This finding suggests that genetic lesions affecting the *CDK12* locus and resulting in aberrant CDK12 expression levels might be involved in breast carcinogenesis.

To experimentally address this possibility, we first needed to identify suitable cellular model systems for investigating the contribution of CDK12 to tumorigenic phenotypes. We therefore screened a panel of commercially available, non-tumorigenic and tumorigenic breast cell lines (Table 8) for their intrinsic CDK12 status, both at the protein and transcript level, by IB and quantitative RT-PCR (Q-PCR) analysis, respectively. In this screening, we observed that CDK12 was overexpressed to varying degrees in the breast tumor cell lines BT474, SKBR3, UACC-812 and AU565 compared with the non-tumorigenic mammary epithelial cell line MCF10A, which is a well-established model for normal mammary epithelial cells (Figure 16).

In particular, MCF10A is a spontaneously immortalized, but non-transformed human mammary epithelial cell line derived from the breast tissue of a patient with fibrocystic changes. MCF10A cells are commonly recognized as a normal breast epithelial cell line because of lack of tumorigenicity in nude mice and lack of anchorage-independent growth. However, some genetic abnormalities have been characterized, in particular the deletion of the locus containing *p16-p14ARF* and amplification of *MYC*. MCF10A cells also express wild-type p53 [128].

Among the breast tumor cell lines analyzed, BT474 and SKBR3 displayed the highest levels of CDK12 protein (Figure 16A), accompanied by the highest levels of CDK12 mRNA (Figure 16B).

We then assessed whether CDK12 overexpression in these breast tumor cell lines was due to gene amplification, thereby eventually recapitulating the genetic lesions observed in biopsy samples of human breast tumors (see section 5.3). We performed a dual color interphase FISH analysis, using a centromeric probe for chromosome 17 and a BAC clone encompassing the *CDK12* gene (CTD3082P18). Based on previous experience accumulated in our laboratory for the analysis of *CDK12* amplification on TMAs (see section 3.1), the *CDK12* gene was considered to be amplified when the cut-off ratio of the *CDK12* signal *vs.* the centromeric signal was higher than 2.25. We observed that *CDK12* was amplified in both BT474 and SKBR3 cells, while no amplification was observed in non-tumorigenic MCF10A cells (Figure 17).

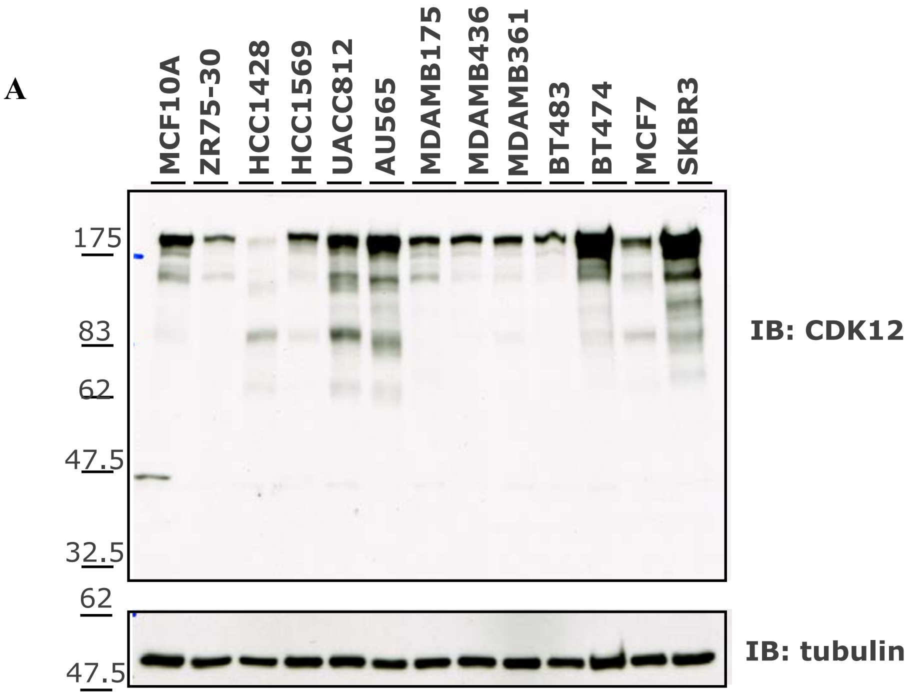
Based on these results, we selected three breast cell lines to be used in the functional characterization of CDK12 according to the following criteria: i) the CDK12-overexpressing breast cancer cell line BT474, which also displays concomitant *ErbB2* amplification; ii) the normal, non-tumorigenic mammary epithelial cell line MCF10A; iii) the breast cancer cell line HCC1569 that expresses CDK12 at levels comparable to normal MCF10A cells. Of note, the comparative phenotypic analysis between HCC1569 cells, that display normal CDK12 levels (see Figure 16) in the presence of *ErbB2* gene amplification (see Table 8), and BT474 cells that harbor concomitant *CDK12* and *ErbB2* gene amplification, provide an ideal setting to dissect, through genetic perturbation of CDK12 expression in silencing and overexpression experiments, the specific contribution of CDK12 in determining tumor phenotypes independently of ErbB2 alterations that frequently coexist in naturally occurring human breast cancers. We excluded the SKBR3 cell line from extensive functional validation because of their lack of tumorigenic potential *in vivo*, following xeno-transplantation into the mammary fat pad of NOD/SCID immunocompromised mice.

Therefore, using a variety of *in vitro* and *in vivo* biological assays, we set out to analyze the consequences of the genetic manipulations of CDK12, either by ablation or overexpression experiments, on several tumor phenotypes displayed by the selected cell-based models (see section 5.4 and 5.5).

Table 8. Source, clinical, and pathological features of breast cancer cell lines used in this study

Cell line	Gene cluster	ER	PR	HER2	TP53	Source	Tumor type	Age (years)	Ethnicity
AU565	Lu	-	[-]	+	+ WT	PE	AC	43	W
BT474	Lu	+	[+]	+	+	P.Br	IDC	60	W
BT483	Lu	+	[+]		-	P.Br	IDC, pap	23	W
HCC1428	Lu	+	[+]		[+]	PE	AC	49	W
HCC1569	BaA	-	[-]	+	-M	P.Br	MC	70	B
MCF10A	BaB	-	[-]		+/-WT	P.Br	F	36	W
MCF7	Lu	+	[+]		+/-WT	PE	IDC	69	W
MDAMB175	Lu	+	[-]		+/-WT	PE	IDC	56	B
MDAMB436	BaB	[-]	[-]		[-]	PE	IDC	43	W
MDAMB361	Lu	+	[-]	+	-WT	P.Br	AC	40	W
SKBR3	Lu	-	[-]	+	+	PE	AC	43	W
UACC812	Lu	+	[-]	+	-WT	P.Br	IDC	43	
ZR7530	Lu	+	[-]	+	-WT	AF	IDC	47	B

AC, adenocarcinoma; BaA, Basal A; BaB, Basal B; F, fibrocystic disease; IDC, invasive ductal carcinoma; Lu, luminal; MC, metaplastic carcinoma; P.Br, primary breast; PE, pleural effusion; W, White; B, Black.ER/PR/HER2/TP53 status: ER/PR positivity, ErbB2 overexpression, and TP53 protein levels and mutational status (obtained from the Sanger web site; M, mutant protein; WT, wild-type protein) are indicated. Expression data are as derived from [19]. Square brackets indicate that levels are inferred from mRNA levels alone where protein data is not available. Table adapted from [19]



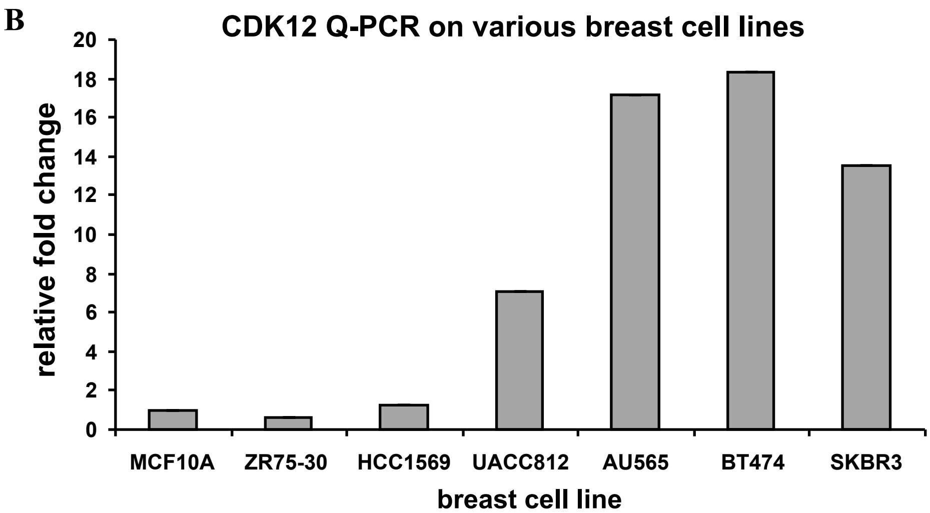


Figure 16. CDK12 expression analysis in human normal and cancer breast cell lines.

(A) Analysis of CDK12 protein expression in the indicated cell lines by immunoblot (IB) analysis. Total cell lysates (40µg of protein) from the indicated cell lines were resolved by SDS-PAGE and immunoblotted with the anti-CDK12 AQ19 monoclonal antibody produced in-house. In the same blot tubulin was detected as a loading control. Molecular weight markers are shown on the left. The blot was performed once.

(B) Analysis of CDK12 mRNA expression in the indicated cell lines by Q-PCR analysis (see Methods for details). The results are normalized to CDK12 mRNA levels in MCF10A cells and are expressed as the mean ± s.dev. Results are representative of 3 independent repeats.

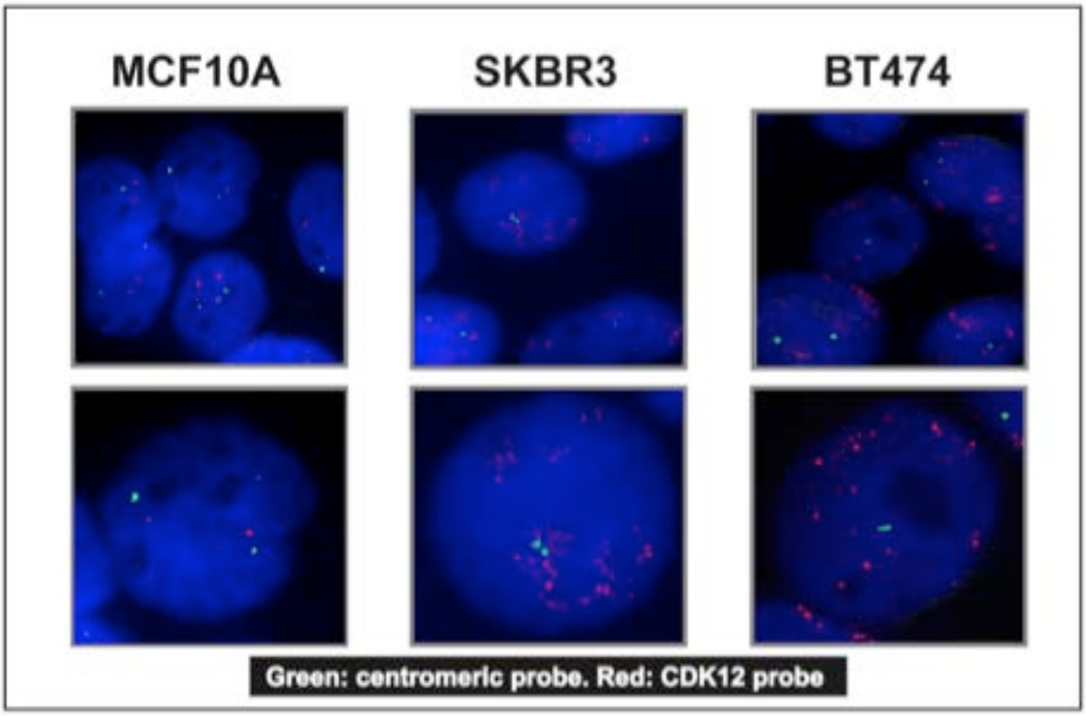


Figure 17. Analysis of *CDK12* amplification in human normal and cancer breast cell lines by FISH analysis.

Dual color interphase FISH analysis of CDK12 in MCF10A, SKBR3 and BT474 cells, using a centromeric probe for chromosome 17 (green) and a BAC clone encompassing the *CDK12* gene (red). The *CDK12* gene was considered to be amplified when the cut-off ratio of the *CDK12* signal (red spots) *vs.* the centromeric signal (green spots) was greater than 2.25. The cell nuclei were counterstained with DAPI (blue signal). Representative FISH images are shown at low (x40, upper panels) and high (x100, lower panels) magnifications.

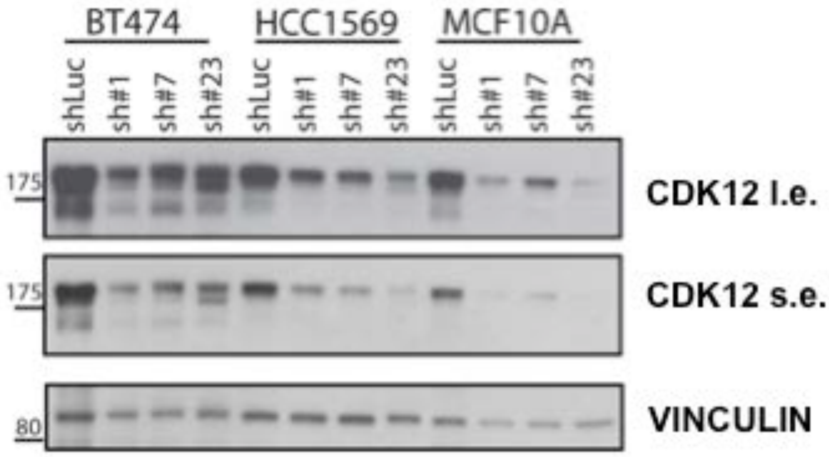
5.4 Investigations on the functional consequences of CDK12 ablation in normal and tumor breast cell lines.

5.4.1 Stable knockdown of CDK12 by lentiviral shRNA transduction

To stably knockdown (KD) CDK12 expression in the selected model cell lines (i.e., BT474, HCC1569 and MCF10A), we used the pSicoR lentiviral system (see Materials and Methods for details). This system allows efficient intracellular delivery and stable expression of small-hairpin(sh)RNA sequences to target genes of interest in mammalian cells. shRNA coding oligos specifically targeting CDK12 mRNA were cloned into the HpaI and XhoI restriction sites of the pSicoR construct. In order to guard against potential off-target effects, we used three different shRNA oligos (sh#1, sh#7, sh#23) for our *in vitro* functional studies. As a control, we used a pSicoR-shRNA construct targeting the Luciferase gene (shLuc), which is not present in mammalian cells.

Lentiviral particles, generated by transient co-transfection of the lentiviral construct with "packaging helper genes" in the recipient HEK293T cells, were used to infect the selected model cell lines. After selection, the infected cells were analyzed for CDK12 expression by IF and IB analysis. We observed that all three anti-CDK12 shRNA oligos yielded an efficient, albeit not complete, ablation of CDK12 expression in the various cell lines (Figure 18).

A



B

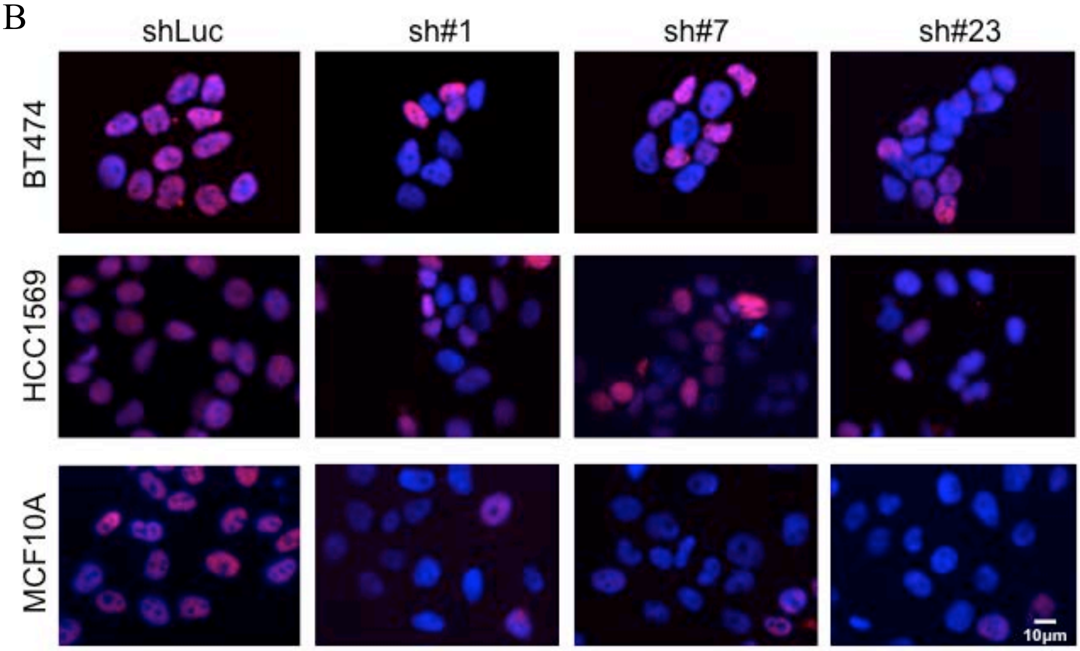


Figure 18. Characterization of CDK12 ablation in breast cell lines.

BT474, HCC1569 and MCF10A cells infected with a control lentiviral vector (shLuc) or with lentiviral vectors expressing shRNAs targeting CDK12 (sh#1, sh#7, sh#23) were analyzed for CDK12 expression both by IB (A) and IF (B). (A) Total cell lysates (40μg of protein) from the indicated cells were resolved by SDS-PAGE and immunoblotted with the anti-CDK12 AQ19 monoclonal antibody. Both short (CDK12 s.e.) and long exposures (CDK12 l.e.) of anti-CDK12 immunoblots are shown. Vinculin was used as a loading control. Molecular weight reference markers are shown on the left of the blots. The blot is representative of 3 independent experiments. (B) The indicated cells were fixed and stained with the AQ19 anti-CDK12 antibody followed by a Cy3-conjugated secondary antibody (Amersham) (red signal). Nuclei were counterstained with DAPI (blue signal). Images were obtained by fluorescence microscopy and show an overlay of red and blue fluorescence signals. Bar, 10 μm.

5.4.2 Effects of CDK12 ablation on the proliferative and clonogenic potential of breast cancer cell lines in 2D-adhesion culture conditions.

We analyzed the effect of stable CDK12 KD on the proliferative and clonogenic potential of BT474 and HCC1569 cells in 2D-adhesion culture conditions.

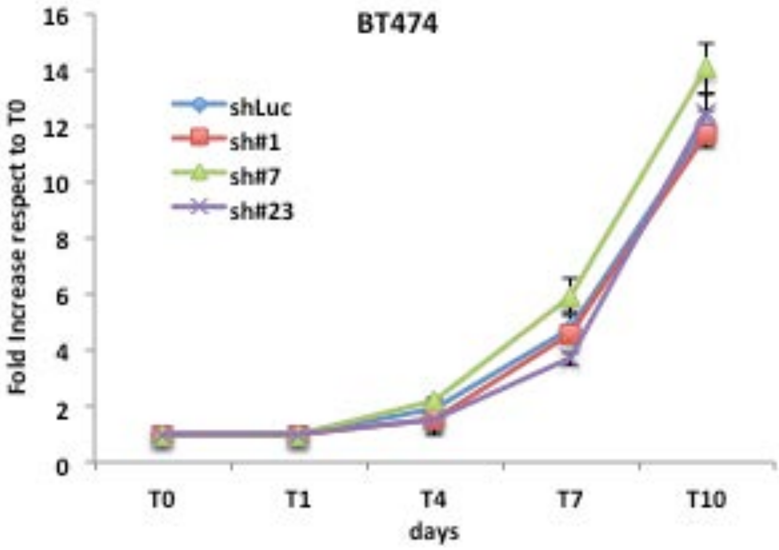
In the proliferation assay, $6x10^4$ cells were seeded in triplicate for each experimental point in 6-well tissue culture plates (T0). Cells were detached and counted after 24 h (T1) and then every 3 days (T4, T7, T10) for a period of ten days. For both BT474 and HCC1569 cells, we observed no significant differences in proliferation rates between controls (shLuc) and CDK12 KD (sh#1, sh#7, sh#23) cells (Figure 19).

For the colony forming assay, CDK12 KD and control cells were plated at clonogenic density ($\sim 1x10^4$ cells) in 10-cm tissue culture dishes in triplicate for each experimental point. After 20 days, colonies were fixed, stained with crystal violet and counted. In contrast to results obtained in the proliferation assay, we could observe a significant impairment of the colony forming efficiency consequent to CDK12 KD in BT474 cells: in particular, compared with the shLuc control cells, sh#1- and sh#7-cells displayed a ~60% decrease in the colony number (p-value < 0.05, Figure 20A), an effect that appeared to be less pronounced (~30%), albeit still statistically significant, in the case of sh#23-cells (p-value < 0.05, Figure 20A). Most likely, these variable effects on the colony forming ability are to be ascribed to partial differences in the silencing efficiency achieved with the various lentiviral vectors. In contrast, no significant differences in the colony forming efficiency were observed between control (shLuc) and CDK12-KD (sh#1, sh#7, sh#23) HCC1569 cells (Figure 20B).

Of note, from the analysis of data concerning the behavior of CDK12-silenced BT474 cells in 2D-functional studies *in vitro*, a discrepancy emerges between the absence of any significant effect in the short-term proliferation kinetics of CDK12-KD BT474 cells and the remarkable decrease observed in their long-term clonogenic potential compared to non-silenced control cells. The most straightforward explanation for these differences is that functional assays in which cells are subjected to more stressful culture conditions, as in the case of a 3D clonogenic assay [129] based on low serum content in the culture medium, very low density plating and prolonged incubation, are better suited to unmask cellular addiction to molecular pathways controlling cellular phenotypes required to increase cellular fitness and survival ability, thereby ultimately resulting in a proliferative advantage. Importantly, our global transcriptome analysis and ensuing pathway reconstruction has led to the identification of some important CDK12-regulated circuitries and phenotypes, such as EMT, proliferation and increased resistance to DNA damage, which might well account for the ability of CDK12 to increase cellular fitness of tumor cells.

Altogether, this first set of results suggest that the clonogenic ability of CDK12-overexpressing BT474 breast cancer cells depends on the presence of CDK12, the silencing of which, by contrast, does not cause any cell reproductive damaging effect in HCC1569 cells in which CDK12 is not overexpressed.

A



B

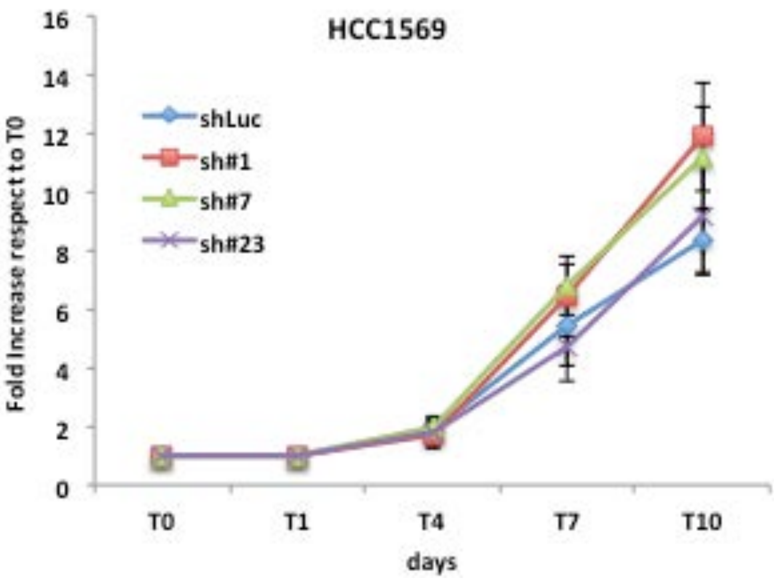
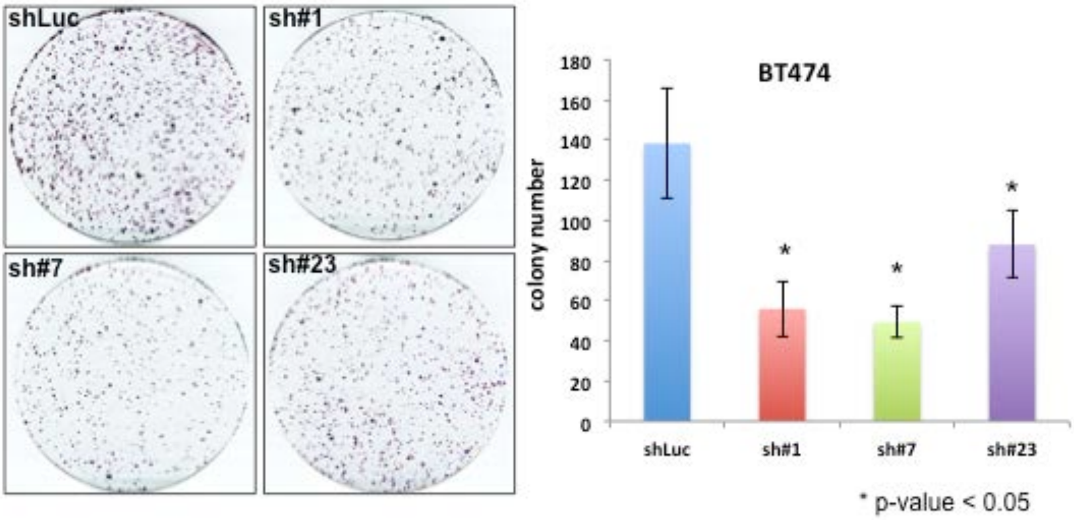


Figure 19. The effect of CDK12 KD on the proliferation of BT474 and HCC1569 cells in 2D-adhesion culture conditions.

