LUND UNIVERSITY

Molecular Analysis of Breast Cancer Transcriptomes, Genomes, and Circulating Tumor DNA

Olsson, Eleonor

2015

Molecular Analysis of Breast Cancer Transcriptomes, Genomes, and Circulating Tumor DNA

Eleonor Olsson
Division of Oncology and Pathology
Department of Clinical Sciences, Lund

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden.
To be defended in Segerfalk lecture hall, Sölvegatan 17, Lund University,
at 2:00 pm, Thursday May 28th, 2015.

Faculty opponent

Lecturer Anita Grigoriadis
Department of Research Oncology, King's College London, UK

Organization:	Document name:
LUND UNIVERSITY Division of Oncology and Pathology, Department of Clinical Sciences, Lund	DOCTORAL DISSERTATION
	Date of issue: May 28th 2015
Author(s): Eleonor Olsson	Sponsoring organization

Title and subtitle:

Molecular Analysis of Breast Cancer Transcriptomes, Genomes, and Circulating Tumor DNA

Abstract

Breast cancer is a very heterogeneous disease in terms of clinical characteristics, genetic aberrations and prognosis. In Paper I, we focused on the CD44 molecule that often is aberrantly expressed in breast cancer and is widely used as a marker for cancer stem cells. Several isoforms of the CD44 molecule were analyzed at the transcriptome level across breast tumors and the expression of individual isoforms was correlated to molecular subtypes, protein expression of clinical markers, and cancer stem cell (CSC) phenotypes in breast tumors and cell lines. The CD44S isoform was associated with expression of the CSC marker ALDH1 and the CSC phenotype $CD44^{+}/CD24^{-}$ was correlated to alternatively spliced isoforms in tumors. The isoforms were differentially expressed in molecular subtypes and HER2 and EGFR positive tumors were associated to CD44S and CD44v8-10, respectively. In Paper II, by using targeted genomic re-sequencing we screened for somatic mutations in 1237 genes in a panel of basal-like breast cancer cell lines, both in coding and surrounding non-coding regions. In total, 658 high confidence SNVs and indels were detected and 315 of these were novel (not in COSMIC). A selection of the variants were validated with Sanger sequencing and, 123 of 130 high confidence variants were confirmed including 111 novel variants. The mutation frequency was higher in coding (CDS) compared to non-coding (non-CDS) regions and in particular G or C base replacements were higher in the CDS compared to non-CDS. The SNVs within the context of T[C]A/T[G]A and T[C]T/A[G]A were significantly more common in the CDS than in the non-CDS regions. Re-sequenced data was used to derive copy number estimations, which correlated well to SNP array data. In Paper III, the potential in using tumor-specific rearrangements present in circulating tumor DNA (ctDNA) to detect occult metastatic breast cancer was evaluated. In total, 14 eventual metastatic (EM) patients and 6 long-term disease free (DF) patients were investigated. We used whole-genome sequencing on the primary tumors to derive patient-specific rearrangements that were confirmed by PCR. Circulating tumor DNA levels across multiple plasma samples during the clinical course were analyzed by quantitative droplet digital PCR. Accurate post-surgical discrimination of EM patients (93%) from DM (100%) was achieved by ctDNA monitoring. The average lead-time to clinical detection of metastatic disease was 11 months (range 0-37 months). Moreover, the ctDNA level was a quantitative predictor for both recurrence (P=0.02) and death (P=0.04). We demonstrated that monitoring of ctDNA can be used for early detection of metastatic breast cancer and is a potential tool for optimization of adjuvant therapy and should be evaluated further in clinical studies.

Key words: breast cancer, RNA splicing, cancer stem cells, mutations, circulating tumor DNA, early detection of recurrent breast cancer, sequencing

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language: English
ISSN and key title 1652-8220	ISBN 978-91-7619-143-9
Recipient's notes	Number of pages: **177** / Price
	Security classification

Molecular Analysis of Breast Cancer Transcriptomes, Genomes, and Circulating Tumor DNA

Eleonor Olsson

LUND
UNIVERSITY

Lund University, Faculty of Medicine Doctoral Dissertation Series 2015:64
ISBN 978-91-7619-143-9
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2015


To my family…

Contents

Appended papers

Paper I: **Olsson E**, Honeth G, Bendahl PO, Saal LH, Gruvberger-Saal S, Ringnér M, Vallon-Christersson J, Jönsson G, Holm K, Lövgren K, Fernö M, Grabau D, Borg Å, Hegardt C. CD44 isoforms are heterogeneously expressed in breast cancer and correlate with tumor subtypes and cancer stem cell markers. *BMC Cancer. 2011 Sep 29;11:418.*

Paper II: **Olsson E**, Winter C, George A, Chen Y, Törngren T, Bendahl PO, Borg Å, Gruvberger-Saal S, Saal LH. Mutation screening of 1237 cancer genes across six model cell lines of basal-like breast cancer. *Submitted.*

Paper III: **Olsson E**[*], Winter C[*], George A, Chen Y, Howlin J, Tang MHE, Dahlgren M, Schultz R, Grabau D, van Westen D, Fernö M, Ingvar C, Rose C, Bendahl PO, Rydén L, Borg Å, Gruvberger-Saal SK, Jernström H, Saal LH. Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease. *Accepted for publication in EMBO Molecular Medicine.*

[*] = shared first authorship

Publications not included in thesis

Augsten M, Sjöberg E, Frings O, Vorrink SU, Frijhoff J, **Olsson E**, Borg Å, Östman A. Cancer-associated fibroblasts expressing CXCL14 rely upon NOS1-derived nitric oxide signaling for their tumor-supporting properties. *Cancer Res. 2014 Jun 1;74(11):2999-3010.*

Harbst K, Lauss M, Cirenajwis H, Winter C, Howlin J, Törngren T, Kvist A, Nodin B, **Olsson E**, Häkkinen J, Jirström K, Staaf J, Lundgren L, Olsson H, Ingvar C, Gruvberger-Saal SK, Saal LH, Jönsson G. Molecular and genetic diversity in the metastatic process of melanoma. *J Pathol. 2014 May;233(1):39-50.*

Frings O, Augsten M, Tobin NP, Carlson J, Paulsson J, Pena C, **Olsson E**, Veerla S, Bergh J, Ostman A, Sonnhammer EL. Prognostic significance in breast cancer of a gene signature capturing stromal PDGF signaling. *Am J Pathol. 2013 Jun;182(6):2037-47.*

Peña C, Céspedes MV, Lindh MB, Kiflemariam S, Mezheyeuski A, Edqvist PH, Hägglöf C, Birgisson H, Bojmar L, Jirström K, Sandström P, **Olsson E**, Veerla S, Gallardo A, Sjöblom T, Chang AC, Reddel RR, Mangues R, Augsten M, Ostman A. STC1 expression by cancer-associated fibroblasts drives metastasis of colorectal cancer. *Cancer Res. 2013 Feb 15;73(4):1287-97.*

Tormin A, Brune JC, **Olsson E**, Valcich J, Neuman U, Olofsson T, Jacobsen SE, Scheding S. Characterization of bone marrow-derived mesenchymal stromal cells (MSC) based on gene expression profiling of functionally defined MSC subsets. *Cytotherapy. 2009;11(2):114-28*.

Augsten M, Hägglöf C, **Olsson E**, Stolz C, Tsagozis P, Levchenko T, Frederick MJ, Borg Å, Micke P, Egevad L, Ostman A. CXCL14 is an autocrine growth factor for fibroblasts and acts as a multi-modal stimulator of prostate tumor growth. *Proc Natl Acad Sci U S A. 2009 Mar 3;106(9):3414-9. Epub 2009 Feb 13.*

Nilsson L, Edén P, **Olsson E**, Månsson R, Astrand-Grundström I, Strömbeck B, Theilgaard-Mönch K, Anderson K, Hast R, Hellström-Lindberg E, Samuelsson J, Bergh G, Nerlov C, Johansson B, Sigvardsson M, Borg Å, Jacobsen SE. The molecular signature of MDS stem cells supports a stem-cell origin of 5q myelodysplastic syndromes. *Blood. 2007 Oct 15;110(8):3005-14. Epub 2007 Jul 6.*

Jönsson G, Staaf J, **Olsson E**, Heidenblad M, Vallon-Christersson J, Osoegawa K, de Jong P, Oredsson S, Ringnér M, Höglund M, Borg Å. High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. *Genes Chromosomes Cancer. 2007 Jun;46(6):543-58.*

Abbreviations

AICDA	activation-induced cytidine deaminase
ALDH	aldehyde dehydrogenase
AML	acute myeologenous leukemia
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide
ASE	alternatively spliced exons
BER	base excision repair
bFGF	basic fibroblast growth factor
BRCA1	breast cancer 1, early onset
BRCA2	breast cancer 2, early onset
cfDNA	cell-free DNA from any cell
CIN	chromosomal instability
CML	chronic myeloid leukemia
CSC	cancer stem cell
CSE	constitutively spliced exons
CTC	circulating tumor cell
ctDNA	circulating tumor DNA
DCIS	ductal carcinoma in situ
ddPCR	droplet digital PCR
DES	diethylstilbestrol
DNA	deoxyribonucleic acid
DSB	DNA double strand break
DTC	disseminated tumor cell
EMT	epithelial-mesenchymal transition
EpCAM	epithelial cell adhesion molecule
ER	estrogen receptor
ERBB2	erb-b2 receptor tyrosine kinase 2 (encodes HER2)
FACS	fluorescent-activated cell sorting
GST	glutathione S-transferase
HB-EGF	epidermal growth factor-like growth factor

HDR	homology-directed repair
HER2	human epidermal growth factor receptor 2
HGF/SF	hepatocyte growth factor/scatter factor
IDC	invasive ductal carcinoma
ILC	invasive lobular carcinoma
Ki67	proliferation marker encoded by MKI67
LCIS	lobular carcinoma in situ
MAPK	mitogen-activated protein kinase
MGMT	DNA methyltransferase
MMEJ	microhomology-mediated end joining
MMR	mismatch repair
mRNA	messenger RNA
MSI/MIN	microsatellite instability
NAHR	non-allelic homologous recombination
NER	nucleotide excision repair
NFQ	non-fluorescent quencher
NHEJ	non-homologous end-joining
NOD/SCID	immunocompromised mice
PARP	poly (ADP-ribose) polymerase
PCR	polymerase chain reaction
PI3K	phosphatidylinositol-3-kinase
RNA	ribonucleic acid
PR	progesterone receptor
RTK	receptor tyrosine kinase
RT-PCR	reverse transcription polymerase chain reaction
SAC	spindle assembly checkpoint
SBS	sequencing by synthesis
snRNP	small nuclear ribonucleoprotein
SP	side population
TP53	tumor protein p53

Abstract

Breast cancer is a very heterogeneous disease in terms of clinical characteristics, genetic aberrations and prognosis. In Paper I, we focused on the CD44 molecule that often is aberrantly expressed in breast cancer and is widely used as a marker for cancer stem cells. Several isoforms of the CD44 molecule were analyzed at the transcriptome level across breast tumors and the expression of individual isoforms was correlated to molecular subtypes, protein expression of clinical markers, and cancer stem cell (CSC) phenotypes in breast tumors and cell lines. The CD44S isoform was associated with expression of the CSC marker ALDH1 and the CSC phenotype $CD44^{+}/CD24^{-}$ was correlated to alternatively spliced isoforms in tumors. The isoforms were differentially expressed in molecular subtypes and HER2 and EGFR positive tumors were associated to CD44S and CD44v8-10, respectively. In Paper II, by using targeted genomic re-sequencing we screened for somatic mutations in 1237 genes in a panel of basal-like breast cancer cell lines, both in coding and surrounding non-coding regions. In total, 658 high confidence SNVs and indels were detected and 315 of these were novel (not in COSMIC). A selection of the variants were validated with Sanger sequencing and, 123 of 130 high confidence variants were confirmed including 111 novel variants. The mutation frequency was higher in coding (CDS) compared to non-coding (non-CDS) regions and in particular G or C base replacements were higher in the CDS compared to non-CDS. The SNVs within the context of T[C]A/T[G]A and T[C]T/A[G]A were significantly more common in the CDS than in the non-CDS regions. Re-sequenced data was used to derive copy number estimations, which correlated well to SNP array data. In Paper III, the potential in using tumor-specific rearrangements present in circulating tumor DNA (ctDNA) to detect occult metastatic breast cancer was evaluated. In total, 14 eventual metastatic (EM) patients and 6 long-term disease free (DF) patients were investigated. We used whole-genome sequencing on the primary tumors to derive patient-specific rearrangements that were confirmed by PCR. Circulating tumor DNA levels across multiple plasma samples during the clinical course were analyzed by quantitative droplet digital PCR. Accurate post-surgical discrimination of EM patients (93%) from DM (100%) was achieved by ctDNA monitoring. The average lead-time to clinical detection of metastatic disease was 11 months (range 0-37 months). Moreover, the ctDNA level was a quantitative predictor for both recurrence (P=0.02) and death (P=0.04). We demonstrated that monitoring of ctDNA can be used for early detection of metastatic breast cancer and is a potential tool for optimization of adjuvant therapy and should be evaluated further in clinical studies.

General background

Introduction to breast cancer

Cancer epidemiology

Worldwide there were 14.1 million new cancer cases reported in 2012 [1]. In total that year, 8.2 million people died from their cancer disease and 32.6 million people were living with cancer (within 5 years of diagnosis). The age-standardized cancer incidence rate is 84% higher in more developed parts of the world, but the mortality rates are only 15% higher in men and 8% higher in women compared to less developed regions [1].

For some decades, breast cancer has been the second most common (12%) cancer in the world, and by far the most frequent malignant disease among women. In total, 1.67 million new breast cancer cases were diagnosed in 2012, which constitutes 25% of all cancers in women. Incidence rates of breast cancer vary in different parts of the word, from 27 per 100,000 in Middle Africa and Eastern Asia to 96 per 100,000 in Western Europe. However, the outcome is more favorable in developed regions than in less developed regions with mortality rates varying from 6 per 100,000 in Eastern Asia to 20 per 100,000 in Western Africa. Of all cancer-associated deaths in the world, breast cancer ranks as the fifth most common with 522,000 deaths per year and for women it is the most common cause of cancer death [1]. In Sweden, breast cancer is the most common malignancy among women and 9,123 new cases were diagnosed in 2013 and the incidence of the disease has increased from 80 to 190 cases per 100,000 during the period from 1970 to 2013 [2].

The regional difference in cancer incidence is related to mainly life style and environmental factors, hence, migration to a more developed region is associated with an increased risk of developing cancer. The prevalence of

carriers of the major susceptibility genes appears to explain only a minor part of the variation in incidence in different world regions [3].

Risk factors for development of breast cancer

Carcinogenic factors that affects the risk of development of breast cancer is evaluated by the International Agency for Research on Cancer (IARC) and there is sufficient evidence that intake of alcoholic beverages, diethylstilbestrol (DES), estrogen-progestin contraceptives, estrogen-progestin menopausal therapy, X-ray radiation, and gamma-radiation increase the risk [4]. According to IARC, there is limited evidence that tobacco smoking may slightly increase the risk of breast cancer, and this was also confirmed in a recent study [4, 5]. The World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) evaluates evidence for risk factors like diet, weight gain, and physical exercise, and found sufficient evidence for the risk of developing post-menopausal breast cancer for both adult attained height and body fatness [6]. IARC reports that the only factor that convincingly reduces risk for breast cancer is breastfeeding longer than 6 months with the risk, by the age of 70 years, reduced by 4% for every 12 months of breastfeeding [4, 7].

Previous benign or malignant breast disease and family breast cancer history increase the risk for developing breast cancer. Other risk factors are advanced age, early menarche, late menopause, endogenous hormonal levels, and reproductive aspects [8]. In pre-menopausal women, only higher serum testosterone levels are unfavorable and in post-menopausal women higher levels of estradiol, estriol, androstenedione, and testosterone all are associated with a higher risk of developing breast cancer [9]. Nulliparity and increasing age at first birth are associated with a higher risk of getting breast cancer, however, this may be limited to estrogen/progesterone receptor positive tumors [7, 10, 11].

Approximately 5-10% of all breast cancer cases are caused by different inherited mutations in certain susceptibility genes and, in general, these are

inherited in an autosomal dominant fashion with limited penetrance. Two genes associated with DNA damage repair, *BRCA1* and *BRCA2,* are both high-penetrance genes and on average 60-65% of *BRCA1* and 45-55% of *BRCA2* female carriers develop breast cancer [12, 13], however, these risks can be modified by presence of other cancer susceptibility alleles [14]. In the average population, *BRCA1* and *BRCA2* mutation carriers are rare (0.11% and 0.12%, respectively), although the prevalence in different ethnicities varies [15, 16]. For women diagnosed with breast cancer at <50 years approximately 6% of the cases have germline mutations in any of these two genes, while for women >50 years only 1-1.5% are affected [15, 16]. Interestingly, women with *BRCA1*/*BRCA2* mutations have been shown to have higher titers of both estradiol and progesterone compared to women known to be negative for the mutations [17].

Other high-penetrance genes are *TP53* (Li-Fraumeni syndrome) and *PTEN* (PTEN hamartoma tumor syndrome), with carriers at increased risk to a variety of different tumor forms including breast cancer; however both of these syndromes are very rare [18]. Several other rare susceptibility genes are known or suspected to confer an intermediate- to low-risk in carriers, including *CHEK2*, *ATM*, *PALB2*, *BRIP1*, *CDH1*, *RAD50*, *RAD51C* and *RAD51D* of which many are involved in DNA repair mechanisms [19-21].

Clinical aspects of breast cancer

The majority of tumors found in the breast are benign which means that they will neither spread outside the breast nor become life-threatening, although they can still be proliferative and cause notable symptoms. However, women diagnosed with benign tumors, especially with atypical hyperplasia, may have a higher risk of developing invasive breast cancer [22].

Tumors that are pathologically classified as breast carcinomas could be either non-invasive, *e.g.* ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS), or invasive, whereof invasive ductal carcinoma (IDC) and invasive

lobular carcinoma (ILC) accounts for 50-75% and 10-15% of all cases, respectively. Other less common types of invasive breast cancers include the mucinous, medullary, papillary and tubular, and these are characterized by different histological and molecular features [23].

Early breast cancer is defined as cancer that has not spread to other parts in the body than the breast or the ipsilateral lymph nodes of the breast. However, all breast carcinomas have the potential to become metastatic and once the tumor has spread to other vital organs of the body, the disease is defined as advance breast cancer and is rarely curable. Developing a local recurrence of the disease is also possible and if the malignancy is inoperable this can also be defined as an advanced stage of the disease. Nevertheless, many different treatment options are available and there is probably much room for improvement in managing patients with advanced breast cancer in terms of individualized and multidisciplinary treatment options [24].

To describe the current clinical status of the disease and aid in determining prognosis, the TNM system for staging is often used, which takes tumor size (T), number of positive lymph nodes (N), and presence of metastasis (M) into account, and is summarized on a scale from Stage I to IV. The Nottingham Prognostic Index (NPI) [25] can also be used to predict outcome of the disease and here the tumor size, the number of positive lymph nodes and the histological grade of the tumor are taken into account. In determination of histological grade, glandular/tubular differentiation, nuclear pleomorphism and the mitotic count are scored by the pathologist. Adjuvant! Online and PREDICT are software tools that utilize these variables and others for predicting outcome and the benefit or risk of adjuvant therapy for breast cancer patients [26, 27].

Three of the most important clinical biomarkers for breast cancer are the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). These serve as important prognostic markers and as drug targets in breast cancer: *e.g.* anti-estrogen therapies such as

tamoxifen and aromatase inhibitors are used for tumors overexpressing ER, which occur in about 80% of cases. Approximately, 60% of all breast tumors overexpress the PR-receptor, and this is most common in ER-positive tumors [28, 29]. The HER2 receptor, whose encoding gene *ERBB2* is amplified in approximately 15% of cases, is the target for antibody therapies such as trastuzumab which bind and inhibit growth of HER2-dependent tumors [30]. Ki67 is a marker for proliferation and can be used to further classify breast cancer, with high proliferation tumors more likely to benefit from chemotherapy [30]. However, due to lack of standardization of methodology for interpretation of staining and scoring of Ki67, the fraction of "Ki67 high" tumors varied from 1% to 28.6% in different studies [31].

More than a decade ago the intrinsic molecular subtypes were proposed in breast cancer, and between 4 and 6 groups were initially suggested based on variation in gene expression [32-34]. More recently, using an integrated approach with both genomic and transcriptomic data, 10 different subtypes were described [34]. Of these, the PAM50 classifier, based on 50 genes separating tumors into 5 classes (luminal A, luminal B, HER2-enriched, basal-like, and normal-like), appears to be the most frequently utilized gene expression signature today. To enable translation of these intrinsic subtypes into clinical use, currently a number of classifications have been proposed using surrogate clinicopathologic markers. For example, the St. Gallen guidelines suggest to divide breast cancers into 5 subtypes using conventional pathological biomarkers [30]. Therein, the basal-like subtype is defined as "triple-negative" for lack of ER/PR hormone receptors and negative for HER2, and in the "HER2 positive (non-luminal)" subtype HER2 is over-expressed or amplified with absence of ER and PR. The luminal A group is defined as "luminal A-like" and positive for ER and PR but negative for HER2 and low expression of Ki67. Moreover, if a multi-gene-expression assay of the recurrence risk (*i.e.* the so-called 21-gene RS or the 70-gene signatures) [35-37] is available, to be luminal A-like the RS score should be defined as low based on the gene expression results. The luminal B subtype could be translated into "luminal B-like (HER2 negative)", which is defined as ER

positive and HER2 negative and either high expression of Ki67 or negative or low expression of PR; furthermore, if a multi-gene-expression assay [35-37] is available the recurrence risk should be defined as high based of the gene expression results. However, some luminal B tumors do express HER2, therefore, this subtype can be described as "luminal B-like (HER2 positive)" if the tumor is ER positive and HER2 is overexpressed or amplified. Together, stage, biomarker status, and the surrogate clinico-pathologic definitions (based on the intrinsic subtypes) add valuable information when it comes to prognosis and treatment recommendations in breast cancer [30].

Survival statistics from Sweden and the UK show that 5 years after diagnosis mortality is 10-13%, and after 10 years this percentage increases to 17-22%. Therefore, the risk for late recurrences is a major concern for women with breast cancer, especially since there is a lack of good predictors for relapse of the disease [2, 38].

Carcinogenesis

Hallmarks of cancer

All cancers acquire various genetic aberrations that alter biological functions from that of the normal (non-cancerous) cells in the originating organ. Typical for breast cancer tumors is that these genetic changes are very heterogeneous [39] compared to for example other cancer forms with pathognomonic aberrations (*e.g.* the Philadelphia chromosome in CML) [40, 41]. Following the transformation of a normal cell to a malignant tumor cell, it has been proposed that several characteristics defined as the "hallmarks of cancer" should be fulfilled [42, 43]. Originally, these hallmarks consisted of six capabilities: sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis, enable replicative immortality, inducing angiogenesis and resisting cell death [43]. Recently, additional hallmarks were proposed and the current hallmarks were updated to also include: avoiding immune destruction, enabling tumor-promoting inflammation, deregulating cellular energetics, and acquired genome instability and mutation [42].

