PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

# A novel approach for the identification of candidate driver lesions in breast cancer based on the comparison of the mutational profiles of a primary tumour and its matched mammospheres and xenograft.

*Valentina Melocchi*

Istituto Europeo di Oncologia (IEO), Milan

Matricola n. R10340

*Supervisor:* Prof. Pier Paolo Di Fiore

Fondazione Istituto FIRC di Oncologia Molecolare (IFOM), Milan
Istituto Europeo di Oncologia (IEO), Milan
Dipartimento di Oncologia ed Emato-Oncologia, Università degli Studi di Milano.

*Added Supervisors:* Prof. Salvatore Pece

Istituto Europeo di Oncologia (IEO), Milan
Dipartimento di Oncologia ed Emato-Oncologia, Università degli Studi di Milano.

Dr. Fabrizio Bianchi
Institute for Stem-cell Biology, Regenerative Medicine and Innovative Therapies (ISBReMIT), San Giovanni Rotondo (FG)

Anno accademico 2015-2016

*Dedicated to Omar*

## TABLE OF CONTENTS

# List of Figures

# List of Tables

# List of Abbreviations

DCIS = Ductal Carcinoma *In Situ*

LCIS = Lobular Carcinoma *In Situ*

IDC = Invasive Ductal Carcinoma

ILC = Invasive Lobular Carcinoma

EMT = Epithelial-to-Mesenchymal Transition

HSC = Hematopoietic Stem Cell

CTC = Circulating Tumour Cell

NGS = Next-Generation Sequencing

TCGA = The Cancer Genome Atlas

ICGC = International Cancer Genome Consortium

PDX = Patient-Derived Xenograft

WGA = Whole-Genome Amplification

SRA = Sequence Read Archive

SNVs = Single Nucleotide Variations

Indels = Insertions and Deletions

IGV = Integrative Genome Viewer

gDNA = Genomic DNA

WGA-DNA = Whole-Genome Amplified DNA

WGS = Whole Genome Sequencing

WES = Whole Exome Sequencing

GVs = Germline Variations

WGA-WES = Whole-Genome Amplification followed by Whole Exome Sequencing

# Abstract

The clinical management of breast cancer patients is complicated by the high genetic heterogeneity of this disease, which makes the standardization of treatments, the prediction of prognosis and therapy response, and the development of personalized therapies difficult. Nevertheless, the advent of high-throughput genomics screenings based on microarray or next-generation sequencing (NGS) technologies has greatly enhanced our understanding of the genomic landscapes underlying breast cancer development and progression. Such discoveries are now allowing clinicians to tailor therapies based on the molecular subtype of the tumour (luminal, basal and HER2).

NGS studies have also started to provide insights into the range of molecular profiles of tumour cells from the same tumour, and have shown that in some breast cancers a high level of intratumoral genetic heterogeneity exists. The findings from these studies support a scenario in which breast tumours can be either: i) monogenomic, comprised of a single clonal cell population; ii) or polygenomic, composed of several related clonal subpopulations. The co-existence of different cancer driver genetic lesions in polygenomic tumours might contribute to treatment failure in some cases, as relapse could be driven by the expansion of a subpopulation of cells intrinsically resistant to the therapy. Importantly, cancer genetic heterogeneity has been recapitulated in experimental settings using cancer stem cells (CSCs) xenografted in mouse models.

We hypothesized that the mutational events that drive the onset and progression of breast tumours lie within the CSC compartment. To explore this possibility, we analysed and compared the mutational profiles of a primary breast tumour and its matched mammospheres (source of CSC-derived population), patient-derived xenograft (PDX) and PDX-derived mammospheres using Whole Exome Sequencing (WES).

We setup a NGS approach to look for rare mutations in the primary tumour that may be present in the CSC compartment using low amounts of DNA input. We optimised an experimental protocol in which the genomic DNA (gDNA) of each sample was

subjected to Whole Genome Amplification (WGA) prior to performing WES. This enabled us to obtain a sufficient amount of DNA (≥ 3 μg) to perform WES. We also introduced a filtering step in our analysis, based on the Xenome software, for PDX-derived samples to eliminate possible contamination from murine DNA.

Our study allowed us to characterize the genetic profiles of CSCs and to identify cancer-relevant mutations that could drive breast cancer onset and progression. We identified 15 candidate driver mutations in 11 genes that were enriched, in terms of mutation frequency, within primary tumour-derived mammospheres and the PDX. Together with these mutations, we identified 4 mutations in 4 genes, not enriched, but shared among all analysed samples, which likely represent "founder" mutations.

Based on our results, we will now endeavour to determine the clinical relevance of the candidate driver mutations identified in our study by determining their prevalence in independent patient cohorts. Having optimised the protocol for NGS of matched primary tumour, PDX and mammosphere populations, we will also extend our mutational analysis to additional breast tumours for the identification of more driver mutations and for the deconvolution of intratumoral genetic heterogeneity of breast cancer.

Understanding the driving mutational forces of breast tumours and relative mechanisms involved is paramount for the development of more effective therapeutic strategies.

# Introduction

## 1. Breast Cancer

### 1.1 Epidemiology

Breast cancer is the second most frequent cancer type worldwide and the first among women (Ferlay et al., 2015). Approximately 250,000 new cases are expected in 2016 in the United States alone (Siegel et al., 2016) (Figure 1). Prognosis of breast cancer largely depends on the tumour stage at diagnosis, i.e. from 26% of survival at 5 years for patients with distant metastases to 99% for patients with localized disease (Siegel et al., 2016). Therefore, breast cancer screening programs (i.e., mammography screening) are paramount for reducing disease burden by augmenting early diagnoses; indeed, mammography screening has been shown to be effective in reducing breast cancer mortality in large randomized screening studies (Gill et al., 2004; Joensuu et al., 2004). However, when a patient presents with advanced disease the chances of a curative treatment decrease due to the frequent propensity of breast cancer to become metastatic and chemoresistant.

**Estimated New Cases**

| | Males | | | Females | |
|---|---|---|---|---|---|
| Prostate | 180,890 | 21% | | Breast | 246,660 | 29% |
| Lung & bronchus | 117,920 | 14% | | Lung & bronchus | 106,470 | 13% |
| Colon & rectum | 70,820 | 8% | | Colon & rectum | 63,670 | 8% |
| Urinary bladder | 58,950 | 7% | | Uterine corpus | 60,050 | 7% |
| Melanoma of the skin | 46,870 | 6% | | Thyroid | 49,350 | 6% |
| Non-Hodgkin lymphoma | 40,170 | 5% | | Non-Hodgkin lymphoma | 32,410 | 4% |
| Kidney & renal pelvis | 39,650 | 5% | | Melanoma of the skin | 29,510 | 3% |
| Oral cavity & pharynx | 34,780 | 4% | | Leukemia | 26,050 | 3% |
| Leukemia | 34,090 | 4% | | Pancreas | 25,400 | 3% |
| Liver & intrahepatic bile duct | 28,410 | 3% | | Kidney & renal pelvis | 23,050 | 3% |
| **All Sites** | **841,390** | **100%** | | **All Sites** | **843,820** | **100%** |

**Estimated Deaths**

| | Males | | | Females | |
|---|---|---|---|---|---|
| Lung & bronchus | 85,920 | 27% | | Lung & bronchus | 72,160 | 26% |
| Prostate | 26,120 | 8% | | Breast | 40,450 | 14% |
| Colon & rectum | 26,020 | 8% | | Colon & rectum | 23,170 | 8% |
| Pancreas | 21,450 | 7% | | Pancreas | 20,330 | 7% |
| Liver & intrahepatic bile duct | 18,280 | 6% | | Ovary | 14,240 | 5% |
| Leukemia | 14,130 | 4% | | Uterine corpus | 10,470 | 4% |
| Esophagus | 12,720 | 4% | | Leukemia | 10,270 | 4% |
| Urinary bladder | 11,820 | 4% | | Liver & intrahepatic bile duct | 8,890 | 3% |
| Non-Hodgkin lymphoma | 11,520 | 4% | | Non-Hodgkin lymphoma | 8,630 | 3% |
| Brain & other nervous system | 9,440 | 3% | | Brain & other nervous system | 6,610 | 2% |
| **All Sites** | **314,290** | **100%** | | **All Sites** | **281,400** | **100%** |

**Figure 1. Estimated New Cancer Cases and Deaths by Sex, United States, 2016.** Taken from Siegel et al., 2016.

## 1.2. Carcinogenesis

Breast cancer like other types of tumours evolves from a normal cell that acquires genetic and epigenetic changes which confer it an advantage in terms of survival, proliferation, migration, and adaptation to different environmental conditions (e.g. different tissues, resistance to cytotoxic agents, etc.). These modifications, together with microenvironment variations determine the establishment of cell clones with the ability to grow and proliferate in an uncontrolled way. Importantly, during tumour initiation and progression the acquirement of chromosomal (Navin et al., 2011) and genomic instability has been observed, which frequently result from the inactivation of signalling pathways involved in the maintenance of chromosomes and genome integrity (i.e. loss of p53/p21, BRCA1-2,

etc.). Genomic and chromosomal instability have been shown to confer tumour cells with an increased potential to acquire alterations, such as DNA mutations, amplification, translocations, and aneuploidy, which ultimately contribute to progression to a metastatic and incurable disease. In addition, loss of heterozygosity and changes in gene copy number were also shown to increase the transition from hyperplasia to ductal carcinoma *in situ* (DCIS) which is the first step during tumour progression (O'Connell et al., 1998).

Many of these alterations ultimately affect key genes (oncogenes and tumour suppressors) involved in cell survival, proliferation, invasiveness, motility and drug resistance (Sherr, 2004; Summy and Gallick, 2003). For example, the *ERBB2*, MYC, *CCND1* and *PI3KCA* oncogenes are typically deregulated in breast cancer. The *ERBB2* gene encodes for a receptor tyrosine kinase that is a member of the epidermal growth factor (EGFR) family. When ERBB2 heterodimerizes with other EGFR family members, it mediates the activation of downstream pathways, such as mitogen-activated protein (MAP) kinase and phosphatidylinositol-3 kinase (PI3K)/Akt. *ERBB2* is found to be amplified in ~20% of breast cancer patients (Zhou et al., 1987). The *MYC* oncogene encodes for a nuclear phosphoprotein that acts as a transcription factor; it regulates targets specific for cell cycle progression, apoptosis and cellular transformation, and is overexpressed in ~20% of breast cancers (Nass and Dickson, 1997). The *CCND1* gene encodes for the Cyclin D1 protein, which is involved in G1-S phase transition during the cell cycle; it is amplified in ~15% of breast cancer cases (Sutherland and Musgrove, 2004) and at high level seems to be associated with an oestrogen receptor (ER)-positive and increased Ki67 levels (Loden et al., 2002). The *PI3KCA* gene encodes for the catalytic subunit of the PI3K, it is mutated in ~35% of breast cancers (Li et al., 2006).

Other genes with a tumour suppressor function are commonly altered in breast cancer, e.g.: *TP53*, *BRCA1*, *BRCA2* and *PTEN*. *TP53* gene encodes for the transcription factor p53, which target genes are involved in cell cycle arrest, DNA repair, apoptosis and senescence. It is mutated in ~20% of breast cancers (Pharoah et al., 1999), and, when it is

not mutated, other mechanisms are in place to inactivate p53 function in breast cancer. For example, alterations in upstream regulators of p53 (*ATM* and *CHEK2*), in p53 co-activators (*ASPP1*, *ASPP2* and *BRCA1*) and in p53 target genes (*SIGMA*, *MDM2* and *CDKN1A*) have been detected in breast cancer (Gasco et al., 2002). *BRCA1* "breast cancer 1 early-onset" and *BRCA2* "breast cancer 2 early-onset" are breast cancer susceptibility genes encoding for nuclear phosphoproteins. *BRCA1* and *BRCA2* genes are mutated in 3-8% of patients affected by breast cancer screened by The Cancer Genome Atlas (TCGA) consortium (http://www.cbioportal.org/). At least 50% of women who inherit a *BRCA1* mutation and around 45% of women who inherit a *BRCA2* mutation will develop breast cancer by the age of 70 years (Antoniou et al., 2003; Chen and Parmigiani, 2007). The *PTEN* gene encodes for a tumour suppressor phosphatase and is mutated in <5% of breast cancer patients (Lu et al., 1999). Loss of heterozygosity of *PTEN* locus at chromosome 10q23 occurs in 10-40% of breast cancers (Bose et al., 1998).

Beyond these genetic alterations of cancer-related genes, epigenetic changes including hypo- and hyper-methylation of gene promoter regions, histone tail modifications, and nucleosome remodelling, have been shown to be relevant for breast cancer onset and progression. For example, it has been shown that only a few genes are hypomethylated in breast cancer (e.g., *FEN1*, *NAT1* and *CDH3*) (Singh et al., 2008; Kim et al., 2008; Paredes et al., 2005), while the majority (>1000) of genes were reported to be hypermethylated (Hinshelwood and Clark, 2008). Among them, *BRCA1* was found hypermethylated in sporadic breast cancer (Chan et al., 2002) and *CDKN2A* has been described to have aberrant CpG island methylation (Herman et al., 1995). Although several studies have been performed to map molecular alterations responsible for breast cancer initiation and progression (Ma et al., 2003; Moinfar et al., 2000; Stephens et al., 2012), a complete picture of the pathogenic events is still lacking, which limits therapeutic options, particularly, for those patients with advanced stage disease.

## 1.3. Classification

Breast cancer can be classified into different subtypes based on histopathological, molecular and functional features:

### *1.3.1. Histopathological Classification*

The most frequent breast cancer subtype is breast adenocarcinoma (~95%), with the remaining cases made up of breast sarcomas and mixed-type breast cancers (Schuur and DeAndrade, 2015). Breast adenocarcinoma can be further categorized into: i) ductal carcinoma *in situ* (DCIS), which grows within the milk duct; ii) lobular carcinoma *in situ* (LCIS), which grows outside the duct within the breast; iii) invasive ductal carcinoma (IDC), which grows within the duct and invades the surrounding tissue in a random manner; iv) invasive lobular carcinoma (ILC), which arises within the end part of the lobule and infiltrates the mammary stroma and adipose tissue in a single-file pattern.

Invasive carcinomas, both IDCs and ILCs, are a heterogeneous group of tumours consisting of different histological subtypes. IDC (Figure 2) is the most common subtype accounting for ~80% of all invasive lesions. IDC is further sub-classified as either well differentiated (grade 1), moderately differentiated (grade 2) or poorly differentiated (grade 3) based on the levels of nuclear pleomorphism, glandular/tubule formation and mitotic index (Elston and Ellis, 1991).

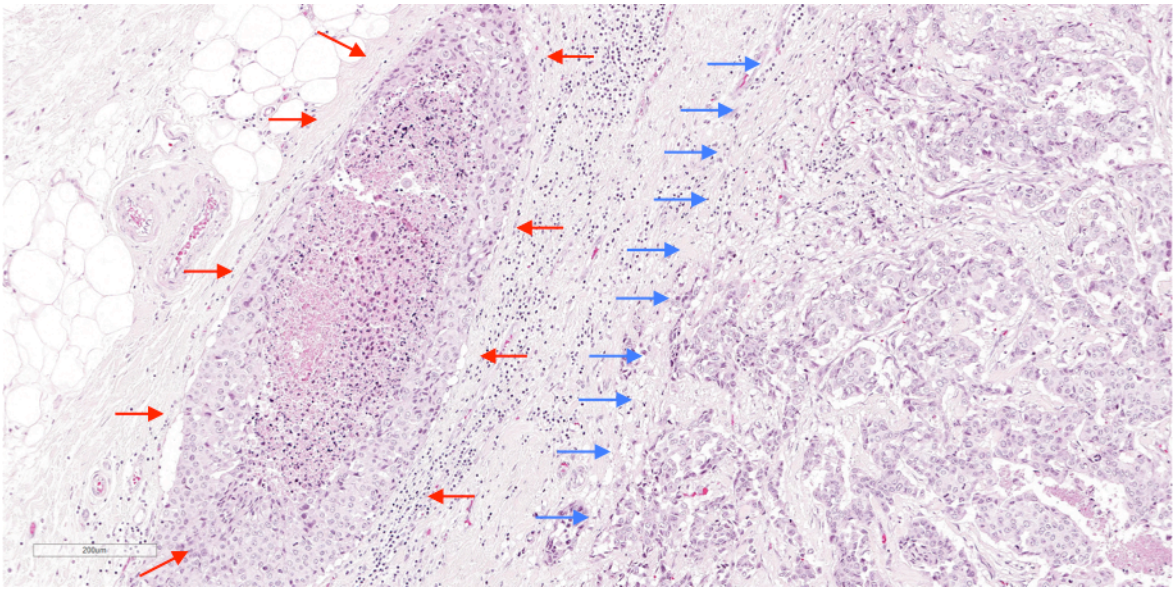

**Figure 2. Infiltrating ductal breast carcinoma**. Haematoxylin and eosin staining of an infiltrating ductal carcinoma in breast tissue (right side of the panel, blue arrows) with an *in situ* component (left side of the panel, red arrows). Taken from IEO archival collection of tumour samples.

Tumour stage is another important clinical and pathological characteristic used to determine treatment options for breast cancer patients. Tumour stage is defined by the TNM system (Figure 3) based on size of the tumour (T), lymph node spread (N) and metastatic dissemination (M). Briefly, the main stages are as follow:

- Stage 0, pre-cancerous condition.
- Stage 1-3, tumour is within the breast and the lymph nodes.
- Stage 4, tumour is already metastatic.

| ANATOMIC STAGE/PROGNOSTIC GROUPS | | | |
|---|---|---|---|
| Stage 0 | Tis | N0 | M0 |
| Stage IA | T1* | N0 | M0 |
| Stage IB | T0 | N1mi | M0 |
| | T1* | N1mi | M0 |
| Stage IIA | T0 | N1** | M0 |
| | T1* | N1** | M0 |
| | T2 | N0 | M0 |
| Stage IIB | T2 | N1 | M0 |
| | T3 | N0 | M0 |
| Stage IIIA | T0 | N2 | M0 |
| | T1* | N2 | M0 |
| | T2 | N2 | M0 |
| | T3 | N1 | M0 |
| | T3 | N2 | M0 |
| Stage IIIB | T4 | N0 | M0 |
| | T4 | N1 | M0 |
| | T4 | N2 | M0 |
| Stage IIIC | Any T | N3 | M0 |
| Stage IV | Any T | Any N | M1 |

**Figure 3. Breast Cancer Staging. TNM staging system.** Adapted from American Joint Committee on Cancer, 7th Edition. Briefly, T refers to primary tumour characteristic: Tis carcinoma *in situ*, T1 Tumour ≤ 20 mm in greatest dimension; T2 Tumour > 20 mm but ≤ 50 mm in greatest dimension; T3 Tumour > 50 mm in greatest dimension; T4 Tumour of any size with direct extension to the chest wall and/or to the skin (ulceration or skin nodules). N refers to lymph node metastatization: N0, No regional lymph node metastases; N1, Metastases to movable axillary lymph node(s); N2, Metastases in axillary lymph nodes that are clinically fixed or matted, or in clinically-detected internal mammary nodes in the absence of clinically evident axillary lymph node metastases; N3 Metastases in infraclavicular lymph node(s), or in clinically-detected internal mammary lymph node(s) with clinically evident axillary lymph node metastases, or metastases in supraclavicular lymph node(s) with or without axillary or internal mammary lymph node involvement. M refers to distant metastasis presence: M0, No clinical or radiographic evidence of distant metastases; M1, Distant detectable metastases larger than 0.2 mm. * Include T1mi, Tumour ≤ 1 mm in greatest dimension; ** T0 and T1 tumours with nodal micrometastases only are excluded from Stage IIA and are classified Stage IB.

### *1.3.2. Molecular Classification*

The identification of peculiar molecular characteristics through microarray analysis led to the classification of breast cancer into distinct molecular subtypes. In particular, the PAM50 study posed the basis for this classification (Parker et al., 2009) and proposed to

classify breast cancer into basal-like (also known as triple negative, TN), HER2, luminal A, and luminal B subtypes. Basal-like/TN breast cancers are negative for three common immunohistological markers: the ER, the progesterone receptor (PR), and HER2. This subtype of breast cancer accounts for 15-20% of all breast tumours and is the most aggressive tumour subtype, associated with a poor prognosis. The HER2 subtype refers to tumours harbouring *ERBB2* amplification but negative for ER and PR. These tumours are usually poorly differentiated (high-grade) and account for 15% of total tumours. Luminal A tumours are ER-positive and/or PR-positive, and HER2-negative with a low proliferation index characterized by a low percentage of Ki67 (a proliferation antigen) positive cells (i.e. with a Ki67<14%). They are low-grade tumours (1 or 2) and are associated with a more favourable prognosis. Finally, luminal B tumours are ER-positive and/or PR-positive, with a Ki67≥14%, or even with a Ki67<14% but HER2-positive. These tumours tend to be poorly differentiated and associated with a poorer prognosis compared with luminal A tumours (Coates et al., 2015).

### *1.3.3. Functional Classification*

Histopathological analyses, and recently also molecular tests, are routinely used to characterize and molecularly categorize breast tumours in order to define the optimal therapeutic strategy to treat patients. However, the identification of curative treatments for more advanced breast cancer is still an unmet clinical need, which will require a deeper understanding of breast tumour biology and the mechanisms of progression and metastatization.

In 2003, Clarke and colleagues first identified tumorigenic breast cancer cells with stem cell characteristics (Al-Hajj et al., 2003). The isolation of tumour-initiating cells with stem cell-like characteristics should allow the identification of specific molecular alterations that can be used to tailor targeted therapies to eradicate the disease. Currently, there are different cell markers that could be used to isolate breast CSCs for further

characterization. In particular, CD44/CD24 (Al-Hajj et al., 2003), CD49f/CD29 (Lim et al., 2009) and ALDH1 (Ginestier et al., 2007) are the markers most commonly used to identify and isolate both normal and cancer breast stem cells. CD44, CD24, CD49 and CD29 are surface proteins, while ALDH1 is an intracellular enzyme. The *CD44* gene encodes for a cell surface glycoprotein involved in cell-cell interactions, cell migration and tumour growth and progression (Alves et al., 2008; Dang et al., 2015; McFarlane et al., 2015). The CD24 surface protein is expressed in B lymphocytes and differentiated neuroblasts, and modulates growth and differentiation signals (Rostoker et al., 2015). The *CD49f* gene encodes for the alpha-6/beta-4 integrin that interacts with the extracellular matrix and stimulates invasion through the EGFR signalling pathway (Carpenter et al., 2015). The CD29 protein is encoded by the *ITGB1* gene and it is an integrin beta subunit. CD29 is involved in cell-cell and cell-extracellular matrix interactions, in signal transduction, and in endothelial-to-mesenchymal transformation (Shi et al., 2015).

### 1.4. Diagnosis

Many efforts have been made to diagnose breast cancer early in order to increase survival of patients and prevent metastatic spread. Typical tools for diagnosis are mammography, echography and magnetic resonance imaging. After the identification of atypical masses through imaging technology, a biopsy is made to understand the nature of the recognized lesion and characterize it. Recent results of more than 20 years of follow-up (range 22-30 years) in the Swedish randomized controlled mammography trials showed that women invited to mammography screenings had a significant 15% relative reduction in breast cancer mortality compared to those that were not invited (Nystrom et al., 2016). Unfortunately, in women aged 40-48, the mammography did not significantly reduce the mortality for breast cancer (Moss et al., 2006). Consequently, young women from age 30, at high-risk of developing breast cancer (i.e. with a family history for breast and ovarian cancer) are screened with magnetic resonance imaging in combination with mammography.

According to the National Institute for Health and Care Excellence, women who have close blood relatives with breast cancer (either first-degree female relative - mother, sister, daughter - or father or brother) have indeed a higher risk of developing the disease, and ~20% of these individuals have inherited mutations in high-penetrant genes, such as *BRCA1*, *BRCA2*, *TP53* and *PTEN*. In this context, testing for mutations in *BRCA1* and *BRCA2* is another important screening strategy to identify women at risk of developing breast cancer that should be enrolled in cancer prevention programs and genetic counselling (King et al., 2003).

### 1.5. Management

Treatment of breast cancer patients varies according to several factors, including tumour stage, ER status, other tumour characteristics and menopausal status.

Surgery alone is the primary treatment for early stage breast cancer and its aim is to completely eradicate the primary tumour to reduce the risk of local recurrences. Pathologic staging of the tumour and sentinel lymph nodes during resection is necessary to obtain prognostic information. There are two main types of surgery for breast cancer: i) breast-conserving surgery, in which only the tumour mass and the surrounding normal tissue are removed; ii) and mastectomy, in which the entire breast is removed to reduce the risk of disease relapse (Schuur and DeAndrade, 2015).

Together with surgery, many patients with more advanced tumours are treated also with radiation and chemotherapy. Radiotherapy is well tolerated by most of the patients and side effects are usually limited to the treated area, although there is a small risk of treatment-induced secondary malignancies (~0.5% in patients affected by breast cancer resident in the North and South Thames regions) (Roychoudhuri et al., 2004). There are two types of radiation therapy: external beam radiation, in which the radiation is focused from a machine outside the body; and brachytherapy (or internal radiation), in which a device containing radioactive seeds is placed for a short time into the breast tissue in the

tumoral area (Polgar and Major, 2009).

Pharmacological therapy comprises of chemotherapy and targeted therapy. Chemotherapy is used to treat cancer with high risk of relapse or metastatic spread to other tissues in the body (adjuvant systemic chemotherapy), to reduce tumour size before surgery, and to treat metastatic breast cancer. Although side effects are drug specific, some are common to all: e.g., nausea, fatigue, hair loss and increased risk of developing infections.

Chemotherapy includes:

- Taxanes, commonly used chemotherapeutic agents for the treatment of early stage breast cancer that inhibit microtubule polymerization, which leads to apoptosis in a subset of the arrested population (Fauzee et al., 2011).
- Anthracyclines, used in the treatment of early stage breast cancer for decades, although concerns regarding anthracycline-associated cardiotoxicity or leukemogenic potential remain (EBCTCG, 2005). These agents target topoisomerase II (TOP2) by intercalating into DNA, binding to TOP2, disrupting its function and inducing the DNA-damage response (Minotti et al., 2004).
- Tamoxifen, used in the treatment of ER-positive breast cancer. This drug decreases the ability of oestrogen to stimulate growth by binding directly to the ER (MacGregor and Jordan, 1998).
- Aromatase inhibitors, used in the treatment of ER-positive breast cancer. These drugs inhibit aromatase, the enzyme responsible for converting adrenal androgen substrate androstenedione into oestrogen (Altundag and Ibrahim, 2006).