(A-B) Growth curves for BT474 (A) and HCC1569 (B) cells infected with lentiviral vectors expressing control shRNA (shLuc) or anti-CDK12 shRNAs (sh#1, sh#7, sh#23) were generated by seeding 6x10^4 (T0) cells in triplicate in 6-well plates and counting cells after 24 h (T1) and then every 3 days (T4, T7, T10) for a period of ten days. Results are plotted as fold increase with respect to T0 and represent the mean ± s.dev. of two independent experiments performed in triplicate.

A



B

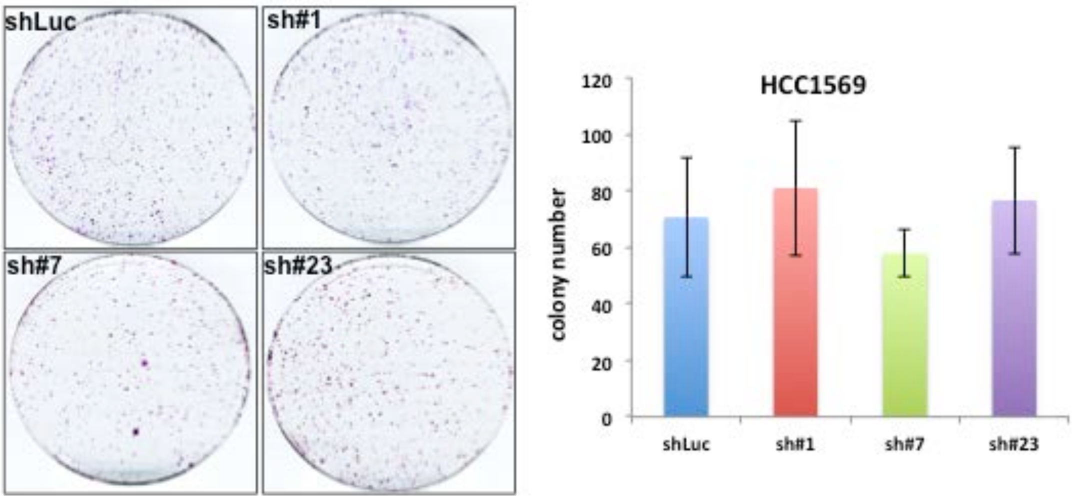


Figure 20. The effect of CDK12 KD on clonogenic potential of BT474 and HCC1569 cells in 2D-adhesion culture conditions.

(A-B) Control (shLuc) and CDK12 KD (sh#1, sh#7, sh#23) BT474 (A) and HCC1569 (B) cells ($1x10^4$) were plated in 10-cm tissue culture plates and stained after 20 days with crystal violet. Colony number was determined using the ImageJ software. Images of colonies from a representative experiment are shown on the left. Quantification of results is shown on the right and is expressed as the mean ± s.dev. from two independent experiments performed in triplicate. P-values were determined using the Student's t-Test.

5.4.3 Effect of CDK12 ablation on organotypic outgrowth of breast epithelial cells in three-dimensional basement membrane culture

Important microenvironmental signals are lost when cells are cultured *in vitro* on plastic substrata. Many of these crucial microenvironmental cues can, however, be restored using 3D-cultures of laminin-rich extracellular matrix extracts, such as Matrigel [130,131].

Mammary epithelial cells (MECs) including primary human and murine cells or human immortalized non-tumorigenic mammary epithelial cell lines, such as MCF10A, undergo a stereotypic morphogenetic program that culminates in the formation of clonally-derived, growth arrested, hollowed acini-like structures, which resemble the typical morphology of the normal mammary gland [132]. In contrast, transformed MECs form filled, overgrown/overbranched structures as typically observed with MCF10A cells transformed with ErbB2 or other known oncogenes [133].

Thus, the 3D-Matrigel assay measures the ability of cells to generate organotypic outgrowths *in vitro* that recapitulate the three-dimensional cytoarchitecture of a tissue, either normal or pathological. In this regard, compared to 2D-cultures, the 3D-Matrigel cultures represent a more physiologically relevant assay to study the involvement of a given protein in cell transformation processes that, besides involving aberrant proliferation, are also able to subvert morphogenetic programs.

We therefore employed the 3D-Matrigel assay to analyze the role of CDK12 in *in vitro* organogensis using the model cell lines, BT474, HCC1569 and MCF10A lentivirally delivered with control shRNA oligos (shLuc) or with different shRNA oligos to silence CDK12 (sh#1, sh#7 and sh#23). As expected, following infection with control shRNA oligos, MCF10A cells gave rise to hollowed acini-like structures, while

BT474 and HCC1569 cells originated filled overgrown structures typical of transformed MECs (Figure 21A).

Stable KD of CDK12 in MCF10A and HCC1569 cells (sh#1, sh#7 and sh#23) had no effect on the number and size of the clonally derived outgrowths typically generated by these cells (Figure 21A,B). In contrast, CDK12 KD BT474 cells, compared to their control counterpart, displayed a ~30 – 50% reduction in the number of organotypic outgrowths, which also appeared to be reduced in size compared to the structures generated by control BT474 cells (Figure 21C). This observation was confirmed by the distribution analysis of outgrowth diameters; we observed a significant shift of the CDK12 KD BT474 outgrowth box-plots towards lower diameter values compared to the box-plots derived from the analysis of their control counterpart, shLuc (Figure 22). This result indicates that CDK12 KD BT474 cells (sh#1, sh#7 and sh#23) form outgrowths that are significantly smaller compared to those formed by control shLuc BT474 cells (Figure 22).

In conclusion, data from the 3D-Matrigel assay indicate that CDK12 is essential for the ability of BT474 cells to sustain 3D-outgrowths *in vitro*, while it is dispensable in the formation of organotypic structures generated by HCC1569 or MCF10A cells.

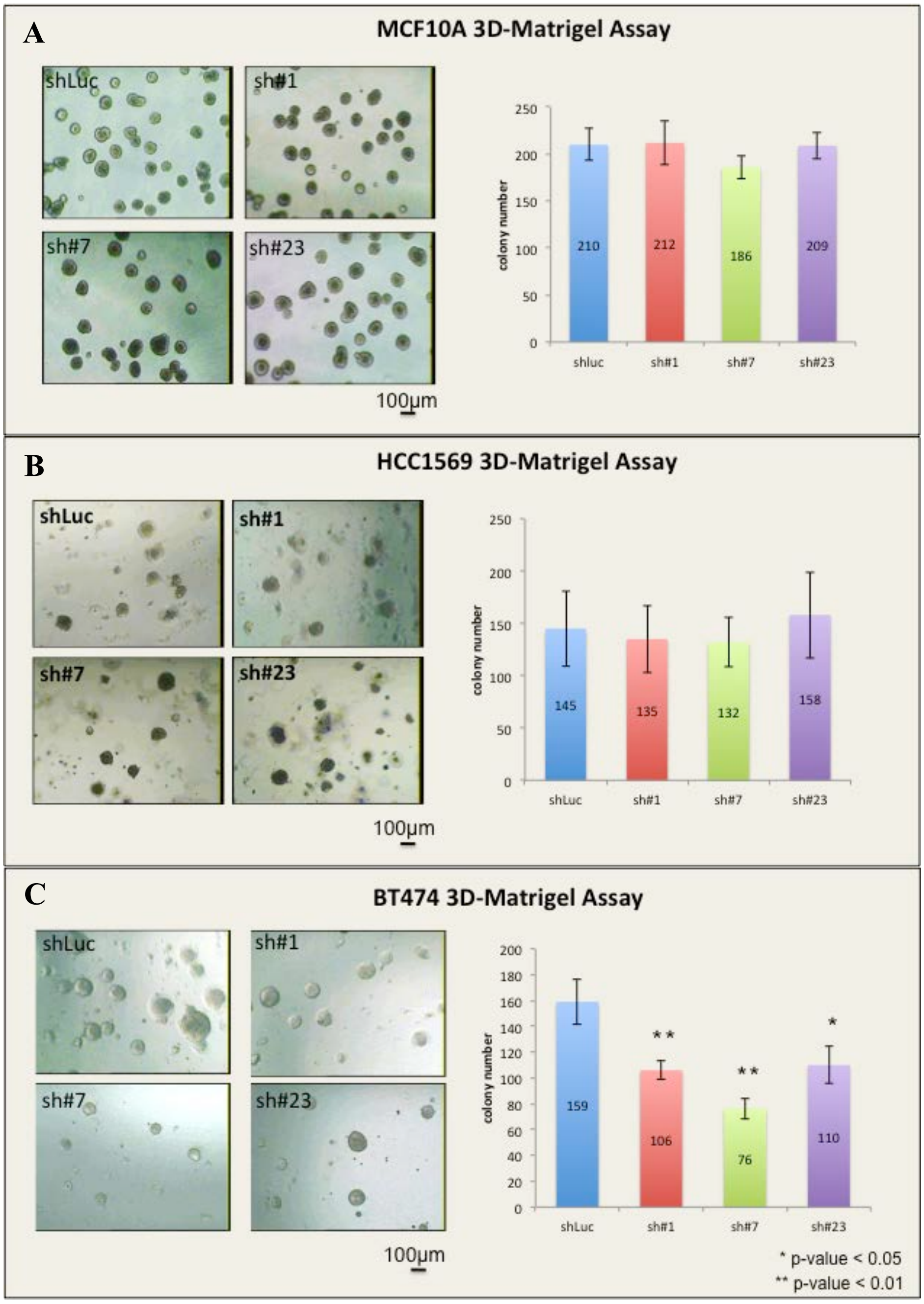
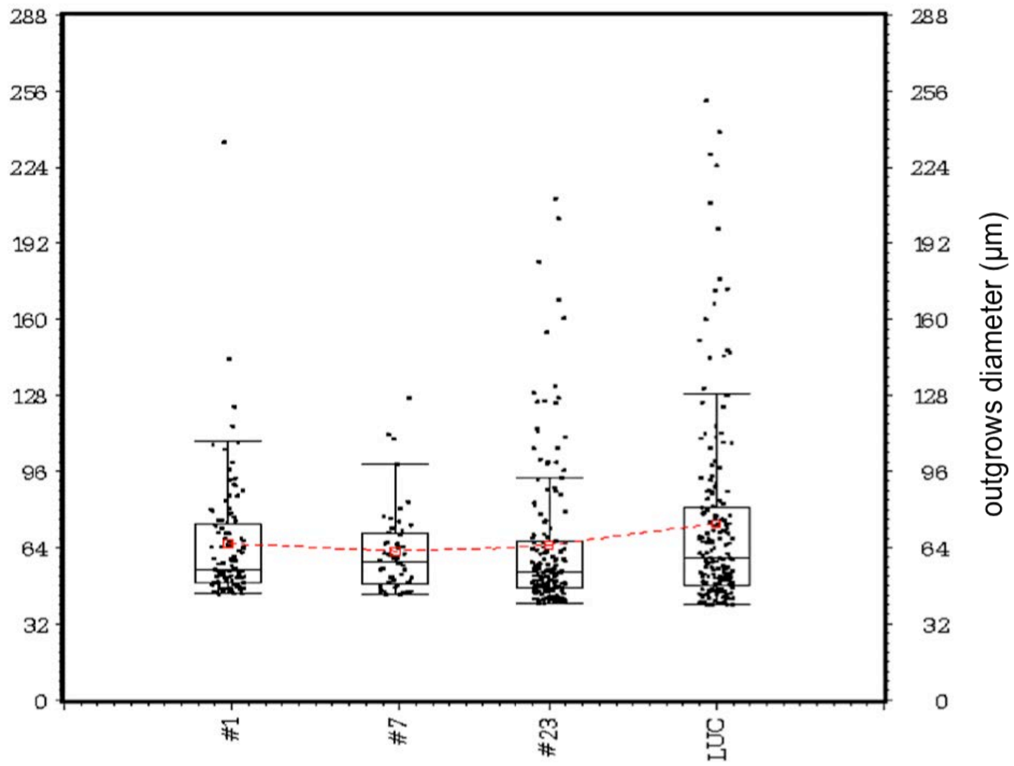


Figure 21. Effects of CDK12 ablation on the ability of BT474, HCC1569 and MCF10A cells to generate organotypic outgrowths in 3D-Matrigel.

(A) MCF10A, (B) HCC1569 and (C) BT474 control (shLuc) or stable CDK12 KD cells (sh#1, sh#7, sh#23) were seeded as single cell suspensions (1000 cells/ml) in Matrigel culture conditions and allowed to form organotypic structures for 20 days (as described in [134]). The number of outgrowths was determined by using the ImageJ software. Left: representative images of outgrowths are shown. Scale bar, 200 μm. Right: bar graphs showing the total number of outgrowths for each experimental condition. Values represent the mean ± s.dev. of two independent experiments performed in triplicate. P-values were determined using Student's t-Test. *-p-value<0.05, **p-value<0.01.



		N	Mean	SD	Median	Q1	Q3	Min	Max
LUC		159	74.2	40.7	59.5	48.2	81.0	40.2	252.0
#23	*	110	65.2	31.5	**53.8**	47.0	66.8	40.6	211.0
#7	*	76	58.2	18.7	**53.8**	44.6	65.6	40.2	122.4
#1	*	106	61.4	27.2	**50.5**	45.1	69.6	40.4	234.6

Figure 22. The effect of CDK12 KD in BT474 cells on outgrowth size in the 3D-Matrigel assay.

The diameter of outgrowths generated by BT474 control (shLuc) and CDK12 KD (sh#1, sh#7, sh#23) cells in 3D-Matrigel cultures was measured by using the image analysis software ImageJ. A) Outgrows diameter distribution is reported for a representative experiment of two independent repeats and represented as box-plot. B) Table of the box-plot values. Number of observations (N); Mean and Standard Deviation (SD); smallest observation (Min), lower quartile (Q1), median, upper quartile (Q3), and largest observation (Max) are reported; The asterisks indicate a significant p-value (<0.05) calculated for the Median values of each experimental point (sh#1, sh#7, sh#23) vs the control (LUC). P-values were calculated with Kruskal-Wallis test. All statistical analysis and box-plot representation are obtained using the statistical analysis software "SAS".

5.4.4 *In vivo* analysis of CDK12 ablation in BT474 cells.

Overall results from the 2D-clonogenic and 3D-Matrigel assay pointed to a crucial role for CDK12 in mediating the tumorigenic potential of BT474 cells. Based on this, we set out to directly prove the contribution of CDK12 in sustaining tumorigenesis *in vivo* by performing hetero-transplantation of control and CDK12 KD BT474 cells into immunodeficient mice. To this aim, $1,5x10^5$ stable CDK12 KD BT474 cells (sh#1 and sh#7) were xenografted orthotopically into the inguinal mammary fat pad of NOD/SCID IL2R gamma-chain null female mice. As a control, the same number of shLuc BT474 cells was injected into the contralateral mammary gland of the same mice. After 6 weeks, mice were sacrificed and tumors were explanted and weighed. Tumor explants were also lysed for protein extraction and IB analysis to control for the efficiency of CDK12 KD. We observed that tumor outgrowths generated by CDK12 KD BT474 cells were ~50% smaller in size compared to tumors generated by control BT474 cells (Figure 23A). IB analysis confirmed that CDK12 was efficiently silenced in the tumor masses generated by CDK12 KD BT474 cells (Figure 23B).

Altogether, these results indicated that CDK12 ablation does not significantly affects the outgrowths of either non-tumorigenic MCF10A cells or HCC1569 cancer cells without CDK12 overexpression. Conversely, CDK12 ablation selectively impaired the growth *in vitro* and the tumorigenic potential *in vivo* of BT474 cells harboring *CDK12* amplification and overexpression (Figure 22 and Figure 23). We therefore concluded that BT474 cells are addicted to CDK12 overexpression for the maintenance of their malignant phenotype.

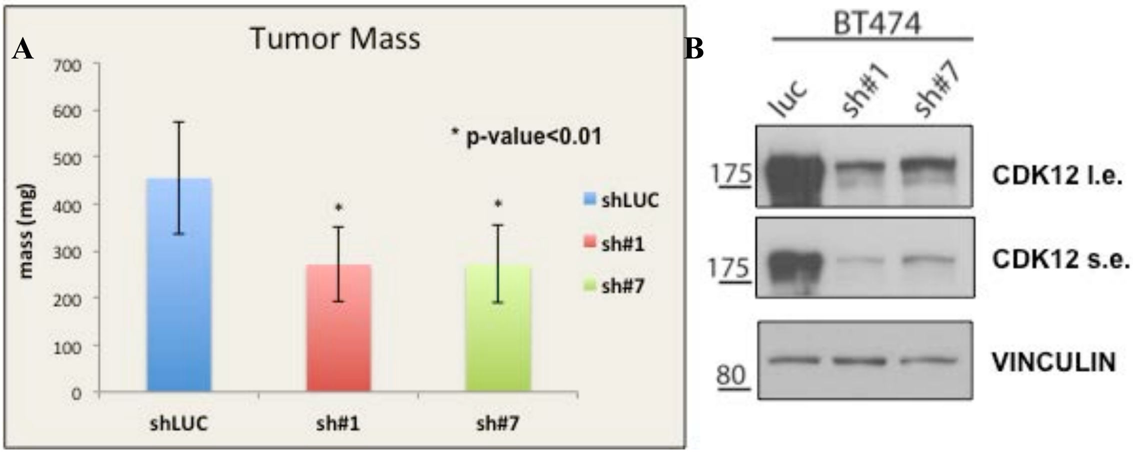


Figure 23. Effect of CDK12 ablation on the ability of BT474 cells to generate tumors *in vivo*.

Mammary tumors were generated by injecting $1,5x10^5$ CDK12 KD (sh#1 and sh#7) and control (shLuc) BT474 cells into the inguinal mammary fat pads of NOD/SCID IL2 gamma-chain null mice. Tumors were grown for 6 weeks before animals were sacrificed and tumors explanted and weighed. (A) Tumor mass is reported in the bar graph. Results represent the mean ± s.dev. (n=12) of 3 independent experiments in which 4 injections per sample were performed in each experiment. P-values were determined using Student's t-Test; * p-value < 0.01. (B) Analysis of CDK12 expression in mammary tumors by IB analysis. Images of both short exposure (CDK12 s.e.) and long exposure (CDK12 l.e.) anti-CDK12 immunoblots are reported. Vinculin was used as a loading control. Molecular weight markers are shown to the left of the blots.

5.5 Functional analysis of CDK12 overexpression in breast cell lines.

In parallel to functional ablation studies, we analyzed the consequences of CDK12 overexpression in either normal or tumor breast epithelial cells by exploiting the same *in vitro* and *in vivo* biological "read-outs" described in the previous section (i.e., 2D-cell proliferation assay, 2D-colony formation assay, 3D-*in vitro* organogenesis assay, *in vivo* tumorigenicity assessed by xenograft experiments). We used MCF10A and HCC1569 cells as cellular model systems of normal and tumor mammary epithelial cells, respectively, that feature low basal levels of CDK12 compared to breast tumor cell lines displaying CDK12 overexpression due to gene amplification (i.e., SKBR3 and BT474) (Figure 16).

5.5.1 Development of a lentiviral transduction-based strategy for the efficient and stable overexpression of CDK12 in target cells

We generated a lentiviral construct to stably overexpress CDK12 in mammalian cells by cloning the human CDK12 cDNA into the pLVX lentiviral vector. This system allows achieving high and constitutive expression of genes of interest, due to the presence of the human cytomegalovirus immediate early promoter. Cells were lentivirally transduced using a pLVX empty vector (EV) as a negative control or with a pLVX-CDK12 construct to yield stable overexpression of the protein. Stably infected MCF10A (MCF10A-EV/-CDK12) and HCC1569 (HCC1569-EV/-CDK12) cells were analyzed for CDK12 expression at the protein level. As evidenced by IB analysis, lentiviral-mediated overexpression of CDK12 yielded, both in MCF10A and in HCC1569 cells, CDK12 protein levels comparable to those observed in *CDK12*-amplified BT474 cells (Figure 24A, C). However, IF analysis of CDK12 expression showed that a percentage of ~30-40%, both for MCF10A and HCC1569 cells, showed high CDK12 expression levels, a result that might have detrimentally affected some of the analysis described below (Figure 24B, D).

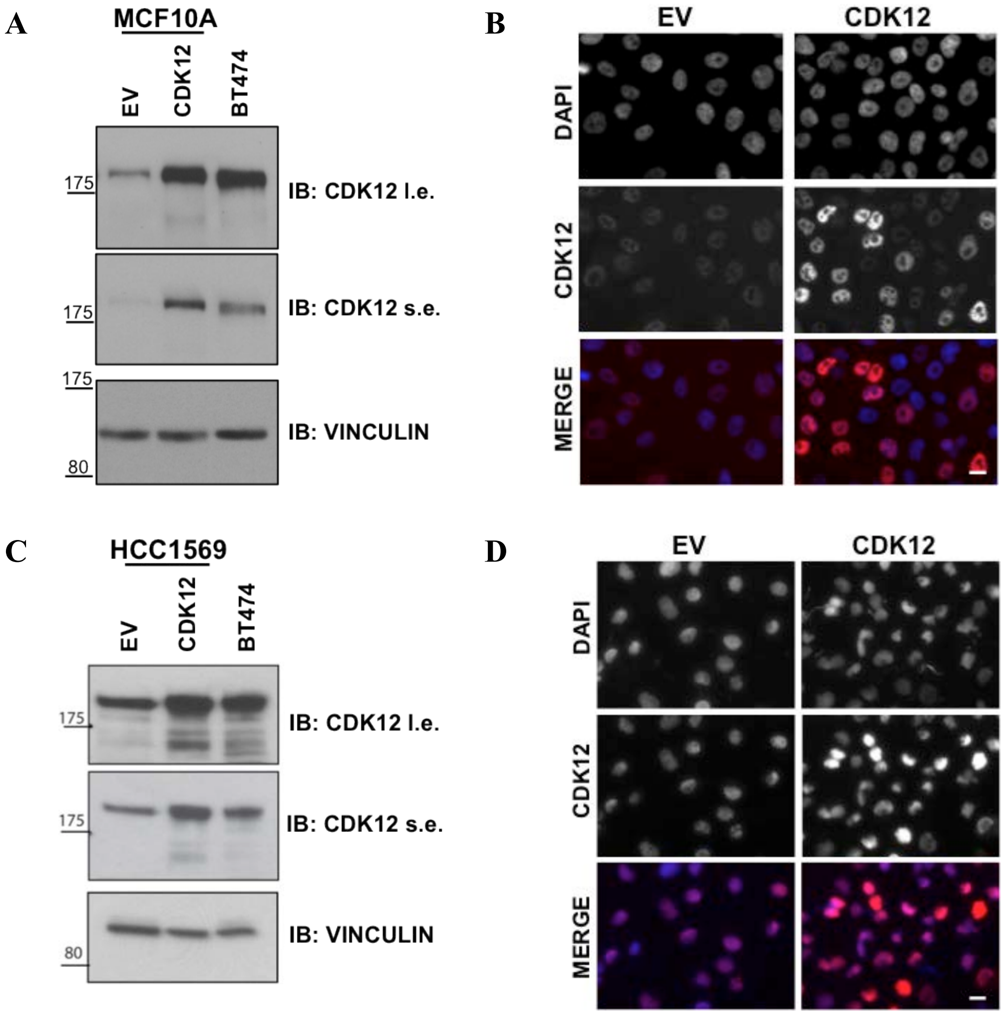


Figure 24. Analysis of CDK12 overexpression in stably transduced MCF10A-CDK12 and HCC1569-CDK12 cells.