Genomic instability in breast cancer

The majority of the cancer hallmark capabilities arise from genomic instability and the acquisition of genetic and epigenetic changes in the genome of the tumor cell during tumor progression. Since most types of cancers have the characteristics of genomic instability, this topic is of great interest in cancer research, but still, at what stage genomic instability arises and the rationale to why it occurs is not fully understood. In pre-cancerous lesions, genomic instability is associated with loss of or dysfunctional telomeres [44]. End-to-end fusions between chromosomes can then lead to breakage-fusion-bridges cycles (see section *Changes in the karyotype of cancer cells*). However, during tumor progression and before transformation to a malignant cell, telomere healing may be observed which is the result of increased telomerase activity [44]. Various measures of genomic instability have been described and the

most common is called chromosomal instability (CIN), which estimates the rate of chromosome structure and number changes occurs over time when comparing cancer cells to normal cells [45]. However, since only some abnormalities can be observed in all cells of a tumor, it has been suggested that all cells in a tumor are the progeny of one single cell, with additional genetic changes occurring over time [46].

Another type of genomic instability has been described called microsatellite instability (MSI or MIN) which refers to the expansion or reduction of the number of oligonucleotide repeats in microsatellite sequences [47] and is common in colorectal cancer [48]. Other forms of genomic instabilities manifest themselves in terms of increased frequencies of base-pair mutations [49].

In hereditary breast cancer, germ line mutations in *BRCA1*, *BRCA2*, *BRIP1* (*BRCA1*-interacting protein), *PALB2* (partner and localizer of *BRCA2*), *RAD50*, *RAD51C* and *RAD51D*, genes involved in double-strand break strand repair, appear to be initiating lesions for genomic instability [50]. Moreover, in the classic mutator hypothesis, it is suggested that an increased mutation rate leads to a diversity of mutations, which is observed in many of these tumors and correlates well with deficiencies in DNA repair mechanisms [51].

In sporadic tumors the initial mechanism of genomic instability is less well understood. The mutational screening studies conducted in sporadic cancer has not detected frequent redundant mutations in genes directly involved in DNA repair and mitotic checkpoint mechanisms [45], however, some of these caretaker genes might still be uncharacterized and effects of other mechanisms like epigenetic silencing may also contribute. However, in hereditary breast tumors often both alleles of inherited breast cancer genes are affected, and this is probably necessary before establishment of genomic instability. In sporadic breast cancer, both alleles in genes associated with hereditary breast cancer are infrequently mutated [52]. Instead, mutations or deletions in a varied constellation of tumor suppressor genes and oncogenes like *TP53*, *PTEN*,

CDKN2A, *CDKN2B*, *RB1*, *PIK3CA*, *MAP3K1*, *MAP2K4* and amplifications in *CCND1*, *ERBB2,* and *MYC* are found [53, 54].

Most sporadic cancers are characterized by genomic instability, especially CIN [45]. Although *TP53* is considered to be a tumor suppressor gene that controls cell proliferation, it is also a DNA damage checkpoint gene and inactivation could be expected to contribute to genomic instability, whereas inactivation of *TP53* alone does not lead to spontaneous genomic instability [55, 56]. Instead, activation of growth signaling pathways controlled by genes that merely function as oncogenes induce genomic instability, and this has led to the formulation of the oncogene-induced DNA replication stress model [57]. In this model, collapse of the DNA replication forks leads to formation of DNA double strand breaks (DSBs), a process that occurs in both pre-cancerous lesions and in established cancers. However, in the pre-cancerous lesions DSBs lead to activation of TP53, which induces apoptosis or senescence. For development of cancer, these DNA repair mechanisms must be reduced, and this is believed to occur by selection for *TP53* mutations caused by oncogene-induced damage. Moreover, inactivation of TP53 can also be achieved by overexpression of MDM2, which negatively regulates TP53 through binding to its transactivation domain or via E3 ligase activity that mediates ubiquitin-dependent degradation of TP53 [58, 59]. *MDM4* is a homologue of *MDM2*, and is also capable of regulating transcriptional activity of TP53 and enhances the E3 ligase activity towards TP53 by forming a heterodimer with MDM2. Another example of a protein that function upstream of TP53 is ATM, a kinase that phosphorylates TP53 in response to DNA damage; therefore, inhibition of ATM can certainly affect the functional role of TP53 [60].

One way of investigating the validity of the model of oncogene-induced DNA replication stress, and that it leads to genomic instability, would be to investigate if mutual exclusive genetic changes actually exist in the TP53 signaling pathway. This hypothesis was recently suggested to be true across different subtypes of breast cancer with aberrant TP53 signaling, including

amplifications of *MDM2* and *MDM4* and mutations in *AKT1*, *ATM*, *CHEK2* and *TP53* [53].

Recurrent mutated genes and activated pathways in breast cancer subtypes

The fact that breast cancer is a genetically heterogeneous disease has been observed in many studies, and it was also confirmed in a mutation screening study including 507 tumors [53]. Notably, only seven genes were mutated at a frequency above 5% (*TP53*, *PIK3CA*, *GATA3*, *MAP3K1*, *MLL3* and *CDH1*) and in total 35 genes were described as significantly mutated in that cohort. Mutation patterns attributed to each of the intrinsic subtypes were also investigated and, on average, the basal-like and the HER2-enriched tumors had a higher load of mutations compared to the luminal A and luminal B subtypes. In the basal-like cancers, 80% of the cases were mutated in *TP53*, and the second most common mutated gene was *PIK3CA* (9%), whereas the HER2-enriched subtype showed high frequencies of both *TP53* (74%) and *PIK3CA* (39%) mutations. However, the luminal subtypes were associated with the highest number of recurrent mutations, and of all luminal A tumors, 47% harbored mutations in *PIK3CA* followed by mutations in *MAP3K1* (13%), *GATA3* (14%), *TP53* (12%), *CDH1* (10%), *MLL3* (8%) and *MAP2K4* (7%). In luminal B cancers, about 31% were mutated in *TP53* and *PIK3CA*, respectively, 15% were mutated in *GATA3* and 5% in *MAP3K1*. Only two genes, *TP53* and *PIK3CA*, were described as significantly recurrently mutated in all four subtypes [53].

As mentioned earlier, genes included in the TP53 pathway were found to have mutually exclusive genetic aberrations across all subtypes. Nevertheless, differences in how frequent the pathway was affected by mutations varied between subtypes. The TP53 pathway was more often affected in the basal-like and HER2-enriched tumors than in the luminal subtypes, most likely owing to the difference in frequency of mutations in *TP53* itself.

Since *PIK3CA* was most frequently mutated in the luminal A tumors, one would anticipate the phosphatidylinositol-3-kinase (PI3K) pathway to be activated in this tumor type. However, the basal-like and HER2-positive tumors showed higher activity in this pathway. This difference was demonstrated both on gene expression and protein level, and may be related to the high rate of PTEN protein loss in basal-like and HER2 subtypes [53, 61, 62]. In the luminal tumors, JNK/JUN mediated apoptosis was suggested to be diminished owing to mutual exclusive mutations found in *MAP3K1* and *MAP2K4*, and *CCND1* amplifications were common [53]. Indeed, evaluation of the significance of genetic aberrations in a biological context is of great importance, and of course there are a multitude of features that differs between the breast cancer subtypes not mentioned here.

Mammary gland biology and tumorigenesis

Models of tumorigenesis and cell heterogeneity

Already in 1976 the clonal evolution model for tumor progression was proposed. In this model genetic instability eventually leads to expansion of tumor cell clones, which could be described as an evolutionary process where different subclones can emerge within individual patients [46]. Thus, this process would be triggered by genetic alterations that confer a cell with growth and survival advantages that will lead to clonal expansion. Subsequently, the descendants of these cells gain a second alteration and another clonal expansion occurs, and this procedure may continue until a cancer has developed. Furthermore, any of the formed tumor subclones could give rise to treatment resistance, become invasive and cause metastasis [46, 63]. Indeed, the presence of different clones within a tumor has been confirmed in various studies [64, 65]. However, it is still far from understood how to elucidate which clone drives the tumor and what triggers the metastatic cascade. Moreover, the reasons for developing therapy resistance and/or late recurrences need to be further investigated.

Another theory, the cancer stem cell model, has more recently been proposed as an alternative explanation for tumorigenesis. In the cancer stem cell model it is proposed that an aberrant differentiation program influences the intra-tumor heterogeneity observed in many cancers. Moreover, even if a differentiated cell perhaps could acquire capacity of renewal, in the cancer stem model it is considered to be more likely that an undifferentiated cell with inherent capacity of renewal actually comprises the cell of origin for the tumor. None of these suggestions are embraced in the clonal evolution model. Furthermore, in the cancer stem cell model only a small proportion of cells sustain tumor progression and only this compartment of cells could, by gaining additional mutations, become more aggressive. This could be related to therapeutic resistance in which cancer stem cells are inherently drug-

resistant while in the clonal evolution model there is a selection for tolerant clones during therapy [63].

Notably, there are similarities between the cancer stem cell model and the clonal evolution model, including the possibility that the tumor arise from a single cell of any differentiation stage and by acquiring multiple mutations, the tumor cell gain unlimited proliferation potential and its characteristics can be influenced by the cell of origin [63]. Yet it is not clear which of these two models that describes the tumor development in epithelial tissues in the best way and it is quite possible that both models are in operation to various extents and under different contexts.

Interestingly, recent evidence suggests that plasticity in terms of bidirectional conversion between stem cells and a more differentiated cell states may exist in epithelial cells [66-69]. If this type of dynamic transition between different cell phenotypes is preserved in epithelial tumors this could to some extent explain the heterogeneity seen in for example breast tumors [70].

Stem and progenitor cells in the normal mammary gland

The mammary gland can be described as a branching tree-like structure, where lobules cluster to form individual lobes that by extralobular ducts converge into main ducts connecting to the nipple. Each lobule consists of a number of ductules connected by intralobular ducts and, notably, the majority of breast malignancies originate from the lobules and not from the extralobular ducts. Both ducts and lobules contain two major epithelial structures defined by an outer layer of myoepithelial cells and inner layer of luminal cells [71, 72].

The cellular composition of the mammary epithelium can in more detail be described as a hierarchy of cells spanning from undifferentiated stem cells to terminally differentiated luminal and myoepithelial cells. The myoepithelial cells do not express hormone receptors, while the luminal cells can stain positive or negative for the estrogen and progesterone receptors [73, 74]. To

improve the understanding of tumor development in breast cancer, it is important to unravel the cellular hierarchy of the mammary gland and to define which cells that have renewing and differentiation capacities. A breast stem cell is characterized by its ability to proliferate and generate a progeny cell that remains a stem cell after symmetric cell division (*i.e.* the cell is self-renewing). The breast stem cell can undergo either symmetric division, creating two identical daughter cells, or asymmetric cell division, in which one of the daughter cells remain a stem cell and the other differentiate into a more mature cell. Daughter cells that enter differentiation are believed to undergo symmetric cell divisions and are termed "transit-amplifying cells" or progenitor cells. The progenitor cells may undergo a large series of symmetrical cell divisions before their descendants eventually are fully differentiated, and thereby become so-called post-mitotic differentiated. There is also a possibility that (committed) progenitor cells might have retained a limited capacity for self-renewal, hence, these cells could possibly divide asymmetrically, but that is still under debate [75].

To outline the phenotype of cells that show enrichment for stem cell properties in the breast, different *in vitro* and *in vivo* assays have been used. By using fluorescent-activated cell sorting (FACS), cell populations can be separated based on expression of different cell surface markers and tested for their ability to differentiate along both luminal and myoepithelial lineages or for branching morphogenesis in three-dimensional (3D) culture. The cell populations can also be tested for engraftment capacity in *in vivo* xenotransplantation experiments. Several different combinations of markers have been tested and as yet there is no clear consensus regarding which markers that best identify the mammary stem cell population, and it has even been suggested that more than one stem cell compartment may be present in the mammary epithelium [74].

Several studies have indicated that $EpCAM^{low}CD49f^{high}$ phenotype characterize mammary epithelial cells with repopulation capacity *in vivo* and capacity for bipotent differentiation *in vitro*. The $EpCAM^{low}CD49f^{high}$ cells express

myoepithelial linage markers like p63 and cytokeratin-14 but stains negative for the estrogen receptor [76-79].

Also the ALDH+ cell population, as defined by the ALDEFLUOR assay that measures enzymatic activity of ALDH, has been suggested to contain breast stem cells [80]. However, the differentiation potential of ALDH+ cells is somewhat controversial. Some studies support that these cells are restricted to the luminal linage when cultured *in vitro,* and provide only short-term engraftment *in vivo* [81, 82], whereas other studies demonstrate both *in vivo* and *in vitro* multilineage potential of ALDH1+ cells [80, 83, 84]. Importantly, is has been shown that ALDH+/ER- cells can generate an ER+ cell of luminal linage as well as cells expressing myoepithelial markers [85].

Interestingly, upon stimulation with progesterone or estrogen, estrogen receptor positive (ER+) epithelial cells (luminal progenitors or differentiated luminal cells) can by paracrine signaling regulate hormone receptor negative cells (stem cells, luminal progenitor cells or differentiated luminal cells) [74]. In a mouse model, it has been shown that the mammary stem cell pool is increased upon stimulation with progesterone and estrogen and this is likely mediated by RANKL, through paracrine signaling. RANKL is a target of progesterone and is demonstrated to be involved in the mouse mammary gland formation [79, 86, 87]. The progesterone/RANKL regulatory mechanism is suggested to be preserved also in the human mammary gland [88]. Moreover, interaction possibly occurs between stromal cells (fibroblasts and adipocytes) and mammary epithelial cells lining the ducts and lobules [74]. Interestingly, it also seems like Notch activation result in increased self-renewal and commitment of the bilineage progenitors to the myoepithelial linage, but appears to have no effect on differentiated cells [89]. Also canonical Wnt/β-catenin pathway and Sonic Hedgehog signaling have been suggested to regulate self-renewal and differentiation in breast epithelial cells [90, 91]

In the immortal strand hypothesis, initially presented in 1975, it was suggested that stem cells could avoid introducing mutations if they, at mitosis, always

keep the chromatid with the older template strand [92]. Consequently, the stem cell compartment would represent the only stable repository of genetic information within a tissue and the stem cells should probably be more protected from genetic damage compared to differentiated cells. In a mouse model, it has been shown that adult mammary stem cells retain their template strand DNA during mitosis [93].

Notably, it seems like luminal progenitor cells have much shorter telomeres than both basal epithelial cells (enriched in cells with both bipotent and myoepithelial clonogenic activity *in vitro*) and mature luminal cells, and perhaps this makes them more susceptible to DNA damage. The luminal progenitor cells also showed the highest telomerase activity, however, the activity declines significantly with age [94]. The telomere shortening in the progenitor cells compared to more primitive stem cells is probably needed to limit the replicative life span [95].

Cell of origin in breast cancer

To better understand the mechanisms behind tumor development, it is important to evaluate the potential relationship between the normal breast epithelial hierarchy and the different breast cancer subtypes. Attempts have been made to link the tumor subtypes to their closest normal epithelial counterpart, by using gene expression profiling and different *in vitro* and *in vivo* models. Although this approach is challenging due to intra- and inter-tumor heterogeneity, results from these analyses indicate that the basal-like tumor subtype may arise from ER+ or ER- luminal progenitors, at least in *BRCA1*-mutant carriers [79]. The breast cancer subtype that has been described as "Claudin-low", is characterized by low expression of cell adhesion markers and luminal differentiation markers (CD24 and EpCAM) and appears to have high gene expression ratios between CD49f/EpCAM and CD44/CD24 (CD44 is a surrogate stem cell marker, see section *Cancer stem cells*) and is therefore suggested to arise from mammary stem cells [96]. Nevertheless, it seems like enrichment for gene signatures that characterize

normal stem cells (in mice) can predict the distant-metastasis free survival for patients with triple-negative breast tumors [97]. Probably this means that a tumor become more aggressive if stem cell properties are preserved during tumor progression.

The HER2-enriched, luminal A and luminal B tumor subtypes are believed to originate from the luminal cell linage, but it is not well understood from which cell populations [74].

Since cancer cells need to be able to self-renew, normal cells with this capacity are conceivable targets for mutagenesis that eventually leads to cancer. The fact that progenitor cells (that may have a limited capacity for self-renewal) divide more often than stem cells could make them a possible target for mutagenic events. Intriguingly, it has also been suggested that breast cancer cells could dedifferentiate through epithelial-mesenchymal transition (EMT), which generates cells with properties of stem cells [75, 98].

Cancer stem cells

A cancer stem cell (CSC) could be defined as a malignant cell that has the capacity of both self-renewal and to generate countless progeny that constitute the tumor bulk. The concept of CSCs was first demonstrated in acute myeologenous leukemia (AML), by showing that different cell populations in the blood possess large differences in tumorigenicity [99]. The presence of CSCs have been demonstrated in other types of leukemia like chronic myeloid leukemia (CML) [100], that is often characterized by the BCR-ABL translocation t(9;22)(q34;q11), which translates into a constitutively active kinase that can be efficiently targeted by the drug imatinib. However, a minimal residual disease can persist for years of treatment with imatinib, and potentially this can be explained by a small subset of imatinib-resistant CSCs that both are slow-cycling and self-renewing [101]. The presence of CSCs has also been suggested in several types of solid tumors like ovary (CD44+CD117+), colon (CD133+) and brain (CD133+) [102-104].

In breast cancer, several methods of isolation and/or enrichment of CSCs have been suggested: separation of the side population (SP) with Hoechst staining, which enriches for cells overexpressing the ATP binding cassette (ABC) family [105]; isolation of cells by fluorescent-activated cell sorting (FACS) of selected cell surface markers associated with CSCs [106]; or measurement of the enzymatic activity of aldehyde dehydrogenase (ALDH), which could be performed by using the ALDEFLUOR assay with subsequent cell sorting [107].

The enrichment for CSCs in different cell populations of breast tumors can be estimated *in vivo* by xenotransplantation of cells into immunocompromised mice (*e.g.* NOD/SCID) and by using this method the tumor-initiating capacity (or tumorigenicity) is determined [106]. To further characterize the enrichment for tumor stem cells in different cell compartments *in vitro*, various methods have been described: anchorage-independent spheroid cultures of so-called tumorspheres in the presence of basic fibroblast growth factors (FGF) or epidermal growth factors (bEGF) [84] as well as colony formation assays [108].

In breast cancer, it has been shown that the combination of the cell surface antigens CD24-/low/CD44+/lineage$^-$ enriches for breast CSCs. Importantly, serially passaging of this cell population *in vivo* could regenerate new tumors with a phenotypically diverse mix including also nontumorigenic cells present in the initial tumor [106]. It has also been shown that ALDH is a marker of breast cancer cells, and a high expression of ALDH is associated with poor outcome [80, 107, 109]. Breast cancers staining positive by immunohistochemistry for either CD44+/CD24- or ALDH1high have been associated with the basal-like subtype although the overlap between the phenotypes seems to be very small, possibly due to distinct levels of differentiation [110, 111]. However, combining the phenotype of CD44+/CD24- and high activity of ALDH as measured by the ALDEFLUOR assay seems to define a subpopulation of breast tumor cells with further

increased tumor-initiating capacity compared to the cells bearing only one of these phenotypes [80].

Activated pathways associated with normal stem and progenitor cells that regulates self-renewal and linage commitment, including Notch [89], Wnt/β-catenin [90, 112] and Hedgehog signaling [91, 113], have been suggested to control the CSC compartment. For example, it has been proposed that inhibition of *NOTCH4* reduce the capability of tumor initiation significantly *in vivo* [114]. Recently, it was shown that inhibition of Wnt/β-catenin signaling with a small molecule decrease the size of both the the ALDH+ CSC compartment and the remaining tumor bulk, both *in vivo* and *in vitro*, using a breast cancer cell line model [115]. Activation of Hedgehog signaling seems to occur through *Bmi-1* in normal stem and progenitor cells and the same has been observed in CSCs [91].

Since intratumoral heterogeneity is a characteristic of breast cancer, there is a need to address this issue in the context of CSCs. It is known that the markers CD44 and CD24 are rarely co-expressed on the same cells [110]. Therefore, cells with the CD44+ phenotype can be considered to be enriched for CSCs, while CD24+ cells define a more differentiated cell type. In a study involving a limited number of tumors, it has been shown that CD44+ and CD24+ cells can be clonally related and it was also proposed that cell CD24+ cells had acquired additional genetic events compared to the CD44+ cells [116].

In a more recent study from the same research group, a more complex pattern was illustrated. Several different clones were present within the pool of CD44+ cells from a single tumor. Moreover, genetic heterogeneity was observed between distinct tumor cell populations that were defined based on markers of cellular phenotypes [117]. It is challenging to interpret these results, and it could be argued that these results are inconsistent with the CSC model. However, it does not exclude the possibility that only the CSC compartment could survive in long-term expansion (even if clonal evolution occurs in all cancer cell populations).

Resistance to therapy in relation to breast cancer stem cells

Established strategies for eradicating breast tumors have been to kill all cancer cells with radiotherapy and systemic therapies. Another emerging strategy is to specifically target the cancer stem cells. Also, to enable targeting of all cells in a tumor by conventional chemotherapy induction of differentiation of the CSCs could be another option.

CSCs have been suggested to mediate therapy resistance by a number of mechanisms. For example, CSCs exhibit greater multidrug resistance, in which overexpression of efflux pumps in the ATP binding cassette family is thought to be responsible for pumping out for example chemotherapeutic agents at a higher rate in the stem cell population than in the remaining tumor bulk. Indeed, in HER2-negative breast cancer, it has been shown that conventional chemotherapy increased the fraction of CD44+/CD24- CSCs [118]. Also, resistance to both apoptosis and senescence has been associated to CSCs [75, 119]. This could lead to survival of the CSC population that eventually could repopulate the tumor, with or without novel mutations induced by the chemotherapeutic treatment [120]. Resistance to irradiation has also been suggested for breast CSCs, which perhaps could indicate that DNA repair mechanisms are differently regulated in CSC [108].

In breast cancer, the stem cell population defined as CD44+CD24- often has low or absent expression of ER. It has been suggested that paracrine signaling towards the CSCs upon estrogen stimulation, is mediated through ER+ breast epithelial cells that do not express the stem cell phenotype. Possibly, downstream signaling of estrogen could be mediated by both epidermal growth factor (EGF) and Notch receptor signals [86]. Therefore, inhibition of the estrogen receptor could perhaps be beneficial for tumors with cells of the CSC phenotype.

The CD44 molecule

The cell adhesion CD44 molecule is a transmembrane glycoprotein that is involved in various functions in normal cells, for example the immune system and embryogenesis. In tumors, a certain epitope of CD44 was first associated with a metastatic potential in an animal model [121]. Later, the CD44 molecule was associated to tumor progression and steps necessary for the metastatic cascade in primary breast cancer [122]. However, association of CD44 expression to survival has given rise to contradictory results [123-126].

The CD44 molecule is often aberrantly expressed and subjected to alternative splicing in tumors and, as described earlier, CD44 is suggested to mark cancer stem cells in different malignant diseases, including breast cancer. There is no clear evidence that CD44 regulates pathways that sustain self-renewal in breast cancer [127]. However, it is known that CD44 is a target gene of Wnt signaling, and recently it was shown in colon cancer that CD44 is a regulator of the canonical Wnt/β-catenin pathway [128].