Therapies that target specific molecular characteristics of cancer cells are called targeted therapies. Since these drugs have a precise molecular target in cancer cells, which is normally unaltered in normal cells, they are usually less toxic and more effective than standard chemotherapy. Some targeted therapies are based on antibodies that work similarly to endogenous antibodies produced during the immune response. These types of

targeted therapies are also called immune targeted therapies. Among the targeted therapies, we find those based on monoclonal antibodies:

- Bevacizumab that works by inhibiting the vascular endothelial growth factor (VEGF), resulting in the block of angiogenesis that cancer cells depend on to grow and function (Ellis, 2006).
- Trastuzumab (Herceptin) and Pertuzumab that both work against HER2-positive breast cancers by blocking the ability of the cancer cells to receive growth signals via HER2. It binds directly the extracellular domain of the tyrosine kinase receptor HER2 (Valabrega et al., 2007).
- T-DM1 or ado-trastuzumab emtansine is a combination of Trastuzumab and the chemotherapy medicine emtansine. T-DM1 was designed to deliver emtansine to cancer cells in a targeted way by attaching emtansine to Trastuzumab. Trastuzumab then carries emtansine to the HER2-positive cancer cells, inducing mitotic arrest, apoptosis, mitotic catastrophe and disruption of intracellular trafficking (Barok et al., 2014).

Targeted therapies also include small molecule inhibitors, such as:

- Palbociclib, a cyclin-dependent kinase 4/6 (CDK4/6) inhibitor that works by inhibiting cancer cells proliferation (Cadoo et al., 2014).
- Everolimus, an mTOR (mammalian target of rapamycin) inhibitor whose effect is on the mTORC1 protein complex. Everolimus binds to its protein receptor FKBP12, which directly interacts with mTORC1 inhibiting its downstream signalling. Everolimus has an important effect on cell growth, cell proliferation and cell survival. (Houghton, 2010).
- Lapatinib that works against HER2-positive breast cancers inhibiting tumour cell growth. Lapatinib enters the cell and binds directly to the ATP-binding pocket of the HER2 intracellular tyrosine kinase domain (Untch and Luck, 2010).

## 2. Breast Cancer Genetics

In the last years, many efforts have been made towards identifying genes involved in the risk of developing, establishment, growth and dissemination of breast cancer. For example, genetic tests involving the analysis of the mutational status of *BRCA1* and *BRCA2* are routinely performed in the case of breast cancer familiarity for breast cancer surveillance. These genes are defined as high-penetrance breast cancer genes. The *BRCA1* and *BRCA2* genes encode for nuclear phosphoproteins that are involved in maintaining genomic stability and act as tumour suppressors (Savage et al., 2014; Shahid et al., 2014) About 40% of inherited breast cancers carry mutations in *BRCA1* gene, and tumours with *BRCA2* mutations generally exhibit loss of heterozygosity of the wild-type allele. In addition to *BRCA* genes, other DNA repair genes that interact with *BRCA1*/*2* and/or with BRCA pathways can confer an increased risk of developing breast cancer. These genes include: *ATM* (Renwick et al., 2006), *CHEK2* (Meijers-Heijboer et al., 2002), *BRIP1* (*BACH1*) (Seal et al., 2006), and *PALB2* (Rahman et al., 2007). ATM is a cell cycle checkpoint kinase that regulates several downstream proteins such as TP53 and BRCA1, and is required for the DNA damage response (Bhoumik et al., 2005). CHEK2, a cell cycle checkpoint regulator and member of the CDS1 subfamily of serine/threonine protein kinases, is able to stabilize p53 protein, controlling cell cycle arrest, and also interacts with BRCA1 (Zannini et al., 2014). BRIP1, DNA-dependent ATPase, 5'-3' DNA helicase and member of the RecQ DEAH helicases family, is involved in DNA double-strand break repair by interacting with BRCT repeats of BRCA1 (Cantor et al., 2001). Finally, PALB2 binds to BRCA2 promoting its stabilization in the nucleus and its checkpoint functions (Park et al., 2014).

Despite the identification of these and other genes involved in breast cancer, it is clear that the molecular complexity of breast cancers cannot be explained by using only a small subset of mutated genes. Indeed, The TCGA, analysing the molecular profiles of

hundreds of patients affected by breast cancer, identified novel significantly mutated genes (including *BX3*, *RUNX1*, *CBFB*, *AFF2*, *PIK3R1*, *PTPN22*, *PTPRD*, *NF1*, *SF3B1* and *CCND3*), together with known cancer genes implicated in breast cancer (Cancer Genome Atlas, 2012). For this reason, breast cancer should be considered as a molecularly heterogeneous disease and more effort should be made to identify novel molecular determinants to refine breast cancer diagnosis and to identify cancer druggable targets.

The molecular heterogeneity of breast cancer can be classified as intertumoral heterogeneity, which describes differences among tumours from different patients, and intratumoral heterogeneity, which describes differences between cancer cells within the same tumour (Russnes et al., 2011). Several hypotheses have been proposed to explain the origin and evolution of this molecular heterogeneity and the impact of heterogeneity on tumour onset and progression. Here below we briefly explain the different proposed theories and their impact on breast tumour aetiology and dissemination.

## 2.1 Tumour Heterogeneity

### *2.1.1. Intertumoral Heterogeneity*

Given the variety of different subtypes in which breast cancer is classified, several hypotheses have been made to define and explain this heterogeneity. Two principal models have been proposed:

- Subtype-specific cell of origin, in which each tumour subtype originates from a specific type of cell (Figure 4A).
- Single cell of origin, in which the cell of origin is the same for each subtype and the tumour phenotype is determined by acquisition of genetic and epigenetic changes (Figure 4B).

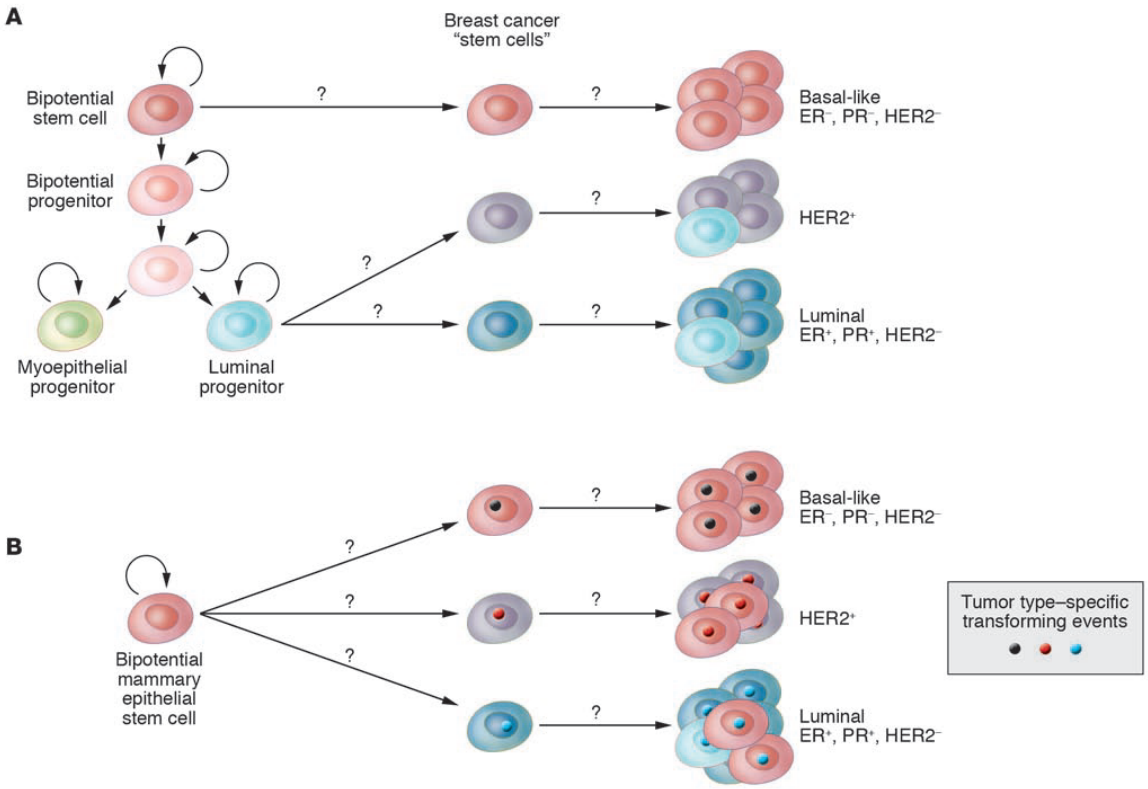


**Figure 4. Inter-tumour Heterogeneity models.** A) Subtype-specific cell of origin model: in dark red progenitor cell of Basal-like, triple negative tumours; in light blue progenitor of both HER2 tumours (grey) and Luminal tumours (dark and light blue); B) Single cell of origin (light red) model: black, red and blue dots represent tumour type-specific transforming events, that originate the different tumour subtypes. Taken from Polyak, 2007.

### *2.1.2. Intratumoral Heterogeneity*

Intertumoral genetic heterogeneity has been recognized for decades (Fidler, 1978; Perou et al., 2000; Sorlie et al., 2001). In contrast, intratumoral heterogeneity was only recently discovered when genome-wide microarray gene expression profiling, chromosome copy number analysis, and Next-Generation Sequencing (NGS) were employed (Gerlinger et al., 2012; Navin et al., 2011). These technologies enabled the molecular characterization of distinct sub-populations of cancer cells within the bulk tumour population, which contributed to the identification of additional cancer genes involved in metastatic spread (Nguyen et al., 2016). Moreover, recent studies have shown that different regions of the same tumour may have distinct genetic alterations, which indicates the existence of different sub-populations confined into geographically distinct sectors of the tumour (Gerlinger et al., 2012). With the advent of NGS (Kim et al., 2008) many studies were

designed toward a better and deeper understanding of the spectrum of mutations in the bulk tumour population and in specific sub-populations of tumour cells (Ding et al., 2010; Navin et al., 2011). Furthermore, single cells can be now captured from the entire tumour and mutations and copy number variations can be detected in order to separate and reconstruct the diversity of sub-populations (Navin et al., 2011).

To explain intratumoral heterogeneity different models have been proposed:

- Clonal Evolution Model, in which clonal tumour progression gives rise to the genetic heterogeneity, starting either from a single monoclonal sub-population or from multiple polyclonal sub-populations (Figure 5A).
- Cancer Stem Cell Model, in which at a certain time a tumour cell acquires stem cell properties and originates the whole population of the tumour. In this model, the final population of the tumour can derive either from a single progenitor or multiple progenitors (Figure 5B).
- Mutator Phenotype Model, in which each cell within the tumour has in principle a distinct genetic and epigenetic profile (Figure 5C).

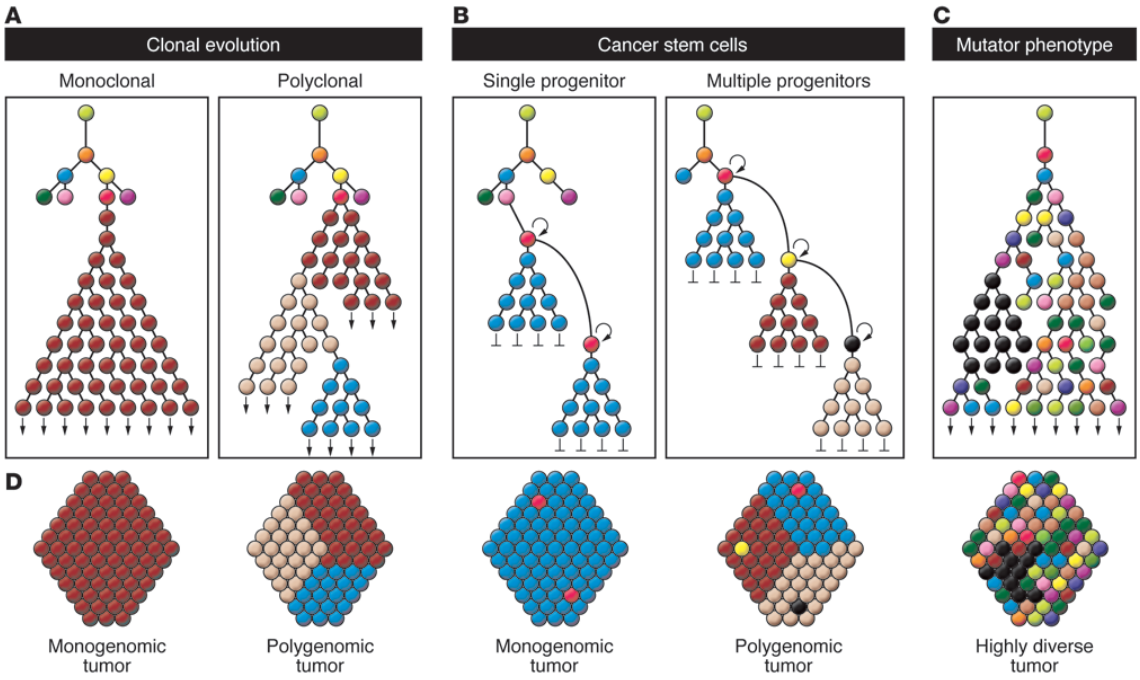


**Figure 5. Model of intratumoral heterogeneity**. A) Clonal Evolution Model, in which either one (monoclonal) or more (polyclonal) tumour clones gives rise to the final bulk tumour population; B) Cancer Stem Cell Model, in which single (red dots) or multiple (red, yellow and black dots) progenitors originate the final tumour; C) Mutator Phenotype, in which every cell has a distinct genetic and epigenetic profiles; D) Phenotype of tumour cells population according to each model. Taken from Russnes et al., 2011.

### 2.2. Cancer Stem Cells and Breast Cancer Stem Cells

According to the hematopoietic stem cell (HSC) paradigm, stem cells are slowly dividing cells able either to divide asymmetrically, which give rise to one new stem cell and one progeny, or to divide symmetrically originating two identical daughter stem cells. HSC have also the capability to originate progeny that enter the irreversible process of differentiation. Despite this model of stem cell has been applied to all tissues, it is not always true and demonstrated for each tissue. For example, in oesophagus, skin, intestinal crypts, stomach gland and testis, stem cells are abundant and divide symmetrically with a not pre-determined life span (Clevers, 2015). In addition, it was demonstrated that differentiated cells in airway epithelium and in stomach gland tissues are able to revert their committed fate to multipontent stem cell fate upon loss of stem cell pools. In contrast, in mammary tissue two main models have been proposed to describe the types of stem cells: i) the model in which rare multipotent cells with the capability to originate both basal

and luminal progenitors are present in breast tissue (Shackleton et al., 2009; Rios et al., 2014); ii) the unipotent cells model, which encompass the presence of specific stem cells for each breast lineage (Van Keymeulen et al., 2011).

In the 90s, it was postulated that CSCs originate from normal stem cells (Bonnet and Dick, 1997), but recent evidence suggests that in many cancers the cancer stem cell compartment arises from progenitor cells that acquire the ability to self-renew (Clarke and Fuller, 2006; Reya et al., 2001). The first theory is supported by evidence that CSCs share features of normal stem cells, such as self-renewal and differentiation capabilities. In addition, normal stem cells could acquire mutations and oncogenic transformation during their long lifespan (Ginestier et al., 2007; Ponti et al., 2005) transforming them into CSCs (Figure 6A).

On the other hand, in the second scenario, progenitors that undergo epithelial-to-mesenchymal transition (EMT) were shown to be prone to transformation and to acquire characteristics and behaviours of neoplastic stem cells (Mani et al., 2008; Morel et al., 2008) (Figure 6B).

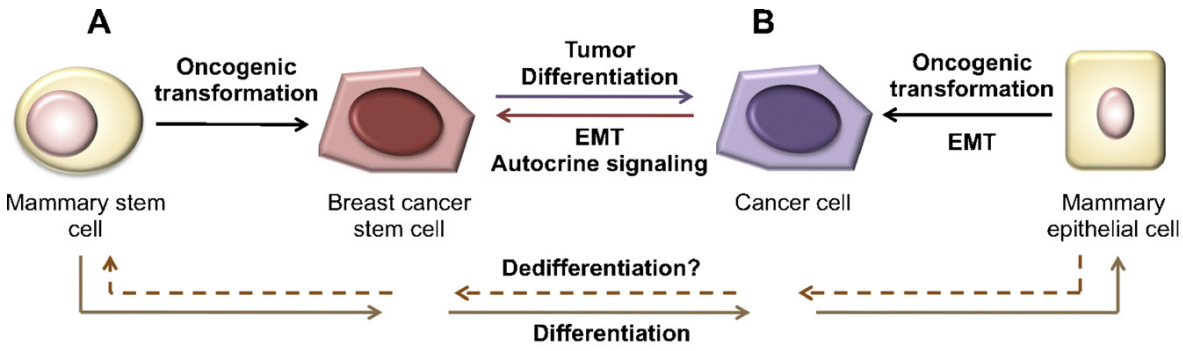


**Figure 6. Origin of Breast CSCs.** A) Normal mammary stem cell that gives rise to a breast cancer stem cell through oncogenic transformation, which then differentiates into the final tumour cell; B) Non-stem cell that, via epithelial-to-mesenchymal transition (EMT) and oncogenic transformation, originates a cancer cell that acquires stem cell abilities. Solid lines represent experimentally demonstrated events; dashed lines represent not yet confirmed events. Taken from Velasco-Velazquez et al., 2012.

EMT is the process in which epithelial cells lose the tight junctions typical of the epithelial phenotype and acquire mesenchymal properties, including fibroblastoid

morphology and increased motility, allowing them to migrate (Kalluri and Weinberg, 2009). EMT was first observed during embryonic development and is essential for normal wound healing.

The ability of mesenchymal cells to spread and self-renew may link EMT to CSCs. Indeed, recent evidence suggests that the EMT process may promote the generation of cancer cells with the mesenchymal characteristics needed for dissemination (Hollier et al., 2009). Therefore, the induction of EMT in immortalized human mammary epithelial cells results in an increased capacity to form mammospheres and generate cells with a stem cell signature (Mani et al., 2008). In addition, EMT was shown to have a role in drug resistance and disease recurrence in patients affected by breast cancer (Oliveras-Ferraros et al., 2012; Creighton et al., 2009). This involvement is supported by the fact that some circulating tumour cells (CTCs) from breast cancer patients possess both epithelial and mesenchymal features and that mesenchymal CTCs are enriched during tumour progression (Yu et al., 2013).

### *2.2.1. Breast Cancer Stem Cells Markers*

Different markers have been identified to isolate breast CSCs. The vast majority are composed of surface markers, such as the CD44 and CD24 proteins. In breast cancer, Al-Hajj *et al.* (Al-Hajj et al., 2003) demonstrated that a subpopulation of the initial tumour, composed of $ESA^{+}/CD44^{+}/CD24^{-/low}$, were the only cells able to generate a tumour when transplanted in NOD/SCID mice. Subsequently, several studies have confirmed their results (Liang et al., 2013; Pece et al., 2010; Ponti et al., 2005). In addition, the intracellular enzyme ALDH1 (aldehyde dehydrogenase 1) was also used to enrich breast CSCs. Indeed, $ALDH1^{+}$ cells were able to give rise to tumours in NOD/SCID mice (Ginestier et al., 2007).

Notably, both populations of cells, $ESA^{+}/CD44^{+}/CD24^{-/low}$ and $ALDH1^{+}$, have the ability to form mammospheres, spherical clonal colonies of cells grown in non-adherent conditions and enriched in CSCs (Dontu et al., 2003).

### *2.2.2. Clinical Implications of Cancer Stem Cells*

Since metastatic breast cancer is one of the major causes of death from breast tumours, it is important to understand which mechanisms are responsible for metastasis. To determine the role of CSCs in the metastatic process, Balic and colleagues (Balic et al., 2006) examined the expression of cancer stem cell markers in metastatic lesions of bone marrow in patients with breast carcinoma. They found an high $CD44^{+}/CD24^{-/low}$ expressing cells in disseminating tumour cells of bone marrow of patients affected by breast cancer indicating an enrichment of the cancer stem cell compartment.

While the presence of micrometastasis is associated with poor prognosis (Braun et al., 2005), ~50% of patients with such metastasis do not develop macrometastasis within a 10-year follow up period.

Since CSCs have both the characteristics to remain quiescent for a long time and to give rise to an offspring of proliferating cancer cells, in the case of a suboptimal surgery or chemotherapy (which does not eliminate all CSCs), the risk of the disease evolving into a metastatic cancer is considerable. Therefore, it is important to understand the molecular features of CSCs in order to develop specific targeted therapies to eradicate CSCs themselves, and inhibit metastasis onset and disease progression.

## 3. Cancer Genomics

During its lifetime, a cell acquires set of differences from its parent cell and this happens also to cancer cells. Such differences encompass several DNA sequence changes:

- Somatic mutations, single nucleotide variations that affect somatic cells.
- Insertions or deletions of small or large portions of DNA.
- Rearrangements, switching of DNA fragments from one position in the genome to another.
- Copy number variations that comprise loss of copy number (from 2 to 1 or zero) and gain of copy number (more than 2 copies).

In addition, a cancer cell may acquire further epigenetic alterations (chromatin structure, methylation and gene-expression alterations) that can be inherited by daughter cells and augment malignancy (GrØnbÆk et al., 2007; Loden et al., 2002).

Somatic mutations can be classified, according to their consequences, as: i) "driver" mutations that confer a growth advantage and are positively selected during the evolution of the cancer; ii) "passenger" mutations that are not selected and which do not confer any growth advantage, thus, not directly involved in cancer development. However, these mutations might have a role in promoting the establishment of "driver" mutations (Pon and Marra, 2015). Driver mutations are present within a subset of genes known as cancer genes. Most of these cancer genes were discovered and studied during the last decades. In 2004, at least 350 (1.6%) of the 22,000 protein-coding genes in the human genome were shown to harbour recurrent somatic mutations in cancer with strong evidence that these contribute to cancer development (Cancer Gene Census) (Futreal et al., 2004). Nowadays, the Cancer Gene Census comprises around 570 cancer genes with somatic mutations causally implied in cancer.

The identification of cancer genes led to the development of genetic tests, which aim to identify whether a particular gene or set of genes is mutated in a tumour sample and

to guide treatment decision-making. In addition, specific therapies have been developed, based on mutated genes, in order to target mutated cancer cells specifically, without affecting normal cells.

Until recently, most genetic tests for cancer focused on testing for either individually inherited mutations or exons of disease-specific cancer genes. However, as more efficient and cheaper DNA sequencing technologies have become available, sequencing of an individual's entire genome or the DNA of an individual's tumour is becoming more common. In this direction, consortia of clinical laboratories have been established.

Moreover, as reported by Dr. Gad Getz, Director of Cancer Genome Computational Analysis at the Broad Institute of the Massachusetts Institute of Technology (MIT), most genes are mutated in a small percentage of patients (at intermediate frequencies, 2 to 20% of patients, or lower) (http://cancergenome.nih.gov/). While frequently mutated genes seem to be easier to be identified and studied, infrequently mutated genes have long eluded researchers. Therefore, several strategies have been recently proposed to prioritize cancer-relevant genes that are mutated at low frequencies. Most of these strategies are based on pathway-oriented analyses of mutated genes and their mutual exclusivity that could indicate their function within a common molecular mechanism (Dees et al., 2012; Leiserson et al., 2015; Vaske et al., 2010).

### 3.1. Genomics in Medicine (Consortia)

Many efforts are being made to identify a more exhaustive map of cancer genes. Such analyses are made difficult by the high genetic heterogeneity of cancer, which complicates the identification of cancer driver genes that frequently appear to be mutated only in a low percentage of patients (i.e. similarly to passenger gene mutations). For this reason, international collaborative studies have been set up to try to overcome such limitations through the launch of high-throughput genetic screening (based on NGS) of

hundreds of tumour samples, from various cohort of patients with different types of cancer. Here below is a brief summary of the most important studies:

### *3.1.1. The Cancer Genome Atlas (TCGA)*

TCGA is a collaboration between the National Cancer Institute and National Human Genome Research Institute (http://cancergenome.nih.gov/). The aim of this consortium is to generate comprehensive, multi-dimensional maps of the key genomic changes in major types and subtypes of cancer. It comprises 33 cancer types regarding most commonly affected tissues, including breast, lung, blood cells, brain, kidney, skin and uterus, with a total of 11,353 analysed cases (https://gdc-portal.nci.nih.gov/). Patients are asked to donate a portion of tumour tissue together with normal tissue, typically blood. These tumour and normal samples are subjected to different kinds of analysis:

- DNA analyses, including whole-exome sequencing, methylation via bisulfite sequencing and array-based and copy number identification.
- RNA analyses, including total RNA sequencing, gene expression analyses through sequencing and array-based, and miRNA sequencing.
- Microsatellite instability (Velasco-Velazquez et al., 2012).
- Protein expression.

### *3.1.2. International Cancer Genome Consortium (ICGC)*

The ICGC (http://icgc.org/) was created to launch and coordinate a large number of research projects with the common aim of elucidating comprehensively the genomic changes present in many forms of cancers. The primary goal of the ICGC is to generate complete catalogues of genomic abnormalities (somatic mutations, abnormal expression of genes, epigenetic modifications) in tumours from 50 different cancer types and/or subtypes that are of clinical importance. Differently from the TCGA, the ICGC gathers together

projects coming from countries worldwide. Almost every continent is represented by at least one research group within the ICGC. It is also made of two different committees, composed of members grouped by specific aims.

Such collaborative studies allowed us to create high-resolution molecular portraits of different cancer types (Alexandrov et al., 2013; Cancer Genome Atlas, 2012; Cancer Genome Atlas Research, 2014; Cancer Genome Atlas Research et al., 2016) and to identify novel cancer genes and cancer pathways with obvious positive implications in the management of cancer patients.

# 4. Next-Generation Sequencing

## 4.1. Brief History of Next-Generation Sequencing

In 2005, the first NGS technology was released, Roche 454, and posed the basis of a new sequencing era. In contrast to the canonical Sanger sequencing method NGS has three major improvements:

- It is based on library preparations instead of bacterial cloning.
- Thousands to millions of sequencing reactions are performed in parallel.
- The output is directly detected without the electrophoresis step.

In 2006, Solexa/Illumina commercialized its own sequencing technology (http://www.illumina.com/), which was followed by Applied Biosystems's SOLiD platform (Life Technologies) in 2007 (http://www.thermofisher.com/it/en/home/brands/applied-biosystems.html) and by Ion Torrent's Personal Genome Machine (Life Technologies) in 2010 (https://www.thermofisher.com/it/en/home/brands/ion-torrent.html).