(A, C) Immunoblot analysis of CDK12 expression in total cell lysates (40μg) of MCF10A and HCC1569 cells after infection with pLVX-CDK12 or with pLVX-EV as a negative control. Both short (CDK12 s.e.) and long (CDK12 l.e.) exposure images of CDK12 immunoblots are reported. BT474 total cell lysates were used as a control for CDK12 overexpressing cells. Vinculin was detected as a protein loading control. Molecular weight reference markers are reported on the left of the blots. (B, D) Analysis of CDK12 expression by IF. Upper panels show the signal relative to the nuclei (DAPI); middle panels show the signal relative to CDK12; bottom panels show the merge of DAPI and CDK12 (red signal: CDK12; Blue signal: DAPI). Scale bar, 10μm. Immunoblots and IF images are representative of 2 experimental repeats.

5.5.2 Functional characterization of CDK12 overexpression in MCF10A cells

To determine whether CDK12 overexpression is sufficient *per se* to malignantly transform normal mammary epithelial cells, we analyzed cell proliferation, clonogenic ability and organotipyc outgrows *in vitro* using MCF10A cells.

We started with the analysis of the proliferation rate of CDK12-overexpressing MCF10A cells compared to MCF10A-EV cells, as a control. Remarkably, MCF10A-CDK12 cells showed a 3-fold increase in the number of cells counted after a period of 7 days in culture, compared to control MCF10A-EV cells (Figure 25A). Likewise, in the 2D-colony forming assay, MCF10A-CDK12 cells showed a ~60% increase in the colony number over control cells (Figure 25B).

By contrast, in 3D-Matrigel organotypic cultures, no evident differences, in terms of gross morphology, size, and number of outgrowths, were detected between CDK12-overexpressing MCF10A cells and their control counterpart. Indeed, organotypic outgrowths formed by MCF10A-CDK12 cells showed no morphological signs of malignant transformation, and closely resembled the stereotypical acini-like structures expected of normal mammary epithelial cells [128]. In contrast, MCF10A cells infected with activated oncogenes such as Neu-T and RasV12, which were used as internal positive controls for genetic lesions able to induce overt malignant conversion, generated aberrant overgrown and hyperbranched structures typical of transformed cells (Figure 26)[133].

Consistent with these results *in vitro*, MCF10A-CDK12 cells did not show any tumorigenic potential *in vivo* following xeno-transplantation into immunocompromised mice. For these xenograft experiments, increasing concentrations (0.1×10^6, 1×10^6, and 10×10^6) of MCF10A-CDK12 and MCF10A-EV cells were injected orthotopically into the controlateral inguinal mammary fat pads of

NOD/SCID IL-2R gamma chain null female immunocompromised mice. Neither MF10A-CDK12 nor MCF10A-EV cells gave origin to any palpable tumor formations up to 6 months from injection.

Based on these data, we concluded that, while CDK12 overexpression is able to confer a proliferative advantage to mammary epithelial cells, it is not sufficient *per se* to promote overt malignant transformation of these cells.

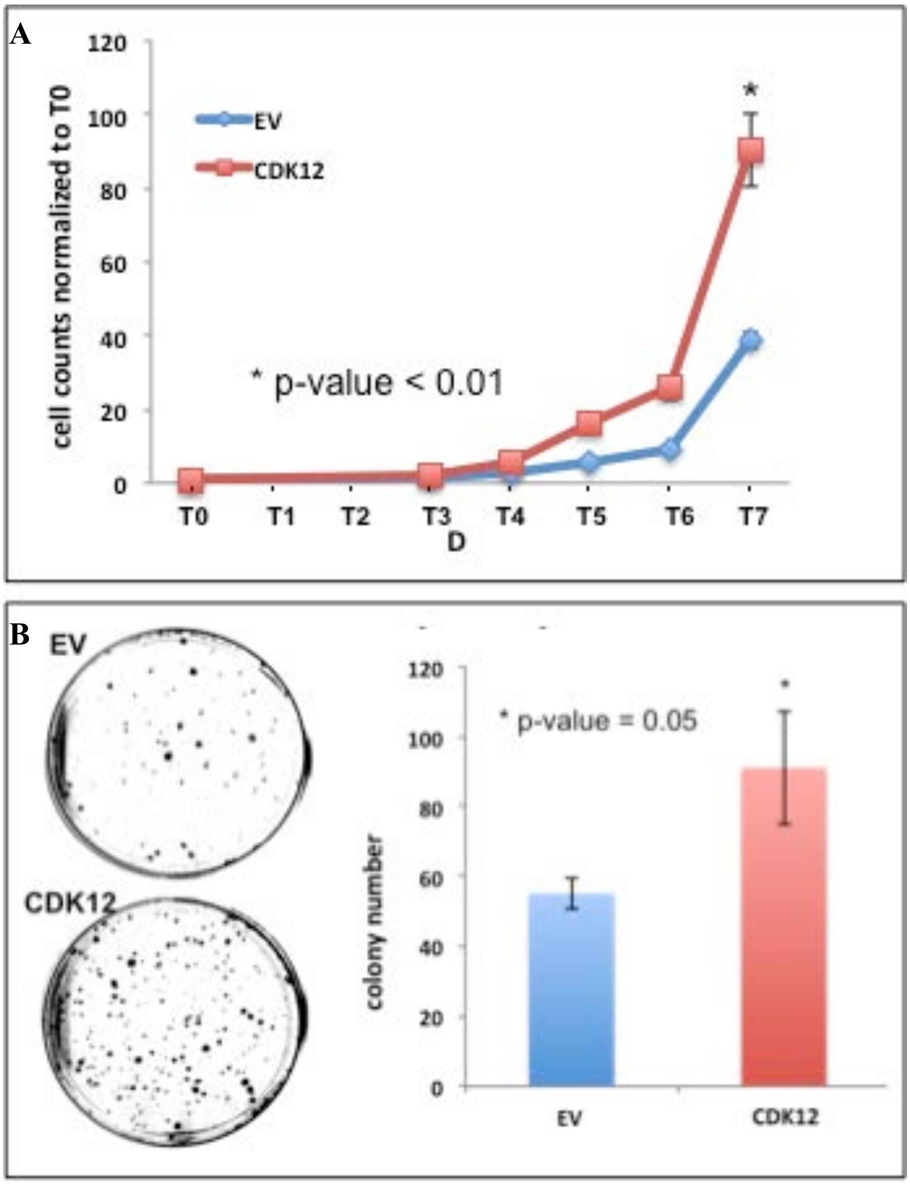


Figure 25. Effects of CDK12 overexpression on proliferation and clonogenic ability of MCF10A cells in 2D culture.

(A) Growth of MCF10A-CDK12 cells compared to control MCF10A-EV cells. Cells ($2x10^4$) were seeded in triplicate, for each time point, in 6-well tissue culture plates and grown over 7 days . At the indicated time-points (D=day), cells were detached and counted using a hemocytometer. Graph reports cell counts normalized to T0.

(B) Colony-forming ability of CDK12-MCF10A cells compared to control MCF10A-EV cells. Cells were plated on tissue culture plates at clonogenic density (500 cells/10-cm plate) and stained after 10 days with crystal violet. Colony number was determined using the ImageJ software. Left: images of a representative experiment are shown. Right: quantification of the colony number/plate across the different experiments.

Results in A and B are shown as the mean ± s.dev. of three independent experiments performed in triplicate. P-values were determined using Student's t-Test.

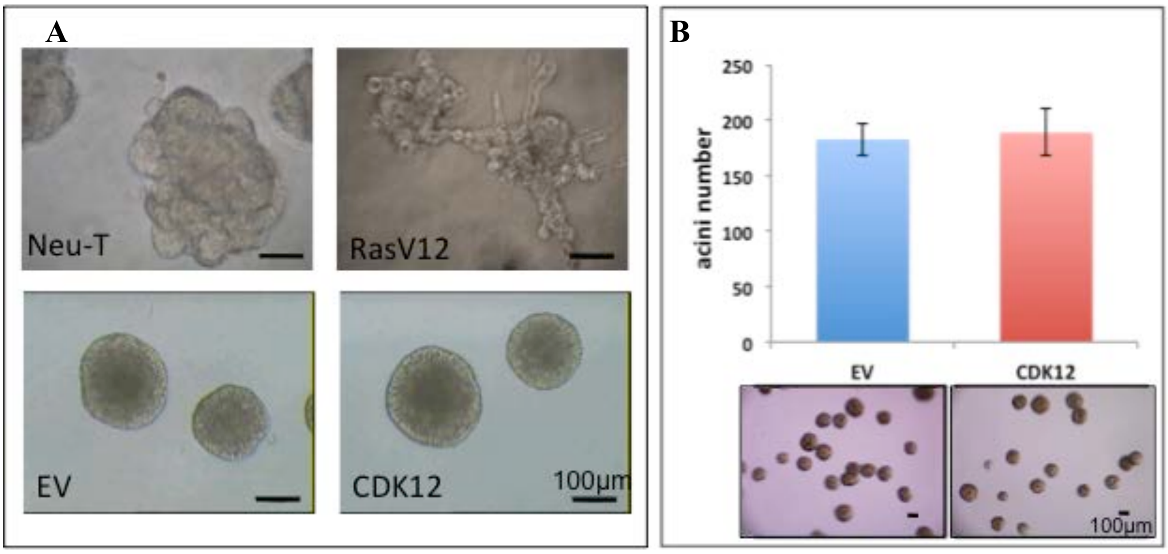


Figure 26. Effects of CDK12 overexpression on the organotypic outgrowth of MCF10A cells in 3D-Matrigel.

MCF10A-CDK12 and –EV cells ($1x10^4$) were plated in 4-well chamber slides in 3D-Matrigel overlay conditions [134] and cultured for 15 days before analysis of outgrowths.

(A) Representative images of outgrowths generated by MCF10A-CDK12 and –EV cells in 3D-Matrigel cultures. MCF10A cells stably infected with the Neu-T or RasV12 oncogene were used as positive controls of transformed mammary epithelial cells. Scale bar, 100 μm.

(B) Top: quantification of the number of outgrowths generated by MCF10A-CDK12 and –EV cells in 3D-Matrigel. Bottom: representative images of outgrowths. Scale bar, 100 μm. Data are reported as the mean ± s.dev. of two independent experiments performed in quadruplicate.

5.5.3 *In vitro* functional characterization of CDK12 overexpression in HCC1569 cells

Data obtained from CDK12 enforced expression in MCF10A cells indicated that, albeit able to increase their proliferation rate, CDK12 is not sufficient, as a single hit, to cause malignant transformation of normal cells. This is consistent with the concept of stepwise carcinogenesis which holds that multiple genetic lesions are required for the onset and progression of a tumor[56]. We therefore set out to investigate whether CDK12 deregulation, once occurred in a tumor background, might represent one of those genetic lesions able to confer a more aggressive behavior to already transformed breast cancer cells.

To this aim, we used HCC1569 cells that, in our initial screening (see also Figure 6 and Figure 8, displayed CDK12 levels comparable to those observed in the normal MCF10A cell line. As described in the case of MCF10A-CDK12 cells (see Figure 15), lentiviral-mediated enforced overexpressiong of CDK12 in HCC1569 cells resulted in a higher proliferation rate and enhanced colony-forming ability in 2D-culture compared with HCC1569 cells infected with an empty vector (HCC1569-EV), as a control (Figure 27A, B).

In particular, in the 2D-proliferation assay, we observed a 4-fold increase in the number of HCC1569-CDK12 cells counted after a period of 10 days in culture, compared to control HCC1569-EV cells (Figure 27A). Likewise, when plated at clonogenic density in adhesion conditions for 20 days, HCC1569-CDK12 cells displayed a ~40% increase both in the number and in the average size of colonies generated compared to control cells (Figure 27B).

Moreover, at variance with the lack of any effect observed in 3D-Matrigel cultures with CDK12-overexpressing MCF10A cells, enforced CDK12 expression in HCC1569 cells resulted in a ~2-fold increase in the number of 3D tumor outgrowths, which also showed a ~2-fold increase in their average colony size when compared to control HCC1569-EV cells (Figure 28).

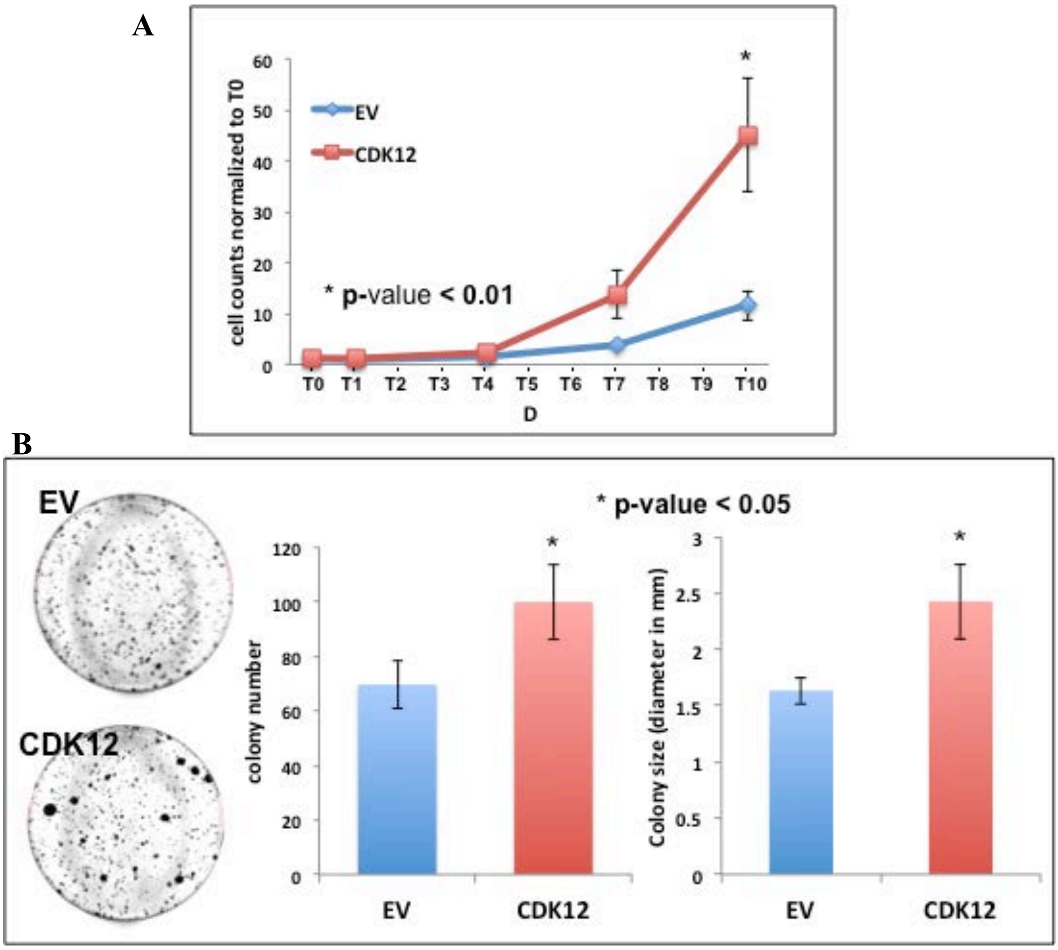


Figure 27. CDK12 overexpression increases the proliferative potential of HCC1569 cells.

(A) Growth of HCC1569-CDK12 cells compared to control HCC1569-EV cells. Cells ($2x10^4$) were seeded in triplicate, for each time-point, in 6-well tissue culture plates and grown in adhesion conditions over a 10-day period (D=days). At the indicated time-points, cells were detached and counted using a hemocytometer. Graph reports cell counts normalized to T0.

(B) Colony-forming ability of HCC1569-CDK12 cells compared to control HCC1569-EV cells. Cells were plated at clonogenic density (5,000 cells/10-cm plate) and stained after 20 days with crystal violet. The colony number and average colony-size were determined using the ImageJ software. Left: images of a representative experiment are shown. Right: quantification of colony number and colony size are shown.

Results in A and B represent the mean ± s.dev. of three independent experiments performed in triplicate. P-values were determined using Student's t-Test.

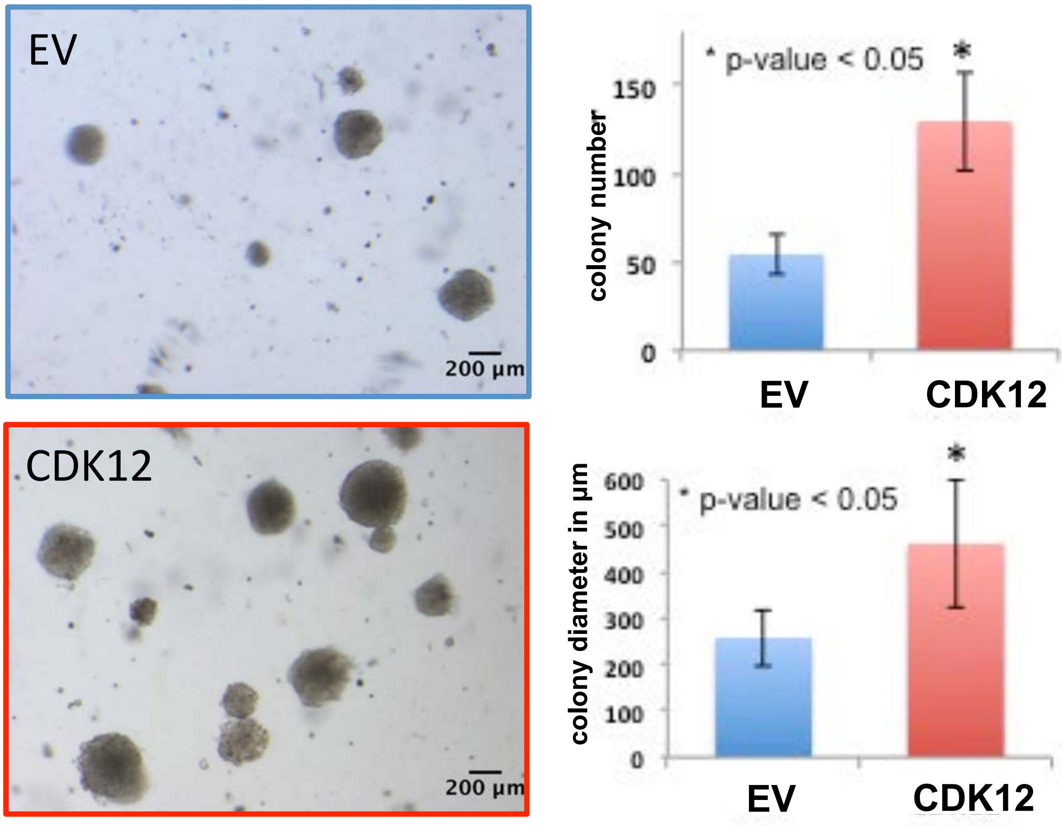


Figure 28. Effects of CDK12 overexpression in HCC1569 cells grown in 3D-Matrigel.

HCC1569-CDK12 and –EV cells (1x10^4) were grown in 4-well chamber slides as Matrigel-embedded 3D cultures for 20 days [134] to generate organotypic outgrowths.

Representative images of 3D structures generated by HCC1569-CDK12 and –EV cells are reported on the left. Scale bar,, 200 µm. Colony number and average colony size were determined by using the ImageJ software and are shown on the right. Results represent the mean ± s.dev. of three independent experiments performed in quadruplicate. P-values were determined using Student's t-Test.

5.5.4 *In vivo* analysis of the effects of CDK12 overexpression in HCC1569 cells

To complement and corroborate the results obtained *in vitro,* we analyzed the effects of CDK12 overexpression on the tumorigenic ability of HCC1569 cells *in vivo*. To this purpose, we injected HCC1569-CDK12 cells orthotopically in one of the two inguinal mammary fat pads of individual NOD/SCID IL-2R gamma chain null mice, while control HCC1569-EV cells were injected into the contralateral gland of the same mice. Tumor growth was monitored over a period of 6 weeks before mice were sacrificed and tumors explanted and weighed. Data analysis from these xenograft experiments showed that CDK12 overexpression in HCC1569 cells resulted in a dramatically enhanced tumor formation, with tumor volume and tumor mass at 6 weeks being, respectively, ~8-fold and ~6-fold greater compared to tumors generated by control HCC1569 cells (Figure 29A-C). Additionally, we confirmed by IHC analysis that high levels of CDK12 expression were maintained in HCC1569-CDK12 tumors (Figure 29D).

Altogether, *in vitro* and *in vivo* experiments converge on the evidence that CDK12 overexpression leads to a worsening of the behavior of fully transformed tumor cells, an event that might be relevant to the tumor progression of naturally occurring human breast tumors, as well.

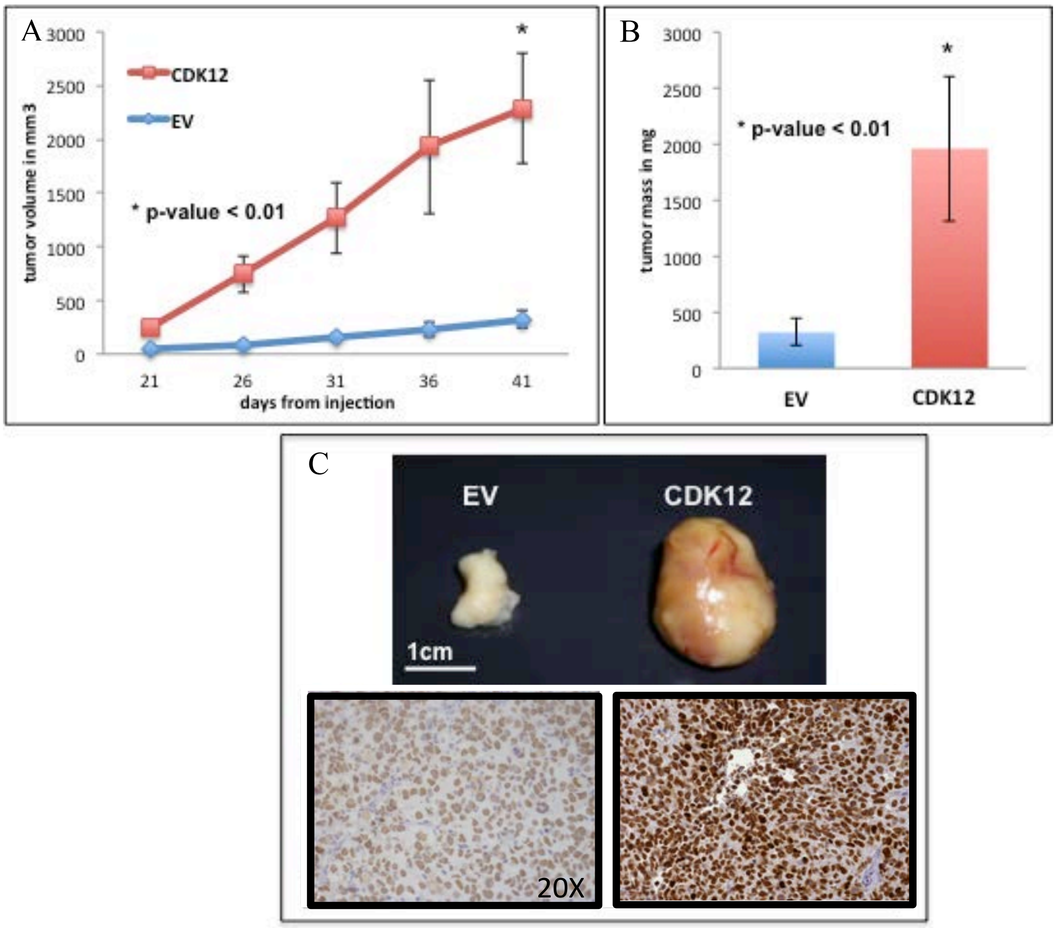


Figure 29. CDK12 overexpression increases the tumorigenic potential of HCC1569 cells *in vivo*.

Mammary tumors were generated by injecting 5x10^5 HCC1569-CDK12 and -EV cells into the inguinal mammary fat pads of NOD/SCID IL-2R gamma chain null mice. Tumors were allowed to develop for 6 weeks before animals were sacrificed and tumors explanted and weighed.

(A) Kinetic evaluation of HCC1569-CDK12 *vs*. HCC1569-EV tumor growth. Tumor volume was assessed by *in vivo* caliper measurements at the indicated time-points after injection.