The pre-mRNA of CD44 is encoded by 20 exons, whereof exons 1-5 and 16-20 are constitutive and exons 6-15 are alternatively spliced (termed variant exons v1-v10) (Figure 1A). The variant exon v1 is not expressed in human, though it is expressed in mice. The CD44 standard (CD44s) isoform contains no variant exons and is ubiquitously expressed in human. Conversely, the isoforms that includes variant exons are expressed in epithelial and often proliferating cells like keratinocytes, dendritic cells, activated lymphocytes and tumor cells [127, 129].

The protein structure in the extracellular region consists of the amino-terminal domain with ligand-binding motifs and a stem structure including the variable exons which can vary between 46 and 381 amino acids in length) (Figure 1B). Moreover, the CD44 molecule comprises a transmembrane region and a cytoplasmic tail with important functions in both cytoskeletal organization

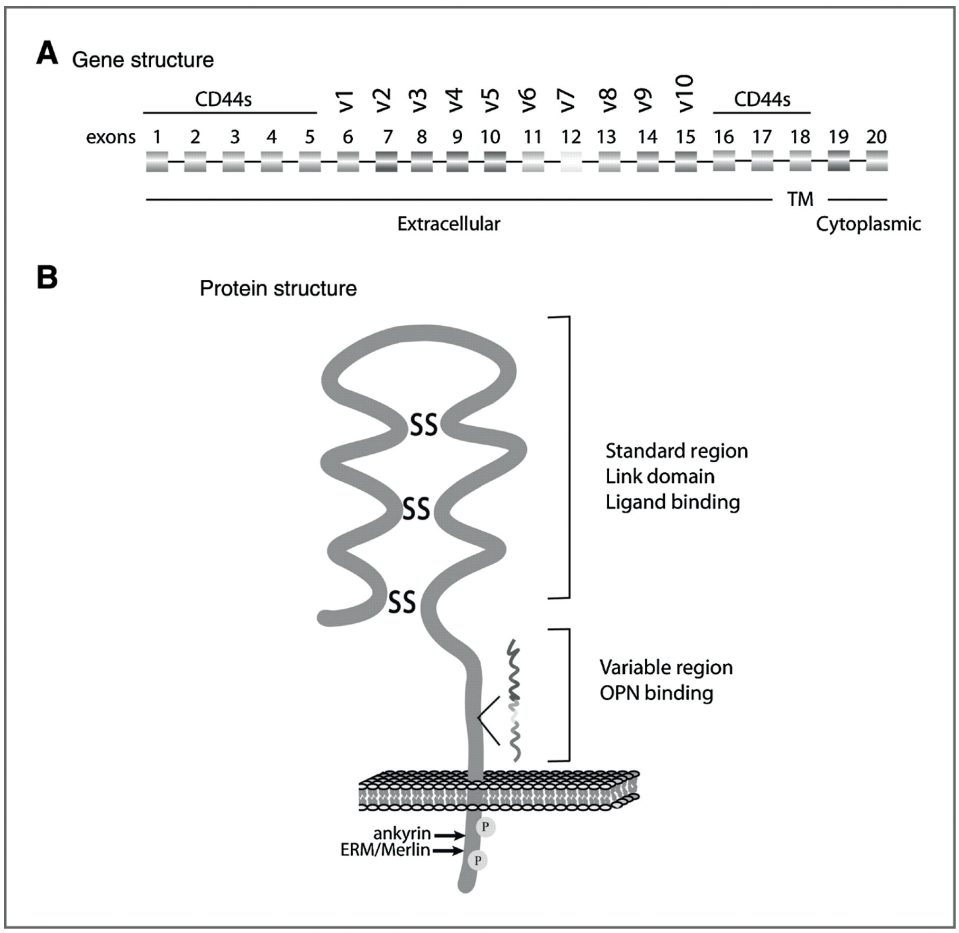

Figure 1. (A) The gene structure of CD44. The gene consists of 20 exons, where exons 1-5 and 16-20 are constitutively expressed and exons 6-15 are alternatively spliced (termed variant exons v1-v10). Exons 1-17 constitute the extracellular regions, exon 18 the transmembrane region and exons 19-20 the cytoplasmic tail. (B) Overview of the protein structure of CD44. The extracellular regions consist of the standard region binding to hyaluronic acid and the variable ligand-binding region and the cytoplasmic tail involved in the cytoskeletal organization and in mediating downstream signaling of CD44. Reprinted from "Molecular Cancer Research, Copyright 2011, Dec;9(12):1573-86, Jeanne M.V. Louderbough, Joyce A. Schroeder, Understanding the dual nature of CD44 in breast cancer progression" with permission from AACR.

and in mediating downstream signaling [127, 129]. Extracellular cleavage of CD44 can occur by membrane type 1 MMP or ADAM proteases and intracellular cleavage of the cytoplasmic tail is accomplished by γ-secretase [130]. Furthermore, the molecule can be both N-glycosylated and O-glycosylated [131]. The extracellular domain of CD44 can interact with

various components of the extracellular matrix such as hyaluronan, collagen, laminin and fibronectin, which all can be associated to matrix dependent cell migration *in vitro* [129].

Hyaluronan (or hyaluronic acid), which is a glycosaminoglycan, is the main ligand for CD44 that can bind to all isoforms of the molecule although the binding capacity can be altered due to post-translational modifications at the ectodomain of the CD44 protein [132]. In addition to hyaluronan, chondroitin sulfate and heparan sulfate are glycosaminoglycans that can interact with CD44 and the binding is dependent on inclusion of alternatively spliced exons [133].

Upon binding of hyaluronan to CD44 in tumor cells, it has been demonstrated that signaling cascades associated with receptor tyrosine kinase (RTKs) activation are initiated, which are suggested to be mediated by both CD44s and CD44v (that include variant exons) isoforms acting as co-factors of ERBB RTK family members. Examples of activated signaling pathways are the phosphatidylinositol-3-kinase (PI3K/Akt) and the mitogen-activated protein kinase (MAPK), which block apoptosis and promote cell survival tumor growth [127, 134]. Increased levels of hyaluronan have been suggested to contribute to resistance to chemotherapeutic drugs like doxorubicin and methotrexate and are possibly caused by hyaluronan/CD44 mediated signaling. Therefore, inhibition/disruption of hyaluronan binding or CD44 blockade could potentially improve the efficacy of this type of treatment [127, 135, 136].

However, since the CD44s molecule is widely expressed in different tissue types, it could be of interest to specifically target the CD44v isoforms in cancer therapy. Some of the CD44 variants that bind specific ligands or are required for signaling activation are mentioned here. For example, the isoforms containing the variant exon v6 is required for c-Met activation by its ligand hepatocyte growth factor/scatter factor (HGF/SF), which leads to activation of the Ras/MAPK pathway [137]. Moreover, it has been suggested

that Ras activation stimulates the transcription of alternatively spliced variants of CD44, and after translation, these CD44 variants can form new complexes with tyrosine kinase receptors and thereby the Ras-dependent proliferation of the cells can be sustained in a positive feedback loop [138]. The CD44 isoforms with the variant exon v3 included can undergo heparan sulfate modification, and can interact with heparin-binding epidermal growth factor-like growth factor (HB-EGF) and basic fibroblast growth factor (bFGF) [139], which leads to activation of ERBB4 [127].

Another ligand to CD44 is osteopontin, whose binding seems to be dependent on exon v3 and v6 [140]. Traditionally, CD44/osteopontin signaling is associated to stimulation of cell motility and chemotaxis [141]. Recently, it was also shown that osteopontin/CD44 signaling promotes cell growth and maintains stemness in glioblastoma [142]. This was mediated through γ-secretase dependent proteolysis of CD44 and induced cleavage of the intracellular domain of CD44, leading to enhanced HIF-2α activity.

The cytoplasmic tail of CD44 provides a link to the cytoskeleton through interactions with ankyrin and ezrin-radixin-moesin (ERM) proteins [127, 129, 143]. For example, ankyrin is involved in hyaluronan dependent cell adhesion and motility and ERM proteins are involved in the cell migration and formation of protein-complexes in the plasma membrane [144, 145].

RNA splicing

During transcription in the nucleus, immature RNA (pre-mRNA) is formed and it has to be further processed into mature messenger RNA (mRNA) before it is transported into the cytoplasm for translation. Processing of the pre-mRNA includes adding a 5′-cap (7-methyl-guanosine triphosphate) to the 5′-end and a 3′-polyA-tail of varying length to the 3'-end and importantly the intronic regions are spliced out. The splicing can be constitutive, which involves only constitutively spliced exons (CSEs), or the selective inclusion or exclusion of alternatively spliced exons (ASEs) occurs, a process termed

alternative splicing. Thus, different mRNA sequences can be formed out of the same pre-mRNA (*i.e.* cis-splicing) [146, 147]. Alternative splicing is estimated to occur in about 92-95% of the genes in the human genome and the frequency of alternatively splicing is associated with organismal complexity (on average seven mRNA isoforms for human) [148, 149]. Splicing is often tissue-specific and is commonly deregulated in cancer [150, 151].

In cancer, mutations can occur in splice sites at exon-intron junctions or in exonic or intronic regulatory elements in addition to genes that encode splicing factors. In the *BRCA1* gene, were germline mutations predispose for hereditary breast cancer, it is also known that mutations of putative splicing enhancer motifs can lead to exon skipping, and thereby change the splicing pattern [152-154]. Recently, it was shown that either by replacing a single motif with all possible hexameres or all possible single nucleotide variants, consecutively, within exon 18 (in *BRCA1*), resulted in distinct effects on transcript abundance. By using this approach for gene editing (the CRISPR/Cas technology) in different genes/exons, the prediction of function of mutations in regulatory elements and other parts of the genome could likely be enhanced [155].

Interestingly, it has recently been shown that silent mutations are more common in oncogenes compared to a set of genes with no association to cancer (1.23 to 1.30-fold enrichment), however, this enrichment tendency was not observed in a set of tumor suppressor genes. In oncogenes, the enrichment of silent mutations was seen near exon-intron boundaries suggesting that there is a selection for mutations in splicing factor-binding sites in pre-mRNA. These somatic mutations also seem to be reflected in transcript level variability when evaluated in RNA-sequencing data [156].

It has also been found that alternative splicing in certain genes can be linked to distinct histone modifications, which through a chromatin-binding protein recruits splicing regulators [157]. Interestingly, it has also been demonstrated that high levels of trimethylation of H3K9me3 is a feature of the alternative

exons in several genes, for example CD44. The chromodomain protein HP1γ facilitates inclusion of the alternative exons in CD44 through a mechanism that slows down the RNA polymerase II elongation rate [158].

It is also known that splicing of the v5 exon in CD44 is regulated by the splicing factor Sam68 (*KHDRBS1*) in cooperation with the splicing coactivator SRm160 (*SRRM1*) [159, 160]. Recently, it was suggested that splicing of v10 seems to be regulated by hnRNP L [161]. Downregulation of the splicing factor epithelial splicing regulatory protein 1 (*ESRP1*), is proposed to promote the splicing of CD44s [162, 163].

Origins of mutations and structural variants in breast cancer

Changes in the karyotype of cancer cells

Several mechanisms are known to give rise to changes in the karyotype of cells in cancer, which could lead to deletions, inversions, translocations, and other types of complex rearrangements that result in aberrant fusions between disparate positions within and between nonhomologous chromosomes. Changes in copy numbers in the genome are also common and appear to be associated with genes that favor proliferation or survival in cancer. Many of these abbreviations derive from defects in the mitotic apparatus and its regulators, for example, centrosomes and proteins that are associated with connection of spindle fibers with chromosomal kinetochores [75].

Breakage-fusion-bridge can occur if two sister chromatids both have unprotected ends in terms of eroded telomeres and thus end-to-end fusion between the two chromatids is possible. During the anaphase of mitosis the fused sister chromatids form a bridge between the two poles of the mitotic spindle. Being pulled in opposite directions during chromosome segregation will cause the two sister chromatids to break apart from each other at some intervening point, resulting in two daughter cells with non-homologous chromosomes. This process can continue with a fusion to another atelomeric chromosome during subsequent mitosis and these breakage-fusion-bridge cycles usually results in genomic rearrangements in terms of translocations [164].

Nondisjunction means that the sister chromatids fail to separate during the chromosome segregation in mitosis. Consequently, one of the daughter cells may become haploid for this chromosome and the other cell triploid. Possibly, a chromatid may also fail to attach to the spindle fiber, which leads to loss of that chromosome in the daughter cell [75].

Spindle assembly checkpoint (SAC), is designed to halt the progress into anaphase if the spindle fibers are not properly attached to all kinetochores of all chromatids. However, in cancer cells this control mechanism does not work properly, and individual kinetochores can be associated with too many spindle fibers. For example, in *merotely*, one single chromatid becomes associated to spindle fibers in opposite directions, which could lead to aneuploidy karyotypes [75, 165, 166].

Introduction to DNA damage

Traditionally, most research on DNA damage has been focused on the effect of exogenous carcinogenic compounds, but during the last decades it has been realized that also endogenous processes in the body eventually can lead to DNA damage. These are natural biochemical processes that unless efficiently repaired can lead to mutagenic events and can therefore be carcinogenic. Interestingly, it is actually believed that endogenous processes contribute to a much larger extent to development of cancer, compared to exogenous agents [75].

Glutathione S-transferases (GSTs) is a group of enzymes that by linking of gluthathione to potential toxic electrophilic compounds protects the DNA against many carcinogens (both endogenous and exogenous agents). However, high levels of GSTs appears to have a role in drug resistance and inhibition of these enzymes could lead to sensitization of tumor cells to anticancer drugs [75, 167].

Endogenous DNA damage

Depurination can be described as a loss of an adenine or a guanine base, and this is occurring in the presence of hydrogen and hydroxyl ions. *Depyrimidation* includes loss of a thymine or cytosine base and is much less frequent than depurination. In one single human genome, the steady-state level of base-free nucleotides is estimated to range from 4000 to 50,000 [75].

Deamination is a biochemical reaction that removes an amine group from a molecule and this occurs at all DNA bases, but at significantly different rates. The most common deamination reactions are *5-methylcytosine→thymine*, which occurs at CpG dinucluotides and seems to be a natural process whose accumulated burden may be related to age and gives rise to C/G→T/A transitions [168]. Another less common deamination process is *cytosine→uracil*, which leads to either C/G→T/A or C/G→G/C base replacements, and is thought to be catalyzed by members of the cytidine deaminase family, including activation-induced cytidine deaminase (AICDA) which shows a preference for cytidines flanked by a 5′ purine (A or G bases) [169], and the apolipoprotein B mRNA editing enzyme, catalytic polypeptide (APOBEC) enzymes of which some show a preference for the TpC sequence context [170-172]. *Adenine* can also deaminate to *hypoxanthine* (but at a much lower rate than that of the cytosine deamination), which can give rise to A/T→G/C transitions during replication and moreover, *guanine* can occasionally deaminate to *xanthine*, but this appears to be a rare event [173, 174].

Another source of endogenous DNA damaging agents are the free radical species (*e.g.* reactive oxygen and nitrogen oxide) that are generated as by-products of normal cellular metabolism during inflammatory response and apoptosis, but also by exposure to exogenous ionizing radiation, and their interaction with DNA can give rise to many different oxidative DNA base lesions [175]. One of these is the 8-oxo-2′-deoxyguanosine lesion, which leads to G/C→T/A transversions in the context of GpGpG sequence [176].

Exogenous DNA damage

Many different chemical compounds are known to cause DNA damage. However, often the specific signatures assigned to these compounds are not well known and therefore remain to be investigated [177].

Intercalating agents such as *benzo[a]pyrene*, which is a carcinogen in cigarette smoke, give rise to exogenous DNA damage in terms of G/C→T/A transversions and have a preference for methylated CpG dinucleotides [178-180]. Paradoxically, chemotherapeutic drugs can in some cases lead to a second cancer due to their DNA damaging properties. Drugs like the daunorubicin and epirubicin (anthracyclines) that are commonly used for cancer treatment could cause DNA damage by intercalation of double-stranded DNA [181]. Furthermore, chemotherapeutic agents like cyclophosphamide and temozolomide give rise to C/G→T/A transitions that is caused by alkylation of guanine and formation of intrastrand and interstrand crosslinks [182, 183]. A similar mechanism of action appears to be valid also for cisplatin, which also could cause bulky adducts interfering with DNA replication [184, 185].

Physical damage could result in exogenous DNA damage and the most well known example is non-ionizing UV radiation, which affects pyrimidines (by formation of covalent bonds between neighboring pyrimidines) with a predominance for C/G→T/A mutations and also CC/GG→TT/AA double substitutions which are characteristic features of cutaneous cancers that are associated with UV exposure [186, 187].

Mutation signatures in breast cancer

Previously, the most common way of describing somatic base substitutions was to use the mutational spectra of C/G→A/T, C/G→G/C, C/G→T/A, T/A→A/T, T/A→C/G and T/A→G/C, and as described above certain types of base replacements have been associated to both exogenous and endogenous DNA damage mutation patterns. However, it is also clear that the context of each type of base replacement is of importance [188], and has therefore been taken into consideration in more recent studies of various tumor types. Advances in high-throughput sequencing platforms have enabled analysis of large data sets of different tumor forms and by using computational methods it has been possible to define distinct mutational patterns and associate these to

different tumor types and underlying mechanisms of both exogenous and endogenous DNA damage [168, 171, 172].

In breast cancer, specific signatures for mutational processes have been suggested and the most common signatures (Signatures 1B, 2 and 3) involved C/G→T/A, C/G→G/C and C/G→A/T replacements with an overall prominence for C/G→T/A transitions [168]. Signature 1B is associated to age at diagnosis and this can be explained by C/G→T/A substitutions at CpG islands that inherently increase with age. Signature 2 is depicted by both C/G→T/A and C/G→G/C substitutions, and this pattern can be associated to mutagenic activity of APOBEC enzymes. Recently, it was proposed that breast tumors belonging to the HER2-enriched subtype often are characterized by the APOBEC mutational patterns in the context of the TpC sequence [171, 172]. Notably, Signature 3 is defined by an overall enrichment for C/G substitutions and this has been associated with cases harboring *BRCA1* and *BRCA2* mutations [168]. Two other less abundant mutational signatures (Signature 8 and 13), have also been associated with breast cancer and, similarly to Signature 2, Signature 13 could be associated to APOBEC editing activity, while Signature 8 was associated to C/G→A/T replacements with some extent of transcriptional strand bias [168] and could possibly be caused by free radicals species or ionizing radiation.

Of the APOBEC enzymes, *in vitro* experiments support that APOBEC1, APOBEC3A and APOBEC3B could be involved in the mutational process of breast cancer [170, 189, 190]. The reason for this is largely unknown, but it is hypothesized that the immune response to exogenous viruses or retrotransposones leads to activation of APOBEC enzymes that subsequently results in chronic DNA damage [168]. The APOBEC enzyme family has also been proposed to have a role in localized substitution hypermutation in the genome, also termed kataegis (thunderstorm in Greek), characterized by clusters of C/G→T/A and C/G→G/C substitutions in the context of at TpCpN trinucleotides and is sometimes associated with genomic rearrangements [170].

Repair mechanisms in DNA

At DNA replication during the S-phase of the cell cycle, occasionally either incorrect or chemically altered bases are introduced in about 1 out of 10^5 polymerized nucleotides. DNA bases can undergo chemical alterations either spontaneously or by chemical species or physical mutagens as described above. The proofreading mechanism (3'-5' exonuclease activity) of DNA polymerase reduces the frequency of inadvertently introduced bases to about 1 out of 10^7 nucleotides and hence they are not completely eliminated. However, these apparently random errors together with more specific endogenous and exogenous DNA damage are the reasons to why mechanisms for DNA repair are needed.

Repair enzymes can restore chemically altered DNA bases and this can be described as enzyme-catalyzed reversal of the chemical reaction that initially introduced the modification. An example is the O^6-methylguanine-DNA methyltransferase (MGMT), that removes ethyl and methyl adducts from guanine. Repression of the MGMT system could potentially have implications in treatment with chemotherapeutic drugs [75, 191].

Mismatch repair (MMR) is an excision repair process that primarily removes mismatched bases or insertion/deletion mispairs introduced in DNA during replication that was missed by the proofreading activity of the DNA polymerases, often in mono- or dinucleotide repeats. This lowers the mutation rate to about 1 per 10^9 nucleotides. Two components, MutSα (a heterodimer of MSH2 and MSH6) and MutLα (a heterodimer of MLH1 and PMS2) initiate the MMR repair. MutSα locates the actual mismatch and MutLα recognizes the recently synthesized strand by scanning for single-stranded nicks (possibly together with under-methylation of that strand). Excision of DNA occurs between the mismatch and a nearby nick by the MMR and is triggered by the MutLα complex. Subsequently the gap is repaired by DNA polymerase Pol δ. Deficiencies in the MMR system results in a high mutation load in highly repeated sequences (*i.e.* microsatellite repeat sequences) in the genome,

a situation called *microsatellite instability*, which is prevalent in some cancers such as colorectal cancer but is rare in breast cancer [177, 192].

Base excision repair (BER) is initiated by DNA glycosylases, which recognizes a single or a few chemically altered or inappropriate bases. In the short patch repair of BER, *e.g.* uracil is removed from the deoxyribose-phosphate backbone of the DNA by uracil DNA-glycosylase (UNG), and results in an abasic site (AP, *i.e.* an apurinic or apyrimidinic site), which is cleaved by AP endonuclease and the 5′-deoxyribose-phosphate (dRP) residue is removed by a dRP lyase and the nucleotide gap is subsequently repaired (usually with a C) by DNA polymerase Pol β and ligase activity. After deamination of 5-methyl cytosine, the resulting T base can also be excised by a certain T/G DNA glycosylase. In the long patch repair of BER, the strand displacement polymerases Pol δ or Pol ε may extend the 3' strand with 4-7 nucleotides, after AP endonuclease cleavage and initiation by Pol β. Base excision repair can for example correct deaminated, alkylated or oxidized bases (*e.g.* 8-oxo-2′-deoxyguanosine). However, replication before completion of repair potentially leads to introduction of mutations [174, 177, 193].

Nucleotide excision repair (NER) can remove various helix distortion lesions introduced by for example cisplatin or repair damage caused by ultraviolet radiation at dipyrimidine sites and is recognized either during global genome repair (by XPC-RAD23B) or during transcription. The helicases XPB or XPD unwinds the DNA and the damaged oligonucleotide lesion is cut by exonuclease and then the introduced nucleotide gap (27-29 nucleotides) is filled by typically Pol δ or Pol ε. When nucleotide excision repair is coupled to transcription (*transcription coupled repair*, TCR), a DNA lesion has stalled the transcription and hence the DNA damage is more efficiently repaired on the transcribed strand than on the non-transcribed strand, and this result in so-called transcriptional strand bias [177, 194].

Error-prone repair occurs during replication when an advancing replication fork encounters an unrepaired DNA lesion, and several error-prone human

DNA polymerases (or bypass polymerases) are responsible for this. Some polymerases can synthesize nucleotides to a growing strand, even if the complementary strand is missing, and another type can extend the DNA strand using a misincorporated base as primer. A third type of enzyme can incorporate a base when the corresponding base on the opposite strand carries a bulky adduct. Seemingly, the error-prone repair does not always restore the wild-type nucleotide sequences [75, 195].