The four technologies listed above have different sequencing chemistries. Roche 454 is based on released pyrophosphate that emits light (pyrosequencing). Illumina is based on sequencing-by-synthesis chemistry, in which each DNA fragment is copied through bridge amplification and then sequenced using fluorescent-tagged nucleotides. The SOLiD platform is based on sequencing by ligation, during which octamers are ligated together and when the ligation takes place a fluorophore is released and the light detected. The Personal Genome Machine sequencer is similar to Roche 454, but instead of pyrophosphate being released, it releases a proton during nucleotide incorporation, which is detected by ion sensors.

### 4.2. Next-Generation Sequencing Platforms

Nowadays, there are three major platforms for sequencing projects that are present in research institutes: Applied Biosystems SOLiD, Illumina HiSeq2000 and Ion Torrent Ion Proton.

#### *4.2.1 Illumina Sequencing*

Illumina sequencing technology is based on sequencing-by-synthesis chemistry. Each DNA strand is copied and when a single nucleotide is incorporated in the new filament, it emits a fluorescent signal that is detected.

Peculiar to Illumina sequencing is the amplification of each DNA filament through the so-called bridge amplification. In this step, each DNA fragment that is attached to the solid surface of the flowcell (Figure 7A), falls over and the adaptor at the end of the fragment then attaches to a probe on the flowcell itself (Figure 7B). The bridge-like structures form the templates for amplification to generate clusters comprising clonally amplified copies of the DNA fragment on the surface of the flowcell (Figure 7C).

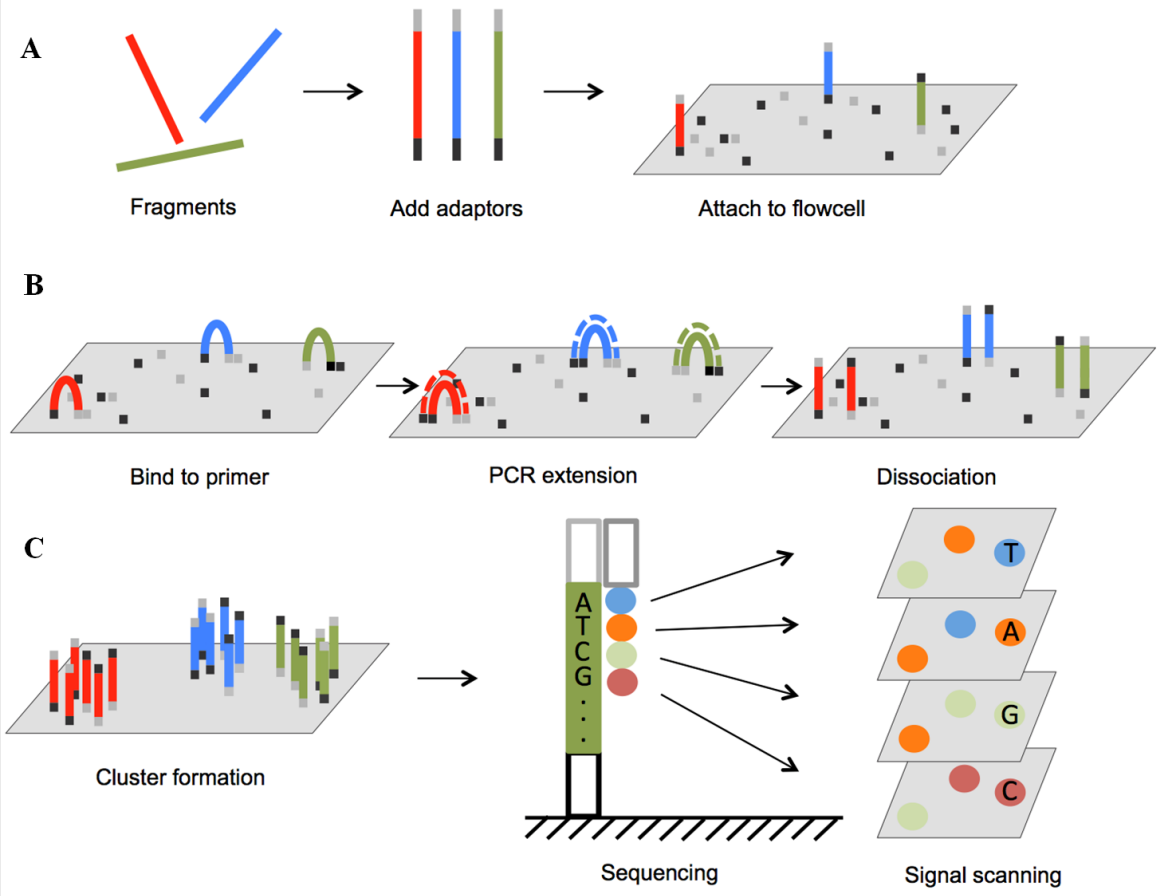


**Figure 7. Illumina Sequencing Workflow**. A) DNA fragments (blue, green and red bars) are ligated with adapters (short black and light grey bars) and then attached to the flowcell surface (grey parallelepiped); B) Each fragment is bridge-amplified, until the generation of clusters; C) After cluster formation reaches a plateau the sequencing-by-synthesis starts (complementary bases are incorporated in the new DNA filament, emitting a fluorescence, which is detected by a light-detector). Adapted from http://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-in-aquatic-models.

Illumina sequencing-by-synthesis is mediated by polymerase enzymes that use four different reversible nucleotides labelled by four different fluorescent colours. The nucleotides are reversible terminators: one terminator nucleotide is incorporated into the synthesis of the complementary strand and then washing steps are applied to remove the extra nucleotides and reagents. The imaging of the fluorescence signals is followed across the whole flowcell, and, after imaging, the 3′ blocking group of the reversible terminator nucleotide is cleaved. These steps are then repeated until the synthesis of the complementary strand is complete.

### *4.2.2. Ion Torrent Sequencing*

The Ion Torrent sequencing technology is based on the detection of $H^+$ ions, that are naturally released when a nucleotide is incorporated into a DNA strand by DNA polymerase.

DNA is fragmented and each fragment is attached to its own bead and then it is copied all over the bead. Since every fragment is attached to one bead, at the end millions of beads are obtained. Beads are then placed into the wells that compose the chip. At each cycle of polymerization, a solution with only one nucleotide type is added to the wells. If the nucleotide is incorporated into the new DNA strand, $H^+$ is released and the pH changes. The variation in pH is converted into a voltage and a base is called (Figure 8). Subsequently, the solution is washed and another nucleotide is added to the wells. This addition, detection and washing workflow is repeated until the completion of synthesis of the complementary DNA strand, in every single well of the chip.

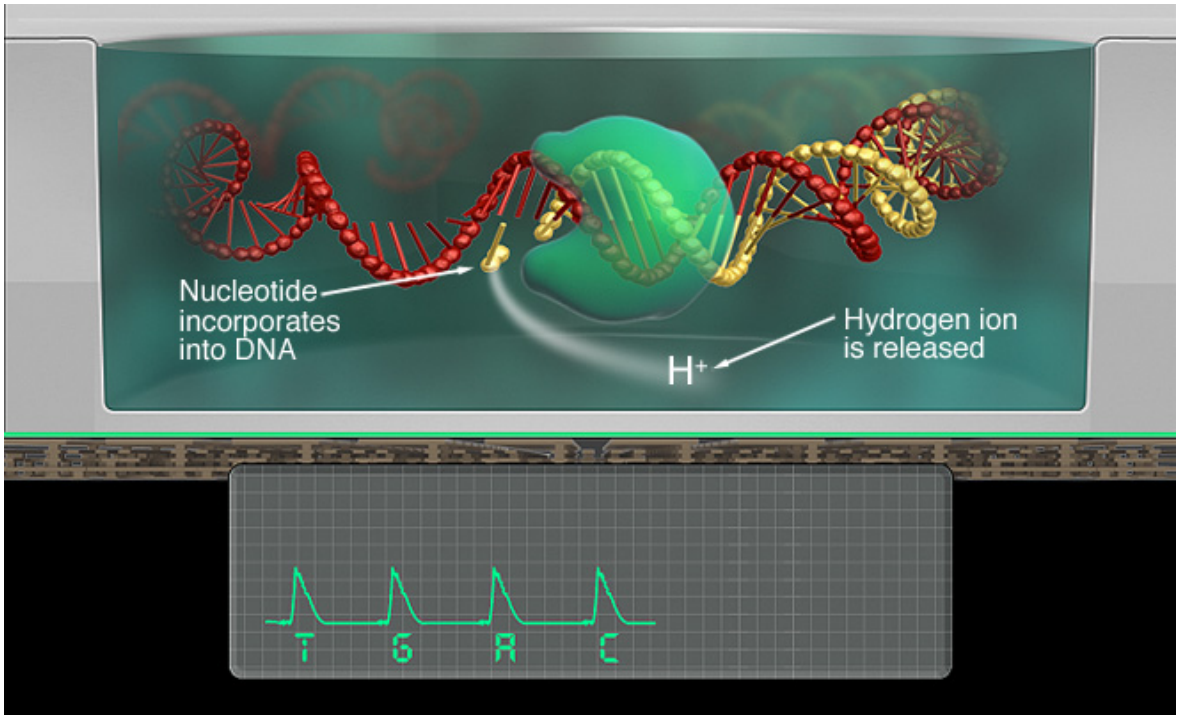


**Figure 8. Principles of the Ion Torrent Sequencing Reaction.** The Ion Torrent Sequencing reaction is based on a sequencing-by-synthesis reaction in which the DNA template being sequenced is immobilized and a complementary DNA strand is synthesized by the sequential addition of single nucleotides. In the synthesis reaction, the incorporation of a nucleotide to a complementary DNA strand is converted into a voltage increase by measuring the release of a hydrogen ($H^+$) ion. Top panel, cartoon depicting the synthesis of a complementary DNA strand during an Ion Torrent sequencing reaction. In red: template strand, yellow: complementary strand; green: DNA polymerase. The release of $H^+$ ion during the incorporation of a nucleotide to the complementary strand is also shown. Bottom panel, depicted the detection of voltage increase induced after the pH change resulting from the release of an $H^+$ ion is shown. Adapted from www.thermofisher.com.

If a nucleotide is not incorporated into the complementary DNA strand, then no $H^+$ ion is released, no voltage change is recorded and no base is called. If two or more bases of the same type are incorporated, two or more $H^+$ ions are released and the voltage changes proportionally (Figure 9).

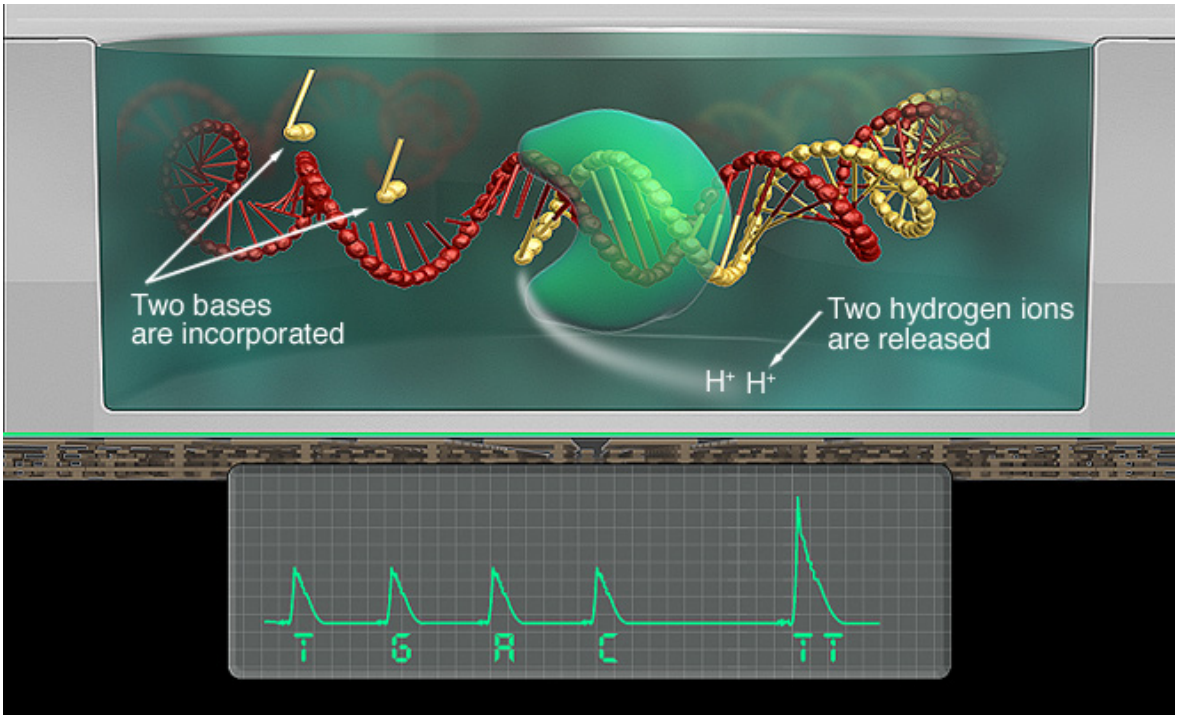


**Figure 9. Ion Torrent Sequencing Reaction in Critical Regions.** In top panel cartoon depicting the synthesis of a complementary DNA strand during an Ion Torrent sequencing reaction when two bases of the same type are incorporated. In red: template strand, yellow: complementary strand; green: DNA polymerase. The release of $H^+$ ions during the incorporation of two nucleotides to the complementary strand is also shown. Bottom panel, depicted the detection of voltage increase induced after the pH change resulting from the release of two $H^+$ ions is shown. Adapted from www.thermofisher.com.

### 4.3. Advantages and Disadvantages

NGS technology has many advantages. First, NGS permits the sequencing of the entire genome of a cell in a shorter timeframe compared with Sanger sequencing, thus enabling the use of NGS platforms in clinics and even in small laboratories. Second, for some applications, it does not require *a priori* knowledge of the genome, since data are *de novo* assembled. Third, it generates billions of data in a very short time, increasing our discovery potential.

Another advantage of NGS is the high versatility of each platform: with only one instrument, it is possible to generate data for different applications, spanning from the basic Whole Genome (WGS) and Whole Exome sequencing (WES), to RNA sequencing (both mRNA and miRNA), to epigenomic applications (e.g. methylation and DNA-protein interactions). NGS also requires less starting material, both for DNA and RNA sequencing,

and permits the sequencing of multiple samples in parallel, with the use of specific barcodes ligated to DNA/RNA from individual samples.

Although NGS has improved our ability in terms of DNA and RNA sequencing, some concerns have been raised. Firstly, sequencing chemistry does not permit the sequencing of very long fragments of DNA/RNA. Second, given the loss in precision of polymerase used for the sequencing itself, the precision and fidelity of the last bases of the reads is lower compared with the other bases, thus, reducing, although not significantly, the throughput. Another problem is the cost of the instrumentation, that is lower compared to the first sequencers, but still remains high (80,000-600,000 euros) and not always affordable. Last but not least, the problem of data storage, since the amount of data produced in a single run of sequencing is huge and typically a laboratory performs several runs. In this scenario, every research centre has invested either in server storage with high costs or in cloud storage which is less costly but more vulnerable.

### 4.4. Next-Generation Sequencing in Cancer Research

The rapid developments and improvements of high-throughput sequencing technologies have not only revolutionized our approach to -omics studies, but they have also changed the genomic medicine context. Besides classical and canonical one/few-gene tests, NGS platforms have appeared in clinical laboratories. This has determined an increase in their translational use.

An important aspect of the use of NGS is the unbiased view of the alterations in the genome. Hence, in cases where there is no prior knowledge of genetic aberrations underlying the cancer phenotype, the approach to sequence either the entire genome or exomes results in a deeper and more accurate identification of modifications that drove the onset and progression of cancer.

### *4.4.1. Targeted Resequencing*

As genomic instabilities that lead to cancerous growth can be due to point mutations in coding regions, exome sequencing has become a natural choice for detecting them. Because mutations in coding regions generally lead to changes in protein structure, exome sequencing is also relevant for developing targeted drug therapies and assessing resistances that may occur during treatment. Targeted resequencing has now entered in the clinic and offers a clear advantage by using small quantities of nucleic acids to screen multiple genes at a time. Although exome sequencing is not yet the main approach in the clinic, a variety of panels of genes are becoming more frequently applied to identify mutations in known cancer-related genes. Indeed, several sequencing companies are producing panels of genes, specific for each cancer type, together with customizable panels, in order to help clinical laboratories not only in their diagnostic routine, but also in clinical research.

## Rationale

Breast cancer is a heterogeneous disease characterized by high inter- and intra-tumoral molecular heterogeneity. This makes diagnosis, the prediction of prognosis and the standardization of treatments more difficult, and ultimately has a negative impact on breast cancer therapy. Thus, deciphering this molecular heterogeneity should lead to improvements in the care of breast cancer patients.

Recent genome-wide profiling studies have analysed in-depth the genomic profiles of several breast tumours using NGS. These studies indicate that breast tumours can be either: i) monogenomic, comprised of a single clonal cell population; ii) or polygenomic, composed of several related clonal subpopulations (Navin et al., 2011). In the first case, the metastatic spread of the disease appears to be driven by cells resulting from a single clonal expansion and which have undergone little further evolution, while, in the second case, metastasis is driven by a minor subpopulation of cells that has acquired 'metastatic' cancer-driving mutations (Ding et al., 2010; Navin et al., 2011). In addition to this temporal dimension, in renal carcinoma, Gerlinger and colleagues (Gerlinger et al., 2012) found that intratumoral heterogeneity can also have a spatial dimension. Indeed, NGS of different regions within the same primary tumour revealed not only shared but also private mutations.

Different models have been proposed to explain the origin and evolution of breast intratumoral heterogeneity: i) the Clonal Evolution Model; ii) the Cancer Stem Cell Model; and iii) the Mutator Phenotype Model (Russnes et al., 2011).

In the Clonal Evolution Model, genetic and epigenetic changes occur over time in individual cancer cells. If such changes confer a selective advantage they will allow individual cancer cell clones to out-compete other clones and give rise to the final tumour mass. Similarly to the Clonal Evolution Model, in the Cancer Stem Cell Model, the cells that undergo advantageous genetic and epigenetic changes may also acquire the ability of

self-renewal, thus transforming these cells into CSCs. The last model is the most dramatic one: the Mutator Phenotype Model, which describes a situation whereby each single cell has a distinct genetic and epigenetic profile. In this scenario, it is extremely hard to identify cancer-driving mutations that sustain tumour growth and progression.

Here, we hypothesized that mutational events that drive the onset and progression of breast tumours lie within the cancer stem cell compartment. Our hypothesis is supported by several experimental observations that demonstrate the ability of CSCs to recapitulate fully malignant breast tumours (Fillmore and Kuperwasser, 2008; Ponti et al., 2005). In particular, tumours with low cancer stem cell content are less able to give rise to PDXs in limiting dilution conditions, and such tumours are frequently low-grade (G1) tumours associated with a less aggressive behaviour compared with tumours with a higher stem cell content (Pece et al., 2010). In addition, it has been demonstrated that the molecular profile of normal breast stem cells can distinguish molecularly breast cancer patients with an adverse prognosis (Ben-Porath et al., 2008; Pece et al., 2010).

To test this hypothesis, we used NGS, specifically, WES, to perform a comparison of the genetic profiles of a primary breast tumour with its matched mammospheres (enriched in CSCs) and PDX. Our goal was to identify mutations enriched in CSCs that might be relevant for breast cancer progression. We setup a NGS protocol to identify rare mutations in the primary tumour that are present in the cancer stem cell compartment, using low amounts of input DNA. Of note, the low frequency mutations in putative novel cancer oncogenes can be diluted within the NGS technical noise and therefore missed. Hence, we proposed that our approach, based on the direct analysis of mammospheres and PDXs matched with primary breast tumours, could be an effective tool for the identification of cancer-relevant, CSC-specific low frequency mutations, which should be enriched and therefore detectable at a higher frequency in the short proliferative history of mammospheres and PDXs. This will enable us, in the long run, to identify new potential

druggable targets for the development of targeted therapies, which is particularly relevant to cases where canonical and known cancer genes are unaltered.

# Material and Methods

## 1. Sample Collection

Primary tumour, embedded in Optimal Cutting Temperature compound, and blood were collected from the IEO Biobank from a patient, aged 55, with infiltrating duct breast carcinoma. The tumour is a Luminal B (HER2-negative, ER-positive, PR-positive assessed by immunohistochemistry) and moderately differentiated (grade II) breast tumour as determined by the Bloom-Richardson score.

## 2. Patient-Derived Xenograft (PDX) and Mammospheres Preparation

PDX and mammospheres were established starting from the same primary tumour. Mammospheres were generated also from PDX sample (PDX-mammospheres). Mammospheres, both human- and mouse-derived, were obtained and cultured as previously described in (Pece et al., 2010). Briefly, epithelial cells from the primary tumour were allowed to adhere for 24 hours in complete SC medium (Dontu et al., 2003). Cells were trypsinized, filtered through a 100 mm cell strainer and then through a 40 mm cell strainer, resuspendend in PBS and plated in suspension (Dontu et al., 2003). After 7–10 days, mammospheres were harvested and dissociated enzymatically.

For the xenotransplant of human breast cancer cells, we proceeded as in (Pece et al., 2010). Briefly, tumour biopsy material was dissociated mechanically and enzymatically and resuspended cells were then injected directly into cleared mammary fat pads of a NOD/SCID mouse. The mouse injected with cancer cells was euthanized when the tumour outgrowth reached approximately 1 cm in the largest diameter, to avoid tumour necrosis and in compliance with regulations for use of vertebrate animal in research. Experiments were performed by Dr Daniela Tosoni in Molecular Medicine for Care Program laboratory at European Institute of Oncology, Milan.

## 3. Sample Processing

Genomic DNA (gDNA) for NGS experiments was extracted from the primary tumour sample, the PDX and mammospheres generated from both the primary tumour and the PDX using the QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's instructions. gDNA for the Whole Genome Amplification (WGA) test was extracted from additional unrelated normal breast tissue, breast cancer primary tumour and primary tumour-derived mammosphere samples with QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's instructions. DNA was extracted from blood using the QIAamp Blood Midi Kit (Qiagen) according to the manufacturer's instructions.

Since we were dealing with mammosphere samples that are composed of very few cells, we tested the smallest amount of gDNA we could use for the genomic amplification step. To do so, 5-10 ng of gDNA of the additional unrelated normal breast tissue, breast cancer primary tumour and primary tumour-derived mammosphere samples were subjected to WGA (Cadwell and Joyce, 1992; Hasmats et al., 2014) with Repli-G Mini Kit (Qiagen).

Having established that 10 ng was the optimum amount of gDNA to obtain a sufficient quantity of DNA for a WES experiment, we subjected 10 ng of gDNA from the primary tumour, blood, primary tumour-derived mammospheres, PDX and PDX-derived mammospheres samples to WGA, using Repli-G Mini Kit (Qiagen). Experiments were performed by Dr Francesca De Santis and Dr Stefania Pirroni in Molecular Medicine for Care Program laboratory at European Institute of Oncology, Milan.

## 4. Whole Exome Sequencing

To test whether WGA introduced artefacts in the NGS analysis, we performed Illumina WES on both gDNA and WGA-DNA obtained from the same tumoral sample test. Three μg of both gDNA and WGA-DNA were used to test for sequencing reproducibility and thereby establish whether there was any possible Whole-Genome Amplification bias.

Based on results from the comparison of gDNA and WGA-DNA, we decided to use 3 µg of WGA-DNA obtained from the primary tumour, primary tumour-derived mammospheres, PDX and PDX-derived mammospheres for Illumina WES.

Libraries were constructed using ligation of Illumina adaptors to sheared WGA- and gDNA. Hybridizations were performed using Agilent SureSelect All Exome v4.0 kit. Paired-end, 100 bp WES (Balic et al., 2006) was performed using HiSeq2000 (Illumina).

For the Ion Torrent sequencing, 100 ng of the same WGA-DNA samples subjected to Illumina sequencing was used as starting material in the AmpliSeq Exome amplification step following manufacturer's protocol (Thermo Fisher Scientific). The final sequencing libraries were quantified using the Bioanalyzer 2100 instrument with DNA HS kit (Agilent Technologies) and Qubit with dsDNA HS kit (Thermo Fisher Scientific). All libraries were diluted to 100pM and then pooled to perform the chip preparation on the Ion Chef instrument (Thermo Fisher Scientific) according to manufacturer's protocols. For our samples, we used a total of 6 chips, in order to obtain a sufficient amount of reads to call variants. Experiments were performed by Cogentech sequencing facility at IFOM-IEO Campus, Milan.

## 5. Bioinformatic Pipeline

### 5.1. Alignment and BAM file generation

#### *5.1.1 Illumina*

The quality of reads was assessed using FastQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Mouse reads from mouse-derived samples (PDX and PDX-derived mammospheres) were identified and removed using Xenome software (Conway et al., 2012).

For each sample, reads were aligned to the NCBI build 37 (hg19) human reference genome sequence using BWA v0.6.2 r126 (Li and Durbin, 2009 We first created .sai files

using the bwa aln command. Then, we generated .sam files using the bwa sampe command. We sorted .sam files using Picard (http://picard.sourceforge.net). BAM files were generated using SAMtools v0.1.18 r982 (Li et al., 2009). Duplicates were marked and removed with Picard (http://picard.sourceforge.net).

#### *5.1.2. Ion Torrent*

Sequence alignment was performed for all samples using the Ion Torrent Suite software version 5.0.4. The alignment was performed using the NCBI build 37 (hg19) as the human reference genome sequence.

### 5.2. Xenome Software Test and Evaluation

We downloaded the Xenome software from http://www.nicta.com.au/bioinformatics and tested it using both human and mouse Illumina reads. We downloaded paired-end 100 bp mouse reads from the Sequence Read Archive (SRA) (code DRX014750 and DRX014751, forward and reverse reads respectively). We used our primary tumour sample reads as pure human reads. To evaluate the performance of the Xenome software in identifying mouse reads within PDX-derived samples, we performed calibration tests as follows:

1. We randomly selected human and mouse reads.
2. We pooled mouse and human reads in order to have different percentage of mouse contamination within the human sample (0-50%, step 5%).
3. We applied the Xenome software to each pool.
4. We evaluated the performance of Xenome software.

Having verified the ability of the Xenome software to identify mouse reads, we applied the software to our experimental samples and calculated mouse contamination in our PDX-derived samples. We removed mouse reads from our samples and retained only those reads flagged as human or both mouse and human.

For Ion Torrent reads, we downloaded from SRA reads derived from a human tumour sample (code SRR1531565) and a PDX sample (code SRR1534102) to evaluate the performance of Xenome software. We then applied the Xenome software to our Ion Torrent PDX-derived samples.