(B) Quantification of the mass of HCC1569-CDK12 and HCC1569-EV tumors at time of explant, 6 weeks after injection.

(C) Upper panel: Representative images of tumor outgrowths obtained with HCC1569-EV/-CDK12 cells. Scale bar, 1cm. Bottom panel: IHC analysis of CDK12 expression in HCC1569-CDK12 and HCC1569-EV tumors. Magnification, x20.

Results reported in A and B are collated from 2 independent experiments and represent the mean ± s.dev., n=10. P-values were determined using Student's t-Test.

5.5.5 CDK12 overexpression in HCC1569 cells induces EMT

Overall results from our functional studies demonstrate that CDK12 overexpression is able to worsen the tumor phenotype of transformed HCC1569 mammary tumor cells.

A variety of mechanisms downstream of CDK12 dysfunction are most likely implicated in this event. However, one major change that we could reproducibly observe was the gradual acquisition in CDK12-overexpressing HCC1569 cells maintained in long-term culture (up to 8 weeks) of phenotypic traits, primarily consisting in the acquisition of a spindle-like appearance at the expense of the typical cobblestone like epithelial morphology, compatible with the occurrence of epithelial-to-mesenchymal transition (EMT). By contrast, none of these morphological alterations reminiscent of EMT could be ever detected in control HCC1569-EV cells (Figure 30A). To directly address the possibility that aberrant CDK12 levels might induce EMT, we tested the expression of two well-established markers, which are typically inversely regulated during EMT, namely E-cadherin and N-cadherin. IF and IB analysis revealed that, compared to their control counterpart, HCC1569-CDK12 cells displayed a gradual increase in the expression of the EMT marker N-cadherin, accompanied by a concomitant decrease in the expression of the epithelial marker E-cadherin (Figure 30B.). Of note, loss of E-cadherin in favor of N-cadherin expression was also observed in freshly dissociated cells from HCC1569-CDK12 tumor xenografts (Figure 30B,C), indicating that EMT, rather than being an artifactual consequence of the long-term culture *in vitro*, represents an event naturally occurring *in vivo*.

Based on these results, we concluded that the induction of EMT might be one of the mechanisms through which CDK12 promotes a more aggressive behavior in transformed mammary epithelial cells, and, as such, could contribute to the more aggressive evolution of CDK12-overexpressing breast cancers.

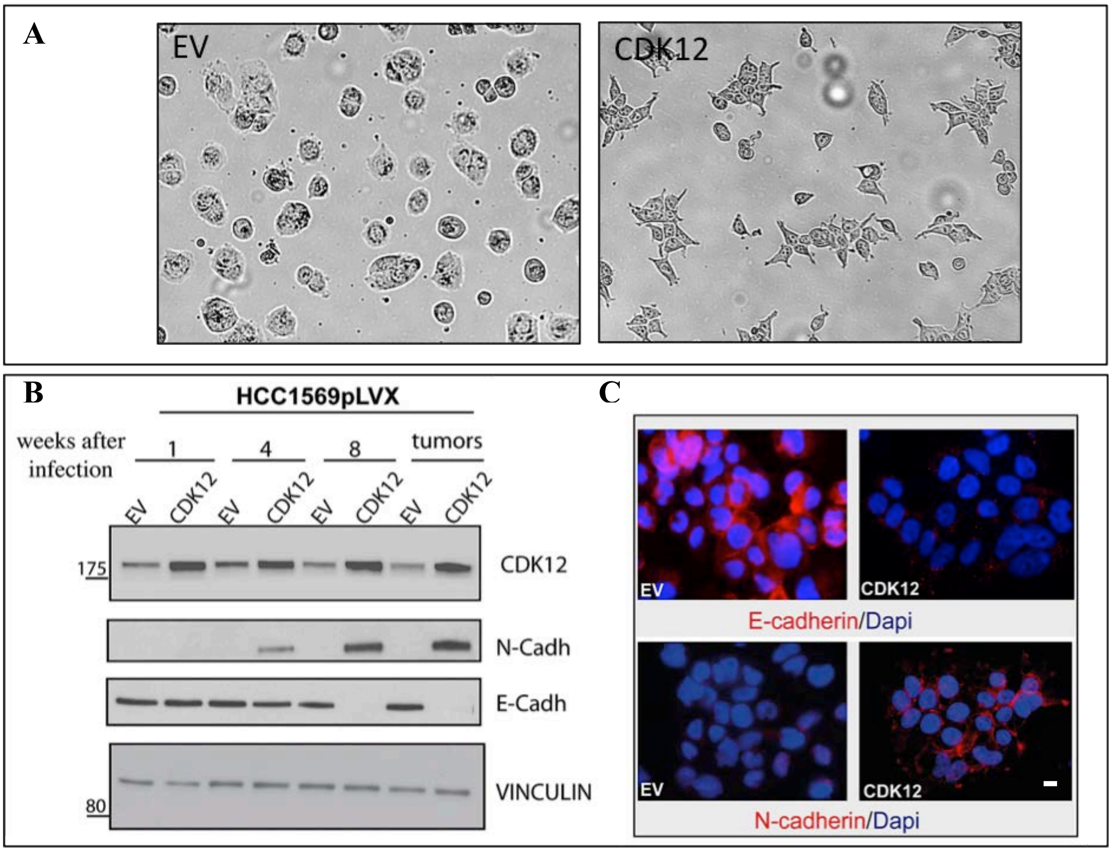


Figure 30. CDK12 overexpression induces EMT in HCC1569 breast tumor cells.

(A) Representative images of HCC1569 cells stably transduced with a lentiviral CDK12 vector (CDK12) or with a control empty vector (EV) and cultivated in adhesion for 8 weeks after infection. Magnification, x20.

(B) WB analysis of CDK12, E-cadherin and N-cadherin expression in HCC1569-CDK12 and -EV cells at 1, 4 and 8 weeks after infection, and in freshly dissociated tumors.

(C) Immunofluorescence analysis of E-cadherin and N-cadherin expression in HCC1569-CDK12 and -EV cells. Images represent the merge of nuclei staining with DAPI (blue signal) and E-cadherin in the upper panel, or N-cadherin in the bottom panel (red). Scale bar, 10μm.

5.5.6 Analysis of the dependency of tumor phenotypes on CDK12 overexpression

The *in vitro* and *in vivo* characterization of the functional consequences of CDK12 overexpression in breast cell lines strongly suggests that CDK12 has an oncogenic role in breast cancer (see Result Sections 5.5.3 and 5.5.4).

To further corroborate this evidence, we decided to perform *in vitro* and *in vivo* tumor reversion phenotype experiments based on CDK12 ablation in CDK12-overexpressing HCC1569 cells. The ultimate goal of this approach was to prove the intrinsic dependency on the continuous presence of CDK12 of the various phenotypic traits associated with CDK12 overexpression. To this aim, we exploited the Cre/Lox-inducible pSico lentiviral system [135]. In this system, Cre-induced recombination results in reconstitution of a functional promoter by the removal of a lox-flanked stop cassette, which permits the transcription of downstream sequences encoding a functional shRNA. In addition, the pSico vector constitutively expresses an enhanced-GFP (EGFP) reporter gene as selection marker. We engineered a pSico vector encoding an shRNA targeting CDK12 (pSico-shCDK12) and a control vector encoding an shRNA targeting Luciferase (pSico-shLuc).

HCC1569-CDK12 cells were infected with these vectors, and then cells were sorted by fluorescence activated cell sorting (FACS) into EGFP-high and EGFP-low subpopulations. EGFP intensity directly correlates with the number of pSico lentiviral transgenes inserted into the genome of infected cells and, as a consequence, with the efficiency of CDK12 silencing obtained upon Cre recombination. To achieve conditional silencing of CDK12 expression, stably shCDK12- and shLuc-infected cells were then transduced with a Tat-Cre fusion protein (kindly provided by the Antibody and Protein Facility at IFOM [17]) to promote recombination. To determine the efficiency of CDK12 KD in the EGFP-high and EGFP-low subpopulations, we analyzed

CDK12 expression by IB. The EGFP-high shCDK12 cell population showed an efficient, but not complete, KD of CDK12 compared to control EGFP-high/-low shLuc cells, while only a minimal KD was achieved in EGFP-low shCDK12 cells (Figure 31). We therefore used these cell subpopulations to investigate whether a reduction in CDK12 expression could induce reversion of tumorigenic phenotypes observed in CDK12-overexpressing HCC1569 cells by performing *in vitro* colony-forming and 3D-Matrigel organogenesis assays, and by xeno-transplantation *in vivo*.

Initially, we analyzed the colony-forming ability of EGFP-high *vs.* -low shCDK12 and shLuc cells *in vitro* in adhesion culture conditions. We observed a ~65% reduction in the number of colonies formed by EGFP-high shCDK12 cells compared with both EGFP-high and -low shLuc cells, while no effects were observed in EGFP-low shCDK12 cells (Figure 32). Similar results were reiterated in the 3D-Matrigel organogenesis assay, in which we observed a ~45% reduction in outgrowths generated by EGFP-high shCDK12 cells compared with EGFP-high/-low shLuc control cells and with EGFP-low shCDK12 cells (Figure 33). These results demonstrate that CDK12 silencing in stably CDK12-overexpressing HCC1569 cells is able to revert, in a dose-dependent fashion, the enhanced colony-forming ability and 3D-outgrowth potential acquired by these cells upon CDK12 overexpression.

Finally, we analyzed whether a reduction in CDK12 levels could revert the increased tumorigenic potential of CDK12-overexpressing HCC1569 cells observed *in vivo*. We injected HCC1569-CDK12 cells stably infected with pSico-shCDK12, or with pSico-shLuc as a control, into the mammary fat pads of NOD/SCID IL-2R-gamma chain null mice. Once tumors had reached a palpable size, mice were treated with Tat-Cre by peritumoral injection to induce CDK12 silencing. After treatment, tumor growth was monitored *in vivo* by caliper measurements for a period of 21 days. Mice were then sacrificed and tumors were explanted and weighed. We verified by IB analysis of tumor lysates that Tat-Cre treatment *in vivo* efficiently silenced CDK12 expression in tumors generated by shCDK12-infected HCC1569-CDK12 cells (Figure 34A). Analysis of the tumor masses showed that CDK12 silenced tumors displayed a significantly slower growth rate compared with control tumors (Figure 34B), which reflected into a greater than 50% reduction in tumor mass (Figure 34C). These results demonstrate that CDK12 silencing reduces the tumorigenic potential of HCC1569-CDK12 cells, confirming that continuous CDK12 overexpression is required to sustain the more aggressive tumorigenic phenotype of these cells compared to the parental HCC1569 cell line.

In conclusion, overall results of the *in vitro* and *in vivo* functional characterization of CDK12 overexpression strongly support our founding hypothesis, demonstrating that CDK12 is a *bona fide* driver oncogene in breast cancer.

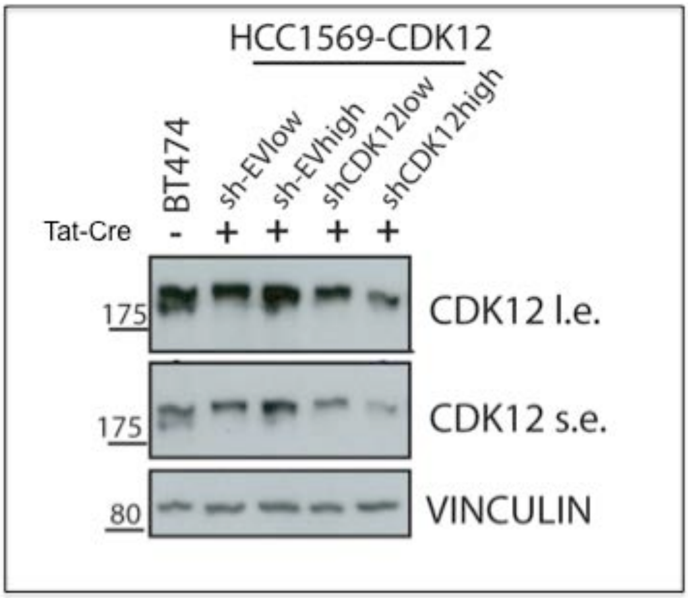


Figure 31. Analysis of the efficiency of conditional CDK12 ablation in HCC1569-CDK12 cells.

HCC1569-CDK12 cells were infected with pSico-shCDK12 or pSico-shLuc, as a negative control, sorted by FACS into EGFP-low (low) and EGFP-high (high) subpopulations and treated with Tat-Cre (100μg/ml). CDK12 protein expression was assessed in the indicated cells by IB analysis of total cell lysates (40 μg protein). CDK12 was detected using the anti-CDK12 AQ19 monoclonal antibody. Both short (CDK12 s.e.) and long exposures (CDK12 l.e.) of anti-CDK12 immunoblots are shown. In the same blot, vinculin was detected as a loading control. Molecular weight markers are shown to the left of the blots. The IB analysis was performed once.

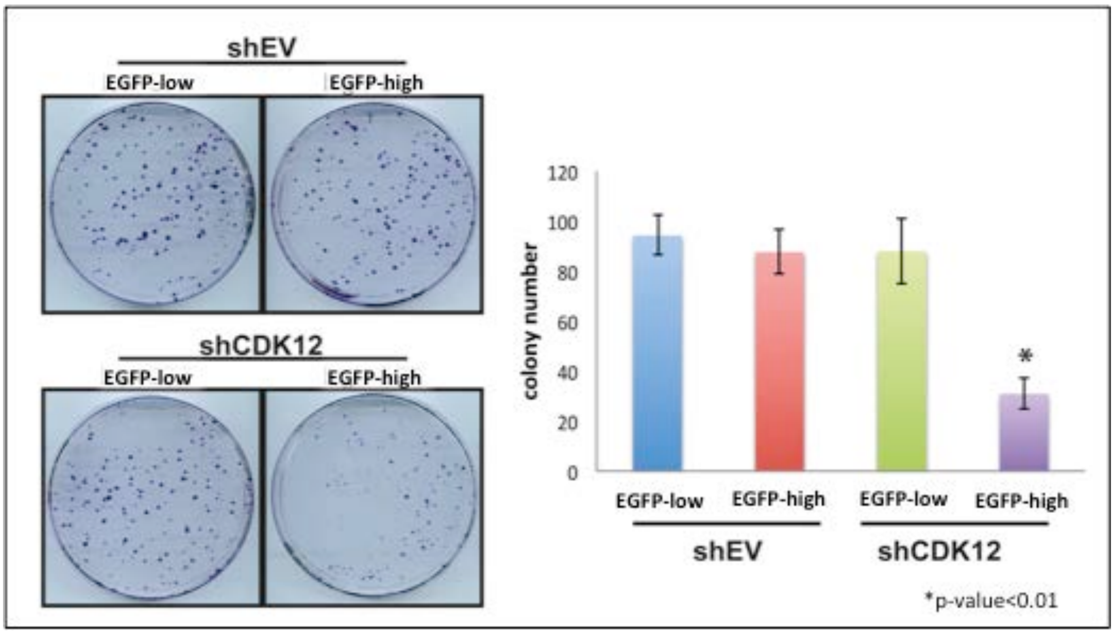


Figure 32. Effects of CDK12 ablation on the clonogenic potential of HCC1569-CDK12 cells in 2D-adhesion culture conditions.

HCC1569-CDK12 cells infected with pSico-shCDK12 or pSico-shLuc, as a negative control, were sorted by FACS into EGFP-low and EGFP-high subpopulations and treated with Tat-Cre (100μg/ml). Cells were then plated at clonogenic density (5000 cells) in 10-cm tissue culture plates and stained after 20 days with crystal violet. Colony number was determined using the ImageJ software. Images of colonies from a representative experiment are shown on the left. Quantification of results is shown on the right and is expressed as the mean ± s.dev. from two independent experiments performed in triplicate (* p-value < 0.01). P-values were determined using the Student's t-Test.

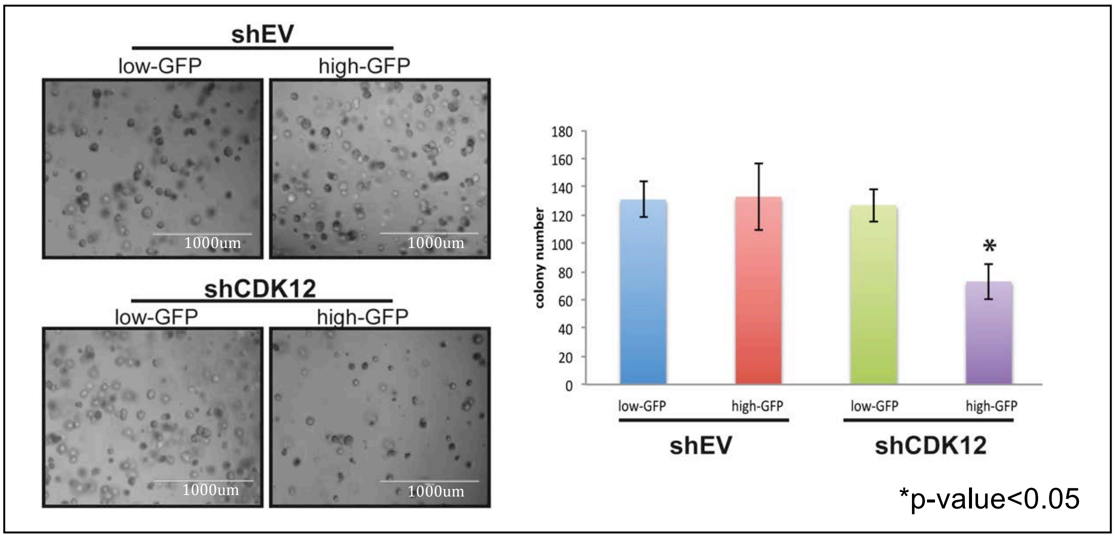


Figure 33. Effects of CDK12 ablation on the ability HCC1569-CDK12 cells to generate organotypic outgrowths in 3D-Matrigel.

HCC1569-CDK12 cells infected with pSico-shCDK12 or pSico-shLuc, as a negative control, were sorted by FACS into EGFP-low and EGFP-high subpopulations and treated with Tat-Cre (100µg/ml). Cells were then embedded as single cell suspensions (4000 cells/ml) in Matrigel, plated at a concentration of 2000 cell per well in 4-well chamber slides and allowed to grow for 20 days to originate organotypic structures. The number of outgrowths/colonies was determined by using the ImageJ software. Left: representative images of outgrowths are shown. Scale bar, 1000 µm. Right: bar graphs showing total number of outgrowths for each experimental condition. Values represent the mean ± s.dev. from two independent experiments performed in triplicate (* p-value < 0.05). P-values were determined using Student's t-Test.

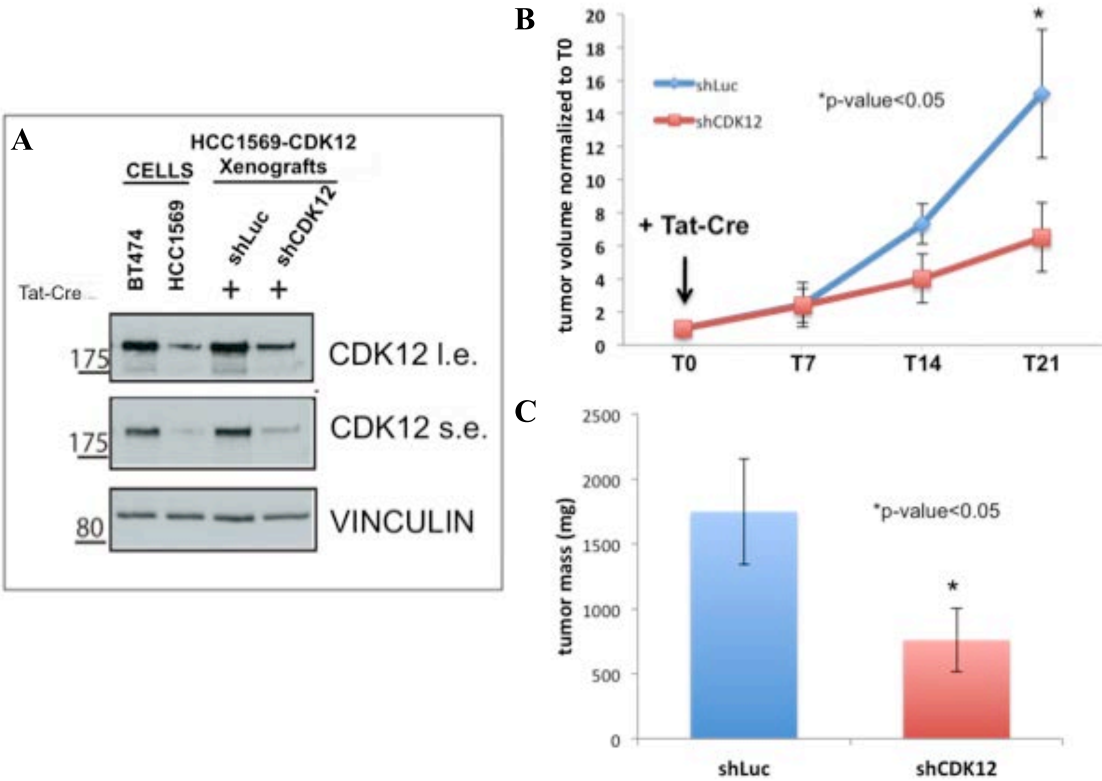


Figure 34. Effects of CDK12 ablation on the ability of HCC1569-CDK12 cells to generate tumors *in vivo*.

Mammary tumors were generated by injecting HCC1569-CDK12 cells ($5x10^5$) infected with pSico-shCDK12 or pSico-shLuc into the contralateral inguinal mammary fat pads of NOD/SCID IL-2R gamma chain null immunocompromised mice. Twenty days after injection tumors were treated with Tat-Cre (0.2 mg) by peritumoral injection and grown for a further 21 day period before animals were sacrificed and tumors explanted and weighed.

(A) Analysis of CDK12 expression in shLuc and shCDK12 tumors by IB analysis of tumor lysates. BT474 and HCC1569 cell lysates were also analyzed in the same blot as reference standards for cells with basally high and low CDK12 expression levels, respectively. CDK12 was detected using the anti-CDK12 AQ19 monoclonal antibody. Images of both short exposure (CDK12 s.e.) and long exposure (CDK12 l.e.) of anti-CDK12 immunoblots are reported. Vinculin was used as a loading control. Molecular weight markers are shown to the left of the blots. The IB was performed once. (B) Growth kinetics of tumors generated by shCDK12- *vs*. shLuc-infected HCC1569-CDK12 cells after exposure to Tat-Cre. Tumor volume was monitored *in vivo* by caliper measurements every 7 days over a period of 21 days. (C) Comparative analysis of the mass of the tumors as in B), performed at the end of the monitoring period. P-values were determined using Student's t-Test; * p-value < 0.05.

5.6 Global profiling analysis of the transcriptional and splicing alterations induced by CDK12 overexpression in breast cancer

Overall results from our functional knockdown and overexpression studies in cell-based model-systems clearly converge on the implication of CDK12 as an oncogene in breast cancer. However, an in-depth understanding of how CDK12 is mechanistically involved in breast carcinogenesis demands the acquisition of a comprehensive picture of the physiological role of CDK12 in cellular homeostasis, which is far from being unequivocally elucidated. The recent literature points to a complex role for CDK12 in RNA processing, with a possible regulatory function in splicing and/or in transcription via RNAPII phosphorylation [103,104,117,127]. With the ultimate goal to couple the physiological CDK12 function in transcription/splicing regulation with its role in breast tumorigenesis, we decided to analyze global gene expression changes induced upon overexpression or silencing of CDK12 in two breast cancer cell lines using the Affymetrix exon array technology.

The Affymetrix Human Exon Array (v1.0) contains more than 5 million of distinct oligonucleotides synthesized on a single array that allows whole-exome analysis. For each of the 1.5 million distinct exons represented, an average of four distinct probes are available, which allows a double-level data analysis: i) at the entire transcript level, for a comprehensive profiling of transcriptional events and ii) at the exon level, for the identification of splicing variants and exon skipping events.

Using this technology, we aimed to: i) identify the entire repertoire of genes whose expression is regulated following perturbation of CDK12 expression levels (i.e., the global gene expression analysis); ii) identify the entire repertoire of splicing variants whose expression is altered following perturbation of CDK12 expression levels (i.e., the global splicing analysis). To tackle these objectives, we employed the following approach: i) a comparative analysis of stably CDK12-overexpressing HCC1569 cells *vs*. their control counterpart (i.e. HCC1569-EV cells; see also Results Section 5.6.1; ii) a comparative analysis of stably CDK12-silenced BT474 cells (BT474-CDK12 KD) *vs*. control BT474-shLuc cells (see Results Section 5.6.1). The ultimate goal of this two-tier approach was to shed light on the possible aberrant regulation of transcription and splicing events occurring in breast cancer cells as a consequence of aberrant CDK12 expression.