Homology-directed repair (HDR) (also termed homologous recombination-mediated repair, HR) is also an alternative for DSB repair, which occurs in late S or G2 phase of the cell cycle, when a double-stranded copy (*i.e.* a sister chromatid) of the sequence is available. For example, a nick on the template strand (causing DSB) leading to a collapse of the replication fork during replication can induce homology-directed repair. HDR begins with resection by an exonuclease of one of the two DNA strands, at each of the ends formed by a double-strand DNA (dsDNA) break (5′ends). Subsequently, strand invasion by each of the resulting single DNA strands (3′ends) of the complementary sequences in the unwound sister chromatid occurs. This is followed by strand extension in a 5′-to-3′-direction by DNA polymerase and release of the newly synthesized strands that then are paired, ligated and, after the repair, the helix is reconstructed. Initiation of HDR can also occur by covalent inter-strand cross-links in the DNA or be triggered by stalling of the replication fork due to a bulky adduct on the leading strand. Hence, the lagging strand is used as template strand during repair, eventually leading to replication fork restart and bypass of the bulky adduct lesion [75, 196, 197].

Non-allelic homologous recombination (NAHR) is another type of homology-directed repair of dsDNA breaks that can occur between two DNA loci with high sequence similarity, which are not alleles. It has been demonstrated that low copy repeats (LCRs or segmental duplications) or transposable elements (TEs) can mediate especially recurrent deletions and translocations by non-allelic homologous recombination [198, 199].

Non-homologous end-joining (NHEJ) repairs DNA double-strand breaks most often in the G_1 phase of the cell cycle (*i.e.* when the sister chromatids are not available) by utilizing microhomologies at the free ends of the DNA strands to guide in the repair process. These microhomologies are often present in single-stranded overhangs on the ends of double-strand breaks and if they are perfectly matched, the break could be repaired correctly. However, ionizing radiation or enzymes that cleave DNA usually result in degradation at the DNA break and, therefore, it cannot be directly ligated. Instead, end-trimming and synthesis of new bases are crucial, but can potentially give rise to mutations (often small insertions and deletions of 1-4 bp) or even translocations. NHEJ is also initiated upon telomere de-protection to promote the formation of chromosome end-to-end fusions. In NHEJ repair, the Ku70–Ku80 heterodimer binds to DNA ends and recruits several factors like the DNA-PKcs–Artemis nuclease and DNA polymerases μ and λ. These proteins process the broken ends in preparation for ligation by DNA ligase IV–XRCC4 [177, 196, 200, 201].

Microhomology-mediated end joining (MMEJ) can also repair double strand breaks by using pairing of microhomologous sequences (5-25 nucleotides). But unlike NHEJ, this mechanism is independent of the Ku proteins. MMEJ is an error-prone pathway and always results in deletions of variable size and is frequently associated to chromosome translocations [201, 202].

Interestingly, a certain type of chromosome shattering was discovered a few years ago, "*chromothripsis"*, and this phenomenon can be described as localized firestorm-like densities of chromosomal rearrangements including deletions, tandem duplications and inversions [203]. It is believed that chromothripsis arise at a single catastrophic chromosome breakage event and that perhaps non-homologous end joining or microhomology-mediated end joining repairs the damage, however, far from error-free [203, 204].

DNA repair proteins deficiencies and implications for therapy

The *BRCA1* and *BRCA2* genes are often mutated in breast cancer, both in hereditary tumors through germ-line mutations with loss of the second allele and in sporadic tumors through somatic mutations. These two genes are involved in various DNA repair mechanisms and are both often found in large protein complexes together with other proteins like RAD50/Mre11 and RAD51 in the cell nucleus. These complexes have been shown to cluster at stalled replication forks and at dsDNA breaks [205-207]. Moreover, genetic deficiencies in the *BRCA1* and *BRCA2* genes can lead to chromosomal translocations due to improperly repaired dsDNA breaks mediated by other repair mechanisms which are less accurate than HDR, and eventually can lead to breast or ovarian cancer.

Therefore, targeted therapies directed against these types of alternate repair mechanisms seem to be an efficient strategy in the treatment of patients. Poly (ADP-ribose) polymerase (PARP) inhibitors are used for this purpose in advanced cancer (*e.g.* olaparib) [208, 209]. Various mechanisms on how these inhibitors target the cells have been suggested [210, 211]. For example, since the PARP1 protein repairs single-strand breaks (nicks) in the DNA, inhibition of this process could lead to dsDNA breaks that eventually kill the cell. A more recent study suggests that PARP inhibitors initiates trapped PARP protein-DNA complexes and are therefore highly toxic since the DNA replication is blocked; moreover, the potency of trapping differs a lot when comparing different inhibitors [212]. It has also been proposed that after stalling of replication forks, PARP in combination with homology directed repair are essential to restart the replication [213]. Another chemotherapeutic drug that has shown promising results in the neo-adjuvant treatment of patients with *BRCA1* mutations is cisplatin, and this drug could possibly be used in advanced disease together with PARP-inhibitors [214, 215].

Breast cancers arising in *BRCA1* mutation carriers are often characterized by a triple-negative phenotype [216]. However, *BRCA1* is infrequently mutated in

sporadic triple-negative tumors. Instead alternative mechanisms for suppression of *BRCA1* in triple-negative tumors have been suggested such as promoter hypermethylation [217, 218], which would implicate that other tumors of this subtype would benefit from therapy with PARP-inhibitors and cisplatin. Interestingly, overexpression of HORMAD1 in tumors seems to select for patients that respond well to this type of treatment. High expression of HORMAD1 was suggested to diminish the homology-directed repair mechanism through suppression of the BRCA1-associated protein RAD51 [219].

Moreover, in cells with compromised homology-directed repair subjected for DNA damage, the DNA repair pathway is sometimes replaced by the more error-prone non-homologous end joining (NHEJ), and therefore, targeting of the polymerase responsible for NHEJ could be a therapeutic approach. Recently, it was shown that the polymerase Pol θ (encoded by POLQ) is active in this process and that inhibition of that enzyme suppresses alternative NHEJ at dysfunctional telomeres and hinders chromosomal translocations at non-telomeric loci [220, 221].

The concept of circulating tumor DNA

Circulating cell-free DNA

The existence of circulating cell-free DNA (cfDNA) was first described in 1948 [222]. In 1977, cfDNA was reported to be elevated in breast cancer patients [223], and studies in the coming decade further associated cfDNA to the stage of malignant disease in several cancer types. The clinical relevance was further corroborated by the detection of mutated RAS genes in the blood of cancer patients in 1994 [224, 225], establishing conclusively the tumor cell origin of circulating tumor DNA (ctDNA). In 1997, circulating fetal DNA was discovered and is now widely used as a prenatal based non-invasive diagnostic method with high sensitivity and specificity [226-228]. Fractional concentrations in maternal plasma seem to be between 3-6% and immediately following delivery, the half-life is less than 1 hour [229-231]. The circulating fetal DNA is highly fragmented and the most common length is around 166 bp, while maternal circulating DNA seems to contain longer fragments [232, 233].

Circulating cell-free DNA can originate both from normal and tumor cells and herein, tumor specific cell-free DNA is termed ctDNA and cfDNA refer to cell-free DNA derived from normal cells. Importantly, healthy individuals always have normal cfDNA in the blood, and cancer patients can have ctDNA mixed with cfDNA in different fractions. The mechanisms behind the appearance of cfDNA or ctDNA in the blood are not fully understood, but it can probably be derived from both apoptotic and necrotic cells and possibly it can also be released by living cells [234] (Figure 2). It is known that degraded ssDNA and dsDNA are present in exosomes representing the genomic DNA, which are likely to be detected in blood after release from both normal and tumor cells [235, 236]. Moreover, during apoptosis endogenous endonuclease activity leads to excision of nucleosome chains of chromatin DNA and this results in fragments of approximately 180 bp and multiplies thereof [237],

which partly consists of "linker DNA" that joins adjacent nucleosomes [238]. DNA from necrotic cells appears to vary more in length compared to DNA from apoptotic cells [238]. Recently, it was shown in hepatocellular carcinoma that the relative abundance of fragments <166 bp was higher in ctDNA compared to cfDNA which suggests that these fragments were derived from the process of apoptosis. However, it was also found that fragments <166 bp could consist of mitochondrial DNA [239]. Moreover, longer ctDNA fragments of ≥166 bp were observed in the same study, in particular if the fraction of ctDNA was low. Notably, hepatocellular carcinoma often carries distinct genomic aberrations, which were used to distinguish between ctDNA and cfDNA [225, 240].

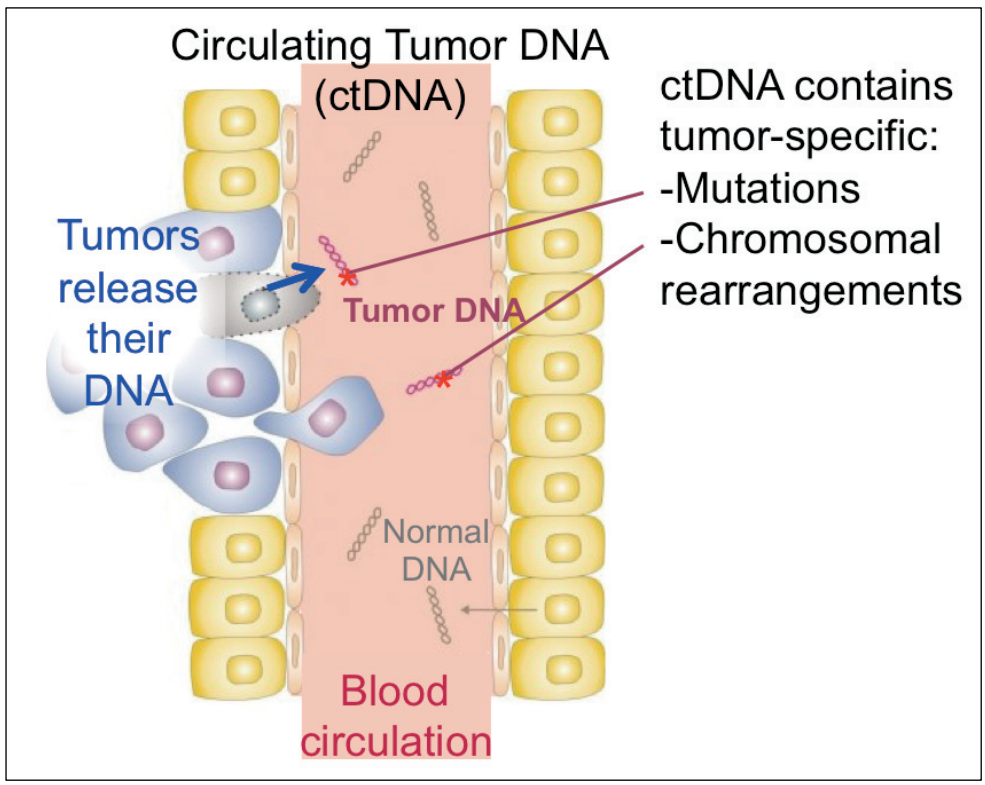

Figure 2. Circulating tumor DNA (ctDNA) can be detected in blood of cancer patients and contains tumor-specific sequences with mutations or chromosomal rearrangements. The ctDNA origins from apoptotic or necrotic cells, or could possibly be released from living cells. Monitoring of the ctDNA levels in the blood can be utilized for diagnostic purposes or as a tool for investigation of therapy response.

Due to the intrinsic heterogeneity in breast cancer, identification of ctDNA is more challenging. There are indications that the integrity of ctDNA is reduced in patients with breast malignancies [241] and an enrichment of the size range between 85-250 bp has been observed [242, 243]. However, also larger fragment sizes seem to be present in ctDNA [244]. After size fractionation of different molecular DNA weights (<1000 bp versus >1000 bp) from breast cancer blood samples, the shorter fragments were enriched for ctDNA, indicating that cfDNA in general contains longer fragments than cfDNA [245]. Differences in size distribution comparing cfDNA and ctDNA have also been observed in other tumor forms [246].

Moreover, when comparing different types of solid tumors, the fractions of ctDNA (compared to cfDNA) seem to vary considerably, with very low levels detected in glioblastoma [247]. The half-time of ctDNA has been reported to be less than two hours for colon cancer patients [248, 249]. Still, more detailed knowledge about the source, sizing, and clearance rate of ctDNA in different types of tumors is needed. For example, studies on cfDNA and ctDNA length may be confounded by differences in source (e.g. plasma versus serum) and the degree of normal cell lysis during preparation, which can falsely increase the level of background wildtype sequence as well as high molecular weight DNA.

Liquid biopsies

Liquid biopsies are non-invasive blood tests that could detect either ctDNA that is shed into the blood or circulating tumor cells (CTCs). A bone marrow sample could also be considered as a liquid biopsy, in which detection of disseminated tumor cells (DTCs) adds information on the prognosis for patients with early breast cancer and possibly on the efficacy of adjuvant therapy [250, 251]. However, in clinical practice, bone marrow aspirations are associated with more risks and inconvenience for the patient compared to a simple blood test and there is also a need for a standardized protocol for analysis of DTCs [252].

CTCs are cells from a primary or metastatic tumor that have entered the bloodstream. Detection of CTCs can be either label-dependent or label-independent. Methods that utilize the label-dependent enrichment (*e.g.* the benchmark CellSearch system), usually defines CTCs as positive for epithelial cell adhesion molecule (EpCAM) and these cells are magnetically separated fromm whole blood [253]. The presence of CTC in peripheral blood can be associated with a poorer prognosis for both early and metastatic breast cancer [254, 255]. The presence of CTCs as an indicator for disease progression after one cycle of initial therapy, followed by switch of chemotherapy seem to have no effect on survival or time to progression [256]. Moreover, not all epithelial tumor cells stain positive for EpCAM and these will be missed by the CellSearch system. Therefore additional markers for CTCs are needed [257]. Conversely, the label-independent detection is based on invasive capacity, size and density selection, but these methods are presently not widely used [253].

In the last years, quantification and genetic analysis of ctDNA isolated from blood samples have arisen as promising tools for examination of the disease state in cancer patients. The concentration of ctDNA seems to reflect the tumor burden, and moreover, the presence of druggable mutations can be determined at diagnosis. By monitoring patients during therapy acquired resistance mutations can be detected [258-261]. In a recent study it was also observed that ctDNA is often present, even if no circulating tumor cells were identified [247].

ctDNA as a predictive biomarker

On average, the total levels of circulating DNA are higher in patients diagnosed with primary breast cancer compared to healthy controls [262]. Moreover, the fractions of ctDNA found in blood are significantly higher in metastatic breast cancer patients compared to patients with localized disease [242, 247].

It is well established that all breast tumors harbor genetic aberrations, both in terms of mutations and genomic rearrangements [53, 54, 263, 264] and these have been utilized to detect ctDNA in breast cancer patients [260]. For example, it has been demonstrated that tumor specific rearrangements can be used to confirm the diagnosis of metastatic disease in breast cancer patients by blood sample analysis [260, 265]. In these studies, the rearrangements were identified by next-generation sequencing of DNA isolated from primary tumors followed by confirmation of breakpoints by PCR and sometimes the exact breakpoints were derived by sequencing of PCR amplicons at base-pair resolution. Quantitative real-time PCR or digital PCR was applied to get estimations on the concentration of ctDNA in blood plasma.

In metastatic breast cancer, whole exome sequencing has also been used to compare the characteristics of plasma derived ctDNA to DNA isolated from a synchronous tumor specimens, showing similarities in both copy number estimations and mutational analysis [259]. The allele frequencies of selected mutations were also followed over time and certain mutations appeared to be selected for during therapy. Moreover, a second whole exome analysis during the clinical course revealed mutations that possibly are associated with therapy resistance. This study demonstrates the feasibility of using mutational screening of ctDNA as a monitoring tool during therapy. Similar conclusions could be drawn in another study by monitoring of somatic genetic alterations in multiple plasma samples during targeted therapy in a patient with metastatic breast cancer [266].

In another study on metastatic breast cancer both targeted screening of mutations in recurrently mutated genes and whole genome sequencing to identify structural variants were performed to find tumor-specific variants suitable for ctDNA monitoring [267]. The selection of mutations or structural variants was based on genetic aberrations detected either in the primary tumor or from a site of metastatic disease. To investigate the disease state, quantification of ctDNA isolated from plasma was performed at multiple time points by using digital PCR and in a majority of cases decreased levels of

ctDNA was a sign on response to treatment. The prognostic value of increasing levels of ctDNA (over time) was found to be significant and showed an advantage over the enumeration of CTCs and the levels of the protein marker CA15-3.

A recent study demonstrated that circulating ctDNA, identified as point mutations or genetic rearrangements, could be detected in about 85% of patients with metastatic breast cancer and in about half of patients with localized disease by using digital PCR [247]. Moreover, across different types of cancers it was shown that the concentration of cfDNA similarly increased with stage of the disease. Thus, by using larger volumes of plasma to increase the amount of input ctDNA in digital PCR analyses, it is likely that higher percentages of patients with different stages of breast cancer could be monitored by detection of ctDNA.

To summarize, the methodologies described above could likely be used for monitoring of therapy response and to elucidate underlying mechanisms of resistance to therapy. Moreover, analysis of ctDNA could likely be used for early detection of advanced breast cancer and as a prognostic tool for patients with breast cancer. Analysis of ctDNA has been successfully used to monitor the disease state in other tumor forms, which further demonstrates its clinical promise [249, 268].

Aims of the thesis

The general aim of the thesis was to better characterize breast cancer on transcriptomic and genomic levels by investigating RNA splicing and mutations in tumors and the dynamics of circulating tumor DNA in blood.

Paper I: To study mRNA levels of CD44 isoforms in breast tumors and to correlate expression of respective isoform to molecular subtypes, clinical characteristics, and cancer stem cell phenotypes.

Paper II: To identify characteristic somatic mutations in a panel of basal-like breast cancer cell lines and to investigate the base replacement patterns and the genomic context of the mutations.

Paper III: To investigate the potential in using detection of tumor-specific chromosomal rearrangements present in circulating tumor DNA as a tool for early detection of symptom-free metastatic breast cancer and to evaluate its utility as a prognostic marker.

Considerations of appended papers

Overview of the main methods

Real-time PCR

Real-time PCR was developed as a method that continuously measures the rate of accumulation of amplified DNA during the PCR, which gives the ability to quantify the amount of DNA during the exponential phase of the PCR when none of the components of the reactions are limiting. Previously, different end-point methods were used to quantify the accumulated amount of DNA after a polymerase chain reaction (PCR). Since the yield of the amplified product might be influenced by small amounts of inhibitors and that the primers used at amplification could affect the efficiency of the PCR, these prior methods are now often regarded as unreliable for DNA quantification [269].

Using the real-time PCR technique, it is feasible to compare the efficiency between assays and to further optimize the PCR reaction. Real-time PCR can be used to quantify the abundance of particular DNA or RNA sequences, mutations or single nucleotide polymorphisms in for example clinical samples [269]. There are several different manufactures of the real-time PCR instruments, but in principal, they all use the same technology including detection of fluorescence in separate wells. In Paper I, we used the Rotor-Gene instrument from Corbett Life Science, which utilize centrifugal rotary design to minimize temperature gradients during the PCR.

To analyze real-time PCR data the manufacturer of the instrument usually provides a software program designed for this purpose. In such software, the baseline is defined as the accumulating fluorescent signal in the initial PCR cycles, which are considered to be beneath the limit of detection of the

instrument. The ΔRn values measure the increment of fluorescent signal at each time point, and are usually plotted versus the cycle number. A threshold has to be set on the basis of the baseline variability and it should be adjusted to include the region of exponential phase in the amplification. The first signal that is detected above the threshold for a sample is considered as a real signal and is defined as the threshold cycle (Ct). The Ct values are used in further calculations for processing of real-time PCR data. Importantly, in quantification assays (*e.g.* gene expression) the relative abundance in a sample is often calculated compared to a reference sample. To compensate for different amounts of input template, one or more endogenous controls must also be used. It is important that the endogenous controls are equally expressed (or abundant) within all analyzed samples. Gene expression analysis using real-time PCR is often referred to as real-time RT-PCR (reverse transcription polymerase chain reaction), since it includes a reverse transcription step converting mRNA into complementary DNA (cDNA) [270]. In Paper I, real-time RT-PCR was used to calculate the relative abundance of different isoforms of the CD44 molecule in different subtypes of breast tumors by using the delta-delta Ct method [271].

Another option for real-time PCR quantification is to use the standard curve method (including samples of known concentrations). The standard curve could be used to determine the absolute abundance of a sequence in a sample of an unknown concentration. However, the latter method is not suitable for high-throughput screening studies [270].

Droplet digital PCR

Digital PCR (dPCR) is a more recent technique for quantification of target sequences of interest. The method works by partitioning the input sample into numerous of individual PCR reactions, and only a fraction of these reactions will contain the target molecule (positive signal) while others will not (no signal; negative), which means that the read-out for each partition after the PCR reactions essentially is binary, *i.e.* digital [272]. In Paper III we have used

the Bio-Rad QX100 Droplet Digital PCR (ddPCR) instrumentation [273] for detection of low abundance tumor-specific DNA sequences in plasma samples from breast cancer patients.

Similar to a real-time PCR experiment, primers in combination with a labeled oligonucleotide probe are used in the setup of digital droplet reactions. However, the input sample is now divided into approximately 20,000 nanoliter-sized droplets by water-oil emulsion technology using the QX100 Droplet Generator. The droplets are then transferred to a 96-well plate and the PCR can be run in a regular high-performance thermal cycler. Following PCR amplification, which only occurs in droplets containing target DNA, the plate containing the droplets is placed in a QX100 Droplet Reader. The autosampler of the droplet reader picks up the droplets from each well of the PCR plate, and each droplet is spaced out individually for fluorescence reading. The detection system is two-color based (set to detect FAM and HEX [or VIC]) which enables multiplexed reactions if desired [273].

Because the partitioning of the PCR reaction, and thus the target templates, into droplets is random and follows a Poisson distribution, the count of positive and negative droplets can be used to calculate the absolute number of target molecules in the sample. Therefore, there is no need for an endogenous control or standard curves. Advantages of sample partitioning in combination with a digital read-out, are that factors like amplification efficiency and PCR inhibitors are of less importance. The digital droplet technology is suitable for quantification of absolute mRNA concentrations or absolute copy numbers of DNA in for example tumor specimens. Moreover the method gives a very good specificity and sensitivity at detection of low abundance molecules [273]. In Paper III, to enhance assay sensitivity and specificity and also reduce the need for per-assay optimization of thermocycling conditions, we designed a touchdown-PCR protocol [274], where the annealing temperature was decreased for every subsequent cycle in the initial part of the PCR program.

Probe technologies used in real-time PCR and digital droplet PCR

Both in the setup of a real-time PCR and in a digital PCR experiment, primers in combination with a labeled oligonucleotide probe are used. The probe consists of a single stranded sequence complementary to one of the strands in the targeted sequence. The probe usually consists of a fluorescent group at its 5'-end (*e.g.* fluorescein, FAM) and a quenching group at its 3'-end (*e.g.* tetramethylrhodamine, TAMRA). These two molecules have overlapping emission-absorption spectra, which means that the emission from FAM after excitement can be absorbed by TAMRA. However, the efficiency of this process is strictly dependent on the distance between the two molecules. During the PCR, the primers and probe binds to the amplicon during at each annealing step. During the elongation step, the probe is displaced at the 5'-end and subsequently the end with the fluorophore is cleaved by a *taq* polymerase. The probe is degraded and the fluorescence can be detected from FAM since the quenching effect of TAMRA is gone. The accumulation of DNA during PCR can be measured in real-time by detecting the fluorescence emitted from the degraded probes, whereas in digital PCR the post-PCR fluorescence emission is used as a read-out [270].