### 5.3. Mutation Detection and Annotation

#### *5.3.1. Illumina*

To identify somatic single nucleotide variants (SNVs) and somatic insertion/deletions (indels) that were present in the primary tumour, primary tumour-derived mammospheres, PDX and PDX-derived mammospheres samples and that were responsible for tumour onset and progression, we used VarScan2 software (Koboldt et al., 2012). SNVs and indels were annotated with Annovar (Wang et al., 2010).

To determine not only high frequency, but also rare variants, we used different frequency thresholds varying the *min-var-freq* parameter of VarScan2. We set thresholds at 0.01, 0.05, 0.1, 0.15 and 0.20. We are aware that at a frequency of 1 to 5%, called variants might be sequencing errors. Therefore, putative somatic mutations were manually reviewed using the Integrative Genomic Viewer (IGV) (Robinson et al., 2011; Thorvaldsdottir et al., 2013). All plots were generated using the statistical software R (R Core Team (2012), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria).

#### *5.3.2. Ion Torrent*

Variant calling and variant annotation were performed on the Ion Reporter website after uploading BAM files. Somatic variants were called using the *AmpliSeq Exome tumour-normal pair* workflow version 5.0 with default parameters. Annotation was performed using the *Annotate variants single sample* workflow version 5.0. Putative somatic mutations were manually reviewed using IGV (Robinson et al., 2011; Thorvaldsdottir et al.,

2013). All plots were generated using the statistical software R (R Core Team (2012), *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria).

### 5.4. Analysis of Mutated Genes

We checked whether the mutated genes identified in this study are present in published databases and, when present, if the mutated base is the same or not as that detected by us. We searched for our mutated genes in cBioPortal (Cerami et al., 2012; Gao et al., 2013) breast cancer studies: we selected all studies included into the *Breast Invasive Carcinoma* category and we searched for *Only Mutation* type of data. We downloaded all the mutations associated with each gene in our list of mutated genes and compared them to the mutations identified in our study. We also downloaded *Mutation Data (Genome Screens)* and *Mutation Data* version 77 from the COSMIC database (Forbes et al., 2015).

Furthermore, we determined whether the mutations identified in our study are present in specific protein domains and whether the amino acid changes are possibly damaging or not. To estimate the damaging potential of a nucleotide substitution, we used freely available web tools that are routinely applied by researchers. We used PROVEAN (Protein Variation Effect Analyzer) and SIFT (Choi et al., 2012; Doerks et al., 2002; Kumar et al., 2009) from J. Creig Venter Institute; FATHMM (Functional Analysis through Hidden Markov Models) MKL algorithm (Shihab et al., 2015); PolyPhen-2 (Polymorphism Phenotyping v2) single and batch query (Adzhubei et al., 2013; Adzhubei et al., 2010); Mutation Assessor (Reva et al., 2011); and CADD (Combined Annotation Dependent Depletion) web application (Kircher et al., 2014).

# Results

## 1. Optimization of Whole Exome Sequencing Protocol

To perform WES on our matched primary tumour, PDX and mammosphere samples, it was first necessary to optimize the procedure. In particular, since we were attempting to perform WES on gDNA derived from mammosphere cultures, which typically contain very few cells after 10 days of growth (~3000 cells/patient sample), we introduced a WGA step and determined whether this step introduced artefacts in the mutational profiles. We also introduced filtering steps to remove potentially contaminating DNA from the mouse host in the PDX samples or from other sources (e.g. unrelated individuals). The following sections describe in detail these optimisation steps.

### 1.1. Validation of the Whole Genome Amplification Step

A critical aspect of our study was the ability to perform WES (Balic et al., 2006) starting from limited amounts of gDNA, such as the gDNA extracted from mammosphere cultures; each mammosphere is composed of ~300 cells. Therefore, we performed several experiments to set up the DNA amplification protocol (WGA) necessary to run WES when starting from small quantities of DNA (5-10 ng). We used a gDNA sample present in the reaction kit (QIAGEN REPLI-g Mini Kit) as a positive control, water as negative control, gDNA extracted from a fresh frozen normal breast tissue sample as a healthy sample, gDNA from a fresh frozen primary tumour-derived mammosphere sample, and gDNA from a frozen breast primary tumour sample. By applying the WGA technique, we were able to obtain a sufficient amount of DNA to perform deep sequencing (~3 μg). In particular, starting from 5-10 ng of gDNA from the different frozen samples, we obtained > 3 μg of DNA, which is sufficient for WES (Figure 10).

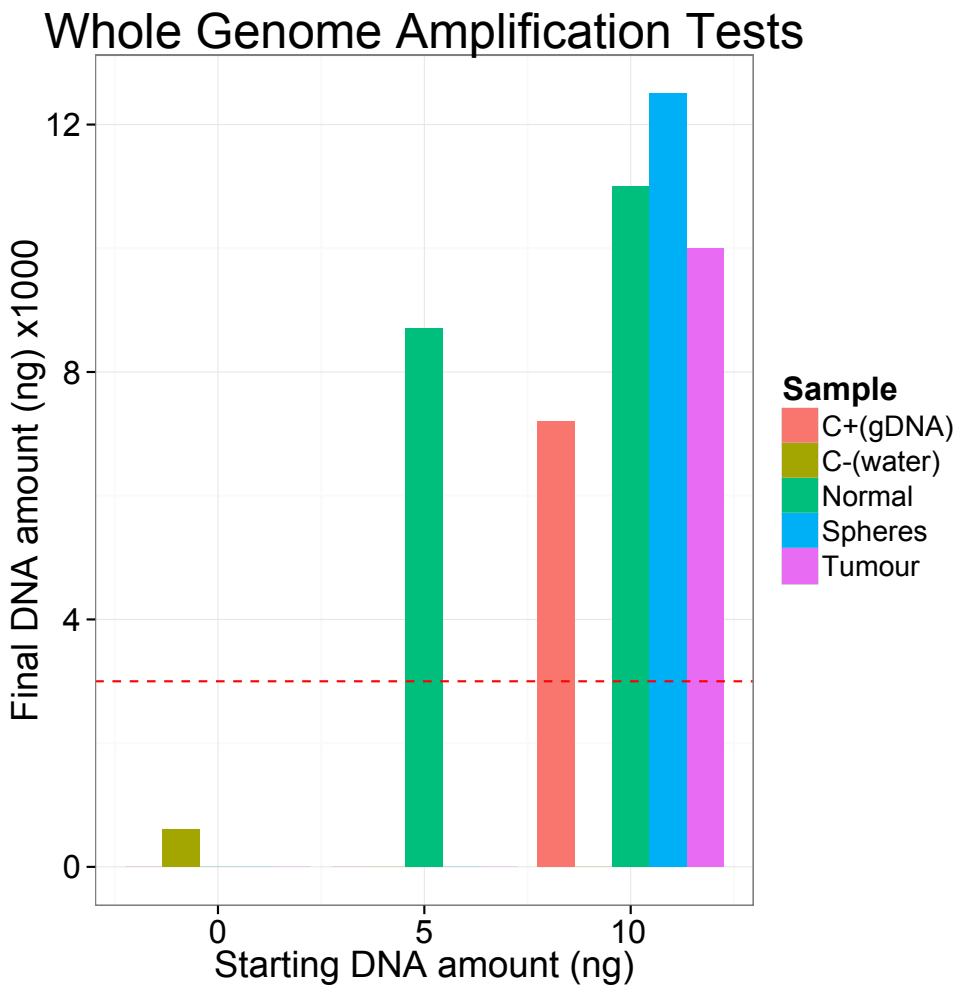


**Figure 10. Optimization of the WGA protocol.** Determination of yield of amplified DNA after WGA using different amounts of starting DNA. WGA was performed on 5 and 10 ng of gDNA obtained from different samples, including: C+, the positive control, gDNA sample present in the reaction kit (QIAGEN REPLI-g Mini Kit); C-, the negative control (water); Normal, gDNA extracted from a fresh frozen normal breast sample; Spheres, gDNA from a fresh frozen primary tumour-derived mammosphere sample; Tumour, gDNA from a frozen breast primary tumour sample. Red dashed line represents the minimum amount of DNA required for an Illumina Whole Exome Sequencing experiment (3 µg).

To establish whether the amplification step introduced a bias, or not, in the final WES profile, we sequenced both a non-amplified gDNA sample and a WGA-amplified DNA sample (WGA-DNA) derived from the same tumoral sample. We compared the two mutational profiles and observed a substantial agreement between the two profiles, both in terms of read alignment (>90% of aligned reads, 79% of these were "on-target reads" with a median depth of coverage of ~120X) (Figure 11) and called variants, with ~80% of variants in common (Figure 12). Our results are consistent with results of a recently

published paper (Hasmats et al., 2014), which showed that 89% of the variations found by WGA-DNA were shared with gDNA. Thus, we concluded that the WGA step did not introduce artefacts in WES experiment, enabling us to apply this protocol to our experimental samples.

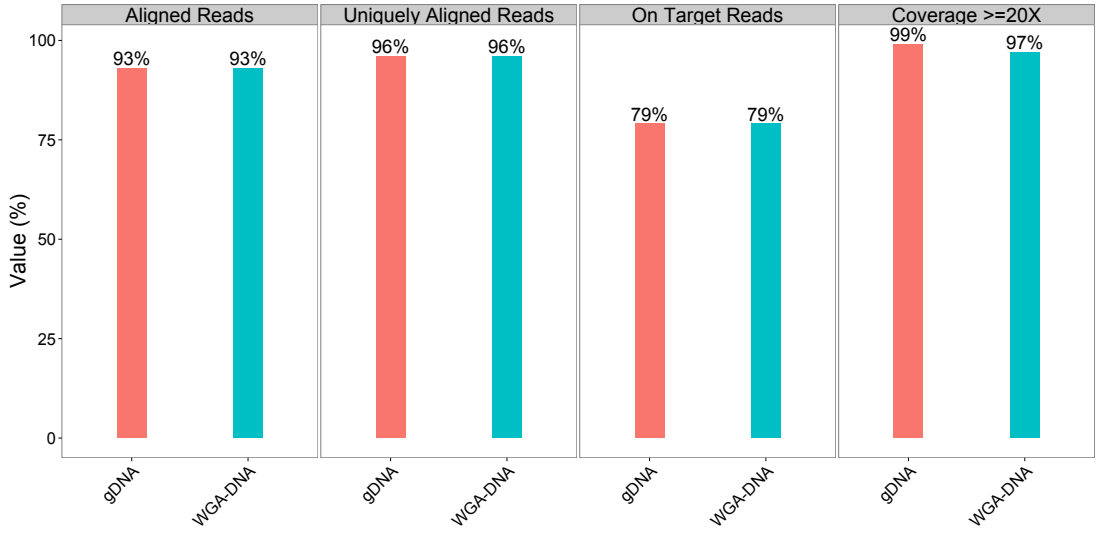


**Figure 11. The effect of WGA on alignment statistics.** Illumina reads of a tumoral test sample were obtained by WES using genomic DNA (gDNA) or whole-genome amplified DNA (WGA-DNA) as starting material. Aligned reads were calculated starting from raw reads of each sample, while uniquely aligned reads, on target reads and coverage >=20X were calculated from aligned reads. Coloured bars indicate different samples, salmon: gDNA; turquoise: WGA-DNA.

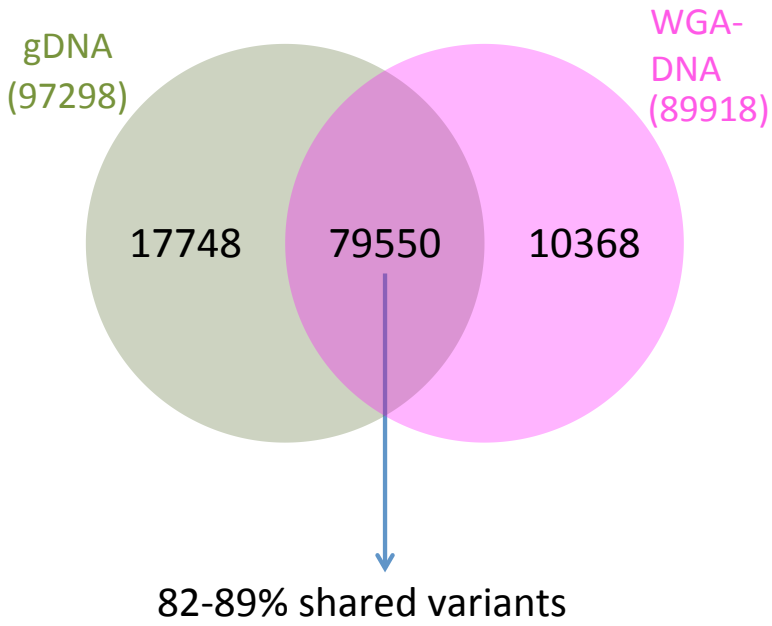


**Figure 12. The effect of WGA on variant calling.** Mutational profiles of a tumoral test sample were obtained by WES using genomic DNA (gDNA) or whole-genome amplified DNA (WGA-DNA) as starting material. Called variants were obtained by applying the VarScan2 software with default parameters. Although more variants were identified in gDNA compared to WGA-DNA, more than 80% of them are shared between the two experimental conditions. In brackets the total number of variants present in each sample.

### 1.2. Whole Exome Sequencing of Matched Primary Tumour, PDX and Mammosphere Samples

Having established that the WGA step did not introduce any substantial bias in the identification of genomic variants, we performed the WGA-WES (Whole-Genome Amplification followed by Whole Exome Sequencing) analysis of the first primary tumour-mammosphere-PDX matched sample from a breast cancer patient, together with a blood sample from the same patient as healthy control. We selected a grade 2, Luminal B/HER2-negative breast cancer as it was moderately aggressive and without known cancer driver events. We have also sequenced the PDX-derived mammospheres to compare their mutational profile with that of primary tumour-derived mammospheres. This latter comparison might allow us to identify specific clones in the cancer stem cell compartment that gave rise to the PDX.

Mammosphere samples were cultured for 10 days in suspension, then were harvested and dissociated enzymatically. The PDX sample was generated starting from dissociated cells from the primary tumour that were injected directly into the mammary fat pad of a NOD/SCID mouse.

Using the WGA step, we obtained on average 4 μg of WGA-DNA (range 3 – 5.5 μg), a sufficient amount of DNA for WES experiments. After amplification, we subjected our WGA-DNA to WES, obtaining an adequate number (>50 million reads) of aligned and on target reads (aligned reads = 80 – 140 million reads, on target reads = 65 – 110 million reads). Therefore, we were able to call variants, both somatic and germline, identifying a comparable number of mutations among all analysed samples (~10,000 mutations per sample).

### 1.3. Validation of the Xenome Filtering Step to Remove Contaminating Mouse DNA

An important aspect when dealing with PDX-derived samples is the potential contamination of mouse DNA because it is not always possible to eliminate all the mouse-contaminant cells before sequencing. To control for this potential contamination, we added an additional filtering step to our analysis pipeline that uses the Xenome software (Conway et al., 2012). This software allows us to evaluate the contamination from mouse DNA and to "clean-up" our samples. Before applying the Xenome software to our samples, we tested it by performing a calibration curve test. In particular, we spiked-in mouse reads into human reads at different percentages and applied the Xenome software. We observed that the calculated contamination was almost equal to that expected from the spike-in ratios; the calibration curve is a straight line with an intercept equal to 0.03 and a slope equal to 0.99 (Figure 13).

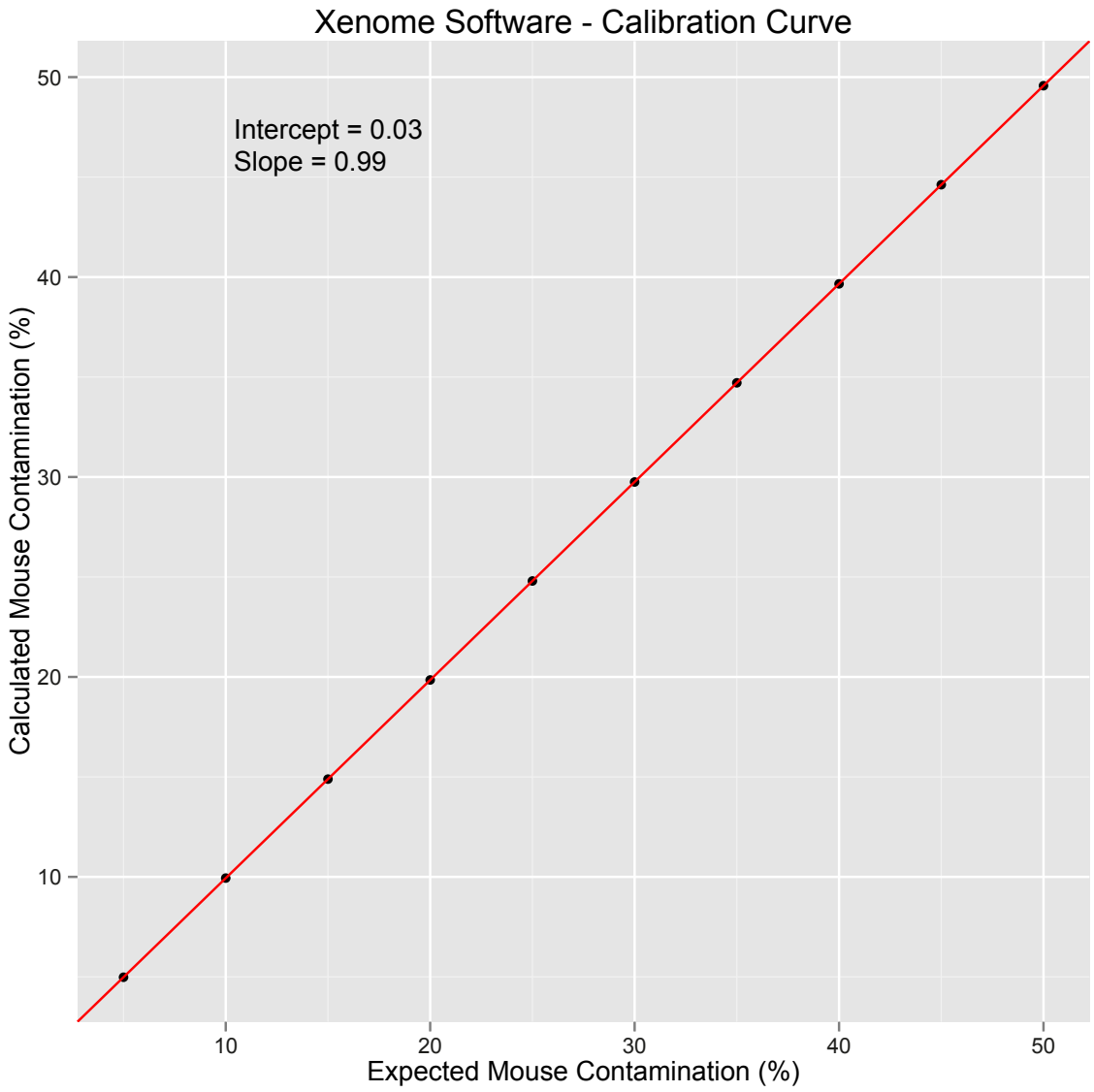


**Figure 13. Xenome Calibration Curve.** We used the Xenome software to evaluate the potential contamination of mouse DNA within human samples. We spiked-in mouse reads, downloaded from SRA (Sequence Read Archive), into human reads of the primary tumour at different percentages. We then calculated the mouse contamination with the Xenome software. On the X-axis, percentage of expected mouse contamination within human reads. On the Y-axis, percentage of calculated mouse contamination (number of reads flagged as mapping on the mouse genome by the Xenome software over the total number of reads) within human reads.

Applying this additional filtering step to the standard WES analysis pipeline, we detected a mouse contamination in the sequencing reads of ~3% in the PDX sample and ~21% in PDX-derived mammosphere sample. Accordingly, we retained only those reads mapping to the human genome or to both the human and mouse genomes. We then realigned these reads against the human genome and called the variants again (Table 1 and Table 2).

**Table 1. Comparison of Alignment Statistics of WES profiles of PDX and PDX-derived Mammospheres before and after Xenome filtering.**

| | Mapped Reads (%) | | Median Depth of Coverage (base) | | Coverage (>=20X) | | Unique Reads (%) | | On Target Reads (%) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Before** | **After** | **Before** | **After** | **Before** | **After** | **Before** | **After** | **Before** | **After** |
| **PDX** | 92.5 | 97.8 | 119 | 118 | 97.4 | 97.3 | 96.0 | 96.2 | 79.0 | 79.1 |
| **PDX-derived Mammospheres** | 68.3 | 92.6 | 89 | 106 | 95.6 | 96.9 | 95.1 | 94.0 | 79.3 | 75.3 |

Table reports the percentage of aligned reads, median depth of coverage, bases with coverage ≥ 20X, uniquely aligned reads and on target reads, before and after Xenome filtering (Before and After in the table).

**Table 2. Comparison of somatic variant calling of PDX and PDX-derived Mammospheres before and after Xenome filtering.**

| | Frequency 1% | | Frequency 5% | | Frequency 10% | | Frequency 15% | | Frequency 20% | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Before** | **After** | **Before** | **After** | **Before** | **After** | **Before** | **After** | **Before** | **After** |
| **PDX** | 276145 | 115668 | 69038 | 12438 | 24107 | 5587 | 10343 | 3428 | 5619 | 2415 |
| **PDX-derived Mammospheres** | 599716 | 105482 | 377350 | 15408 | 259016 | 6906 | 184608 | 4377 | 135668 | 3214 |

Somatic variants were called using the VarScan2 software. We used different frequency thresholds to call somatic mutations (1-5-10-15-20%), before and after Xenome filtering.

Despite filtering out many murine reads, we were able to obtain comparable WES profiles in all the samples analysed (primary tumour, primary tumour-derived mammospheres, PDX, PDX-derived mammospheres and blood) in terms of total aligned reads (93-98%), uniquely aligned reads (93-96%), on target reads (76-79%) and coverage (92-98% with coverage ≥ 20X) (Figure 14A). However, we did observe a lower median depth of coverage of aligned reads in samples with lower numbers of cells (primary tumour-derived mammospheres and PDX-derived mammospheres) compared with the other samples (Figure 14B). This might be a result of the lower heterogeneity of cells in mammosphere samples compared with primary tumours and PDXs, or a technical problem due to the lower amount of cells and therefore to a lower enrichment of some exonic regions. Although the median depth of coverage in mammosphere samples was sufficient to call variants, to overcome this problem in future experiments it might be possible either to increase the size of the mammosphere culture or to modify the incubation time of the WGA step.

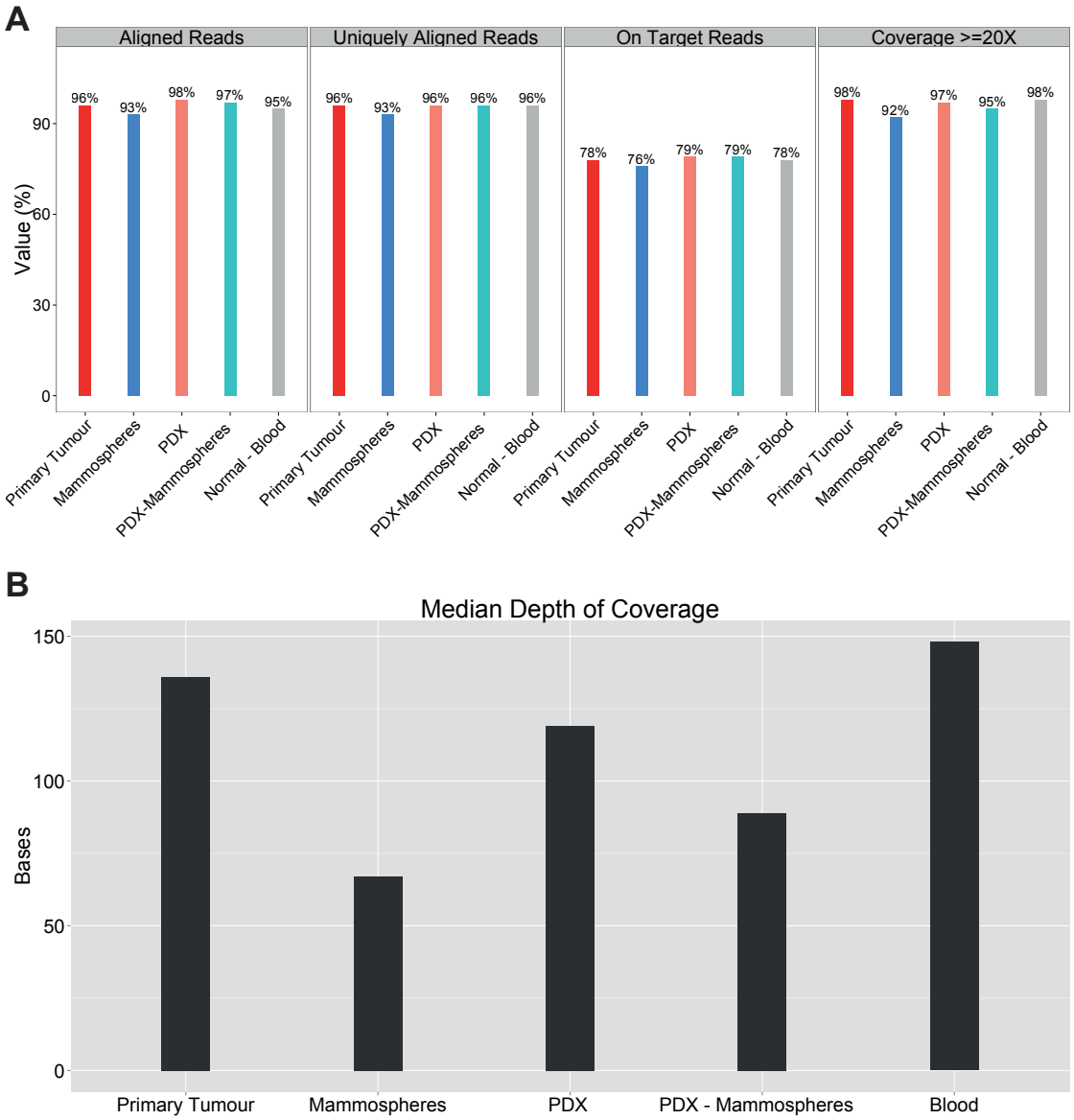


**Figure 14. Alignment statistics of WES experiments of the analysed samples.** WES experiments were performed using an Illumina HiSeq2000 instrument on the primary tumour, primary tumour-derived mammospheres, PDX, PDX-derived mammospheres and blood samples. A Xenome filtering step was applied to PDX and PDX-derived mammosphere samples. Reads were aligned using bwa software and statistics were performed by custom R pipeline. **A)** Percentage of aligned reads (based on raw number of reads), uniquely aligned reads, on target reads and bases with coverage $\geq$ 20X (based on number of aligned reads). Coloured bars indicate different samples. **B)** Median depth of coverage: on Y-axis the value of the median depth of coverage; on X-axis the analysed samples.