5.6.1 Global transcriptome analysis

The global gene expression analysis was performed following two independent approaches. We first analyzed global gene expression changes by selecting those genes differentially expressed in CDK12-perturbed cells (i.e. CKD12- overexpressing HCC1569 or BT474–silenced cells) *vs.* their control counterpart (respectively, HCC1569-EV and BT474-shLUC) by t-test analysis (p-value < 0.05, Welch's t-test). We found a list of 641 genes significantly regulated upon CDK12 overexpression in HCC1569 cells, and a list of 1431 genes significantly regulated upon silencing of CDK12 in BT474 cells. For subsequent validation, we focused on a list of 23 genes that were differentially and inversely regulated in the two matched CDK12-perturbed *vs.* -unperturbed conditions (Table 9).

We selected 8 genes for QPCR validation from the 23-gene list. Four out of 8 genes (50%) were validated with a significant p-value (p-value < 0.05) in CDK12-overexpressing HCC1569 cells: *CDK12*, *BRCC3*, *RHOU* and *MYBL2*. Five out of 8 genes (62%) were validated with a significant p-value (p-value < 0.05) in CDK12 KD BT474 cells: *CDK12*, *BRCC3*, *RHOU, ITF27* and *RASL11B*. Interestingly, most of these genes are known to be involved in cancer deregulated cellular processes such as *MYBL2,* involved in cell proliferation [136]; *RHOU,* involved in focal adhesion formation and cell migration [137]; *ITF27*, involved in cell-cycle control [138]; *BRCC3* involved in DNA damage response (DDR) and angiogenesis [139].

However, with the aim of identifying and functionally validating genes of potential relevance to the CDK12-related phenotypes observed in our *in vitro* and *in vivo* biological studies, special attention was placed towards those genes that appeared to be the most differentially upregulated upon CDK12 overepression. In particular, we selected 7 genes that were upregulated upon CDK12 overexpression in HCC1569: *HIST1H3H*, *LIN28B, TP53*, *CDH2*, *MCM9*, *HEY1*, and *ZEB1*. Five out of the 7 genes (~70%), namely *HIST1H3H*, *HEY1*, *ZEB1*, *CDH2* and *MCM9*, were validated as CDK12-upregulated genes by Q-PCR analysis (p-value <0.05). One of the validated genes that merits further attention is the transcription factor *ZEB1* (zinc finger E-box binding homeobox 1) that is a crucial mediator of EMT. *ZEB1* induces EMT by inhibiting the expression of E-cadherin and miRNAs, which are responsible for maintaining the epithelial phenotype [140]. The upregulation of *ZEB1*, therefore, could be one of the key events determining the EMT observed in HCC1569 cells upon CDK12 overexpression. We are planning to address this hypothesis through further functional validation studies.

Complementary to the above approach of selecting and validating individual genes as *bona fide* CDK12 transcriptional targets, we also performed a pathway-oriented analysis of the global transcriptome changes induced upon CDK12 perturbation with the ultimate goal to shed light on cancer pathways mechanistically linked to CDK12 dysfunction. To this aim, we performed gene set enrichment analyses (GSEA) by comparing gene expression data obtained upon CDK12 overexpression or ablation with a large number of gene sets (a total of 3272 gene sets) available from the Molecular Signatures Database (MSIGDB) (http://www.broadinstitute.org/gsea/msigdb/index.jsp). MSIGDB is a collection of annotated genes, which represent curated canonical pathways or were obtained by several chemical or genetic perturbations of normal and cancer cells. By GSEA analysis, a total of 356 gene sets induced by CDK12 overexpression, and, respectively, of 97 gene sets linked to CDK12 ablation were found to be significantly enriched in the ranked gene expression. Of note, 30 gene sets were common ($P = 7.5*10^{-5}$, Fisher's exact test) and enriched in an opposite manner in the CDK12 overexpressing and silencing experiments.

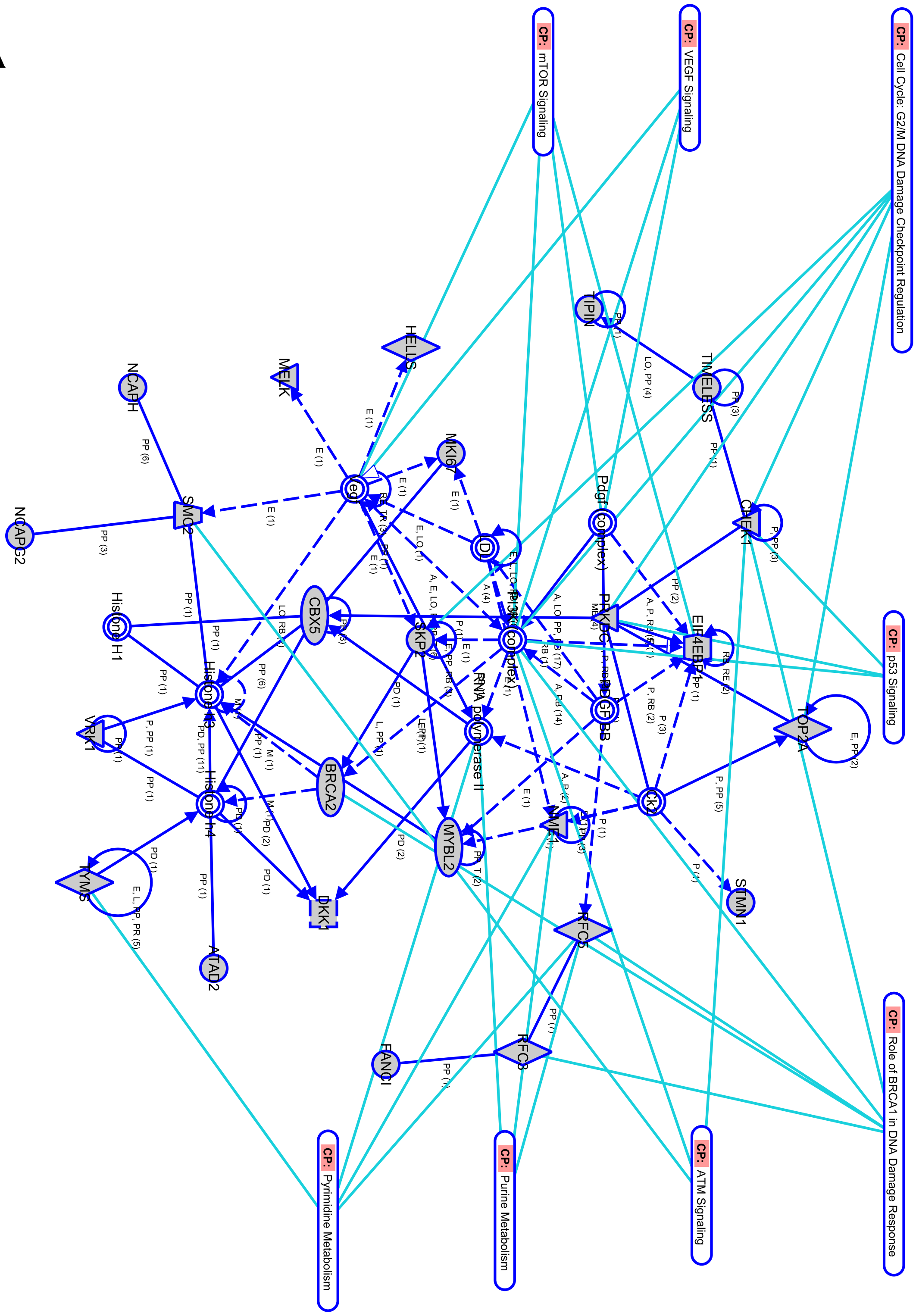
We used the common genes (i.e., the "core" genes) present in more than one gene set and identified by leading-edge analysis (GSEA, see Methods section 7.14) to perform an Ingenuity Pathways Analysis (IPA) to investigate the biological relevance of these genes. The top gene networks obtained through this analysis were highly enriched in genes with known functions in the DNA damage response (DDR) or in cell cycle control (Figure 35). Of note, the identification of a DDR network suggests a function of CDK12 in the genotoxic stress response, a notion that is consistent with a recent study highlighting a role for CDK12 in the transcriptional regulation of genes involved in the DDR, such as BRCA1, ATR, FANCI and FANCD2 [104]. On the other hand, the enrichment of genes involved in cell cycle regulation among those upregulated upon CDK12 overexpression is in a good agreement with our functional studies demonstrating that CDK12 deregulation reflects into the acquisition of a markedly increased proliferative potential by cancer cells (see Results Sections 5.5).

Table 9. Genes differentially regulated upon manipulation of CDK12 expression in BT474 and HCC1569 cells.

	KD vs CTR		OE vs CTR	
Symbol	**Log2 DIFF**	**P-value**	**Log2 DIFF**	**P-value**
CDK12	-1.24	0.00123	1.15	0.02939
CCDC113	-0.81	0.00122	0.18	0.03435
MYBL2	-0.42	0.03640	0.19	0.03033
BRCC3	-0.41	0.03087	0.25	0.00460
RHOU	-0.33	0.00525	0.23	0.01085
CSDA	-0.30	0.00136	0.13	0.03812
IPO4	-0.29	0.00045	0.17	0.01506
BAX	-0.25	0.00157	0.12	0.04601
METT11D1	-0.12	0.00327	0.15	0.03582
SLC25A19	-0.10	0.01165	0.18	0.02730
UBA1	-0.08	0.01026	0.16	0.00582
WDR91	0.18	0.01102	-0.14	0.04245
TSPAN17	0.18	0.00487	-0.08	0.03584
PHC2	0.22	0.00091	-0.11	0.04216
ATF5	0.28	0.00110	-0.16	0.00514
RASL11B	0.30	0.02571	-0.31	0.01511
AIF1L	0.33	0.04018	-0.15	0.03252
IFT27	0.36	0.00178	-0.18	0.01329
ORMDL3	0.39	0.00049	-0.31	0.03378
CNNM3	0.39	0.00715	-0.17	0.04288
IL7	0.50	0.00599	-0.24	0.01037
ZNF596	0.60	0.00346	-0.49	0.00069
ANKRD39	0.73	0.00003	-0.15	0.02869

Gene symbols are indicated in the first column. For each gene, the log difference (Log_2 DIFF) in expression and p-value for both the CDK12 knockdown (KD) and overexpression (OE) conditions are indicated. For the KD experiment, the data were derived by comparing 6 biological replicas of CDK12 KD-BT474 cells, obtained by infection with pSicoR vectors encoding CDK12-specific shRNAs (sh#1, sh#7) *vs.* 3 biological replicas of BT474 cells infected with pSicoR-shLuc as control. For the OE experiment, the results were obtained by comparing 3 biological replicas of HCC1569-CDK12 cells *vs.* 3 biological replicas of control HCC1569-EV cells. Upregulated genes are highlighted in red; downregulated genes in brown.

A



CP: mTOR Signaling
CP: VEGF Signaling
CP: Cell Cycle: G2/M DNA Damage Checkpoint Regulation
CP: p53 Signaling
CP: Role of BRCA1 in DNA Damage Response
CP: ATM Signaling
CP: Purine Metabolism
CP: Pyrimidine Metabolism
TIPIN
TIMELESS
CHEK1
HELLS
MELK
NCAPH
NCAPG2
SMC2
MKI67
CBX5
Histone H1
Histone H3
Histone H4
VRK1
TYMS
ATAD2
DKK1
BRCA2
SKP2
MYBL2
RNA polymerase II
EIF4EBP1
TOP2A
STMN1
RFC5
RFC3
FANCI

B

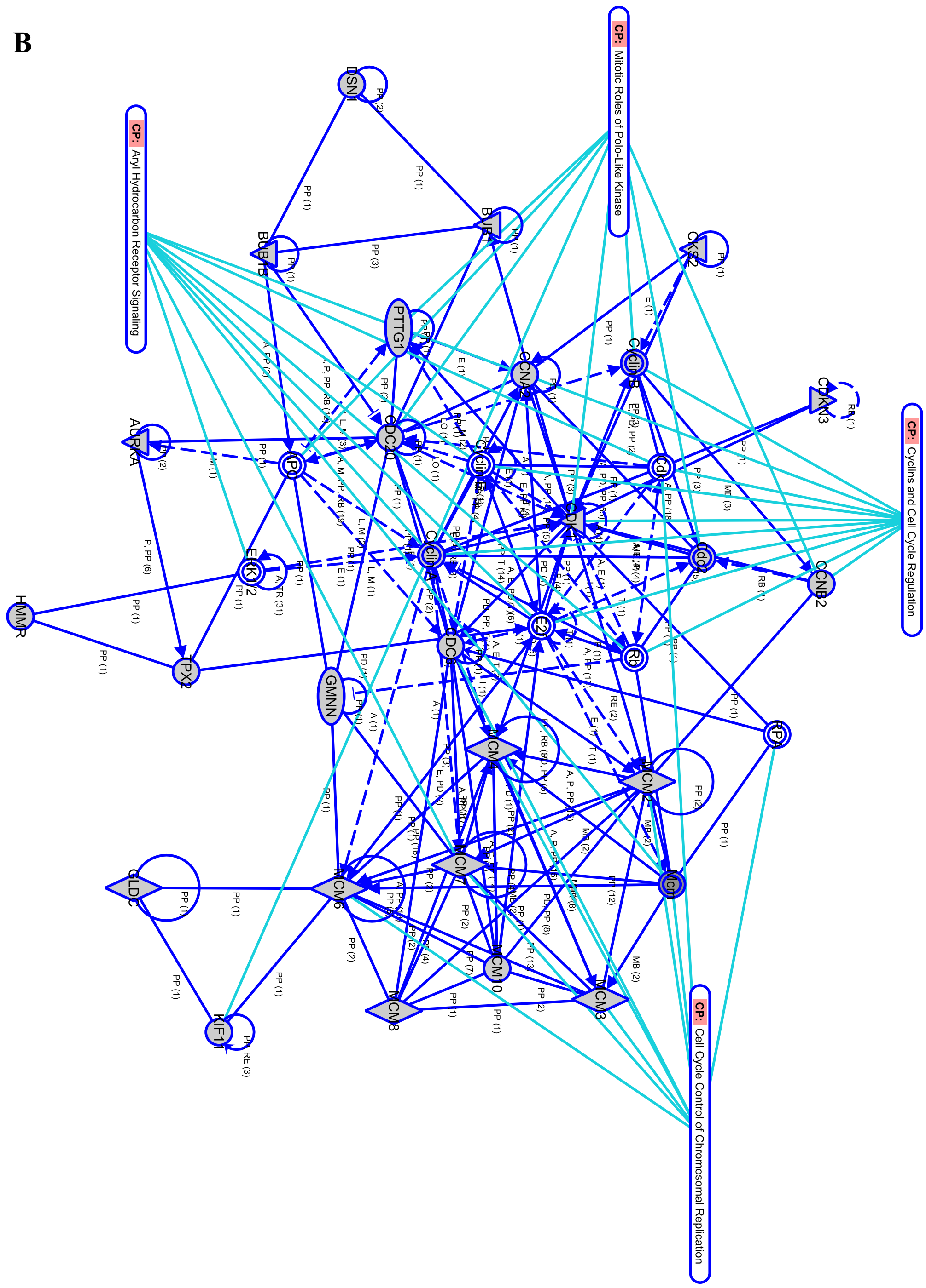


Figure 35. CDK12-related gene networks identified by IPA.

Global gene networks involving “core” genes that were differentially regulated in CDK12-OE or -KD experiments were identified using IPA. The top networks in which CDK12 appears to be implicated are enriched in genes with known functions in the A) DNA damage response (DDR) and in B) cell cycle control. Each line and arrow represents a functional and physical interaction and direction of regulation, respectively, demonstrated in the literature. Cellular processes (CP), in which genes are known to be involved, are also reported.

5.6.2 Global splicing analysis

For the global splicing analysis, we used two different algorithms, the Splicing Index (SI) and the FIRMA method using AltAnalyze software (http://www.altanalyze.org/) to identify alternatively regulated genes (ARGs) in CDK12-OE and -KD conditions. Respectively, in these two experimental conditions, we identified 211 and 160 ARGs that resulted to be significantly regulated by both analytical methods. Of note, we did not observe a significant overlap between the ARGs identified in the two experimental model systems (i.e. BT474 and HCC1569), which could be in part due to differences in the genetic landscape of the two model cell lines.

Of note, pathway reconstruction analysis using the 211 CDK12-upregulated ARGs as an input for the IPA resulted in the identification of networks highly enriched in genes implicated in EMT and cell adhesion regulation (Figure 36).

Remarkably, several of the identified ARGs were also found to be regulated in the global transcriptome analysis. Among these, one of the most interesting genes for its well-established pathogenetic relevance as an oncogene in diverse types of cancer, including breast cancer [141], is CCND1 (Cyclin D1).

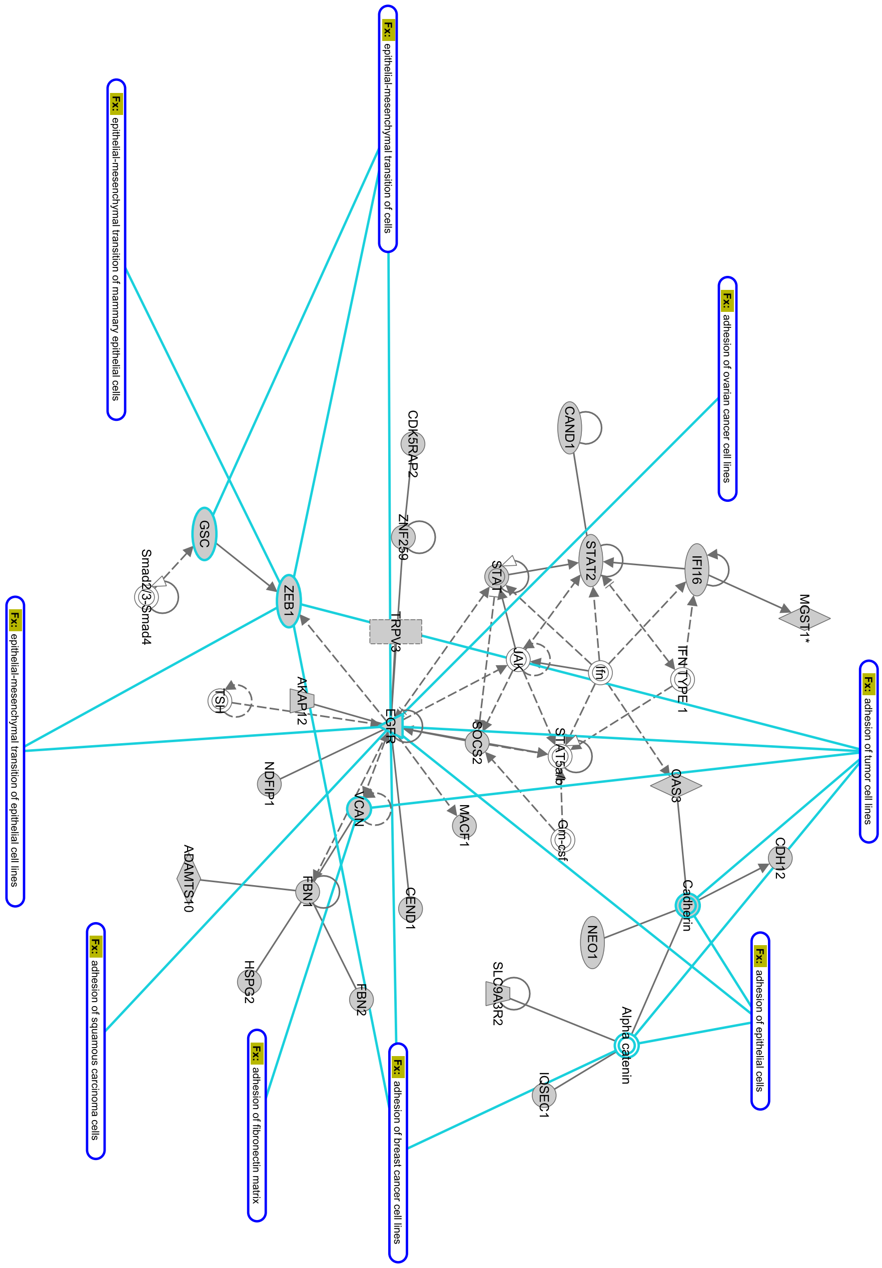


Figure 36. Network of genes alternatively regulated by CDK12 overexpression, identified by IPA.

The list of 211 ARGs, identified in CDK12 OE experiment by the SI and FIRMA algorithms, were used as input to perform IPA. The analysis identified a network highly enriched in genes involved in EMT and cell adhesion processes. Each line and arrow represents a functional and physical interaction, and direction of regulation, respectively, as reported in the literature. Cellular processes, indicated as Fx, in which genes are known to be involved, are also indicated.

5.6.3 Validation of cyclinD1 as a CDK12 transcriptional and splicing target.

The above described global gene expression analysis identified cyclin D1 as a putative transcriptional target upregulated in conditions of CDK12 overexpression. With a potential relevance to the oncogenic function of CDK12, a parallel global splicing analysis predicted that a non-constitutively expressed splicing variant of cyclin D1 was also upregulated upon CDK12 overexpression. We therefore set out to validate *cyclinD1* transcription and splicing as possible molecular events downstream of CDK12 overexpression.

To this aim, we assessed the relative expression of each single exon of the cyclin D1 transcript from the Affymetrix data. A log ratio of the signals obtained from HCC1569-CDK12 vs. HCC1569-EV cells, showed an increase of 1.0 – 1.8 log in the levels of exons 1, 2, 3 and 4, while exon 5 showed a modest increase of only 0.2 – 0.3 log (Figure 37). These data are compatible with a general upregulation of cyclin D1 expression, and in particular with a marked increase in the expression of a short cyclin D1 isoform that lacks exon 5.

The existence of a short non-canonical cyclin D1 isoform, the cyclin D1b splicing variant (CycD1b), has been described in literature [142]. This isoform is generated when splicing at the exon-4/intron-4 boundary fails to take place. This results in complete loss of exon-5 encoded sequences and in the acquisition of a novel 33-amino acid stretch at the C-terminal end due to intron-4 translation and to the presence of an intronic stop codon (Figure 38). Of note, CycD1b lacks both the C-terminal PEST sequence and the Thr286 residue: importantly, the lack of these domains predicts that the protein would reside in the nucleus and be intrinsically more stable than the full-length cyclin D1 (CycD1a) [143]. Consistent with this prediction, functional analyses of CycD1b revealed that this isoform harbors a stronger oncogenic potential than the canonical full-length variant CycD1a [143,144]. However, the molecular mechanisms by which CycD1b drives tumorigenesis have not been fully elucidated.

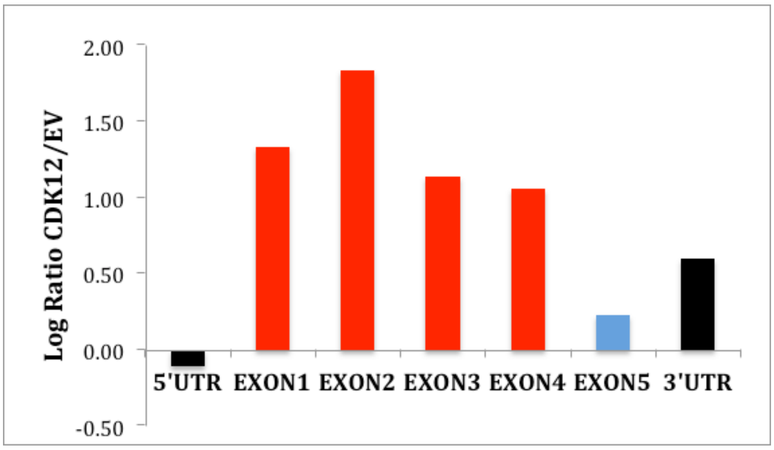


Figure 37. CCND1 transcript analysis

Relative expression of UTR regions and exons of the CCND1 transcript derived from the exon array analysis expressed as a log ratio of signals from HCC1569-CDK12 *vs.* HCC1569-EV cells. Values reported represent the mean of the 3 biological replicas used in the exon array experiment.