The TaqMan probe from Life Technologies is one example of a double-dye oligonucleotide probe that can be designed by using different fluorophores and quenchers. Traditionally, FAM has been used as fluorophore and TAMRA as quencher in this type of probe [270]. The TaqMan MGB probe is a further development of the original TaqMan probe and is always labeled with a non-fluorescent quencher (NFQ) at the 3'-end conjugated to a minor groove binder. The 5'-end can be labeled with different fluorophores, as for example FAM or VIC. Due to formation of stable duplexes with the template, the minor groove binder increases the melting temperature (T_m), allowing the use of shorter probes (<13 bp) [275]. In Paper I, TaqMan MGB probes were used (after reverse transcription of mRNA) in different exon-exon spanning assays to detect different isoforms of the CD44 gene.

Another type of probe system utilizes a double-quenched probe (provided by Integrated DNA Technologies), which has an internal quencher (ZEN) in close proximity (9 bp) to the 5'-end in addition to the 3'-end quencher (Iowa Black FQ) [276], and the 5'-end is labeled with the fluorophore (e.g FAM). This type of quenching is designed to give a lower background level of fluorescence. This probe system was successfully used in Paper III, by using the similar guidelines as recommended for the TaqMan probe design. The double-quenched probes generally need to be between 18-30 bp to have an optimal T_m. This issue is worth taking into consideration since a shorter probe length simplify the assay design of small amplicons, which means that the TaqMan MGB probe design could be an alternative choice for detection of short DNA fragments that is a common feature for circulating tumor DNA. However, in large-scale projects the double-quenched probes are considerably more affordable compared to the TaqMan MGB probe.

DNA sequencing

In 1977, a DNA sequencing technology was introduced that would revolutionize molecular biology and cancer research, the "Sanger" sequencing method [277]. Named after inventor, Fred Sanger who won his second Nobel Prize in Chemistry in 1980 for this work, the method utilizes modified di-deoxynucleotidetriphosphates (ddNTPs) lacking a 3'-OH group, which terminate the DNA strand elongation during PCR. Originally, four different PCR reactions per sample were needed, including only one out of four possible modified ddNTPs (ddATP, ddGTP, ddCTP or dTTP) per reaction, in addition to ordinary PCR reagents. This results in a number of DNA fragments of varying length in each reaction. By using gel electrophoresis, PCR products of different sizes can be separated and the DNA sequence can be resolved with respect to the relative positions of the different bands.

Later, the technology has developed into using various fluorescent tags to separate signals from different bases (dye-terminator sequencing), which means that one reaction is sufficient per sample [278, 279]. Automated multi-

sample sequencing instruments have been developed that utilize capillary electrophoresis for size separation, detection and recording of fluorescence results in chromatograms with individual peaks for each base in the sequence of interest [280]. A commonly used automated Sanger sequencing instrument is the ABI 3730xl from Life Technologies that has a 96-capillary array format and is capable of producing up to 96 kb in a 3 hour run, and fragments up to 900 bp can be sequenced. Sanger sequencing was used in the human genome project (HUGO) that sequenced the entire human genome and took several years to finish [281]. In Paper II, Sanger sequencing was used for validation of novel candidate mutations in different breast cancer cell lines.

High-throughput DNA sequencing

In 2007 Illumina Inc. acquired the company Solexa that had developed and commercialized the sequencing by synthesis (SBS) technology for short-read sequencing on a solid substrate. In 2008 Illumina released the Genome Analyzer II, and in 2009 announced a service for sequencing a human genome for $48000. Progressively, the cost of sequencing has been decreasing significantly and in 2014 Illumina launched machines that can sequence a whole genome for about $1000. Today, the Illumina sequencing technology is dominating the market of high-throughput sequencing and this technology has enabled generation of whole genome, exome, and RNA sequencing data in different types of large-scale studies.

The performance of Illumina sequencing varies depending on which sequencing system that is used. The Illumina Genome Analyzer IIx, can generate up to 95 GB of data in two weeks and in Paper II, this machine was used. In more recently developed machines, such as the HiSeq 2000 and HiSeq 2500 which were used in Paper III, generation of 300 GB data of reads up to 2×250 bp in one single run is possible. The HiSeq X Ten system launched last year can yield up to 1800 GB of data of reads up to 2×150 bp in a run that takes about 3 days. The MiSeq is an instrument that is more suited

for small-scale projects like targeted re-sequencing with potential clinical utility and which require a rapid run time.

After preparation of the input sample, a clustering step is included in the Illumina workflow where single stranded molecules in the library are hybridized to oligonucleotides attached to the glass surface of a sequencing flow cell [282]. The complementary strand of the hybridized strand is synthesized by a polymerase and the original template is denatured and washed away. There are two types of oligonucleotides attached to the surface, and during a step called bridge amplification, the free end of the newly synthesized strand falls over and attaches to the second type of oligonucleotide. Then the complementary strand is synthesized to form a dsDNA bridge. The two strands are denatured and the amplification process is repeated under isothermal conditions until a dense clonal cluster has formed. Subsequently, each cluster of dsDNA bridges is denatured, and the reverse strand is removed, leaving only the forward strand. In total, millions of clusters are formed, each cluster representing a single molecule in the original sample library.

At sequencing, a primer is hybridized to the forward strand, by binding to the adaptor added during the sample preparation. Illumina utilize sequencing by synthesis (SBS) technology, which means that a reversible terminator that is fluorescently labeled and bound to each dNTP is imaged, before it is cleaved to allow incorporation of the next base. All four terminator-bound dNTPs are present during each sequencing cycle to minimize incorporation bias. This method seems to handle homopolymer regions better than the 454 sequencing technology and the sequencing errors appears to be less frequent (1 substitution per 1000 bases) [283].

In Paper II, we sequenced matched tumor-normal cell line samples covering 6.5 Mbases of respective genome, including exons and flanking regions of 1237 genes. These regions were captured by using a custom SureSelect library together with the SureSelect target enrichment system from Agilent Technologies. First, the genomic DNA of the samples was sheared to an

average length of 180 bp followed by end-repair, dA-tailing and ligation of sequencing adaptors. Then size selection was done using Agencourt AMPure beads from Beckman Coulter to keep only fragments between 200-350 bp and PCR amplification of the adaptor-ligated genomic DNA was performed. The SureSelect library was used to capture the sequence of interest by hybridization of biotinylated cRNA baits of 120 bp in length to the targeted regions of genomic DNA. The targeted regions can then be selected by using magnetic streptavidin beads followed by PCR amplification and quantification before sequencing of the DNA.

The SureSelect system is also available as pre-designed libraries including all human exons and selected panels of for example cancer related genes. Other strategies for enrichment of DNA regions of interest are the SeqCap EZ Libraries from Roche NimbleGen that includes options for targeting whole exome, miRNA exons, untranslated regions or custom regions. Moreover, Illumina provides whole exome capturing kits and different cancer panels for targeted enrichment [284, 285]. It is not a trivial task to make a comparison between these different systems since many different parameters have to be considered like the sensitivity, sequencing depth, targeted sequence, density and length of baits as well as the quality of input DNA. The technology is still under continuous development, but it has been indicated that overlapping baits originally used by SeqCap EZ cover the largest fraction of the targeted regions with the least amount of sequencing. In Paper II, an end-to-end design was used, but if the study was to be repeated today an overlapping design would have been preferred. Conversely, the detection of indels could be more efficient using baits of approximately 120 bp (as for SureSelect), whereas, SeqCap EZ bait lengths varies between 55-105 bp. The selection of database used for bait design could also influence which regions that are covered, as the definition of coding sequence differs between the RefSeq and Ensembl databases [284].

In Paper III, whole-genome sequencing was performed on specimens from 20 patients and the TruSeq DNA Sample Preparation Kit from Illumina was used

for library construction. In short, the genomic DNA samples (2.4 μg of each) were first sheared to an average of 700 bp by using the S220 Focused Ultrasonicator Instrument from Covaris and 1 μg of each sample was used for library preparation followed by end repair, dA-tailing and ligation of paired end sequencing adaptors (TruSeq DNA adaptors). Each library was size separated by agarose gel electrophoresis and fragments between 550 and 950 bp were cut out and purified prior to PCR amplification. To increase the physical coverage, and thus the sensitivity to detect rearrangements for a given sequencing depth, the desired fragment length was increased significantly compared to the original protocol. In later versions of the protocol provided by Illumina, the PCR amplification has been omitted and size selection beads provided to replace the gel-based size-selection. These changes in protocol lower the hands-on time and also reduce biases introduced by PCR (*i.e.* in G/C-rich regions).

Analysis of sequencing data

The next generation sequencing methodology easily generates massive amounts of sequencing data, which means that the flow of data analysis has changed considerably since the Sanger sequencing era. Today both computational resources and bioinformatics expertise are necessary to perform these often large-scale projects. In total, DNA sequencing can give information on point mutations, indels, copy number variation and structural variants (Figure 3).

In Study II, the mutational analysis of next-generation sequencing data can be divided into the following main steps: read alignment, deduplication of reads, realignment and recalibration, variant calling, filtering of the called variants and determination of somatic variants. For alignment, we used the Burrows-Wheeler aligner (bwa) [286], which is a fast and accurate tool for alignment of short reads to the reference genome and it can handle alignment of gapped sequences. Picard Tools (http://picard.sourceforge.net/) was used to flag reads

with identical mapping positions (*i.e.* deduplication), likely to be PCR duplicates introduced before sequencing.

These are ignored in further analysis to reduce the impact of amplification biases in PCR. In the next steps, the Genome Analysis Toolkit (GATK) was used [287]. First, local realignment around indels was performed since indels within reads often lead to false positive SNPs at the end of sequence reads. To prevent this artifact, local realignment around indels is done using the local realignment tool in GATK. Next base quality score recalibration was performed for adjustment of the sequencing base quality scores. The variant calling of SNPs and indels were done with Unified Genotyper followed by variant filtering according to GATK v3 best practices recommendations. Unified Genotyper defined the genotypes in the tumor respective normal samples, and only variants homozygous for the reference allele in the normal sample and variants hetero- or homozygous for the variant allele in the tumor were considered as somatic. The informative read depths for the normal and tumor sample, respectively, were used to define "high confidence" somatic variants. Annovar was used to annotate all somatic variants for the tumor [288].

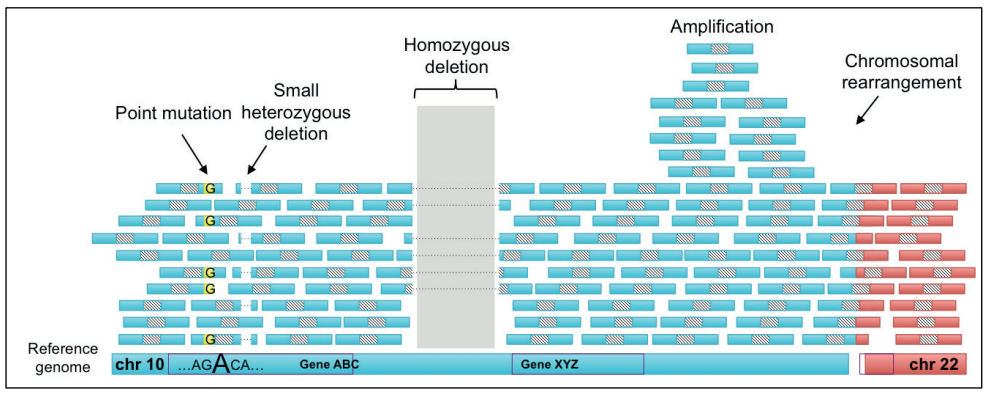

Figure 3. Genomic aberrations detected in next-generation DNA sequencing data. Often mutation screening is performed by using targeted resequencing at an average coverage >100x, and both point mutations and smaller indels can be detected in analysis of this type of data. Whole genome sequencing at significantly lower coverage can be used for detection of larger indels, amplifications and chromosomal rearrangements.

Copy number variation (CNVs) was derived by the CONTRA software using realigned data (see above) with matched tumor-normal samples as input. Minor modifications to default settings were done to improve the resolution at exon level [289].

In Paper III, alignment was performed with Novoalign from Novocraft Technologies, which is a highly accurate commercial mapping software. Soft-clipping was used to preserve reads that may contain sequences specific for translocations (*i.e.* only a part of the read was aligned to the reference genome). Deduplication was performed as described above. To identify chromosomal rearrangements, BreakDancer [290] was used with default options for discordant read-pairs (to predict translocations, duplications, inversions and deletions). Discordant read-pairs in predicted translocations were re-aligned to the reference genome using Novoalign, using an exhaustive re-alignment step. Discordant read pairs of translocations that became concordant after this step were discarded. To minimize the false-positive rate of predicted translocations, filtering on repetitive regions and sequence gaps (in the reference genome) and distance between ends of reads was performed. Moreover, rearrangements detected also in other tumors or normal samples were removed as they were assumed to represent likely false alignments or germline events.

To facilitate the primer design for short amplicons (due to fragmented ctDNA), the Splitseq pipeline was developed which utilize the potential breakpoint sequences in soft-clipped reads and also searches for these sequences in unmapped reads to finally reconstruct the exact breakpoint sequence, if possible.

Summary and discussion of papers

Implications of CD44 isoforms in breast cancer, Paper I

The aim of Paper I was to investigate the potential association between expression of distinct isoforms of the CD44 molecule and cancer stem cell markers, subtypes and clinical markers in breast cancer. This is of interest since CD44 has been associated with tumor progression and outcome of breast cancer. However, different studies have shown conflicting results and this could to some extent be related to which isoforms of CD44 that were investigated [123-125, 291]. Since the CD44 molecule is widely used as a stem cell marker, it would also be interesting to characterize which isoforms that are expressed in breast cancer cells displaying characteristics of cancer stem cells, as information on this is very limited in the literature.

The CD44 molecule can be alternatively spliced into several different isoforms and is often subject to post-translational modifications. In many studies on CD44, breast tumors have been stained with the CD44 standard (CD44S) antibody that binds to the constant extracellular region. In other words, the CD44S antibody most likely stains for many different isoforms with diverse biological properties. Other studies have used antibodies specific for different epitopes of CD44, for example CD44v6, and even if this isoform binds to the variable region of the molecule it is not specific for only one isoform.

In Paper I, we first compared the total gene expression (mRNA) of all CD44 isoforms to the protein expression of CD44 analyzed by flow cytometry, in breast cell lines of different molecular subtypes. An antibody binding to the constant region of CD44 was used to measure the fraction of CD44 positive cells. In general, a low gene expression was associated with a low protein expression (*i.e.* a low fraction of positive cells) and vice versa. However, one cell line showed a considerably higher protein expression than expected, compared to the gene expression levels. Possibly, this is caused by a post-

translational mechanism that inhibits degradation of CD44. However, there is a small risk that the analyzed transcripts do not correspond to all protein variants of CD44 detected by the antibody.

The total expression of CD44 was in general higher in cell lines classified as basal-like compared to luminal cell lines. Interestingly, the cell lines classified as basal B, showed a higher gene expression of the CD44 standard isoform (not to be confused with the antibody CD44S), compared to cell lines classified as basal A. Conversely, the basal B cell lines showed a preference for the isoforms including variant exons. Even if only a few cell lines of each subtype were analyzed, one could speculate that these different isoforms have distinct functions in respective subtype and/or might reflect the cell of origin. Basal B cell lines have earlier been associated to mesenchymal and stem/progenitor-cell characteristics [96, 292]. Moreover, in mammosphere (or tumorsphere) culture of the cell lines, a switch in CD44 expression was observed, and the expression of isoforms including variant exons increased. This highlights the importance of the cellular context in *in vitro* studies and possibly this shift in expression of CD44 variants can be observed also *in vivo* [162].

In a tumor material of 187 primary breast tumors, we correlated the gene expression (mRNA) of respective isoform of CD44 to the level of protein expression detected by immunohistochemistry. Notably, all analyzed isoforms except for CD44 standard were positively correlated to CD44 protein expression. Moreover, tumors harboring the cancer stem cell phenotype CD44+/CD24- were significantly associated with a higher expression of alternatively spliced variants. Conversely, tumors with strong staining of another cancer stem cell marker, ALDH1A1, displayed a higher expression of CD44 standard. These findings deserve further investigations in, for example, cell sorted tumor compartments enriched for cancer stem cell properties.

Clinical characteristics of these primary breast tumors were compared to expression of CD44 isoforms. Strikingly, the variants including more variant

exons (CD44v2-v10 and CD44v3-v10) could be associated to hormone receptor positive breast tumors, whereas high expression of CD44v8-v10 and CD44S, respectively, were associated with strong protein staining of EGFR or HER2. Similar differences were also apparent in the different molecular subtypes, and upon hierarchical clustering of the expression values of the CD44 isoforms, the tumors were divided in four distinct clusters representing various tumor characteristics and subtypes. Cluster A and Cluster B were associated to hormone receptor positive tumors (luminal subtypes). Moreover, HER2 positive tumors were more common in Cluster A, whereas Cluster B was enriched for tumors with *PIK3CA* mutations. In Cluster C and D the basal-like tumors were more frequent, although Cluster C also included tumors expressing both CD44 standard and HER2 and Cluster D was represented by tumors with high expression of CD44v8-v10. Significant difference in 10-year overall survival between the different clusters were found, with the best outcome in Cluster B and the worst in Cluster D. Interestingly, the presence of tumors with the CD44+/CD24- phenotype varied significantly in the subgroups, being more common in Cluster B and D.

To summarize, a multitude of different isoforms of CD44 are most likely detected by the antibodies used for detection of tumor cells with a cancer stem cell phenotype. Moreover, our results suggest that specific isoforms may cooperate with clinical markers like HER2 and EGFR, and that the expression pattern of CD44 is associated to different molecular subtypes. Possibly, breast cancer cells harboring an undifferentiated phenotype may splice out the variable exons while more variable exons are retained in more mature cells. When considering CD44 as a therapeutic target in breast cancer these conclusions are of great importance.

Characteristic mutations in basal-like breast cancer, Paper II

In Paper II our aim was to identify mutations present in a panel of basal-like breast cancer cell lines that are common experimental models. Breast tumors of the basal-like subtype are often characterized as triple-negative, which

means that they typically lack expression of ER, PR and HER2. Therefore, no targeted treatment options exist for this type of tumors and instead chemotherapy is used as standard treatment, with a high risk of relapse of the disease. Tumors of this subtype harbor a high mutational load, most commonly in *TP53* and *PIK3CA*, and they are characterized by frequent genomic amplifications and deletions as well as loss of PTEN expression. However, compared to for example luminal A tumors, only a few mutations are known to be recurrently mutated in basal-like tumors [53].

A selection of 1237 genes was analyzed in Paper II, which previously have been found to be mutated in breast cancer. Mutation analysis was performed using targeted resequencing of a region of 6.5 Mbases including exons, 3'UTR and portions of the 5'UTR and upstream regions in six basal-like cell lines and their matched lymphocyte DNA. On average, a sequence coverage of 127-fold was achieved for the cancer cell lines and 98-fold for the paired normal samples. The Burrows-Wheeler aligner (bwa) was used for the alignment of sequencing reads to the reference human genome, and Genome Analysis ToolKit (GATK) UnifiedGenotyper was used for the calling of single nucleotide variants (SNVs) and indels [287]. To derive copy number estimations of the targeted regions, the CONTRA software was used [289].

We detected 658 high confidence somatic variants and in total 315 of these were not present in the database COSMIC and therefore considered as novel and of these 110 were coding and the remaining portion was non-coding. We used Sanger sequencing to confirm 125 of the novel variants. Of the high-confidence variants, 98% of the exonic variants could be confirmed, whereas the results in non-coding regions varied slightly more but in general the validation rate was very high. This show that the specificity of our analysis pipeline is satisfactory, and that the used filters can be applied in similar studies. However, since we actually could validate variants among the low confidence variant calls with Sanger sequencing, the sensitivity of the analysis could be improved. The easiest way achieve this would be by increasing the sequencing coverage. The results from the copy number analysis showed good

correlations to SNP array data, which demonstrate the utility of using targeted resequencing data to derive copy number estimations without using additional platforms, which could significantly reduce the costs in large-scale projects. Increasing the sequencing coverage in the normal sample would decrease the number of genes with missing copy number estimates, since a filter on coverage in the normal sample was applied to get reliable data.

Other analysis tools that could be used for calling of SNPs and short indels are for example VarScan [293] and Mutect [294], which by statistical methods determine the likelihood for a variant to be somatic based on the number of aligned reads supporting each allele per sample. These programs might discover more low allele frequency mutations, which potentially were been missed by our pipeline. However, independent on what analysis method that is used, the filtering settings are of importance and there is always a balance between specificity and sensitivity that has to be taken into consideration.

Notably, for analysis of primary tumor material stricter thresholds combined with higher sequencing coverage are required since the allele frequencies are affected by normal cell contamination and the presence of multiple clones. Using paired-end sequencing data would also be preferred to enable a better detection of shorter indels.

As have been reported earlier, *TP53* was found to be mutated in all six cancer cell lines. In addition to this, 17 genes harbored mutations in more than one cell line. To get a more comprehensive overview of the genetic aberrations in these cell lines, all somatic mutations, high level amplifications and exonic deletions (*i.e.*, |log2 ratios|>2), and COSMIC variants detected per gene were summarized, and we found that 34 genes were affected in more than one cell line. In a cohort of basal-like breast tumors, 91% of these 34 genes had somatic mutations or copy number aberrations. A large fraction of these tumors were as expected *TP53* mutated, but after excluding *TP53*, we could still find mutations in the other 33 genes in almost 50% of the tumors.

Furthermore, we investigated patterns of base replacements and the genomic context of the SNVs in our data, since potential underlying mechanisms leading to non-random variations of this kind have been observed in previous mutational screening studies. We decided to focus on differences between coding and non-coding regions and, interestingly, the average mutational rate was considerably higher in the coding regions (20.6 mutations/Mbp) than in the non-coding regions (8.7 mutations/Mbp). Moreover, G and C base replacements were more frequent in the coding regions (76.4%) than in the non-coding regions (61.0%). Notably, the base replacements C/G→A/T, C/G→G/C, C/G→T/A and A/T→G/C were significantly more common in the coding regions than in the non-coding regions, taking differences in GC-content in to consideration in the calculations. In a previously published signature for mutational processes in breast cancer (Signature 3, described above under section *Mutation signatures in breast cancer*), an enrichment for C/G substitutions was observed [168], but to our knowledge it has not been reported earlier that this signature show a preference for coding regions. The substitutions C/G→T/A and C/G→G/C are results of natural deamination processes or possibly caused by APOBEC editing, whereas C/G→A/T could be caused by free radical species or benzo[a]pyrene in cigarette smoke.

Interestingly, the SNVs within the context of T[C]A/T[G]A and T[C]T/A[G]A were significantly more common in the coding than in the non-coding regions. APOBEC enzyme activity has been suggested to enrich for substitutions in the context of at T[C]N trinucleotides [170].