### 1.4. Control for Cross Contamination Germline Variation Analysis

To assess the level of cross contamination of the DNA samples analysed with other unrelated DNA samples, as recently reported in a study on human DNA samples (Jun et al., 2012), we identified germline variations (GVs) in all tumour samples. We compared GVs of primary tumour-derived mammospheres and PDX-derived samples with GVs of

the primary tumour. In all cases, we obtained a concordance in GVs of at least 90%. As a negative control, we also compared GVs of our samples with GVs of an unrelated healthy individual and obtained an overlap of 0.7%. Based on these results, we concluded that our samples were derived from the patient being analysed and that the level of cross-contamination was zero.

## 2. Mutational Profile Analysis

Using the mutational profiles of the matched primary tumour-PDX-mammosphere samples described above, we performed a series of analyses in an attempt to distillate patterns of mutations that could be informative either in terms of cancer relevant mutations or mutations related to CSCs.

Our analyses were conducted according the following workflow:

- Comparison of the mutational profile of primary tumour-derived mammospheres with that of the bulk primary tumour to identify mutations that were either "common" between the two samples or "enriched" in primary tumour-derived mammospheres compared to the primary tumour.
- Comparison of the mutational profile of primary tumour-derived mammospheres with that of the PDX to identify mutations shared between the two samples, but undetectable within the primary tumour due to the detection limit of WES.
- Analysis of the behaviour of the mutations identified in the comparison between primary tumour-derived mammospheres and the primary tumour with respect to the mutational profile of PDX and PDX-derived mammospheres.
- Analysis of the behaviour of the mutations identified in the comparison between primary tumour-derived mammospheres and PDX with respect to the mutational profile of PDX-derived mammospheres.

We also tested the significance of the overlapping mutations. Finally, we performed WES with an orthogonal sequencing technology (Ion Torrent platform) to validate the identified mutations.

### 2.1. Identification of Shared and Enriched Mutations

The analyses described above allowed the identification of mutations that are either common to all the different samples or enriched in the primary tumour-derived mammosphere or PDX samples compared with the primary tumour. The shared mutations would have appeared in the initial population of cells, be responsible for the tumour onset and be present in nearly all cells of the tumour. They could represent a pool of variants that permitted the development of the primary tumour, i.e. the founder (or so-called "truncal") somatic mutations, and would be expected to have a high frequency in all the matched samples. In addition, somatic mutations with low frequency shared between all samples may represent non-founder somatic mutations that appeared late in the evolution of the tumour and were selected for in a subpopulation of tumour cells.

In contrast, the set of mutations enriched or appeared *de* novo in mammospheres or in the PDX sample compared with the primary tumour (i.e. mutations present at a lower frequency or undetectable in the primary tumour, respectively) could be part of the genetic make-up of rare sub-clones of cells dispersed within the bulk tumour population. Thus, such mutations would be diluted in the context of the molecular heterogeneity of the entire tumour mass and would be detected either at very low frequency or be undetectable by NGS in the primary tumour, due to the technical limitation of the deep sequencing technology. The identification of such enriched/ appeared *de novo* mutations is compatible with the idea of the existence of sub-clones of cells that harbour private somatic mutations that represent the driving force of tumour progression and metastasis formation.

For the identification of mutations in the different samples, we used the *somatic* command of VarScan2 (Koboldt et al., 2012). Since VarScan2 permits the user to set the

minimum variant frequency to call and report a variant, we used different thresholds: 1%, 5%, 10%, 15% and 20% (default value). Overall, we observed comparable numbers of mutations in the matched tumour samples, i.e. the primary tumour, PDX and corresponding mammosphere cultures (~20,000 mutations at 5% of frequency threshold; Figure 15).

Since our aim was to identify "rare" mutations with a biological and functional role in CSCs and in the tumour, and to minimize the number of false positives, we focused our attention on mutations called at 5% frequency threshold that were exonic and non-synonymous.

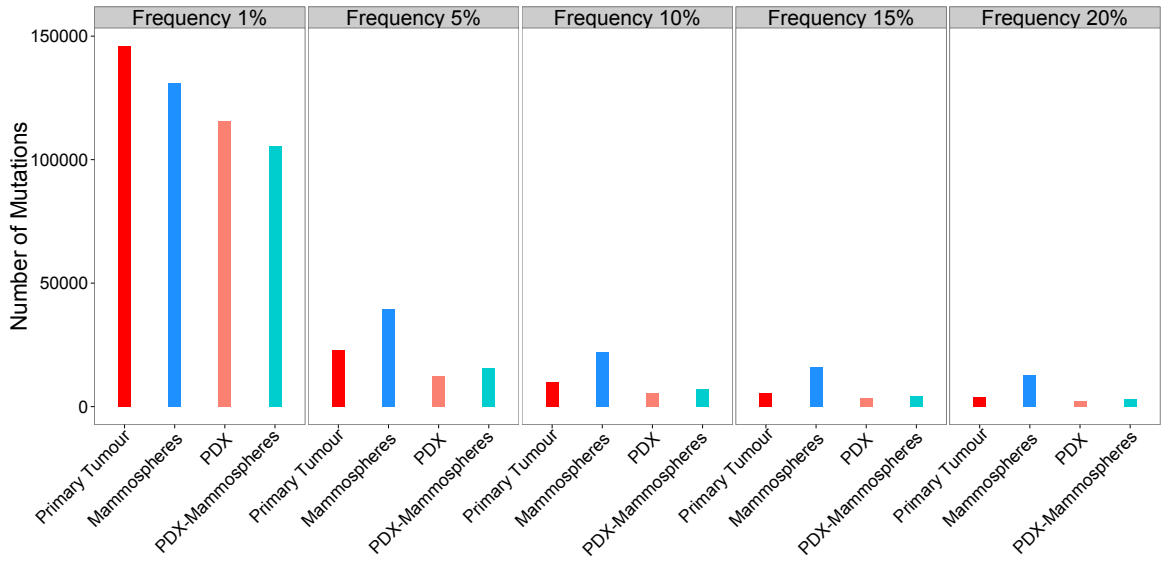


**Figure 15. Number of somatic variants identified by VarScan2 of analysed samples.** WES reads obtained from the matched tumour samples were analysed using VarScan2 software to identify somatic variants. Different frequency thresholds were used to call somatic mutations (Frequency 1-5-10-15-20%). On the Y-axis, number of detected mutations. On the X-axis, the analysed samples.

By the independent comparison of the mutational profile of the primary tumour-derived mammospheres with that of the primary tumour and PDX, we classified the called mutations into two main groups according to how they were shared between the different samples:

1. “Primary tumour-shared” mutations, corresponding to those mutations shared between the primary tumour and primary tumour-derived mammospheres (LIST1, Table 3).
2. “Mammosphere- and PDX-shared” mutations, corresponding to those mutations that were below the WES detection limit in the primary tumour but present in primary tumour-derived mammospheres and PDX (LIST2, Table 4).

We also performed further analyses (described below in 2.2) by cross-comparing: i) mutations of LIST1 (shared between primary tumour-derived mammospheres and primary tumour) with the mutational profiles of PDX and PDX-derived mammospheres; ii) mutations of LIST2 (shared between primary tumour-derived mammospheres and PDX compared to primary tumour) with the mutational profile of PDX-derived mammospheres. These analyses led to the identification of mutations shared across all the samples (primary tumour, primary tumour-derived mammospheres, PDX, and PDX-derived mammospheres), which likely represent “founder” or “truncal” mutations maintained throughout the branching evolution of the tumorigenic process (LIST3, Table 5 in red). In contrast, the set of mutations shared between primary tumour-derived mammospheres and the PDX that were also found in the profile of PDX-derived mammospheres (described in 2.3 below) are likely to represent mutations belonging to sub-clones of cells (likely CSCs) involved in tumour progression towards metastasis (LIST4, Table 5 in black).

**Table 3. "Primary tumour-shared" mutations present in LIST1.**

| Mutation Name | Chr | Position | Ref Allele | Var Allele | Mutation Frequency in Primary Tumour | Mutation Frequency in Primary Tumour-derived Mammospheres |
|---|---|---|---|---|---|---|
| *AEN* | 15 | 89173353 | T | G | 8% | 14% |
| *AKNA* | 9 | 117110115 | T | G | 14% | 7% |
| *ANKRD55* | 5 | 55412572 | C | T | 54% | 20% |
| *CHD3* | 17 | 7796815 | G | C | 6% | 12% |
| *CYP4F2* | 19 | 15989661 | G | A | 6% | 20% |
| *GALNT15* | 3 | 16250069 | C | T | 42% | 67% |
| *GOLGA6L1* | 15 | 22743086 | C | T | 15% | 19% |
| *GPATCH3* | 1 | 27226921 | C | A | 5% | 10% |
| *IRS1* | 2 | 227659828 | T | G | 7% | 8% |
| *KDM4E* | 11 | 94758846 | A | G | 6% | 19% |
| *KLRF1* | 12 | 9994450 | G | A | 10% | 14% |
| *KRT10*_1 | 17 | 38975276 | G | C | 8% | 18% |
| *KRT10*_2 | 17 | 38975277 | A | C | 8% | 17% |
| *KRT10*_3 | 17 | 38975279 | C | T | 7% | 17% |
| *KRT10*_4 | 17 | 38975280 | T | A | 7% | 17% |
| *KRTAP4-8* | 17 | 39254054 | A | T | 14% | 15% |
| *LHCGR* | 2 | 48982749 | A | G | 5% | 11% |
| *MUC6*_1 | 11 | 1016585 | G | C | 6% | 7% |
| *MUC6*_2 | 11 | 1030228 | A | C | 25% | 14% |
| *OR13C2* | 9 | 107367653 | G | C | 6% | 20% |
| *PABPC3*_1 | 13 | 25670877 | G | A | 8% | 5% |
| *PABPC3*_2 | 13 | 25670988 | T | G | 7% | 17% |
| *PABPC3*_3 | 13 | 25671089 | G | T | 18% | 18% |
| *TAS2R19*_1 | 12 | 11174467 | A | C | 8% | 13% |
| *TAS2R19*_2 | 12 | 11174476 | G | A | 11% | 11% |
| *TAS2R30*_1 | 12 | 11285909 | G | C | 8% | 10% |
| *TAS2R30*_2 | 12 | 11285975 | A | T | 6% | 7% |
| *TAS2R30*_3 | 12 | 11285978 | A | G | 6% | 6% |
| *TAS2R30*_4 | 12 | 11286002 | A | C | 5% | 7% |
| *TAS2R30*_5 | 12 | 11286024 | T | C | 5% | 7% |
| *TUBGCP3* | 13 | 113210444 | G | T | 6% | 6% |

Table reports annotations of LIST1 mutations: mutation name (composed of gene name alone when only one mutation is present in the corresponding gene or gene name_ordinal number when more than one mutation was present in the same gene), chromosomal coordinates, reference and variant alleles, mutation frequency in primary tumour-derived mammospheres and mutation frequency in primary tumour. Mutation frequency for each variant was calculated by VarScan2 software as the number of mutated reads over the coverage in each position. LIST1 is composed of mutation shared between primary tumour and primary tumour-derived mammospheres.

**Table 4. "Mammosphere- and PDX-shared" mutations present in LIST2.**

| Mutation Name | Chr | Position | Ref Allele | Var Allele | Mutation Frequency in Primary Tumour | Mutation Frequency in Primary Tumour-derived Mammospheres | Mutation Frequency in PDX |
|---|---|---|---|---|---|---|---|
| *ACOT1* | 14 | 74004547 | C | T | ND | 16% | 12% |
| *AMOTL1* | 11 | 94592720 | C | T | ND | 13% | 5% |
| *GOLGA6L1* | 15 | 22743398 | T | C | ND | 25% | 17% |
| *GRIA3* | X | 122336604 | C | G | ND | 8% | 8% |
| *MAL2* | 8 | 120220749 | A | C | ND | 7% | 6% |
| *MUC4*_1 | 3 | 195515470 | T | G | ND | 13% | 14% |
| *MUC4*_2 | 3 | 195515459 | C | T | ND | 21% | 20% |
| *MUC4*_3 | 3 | 195515449 | A | T | ND | 29% | 29% |
| *OR13C2* | 9 | 107367674 | G | A | ND | 20% | 25% |
| *PABPC3* | 13 | 25670907 | C | A | ND | 5% | 7% |
| *RETSAT*_1 | 2 | 85570849 | C | T | ND | 44% | 12% |
| *RETSAT*_2 | 2 | 85570857 | G | A | ND | 47% | 10% |
| *USP20* | 9 | 132630666 | A | T | ND | 33% | 17% |
| *ZNF28* | 19 | 53303182 | T | C | ND | 8% | 5% |

Table reports annotations of LIST2 mutations: mutation name (composed of gene name alone when present only one mutation for the corresponding gene and gene name_ordinal number when more than one mutation was present for the same gene), chromosomal coordinates, reference and variant alleles, mutation frequency in primary tumour, mutation frequency in primary tumour-derived mammospheres and mutation frequency in PDX. Mutation frequency for each variant was calculated by VarScan2 software as the number of mutated reads over the coverage in each position. ND = not detectable by WES. LIST2 is composed of mutations that were not detectable in the primary tumour, but present in primary tumour-derived mammospheres and PDX.

**Table 5. Candidate "founder" mutations present in LIST3 (in red) and candidate CSC mutations present in LIST4 (in black).**

| Mutation Name | Chromosome | Position | Ref Allele | Var Allele | Mutation Frequency in Primary Tumour | Mutation Frequency in Primary Tumour-derived Mammospheres | Mutation Frequency in PDX | Mutation Frequency in PDX-derived Mammospheres |
|---|---|---|---|---|---|---|---|---|
| *AKNA* | 9 | 117110115 | T | G | 14% | 7% | 19% | 11% |
| *ANKRD55* | 5 | 55412572 | C | T | 54% | 20% | 48% | 56% |
| *GALNT15* | 3 | 16250069 | C | T | 42% | 67% | 42% | 53% |
| *PABPC3_3* | 13 | 25671089 | G | T | 18% | 18% | 9% | 19% |
| *ACOT1* | 14 | 74004547 | C | T | ND | 16% | 12% | 18% |
| *MUC4_2* | 3 | 195515459 | C | T | ND | 21% | 20% | 17% |
| *PABPC3* | 13 | 25670907 | C | A | ND | 5% | 7% | 6% |
| *RETSAT_1* | 2 | 85570849 | C | T | ND | 44% | 12% | 21% |
| *RETSAT_2* | 2 | 85570857 | G | A | ND | 47% | 10% | 26% |

Table reports annotations of LIST3 and LIST4 mutations: mutation name (composed of gene name alone when only one mutation is present in the corresponding gene and gene name_ordinal number when more than one mutation is present for the same gene), chromosomal coordinates, reference and variant alleles, mutation frequency in primary tumour-derived mammospheres, mutation frequency in primary tumour, mutation frequency in PDX and mutation frequency in PDX-derived mammospheres. Mutation frequency for each variant was calculated by VarScan2 software as the number of mutated reads over the coverage in each position. In red: mutations that were present in all four analysed samples, the founder mutations (LIST3); in black: mutations enriched in primary tumour-derived mammospheres, PDX and PDX-derived mammospheres (LIST4). ND = Not Detectable.

To verify the quality and significance of these mutations within the context of our experimental design, we manually checked a representative pool for each list that was: i) supported by at least two reads, one in forward and one in reverse; ii) present in at least two samples; iii) present in the normal blood sample with a frequency ≤ 5%. We did not exclude mutations annotated with a dbSNP entry, due to the fact that some somatic point mutations present in dbSNP database should not be excluded *a priori* as naturally occurring in the human genome because they have been experimentally associated with cancer (Jung et al., 2013). This manual control enabled us to verify that mutations in LIST1, LIST2, LIST3 and LIST4 were supported by good quality data in terms of supporting reads and coverage.

To evaluate the significance of our gene lists, we compared the number of mutations in our lists with those present in random lists (LIST1_rand, LIST2_rand, LIST3_rand and LIST4_rand). We observed that our lists of genes showed a different distribution compared with the random lists (Figure 16A-D). We also tested the differences using the phyper function of R for LIST1 and LIST2, obtaining a p-value < 0.001 for both lists. The analysis indicated that shared mutations were not due to random mutational events, but rather they were the result of positive evolutionary selection. Thus, these mutations represent good candidates for the identification of cancer-relevant mutations involved in disease onset and progression, and therefore could be considered mutations for further analysis.

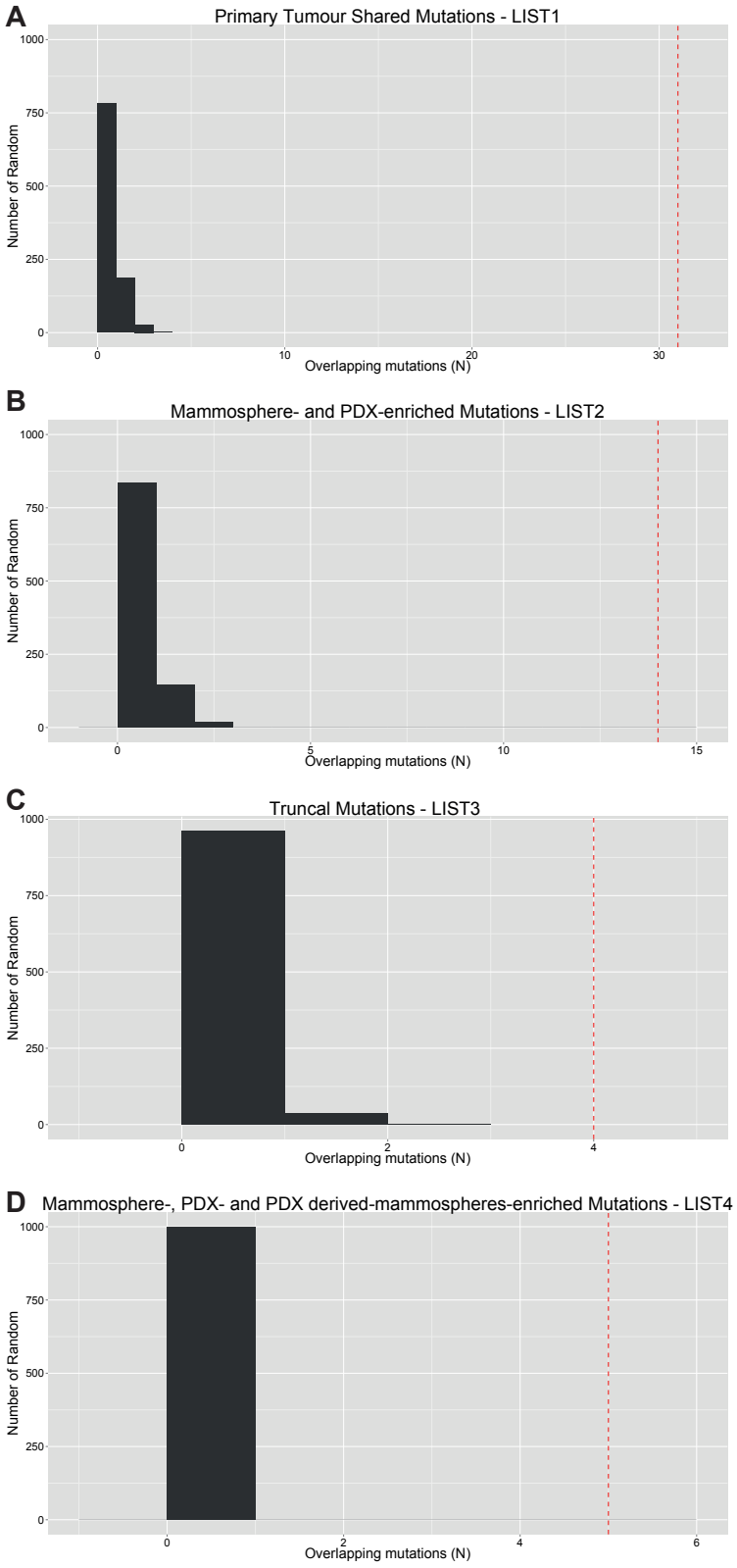


**Figure 16. Comparison of our lists of mutations with random lists of mutations.** Histograms showing the distribution of random overlaps in the four groups of mutations: LIST1, LIST2, LIST3 and LIST4. Y-axis: the number of random sets of overlapping mutations. X-axis: the number of overlapping mutations. The dashed red line represents the number of mutations present in our lists. **A**) LIST1: primary-shared mutations present in the primary tumour and the primary tumour-derived mammospheres; **B**) LIST2: mammosphere-PDX-enriched mutations present in primary tumour-derived mammospheres and PDX; **C**) LIST3: truncal mutations present in primary tumour, primary tumour-derived mammospheres, PDX and PDX-derived mammospheres; **D**) LIST4:

mammospheres- PDX- and PDX-derived mammospheres enriched mutations present in, primary tumour-derived mammospheres, PDX and PDX-derived mammospheres.

### 2.2. Identification of Candidate Founder Mutations

As mentioned above, in an attempt to identify candidate founder mutations, we assessed the presence of mutations shared between the primary tumour and primary tumour-derived mammospheres (mutations comprised in LIST1) in the PDX and in the PDX-derived mammospheres. Among the 31 variants common to primary tumour and primary tumour-derived mammospheres, 11 were also present in the PDX, and 4 of these 11 mutations, were also present in the PDX-derived mammospheres (LIST3, Table 5, in red).

We reasoned that the 11 mutations shared between the primary tumour, primary tumour-derived mammospheres and PDX could represent a cluster of mutations of key importance for the tumorigenic process. Indeed, we note that 4 of these mutations were also conserved in PDX-derived mammospheres arguing that they were positively selected for during tumour formation in a host organism. Notably, only two mutations, the *ANKRD55* and the *GALNT15* mutations, could be considered as founder mutations as they appeared at high frequency in all analysed samples. Based on results from the cross-comparison of the different mutational profiles, we concluded that our approach was able to identify a minimal set of mutations with a likely driver function in tumour onset and expansion: the *AKNA*, the *ANKRD55*, the *GALNT15* and the *PABPC3*_3 mutations (LIST3, Table 5, in red).

The *AKNA* gene (AT-hook-containing transcription factor), whose protein product is a transcription factor that activates the expression of CD40 and its ligand CD40L, has been described to be a cervical cancer susceptibility gene (Martinez-Nava et al., 2015). Moreover, nucleotide variations in this gene have been identified in relapsed tumours of children affected by high hyperdiploid acute lymphoblastic leukaemia (Chen et al., 2015).

The *ANKRD55* (ankyrin repeat domain-containing protein 55) gene is still uncharacterized in terms of function and implication in cancer.

The *GALNT15* (polypeptide N-acetylgalactosaminyltransferase 15) gene encodes a protein member of the pp-GalNAc-T family, which is involved in the initiation of mucin-type O-glycosylation. The *GALNT15* gene has a similar expression pattern to *MUC5AC* and is able to transfer to *MUC5AC* up to seven GalNAc residues (Cheng_et_al-2004-FEBS_Letters). The function of *GALNT15* in cancer is still uncharacterized.

The *PABPC3* (poly(A)-binding protein cytoplasmic 3) gene encodes a protein involved in poly(A) tail of mRNAs. *PABPC3* implication in cancer has not been characterized yet.

In conclusion, of the 4 identified genes harbouring candidate founder mutations, only the *AKNA* gene has been demonstrated to be associated with cancer. Therefore, functional validation of the other 3 genes is warranted to demonstrate their possible involvement in cancer.

### 2.3. Identification of Candidate CSC-specific Mutations

CSCs are very rare cells within the primary tumour and therefore mutations present in the CSC compartment are predictably very difficult to identify when performing deep sequencing on the primary tumour. Hence, in an attempt to detect candidate CSC-specific mutations, we focused our attention on variants that were enriched in primary tumour-derived mammospheres or in the PDX compared with the primary tumour. In particular, we classified as "enriched" those mutations that appeared in the primary tumour-derived mammospheres or in PDX with a frequency of at least 5% higher than that observed in the primary tumour.

We identified 14 mutations, within LIST1 mutations, that, although detectable in the primary tumour, were enriched in primary tumour-derived mammospheres (*AEN*, *CHD3*, *CYP4F2*, *GALNT15*, *GPATCH3*, *KDM4E*, *KRT10*_1, *KRT10*_2, *KRT10*_3,

*KRT10*_4, *LHCGR*, *OR13C2*, *PABPC3*_1, *TAS2R19*_1). Of these 14 mutations, only 1, *OR13C2,* was also present and enriched in the PDX (Figure 17), while additional 4 mutations (*LHCGR*, *PABPC3*_1, *TAS2R19*_1 and *GALNT15*) were present in the PDX, but not enriched (i.e., with a frequency in the PDX comparable to that observed in the primary tumour).

From the comparative analysis of the mutational profiles of the primary tumour, primary tumour-derived mammospheres and PDX, we were therefore able to characterize a first set of 5 mutations (the *OR13C2*, *LHCGR*, *PABPC3*_1, *TAS2R19*_1 and *GALNT15* mutations mentioned above), which were common to the primary tumour and mammospheres, but also enriched in mammospheres and maintained in the PDX outgrowth. We argue that these mutations are harboured in sub-clones of cells endowed with a unique tumorigenic ability and therefore operationally classifiable as CSCs, a hypothesis in keeping with the observation that these mutations are part of the molecular profiles of the mammospheres and of the PDX derived from the primary tumour.

We also noted that the *GALNT15* mutation, besides being shared and expressed at high frequency across all the samples analysed, was further enriched in primary tumour-derived mammospheres and PDX-derived mammospheres. We therefore speculated that this mutation could represent a lesion characteristic of the original CSC sub-clone present from the beginning of, and possibly responsible for, the tumorigenic process.

In contrast, the 14 mutations enriched in primary tumour-derived mammospheres and PDX compared to the primary tumour counterpart (LIST2), likely constitute a set of mutations present in sub-clones of cells with a putative CSC activity, which appeared in the course of tumorigenesis as a consequence of genomic instability and became "dominant" in the evolutionary history of the primary tumour.

In our analysis, we also identified a set of mutations that were shared between primary tumour-derived mammospheres and PDX, and appeared *de novo* in these samples

compared to the primary tumour (*MAL2, GRIA3, ZNF28, AMOTL1, MUC4*_1, *MUC4*_3, *OR13C2, GOLGA6L1, USP20, PABPC3*, *MUC4*_2, *RETSAT*_1*, RETSAT*_2 and *ACOT1*) (see LIST2 and Figure 18). We reasoned that the appearance of these mutations may be due to the existence of very rare sub-clones of cells, likely endowed with CSC activity, that became predominant in the short proliferative history of the primary tumour-derived mammospheres and of the PDX. In keeping with this hypothesis, we found that, in the set of mutations shared between primary tumour-derived mammospheres and PDX, it was possible to identify 5 mutations (*PABPC3*, *MUC4*_2, *RETSAT*_1*, RETSAT*_2 and *ACOT1*) (LIST4, Table 5, in black) common to PDX-derived mammospheres, arguing that these mutations likely represent candidate CSC-specific mutations.