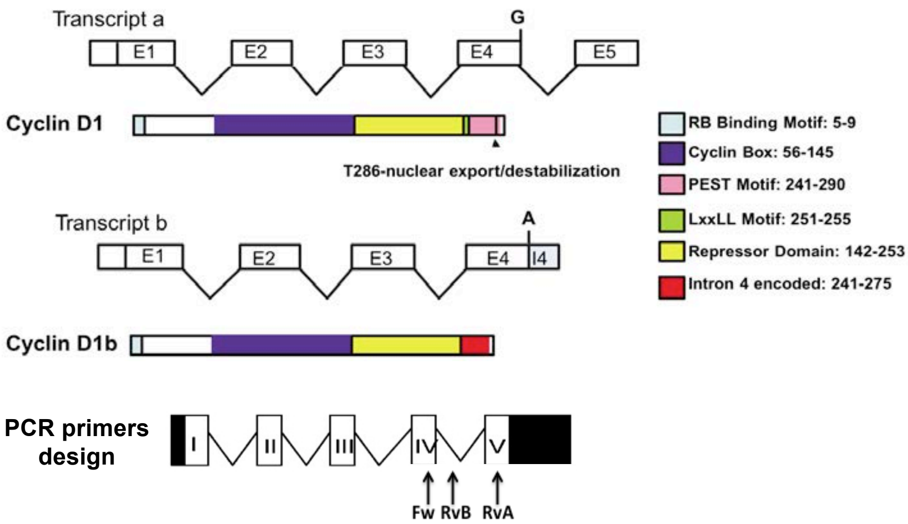


Figure 38. Schematic representation of Cyclin D1 isoforms

Domain structure of CycD1a and CycD1b. The predominant domains and their localization within the primary sequence of CycD1a and CycD1b are shown. Approximate position of the primers used to specifically amplify the CycD1a (Fw-RvA) or CycD1b transcripts (Fw-RvB) are also indicated by black arrows. Figure adapted from [142].

5.6.3.1 Q-PCR and WB validation

In order to validate the exon array prediction of cyclin D1 alterations, we performed a Q-PCR analysis using specific primers designed to specifically detect the two isoforms (Figure 38). The SYBR Green Q-PCR reaction, using the in-house designed primers specific for full-length cyclin D1 gave rise to a single band of the expected size in both HCC1569-CDK12 and –EV cells (Figure 39). However, in the quantitative analysis, in which cyclin D1 levels were normalized to the housekeeping genes GAPDH and GUSB, a 30-fold increase in the expression of full-length cyclin D1 mRNA in CDK12 overexpressing cells compared with control cells was observed (Figure 39). In contrast, the cyclin D1b isoform was not detected in HCC1569-EV cells by SYBR Green Q-PCR amplification, while a single band of the expected size appeared in HCC1569-CDK12 overexpressing cells. Of note, CycD1a was expressed at higher levels (~250-fold) compared with the CycD1b isoform in CDK12 overexpressing cells (Figure 39). Since the CycD1b isoform is not expressed at detectable levels in control cells, it was not possible to quantify the relative increase in expression of this short isoform over the long CycD1a isoform in HCC1569-CDK12 *vs.* HCC1569-EV cells.

In conclusion, by Q-PCR analysis, we were able to confirm that CDK12 overexpression leads to altered transcriptional regulation of cyclin D1. In particular, we established that CDK12 overexpression in HCC1569 cells results in both increased expression of the full-length isoform and in enhanced expression of an alternative short isoform that is not expressed in parental cells.

We also tried to validate gene expression data at the protein level by performing IB analysis with antibodies that specifically recognize either the full-length cyclin D1 alone (AB3) or both the full-length and the short cyclin D1 isoforms (DCS6). The CycD1a and CycD1b proteins have an expected difference in molecular weight of ~3KDa, being respectively of 33 and 30 kDa in size. Both antibodies recognized only a single band with an apparent molecular weight of ~33 kDa, consistent with the full-length cyclin D1 isoform, in HCC1569-CDK12 overexpressing cells, that was not detected in HCC1569-EV control cells (Figure 40). Thus, the increase in CycD1a mRNA levels upon CDK12 overexpression is accompanied by an increase in CycD1a protein levels. However, at protein level, we were unable to detect the presence of the short cyclin D1 isoform. The most likely explanation for this rests on the observation that, according to the literature (Wang et al. 2008) and consistent with our Q-PCR experiments, even when expressed in cancer cell lines, the CycD1b isoform constitutes only a minor portion of the total cyclin D1 (Figure 39). Thus, it might be intrinsically difficult to detect the expression of CycD1b in the absence of an antibody specifically directed against this short isoform, with no cross-reaction with the full-length protein. For this reason, we are now generating an antibody in-house against an epitope in the translated intron-4 region, which predictably should be instrumental to study the D1b isoform.

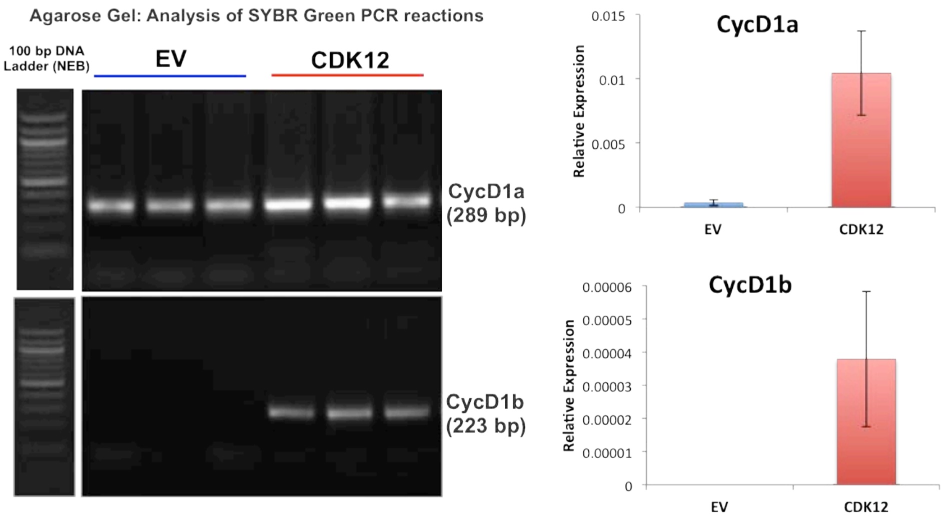


Figure 39. Q-PCR analysis to evaluate the expression of cyclin D1a and cyclin D1b mRNA transcript levels in HCC1569 cells.

RT-Q-PCR evaluation of cyclin D1 and cyclin D1b transcripts in HCC1569-EV and -CDK12 cells. Left: PCR products were resolved on agarose gels and size of the amplified sequences was determined using the DNA base-pair (BP) marker. Right: quantification of the levels of expression of full-length cyclin D1 and the cyclin D1b isoform relative to two housekeeping genes (GAPDH and GUSB) is reported. Values represent the mean of 3 biological replicas ± s.dev.

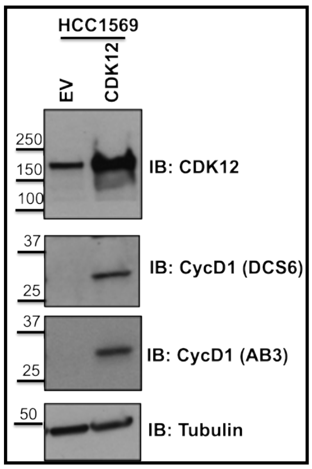


Figure 40. Effects of CDK12 overexpression on cyclin D1 protein levels in HCC1569 cells.

Analysis of cyclin D1 protein expression by IB of total cell lysates from HCC1569 cells after infection with pLVX-CDK12 or pLVX-EV as a negative control. Cyclin D1 was detected with antibodies that recognize either full-length cyclin D1 alone (AB3) or both the full-length and cyclin D1b isoforms (DCS6). CDK12 protein was also detected to control for its overexpression. Tubulin was used as a protein loading control. Molecular weight reference markers are reported to the left of the blots. The blot is representative of three biological repeats.

6 Discussion

6.1 CDK12 is a novel prognostic biomarker in breast cancer

In addition to scientific robustness, technical and economical considerations need to be taken into account in the incorporation of new biomarkers into routine clinical practice. Prognostic and predictive gene expression profiles that have been developed during the last decade, such as the Mammaprint (Agendia) and Oncotype DX (Genomic Health) (see Introduction Section 20), have found poor clinical applications because of difficulties in the standardization of procedures, high costs and little improvement over the traditional clinicopathological markers [97]. For this reason it is important that researchers provide the clinic with new biomarkers that can be assessed in large patient cohorts using rapid, standardized and low cost tests, such as IHC, which is routinely used in clinical practice to assess the expression of molecular markers (e.g., ErbB2, ER and PR).

In a preliminary unpublished study performed in our lab (see section 48), we analyzed CDK12 expression by ISH on TMA. CDK12 overexpression at the transcript level appeared to be a strong prognostic factor associated with an increased risk of relapse and metastasis. However, to clearly establish CDK12 as a prognostic biomarker in breast cancer it was necessary to determine the correlation between CDK12 expression at the protein level and clinical outcome. To perform this analysis, we generated in-house a monoclonal antibody that was able to specifically recognize the human CDK12 protein in a number of techniques including IHC on FFPE samples (see results section 5.1). We used this reagent to analyze CDK12 expression by IHC on TMA in large cohorts of breast cancer patients, including a case-control group of ~350 cases and a validation group of ~1000 cases. In both study-groups, CDK12 overexpression was found to be significantly associated with traditional prognostic markers of aggressive disease, namely, high KI67 proliferation index, high ErbB2 status, negative ER and PR status, and high tumor grade (G3), confirming and extending our previous results [2]. Furthermore, by regression analysis we observed that CDK12 expression significantly predicted the risk of disease recurrence and death up to 15 years after surgery. We showed that the higher the level of CDK12 expression, the poorer the prognosis (see results section 5.2). Of note, we also found that CDK12 overexpression retains its prognostic value in ErbB2-negative patients, indicating a role of CDK12 in breast cancer that is independent of ErbB2. This finding is particular relevant as the *CDK12* gene is located in close proximity to the oncogene *ERBB2*, which is frequently amplified in breast cancer.

In conclusion, we have clearly established CDK12 as a novel biomarker in breast cancer that is able to stratify patients and identify those patients with a higher risk of poor clinical outcome. Moreover, the generation of a monoclonal antibody that can be used to detect CDK12 expression by IHC on FFPE samples is an important step forward in the development of a clinical test that can be used to routinely assess CDK12 expression in breast cancer patients.

6.2 *CDK12* is amplified in breast cancer

Amplification is a well-known mechanism of oncogene activation that is selected for during cancer progression. *CDK12* is located on chromosome 17 (17q12) in a region known as the "*ERBB2*-amplicon", and previous studies have reported that *CDK12* is frequently co-amplified with *ERBB2* [60,63].

We therefore investigated whether amplification was the genetic lesion underpinning CDK12 overexpression in breast cancer. We assessed both *CDK12* and *ERBB2* amplification by FISH analysis in a cohort of ~350 breast cancer patients, of which 132 cases gave evaluable results for both the *CDK12* and *ERBB2* genes. *CDK12* was amplified in 32% of cases. It is worth noting that although the genomic localization of *CDK12* is relatively close to *ERBB2* (300 kb apart), in most of the *CDK12* positive cases (74%), *CDK12* amplification occurred as a single event, in the absence of *ERBB2* amplification (see section 3)

This observation, together with the notion that the predictive prognostic power of CDK12 overexpression is maintained in ErbB2-negative patients, strongly argues that *CDK12* amplification could represent a driver lesion in breast carcinogenesis that is independent from *ERBB2* amplification. On the other hand, *CDK12*, when co-amplified with *ERBB2,* could confer a synergistic or additive growth advantage to breast cancer cells, and might be a determinant of resistance to anti-ErbB2 based cancer therapies, a possibility with enormous clinical implication that warrants future investigations.

6.3 CDK12 is a novel oncogene in breast cancer

Over the past few years, the introduction of sophisticated post-genomic platforms has determined a constant increase in the identification of molecular alterations in cancer. However, it is often disputed whether a given molecular alteration constitutes a causal lesion, selected for during tumorigenesis (i.e., a driver mutation), or whether it is just part of the broad repertoire of molecular alterations occurring in the natural history of the tumor due to the intrinsic genomic instability of tumor cells (i.e., a passenger mutation). In the former case, the expectation is that tumor cells harboring a given alteration are dependent on it for the maintenance of the malignant phenotype.

Several lines of evidence presented in this thesis indicate that *CDK12* amplification is a driver lesion in breast cancer. We demonstrated that silencing CDK12 expression in a breast cancer cell line harboring *CDK12* amplification and protein overexpression, BT474, results in a significant decrease in the colony forming ability *in vitro* and tumorigenic potential *in* vivo of these cells (see results section 5.4). In contrast, downregulation of CDK12 expression in the quasi-normal mammary epithelial MCF10A cells, as well as in breast tumor HCC1569 cells with low constitutive CDK12 levels, does not affect their growth ability.

Additionally, enforced CDK12 overexpression, albeit unable to malignantly convert quasi-normal MCF10A cells increases their proliferative potential *in vitro*, an effect that was also reiterated in HCC1569 tumor cells which display low basal CDK12 levels. However, in HCC1569 cells, most likely as a consequence of a permissive tumor background, stable CDK12 overexpression results in a dramatic increase of their *in vitro* and *in vivo* tumorigenic potential. Of note, the enhanced tumorigenic phenotype conferred by stable CDK12 overexpression in HCC1569 cells can be reverted through the restoration of basal CDK12 levels in silencing experiments, indicating that these cells are dependent on CDK12 overexpression to maintain the worsened tumor phenotype.

Further supporting a causative role for CDK12 in breast carcinogenesis, we showed that enforced CDK12 overexpression confers mesenchymal traits to cancer cells, such as the acquisition of a spindle-like morphology and the molecular switch from E-cadherin to N-cadherin expression (see results section 5.5.5), which strongly suggest that EMT might be one of the mechanisms through which CDK12 confers a more aggressive behavior to tumor cells and contributes to a more adverse natural evolution of the breast cancer disease. It is worth noting that this scenario is consistent with results from our pathway reconstruction analysis on the list of genes identified as transcriptional CDK12 targets, which showed, among other CDK12-related cancer relevant pathways, a significant enrichment in genes implicated in EMT.

In conclusion, overall results from our functional studies implicate CDK12 as a *bona fide* oncogene in breast cancer. These results are of particular relevance when viewed in the broader perspective of the personalized treatment of the heterogeneous breast cancer disease for a two-fold reason; i) they indicate that targeting CDK12 may represent a rationale therapeutic strategy in CDK12-overexpressing tumors, if correctly stratified based on their intrinsic CDK12 expression status (CDK12 ablation does not affect the HCC1569 tumor cell line with constitutively low CDK12 levels); ii) they demonstrate that targeting CDK12 has no detrimental effects on the normal mammary tissue physiology: two criteria that are at the basis of the concept of rationale targeted therapy in cancer.

A major question that remains to be addressed is whether the STK activity of CDK12 underpins its oncogenic activity. To address this question, we plan to perform *in vitro* and *in vivo* experiments comparing overexpression of wild-type CDK12 with either dominant-active or kinase-dead mutants of the protein in HCC1569 cells. If these experiments succeed in revealing the direct contribution of the CDK12 STK activity in worsening the tumor phenotype of HCC16569 cells, the next important step towards the clinical translation of these results will be the development of CDK12 specific kinase inhibitors to test the suitability of CDK12 as a molecular target for anti-cancer therapy.

6.4 Molecular consequences of CDK12 overexpression

An important question that we started to address in this thesis is how CDK12 deregulation is mechanistically implicated in cancer and which cellular processes are deregulated by its overexpression. So far, few studies have addressed the role of CDK12 in cellular homeostasis. Recent reports point to a complex role of CDK12 in RNA processing. Indeed, it is thought that CDK12 has a regulatory function in both the splicing and transcription processes, which is mediated by its ability to phosphorylate the CTD of RNAPII [103,104,117,127]. We therefore investigated whether alteration of CDK12 expression in breast cancer cell lines induces changes in the transcriptomic landscapes of these cells. We therefore performed a genome-wide analysis using Affymetrix Exon Array technology to identify alterations in both transcription and splicing linked to CDK12 overexpression in HCC1569 cells or CDK12 ablation in BT474 cells.

By global gene expression analysis, we identified a total of 1431 genes whose expression was altered by CDK12 ablation and 641 genes whose expression was altered upon CDK12 overexpression. Of these genes, 23 were found to be in common between the two experimental systems, albeit with an expected opposite modes of regulation. So far, we have validated 5 of these genes by Q-PCR analysis, namely, *BRCC3*, *RHOU, ITF27, RASL11B* and *MYBL2.* Interestingly, most of the common genes are involved in cellular processes that are typically deregulated in cancer. For instance *MYBL2* is involved in cell proliferation [136]; *RHOU* controls focal adhesion formation and cell migration [137]; *ITF27* is involved in cell cycle control [138]; *BRCC3* is implicated in DDR and angiogenesis [139]. These genes therefore represent interesting candidates for high-resolution functional studies investigating their involvement in breast tumorigenesis.

To gain insights into the biological relevance of the global transcriptional changes induced by CDK12 overexpression we performed a pathway-oriented analysis. This analysis resulted in the identification of both a cell cycle control and a DDR gene-network, suggesting an important role of CDK12 in the regulation of these processes. A number of considerations support the patho-physiological relevance of the pathways identified in our analysis. Increased proliferative ability is one of the main hallmarks of cancer [56]. Since CDKs have a specific role in the control of cell cycle, it is not surprising that we found a cell cycle gene network regulated by CDK12 modulation. This observation also reflects the results obtained in our functional studies, which revealed that CDK12 overexpression both in normal and cancer cells causes an increase in cell proliferation. It will be interesting to investigate the precise role of CDK12 in the regulation of the cell cycle, both in normal and cancer cell model systems, and to identify the main downstream players that mediate CDK12-induced proliferative effects.

Furthermore, the involvement of CDK12 in the DDR process is in a good agreement with a recent study by Blazek *et al.* [104]. Using HeLa cells as a model system to KD CDK12 expression, these authors reported that CDK12 directs DDR and maintenance of genomic stability via regulation of expression of key DDR genes, such as BRCA1, ATR, FANC1 and FANCD2. The role of DDR genes in the process of carcinogenesis is conceptually dualistic because, on the one hand, they protect normal cells from transformation by avoiding genomic instability, which is one of the first hallmarks detected in cancer [56]. On the other hand, DDR genes protect cancer cells from genotoxic stress that could lead to apoptosis [145]. A recent study reported that the overexpression of DDR genes in a number of human tumors including melanoma, breast and bladder cancers is associated with metastasis and poor patient survival [146]. These results could explain the high resistance of metastases and advanced tumors towards chemo- and radio-therapies. Indeed, it has been proposed by Sarasin and Kauffmann that genomic instability is absolutely necessary to transform normal cells, but cancer cells need some genetic stabilization, which can be obtained by overexpressing specific DNA repair genes, to be able to progress and eventually metastasize [147]. It would be interesting to investigate further the role of CDK12 in the maintenance of genomic stability and progression of breast cancer, for example by evaluating whether CDK12 overexpression could be a determinant of resistance to chemotherapeutic treatments.

The second goal of our genome-wide expression analysis was to identify alternative splicing variants linked to CDK12 deregulation. By global splicing analysis, we identified 211 ARGs by CDK12 overexpression in HCC1569 cells and 160 ARGs by CDK12 KD in BT474 cells. In contrast to the gene expression analysis, no significant overlap was observed between the ARGs in the two experimental model systems. A possible explanation for this lack of overlap could be that the splicing effects linked to CDK12 are dependent on the cellular background. Nevertheless, we observed that several of the genes identified in the splicing analysis were also identified in the gene expression analysis, such as Cyclin D1 (see below for an extensive characterization of Cyclin D1 as a CDK12 transcriptional target) and ZEB1. Of note, ZEB1 is a well-characterized EMT transcription factor [140] .It is therefore tempting to speculate that upregulation of ZEB1 transcription/splicing might mediate the occurrence of EMT observed in HCC1569 cells upon stable CDK12 overexpression. Future high-resolution studies are required for the thorough characterization of some of these interesting candidates, among which Cyclin D1 and ZEB1, as downstream effectors involved in CDK12 oncogenic activity.

As yet, we do not know whether the effects we observed on transcription and alternative splicing of endogenous genes are directly or indirectly mediated by CDK12. However, based on the current literature and the data obtained in our exon array analysis, a possible scenario emerges in which CDK12, through phosphorylation of the CTD of RNAPII, could have a role in coupling transcription and splicing [104,117,127]. The notion that transcription and splicing are intimately linked processes came from the observation that most of the primary mRNA-processing events, including intron removal by splicing, occur co-transcriptionally [148-150]. The CTD of

RNAPII has a central role in coupling the two processes [148], both by recruiting splicing factors to the active site of transcription and by modulating the rate of transcription [151]. Thus, CDK12, by phosphorylating the CTD of RNAPII could be a key player in the regulation of transcription and splicing. This hypothesis is supported by the presence of a RS-rich domain, a typical feature of SR proteins, in CDK12[101]. SR proteins have been primarily characterized as splicing regulators, but have also been proposed to be transcription-splicing coupling factors [89,106]. Interestingly, the SR protein, SF2/ASF, has been shown to be a proto-oncogene that is amplified in breast cancer[152,153].

Recent studies suggest that both transcription and splicing are intimately linked to the chromatin organization [154] and that splicing is regulated by histone modifications [155] and *vice versa* [26,156,157]. Of note, CTD-Ser2 phosphorylation, an event in part mediated by CDK12, temporally correlates with levels of H3K36 trimethylation[158], considered to be a marker of active transcription/splicing[156,157]. It will be interesting to address the role of CDK12 in this context by looking at possible effects of CDK12 overexpression on H3K36 methylation status.

6.5 Cyclin D1 is a putative downstream effector of CDK12 overexpression

CDK12 overexpression could affect transcription and splicing through several possible mechanisms. According to the model proposed by Rodrigues *et al.*, CDK12 overexpression, via enhanced CTD-Ser2 phosphorylation, could increase/alter the recruitment of the splicing machinery to active sites of transcription, thereby altering the transcriptome. An alternative hypothesis is that CDK12 overexpression, through an enhanced CTD-Ser2 phosphorylation, could increase the RNAPII elongation rate, resulting in a deregulated or inaccurate splicing processivity. Our results on cyclin D1 support this latter scenario.

Cyclin D1 was one of the most highly upregulated genes in the expression analysis, as well as one of the most alternatively regulated genes in the splicing analysis, in HCC1569 cells following CDK12 overexpression. We verified by Q-PCR analysis with specific primers in HCC1569-CDK12 overexpressing cells, the presence of a short non-canonical isoform of cyclin D1, CycD1b, which was not present in parental cells. This isoform results from an exon-skipping event at the exon4/intron4 boundary and in a premature termination of the transcript within the intron 4[142]. This event is accompanied by an upregulation of the canonical CycD1a isoform, indicating a general increase of the RNAPII processivity rate at level of the cyclin D1 gene. These events are consistent with an increased phosphorylation of the CTD of RNAPII, most likely due to the upstream overexpression of CDK12. However, additional studies are needed to verify this hypothesis.

From a functional point of view, the upregulation of CycD1a and CycD1b, downstream of CDK12 overexpression, could explain some of the oncogenic effects we observed in our functional characterization study. Firstly, based on the known role of CycD1a in controlling cell cycle[141], its upregulation could cause the increased proliferative potential observed in CDK12 overexpressing cells. On the other hand, CycD1b expression has been linked to increased tumorigenic potential[143,144], although the molecular mechanisms through which this isoform drives tumorigenesis have not been fully elucidated. However, based on the absence of the C-terminal PEST sequence and of the Thr286 residue in the short CycD1b isoform, the predicted mechanism underpinning its oncogenic activity is a prolonged stabilization and persistence of this protein at the nuclear level, which is also consistent with the observation that this short isoform harbors a stronger oncogenic potential than the canonical full-length variant CycD1a [143,144].

To clarify the role of cyclin D1 in mediating the effects of CDK12 overexpression, we plan to assess the effects of CycD1a- and CycD1b-KD on tumor phenotypes of CDK12-overexpressing breast cancer cells.

Of note, cyclin D1 is overexpressed in 40 – 50% of human breast cancers and its gene, which is located on chromosome 11q13, is amplified in 10 – 20% of breast cancer cases [77]. Thus, amplification accounts for only a minor percentage of cases in which cyclin D1 is overexpressed. No precise and exhaustive events have been described to explain cyclin D1 upregulation in the remaining percentage of cases. To verify whether CDK12 overexpression could be responsible for cyclin D1 overexpression in the remaining 30 - 40% of cases, we are planning to perform IHC analysis to detect both CycD1a and CycD1b expression in large cohorts of human breast cancers and to correlate their expression with CDK12 expression.