To conclude, we have identified a panel of 34 genes, which eventually could be included in a panel of genes for monitoring of the disease and it could be worth to investigate their function in tumorigenesis and whether some of these genes could be druggable targets. We also observed significant differences in the mutational patterns between coding and non-coding regions that would be interesting to further examine in a larger material of primary breast tumors. Underlying mechanisms for these differences remain to be elucidated.

Early detection of occult metastatic breast cancer, Paper III

In Paper III, we wanted to investigate the possibility of detecting ctDNA in the blood of breast cancer patients before clinical presentation of metastatic disease (Figure 4). A major issue for women with breast cancer is the fact that about half of recurrences occur more than five years after the initial diagnosis. Today we lack reliable markers to predict which women that eventually will develop metastatic disease. Therefore, it would be desirable with a simple blood-based test for monitoring of the patient's status during therapy and for detection of relapse of the disease. This would enable a switch or onset of therapy at the earliest moment, to possibly improve the patient survival. However, currently there is no evidence that an early detection of metastatic disease would change the outcome for those patients. Circulating tumor DNA may better reflect the tumor burden in a patient than other markers like CA15-3 [267], which could lead to an improved specificity and sensitivity for an early detection of metastases and could possibly help in optimization of therapy. To corroborate a cancer free status after initial surgery or after adjuvant therapy could also be of interest during the clinical follow-up, which potentially could be determined if no ctDNA is present.

Low-coverage whole genome sequencing was performed on 21 primary tumor specimens from 20 patients diagnosed with non-metastatic breast cancer. Six patients had a long-term disease free survival (median follow up >9 years), and 14 patients were eventually diagnosed with clinical metastasis (range 1.2-5.1 years after primary surgery). Tumors were sequenced to an average sequence coverage of 5.3-fold, and owing to relatively large insert sizes at sequencing, an average physical coverage of 15.6-fold was obtained.

A novel pipeline, SplitSeq, was developed for detection of the exact breakpoint sequence in inter- and intra-chromosomal rearrangements, which simplify the design of primers for short amplicons needed for detection of fragmented ctDNA. In total, 85% of the breakpoints could be confirmed by PCR in primary tumor DNA, using matched normal DNA as negative control.

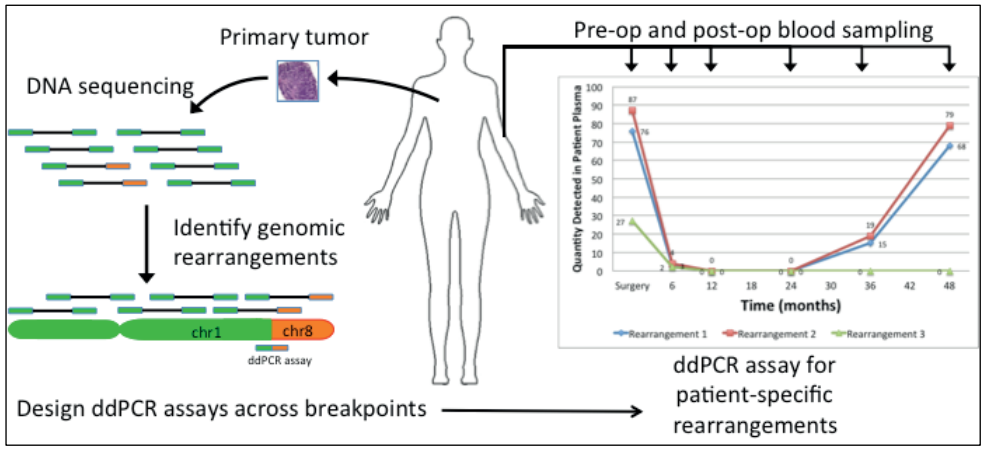

Figure 4. Schematic workflow used in Study III. Patients diagnosed with primary breast cancer were included in the study and DNA isolated from each patient's tumor was sequenced to identify patient-specific genomic rearrangements. Subsequently, liquid biopsies in terms of multiple blood samples were analyzed with droplet digital PCR (ddPCR) during the clinical course. By detection of these individual genomic rearrangements in the blood samples, early detection of recurrent metastatic breast cancer was possible.

Digital droplet PCR (ddPCR) was used for analysis of patient plasma samples, both pre-surgery and post-surgery at multiple time points. In total, 93 plasma samples were studied for the 20 patients using 4-6 validated assays per patient. The size of the PCR amplicons was kept as short as possible owing to the short fragment sizes of ctDNA in plasma. In ddPCR, the reactions occur in nanoliter-sized droplets and only droplets containing template DNA result in positive signals. This makes this method suitable for analysis of input samples of low concentration, since it yields a very good specificity and sensitivity. In our study, we could detect ≤0.01% of tumor-specific DNA with ddPCR, and no positive droplets were obtained from the negative control reactions. All droplet-derived intensity values were normalized with respect to the negative droplet intensities (2*[max intensity of negative control]), and a cutoff at 50% on the normalized range of values was used to discriminate between positive and negative droplets.

The fraction of ctDNA was calculated relative to the measured concentration of a normal region at 2p14 (that rarely undergoes copy number changes in

breast cancer), at each analyzed time-point. In total, tumor-specific DNA was detected in 29 samples corresponding to 13 of 14 of the eventual metastatic patients. Importantly, no positive droplets were detected in any of the long-term disease free patients.

Detection of ctDNA preceded clinical detection of metastatic disease in 12 out of 14 patients, with an average lead-time of 11 months. Interestingly, tumor-specific rearrangements were found in some patients up to 3 years before diagnosis of recurrent disease. The response to treatment or progression of disease could be followed by changes in ctDNA concentrations, pointing at the utility of using this method to evaluate the clinical response or as a measure of tumor burden. Sometimes only a fraction of the analyzed rearrangements was apparent in plasma before clinical detection of metastases, possibly indicating the existence of more than one clone in the primary tumor. One of the patients was diagnosed with bilateral primary breast tumors; ctDNA analysis detected only rearrangements matching one of the primary tumors was detected. This shows the utility of using ctDNA to elucidate which primary tumor that has given rise to metastatic disease and could be used as a guide in optimization of therapy.

Importantly, we found the fraction of ctDNA, detected at the first positive time-point, to be predictive of the outcome for the patients and each doubling of ctDNA level was associated with poor recurrence-free survival and poor overall survival.

In previous studies it has been demonstrated that ctDNA could be used for monitoring of metastatic disease in breast cancer patients [259, 267], and that the levels of ctDNA reflect the response to therapy better than other biomarkers. In this study, we showed for the first time that by monitoring breast cancer patients that has not (yet) developed metastasis, it is possible to identify metastatic breast cancer months to years prior to clinical presentation. Furthermore, we showed ctDNA levels to be a quantitative factor that predicted poor outcome. Since our analysis pipeline very accurately distinguish

the eventual metastatic patients from long-term disease free patients it could be utilized in prospective studies of adjuvant therapy. For example, it has been suggested that 10-year, instead of 5-year, Tamoxifen treatment would improve survival rates for patients with ER-positive tumors [295]. It would be interesting to see if the effect of additional endocrine treatment is related to levels of ctDNA in the plasma. Perhaps some patients would benefit from additional therapy and some patients can be considered as cured, if no ctDNA can be detected in blood after 5 years, and therefore overtreatment can be avoided. This is especially interesting as both overtreatment [296] and late recurrences [297] are significant problems in breast cancer, especially for women with hormone receptor positive tumors.

Concluding remarks

As has been touched upon in this thesis, breast cancer is a very heterogeneous disease where both clinical characteristics and prognosis can be linked different molecular subtypes and genetic aberrations. A major concern for women diagnosed with breast cancer is the development of metastatic disease, and the risk for this is believed to be strongly influenced by the characteristics of the primary tumor. Some women will develop early or late recurrences of the disease, while others remain cancer-free for many years after the diagnosis and will eventually die of other causes. The clinical markers currently in use as guidance for decision of therapy of breast cancer have been improved during the last decades. Still, both overtreatment and lack of accurate tools to predict which patients that eventually will develop metastases are significant problems. Therapy resistance is another major problem where the underlying mechanisms are less well understood. This thesis have addressed issues that potentially can be related to therapy resistance and tumor heterogeneity due to characteristics associated with cells of cancer stem cell phenotype. The potential inherent plasticity in these types of cells should be further evaluated as a factor that contributes to relapse of the disease. Moreover, the mutational background in an aggressive phenotype of tumors has been investigated and the recurrent mutations in this subtype of tumors should be evaluated in terms of therapy resistance or as potential druggable targets. Importantly, by utilizing non-invasive blood tests for disease monitoring of patients diagnosed with primary breast cancer, we have managed to detect metastatic disease when the disease burden is smaller and prior to clinical symptoms. This opens up possibilities to in the earliest moment change or onset therapy that in the end could lead to an improved survival of women with breast cancer.

To conclude, this thesis has contributed to increased knowledge in important fields of breast cancer research. However, due to the adaptable nature of cancer, the need for research may indeed be perpetual as we continue to work towards reducing incidence of cancer and increasing the rate of cures.

Sammanfattning på svenska

Årligen diagnostiseras 1,7 miljoner kvinnor med bröstcancer i världen, och trots att stora framsteg gjorts när det gäller behandling så avlider över en halv miljon kvinnor varje år i sjukdomen. I majoriteten av de fall som har dödlig utgång beror på att cancern har metastaserat, vilket innebär att sjukdomen har spridit sig till andra organ i kroppen. Även om det finns flertalet faktorer som bidrar till hur prognosen ser ut och vilken behandling som kan vara aktuell, så finns det ändå inget säkert sätt att förutspå vem som kommer att få återfall i sjukdomen. Anmärkningsvärt är att återfall kan ske många år (ibland över ett decennium) efter den ursprungliga diagnosen medan andra uppnår en hög ålder utan några spår av sjukdomen efter avslutad behandling. Tyvärr innebär detta att många kvinnor i dagens läge överbehandlas med läkemedel när enbart kirurgi skulle vara botande, medan vissa kvinnor skulle få en bättre prognos om det fanns en mer effektiv och skräddarsydd medicinsk behandling att tillgå.

I de fall behandlingen inte är botande så finns det troligen flera anledningar till detta. En teori till varför detta sker är att det kan finnas en viss typ av celler i tumören som är mer motståndskraftiga mot traditionell behandling som till exempel cytostatika och dessa celler troligtvis kan ha egenskaper liknande de som är karaktäristiska för stamceller. Denna typ av celler benämns ofta cancerstamceller och har specifika proteiner på cellytan, så kallade markörer. I ett av delarbetena undersöks hur genuttrycket för olika varianter av en markör för cancerstamceller korrelerar med proteinuttrycket av stamcellsmarkörer samt med prognostiska och behandlingsprediktiva markörer. Resultaten från detta arbete ger indikationer på att olika varianter av denna cancerstamcellsmarkör troligen interagerar med specifika proteiner som är viktiga för cancercellernas tillväxt och överlevnad.

En annan anledning till att behandling av cancer inte fungerar är att det finns eller uppstår mutationer som ger upphov till resistens under behandlingens

gång. Därför har vi i det andra delarbetet karaktäriserat nya och kända mutationer i cellmodeller för en aggressiv typ av bröstcancer som i dagsläget saknar målinriktad terapi. Dessa cellmodeller används ofta i experimentella försök för att bland annat testa olika läkemedel. Totalt analyserades 1237 gener och i dessa hittades hundratals mutationer både i proteinkodande och i icke-kodande DNA sekvenser. Vetskapen om dessa mutationer kommer att utgöra en värdefull grund i framtida försök där dessa cellmodeller används. Dessutom fann vi att de proteinkodande regionerna har en högre frekvens av mutationer än de icke-kodande och att det går att urskilja ett mönster av dessa skillnader i form av vilka basutbyten som sker.

För att kunna förbättra diagnostiken och följa sjukdomsutvecklingen hos en cancerpatient så finns det ett stort utrymme för att utveckla och öka användningen av biomarkörer. Ur patientsynpunkt så är en icke-invasiv analysmetod att föredra, företrädesvis i form av ett enkelt blodprov. Vi vet att DNA från cancerceller kan detekteras i blodcirkulationen, och att det därför går att återfinna mutationer och kromosomala rearrangemang som är specifika för tumören i blodet. Vi har använt oss av en sekvenseringsmetod för att ta fram information om vilka kromosomala rearrangemang som fanns i primärtumören, dvs. den tumör som avlägsnades kirurgiskt från bröstet vid diagnos. Ett kromosomalt rearrangemang kan exempelvis betyda att delar från två olika kromosomer felaktigt sammanfogats, vilket ibland kan ske i samband med celldelning. Dessa rearrangemang kunde sedan återfinnas i blodet hos cancerpatienter efter flera månader, ibland flera år, innan den kliniska diagnosen för metastaserande sjukdom kunde ställas. Detta öppnar upp möjligheter att ändra eller återuppta behandling av patienter tidigare än vad som är görligt idag. Med anpassade behandlingsalternativ så finns det troligtvis stora möjligheter att förlänga överlevnaden för patienter med metastaserande sjukdom. I en kontrollgrupp av patienter som inte fått återfall i sjukdomen, så återfanns inga spår av tumör-specifika rearrangemang i blodet efter operation, vilket bekräftar att all cancer är borta från patienten.

Acknowledgements

First of all, I would like to thank Åke Borg for accepting me as a PhD student and for introducing me into the fascinating and challenging field of breast cancer. Your support and scientific skills have meant a lot to me over the years and your contribution in this field of research is extremely impressive.

Second, I really would like to thank Lao Saal for being my main supervisor during the second part of my PhD period. Your excellent scientific skills, positive attitude and endless optimism have helped very much during the last years. Without your support this would not have been possible!

I would like to thank Cecilia Hegardt for introducing me into the field of cancer stem cells in the first period of my PhD period and for being by co-supervisor.

Thanks to all in our great team at work, Sofia, Christof, Jill, Barbara, Malin, Chris, Alan and Tony. Sofia, for all support and encouragement, Christof, for your excellent bioinformatics skills, Malin for all lunch dates and most interesting discussions, Chris for helping with bioinformatics, Jill and Barbara for your positive attitude and last but not least Alan and Tony for invaluable help in the lab and for your gentle attitude.

Many thanks to my roommates and friends at work, you have meant a lot during these years, especially Gabriella, Karolina and Tamara. Thanks for all fun moments we have shared and for cheering me up whenever that was needed. Special thanks to Gabriella for all help with the thesis!

Thanks to Pär-Ola Bendahl for your excellent statistical help and pedagogic skills, and thanks to Anders Kvist for kindly introducing me into the world of bioinformatics.

Thanks to the head of the department, Lars Ekblad and deputy of department, Ingrid Wilson, for creating a nice working atmosphere and thanks to Susanne André and Björn Frostner for your great administrative support.

A special thanks to my beloved family, our adorable son Gustav and my soulmate Martin for dealing with me during these years. You put things in another perspective, which I am very grateful for and this would for sure not have been possible without your support!

Thanks to my brother Ted, Mum and Ove for all your encouragement and practical help during my PhD period.

Thanks to all friends and relatives for all of your support!

Finally, thanks to all patients for participating in these studies!

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F: Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015, 136(5):E359-386.
2. Engholm G FJ, Christensen N, Kejs AMT, Johannesen TB, Khan S, Milter MC, Ólafsdóttir E, Petersen T, Pukkala E, Stenz F, Storm HH. : NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 7.0 (17.12.2014). Association of the Nordic Cancer Registries. Danish Cancer Society. *Available from* http://www.ancr.nu *, accessed April 2015*
3. Parkin DM: Global cancer statistics in the year 2000. *Lancet Oncol* 2001, 2(9):533-543.
4. Cancer IAfRo: List of Classifications by cancer sites with sufficient or limited evidence in humans, http://monographs.iarc.fr/ENG/Classification/index.php, Accessed January 2015. In., vol. Volumes 1 to 105; 2015.
5. Carter BD, Abnet CC, Feskanich D, Freedman ND, Hartge P, Lewis CE, Ockene JK, Prentice RL, Speizer FE, Thun MJ *et al*: Smoking and mortality--beyond established causes. *N Engl J Med* 2015, 372(7):631-640.
6. Research WCRFAIfC: Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective, Washington DC: AICR; 2007, http://www.dietandcancerreport.org/cancer_resource_center/downloads/Second_Expert_Report_full.pdf.
7. Collaborative Group on Hormonal Factors in Breast C: Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. *Lancet* 2002, 360(9328):187-195.
8. McPherson K, Steel CM, Dixon JM: ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. *BMJ* 2000, 321(7261):624-628.
9. Stapelkamp C, Holmberg L, Tataru D, Moller H, Robinson D: Predictors of early death in female patients with breast cancer in the UK: a cohort study. *BMJ Open* 2011, 1(2):e000247.
10. Ewertz M, Duffy SW, Adami HO, Kvale G, Lund E, Meirik O, Mellemgaard A, Soini I, Tulinius H: Age at first birth, parity and risk of breast cancer: a meta-analysis of 8 studies from the Nordic countries. *Int J Cancer* 1990, 46(4):597-603.
11. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, Gaudet M, Schmidt MK, Broeks A, Cox A *et al*: Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. *J Natl Cancer Inst* 2011, 103(3):250-263.
12. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A *et al*: 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. *Am J Hum Genet* 2003, 72(5):1117-1130.

13. Mavaddat N, Peock S, Frost D, Ellis S, Platte R, Fineberg E, Evans DG, Izatt L, Eeles RA, Adlard J *et al*: Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. *J Natl Cancer Inst* 2013, 105(11):812-822.
14. Antoniou AC, Beesley J, McGuffog L, Sinilnikova OM, Healey S, Neuhausen SL, Ding YC, Rebbeck TR, Weitzel JN, Lynch HT *et al*: Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers: implications for risk prediction. *Cancer Res* 2010, 70(23):9742-9754.
15. Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N, Easton DF, Evans C, Deacon J, Stratton MR: Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *J Natl Cancer Inst* 1999, 91(11):943-949.
16. Manchanda R, Loggenberg K, Sanderson S, Burnell M, Wardle J, Gessler S, Side L, Balogun N, Desai R, Kumar A *et al*: Population Testing for Cancer Predisposing BRCA1/BRCA2 Mutations in the Ashkenazi-Jewish Community: A Randomized Controlled Trial. *J Natl Cancer Inst* 2015, 107(1).
17. Widschwendter M, Rosenthal AN, Philpott S, Rizzuto I, Fraser L, Hayward J, Intermaggio MP, Edlund CK, Ramus SJ, Gayther SA *et al*: The sex hormone system in carriers of BRCA1/2 mutations: a case-control study. *Lancet Oncol* 2013, 14(12):1226-1232.
18. Malkin D, Li FP, Strong LC, Fraumeni JF, Jr., Nelson CE, Kim DH, Kassel J, Gryka MA, Bischoff FZ, Tainsky MA *et al*: Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 1990, 250(4985):1233-1238.
19. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T *et al*: PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* 2007, 39(2):165-167.
20. Kuusisto KM, Bebel A, Vihinen M, Schleutker J, Sallinen SL: Screening for BRCA1, BRCA2, CHEK2, PALB2, BRIP1, RAD50, and CDH1 mutations in high-risk Finnish BRCA1/2-founder mutation-negative breast and/or ovarian cancer individuals. *Breast Cancer Res* 2011, 13(1):R20.
21. Couch FJ, Hart SN, Sharma P, Toland AE, Wang X, Miron P, Olson JE, Godwin AK, Pankratz VS, Olswold C *et al*: Inherited Mutations in 17 Breast Cancer Susceptibility Genes Among a Large Triple-Negative Breast Cancer Cohort Unselected for Family History of Breast Cancer. *J Clin Oncol* 2014.
22. London SJ, Connolly JL, Schnitt SJ, Colditz GA: A prospective study of benign breast disease and the risk of breast cancer. *JAMA* 1992, 267(7):941-944.
23. Harris JR LM, Morrow M, Osborne CK: Diseases of the Breast - Fourth Edition, Lippincott Williams & Wilkins; Dillon DA, Guidi AJ, Schnitt SJ. Chapter 28: Pathology of Invasive Breast Cancer,. 2009.
24. Cardoso F, Costa A, Norton L, Senkus E, Aapro M, Andre F, Barrios CH, Bergh J, Biganzoli L, Blackwell KL *et al*: ESO-ESMO 2nd international consensus guidelines for advanced breast cancer (ABC2)dagger. *Ann Oncol* 2014, 25(10):1871-1888.
25. Galea MH, Blamey RW, Elston CE, Ellis IO: The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* 1992, 22(3):207-219.

26. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, Parker HL: Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J Clin Oncol* 2001, 19(4):980-991.
27. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, Caldas C, Pharoah PD: PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010, 12(1):R1.
28. Burstein HJ, Temin S, Anderson H, Buchholz TA, Davidson NE, Gelmon KE, Giordano SH, Hudis CA, Rowden D, Solky AJ *et al*: Adjuvant endocrine therapy for women with hormone receptor-positive breast cancer: american society of clinical oncology clinical practice guideline focused update. *J Clin Oncol* 2014, 32(21):2255-2269.
29. Arpino G, Weiss H, Lee AV, Schiff R, De Placido S, Osborne CK, Elledge RM: Estrogen receptor-positive, progesterone receptor-negative breast cancer: association with growth factor receptor expression and tamoxifen resistance. *J Natl Cancer Inst* 2005, 97(17):1254-1261.
30. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, Senn HJ, Panel m: Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann Oncol* 2013, 24(9):2206-2223.
31. Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, Ellis M, Henry NL, Hugh JC, Lively T *et al*: Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst* 2011, 103(22):1656-1664.
32. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: Molecular portraits of human breast tumours. *Nature* 2000, 406(6797):747-752.
33. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001, 98(19):10869-10874.
34. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y *et al*: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012, 486(7403):346-352.
35. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J *et al*: Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* 2006, 24(23):3726-3734.
36. Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT, Ravdin P, Bugarini R, Baehner FL, Davidson NE *et al*: Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncol* 2010, 11(1):55-65.
37. Drukker CA, Bueno-de-Mesquita JM, Retel VP, van Harten WH, van Tinteren H, Wesseling J, Roumen RM, Knauer M, van 't Veer LJ, Sonke GS *et al*: A prospective

evaluation of a breast cancer prognosis signature in the observational RASTER study. *Int J Cancer* 2013, 133(4):929-936.
38. Quaresma M, Coleman MP, Rachet B: 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study. *Lancet* 2015, 385(9974):1206-1218.
39. Polyak K: Heterogeneity in breast cancer. *J Clin Invest* 2011, 121(10):3786-3788.
40. Rudkin CT, Hungerford DA, Nowell PC: DNA Contents of Chromosome Ph1 and Chromosome 21 in Human Chronic Granulocytic Leukemia. *Science* 1964, 144(3623):1229-1231.
41. Nowell PC, Hungerford DA: Chromosome studies in human leukemia. II. Chronic granulocytic leukemia. *J Natl Cancer Inst* 1961, 27:1013-1035.
42. Hanahan D, Weinberg RA: Hallmarks of cancer: the next generation. *Cell* 2011, 144(5):646-674.
43. Hanahan D, Weinberg RA: The hallmarks of cancer. *Cell* 2000, 100(1):57-70.
44. Martinez P, Blasco MA: Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins. *Nat Rev Cancer* 2011, 11(3):161-176.
45. Negrini S, Gorgoulis VG, Halazonetis TD: Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 2010, 11(3):220-228.
46. Nowell PC: The clonal evolution of tumor cell populations. *Science* 1976, 194(4260):23-28.
47. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R: The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* 1994, 77(1):1 p following 166.
48. Boland CR, Goel A: Microsatellite instability in colorectal cancer. *Gastroenterology* 2010, 138(6):2073-2087 e2073.
49. Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, Hodges AK, Davies DR, David SS, Sampson JR *et al*: Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* 2002, 30(2):227-232.
50. Deng CX: BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic Acids Res* 2006, 34(5):1416-1426.
51. Loeb LA: A mutator phenotype in cancer. *Cancer Res* 2001, 61(8):3230-3239.
52. Bodmer W, Bielas JH, Beckman RA: Genetic instability is not a requirement for tumor development. *Cancer Res* 2008, 68(10):3558-3560; discussion 3560-3551.
53. Cancer Genome Atlas N: Comprehensive molecular portraits of human breast tumours. *Nature* 2012, 490(7418):61-70.
54. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G *et al*: The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 2012, 486(7403):395-399.
55. Bunz F, Fauth C, Speicher MR, Dutriaux A, Sedivy JM, Kinzler KW, Vogelstein B, Lengauer C: Targeted inactivation of p53 in human cells does not result in aneuploidy. *Cancer Res* 2002, 62(4):1129-1133.
56. Gorgoulis VG, Vassiliou LV, Karakaidos P, Zacharatos P, Kotsinas A, Liloglou T, Venere M, Ditullio RA, Jr., Kastrinakis NG, Levy B *et al*: Activation of the DNA

damage checkpoint and genomic instability in human precancerous lesions. *Nature* 2005, 434(7035):907-913.
57. Halazonetis TD, Gorgoulis VG, Bartek J: An oncogene-induced DNA damage model for cancer development. *Science* 2008, 319(5868):1352-1355.
58. Momand J, Zambetti GP, Olson DC, George D, Levine AJ: The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell* 1992, 69(7):1237-1245.
59. Haupt Y, Maya R, Kazaz A, Oren M: Mdm2 promotes the rapid degradation of p53. *Nature* 1997, 387(6630):296-299.
60. Jiang H, Reinhardt HC, Bartkova J, Tommiska J, Blomqvist C, Nevanlinna H, Bartek J, Yaffe MB, Hemann MT: The combined status of ATM and p53 link tumor development with therapeutic response. *Genes Dev* 2009, 23(16):1895-1909.
61. Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, Maurer M, Koujak S, Ferrando AA, Malmstrom P, Memeo L *et al*: Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc Natl Acad Sci U S A* 2007, 104(18):7564-7569.
62. Saal LH, Holm K, Maurer M, Memeo L, Su T, Wang X, Yu JS, Malmstrom PO, Mansukhani M, Enoksson J *et al*: PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer Res* 2005, 65(7):2554-2559.
63. Campbell LL, Polyak K: Breast tumor heterogeneity: cancer stem cells or clonal evolution? *Cell Cycle* 2007, 6(19):2332-2338.
64. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J *et al*: Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 2009, 461(7265):809-813.
65. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H *et al*: Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 2014, 512(7513):155-160.
66. Schwitalla S, Fingerle AA, Cammareri P, Nebelsiek T, Goktuna SI, Ziegler PK, Canli O, Heijmans J, Huels DJ, Moreaux G *et al*: Intestinal tumorigenesis initiated by dedifferentiation and acquisition of stem-cell-like properties. *Cell* 2013, 152(1-2):25-38.
67. Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C, Lander ES: Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* 2011, 146(4):633-644.
68. He K, Xu T, Goldkorn A: Cancer cells cyclically lose and regain drug-resistant highly tumorigenic features characteristic of a cancer stem-like phenotype. *Mol Cancer Ther* 2011, 10(6):938-948.
69. Chaffer CL, Marjanovic ND, Lee T, Bell G, Kleer CG, Reinhardt F, D'Alessio AC, Young RA, Weinberg RA: Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. *Cell* 2013, 154(1):61-74.
70. Vermeulen L, de Sousa e Melo F, Richel DJ, Medema JP: The developing cancer stem-cell model: clinical challenges and opportunities. *Lancet Oncol* 2012, 13(2):e83-89.

71. Gusterson BA, Stein T: Human breast development. *Semin Cell Dev Biol* 2012, 23(5):567-573.
72. Gusterson BA, Ross DT, Heath VJ, Stein T: Basal cytokeratins and their relationship to the cellular origin and functional classification of breast cancer. *Breast Cancer Res* 2005, 7(4):143-148.
73. Visvader JE: Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. *Genes Dev* 2009, 23(22):2563-2577.
74. Visvader JE, Stingl J: Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev* 2014, 28(11):1143-1158.
75. Weinberg RA: The biology of cancer - Second edition: Garland Science, Taylor & Francis Group; 2011.
76. Stingl J, Eaves CJ, Kuusk U, Emerman JT: Phenotypic and functional characterization in vitro of a multipotent epithelial cell present in the normal adult human breast. *Differentiation* 1998, 63(4):201-213.
77. Stingl J, Eaves CJ, Zandieh I, Emerman JT: Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res Treat* 2001, 67(2):93-109.
78. Eirew P, Stingl J, Raouf A, Turashvili G, Aparicio S, Emerman JT, Eaves CJ: A method for quantifying normal human mammary epithelial stem cells with in vivo regenerative ability. *Nat Med* 2008, 14(12):1384-1389.
79. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A *et al*: Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* 2009, 15(8):907-913.
80. Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, Dutcher J, Brown M, Jacquemier J, Viens P, Kleer CG, Liu S *et al*: ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell* 2007, 1(5):555-567.
81. Eirew P, Kannan N, Knapp DJ, Vaillant F, Emerman JT, Lindeman GJ, Visvader JE, Eaves CJ: Aldehyde dehydrogenase activity is a biomarker of primitive normal human mammary luminal cells. *Stem Cells* 2012, 30(2):344-348.
82. Shehata M, Teschendorff A, Sharp G, Novcic N, Russell IA, Avril S, Prater M, Eirew P, Caldas C, Watson CJ *et al*: Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res* 2012, 14(5):R134.
83. Liu S, Cong Y, Wang D, Sun Y, Deng L, Liu Y, Martin-Trevino R, Shang L, McDermott SP, Landis MD *et al*: Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem Cell Reports* 2014, 2(1):78-91.
84. Dontu G, Abdallah WM, Foley JM, Jackson KW, Clarke MF, Kawamura MJ, Wicha MS: In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev* 2003, 17(10):1253-1270.
85. Honeth G, Lombardi S, Ginestier C, Hur M, Marlow R, Buchupalli B, Shinomiya I, Gazinska P, Bombelli S, Ramalingam V *et al*: Aldehyde dehydrogenase and estrogen receptor define a hierarchy of cellular differentiation in the normal human mammary epithelium. *Breast Cancer Res* 2014, 16(3):R52.

86. Harrison H, Simoes BM, Rogerson L, Howell SJ, Landberg G, Clarke RB: Oestrogen increases the activity of oestrogen receptor negative breast cancer stem cells through paracrine EGFR and Notch signalling. *Breast Cancer Res* 2013, 15(2):R21.
87. Asselin-Labat ML, Vaillant F, Sheridan JM, Pal B, Wu D, Simpson ER, Yasuda H, Smyth GK, Martin TJ, Lindeman GJ *et al*: Control of mammary stem cell function by steroid hormone signalling. *Nature* 2010, 465(7299):798-802.
88. Tanos T, Sflomos G, Echeverria PC, Ayyanan A, Gutierrez M, Delaloye JF, Raffoul W, Fiche M, Dougall W, Schneider P *et al*: Progesterone/RANKL is a major regulatory axis in the human breast. *Sci Transl Med* 2013, 5(182):182ra155.
89. Dontu G, Jackson KW, McNicholas E, Kawamura MJ, Abdallah WM, Wicha MS: Role of Notch signaling in cell-fate determination of human mammary stem/progenitor cells. *Breast Cancer Res* 2004, 6(6):R605-615.
90. Reya T, Clevers H: Wnt signalling in stem cells and cancer. *Nature* 2005, 434(7035):843-850.
91. Liu S, Dontu G, Mantle ID, Patel S, Ahn NS, Jackson KW, Suri P, Wicha MS: Hedgehog signaling and Bmi-1 regulate self-renewal of normal and malignant human mammary stem cells. *Cancer Res* 2006, 66(12):6063-6071.
92. Cairns J: Mutation selection and the natural history of cancer. *Nature* 1975, 255(5505):197-200.
93. Smith GH: Label-retaining epithelial cells in mouse mammary gland divide asymmetrically and retain their template DNA strands. *Development* 2005, 132(4):681-687.
94. Kannan N, Huda N, Tu L, Droumeva R, Aubert G, Chavez E, Brinkman RR, Lansdorp P, Emerman J, Abe S *et al*: The luminal progenitor compartment of the normal human mammary gland constitutes a unique site of telomere dysfunction. *Stem Cell Reports* 2013, 1(1):28-37.
95. Gunes C, Rudolph KL: The role of telomeres in stem cells and cancer. *Cell* 2013, 152(3):390-393.
96. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM: Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* 2010, 12(5):R68.
97. Soady KJ, Kendrick H, Gao Q, Tutt A, Zvelebil M, Ordonez LD, Quist J, Tan DW, Isacke CM, Grigoriadis A *et al*: Mouse mammary stem cells express prognostic markers for triple-negative breast cancer. *Breast Cancer Res* 2015, 17(1):539.
98. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipitsin M *et al*: The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* 2008, 133(4):704-715.
99. Bonnet D, Dick JE: Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* 1997, 3(7):730-737.
100. de Klein A, van Kessel AG, Grosveld G, Bartram CR, Hagemeijer A, Bootsma D, Spurr NK, Heisterkamp N, Groffen J, Stephenson JR: A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* 1982, 300(5894):765-767.

101. Bhatia R, Holtz M, Niu N, Gray R, Snyder DS, Sawyers CL, Arber DA, Slovak ML, Forman SJ: Persistence of malignant hematopoietic progenitors in chronic myelogenous leukemia patients in complete cytogenetic remission following imatinib mesylate treatment. *Blood* 2003, 101(12):4701-4707.
102. Zhang S, Balch C, Chan MW, Lai HC, Matei D, Schilder JM, Yan PS, Huang TH, Nephew KP: Identification and characterization of ovarian cancer-initiating cells from primary human tumors. *Cancer Res* 2008, 68(11):4311-4320.
103. O'Brien CA, Pollett A, Gallinger S, Dick JE: A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature* 2007, 445(7123):106-110.
104. Singh SK, Clarke ID, Terasaki M, Bonn VE, Hawkins C, Squire J, Dirks PB: Identification of a cancer stem cell in human brain tumors. *Cancer Res* 2003, 63(18):5821-5828.
105. Britton KM, Eyre R, Harvey IJ, Stemke-Hale K, Browell D, Lennard TW, Meeson AP: Breast cancer, side population cells and ABCG2 expression. *Cancer Lett* 2012, 323(1):97-105.
106. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF: Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A* 2003, 100(7):3983-3988.
107. Charafe-Jauffret E, Ginestier C, Iovino F, Wicinski J, Cervera N, Finetti P, Hur MH, Diebel ME, Monville F, Dutcher J *et al*: Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. *Cancer Res* 2009, 69(4):1302-1313.
108. Phillips TM, McBride WH, Pajonk F: The response of CD24(-/low)/CD44+ breast cancer-initiating cells to radiation. *J Natl Cancer Inst* 2006, 98(24):1777-1785.
109. Charafe-Jauffret E, Ginestier C, Bertucci F, Cabaud O, Wicinski J, Finetti P, Josselin E, Adelaide J, Nguyen TT, Monville F *et al*: ALDH1-positive cancer stem cells predict engraftment of primary breast tumors and are governed by a common stem cell program. *Cancer Res* 2013, 73(24):7290-7300.
110. Honeth G, Bendahl PO, Ringner M, Saal LH, Gruvberger-Saal SK, Lovgren K, Grabau D, Ferno M, Borg A, Hegardt C: The CD44+/CD24- phenotype is enriched in basal-like breast tumors. *Breast Cancer Res* 2008, 10(3):R53.
111. Ricardo S, Vieira AF, Gerhard R, Leitao D, Pinto R, Cameselle-Teijeiro JF, Milanezi F, Schmitt F, Paredes J: Breast cancer stem cell markers CD44, CD24 and ALDH1: expression distribution within intrinsic molecular subtype. *J Clin Pathol* 2011, 64(11):937-946.
112. Li Y, Welm B, Podsypanina K, Huang S, Chamorro M, Zhang X, Rowlands T, Egeblad M, Cowin P, Werb Z *et al*: Evidence that transgenes encoding components of the Wnt signaling pathway preferentially induce mammary cancers from progenitor cells. *Proc Natl Acad Sci U S A* 2003, 100(26):15853-15858.
113. Merchant AA, Matsui W: Targeting Hedgehog--a cancer stem cell pathway. *Clin Cancer Res* 2010, 16(12):3130-3140.
114. Harrison H, Farnie G, Howell SJ, Rock RE, Stylianou S, Brennan KR, Bundred NJ, Clarke RB: Regulation of breast cancer stem cell activity by signaling through the Notch4 receptor. *Cancer Res* 2010, 70(2):709-718.

115. Jang GB, Hong IS, Kim RJ, Lee SY, Park SJ, Lee ES, Park JH, Yun CH, Chung JU, Lee KJ *et al*: Wnt/beta-catenin small molecule inhibitor CWP232228 preferentially inhibits the growth of breast cancer stem-like cells. *Cancer Res* 2015.
116. Shipitsin M, Campbell LL, Argani P, Weremowicz S, Bloushtain-Qimron N, Yao J, Nikolskaya T, Serebryiskaya T, Beroukhim R, Hu M *et al*: Molecular definition of breast tumor heterogeneity. *Cancer Cell* 2007, 11(3):259-273.
117. Park SY, Gonen M, Kim HJ, Michor F, Polyak K: Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J Clin Invest* 2010, 120(2):636-644.
118. Li X, Lewis MT, Huang J, Gutierrez C, Osborne CK, Wu MF, Hilsenbeck SG, Pavlick A, Zhang X, Chamness GC *et al*: Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *J Natl Cancer Inst* 2008, 100(9):672-679.
119. Wicha MS, Liu S, Dontu G: Cancer stem cells: an old idea--a paradigm shift. *Cancer Res* 2006, 66(4):1883-1890; discussion 1895-1886.
120. Dean M, Fojo T, Bates S: Tumour stem cells and drug resistance. *Nat Rev Cancer* 2005, 5(4):275-284.
121. Gunthert U, Hofmann M, Rudy W, Reber S, Zoller M, Haussmann I, Matzku S, Wenzel A, Ponta H, Herrlich P: A new variant of glycoprotein CD44 confers metastatic potential to rat carcinoma cells. *Cell* 1991, 65(1):13-24.
122. Marhaba R, Zoller M: CD44 in cancer progression: adhesion, migration and growth regulation. *J Mol Histol* 2004, 35(3):211-231.
123. Friedrichs K, Franke F, Lisboa BW, Kugler G, Gille I, Terpe HJ, Holzel F, Maass H, Gunthert U: CD44 isoforms correlate with cellular differentiation but not with prognosis in human breast cancer. *Cancer Res* 1995, 55(22):5424-5433.
124. Kaufmann M, Heider KH, Sinn HP, von Minckwitz G, Ponta H, Herrlich P: CD44 variant exon epitopes in primary breast cancer and length of survival. *Lancet* 1995, 345(8950):615-619.
125. Joensuu H, Klemi PJ, Toikkanen S, Jalkanen S: Glycoprotein CD44 expression and its association with survival in breast cancer. *Am J Pathol* 1993, 143(3):867-874.
126. Tokue Y, Matsumura Y, Katsumata N, Watanabe T, Tarin D, Kakizoe T: CD44 variant isoform expression and breast cancer prognosis. *Jpn J Cancer Res* 1998, 89(3):283-290.
127. Zoller M: CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nat Rev Cancer* 2011, 11(4):254-267.
128. Schmitt M, Metzger M, Gradl D, Davidson G, Orian-Rousseau V: CD44 functions in Wnt signaling by regulating LRP6 localization and activation. *Cell Death Differ* 2015, 22(4):677-689.
129. Ponta H, Sherman L, Herrlich PA: CD44: from adhesion molecules to signalling regulators. *Nat Rev Mol Cell Biol* 2003, 4(1):33-45.
130. Louderbough JM, Schroeder JA: Understanding the dual nature of CD44 in breast cancer progression. *Mol Cancer Res* 2011, 9(12):1573-1586.
131. Bartolazzi A, Nocks A, Aruffo A, Spring F, Stamenkovic I: Glycosylation of CD44 is implicated in CD44-mediated cell adhesion to hyaluronan. *J Cell Biol* 1996, 132(6):1199-1208.

132. Orian-Rousseau V: CD44, a therapeutic target for metastasising tumours. *Eur J Cancer* 2010, 46(7):1271-1277.
133. Sleeman JP, Kondo K, Moll J, Ponta H, Herrlich P: Variant exons v6 and v7 together expand the repertoire of glycosaminoglycans bound by CD44. *J Biol Chem* 1997, 272(50):31837-31844.
134. Toole BP: Hyaluronan: from extracellular glue to pericellular cue. *Nat Rev Cancer* 2004, 4(7):528-539.
135. Misra S, Ghatak S, Zoltan-Jones A, Toole BP: Regulation of multidrug resistance in cancer cells by hyaluronan. *J Biol Chem* 2003, 278(28):25285-25288.
136. Ghatak S, Misra S, Toole BP: Hyaluronan oligosaccharides inhibit anchorage-independent growth of tumor cells by suppressing the phosphoinositide 3-kinase/Akt cell survival pathway. *J Biol Chem* 2002, 277(41):38013-38020.
137. Orian-Rousseau V, Chen L, Sleeman JP, Herrlich P, Ponta H: CD44 is required for two consecutive steps in HGF/c-Met signaling. *Genes Dev* 2002, 16(23):3074-3086.
138. Cheng C, Yaffe MB, Sharp PA: A positive feedback loop couples Ras activation and CD44 alternative splicing. *Genes Dev* 2006, 20(13):1715-1720.
139. Bennett KL, Jackson DG, Simon JC, Tanczos E, Peach R, Modrell B, Stamenkovic I, Plowman G, Aruffo A: CD44 isoforms containing exon V3 are responsible for the presentation of heparin-binding growth factor. *J Cell Biol* 1995, 128(4):687-698.
140. Weber GF: Molecular mechanisms of metastasis. *Cancer Lett* 2008, 270(2):181-190.
141. Murakami D, Okamoto I, Nagano O, Kawano Y, Tomita T, Iwatsubo T, De Strooper B, Yumoto E, Saya H: Presenilin-dependent gamma-secretase activity mediates the intramembranous cleavage of CD44. *Oncogene* 2003, 22(10):1511-1516.
142. Pietras A, Katz AM, Ekstrom EJ, Wee B, Halliday JJ, Pitter KL, Werbeck JL, Amankulor NM, Huse JT, Holland EC: Osteopontin-CD44 signaling in the glioma perivascular niche enhances cancer stem cell phenotypes and promotes aggressive tumor growth. *Cell Stem Cell* 2014, 14(3):357-369.
143. Tsukita S, Oishi K, Sato N, Sagara J, Kawai A, Tsukita S: ERM family members as molecular linkers between the cell surface glycoprotein CD44 and actin-based cytoskeletons. *J Cell Biol* 1994, 126(2):391-401.
144. Lokeshwar VB, Fregien N, Bourguignon LY: Ankyrin-binding domain of CD44(GP85) is required for the expression of hyaluronic acid-mediated adhesion function. *J Cell Biol* 1994, 126(4):1099-1109.
145. Fehon RG, McClatchey AI, Bretscher A: Organizing the cell cortex: the role of ERM proteins. *Nat Rev Mol Cell Biol* 2010, 11(4):276-287.
146. Chen M, Manley JL: Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* 2009, 10(11):741-754.
147. Keren H, Lev-Maor G, Ast G: Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 2010, 11(5):345-355.
148. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008, 456(7221):470-476.

149. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008, 40(12):1413-1415.
150. David CJ, Manley JL: Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev* 2010, 24(21):2343-2364.
151. Li H, Wang J, Ma X, Sklar J: Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell Cycle* 2009, 8(2):218-222.
152. Liu HX, Cartegni L, Zhang MQ, Krainer AR: A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet* 2001, 27(1):55-58.
153. Mazoyer S, Puget N, Perrin-Vidoz L, Lynch HT, Serova-Sinilnikova OM, Lenoir GM: A BRCA1 nonsense mutation causes exon skipping. *Am J Hum Genet* 1998, 62(3):713-715.
154. Wappenschmidt B, Becker AA, Hauke J, Weber U, Engert S, Kohler J, Kast K, Arnold N, Rhiem K, Hahnen E *et al*: Analysis of 30 putative BRCA1 splicing mutations in hereditary breast and ovarian cancer families identifies exonic splice site mutations that escape in silico prediction. *PLoS One* 2012, 7(12):e50800.
155. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J: Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* 2014, 513(7516):120-123.
156. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B: Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 2014, 156(6):1324-1335.
157. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T: Regulation of alternative splicing by histone modifications. *Science* 2010, 327(5968):996-1000.
158. Saint-Andre V, Batsche E, Rachez C, Muchardt C: Histone H3 lysine 9 trimethylation and HP1gamma favor inclusion of alternative exons. *Nat Struct Mol Biol* 2011, 18(3):337-344.
159. Matter N, Herrlich P, Konig H: Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature* 2002, 420(6916):691-695.
160. Cheng C, Sharp PA: Regulation of CD44 alternative splicing by SRm160 and its potential role in tumor cell invasion. *Mol Cell Biol* 2006, 26(1):362-370.
161. Loh TJ, Cho S, Moon H, Jang HN, Williams DR, Jung DW, Kim IC, Ghigna C, Biamonti G, Zheng X *et al*: hnRNP L inhibits CD44 V exon splicing through interacting with its upstream intron. *Biochim Biophys Acta* 2015.
162. Brown RL, Reinke LM, Damerow MS, Perez D, Chodosh LA, Yang J, Cheng C: CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *J Clin Invest* 2011, 121(3):1064-1074.
163. Yae T, Tsuchihashi K, Ishimoto T, Motohara T, Yoshikawa M, Yoshida GJ, Wada T, Masuko T, Mogushi K, Tanaka H *et al*: Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. *Nat Commun* 2012, 3:883.
164. Bunting SF, Nussenzweig A: End-joining, translocations and cancer. *Nat Rev Cancer* 2013, 13(7):443-454.