In conclusion, the bulk of our results indicate the existence of sub-clones of tumour cells that can harbour either shared (mutations comprised in LIST1) or private (mutations comprised in LIST4) mutations, which likely reflect set of mutations present in sub-clones of cells endowed with stemness activity. We also speculated that the frequency of these CSC sub-clones in the bulk tumour mass might be a function of their order of appearance throughout the evolutionary history of the tumour or of the specific molecular mechanisms controlling the dynamics of CSC expansion. Regardless, mutations harboured in rare sub-clones of cells would appear at very low frequency or be undetectable by NGS in the primary tumour, due to the technical limitation of the deep sequencing technology.

Together, these data indicate that our approach of performing WES on matched primary tumour, PDX and mammosphere samples is a suitable strategy for identifying rare mutations that are expressed at very low levels in the primary tumour, but become enriched during mammosphere or xenograft formation due to their presence in cells featuring CSC activity. These mutations could represent candidate driver mutations that are important in tumour onset and progression, including evolution towards metastasis.

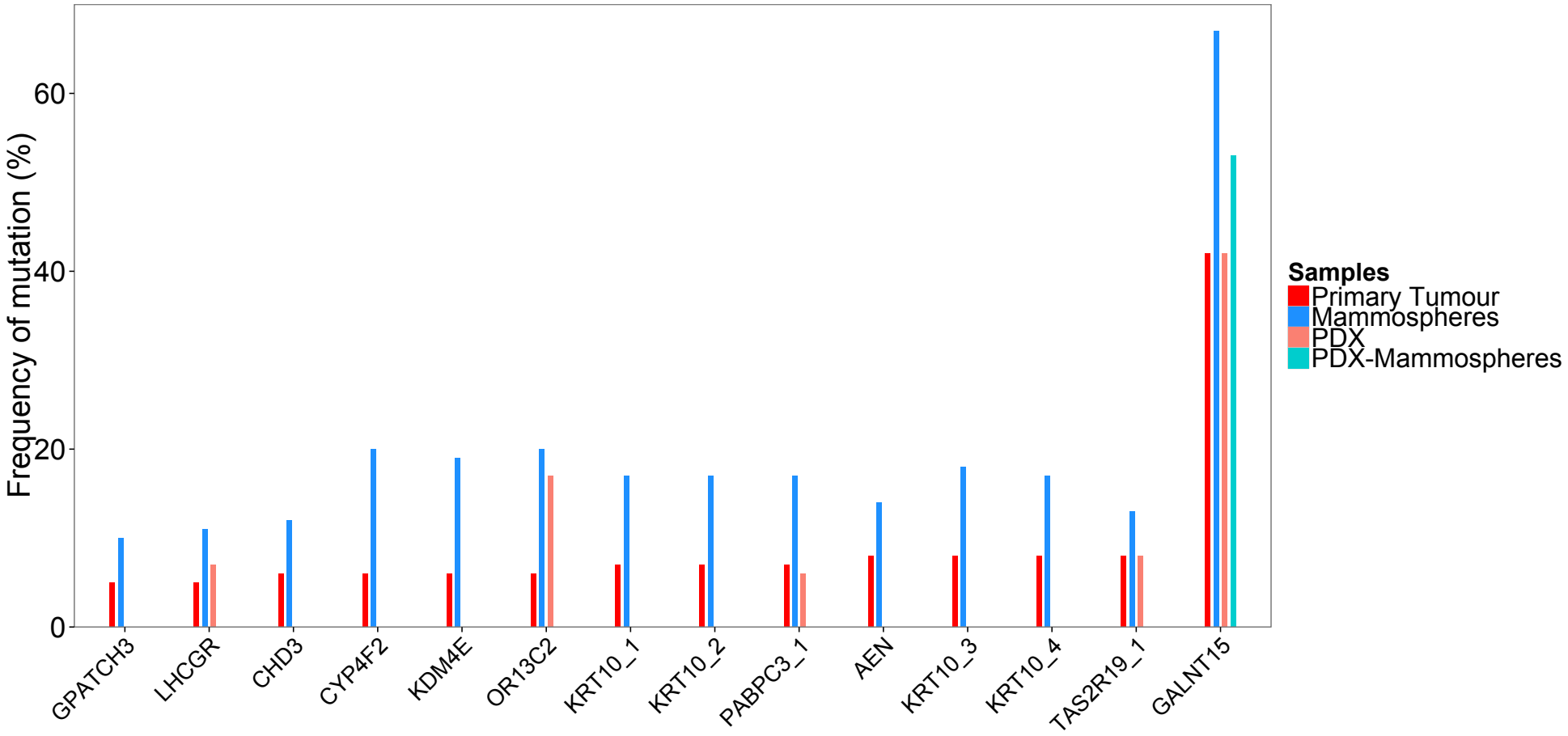


**Figure 17. Mutations in LIST1 with a higher frequency in primary tumour-derived mammospheres, PDX and PDX derived-mammospheres compared with the primary tumour.** Mutations that were found to be enriched in terms of frequency in primary tumour-derived mammospheres and/or in the PDX and PDX-derived mammospheres compared with the primary tumour. Y-axis: frequency of detected mutations. X-axis: the mutated genes - different mutated positions within the same gene are reported as GeneName_#.

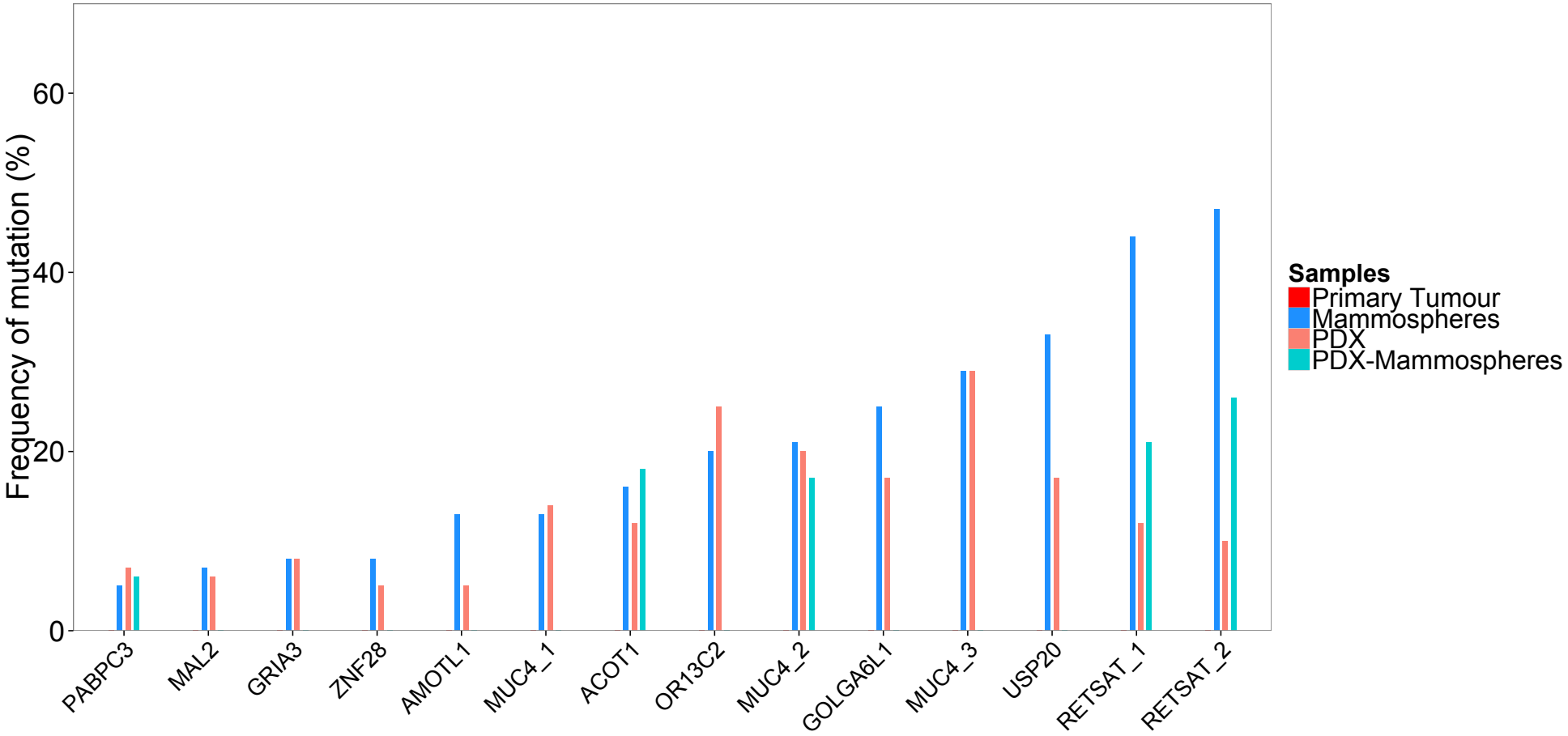


**Figure 18. Mutation of LIST2 with a higher frequency in primary tumour-derived mammospheres, PDX and PDX derived-mammospheres compared with the primary tumour.** Mutations that were found to be enriched in terms of frequency in primary tumour-derived mammospheres, in the PDX and PDX-derived mammospheres compared with the primary tumour. Y-axis: frequency of detected mutations. X-axis: the mutated genes - different mutated positions within the same gene are reported as GeneName_#.

### 2.4. Validation of Candidate CSC-specific Mutations by Ion Torrent WES

To validate our results from the WES analysis using the Illumina platform, we decided to perform WES experiments applying an independent sequencing technology, namely the Ion Torrent platform. We therefore performed WES of WGA-DNA of the primary tumour, primary tumour-derived mammospheres, PDX, PDX-derived mammospheres and blood samples using the Ion Torrent sequencing platform.

With the idea to validate CSC-specific mutations that are part of sub-clones of cells characterized by CSC activity, we focused our analyses on mutations enriched in LIST1 and present in LIST2. Within enriched mutations in LIST1, we were able to confirm only two of them: *GALNT15* and *OR13C2*. Despite a coverage >20X, the *GPATCH3*, *LHCGR*, *KDM4E*, *PABPC3*_1, *AEN* and *CHD3* mutations failed to be confirmed by Ion Torrent. In addition, the other mutations (*TAS2R19*_1, *KRT10*_1, *KRT10*_2, *KRT10*_3, *KRT10*_4 and *CYP4F2*) lacked sufficient coverage (<20X) in the Ion Torrent experiments (Table 7).

Of the mutations in LIST2, we were able to confirm mutations in the *ACOT1*, *OR13C2* and *RETSAT* genes (both mutations), which all had a frequency higher than 10% in primary tumour-derived mammospheres, PDX and PDX-derived mammospheres. All 3 *MUC4* mutations lacked sufficient coverage in the Ion Torrent analysis, with only a few reads (<20) at the mutated position, and therefore could not be confirmed. Instead, the *AMOTL1*, *PABPC3* and *ZNF28* mutations were not present in the Ion Torrent experiment despite a coverage >20X in primary tumour-derived mammospheres, PDX and PDX-derived mammospheres (Table 6).

A possible explanation for this discrepancy between results from Illumina and Ion Torrent may lie in the different sequencing chemistries implemented in the two systems. Although we obtained less aligned reads in the Ion Torrent WES compared to the Illumina WES, the on target reads were higher in the Ion Torrent WES compared to the Illumina WES. In addition, while the mean depth of coverage and coverage >20X for the primary

tumour, primary tumour-derived mammospheres and PDX were comparable between Illumina and Ion Torrent WES, they were sub-optimal in the Ion Torrent analysis compared to the Illumina analysis for the PDX-derived mammospheres (Table 7). Hence, we were unable to obtain similar mutational profiles for the same sample using the two different sequencing approaches.

Based on these considerations, we decided not to exclude *a priori* from further analyses those mutations that were not validated by Ion Torrent. Hence, we decided to characterize the enriched mutated genes in both primary tumour-derived mammospheres and the PDX compared with the primary tumour (in both LIST1 and LIST2) found in the Illumina experiments in terms of functional involvement in cancer.

**Table 6. Candidate CSC-specific mutations identified by Illumina and Ion Torrent WES.**

| | | | | Illumina WES | | | Ion Torrent WES | | |
|---|---|---|---|---|---|---|---|---|---|
| **Mutation Name** | **List** | **Chr** | **Position** | **Primary tumour-derived Mammospheres** | **PDX** | **PDX-derived Mammospheres** | **Primary tumour-derived Mammospheres** | **PDX** | **PDX-derived Mammospheres** |
| *AEN* | LIST1 | 15 | 89173353 | P | NP | NP | NP | NP | NP |
| *CHD3* | LIST1 | 17 | 7796815 | P | NP | NP | NP | NP | NP |
| *CYP4F2* | LIST1 | 19 | 15989661 | P | NP | NP | NP | NP | NP |
| *GALNT15* | LIST1 | 3 | 16250069 | P | P | P | P | P | P |
| *GPATCH3* | LIST1 | 1 | 27226921 | P | NP | NP | NP | NP | NP |
| *KDM4E* | LIST1 | 11 | 94758846 | P | NP | NP | NP | NP | NP |
| *KRT10*_1 | LIST1 | 17 | 38975279 | P | NP | NP | NP | NP | NP |
| *KRT10*_2 | LIST1 | 17 | 38975280 | P | NP | NP | NP | NP | NP |
| *KRT10*_3 | LIST1 | 17 | 38975276 | P | NP | NP | NP | NP | NP |
| *KRT10*_4 | LIST1 | 17 | 38975277 | P | NP | NP | NP | NP | NP |
| *LHCGR* | LIST1 | 2 | 48982749 | P | P | NP | NP | NP | NP |
| *OR13C2* | LIST1 | 9 | 107367653 | P | P | NP | P | P | P |
| *PABPC3*_1 | LIST1 | 13 | 25670988 | P | P | NP | NP | NP | NP |
| *TAS2R19*_1 | LIST1 | 12 | 11174467 | P | P | NP | NP | NP | NP |
| *ACOT1* | LIST2 | 14 | 74004547 | P | P | P | P | P | P |
| *AMOTL1* | LIST2 | 11 | 94592720 | P | P | NP | NP | NP | NP |
| *GOLGA6L1* | LIST2 | 15 | 22743398 | P | P | NP | NP | NP | NP |
| *GRIA3* | LIST2 | X | 122336604 | P | P | NP | NP | NP | NP |
| *MAL2* | LIST2 | 8 | 120220749 | P | P | NP | NP | NP | NP |
| *MUC4*_1 | LIST2 | 3 | 195515470 | P | P | NP | NP | NP | NP |
| *MUC4*_2 | LIST2 | 3 | 195515459 | P | P | P | NP | NP | NP |
| *MUC4*_3 | LIST2 | 3 | 195515449 | P | P | NP | NP | NP | NP |
| *OR13C2* | LIST2 | 9 | 107367674 | P | P | NP | P | P | P |
| *PABPC3* | LIST2 | 13 | 25670907 | P | P | P | NP | NP | NP |
| *RETSAT*_1 | LIST2 | 2 | 85570849 | P | P | P | P | P | P |
| *RETSAT*_2 | LIST2 | 2 | 85570857 | P | P | P | P | P | P |
| *USP20* | LIST2 | 9 | 132630666 | P | P | NP | NP | NP | NP |
| *ZNF28* | LIST2 | 19 | 53303182 | P | P | NP | NP | NP | NP |

The table reports the mutation names, the chromosomal coordinates and the presence of the mutation in the Illumina and Ion Torrent WES profiles of the primary tumour-derived mammospheres, PDX and PDX-derived mammospheres samples for candidate CSC-specific mutations. P= present; NP = not present. In red: mutations considered for further analysis.

**Table 7. Comparison of Illumina and Ion Torrent Alignment Statistics.**

| | Illumina WES | | | | Ion Torrent WES | | | |
|---|---|---|---|---|---|---|---|---|
| | **Number of Mapped Reads** | **Reads On Target (%)** | **Mean Depth of Coverage (base)** | **Coverage >20X (%)** | **Number of Mapped Reads** | **Reads On Target (%)** | **Mean Depth of Coverage (base)** | **Coverage >20X (%)** |
| **Primary Tumour** | 125,386,838 | 78 | 136 | 98 | 31,624,977 | 86 | 82 | 89 |
| **Mammospheres** | 111,268,203 | 76 | 67 | 92 | 41,617,735 | 84 | 102 | 71 |
| **PDX** | 107,663,534 | 79 | 119 | 97 | 22,289,307 | 90 | 60 | 84 |
| **PDX-derived Mammospheres** | 82,380,904 | 79 | 89 | 95 | 8,615,822 | 84 | 21 | 39 |
| **Blood** | 142,144,133 | 78 | 148 | 98 | 30,934,336 | 92 | 84 | 90 |

WES experiments were performed using an Ion Torrent instrument on the primary tumour, primary tumour-derived mammospheres, PDX, PDX-derived mammospheres and blood samples. Number of total mapped reads, percentage of on target reads, mean depth of coverage and bases with coverage > 20X are reported.

### 2.5. Determination of Functional Impact of Candidate CSC-specific Mutations

To understand the potential functional impact of our candidate CSC-specific mutations, we first determined whether the amino acid changes resulting from the identified mutations in our candidate CSC-specific genes were within particular functional domains of the corresponding proteins. To this end, we used UniProt website (http://www.uniprot.org/) and cBioPortal website (http://www.cbioportal.org/) to map the chromosomal coordinates of the mutated positions onto protein coordinates. We also used the TCGA breast invasive carcinoma dataset to screen for mutations previously identified in our candidate CSC-specific genes in breast cancer.

We found that the mutations present in *GOLGA6L1*, *MAL2*, *MUC4*, *PABPC3*, *RETSAT and ZNF28* do not fall within specific protein domains (Figure 19), while the mutations in *ACOT1, AMOTL1, GRIA3, OR13C2* (both two mutations present in the two lists) and *USP20* were present in specific protein domains (Figure 20). In particular, the *ACOT1* mutation is in the Acyl-CoA thioester hydrolase domain of the protein. The *AMOTL1* mutation is in the Angiomotin domain. The *GRIA3* mutation is in the ANF receptor ligand domain. Both *OR13C2* mutations fall within the transmembrane domain of the protein. Finally, the *USP20* mutation is in the ubiquitin carboxyl-terminal hydrolase domain.

Notably, within the COSMIC database v77, the same mutations of the *ACOT1*, *GOLGA6L1*, *MUC4*, *PABPC3*, *RETSAT and ZNF28* genes identified in our analysis were present, although not in breast cancer tissue. See Table 8 for detailed COSMIC annotations of these mutations.

**Table 8. COSMIC annotation of candidate CSC-specific mutations.**

| Mutation Name | Tumour Tissue | COSMIC ID |
|---|---|---|
| *ACOT1* | Prostate | COSM5423118 |
| *GOLGA6L1* | Large Intestine<br>Pancreas<br>Upper Aerodigestive Tract | COSM4443866 |
| *MUC4*_1 | Biliary Tract<br>Large Intestine<br>Prostate | COSM5463888 |
| *MUC4*_2 | Bile Duct<br>Head And Neck | COSM4322884 |
| *MUC4*_3 | Biliary Tract<br>Central Nervous System<br>Cervix<br>Eye<br>Hematopoietic And Lymphoid Tissue<br>Kidney<br>Large Intestine<br>Liver<br>Prostate<br>Upper Aerodigestive Tract | COSM1131502 |
| *PABPC3* | Kidney<br>Large Intestine<br>Lung<br>Upper Aerodigestive Tract | COSM3773669 |
| *RETSAT*_1 | Biliary Tract<br>Liver<br>Thyroid | COSM3746549 |
| *RETSAT*_2 | Biliary Tract<br>Haematopoietic And Lymphoid Tissue<br>Liver<br>Thyroid | COSM3746550 |
| *ZNF28* | Large Intestine<br>Liver<br>Pancreas | COSM4286443 |

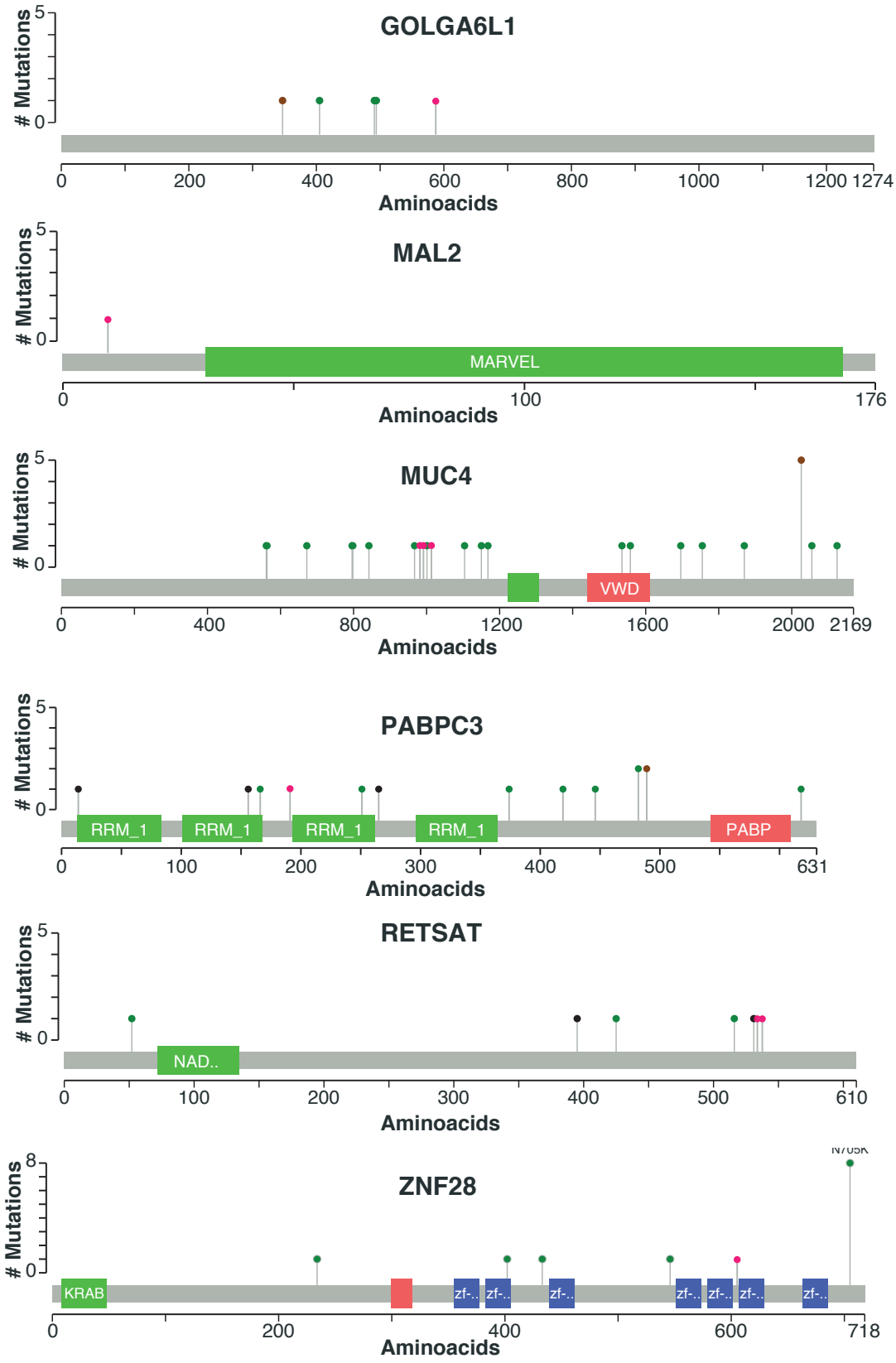


**Figure 19. Positions of mutations relative to protein domains (Part1).** Schematic structures of proteins encoded by candidate CSC-specific genes identified by Illumina WES (*GOLGA6L1*, *MAL2*, *MUC4 PABPC3*, *RETSAT* and *ZNF28*). The positions of known protein domains are indicated by coloured rectangles and the positions of mutations identified in our study or listed in the TCGA breast invasive carcinoma dataset are shown by dots. Pink dots represent the mutation identified in this study. Green dot represents missense mutations. Black dots represent nonsense mutations. Brown dots represent in-frame insertions. Y-axis: number of mutations present into TCGA breast invasive carcinoma dataset. X-axis: amino acid positions. Images were adapted from cBioPortal.