6.6 Concluding Remarks

Breast cancer is a heterogeneous disease and currently available diagnostic tools are insufficient for accurate prediction of patient outcome and for tailoring of personalized treatments. Additionally, although adjuvant systemic therapy substantially improves disease-free and overall survival in both premenopausal and postmenopausal women with breast cancer, not all breast cancer patients respond to the treatment and certain adjuvant therapy regimens are not appropriate for some patients [159,160]. Therefore, the identification of novel biomarkers and therapeutic targets for a more refined prognostic, clinical and therapeutic stratification of breast cancer patients remains a major unmet clinical need. Protein kinases, frequently implicated in malignant transformation, hold great promise for the development of targeted therapies, as witnessed by the increasing number of kinase inhibitors that have already found application in cancer therapy or are currently under clinical trial

[126]. In a recent high-throughput analysis to assess expression of 125 STKs in different types of human cancers using ISH on TMA, we identified CDK12 as a STK overexpressed in breast cancer [2].

In the present study, we have established CDK12 overexpression as a new breast cancer biomarker associated with aggressive disease and a higher risk of poor prognosis. Moreover, we have identified CDK12 as a novel proto-oncogene that is activated by amplification in human breast tumors and which induces its oncogenic effects probably by altering the transcription and splicing of specific cancer genes.

7 MATERIALS AND METHODS

7.1 Generation of an anti-CDK12 monoclonal antibody

Using the bioinformatics tool GlobPlot, available online (http://globplot.embl.de/), we selected a specific globular region of the CDK12 protein, outside of the kinase domain. An antigenic peptide corresponding to a unique sequence within CDK12 (amino acid residues 400 – 524) was selected and used to immunize mice in the form of a GST-fusion protein, in collaboration with the Antibody Biochemistry Facility at the IFOM-IEO Campus, Milan. The AQ19 monoclonal antibody was selected and affinity-purified using GE Healthcare columns . The final concentration of the antibody was 250 μg/ml.

7.2 TMA

7.2.1 Patient selection and study design.

Case-control study group (or 'training' set): we established a case-control study on a large series of breast cancer patients (n = 349) who had undergone surgery at the European Institute of Oncology (IEO) in Milan for the removal of primary breast cancer between 1994 and 1997. Disease recurrence (any relapse and distant relapse) was within 18 years (median 9.2 years and 10.6 years respectively). For some patients not all information was available. (See Table 5 for further details on patient characteristics)

Validation study group (or 'test' set): this group included 970 breast cancer patients who underwent surgery at IEO for the removal of primary breast cancer between 1997 and 2000. For some patients not all information was available. (See Table 6 for further details on patient characteristics).

Written informed consent for research use of biological samples was obtained from all patients, and the research project was approved by IEO's Institutional Ethical Committee.

7.2.2 Analysis of CDK12 expression in breast cancers by immunohistochemistry on tissue microarray

Tissue sample preparation, IHC and TMA-IHC analyses were performed in collaboration with the Molecular Pathology Unit at the IFOM-IEO Campus, Milan.

TMAs were prepared [161] and analyzed as previously described [162]. Briefly, two representative normal (when available) and tumor areas (diameter 0.6 mm) from each biopsy sample, previously identified on haematoxylin-eosin stained sections, were removed from the paraffin donor blocks and deposited on the recipient block using a custom-built precision instrument (Tissue Arrayer - Beecher Instruments, Sun Prairie, WI 53590, USA). Two-μm sections of the resulting recipient block were cut, mounted on glass slides, and used for IHC.

TMAs were analyzed for CDK12 protein expression by IHC (IHC-CDK12). TMA sections were routinely processed, placed for 30 minutes in 0.25 mM EDTA at 95°C for antigen retrieval and incubated for 3 hours with the AQ19 anti-CDK12 monoclonal antibody (1:1000, produced in-house); bound antibody was revealed using the EnVision Plus/HRP detection system (DAKO) and diaminobenzidine as a chromogenic substrate. TMA sections were finally counterstained with hematoxylin and mounted. Positive and negative controls were included in each experiment and only clear staining of the tumor cell nuclei was considered positive for CDK12 expression. A semiquantitative approach was used to generate a score for each tissue core, ranging from 0 to 3 according to the signal intensity. Scores 1 (weak), 2 (moderate) and 3 (strong) were assigned when at least 30% of tumor cells in the sample were positive. IHC values of each duplicate core were then averaged. Tumors showing IHC scores > 1.0 were assigned to the CDK12-HIGH group, whereas those with IHC scores ≤ 1.0 were considered as the CDK12-LOW group.

ER and PR proteins, measured by IHC on whole tissue sections, were retrieved from histopathology reports. The rate of proliferation was measured by determining the percentage of nuclei in which labeled antigen Ki-67, a marker of cell division, was expressed (MIB1 antibody DAKO, Cytomation). ErbB2 expression was measured by IHC on TMAs, processed as previously described, using an anti-ErbB2 polyclonal antibody (1:160, DAKO, Cytomation). ErbB2 overexpression was evaluated according to the scoring system recommended by the DAKO HercepTest: score 0, no staining or membrane staining in <10% of the tumor cells; score 1, barely perceptible membrane staining in >10% of the tumor cells; score 2, weak-to-moderate staining of the entire membrane in >10% of the tumor cells; score 3, strong staining of the entire membrane in >10% of the tumor cells. Scores of 2 and 3 were considered to represent overexpression.

7.2.3 Statistical Analysis

Association between the clinical/pathological features of the tumors and CDK12 expression was evaluated by the Fisher's exact test. In the breast cancer validation cohort, logistic regression was used to assess the association between CDK12 expression and the presence or absence of events (relapse or death) from the date of surgery to the date of the event or the date of last follow-up. Follow-up was updated to 2012. Kaplan-Meier plots were carried out by means of the proportional hazards Cox-model. All *P*-values were two-sided. A *P*-value equal to or less than 0.05 was considered significant. All statistical analyses were carried out using SAS statistical software (SAS Institute, Inc., Cary, NC).

7.3 Cell Lines

All human breast cell lines were from the American Type Culture Collection (ATCC). The MDA-MB-361, BT474, MCF7 and SKBR3 cell lines were cultured in DMEM medium (from Lonza), supplemented with 10% fetal bovine serum (FBS, HyClone) and 4 mM L-glutamine (Euroclone). The HCC1569, HCC1428, ZR7530 and AU565 cell lines were cultured in RPMI-1640 medium (from Lonza), supplemented with 10% FBS and 4 mM L-glutamine. The BT483 cell line was cultured in RPMI-1640 medium supplemented with 20% FBS and 0.01 mg/ml insulin (Sigma). MDA-MB-175 cell line was cultured in Leibovitz's L-15 medium (Invitrogen, Life Science Technologies) supplemented with 10% FBS. MDA-MB-436 cell line was cultured in Leibovitz's L-15 medium supplemented with 10% FBS, 10 μg/ml insulin, 16 μg/ml glutathione (from Sigma). The MCF10A cell line was cultured in a 1:1 mixture of DMEM and Ham's F12 medium (Gibco, Life Technologies), supplemented with 5% Horse Serum (Invitrogen), 20 ng/ml human epidermal growth factor (EGF; Invitrogen), and 100 ng/ml cholera toxin, 10 μg/ml insulin and 500 ng/ml hydrocortisone (from Sigma). The UACC-812 cell line was cultured in Leibovitz's L-15 medium supplemented with 20% FBS, 2 mM L-glutamine and 20 ng/ml human EGF. All cells were cultured at 37°C in a humidified atmosphere containing 5% CO_2.

7.4 Cell transfection

Transfections were performed using calcium phosphate or Oligofectamine™ (Invitrogen) reagents, according to manufacturer's instructions. For lentiviral production cells were transfected with calcium phosphate (293T cells) or Oligofectamine™ (MCF10A cells) for siRNA experiments.

7.5 Silencing CDK12 expression by siRNA

Transient KD of CDK12 was achieved using a specific pre-designed Stealth siRNA and the corresponding non-targeting universal control siRNA (Invitrogen). Briefly, 60% confluent cells were seeded in 6-well plates for 24 hours prior to siRNA transfection. siRNA oligos were diluted at a final concentration of 10 nM in 175 µL OptiMEM (Invitrogen-Gibco Carlsbad, CA), and 4 µL Oligofectamine (Invitrogen, Carlsbad, CA) were diluted in 15 µl OptiMEM for each well to be transfected. These mixes were incubated at room temperature for 5 min, combined, and incubated for an additional 20 min. Growth medium was washed away from the cells and replaced with OptiMEM. The transfection mixture was then added. Following a 5 hour incubation at 37°C, 0.5 mL medium containing 10% FBS was added. After 72 hours, the cells were harvested for WB or IF analysis.

7.6 Infections

Lentivirus was generated by co-transfection of third generation helper vectors together with lentiviral vectors, pSico, pSicoR or pLVX, in 293T cells. Twenty-four hours after transfection the supernatant was concentrated to 5 ml for each 10 cm plate. After an additional 24 hours, supernatant was collected, filtered through 45 µm filters, and added to the target cells at 40 – 50% confluency. Cells were then incubated at 37°C for 12 hours. Forty-eight hours after infection, cells were split and puromycin (1 µg/ml) was added to select infected cells.

7.7 mRNA extraction and cDNA synthesis

mRNA from control and test cell lines was extracted using the RNeasy kit from Qiagen, according to the manufacturer's protocol.

Single stranded cDNA synthesis was performed using the SuperScriptIII reverse trancriptase (Invitrogen) following manufacturer's instructions. Briefly, 1 μg of total RNA was mixed with 250 ng of random primers in RNase-free water and then incubated at 70°C for 5 min. Following the incubation, 10X reaction buffer, dNTPs mix (0.5 mM final concentration), and 1 μl of reverse transcriptase were added to the mix (20 μl final volume) and the reaction was incubated at 42°C for 1 hour. Finally, the reaction was inactivated by heating at 70°C for 15 min.

7.8 Q-PCR

For Q-PCR experiments the Taqman chemistry was used. In the table below, the Taqman assays employed are listed:

gene symbol	Assay ID	Ref Seq
CDK12	Hs00212914_m1	NM_015083.1
ATF5	Hs01119208_m1	NM_001193646
BRCC3	Hs02386484_g12	NM_001018055.2
CDH2	Hs00983056_m1	NM_001792.3
CSDA	Hs01124964_m1	NM_001145426.1
HIST1H3A	Hs00543854_s1	NM_003529.2
IFT27	Hs01017625_m1	NM_001177701.1

gene symbol	Assay ID	Ref Seq
MCM9	Hs01024285_m1	NM_017696.2
MYBL2	Hs00942543_m1	NM_002466.2
OR3A2	Hs01635424_s1	NM_002551
HEY1	Hs00232618-m1	NM_012258
RASL11B	Hs00225132_m1	NM_023940.2
RHOU	Hs00221873_m1	NM_021205.5
SSX5	Hs00820186_m1	NM_021015.3
TUBA3E	Hs01941853_g1	NM_207312.2
ZEB1	Hs01566410_m1	NM_001128128.2
HES5	Hs01387464_g1	NM_001010926.1

7.9 FISH Analysis

FISH on breast TMAs and breast cell lines was performed in collaboration with the Molecular Pathology Unit at the IFOM-IEO Campus.

DNA probes were labeled with a fluorescent dye (Cy3-dUTP or Green-dUTP) that is incorporated into the DNA by nick translation. The following reaction mixture was used: 3µl Buffer 10X, 0.6µl dAGC, 0.3µl dUTP/Cy-3, 3µl ß-mercaptoethanol, 0.3µl DNA polymerase, 6µl DNase (1:700 in H_2O), and5 µg DNA in 30 µl final volume. The reaction mix was incubated at 16°C for 2 hours. The probe was precipitated using the following reaction mixture: 3µl salmon sperm DNA, 10µl Cot-1 DNA, 1/10 volume NaAcetate, 3 volumes of cold 100% ethanol. The reaction was then placed at -80°C for 15min (or at -20°C for 30 min) and then centrifuged at 13000 rpm for 20 min at 4°C. The supernatant was removed and the pellet dried and resuspended in 15 µl of hybridization mix (for 15ml: 7.5 ml ultrapure formamide, 6 ml dextran solphate 25%, 1.5 ml 20 x SSC). The mixture was then incubated with shaking for 10 min at room temperature.

To prepare the cells, the cell suspension was centrifuged at 1500 rpm for 10 min at room temperature. The supernatant was removed and the cell pellet was resuspended in 10 ml hypotonic solution (0.075M KCL) and incubated at 37°C for 18 min. Cells were then fixed in 3:1 methanol:acetic acid.

For the probe hybridyzation, probes, dissolved in hybridization mix, were placed on pretreated slides, and covered with a coverslip and sealed with rubber cement. Slides were placed in a Hybrite machine (Vysis) and incubated at 73°C for 3 min and then at 37°C overnight. Slides were then washed 3 times in 0.1x SSC for 5 min each wash, then incubated with DAPI (DAPI in 2 X SSC) for 5 min.

7.10 Protein Procedures

7.10.1 Cell lysis and protein purification

Cells were washed in PBS and lysed in RIPA lysis buffer [50 mmol/L Tris (pH 8), 120 mmol/L NaCl, 0.5% NP40, Phosphatase and protease inhibitors were added freshly to lysis and wash buffers: 20 mM Na pyrophosphate pH 7.5, 50 mM NaF, 2 mM PMSF in ethanol, 10mM Na vanadate, Protease Inhibitor Cocktail (Calbiochem). Cells were harvested directly on the plates using a cell scraper. About 300 μl of RIPA lysis buffer/10-cm plates and 50 μl RIPA buffer/for one well of a 6-well plate were used. Lysates were incubated on ice for 10 min and centrifuged at 12,000 rpm for 15 min at 4ºC. The supernatant was transferred to a new Eppendorf tube and protein concentration was measured by the Bradford assay (Biorad), following manufacturer’s instructions.

7.10.2 SDS polyacrylamide gel electrophoresis (SDS-PAGE)

Gels for resolution of proteins were made from a 30%, 29.1:1 mix of acrylamide:bisacrylamide (Sigma). As polymerization catalysts, 10% ammonium persulphate (APS) and TEMED were used.

Separating gel mix

	8%	15%
• acrylamide mix (ml)	2.7	5
• 1.5M Tris pH 8.8 (ml)	2.5	2.5
• distilled water (ml)	4.6	2.5
• 10% SDS (ml)	0.1	0.1
• 10% APS (ml)	0.1	0.1
• TEMED (ml)	0.006	0.004
• TOTAL (ml)	10	10

Stacking gel mix

• acrylamide mix	1.68 ml
• 1M Tris pH 6.8	1.36 ml
• distilled water	6.8 ml
• 10% SDS	0.1 ml
• 10% APS	0.1 ml
• TEMED	0.01 ml
• TOTAL	10 ml

7.10.3 Immunoblotting

Desired amounts of proteins were loaded onto 0.75 - 1.5 mm thick polyacrylamide gels for electrophoresis (Biorad). Proteins were transferred in western transfer tanks (Biorad) to nitrocellulose (Schleicher and Schnell) in 1 x Western Transfer buffer (diluted in 20% methanol) at 30 V overnight, or 100 V for 1 hour for small gels and at 70 V for 3 hours for large gels. Ponceau coloring was used to reveal the amount of protein transferred to the filters. Filters were blocked 1 hour (or overnight) in 5% milk or 5% BSA in TBS 0.1% Tween (TBS-T). After blocking, filters were incubated with the primary antibody, diluted in TBS-T with 5% milk or BSA, for 1 hour at room temperature, or overnight at 4ºC, followed by 3 washes of 5 min each in TBS-T. Filters were then incubated with the appropriate horseradish peroxidase (HRP)-conjugated secondary antibody diluted in TBS-T with 5% milk or BSA for 30 min. The primary antibody used were anti-CDK12 (AQ19) anti-vinculin and anti tubulin produced in-house; anti-cyclin D1 (DCS-6 Abcam) (Ab-3 NeoMarkers); anti-e-cadherin (BD); anti n-cadherin (BD). After the incubation with the secondary antibody, the filter was washed 3 times in TBS-T and the bound secondary antibody was revealed using the ECL (enhanced chemiluminescence) method (Amersham).

7.10.4 Immunoprecipitation

Lysates prepared in RIPA buffer were incubated in the presence of specific antibodies for 2 hours at 4ºC with rocking. Then, protein G Sepharose beads (Zymed) were added, and samples were left for an additional hour at 4ºC, rocking. Immunoprecipitates were then washed 4 times in RIPA buffer. After washing, beads were resuspended in 1:1 volume of 2 x SDS-PAGE Sample Buffer, boiled for 5 min at 95ºC, centrifuged for 1 minute and then loaded onto polyacrylamide gels.

7.10.5 Immunofluorescence

Cells were plated on glass coverslips pre-incubated with 0.5% gelatin in PBS at 37ºC for 30 min. Cells were fixed in 4% paraformaldehyde (in Pipes Buffer) for 10 min, washed with PBS and permeabilized in PBS 0.1% Triton X-100 for 10 min at room temperature. To prevent non-specific binding of the antibodies, cells were incubated with PBS in the presence of 5% BSA for 30 min. The coverslips were incubated with primary antibodies diluted in PBS 0.2 % BSA. After 1 hour of incubation at room temperature, coverslips were washed 3 times with PBS. Cells were then incubated for 30 min at room temperature with the appropriate secondary antibody Cy3 (Amersham), Alexa 488-conjugated (Molecular Probes).

After three washes in PBS, coverslips were mounted in a 90% glycerol solution containing diazabicyclo-(2.2.2)octane antifade (Sigma) and examined under a wild-field immunofluorescence microscope (Leica). Images were further processed with the Adobe Photoshop software (Adobe) or with Image J to merge the images of the single channels.

7.11 Constructs and plasmids

The pSicoR and pSico lentiviral vectors (Addgene), developed by Ventura *et al.* [135], were used for constitutive and Cre-inducible shRNA expression, respectively. Vectors were engineered to express shRNA specifically targeting CDK12 expression (sh#1, sh#7, sh#23) or luciferase (shLuc) as a control. The shRNA oligonucleotides were expressed under the control of a PGK promoter. The pSicoR vector expresses the puromycin resistance gene as selection marker. The pSico vector expresses EGFP as a selection marker. Oligos were designed and cloned according to the protocols available at: http://web.mit.edu/jacks-lab/protocols/pSico.html.

The pLVX-puro lentiviral vector was used to generate a construct to overexpress CDK12 in mammalian cells. The human CDK12 coding sequence, already present in the lab, was inserted by digestion with the restriction enzymes BamH1 and Sal1 and ligation into the pLVX-puro vector. With this system, expression of CDK12 is driven by the human cytomegalovirus immediate early promoter, located just upstream of the multiple cloning site. pLVX contains a puromycin resistance gene under the control of the murine phosphoglycerate kinase promoter for the selection of stable transductants.

7.12 Basic cloning techniques

7.12.1 Agarose gel electrophoresis

DNA samples were loaded onto 0.8 - 2% agarose gels along with DNA markers. Gels were made in TAE buffer containing 0.3 μg/ml ethidium bromide and run at 80 V until the desired separation was achieved. DNA bands were visualized under a UV lamp.

7.12.2 Transformation of competent cells

Fresh competent cells (50 μl), Top10 (Invitrogen) for cloning and DNA preparation or electro-competent DH10B cells (produced in-house), were thawed on ice for approximately 10 min prior to the addition of plasmid DNA. Cells were incubated with DNA on ice for 30 min and then subjected to a heat shock for 1 min at 42°C. Cells were then returned to ice for 2 min. SOC medium (300 μl) was then added and the cells were left at 37°C for 1 hour before plating them onto LB-agar plates with the appropriate antibiotic. Two plates for each reaction were used, one plated with 2/3 of the transformed bacterial cells and the other one with the rest. Plates were incubated overnight at 37°C.

7.12.3 Minipreps

Clones picked from individual colonies were used to inoculate 5 ml LB (containing the appropriate antibiotic) and grown overnight at 37°C. Bacteria cells were transferred to Eppendorf tubes and pelleted for 5 min at 12,000 rpm. Minipreps were performed with the Wizard® Plus SV Minipreps Kit (Promega) following the manufacturer's instructions. The plasmids were eluted in 50 μl nuclease free H_2O.

7.12.4 Diagnostic DNA restriction

Between 0.5 and 5 μg DNA were digested for 2 hours at 37°C with 10 – 20 units of restriction enzyme (New England Biolabs). For digestion, the volume was made up depending on the DNA volume to 20 – 50 μl with the appropriate buffer and ddH_2O.

7.12.5 Large scale plasmid preparation

Cells containing transfected DNA were expanded into 200 ml cultures overnight. Plasmid DNA was isolated from these cells using the Qiagen Maxi-prep kit according to the manufacturer's instructions.

7.13 Biological assays

7.13.1 Proliferation assay

Cells were seeded in triplicate, per cell line, per time point, in 6-well tissue culture plates (T0). Cells were detached by trypsin, resuspended in serum containing medium and counted using a hemocytometer at defined time-points for 7 days or 10 days, according to the time taken to reach confluency.

7.13.2 Colony forming assay

Cells (500 cells for MCF10A; 10,000 cells for BT474 and HCC1569) were seeded under adhesion conditions in 10-cm tissue culture plates and cultured for 10 (MCF10A) or 20 (BT474 and HCC1569) days. Colonies were then fixed and stained using a crystal violet solution (crystal violet "Sigma" 1%, ethanol 35%). The number and size of colonies were determined using the ImageJ software

7.13.3 3D-Matrigel Assay

Cells were seeded in 4-well chamber slides (LabTek: 2000 cells/ml/well) in Matrigel (BD-Biosciences) overlay or embedding conditions as described by Lee *et al.* [134], and then cultured for 15 (MCF10A) or 20 (BT474 and HCC1569) days. The resulting outgrowths were photographed and analyzed. All measurements were made with ImageJ software.

7.13.4 *In vivo* xenograft assays

Six to 8 week-old NOD/SCID IL2R gamma-chain null female mice were injected in the inguinal mammary fat pad with 150,000 BT474 and 500,000 HCC1569 cells resuspended in 40 μL of a 1:1 Matrigel-PBS solution. Mice were monitored by hand-palpation for tumor development. Tumor growth was measured by using a vernier caliper and applying the standard formula: tumor volume = $(a \times b^2)/2$. Mice were sacrificed when tumors reached a dimension of 1.5 – 2 mm^3. Tumors were explanted, weighed, and processed for formalin-fixing and paraffin embedding.

7.14 Exon Array

Total RNA from control and test cell lines was extracted using the RNeasy kit from Qiagen, according to the manufacturer's protocol. The Microarray Facility at IFOM performed all the microarray hybridization and signal detection processes.

7.14.1 Affymetrix gene expression analysis

All the bioinformatic analyses were performed in collaboration with Fabrizio Bianchi, IEO, Milan. Data were normalized using the Robust Multiarray Average algorithm (RMA) [18] and performed by AltAnalyze software [21]. RMA chip summarization data were then analyzed to identify probesets that were regulated in a statistically significant manner in the different experimental conditions (i.e., CDK12 overexpressed or ablated *vs.* controls). Probesets with a p-value < 0.05 (Welch's t-test) were considered as differentially expressed. All statistical analyses and CDK12 over/down-regulated gene set comparisons were carried out using SAS statistical software (SAS Institute, Inc., Cary, NC). Gene set enrichment analysis and leading-edge analysis were performed using GSEA software [163] [24]. Gene sets (total of 3272 gene sets, C2 category) were downloaded from the MSIGDB database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). A gene set was considered as significantly enriched when the false discovery rate (FDR) was less than 25%. Ingenuity pathway analysis (IPA; http://www.ingenuity.com/) was performed using CDK12 over- or down-regulated genes that were overlapping in more than one MSIGDB gene sets.

7.14.2 Affymetrix differential splicing analysis

We used two different algorithms to detect splicing variants that were available in AltAnalyze software [21]. The first algorithm was the Splicing Index (SI) that calculates a normalised index (NI) for each exon, which is the ratio of the exon-level signal to the gene-level signal. NI represents the exon inclusion rate and can be used in statistical testing to detect differential splicing between sample groups. This strategy eliminates differential gene-level expression in a simple manner, however, it relies heavily on the correct estimation of the gene-level expression. After the calculation of NI, one can calculate the SI, which measures the difference in NI between two samples as follows: SI = $\log_2$(NIsample1/NIsample2). After the calculation of SI one can use statistical methods, such as ANOVA, to determine which genes show alternative splicing.