165. Anand S, Penrhyn-Lowe S, Venkitaraman AR: AURORA-A amplification overrides the mitotic spindle assembly checkpoint, inducing resistance to Taxol. *Cancer Cell* 2003, 3(1):51-62.
166. Thompson SL, Compton DA: Chromosome missegregation in human cells arises through specific types of kinetochore-microtubule attachment errors. *Proc Natl Acad Sci U S A* 2011, 108(44):17974-17978.
167. Townsend DM, Tew KD: The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* 2003, 22(47):7369-7375.
168. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL *et al*: Signatures of mutational processes in human cancer. *Nature* 2013, 500(7463):415-421.
169. Pham P, Bransteitter R, Petruska J, Goodman MF: Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* 2003, 424(6944):103-107.
170. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA *et al*: Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012, 149(5):979-993.
171. Suspene R, Aynaud MM, Guetard D, Henry M, Eckhoff G, Marchio A, Pineau P, Dejean A, Vartanian JP, Wain-Hobson S: Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc Natl Acad Sci U S A* 2011, 108(12):4858-4863.
172. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G *et al*: An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* 2013, 45(9):970-976.
173. Karran P, Lindahl T: Hypoxanthine in deoxyribonucleic acid: generation by heat-induced hydrolysis of adenine residues and release in free form by a deoxyribonucleic acid glycosylase from calf thymus. *Biochemistry* 1980, 19(26):6005-6011.
174. Lindahl T: Instability and decay of the primary structure of DNA. *Nature* 1993, 362(6422):709-715.
175. Hussain SP, Hofseth LJ, Harris CC: Radical causes of cancer. *Nat Rev Cancer* 2003, 3(4):276-285.
176. Oikawa S, Tada-Oikawa S, Kawanishi S: Site-specific DNA damage at the GGG sequence by UVA involves acceleration of telomere shortening. *Biochemistry* 2001, 40(15):4763-4768.
177. Helleday T, Eshtad S, Nik-Zainal S: Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 2014, 15(9):585-598.
178. Hess MT, Gunz D, Luneva N, Geacintov NE, Naegeli H: Base pair conformation-dependent excision of benzo[a]pyrene diol epoxide-guanine adducts by human nucleotide excision repair enzymes. *Mol Cell Biol* 1997, 17(12):7069-7076.
179. Schiltz M, Cui XX, Lu YP, Yagi H, Jerina DM, Zdzienicka MZ, Chang RL, Conney AH, Wei SJ: Characterization of the mutational profile of (+)-7R,8S-dihydroxy-9S, 10R-epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene at the hypoxanthine (guanine) phosphoribosyltransferase gene in repair-deficient Chinese hamster V-H1 cells. *Carcinogenesis* 1999, 20(12):2279-2286.

180. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P: Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 2002, 21(48):7435-7451.
181. Topkaya SN, Aydinlik S, Aladag N, Ozsoz M, Ozkan-Ariksoysal D: Different DNA immobilization strategies for the interaction of anticancer drug irinotecan with DNA based on electrochemical DNA biosensors. *Comb Chem High Throughput Screen* 2010, 13(7):582-589.
182. Karran P, Offman J, Bignami M: Human mismatch repair, drug-induced DNA damage, and secondary cancer. *Biochimie* 2003, 85(11):1149-1160.
183. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C *et al*: Patterns of somatic mutation in human cancer genomes. *Nature* 2007, 446(7132):153-158.
184. Siddik ZH: Cisplatin: mode of cytotoxic action and molecular basis of resistance. *Oncogene* 2003, 22(47):7265-7279.
185. Malinge JM, Giraud-Panis MJ, Leng M: Interstrand cross-links of cisplatin induce striking distortions in DNA. *J Inorg Biochem* 1999, 77(1-2):23-29.
186. Cadet J, Sage E, Douki T: Ultraviolet radiation-mediated damage to cellular DNA. *Mutat Res* 2005, 571(1-2):3-17.
187. Hendriks G, Calleja F, Besaratinia A, Vrieling H, Pfeifer GP, Mullenders LH, Jansen JG, de Wind N: Transcription-dependent cytosine deamination is a novel mechanism in ultraviolet light-induced mutagenesis. *Curr Biol* 2010, 20(2):170-175.
188. Ellegren H, Smith NG, Webster MT: Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* 2003, 13(6):562-568.
189. Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, Rada C, Stratton MR, Neuberger MS: DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* 2013, 2:e00534.
190. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB *et al*: APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* 2013, 494(7437):366-370.
191. Esteller M, Hamilton SR, Burger PC, Baylin SB, Herman JG: Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer Res* 1999, 59(4):793-797.
192. Li GM: Mechanisms and functions of DNA mismatch repair. *Cell Res* 2008, 18(1):85-98.
193. Croteau DL, Bohr VA: Repair of oxidative damage to nuclear and mitochondrial DNA in mammalian cells. *J Biol Chem* 1997, 272(41):25409-25412.
194. Reardon JT, Sancar A: Nucleotide excision repair. *Prog Nucleic Acid Res Mol Biol* 2005, 79:183-235.
195. Goodman MF: Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu Rev Biochem* 2002, 71:17-50.
196. Long DT, Raschle M, Joukov V, Walter JC: Mechanism of RAD51-dependent DNA interstrand cross-link repair. *Science* 2011, 333(6038):84-87.

197. Izhar L, Ziv O, Cohen IS, Geacintov NE, Livneh Z: Genomic assay reveals tolerance of DNA damage by both translesion DNA synthesis and homology-dependent repair in mammalian cells. *Proc Natl Acad Sci U S A* 2013, 110(16):E1462-1469.
198. Chen JM, Cooper DN, Ferec C, Kehrer-Sawatzki H, Patrinos GP: Genomic rearrangements in inherited disease and cancer. *Semin Cancer Biol* 2010, 20(4):222-233.
199. Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, Stankiewicz P, Gambin A: Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res* 2015, 43(4):2188-2198.
200. Ghezraoui H, Piganeau M, Renouf B, Renaud JB, Sallmyr A, Ruis B, Oh S, Tomkinson AE, Hendrickson EA, Giovannangeli C *et al*: Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. *Mol Cell* 2014, 55(6):829-842.
201. McVey M, Lee SE: MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet* 2008, 24(11):529-538.
202. Truong LN, Li Y, Shi LZ, Hwang PY, He J, Wang H, Razavian N, Berns MW, Wu X: Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proc Natl Acad Sci U S A* 2013, 110(19):7720-7725.
203. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA *et al*: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011, 144(1):27-40.
204. Forment JV, Kaidi A, Jackson SP: Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat Rev Cancer* 2012, 12(10):663-670.
205. Scully R, Chen J, Plug A, Xiao Y, Weaver D, Feunteun J, Ashley T, Livingston DM: Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell* 1997, 88(2):265-275.
206. Scully R, Chen J, Ochs RL, Keegan K, Hoekstra M, Feunteun J, Livingston DM: Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. *Cell* 1997, 90(3):425-435.
207. Greenberg RA, Sobhian B, Pathania S, Cantor SB, Nakatani Y, Livingston DM: Multifactorial contributions to an acute DNA damage response by BRCA1/BARD1-containing complexes. *Genes Dev* 2006, 20(1):34-46.
208. Tutt A, Robson M, Garber JE, Domchek SM, Audeh MW, Weitzel JN, Friedlander M, Arun B, Loman N, Schmutzler RK *et al*: Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. *Lancet* 2010, 376(9737):235-244.
209. Ledermann J, Harter P, Gourley C, Friedlander M, Vergote I, Rustin G, Scott CL, Meier W, Shapira-Frommer R, Safra T *et al*: Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. *Lancet Oncol* 2014, 15(8):852-861.
210. Underhill C, Toulmonde M, Bonnefoi H: A review of PARP inhibitors: from bench to bedside. *Ann Oncol* 2011, 22(2):268-279.

211. Helleday T: The underlying mechanism for the PARP and BRCA synthetic lethality: clearing up the misunderstandings. *Mol Oncol* 2011, 5(4):387-393.
212. Murai J, Huang SY, Das BB, Renaud A, Zhang Y, Doroshow JH, Ji J, Takeda S, Pommier Y: Trapping of PARP1 and PARP2 by Clinical PARP Inhibitors. *Cancer Res* 2012, 72(21):5588-5599.
213. Bryant HE, Petermann E, Schultz N, Jemth AS, Loseva O, Issaeva N, Johansson F, Fernandez S, McGlynn P, Helleday T: PARP is activated at stalled forks to mediate Mre11-dependent replication restart and recombination. *EMBO J* 2009, 28(17):2601-2615.
214. Byrski T, Huzarski T, Dent R, Marczyk E, Jasiowka M, Gronwald J, Jakubowicz J, Cybulski C, Wisniowski R, Godlewski D *et al*: Pathologic complete response to neoadjuvant cisplatin in BRCA1-positive breast cancer patients. *Breast Cancer Res Treat* 2014, 147(2):401-405.
215. Balmana J, Tung NM, Isakoff SJ, Grana B, Ryan PD, Saura C, Lowe ES, Frewer P, Winer E, Baselga J *et al*: Phase I trial of olaparib in combination with cisplatin for the treatment of patients with advanced breast, ovarian and other solid tumors. *Ann Oncol* 2014, 25(8):1656-1663.
216. Jonsson G, Naylor TL, Vallon-Christersson J, Staaf J, Huang J, Ward MR, Greshock JD, Luts L, Olsson H, Rahman N *et al*: Distinct genomic profiles in hereditary breast tumors identified by array-based comparative genomic hybridization. *Cancer Res* 2005, 65(17):7612-7621.
217. Garcia AI, Buisson M, Bertrand P, Rimokh R, Rouleau E, Lopez BS, Lidereau R, Mikaelian I, Mazoyer S: Down-regulation of BRCA1 expression by miR-146a and miR-146b-5p in triple negative sporadic breast cancers. *EMBO Mol Med* 2011, 3(5):279-290.
218. Birgisdottir V, Stefansson OA, Bodvarsdottir SK, Hilmarsdottir H, Jonasson JG, Eyfjord JE: Epigenetic silencing and deletion of the BRCA1 gene in sporadic breast cancer. *Breast Cancer Res* 2006, 8(4):R38.
219. Watkins J, Weekes D, Shah V, Gazinska P, Joshi S, Sidhu B, Gillett C, Pinder S, Vanoli F, Jasin M *et al*: Genomic complexity profiling reveals that HORMAD1 overexpression contributes to homologous recombination deficiency in triple-negative breast cancers. *Cancer Discov* 2015.
220. Ceccaldi R, Liu JC, Amunugama R, Hajdu I, Primack B, Petalcorin MI, O'Connor KW, Konstantinopoulos PA, Elledge SJ, Boulton SJ *et al*: Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair. *Nature* 2015, 518(7538):258-262.
221. Mateos-Gomez PA, Gong F, Nair N, Miller KM, Lazzerini-Denchi E, Sfeir A: Mammalian polymerase theta promotes alternative NHEJ and suppresses recombination. *Nature* 2015, 518(7538):254-257.
222. Mandel P, Metais P: [Not Available]. *C R Seances Soc Biol Fil* 1948, 142(3-4):241-243.
223. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ: Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res* 1977, 37(3):646-650.

224. Sorenson GD, Pribish DM, Valone FH, Memoli VA, Bzik DJ, Yao SL: Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol Biomarkers Prev* 1994, 3(1):67-71.
225. Vasioukhin V, Anker P, Maurice P, Lyautey J, Lederrey C, Stroun M: Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br J Haematol* 1994, 86(4):774-779.
226. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, Wainscoat JS: Presence of fetal DNA in maternal plasma and serum. *Lancet* 1997, 350(9076):485-487.
227. Peters D, Chu T, Yatsenko SA, Hendrix N, Hogge WA, Surti U, Bunce K, Dunkel M, Shaw P, Rajkovic A: Noninvasive prenatal diagnosis of a fetal microdeletion syndrome. *N Engl J Med* 2011, 365(19):1847-1848.
228. Sehnert AJ, Rhees B, Comstock D, de Feo E, Heilek G, Burke J, Rava RP: Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood. *Clin Chem* 2011, 57(7):1042-1049.
229. Lo YM, Tein MS, Lau TK, Haines CJ, Leung TN, Poon PM, Wainscoat JS, Johnson PJ, Chang AM, Hjelm NM: Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am J Hum Genet* 1998, 62(4):768-775.
230. Lo YM, Zhang J, Leung TN, Lau TK, Chang AM, Hjelm NM: Rapid clearance of fetal DNA from maternal plasma. *Am J Hum Genet* 1999, 64(1):218-224.
231. Yu SC, Lee SW, Jiang P, Leung TY, Chan KC, Chiu RW, Lo YM: High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing. *Clin Chem* 2013, 59(8):1228-1237.
232. Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, Zheng YW, Leung TY, Lau TK, Cantor CR *et al*: Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* 2010, 2(61):61ra91.
233. Li Y, Zimmermann B, Rusterholz C, Kang A, Holzgreve W, Hahn S: Size separation of circulatory DNA in maternal plasma permits ready detection of fetal DNA polymorphisms. *Clin Chem* 2004, 50(6):1002-1011.
234. van der Vaart M, Pretorius PJ: The origin of circulating free DNA. *Clin Chem* 2007, 53(12):2215.
235. Thakur BK, Zhang H, Becker A, Matei I, Huang Y, Costa-Silva B, Zheng Y, Hoshino A, Brazier H, Xiang J *et al*: Double-stranded DNA in exosomes: a novel biomarker in cancer detection. *Cell Res* 2014, 24(6):766-769.
236. Balaj L, Lessard R, Dai L, Cho YJ, Pomeroy SL, Breakefield XO, Skog J: Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. *Nat Commun* 2011, 2:180.
237. Wyllie AH: Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation. *Nature* 1980, 284(5756):555-556.
238. Nakano H, Shinohara K: X-ray-induced cell death: apoptosis and necrosis. *Radiat Res* 1994, 140(1):1-9.
239. Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, Wong GL, Chan SL, Mok TS, Chan HL *et al*: Lengthening and shortening of plasma DNA in

hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* 2015, 112(11):E1317-1325.
240. Tornillo L, Carafa V, Sauter G, Moch H, Minola E, Gambacorta M, Vecchione R, Bianchi L, Terracciano LM: Chromosomal alterations in hepatocellular nodules by comparative genomic hybridization: high-grade dysplastic nodules represent early stages of hepatocellular carcinoma. *Lab Invest* 2002, 82(5):547-553.
241. Madhavan D, Wallwiener M, Bents K, Zucknick M, Nees J, Schott S, Cuk K, Riethdorf S, Trumpp A, Pantel K *et al*: Plasma DNA integrity as a biomarker for primary and metastatic breast cancer and potential marker for early diagnosis. *Breast Cancer Res Treat* 2014, 146(1):163-174.
242. Heidary M, Auer M, Ulz P, Heitzer E, Petru E, Gasch C, Riethdorf S, Mauermann O, Lafer I, Pristauz G *et al*: The dynamic range of circulating tumor DNA in metastatic breast cancer. *Breast Cancer Res* 2014, 16(4):421.
243. Klevebring D, Neiman M, Sundling S, Eriksson L, Darai Ramqvist E, Celebioglu F, Czene K, Hall P, Egevad L, Gronberg H *et al*: Evaluation of exome sequencing to estimate tumor burden in plasma. *PLoS One* 2014, 9(8):e104417.
244. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, Knippers R: DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res* 2001, 61(4):1659-1665.
245. Schwarzenbach H, Eichelser C, Kropidlowski J, Janni W, Rack B, Pantel K: Loss of heterozygosity at tumor suppressor genes detectable on fractionated circulating cell-free tumor DNA as indicator of breast cancer progression. *Clin Cancer Res* 2012, 18(20):5719-5730.
246. Giacona MB, Ruben GC, Iczkowski KA, Roos TB, Porter DM, Sorenson GD: Cell-free DNA in human blood plasma: length measurements in patients with pancreatic cancer and healthy controls. *Pancreas* 1998, 17(1):89-97.
247. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM *et al*: Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* 2014, 6(224):224ra224.
248. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, Thornton K, Agrawal N, Sokoll L, Szabo SA *et al*: Circulating mutant DNA to assess tumor dynamics. *Nat Med* 2008, 14(9):985-990.
249. Diaz LA, Jr., Bardelli A: Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol* 2014, 32(6):579-586.
250. Naume B, Synnestvedt M, Falk RS, Wiedswang G, Weyde K, Risberg T, Kersten C, Mjaaland I, Vindi L, Sommer HH *et al*: Clinical outcome with correlation to disseminated tumor cell (DTC) status after DTC-guided secondary adjuvant treatment with docetaxel in early breast cancer. *J Clin Oncol* 2014, 32(34):3848-3857.
251. Domschke C, Diel IJ, Englert S, Kalteisen S, Mayer L, Rom J, Heil J, Sohn C, Schuetz F: Prognostic value of disseminated tumor cells in the bone marrow of patients with operable primary breast cancer: a long-term follow-up study. *Ann Surg Oncol* 2013, 20(6):1865-1871.
252. Fehm T, Braun S, Muller V, Janni W, Gebauer G, Marth C, Schindlbeck C, Wallwiener D, Borgen E, Naume B *et al*: A concept for the standardized detection of

disseminated tumor cells in bone marrow from patients with primary breast cancer and its clinical implementation. *Cancer* 2006, 107(5):885-892.
253. Joosse SA, Gorges TM, Pantel K: Biology, detection, and clinical implications of circulating tumor cells. *EMBO Mol Med* 2015, 7(1):1-11.
254. Zhang L, Riethdorf S, Wu G, Wang T, Yang K, Peng G, Liu J, Pantel K: Meta-analysis of the prognostic value of circulating tumor cells in breast cancer. *Clin Cancer Res* 2012, 18(20):5701-5710.
255. Bidard FC, Peeters DJ, Fehm T, Nole F, Gisbert-Criado R, Mavroudis D, Grisanti S, Generali D, Garcia-Saenz JA, Stebbing J *et al*: Clinical validity of circulating tumour cells in patients with metastatic breast cancer: a pooled analysis of individual patient data. *Lancet Oncol* 2014, 15(4):406-414.
256. Raimondi C, Gradilone A, Naso G, Cortesi E, Gazzaniga P: Clinical utility of circulating tumor cell counting through CellSearch((R)): the dilemma of a concept suspended in Limbo. *Onco Targets Ther* 2014, 7:619-625.
257. Lustberg MB, Balasubramanian P, Miller B, Garcia-Villa A, Deighan C, Wu Y, Carothers S, Berger M, Ramaswamy B, Macrae ER *et al*: Heterogeneous atypical cell populations are present in blood of metastatic breast cancer patients. *Breast Cancer Res* 2014, 16(2):R23.
258. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE *et al*: An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* 2014, 20(5):548-554.
259. Murtaza M, Dawson SJ, Tsui DW, Gale D, Forshew T, Piskorz AM, Parkinson C, Chin SF, Kingsbury Z, Wong AS *et al*: Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* 2013, 497(7447):108-112.
260. Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, Duncan C, Antipova A, Lee C, McKernan K, De La Vega FM *et al*: Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med* 2010, 2(20):20ra14.
261. Diaz LA, Jr., Williams RT, Wu J, Kinde I, Hecht JR, Berlin J, Allen B, Bozic I, Reiter JG, Nowak MA *et al*: The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 2012, 486(7404):537-540.
262. Agassi R, Czeiger D, Shaked G, Avriel A, Sheynin J, Lavrenkov K, Ariad S, Douvdevani A: Measurement of circulating cell-free DNA levels by a simple fluorescent test in patients with breast cancer. *Am J Clin Pathol* 2015, 143(1):18-24.
263. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J *et al*: The genomic landscapes of human breast and colorectal cancers. *Science* 2007, 318(5853):1108-1113.
264. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR *et al*: The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012, 486(7403):400-404.
265. McBride DJ, Orpana AK, Sotiriou C, Joensuu H, Stephens PJ, Mudie LJ, Hamalainen E, Stebbings LA, Andersson LC, Flanagan AM *et al*: Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosomes Cancer* 2010, 49(11):1062-1069.
266. De Mattos-Arruda L, Weigelt B, Cortes J, Won HH, Ng CK, Nuciforo P, Bidard FC, Aura C, Saura C, Peg V *et al*: Capturing intra-tumor genetic heterogeneity by de

novo mutation profiling of circulating cell-free tumor DNA: a proof-of-principle. *Ann Oncol* 2014, 25(9):1729-1735.
267. Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, Dunning MJ, Gale D, Forshew T, Mahler-Araujo B *et al*: Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 2013, 368(13):1199-1209.
268. Reinert T, Scholer LV, Thomsen R, Tobiasen H, Vang S, Nordentoft I, Lamy P, Kannerup AS, Mortensen FV, Stribolt K *et al*: Analysis of circulating tumour DNA to monitor disease burden following colorectal cancer surgery. *Gut* 2015.
269. Sambrook J. RDW: Molecular cloning, a laboratory manual, Third Edition. 2001.
270. Heid CA, Stevens J, Livak KJ, Williams PM: Real time quantitative PCR. *Genome Res* 1996, 6(10):986-994.
271. Livak KJ, Schmittgen TD: Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001, 25(4):402-408.
272. Pohl G, Shih Ie M: Principle and applications of digital PCR. *Expert Rev Mol Diagn* 2004, 4(1):41-47.
273. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC *et al*: High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem* 2011, 83(22):8604-8610.
274. Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS: 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res* 1991, 19(14):4008.
275. Kutyavin IV, Afonina IA, Mills A, Gorn VV, Lukhtanov EA, Belousov ES, Singer MJ, Walburger DK, Lokhov SG, Gall AA *et al*: 3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res* 2000, 28(2):655-661.
276. Moreira BG, You Y, Behlke MA, Owczarzy R: Effects of fluorescent dyes, quenchers, and dangling ends on DNA duplex stability. *Biochem Biophys Res Commun* 2005, 327(2):473-484.
277. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977, 74(12):5463-5467.
278. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE: Fluorescence detection in automated DNA sequence analysis. *Nature* 1986, 321(6071):674-679.
279. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K: A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 1987, 238(4825):336-341.
280. Huang XC, Quesada MA, Mathies RA: DNA sequencing using capillary array electrophoresis. *Anal Chem* 1992, 64(18):2149-2154.
281. International Human Genome Sequencing C: Finishing the euchromatic sequence of the human genome. *Nature* 2004, 431(7011):931-945.
282. Metzker ML: Sequencing technologies - the next generation. *Nat Rev Genet* 2010, 11(1):31-46.

283. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012, 30(5):434-439.
284. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M: Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011, 29(10):908-914.
285. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010, 7(2):111-118.
286. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009, 25(14):1754-1760.
287. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010, 20(9):1297-1303.
288. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010, 38(16):e164.
289. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL: CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012, 28(10):1307-1313.
290. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP *et al*: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* 2009, 6(9):677-681.
291. Diaz LK, Zhou X, Wright ET, Cristofanilli M, Smith T, Yang Y, Sneige N, Sahin A, Gilcrease MZ: CD44 expression is associated with increased survival in node-negative invasive breast carcinoma. *Clin Cancer Res* 2005, 11(9):3309-3314.
292. Kao J, Salari K, Bocanegra M, Choi YL, Girard L, Gandhi J, Kwei KA, Hernandez-Boussard T, Wang P, Gazdar AF *et al*: Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One* 2009, 4(7):e6146.
293. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012, 22(3):568-576.
294. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013, 31(3):213-219.
295. Davies C, Pan H, Godwin J, Gray R, Arriagada R, Raina V, Abraham M, Medeiros Alencar VH, Badran A, Bonfill X *et al*: Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. *Lancet* 2013, 381(9869):805-816.
296. Eccles SA, Aboagye EO, Ali S, Anderson AS, Armes J, Berditchevski F, Blaydes JP, Brennan K, Brown NJ, Bryant HE *et al*: Critical research gaps and translational

priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Res* 2013, 15(5):R92.
297. Saphner T, Tormey DC, Gray R: Annual hazard rates of recurrence for breast cancer after primary therapy. *J Clin Oncol* 1996, 14(10):2738-2746.