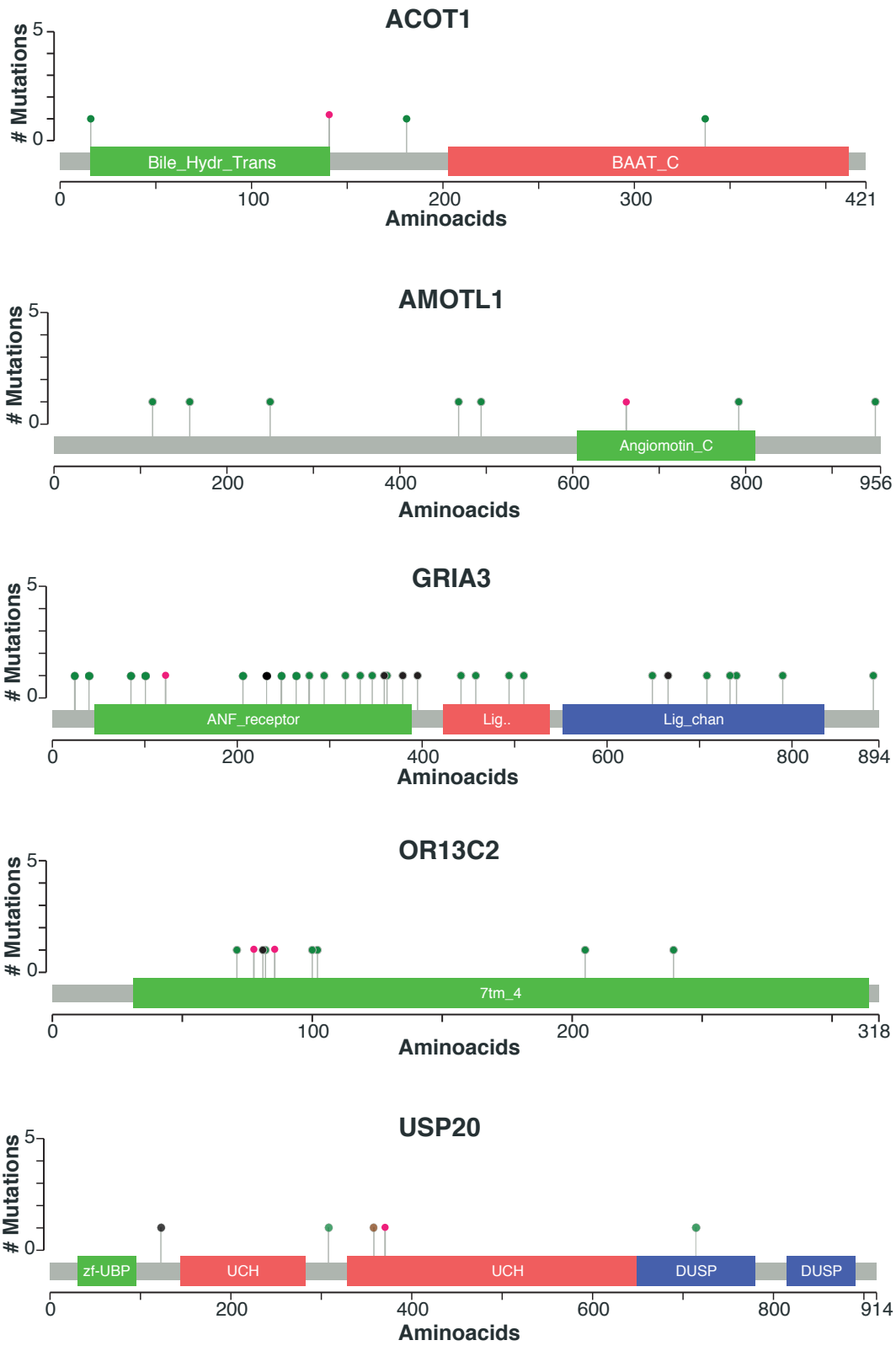


**Figure 20. Positions of mutations relative to protein domains (Part2).** Schematic structures of proteins encoded by candidate CSC-specific genes identified by Illumina WES (*ACOT1*, *AMOTL1*, *GRIA3*, *OR13C2* and *USP20*). The positions of known protein domains are indicated by coloured rectangles and the positions of mutations identified in our study or listed in the TCGA breast invasive carcinoma dataset are shown by dots. Pink dots represent the mutation identified in this study. Green dot represents missense mutations. Black dots represent nonsense mutations. Brown dots represent in-frame insertions. Y-axis: number of mutations present into TCGA breast invasive carcinoma dataset. X-axis: amino acid positions. Images were adapted from cBioPortal.

To understand if the mutations have a functional impact on the expressed proteins, we used several website tools for prediction (the results of the predictive tools we used are reported in Table 9), including:

- PROVEAN, which is based on two main steps. In the first one, PROVEAN collects a set of homologous and distantly related sequences and clusters them by sequence identity to obtain a supporting sequence set with the most similar sequence to query. In the second step, a delta score (named also PROVEAN score), which is based on the similarity matrix BLOSUM62, is calculated for each sequence of the supporting set. If the PROVEAN score is smaller than or equal to a given threshold, the variation is predicted as deleterious.
- SIFT that identifies homologous functional proteins, performs multiple sequence alignment and calculates the probability for all 20 amino acids at query position. Hence this probability is normalized by the most frequent amino acid to obtain a scaled probability, which is used to classify the damaging potential of the substitution.
- FATHMM, which is based on 10 feature groups to annotate variants from HGMD (pathogenic dataset) and 1000 Genome Project (control dataset). These 10 groups include 46-way sequence conservation, histone modifications, transcription factor binding, open chromatin (based on DNase-Seq), 100-way sequence, GC-content, open chromatin (based on FAIRE), transcription factor binding sites, genome segmentation and footprints. FATHMM then integrates all these features using a classifier based on multiple kernel learning (MKL) to score functional impact of all variants.
- PolyPhen2, which extracts different sequence- and structure-based features to be subjected to a probabilistic classifier, based on machine-learning method. Sequence-based features encompass the occurrence of the substitution within a specific site and the conservation of the substituted nucleotide in a family of

homologous proteins. Structure-based parameters, which is related to the mapping of the variant amino acid either on known 3D structure or on 3D structure of homologous proteins, include accessible surface area, hydrophobic propensity and B-factor (index of atom mobility).

- Mutation Assessor , which is based on evolutionary conservation of the affected amino acid in protein homologs. Mutation Assessor uses a multiple sequence alignment to calculate a functional score. This score is a combination of conservation score (conserved residues across the entire family) and specificity score (conserved residues within each subfamily, but vary between subfamilies).
- CADD that integrates multiple annotations into one metric to all variants that survived natural selection (present in the 1000 Genome Project catalogue; The 1000 Genomes Project Consortium, 2012) and to simulated mutations. To simulate SNVs and INDELs, CADD applies a genome-wide simulator of *de novo* variations. Then, CADD annotates each variant using the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2010), containing information taken from i.e. GERP (Cooper et al., 2005), phastCons (Siepel et al., 2005), Grantham (Grantham, 1974), SIFT (Kumar et al., 2009) and PolyPhen (Adzhubei et al., 2013).

From Table 9, it is clear that it was not always possible to achieve a consensus among all the six predictors. Of note, FATHMM and CADD report a score of deleteriousness, and in particular, FATHMM considers as damaging all those mutations with a score >0.5. Although CADD developers do not indicate a defined score to assess the deleterious impact of the mutations, they suggest using 15 as a threshold.

The mutation that consistently was predicted to have deleterious impact on protein structure and function was the *RETSAT*_1 mutation, which scored as damaging by all tools, except for Mutation Assessor, which predicted a medium impact. Instead, several mutations were predicted to be damaging by at least 4 tools, including *RETSAT*_2,

*OR13C2* (LIST2) and *ZNF28*. While other mutations were predicted to be damaging by less than 4 tools, e.g. *MAL2*, *GRIA3*, *MUC4*_2, *OR13C2* (LIST1) and *USP20*. In contrast, the *ACOT1*, *AMOTL1*, *GOLGA6L1*, *MUC4_1*, *MUC4_3* and *PABPC3* mutations were predicted not to have a deleterious impact on the protein structure and function according to all six tools, and thus their possible impact on protein structure and function is still unclear.

In an attempt to assess the relevance of the candidate CSC-specific mutations to cancer in general and, in particular, to breast cancer, we performed a literature search. For the *ACOT1*, *GOLGA6L1*, *OR13C2*, *PABPC3*, *RETSAT* and *ZNF28* genes no reports of an involvement in cancer were found in the literature. In contrast, for the *AMOTL1*, *GRIA3*, *MAL2*, *MUC4* and *USP20* genes a direct or indirect role in cancer has been previously described.

In particular, the *AMOTL1* (acyl-CoA thioesterase 1) gene encodes for a peripheral membrane protein that is a component of tight junctions. AMOTL1 mRNA has been identified as differentially expressed between ER- and ER+ breast cancer and AMOTL1 protein has been shown to participate in EMT and stimulate breast cancer cell proliferation by inducing Src activity (Couderc et al., 2016).

The *GRIA3* (glutamate ionotropic receptor AMPA type subunit 3) gene, whose protein product is a subunit of the glutamate receptor family, is a downstream target of *CUX1* (Cut Like Homeobox 1). *CUX1* effects on tumour proliferation, survival and migration are mediated by *GRIA3* in pancreatic cells *in vitro* and *in vivo* (Ripka et al., 2010).

The *MAL2* (mal, T-cell differentiation protein 2) gene encodes a multi-span transmembrane protein belonging to the MAL proteolipid family. MAL2 mRNA has been shown to be significantly upregulated in pancreatic metastatic cell line compared with parent non-metastatic cells (Eguchi et al., 2013), and MAL2 protein is a heterologous

partner for TPD52-like proteins (Wilson et al., 2001), which are overexpressed in breast cancer (Shehata et al., 2008).

The *MUC4* (mucin 4) gene encodes an integral membrane glycoprotein that acts as a ligand for ERBB2 and has been shown to mask the Trastuzumab binding site, inducing Trastuzumab resistance both *in vitro* and *in vivo* (Elster et al., 2015). Reduced MUC4 protein expression in invasive breast carcinoma is involved in tumour progression and correlates with increased tumour-infiltrating immune cells and promotes hyper-methylation (Cho et al., 2015).

The *USP20* (ubiquitin specific peptidase 20, also called *VDU2*) gene encodes an ubiquitin specific protease that deubiquitinates hypoxia-inducible factor (HIF)-1 alpha causing an increased expression of HIF1α target genes and enhancing HIF1α mediated activities (i.e. angiogenesis, cell proliferation and metastasis) (Li et al., 2005).

**Table 9. Functional impact prediction of candidate CSC-specific mutations.**

| Mutation Name | PROVEAN | SIFT | FATHMM | PolyPhen2 | Mutation Assessor | CADD |
|---|---|---|---|---|---|---|
| *OR13C2* (LIST1*)* | N | T | 0.13718 | Poss Dam | Low | 11.75 |
| *ACOT1* | N | T | 0.22913 | B | Low | 13.61 |
| *AMOTL1* | N | T | 0.38461 | B | Low | 5.04 |
| *GOLGA6L1* | N | T | 0.00072 | B | Neutral | 0.07 |
| *GRIA3* | Del | Dam | 0.02938 | B | NA | 2.30 |
| *MAL2* | NA | NA | 0.82925 | Poss Dam | NA | 21.50 |
| *MUC4*_1 | N | T | 0.00119 | B | NA | 0.00 |
| *MUC4*_2 | N | Dam | 0.00475 | Poss Dam | NA | 0.27 |
| *MUC4*_3 | N | T | 0.00106 | B | NA | 3.07 |
| *OR13C2* (LIST2) | Del | Dam | 0.94326 | B | Medium | 11.63 |
| *PABPC3* | N | T | 0.05858 | B | Neutral | 0.94 |
| *RETSAT*_1 | Del | Dam | 0.85171 | Prob Dam | Medium | 29.50 |
| *RETSAT*_2 | N | T | 0.84738 | Poss Dam | Medium | 22.70 |
| *USP20* | N | T | 0.88918 | B | Neutral | 4.79 |
| *ZNF28* | Del | Dam | 0.03656 | Prob Dam | Low | 23.80 |

Six prediction tools (PROVEAN, SIFT, FATHMM, PolyPhen2, Mutation Assessor and CADD) were used to predict the functional consequences of the candidate CSC-specific mutations. N = Neutral; T = Tolerated; B= Benign; Del = Deleterious; Dam = Damaging; Poss Dam = Possibly Damaging; Prob Dam = Probably Damaging. For FATHMM and CADD, values >0.5 and >15 represent deleterious mutations, respectively. NA = result not available, in particular the prediction tool Mutation Assessor was not able to retrieve a protein structure for the *MUC4* gene and the *GRIA3* gene.

# Discussion

Over the past few years, our knowledge on the on the clinical and pathological heterogeneity of human breast cancers has considerably expanded. On the one hand, major molecular subtypes of breast cancer (Luminal-A, Luminal-B, Basal-like, HER2) have been recognized, which have become critical for addressing inter-tumour heterogeneity and for informing patient management in the clinical practice (Perou et al., 2000; Sorlie et al., 2001). On the other, parallel sequencing studies have provided evidence that, like other solid tumours, breast cancers can also show varying degrees of intra-tumour heterogeneity, with the presence of subpopulations of cells with profound genomic and phenotypic differences (Cottu et al., 2008; Shah et al., 2009; Navin et al., 2010; Navin et al., 2011; Shah et al., 2012; Nik-Zainal et al., 2012). Remarkably, these differences can be detected either in the context of the bulk primary tumour and between the primary tumour and its metastasis (Torres et al., 2007; Ding et al., 2010). It is also becoming increasingly evident that the two models historically invoked to explain intra-tumour heterogeneity, i.e. the clonal evolution and the CSC model (Reya et al., 2001; Shackleton et al., 2009; Visvader, 2011; Meacham and Morrison, 2013), are not necessarily mutually exclusive, at least for those tumours characterized by high levels of genomic instability (Torres et al., 2007; Kreso and Dick, 2014; Meacham and Morrison, 2013). In the clonal evolution model, the clone with the highest degree of fitness, due to the stochastic occurrence of genetic and epigenetic alterations, is selected for and outcompetes other possible clones in a spatial and temporal fashion according to a Darwinian evolutionary process (Merlo et al., 2006; Gerlinger et al., 2012; Bedard et al., 2013). This scenario also applies to the acquisition of therapy resistance and metastatic ability by tumour cells (Marusyk and Polyak, 2010; Greaves and Maley, 2012). Therefore, in the clonal evolution model, tumour eradication

can be achieved only when the entire subset of cells showing similar genetic make-up and phenotypical behaviour are targeted by therapies.

In contrast, the CSC model holds that tumours, like normal tissues, are hierarchically organized and that, therefore, the nature of intratumoral heterogeneity is largely phenotypic. The hierarchical organization presupposes the existence of CSCs, either derived from malignant transformation of normal stem cells or from progenitors that have re-acquired stemness traits, which are responsible for the tumorigenic process, while giving origin to a progeny of non-tumorigenic and phenotypically heterogeneous progenitors (Reya et al., 2001; Dick, 2008; Clevers, 2011; Beck and Blanpain, 2013). From a therapeutic point of view, a corollary of this model is that tumour eradication can be achieved only when CSCs are eliminated (Creighton et al., 2009; Diehn et al., 2009; Bedard et al., 2013). As mentioned above, the clonal evolution and the CSC theory are not necessarily mutually exclusive, considering the possibility that, in tumours with high genetic instability, CSCs may also contribute to intra-tumour heterogeneity with the presence of CSC clones that are genetically heterogeneous (Cottu et al., 2008; Shah et al., 2009; Navin et al., 2010; Navin et al., 2011; Shah et al., 2012; Nik-Zainal et al., 2012). The co-existence of genetically different CSC clones would, for instance, account for the co-existence of multiple genetically distinct lineages of tumour cells typical of poly-genomic breast cancers (Navin et al., 2011; Shah et al., 2012).

One important implication of the CSC model is that mutations putatively involved in the emergence and expansion of CSCs might be present at very low frequency and, especially in tumours with a high degree of genomic instability, “diluted” in the wide heterogeneous landscape of mutations progressively accumulated in the bulk progenitor population during the proliferative history of the tumour. In more general terms, distinguishing the true “driver” mutations of the tumorigenic process from the “passenger” mutations that represent a by-product of tumour development constitutes a major challenge in massive parallel sequencing studies. In this regard, recent studies have demonstrated

that it is possible to resolve the initial heterogeneity of the primary tumour by xenografting it into an immunocompromised host. In the original study from Ding *et al.* (Ding et al., 2010), this approach resulted in the identification of mutations expressed at very low or even undetectable levels in the primary tumour, which were enriched in the mutational profile of the xenograft and also of the distant metastasis from the same breast cancer patient. These findings argue that in the relatively short proliferative history of the xenograft and metastatic outgrowths, it is possible to identify mutations with a likely driver role in the tumorigenic process.

Based on these premises, we decided to perform a comparative analysis of a primary breast tumour and of its matched mammosphere and xenograft samples, based on the dual assumption that: i) the ability to generate mammospheres and to drive xenograft outgrowth resides in sub-clones of cells characterised by a unique tumorigenic ability; ii) the short proliferative history required for the formation of mammospheres and of the PDX should minimize the confounding effect of passenger lesions appearing as by-products of the intrinsic genomic instability of tumour cells. We therefore used NGS data derived from the analysis of the primary tumour and its matched mammosphere and PDX samples to identify a subset of mutations either shared or enriched in mammospheres and in the xenograft, which might represent driver mutations likely belonging to the profile of tumour cells with an intrinsic stemness ability.

## 1. Optimisation of the WES Protocol

Before testing our hypothesis, we set out to optimise the protocol for the application of deep sequencing to samples composed of very few cells. To this end, we tested whether the DNA amplification step, the WGA technique, required to obtain sufficient DNA material from small amount of starting DNA, introduced any bias in the WES profiles. Therefore, we performed WES on gDNA and WGA-DNA from the same tumour test sample. After

the comparison of WES alignment and mutational profiles of gDNA and WGA-DNA, we established that the WGA did not introduce artefacts within the WES experiments. In particular, we obtained an overlap of variants of ~80% and comparable alignment statistics. Hence, we applied the WGA-WES strategy to our experimental samples: a primary breast tumour and matched primary tumour-derived mammospheres, composed mostly of early progenitors that are representative of the CSC compartment, PDX, PDX-derived mammospheres and blood samples. Starting from a limited amount of gDNA (approximately 10 ng/µl) and applying the WGA step to all experimental samples, we were able to perform a NGS analysis. This allowed us to compare the mutational profile of primary tumour-derived mammospheres with that of the primary tumour and also of the PDX and PDX-derived mammospheres.

Furthermore, we have also implemented a bioinformatics strategy (based on the Xenome software), which allowed us to successfully eliminate mouse contamination within PDX-derived samples (PDX and PDX-derived mammospheres). This was necessary to reduce false positives when calling mutations and to obtain a reliable genetic profile of the PDX-derived samples.

## 2. Identification of Candidate Founder Mutations

By comparing the mutational profiles of primary tumour, primary tumour-derived mammospheres and PDX, we identified 11 mutations that were present in primary tumour, primary tumour-derived mammospheres and PDX, with a comparable mutation frequency among all analysed samples. We reasoned that this set of mutations most likely represents a cluster of founder genetic lesions required for the onset and progression of tumorigenesis. Supporting this idea, we found that 4 of these 11 mutations (*AKNA*, *ANKRD55*, *GALNT15* and *PABPC3*_3 mutations, LIST3 in the Results) were also present in PDX-derived mammospheres, indicating a strong selective pressure towards the

maintenance of these mutations as a part of the genetic make-up of tumour sub-clones with high tumorigenic potential. We also noted that two mutations, *ANKRD55* and *GALNT15*, are present in all samples at high frequency, further arguing for their role as "founder" (or "truncal") mutations (De Grassi et al., 2014). These results indicate that the bulk tumour cells, due to their intrinsic genetic instability, may acquire distinct genetic profiles during tumour formation and that shortening the proliferative history of the tumorigenic process, as recapitulated in the xenograft or mammosphere setting, is instrumental to identify mutations that are likely to be the real "drivers" of tumour growth.

Of note, the 4 mutated genes carrying the 4 mutations common to all samples have never been previously described as relevant to cancer, with the exception of the *AKNA* gene, which has been described to have a function in cervical cancer (Martinez-Nava et al., 2015) and leukaemia (Chen et al., 2015).

### 3. Identification of Candidate CSC-specific Mutations

From the comparative analysis of the primary tumour and mammospheres, it was possible to identify 14 mutations that were enriched in primary tumour-derived mammospheres, compared to the primary tumour (mutations present within LIST1 in the Results). Of note, one of these mutations (*OR13C2*, LIST1 in the Results) was also enriched in the PDX, while only the *GALNT15* mutation was present also in PDX-derived mammospheres. We also performed a comparative analysis of the mutational profile of primary tumour-derived mammospheres with that of the PDX. Remarkably, by this approach, we identified 14 overlapping mutations (LIST2 in the Results) that were not detectable within the bulk population of the primary tumour. Thus, direct NGS analysis of mammospheres could help to find a core-component of cancer-relevant genes that were not immediately identifiable in the mutational profile of the primary tumour. The most straightforward explanation is that the set of mutations appeared *de novo* in mammospheres or in the PDX compared to

the primary tumour likely belong to sub-clones of CSCs scarcely represented in the bulk tumour mass. These mutations would not be identified by NGS because of the detection limit of the procedure. Arguing for the actual relevance of some of these mutations to the ability of CSCs to drive tumorigenesis, we found that 5 of these 14 LIST2 mutations were maintained in PDX-mammospheres.

We were able to identify at least three mutated genes (*AMOTL1*, *MAL2* and *MUC4*) whose involvement in breast cancer were also described in other independent studies. In particular, *AMOTL1* gene was found upregulated in oestrogen receptor negative breast tumours and its protein level was negatively regulated by the direct interaction with tumour suppressor *NF2*. In addition, NF2 is able to induce AMOTL1 phosphorilation, which leads to AMOTL1 degradation in BC52 cells by increasing its binding to NEDD4 family of ubiquitin ligases (Couderc et al., 2016). In contrast, in normal endothelial cells, Choi and colleagues demonstrated that HECW2 E3 ligases was able to stabilize AMOTL1 through lysine 63-linked polyubiquitination and to increase its localization to junction area of human umbilical vein endothelial cells (Choi et al., 2016). Thus far, a complete and clear function of AMOTL1 ubiquitination is still missing and therefore a possible role of nucleotide mutations in the *AMOTL1* gene in its interactions with Merlin and ubiquitin ligases.

MAL2 protein is a heterologous partner of all three TPD52-like proteins (TPD52, TPD52L1 and TPD52L2) (Wilson et al,. 2001), and is overexpressed in metastatic pancreatic cell line compared to non-metastatic cells (Eguchi et al., 2013). In breast cancer *MAL2* and *TPD52*-like genes were also found overexpressed (Shehata et al., 2008) indicating a potential oncogenic role of these genes. Additionally, Li and colleagues (Li et al., 2016) demonstrated that miR-34a overexpression significantly inhibited the expression of TPD52 and block invasion/migration phenotypes in breast cancer cells. Therefore,

analysis of miR-34a and TPD52 expression in the context of *MAL2* mutations could eventually clarify the functional role of MAL2 in breast cancer.

Finally, overexpression of MUC4 in breast cancer promotes metastasis formation by the disruption of cell-cell and cell-extracellular matrix interactions, preventing cell anoikis (Workman et al., 2009), and by the sustained expression of EGFR family proteins and β-catenin (Mukhopadhyay et al., 2013). Thus, characterization of the functional roles of the mutated forms of *MUC4* gene in its overexpression may lead to the identification of its potential contribute to breast cancer formation and progression.

Lastly, although the *RETSAT* gene has not been reported so far to be implicated in cancer, all the six predictive tools we used in our analysis have predicted a damaging role of the mutation we identified, which warrants further investigation of an eventual cancer role for the *RETSAT* gene.

## 4. Limitations of Our Study Design

Although we obtained encouraging results, we are aware that our study may suffer from a number of limitations. Firstly, for obvious reasons, we could sequence only a fraction of the primary tumour from the patient, i.e. the amount of tissue available through the biopsy specimen that was destined to molecular studies. The possibility exists therefore that the molecular profile we have obtained from the biopsy specimen of the primary tumour might not be entirely representative of the entire collection of mutations of the bulk primary tumour mass. This scenario has been previously described in the work of Gerlinger and colleagues (Gerlinger et al., 2012) which showed the existence of a spatial intratumoral heterogeneity with the presence of private mutations present only in a few areas of the tumour in addition to mutations shared by most of the tumour areas analysed.

Another important point is that, if it is true that by the mutational profiling of mammospheres and PDX we have likely identified mutations belonging to sub-clones of cells endowed with stemness ability, it is difficult to predict whether these mutations are present in individual CSC clones or shared across more than one clone. In other words, if our approach revealed to be instrumental to deconvoluting tumour heterogeneity with the identification of mutations with a likely tumorigenic function, it does not allow understand whether multiple CSC clones co-exist in the bulk tumour mass, which appears to be the case of polygenomic tumours.

## 5. Conclusion

We developed an experimental and bioinformatics strategy to identify candidate cancer-relevant genes in breast cancer samples. We applied the WGA-WES protocol to matched primary tumour, mammosphere and PDX-derived samples. While the comparative analysis of a primary tumour with its matched xenograft has already been exploited as a strategy to deconvolute the heterogeneity of the primary tumour, the use of mammospheres derived from the primary tumour constitutes a novel approach allowing the identification of mutations with a likely relevance to the biology of CSC. The bulk of our results indicate the existence in the context of the primary tumour of sub-clones of cells that may harbour either shared and private somatic mutations likely at the basis of the masterplan of the tumorigenic process. Whether the shared and private mutations simply indicate the co-existence of several heterogeneous CSC clones or their hierarchical organization according to a spatial or temporal evolutionary tree remains to be elucidated.

## 6. Future plans

In the forthcoming future, we plan to extend the WGA-WES analysis of matched primary tumour, mammosphere and PDX samples to a set of additional breast tumours belonging to

different breast tumour subtypes. An important step in future studies will be the validation by targeted re-sequencing of the most relevant mutations found in our experimental model, using an alternative NGS platform (i.e. Ion Torrent or miSeq).

We also plan to perform high resolution studies on candidate mutated genes to assess their functional relevance to CSC biology and to tumour progression and metastasis. To this aim, we plan to use strategies to over-express or silence candidate mutated genes in relevant model systems. The genes to be subjected to functional validation studies will be selected based on results from a preliminary analysis of a sizable cohort of human breast cancers.

# References


Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics Chapter *7*, 1-41.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature Methods *7*, 248-249.

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., and Clarke, M.F. (2003). Prospective identification of tumorigenic breast cancer cells. Proceedings of the National Academy of Sciences of the United States of America *100*, 3983-3988.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L.*, et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

Altundag, K., and Ibrahim, N.K. (2006). Aromatase inhibitors in breast cancer: an overview. The Oncologist *11*, 553-562.

Alves, C.S., Burdick, M.M., Thomas, S.N., Pawar, P., and Konstantopoulos, K. (2008). The dual role of CD44 as a functional P-selectin ligand and fibrin receptor in colon carcinoma cell adhesion. American Journal of Physiology Cell Physiology *294*, C907-916.

Antoniou, A., Pharoah, P.D., Narod, S., Risch, H.A., Eyfjord, J.E., Hopper, J.L., Loman, N., Olsson, H., Johannsson, O., Borg, A.*, et al.* (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. American Journal of Human Genetics *72*, 1117-1130.

Balic, M., Lin, H., Young, L., Hawes, D., Giuliano, A., McNamara, G., Datar, R.H., and Cote, R.J. (2006). Most early disseminated cancer cells detected in bone marrow of breast cancer patients have a putative breast cancer stem cell phenotype. Clinical cancer research *12*, 5615-5621.

Barok, M., Joensuu, H., and Isola, J. (2014). Trastuzumab emtansine: mechanisms of action and drug resistance. Breast Cancer Research *16*, 1-12.

Beck B. and Blanpain C. (2013). Unravelling cancer stem cell potential. Nature Reviews Cancer *13*, 727-738.

Bedard P.L., Hansen A.R., Ratain M.J. and Siu L.L. (2013). Tumour heterogeneity in the clinic. Nature *501*, 355-364.

Ben-Porath, I., Thomson, M.W., Carey, V.J., Ge, R., Bell, G.W., Regev, A., and Weinberg, R.A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nature Genetics *40*, 499-507.