The second algorithm we used was the FIRMA (finding isoforms using robust multichip analysis) that frames the problem of detecting alternative splicing as a problem of outlier detection. FIRMA is an alternative to the SI approach to calculate alternative splicing statistics. Rather than using the probe set expression values to determine differences in the relative expression of an exon for two or more conditions, FIRMA uses the residual values produced by the RMA algorithm for each probe, corresponding to a gene. The median of the residuals for each probe set, for each array sample, is compared to the median absolute deviation for all residuals and samples for the gene [23].

We used UCSC Genome Browser (http://genome.ucsc.edu/) to align Affymetrix exons probe sets and relative expression value on the human genome for further analysis of alternative splicing events.

8 Bibliography

1. Ferlay, J.*, et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* **127**, 2893-2917 (2010).
2. Capra, M.*, et al.* Frequent alterations in the expression of serine/threonine kinases in human cancers. *Cancer Res* **66**, 8147-8154 (2006).
3. Daniel, C.W. & Smith, G.H. The mammary gland: a model for development. *J Mammary Gland Biol Neoplasia* **4**, 3-8 (1999).
4. Visvader, J.E. Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. *Genes Dev* **23**, 2563-2577 (2009).
5. Sargent, D.J., Conley, B.A., Allegra, C. & Collette, L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* **23**, 2020-2027 (2005).
6. Abner, A.L.*, et al.* Correlation of tumor size and axillary lymph node involvement with prognosis in patients with T1 breast carcinoma. *Cancer* **83**, 2502-2508 (1998).
7. Kim, J.Y.*, et al.* The prognostic significance of the lymph node ratio in axillary lymph node positive breast cancer. *J Breast Cancer* **14**, 204-212 (2011).
8. Solomayer, E.F., Diel, I.J., Meyberg, G.C., Gollan, C. & Bastert, G. Metastatic breast cancer: clinical course, prognosis and therapy related to the first site of metastasis. *Breast Cancer Res Treat* **59**, 271-278 (2000).
9. Rakha, E.A.*, et al.* Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res* **12**, 207 (2010).
10. Jones, R.L.*, et al.* The prognostic significance of Ki67 before and after neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res Treat* **116**, 53-68 (2009).
11. Ross, J.S.*, et al.* The Her-2/neu gene and protein in breast cancer 2003: biomarker and target of therapy. *Oncologist* **8**, 307-325 (2003).
12. Lim, E., Metzger-Filho, O. & Winer, E.P. The natural history of hormone receptor-positive breast cancer. *Oncology (Williston Park)* **26**, 688-694, 696 (2012).
13. Sinn, H.P., Helmchen, B. & Wittekind, C.H. [TNM classification of breast cancer: changes and comments on the 7th edition]. *Pathologe* **31**, 361-366 (2010).
14. Mook, S.*, et al.* Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study. *Lancet Oncol* **10**, 1070-1076 (2009).
15. Goldhirsch, A.*, et al.* Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. *Ann Oncol* **20**, 1319-1329 (2009).
16. Galea, M.H., Blamey, R.W., Elston, C.E. & Ellis, I.O. The Nottingham Prognostic Index in primary breast cancer. *Breast cancer research and treatment* **22**, 207-219 (1992).
17. Veronesi, U., Stafyla, V., Luini, A. & Veronesi, P. Breast cancer: from "maximum tolerable" to "minimum effective" treatment. *Front Oncol* **2**, 125 (2012).
18. Veronesi, U.*, et al.* Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *The New England journal of medicine* **347**, 1227-1232 (2002).
19. Veronesi, U.*, et al.* Comparing radical mastectomy with quadrantectomy, axillary dissection, and radiotherapy in patients with small cancers of the breast. *The New England journal of medicine* **305**, 6-11 (1981).

20. Veronesi, U., Boyle, P., Goldhirsch, A., Orecchia, R. & Viale, G. Breast cancer. *Lancet* **365**, 1727-1741 (2005).
21. Veronesi, U*., et al.* Radiotherapy after breast-preserving surgery in women with localized cancer of the breast. *The New England journal of medicine* **328**, 1587-1591 (1993).
22. Jensen, E.V., Block, G.E., Smith, S., Kyser, K. & DeSombre, E.R. Estrogen receptors and breast cancer response to adrenalectomy. *Natl Cancer Inst Monogr* **34**, 55-70 (1971).
23. Bardou, V.J., Arpino, G., Elledge, R.M., Osborne, C.K. & Clark, G.M. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. *J Clin Oncol* **21**, 1973-1979 (2003).
24. Chang, J*., et al.* Prediction of clinical outcome from primary tamoxifen by expression of biologic markers in breast cancer patients. *Clin Cancer Res* **6**, 616-621 (2000).
25. Fang, L., Barekati, Z., Zhang, B., Liu, Z. & Zhong, X. Targeted therapy in breast cancer: what's new? *Swiss Med Wkly* **141**, w13231 (2011).
26. Tavassoli, F.A. *Pathology of the breast*, (Appleton & Lange, Stamford, Conn., 1999).
27. Brown, T.M. & Fee, E. Rudolf Carl Virchow: medical scientist, social reformer, role model. *Am J Public Health* **96**, 2104-2105 (2006).
28. Komaki, K., Sano, N. & Tangoku, A. Problems in histological grading of malignancy and its clinical significance in patients with operable breast cancer. *Breast Cancer* **13**, 249-253 (2006).
29. Malhotra, G.K., Zhao, X., Band, H. & Band, V. Histological, molecular and functional subtypes of breast cancers. *Cancer biology & therapy* **10**, 955-960 (2010).
30. Li, C.I., Uribe, D.J. & Daling, J.R. Clinical characteristics of different histologic types of breast cancer. *British journal of cancer* **93**, 1046-1052 (2005).
31. Sims, A.H., Howell, A., Howell, S.J. & Clarke, R.B. Origins of breast cancer subtypes and therapeutic implications. *Nat Clin Pract Oncol* **4**, 516-525 (2007).
32. Wellings, S.R. A hypothesis of the origin of human breast cancer from the terminal ductal lobular unit. *Pathol Res Pract* **166**, 515-535 (1980).
33. Lee, S*., et al.* Hormones, receptors, and growth in hyperplastic enlarged lobular units: early potential precursors of breast cancer. *Breast Cancer Res* **8**, R6 (2006).
34. Perou, C.M*., et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000).
35. Sorlie, T*., et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869-10874 (2001).
36. Weigelt, B*., et al.* Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol* **11**, 339-349 (2010).
37. Sorlie, T*., et al.* Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics* **7**, 127 (2006).
38. Sorlie, T*., et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418-8423 (2003).
39. Sotiriou, C. & Pusztai, L. Gene-expression signatures in breast cancer. *N Engl J Med* **360**, 790-800 (2009).
40. Russnes, H.G*., et al.* Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci Transl Med* **2**, 38ra47 (2010).

41. Curtis, C., *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352 (2012).
42. van 't Veer, L.J., *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).
43. Wang, Y., *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-679 (2005).
44. Buyse, M., *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* **98**, 1183-1192 (2006).
45. Paik, S., *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817-2826 (2004).
46. Harris, L., *et al.* American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol* **25**, 5287-5312 (2007).
47. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685-696 (2010).
48. Ding, L., *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005 (2010).
49. Shah, S.P., *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809-813 (2009).
50. Visvader, J.E. Cells of origin in cancer. *Nature* **469**, 314-322 (2011).
51. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).
52. Navin, N., *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
53. Clarke, M.F., *et al.* Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells. *Cancer Res* **66**, 9339-9344 (2006).
54. Bielas, J.H. & Loeb, L.A. Mutator phenotype in cancer: timing and perspectives. *Environ Mol Mutagen* **45**, 206-213 (2005).
55. (!!! INVALID CITATION !!!).
56. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).
57. Croce, C.M. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* **10**, 704-714 (2009).
58. Yao, J., *et al.* Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res* **66**, 4065-4078 (2006).
59. Morrison, L.E., *et al.* Effects of ERBB2 amplicon size and genomic alterations of chromosomes 1, 3, and 10 on patient response to trastuzumab in metastatic breast cancer. *Genes Chromosomes Cancer* **46**, 397-405 (2007).
60. Luoh, S.W. Amplification and expression of genes from the 17q11 approximately q12 amplicon in breast cancer cells. *Cancer Genet Cytogenet* **136**, 43-47 (2002).
61. Kauraniemi, P., Barlund, M., Monni, O. & Kallioniemi, A. New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Res* **61**, 8235-8240 (2001).
62. Kauraniemi, P., Kuukasjarvi, T., Sauter, G. & Kallioniemi, A. Amplification of a 280-kilobase core region at the ERBB2 locus leads to activation of two hypothetical proteins in breast cancer. *The American journal of pathology* **163**, 1979-1984 (2003).
63. Sircoulomb, F., *et al.* Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer* **10**, 539 (2010).

64. Kauraniemi, P. & Kallioniemi, A. Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer. *Endocr Relat Cancer* **13**, 39-49 (2006).
65. Kao, J. & Pollack, J.R. RNA interference-based functional dissection of the 17q12 amplicon in breast cancer reveals contribution of coamplified genes. *Genes Chromosomes Cancer* **45**, 761-769 (2006).
66. Katz, E.*, et al.* A gene on the HER2 amplicon, C35, is an oncogene in breast cancer whose actions are prevented by inhibition of Syk. *Br J Cancer* **103**, 401-410 (2010).
67. Cui, J.*, et al.* Cross-talk between HER2 and MED1 Regulates Tamoxifen Resistance of Human Breast Cancer Cells. *Cancer Res* **72**, 5625-5634 (2012).
68. Hynes, N.E. & Stern, D.F. The biology of erbB-2/neu/HER-2 and its role in cancer. *Biochim Biophys Acta* **1198**, 165-184 (1994).
69. Oved, S. & Yarden, Y. Signal transduction: molecular ticket to enter cells. *Nature* **416**, 133-136 (2002).
70. Slamon, D.J.*, et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177-182 (1987).
71. Shalaby, M.R.*, et al.* Development of humanized bispecific antibodies reactive with cytotoxic lymphocytes and tumor cells overexpressing the HER2 protooncogene. *J Exp Med* **175**, 217-225 (1992).
72. Dillon, R.L., White, D.E. & Muller, W.J. The phosphatidyl inositol 3-kinase signaling network: implications for human breast cancer. *Oncogene* **26**, 1338-1345 (2007).
73. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
74. Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
75. Mita, M.M., Mita, A. & Rowinsky, E.K. Mammalian target of rapamycin: a new molecular target for breast cancer. *Clin Breast Cancer* **4**, 126-137 (2003).
76. Sarbassov, D.D., Guertin, D.A., Ali, S.M. & Sabatini, D.M. Phosphorylation and regulation of Akt/PKB by the rictor-mTOR complex. *Science* **307**, 1098-1101 (2005).
77. Steeg, P.S. & Zhou, Q. Cyclins and breast cancer. *Breast Cancer Res Treat* **52**, 17-28 (1998).
78. Eilers, M. & Eisenman, R.N. Myc's broad reach. *Genes Dev* **22**, 2755-2766 (2008).
79. Meyer, N. & Penn, L.Z. Reflecting on 25 years with MYC. *Nat Rev Cancer* **8**, 976-990 (2008).
80. Nass, S.J. & Dickson, R.B. Defining a role for c-Myc in breast tumorigenesis. *Breast Cancer Res Treat* **44**, 1-22 (1997).
81. Hartmann, A., Blaszyk, H., Kovach, J. & Sommer, S. The molecular epidemiology of p53 gene mutations in human breast cancer. *Trends in genetics : TIG* **13**, 27-33 (1997).
82. Lane, D. Cancer. p53, guardian of the genome. *Nature* **358**, 15-16 (1992).
83. Levine, A. p53, the cellular gatekeeper for growth and division. *Cell* **88**, 323-331 (1997).
84. Venkitaraman, A. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**, 171-182 (2002).
85. Scully, R.*, et al.* Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. *Cell* **90**, 425-435 (1997).
86. Scully, R.*, et al.* Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell* **88**, 265-275 (1997).

87. Wong, A., Pero, R., Ormonde, P., Tavtigian, S. & Bartel, P. RAD51 interacts with the evolutionarily conserved BRC motifs in the human breast cancer susceptibility gene brca2. *The Journal of biological chemistry* **272**, 31941-31944 (1997).
88. Deng, C.-X. BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic acids research* **34**, 1416-1426 (2006).
89. Bosco, E.E. & Knudsen, E.S. RB in breast cancer: at the crossroads of tumorigenesis and treatment. *Cell Cycle* **6**, 667-671 (2007).
90. Osborne, C., Wilson, P. & Tripathy, D. Oncogenes and tumor suppressor genes in breast cancer: potential diagnostic and therapeutic applications. *Oncologist* **9**, 361-377 (2004).
91. Hirohashi, S. Inactivation of the E-cadherin-mediated cell adhesion system in human cancers. *The American journal of pathology* **153**, 333-339 (1998).
92. van Wezel, T.*, et al.* Expression analysis of candidate breast tumour suppressor genes on chromosome 16q. *Breast Cancer Res* **7**, R998-1004 (2005).
93. Burkhart, D. & Sage, J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat Rev Cancer* **8**, 671-682 (2008).
94. Mootha, V.K.*, et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-273 (2003).
95. Knudson, A.G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-823 (1971).
96. Andersen, T.I.*, et al.* Genetic alterations of the tumour suppressor gene regions 3p, 11p, 13q, 17p, and 17q in human breast carcinomas. *Genes Chromosomes Cancer* **4**, 113-121 (1992).
97. Emig, D.*, et al.* AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res* **38**, W755-762 (2010).
98. Besson, A., Dowdy, S.F. & Roberts, J.M. CDK inhibitors: cell cycle regulators and beyond. *Dev Cell* **14**, 159-169 (2008).
99. Cantley, L. & Neel, B. New insights into tumor suppression: PTEN suppresses tumor formation by restraining the phosphoinositide 3-kinase/AKT pathway. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 4240-4245 (1999).
100. Mills, G.*, et al.* The role of genetic abnormalities of PTEN and the phosphatidylinositol 3-kinase pathway in breast and ovarian tumorigenesis, prognosis, and therapy. *Seminars in oncology* **28**, 125-141 (2001).
101. Ko, T.K., Kelly, E. & Pines, J. CrkRS: a novel conserved Cdc2-related protein kinase that colocalises with SC35 speckles. *J Cell Sci* **114**, 2591-2603 (2001).
102. Chen, H.H., Wang, Y.C. & Fann, M.J. Identification and characterization of the CDK12/cyclin L1 complex involved in alternative splicing regulation. *Mol Cell Biol* **26**, 2736-2745 (2006).
103. Bartkowiak, B.*, et al.* CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev* **24**, 2303-2316 (2010).
104. Blazek, D.*, et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev* **25**, 2158-2172 (2011).
105. Fu, X.D. The superfamily of arginine/serine-rich splicing factors. *RNA* **1**, 663-680 (1995).
106. Busch, A. & Hertel, K.J. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA* **3**, 1-12 (2012).
107. Williamson, M.P. The structure and function of proline-rich regions in proteins. *Biochem J* **297 (Pt 2)**, 249-260 (1994).

108. Ball, L.J., Kuhne, R., Schneider-Mergener, J. & Oschkinat, H. Recognition of proline-rich motifs by protein-protein-interaction domains. *Angew Chem Int Ed Engl* **44**, 2852-2869 (2005).
109. Sotiroudis, T.G. & Xenakis, A. Pest sequences present in the subunits of phosphorylase kinase. *Biochem Int* **21**, 941-947 (1990).
110. Rechsteiner, M. & Rogers, S.W. PEST sequences and regulation by proteolysis. *Trends Biochem Sci* **21**, 267-271 (1996).
111. Mintz, P.J., Patterson, S.D., Neuwald, A.F., Spahr, C.S. & Spector, D.L. Purification and biochemical characterization of interchromatin granule clusters. *EMBO J* **18**, 4308-4320 (1999).
112. Malumbres, M*., et al.* Cyclin-dependent kinases: a family portrait. *Nat Cell Biol* **11**, 1275-1276 (2009).
113. Buratowski, S. The CTD code. *Nat Struct Biol* **10**, 679-680 (2003).
114. Egloff, S. & Murphy, S. Cracking the RNA polymerase II CTD code. *Trends Genet* **24**, 280-288 (2008).
115. Price, D.H. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol Cell Biol* **20**, 2629-2634 (2000).
116. Peterlin, B.M. & Price, D.H. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell* **23**, 297-305 (2006).
117. Rodrigues, F., Thuma, L. & Klambt, C. The regulation of glial-specific splicing of Neurexin IV requires HOW and Cdk12 activity. *Development* **139**, 1765-1776 (2012).
118. Morris, D.P. & Greenleaf, A.L. The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J Biol Chem* **275**, 39935-39943 (2000).
119. Chung, S., McLean, M.R. & Rymond, B.C. Yeast ortholog of the Drosophila crooked neck protein promotes spliceosome assembly through stable U4/U6.U5 snRNP addition. *RNA* **5**, 1042-1054 (1999).
120. Knuutila, S*., et al.* DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. *Am J Pathol* **152**, 1107-1123 (1998).
121. Kallioniemi, O.P*., et al.* ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. *Proc Natl Acad Sci U S A* **89**, 5321-5325 (1992).
122. Kallioniemi, A*., et al.* Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc Natl Acad Sci U S A* **91**, 2156-2160 (1994).
123. Revillion, F., Bonneterre, J. & Peyrat, J.P. ERBB2 oncogene in human breast cancer and its clinical significance. *Eur J Cancer* **34**, 791-808 (1998).
124. Isola, J.J*., et al.* Genetic aberrations detected by comparative genomic hybridization predict outcome in node-negative breast cancer. *Am J Pathol* **147**, 905-911 (1995).
125. Purdom, E*., et al.* FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics* **24**, 1707-1714 (2008).
126. Eglen, R.M. & Reisine, T. The current status of drug discovery against the human kinome. *Assay Drug Dev Technol* **7**, 22-43 (2009).
127. Cheng, S.W*., et al.* The interaction of CDK12/CrkRS with CYCLIN K1 is required for the phosphorylation of the C-terminal domain of RNA Pol II. *Mol Cell Biol* (2012).
128. Debnath, J., Muthuswamy, S.K. & Brugge, J.S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* **30**, 256-268 (2003).
129. Franken, N.A., Rodermond, H.M., Stap, J., Haveman, J. & van Bree, C. Clonogenic assay of cells in vitro. *Nat Protoc* **1**, 2315-2319 (2006).

130. Bissell, M.J., Kenny, P.A. & Radisky, D.C. Microenvironmental regulators of tissue structure and function also regulate tumor induction and progression: the role of extracellular matrix and its degrading enzymes. *Cold Spring Harb Symp Quant Biol* **70**, 343-356 (2005).
131. Kenny, P.A.*, et al.* The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. *Mol Oncol* **1**, 84-96 (2007).
132. Shaw, K.R., Wrobel, C.N. & Brugge, J.S. Use of three-dimensional basement membrane cultures to model oncogene-induced changes in mammary epithelial morphogenesis. *J Mammary Gland Biol Neoplasia* **9**, 297-310 (2004).
133. Debnath, J. & Brugge, J.S. Modelling glandular epithelial cancers in three-dimensional cultures. *Nat Rev Cancer* **5**, 675-688 (2005).
134. Lee, G.Y., Kenny, P.A., Lee, E.H. & Bissell, M.J. Three-dimensional culture models of normal and malignant breast epithelial cells. *Nat Methods* **4**, 359-365 (2007).
135. Ventura, A.*, et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 10380-10385 (2004).
136. Martinez, I. & Dimaio, D. B-Myb, cancer, senescence, and microRNAs. *Cancer Res* **71**, 5370-5373 (2011).
137. Chuang, Y.Y., Valster, A., Coniglio, S.J., Backer, J.M. & Symons, M. The atypical Rho family GTPase Wrch-1 regulates focal adhesion formation and cell migration. *J Cell Sci* **120**, 1927-1934 (2007).
138. Qin, H., Wang, Z., Diener, D. & Rosenbaum, J. Intraflagellar transport protein 27 is a small G protein involved in cell-cycle control. *Curr Biol* **17**, 193-202 (2007).
139. Miskinyte, S.*, et al.* Loss of BRCC3 deubiquitinating enzyme leads to abnormal angiogenesis and is associated with syndromic moyamoya. *Am J Hum Genet* **88**, 718-728 (2011).
140. Manning, A.L. & Dyson, N.J. RB: mitotic implications of a tumour suppressor. *Nat Rev Cancer* **12**, 220-226 (2012).
141. Velasco-Velazquez, M.A.*, et al.* Examining the role of cyclin D1 in breast cancer. *Future Oncol* **7**, 753-765 (2011).
142. Knudsen, K.E., Diehl, J.A., Haiman, C.A. & Knudsen, E.S. Cyclin D1: polymorphism, aberrant splicing and cancer risk. *Oncogene* **25**, 1620-1628 (2006).
143. Solomon, D.A.*, et al.* Cyclin D1 splice variants. Differential effects on localization, RB phosphorylation, and cellular transformation. *J Biol Chem* **278**, 30339-30347 (2003).
144. Wei, M.*, et al.* Knocking down cyclin D1b inhibits breast cancer cell growth and suppresses tumor development in a breast cancer model. *Cancer Sci* **102**, 1537-1544 (2011).
145. Lord, C.J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287-294 (2012).
146. Winnepenninckx, V.*, et al.* Gene expression profiling of primary cutaneous melanoma and clinical outcome. *Journal of the National Cancer Institute* **98**, 472-482 (2006).
147. Sarasin, A. & Kauffmann, A. Overexpression of DNA repair genes is associated with metastasis: a new hypothesis. *Mutat Res* **659**, 49-55 (2008).
148. McCracken, S.*, et al.* The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**, 357-361 (1997).
149. Perales, R. & Bentley, D. "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. *Mol Cell* **36**, 178-191 (2009).

150. Moore, M.J. & Proudfoot, N.J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**, 688-700 (2009).
151. Munoz, M.J., de la Mata, M. & Kornblihtt, A.R. The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends Biochem Sci* **35**, 497-504 (2010).
152. Depowski, P.L., Rosenthal, S.I. & Ross, J.S. Loss of expression of the PTEN gene protein product is associated with poor outcome in breast cancer. *Mod Pathol* **14**, 672-676 (2001).
153. Irizarry, R.A.*, et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).
154. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16**, 990-995 (2009).
155. Luco, R.F.*, et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000 (2010).
156. de Almeida, S.F.*, et al.* Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nat Struct Mol Biol* **18**, 977-983 (2011).
157. Kim, S., Kim, H., Fong, N., Erickson, B. & Bentley, D.L. Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proc Natl Acad Sci U S A* **108**, 13564-13569 (2011).
158. Fuchs, S.M., Kizer, K.O., Braberg, H., Krogan, N.J. & Strahl, B.D. RNA polymerase II carboxyl-terminal domain phosphorylation regulates protein stability of the Set2 methyltransferase and histone H3 di- and trimethylation at lysine 36. *The Journal of biological chemistry* **287**, 3249-3256 (2012).
159. Fisher, B.*, et al.* Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. *J Clin Oncol* **16**, 2672-2685 (1998).
160. Vogel, C.L.*, et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J Clin Oncol* **20**, 719-726 (2002).
161. Kononen, J.*, et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* **4**, 844-847 (1998).
162. Capra, M.*, et al.* Frequent alterations in the expression of serine/threonine kinases in human cancers. *Cancer Res* **66**, 8147-8154 (2006).
163. Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005).

Acknowledgments

First and foremost I am very grateful to my PhD supervisor, Prof. Pier Paolo Di Fiore, for giving me the great opportunity to work in his exciting laboratory and for his guidance during the years of my PhD study.

A very special thanks goes to Prof. Salvatore Pece for his constant support during these years, for the passion he put in science and especially for teaching me to enjoy my job and that sacrifice is the first step toward success.

I want also to express my gratitude to Dr. Manuela Vecchi for her precious work and for guiding me during my first steps in the lab.

I thank all the people from Pier Paolo Di Fiore's lab for their friendship and the numerous scientific discussions, all the IFOM-IEO Campus Facilities, and Dr. Rosalind Gunby for critically reading of my thesis.

I thank Prof. Kristian Helin at BRIC of Copenaghen and Prof. Saverio Minucci at IFOM-IEO-Campus for co-supervising this thesis.

Infine un grazie immenso alla mia famiglia, Francesca, Michele, Enrichetta, Mariano e Aniello, senza i quali non avrei mai potuto raggiungere questo obiettivo e alla mia compagna Marta che mi ha sopportato e mi e' stata vicina durante tutti gli anni di questo percorso.