Bhoumik, A., Takahashi, S., Breitweiser, W., Shiloh, Y., Jones, N., and Ronai, Z. (2005). ATM-dependent phosphorylation of ATF2 is required for the DNA damage response. Molecular Cell *18*, 577-587.

Bonnet, D., and Dick, J. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nature Medicine *3*, 730-737.

Bose, S., Wang, S.I., Terry, M.B., Hibshoosh, H., and Parsons, R. (1998). Allelic loss of chromosome 10q23 is associated with tumor progression in breast carcinomas. Oncogene *17*, 123-127.

Braun, S., Vogl, F.D., Naume, B., Janni, W., Osborne, M.P., Coombes, R.C., Schlimok, G.n., Diel, I.J., Gerber, B., Gebauer, G.*, et al.* (2005). A pooled analysis of bone marrow micrometastasis in breast cancer. The New England Journal of Medicine *353*, 793-802.

Cadoo, K.A., Gucalp, A., and Traina, T.A. (2014). Palbociclib: an evidence-based review of its potential in the treatment of breast cancer. Breast Cancer *6*, 123-133.

Cadwell, R.C., and Joyce, G.F. (1992). Randomization of genes by PCR mutagenesis. PCR Methods and Applications *2*, 28-33.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Cancer Genome Atlas Research, N. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543-550.

Cancer Genome Atlas Research, N., Linehan, W.M., Spellman, P.T., Ricketts, C.J., Creighton, C.J., Fei, S.S., Davis, C., Wheeler, D.A., Murray, B.A., Schmidt, L.*, et al.* (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. The New England Journal of Medicine *374*, 135-145.

Cantor, S.B., Bell, D.W., Ganesan, S., Kass, E.M., Drapkin, R., Grossman, S., Wahrer, D.C.R., Sgroi, D.C., Lane, W.S., Daniel A. Haber*, et al.* (2001). BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. Cell *105*, 149-160.

Carpenter, B.L., Chen, M., Knifley, T., Davis, K.A., Harrison, S.M., Stewart, R.L., and O'Connor, K.L. (2015). Integrin alpha6beta4 promotes autocrine Epidermal Growth Factor Receptor (EGFR) signaling to stimulate migration and invasion toward Hepatocyte Growth Factor (HGF). The Journal of Biological Chemistry *290*, 27228-27238.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E.*, et al.* (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery *2*, 401-404.

Chan, K.Y.K., Ozcelik, H., Cheung, A.N.Y., Ngan, H.Y.S., and Khoo, U.-S. (2002). Epigenetic factors controlling the BRCA1 and BRCA2 genes in sporadic ovarian cancer. Cancer Research *62*, 4151-4156.

Chen, C., Bartenhagen, C., Gombert, M., Okpanyi, V., Binder, V., Rottgers, S., Bradtke, J., Teigler-Schlegel, A., Harbott, J., Ginzel, S.*, et al.* (2015). Next-generation-sequencing of recurrent childhood high hyperdiploid acute lymphoblastic leukemia reveals mutations typically associated with high risk patients. Leukemia Research *39*, 990-1001.

Chen, S., and Parmigiani, G. (2007). Meta-analysis of BRCA1 and BRCA2 penetrance. Journal of Clinical Oncology *25*, 1329-1333.

Cho, J.S., Park, M.H., Lee, J.S., and Yoon, J.H. (2015). Reduced MUC4 expression is a late event in breast carcinogenesis and is correlated with increased infiltration of immune cells as well as promoter hypermethylation in invasive breast carcinoma. Applied Immunohistochemistry and Molecular Morphology *23*, 44-53.

Choi, K.-S., Choi, H.-J., Lee, J.-K., Ima, S., Zhang, H., Jeong, Y. *et al*. (2016) The endothelial E3 ligase HECW2 promotes endothelial cell junctions by increasing AMOTL1 protein stability via K63-linked ubiquitination. Cellular Signaling *28*, 1642-1651.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PloS one *7*, e46688.

Clarke, M.F., and Fuller, M. (2006). Stem cells and cancer: two faces of eve. Cell *124*, 1111-1115.

Clevers H. (2011). The cancer stem cell: premises, promises and challenges. Nature Medicine *17*, 313-319.

Coates, A.S., Winer, E.P., Goldhirsch, A., Gelber, R.D., Gnant, M., Piccart-Gebhart, M., Thurlimann, B., Senn, H.J., and Panel, M. (2015). Tailoring therapies--improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Annals of Oncology *26*, 1533-1546.

Cooper, G M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S. and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Research *15*, 901-913.

Conway, T., Wazny, J., Bromage, A., Tymms, M., Sooraj, D., Williams, E.D., and Beresford-Smith, B. (2012). Xenome--a tool for classifying reads from xenograft samples. Bioinformatics *28*, i172-178.

Cottu P.H., Asselah J., Lae M., Pierga J.Y., Diéras V., Mignot L., Sigal-Zafrani B., Vincent-Salomon A.(2008). Intratumoral heterogeneity of HER2/neu expression and its consequences for the management of advanced breast cancer. Annals of Oncology *19*, 595-597.

Couderc, C., Boin, A., Fuhrmann, L., Vincent-Salomon, A., Mandati, V., Kieffer, Y., Mechta-Grigoriou, F., Del Maestro, L., Chavrier, P., Vallerand, D.*, et al.* (2016). AMOTL1 promotes breast cancer progression and is antagonized by Merlin. Neoplasia *18*, 10-24.

Creighton C.J., Li X., Landis M., Dixon J.M., Neumeister V.M., Sjolund A., *et al.* (2009). Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. Proceedings of the National Academy of Sciences of the United States of America *106*, 13820-13825.

Dang, H., Steinway, S.N., Ding, W., and Rountree, C.B. (2015). Induction of tumor initiation is dependent on CD44s in c-Met$^{+}$ hepatocellular carcinoma. BMC Cancer *15*, 161-172.

De Grassi A., Iannelli F., Cereda M., Volorio S., Melocchi V., Viel A., *et al.* (2014). Deep sequencing of the X chromosome reveals the proliferation history of colorectal adenomas. Genome Biology *15*, 437-454.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R.*, et al.* (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Research *22*, 1589-1598.

Dick J.E. (2008). Stem cell concepts renew cancer research. Blood *112*, 4793-4807.

Diehn M., Cho R.W., Lobo N.A., Kalisky T., Dorie M.J., Kulp A.N., *et al.* (2009). Association of reactive oxygen species levels and radioresistance in cancer stem cells. Nature 458, 780-783.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L.*, et al.* (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999-1005.

Doerks, T., Copley, R.R., Schultz, J., Ponting, C.P., and Bork, P. (2002). Systematic identification of novel protein domain families associated with nuclear functions. Genome Research *12*, 47-56.

Dontu, G., Abdallah, W.M., Foley, J.M., Jackson, K.W., Clarke, M.F., Kawamura, M.J., and Wicha, M.S. (2003). In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. Genes & Development *17*, 1253-1270.

EBCTCG, E.B.C.T.C.G.-. (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet *365*, 1687-1717.

Eguchi, D., Ohuchida, K., Kozono, S., Ikenaga, N., Shindo, K., Cui, L., Fujiwara, K., Akagawa, S., Ohtsuka, T., Takahata, S.*, et al.* (2013). MAL2 expression predicts distant metastasis and short survival in pancreatic cancer. Surgery *154*, 573-582.

Ellis, L.M. (2006). Mechanisms of action of bevacizumab as a component of therapy for metastatic colorectal cancer. Seminars in Oncology *33*, S1-7.

Elster, N., Collins, D.M., Toomey, S., Crown, J., Eustace, A.J., and Hennessy, B.T. (2015). HER2-family signalling mechanisms, clinical implications and targeting in breast cancer. Breast Cancer Research and Treatment *149*, 5-15.

Elston, C.W., and Ellis, I.O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology *19*, 403-410.

Fauzee, N.J.S., Dong, Z., and Wang, Y.-l. (2011). Taxanes: promising anti-cancer drugs. Asian Pacific Journal of Cancer Prevention *12*, 837-851.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer *136*, E359-386.

Fidler, I.J. (1978). Tumor heterogeneity and the biology of cancer invasion and metastasis. Cancer Research *38*, 2651-2660.

Fillmore, C.M., and Kuperwasser, C. (2008). Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. Breast Cancer Research *10*, R25-38.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S.*, et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research *43*, D805-811.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nature Reviews Cancer *4*, 177-183.

Gao, J., Aksoy, B.l.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E.*, et al.* (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling *6*, 2-19.

Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. Breast Cancer Research *4*, 70-76.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P.*, et al.* (2012). Intratumor

heterogeneity and branched evolution revealed by multiregion sequencing. The New England Journal of Medicine *336*, 883-892.

Gill, P.G., Farshid, G., Luke, C.G., and Roder, D.M. (2004). Detection by screening mammography is a powerful independent predictor of survival in women diagnosed with breast cancer. The Breast *13*, 15-22.

Ginestier, C., Hur, M.H., Charafe-Jauffret, E., Monville, F., Dutcher, J., Brown, M., Jacquemier, J., Viens, P., Kleer, C.G., Liu, S.*, et al.* (2007). ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. Cell Stem Cell *1*, 555-567.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. Science *185*, 862-864.

Greaves, M. and Maley, C.C. (2012). Clonal evolution in cancer. Nature 481, 306-313.

GrØnbÆk, K., Hother, C., and Jones, P.A. (2007). Epigenetic changes in cancer. APMIS *115*, 1039-1059.

Hasmats, J., Green, H., Orear, C., Validire, P., Huss, M., Kaller, M., and Lundeberg, J. (2014). Assessment of whole genome amplification for sequence capture and massively parallel sequencing. PloS one *9*, e84785.

Herman, J.G., Merlo, A., Mao, L., Lapidus, R.G., Issa, J.-P.J., Davidson, N.E., Sidransky, D., and Baylin, S.B. (1995). Inactivation of the CDKN2/p16/MTSJ gene is frequently associated with aberrant DNA methylation in all common human cancers. Cancer Research *55*, 4525-4530.

Hinshelwood, R.A., and Clark, S.J. (2008). Breast cancer epigenetics: normal human mammary epithelial cells as a model system. Journal of Molecular Medicine *86*, 1315-1328.

Hollier, B.G., Evans, K., and Mani, S.A. (2009). The epithelial-to-mesenchymal transition and cancer stem cells: a coalition against cancer therapies. Journal of Mammary Gland Biology and Neoplasia *14*, 29-43.

Houghton, P.J. (2010). Everolimus. Clinical cancer research *16*, 1368-1372.

Joensuu, H., Lehtimaki, T., Holli, K., Elomaa, L., Turpeenniemi-Hujanen, T., Kataja, V., Anttila, A., Lundin, M., Isola, J., and Lundin, J. (2004). Risk for distant recurrence of breast cancer detected by mammography screening or other methods. JAMA *292*, 1064-1073.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. American Journal of Human Genetics *91*, 839-848.

Jung, H., Bleazard, T., Lee, J., and Hong, D. (2013). Systematic investigation of cancer-associated somatic point mutations in SNP databases. Nature Biotechnology *31*, 787-789.

Kalluri, R., and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. The Journal of Clinical Investigation *119*, 1420-1428.

Kim, S.J., Kang, H.-S., Chang, H.L., Jung, Y.C., Sim, H.-B., Lee, K.S., Ro, J., and Lee, E.S. (2008). Promoter hypomethylation of the N-acetyltransferase 1 gene in breast cancer. Oncology Report *19*, 663-668.

King, M.-C., Marks, J.H., and Mandell, J.B. (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science *302*, 643-646.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics *46*, 310-315.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research *22*, 568-576.

Kreso A. and Dick J.E. (2014). Evolution of the cancer stem cell model. Cell Stem Cell *14*, 275-291.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols *4*, 1073-1081.

Leiserson, M.D., Wu, H.T., Vandin, F., and Raphael, B.J. (2015). CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biology *16*, 160-180.

Li, G., Yao, L., Zhang, J., Li, X., Dang, S., Zeng, K. *et al.* (2016). Tumor-suppressive microRNA-34a inhibits breast cancer cell migration and invasion via targeting oncogenic TPD52. Tumor Biology *37*, 7481-7491.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, S.Y., Rong, M., Grieu, F., and Iacopetta, B. (2006). PIK3CA mutations in breast cancer are associated with poor outcome. Breast Cancer Research and Treatment *96*, 91-95.

Li, Z., Wang, D., Messing, E.M., and Wu, G. (2005). VHL protein-interacting deubiquitinating enzyme 2 deubiquitinates and stabilizes HIF-1alpha. EMBO Reports *6*, 373-378.

Liang, S., Furuhashi, M., Nakane, R., Nakazawa, S., Goudarzi, H., Hamada, J., and Iizasa, H. (2013). Isolation and characterization of human breast cancer cells with SOX2 promoter activity. Biochemical and Biophysical Research Communications *437*, 205-211.

Lim, E., Vaillant, F., Wu, D., Forrest, N.C., Pal, B., Hart, A.H., Asselin-Labat, M.L., Gyorki, D.E., Ward, T., Partanen, A.*, et al.* (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nature Medicine *15*, 907-913.

Loden, M., Stighall, M., Nielsen, N.H., Roos, G., Emdin, S.O., Ostlund, H., and Landberg, G. (2002). The cyclin D1 high and cyclin E high subgroups of breast cancer: separate pathways in tumorogenesis based on pattern of genetic aberrations and inactivation of the pRb node. Oncogene *21*, 4680-4690.

Lu, Y., Lin, Y.-Z., LaPushin, R., Cuevas, B., Fang, X., Yu, S.X., Davies, M.A., Khan, H., Furui, T., Mao, M.*, et al.* (1999). The PTEN/MMAC1/TEP tumor suppressor gene decreases cell growth and induces apoptosis and anoikis in breast cancer cells. Oncogene *18*, 7034-7045.

Ma, X.J., Salunga, R., Tuggle, J.T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B.M.*, et al.* (2003). Gene expression profiles of human breast cancer progression. Proceedings of the National Academy of Sciences of the United States of America *100*, 5974-5979.

MacGregor, J.I., and Jordan, V.C. (1998). Basic guide to the mechanisms of antiestrogen action. Pharmacological Reviews *50*, 151-196.

Mani, S.A., Guo, W., Liao, M.J., Eaton, E.N., Ayyanan, A., Zhou, A.Y., Brooks, M., Reinhard, F., Zhang, C.C., Shipitsin, M.*, et al.* (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell *133*, 704-715.

Martinez-Nava, G.A., Torres-Poveda, K., Lagunas-Martinez, A., Bahena-Roman, M., Zurita-Diaz, M.A., Ortiz-Flores, E., Garcia-Carranca, A., Madrid-Marina, V., and Burguete-Garcia, A.I. (2015). Cervical cancer-associated promoter polymorphism affects akna expression levels. Genes and Immunity *16*, 43-53.

Marusyk A. and Polyak K. (2010). Tumor heterogeneity: causes and consequences. Biochimica Biophysica Acta *1805*, 105-117.

McFarlane, S., Coulter, J.A., Tibbits, P., O'Grady, A., McFarlane, C., Montgomery, N., Hill, A., McCarthy, H.O., Young, L.S., Kay, E.W.*, et al.* (2015). CD44 increases the efficiency of distant metastasis of breast cancer. Oncotarget *6*, 11465-11476.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069-2070.

Meacham C.E. and Morrison S.J. (2013). Tumour heterogeneity and cancer cell plasticity. Nature 501, 328-337.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M.*, et al.* (2002). Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nature Genetics *31*, 55-59.

Merlo L.M., Pepper J.W., Reid B.J. and Maley C.C. (2006). Cancer as an evolutionary and ecological process. Nature Reviews Cancer *6*, 924-935.

Minotti, G., Menna, P., Salvatorelli, E., Cairo, G., and Gianni, L. (2004). Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. Pharmacological Reviews *56*, 185-229.

Moinfar, F., Man, Y.G., Arnould, L., Bratthauer, G.L., Ratschek, M., and Tavassoli, F.A. (2000). Concurrent and independent genetic alterations in the stromal and epithelial cells of mammary carcinoma: implications for tumorigenesis. Cancer Research *60*, 2562-2566.

Morel, A.P., Lievre, M., Thomas, C., Hinkal, G., Ansieau, S., and Puisieux, A. (2008). Generation of breast cancer stem cells through epithelial-mesenchymal transition. PloS one *3*, e2888.

Moss, S.M., Cuckle, H., Evans, A., Johns, L., Waller, M., and Bobrow, L. (2006). Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. The Lancet *368*, 2053-2060.

Mukhopadhyay, P., Lakshmanan, I., Ponnusamy, M.P., Chakraborty, S., Jain, M., Pai, P. *et al.* (2013). MUC4 overexpression augments cell migration and metastasis through EGFR family proteins in triple negative breast cancer cells. PloS one *8*, e54455.

Nass, S.J., and Dickson, R.B. (1997). Defining a role for c-Myc in breast tumorigenesis. Breast Cancer Research and Treatment *44*, 1-22.

Navin N., Krasnitz A., Rodgers L., Cook K., Meth J., Kendall J., *et al.* (2010). Inferring tumor progression from genomic heterogeneity. Genome Research *20*, 68-80.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D.*, et al.* (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90-94.

Nguyen, A., Yoshida, M., Goodarzi, H., and Tavazoie, S.F. (2016). Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. Nature Communications *7*, 11246.

Nik-Zainal S., Davies H., Staaf J., Ramakrishna M., Glodzik D., Zou X., *et al.* (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47-54.

Nystrom, L., Bjurstam, N., Jonsson, H., Zackrisson, S., and Frisell, J. (2016). Reduced breast cancer mortality after 20+ years of follow-up in the Swedish randomized controlled mammography trials in Malmo, Stockholm, and Goteborg. Journal of Medical Screening 1-9.

O'Connell, P., Pekkel, V., Fuqua, S.A.W., Osborne, C.K., Clark, G.M., and Allred, D.C. (1998). Analysis of loss of heterozygosity in 399 premalignant breast lesions at 15 genetic loci. Journal of the National Cancer Institute *90*, 697-703.

Paredes, J., Albergaria, A., Oliveira, J.T., Jeronimo, C., Milanezi, F., and Schmitt, F.C. (2005). P-cadherin overexpression is an indicator of clinical outcome in invasive breast carcinomas and is associated with CDH3 promoter hypomethylation. Clinical Cancer Research *11*, 5869-5877.

Park, J.Y., Zhang, F., and Andreassen, P.R. (2014). PALB2: the hub of a network of tumor suppressors involved in DNA damage responses. Biochimica et Biophysica Acta *1846*, 263-275.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z.*, et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology *27*, 1160-1167.

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

Perou, C.M., Sùrlie, T., Eisen, M.B., Rijn, M.v.d., Jeffreyk, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A.*, et al.* (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752.

Pharoah, P., Day, N., and Caldas, C. (1999). Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. British Journal of Cancer *80*, 1968-1973.

Polgar, C., and Major, T. (2009). Current status and perspectives of brachytherapy for breast cancer. International Journal of Clinical Oncology *14*, 7-24.

Polyak, K. (2007). Breast cancer: origins and evolution. The Journal of Clinical Investigation *117*, 3155-3163.

Pon, J.R., and Marra, M.A. (2015). Driver and passenger mutations in cancer. Annual Review of Pathology *10*, 25-50.

Ponti, D., Costa, A., Zaffaroni, N., Pratesi, G., Petrangolini, G., Coradini, D., Pilotti, S., Marco A. Pierotti, and Daidone, M.G. (2005). Isolation and in vitro propagation of tumorigenic breast cancer cells with stem/progenitor cell properties. Cancer Research *65*, 5506-5511.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., *et al.* (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nature Genetics *39*, 165-167.

Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., North, B., Jayatilake, H., Barfoot, R., Spanova, K., *et al.* (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genetics *38*, 873-875.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Research *39*, e118-132.

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

Ripka, S., Riedel, J., Neesse, A., Griesmann, H., Buchholz, M., Ellenrieder, V., Moeller, F., Bartht, P., Gress, T.M., and Michl, P. (2010). Glutamate Receptor GRIA3—target of CUX1 and mediator of tumor progression in pancreatic cancer. Neoplasia *12*, 659-667.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature *29*, 24-26.

Rostoker, R., Abelson, S., Genkin, I., Ben-Shmuel, S., Sachidanandam, R., Scheinman, E.J., Bitton-Worms, K., Orr, Z.S., Caspi, A., Tzukerman, M., *et al.* (2015). $CD24^{+}$ cells fuel rapid tumor growth and display high metastatic capacity. Breast Cancer Research *17*, 78-92.

Roychoudhuri, R., Evans, H., Robinson, D., and Moller, H. (2004). Radiation-induced malignancies following radiotherapy for breast cancer. British Journal of Cancer *91*, 868-872.

Russnes, H.G., Navin, N., Hicks, J., and Borresen-Dale, A.L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. The Journal of Clinical Investigation *121*, 3810-3818.

Savage, K.I., Gorski, J.J., Barros, E.M., Irwin, G.W., Manti, L., Powell, A.J., Pellagatti, A., Lukashchuk, N., McCance, D.J., McCluggage, W.G.*, et al.* (2014). Identification of a BRCA1-mRNA splicing complex required for efficient DNA repair and maintenance of genomic stability. Molecular Cell *54*, 445-459.

Schuur, E.R., and DeAndrade, J.P. (2015). Breast cancer: molecular mechanisms, diagnosis, and treatment. International Manual of Oncology Practice *Chapter 9*, 155-200.

Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K.*, et al.* (2006). Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nature Genetics *38*, 1239-1241.

Shackleton M., Vaillant F., Simpson K.J., Stingl J., Smyth G.K., Asselin-Labat M.L., *et al.* (2006). Generation of a functional mammary gland from a single stem cell. Nature *439*, 84-88.

Shah S.P., Morin R.D., Khattra J., Prentice L., Pugh T., Burleigh A. , *et al.* (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809-813.

Shah S.P., Roth A., Goya R., Oloumi A., Ha G., Zhao Y., *et al.* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395-399.

Shahid, T., Soroka, J., Kong, E.H., Malivert, L., McIlwraith, M.J., Pape, T., West, S.C., and Zhang, X. (2014). Structure and mechanism of action of the BRCA2 breast cancer tumor suppressor. Nature Structural & Molecular Biology *21*, 962-968.

Shehata, M., Bieche, I., Boutros, R., Weidenhofer, J., Fanayan, S., Spalding, L., Zeps, N., Byth, K., Bright, R.K., Lidereau, R.*, et al.* (2008). Nonredundant functions for tumor protein D52-like proteins support specific targeting of TPD52. Clinical Cancer Research *14*, 5050-5060.

Sherr, C.J. (2004). Principles of tumor suppression. Cell *116*, 235-246.

Shi, S., Srivastava, S.P., Kanasaki, M., He, J., Kitada, M., Nagai, T., Nitta, K., Takagi, S., Kanasaki, K., and Koya, D. (2015). Interactions of DPP-4 and integrin beta1 influences endothelial-to-mesenchymal transition. Kidney International *88*, 479-489.

Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536-1543.

Siegel, R.L., Miller, K.D., and Jemal, A. (2016). Cancer statistics, 2016. CA: a Cancer Journal for Clinicians *66*, 7-30.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., *et al*. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research *15*, 1034-1050.

Singh, P., Yang, M., Dai, H., Yu, D., Huang, Q., Tan, W., Kernstine, K.H., Lin, D., and Shen, B. (2008). Overexpression and hypomethylation of flap endonuclease 1 gene in breast and other cancers. Molecular Cancer Research *6*, 1710-1717.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S.*, et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America *98*, 10869-10874.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R.*, et al.* (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400-404.

Summy, J.M., and Gallick, G.E. (2003). Src family kinases in tumor progression and metastasis. Cancer and Metastasis Reviews *22*, 337-358.

Sutherland, R.L., and Musgrove, E.A. (2004). Cyclins and breast cancer. Journal of Mammary Gland Biology and Neoplasia *9*, 95-104.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics *14*, 178-192.

Torres L., Ribeiro F.R., Pandis N., Andersen J.A., Heim S. and Teixeira M.R. (2007). Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. Breast Cancer Research and Treatment *102*, 143-155.

Untch, M., and Luck, H.J. (2010). Lapatinib - Member of a new generation of ErbB-targeting drugs. Breast Care *5*, 8-12.

Valabrega, G., Montemurro, F., and Aglietta, M. (2007). Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. Annals of oncology *18*, 977-984.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237-245.

Velasco-Velazquez, M.A., Homsi, N., De La Fuente, M., and Pestell, R.G. (2012). Breast cancer stem cells. The International Journal of Biochemistry & Cell Biology *44*, 573-577.

Visvader, J. E. (2011). Cells of origin in cancer. Nature *469*, 314-322.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research *38*, e164-171.

Wilson, S.H., Bailey, A.M., Nourse, C.R., Mattei, M.G., and Byrne, J.A. (2001). Identification of MAL2, a novel member of the mal proteolipid family, though interactions with TPD52-like proteins in the yeast two-hybrid system. Genomics *76*, 81-88.

Workman, H.C., Miller, J.K., Ingalla, E.Q., Kaur, R.P., Yamamoto, D.I., Beckett, L.A. *et al.* (2009). The membrane mucin MUC4 is elevated in breast tumor lymph node metastases relative to matched primary tumors and confers aggressive properties to breast cancer cells. Breast Cancer Research *11*, R70.

Zannini, L., Delia, D., and Buscemi, G. (2014). CHK2 kinase in the DNA damage response and beyond. Journal of Molecular Cell Biology *6*, 442-457.

Zhou, D., Hector Battifora, Yokota, J., Yamamoto, T., and Cline, M.J. (1987). Association of Multiple Copies of the c-erbB-2 Oncogene with Spread of Breast Cancer. Cancer Research *47*, 6123-6125.

## Acknowledgements

At the end of this long and hard journey I'd like to thank the people in MolMed who have supported me and who have seen me grow up into the world of research. These have been hard and intense years full of joys and sorrows that have made me able to increase my great will of knowledge and my determination in doing this job.

Thanks to Prof. Pier Paolo Di Fiore, my boss, who gave me the opportunity to join his group. Many thanks to Dr. Fabrizio Bianchi and Prof. Salvatore Pece who followed me during the last four years, teaching me the importance of a constant research for knowledge in all its forms.

Inoltre, vorrei ringraziare la mia famiglia che mi ha supportato in tutti i modi possibili durante gli anni di studio. Gli amici, che sono sempre stati presenti e che mi hanno strappato un sorriso anche nei momenti più difficili.

Infine, ma non meno importante, ringrazio Omar, l'altra metà della mela, la persona che più di tutte mi è stata vicino durante gli ultimo sei anni. A lui devo la serenità che ho raggiunto ed è a lui che dedico questa tesi di dottorato.